BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN THỊ THANH THỦY

NGHIÊN CỨU CÁC PHƯƠNG PHÁP HỌC MÁY CHO TRÍCH XUẤT THÔNG TIN TỰ ĐỘNG TỪ VĂN BẢN

LUẬN ÁN TIẾN SĨ KỸ THUẬT

HÀ NỘI – 2023

BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN THỊ THANH THỦY

NGHIÊN CỨU CÁC PHƯƠNG PHÁP HỌC MÁY CHO TRÍCH XUẤT THÔNG TIN TỰ ĐỘNG TỪ VĂN BẢN

CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN
MÃ SỐ: 9.48.01.04

LUẬN ÁN TIẾN SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC:
1. GS.TS. TỪ MINH PHƯƠNG
2. PGS.TS. NGÔ XUÂN BÁCH

HÀ NỘI – 2023

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi, dưới sự hướng dẫn của GS.TS. Từ Minh Phương và PGS.TS. Ngô Xuân Bách. Các kết quả được viết chung với các tác giả khác đều được sự đồng ý của đồng tác giả trước khi đưa vào luận án. Các kết quả nêu trong luận án là trung thực và chưa từng được công bố trong các công trình nào khác.

Hà Nội, ngày .... tháng .... năm 2023

Nghiên cứu sinh

Nguyễn Thị Thanh Thủy

LỜI CẢM ƠN

Trong quá trình học tập, nghiên cứu để hoàn thành đề tài luận án "Nghiên cứu các phương pháp học máy cho trích xuất thông tin tự động từ văn bản", ngoài sự nỗ lực của cá nhân, tôi đã nhận được rất nhiều sự giúp đỡ, tạo điều kiện của các Thầy hướng dẫn, nhà trường, các nhà khoa học, đơn vị công tác và gia đình. Tôi xin bày tỏ lòng biết ơn chân thành về sự giúp đỡ đó.

Đầu tiên, tôi xin bày tỏ lòng biết ơn sâu sắc tới Thầy GS.TS. Từ Minh Phương và Thầy PGS.TS. Ngô Xuân Bách đã tận tình hướng dẫn, chỉ bảo, giúp đỡ và đồng hành cùng tôi trong suốt quá trình thực hiện nghiên cứu và hoàn thành luận án.

Tôi xin trân trọng cảm ơn Lab Học máy và Ứng dụng, Khoa Quốc tế và Đào tạo Sau Đại học và Lãnh đạo Học viện Công nghệ Bưu chính Viễn thông đã tạo điều kiện thuận lợi cho tôi trong suốt quá trình thực hiện luận án. Tôi xin cảm ơn các Thầy Lãnh đạo và tập thể cán bộ, giảng viên Khoa Công nghệ thông tin 1, Học viện Công nghệ Bưu chính Viễn thông đã luôn cổ vũ, động viên tôi trong quá trình nghiên cứu.

Tôi xin trân trọng cảm ơn Quỹ Đổi mới sáng tạo Vingroup (VINIF), Viện nghiên cứu VINBIGDATA, Tập đoàn Vingroup đã trao học bổng học tập cho tôi trong thời gian tôi làm nghiên cứu luận án.

Tôi xin gửi lời cảm ơn chân thành tới tất cả những người bạn luôn chia sẻ và động viên tôi trong những lúc khó khăn. Cuối cùng, tôi xin bày tỏ lòng biết ơn đối với gia đình đã luôn bên cạnh ủng hộ, động viên, tạo mọi điều kiện hỗ trợ tôi.

Hà Nội, ngày .... tháng .... năm 2023

Nghiên cứu sinh

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ

DANH MỤC CÁC TỪ VIẾT TẮT

TỪ VIẾT TẮT	DIỄN GIẢI	
	TIẾNG ANH	TIẾNG VIỆT
BERT	Bidirectional Encoder Representations from Transformers	Biểu diễn thể hiện mã hóa hai chiều từ Transformer
BiLSTM	Bidirectional long short-term memory	Mô hình mạng bộ nhớ dài ngắn hai chiều
CNN	Convolutional neural network	Mạng nơ-ron tích chập
CRF	Conditional random field	Trường ngẫu nhiên có điều điều kiện
CRL	Cross language	
FN	False negative	Âm tính giả (mẫu mang nhãn dương được phân lớp vào lớp âm)
FNR	False negative rate	Tỉ lệ âm tính giả
FP	False positive	Dương tính giả (mẫu mang nhãn âm được phân lớp vào lớp dương)
FPR	False positive rate	Tỉ lệ dương tính giả
HMM	Hidden Markov model	Mô hình Markov ẩn
IE	Information Extraction	Trích xuất thông tin
LSTM	Long short-term memory	Mô hình mạng bộ nhớ dài ngắn
MEMM	Maximum Entropy Markov model	Mô hình Markov entropy cực đại

MLP	Multilayer perceptron	Mô hình Perceptron nhiều lớp
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
NN	Neural Network	Mạng nơ-ron
PhoBERT	Pho Bidirectional Encoder Representations from Transformers	Biểu diễn thể hiện mã hóa hai chiều từ Transformer cho tiếng Việt.
RNN	Recurrent Neural Networks	Mạng nơ-ron hồi quy
SPN	Set Prediction Networks	Mạng dự đoán theo tập hợp
SVM	Support Vector Machines	Máy véc-tơ tựa
TF-IDF	Term Frequency - Inverse Document Frequency	Tần số xuất hiện của một từ trong một văn bản - Tần số nghịch đảo của một từ trong tập văn bản
TN	True Negative	Âm tính thật (mẫu mang nhãn âm được phân lớp đúng vào lớp âm)
TP	True Positive	Dương tính thật (mẫu mang nhãn dương được phân lớp đúng vào lớp dương)
VLSP	Vietnamese Language and Speech Processing	Hội nghị thường niên về Xử lý ngôn ngữ tự nhiên và tiếng nói cho tiếng Việt

PHẦN MỞ ĐẦU

1. TÍNH CẤP THIẾT CỦA LUẬN ÁN

Ngày nay, dữ liệu được coi là một nguồn tài nguyên vô cùng quan trọng với sự gia tăng nhanh chóng theo thời gian. Một phần rất lớn dữ liệu thường được trình bày dưới các dạng văn bản, tài liệu không có cấu trúc hoặc bán cấu trúc và hoàn toàn miễn phí. Tuy nhiên, việc tìm kiếm và trích chọn ra được các thông tin người dùng cần từ những nguồn dữ liệu này là điều không dễ dàng. Việc này đã thúc đẩy những nghiên cứu về các phương pháp, kỹ thuật nhằm phân tích dữ liệu và trích xuất thông tin từ văn bản một cách hiệu quả.

Trích xuất thông tin *(Information Extraction)* thực hiện trích xuất tự động những thông tin có cấu trúc như các thực thể, các ý kiến/quan điểm mô tả thực thể, mối quan hệ giữa các thực thể, hay các sự kiện từ các nguồn dữ liệu không có cấu trúc hoặc bán cấu trúc. Mục tiêu cuối cùng là chuyển thông tin trong văn bản sang một hình thức dễ tiếp cận (/truy xuất) hơn để có thể tiếp tục xử lý, nhằm hỗ trợ tốt hơn cho người dùng.

Hiện tại trên thực tế có khá nhiều ứng dụng của trích xuất thông tin, từ các ứng dụng quản lý thông tin cá nhân, tới các ứng dụng trong doanh nghiệp (như theo dõi tin tức, chăm sóc khách hàng, làm sạch dữ liệu), đến các ứng dụng trong các lĩnh vực khoa học (ví dụ, tin sinh học), và đặc biệt là sự phát triển mạnh mẽ của các ứng dụng hướng web (như cơ sở dữ liệu trích dẫn, cơ sở dữ liệu ý kiến/quan điểm, các trang web cộng đồng, so sánh khi mua sắm) [40,101].

Có hai nhóm phương pháp tiếp cận chính được sử dụng để giải quyết các nhiệm vụ trích xuất thông tin là các phương pháp dựa trên luật *(rule-based)* và các phương pháp dựa trên học máy *(learning-based)*. Các phương pháp dựa trên luật đòi hỏi người thực hiện phải là các chuyên gia có kiến thức sâu về các miền lĩnh vực và ngôn ngữ để có thể phát triển các luật trích xuất hiệu quả. Phương pháp này không

có khả năng tự động cập nhật các luật do nguồn dữ liệu đầu vào thường ở dạng không có cấu trúc và thường xuyên thay đổi, ngoài ra phương pháp cũng không có khả năng xử lý những thông tin tạm thời và không tường minh. Các phương pháp dựa trên học máy được thực hiện bằng cách sử dụng các mẫu không có cấu trúc được gán nhãn/chú thích thủ công để huấn luyện mô hình học máy cho việc trích xuất thông tin. Phương pháp này cũng cần có kiến thức chuyên gia về các miền lĩnh vực để xác định và gán nhãn cho các mẫu đại diện, đồng thời cần có kiến thức về học máy để có thể lựa chọn giữa các mô hình khác nhau, cũng như xác định được các đặc trưng tốt trong nguồn dữ liệu. Ưu điểm lớn của phương pháp này là đảm bảo được việc cập nhật tự động các luật mà không phụ thuộc vào chuyên gia, hơn nữa phương pháp có khả năng thích nghi cao và tận dụng được các nguồn dữ liệu có sẵn. *Do vậy, nội dung luận án định hướng nghiên cứu các phương pháp học máy để giải quyết một số nhiệm vụ trong trích xuất thông tin tự động từ văn bản.*

Khảo sát các phương pháp tiếp cận dựa trên học máy để giải quyết các nhiệm vụ trong lĩnh vực trích xuất thông tin từ văn bản, chúng tôi nhận thấy có một số vấn đề còn tồn tại như sau:

1) Các phương pháp học máy đã được chứng minh là có hiệu quả trong nhiều nghiên cứu về trích xuất thông tin trước đây [1,30,52,94,122,123], nhưng các phương pháp này thường yêu cầu cần phải chú thích (gán nhãn) thủ công một lượng lớn dữ liệu cho giai đoạn huấn luyện, việc này rất tốn kém thời gian và chi phí. Ngoài ra, các phương pháp học máy thường phụ thuộc vào miền lĩnh vực. Do đó, việc áp dụng các phương pháp học máy vào một miền lĩnh vực mới hoặc một ngôn ngữ mới, đặc biệt là ngôn ngữ có ít tài nguyên dữ liệu đã được gán nhãn sẵn (như tiếng Việt) trong một số bài toán trích xuất thông tin, vẫn còn rất nhiều khó khăn. Ví dụ, tập dữ liệu cho xử lý các nhiệm vụ trích xuất thông tin trong bài toán khai phá quan điểm dựa trên khía cạnh cho tiếng Việt (chú thích ở mức câu văn bản) hiện nay chưa thấy có công bố nào (theo khảo sát của chúng tôi). Vậy liệu có thể sử dụng dữ liệu từ ngôn ngữ này để bổ sung vào cho dữ liệu của ngôn ngữ khác được không?

2) Các mô hình học máy truyền thống thường cần sử dụng các phương pháp, kỹ thuật khác nhau để chọn ra được tập các đặc trưng tốt cho các mô hình học, được gọi là kỹ thuật trích chọn đặc trưng (*feature engineering*). Các phương pháp này thường được thực hiện theo cách thủ công, do vậy cũng rất tốn kém thời gian và công sức, đồng thời cần có kiến thức chuyên gia về miền lĩnh vực nghiên cứu. Hơn nữa, trong nhiều trường hợp, tập đặc trưng thu được vẫn có thể không được đầy đủ (còn thiếu đặc trưng quan trọng cho bài toán), các đặc trưng rời rạc (không có mối liên hệ với nhau), và có thể xuất hiện lỗi trong quá trình chọn và trích xuất đặc trưng. Những vấn đề này dẫn đến giảm hiệu quả của các hệ thống trích xuất thông tin. Vậy có thể sử dụng phương pháp nào để hỗ trợ trích chọn đặc trưng tự động và giúp tăng hiệu quả cho trích xuất thông tin?

Đề tài "*Nghiên cứu các phương pháp học máy cho trích xuất thông tin tự động từ văn bản"* được thực hiện trong khuôn khổ luận án tiến sĩ chuyên ngành Hệ thống thông tin nhằm góp phần giải quyết một số vấn đề còn tồn tại khi sử dụng các phương pháp học máy để giải quyết các nhiệm vụ trích xuất thông tin từ văn bản.

2. MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU LUẬN ÁN

Mục tiêu của luận án là nghiên cứu và đề xuất một số phương pháp học máy nhằm giải quyết và nâng cao hiệu quả cho trích xuất thông tin tự động từ văn bản, bao gồm hai mục tiêu cụ thể như sau:

1) Để giải quyết vấn đề thứ nhất (nêu trên): tiết kiệm thời gian và công sức gán nhãn thủ công trong quá trình xây dựng tập dữ liệu huấn luyện các mô hình trích xuất thông tin cho các ngôn ngữ ít tài nguyên (như tiếng Việt), mục tiêu của luận án là nghiên cứu đề xuất giải pháp giải quyết một số nhiệm vụ trích xuất thông tin bằng cách khai thác dữ liệu đã được gán nhãn sẵn từ các ngôn ngữ khác (ví dụ tiếng Anh, giàu tài nguyên hơn) để bổ sung dữ liệu cho tập huấn luyện. Giải pháp này khá tổng quát và linh hoạt do không phụ thuộc vào

ngôn ngữ và các thuật toán học máy, giảm thời gian và chi phí gán nhãn dữ liệu thủ công, đồng thời giúp nâng cao hiệu quả trích xuất.

2) Để giải quyết vấn đề thứ hai (nêu trên): khó khăn trong việc trích chọn đặc trưng thủ công trong các phương pháp học máy truyền thống, mục tiêu của luận án là nghiên cứu đề xuất giải quyết một số nhiệm vụ trích xuất thông tin với các phương pháp tiên tiến dựa trên học sâu, là các phương pháp được đánh giá có độ chính xác cao và được ứng dụng hiệu quả trong rất nhiều lĩnh vực khác nhau. Các phương pháp học sâu có ưu điểm là có khả năng tự động tạo ra các biểu diễn đặc trưng hiệu quả từ dữ liệu, do vậy giảm được thời gian và công sức trong việc trích chọn đặc trưng thủ công. Trong xử lý ngôn ngữ tự nhiên nói chung, các phương pháp này sẽ tạo ra những biểu diễn chung cho các từ trong tập văn bản, từ đó có thể giúp nắm bắt được những đặc trưng về ngữ nghĩa cũng như các ràng buộc về cú pháp trong các câu văn bản.

   Ngoài ra, luận án cũng tập trung nghiên cứu và đề xuất các phương pháp kết hợp ưu điểm giữa các phương pháp học máy truyền thống với các phương pháp học sâu nhằm cải thiện hiệu quả hơn nữa cho các nhiệm vụ trích xuất thông tin.

Với các mục tiêu này, phạm vi nghiên cứu luận án tập trung vào hai nội dung cụ thể như sau:

1) Nghiên cứu đề xuất phương pháp trích xuất thông tin cho ngôn ngữ tiếng Việt bằng cách khai thác nguồn dữ liệu đã được gán nhãn từ ngôn ngữ khác trong bài toán khai phá quan điểm dựa trên khía cạnh tiếng Việt. Các thông tin được trích xuất bao gồm khía cạnh và ý kiến/quan điểm về khía cạnh. Đây là một bài toán rất có ý nghĩa trong thực tế và mang tính ứng dụng cao, do có thể cung cấp thông tin về ý kiến/quan điểm chi tiết đến từng khía cạnh cụ thể của sản phẩm/dịch vụ được đề cập trong câu (thay vì chỉ xác định một ý kiến/quan điểm tổng thể chung cho toàn bộ văn bản đầu vào). Các thông tin khía cạnh và quan điểm về khía cạnh được trích xuất đều rất quan trọng với các đối tượng

người dùng là khách hàng, người bán hàng và nhà cung cấp dịch vụ/sản phẩm: giúp khách hàng lựa chọn được sản phẩm/dịch vụ tốt, phù hợp với các đặc điểm cụ thể khách hàng mong muốn; giúp người bán hàng và nhà cung cấp dịch vụ/sản phẩm nắm được thị hiếu của khách hàng, xu hướng thị trường; cũng từ đó, giúp nhà cung cấp dịch vụ/sản phẩm định hướng thiết kế, phát triển các dòng sản phẩm/dịch vụ tiếp theo.

Nghiên cứu thực hiện trích xuất thông tin với hai nhiệm vụ cụ thể: (1) trích xuất các loại khía cạnh *(aspect category)* và (2) phân loại ý kiến/quan điểm *(sentiment classification)* cho khía cạnh đã được trích xuất. Trích xuất các loại khía cạnh thực hiện xác định các loại khía cạnh (bao gồm thực thể và thuộc tính), mà có một ý kiến được thể hiện trong văn bản. Phân loại quan điểm nhằm xác định ý kiến/quan điểm (tích cực, tiêu cực hay trung tính) cho từng loại khía cạnh đã được xác định trong nhiệm vụ trước. Ví dụ có một nhận xét về nhà hàng như sau: "*Nhân viên rất thân thiện, nhưng đồ ăn không ngon.*", thì đầu ra mong muốn của hệ thống khai phá quan điểm dựa trên khía cạnh bao gồm: hai khía cạnh được xác định là *dịch vụ* và *chất lượng thực phẩm* của nhà hàng, và phân loại quan điểm đối với hai khía cạnh là *tích cực* đối với *dịch vụ*, và *tiêu cực* đối với *chất lượng thực phẩm*.

2) Nghiên cứu đề xuất phương pháp dựa trên học sâu để giải quyết và nâng cao hiệu quả cho một số nhiệm vụ trích xuất thông tin trong lĩnh vực xử lý văn bản pháp quy tiếng Việt. Văn bản pháp quy (hay còn gọi là văn bản quy phạm pháp luật) như hiến pháp, luật, nghị định, thông tư là những văn bản do cơ quan Nhà nước ban hành để điều tiết hoạt động của Nhà nước và xã hội. Với số lượng văn bản pháp quy lớn, được gia tăng và cập nhật theo thời gian, việc tiếp cận và chọn lọc thông tin từ hệ thống văn bản pháp quy là một việc rất khó khăn với những người bình thường không có chuyên môn về pháp luật, và thậm chí cả những người có chuyên môn như các chuyên gia về luật, luật sư. Do vậy, nhu cầu thực tế là cần phải có các công cụ/hệ thống xử lý văn bản pháp quy tự động, như tìm kiếm, tra cứu, phân tích, truy vấn (hỏi/đáp) nhằm hỗ trợ tốt hơn

cho người dùng. Trích xuất thông tin trong văn bản pháp quy là bước quan trọng đầu tiên để có thể xây dựng các công cụ/hệ thống xử lý văn bản này.

Nghiên cứu thực hiện trích xuất thông tin trong văn bản pháp quy tiếng Việt, với 2 nhiệm vụ cụ thể: (1) trích xuất thực thể tham chiếu từ văn bản pháp quy, và (2) phân loại quan hệ giữa các thực thể văn bản pháp quy. Trích xuất thực thể tham chiếu từ văn bản pháp quy là việc trích xuất ra được các tham chiếu là tên của văn bản được đề cập/nhắc đến trong văn bản pháp quy đang xem xét (đang đọc). Phân loại quan hệ giữa các thực thể văn bản pháp quy là việc phân loại mối liên quan giữa thực thể là văn bản tham chiếu (văn bản được đề cập) và thực thể là văn bản đang xem xét. Ví dụ, xem xét văn bản "Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004 của Bộ Tài chính", có đoạn như sau: "*Căn cứ Nghị định số 60/2003/NĐ-CP ngày 6/6/2003 của Chính phủ quy định chi tiết và hướng dẫn thi hành…*". Có hai thực thể được xác định là: văn bản đang xem xét "*Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004*", và văn bản được đề cập đến trong nội dung của văn bản đang xem xét "*Nghị định số 60/2003/NĐ-CP ngày 6/6/2003*". Ngữ nghĩa ở đây là, văn bản "*Nghị định số 60/2003/NĐ-CP ngày 6/6/2003*" có quan hệ **"*căn cứ*"** với văn bản "*Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004*".

3. CÁC ĐÓNG GÓP CỦA LUẬN ÁN

Đóng góp thứ nhất của luận án là *đề xuất giải pháp nâng cao hiệu quả cho trích xuất khía cạnh và phân loại quan điểm trong ngôn ngữ tiếng Việt bằng cách khai thác nguồn dữ liệu đã được gán nhãn sẵn từ ngôn ngữ khác* [4, 6] (Theo danh mục các công trình công bố). Phương pháp đề xuất khá tổng quát và linh hoạt do không phụ thuộc vào ngôn ngữ và các thuật toán học máy. Việc xác định các loại khía cạnh và phân loại quan điểm được thực hiện theo từng câu thay vì toàn bộ bài đánh giá, sẽ thực tế hơn và có thể áp dụng trong các ứng dụng. Để chứng minh tính hiệu quả của phương pháp đề xuất, chúng tôi đã xây dựng một tập dữ liệu văn bản tiếng Việt có chú thích về các loại khía cạnh và quan điểm được trích từ các bài đánh

giá về lĩnh vực nhà hàng bằng ngôn ngữ tiếng Việt, bao gồm 575 bài đánh giá với 3.796 câu, và tiến hành các thực nghiệm. Kết quả thực nghiệm cho thấy với việc sử dụng thêm dữ liệu (đã được gán nhãn sẵn) dịch từ tiếng Anh, phương pháp đề xuất đã cải thiện hiệu năng của cả hai nhiệm vụ trích xuất khía cạnh và phân loại quan điểm.

Đóng góp thứ hai của luận án là *đề xuất phương pháp trích xuất thông tin sử dụng học máy truyền thống và học sâu cho văn bản pháp quy tiếng Việt. Các thông tin được trích xuất bao gồm thực thể tham chiếu và mối quan hệ giữa các thực thể văn bản pháp quy* [1, 5] (Theo danh mục các công trình công bố). Với nhiệm vụ *trích xuất thực thể tham chiếu*, nghiên cứu sử dụng phương pháp kết hợp lợi thế của mô hình học sâu và các đặc trưng được thiết kế thủ công (theo phương pháp học máy truyền thống). Mô hình trích xuất bao gồm một số lớp LSTM hai chiều (BiLSTM) tạo ra biểu diễn câu từ các từ, ký tự và các đặc trưng nhúng thủ công, và một trường ngẫu nhiên có điều kiện (CRF) ở lớp suy diễn. Với nhiệm vụ *phân loại quan hệ* giữa các thực thể tham chiếu (đã được trích xuất ở trên) với thực thể là văn bản pháp quy đang xem xét, ngoài việc sử dụng phương pháp học máy truyền thống, nghiên cứu sử dụng mô hình học sâu bao gồm một số lớp LSTM hai chiều (BiLSTM) để học cách biểu diễn từ, biểu diễn câu và một lớp softmax để suy diễn. Để chứng minh tính hiệu quả của các phương phương pháp đề xuất, chúng tôi đã xây dựng một tập dữ liệu gồm 5.031 văn bản pháp quy tiếng Việt được gán nhãn thực thể tham chiếu và quan hệ giữa các thực thể, và tiến hành các thực nghiệm. Kết quả thực nghiệm cho thấy phương pháp đề xuất cho kết quả khả quan với cả hai nhiệm vụ trích xuất thực thể tham chiếu và phân loại quan hệ, với độ đo F_1 đều đạt trên 95%.

Đóng góp thứ ba của luận án là *đề xuất phương pháp trích xuất kết hợp thực thể và quan hệ trong văn bản pháp quy tiếng Việt sử dụng mô hình dựa trên học sâu* [2, 3] (Theo danh mục các công trình công bố). Phương pháp đề xuất thực hiện trích xuất đồng thời cả hai thông tin thực thể tham chiếu và quan hệ, khác với đóng góp thứ hai (nêu trên) thực hiện trích xuất các thông tin này theo cách tuần tự. Mô hình trích xuất kết hợp sử dụng kiến trúc bộ mã hóa-giải mã dựa trên Transformer với cơ

chế giải mã song song không tự hồi quy *(non-autoregressive decoding mechanism)* để trích xuất đồng thời các thực thể tham chiếu và quan hệ trong văn bản pháp quy. Nhằm cải thiện hiệu quả của mô hình trích xuất kết hợp, nghiên cứu sử dụng phương pháp tăng cường đầu vào bộ giải mã với các thông tin đầu mối quan trọng của văn bản tham chiếu. Kết quả thử nghiệm trên tập dữ liệu đã được xây dựng (trong đóng góp thứ hai) cho thấy phương pháp đề xuất có hiệu quả tốt hơn so với một số mô hình đã đạt được kết quả tốt trong các nghiên cứu trước đây.

4. BỐ CỤC CỦA LUẬN ÁN

Nội dung luận án được tổ chức thành bốn chương như sau.

Chương 1. Tổng quan về trích xuất thông tin tự động từ văn bản

Chương 1 trình bày khái quát về trích xuất thông tin, các dạng bài toán trong trích xuất thông tin, cùng với các lĩnh vực ứng dụng đa dạng của trích xuất thông tin. Nội dung chương trình bày các phương pháp tiếp cận dựa trên học máy để giải quyết các bài toán trích xuất thông tin và giới thiệu tóm tắt một số phương pháp học máy được sử dụng trong nghiên cứu đề tài luận án. Từ mục tiêu và phạm vi nghiên cứu luận án, nội dung Chương 1 cũng trình bày khảo sát những nghiên cứu liên quan đến các nội dung thực hiện trong đề tài luận án. Các chương 2, 3 và 4 tiếp theo sẽ trình bày những đóng góp cụ thể của nghiên cứu luận án.

Chương 2. Trích xuất khía cạnh và phân loại quan điểm cho tiếng Việt tận dụng nguồn dữ liệu đã được gán nhãn từ ngôn ngữ khác

Trình bày đề xuất phương pháp trích xuất khía cạnh và phân loại quan điểm cho tiếng Việt bằng cách khai thác nguồn dữ liệu đã được gán nhãn từ ngôn ngữ khác (trong luận án sử dụng là ngôn ngữ tiếng Anh), bao gồm hai nhiệm vụ: (1) trích xuất các loại khía cạnh và (2) phân loại quan điểm. Nội dung trình bày trong chương này được tổng hợp dựa trên kết quả các công trình nghiên cứu [4, 6] (Theo danh mục các công trình công bố).

Chương 3. Trích xuất thực thể và quan hệ trong văn bản pháp quy tiếng Việt sử dụng học máy truyền thống và học sâu

Trình bày đề xuất phương pháp trích xuất thông tin sử dụng học máy truyền thống và học sâu trong lĩnh vực văn bản pháp quy tiếng Việt, bao gồm 2 nhiệm vụ riêng: (1) trích xuất thực thể tham chiếu từ văn bản pháp quy, và (2) phân loại quan hệ giữa các thực thể là tham chiếu (đã được trích xuất ở trên) và thực thể là văn bản pháp quy đang xem xét. Nội dung trình bày trong chương này được tổng hợp dựa trên kết quả các công trình nghiên cứu [1, 5] (Theo danh mục các công trình công bố).

Chương 4. Trích xuất kết hợp đồng thời thực thể và quan hệ trong văn bản pháp quy tiếng Việt sử dụng phương pháp học sâu

Trình bày đề xuất phương pháp trích xuất thông tin kết hợp sử dụng phương pháp học sâu cho văn bản pháp quy tiếng Việt. Mô hình đề xuất sử dụng kiến trúc bộ mã hóa-giải mã dựa trên Transformer với cơ chế giải mã song song không tự hồi quy cho trích xuất đồng thời cả hai thông tin là thực thể tham chiếu và quan hệ giữa các thực thể trong văn bản pháp quy. Nội dung trình bày trong chương này được tổng hợp dựa trên kết quả các công trình nghiên cứu [2, 3] (Theo danh mục các công trình công bố).

Cuối cùng là một số kết luận về luận án và định hướng phát triển nghiên cứu tiếp theo.

CHƯƠNG 1. TỔNG QUAN VỀ TRÍCH XUẤT THÔNG TIN TỰ ĐỘNG TỪ VĂN BẢN

Chương 1 giới thiệu khái quát về trích xuất thông tin từ văn bản, các dạng bài toán trích xuất thông tin và các phương pháp tiếp cận dựa trên học máy để giải quyết các bài toán trong lĩnh vực này. Từ những nghiên cứu cơ bản, nội dung chương sẽ trình bày mục tiêu và phạm vi nghiên cứu đề tài luận án, cùng với những khảo sát về các nghiên cứu liên quan đến mục tiêu luận án thực hiện. Các phương pháp đề xuất và những kết quả nghiên cứu của đề tài luận án sẽ được trình bày chi tiết trong các chương 2, 3 và 4 tiếp theo.

1.1. GIỚI THIỆU VỀ TRÍCH XUẤT THÔNG TIN

Trích xuất thông tin *(Information Extraction, IE)* là việc phát hiện và chọn ra được các thông tin có cấu trúc một cách tự động từ những nguồn không có cấu trúc hoặc bán cấu trúc (ví dụ, các bài báo, văn bản trên web, các bài đánh giá sản phẩm trên mạng xã hội, các ấn phẩm khoa học, hồ sơ y tế,…). Trong hầu hết các trường hợp, trích xuất thông tin thường liên quan đến việc xử lý ngôn ngữ văn bản thuộc lĩnh vực xử lý ngôn ngữ tự nhiên *(Natural Language Processing, NLP)*.

Trích xuất thông tin thực hiện trích xuất tự động những thông tin có cấu trúc như các thực thể, các ý kiến/quan điểm mô tả thực thể, mối quan hệ giữa các thực thể, hay các sự kiện/kịch bản từ các nguồn không có cấu trúc/bán cấu trúc. Việc này cho phép đa dạng hóa hơn khi thực hiện truy vấn với các nguồn dữ liệu không có cấu trúc/bán cấu trúc, thay vì chỉ có thể thực hiện tìm kiếm được với các từ khóa riêng lẻ. Hiểu một cách đơn giản, trích xuất thông tin là việc làm cho thông tin trong văn bản dễ tiếp cận (/truy xuất) hơn để có thể tiếp tục xử lý. Ví dụ, lấy một số thông tin về tiểu sử cá nhân của một người nào đó từ một tập các tài liệu/bài báo là một việc khó

làm, nhưng nếu có sẵn một bảng tổng hợp thông tin về người đó thì việc này trở nên dễ dàng hơn rất nhiều.

Các bài toán trích xuất thông tin trong giai đoạn đầu tập trung vào việc xác định các thực thể có tên, như người, tổ chức và mối quan hệ giữa các thực thể này trong các văn bản ngôn ngữ tự nhiên. Cùng với sự phát triển của mạng Internet, người dùng ngày càng có nhu cầu khai thác thông tin nhiều hơn, với nhiều mức độ và phương pháp đa dạng, dẫn đến ngày càng gia tăng nhu cầu nghiên cứu và thương mại liên quan đến lĩnh vực này. Có thể chia thành bốn nhóm bài toán trích xuất thông tin như sau [101] (Hình 1.1):

1) *Trích xuất thực thể có tên:* Các thực thể thường là cụm danh từ và bao gồm một hoặc một số từ trong văn bản không có cấu trúc. Các thực thể có tên phổ biến nhất là tên người, tên địa danh, tên tổ chức. Ngoài ra hiện nay, tên thực thể được mở rộng bao gồm cả tên thuốc, tên bệnh, tên protein, tên các loại văn bản, bài báo, tạp chí,...

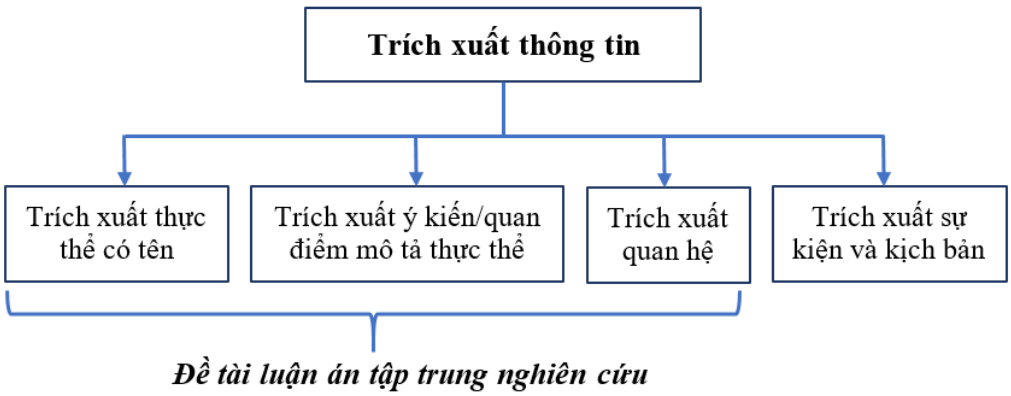
   Ví dụ: Giả sử cho văn bản: "Từ quý II năm 2017, điện thoại Samsung Galaxy S8 chính hãng sẽ được bán tại Viettablet, thành phố Hà Nội.". Một hệ thống trích xuất thực thể có tên sẽ cho kết quả như sau: "Từ quý II năm **2017**[Date], điện thoại **Samsung Galaxy S8**[Product] chính hãng sẽ được bán tại **Viettablet**[Organization], thành phố **Hà Nội**[Location].". Trong đó, có bốn thực thể có tên được trích xuất là "2017" (thuộc thực thể "Date", ngày tháng), "Samsung Galaxy S8" (thuộc thực thể "Product", sản phẩm), "Viettablet" (thuộc thực thể "Organization", tổ chức), và "Hà Nội" (thuộc thực thể "Location", địa danh).

2) *Trích xuất ý kiến/quan điểm mô tả thực thể:* Đây là dạng bài toán khai phá quan điểm và hiện là một chủ đề được quan tâm nghiên cứu tích cực trong nhiều cộng đồng khác nhau. Trong nhiều ứng dụng, có thể cần liên kết một thực thể với tính từ mô tả thực thể đó, thể hiện ý kiến/quan điểm về thực thể. Các ý kiến/quan điểm có thể được thể hiện theo các mức: tích cực, trung tính hoặc tiêu cực.

Ví dụ có một bài bình luận về một nhà hàng trên mạng xã hội: "Các món của nhà hàng này khá ngon!". Với thực thể là nhà hàng (đang được quan tâm trong luồng bình luận), ý kiến về nhà hàng trong bài bình luận này là tốt (quan điểm tích cực).

3) *Trích xuất quan hệ:* Mối quan hệ được xác định qua hai hoặc nhiều thực thể có liên quan đến nhau theo cách được xác định trước. Ví dụ, một nhân viên có mối quan hệ với một tổ chức, một sản phẩm có mối liên hệ với một lượng giá tiền trên một trang web, một văn bản pháp quy là sửa đổi, bổ sung của một văn bản pháp quy khác đã có từ trước,… Như vậy, có thể thấy có trường hợp cần trích xuất mối quan hệ giữa các thực thể trên cùng một văn bản, nhưng có trường hợp cần trích xuất mối quan hệ giữa các thực thể trên nhiều văn bản khác nhau (phức tạp hơn rất nhiều).

4) *Trích xuất sự kiện và kịch bản:* Ở mức độ cao hơn, nhu cầu là cần xác định các loại sự kiện xảy ra trong một chuỗi các bài tin tức liên quan đến một chủ đề cụ thể. Trích xuất sự kiện và kịch bản nhằm xây dựng ra một cấu trúc lớn hơn về các mối liên hệ/kết nối của nhiều thực thể.



Hình 1.1. Các nhóm bài toán trích xuất thông tin

Việc trích xuất thông tin từ các nguồn không có cấu trúc và có nhiễu là một nhiệm vụ đầy thách thức. Với nguồn gốc ban đầu là cộng đồng các nhà nghiên cứu

xử lý ngôn ngữ tự nhiên, cho đến nay trích xuất thông tin đã thu hút được nhiều cộng đồng các nhà nghiên cứu khác nhau như học máy, truy xuất thông tin *(Information Retrieval, IR)*, cơ sở dữ liệu, web và phân tích tài liệu.

1.2. ỨNG DỤNG CỦA TRÍCH XUẤT THÔNG TIN

Hiện tại trên thực tế có khá nhiều ứng dụng của trích xuất thông tin, từ các ứng dụng quản lý thông tin cá nhân, tới các ứng dụng trong doanh nghiệp, các ứng dụng trong các lĩnh vực khoa học, đặc biệt là sự phát triển mạnh mẽ của các ứng dụng hướng web [40,101].

1) *Ứng dụng quản lý thông tin cá nhân:* Các hệ thống quản lý thông tin cá nhân thường tìm cách tổ chức dữ liệu cá nhân (tên người, tên dự án, địa chỉ email, tên văn bản/tài liệu,…) theo các định dạng liên kết có cấu trúc [17,29]. Để làm được việc này, các hệ thống cần phải có khả năng tự động trích xuất các thông tin có cấu trúc từ những nguồn không có cấu trúc (ví dụ như từ các tệp văn bản hiện có). Ví dụ, từ một tệp báo cáo dưới dạng văn bản trình chiếu, có thể trích xuất ra được tên người trình bày, địa chỉ email. Sau đó, từ địa chỉ email có thể thực hiện một số truy xuất thông tin khác liên quan đến người đó, ví dụ như số điện thoại, các dịch vụ người đó sử dụng.
2) *Ứng dụng trong doanh nghiệp:* Trích xuất thông tin có rất nhiều các ứng dụng trong doanh nghiệp.
    - Làm sạch dữ liệu: ví dụ thực hiện chuyển đổi địa chỉ viết dưới dạng một dòng liên tục thành dạng có cấu trúc như tên đường, thành phố, tỉnh,... giúp truy vấn tốt hơn và tránh trùng lặp dữ liệu [8,102].
    - Ứng dụng chăm sóc khách hàng: trích xuất tên sản phẩm và đặc tính sản phẩm khách hàng quan tâm nhằm giới thiệu hoặc cung cấp thêm các dịch vụ liên quan cho khách hàng; liên kết địa chỉ email của khách hàng với giao dịch cụ thể mà khách hàng thực hiện [15].

- Phân loại quảng cáo: thực hiện xây dựng danh sách nhà hàng và các bản tin quảng cáo để trợ giúp cho việc truy vấn liên quan đến quảng cáo, bán hàng [75,79].
- Ứng dụng theo dõi tin tức: tự động theo dõi các loại sự kiện cụ thể từ các nguồn tin tức, ví dụ theo dõi thông tin dịch bệnh, theo dõi các sự kiện khủng bố từ các nguồn tin tức [41,114].

3) *Các ứng dụng trong các lĩnh vực khoa học:* ví dụ như trong lĩnh vực tin sinh học: trích xuất các thông tin như thực thể, tên gọi, các đối tượng sinh học như protein và gen,... và các mối quan hệ giữa chúng [12,93].

4) *Các ứng dụng hướng web:* Trích xuất thông tin trong các ứng dụng hướng web là đa dạng nhất trong các loại ứng dụng. Có thể liệt kê ra một số loại ứng dụng cụ thể như sau:

   - Xây dựng cơ sở dữ liệu ý kiến/quan điểm: từ những nguồn thông tin đa dạng về các loại chủ đề có thể trích xuất và tổ chức thông tin có cấu trúc theo lĩnh vực, cùng với những ý kiến/quan điểm trên các blog, nhóm tin, các bài đánh giá,... từ đó tìm ra được những thông tin có ích. Ví dụ như trích xuất tên sản phẩm, tính năng của sản phẩm và các ý kiến/quan điểm phổ biến cho sản phẩm hoặc tính năng của sản phẩm [91,95].
   - Đặt vị trí quảng cáo trên trang web: giả sử một trang web muốn đặt quảng cáo về một sản phẩm bên cạnh văn bản vừa đề cập đến sản phẩm đó, và bày tỏ ý kiến tích cực về sản phẩm. Để làm được việc này cần phải trích xuất được thông tin đề cập đến sản phẩm và trích chọn được loại ý kiến về sản phẩm [9].
   - So sánh khi mua sắm: tự động thu thập thông tin từ những trang web của người bán hàng để tìm sản phẩm và giá của sản phẩm, và sau đó có thể sử dụng những thông tin này để so sánh khi mua sắm [32,45].

- Xây dựng cơ sở dữ liệu trích dẫn: trích xuất thông tin từ các nguồn khác nhau có cấu trúc phức tạp, như các trang web hội nghị hay các trang web riêng, để tạo ra các trang cơ sở dữ liệu trích dẫn trên web [61,74].
- Các trang web cộng đồng: tạo ra cơ sở dữ liệu từ các tài liệu web là các trang web cộng đồng. Ví dụ theo dõi thông tin về các nhà nghiên cứu, các hội nghị, dự án, các sự kiện liên quan đến một cộng đồng cụ thể: trích xuất tên tác giả, tiêu đề bài viết,... [49,103].
- Tìm kiếm trên web có cấu trúc: cho phép các truy vấn tìm kiếm có cấu trúc liên quan đến các thực thể và các mối quan hệ của chúng trên mạng thông tin toàn cầu *(World Wide Web)*. Việc tìm kiếm dưới dạng từ khóa có thể giúp nhận ra được thông tin về các thực thể, thường là danh từ hoặc cụm danh từ. Tuy nhiên, việc tìm kiếm sẽ khó khăn hơn nhiều khi muốn biết mối quan hệ giữa các thực thể với nhau [16].

1.3. CÁC PHƯƠNG PHÁP TIẾP CẬN

Có hai nhóm phương pháp tiếp cận chính được sử dụng để giải quyết các bài toán trích xuất thông tin là các phương pháp dựa trên luật *(rule-based)* và các phương pháp dựa trên học máy *(learning-based)*. Trong đó, các phương pháp dựa trên học máy có nhiều ưu điểm hơn các phương pháp dựa trên luật do có độ chính xác cao, đảm bảo được việc cập nhật tự động các luật mà không cần phụ thuộc vào chuyên gia, đồng thời cũng có khả năng thích nghi cao và tận dụng được nguồn dữ liệu có sẵn. *Do vậy, đề tài luận án tập trung nghiên cứu các phương pháp học máy để giải quyết một số nhiệm vụ trong trích xuất thông tin tự động từ văn bản.*

Có hai hướng tiếp cận dựa trên các phương pháp học máy truyền thống được sử dụng cho các bài toán trích xuất thông tin là các phương pháp tiếp cận dựa trên phân loại và các phương pháp tiếp cận dựa trên gán nhãn chuỗi [40,94,101]. Với phương pháp tiếp cận dựa trên phân loại, bài toán trích xuất thông tin được quy về bài toán phân loại sử dụng các phương pháp học có giám sát, trong đó mỗi mẫu dữ liệu sẽ được phân loại độc lập. Ví dụ, khi phân loại các khía cạnh của một sản phẩm,

mỗi khía cạnh sẽ được phân loại riêng biệt. Tuy nhiên, trong thực tế có nhiều trường hợp cần phân loại các mẫu theo một trình tự xuất hiện nào đó, khi đó việc phân loại một mẫu sẽ ảnh hưởng đến việc phân loại các mẫu khác. Ví dụ trong bài toán gán nhãn từ loại trong câu văn bản, việc xác định từ xuất hiện sau sẽ liên quan đến từ xuất hiện trước đó. Do vậy, có thể xử lý bài toán trích xuất thông tin theo cách tiếp cận thứ hai là coi bài toán này như một nhiệm vụ gán nhãn chuỗi. Ngoài ra, hiện nay cách tiếp cận dựa trên các phương pháp học sâu cũng được nghiên cứu và áp dụng cho các bài toán này. Các phương pháp dựa trên học sâu sẽ được sử dụng với cả hai nhóm phương pháp dựa trên phân loại và dựa trên gán nhãn chuỗi. Phần sau đây sẽ giới thiệu về các phương pháp tiếp cận giải quyết các bài toán trích xuất thông tin.

1.3.1. Phương pháp tiếp cận dựa trên phân loại

Trích xuất thông tin dựa trên phân loại là cách tiếp cận quy bài toán trích xuất thông tin về bài toán phân loại sử dụng các phương pháp học có giám sát. Phần này sẽ trình bày mô hình phân loại và giới thiệu tóm tắt một phương pháp học máy hiệu quả được sử dụng cho bài toán trích xuất thông tin trong nghiên cứu luận án.

1.3.1.1. Mô hình phân loại

Có hai dạng bài toán phân loại, là phân loại nhị phân và phân loại đa lớp. Trong nhiều trường hợp, chúng ta có thể quy bài toán phân loại đa lớp về bài toán phân loại nhị phân để thuận tiện xử lý.

Xét bài toán phân loại nhị phân như sau:

Tập dữ liệu huấn luyện: $\{(x_1, y_1), \dots, (x_n, y_n)\}$,

trong đó: x_i là ký hiệu một mẫu (là một véc-tơ đặc trưng),

và: $y_i \in \{-1, +1\}$ là ký hiệu cho nhãn phân lớp.

Một mô hình phân loại thường bao gồm hai giai đoạn: huấn luyện và dự đoán. Giai đoạn huấn luyện sẽ cố gắng tìm một mô hình từ dữ liệu đã được gán nhãn mà mô hình này có thể phân tách dữ liệu huấn luyện thành hai lớp. Giai đoạn dự đoán sẽ

sử dụng mô hình đã học được trong giai đoạn huấn luyện để xác định xem liệu một mẫu chưa được gán nhãn sẽ thuộc vào lớp +1 (dương) hay lớp -1 (âm).

Trong trường hợp kết quả dự đoán là các giá trị số, ví dụ trong khoảng từ 0 đến 1, thì một mẫu sẽ được phân lớp theo một số quy tắc. Ví dụ: mẫu được phân vào lớp +1 khi có giá trị dự đoán lớn hơn (hoặc bằng) 0,5; ngược lại các mẫu được phân vào lớp -1.

Một số phương pháp học máy được sử dụng nhiều và rất hiệu quả trong các bài toán phân loại bao gồm: Phân loại Bayes đơn giản *(Naïve Bayes)* [100], Cây quyết định *(Decision Tree)* [98], Máy véc-tơ tựa *(Support Vector Machines, SVM)* [116]. Phần sau sẽ trình bày tóm tắt về Máy véc-tơ tựa, là một kỹ thuật phân lớp có giám sát rất hiệu quả (có độ chính xác cao) đối với nhiều bài toán phân loại khác nhau trong xử lý ngôn ngữ tự nhiên [52,94].

1.3.1.2. Máy véc-tơ tựa

Máy véc-tơ tựa *(Support Vector Machines, SVM)* [116] là kỹ thuật phân lớp có giám sát rất hiệu quả, dựa trên hai nguyên tắc chính. Thứ nhất, SVM thực hiện phân tách các mẫu theo các nhãn khác nhau bằng một siêu phẳng sao cho khoảng cách từ siêu phẳng đến các mẫu có nhãn khác nhau là lớn nhất. Nguyên tắc này được gọi là lề cực đại. Trong quá trình huấn luyện, thuật toán SVM xác định một siêu phẳng có lề cực đại bằng cách giải bài toán tối ưu cho hàm mục tiêu bậc hai. Thứ hai, để giải quyết các trường hợp mẫu không phân tách được bởi siêu phẳng, phương pháp SVM ánh xạ không gian ban đầu của mẫu sang không gian mới nhiều chiều hơn, sau đó tìm siêu phẳng có lề cực đại trong không gian mới này. Để tăng hiệu năng của ánh xạ, SVM sử dụng một kỹ thuật được gọi là hàm nhân, ví dụ, hàm nhân tuyến tính, hàm nhân đa thức, hàm nhân RBF, hàm nhân Gauss.

Phân lớp tuyến tính với lề cực đại. SVMs là các hàm tuyến tính có dạng:

$$f(x) = w^T x + b \tag{1.1}$$

trong đó: $w^T x$ là thành phần bên trong giữa véc-tơ trọng số w và véc-tơ đầu vào x,

b là độ lệch.

Ý tưởng chính của SVM là tìm ra một siêu phẳng phân tách tối ưu, có thể phân tách tối đa hai lớp mẫu huấn luyện (chính xác hơn là tối đa hóa lề giữa hai lớp mẫu). Siêu phẳng sau đó tương ứng với một bộ phân lớp (trong trường hợp này là SVM tuyến tính). Bài toán tìm siêu phẳng có lề cực đại (tối đa hóa lề) có thể được đưa về bài toán tối ưu hóa, tìm w và b sao cho:

$$\frac{1}{2} w^T w \quad \text{là nhỏ nhất}$$

với: $y_i(w^T x_i + b) \geq 1$, với mọi i.

Phân lớp với lề mềm cho dữ liệu không phân chia tuyến tính. Để xử lý các trường hợp không tìm được siêu phẳng nào có thể phân tách dữ liệu huấn luyện thành hai lớp, do có các nhãn nhiễu của cả hai trường hợp mẫu huấn luyện và âm và dương, phương pháp lề mềm *(soft margin)* SVM được đề xuất. Ý tưởng của phương pháp này là tìm siêu phẳng với lề rộng có chấp nhận phân lớp sai một số mẫu. Khi đó, bài toán tìm siêu phẳng có lề cực đại được thể hiện như sau:

Tìm w và b sao cho:

$$\frac{1}{2} w^T w - C \sum_{i=1}^{n} \xi_i \tag{1.2}$$

là nhỏ nhất, với: $y_i(w^T x_i + b) \geq 1 - \xi_i$, với mọi i.

trong đó: $\xi_i \geq 0$ là các biến phụ, được thêm vào để cho phép các mẫu nằm bên trong lề ($1 \geq \xi_i > 0$) hoặc có thể bị phân lớp sai ($\xi_i \geq 1$).

$C \geq 0$ là tham số chi phí kiểm soát lượng lỗi huấn luyện được phép (cố gắng tìm được lề lớn nhất và giảm được số lỗi càng nhiều càng tốt).

Về mặt lý thuyết cần đảm bảo rằng bộ phân lớp tuyến tính thu được theo cách này có các lỗi tổng quát nhỏ. SVM tuyến tính có thể được mở rộng hơn thành các SVM phi tuyến tính bằng cách sử dụng các hàm nhân như hàm Gauss và hàm đa thức [105,116]. Khi có nhiều hơn hai lớp, có thể đưa bài toán về dạng "một lớp so với tất cả các lớp còn lại", nghĩa là lấy một lớp là dương và các lớp khác còn lại là âm.

1.3.2. Phương pháp tiếp cận dựa trên gán nhãn chuỗi

Phần này sẽ trình bày mô hình gán nhãn chuỗi và giới thiệu tóm tắt một phương pháp học máy hiệu quả được sử dụng cho bài toán trích xuất thông tin trong nghiên cứu luận án.

1.3.2.1. Mô hình gán nhãn chuỗi

Trích xuất thông tin có thể được thực hiện như là một nhiệm vụ gán nhãn chuỗi (*sequential labeling*). Trong gán nhãn chuỗi, một văn bản được coi là một chuỗi các từ và một chuỗi nhãn sẽ được gán cho mỗi từ để chỉ ra thuộc tính của từ. Một dạng cấu trúc phổ biến thường được sử dụng trong nhiệm vụ gán nhãn chuỗi là BIO, với ý nghĩa nhãn B là từ bắt đầu (ví dụ từ bắt đầu của một thực thể), nhãn I là từ bên trong (là các từ tiếp theo bên trong của thực thể), và nhãn O là từ bên ngoài (không thuộc thực thể). Ví dụ, xem xét việc gán nhãn trong bài toán xác định thực thể tham chiếu trong văn bản pháp quy, với một câu văn bản pháp quy đầu vào: "*Căn_cứ Luật đất_đai năm 2011;*", sẽ có kết quả chuỗi nhãn đầu ra như sau: [O B I I I O], tương ứng là:

[**O** *Căn_cứ*] [**B** *Luật*] [**I** *đất_đai*] [**I** *năm*] [**I** *2011*] [**O** *;*]

Để hình thức hóa, cho một chuỗi quan sát $x = (x_1, x_{2,\dots}, x_n)$, nhiệm vụ trích xuất thông tin dựa trên gán nhãn chuỗi là việc tìm một chuỗi nhãn $y^* = (y_1, y_{2,\dots}, y_n)$ mà làm cực đại hóa xác suất có điều kiện $p(y|x)$, nghĩa là:

$$y^* = argmax_y p(y|x) \quad (1.3)$$

Khác với các phương pháp học dựa trên luật và phương pháp dựa trên phân loại, gán nhãn chuỗi cho phép mô tả sự phụ thuộc giữa các thông tin đích. Các phụ thuộc có thể được sử dụng để cải thiện độ chính xác của việc trích xuất thông tin. Một số mô hình gán nhãn chuỗi được sử dụng rộng rãi bao gồm: mô hình Markov ẩn *(Hidden Markov Model – HMMs)* [99], Mô hình Markov cực đại hóa Entropy *(Maximum Entropy Markov Models – MEMMs)* [73] và Trường ngẫu nhiên có điều kiện *(Conditional Random Fields – CRFs)* [60]. Phần sau sẽ trình bày tóm tắt về phương pháp được sử dụng phổ biến nhất và rất hiệu quả trong các bài toán gán nhãn chuỗi là phương pháp Trường ngẫu nhiên có điều kiện.

1.3.2.2. Trường ngẫu nhiên có điều kiện

Các mô hình Markov ẩn và Mô hình Markov cực đại hóa Entropy có một số hạn chế như sau:

- Trong mô hình Markov ẩn [99], việc quan sát tại một thời điểm được giả định chỉ phụ thuộc vào trạng thái tại thời điểm đó, để đảm bảo mỗi phần tử quan sát được coi là một phần tử riêng biệt, độc lập với tất cả các phần tử khác trong chuỗi. Giả định này không được thực tế lắm do hầu hết các dữ liệu dạng chuỗi đều không thể biểu diễn chính xác như một phần tử độc lập, mà thông thường mỗi phần tử thường có liên quan (có phụ thuộc) vào các phần tử khác (liền kề hoặc có thể ở xa hơn) trong các chuỗi quan sát.
- Mô hình Markov cực đại hóa Entropy [73] xác định một tập các phân phối xác suất theo trạng thái được huấn luyện riêng. Điều này dẫn đến vấn đề sai lệch nhãn *(label bias)* trong một số tình huống. Vấn đề này xảy ra do MEMM sử dụng mô hình hàm mũ theo trạng thái cho xác suất có điều kiện của các trạng thái tiếp theo với trạng thái hiện tại.

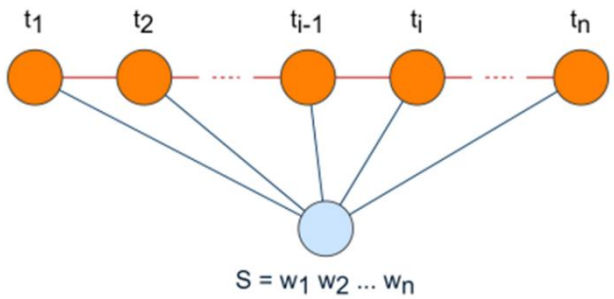
Các hạn chế này sẽ được khắc phục trong mô hình Trường ngẫu nhiên có điều kiện [60]. CRF là mô hình dựa trên xác suất có điều kiện, có thể tích hợp được các thuộc tính đa dạng của chuỗi dữ liệu quan sát để hỗ trợ cho quá trình phân lớp. Tuy vậy, khác với MEMM, CRF là mô hình đồ thị vô hướng. Điều này cho phép CRF có

thể định nghĩa phân phối xác suất của toàn bộ chuỗi trạng thái với điều kiện biết chuỗi quan sát cho trước. Trong khi đó, mô hình MEMM phân phối theo mỗi trạng thái với điều kiện biết trạng thái trước đó và quan sát hiện tại. Với cách mô hình hóa như vậy, CRF có thể giải quyết được vấn đề sai lệch nhãn của mô hình MEMM.

Mô hình đồ thị vô hướng của CRF được huấn luyện để tối đa hóa xác suất có điều kiện, được định nghĩa như sau:

Cho $G = (V, E)$ là đồ thị dạng $Y = (Y_v)_{v \in V}$, sao cho Y được đánh chỉ số theo các đỉnh của G. Khi đó (X, Y) là trường ngẫu nhiên có điều kiện trong trường hợp, khi có điều kiện X, biến ngẫu nhiên Y_v tuân theo thuộc tính Markov theo biểu đồ: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, trong đó $w \sim v$ có nghĩa là w và v là 2 nút kề nhau (hàng xóm) trong G.

CRF là một trường ngẫu nhiên có điều kiện toàn cục dựa trên quan sát X. CRF chuỗi tuyến tính *(linear-chain CRFs)* được giới thiệu lần đầu tiên bởi tác giả Lafferty và các cộng sự vào năm 2001 [60] (Hình 1.2).



Hình 1.2. Trường ngẫu nhiên có điều kiện chuỗi tuyến tính

Theo định lý cơ bản của các trường ngẫu nhiên, phân phối có điều kiện của các nhãn y cho dữ liệu quan sát x có dạng:

$$p_\lambda(y|x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k . f_k(y_{t-1}, y_t, x, t)\right) \quad (1.4)$$

Trong đó $Z_\lambda(x)$ là hệ số chuẩn hóa, còn được gọi là hàm phân vùng, có dạng:

$$Z_\lambda(x) = \sum_y \exp\left(\sum_{t=1}^{T}\sum_k \lambda_k . f_k(y_{t-1}, y_t, x, t)\right) \quad (1.5)$$

trong đó: $f_k(y_{t-1}, y_t, x, t)$ là một hàm đặc trưng có thể có giá trị thực hoặc giá trị nhị phân. Các hàm đặc trưng có thể đo bất kỳ khía cạnh nào của quá trình chuyển đổi trạng thái (từ y_{t-1} sang y_t), và chuỗi quan sát (x), tại bước thời gian hiện tại t.

λ_k tương ứng với trọng số của đặc trưng f_k.

Chuỗi nhãn tiềm năng nhất cho một đầu vào *x*:

$$y^* = argmax_y p_\lambda(y|x) \quad (1.6)$$

có thể được tính toán hiệu quả bằng lập trình quy hoạch động theo thuật toán Viterbi. Các tham số $\lambda = (1, 2, ...)$ có thể được huấn luyện bằng cách tối đa hóa khả năng của tập huấn luyện đã cho.

CRF tránh được vấn đề sai lệch nhãn vì nó có một mô hình hàm mũ duy nhất cho xác suất có điều kiện của toàn bộ chuỗi nhãn cho một chuỗi quan sát. Do đó, trọng số của các đặc trưng khác nhau tại các trạng thái khác nhau có thể được cân bằng với nhau.

1.3.3. Phương pháp tiếp cận dựa trên học sâu

Học sâu là một bước tiến vượt bậc của học máy và được ứng dụng hiệu quả trong rất nhiều lĩnh vực khác nhau. Ưu điểm của phương pháp này là có khả năng mô hình hóa nhiều loại dữ liệu, kết hợp được nhiều nguồn thông tin và có độ chính xác cao. Nhược điểm của phương pháp là yêu cầu tính toán lớn khi huấn luyện mô hình, tuy nhiên vấn đề này có thể khắc phục được nhờ sự phát triển nhanh chóng của các hệ thống phần cứng. Học sâu yêu cầu sử dụng một lượng lớn (/rất lớn) dữ liệu cho quá trình huấn luyện để có thể đưa ra quyết định cho các mẫu dữ liệu mới. Dữ liệu được nạp vào hệ thống để tự huấn luyện thông qua các mạng nơ-ron nhân tạo có nhiều lớp (còn gọi là mạng nơ-ron sâu, *Deep Neural Network*). Hiện nay, học sâu cũng đã

và đang được áp dụng cho nhiều bài toán xử lý ngôn ngữ tự nhiên dạng văn bản. Phần này sẽ trình bày một số phương pháp học sâu được sử dụng cho trích xuất thông tin.

1.3.3.1. Kỹ thuật nhúng từ

Nhúng từ *(Word embeddings)* [76,131] là một bước đột phá quan trọng trong nghiên cứu về xử lý ngôn ngữ tự nhiên. Kỹ thuật này cho phép sử dụng nhúng từ vựng làm đầu vào, thay vì sử dụng các từ và các ký tự trong tập văn bản gốc. Kỹ thuật nhúng từ biểu diễn từ theo phân phối, trong đó các từ được ánh xạ vào không gian véc-tơ có số chiều thấp. So với phương pháp biểu diễn từ truyền thống, nghĩa là biểu diễn theo dạng véc-tơ *one-hot*, phương pháp nhúng từ có hai ưu điểm chính. Thứ nhất là, trong khi các véc-tơ *one-hot* có có số chiều lớn và thưa, thì các véc-tơ nhúng từ có số chiều thấp và dày đặc. Do đó, véc-tơ nhúng từ hiệu quả hơn trong việc biểu diễn và tính toán. Thứ hai là, nhúng từ có khả năng khái quát hóa do các từ giống nhau về mặt ngữ nghĩa được biểu diễn bằng các điểm gần nhau (các véc-tơ tương tự nhau) trong không gian véc-tơ.

Có một số phương pháp tạo nhúng từ từ một kho dữ liệu lớn các văn bản, trong đó phương pháp phổ biến được sử dụng đó là giảm kích thước trên ma trận đồng xuất hiện từ và mạng nơ-ron [76,131].

Trong một kiến trúc mạng nơ-ron phổ biến để học nhúng từ được gọi là mô hình "skip-gram" [76], việc tối ưu hóa xác suất của dữ liệu quan sát được thực hiện như sau:

$$\mathcal{L}(X) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0} \log p(x_{t+j}|x_t) \tag{1.7}$$

với xác suất có điều kiện $p(x_{t+j}|x_t)$ được mô hình hóa như sau:

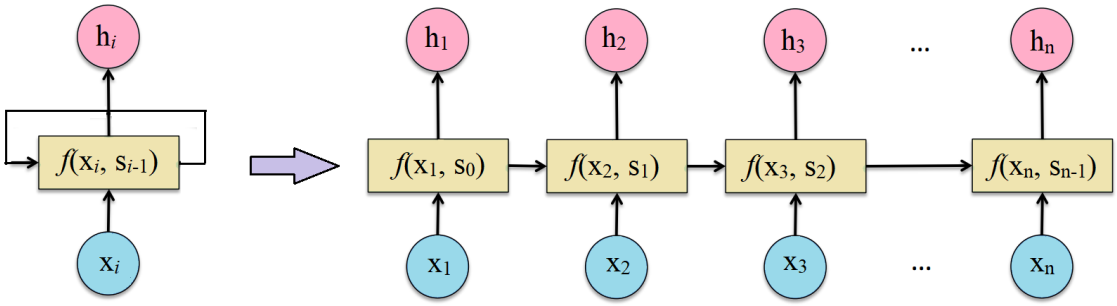
$$p(x_0|x_i) = \frac{\exp(e_{x_0}^T e_{x_i})}{\sum_{x\in W} \exp e_x^T e_{x_i}} \tag{1.8}$$

Trong đó, e_{x_i} và e_{x_o} là nhúng đầu vào và nhúng đầu ra cho từ x. W là tập tất cả các từ.

Nhúng từ có thể được sử dụng trực tiếp như là một đặc trưng rất hiệu quả trong các bài toán xử lý ngôn ngữ tự nhiên [88,113], hoặc có thể đóng vai trò là lớp đầu tiên trong một kiến trúc học sâu [3,87].

1.3.3.2. Mạng nơ-ron hồi quy

Trong các mạng nơ-ron truyền thống, tất cả các đầu vào và đầu ra là hoàn toàn độc lập với nhau, không có sự liên kết nào. Tuy nhiên, trong một số bài toán thực tế, mô hình dạng này nhiều khi không phù hợp. Ví dụ, trong bài toán dự đoán từ trong một câu: nếu muốn dự đoán từ xuất hiện tiếp theo trong câu, thì thông thường cần biết thông tin về từ xuất hiện trước đó. Ý tưởng chính của mạng nơ-ron hồi quy *(Recurrent neural networks, RNNs)* [35] khác với mạng nơ-ron truyền thống là mạng nơ-ron hồi quy sử dụng một bộ nhớ để lưu lại thông tin từ những bước xử lý tính toán trước đó để có thể đưa ra kết quả dự đoán tốt nhất cho bước hiện tại.



Hình 1.3. Minh họa một mạng nơ-ron hồi quy cơ bản

Hình 1.3 minh họa một mạng nơ-ron hồi quy cơ bản, nhận đầu vào từ sự kiện x_i, tiến hành xử lý và đưa ra đầu ra h_i. Điểm khác biệt là mạng nơ-ron hồi quy sẽ lưu giữ lại kết quả h_i để sử dụng cho đầu vào bước tiếp theo. Như vậy, có thể coi mạng nơ-ron hồi quy là một chuỗi các mạng con giống nhau, trong đó mỗi mạng con sẽ truyền thông tin vừa xử lý cho mạng phía sau nó. Khi triển khai mô hình của mạng nơ-ron hồi quy thành từng mạng con, thì mô hình mạng sẽ có kiến trúc bao gồm n

phần tử, tương ứng với *n* tầng nơ-ron. Các bước tính toán bên trong mạng nơ-ron hồi quy được thực hiện như sau:

- x_i là đầu vào tại bước thứ *i*.
- s_i là trạng thái ẩn tại bước thứ *i*. s_i được tính toán dựa trên tất cả các trạng thái ẩn phía trước và đầu vào tại bước *i*.

$$s_i = f(Ux_i + Ws_{i-1}) \tag{1.9}$$

  Hàm f thường là một hàm phi tuyến tính như *Tanh* hoặc ReLU. s_i mang cả thông tin từ trạng thái trước và đầu vào của trạng thái hiện tại nên s_i giống như một bộ nhớ, nhớ được các đặc điểm của các đầu vào x_i. Để thực hiện được phép toán cho phần tử ẩn đầu tiên, cần khởi tạo một giá trị s_0 bằng 0 (hoặc một giá trị ngẫu nhiên), với ý nghĩa là ban đầu bộ nhớ sẽ có giá trị là rỗng do chưa có dữ liệu gì để học.

- h_i là đầu ra tại bước thứ *i*.

$$h_i = O(s_i) = s_i \tag{1.10}$$

  Nếu trong bài toán phân loại thì O có thể được sử dụng là một hàm softmax.

1.3.3.3. Long Short-Term Memory

Về mặt lý thuyết, RNN có khả năng nhớ được những thông tin (kết quả tính toán) từ các bước trước, nhưng thực tế thì mô hình RNN truyền thống không thể nhớ được thông tin từ những bước ở xa do bị mất đạo hàm. Một đề xuất cải tiến cho vấn đề này là LSTM *(Long Short-Term Memory)* [38]. Cơ bản LSTM giống với RNN truyền thống, chỉ khác là LSTM bổ sung thêm các cổng tính toán ở tầng ẩn để giúp cho việc xác định các thông tin cần lưu giữ lại cho bước tiếp theo.

Giả sử cho một chuỗi các véc-tơ đầu vào $(x_1, x_{2,\dots}, x_n)$. LSTM sử dụng một số cổng, bao gồm một cổng vào $\mathrm{I_i}$, một cổng quên $\mathrm{F_i}$, một cổng ra $\mathrm{O_i}$ và một ô nhớ $\mathrm{C_i}$ để cập nhật trạng thái ẩn h_i tại mỗi vị trí *i*. Trạng thái ẩn h_i sau đó được sử dụng để tạo đầu ra y_i như sau:

$$
\begin{aligned}
\mathrm{I_i} &= \sigma(\mathrm{W_I} x_i + \mathrm{V_I} h_{i-1} + \mathrm{b_I}),\\
\mathrm{F_i} &= \sigma(\mathrm{W_F} x_i + \mathrm{V_F} h_{i-1} + \mathrm{b_F}),\\
\mathrm{O_i} &= \sigma(\mathrm{W_O} x_i + \mathrm{V_O} h_{i-1} + \mathrm{b_O}),\\
\mathrm{C_i} &= \mathrm{F_i} \odot \mathrm{C_{i-1}} + \mathrm{I_i} \odot \tanh(\mathrm{W_C} x_i + \mathrm{V_C} h_{i-1} + \mathrm{b_C}),\\
h_i &= \mathrm{O_t} \odot \tanh(\mathrm{C_t}),\\
y_i &= \sigma\big(\mathrm{W_y} h_i + \mathrm{b_y}\big),
\end{aligned}
\tag{1.11}
$$

trong đó σ biểu thị hàm dự đoán (ví dụ trong bài toán phân loại σ có thể là một hàm softmax), ⊙ biểu thị hàm nhân các cặp phần tử, và W, V và b là các ma trận trọng số và vec-tơ cần học.

1.3.3.4. Mô hình Seq2Seq

Mô hình Seq2Seq *(Sequence-to-Sequence)* được đề xuất năm 2014 bởi Sutskever và cộng sự [111]. Từ một chuỗi các từ của câu đầu vào $x = \{x_1, \dots, x_n\}$, mô hình sẽ tạo ra một chuỗi các từ của câu đầu ra $y = \{y_1, \dots, y_m\}$ tương ứng. Mục tiêu của việc huấn luyện là tối ưu hóa xác suất có điều kiện $p(y_1, \dots, y_m | x_1, \dots, x_n)$ với *n* là độ dài của chuỗi đầu vào, *m* là độ dài của chuỗi đầu ra. Mô hình seq2seq sử dụng kiến trúc bộ mã hóa-bộ giải mã *(encoder-decoder)*, trong đó mạng RNN (hoặc các biến thể của nó như LSTM hoặc GRU) sẽ được sử dụng cho cả bộ mã hóa và bộ giải mã. Mạng LSTM có ưu thế hơn do có khả năng giải quyết được các vấn đề phụ thuộc dài, có khả năng ghi nhớ và biểu diễn mối quan hệ của các thông tin phụ thuộc vào ngữ cảnh trong câu văn bản.

- Bộ mã hóa: ánh xạ chuỗi từ đầu vào thành một véc-tơ có kích thước cố định. Tại mỗi bước, bộ mã hóa sẽ nhận véc-tơ tương ứng với mỗi từ trong chuỗi đầu vào để tạo ra véc-tơ trạng thái ẩn s đại diện cho chuỗi đầu vào tại bước mã hóa cuối cùng.
- Bộ giải mã: sử dụng véc-tơ s là khởi tạo cho trạng thái ẩn đầu tiên và tạo ra chuỗi các từ đích tại mỗi bước giải mã. Như vậy, hàm xác suất có điều kiện được viết như sau:

$$p(y_1, \dots, y_m | x_1, \dots, x_n) = \prod_{j=1}^{m} p\left(y_j | s, y_1, \dots, y_{j-1}\right) \quad (1.12)$$

trong đó, mỗi phân bố $p\left(y_j | s, y_1, \dots, y_{j-1}\right)$ là xác suất xuất hiện của từ y_j với véc-tơ đại diện cho câu đầu vào s và các từ đứng trước nó trong chuỗi đầu ra. Hàm softmax sẽ được sử dụng để biểu diễn phân bố này trên tất cả các từ trong tập từ đích.

Sau khi quá trình huấn luyện hoàn thành, có thể tạo ra chuỗi $\hat{y}$ từ một chuỗi đầu vào $\hat{x}$ bằng cách tính toán và chọn chuỗi có khả năng xuất hiện cao nhất (dựa vào mô hình thu được sau khi huấn luyện):

$$\hat{y} = argmax_y p(y|\hat{x}) \quad (1.13)$$

1.3.3.5. Cơ chế Attention

Nhược điểm của mô hình seq2seq là các bộ mã hóa phải ánh xạ chuỗi từ đầu vào thành một véc-tơ cố định (việc này tương đối khó và có thể bỏ sót thông tin) và bộ giải mã chỉ thấy được một véc-tơ đầu vào duy nhất (trong khi tại mỗi time-step thì các phần khác nhau của chuỗi đầu vào có thể có độ hữu dụng khác nhau). Để khắc phục nhược điểm này, tác giả Bahdanau và cộng sự đề xuất cơ chế attention (cơ chế tập trung) [7]. Trong cơ chế này, tại mỗi bước thời gian *(time-step)*, mô hình sẽ tập trung vào các thành phần đầu vào khác nhau. Nghĩa là, cơ chế attention sẽ sử dụng một véc-tơ ngữ cảnh có thể tương tác với toàn bộ véc-tơ trạng thái ẩn của bộ mã hóa thay vì chỉ sử dụng trạng thái ẩn cuối cùng để tạo ra véc-tơ biểu diễn đầu vào cho bộ giải mã.

Tại mỗi bước thời gian *t* ở phía bộ giải mã:

- Nhận véc-tơ trạng thái ẩn của bộ giải mã h_t và tất cả các véc-tơ trạng thái ẩn của bộ mã hóa h_s.
- Tính điểm số attention: Với mỗi véc-tơ trạng thái ẩn của bộ mã hóa cần tính điểm số thể hiện mức độ liên quan với véc-tơ trạng thái ẩn h_t của bộ giải mã, kết quả là một giá trị $score\left(h_t, \overline{h_s}\right)$.

Áp dụng hàm softmax, thu được giá trị trọng số attention:

$$\alpha_{ts} = \frac{\exp\left(score\left(h_t, \overline{h_s}\right)\right)}{\exp\left(\sum_{s'=1}^{S} score\left(h_t, \overline{h_{s'}}\right)\right)} \quad (1.14)$$

- Tính véc-tơ ngữ cảnh c_t tại time-step tương ứng:

$$c_t = \sum_{s'=1}^{S} \alpha_{ts}\, \overline{h_{s'}} \quad (1.15)$$

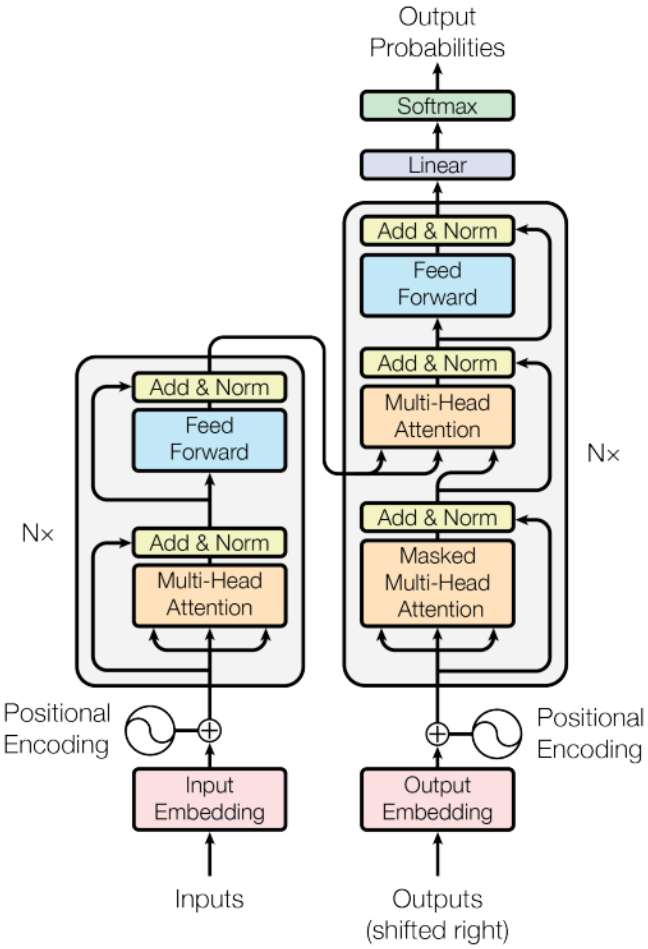
Sau cùng, các véc-tơ attention a_t (được tính dựa trên véc-tơ ngữ cảnh c_t và véc-tơ trạng thái ẩn ở bộ giải mã h_t) được dùng để đưa ra đầu ra.

1.3.3.6. Transformer

Các mô hình mạng RNN tuy có khả năng nắm bắt các phụ thuộc chuỗi có độ dài biến thiên nhưng có nhược điểm là tốc độ huấn luyện chậm, do phải xử lý các phần tử trong chuỗi một cách tuần tự. Trong khi đó các mô hình mạng tích chập CNN *(Convolutional neural networks)* có khả năng thực hiện song song tại từng lớp nhưng lại không thể nắm bắt các phụ thuộc chuỗi có độ dài biến thiên. Tác giả Vaswani và cộng sự [117] đã thiết kế một kiến trúc mới là Transformer có được cả hai ưu điểm trên bằng cách sử dụng cơ chế attention. Kiến trúc này có khả năng song song hóa thông qua việc học chuỗi hồi tiếp với cơ chế attention, đồng thời mã hóa vị trí của từng phần tử trong chuỗi. Kết quả là có một mô hình học sâu tương thích với thời gian huấn luyện ngắn hơn đáng kể. Nếu dữ liệu đầu vào là một câu văn bản, Transformer không cần phải xử lý lần lượt từ phần đầu tới phần cuối của câu. Transformer có thể tận dụng khả năng tính toán song song của GPU và giảm bớt thời gian xử lý, do đó cho phép huấn luyện trên các tập dữ lớn hơn. Kể từ khi xuất hiện vào năm 2017, càng ngày mô hình này càng được lựa chọn nhiều hơn cho các bài toán xử lý ngôn ngữ tự nhiên để thay thế các mô hình dạng RNN như LSTM.

Tương tự như mô hình seq2seq, Transformer cũng dựa trên kiến trúc bộ mã hóa-bộ giải mã. Tuy nhiên, kiến trúc này thay thế các lớp hồi tiếp trong seq2seq bằng các lớp attention đa đầu *(multi-head attention)*, kết hợp thông tin vị trí thông qua biểu

diễn vị trí *(positional encoding)* và áp dụng chuẩn hóa lớp *(layer normalization)* (xem kiến trúc của mô hình Transformer trong Hình 1.4). Giống như seq2seq, trong Transformer, các embeddings của chuỗi nguồn được đưa vào *n* khối lặp lại. Sau đó, đầu ra của khối mã hóa cuối cùng sẽ được sử dụng làm bộ nhớ tập trung cho bộ giải mã. Tương tự như vậy, các embeddings của chuỗi đích được đưa vào *n* khối lặp lại trong bộ giải mã. Đầu ra cuối cùng thu được bằng cách áp dụng một tầng kết nối đầy đủ có kích thước bằng kích thước bộ từ vựng lên các đầu ra của khối giải mã sau cùng.



Hình 1.4. Kiến trúc của mô hình Transformer [117]

Khác với mô hình seq2seq, Transformer sử dụng cơ chế attention (cơ chế tập trung). Một số điểm khác biệt như sau:

- Khối Transformer: một lớp hồi tiếp trong seq2seq được thay bằng một khối Transformer. Với bộ mã hóa, khối này chứa một lớp tập trung đa đầu và một mạng truyền xuôi theo vị trí *(position-wise feed-forward network)* gồm hai lớp kết nối đầy đủ. Với bộ giải mã, khối này có thêm một lớp tập trung đa đầu khác để nhận vào trạng thái bộ mã hóa.
- Cộng và chuẩn hóa: đầu vào và đầu ra của cả lớp tập trung đa đầu hoặc mạng truyền xuôi theo vị trí được xử lý bởi hai lớp "cộng và chuẩn hóa" bao gồm cấu trúc phần dư và lớp chuẩn hóa theo lớp *(layer normalization)*.
- Biểu diễn vị trí: do lớp self-attention (tự tập trung) không phân biệt thứ tự của các phần tử trong chuỗi, nên lớp biểu diễn vị trí được sử dụng để thêm thông tin vị trí vào từng phần tử trong chuỗi.

Self-attention

Cơ chế tự tập trung (hay *self-attention*) cho phép mô hình tìm ra mức độ quan trọng của mỗi từ nên chú ý tới trong dữ liệu đầu vào, bằng cách tính toán trọng số cho mỗi từ (hoặc vị trí) trong câu dựa trên sự tương quan giữa từ đó và các từ khác trong câu, bao gồm cả chính nó *(self)*. Điều này cho phép mô hình tự động học cách lựa chọn thông tin quan trọng và định hình các mối quan hệ ngữ nghĩa giữa các từ trong câu.

Cách tính toán self-attention bắt đầu bằng việc biến đổi đầu vào thành ba ma trận: ma trận truy vấn *(query)*, ma trận khóa *(key)* và ma trận giá trị *(value)*. Mỗi từ trong câu được ánh xạ thành các véc-tơ truy vấn, khóa và giá trị. Tiếp theo, self-attention tính toán điểm tương đồng *(dot product)* giữa mỗi cặp truy vấn-khóa và sau đó áp dụng hàm softmax để có được trọng số cho mỗi từ trong câu. Cuối cùng, thực hiện tính toán trọng số cho ma trận giá trị và kết quả của self-attention là tổng trọng số của các véc-tơ giá trị.

Quá trình self-attention được thực hiện nhiều lần nhằm cho phép mô hình tập trung vào các thông tin khác nhau trong câu. Các trọng số self-attention được học

thông qua huấn luyện mô hình, cho phép học được các mối quan hệ phức tạp giữa các từ trong câu.

Multi-head self-attention (tự tập trung đa đầu)

Trong kiến trúc của Transformer, mỗi một mô-đun self-attention được gọi là một head. Việc sử dụng nhiều head đồng thời gọi là multi-head. Trong multi-head self-attention, quá trình self-attention được áp dụng đồng thời cho nhiều nhóm (head) của truy vấn, khóa và giá trị, từ đó cho phép mô hình học được các mẫu tương quan phức tạp hơn giữa các từ trong câu.

Ý tưởng của multi-head self-attention là thực hiện nhiều phép self-attention song song và độc lập trên các không gian biểu diễn khác nhau của truy vấn, khóa và giá trị. Mỗi head sẽ có các ma trận truy vấn, khóa và giá trị riêng, được tạo ra từ việc ánh xạ từ nguồn ban đầu thông qua các ma trận trọng số học được. Sau đó, quá trình self-attention sẽ được áp dụng độc lập trên mỗi head, và kết quả của các head sẽ được kết hợp thông qua phép tuyến tính để tạo ra đầu ra cuối cùng của multi-head self-attention.

1.3.4. Phương pháp thực hiện thực nghiệm và đánh giá kết quả

Thực nghiệm được tiến hành với các bước như sau: thu thập và gán nhãn dữ liệu, trích chọn đặc trưng, huấn luyện mô hình học máy, kiểm tra mô hình với các mẫu dữ liệu mới, và đánh giá kết quả.

- Thu thập dữ liệu, thực hiện các bước tiền xử lý dữ liệu, gán nhãn cho các mẫu dữ liệu.
- Trích chọn đặc trưng: trích xuất đặc trưng và lựa chọn đặc trưng phù hợp cho bài toán.
- Sử dụng các thuật toán học máy để xây dựng mô hình huấn luyện trích xuất thông tin từ tập dữ liệu đã được gán nhãn.

- Sử dụng mô hình đã được huấn luyện để thực hiện trích xuất thông tin với các mẫu dữ liệu mới.
- Đánh giá kết quả thực nghiệm: Để có thể đánh giá kết quả thực nghiệm cho các nghiên cứu, việc thực nghiệm sẽ được tiến hành nhiều lần trên tập dữ liệu, theo phương pháp kiểm tra chéo *(cross-validation)*. Kết quả được tính trung bình trên số lần thực nghiệm. Ngoài độ chính xác chung *(accuracy)*, kết quả được tính trên các độ đo là độ chính xác *(precision)*, độ phủ *(recall)* và độ đo F_1 [40]. Độ chính xác là tỷ lệ phần trăm các mục được trích xuất chính xác. Độ phủ là tỷ lệ phần trăm các mục thực tế được trích xuất chính xác. Một phương pháp được đánh giá là tốt khi có cả hai giá trị độ chính xác và độ phủ cao. Thực tế có những hệ thống có độ chính xác cao nhưng độ phủ thấp và ngược lại. Độ đo F_1 là giá trị trung bình điều hòa của hai giá trị độ chính xác và độ phủ.

  Các công thức tính độ đo [40] như sau:

$$Accuracy = \frac{|A|}{|B|} \tag{1.16}$$

$$Precision = \frac{|A \cap B|}{|A|} \tag{1.17}$$

$$Recall = \frac{|A \cap B|}{|B|} \tag{1.18}$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{1.19}$$

trong đó, A biểu thị cho tập các giá trị nhãn đầu ra đúng được xác định bởi mô hình, và B biểu thị cho tập các giá trị nhãn gốc (được gán nhãn bởi người gán nhãn). Các tham số này sẽ được giải thích chi tiết hơn trong các phần thiết lập thực nghiệm ở các Chương 2, 3 và 4 của luận án khi xây dựng các mô hình thực nghiệm cụ thể.

1.4. KHẢO SÁT CÁC NGHIÊN CỨU LIÊN QUAN

Mục tiêu của đề tài luận án là nghiên cứu các phương pháp học máy cho trích xuất thông tin tự động từ văn bản, bao gồm hai nội dung cụ thể như sau:

1) Nghiên cứu đề xuất phương pháp trích xuất thông tin cho ngôn ngữ ít tài nguyên bằng cách khai thác nguồn dữ liệu đã được gán nhãn từ ngôn ngữ khác trong bài toán khai phá quan điểm dựa trên khía cạnh tiếng Việt, với hai nhiệm vụ cụ thể: (1) trích xuất các loại khía cạnh và (2) phân loại quan điểm cho khía cạnh (đã được trích xuất).
2) Nghiên cứu đề xuất phương pháp học sâu tiên tiến để giải quyết và nâng cao hiệu quả cho một số nhiệm vụ trích xuất thông tin trong lĩnh vực xử lý văn bản pháp quy. Cụ thể là nghiên cứu đề xuất phương pháp trích xuất thực thể tham chiếu và quan hệ giữa các thực thể trong văn bản pháp quy tiếng Việt, với 2 nhiệm vụ: (1) trích xuất thực thể tham chiếu từ văn bản pháp quy, và (2) phân loại quan hệ giữa các thực thể là tham chiếu và thực thể là văn bản pháp quy đang xem xét.

Sau đây sẽ trình bày một số khảo sát về các nghiên cứu liên quan đến các nội dung trên. *Phần đầu tiên* trình bày các phương pháp khai phá quan điểm dựa trên khía cạnh và những nghiên cứu về phân tích cảm xúc và khai phá quan điểm trong tiếng Việt. *Phần thứ hai* sẽ trình bày các nghiên cứu liên quan đến trích xuất thông tin trong văn bản pháp quy nói chung, các nghiên cứu về trích xuất tham chiếu và quan hệ trong văn bản pháp quy, và một số nghiên cứu xử lý văn bản pháp quy tiếng Việt. Và cuối cùng, *phần thứ ba* sẽ trình bày các nghiên cứu liên quan về mô hình trích xuất kết hợp thực thể và quan hệ dựa trên học sâu.

1) Các phương pháp khai phá quan điểm dựa trên khía cạnh

Trích xuất khía cạnh là nhiệm vụ con đầu tiên trong khai phá quan điểm dựa trên khía cạnh. Nếu khía cạnh bị trích xuất sai, thì việc xác định ý kiến/quan điểm với khía cạnh đó là không có ý nghĩa. Mặt khác, nếu không trích xuất được khía cạnh, thì cũng không xác định được quan điểm về khía cạnh. Do vậy, việc trích xuất được

chính xác khía cạnh của sản phẩm/dịch vụ là việc rất quan trọng. Các phương pháp hiện có cho trích xuất khía cạnh có thể được chia thành hai cách tiếp cận chính, dựa trên phân loại và dựa trên phân cụm.

Đối với cách tiếp cận dựa trên phân cụm, các thuật toán học không giám sát hoặc bán giám sát với kỹ thuật bootstrapping [46,78] là một lựa chọn thích hợp vì phần lớn nghiên cứu ban đầu không có nguồn dữ liệu đã được chú thích, hoặc trong một số ngữ cảnh có thể không xác định được chính xác danh sách các loại khía cạnh. Các phương pháp theo cách tiếp cận này thường trích xuất các loại khía cạnh với hai bước: (1) trích xuất tất cả các thuật ngữ khía cạnh từ một văn bản được cho; và (2) phân cụm các thuật ngữ khía cạnh có ý nghĩa tương tự nhau thành các nhóm danh mục (hay loại) khía cạnh.

Các phương pháp theo cách tiếp cận dựa trên phân loại được sử dụng trong trường hợp biết được chính xác danh sách các loại khía cạnh muốn trích xuất và có được nguồn dữ liệu đã được chú thích với các nhãn loại khía cạnh. Các phương pháp theo cách tiếp cận này sử dụng các thuật toán học có giám sát và thường mô hình hóa nhiệm vụ như một bài toán phân loại đa nhãn trong đó mỗi nhãn tương ứng với một loại khía cạnh, hoặc nhiều bài toán phân loại nhị phân trong đó mỗi bài toán xử lý một loại khía cạnh cụ thể [52,94,126]. So với cách tiếp cận dựa trên phân cụm, cách tiếp cận dựa trên phân loại chính xác hơn do có thể sử dụng học có giám sát với dữ liệu đã được chú thích.

Một số phương pháp khác nhau đã được đề xuất để giải quyết bài toán phân loại quan điểm dựa trên khía cạnh, từ các thuật toán học truyền thống đến các kỹ thuật tiên tiến hơn với các mạng nơ-ron. Một số mô hình mạng nơ-ron đã thành công với phân loại quan điểm dựa trên khía cạnh, ví dụ như các mạng LSTM [84,123], mạng nơ-ron tích chập [129], mạng hồi quy [89,122]. Tuy nhiên, các mô hình học sâu thường được yêu cầu huấn luyện với một lượng lớn dữ liệu được gán nhãn, đây là một khó khăn lớn đối với các nhiệm vụ NLP trong các ngôn ngữ có ít tài nguyên dữ liệu được gán nhãn sẵn.

Phân tích cảm xúc và khai phá quan điểm trong tiếng Việt

Tương tự như trong các ngôn ngữ khác, hầu hết các nhiệm vụ trước đây về phân tích cảm xúc và khai phá quan điểm tiếng Việt thường tập trung vào phân loại quan điểm. Các phương pháp tiếp cận hiện tại bao gồm từ các phương pháp dựa trên luật [57] đến các thuật toán học có giám sát/bán giám sát yếu [5,43]. Một nghiên cứu thực nghiệm về phân loại quan điểm dựa trên học máy cho tiếng Việt được tác giả Duyen và các cộng sự mô tả trong [33]. Các tác giả giới thiệu một tập dữ liệu có chú thích trong miền lĩnh vực khách sạn và thực hiện một loạt các thực nghiệm trên tập dữ liệu bằng cách sử dụng một số phương pháp trích xuất đặc trưng và thuật toán học có giám sát. Tác giả Ha và cộng sự [43] mô tả một phương pháp sử dụng các đặc trưng bag-of-bigram trong lifelong learning framework để phân loại quan điểm tiếng Việt chéo miền (*cross-domain*). Tác giả Bach và Phuong [5] trình bày một phương pháp giám sát yếu để phân loại quan điểm trong các ngôn ngữ có ít tài nguyên dữ liệu được gán nhãn sẵn. Kết quả thực nghiệm trên hai tập dữ liệu tiếng Nhật và tiếng Việt cho thấy hiệu quả của phương pháp tác giả đề xuất. Tác giả Phu và cộng sự [92] đề xuất một mô hình valence-totaling cho phân loại quan điểm tiếng Việt, đạt được độ chính xác 63,9% trên một tập có 30.000 tài liệu tiếng Việt.

Trong số các nghiên cứu trước đây về khai phá quan điểm dựa trên khía cạnh tiếng Việt, tác giả Vu và cộng sự [118] trình bày một phương pháp dựa trên luật để khai thác đánh giá sản phẩm. Trong đó, từ thể hiện khía cạnh và từ thể hiện quan điểm được trích xuất bằng cách sử dụng một bộ luật cú pháp tiếng Việt. Tác giả Le và cộng sự [62] trình bày một phương pháp bán giám sát để trích xuất và phân loại các thuật ngữ khía cạnh trong văn bản tiếng Việt. Phương pháp của họ có thể được phân vào nhóm phương pháp dựa trên phân cụm. Đầu tiên, phương pháp này trích xuất tất cả các từ *(token)* từ một kho dữ liệu văn bản thuần túy và chọn ra những từ có tần suất xuất hiện nhiều nhất. Các từ này sẽ được gán nhãn thủ công và được chọn làm các từ hạt giống cho thuật toán. Sau đó, tác giả xây dựng một cây quyết định để phân loại khía cạnh.

Gần đây, một số nghiên cứu đã được trình bày bằng các thuật toán học có giám sát [30,85] trong một cuộc thi giải quyết bài toán phân tích quan điểm dựa trên khía cạnh tiếng Việt tại hội thảo quốc tế lần thứ năm về Xử lý ngôn ngữ và tiếng Việt (VLSP 2018) [85].

Nghiên cứu trong luận án khác với các nghiên cứu trước ở chỗ sẽ xác định trực tiếp các loại khía cạnh mà không cần trích xuất ra các thuật ngữ khía cạnh. Hơn nữa, phương pháp đề xuất sẽ tận dụng dữ liệu đã được chú thích (gán nhãn) sẵn từ ngôn ngữ khác để cải tiến hiệu quả cho trích xuất khía cạnh và quan điểm về khía cạnh trong tiếng Việt.

2) Trích xuất thông tin trong văn bản pháp quy và các phương pháp

Trích xuất thông tin trong văn bản pháp quy: Nghiên cứu của tác giả Walter [120] trình bày một phương pháp dựa trên luật cho phép sử dụng cây phân tích cú pháp phụ thuộc để trích xuất các định nghĩa trong văn bản pháp quy từ các quyết định của tòa án Đức, nhằm trợ giúp cho việc phân tích và hiểu các quyết định của tòa án dễ dàng hơn. Kết quả thực nghiệm cho thấy phương pháp đề xuất đạt được độ chính xác tương đối cao và có thể áp dụng để trích xuất định nghĩa trong các văn bản pháp quy thuộc các lĩnh vực khác nhau. Nghiên cứu [26] xây dựng hệ thống Legal TRUTHS để trích xuất thông tin liên quan từ các quyết định của tòa án tối cao Philippines về các vụ án hình sự. Các thông tin trích xuất bao gồm loại tội phạm, ngày giờ phạm tội, nguyên đơn và hình phạt, được xác định từ một tập văn bản mẫu. Nghiên cứu [97] sử dụng cách tiếp cận kết hợp học máy và đặc trưng về ngôn ngữ để trích xuất thông tin từ văn bản pháp quy, kết quả đạt được độ chính xác tương đối cao trên tập dữ liệu văn bản pháp quy có sẵn tại EUR-Lex. Nghiên cứu này sử dụng bộ phân lớp SVM để liên kết các khái niệm với văn bản pháp quy và bộ phân tích cú pháp ngôn ngữ tự nhiên để xác định các thực thể có tên, gồm vị trí, tổ chức, ngày tháng và tham chiếu đến các văn bản khác. Nghiên cứu [2] trình bày một phương pháp tự động xác định và chú thích các thực thể như tên người, tổ chức, vai trò và chức năng của người, cùng với các quan hệ giữa các thực thể trong văn bản pháp quy.

Phương pháp sử dụng kết hợp cả kỹ thuật dựa trên luật (các biểu thức chính quy) và kỹ thuật học máy (trường ngẫu nhiên có điều kiện, CRF), kết quả cũng thu được độ chính xác khá cao.

Hiện nay, các mô hình mới cho trích xuất thực thể có tên chủ yếu dựa trên mạng nơ-ron như BiLSTM, Transformer [63,64]. Trong nghiên cứu [14], các tác giả trình bày một mô hình nhận dạng thực thể cho văn bản pháp quy tiếng Thổ Nhĩ Kỳ dựa trên các mô hình CRF và BiLSTMs, kết hợp các yếu tố đặc trưng riêng của ngôn ngữ Thổ Nhĩ Kỳ. Nghiên cứu thử nghiệm với một số phương pháp kết hợp khác nhau, bao gồm các phương pháp nhúng từ (GloVe, Morph2Vec) và các kỹ thuật trích xuất đặc trưng từ ký tự dựa trên các mạng nơ-ron (BiLSTM, mạng nơ-ron tích chập), cho kết quả trích xuất thực thể tương đối hiệu quả. Tác giả Mandal và các cộng sự [70] đã trình bày một nghiên cứu trích xuất các cụm từ quan trọng trong văn bản pháp quy sử dụng các phương pháp học sâu khác nhau như CNN và GRU. Trong nghiên cứu [28], các tác giả cũng giới thiệu mô hình dựa trên CRF và BiLSTM để nhận dạng các thực thể có tên trong văn bản pháp quy tiếng Bồ Đào Nha.

Để nâng cao hiệu quả của các mô hình trích xuất thông tin, một số nghiên cứu đã xây dựng các nhúng từ riêng cho miền lĩnh vực văn bản pháp quy khi triển khai các mô hình học sâu tiên tiến [19], trong đó có Law2Vec [20], LEGAL-BERT [18], Trinh và cộng sự [138], Song và cộng sự [107], VNLawBERT cho văn bản tiếng Việt [23], Lawformer cho tiếng Trung Quốc [127]. Tác giả Filtz và cộng sự [36] đã tiến hành thử nghiệm trích xuất các thực thể pháp quy dựa trên các mô hình ngôn ngữ khác nhau và đánh giá, so sánh với các kỹ thuật cổ điển như CRF. Trong nghiên cứu [20], các tác giả Chalkidis và Kampas cũng đã tiến hành đánh giá về hiệu quả của việc sử dụng nhúng từ trong các nhiệm vụ xử lý văn bản pháp quy, bao gồm cả trích xuất thông tin. Các nghiên cứu này đều cho thấy tính hiệu quả cao của các mô hình ngôn ngữ huấn luyện trước trên miền lĩnh vực văn bản pháp quy.

Trích xuất tham chiếu trong văn bản pháp quy: Trích xuất tham chiếu từ các văn bản pháp quy đã được nghiên cứu với nhiều ngôn ngữ khác nhau, bao gồm tiếng

Ý [90], tiếng Tây Ban Nha [71], tiếng Hà Lan [69] và tiếng Nhật [112]. Tác giả Palmirani và cộng sự [90] giới thiệu một phương pháp trích xuất các tham chiếu quy phạm từ các văn bản pháp quy của Ý bao gồm 5 bước bằng cách sử dụng các luật và biểu thức chính quy. Kết quả cho phép xác định các phần khác nhau của một văn bản pháp quy bao gồm: thông tin nhận dạng (loại văn bản, số văn bản, ngày gửi và các thông tin tương tự), các phân vùng (ví dụ như các điều khoản và mục tạo thành bố cục của văn bản) và các tham chiếu pháp quy có trong văn bản. Tác giả Mercedes và cộng sự [71] trình bày một phương pháp dựa trên luật để trích xuất và phân giải tham chiếu từ các văn bản pháp quy Tây Ban Nha, để lưu trữ và sử dụng trong các thư viện kỹ thuật số. Các tham chiếu được trích xuất bằng cách so khớp (sử dụng ngữ pháp) văn bản trong bộ dữ liệu với các tập mẫu. Tác giả Maat và cộng sự [69] mô tả một phương pháp dựa theo ngữ pháp để phát hiện tự động các cấu trúc tham chiếu trong văn bản pháp quy Hà Lan. Để thực hiện điều này, các tác giả khảo sát chi tiết về loại và cấu trúc của các văn bản pháp quy, từ đó phát triển phương pháp tìm kiếm tham chiếu dựa theo ngữ pháp. Phương pháp đề xuất mặc dù tương đối đơn giản nhưng có độ chính xác khá cao.

Không giống như các nghiên cứu khác chỉ thực hiện phân giải tham chiếu ở mức văn bản, tác giả Tran và các cộng sự [112] giới thiệu một phương pháp có thể trích xuất tham chiếu đến các mục con trong văn bản pháp quy Nhật Bản, sử dụng cả hai cách tiếp cận dựa trên luật và học máy, với bốn bước là phát hiện đề cập *(mention),* trích xuất thông tin theo ngữ cảnh, trích xuất ứng viên tiền đề *(antecedent candidate)* và xác định tiền đề. Báo cáo kết quả nghiên cứu đạt độ đo F_1 là 80,06% trong việc phát hiện tham chiếu và độ chính xác là 85,61% cho việc phân giải tham chiếu trên một tập văn bản đã được chú thích từ Luật hưu trí quốc gia Nhật Bản.

Trích xuất quan hệ: Các nghiên cứu trước đây về trích xuất quan hệ thường sử dụng phương pháp tiếp cận dựa trên luật, như trong các nghiên cứu [81,136]. Các phương pháp này thường cần phải xác định trước các luật mô tả cấu trúc của các thực thể liên quan. Phương pháp dựa trên luật yêu cầu người tạo ra luật cần có những hiểu biết sâu về nền tảng và đặc điểm của miền lĩnh vực văn bản xử lý. Do vậy, nhược

điểm chính của phương pháp tiếp cận này là cần phải có sự tham gia của chuyên gia và khó chuyển đổi giữa các miền lĩnh vực khác nhau.

Một cách tiếp cận phổ biến hiện nay cho trích xuất quan hệ là dựa trên học máy thống kê. Trong đó, có một số nghiên cứu dựa trên các phương pháp học không giám sát và bán giám sát như [44,109]. Tuy nhiên, phổ biến nhất là các nghiên cứu dựa trên học có giám sát để trích xuất quan hệ với độ chính xác tương đối cao. Trong mô hình học có giám sát, trích xuất quan hệ được coi là bài toán phân loại. Nghiên cứu của tác giả Kambhatla [54] sử dụng các đặc trưng từ vựng, cú pháp và ngữ nghĩa khác nhau cùng với bộ phân loại entropy cực đại để trích xuất các loại quan hệ. Nghiên cứu [13] đề xuất các nhân *(kernel)* dựa trên đường đi ngắn nhất, từ đó xác định độ đo tương tự hiệu quả giữa các đối tượng trong một không gian nhiều chiều hơn.

Gần đây, các nghiên cứu về trích xuất quan hệ dựa trên mô hình học sâu đang dần được quan tâm nhiều hơn do các mô hình này có khả năng tự học đặc trưng và đã thu được nhiều kết quả đáng khích lệ. Các nghiên cứu [51,66,133] dựa trên các cấu trúc mạng đa dạng, như mạng nơ-ron tích chập (CNN), mạng nơ-ron hồi quy (RNN), kết hợp với cơ chế tập trung giúp trích xuất các quan hệ hiệu quả và có độ chính xác cao. Tuy nhiên, hạn chế chính của cách tiếp cận này so với các phương pháp học máy thống kê là tốc độ, cùng với yêu cầu cần phải có tập dữ liệu huấn luyện đủ lớn.

Xử lý văn bản pháp quy tiếng Việt: Có rất ít nghiên cứu về xử lý văn bản pháp quy tiếng Việt. Tác giả Bui và cộng sự [11] mô tả một công cụ tìm kiếm cho các văn bản pháp quy tiếng Việt. Ý tưởng chính là các văn bản có thể được lập chỉ mục theo một số khía cạnh, bao gồm toàn bộ văn bản, cấu trúc logic và ontology của các văn bản pháp quy tiếng Việt. Nhóm tác giả Bui và Ho [10] và nhóm tác giả Son và các cộng sự [106] đề cập đến nhiệm vụ nhận dạng các phần logic/cấu trúc logic trong các văn bản pháp quy tiếng Việt. Trong khi nhóm tác giả Bui và Ho [10] sử dụng cách tiếp cận dựa trên luật, thì nhóm tác giả Son và các cộng sự [106] sử dụng một phương

pháp học máy thống kê, là trường ngẫu nhiên có điều kiện (CRF), và đạt được kết quả tốt hơn. Trong nghiên cứu [115], các tác giả đề xuất một phương pháp dựa trên đặc điểm phụ thuộc cú pháp kết hợp bộ phân lớp SVM để trích xuất các mối quan hệ ngữ nghĩa giữa các thực thể trong văn bản pháp quy tiếng Việt. Các thực thể được trích xuất bằng cách sử dụng biểu thức chính quy. Tập dữ liệu thử nghiệm rất ít, chỉ bao gồm 200 văn bản pháp quy tiếng Việt. Tác giả Nguyen và cộng sự [83] khai thác các đặc tính của văn bản pháp quy tiếng Việt và đề xuất một phương pháp biểu diễn cơ sở tri thức cho các văn bản và mối quan hệ của chúng, sử dụng 3 phương pháp là Mạng ngữ nghĩa, Các luật, và Ngôn ngữ khung, và tận dụng các ưu điểm của các phương pháp này. Trong nghiên cứu [6], tác giả Bach và cộng sự trích xuất thông tin quan trọng trong các câu hỏi pháp luật Việt Nam, để sử dụng trong một hệ thống trả lời câu hỏi tự động. Kết quả nghiên cứu đạt độ đo F_1 là 92,66% trên một tập dữ liệu bao gồm 1678 câu hỏi về luật giao thông bằng tiếng Việt. Gần đây, tác giả Chau và công sự [23] đề xuất một phương pháp lựa chọn câu trả lời bằng cách tinh chỉnh và huấn luyện mô hình ngôn ngữ BERT trên bộ dữ liệu cặp câu hỏi-câu trả lời pháp quy tiếng Việt. Mô hình BERT được huấn luyện trước trên bộ dữ liệu đã thu thập cho kết quả cao hơn 3,6% (tính theo độ đo F_1) so với mô hình BERT tinh chỉnh, cho thấy tiềm năng của mô hình ngôn ngữ được huấn luyện trước đối với lĩnh vực văn bản pháp quy.

Như vậy, theo khảo sát của chúng tôi, đây là nghiên cứu đầu tiên về trích xuất thực thể tham chiếu và quan hệ cho văn bản pháp quy tiếng Việt sử dụng cách tiếp cận dựa trên học máy. Hơn nữa, nghiên cứu luận án sử dụng các phương pháp học sâu tiên tiến, bên cạnh các phương pháp học máy truyền thống, cho cả hai nhiệm vụ trích xuất thực thể tham chiếu và phân loại quan hệ giữa các thực thể trong lĩnh vực văn bản pháp quy.

3) Các mô hình trích xuất kết hợp thực thể và quan hệ dựa trên học sâu

Trích xuất thực thể và quan hệ là một nhiệm vụ nhận được nhiều sự quan tâm từ cộng đồng nghiên cứu trong những năm gần đây. Cách tiếp cận ban đầu là trích

xuất thông tin tuần tự như trong nghiên cứu [21]. Phương pháp thực hiện là, đầu tiên trích xuất tất cả các thực thể trong một câu và sau đó phân loại mối quan hệ giữa các thực thể đã được trích xuất. Cách tiếp cận này coi việc trích xuất thực thể và phân loại quan hệ là hai nhiệm vụ con riêng biệt. Mặc dù phương pháp này dễ thực hiện, nhưng lại có khả năng lan truyền lỗi (ví dụ, trích xuất thực thể sai sẽ dẫn đến phân loại quan hệ bị sai) và có thể bỏ qua mối liên hệ tương quan giữa hai nhiệm vụ con [65].

Các phương pháp tiên tiến sử dụng các mô hình trích xuất kết hợp đã được đề xuất để giải quyết khó khăn này. Các nghiên cứu ban đầu hiệu quả như [77,139] là các mô hình kết hợp giúp giảm sự lan truyền lỗi bằng cách sử dụng phương pháp chia sẻ tham số để đạt được sự tương tác giữa nhận dạng thực thể và phân loại quan hệ. Tuy nhiên, các mô hình này không thực hiện giải mã kết hợp và vẫn chuyển cặp thực thể được xác định tới bộ phân loại quan hệ để xác định mối quan hệ. Việc giải mã riêng biệt khiến cho sự phụ thuộc giữa các thực thể và mối quan hệ dự đoán không được khai thác đầy đủ. Các phương pháp trong các nghiên cứu [25,125,132,137] đã thực hiện giải mã kết hợp bằng cách sử dụng gán nhãn/thẻ (tagging) và chuyển đổi nhiệm vụ trích xuất thực thể và quan hệ thành bài toán gán nhãn chuỗi với việc trích xuất đồng thời các thực thể và quan hệ.

Ngoài ra, một số nghiên cứu gần đây đã đề xuất và sử dụng các mô hình trích xuất dựa trên kiến trúc bộ mã hóa-giải mã seq2seq như [80,124,135]. Các phương pháp này đã cải thiện độ chính xác của việc trích xuất thực thể và quan hệ đồng thời tăng tốc độ xử lý. Trong [135], các tác giả đề xuất một mô hình đầu cuối-đến-đầu cuối *(end-to-end)* dựa trên việc học từ chuỗi này sang chuỗi khác *(sequence-to-sequence)* với cơ chế sao chép trong bộ giải mã, có thể cùng trích xuất các mối quan hệ từ các câu. Cơ chế sao chép cho phép mô hình tránh được vấn đề ngoài từ vựng *(out-of-vocabulary)*. Tuy nhiên, độ chính xác trích xuất thực thể chưa cao do khả năng phân biệt thực thể đứng đầu và thực thể cuối còn chưa tốt. CopyMTL [134] giải quyết những vấn đề này bằng cách sử dụng khung học tập đa tác vụ được trang bị cơ chế sao chép và quy trình ghi nhãn tuần tự. Trong [80], các tác giả đề xuất một phương

pháp giải mã dựa trên mạng con trỏ trong đó toàn bộ tập dữ liệu được tạo tại mỗi bước thời gian (time-step). Mô hình này có độ đo F_1 cải thiện hơn. Nghiên cứu trong [124] coi nhận dạng thực thể và phân loại quan hệ là một bài toán điền thông tin phù hợp vào bảng *(table-filling)* chỉ với một không gian nhãn, trong đó mỗi mục trong bảng đầu vào (chứa tất cả các cặp từ trong một câu) biểu thị sự tương tác giữa hai từ riêng lẻ.

Tuy nhiên, các mô hình này sử dụng bộ giải mã tự hồi quy (*autoregressive decoder*) dựa trên dự đoán các giá trị tương lai từ các giá trị trong quá khứ và cơ chế bước thời gian. Hạn chế của dạng mô hình này là phức tạp, tốn thời gian vì đầu ra của các cặp thực thể và quan hệ trong câu đầu vào phải theo thứ tự, khiến mô hình không chỉ cần học cách sinh bộ ba (gồm 2 thực thể và quan hệ giữa chúng), mà còn phải xem xét đến thứ tự của chuỗi đầu ra. Để giải quyết khó khăn này, nghiên cứu [108] đã đề xuất một mô hình bộ mã hóa-giải mã hiệu quả hơn. Mô hình này có thể giải mã đồng thời các thực thể và quan hệ, đồng thời coi việc trích xuất tổ hợp thực thể và quan hệ là một bài toán dự đoán theo tập hợp, mà không cần quan tâm đến thứ tự của các bộ ba trong tập hợp. Mặc dù thành công nhưng phương pháp này thiếu việc sử dụng thông tin ngữ cảnh rõ ràng. Cụ thể, [108] chỉ sử dụng cố định các đầu vào của bộ giải mã mà không tận dụng các thông tin bổ trợ từ câu đầu vào để hỗ trợ bộ giải mã trích xuất đồng thời các bộ ba.

Nghiên cứu trong luận án giải quyết vấn đề này bằng cách mã hóa thông tin có sẵn để giúp xây dựng các đầu vào bộ giải mã tốt hơn. Ý tưởng là cải thiện đầu vào bộ giải mã với thông tin có trước về vị trí bắt đầu của thực thể tham chiếu trong câu đầu vào. Điều này làm cho mô hình trích xuất kết hợp thực thể tham chiếu và quan hệ giữa các thực thể trong văn bản pháp quy đạt hiệu quả tốt hơn.

1.5. KẾT LUẬN CHƯƠNG 1

Chương 1 giới thiệu khái quát về trích xuất thông tin từ văn bản. Có 2 nhóm phương pháp tiếp cận dựa trên học máy thường được sử dụng cho trích xuất thông tin, đó là phương pháp tiếp cận dựa trên phân loại và phương pháp tiếp cận dựa trên

gán nhãn chuỗi. Ngoài ra, hiện nay trích xuất thông tin cũng được xử lý theo phương pháp tiếp cận tiên tiến hơn là dựa trên học sâu.

Từ những cơ sở nghiên cứu cơ bản, nội dung Chương 1 đã trình bày mục tiêu và phạm vi nghiên cứu đề tài luận án, đó là nghiên cứu các phương pháp học máy cho trích xuất thông tin tự động trong văn bản, bao gồm hai nội dung chính:

1) Nghiên cứu đề xuất phương pháp trích xuất thông tin cho ngôn ngữ ít tài nguyên bằng cách tận dụng nguồn dữ liệu đã được gán nhãn từ ngôn ngữ khác trong bài toán khai phá quan điểm dựa trên khía cạnh tiếng Việt, với hai nhiệm vụ cụ thể: (1) trích xuất các loại khía cạnh và (2) phân loại quan điểm cho khía cạnh (đã được trích xuất).
2) Nghiên cứu đề xuất phương pháp học sâu tiên tiến để giải quyết và nâng cao hiệu quả cho một số nhiệm vụ trích xuất thông tin trong lĩnh vực xử lý văn bản pháp quy. Cụ thể là nghiên cứu đề xuất phương pháp trích xuất thực thể tham chiếu và quan hệ trong các văn bản pháp quy tiếng Việt, với 2 nhiệm vụ: (1) trích xuất thực thể tham chiếu từ văn bản pháp quy, và (2) phân loại quan hệ giữa các thực thể là tham chiếu và thực thể là văn bản pháp quy đang xem xét.

Đi kèm với những mục tiêu này, nội dung Chương 1 cũng đã trình bày khảo sát và phân tích các nghiên cứu liên quan: (1) Các phương pháp khai phá quan điểm dựa trên khía cạnh và những nghiên cứu về phân tích cảm xúc và khai phá quan điểm trong tiếng Việt; (2) Các nghiên cứu liên quan đến trích xuất thông tin trong văn bản pháp quy nói chung, các nghiên cứu về trích xuất tham chiếu và quan hệ trong văn bản pháp quy, và một số nghiên cứu xử lý văn bản pháp quy tiếng Việt; và (3) Các nghiên cứu liên quan về mô hình trích xuất kết hợp thực thể và quan hệ dựa trên học sâu.

Các phương pháp đề xuất và những kết quả nghiên cứu của đề tài luận án sẽ được trình bày chi tiết trong các Chương 2, 3 và 4 tiếp theo.

CHƯƠNG 2. TRÍCH XUẤT KHÍA CẠNH VÀ PHÂN LOẠI QUAN ĐIỂM CHO TIẾNG VIỆT TẬN DỤNG NGUỒN DỮ LIỆU ĐÃ ĐƯỢC GÁN NHÃN TỪ NGÔN NGỮ KHÁC

Các phương pháp học máy đã được chứng minh là có hiệu quả trong nghiên cứu về trích xuất thông tin trước đây, nhưng các phương pháp này thường yêu cầu cần phải chú thích (gán nhãn) thủ công một lượng lớn dữ liệu cho giai đoạn huấn luyện, việc này rất tốn kém thời gian và chi phí. Ngoài ra, trong xử lý ngôn ngữ tự nhiên, các phương pháp học máy thường phụ thuộc vào miền lĩnh vực. Do đó, việc áp dụng các phương pháp học máy vào một miền lĩnh vực mới hoặc một ngôn ngữ mới, đặc biệt là ngôn ngữ có ít tài nguyên dữ liệu được gán nhãn sẵn (như tiếng Việt) trong một số bài toán, vẫn còn rất nhiều khó khăn. Để giải quyết vấn đề này, nội dung Chương 2 luận án tập trung nghiên cứu một số nhiệm vụ trích xuất thông tin cho ngôn ngữ ít tài nguyên bằng cách khai thác dữ liệu đã được gán nhãn sẵn từ các ngôn ngữ khác để bổ sung dữ liệu cho tập huấn luyện khi xây dựng các mô hình học máy. Giải pháp này khá tổng quát và linh hoạt do không phụ thuộc vào ngôn ngữ và các thuật toán học máy, giảm thời gian và chi phí gán nhãn dữ liệu thủ công, đồng thời giúp nâng cao hiệu quả trích xuất.

Để thực hiện nội dung nghiên cứu này, luận án tập trung nghiên cứu trích xuất khía cạnh và phân loại quan điểm trong bài toán khai phá quan điểm dựa trên khía cạnh tiếng Việt. Trích xuất các loại khía cạnh nhằm phát hiện và phân loại các mục thể hiện ý kiến/quan điểm, ví dụ như các tính năng của sản phẩm/dịch vụ. Phân loại quan điểm thực hiện gán nhãn ý kiến/quan điểm, bao gồm tích cực, tiêu cực hoặc trung tính cho từng mục khía cạnh đã được xác định. Nội dung Chương 2 luận án sẽ trình bày đề xuất phương pháp trích xuất khía cạnh và phân loại quan điểm sử dụng học máy có giám sát cho tiếng Việt với việc tận dụng dữ liệu đã được gán nhãn từ

ngôn ngữ khác cùng miền lĩnh vực (trong trường hợp này sử dụng tiếng Anh). Dữ liệu từ ngôn ngữ tiếng nước ngoài sẽ được dịch sang tiếng Việt bằng một công cụ dịch tự động (Google Translate).

Chương 2 trình bày tổng hợp kết quả nghiên cứu của luận án dựa trên các công trình nghiên cứu số [4, 6] của tác giả (Theo danh mục các công trình công bố), bao gồm các nội dung sau:

- Đề xuất phương pháp trích xuất khía cạnh, quan điểm từ văn bản có ý kiến/quan điểm trong tiếng Việt, bao gồm hai nhiệm vụ con: trích xuất các loại khía cạnh và phân loại quan điểm. Phương pháp đề xuất tận dụng dữ liệu đã được gán nhãn từ ngôn ngữ khác (ví dụ, tiếng Anh) để giải quyết khó khăn cho việc thiếu tài nguyên dữ liệu của ngôn ngữ tiếng Việt. Ngoài các đặc trưng phổ biến, nghiên cứu sử dụng đặc trưng nhúng từ để giảm sự khác biệt về từ vựng giữa văn bản gốc và văn bản dịch từ ngôn ngữ tiếng nước ngoài, nhằm cải thiện hiệu quả của các quá trình trích xuất.
- Xây dựng một tập dữ liệu văn bản tiếng Việt có chú thích về các loại khía cạnh và quan điểm được trích từ các bài đánh giá về lĩnh vực nhà hàng bằng ngôn ngữ tiếng Việt, bao gồm 575 bài đánh giá với 3.796 câu.
- Thực hiện các thử nghiệm trên tập dữ liệu đã được xây dựng và phân tích kết quả thử nghiệm, chứng minh tính hiệu quả của phương pháp đề xuất.

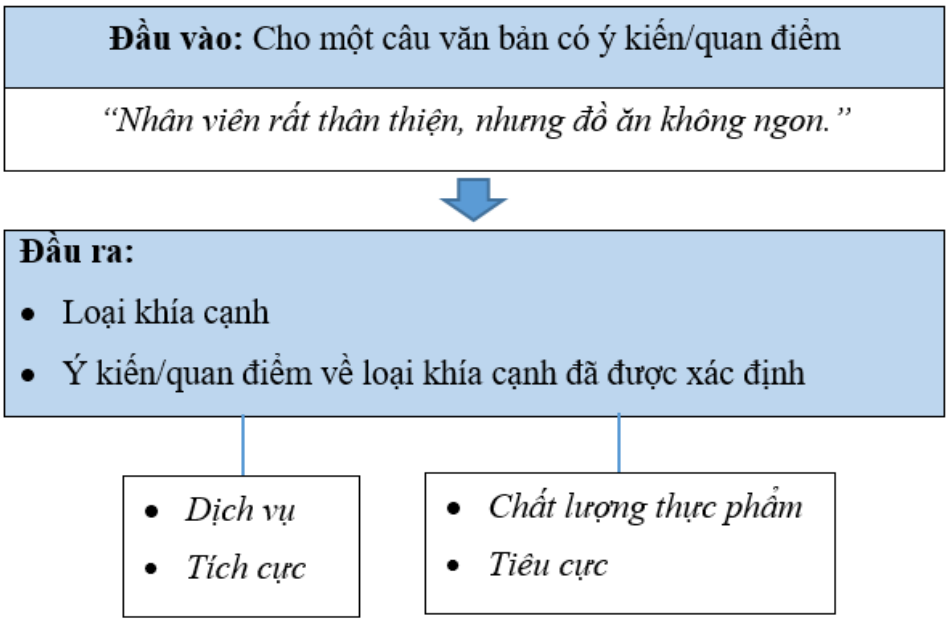
2.1. ĐẶT VẤN ĐỀ

Phân tích cảm xúc và khai phá quan điểm [67] là lĩnh vực con của phân tích văn bản có ý kiến/quan điểm trong đánh giá sản phẩm trực tuyến, mạng xã hội, blog, diễn đàn,... đã trở thành một chủ đề nghiên cứu quan trọng và được quan tâm trong xử lý ngôn ngữ tự nhiên (NLP) và khai phá dữ liệu (DM). Các hệ thống khai phá quan điểm cung cấp nhiều thông tin hữu ích cho không chỉ khách hàng mà còn cho các nhà cung cấp và sản xuất dịch vụ. Đối với khách hàng, việc biết được ý kiến/quan điểm của những người dùng khác là một điều quan trọng khi họ muốn lựa chọn sản phẩm

hoặc dịch vụ phù hợp. Đối với nhà cung cấp dịch vụ và nhà sản xuất, phân tích ý kiến/quan điểm giúp hiểu mong muốn, xu hướng của khách hàng, từ đó có thể quảng cáo sản phẩm/dịch vụ phù hợp cho khách hàng, đồng thời giúp lựa chọn chiến lược phát triển sản phẩm mới.

Nhiệm vụ được tập trung và nghiên cứu nhiều nhất trong nghiên cứu phân tích cảm xúc và khai thác quan điểm là phân loại quan điểm, đó là việc gán một nhãn (tích cực, tiêu cực hay trung tính) cho một văn bản có ý kiến/quan điểm được đưa ra (ví dụ: một câu hoặc một bài đánh giá của khách hàng). Do phân loại quan điểm chỉ xác định một ý kiến/quan điểm tổng thể cho văn bản đầu vào, nên nó sẽ không phù hợp với các tình huống phức tạp. Ví dụ với trường hợp có một nhận xét về nhà hàng như sau: "*Nhân viên rất thân thiện, nhưng đồ ăn không ngon.*". Trường hợp này có thể có khó khăn khi muốn xác định ý kiến/quan điểm của người nói vì họ hài lòng với dịch vụ của nhà hàng nhưng lại phàn nàn (không hài lòng) về chất lượng thực phẩm. Một số hệ thống phân loại quan điểm sẽ gán nhãn trung tính cho các tình huống như vậy. Tuy nhiên, thông tin này không có ý nghĩa lắm vì không thể biết được chính xác những gì họ hài lòng và/hoặc không hài lòng.

Khai phá quan điểm dựa trên khía cạnh (ABOM) [85,94,122] giải quyết hạn chế của phân loại quan điểm bằng cách nhận ra ý kiến/quan điểm theo từng khía cạnh của sản phẩm/dịch vụ được thể hiện trong văn bản có ý kiến/quan điểm. Khai phá quan điểm dựa trên khía cạnh bao gồm hai nhiệm vụ: (1) Trích xuất các loại khía cạnh, nghĩa là thực hiện xác định danh mục khía cạnh (cặp thực thể và thuộc tính), mà có một ý kiến/quan điểm được thể hiện trong văn bản; và (2) Phân loại quan điểm, nghĩa là thực hiện gán nhãn quan điểm cho từng loại khía cạnh đã được xác định trong nhiệm vụ (1). Ví dụ trong câu văn bản phía trên "*Nhân viên rất thân thiện, nhưng đồ ăn không ngon.*", thì đầu ra mong muốn của quá trình khai phá quan điểm dựa trên khía cạnh bao gồm *hai khía cạnh* (*dịch vụ* và *chất lượng thực phẩm* của nhà hàng) và *hai ý kiến/quan điểm* đối với hai khía cạnh (*tích cực* đối với *dịch vụ* và *tiêu cực* đối với *chất lượng thực phẩm*) (Hình 2.1).



Hình 2.1. Trích xuất khía cạnh và phân loại quan điểm

Các nghiên cứu hiện tại [1,47,94,122,123] về khai phá quan điểm dựa trên khía cạnh thường sử dụng học có giám sát, đây là phương pháp tiếp cận chủ yếu nhất trong xử lý ngôn ngữ tự nhiên (NLP) và khai phá dữ liệu (DM). Mặc dù các phương pháp học máy có giám sát đã được chứng minh là có hiệu quả trong các nghiên cứu trước đây [1,30,52,94,122,123], nhưng các phương pháp này thường yêu cầu cần phải chú thích thủ công một lượng lớn dữ liệu cho giai đoạn huấn luyện, việc này rất tốn kém thời gian và chi phí. Hơn nữa, các phương pháp học máy có giám sát thường phụ thuộc vào miền lĩnh vực. Việc chú thích dữ liệu huấn luyện cho tất cả các lĩnh vực trong mọi ngôn ngữ là điều không thể. Do đó, việc áp dụng các phương pháp học có giám sát vào một miền lĩnh vực mới hoặc một ngôn ngữ mới, đặc biệt là ngôn ngữ có ít tài nguyên dữ liệu được gán nhãn sẵn, vẫn còn rất nhiều khó khăn.

Nghiên cứu trình bày trong Chương 2 của luận án đề cập đến cả hai nhiệm vụ trích xuất thông tin trong khai phá quan điểm dựa trên khía cạnh cho tiếng Việt, một ngôn ngữ có ít tài nguyên cho bài toán này. Để giảm bớt sự phụ thuộc vào lượng dữ liệu gán nhãn, nghiên cứu giới thiệu một phương pháp sử dụng các tập dữ liệu đã

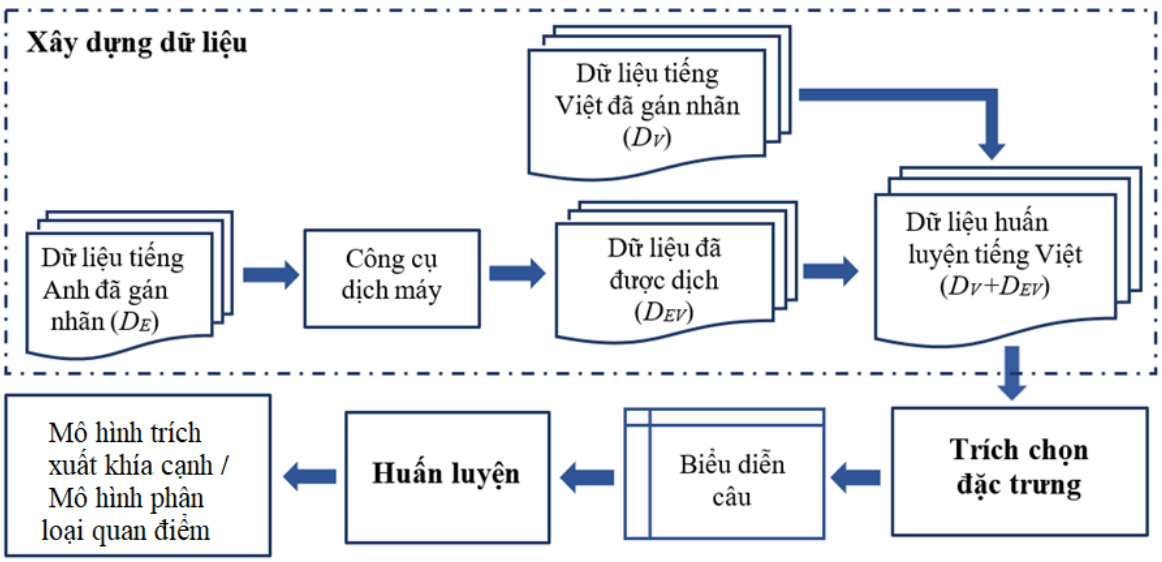
được gán nhãn sẵn từ các ngôn ngữ khác để cải thiện hiệu quả của các mô hình học có giám sát. Dữ liệu gán nhãn sẵn từ ngôn ngữ khác trước tiên được dịch tự động sang tiếng Việt và sau đó được hợp nhất với dữ liệu gán nhãn bằng tiếng Việt thành một tập dữ liệu, sẽ được sử dụng để huấn luyện mô hình trích xuất khía cạnh và mô hình phân loại quan điểm. Để làm giảm tác động tiêu cực về sự khác biệt trong cách diễn đạt và sự khác biệt về từ trong các ngôn ngữ khác nhau do quá trình thu thập dữ liệu gây ra, nghiên cứu đề xuất sử dụng kỹ thuật nhúng từ [76], trong đó các từ tương tự về ngữ nghĩa được biểu thị bằng các véc-tơ có độ tương tự gần nhau. Do đó, các đặc trưng nhúng từ có thể giúp các mô hình phân loại nhận ra từ đồng nghĩa và từ có nghĩa bao hàm một nhóm các từ có nghĩa hẹp hơn *(hypernyms)*.

Đóng góp của nghiên cứu được trình bày trong Chương 2 *là đề xuất giải pháp nâng cao hiệu quả cho trích xuất khía cạnh và phân loại quan điểm trong ngôn ngữ tiếng Việt bằng cách khai thác nguồn dữ liệu đã được gán nhãn sẵn từ ngôn ngữ khác* (tiếng Anh). Kết quả thử nghiệm trên tập dữ liệu tự xây dựng bao gồm 575 bài đánh giá với 3.796 câu cho thấy, với việc sử dụng thêm dữ liệu dịch từ tiếng Anh, phương pháp đề xuất đã cải thiện hiệu năng của hệ thống khai phá quan điểm dựa trên khía cạnh tiếng Việt trong cả hai nhiệm vụ trích xuất khía cạnh và phân loại quan điểm.

Nội dung còn lại của Chương 2 được cấu trúc như sau. Mục 2.2 giới thiệu đề xuất phương pháp trích xuất khía cạnh và phân loại quan điểm cho tiếng Việt bằng cách tận dụng nguồn dữ liệu đã được gán nhãn từ ngôn ngữ khác (giàu tài nguyên hơn). Mục 2.3 trình bày quá trình xây dựng tập dữ liệu thử nghiệm. Mục 2.4 trình bày các thực nghiệm cho phương pháp đề xuất, với hai nhiệm vụ trích xuất khía cạnh và phân loại quan điểm tiếng Việt trên tập dữ liệu văn bản đã xây dựng. Cuối cùng, Mục 2.5 trình bày kết luận chương.

2.2. ĐỀ XUẤT PHƯƠNG PHÁP TRÍCH XUẤT KHÍA CẠNH VÀ PHÂN LOẠI QUAN ĐIỂM CHO TIẾNG VIỆT

Phần này trình bày một phương pháp trích xuất thông tin cho ngôn ngữ ít tài nguyên bằng cách tận dụng nguồn dữ liệu đã được gán nhãn từ ngôn ngữ khác. Các thông tin thực hiện trích xuất bao gồm khía cạnh và quan điểm về khía cạnh từ văn bản có quan điểm tiếng Việt. Nhiệm vụ này bao gồm hai nhiệm vụ con cụ thể: (1) trích xuất các loại khía cạnh và (2) phân loại quan điểm. Đối với mỗi câu thể hiện quan điểm, nhiệm vụ trích xuất các loại khía cạnh sẽ xác định mục tiêu thể hiện quan điểm, chẳng hạn như các đặc trưng của sản phẩm hoặc dịch vụ và các loại danh mục của chúng. Trong nhiệm vụ này, phương pháp đề xuất xác định cả khía cạnh tường minh và khía cạnh ẩn. Nhiệm vụ phân loại quan điểm sẽ xác định một ý kiến/quan điểm (tích cực, tiêu cực hoặc trung tính) cho mỗi loại khía cạnh đã được xác định.



Hình 2.2. Phương pháp đề xuất cho trích xuất khía cạnh và phân loại quan điểm tiếng Việt

Đối với cả hai nhiệm vụ, giả thiết là có ba tập dữ liệu từ cùng một miền lĩnh vực: (1) tập dữ liệu tiếng Việt đã được chú thích, ký hiệu là D_V, trong đó mỗi câu thể hiện quan điểm được gán nhãn với các loại khía cạnh và loại ý kiến/quan điểm, hoặc

gán là "*Null*" nếu câu không chứa bất kỳ khía cạnh nào; (2) tập dữ liệu đã được gán nhãn bằng một ngôn ngữ khác (ở đây sử dụng tập dữ liệu đã được gán nhãn bằng tiếng Anh), ký hiệu là D_E; và (3) tập dữ liệu tiếng Việt dùng để kiểm tra, ký hiệu là T_V. Tập dữ liệu D_E được dịch sang ngôn ngữ gốc (ở đây là tiếng Việt), ký hiệu là D_{EV}, bằng một công cụ dịch tự động. Mục tiêu của bài toán là sử dụng D_{EV} cùng với D_V để cải thiện độ chính xác của việc trích xuất các thông tin về loại khía cạnh và loại quan điểm từ T_V.

Phương pháp tổng thể đề xuất để giải quyết cả hai nhiệm vụ bao gồm ba bước chính: (1) xây dựng dữ liệu huấn luyện, (2) trích chọn đặc trưng, và (3) huấn luyện mô hình trích xuất các loại khía cạnh và mô hình phân loại quan điểm. Sự khác biệt trong phương pháp đề xuất là, thay vì chỉ sử dụng tập dữ liệu tiếng Việt có chú thích D_V để huấn luyện các mô hình, nghiên cứu thực hiện bổ sung tập dữ liệu huấn luyện tiếng Việt với tập dữ liệu D_{EV}, được dịch từ tập dữ liệu tiếng Anh đã được gán nhãn D_E bằng một công cụ dịch tự động. Hình 2.2 trình bày ba bước của phương pháp đề xuất.

2.2.1. Xây dựng dữ liệu huấn luyện

Tập dữ liệu huấn luyện được xây dựng từ hai nguồn: (1) dữ liệu được gán nhãn bằng tiếng Việt và (2) dữ liệu được gán nhãn bằng tiếng nước ngoài (trong trường hợp này là tiếng Anh). Đối với tập dữ liệu tiếng Anh, nghiên cứu sử dụng công cụ dịch tự động Google Translate để dịch các câu với các loại khía cạnh và các nhãn ý kiến/quan điểm tương ứng sang tiếng Việt. Tập dữ liệu được dán nhãn D_{EV} sau đó được thêm vào tập dữ liệu tiếng Việt D_V được chú thích để tạo thành tập dữ liệu huấn luyện mới ($D_V + D_{EV}$). Trên thực tế, việc sử dụng phương pháp dịch thủ công sẽ mang lại chất lượng dịch tốt hơn cho tập dữ liệu huấn luyện, nhưng việc này rất tốn kém thời gian và chi phí. Hơn nữa, dịch thủ công cũng khó áp dụng khi muốn sử dụng các tập dữ liệu được gán nhãn từ các ngôn ngữ khác nhau.

2.2.2. Trích chọn đặc trưng

Trước khi trích chọn đặc trưng, tất cả các câu tiếng Việt, trong cả tập dữ liệu gốc và dữ liệu dịch, đều được phân đoạn thành các từ tiếng Việt [119]. Phân đoạn từ là một bước tiền xử lý quan trọng trong hầu hết các bài toán xử lý ngôn ngữ tự nhiên cho tiếng Việt do đặc trưng mỗi từ tiếng Việt có thể có một hoặc nhiều âm tiết. Các âm tiết trong một từ tiếng Việt được phân cách bằng dấu khoảng trắng. Phần sau đây sẽ trình bày các phương pháp trích chọn đặc trưng cho cả nhiệm vụ trích xuất các loại khía cạnh và phân loại quan điểm.

1) Trích xuất các loại khía cạnh

Để trích xuất các loại khía cạnh, nghiên cứu thực hiện trích chọn hai loại đặc trưng là các đặc trưng cơ bản và các đặc trưng nhúng từ, để biểu diễn các câu văn bản.

a) Đặc trưng cơ bản

Các đặc trưng cơ bản được sử dụng là *n*-grams tiếng Việt, bao gồm unigrams, bigrams, và trigrams được trích chọn trong các câu tiếng Việt đã được phân đoạn thành các từ tiếng Việt. Ví dụ trong câu: "*Đồ_ăn rất ngon*", *n*-grams (unigrams, bigrams, và trigrams) được trích chọn như sau: *đồ_ăn, rất, ngon, đồ_ăn rất, rất ngon, đồ_ăn rất ngon*. Mặc dù các đặc trưng cơ bản tương đối đơn giản nhưng chúng đã được chứng minh là có hiệu quả đối với hầu hết các nhiệm vụ xử lý ngôn ngữ tự nhiên trong tiếng Việt.

b) Đặc trưng nhúng từ

Khi so sánh các từ giữa dữ liệu tiếng Việt (D_V) và dữ liệu được dịch (D_{EV}) (từ tiếng Anh), kết quả cho thấy có ít hơn 50% số lượng từ vựng trùng nhau. Điều này có thể dễ dàng thấy được là do người ở các quốc gia khác nhau thường dùng từ và thể hiện ý tưởng theo những phong cách khác nhau. Ví dụ, có một câu trong tiếng Anh như sau: "*The food was not great, the waiters were rude.*", khi dịch ra tiếng Việt: "*Thức_ăn không tuyệt_vời và bồi_bàn thô_lỗ.*". Cũng nội dung này, trong ngôn ngữ

tiếng Việt mọi người thường nói: "*Thức_ăn không ngon và phục_vụ bàn hơi mất lịch_sự.*". Như vậy, để đạt được hiệu quả tốt hơn, cần giảm bớt sự khác biệt về từ vựng giữa văn bản gốc và văn bản dịch bằng cách trích xuất một số đặc đặc trưng có khả năng nắm bắt được sự giống nhau và mối quan hệ giữa các từ. Vì lý do này, nghiên cứu đề xuất bổ sung thêm đặc trưng nhúng từ cùng với các đặc trưng cơ bản nhằm làm tăng hơn nữa hiệu quả của quá trình trích xuất các loại khía cạnh.

Nhúng từ [76,131], một bước đột phá quan trọng trong nghiên cứu về xử lý ngôn ngữ tự nhiên, là một loại kỹ thuật biểu diễn từ phân phối trong đó các từ được ánh xạ vào không gian véc-tơ có số chiều thấp. So với phương pháp biểu diễn từ truyền thống, nghĩa là biểu diễn theo dạng *one-hot*, phương pháp nhúng từ có hai ưu điểm chính. Thứ nhất, trong khi các véc-tơ *one-hot* có số chiều lớn và thưa, thì các véc-tơ nhúng từ có số chiều thấp và dày đặc, do đó véc-tơ nhúng từ hiệu quả hơn trong việc biểu diễn và tính toán. Ưu điểm thứ hai, và quan trọng hơn, là nhúng từ có khả năng khái quát hóa do các từ giống nhau về mặt ngữ nghĩa được biểu diễn bằng các điểm gần nhau (các véc-tơ tương tự nhau) trong không gian véc-tơ. Trong số các phương pháp tạo nhúng từ từ một kho lớn các văn bản thuần túy, phương pháp phổ biến được sử dụng đó là giảm kích thước trên ma trận đồng xuất hiện từ và mạng nơ-ron [76,131]. Nhúng từ có thể được sử dụng trực tiếp như là một đặc trưng trong bài toán phân loại câu/văn bản hoặc đóng vai trò là lớp đầu tiên trong một kiến trúc học sâu [3].

Cho v là một hàm ánh xạ một từ w thành biểu diễn véc-tơ từ $v(w)$ của nó, đặc trưng nhúng từ của một câu s được biểu diễn dưới dạng một chuỗi các từ $(w_1, w_2, \dots, w_{|s|})$ *(*có thể được tính là tổng phần tử của các véc-tơ từ:

$$v_{em}(s) = \sum_{i=1}^{|s|} v(w_i) \tag{2.1}$$

trong đó $|s|$ ký hiệu là độ dài của câu s.

c) Ghép nối đặc trưng

Một phương pháp đơn giản được sử dụng để kết hợp các đặc trưng cơ bản và các đặc trưng nhúng từ của một câu s là ghép nối hai loại đặc trưng với nhau như sau:

- Biểu diễn s bằng một véc-tơ one-hot $v_{oh}(s)$ của *n*-grams;
- Biểu diễn s bởi một véc-tơ nhúng từ $v_{em}(s)$; và
- Ghép nối hai véc-tơ $v_{oh}(s)$ và $v_{em}(s)$.

2) Phân loại quan điểm

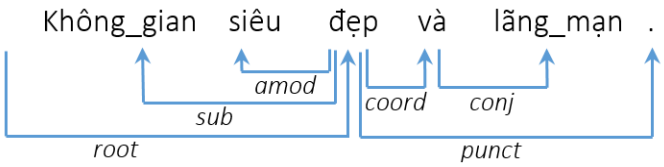
Với nhiệm vụ phân loại quan điểm, ba loại đặc trưng được sử dụng bao gồm: đặc trưng từ quan trọng, đặc trưng nhúng từ, và đặc trưng loại khía cạnh.

a) Đặc trưng từ quan trọng

Mỗi câu có thể có một vài loại khía cạnh. Do đó, để có được các đặc trưng có ý nghĩa hơn cho huấn luyện mô hình phân loại quan điểm tương ứng với từng loại khía cạnh được xác định, nghiên cứu thực hiện trích xuất các từ quan trọng đối với mỗi loại khía cạnh trong câu. Các từ quan trọng được định nghĩa là những từ chỉ ra loại khía cạnh và có thể được xác định thông qua các hệ số SVM *(Support Vector Machine)* khi huấn luyện mô hình trích xuất cho loại khía cạnh đó. Giá trị tuyệt đối của các hệ số SVM cho thấy mức độ quan trọng của các đặc trưng (nghĩa là mức độ quan trọng của các từ tương ứng) [42]. Đầu tiên chọn *k* từ (gọi là các từ hạt giống) có giá trị tuyệt đối của các hệ số SVM cao nhất, trong đó *k* là một tham số. Sau khi khảo sát và thực hiện một số thử nghiệm, *k* được chọn trong các báo cáo kết quả thực nghiệm của nghiên cứu ở đây là 5. Sau đó, để trích xuất các từ thể hiện ý kiến/quan điểm, nghiên cứu thực hiện mở rộng tập hợp các từ bằng cách sử dụng cây phụ thuộc *(dependency tree)* để chọn ra các từ liên quan. Các từ có mối quan hệ với một từ hạt giống, nghĩa là cha hoặc con, sẽ được chọn.

Hình 2.3 biểu diễn cây phụ thuộc của câu "*Không_gian siêu đẹp và lãng_mạn.*". Một mối quan hệ được thể hiện bằng một liên kết được định hướng từ từ cha đến từ con theo hướng mũi tên và loại quan hệ được hiển thị trên liên kết. Ví dụ, "*không_gian*" có mối quan hệ *sub* (subject - chủ ngữ) với "*đẹp*", hay "*siêu*" có

mối quan hệ *amod* (adjectival modifier – bổ nghĩa tính từ) với "*đẹp*". Giả sử "*không_gian*" là một từ hạt giống được trích xuất trong bước đầu tiên, thì từ "*đẹp*" sẽ được chọn trong bước mở rộng.



Hình 2.3. Một ví dụ của cây phụ thuộc

a) Đặc trưng nhúng từ

Đặc trưng nhúng từ được trích xuất theo cách tương tự như trong nhiệm vụ trích xuất khía cạnh. Tuy nhiên, các từ được chọn sử dụng ở phần này chỉ bao gồm các từ quan trọng thay vì tất cả các từ trong mỗi câu.

a) Đặc trưng loại khía cạnh

Con người có thể sử dụng các diễn đạt tương tự khi bình luận về các loại khía cạnh khác nhau cùng với các ý kiến/quan điểm khác nhau. Do đó, một cụm từ có thể thể hiện quan điểm tích cực về một loại khía cạnh nhưng lại thể hiện quan điểm tiêu cực về một loại khía cạnh khác. Ví dụ, xem xét hai câu sau: 1) "*Kim chi không cay mà lại hơi ngọt.*", 2) "*Đồ uống có vị hơi ngọt.*". Cụm từ "*hơi ngọt*" thể hiện sự không hài lòng (tiêu cực) về thực phẩm (kim chi) nhưng lại thể hiện sự hài lòng (tích cực) về đồ uống. Bởi lý do này, loại khía cạnh được khai thác như một loại đặc trưng cho nhiệm vụ phân loại quan điểm.

2.2.3. Các mô hình huấn luyện

Cho *N* là số lượng các loại khía cạnh muốn trích xuất. Đối với nhiệm vụ *trích xuất các loại khía cạnh*, nghiên cứu thực hiện huấn luyện từng bộ phân loại nhị phân cho từng loại khía cạnh để dự đoán xem một câu có chứa danh mục khía cạnh hay không. Đối với nhiệm vụ *phân loại quan điểm*, nghiên cứu thực hiện huấn luyện một

bộ phân loại để dự đoán ý kiến/quan điểm cho một loại khía cạnh đã được xác định trong câu. Như vậy, có *N* bộ phân loại cho *N* loại khía cạnh và một bộ phân loại để xác định loại quan điểm.

Mặc dù ở đây có thể sử dụng bất kỳ thuật toán học có giám sát nào, nhưng nghiên cứu lựa chọn Support Vector Machines (SVM) [116] do thuật toán này đã được chứng minh là rất hiệu quả với nhiều nhiệm vụ phân loại khác nhau trong lĩnh vực xử lý ngôn ngữ tự nhiên [30,33,52,94]. (Nội dung giới thiệu tóm tắt về SVM có tại Mục 1.3.1.2).

Nghiên cứu thực hiện các thực nghiệm với mô hình SVM tuyến tính (Linear SVM), sử dụng các tham số mặc định như sau: *penalty='l2', loss='squared_hinge', C=1.0.*

2.3. XÂY DỰNG TẬP DỮ LIỆU

Để kiểm tra tính hiệu quả của phương pháp được đề xuất, nghiên cứu đã xây dựng một tập dữ liệu tiếng Việt có chú thích và tiến hành một loạt các thực nghiệm. Tập dữ liệu tiếng Anh được tận dụng làm dữ liệu bổ sung được trích xuất từ nhiệm vụ 5 trong SemEval-2016 [94]. Phần sau sẽ mô tả hai tập dữ liệu và các kết quả thực nghiệm cũng như phân tích.

a) Tập dữ liệu tiếng Việt

Để xây dựng tập dữ liệu tiếng Việt, nghiên cứu thực hiện theo 3 bước sau: thu thập dữ liệu thô, tiền xử lý và gán nhãn dữ liệu.

1) Thu thập dữ liệu thô: Dữ liệu thô được thu thập từ trang web Foody (có tại: https://www.foody.vn/). Đây là một trang web lớn và phổ biến nhất Việt Nam, nơi người dùng có thể tìm kiếm, đánh giá/bình luận và đặt hàng, không chỉ riêng với đồ ăn uống mà còn có cả dịch vụ du lịch, làm đẹp và các dịch vụ khác. Dữ liệu thô được trích xuất các bài đánh giá từ một số nhà hàng ở Việt Nam (hầu hết ở Hà Nội và thành phố Hồ Chí Minh).

2) Tiền xử lý dữ liệu: Một số bước tiền xử lý dữ liệu được tiến hành trên tập dữ liệu thô thu thập được, bao gồm làm sạch dữ liệu, thực hiện phân tách câu. Kết quả thu được tập dữ liệu bao gồm 575 bài đánh giá với 3.796 câu tiếng Việt.

3) Gán nhãn dữ liệu: Dữ liệu sau khi tiền xử lý sẽ được gán nhãn theo loại khía cạnh và loại quan điểm tương ứng. Đầu tiên phân công hai người thực hiện việc gán nhãn. Sau đó, một người thứ ba sẽ thực hiện kiểm tra, so khớp dữ liệu đã được gán nhãn. Nếu hai người gán nhãn trước bất đồng ý kiến về một nhãn được gán, thì người thứ ba sẽ đưa ra quyết định cuối cùng. Những người thực hiện gán nhãn là sinh viên chuyên ngành Công nghệ thông tin của Học viện Công nghệ Bưu chính Viễn thông (hai sinh viên đại học và một học viên sau đại học) có kiến thức nền tảng cơ bản về xử lý ngôn ngữ tự nhiên và học máy. Nghiên cứu sử dụng hệ số *Kappa* của Cohen [27] để đo mức độ tương đồng ý kiến giữa các nhãn được gán:

$$Kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{2.2}$$

trong đó $\Pr(a)$ là độ tương đồng ý kiến giữa hai người chú thích, và $\Pr(e)$ là xác suất giả thiết có ý kiến khác nhau. Nghiên cứu đã tính toán hệ số *Kappa* cho 2 loại nhãn, bao gồm loại khía cạnh và loại quan điểm. Do mỗi câu có thể được gán với nhiều nhãn khía cạnh và quan điểm, nên ở đây hệ số *Kappa* được tính riêng cho từng loại khía cạnh và loại quan điểm, sau đó sẽ tính kết quả trung bình. Hệ số *Kappa* đo được của tập dữ liệu tiếng Việt đã thực hiện gán nhãn là 0,83 cho loại khía cạnh và 0,86 cho loại quan điểm. Các giá trị này được coi là có độ tương đồng ý kiến khá tốt [27].

Tương tự như dữ liệu trong cuộc thi SemEval-2016 Task 5 [94], nghiên cứu ở đây quan tâm đến 12 loại khía cạnh được đại diện bởi 12 bộ (thực thể, thuộc tính). Hình 2.4 trình bày một ví dụ gán nhãn cho một bài đánh giá trong tập dữ liệu tiếng Việt thu thập được. Câu đầu tiên bình luận về chất lượng và giá cả của đồ ăn (loại khía cạnh FOOD#QUALITY và FOOD#PRICES), trong đó quan điểm về chất lượng là tích cực *(positive)*, còn quan điểm về giá cả là tiêu cực *(negative)*. Câu thứ hai bình

luận về thái độ phục vụ của nhân viên (loại khía cạnh SERVICE#GENERAL) với quan điểm tích cực. Câu cuối cùng nhận xét về không gian (loại khía cạnh AMBIENCE#GENERAL) với quan điểm tiêu cực.

```
<Review rid="bc8">
    <Sentence id="bc8:0">
        <Text> Đồ_ăn ngon nhưng khá đắt . </Text>
        <Opinions>
            <Opinion category="FOOD#QUALITY" polarity="positive"/>
            <Opinion category="FOOD#PRICES" polarity="negative"/>
        </Opinions>
    </Sentence>
    <Sentence id="bc8:1">
        <Text> Nhân_viên lịch_sự , phục_vụ nhanh . </Text>
        <Opinions>
            <Opinion category=" SERVICE#GENERAL " polarity="positive"/>
        </Opinions>
    </Sentence>
    <Sentence id="bc8:2">
        <Text> Tuy_nhiên , không_gian ở đây hơi chật_chội . </Text>
        <Opinions>
            <Opinion category=" AMBIENCE#GENERAL" polarity="negative"/>
        </Opinions>
    </Sentence>
</Review>
```

Hình 2.4. Các câu trong một bài đánh giá được gán nhãn trong tập dữ liệu tiếng Việt

b) Tập dữ liệu tiếng Anh

Nghiên cứu sử dụng tập dữ liệu tiếng Anh về lĩnh vực nhà hàng (cả dữ liệu huấn luyện và dữ liệu kiểm tra) từ SemEval-2016 Task 5 [94] làm nguồn dữ liệu bổ sung cho phương pháp đề xuất. Tập dữ liệu tiếng Anh bao gồm 440 bài đánh giá với 2.676 câu. Như vậy tổng cộng có 1.015 bài đánh giá với 6.472 câu cho cả bộ dữ liệu tiếng Việt và tiếng Anh. Bảng 2.1 trình bày thông tin thống kê trên cả 2 tập dữ liệu.

Bảng 2.1. Thông tin thống kê trên hai tập dữ liệu

	Tiếng Việt	**Tiếng Anh**	**Tổng số**
Số bài đánh giá	575	440	1.015
Số câu	3.796	2.676	6.472

Bảng 2.2 trình bày chi tiết 12 loại khía cạnh và quan điểm của chúng (tích cực, tiêu cực, trung tính) trong tập dữ liệu tiếng Việt và tiếng Anh. Trong cả 2 tập dữ liệu, loại khía cạnh có tần số xuất hiện nhiều nhất là FOOD#QUALITY (với 1.357 và 1.162 lần tương ứng với bộ dữ liệu tiếng Việt và tiếng Anh), trong khi các loại khía cạnh có tần số xuất hiện thấp nhất là DRINKS#STYLE_OPTIONS và DRINKS#PRICES (với dưới 80 lần cho mỗi loại khía cạnh). Và có một điều cần lưu ý là số lượng mẫu có quan điểm tích cực thường nhiều hơn số lượng mẫu có quan điểm tiêu cực trong cả hai tập dữ liệu.

Bảng 2.2. Loại khía cạnh và quan điểm tương ứng trên hai tập dữ liệu

Loại khía cạnh	**Tiếng Việt**				**Tiếng Anh**			
	Tích cực	***Tiêu cực***	***Trung tính***	***Tổng***	***Tích cực***	***Tiêu cực***	***Trung tính***	***Tổng***
RESTAURANT#GENERAL	178	30	25	**233**	420	135	9	**564**
RESTAURANT#PRICES	71	44	17	**132**	40	53	8	**101**
RESTAURANT#MISCELLANEOUS	157	22	15	**194**	74	40	17	**131**
FOOD#QUALITY	957	247	153	**1.357**	886	235	41	**1.162**
FOOD#STYLE_OPTIONS	475	80	31	**586**	114	60	18	**192**
FOOD#PRICES	115	76	16	**207**	47	62	4	**113**
DRINKS#QUALITY	213	56	38	**307**	61	6	2	**69**
DRINKS#STYLE_OPTIONS	56	14	5	**75**	40	4	0	**44**
DRINKS#PRICES	29	19	8	**56**	13	11	0	**24**
SERVICE#GENERAL	345	100	42	**487**	283	302	19	**604**
AMBIENCE#GENERAL	371	92	53	**516**	258	44	19	**321**
LOCATION#GENERAL	95	101	19	**215**	32	1	8	**41**
Tổng	**3.062**	**881**	**422**	**4.365**	**2.268**	**953**	**145**	**3.366**

2.4. THỰC NGHIỆM VÀ PHÂN TÍCH KẾT QUẢ

Phần này trình bày thực nghiệm và phân tích kết quả cho trích xuất thông tin khía cạnh, quan điểm trong ngôn ngữ có ít tài nguyên (tiếng Việt) từ văn bản có quan điểm, bao gồm hai nhiệm vụ con: (1) trích xuất các loại khía cạnh và (2) phân loại quan điểm.

2.4.1. Thiết lập thực nghiệm

Tập dữ liệu tiếng Việt được chia ngẫu nhiên thành 10 phần và được tiến hành kiểm tra chéo *(cross-validation)*. Hiệu năng của các mô hình trích xuất khía cạnh và phân loại quan điểm được đo bởi độ chính xác *(precision)*, độ phủ *(recall)* và độ đo F_1 *(F_1 score)* trên mỗi loại khía cạnh và mỗi loại quan điểm (tích cực hoặc tiêu cực).

Với bài toán phân loại quan điểm, có thể thực hiện phân thành 3 loại ý kiến/quan điểm (tích cực, tiêu cực, trung lập), hoặc 2 loại ý kiến/quan điểm (tích cực và tiêu cực, như trong [5, 33, 89, 92]). Mục tiêu của Chương 2 luận án là nghiên cứu giải pháp nâng cao hiệu quả cho trích xuất khía cạnh và phân loại quan điểm trong ngôn ngữ tiếng Việt bằng cách khai thác nguồn dữ liệu đã được gán nhãn sẵn từ ngôn ngữ khác. Để thực hiện các thực nghiệm, nghiên cứu lựa chọn sử dụng bài toán phân lớp quan điểm với 2 loại ý kiến/quan điểm (tích cực, tiêu cực) làm minh chứng cho giải pháp đề xuất do việc này không làm ảnh hưởng đến mục tiêu nghiên cứu.

Trích xuất các loại khía cạnh: Xét khía cạnh RESTAURANT#GENERAL (nhận xét chung về nhà hàng) làm ví dụ. Các độ đo *độ chính xác*, *độ phủ*, và độ đo F_1 cho loại khía cạnh này được tính theo các công thức (1.17), (1.18) và (1.19) trong Mục 1.3.4 Chương 1. Trong đó, A biểu thị cho tập các câu có khía cạnh RESTAURANT#GENERAL được xác định bởi mô hình, và B biểu thị cho tập các câu có khía cạnh này được gán nhãn bởi người chú thích. Các độ đo này được xác định tương tự cho các loại khía cạnh khác.

Phân loại quan điểm: Xét khía cạnh RESTAURANT#GENERAL. Xét loại quan điểm *tích cực (positive)* của khía cạnh này làm ví dụ. *Độ chính xác*, *độ phủ* và

độ đo F_1 cho loại quan điểm *tích cực* của khía cạnh RESTAURANT#GENERAL được tính theo các công thức (1.17), (1.18) và (1.19) trong Mục 1.3.4 Chương 1. Trong đó, *A* biểu thị tập các câu có khía cạnh RESTAURANT#GENERAL tích cực được xác định bởi mô hình, và *B* biểu thị tập các câu có khía cạnh RESTAURANT#GENERAL tích cực được xác định bởi người chú thích. Các độ đo này được xác định tương tự cho loại quan điểm *tiêu cực (negative)* của khía cạnh RESTAURANT#GENERAL, và tương tự cho các khía cạnh khác cần trích xuất.

Trong quá trình thực hiện thực nghiệm, tất cả các câu trong tập dữ liệu tiếng Anh được dịch sang tiếng Việt bằng công cụ Google Translate *(https://translate.google.com/)*. Với đặc trưng nhúng từ, nghiên cứu sử dụng các véc-tơ từ 50 chiều được huấn luyện với Word2Vec [76] (dùng phần mềm có tại *https://code.google.com/archive/p/word2vec/*) trên một tập dữ liệu văn bản từ Baomoi *(https://baomoi.com/)* với hơn 433.000 từ tiếng Việt.

2.4.2. Triển khai các mô hình thực nghiệm

Với cả hai nhiệm vụ, trích xuất các loại khía cạnh và phân loại quan điểm, nghiên cứu tiến hành thực nghiệm nhằm so sánh hiệu năng của ba mô hình, đó là *Cơ sở* (*baseline), CRL (Cross-Language),* và *WEmb (Word Embedding).* Bảng 2.3 trình bày tóm tắt các mô hình thực nghiệm. Cả ba mô hình này đều được huấn luyện với SVM tuyến tính (dùng phần mềm có tại *https://www.csie.ntu.edu.tw/~cjlin/libsvm/*.) [22].

- **Mô hình Cơ sở:** Mục đích của mô hình này tập trung xem xét hiệu quả của phương pháp được đề xuất khi chỉ sử dụng bộ dữ liệu tiếng Việt với các đặc trưng cơ bản, do vậy, thực nghiệm với mô hình này chỉ sử dụng bộ dữ liệu tiếng Việt. Tập đặc trưng cho trích xuất các loại khía cạnh là *n*-grams, và tập đặc trưng cho phân loại quan điểm là đặc trưng từ quan trọng và đặc trưng danh mục khía cạnh.

- **Mô hình CRL (Cross-Language Model):** Mục đích của mô hình này tập trung xem xét ảnh hưởng về hiệu năng của việc sử dụng dữ liệu dịch bổ sung từ một ngôn ngữ khác đối với các nhiệm vụ trích xuất các loại khía cạnh và phân loại quan điểm. Do vậy, thực nghiệm với mô hình này sử dụng cả hai tập dữ liệu tiếng Việt và tập dữ liệu dịch từ tiếng Anh để huấn luyện mô hình. Các tập đặc trưng sử dụng cho cả hai nhiệm vụ tương tự như trong mô hình cơ sở.
- **Mô hình WEmb (Word Embedding Model):** Mục đích của mô hình này tập trung xem xét ảnh hưởng về hiệu năng của đặc trưng nhúng từ với các mô hình trích xuất các loại khía cạnh và phân loại quan điểm, do vậy mô hình này được xây dựng tương tự với mô hình CRL nhưng bổ sung thêm đặc trưng nhúng từ.

Bảng 2.3. Các mô hình thực nghiệm

Mô hình	**Dữ liệu**	**Đặc trưng trích chọn**	
		Trích xuất các loại khía cạnh	***Phân loại quan điểm***
Cơ sở	Tiếng Việt	*n*-grams	Từ quan trọng + Loại khía cạnh
CRL	Tiếng Việt + Tiếng Anh	*n*-grams	Từ quan trọng + Loại khía cạnh
WEmb	Tiếng Việt + Tiếng Anh	*n*-grams + Nhúng từ	Từ quan trọng + Nhúng từ + Loại khía cạnh

2.4.3. Kết quả thực nghiệm và phân tích

a) Kết quả trích xuất các loại khía cạnh

Trước tiên, nghiên cứu tiến hành thực nghiệm với mô hình cơ sở, tập trung vào ba tập đặc trưng: 1) unigrams; 2) unigrams và bigrams; và 3) unigrams, bigrams và trigrams. Kết quả cho thấy với tập đặc trưng thứ hai, sử dụng unigrams và bigrams, là tốt nhất. Do đó, báo cáo kết quả luận án sẽ trình bày kết quả thực nghiệm với tập đặc trưng này, và sẽ bỏ qua các tập đặc trưng còn lại.

Bảng 2.4. Kết quả thực nghiệm trích xuất các loại khía cạnh với mô hình cơ sở

STT	Loại khía cạnh	Độ chính xác (%)	Độ phủ (%)	F_1 (%)
1	RESTAURANT#GENERAL	53,97	35,53	42,20
2	RESTAURANT#PRICES	52,57	33,53	40,39
3	RESTAURANT#MISCELLANEOUS	76,85	57,65	64,17
4	FOOD#QUALITY	82,95	78,42	80,40
5	FOOD#STYLE_OPTIONS	79,82	64,51	69,32
6	FOOD#PRICES	50,55	30,61	36,85
7	DRINKS#QUALITY	72,65	56,03	62,20
8	DRINKS#STYLE_OPTIONS	15,00	4,11	5,86
9	DRINKS#PRICES	33,33	11,00	16,27
10	SERVICE#GENERAL	93,32	76,00	83,52
11	AMBIENCE#GENERAL	89,43	72,54	79,58
12	LOCATION#GENERAL	90,07	68,26	76,90
	Trung bình	**78,00**	**64,52**	**70,62**

Bảng 2.4 trình bày kết quả thực nghiệm trích xuất các loại khía cạnh của mô hình cơ sở sử dụng các đặc trưng unigrams và bigrams với tập dữ liệu tiếng Việt. Các loại khía cạnh trích xuất đạt được độ chính xác cao nhất với độ đo F_1 khoảng 80%, bao gồm SERVICE#GENERAL (83,52%), FOOD#QUALITY (80,40%), và AMBIENCE#GENERAL (79,58%). Cả ba loại khía cạnh này có tần số xuất hiện khá thường xuyên trong tập dữ liệu: 487, 1357 và 516 lần tương ứng với SERVICE#GENERAL, FOOD#QUALITY, và AMBIENCE#GENERAL. Các loại khía cạnh có độ đo F_1 thấp nhất là DRINKS#STYLE_OPTIONS (5,86%) và DRINKS#PRICES (16,27%). Các loại khía cạnh này xuất hiện tương đối ít trong tập dữ liệu, với 75 lần cho loại khía cạnh DRINKS#STYLE_OPTIONS và 56 lần cho loại khía cạnh DRINKS#PRICES. Loại khía cạnh LOCATION#GENERAL là một trường hợp khá đặc biệt. Mặc dù khía cạnh này có tần số xuất hiện trong các câu tương đối ít, nhưng kết quả trích xuất lại có F_1 tương đối cao là 76,90%. Lý do có thể

là chỉ có một khía cạnh liên quan đến vị trí và các câu mô tả vị trí thường chứa một số cụm từ đặc biệt như "vị trí", "nằm tại", "trung tâm", "phố" và "đường". Tính trung bình, mô hình cơ sở đạt được độ chính xác là 78,00%, độ phủ là 64,52% và độ đo F_1 là 70,62%.

Bảng 2.5. Kết quả trích xuất các loại khía cạnh của các mô hình đề xuất (tính theo % độ đo F_1)

STT	Loại khía cạnh	Mô hình cơ sở	CRL	WEmb
1	RESTAURANT#GENERAL	42,20	52,66	53,48
2	RESTAURANT#PRICES	40,39	48,97	49,83
3	RESTAURANT#MISCELLANEOUS	64,17	61,84	61,19
4	FOOD#QUALITY	80,40	80,47	80,99
5	FOOD#STYLE_OPTIONS	69,32	68,90	68,45
6	FOOD#PRICES	36,85	39,98	37,21
7	DRINKS#QUALITY	62,20	61,38	63,18
8	DRINKS#STYLE_OPTIONS	5,86	10,86	11,81
9	DRINKS#PRICES	16,27	16,52	24,21
10	SERVICE#GENERAL	83,52	84,30	84,78
11	AMBIENCE#GENERAL	79,58	80,46	81,59
12	LOCATION#GENERAL	76,90	77,09	79,76
	Trung bình	**70,62**	**71,77**	**72,33**

Sau đó, nghiên cứu tiến hành các thực nghiệm nhằm so sánh các mô hình đề xuất với mô hình cơ sở. Bảng 2.5 trình bày kết quả thực nghiệm của các mô hình cơ sở, CRL và WEmb theo độ đo F_1. So với mô hình cơ sở, mô hình CRL đạt được 9/12 danh mục khía cạnh cao hơn, cho thấy hiệu quả của việc sử dụng dữ liệu dịch bổ sung cho trích xuất khía cạnh. Các danh mục có kết quả cải thiện tốt như RESTAURANT#GENERAL (10,46%), RESTAURANT#PRICES (8,58%), DRINKS#STYLE_OPTIONS (5,00%), và FOOD#PRICES (3,13%). Tính trung

bình, mô hình CRL đạt được độ đo F_1 là 71,77%, cải thiện hơn 1,15% so với mô hình cơ sở.

Bằng cách thêm các đặc trưng nhúng từ, WEmb đạt được kết quả với 9/12 loại khía cạnh tốt hơn so với mô hình CRL. Một số kết quả cao hơn đáng kể tính theo độ đo F_1 bao gồm DRINKS#PRICES (7,69%), LOCATION#GENERAL (2,67%), DRINKS#QUALITY (1,8%) và AMBIENCE#GENERAL (1,13%). Tính trung bình, mô hình WEmb có độ đo F_1 là 72,33%, cải tiến hơn 1,71% và 0,56% so với mô hình cơ sở và mô hình CRL tương ứng.

b) Kết quả phân loại quan điểm

Để trích chọn đặc trưng từ quan trọng cho nhiệm vụ phân loại quan điểm, số các từ hạt giống được chọn ban đầu là *k* từ. Đây là các từ có giá trị tuyệt đối của các hệ số SVM cao nhất (thể hiện mức độ quan trọng của từ) khi huấn luyện mô hình trích xuất loại khía cạnh (xem Mục 2.2.2). Nếu chọn số từ hạt giống quá ít, sẽ không bao phủ được đặc trưng của câu, ngược lại, nếu chọn số từ hạt giống quá nhiều, sẽ dẫn đến việc mở rộng số từ thành toàn bộ câu ban đầu (sẽ không có ý nghĩa từ quan trọng nữa). Sau khi khảo sát và thực hiện một số thử nghiệm, chúng tôi chọn $k=5$ trong các báo cáo kết quả thực nghiệm nghiên cứu ở đây.

Bảng 2.6. Kết quả thực nghiệm phân loại quan điểm (với k=5 từ)

Loại quan điểm	**Mô hình cơ sở**			**CLR**			**WEmb**		
	***Pre.* (%)**	***Rec.* (%)**	**F_1 (%)**	***Pre.* (%)**	***Rec.* (%)**	**F_1 (%)**	***Pre.* (%)**	***Rec.* (%)**	**F_1 (%)**
Tích cực	82,97	80,09	81,45	80,99	86,12	83,43	81,55	85,93	**83,63**
Tiêu cực	45,23	50,27	47,33	49,52	47,56	48,20	50,86	50,17	**50,19**

Bảng 2.6 trình bày kết quả thực nghiệm nhiệm vụ phân loại quan điểm với cả ba mô hình. Quan sát đầu tiên nhận thấy được đó là với tất cả các mô hình, độ đo F_1 của nhãn tích cực (positive) cao hơn nhiều so với nhãn tiêu cực (negative): 81,45%

so với 47,33% (mô hình cơ sở), 83,43% so với 48,20% (mô hình CRL) và 83,63% so với 50,19% (mô hình WEmb). Một lý do có thể được giải thích ở đây là do số lượng các mẫu tích cực trong các tập dữ liệu đều cao hơn nhiều so với số lượng các mẫu tiêu cực, tổng số có 5330 mẫu tích cực và 1834 mẫu tiêu cực. Một lý do khác có thể có là ý kiến/quan điểm tích cực thường được nêu trực tiếp và rõ ràng, trong khi ý kiến/quan điểm tiêu cực thường tiềm ẩn. Ví dụ, khá dễ dàng để xác định quan điểm tích cực trong câu "*Chúng tôi rất hài lòng với thức ăn của nhà hàng này.*" vì câu này chứa một dấu hiệu tích cực rất rõ ràng qua từ "*hài lòng*". Tuy nhiên, việc xác định quan điểm tiêu cực khó khăn hơn nhiều trong các câu sau "*Chúng tôi phải đợi thức ăn khoảng nửa tiếng*.", hay "*Kim chi không cay mà lại hơi ngọt.*". Các ý kiến/quan điểm ở đây đều được thể hiện ở dạng ẩn (không tường minh).

Quan sát thứ hai là hai mô hình được đề xuất, CRL và WEmb, vượt trội hơn so với mô hình cơ sở trên cả hai lớp tích cực và tiêu cực. Đối với mô hình CRL, kết quả đạt được độ đo F_1 là 83,43% (đối với nhãn tích cực) và 48,20% (đối với nhãn tiêu cực), cải thiện 1,98% (tích cực) và 0,87% (tiêu cực). Bằng cách thêm các đặc trưng nhúng từ, mô hình WEmb có độ đo F_1 cao nhất trên cả hai nhãn với 83,63% (tích cực) và 50,19% (tiêu cực), cải thiện 2,18% và 2,86% so với mô hình cơ sở. Kết quả cho thấy hiệu quả của việc sử dụng bộ dữ liệu dịch từ tiếng Anh và các đặc trưng nhúng từ cho phân loại quan điểm tiếng Việt.

Trong các thực nghiệm ở trên, một bộ phân loại quan điểm duy nhất được huấn luyện cho tất cả các loại khía cạnh, trong đó thông tin khía cạnh đóng vai trò là một đặc trưng cho mô hình huấn luyện. Ưu điểm của phương pháp này là có thể khai thác tất cả các mẫu huấn luyện có sẵn, bất kể là loại khía cạnh nào. Trong thử nghiệm tiếp theo, nghiên cứu xây dựng nhiều bộ phân loại quan điểm, mỗi bộ cho một loại khía cạnh. Phương pháp này sẽ làm giới hạn số lượng mẫu huấn luyện do mỗi bộ phân loại chỉ tập trung vào một loại khía cạnh cụ thể. Bảng 2.7 trình bày độ đo F_1 của ba mô hình theo chiến lược huấn luyện mới. Với nhãn tích cực, tất cả các mô hình đều đạt được kết quả tốt (cao hơn 80%) trên hầu hết các loại khía cạnh ngoại trừ DRINKS#PRICES. So sánh với mô hình cơ sở, CRL và WEmb đạt được kết quả tốt

hơn trên một số loại khía cạnh, bao gồm RESTAURANT#PRICES, RESTAURANT#MISCELLANEOUS, FOOD#STYLE_OPTIONS, FOOD#QUALITY, DRINKS#STYLE_OPTIONS, và SERVICE#GENERAL.

Bảng 2.7. Kết quả độ đo F_1 (%) cho phân loại quan điểm (mỗi bộ phân loại cho một loại khía cạnh) với k=5 từ

Loại khía cạnh	Mô hình cơ sở		CRL		WEmb	
	Tích cực	*Tiêu cực*	*Tích cực*	*Tiêu cực*	*Tích cực*	*Tiêu cực*
RESTAURANT#GENERAL	88,12	32,00	87,51	36,38	85,69	32,97
RESTAURANT#PRICES	84,11	70,86	84,45	65,65	84,57	66,44
RESTAURANT#MISCELLANEOUS	92,81	28,33	91,32	26,33	92,14	31,38
FOOD#QUALITY	87,30	48,44	85,75	45,65	85,64	51,35
FOOD#STYLE_OPTIONS	90,53	36,86	90,95	45,22	89,51	37,80
FOOD#PRICES	85,40	75,78	86,34	76,84	79,98	69,61
DRINKS#QUALITY	84,74	35,81	83,39	24,38	83,23	30,55
DRINKS#STYLE_OPTIONS	85,61	10,00	87,79	10,00	86,93	20,00
DRINKS#PRICES	72,29	40,00	77,40	53,00	67,85	48,33
SERVICE#GENERAL	84,93	47,38	87,05	59,69	86,99	61,09
AMBIENCE#GENERAL	89,03	51,55	89,03	50,60	88,15	51,66
LOCATION#GENERAL	90,24	90,59	86,36	85,00	83,71	82,17

2.5. KẾT LUẬN CHƯƠNG 2

Chương 2 luận án đã trình bày một phương pháp dựa trên phân loại để khai thác ý kiến/quan điểm dựa trên khía cạnh cho tiếng Việt, một ngôn ngữ ít tài nguyên. Phương pháp đề xuất được thực hiện bằng cách tận dụng các tập dữ liệu đã được gán nhãn sẵn từ các ngôn ngữ khác (giàu tài nguyên hơn), kết hợp với tập dữ liệu tự gán nhãn cho ngôn ngữ tiếng Việt, để huấn luyện mô hình học nhằm nâng cao hiệu quả cho các bộ phân loại có giám sát. Phương pháp này khá tổng quát và linh hoạt do tính chất không phụ thuộc vào ngôn ngữ và có thể được sử dụng với bất kỳ thuật toán học

có giám sát nào. Kết quả thử nghiệm trên tập dữ liệu tác giả xây dựng cho tiếng Việt (trong miền lĩnh vực nhà hàng) cho bài toán khai thác ý kiến/quan điểm dựa trên khía cạnh cho thấy việc làm phong phú tập dữ liệu huấn luyện với dữ liệu dịch từ tiếng Anh đã làm tăng đáng kể hiệu năng của các mô hình trích xuất các loại khía cạnh và phân loại quan điểm. Ngoài ra, việc sử dụng các đặc trưng nhúng từ đã làm cải thiện hơn nữa hiệu quả của quá trình trích xuất các loại khía cạnh và phân loại quan điểm.

CHƯƠNG 3. TRÍCH XUẤT THỰC THỂ VÀ QUAN HỆ TRONG VĂN BẢN PHÁP QUY TIẾNG VIỆT SỬ DỤNG HỌC MÁY TRUYỀN THỐNG VÀ HỌC SÂU

Các mô hình học máy truyền thống thường cần sử dụng các phương pháp, kỹ thuật khác nhau để chọn ra được tập các đặc trưng tốt cho các mô hình học, được gọi là kỹ thuật trích chọn đặc trưng (*feature engineering*). Các phương pháp này thường được thực hiện theo cách thủ công, do vậy rất tốn kém thời gian và công sức, đồng thời cần có kiến thức chuyên gia về miền lĩnh vực nghiên cứu. Hơn nữa, trong nhiều trường hợp, tập đặc trưng thu được vẫn có thể không được đầy đủ (còn thiếu đặc trưng quan trọng cho bài toán), các đặc trưng rời rạc (không có mối liên hệ với nhau), và có thể xuất hiện lỗi trong quá trình chọn và trích xuất đặc trưng. Những vấn đề này dẫn đến giảm hiệu quả của các hệ thống trích xuất thông tin.

Trong các năm gần đây, học sâu là một hướng tiếp cận được cộng đồng các nhà khoa học quan tâm nghiên cứu, được coi là một bước tiến vượt bậc của học máy, và được ứng dụng hiệu quả trong rất nhiều lĩnh vực khác nhau. Ưu điểm của các phương pháp này là có khả năng tự động tạo ra các biểu diễn đặc trưng hiệu quả từ dữ liệu, kết hợp được nhiều nguồn thông tin và có độ chính xác cao. Trong xử lý ngôn ngữ tự nhiên, các phương pháp dựa trên học sâu sẽ tạo ra những biểu diễn chung cho các từ trong tập văn bản, từ đó có thể giúp nắm bắt được những đặc trưng về ngữ nghĩa cũng như các ràng buộc về cú pháp trong các câu văn bản. Do vậy, Chương 3 luận án tập trung nghiên cứu đề xuất giải quyết và nâng cao hiệu quả cho một số nhiệm vụ trích xuất thông tin sử dụng các phương pháp học máy truyền thống và học sâu tiên tiến, kết hợp lợi thế của các phương pháp này. Lĩnh vực lựa chọn áp dụng cho nghiên cứu ở đây là xử lý văn bản pháp quy.

Văn bản pháp quy (hay văn bản quy phạm pháp luật) như hiến pháp, luật, nghị định, thông tư là các văn bản do cơ quan nhà nước ban hành để điều tiết hoạt động

của nhà nước và xã hội. Với số lượng văn bản pháp quy lớn, được gia tăng và cập nhật theo thời gian, việc tiếp cận và chọn lọc thông tin từ hệ thống văn bản pháp quy là một việc rất khó khăn với những người bình thường không có chuyên môn về pháp luật, và thậm chí cả những người có chuyên môn như các chuyên gia về luật, luật sư. Từ đó dẫn tới nhu cầu cần phải có các công cụ xử lý văn bản pháp quy tự động, như tìm kiếm, tra cứu, phân tích, truy vấn (hỏi/đáp) nhằm hỗ trợ tốt hơn cho người dùng. Nội dung Chương 3 luận án đề cập đến vấn đề trích xuất thông tin trong các văn bản pháp quy tiếng Việt, với hai nhiệm vụ chính: (1) trích xuất thực thể tham chiếu từ văn bản pháp quy, và (2) phân loại quan hệ giữa các thực thể văn bản pháp quy. Trích xuất thực thể tham chiếu từ văn bản pháp quy là việc trích xuất ra được các tham chiếu là tên của văn bản được đề cập/nhắc đến trong văn bản pháp quy đang xem xét. Phân loại quan hệ giữa các thực thể văn bản pháp quy là việc phân loại mối liên quan giữa thực thể là văn bản tham chiếu được đề cập (đã trích xuất được ở nhiệm vụ trước) và thực thể là văn bản đang xem xét. Việc xác định được thực thể tham chiếu là một yêu cầu cần thiết để nhận ra mối quan hệ giữa các văn bản và các phần của văn bản, đồng thời cũng có thể sử dụng cho các bài toán khác. Việc xác định được mối quan hệ giữa các thực thể giúp người dùng thuận tiện trong việc tìm kiếm, tra cứu, phân tích, hay truy vấn nội dung văn bản pháp quy.

Chương 3 trình bày tổng hợp kết quả nghiên cứu của luận án dựa trên các công trình nghiên cứu số [1, 5] của tác giả (Theo danh mục các công trình công bố), bao gồm các nội dung sau:

- Đề xuất phương pháp trích xuất thực thể tham chiếu trong các văn bản pháp quy tiếng Việt sử dụng phương pháp kết hợp lợi thế của mô hình học sâu và các đặc trưng được thiết kế thủ công (theo phương pháp học máy truyền thống). Mô hình đề xuất bao gồm một số lớp LSTM hai chiều (BiLSTM) tạo ra biểu diễn câu từ các từ, ký tự và các đặc trưng nhúng thủ công, và một trường ngẫu nhiên có điều kiện (CRF) ở lớp suy diễn. Các đặc trưng được thiết kế thủ công là những đặc trưng được trích chọn phù hợp cho miền lĩnh vực văn bản riêng (trong trường hợp này là văn bản pháp quy).

- Đề xuất phương pháp phân loại quan hệ giữa các thực thể tham chiếu (đã được trích xuất ở phần trên) với thực thể là văn bản pháp quy đang xem xét sử dụng các phương pháp học máy truyền thống và học sâu. Mô hình học sâu đề xuất bao gồm một số lớp LSTM hai chiều (BiLSTM) để học cách biểu diễn từ, biểu diễn câu và một lớp softmax để suy diễn.
- Xây dựng một tập dữ liệu gồm 5.031 văn bản pháp quy tiếng Việt được gán nhãn thực thể tham chiếu và quan hệ giữa các thực thể.
- Thực hiện các thử nghiệm trên tập dữ liệu đã được xây dựng và phân tích kết quả thử nghiệm, chứng minh tính hiệu quả của phương pháp đề xuất.

3.1. ĐẶT VẤN ĐỀ

Xử lý văn bản pháp quy là một lĩnh vực nhận được nhiều sự quan tâm của các nhà nghiên cứu trên toàn cầu trong những năm vừa qua. Mục tiêu của xử lý văn bản pháp quy là hỗ trợ người dùng truy cập, truy xuất và thu thập thông tin pháp quy cần thiết. Nhiều nghiên cứu đã được thực hiện bao gồm các chủ đề khác nhau, như trích xuất thông tin pháp quy, truy xuất thông tin pháp quy, tự động xác định các thuật ngữ pháp quy, phân tích cấu trúc logic của văn bản pháp quy, dịch văn bản pháp quy, tóm tắt văn bản quy, trả lời câu hỏi pháp luật và mô hình hóa kiến thức pháp luật [4,24,50,53,55,56,112].

Có thể nhận thấy một đặc tính quan trọng trong các văn bản pháp quy đó là nội dung của văn bản thường đề cập đến các văn bản khác có từ trước, có mối liên quan đến văn bản hiện tại. Ví dụ, xem xét văn bản "Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004 của Bộ Tài chính", có đoạn như sau: "*Căn cứ Nghị định số 60/2003/NĐ-CP ngày 6/6/2003 của Chính phủ quy định chi tiết và hướng dẫn thi hành…*". Ngữ nghĩa ở đây là, văn bản "Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004" có quan hệ ***"căn cứ"*** với văn bản "*Nghị định số 60/2003/NĐ-CP ngày 6/6/2003*" được đề cập đến trong nội dung văn bản. Một số dạng quan hệ hay gặp khác bao gồm: "*dẫn chiếu*", "*thay thế*", "*hết hiệu lực*", "*sửa đổi hoặc bổ sung*",…

Như vậy, để có thể xây dựng được các công cụ xử lý văn bản pháp quy tự động, việc trích xuất ra được các thông tin cần thiết về tên/tham chiếu của văn bản, cũng như mối quan hệ giữa các văn bản là một phần công việc quan trọng.

"Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004 của Bộ Tài chính"

Căn cứ [Nghị định số 60/2003/NĐ-CP ngày 6/6/2003]Căn_cứ của Chính phủ quy định chi tiết và hướng dẫn thi hành [Luật Ngân sách nhà nước]None, [Thông tư số 59/TT-BTC ngày 23/6/2003]Căn_cứ của Bộ Tài chính hướng dẫn thực hiện [Nghị định số 60/2003/NĐ-CP ngày 6/6/2003]None của Chính phủ và hướng dẫn tại Thông tư này, Chủ tịch UBND tỉnh, thành phố trực thuộc trung ương quy định, hướng dẫn cụ thể cho phù hợp.

Hình 3.1. Ví dụ thực thể tham chiếu và mối quan hệ giữa các thực thể tham chiếu với văn bản pháp quy đang xem xét

Nghiên cứu Chương 3 của luận án đề cập đến trích xuất thông tin trong văn bản pháp quy tiếng Việt với hai nhiệm vụ: (1) trích xuất thực thể tham chiếu được đề cập đến trong nội dung của các văn bản pháp quy như văn bản luật, nghị định, thông tư,…; và, (2) phân loại mối quan hệ giữa các thực thể tham chiếu (đã được trích xuất) và thực thể văn bản pháp quy đang xem xét thành các loại như "căn cứ", "dẫn chiếu", "thay thế", "hết hiệu lực", "sửa đổi hoặc bổ sung",... (sau đây sẽ gọi tắt là phân loại quan hệ giữa các thực thể). Hình 3.1 trình bày một ví dụ kết quả trích xuất thực thể tham chiếu và xác định quan hệ giữa các thực thể từ một đoạn trong văn bản "Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004" (ví dụ được nêu ở trên). Có bốn thực thể tham chiếu được trích xuất trong đoạn văn bản là (**1**) "Nghị định số 60/2003/NĐ-CP ngày 6/6/2003", (**2**) "Luật Ngân sách nhà nước", (**3**) "Thông tư số 59/TT-BTC ngày 23/6/2003", và (**4**) "Nghị định số 60/2003/NĐ-CP ngày 6/6/2003". Văn bản đang xem xét, "Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004", được xác định có quan hệ "***căn cứ***" với thực thể tham chiếu (**1**) và thực thể tham

chiếu (**3**), và không có quan hệ với thực thể tham chiếu (**2**) và (**4**) (trong Hình 3.1 giá trị quan hệ là "***none***").

Mặc dù các tài liệu/văn bản pháp quy được viết theo cách chính thống để duy trì tính rõ ràng và tránh mơ hồ, tuy nhiên, việc trích xuất tự động các thực thể tham chiếu từ văn bản pháp quy vẫn là thách thức lớn. Ví dụ, cùng một thực thể tham chiếu có thể được đưa ra theo nhiều cách, như ở dạng dài là mã văn bản, tên đầy đủ của văn bản và ngày văn bản được phát hành, hoặc ở dạng ngắn gọn chỉ là mã văn bản. Trong nhiều trường hợp, tên của một văn bản được đưa ra cũng có thể chứa một tham chiếu đến một văn bản khác, ví dụ: "Quốc hội ban hành **[**Luật sửa đổi, bổ sung một số điều của **[**Luật chứng khoán số 70/2006/QH11**]]**" (đây là trường hợp một văn bản là để sửa đổi một văn bản trước đó). Do có các biến thể như vậy trong việc thể hiện các thực thể tham chiếu và cấu trúc đa dạng của các văn bản làm cho việc sử dụng các quy tắc đơn giản (ví dụ, dùng biểu thức chính quy) để trích xuất thực thể tham chiếu trở nên khó khăn hơn.

Trích xuất tự động quan hệ giữa các thực thể từ văn bản pháp quy có một số khó khăn do không có định nghĩa rõ ràng về các thực thể cũng như mối quan hệ giữa các thực thể từ văn bản pháp quy. Xét ví dụ trong Hình 3.1, thực thể tham chiếu thứ nhất có thể có một trong các định dạng sau: "*Nghị định số 60/2003/NĐ-CP*", "*Nghị định số 60/2003/NĐ-CP ngày 6/6/2003*", hay "*Nghị định số 60/2003/NĐ-CP ngày 6/6/2003 của Chính phủ*". Như vậy, để trích xuất được thực thể tham chiếu cần phải có quy định về định dạng nhận diện thực thể. Về vấn đề xác định các mối quan hệ, thực thể văn bản "Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004" (đang xem xét) có quan hệ *"căn cứ"* với hai thực thể tham chiếu (**1**) và (**3**) được đề cập trong nội dung. Tuy nhiên, có thể suy luận theo cách khác là hai thực thể tham chiếu (**1**) và (**3**) có quan hệ "*dẫn chiếu*" với thực thể văn bản đang xem xét. Ngoài ra, thực thể tham chiếu (**3**) cũng có thể bị xác định nhầm là không có quan hệ với thực thể văn bản đang xem xét, do đứng liền sau thực thể tham chiếu (**1**) và (**2**) trong cùng một câu.

Nghiên cứu Chương 3 luận án đề xuất các phương pháp nhằm giải quyết những khó khăn nêu trên. Với nhiệm vụ thứ nhất, *trích xuất thực thể tham chiếu*, nghiên cứu đề xuất xử lý như là một bài toán gán nhãn chuỗi *(sequence labeling)*. Các mô hình đề xuất thử nghiệm trích xuất thực thể tham chiếu bao gồm cả phương pháp truyền thống (trường ngẫu nhiên có điều kiện, CRF) và phương pháp tiên tiến hơn (các mạng nơ-ron sâu), đồng thời kết hợp các mô hình này. Ngoài các đặc trưng được học bởi các mạng sâu, nghiên cứu khai thác thêm các loại đặc trưng thủ công riêng thể hiện đặc tính của tài liệu văn bản pháp quy. Với nhiệm vụ thứ hai, *trích xuất quan hệ* giữa các thực thể văn bản pháp quy, nghiên cứu đề xuất xử lý như là một bài toán phân loại. Các phương pháp đề xuất thử nghiệm phân loại quan hệ bao gồm cả các phương pháp học máy có giám sát truyền thống (là các phương pháp phổ biến và đạt được độ chính xác cao trong các nghiên cứu về trích xuất quan hệ) và các phương pháp dựa trên học sâu nhằm tận dụng các ưu điểm về khả năng tự động tạo ra các biểu diễn đặc trưng hiệu quả từ dữ liệu của các phương pháp này.

Đóng góp của nghiên cứu được trình bày trong Chương 3 *là đề xuất phương pháp trích xuất thông tin sử dụng học máy truyền thống và học sâu cho văn bản pháp quy tiếng Việt. Các thông tin được trích xuất bao gồm thực thể tham chiếu và mối quan hệ giữa các thực thể văn bản pháp quy.* Với nhiệm vụ *trích xuất thực thể tham chiếu,* nghiên cứu sử dụng một số mô hình trích xuất kết hợp lợi thế của các mô hình học sâu và các đặc trưng được thiết kế thủ công trong một khung làm việc thống nhất. Kết quả trích xuất thực thể tham chiếu thu được tốt nhất với một mô hình bao gồm một số lớp LSTM hai chiều (BiLSTM) tạo ra biểu diễn câu từ các từ, ký tự và các đặc trưng nhúng thủ công, và một CRF ở lớp suy diễn. Sự vượt trội của mô hình này so với phương pháp dựa trên CRF khi sử dụng cùng một tập đặc trưng và các phương pháp học sâu mà không có các đặc trưng được thiết kế thủ công cho thấy lợi ích khi bổ sung các đặc trưng cụ thể của bài toán vào các mô hình học sâu. Với nhiệm vụ *phân loại quan hệ* giữa các thực thể trong văn bản pháp quy, ngoài việc sử dụng phương pháp học máy truyền thống, nghiên cứu sử dụng mô hình học sâu với các mạng BiLSTM để học cách biểu diễn từ, biểu diễn câu và một lớp softmax để suy

diễn. Kết quả thử nghiệm trên tập dữ liệu bao gồm 5.031 văn bản pháp quy tiếng Việt cho thấy tính hiệu quả của các phương pháp đề xuất, với các kết quả tốt nhất là độ đo F_1 đạt 95,35% cho trích xuất thực thể tham chiếu và 97,03% cho phân loại quan hệ giữa các thực thể văn bản pháp quy.

Nội dung còn lại của Chương 3 được cấu trúc như sau. Mục 3.2 trình bày đề xuất phương pháp thực hiện trích xuất thực thể tham chiếu và quan hệ trong văn bản pháp quy tiếng Việt. Việc xây dựng bộ dữ liệu và các thực nghiệm được trình bày trong Mục 3.3 và Mục 3.4. Cuối cùng, Mục 3.5 là kết luận chương.

3.2. ĐỀ XUẤT PHƯƠNG PHÁP TRÍCH XUẤT THỰC THỂ VÀ QUAN HỆ

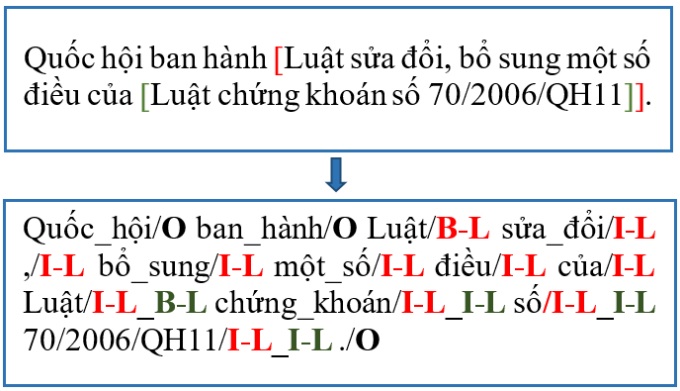
Phần này trình bày đề xuất phương pháp trích xuất thực thể tham chiếu và quan hệ trong văn bản pháp quy tiếng Việt, bao gồm hai nhiệm vụ: (1) trích xuất thực thể tham chiếu từ văn bản pháp quy, và (2) phân loại quan hệ giữa các thực thể tham chiếu và thực thể văn bản pháp quy đang xem xét.

3.2.1. Trích xuất thực thể tham chiếu

Cho một câu thuộc văn bản pháp quy tiếng Việt được biểu diễn dưới dạng một chuỗi các từ $S = w_1 w_2 \dots w_n$ trong đó n biểu thị độ dài các từ trong câu S, mục tiêu là trích xuất tất cả các tham chiếu (và loại của chúng) có trong S. Thực thể tham chiếu là một chuỗi các từ liên tiếp đề cập đến một văn bản pháp quy khác như luật, nghị định, hoặc thông tư. Nghiên cứu xem xét các tham chiếu ở cả dạng đầy đủ và dạng rút gọn, và các tham chiếu có chứa các tham chiếu khác bên trong chúng (được gọi là các tham chiếu lồng nhau).

Có thể thấy, nhiệm vụ trích xuất thực thể tham chiếu ở đây chính là nhận dạng thực thể có tên, thuộc lớp các bài toán trích xuất thông tin. Để xử lý nhiệm vụ này, có thể sử dụng một cách tiếp cận khá phổ biến là dựa trên mô hình gán nhãn chuỗi. Cách thực hiện như sau: gán cho mỗi từ một nhãn cho biết từ đó có là bắt đầu của một thực thể tham chiếu (nhãn B) hay không, nằm ở bên trong (không phải ở đầu) của một thực thể tham chiếu (nhãn I), hoặc không thuộc thực thể tham chiếu nào (nằm

ngoài, nhãn O). Hình 3.2 trình bày ví dụ một câu trong văn bản pháp quy và chuỗi nhãn tương ứng của nó. Trong trường hợp thực thể tham chiếu lồng nhau sẽ thực hiện ghép hai nhãn để tạo thành một nhãn mới. Ví dụ: nhãn I-L_B-L sẽ chỉ ra từ đó nằm trong một thực thể tham chiếu đề cập đến một luật (I-L). Mặt khác, từ này cũng bắt đầu một thực thể tham chiếu khác đề cập đến một luật (B-L).



Hình 3.2. Ví dụ một câu trong văn bản pháp quy và chuỗi nhãn được gán tương ứng

Để giải quyết nhiệm vụ gán nhãn chuỗi, nghiên cứu đề xuất sử dụng các mô hình sau: (1) Mô hình dựa trên CRF, và (2) Mô hình BiLSTM và BiLSTM-CRF.

a) Mô hình dựa trên CRF

Mô hình đầu tiên được sử dụng cho nhiệm vụ trích xuất thực thể tham chiếu là mô hình dựa trên CRF, trong đó, đầu vào là một chuỗi các từ (một câu) và đầu ra là một chuỗi các nhãn (như B-L, I-L, B-D, O) có cùng độ dài. Xác suất có điều kiện của một chuỗi nhãn $T = t_1 t_2 \dots t_n$ với một câu đầu vào $S = w_1 w_2 \dots w_n$ có thể được viết như sau:

$$p(T|S,\lambda,\mu) = \frac{1}{Z(S)} exp\left(\sum_j \lambda_j . f_j(t_{i-1}, t_i, S, i) + \sum_k \mu_k . g_k(t_i, S, i)\right) \quad (3.1)$$

Trong đó λ và μ là các tham số của mô hình, $f_j(t_{i-1}, t_i, S, i)$ là một hàm đặc trưng chuyển tiếp được xác định tại các vị trí i và $i-1$, và $g_k(t_i, S, i)$ là một hàm đặc trưng trạng thái được xác định tại vị trí i.

Có hai loại đặc trưng được trích chọn để phục vụ cho quá trình huấn luyện mô hình, bao gồm các đặc trưng *n*-grams (n = 1, 2, 3) và các đặc trưng thủ công. Các đặc trưng *n*-grams là những đặc trưng đơn giản, nhưng đã được sử dụng rất hiệu quả trong nhiều bài toán xử lý ngôn ngữ tự nhiên tiếng Việt. Các đặc trưng thủ công được trích chọn dựa trên những quan sát và phân tích về thông tin cần trích xuất, ở đây là thực thể tham chiếu, trong dữ liệu văn bản pháp quy thu thập được. Trong nghiên cứu này, chúng tôi đề xuất chọn các đặc trưng thủ công như sau:

- Đặc trưng kiểm tra xem một từ có chứa các chữ số hay không. Trong nhiều trường hợp, thực thể tham chiếu được đề cập chứa thông tin về số hiệu của văn bản, hoặc có chứa thông tin về ngày tháng phát hành văn bản. Ví dụ như *"Nghị định số 60/2003/NĐ-CP"*, hoặc *"Nghị định số 60/2003/NĐ-CP ngày 6/6/2003",...*
- Đặc trưng kiểm tra xem một từ có chứa một ký tự đặc biệt không (gạch nối, dấu gạch ngang,…). Khi thực thể tham chiếu chứa thông tin về số hiệu của văn bản, nó thường chứa các ký tự đặc biệt, ví dụ như *"Nghị định số 60/2003/NĐ-CP"*.
- Đặc trưng kiểm tra xem ký tự đầu tiên của từ có được viết hoa hay không. Trong các văn bản pháp quy, khi đề cập đến tên một văn bản, thường tên này sẽ được viết hoa, hoặc là từ đầu tiên, hoặc là một số âm tiết trong từ. Ví dụ: "*Căn cứ [**N**ghị định số 60/2003/NĐ-CP ngày 6/6/2003] ...".*
- Đặc trưng kiểm tra xem các ký tự đầu tiên của tất cả các âm tiết của từ có được viết hoa hay không. Ví dụ, *"Bộ Luật Lao động năm 2019"*.
- Đặc trưng kiểm tra xem một từ có phải là từ khóa hay không. Đây là các từ đại diện cho một loại văn bản pháp quy như hiến pháp, bộ luật, luật, nghị định, quyết định, thông tư và thông tư liên tịch. Các từ này thường xuất hiện tại vị

trí đầu tiên của thực thể tham chiếu. Ví dụ: *"**Nghị định** số 60/2003/NĐ-CP ngày 6/6/2003"*.

b) Mô hình BiLSTM và BiLSTM-CRF

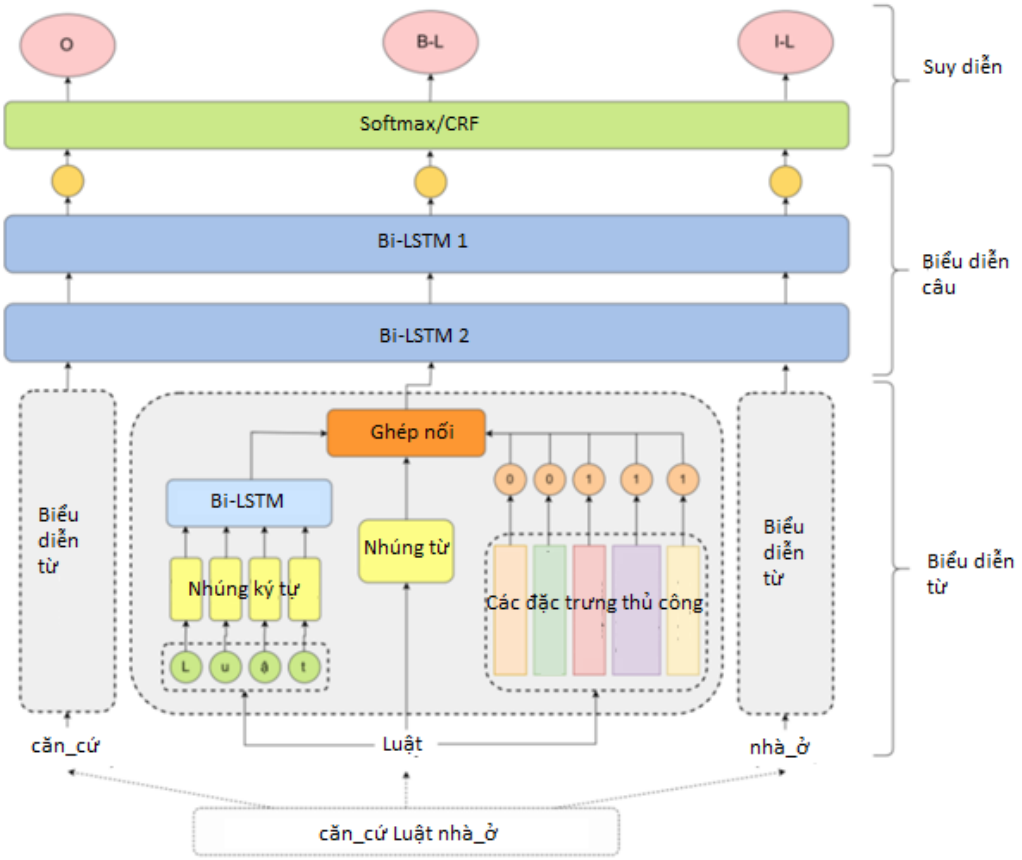
Các mô hình tiếp theo được sử dụng là BiLSTM và BiLSTM-CRF, được minh họa trong Hình 3.3, bao gồm ba lớp: biểu diễn từ, biểu diễn câu và suy diễn.

- **Biểu diễn từ:** lớp này tạo ra cách biểu diễn từ bằng cách ghép các nguồn thông tin khác nhau: nhúng từ *(word embeddings)*, nhúng ký tự *(character embeddings)* và bổ sung các đặc trưng được thiết kế thủ công. Mạng BiLSTM được sử dụng để mô hình hóa mối quan hệ giữa các ký tự.
- **Biểu diễn câu:** lớp này sử dụng BiLSTM để mô hình hóa mối quan hệ giữa các từ. Các nghiên cứu trước đây cho thấy rằng việc xếp chồng một số mạng BiLSTM có thể tạo ra các biểu diễn tốt hơn [37]. Do đó trong nghiên cứu này, mô hình đề xuất cũng sử dụng hai mạng BiLSTM.
- **Suy diễn:** trong khi BiLSTM sử dụng các hàm softmax thì BiLSTM-CRF sử dụng CRF để suy diễn. Ưu điểm của việc sử dụng CRF để suy diễn là có thể sử dụng mối tương quan giữa nhãn hiện tại và các nhãn lân cận.

Ý tưởng chính của các mô hình này là tích hợp các đặc trưng được thiết kế thủ công vào kiến trúc mạng nơ-ron sâu. Ưu điểm của các mô hình truyền thống như CRF là có thể được huấn luyện trên một bộ dữ liệu khá nhỏ với các đặc trưng được thiết kế thủ công. Ngược lại, kiến trúc mạng nơ-ron sâu có thể học đặc trưng tự động từ dữ liệu. Tuy nhiên, nhược điểm của kiến trúc sâu là yêu cầu dữ liệu lớn. Hiệu năng của kiến trúc này thường bị suy giảm khi chỉ được huấn luyện trên một tập dữ liệu nhỏ. Do đó, nghiên cứu đề xuất kết hợp cả hai phương pháp truyền thống (CRF) và tiên tiến hơn (BiLSTM) vào một kiến trúc duy nhất, sử dụng cả hai loại đặc trưng, là các đặc trưng được học tự động và các đặc trưng được thiết kế thiết kế thủ công, để đạt được một mô hình tiềm năng hơn.

Long-short term memory (LSTM), một biến thể của RNN, đã được đề xuất bởi [38] với ý tưởng chính là sử dụng một số cổng để kiểm soát việc truyền thông tin

dọc theo chuỗi (đã được tóm tắt giới thiệu trong Mục 1.3.3 Chương 1). Mạng LSTM hai chiều (BiLSTM) [39] kết hợp hai LSTM: một di chuyển về phía trước từ phía bên trái và một di chuyển ngược từ phía bên phải của câu. Bằng cách này, mạng BiLSTM có thể nắm bắt thông tin phong phú trên cả hai hướng cho việc mô hình hóa các câu. Việc sử dụng BiLSTM để mô hình hóa các từ từ các ký tự được thực hiện tương tự. BiLSTM đã được áp dụng thành công cho nhiều nhiệm vụ NLP khác nhau, như gán nhãn từ loại, nhận dạng thực thể có tên, suy luận ngôn ngữ tự nhiên.



Hình 3.3. Các mô hình BiLSTM và BiLSTM-CRF cho trích xuất thực thể tham chiếu

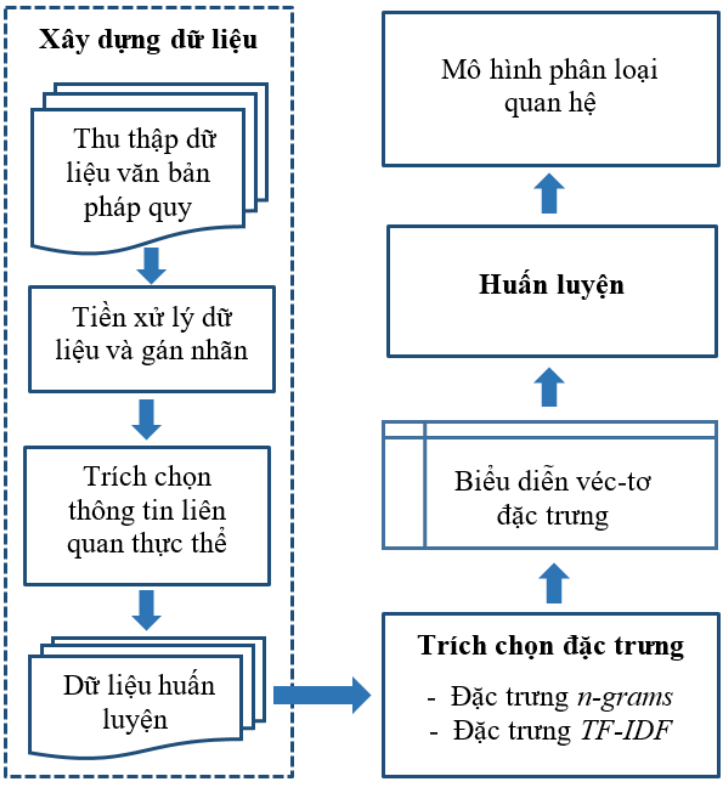
3.2.2. Phân loại quan hệ giữa các thực thể văn bản pháp quy

Phần này trình bày đề xuất phương pháp phân loại quan hệ giữa các thực thể là văn bản tham chiếu và thực thể văn bản pháp quy đang xem xét. Các loại quan hệ

được xác định bao gồm: căn cứ, dẫn chiếu, hết hiệu lực, hướng dẫn, sửa đổi hoặc bổ sung, thay thế. Nghiên cứu đề xuất sử dụng cả hai cách tiếp cận sử dụng học máy truyền thống và học sâu để giải quyết nhiệm vụ này. Phần sau sẽ trình bày chi tiết về các phương pháp đề xuất.

3.2.2.1. Phân loại quan hệ giữa các thực thể văn bản pháp quy sử dụng học máy truyền thống

Giả sử cho một tập dữ liệu văn bản pháp quy *D* đã xác định được các thực thể tham chiếu. Xét *A* là một văn bản trong tập *D*, *A* có thể có một hoặc nhiều thực thể tham chiếu, được ký hiệu là B_k.



Hình 3.4. Sơ đồ các bước đề xuất giải quyết nhiệm vụ phân loại quan hệ giữa các thực thể trong văn bản pháp quy

Với mỗi thực thể tham chiếu B_k, xét đoạn văn bản chứa thực thể tham chiếu này. Mỗi đoạn văn bản trên sẽ được sử dụng làm đầu vào cho bài toán phân loại. Mục

tiêu là, với mỗi thực thể tham chiếu B_k, cần phải xác định quan hệ giữa thực thể B_k với thực thể văn bản A đang xem xét, dựa trên các thông tin đầu vào từ đoạn văn bản chứa thực thể tham chiếu B_k.

Hình 3.4 trình bày sơ đồ các bước đề xuất giải quyết bài toán phân loại giữa các thực thể trong văn bản pháp quy, bao gồm 3 bước chính: xây dựng dữ liệu huấn luyện, trích chọn đặc trưng và huấn luyện mô hình phân loại quan hệ.

a) Xây dựng dữ liệu huấn luyện

Mỗi văn bản pháp quy A có chứa một hoặc nhiều thực thể tham chiếu B_k có mối quan hệ với văn bản đang xem xét A. Giả thiết là đã xác định được tất cả các thực thể tham chiếu B_k trong văn bản A. Để có thể xây dựng dữ liệu huấn luyện mô hình xác định quan hệ giữa thực thể A và từng thực thể B_k đã được xác định, nghiên cứu thực hiện trích chọn các phần nội dung văn bản có liên quan đến các thực thể. Các thông tin trích chọn là thông tin về các thực thể và thông tin ngữ cảnh xung quanh thực thể tham chiếu thuộc đoạn văn bản chứa thực thể tham chiếu đó. Cụ thể, xét một thực thể tham chiếu B_k đã được xác định trong văn bản A, các thông tin được trích chọn để tạo thành một mẫu dữ liệu huấn luyện sẽ bao gồm như sau:

- Thực thể tham chiếu B_k,
- Phần văn bản ở phía trước thực thể tham chiếu B_k (trong cùng câu với B_k),
- Phần văn bản ở phía sau thực thể tham chiếu B_k (trong cùng câu với B_k),
- Tên của thực thể văn bản A,
- Tên điều khoản (nếu có) của đoạn văn bản chứa thực thể tham chiếu B_k

Mỗi phần thông tin (văn bản) trên sẽ được trích chọn đặc trưng riêng và biểu diễn dưới dạng véc-tơ, sau đó, các véc-tơ đặc trưng này sẽ được ghép nối để tạo thành một véc-tơ đặc trưng kết hợp, làm đầu vào cho quá trình huấn luyện mô hình phân loại quan hệ, như được trình bày trong phần sau đây.

b) Trích chọn đặc trưng

Để trích chọn đặc trưng, các văn bản pháp quy được thực hiện phân đoạn từ tiếng Việt. Do mỗi từ tiếng Việt bao gồm một âm tiết (trong các từ đơn) hoặc nhiều âm tiết (trong các từ ghép và từ láy) được phân tách nhau bởi các ký tự trống. Vì thế, phân đoạn từ là một bước tiền xử lý quan trọng trong hầu hết các bài toán xử lý ngôn ngữ tự nhiên tiếng Việt.

Trong nghiên cứu này, hai loại đặc trưng được đề xuất trích chọn là đặc trưng *n*-grams và đặc trưng TF-IDF. Đây là những đặc trưng phổ biến được sử dụng trong các bài toán xử lý ngôn ngữ tự nhiên và đã được chứng minh đạt kết quả tốt trong nhiều nghiên cứu trước đây. Phần sau sẽ giới thiệu ngắn gọn về hai loại đặc trưng này và mô tả các kết hợp chúng để biểu diễn các mẫu dữ liệu đầu vào cho nhiệm vụ.

1) ***Đặc trưng n-grams***: Các đặc trưng *n*-grams của từ được trích xuất từ các văn bản pháp quy đã được phân đoạn từ tiếng Việt. Mặc dù các đặc trưng này rất đơn giản, nhưng chúng có hiệu quả tốt đối với hầu hết các bài toán phân loại văn bản. Ở đây, các đặc trưng *n*-grams được trích chọn là unigrams và bigrams của từ được trích xuất từ văn bản pháp quy đã được phân đoạn từ tiếng Việt.

2) ***Đặc trưng TF-IDF*** *(Term Frequency – Inverse Document Frequency)*: Cho một tập các văn bản D. Xét một từ w trong văn bản d thuộc tập D. $TF\text{-}IDF$ của từ w là giá trị thể hiện mức độ quan trọng của từ w trong văn bản d trên tập D, được tính toán dựa trên hai thành phần là TF và IDF như sau:

$$TF\text{-}IDF(w, d, D) = TF(w, d) * IDF(w, D) \tag{3.2}$$

trong đó, $TF(w, d)$ là tần số xuất hiện của từ w trong văn bản d:

$$TF(w, d) = \frac{\text{Số lần từ } w \text{ xuất hiện trong văn bản } d}{\text{Tổng số từ trong văn bản } d} \tag{3.3}$$

và, $IDF(w, D)$ là tần số nghịch đảo của từ w trong tập văn bản D:

$$IDF(w, D) = log \frac{\text{Tổng số văn bản có trong } D}{\text{Số văn bản có chứa từ } w} \tag{3.4}$$

Giá trị $TF\text{-}IDF(w, d, D)$ cao thể hiện w xuất hiện nhiều trong văn bản d và ít xuất hiện trong các văn bản khác trong tập D. Nghĩa là, w là từ có giá trị cao (từ khóa)

của văn bản d. Giá trị $TF\text{-}IDF(w, d, D)$ thấp chỉ ra w là từ phổ biến với tất cả các văn bản, nên sẽ ít có giá trị với văn bản d.

Trong nghiên cứu này, giá trị $TF\text{-}IDF$ sẽ được tính với *n*-grams (unigrams, bigrams) của từ được trích xuất từ văn bản pháp quy đã được phân đoạn từ tiếng Việt.

3) ***Kết hợp đặc trưng***: Gọi d_i là một phần thông tin thuộc 5 phần thông tin được trích chọn như trong mục (a) ở phần trên. Việc kết hợp đặc trưng *n*-grams với đặc trưng $TF\text{-}IDF$ cho đoạn văn bản d_i được thực hiện bằng cách ghép nối các véc-tơ đặc trưng như sau:

- Biểu diễn d_i bằng một véc-tơ *one-hot* $v_{oh}(d_i)$ theo *n-grams.*
- Biểu diễn d_i bằng một véc-tơ *TF-IDF* $v_{tf-idf}(d_i)$ cho tất cả các từ *w* (là *n-grams*) trong phần văn bản d_i trong tập văn bản D.
- Ghép nối 2 véc-tơ $v_{oh}(d_i)$ và $v_{tf-idf}(d_i)$ tạo thành véc-tơ $v(d_i)$ (là véc-tơ đặc trưng của đoạn văn bản d_i)

Cuối cùng, ghép nối 5 véc-tơ $v(d_i)$ để tạo thành véc-tơ đặc trưng cho một mẫu dữ liệu huấn luyện.

c) Huấn luyện mô hình

Giả sử *N* là số lượng quan hệ muốn trích xuất. Nhiệm vụ là cần huấn luyện một bộ phân loại đa lớp để dự đoán nhãn quan hệ giữa các thực thể văn bản pháp quy đã được xác định. Để huấn luyện mô hình, nghiên cứu sử dụng ba thuật toán học máy khác nhau là Phân loại Bayes đơn giản (Naïve Bayes) [100], Cây quyết định [98] và Máy véc-tơ tựa [116], đại diện cho ba nhóm thuật toán khác nhau: dựa trên trên mô hình xác suất, dựa trên mô hình cây và dựa trên hàm nhân.

3.2.2.2. Phân loại quan hệ giữa các thực thể văn bản pháp quy sử dụng mô hình dựa trên học sâu

Giả sử cho một tập dữ liệu văn bản pháp quy *D* đã xác định được các thực thể tham chiếu. Xét *S* là một câu trong văn bản *A* thuộc tập văn bản *D*, *S* có thể có một

hoặc nhiều thực thể tham chiếu, được ký hiệu là B_k. Nhiệm vụ là với mỗi thực thể tham chiếu B_k trong câu *S*, cần xác định mối quan hệ của thực thể này với thực thể là văn bản đang xem xét *A*. Nhiệm vụ này có thể được thực hiện theo cách tiếp cận dựa trên phân loại (như đề xuất ở phần trên) hoặc có thể được thực hiện theo cách tiếp cận dựa trên gán nhãn chuỗi, trong đó, với mỗi câu đầu vào *S*, sẽ cho đầu ra là một chuỗi nhãn quan hệ tương ứng với các thực thể tham chiếu có trong *S*.

Cụ thể, ký hiệu các nhãn quan hệ có thể có: căn cứ (CC), dẫn chiếu (DaC), hết hiệu lực (HHL), hướng dẫn (HD), sửa đổi hoặc bổ sung (DSD), thay thế (BTT). Ngoài ra, trong trường hợp thực thể tham chiếu không có quan hệ với văn bản đang xem xét, nhãn quan hệ sẽ được gán là "*none*". Các từ còn lại trong câu đầu vào *S* không thuộc thực thể tham chiếu nào sẽ được gán nhãn O.

Do mỗi thực thể tham chiếu thường bao gồm nhiều từ trong câu đầu vào, nên nếu gán nhãn quan hệ theo từng từ thuộc thực thể thì sẽ bị lặp nhãn quan hệ, việc này không có ý nghĩa, đồng thời việc sử dụng nhãn lặp cũng có thể gây nhiễu cho quá trình trích xuất. Do vậy, nghiên cứu đề xuất xử lý câu đầu vào *S* như sau: với mỗi thực thể tham chiếu trong câu *S* sẽ được chuyển thành một từ đại diện cho loại thực thể. Việc xử lý này hoàn toàn không gây ảnh hưởng tới ngữ nghĩa của câu về mối quan hệ giữa thực thể tham chiếu với thực thể văn bản pháp quy đang xem xét.

Ví dụ:

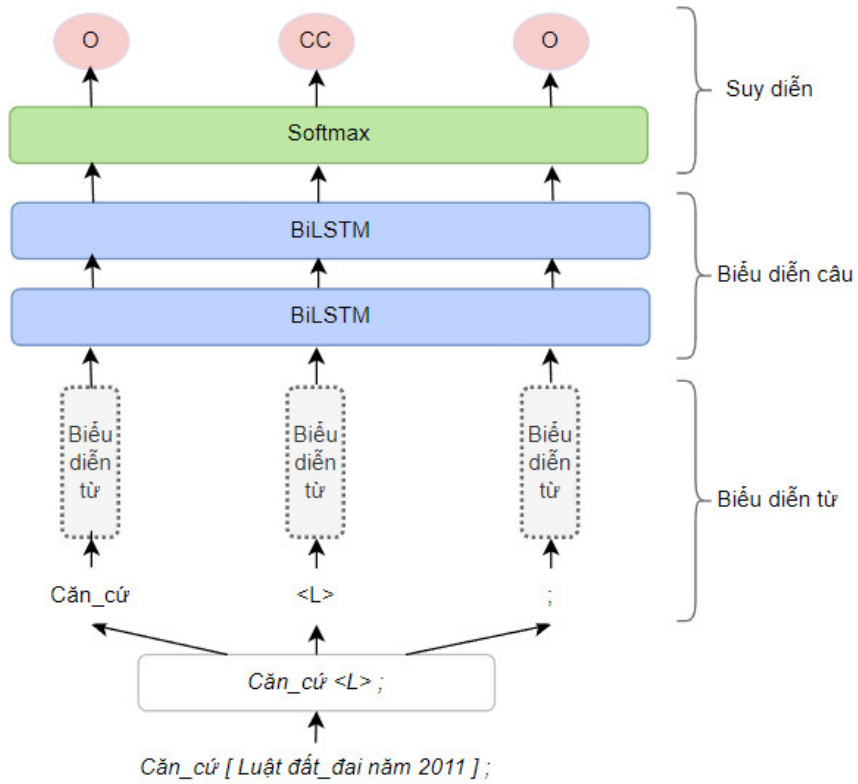
Đầu vào: Cho câu đầu vào "*Căn_cứ Luật đất_đai năm 2011 ;*".

Xử lý câu đầu vào thành chuỗi: "*Căn_cứ <L> ;*", trong đó <*L*> đại diện cho loại thực thể "*luật*", thay thế cho thực thể "*Luật đất_đai năm 2011"*.

Đầu ra: Chuỗi nhãn đầu ra tương ứng: "*O, CC, O*"

Mô hình đề xuất để giải quyết nhiệm vụ gán nhãn quan hệ cho chuỗi *S* đầu vào trong nghiên cứu ở đây là sử dụng mô hình học sâu dựa trên BiLSTM. Mô hình này gồm ba lớp: biểu diễn từ, biểu diễn câu và suy diễn. Lớp biểu diễn từ tạo ra cách biểu diễn từ với kỹ thuật nhúng ký tự. Lớp biểu diễn câu được thực hiện với kỹ thuật nhúng

từ. Lớp suy diễn sử dụng hàm softmax để suy diễn. Dựa trên kiến trúc mạng nơ-ron sâu này, mô hình có thể tự động học đặc trưng từ dữ liệu đầu vào. Hình 3.5 thể hiện kiến trúc của mạng nơ-ron đề xuất.



Hình 3.5. Mô hình BiLSTM cho phân loại quan hệ giữa các thực thể

3.3. XÂY DỰNG TẬP DỮ LIỆU

Phần này sẽ mô tả về việc xây dựng tập dữ liệu để sử dụng cho các thực nghiệm, bao gồm hai bước chính: thu thập và tiền xử lý dữ liệu, và gán nhãn dữ liệu.

a) Thu thập và tiền xử lý dữ liệu

Nguồn dữ liệu được thu thập từ Cổng thông tin "Cơ sở dữ liệu Quốc gia về Văn bản pháp luật" của Nhà nước, tại *http://vbpl.vn*. Trong đó, dữ liệu được lựa chọn từ ba loại văn bản pháp quy quan trọng và phổ biến nhất, là luật, nghị định và thông tư, và chọn ngẫu nhiên một tập hợp con trong nguồn này để xây dựng tập dữ liệu. Một số bước tiền xử lý được thực hiện trước khi gán nhãn dữ liệu như sau:

- Loại bỏ các phần văn bản không liên quan, như phần đầu trang, chân trang
- Tách các âm tiết bị lỗi dính liền nhau
- Chuẩn hóa dấu từ (thanh điệu)
- Tách câu, tách từ tiếng Việt.

Việc tách từ tiếng Việt được thực hiện bằng cách sử dụng Pyvi, là một bộ công cụ xử lý ngôn ngữ tự nhiên của Python cho tiếng Việt, có tại: *https://github.com/trungtv/pyvi*.

Kết quả sau khi tiền xử lý thu được tập dữ liệu gồm 5.031 văn bản pháp quy tiếng Việt. Bước tiếp theo là thực hiện gán nhãn cho tập dữ liệu này.

b) Gán nhãn dữ liệu

Quy trình gán nhãn thực thể tham chiếu và quan hệ giữa các thực thể văn bản pháp quy được thực hiện theo 2 bước: gán nhãn tự động và gán nhãn thủ công.

Gán nhãn tự động: Việc gán nhãn tự động nhằm mục đích làm tăng tốc độ gán nhãn bằng cách sử dụng các biểu thức chính quy. Có một số quan sát và thảo luận như sau:

- Thực thể tham chiếu của văn bản pháp quy thường bắt đầu bằng một từ khóa về loại văn bản pháp quy. Do vậy, nghiên cứu xây dựng một từ điển các từ khóa về loại văn bản pháp quy, bao gồm: Hiến pháp, Bộ luật, Luật, Pháp lệnh, Nghị định, Nghị quyết, Quyết định, Thông tư, Thông tư liên tịch.
- Thực thể tham chiếu của văn bản pháp quy thường kết thúc theo một trong các dạng sau:
  - Ngày tháng năm: có các dạng như năm *yyyy* (ví dụ: 2015) hoặc ngày *dd* tháng *mm* năm *yyyy* (ví dụ: ngày 23 tháng 12 năm 2013).
  - Mã số văn bản pháp quy (ví dụ như *215/2013/NĐ-CP*)

- Các loại thực thể tham chiếu cũng thường liên quan đến từ đầu tiên trong một thực thể tham chiếu. Do vậy, nghiên cứu đã xây dựng một từ điển tương tự để tự động phân loại thực thể tham chiếu.
- Các loại quan hệ: Các loại quan hệ giữa các thực thể văn bản pháp quy thường được mô tả bằng một số từ khóa hoặc cụm từ khóa xung quanh các thực thể tham chiếu như "căn cứ" và "thay thế". Nghiên cứu cũng xây dựng một từ điển các từ/cụm từ khóa như vậy để xác định các loại quan hệ, bao gồm: căn cứ, dẫn chiếu, hết hiệu lực, thay thế, sửa đổi hoặc bổ sung và hướng dẫn.

Gán nhãn thủ công: Trong bước này, các thực thể tham chiếu và quan hệ được gán nhãn tự động sẽ được kiểm tra và chỉnh sửa thủ công bởi hai người gán nhãn độc lập. Hai người này là sinh viên ngành Công nghệ thông tin, học tại Học viện Công nghệ Bưu chính Viễn thông, có kiến thức cơ bản về trích xuất thông tin, xử lý văn bản pháp quy và học máy. Trong trường hợp hai người gán nhãn bất đồng ý kiến, việc gán nhãn sẽ được kiểm ra và đưa ra quyết định cuối cùng bởi một người thứ ba là Cử nhân ngành Luật. Các thông tin gán nhãn được yêu cầu chỉnh sửa bao gồm: vị trí bắt đầu và vị trí kết thúc của thực thể tham chiếu, loại thực thể tham chiếu và loại quan hệ.

Nghiên cứu sử dụng hệ số Kappa của Cohen để đo mức độ tương đồng ý kiến giữa các nhãn được gán, tương tự như trình bày trong nghiên cứu Chương 2 luận án (Mục 2.3). Hệ số Kappa được tính toán theo công thức (2.2) cho 2 loại nhãn là thực thể tham chiếu và quan hệ tương ứng là 0,92 và 0,94. Các giá trị này được coi là có độ tương đồng ý kiến khá tốt [27].

Hình 3.6 trình bày ví dụ cụ thể một đoạn văn bản pháp quy "*Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004 của Bộ tài chính*" được gán nhãn thực thể tham chiếu và quan hệ theo định dạng XML trong tập dữ liệu. Các cặp thẻ chứa thực thể tham chiếu: thông tư (*<TT>,</TT>*), nghị định (*<NĐ>,</NĐ>*),…; thuộc tính "*rel*" xác định loại quan hệ: căn cứ "*CC*", dẫn chiếu "*DaC*",… của thực thể văn bản tham chiếu trong nội dung với thực thể văn bản đang xem xét.

"Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004 của Bộ Tài chính"

Căn cứ <NĐ rel="CC"> Nghị định số 60/2003/NĐ-CP ngày 6/6/2003 </NĐ> của Chính phủ quy định chi tiết và hướng dẫn thi hành <L rel="none"> Luật Ngân sách nhà nước </L>, <TT rel="CC"> Thông tư số 59/TT-BTC ngày 23/6/2003 </TT> của Bộ Tài chính hướng dẫn thực hiện <NĐ rel="none"> Nghị định số 60/2003/NĐ-CP ngày 6/6/2003 </NĐ> của Chính phủ và hướng dẫn tại Thông tư này, Chủ tịch UBND tỉnh, thành phố trực thuộc trung ương quy định, hướng dẫn cụ thể cho phù hợp.

Hình 3.6. Văn bản pháp quy được gán nhãn thực thể tham chiếu và quan hệ

Bảng 3.1. Thông tin thống kê về các loại thực thể tham chiếu và số lượng

STT	Loại thực thể tham chiếu	Nhãn	Số lượng thực thể	Số lượng thực thể lồng nhau	Tổng số
1	Hiến pháp	HP	103	0	103
2	Bộ luật	BL	878	82	960
3	Luật	L	19.931	1.226	21.157
4	Nghị định	NĐ	22.901	16	22.917
5	Thông tư	TT	7.027	6	7.033
6	Thông tư liên tịch	TTLT	424	0	424
7	Quyết định	QĐ	4.036	0	4.036
8	Pháp lệnh	PL	3.617	309	3.926
9	Nghị quyết	NQ	890	0	890
	Tổng				**61.446**

Kết quả thống kê cuối cùng về tập dữ liệu được trình bày trong các Bảng 3.1 và Bảng 3.2. Tổng cộng có 61.446 thực thể tham chiếu thuộc 9 loại: Hiến pháp (103), Bộ luật (960), Luật (21.157), Nghị định (22.917), Thông tư (7.033), Thông tư liên tịch (424), Quyết định (4.036), Pháp lệnh (3.926) và Nghị quyết (890). Các thực thể tham chiếu có một trong 7 loại quan hệ sau: Căn cứ (18.540), Dẫn chiếu (27.783),

Hết hiệu lực (1.618), Thay thế (1.765), Sửa đổi hoặc bổ sung (1.203), Hướng dẫn (320) và Không có mối quan hệ (10.217).

Bảng 3.2. Thông tin thống kê về các loại quan hệ và số lượng

STT	Loại quan hệ	Nhãn	Ngữ nghĩa mối quan hệ	Số lượng
1	Căn cứ	CC	Văn bản hiện tại căn cứ theo văn bản tham chiếu	18.540
2	Dẫn chiếu	DaC	Văn bản hiện tại đề cập đến văn bản tham chiếu	27.783
3	Hết hiệu lực	HHL	Văn bản tham chiếu đã hết hiệu lực	1.618
4	Thay thế	BTT	Văn bản hiện tại thay thế văn bản tham chiếu	1.765
5	Sửa đổi hoặc bổ sung	DSD	Văn bản hiện tại sửa đổi, bổ sung văn bản tham chiếu	1.203
6	Hướng dẫn	DHD	Văn bản hiện tại hướng dẫn văn bản tham chiếu	320
7	Không có quan hệ	none	Văn bản hiện tại không có quan hệ với văn bản tham chiếu	10.217
	Tổng			**61.466**

3.4. THỰC NGHIỆM VÀ PHÂN TÍCH KẾT QUẢ

Phần này trình bày thực nghiệm và phân tích kết quả cho trích xuất thực thể tham chiếu và quan hệ trong văn bản pháp quy tiếng Việt, bao gồm hai nhiệm vụ: (1) trích xuất thực thể tham chiếu từ văn bản pháp quy, và (2) phân loại quan hệ giữa các thực thể tham chiếu và văn bản pháp quy đang xem xét.

3.4.1. Thiết lập thực nghiệm

Tập dữ liệu được chia ngẫu nhiên thành 5 phần *(folds)* và thực hiện kiểm tra chéo *(cross-validation tests)*. Hiệu năng của mô hình trích xuất thực thể tham chiếu và quan hệ được đo bằng độ chính xác chung (*accuracy*), và các độ đo: độ chính xác (*precision*), độ phủ (*recall*) và độ đo F_1.

Trích xuất thực thể tham chiếu: Hiệu năng của mô hình trích xuất thực thể tham chiếu được đo bằng độ chính xác, độ phủ và độ đo F_1 cho từng loại thực thể tham

chiếu và cho tất cả các loại thực thể tham chiếu theo các công thức (1.17), (1.18) và (1.19) trong Mục 1.3.4 Chương 1. Trong đó, *A* và *B* tương ứng là tập các thực thể tham chiếu được nhận ra (theo mô hình) và tập các thực thể tham chiếu gốc (đã được gán nhãn) của một loại tham chiếu cụ thể (ví dụ như luật, quyết định, nghị quyết,…).

Phân loại quan hệ giữa các thực thể văn bản pháp quy: Hiệu năng của mô hình phân loại quan hệ giữa các thực thể văn bản pháp quy được đo bằng độ chính xác chung theo công thức (1.16), và các độ đo: độ chính xác, độ phủ và độ đo F_1 cho từng loại quan hệ theo các công thức (1.17), (1.18) và (1.19) trong Mục 1.3.4 Chương 1. Trong đó, *A* và *B* tương ứng là tập các quan hệ được nhận ra (bởi mô hình) và tập các quan hệ gốc (đã được gán nhãn) của một loại quan hệ cụ thể (ví dụ như "căn cứ", "dẫn chiếu",…).

3.4.2. Trích xuất thực thể tham chiếu

a) Các mô hình thực nghiệm

Thực nghiệm được xây dựng nhằm so sánh hiệu năng của các mô hình trích xuất thực thể tham chiếu, như sau:

- **CRF:** Mô hình này sử dụng phương pháp CRF cho quá trình huấn luyện. Nghiên cứu đã tiến hành thử nghiệm với hai biến thể của mô hình: 1) chỉ sử dụng các đặc trưng *n*-grams; và 2) sử dụng cả đặc trưng *n*-grams và bổ sung các đặc trưng thủ công. Thử nghiệm được thực hiện với công cụ CRF++ [58], có sẵn tại: *https://taku910.github.io/crfpp/*.
- **BiLSTM:** Mô hình này sử dụng các mạng BiLSTM để học cách biểu diễn từ và câu, và một lớp softmax để suy diễn. Nghiên cứu cũng đã tiến hành thử nghiệm với hai biến thể của mô hình: 1) chỉ sử dụng các đặc trưng được học tự động bao gồm các nhúng từ và nhúng ký tự; và 2) sử dụng cả các đặc trưng học tự động và bổ sung các đặc trưng thủ công.
- **BiLSTM-CRF:** Mô hình này tương tự như mô hình BiLSTM nhưng sử dụng CRF ở lớp suy diễn thay vì sử dụng các hàm softmax. Nghiên cứu cũng đã tiến

hành thử nghiệm với hai biến thể của mô hình giống như với mô hình BiLSTM.

b) Huấn luyện mạng

Cả hai mô hình BiLSTM và BiLSTM-CRF đều được huấn luyện bằng NCRF++, bộ công cụ gán nhãn chuỗi nơ-ron mã nguồn mở [130]. Kích thước của các véc-tơ nhúng từ và nhúng ký tự tương ứng là 100 và 50, được khởi tạo ngẫu nhiên và cùng học với các mạng. Thực nghiệm đã sử dụng độ dốc ngẫu nhiên tiêu chuẩn *(standard stochastic gradient descent, SGD)* với các lô kích thước 32. Tốc độ học được khởi tạo và cập nhật trên mỗi epoch huấn luyện như đã được đề xuất trong các nghiên cứu trước [128,130]: $\eta_0 = 0{,}003$, $\eta_i = \eta_0/(1 + \rho_i)$, với sự phân rã tỷ lệ $\rho = 0{,}05$ và i biểu thị số epoch đã hoàn thành. Dropout được áp dụng cho các kết quả đầu ra của cả hai giai đoạn biểu diễn từ và biểu diễn câu với tỷ lệ dropout là 0,5 để giảm tình trạng quá khớp *(overfiting)*.

c) Kết quả

Các thử nghiệm cho các mô hình trích xuất tham chiếu được thực hiện trên tập dữ liệu đã được xây dựng ở trên (trong Mục 3.3). Bảng 3.3 trình bày kết quả so sánh hiệu năng của các mô hình trích xuất tham chiếu khác nhau, bao gồm CRF, BiLSTM và BiLSTM-CRF. Quan sát đầu tiên là tất cả các mô hình đều có kết quả khá cao (từ 95,78% đến 96,62% tính theo độ đo F_1). Kết quả như vậy cho thấy các mô hình nghiên cứu lựa chọn là hợp lý với nhiệm vụ. Quan sát thứ hai là, đối với tất cả các mô hình, biến thể sử dụng các đặc trưng thủ công bổ sung cho kết quả trích xuất tốt hơn so với phiên bản chỉ có các đặc trưng cơ bản (*n*-grams hoặc đặc trưng học tự động). Điều này khẳng định tầm quan trọng của các đặc trưng thủ công trong việc trích xuất tham chiếu từ văn bản pháp quy tiếng Việt. Mô hình tốt nhất nghiên cứu đề xuất là BiLSTM-CRF với các đặc trưng thủ công, đạt 96,62% tính theo độ đo F_1, cải thiện 0,60% (giảm tỷ lệ lỗi 15,01%) so với mô hình CRF, và cải thiện 0,39% (giảm tỷ lệ lỗi 10,34%) so với mô hình BiLSTM.

Bảng 3.3. So sánh hiệu năng của các mô hình trích xuất thực thể tham chiếu

Mô hình	**Các biến thể**	**Độ chính xác (%)**	**Độ phủ (%)**	**Độ đo F_1 (%)**
CRF	*n*-grams	95,88	95,93	95,91
CRF	*n*-grams + đặc trưng thủ công	96,02	96,01	96,02
BiLSTM	Đặc trưng học tự động	95,78	95,78	95,78
BiLSTM	Đặc trưng học tự động + đặc trưng thủ công	96,23	96,22	96,23
BiLSTM-CRF	Đặc trưng học tự động	96,47	96,51	96,48
BiLSTM-CRF	Đặc trưng học tự động + đặc trưng thủ công	**96,63**	**96,62**	**96,62**

Bảng 3.4. Hiệu năng của mô hình BiLSTM-CRF trên mỗi loại thực thể tham chiếu được trích xuất

Các loại thực thể	**Độ chính xác (%)**	**Độ phủ (%)**	**Độ đo F_1 (%)**
Hiến pháp	99,14	99,32	99,23
Bộ luật	95,56	93,48	94,51
Luật	97,20	98,04	97,62
Nghị định	97,29	98,36	97,82
Thông tư	96,44	96,44	96,44
Thông tư liên tịch	89,19	92,96	91,03
Nghị quyết	92,12	90,48	91,29
Pháp lệnh	93,59	95,64	94,60
Quyết định	93,33	96,02	94,66

Tiếp theo, nghiên cứu thực hiện đo hiệu năng của mô hình BiLSTM-CRF trên từng loại thực thể tham chiếu. Như được trình bày trong Bảng 3.4, kết quả thu được tương đối tốt trên hầu hết các loại thực thể tham chiếu, thấp nhất là loại "Thông tư liên tịch" (91,03% tính theo độ đo F_1), có tần suất xuất hiện rất ít trong toàn bộ tập

dữ liệu (424 lần). Các loại thực thể tham chiếu khác có kết quả F_1 thấp là "Bộ luật" (94,51%) và "Nghị quyết" (91,29%), đều là các loại thực thể có tần số xuất hiện thấp trong tập dữ liệu. "Hiến pháp" là một ngoại lệ. Mặc dù có tần suất xuất hiện rất ít trong tập dữ liệu (103 lần), nhưng kết quả đạt được độ đo F_1 rất cao (99,23%). Điều này có thể được giải thích là do thực tế số lượng văn bản "Hiến pháp" trong hệ thống văn bản pháp quy là rất nhỏ so với các loại văn bản pháp quy khác, nhưng các thực thể tham chiếu của loại văn bản này có định dạng giống nhau trong hầu hết các câu.

Bảng 3.5. Hiệu năng trên các loại thực thể lồng nhau

Các loại thực thể	**Độ chính xác (%)**	**Độ phủ (%)**	**Độ đo F_1 (%)**
Bộ luật	96,47	97,97	97,16
Luật	90,71	90,22	90,46
Nghị định	73,15	71,28	72,85
Thông tư	0,00	0,00	0,00
Pháp lệnh	78,52	81,69	80,14

Nghiên cứu cũng thực hiện đo hiệu năng của mô hình trích xuất đề xuất trên các thực thể tham chiếu lồng nhau. Như trình bày trong Bảng 3.5, kết quả đạt được rất khả quan, với độ đo F_1 là 97,16% cho "Bộ luật", 90,46% cho "Luật", 72,85% cho "Nghị định" và 80,14% cho "Pháp lệnh", ngoại trừ "Thông tư" (lồng nhau) có tần suất xuất hiện rất ít trong tập văn bản (6 lần).

d) Phân tích lỗi

Một câu hỏi quan trọng là loại thực thể nào thường được gán cho các thực thể tham chiếu trong các câu dự đoán sai. Để trả lời câu hỏi này, nghiên cứu đã tiến hành thống kê về hầu hết các lỗi của từng loại thực thể tham chiếu, được trình bày trong Bảng 3.6. Đối với hầu hết các dự đoán sai, mô hình không thể nhận ra các thực thể tham chiếu hoặc thường nhận nhầm với các loại thực thể "Luật" hoặc "Nghị định" là những loại thực thể xuất hiện nhiều nhất trong tập dữ liệu (Luật: 21.157 và Nghị định:

22.917 lần). Loại văn bản “Thông tư liên tịch” thường được dự đoán là “Thông tư" vì cả hai loại đều bắt đầu bằng cùng một từ trong tiếng Việt.

Bảng 3.6. Thống kê lỗi nhiều nhất theo từng thực thể tham chiếu

Các loại thực thể	**Độ đo F_1 (%)**	**Lỗi nhiều nhất**
Hiến pháp	99,23	Luật
Bộ luật	94,51	Không nhận ra, Luật
Luật	97,62	Không nhận ra, Nghị định
Nghị định	97,82	Không nhận ra, Luật
Thông tư	96,44	Không nhận ra, Nghị định
Thông tư liên tịch	91,03	Không nhận ra, Thông tư
Nghị quyết	91,29	Không nhận ra, Nghị định
Pháp lệnh	94,60	Không nhận ra, Luật
Quyết định	94,66	Không nhận ra, Luật

Như được trình bày trong Bảng 3.3, mô hình tiên tiến với các mạng nơ-ron sâu (BiLSTM-CRF) hoạt động tốt hơn so với mô hình truyền thống (CRF). Thực hiện phân tích và so sánh kết quả đầu ra của hai mô hình cho thấy BiLSTM-CRF có thể trích xuất các thực thể tham chiếu chính xác trong nhiều trường hợp trong khi CRF không thể. Bảng 3.7 cho thấy một số ví dụ về các trường hợp như vậy.

Trong ví dụ đầu tiên, CRF không trích xuất được một tham chiếu loại “Thông tư”. Trong ví dụ thứ hai, CRF nhận ra được một thực thể tham chiếu “Luật” nhưng không trích xuất được toàn bộ tham chiếu (thiếu một số từ). Trong ví dụ thứ ba, CRF trích xuất ra một thực thể tham chiếu có chứa một số từ nhiễu. Trong ví dụ thứ tư, CRF kết hợp hai thực thể tham chiếu “Luật” đứng kề nhau thành một thực thể tham chiếu “Luật” duy nhất. Trong ví dụ cuối cùng, CRF không nhận ra được một câu có chứa thực thể tham chiếu “Nghị định” bên ngoài và thực thể tham chiếu “Luật” bên trong (tham chiếu lồng nhau).

Bảng 3.7. Một số trường hợp mô hình BiLSTM-CRF trích xuất được đúng trong khi mô hình CRF trích xuất sai

STT	BiLSTM-CRF (trích xuất đúng)	CRF (trích xuất sai)	Ghi chú
1	Bộ trưởng Bộ giao thông vận tải ban hành [Thông tư quy định về quản lý tiếp nhận, truyền phát và xử lý thông tin an ninh hàng hải].	Bộ trưởng Bộ giao thông vận tải ban hành Thông tư quy định về quản lý tiếp nhận, truyền phát và xử lý thông tin an ninh hàng hải.	CRF không trích xuất được thực thể tham chiếu.
2	[Luật cải cách ruộng đất] này quy định cho toàn quốc.	[Luật cải cách] ruộng đất này quy định cho toàn quốc.	CRF nhận thiếu từ trong tham chiếu đúng.
3	… [Luật đầu tư nước ngoài tại Việt Nam] là 32%.	… [Luật đầu tư nước ngoài tại Việt Nam là 32%].	CRF nhận ra thực thể tham chiếu có chứa các từ nhiễu.
4	Giấy phép kinh doanh theo quy định của [Luật doanh nghiệp] và [Luật đầu tư].	Giấy phép kinh doanh theo quy định của [Luật doanh nghiệp và Luật đầu tư].	CRF kết hợp 2 thực thể gần nhau thành 1 thực thể.
5	[Nghị định quy định chi tiết và hướng dẫn thi hành một số điều của [Luật nhà ở]];	[Nghị định quy định chi tiết và hướng dẫn thi hành một số điều của] [Luật nhà ở];	CRF không nhận được đúng tham chiếu lồng.

3.4.3. Phân loại quan hệ giữa các thực thể văn bản pháp quy

Mục đích xây dựng các thực nghiệm trong phần này là giải quyết bài toán phân loại quan hệ giữa các thực thể văn bản pháp quy bằng các phương pháp học máy khác nhau. Phần sau sẽ mô tả các thực nghiệm và kết quả.

3.4.3.1. Phân loại quan hệ giữa các thực thể văn bản pháp quy sử dụng học máy truyền thống

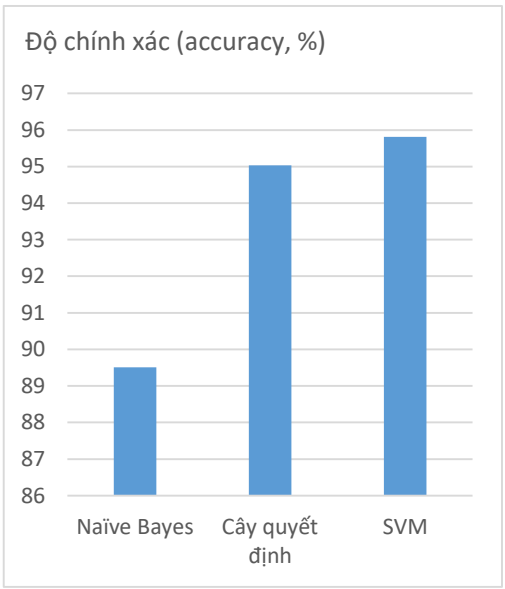
Các thực nghiệm được xây dựng với 3 mục tiêu sau:

- So sánh hiệu năng của các bộ phân loại sử dụng học máy truyền thống.

- So sánh hiệu quả của các phương pháp trích chọn thông tin liên quan đến thực thể khi xây dựng dữ liệu huấn luyện.
- So sánh hiệu quả của các phương pháp trích chọn đặc trưng khi xây dựng mô hình phân loại quan hệ.

a) So sánh hiệu năng của các bộ phân loại

Các thử nghiệm đầu tiên được thực hiện nhằm so sánh hiệu năng của ba bộ phân loại Bayes đơn giản, Cây quyết định (C4.5) và Máy véc-tơ tựa (SVM). Với mỗi phương pháp, nghiên cứu thực hiện các thử nghiệm với từng loại đặc trưng riêng (*n*-grams và TF-IDF), và sau đó thực nghiệm kết hợp các đặc trưng này. Dữ liệu huấn luyện được trích xuất từ các câu có chứa thực thể tham chiếu đã được xác định (thông tin ngữ cảnh gần nhất liên quan với thực thể).



Hình 3.7. So sánh các bộ phân loại khác nhau

Hình 3.7 trình bày kết quả tốt nhất thực nghiệm được với ba bộ phân loại đề xuất. Nhìn chung, cả ba đều có kết quả phân loại quan hệ tương đối tốt, với độ chính xác chung *(accuracy)* đạt trên 89%. Trong đó, phương pháp SVM cho kết quả tốt nhất, có độ chính xác đạt 95,81%. Phương pháp Cây quyết định đạt được độ chính

xác 95,03%. Còn phân loại Bayes đơn giản có độ chính xác kém nhất, đạt 89,51%. Trong các phần sau, nghiên cứu sẽ thực hiện thực nghiệm sử dụng bộ phân loại tốt nhất là SVM.

b) So sánh các phương pháp trích chọn thông tin liên quan đến thực thể

Để phân loại quan hệ giữa thực thể là văn bản đang xem xét với thực thể tham chiếu đã được xác định trong nội dung của văn bản, cần trích chọn một số thông tin liên quan thực thể. Thông tin trích chọn là thông tin về các thực thể và các thông tin ngữ cảnh xung quanh thực thể tham chiếu, bao gồm: thực thể tham chiếu đã xác định trong nội dung (gọi là "*thực thể* B_k"), phần văn bản trong cùng câu ở phía trước thực thể tham chiếu (gọi là "*văn bản trước*"), phần văn bản trong cùng câu ở phía sau thực thể tham chiếu (gọi là "*văn bản sau*"), tên của thực thể văn bản đang xem xét (gọi là "*thực thể A*"), và tên điều khoản (nếu có) của đoạn văn bản chứa thực thể tham chiếu đã được xác định trong nội dung văn bản đang xem xét (gọi là "*điều*"). Bảng 3.8 trình bày một ví dụ về các thông tin được trích chọn trong một đoạn văn bản pháp quy có chứa thực thể tham chiếu, thuộc Nghị định "*Quy định chi tiết thi hành một số điều của pháp lệnh xử lý vi phạm hành chính năm 2002 và pháp lệnh sửa đổi, bổ sung một số điều của pháp lệnh xử lý vi phạm hành chính năm 2008 của Chính phủ*".

Nghiên cứu đề xuất ba phương pháp trích chọn thông tin liên quan đến thực thể được sử dụng khi xây dựng dữ liệu huấn luyện, tương ứng được thực hiện trong 3 thử nghiệm sau:

- ***Thử nghiệm 1:*** Trích chọn thông tin ngữ cảnh gần nhất với thực thể tham chiếu đã được xác định trong nội dung của văn bản, là phần nội dung phía trước và phía sau tham chiếu đó trong cùng câu.
- ***Thử nghiệm 2:*** Trích chọn thông tin về hai thực thể, là thực thể tham chiếu được đề cập và tên của thực thể văn bản pháp quy; và thông tin ngữ cảnh gần nhất với thực thể tham chiếu đã được xác định trong nội dung của văn bản, là phần nội dung phía trước và phía sau tham chiếu đó trong cùng câu.

- ***Thử nghiệm 3:*** Trích chọn thông tin về hai thực thể, là thực thể tham chiếu được đề cập và tên của thực thể văn bản pháp quy; thông tin ngữ cảnh gần nhất với thực thể tham chiếu đã được xác định trong nội dung của văn bản, là phần nội dung phía trước và phía sau tham chiếu đó trong cùng câu; và thông tin ngữ cảnh xa hơn có liên quan đến thực thể tham chiếu đã được xác định trong nội dung của văn bản, là tên điều khoản (nếu có) của đoạn văn bản chứa thực thể tham chiếu đó.

Bảng 3.8. Ví dụ trích chọn thông tin liên quan đến thực thể trong một đoạn văn bản

Thông tin	**Nội dung**
Văn bản hiện tại đang xem xét	Nghị định *Quy định chi tiết thi hành một số điều của pháp lệnh xử lý vi phạm hành chính năm 2002 và pháp lệnh sửa đổi, bổ sung một số điều của pháp lệnh xử lý vi phạm hành chính năm 2008* của Chính phủ
Đoạn văn bản chứa thực thể tham chiếu	**Điều 39. Hiệu lực của Nghị định** Nghị định này có hiệu lực thi hành kể từ ngày 01 tháng 01 năm 2009 và thay thế Nghị định số 134/2003/NĐ-CP ngày 14 tháng 11 năm 2003 quy định chi tiết thi hành một số điều của Pháp lệnh Xử lý vi phạm hành chính năm 2002.
Thực thể A	Nghị định Quy định chi tiết thi hành một số điều của pháp lệnh xử lý vi phạm hành chính năm 2002 và pháp lệnh sửa đổi, bổ sung một số điều của pháp lệnh xử lý vi phạm hành chính năm 2008
Thực thể B_k	Nghị định số 134/2003/NĐ-CP ngày 14 tháng 11 năm 2003
Văn bản trước	Nghị định này có hiệu lực thi hành kể từ ngày 01 tháng 01 năm 2009 và thay thế
Văn bản sau	quy định chi tiết thi hành một số điều của Pháp lệnh Xử lý vi phạm hành chính năm 2002
Điều	Điều 39. Hiệu lực của Nghị định

Bảng 3.9 trình bày tóm tắt các phương pháp trích chọn thông tin liên quan đến thực thể.

Bảng 3.9. Các phương pháp trích chọn thông tin liên quan đến thực thể

Thử nghiệm	Phương pháp trích chọn thông tin
1	Văn bản trước + Văn bản sau
2	Văn bản trước + Thực thể B_k + Văn bản sau + Thực thể A
3	Điều + Văn bản trước + Thực thể B_k + Văn bản sau + Thực thể A

Bảng 3.10. Kết quả phân loại quan hệ với các phương pháp trích chọn thông tin liên quan thực thể (tính theo % độ đo F_1)

Quan hệ	Thử nghiệm 1	Thử nghiệm 2	Thử nghiệm 3
CC	99,05	99,02	99,05
DaC	95,06	95,90	96,13
HHL	76,75	82,23	82,56
BTT	80,77	83,19	83,46
DSD	89,27	89,07	88,62
DHD	65,47	66,51	68,73
none	89,28	90,24	91,07
Trung bình	**94,16**	**94,90**	**95,16**

Bảng 3.10 trình bày kết quả phân loại quan hệ với các phương pháp trích chọn thông tin liên quan thực thể khác nhau. Hiệu năng được đo bằng độ đo F_1 cho từng loại quan hệ. Kết quả trong Bảng 3.10 cho thấy độ chính xác của trích xuất cho từng loại quan hệ tương đối tốt. Kết quả tốt nhất với hầu hết các quan hệ đều đạt trên 82% tính theo độ đo F_1, trừ trường hợp quan hệ "*hướng dẫn*" (DHD) đạt 68,73%. Một trong những lý do là quan hệ DHD có tần số xuất hiện rất ít (và ít hơn nhiều so với các loại quan hệ khác) trong tập dữ liệu, chỉ có 320 lần (trên tổng số 61.466 quan hệ, xem Bảng 3.2). Điều này dẫn đến thiếu dữ liệu học cho mô hình học máy, từ đó làm giảm độ chính xác của dự đoán. Hai loại quan hệ "*căn cứ*" và "*dẫn chiếu*" cho kết

quả cao nhất, lần lượt là 99,05% và 96,13% (tính theo độ đo F_1). Hai loại quan hệ này có tần số xuất hiện nhiều nhất trong tập dữ liệu, tương ứng là 18.540 lần (*căn cứ*) và 27.783 (*dẫn chiếu*).

Về kết quả của ba phương pháp trích chọn thông tin liên quan thực thể được sử dụng để xây dựng dữ liệu huấn luyện, phương pháp thứ ba sử dụng thông tin về hai thực thể (tham chiếu được đề cập và tên của thực thể văn bản pháp quy), phần nội dung phía trước và phía sau thực thể tham chiếu (đã được xác định) trong cùng câu, và tên điều khoản của đoạn văn bản chứa thực thể tham chiếu, đạt được độ chính xác cao nhất so với hai phương pháp còn lại. Kết quả tính trung bình theo độ đo F_1, phương pháp thứ nhất đạt được 94,16%, phương pháp thứ hai đạt 94,90%, và phương pháp thứ ba đạt 95,16%. Cụ thể, phương pháp thứ ba có 6 (trên tổng số 7) loại quan hệ có kết quả trích xuất chính xác tốt hơn hai phương pháp còn lại. Đặc biệt, phương pháp phương pháp thứ ba có hiệu quả trích xuất tốt hơn hẳn với các quan hệ có số mẫu ít trong tập dữ liệu, như HHL tăng 5,81%, DHD tăng 3,26%, hay BTT tăng 2,69% (tính theo độ đo F_1), so với phương pháp thứ nhất chỉ dựa trên thông tin phần nội dung phía trước và phía sau thực thể tham chiếu trong cùng câu.

c) So sánh các phương pháp trích chọn đặc trưng

Để thực nghiệm với các phương pháp trích chọn đặc trưng khác nhau, nghiên cứu sử dụng phương pháp học máy SVM với dữ liệu huấn luyện được xây dựng theo phương pháp trích chọn thông tin liên quan thực thể thứ ba trong Bảng 3.9. Phương pháp này sử dụng thông tin về hai thực thể, là tham chiếu được đề cập và tên của thực thể văn bản pháp quy; thông tin ngữ cảnh gần nhất với thực thể tham chiếu đã được xác định trong nội dung của văn bản, là phần nội dung phía trước và phía sau tham chiếu đó trong cùng câu; và thông tin ngữ cảnh xa hơn có liên quan đến thực thể tham chiếu đã được xác định trong nội dung của văn bản, là tên điều khoản (nếu có) của đoạn văn bản chứa thực thể đó. Nghiên cứu đề xuất hai phương pháp trích chọn đặc trưng cho các thử nghiệm, đó là đặc trưng *n*-grams, và kết hợp đặc trưng *n*-grams với đặc trưng TF-IDF. Mỗi loại văn bản pháp quy thường có từ khóa riêng, ví dụ văn bản

là Nghị định, Luật, Thông tư,… Do vậy, việc sử dụng đặc trưng thể hiện mức độ quan trọng của từ trong văn bản, như TF-IDF, sẽ làm tăng khả năng trích xuất thông tin từ văn bản pháp quy.

Bảng 3.11. Kết quả phân loại quan hệ với các phương pháp trích chọn đặc trưng (%)

Quan hệ	***n*-grams + TF-IDF**			***n*-grams (F_1)**
	Độ chính xác	**Độ phủ**	**Độ đo F_1**	
CC	99,70	98,50	99,10	99,05
DaC	94,36	98,57	96,42	96,13
HHL	89,16	78,68	83,28	82,56
BTT	96,29	76,96	85,46	83,46
DSD	91,85	86,31	88,94	88,62
DHD	93,37	54,94	68,87	68,73
none	93,35	90,98	92,15	91,07
Trung bình	**95,68**	**95,67**	**95,57**	**95,16**

Bảng 3.11 trình bày kết quả thực nghiệm với các phương pháp trích chọn đặc trưng đã đề xuất. Kết quả trích xuất được đo trên từng quan hệ theo độ chính xác, độ bảo phủ và độ đo F_1. Có thể thấy, việc kết hợp đặc trưng *n*-grams và TF-IDF cho kết quả phân loại quan hệ giữa các thực thể văn bản pháp quy tốt hơn khi chỉ sử dụng đặc trưng *n*-grams. Tính trung bình, phương pháp kết hợp đặc trưng *n*-grams và TF-IDF đạt được độ chính xác là 95,68%, độ phủ là 95,67% và độ đo F_1 là 95,57%. So với phương pháp trích chọn đặc trưng chỉ sử dụng *n*-grams, phương pháp kết hợp đặc trưng *n*-grams và TF-IDF đạt kết quả cao hơn 0,41% tính theo độ đo F_1.

d) Phân tích lỗi

Các lỗi được chia thành hai loại, đó là FP (dương tính giả) và FN (âm tính giả). Lỗi FP đề cập tới việc một mối quan hệ khác bị nhận nhầm thành một quan hệ đang quan tâm, còn lỗi FN đề cập đến việc một quan hệ đang quan tâm bị nhận nhầm thành

một quan hệ khác. Để phân tích lỗi, Bảng 3.12 được xây dựng với thống kê về các giá trị của tỉ lệ FP (FPR) và tỉ lệ FN (FNR), tương ứng đại diện cho tỉ lệ nhận nhầm và tỉ lệ bỏ sót của các loại quan hệ được trích xuất, và các lỗi chính tương ứng (các quan hệ là nguyên nhân gây ra lỗi chính). Tỉ lệ bỏ sót trả lời được cho câu hỏi là các quan hệ trong các câu dự đoán sau thường bị gán nhầm thành các loại nhãn nào. Như trình bày trong Bảng 3.12, FNR khá thấp nên ở đây chỉ tập trung phân tích cho FPR. Nghĩa là trả lời cho câu hỏi là loại nhãn nào thường được gán cho các quan hệ trong các câu dự đoán sai.

Bảng 3.12. Phân tích lỗi phân loại quan hệ

Quan hệ	F_1 **(%)**	**FPR (%)**	**FNR (%)**	**Các lỗi chính**
CC	99,10	1,44	0,39	DaC, none
DaC	96,42	1,28	5,63	none, HHL
HHL	83,28	25,67	5,00	DaC, BTT
BTT	85,46	14,38	3,59	None
DSD	88,94	8,86	6,33	None
DHD	68,87	42,25	1,41	DaC
none	92,15	8,67	5,91	DaC, CC

Đối với hầu hết các dự đoán sai kiểu FP, mô hình không thể nhận ra các quan hệ CC, DaC và *none*, xuất hiện nhiều nhất trong tập dữ liệu với lần lượt là 18.540, 27.783 và 10.217 lần. Các quan hệ này bị nhận nhầm tạo nên 3 giá trị FPR cao nhất trong bảng, cho 3 nhãn là DHD, HHL, BTT, kéo theo độ chính xác trung bình của mô hình bị giảm xuống khá nhiều. Cụ thể, quan hệ DaC gây ra ảnh hưởng lớn tới quan hệ DHD, khiến cho số lỗi sai FP có tỉ lệ lên tới 42,25%. Thực tế số lỗi nhận nhầm thành DHD là không nhiều nhưng nghiêm trọng do số mẫu quan hệ DHD ít hơn rất nhiều so với các quan hệ khác. Tương tự, DaC cũng bị nhận nhầm sang HHL và cũng gây ra tỉ lệ lỗi sai FP cao. Quan hệ BTT cũng có tỉ lệ lỗi FP cao do quan hệ *none* bị nhận nhầm thành BTT. Quan hệ *none* cũng bị nhận nhầm thành DSD khá nhiều, còn DaC và CC lại bị nhận nhầm thành *none*. Thống kê trên bảng cũng phản ánh đúng độ khó trong việc phân biệt của 3 quan hệ có số lượng mẫu lớn nhất là CC,

DaC và *none*. CC chỉ có tỉ lệ bỏ sót (FNR) bằng 0,39%, trong khi DaC và *none* đều trên 5%.

Như vậy, để làm tăng độ chính xác của mô hình phân loại quan hệ thì cần phải xây dựng các đặc trưng phân biệt rõ các quan hệ hiện có, trong đó cần tập trung nhiều nhất vào các quan hệ DaC với DHD và HHL; BTT và HHL; và *none* với DaC, CC, BTT (xem Bảng 3.12). Khảo sát cụ thể các câu có lỗi sai dạng FP vì nhận nhầm từ các quan hệ DaC, BTT cho thấy, nhiều câu bị nhận nhầm do trong câu có một số các từ thường thấy trong đặc trưng đại diện cho quan hệ gây nên sự nhầm lẫn. Ví dụ như trong hai trường hợp sau:

- Trường hợp 1: "*Điều 2. Đổi các cụm từ "Bộ Nội vụ" quy định tại Nghị định số 51/CP ngày 10 tháng 5 năm 1997 của Chính phủ thành cụm từ " Bộ Công an "* ." chứa từ "*quy định tại*" dễ gây nhầm từ DSD sang DaC.
- Trường hợp 2: "*2. Kể từ ngày Thông tư này có hiệu lực thi hành, các quy định về cấp Giấy phép, tổ chức và hoạt động tại Thông tư số 02/2008/TT-NHNN ngày 02/4/2008 của Thống đốc Ngân hàng Nhà nước hướng dẫn thực hiện Nghị định số 28/2005/NĐ-CP ngày 09/3/2005 của Chính phủ về tổ chức và hoạt động của tổ chức tài chính quy mô nhỏ tại Việt Nam và Nghị định số 165/2007/NĐ-CP ngày 15/11/2007 của Chính phủ sửa đổi, bổ sung, bãi bỏ một số điều của Nghị định số 28/2005/NĐ-CP ngày 09/3/2005 của Chính phủ về tổ chức và hoạt động của tổ chức tài chính quy mô nhỏ tại Việt Nam hết hiệu lực thi hành* ." gây nhầm từ HHL thành BTT.

3.4.3.2. Phân loại quan hệ giữa các thực thể văn bản pháp quy sử dụng mô hình dựa trên học sâu

Mô hình đề xuất cho thử nghiệm cho phân loại quan hệ là sử dụng các mạng BiLSTM để học cách biểu diễn từ, và một lớp softmax để suy diễn. Lớp biểu diễn từ tạo ra cách biểu diễn từ với kỹ thuật nhúng từ. Dựa trên kiến trúc mạng nơ-ron sâu này, mô hình có thể tự động học đặc trưng từ dữ liệu đầu vào.

Trong mô hình học sâu đề xuất, kích thước của các véc-tơ nhúng từ là 100, được khởi tạo ngẫu nhiên và cùng học với các mạng. Mỗi lớp BiLSTM có 100 unit. Thực nghiệm đã sử dụng bộ tối ưu RMSprop với các lô kích thước 32. Tốc độ học được khởi tạo và cập nhật trên mỗi epoch huấn luyện như đã được đề xuất trong các nghiên cứu trước [128,130]: $\eta_0 = 0{,}003$, $\eta_i = \eta_0/(1 + \rho_i)$, với sự phân rã tỷ lệ $\rho = 0{,}05$ và i biểu thị số epoch đã hoàn thành. Dropout được áp dụng cho các kết quả đầu ra của cả hai giai đoạn biểu diễn từ và biểu diễn câu với tỷ lệ dropout là 0,2 để giảm tình trạng quá khớp *(overfiting)*.

Bảng 3.13. Kết quả phân loại quan hệ với mô hình BiLSTM (%)

Quan hệ	***BiLSTM***			***SVM*** ***(F_1)***
	Độ chính xác	**Độ phủ**	**Độ đo *F_1***	
CC	99,05	98,89	98,97	99,10
DaC	97,42	98,09	97,76	96,42
HHL	88,19	84,15	86,13	83,28
BTT	88,09	93,67	90,79	85,46
DSD	94,72	91,27	92,96	88,94
DHD	71,23	80,00	75,36	68,87
none	96,03	94,21	95,11	92,15
Trung bình	**97,03**	**97,03**	**97,03**	**95,57**

Bảng 3.13 trình bày kết quả thực nghiệm phân loại quan hệ giữa các thực thể văn bản pháp quy với mô hình BiLSTM đề xuất. Kết quả phân loại được đo trên từng quan hệ theo độ chính xác, độ bảo phủ và độ đo F_1. Có thể thấy, việc phân loại quan hệ sử dụng học sâu dựa trên mô hình BiLSTM cho kết quả tốt hơn so với phương pháp học máy truyền thống tốt nhất (SVM, cột cuối cùng trong Bảng 3.13). Tính trung bình, phương pháp phân loại dựa trên BiLSTM đạt được độ chính xác là 97,03%, độ phủ là 97,03% và độ đo F_1 là 97,03%. So với phương pháp sử dụng SVM, phương pháp dựa trên BiLSTM đạt kết quả cao hơn 1,46% tính theo độ đo F_1.

3.5. KẾT LUẬN CHƯƠNG 3

Chương 3 đã trình bày nghiên cứu luận án về trích xuất các thông tin là thực thể tham chiếu và quan hệ trong văn bản pháp quy tiếng Việt, là những nhiệm vụ quan trọng trong xử lý văn bản pháp quy. Nội dung chương trình bày nghiên cứu đề xuất một số phương pháp học máy cho hai nhiệm vụ riêng: (1) trích xuất thực thể tham chiếu và (2) phân loại quan hệ giữa các thực thể trong văn bản pháp quy. Với nhiệm vụ thứ nhất, trích xuất thực thể tham chiếu, nghiên cứu đề xuất một số mô hình trích xuất, bao gồm cả các phương pháp truyền thống (như CRF) và các phương pháp tiên tiến hơn với mạng nơ-ron (BiLSTM và BiLSTM-CRF). Nghiên cứu đã sử dụng cả các đặc trưng học tự động và các đặc trưng được trích chọn thủ công cho nhiệm vụ. Với nhiệm vụ thứ hai, phân loại quan hệ giữa thực thể là tham chiếu với thực thể là văn bản pháp quy hiện tại đang xem xét, nghiên cứu đề xuất sử dụng cả phương pháp học máy truyền thống, với các đặc trưng phù hợp được trích chọn (dựa trên sự kết hợp của các thông tin về các thực thể cùng các thông tin ngữ cảnh liên quan), và phương pháp sử dụng học sâu (dựa trên mô hình BiLSTM) để giải quyết nhiệm vụ.

Nội dung chương cũng đã trình bày việc xây dựng tập dữ liệu hơn 5000 văn bản pháp quy tiếng Việt, với các thực thể và mối quan hệ giữa các thực thể được gán nhãn, để sử dụng cho một loạt các thử nghiệm nhằm giải quyết bài toán cũng như đánh giá hiệu năng cho các phương pháp đã đề xuất. Kết quả cho thấy các mô hình đề xuất đều có kết quả khả quan với cả hai nhiệm vụ trích xuất thực thể tham chiếu và phân loại quan hệ giữa các thực thể trong văn bản pháp quy. Với nhiệm vụ trích xuất thực thể tham chiếu, bằng cách sử dụng phương pháp kết hợp BiLSTM-CRF với tập đặc trưng phong phú, mô hình đề xuất đã đạt được kết quả ấn tượng với độ đo F_1 đạt hơn 95% cho trích xuất cả các thực thể tham chiếu bên ngoài và lồng nhau. Với nhiệm vụ phân loại quan hệ giữa các thực thể văn bản pháp quy, phương pháp dựa trên mô hình BiLSTM cho kết quả với hầu hết các quan hệ đều đạt độ đo F_1 trên 86%. Trong đó, hầu hết các quan hệ có tần số xuất hiện càng nhiều trong tập dữ liệu thì đạt độ chính xác càng cao.

CHƯƠNG 4. TRÍCH XUẤT KẾT HỢP ĐỒNG THỜI THỰC THỂ VÀ QUAN HỆ TRONG VĂN BẢN PHÁP QUY TIẾNG VIỆT SỬ DỤNG PHƯƠNG PHÁP HỌC SÂU

Việc trích xuất thông tin về thực thể và quan hệ trong các bài toán trích xuất thông tin thường được thực hiện theo phương pháp trích xuất tuần tự, các thông tin được trích xuất riêng rẽ: đầu tiên là trích xuất thực thể và sau đó mới thực hiện trích xuất mối quan hệ giữa các thực thể. Phương pháp này có ưu điểm là dễ dàng thực hiện, tuy nhiên có nhược điểm là dễ lan truyền lỗi (khi trích xuất thực thể sai dẫn đến trích xuất quan hệ sẽ bị sai) và trong nhiều trường hợp có thể bỏ qua mối liên hệ (hay mối ràng buộc) giữa các thông tin về thực thể và quan hệ trong các bài toán thực tế.

Nghiên cứu trong Chương 4 khắc phục các nhược điểm kể trên của phương pháp trích xuất thông tin thực thể và quan hệ theo cách tuần tự. Nghiên cứu ở đây sẽ đề xuất xây dựng một mô hình học chung để giải quyết đồng thời cả hai nhiệm vụ trích xuất thực thể và quan hệ trong lĩnh vực văn bản pháp quy. Mô hình đề xuất sử dụng kiến trúc bộ mã hóa-giải mã dựa trên Transformer với cơ chế giải mã không tự hồi quy (*non-autoregressive decoding*) cho phép giảm bớt tính tuần tự của các mô hình seq2seq truyền thống và có khả năng trích xuất các thực thể và quan hệ đồng thời. Nghiên cứu cũng đề xuất một phương pháp nhằm tăng cường đầu vào bộ giải mã với thông tin có ý nghĩa có thể học được và do đó, cải thiện độ chính xác của mô hình. Kết quả thực nghiệm trên bộ dữ liệu đã được sử dụng trong nghiên cứu Chương 3 (gồm 5.031 văn bản pháp quy tiếng Việt) cho thấy mô hình đề xuất cho kết quả tốt, đạt 99,4% tính theo độ đo F_1 cho nhiệm vụ trích xuất đồng thời cả hai thông tin.

Chương 4 trình bày tổng hợp kết quả nghiên cứu của luận án dựa trên các công trình nghiên cứu số [2, 3] của tác giả (Theo danh mục các công trình công bố), bao gồm các nội dung sau:

- Đề xuất một mô hình trích xuất kết hợp sử dụng kiến trúc bộ mã hóa-giải mã dựa trên Transformer với cơ chế giải mã song song không tự hồi quy để trích xuất đồng thời các thực thể tham chiếu và quan hệ từ các văn bản pháp quy. Để cải thiện hiệu năng của mô hình trích xuất kết hợp, nghiên cứu giới thiệu phương pháp tăng cường đầu vào bộ giải mã bằng cách sử dụng các thông tin đầu mối quan trọng của văn bản tham chiếu.
- Thực hiện các thử nghiệm trên tập dữ liệu đã được xây dựng và phân tích kết quả thử nghiệm, so sánh tính hiệu quả của phương pháp đề xuất với một số mô hình cơ sở đã đạt được kết quả tốt.

4.1. ĐẶT VẤN ĐỀ

Trong những năm gần đây, các nhà nghiên cứu đã phát triển một số mô hình học chung (mô hình kết hợp) và đã đạt được hiệu quả tốt hơn trên các nhiệm vụ xử lý ngôn ngữ tự nhiên và xử lý văn bản pháp quy khác nhau, như trích xuất thực thể và quan hệ [34,80,108,121,125,132]; phát hiện ý định người dùng *(intent detection)* và điền thông tin vào vị trí *(slot filling)* [48,96,110]; các bài toán NLP kết hợp cơ bản [86]; nhận dạng thực thể và liên kết thực thể [72]; trích xuất thông tin từ văn bản quy phạm pháp luật [25].

Như đã đề cập đến trong nội dung Chương 3 của luận án, việc trích xuất được các thông tin về thực thể tham chiếu và quan hệ giữa các thực thể trong văn bản pháp quy là một phần quan trọng trong các hệ thống xử lý văn bản pháp quy tự động nhằm hỗ trợ tốt hơn cho người dùng. Trong đó, việc xác định được thực thể tham chiếu là một yêu cầu cần thiết để nhận ra mối quan hệ giữa các văn bản và các phần của văn bản, đồng thời cũng có thể sử dụng cho các bài toán khác; việc xác định được mối quan hệ giữa các thực thể giúp người dùng thuận tiện trong việc tìm kiếm, tra cứu, phân tích, hay truy vấn nội dung văn bản pháp quy. Nghiên cứu trong Chương 3 đề xuất phương pháp trích xuất các thông tin về thực thể tham chiếu và quan hệ giữa các thực thể văn bản pháp quy theo cách tuần tự, đầu tiên (1) trích xuất thực thể tham chiếu, và sau đó (2) phân loại quan hệ giữa thực thể tham chiếu đã được trích xuất và

thực thể văn bản đang xem xét. Phương pháp này dễ thực hiện do tách bài toán thành hai nhiệm vụ trích xuất thực thể tham chiếu và phân loại quan hệ riêng rẽ. Tuy nhiên, thực tế có thể thấy, với phương pháp trích xuất tuần tự có thể dẫn đến việc lan truyền lỗi trích xuất thông tin, nghĩa là khi xác định thực thể tham chiếu hoặc loại thực thể tham chiếu sai sẽ dẫn đến xác định mối quan hệ giữa thực thể tham chiếu này và thực thể văn bản đang xem xét bị sai. Mặt khác, việc xác định mối quan hệ giữa các thực thể tham chiếu có thể liên quan đến loại thực thể: ví dụ một nghị định thường thay thế một nghị định khác, không phải là luật, hoặc nghị định thường căn cứ dựa trên luật, nhưng điều ngược lại là không đúng. Như vậy, về bản chất có thể thấy hai nhiệm vụ trích xuất thực thể tham chiếu và phân loại quan hệ giữa các thực thể trong văn bản pháp quy có sự liên quan và có chia sẻ thông tin chung với nhau.

Nghiên cứu trong Chương 4 khắc phục các vấn đề kể trên trong phương pháp trích xuất thông tin thực thể và quan hệ theo cách tuần tự bằng cách đề xuất xây dựng một mô hình trích xuất kết hợp, sử dụng các kết quả gần đây trong nghiên cứu học sâu, để xử lý đồng thời cả hai nhiệm vụ con trích xuất thực thể tham chiếu và xác định quan hệ giữa các thực thể trong văn bản pháp quy.

Đóng góp của nghiên cứu được trình bày trong Chương 4 *là đề xuất phương pháp trích xuất kết hợp thực thể và quan hệ trong văn bản pháp quy tiếng Việt sử dụng mô hình dựa trên học sâu.* Mô hình trích xuất kết hợp sử dụng kiến trúc bộ mã hóa-giải mã dựa trên Transformer với cơ chế giải mã song song không tự hồi quy để trích xuất đồng thời các thực thể tham chiếu và quan hệ trong văn bản pháp quy. Ưu điểm là mô hình không chỉ độc lập với thứ tự của các thực thể tham chiếu mà còn có thể trích xuất các thực thể tham chiếu lồng nhau. Để cải thiện hiệu năng của mô hình trích xuất kết hợp, nghiên cứu sử dụng phương pháp tăng cường đầu vào bộ giải mã với các thông tin đầu mối quan trọng của văn bản tham chiếu. Kết quả thử nghiệm trên tập dữ liệu đã được xây dựng (trong Chương 3 luận án, bao gồm 5.031 văn bản pháp quy tiếng Việt) cho thấy phương pháp đề xuất có hiệu quả tốt hơn so với một số mô hình đã đạt được kết quả tốt trong các nghiên cứu trước đây, đặc biệt trên cả các câu phức tạp có nhiều thực thể tham chiếu.

Nội dung còn lại của Chương 4 được cấu trúc như sau. Mục 4.2 mô tả mô hình trích xuất đề xuất kết hợp thực thể và quan hệ trong văn bản pháp quy, bao gồm kiến trúc của mô hình và các thành phần. Các kết quả thực nghiệm được trình bày trong Mục 4.3. Cuối cùng, Mục 4.4 là kết luận chương.

4.2. ĐỀ XUẤT MÔ HÌNH TRÍCH XUẤT KẾT HỢP THỰC THỂ VÀ QUAN HỆ

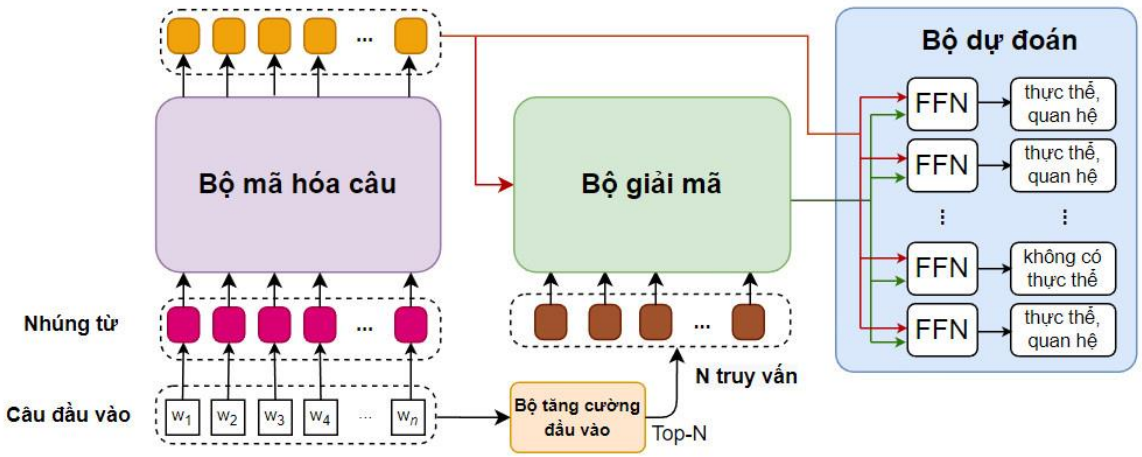
Phần này trình bày đề xuất phương pháp trích xuất kết hợp thực thể tham chiếu và quan hệ trong văn bản pháp quy tiếng Việt, bao gồm kiến trúc mô hình, chi tiết các thành phần chính và các thủ tục huấn luyện trích xuất kết hợp. Một số ký hiệu được sử dụng như sau: các ma trận được ký hiệu bởi chữ hoa in đậm (ví dụ: **W**, **H**), các véc-tơ được ký hiệu bởi chữ thường in đậm (ví dụ: **h**, $\mathbf{v}_i$), và các đại lượng vô hướng được ký hiệu bởi chữ thường in nghiêng (ví dụ: *k*, α_i).

4.2.1. Kiến trúc mô hình

Cho một văn bản pháp quy *x*, mục tiêu của mô hình đề xuất là trích xuất: (1) thực thể tham chiếu, là các đoạn (nhóm) từ cùng với nhãn thực thể tham chiếu, trong *x*; và (2) mối quan hệ giữa từng thực thể tham chiếu đã được trích xuất với *x*. Do thực tế là mỗi thực thể tham chiếu thường nằm trong một câu và mối quan hệ có thể được xác định thông qua ngữ cảnh của câu, nên mô hình đề xuất sẽ thực hiện xử lý theo từng câu *s* (được biểu diễn dưới dạng một chuỗi *n* từ $s = t_1 t_2 \dots t_n$) trong văn bản *x*. Đầu ra của mô hình bao gồm *m* bộ ba (không có thứ tự), mỗi bộ ba tương ứng với một thực thể tham chiếu theo mẫu $(r_{start}, r_{end}, rel)$, trong đó r_{start} và r_{end} biểu thị vị trí bắt đầu/kết thúc của thực thể tham chiếu trong câu đầu vào và *rel* là một nhãn được kết hợp bởi một loại thực thể tham chiếu và một loại quan hệ "*reference_type/relation_type*".

Hình 4.2 trình bày kiến trúc tổng thể của mô hình bao gồm bốn thành phần chính: bộ mã hóa câu, bộ tăng cường đầu vào, bộ giải mã và bộ dự đoán.

- Bộ mã hóa câu: Bộ mã hóa câu sử dụng mô hình ngôn ngữ được huấn luyện trước để tạo ra các biểu diễn từ theo ngữ cảnh.
- Bộ tăng cường đầu vào: Như đã chỉ ra trong các nghiên cứu trước đây, chất lượng của đầu vào bộ giải mã ảnh hưởng lớn đến độ chính xác của mô hình. Do đó, thay vì sử dụng cùng một truy vấn ngẫu nhiên cho tất cả các câu đầu vào, bộ tăng cường đầu vào ở đây học *N* truy vấn, mỗi truy vấn chứa các đầu mối quan trọng về thực thể tham chiếu có thể có trong câu đầu vào. Ở đây, *N* là một siêu tham số, được đặt lớn hơn số lượng thực thể tham chiếu tối đa trong một câu.
- Bộ giải mã: Bộ giải mã lấy đầu ra của bộ mã hóa câu và bộ tăng cường đầu vào làm đầu vào và tạo ra *N* phần nhúng đầu ra, sau đó được đưa vào bộ dự đoán để trích xuất các thực thể tham chiếu và quan hệ.
- Bộ dự đoán: Bộ dự đoán sử dụng các mạng nơ-ron truyền xuôi *(feed forward networks)* để xác định thực thể tham chiếu và mối quan hệ (hoặc xác định là không có thực thể tham chiếu) từ mỗi nhúng đầu ra của bộ giải mã.



Hình 4.1. Minh họa kiến trúc của mô hình đề xuất

Phần sau đây sẽ mô tả chi tiết các thành phần của mô hình.

4.2.2. Bộ mã hóa câu

Bộ mã hóa câu trong nghiên cứu sử dụng mô hình ngôn ngữ được huấn luyện trước dựa trên Transformer như BERT [31], để tạo ra các nhúng từ theo ngữ cảnh từ các nhúng tĩnh và nhúng theo vị trí:

$$\mathbf{c}_i = PretrainedLM\ (t_{1:n}, i), \tag{4.1}$$

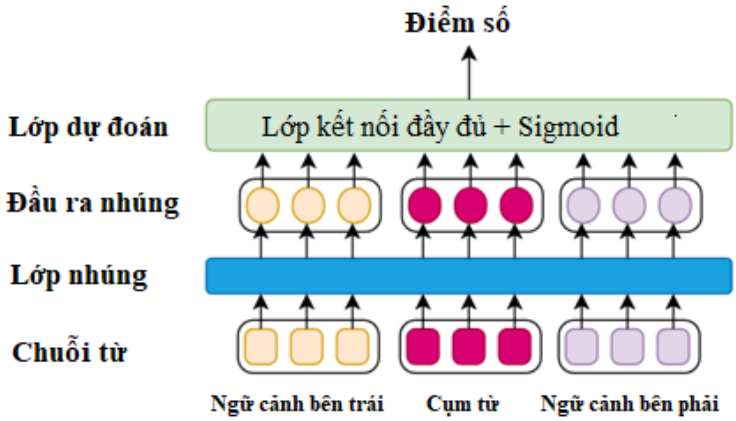
trong đó $t_{1:n}$ biểu thị chuỗi đầu vào có n từ và $\mathbf{c}_i \in \mathbb{R}^{1\times d}$ $(1 \leq i \leq n)$ là nhúng từ theo ngữ cảnh của từ thứ i, nghĩa là t_i, và d biểu thị kích thước nhúng.

4.2.3. Bộ tăng cường đầu vào

Việc xác định cụm từ bắt đầu của một thực thể tham chiếu (gọi tắt là các cụm từ bắt đầu) đóng vai trò quan trọng trong việc giải quyết nhiệm vụ, không những để trích xuất thực thể tham chiếu mà còn xác định loại thực thể tham chiếu và loại quan hệ. Do thực tế là một số cụm từ có xu hướng xuất hiện ở phần đầu của thực thể tham chiếu thường xuyên hơn các cụm từ khác, nên nghiên cứu ở đây đề xuất một phương pháp tăng cường đầu vào bộ giải mã nhằm kiểm tra các cụm từ trong câu đầu vào và ước lượng về khả năng mà mỗi cụm từ có thể là cụm từ bắt đầu của thực thể tham chiếu. Các cụm từ bắt đầu có khả năng cao nhất sẽ được sử dụng để cung cấp các gợi ý quan trọng cho bộ giải mã trong việc trích xuất các thực thể tham chiếu và quan hệ.

Có một số phương pháp xây dựng bộ tăng cường đầu vào. Một phương pháp đơn giản là xây dựng một từ điển chứa các cụm từ bắt đầu thường xuyên nhất của từng loại thực thể tham chiếu. Bộ tăng cường đầu vào tìm kiếm các cụm từ trong câu đầu vào xuất hiện trong từ điển và coi chúng là các cụm từ bắt đầu của văn bản tham chiếu tiềm năng. Ngoài cách tiếp cận dựa trên từ điển như trên, nghiên cứu cũng đề xuất một phương pháp linh hoạt hơn dựa trên phân loại, có xét đến ngữ cảnh của câu đầu vào. Cụ thể, bộ tăng cường đầu vào lấy một cụm từ bao gồm m từ liên tiếp cùng với ngữ cảnh của nó, nghĩa là các từ xung quanh, làm đầu vào và trả về khả năng nó là cụm từ bắt đầu.

Hình 4.3 minh họa kiến trúc của bộ tăng cường đầu vào. Đầu vào bao gồm ba thành phần: một cụm từ (*m* từ liên tiếp), ngữ cảnh bên trái (*q* từ ở phía bên trái) và ngữ cảnh bên phải (*q* từ ở phía bên phải). Như vậy, đầu vào có thể được coi là một chuỗi (*m+2q*) từ, trong đó *m* và *q* là các siêu tham số. Đầu vào được đưa vào một lớp nhúng và sau đó là một lớp kết nối đầy đủ (*Fully-connected layer*) với hàm sigmoid để tạo ra một giá trị điểm số. Điểm số càng cao thì cụm từ càng có nhiều khả năng là cụm từ bắt đầu.



Hình 4.2. Bộ tăng cường đầu vào

Bộ tăng cường đầu vào được huấn luyện độc lập với mô hình trích xuất kết hợp. Tập dữ liệu huấn luyện bao gồm các mẫu dương (+) và mẫu âm (-) như sau.

- Mẫu dương: Mỗi mẫu dương được tạo ra bằng cách sử dụng cụm từ bắt đầu của thực thể tham chiếu.
- Mẫu âm: Về lý thuyết, tất cả các cụm từ ngoại trừ cụm từ bắt đầu đều có thể được sử dụng để tạo mẫu âm. Tuy nhiên, để giải quyết vấn đề mất cân bằng dữ liệu, nghiên cứu đề xuất một chiến lược lấy mẫu âm, trong đó chỉ chọn ngẫu nhiên 2 mẫu âm cho mỗi mẫu dương. Trong 2 mẫu âm, 1 mẫu nằm bên trong và 1 mẫu nằm bên ngoài thực thể tham chiếu. Lưu ý là vị trí tương đối của cụm từ (bên trong/bên ngoài) đối với thực thể tham chiếu được xác định dựa trên từ đầu tiên của cụm từ.

Bộ tăng cường đầu vào đã được huấn luyện sau đó được sử dụng trong cả giai đoạn huấn luyện và giai đoạn dự đoán của mô hình trích xuất kết hợp. Cho một câu đầu vào được biểu thị dưới dạng một chuỗi các từ, bộ tăng cường đầu vào tiến hành thực hiện các bước sau:

- Trích xuất tất cả các ứng cử viên cụm từ có thể có trong câu đầu vào. Mỗi ứng viên là một chuỗi con của *m* từ liên tiếp.
- Xếp hạng các ứng cử viên cụm từ và chọn *N* cụm từ có xác suất cao nhất (*Top-N*) là cụm từ bắt đầu.
- Đối với mỗi cụm từ đã chọn, truy xuất từ đầu tiên.
- Trả về các nhúng theo vị trí của *N* từ đã truy xuất: $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{N}$, $\mathbf{P} \in \mathbb{R}^{N \times d}$, trong đó $\mathbf{p}_i \in \mathbb{R}^{1 \times d}$ là nhúng theo vị trí của từ đầu tiên của cụm từ được chọn thứ *i*.

Các nhúng vị trí của từ đầu tiên của các cụm từ được chọn sẽ cung cấp các gợi ý quan trọng cho bộ giải mã trong trích xuất thực thể tham chiếu và quan hệ.

4.2.4. Bộ giải mã

Bộ giải mã ở đây sử dụng kiến trúc Transformer chuẩn [117] với *K* lớp transformers giống hệt nhau. Bằng cách sử dụng cơ chế giải mã song song không tự hồi quy *(non-autoregressive decoding mechanism)*, mô hình đề xuất sẽ giải mã một tập bao gồm *N* bộ ba cho cả thực thể tham chiếu và quan hệ đồng thời tại mỗi lớp giải mã thay vì giải mã tuần tự. *N* là một siêu tham số (*N* lớn hơn số tham chiếu tối đa trong một câu). Không giống như các mô hình bộ mã hóa-giải mã trước đây [108] sử dụng truy vấn cố định cho tất cả đầu vào, nghiên cứu này coi đầu ra của bộ tăng cường đầu vào là các truy vấn linh hoạt thay đổi theo câu đầu vào và cung cấp gợi ý về thực thể tham chiếu.

Bộ giải mã lấy chuỗi *n* nhúng từ $\mathbf{c}_i (1 \leq i \leq n)$ và *N* nhúng vị trí $\mathbf{p}_i (1 \leq i \leq N)$ làm đầu vào, biến đổi chúng và tạo ra *N* véc-tơ nhúng đầu ra. Các véc-tơ nhúng đầu ra này sẽ được đưa vào một bộ dự đoán, trong đó mỗi véc-tơ nhúng sẽ được giải mã thành một bộ ba đại diện cho một thực thể tham chiếu và mối quan hệ của nó.

$$\mathbf{H} = \mathbf{Decoder}(\mathbf{c}_{1:n}, \mathbf{p}_{1:N}) \quad (4.2)$$

trong đó $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^{N}$, $\mathbf{H} \in \mathbb{R}^{N \times d}$, và $\mathbf{h}_i \in \mathbb{R}^{1 \times d}$ ký hiệu véc-tơ nhúng đầu ra thứ *i* của bộ giải mã.

4.2.5. Bộ dự đoán

Bộ dự đoán lấy đầu ra $\mathbf{H}$ của bộ giải mã và đầu ra $\mathbf{C} = \{\mathbf{c}_i\}_{i=1}^{n}$ ($\mathbf{C} \in \mathbb{R}^{n \times d}$) của bộ mã hóa câu làm đầu vào và dự đoán bộ *N* bộ ba *(r_{start}, r_{end}, rel)*. Mỗi bộ ba này tương ứng với một véc-tơ đầu ra $\mathbf{h}_i$ và đại diện cho một thực thể tham chiếu và loại quan hệ của nó. Nhãn *rel* chứa thông tin về cả loại thực thể tham chiếu và loại quan hệ. Có một mối quan hệ đặc biệt "*no reference*" đối với trường hợp không có thực thể tham chiếu, do vậy số lượng thực thể tham chiếu được trích xuất thực tế có thể ít hơn *N*.

Nghiên cứu sử dụng ba mạng nơ-ron truyền xuôi (FFN) để phát hiện vị trí bắt đầu (r_{start}), vị trí kết thúc của thực thể tham chiếu (r_{end}), và loại thực thể tham chiếu và loại quan hệ (*rel*) như sau:

$$\begin{aligned}
\mathbf{p}_i^{start} &= softmax(\mathbf{v}_1 \tanh(\mathbf{W}_1 \mathbf{H}_i^{\mathrm{T}} + \mathbf{W}_2 \mathbf{C}^{\mathrm{T}})) \\
\mathbf{p}_i^{end} &= softmax(\mathbf{v}_2 \tanh(\mathbf{W}_3 \mathbf{H}_i^{\mathrm{T}} + \mathbf{W}_4 \mathbf{C}^{\mathrm{T}})) \\
\mathbf{p}_i^{rel} &= softmax(\mathbf{h}_i \mathbf{W}_r^{\mathrm{T}})
\end{aligned} \quad (4.3)$$

trong đó $\mathbf{W}_* \in \mathbb{R}^{d \times d}$, $\mathbf{v}_* \in \mathbb{R}^{1 \times d}$, và $\mathbf{W}_r \in \mathbb{R}^{t \times d}$ là các tham số có thể học được và *t* biểu thị số lượng nhãn thực thể tham chiếu (bao gồm cả nhãn đặc biệt cho các trường hợp không có tham chiếu), và $\mathbf{H}_i \in \mathbb{R}^{n \times d}$ là ma trận có *n* dòng, mỗi dòng bằng $\mathbf{h}_i$. Khi đó, r_{start} , r_{end} và *rel* lần lượt được định nghĩa là các chỉ số của các phần tử có giá trị lớn nhất trong các véc-tơ xác suất $\mathbf{p}_i^{start}$, $\mathbf{p}_i^{end}$ và $\mathbf{p}_i^{rel}$.

$$\begin{aligned}
r_{start} &= \arg\max_{1 \le j \le n} \mathbf{p}_i^{start}(j), \\
r_{end} &= \arg\max_{1 \le j \le n} \mathbf{p}_i^{end}(j) \\
rel &= \arg\max_{1 \le j \le t} \mathbf{p}_i^{rel}(t)
\end{aligned} \quad (4.4)$$

4.2.6. Huấn luyện trích xuất kết hợp

Gọi $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N}$ là ký hiệu tập các nhãn gốc, trong đó $\mathbf{y}_i$ là các bộ ba có dạng $(e_i^{start}, e_i^{end}, r_i)$. Ở đây, e_i^{start} và e_i^{end} là vị trí bắt đầu và kết thúc của thực thể tham chiếu thứ *i* và r_i là loại thực thể tham chiếu. Gọi $\widehat{\mathbf{Y}} = \{\hat{\mathbf{y}}_i\}_{i=1}^{N}$ là nhãn dự đoán của mô hình trích xuất kết hợp, trong đó $\hat{\mathbf{y}}_i$ là các bộ ba có dạng ($\mathbf{p}_i^{start}$, $\mathbf{p}_i^{end}$, $\mathbf{p}_i^{rel}$). Trong trường hợp số lượng thực thể tham chiếu trong tập các nhãn gốc ít hơn *N* thì sẽ thêm các thực thể tham chiếu giả (là ∅). Đầu tiên, cần xác định hàm mất mát đối sánh theo cặp, giữa một cặp là giá trị nhãn gốc/nhãn dự đoán ($\mathbf{y}_i$, $\hat{\mathbf{y}}_i$) như sau:

$$\mathcal{C}_{match}(\mathbf{y}_i, \hat{\mathbf{y}}_i) = -\mathbb{1}_{\{r_i \neq \emptyset\}}[\mathbf{p}_i^{start}(e_i^{start}) + \mathbf{p}_i^{end}(e_i^{end}) + \mathbf{p}_i^{rel}(r_i)] \quad (4.5)$$

Gọi π^* là phép khớp tối ưu giữa hai tập $\mathbf{Y}$ và $\widehat{\mathbf{Y}}$:

$$\pi^\star = \arg\min_{\pi \in \Pi(N)} \sum_{i=1}^{N} \mathcal{C}_{match}(\mathbf{y}_i, \hat{\mathbf{y}}_{\pi(i)}) \quad (4.6)$$

trong đó $\Pi(N)$ là tập $N!$ các hoán vị của $\{1,2,\ldots,N\}$. Phép khớp tối ưu π^* này có thể được tính toán trong thời gian đa thức $(O(N^3))$ sử dụng thuật toán Hungary [59].

Cuối cùng, định nghĩa một hàm mất mát để huấn luyện mô hình trích xuất kết hợp như sau:

$$\mathcal{L}(\mathbf{Y}, \widehat{\mathbf{Y}}) = \sum_{i=1}^{N}\{-\log \mathbf{p}_{\pi^\star(i)}^{rel}(r_i) + \mathbb{1}_{\{r_i \neq \emptyset\}}[-\log \mathbf{p}_{\pi^\star(i)}^{start}(e_i^{start}) - \log \mathbf{p}_{\pi^\star(i)}^{end}(e_i^{end})]\} \quad (4.7)$$

4.3. THỰC NGHIỆM VÀ PHÂN TÍCH KẾT QUẢ

Phần này sẽ trình bày chi tiết về các thực nghiệm theo phương pháp đề xuất và phân tích kết quả. Tập dữ liệu được sử dụng cho thực nghiệm là tập dữ liệu đã được xây dựng trong Chương 3 (xem thông tin trong Mục 3.3 Chương 3 của luận án).

4.3.1. Thiết lập thực nghiệm

Để tiến hành thử nghiệm, thực hiện chia ngẫu nhiên tập dữ liệu thành các tập huấn luyện/kiểm chứng/kiểm tra *(training/validation/test sets)* với tỷ lệ tương ứng là

70/10/20, cho mục đích huấn luyện các mô hình trích xuất, điều chỉnh siêu tham số và thử nghiệm các mô hình.

Để đánh giá hiệu năng của các mô hình trích xuất, nghiên cứu sử dụng độ chính xác *(precision)*, độ phủ *(recall)* và độ đo F_1 cho xác định thực thể tham chiếu, cũng như cho xác định đồng thời cả thực thể tham chiếu và quan hệ. Công thức tính các độ đo này đã được trình bày trong Mục 1.3.4 Chương 1: công thức (1.17), (1.18) và (1.19).

Trích xuất riêng thực thể tham chiếu: Hiệu năng của mô hình trích xuất thực thể tham chiếu được đo bằng độ chính xác, độ phủ và độ đo F_1 cho từng loại thực thể tham chiếu và cho tất cả các loại thực thể tham chiếu. Việc trích xuất thực thể tham chiếu được coi là chính xác khi mô hình phát hiện được đúng vị trí bắt đầu, vị trí kết thúc và loại thực thể tham chiếu. Trong các công thức (1.17), (1.18), (1.19), A và B tương ứng là tập các thực thể tham chiếu được nhận ra (theo mô hình) và tập các thực thể tham chiếu gốc (đã được gán nhãn) của một loại tham chiếu cụ thể.

Trích xuất kết hợp thực thể tham chiếu và quan hệ: Hiệu năng của mô hình trích xuất kết hợp thực thể tham chiếu và quan hệ được đo bằng độ chính xác, độ phủ và độ đo F_1 cho từng bộ thực thể và quan hệ. Việc trích xuất kết hợp đồng thời cả thực thể tham chiếu và quan hệ được coi là chính xác khi mô hình phát hiện được đúng thực thể tham chiếu (với 3 thông tin: vị trí bắt đầu, vị trí kết thúc và loại thực thể tham chiếu) và đồng thời phát hiện được đúng loại quan hệ. Trong các công thức (1.17), (1.18), (1.19), A và B tương ứng là bộ các thực thể tham chiếu và quan hệ được nhận ra (bởi mô hình) và bộ các thực thể tham chiếu và quan hệ gốc (đã được gán nhãn).

4.3.2. Các mô hình thực nghiệm

Đầu tiên, nghiên cứu tiến hành các thử nghiệm để so sánh mô hình đề xuất với các phương pháp đã được thực hiện trong các nghiên cứu trước như dưới đây. Các kết quả trong các nghiên cứu này sẽ được sử dụng làm kết quả cơ sở để so sánh. Đây đều là các mô hình trích xuất kết hợp dựa trên học sâu.

- CasRel [125]: Sử dụng mô hình gán nhãn nhị phân phân tầng *(cascade binary tagging)* để trích xuất ra các bộ ba quan hệ. Mô hình này cũng sử dụng bộ mã hóa Transformer để mã hóa câu đầu vào.
- SPERT [34]: một mô hình tập trung *(attention)* với các nhúng BERT cho trích xuất kết hợp thực thể và quan hệ dựa trên vị trí *(span-based)*.
- JointER [132]: Mô hình trích xuất kết hợp thực thể và quan hệ dựa trên chiến lược phân tách *(decomposition strategy)*.
- SPN [108]: Mô hình trích xuất kết hợp thực thể và quan hệ với các mạng dự đoán theo tập hợp *(set prediction networks)*.

a) Mô hình CasRel [125]

Đây là một mô hình gán nhãn phân tầng dựa trên kiến trúc Transformer để trích xuất ra bộ ba gồm 2 thực thể và quan hệ giữa hai thực thể này. Mô hình có một Bộ mã hóa câu *(Sentence Encoder)* sử dụng BERT encoder và một Bộ giải mã đối tượng *(Object Decoder)* với hai khối gồm Bộ gán thực thể *(Entity tagger)* và Bộ gán quan hệ *(Relation tagger)*. Mô hình sử dụng một bộ phân lớp với hàm softmax để xác định loại quan hệ tương ứng với các thực thể đã được xác định, với đầu vào là kết hợp của véc-tơ ẩn trong lớp cuối cùng từ bộ mã hóa BERT và đặc trưng của các thực thể (lấy từ Bộ gán thực thể). Mô hình được huấn luyện bằng việc tối đa hóa tổng của ước lượng cực đại cho mỗi bộ gán thực thể và gán quan hệ, sử dụng với bộ tối ưu Adam. Hàm mất mát dùng trong mô hình là Cross Entropy.

Để phù hợp với mục tiêu của bài toán đặt ra trong nghiên cứu ở đây (trích xuất thực thể tham chiếu và quan hệ trong văn bản pháp quy), "thực thể đối tượng" *(object entity)* như mô tả trong nghiên cứu [125] được bỏ đi. Như vậy đầu ra mô hình lúc này là các bộ gồm 2 thông tin: thực thể tham chiếu và quan hệ tương ứng của thực thể tham chiếu với thực thể văn bản pháp quy hiện đang xem xét.

Bộ mã hóa câu được sử dụng để chuyển đổi các câu đầu vào thành các véc-tơ ngữ nghĩa. Do bài toán hiện tại sử dụng dữ liệu đầu vào là văn bản tiếng Việt, nên PhoBERT [82] được sử dụng để mã hóa thông tin ngữ cảnh cho câu tiếng Việt đầu

vào. Kiến trúc của mô hình được huấn luyện trước này tương tự như kiến trúc của BERT [31], là bộ mã hóa Transformer hai chiều đa lớp. Các véc-tơ ẩn của lớp cuối cùng trong PhoBERT được sử dụng làm biểu diễn chung của mỗi từ *(token)* trong câu đầu vào.

Bộ giải mã được sử dụng để dự đoán các cặp thực thể tham chiếu và quan hệ từ câu đầu vào. Các cặp thực thể và quan hệ được trích xuất thông qua hai bước liên tiếp. Đầu tiên, các thực thể tham chiếu được xác định từ câu đầu vào. Sau đó, đối với mỗi ứng viên thực thể tham chiếu đã được xác định trước đó, tiếp tục xác định quan hệ với văn bản đang xem xét. Các thực thể tham chiếu được xác định bằng cách giải mã trực tiếp véc-tơ đầu ra được tạo từ bộ mã hóa BERT.

Độ chính xác của mô hình được tính toán lại theo thực thể tham chiếu và quan hệ gốc cần trích xuất, như yêu cầu của thực nghiệm. Các tham số của mô hình được sử dụng theo mặc định trong nghiên cứu [125].

b) SPERT [34]

SPERT [34] là một mô hình tập trung *(attention)* để trích xuất đồng thời thực thể và quan hệ dựa trên vị trí *(span-based)*. Mô hình được huấn luyện sử dụng các mẫu phủ định mạnh trong câu để việc tìm kiếm trên các vị trí trong câu đạt hiệu quả cao hơn.

SPERT sử dụng mô hình BERT huấn luyện trước để mã hóa chuỗi đầu vào dạng các từ mã hóa BPE *(byte-pair encoding)*, chuyển thành chuỗi nhúng. Khác với kiểu phân loại quan hệ cổ điển, mô hình sẽ phát hiện các thực thể trong số tất cả các chuỗi con các từ (hoặc các vị trí). Ví dụ: chuỗi từ (we, will, rock, you) ánh xạ tới các chuỗi con (we), (we, will), (will, rock, you), … Mỗi chuỗi con này được phân loại thành: 1) loại thực thể, 2) không phải thực thể, và 3) các quan hệ. Việc phân loại quan hệ được thực hiện dựa trên các ngữ cảnh (lấy từ các chuỗi con giữa các thực thể) để phân loại tất cả các cặp thực thể thành các quan hệ.

Để phù hợp với mục tiêu của bài toán đặt ra là trích xuất thực thể tham chiếu và quan hệ giữa các thực thể trong văn bản pháp quy, số lượng thực thể và quan hệ

cần trích xuất trong một bộ chỉ là 1 thực thể (tham chiếu) thay vì 2 thực thể như bài toán ban đầu. Quan hệ cần trích xuất là quan hệ tương ứng của thực thể tham chiếu với thực thể văn bản pháp quy hiện đang xem xét.

Tương tự như cách thức sửa đổi mô hình CasRel cho phù hợp với bài toán đề xuất cho dữ liệu pháp quy tiếng Việt, mô hình SPERT được sửa đổi trong thực nghiệm này cũng sử dụng PhoBERT [82] để mã hóa thông tin ngữ cảnh cho câu tiếng Việt đầu vào thay vì mô hình BERT được huấn luyện trước. Từng chuỗi con đầu ra của PhoBERT được đưa vào bộ phân lớp để xác định là thực thể tham chiếu hoặc không phải là thực thể tham chiếu. Mô hình ban đầu SPERT [34] được sử dụng để xác định các bộ ba trong câu với 2 thực thể và quan hệ giữa hai thực thể này. Tuy nhiên, do bài toán đề xuất chỉ có một thực thể tham chiếu, nên để phù hợp với bài toán, một thực thể tham chiếu mặc định được thêm vào mỗi câu đầu vào ban đầu. Nhãn đầu vào của câu sẽ được sửa đổi để phù hợp với bài toán, gồm: các thực thể tham chiếu, thực thể tham chiếu được thêm vào cuối câu, và quan hệ của mỗi thực thể tham chiếu hiện có với thực thể tham chiếu được thêm vào (thực chất là quan hệ của tham chiếu đã có với thực thể văn bản pháp quy đang xem xét). Khi đó đầu ra của mô hình là các bộ ba gồm 1 thực thể tham chiếu đã có, một thực thể tham chiếu mặc định và quan hệ giữa 2 thực thể tham chiếu này.

Dựa trên mỗi cặp tham chiếu ứng viên đã được xác định trước đó nhờ bộ phân lớp thực thể, kết hợp với các ngữ cảnh (lấy từ các chuỗi con giữa 2 thực thể ứng viên này) để phân loại tất cả các cặp thực thể thành các quan hệ (của thực thể tham chiếu với thực thể văn bản pháp quy đang xem xét).

Trong quá trình huấn luyện, mỗi batch huấn luyện bao gồm tập các câu, và các mẫu huấn luyện được lấy ra từ mỗi câu. Đối với bộ phân lớp chuỗi con các từ, tất cả các thực thể được gán nhãn làm mẫu dương, cộng với một lượng cố định các chuỗi con các từ ngẫu nhiên không phải là thực thể làm mẫu âm. Để huấn luyện bộ phân lớp quan hệ, các quan hệ nhãn gốc được sử dụng làm mẫu dương và rút ra một số cố định các mẫu âm từ các cặp thực thể không được gán nhãn với bất kỳ quan hệ nào.

Các mẫu âm này gọi là mẫu âm mạnh và có tầm quan trọng cao đối với hiệu năng của mô hình.

Độ chính xác của mô hình được tính toán lại theo thực thể tham chiếu và quan hệ gốc cần trích xuất, như yêu cầu của thực nghiệm. Các tham số của mô hình được sử dụng theo mặc định trong nghiên cứu [34].

c) JointER [132]

Mô hình JointER [132] là một mô hình trích xuất kết hợp thực thể và quan hệ dựa trên chiến lược phân tách *(decomposition strategy)*. Mô hình thực hiện phân tách tác vụ trích xuất đồng thời thành hai nhiệm vụ con có liên quan với nhau, đó là trích xuất thực thể và trích xuất quan hệ. Nhiệm vụ con thứ nhất trích xuất tất cả các thực thể đứng đầu có thể liên quan đến quan hệ đích, và nhiệm vụ con thứ hai xác định các thực thể đứng cuối tương ứng và các mối quan hệ cho mỗi thực thể đứng đầu được trích xuất. Tiếp theo, hai nhiệm vụ con này được tiếp tục phân giải thành các nhiệm vụ gán nhãn tuần tự dựa trên cơ chế gán nhãn cho các chuỗi con, với bộ gán nhãn ranh giới phân cấp và thuật toán giải mã nhiều chuỗi con. Mô hình này có thể nắm bắt đầy đủ sự phụ thuộc lẫn nhau về ngữ nghĩa giữa các nhiệm vụ khác nhau, cũng như giảm nhiễu trong trường hợp các cặp thực thể không có quan hệ.

Kiến trúc của mô hình được tác giả đề xuất với một bộ mã hóa dùng chung sử dụng một lớp BiLSTM, một bộ trích xuất thực thể đứng đầu và một bộ trích xuất thực thể đứng cuối, cùng với quan hệ của thực thể này với thực thể đứng đầu.

Đầu vào của chuỗi bộ mã hóa dùng chung là biểu diễn của các từ trong câu đầu vào, bao gồm véc-tơ nhúng từ được huấn luyện trước và biểu diễn từ dựa trên ký tự sử dụng CNN trên chuỗi ký tự trong từ cho trước, cùng với véc-tơ nhúng từ loại. Đầu vào của bộ trích xuất thực thể là véc-tơ đầu ra của bộ mã hóa dùng chung sử dụng một lớp BiLSTM như mô tả ở trên, cộng với véc-tơ biểu diễn ngữ cảnh toàn cục được tính bằng phương pháp gộp cực đại *(max pooling)* trên tất cả các trạng thái ẩn. Bộ trích xuất thực thể đứng đầu sẽ trả lại các thực thể đứng đầu và loại thực thể tương ứng có trong câu đầu vào.

Bộ trích xuất thực thể đứng cuối và quan hệ cũng sử dụng dữ liệu đầu vào của bộ trích xuất thực thể đứng đầu. Tuy nhiên, với mục đích là vừa phát hiện các thực thể đứng cuối vừa phát hiện các mối quan hệ với thực thể đứng đầu, dữ liệu đầu vào của bộ trích xuất này không chỉ là dữ liệu ghép nối hai véc-tơ biểu diễn mà sẽ bao gồm: 1) các từ bên trong thực thể đứng cuối; 2) thực thể đứng đầu tương ứng; 3) ngữ cảnh chỉ ra mối quan hệ; 4) khoảng cách giữa thực thể đứng cuối và thực thể đứng đầu. Đầu ra của bộ trích xuất này là thực thể đứng cuối và quan hệ của nó với thực thể đứng đầu tương ứng. Kết hợp với đầu ra của bộ trích xuất thực thể đứng đầu, kết quả có bộ ba gồm 2 thực thể và quan hệ giữa hai thực thể này.

Để huấn luyện mô hình trích xuất kết hợp này, hàm mất mát kết hợp là tổng của hai hàm mất mát của từng bộ trích xuất. Mô hình huấn luyện được tối ưu sử dụng thuật toán SGD *(Stochastic Gradient Descent)*. Tuy nhiên, để phù hợp với nghiên cứu đề xuất trong chương này, dữ liệu đầu vào được bổ sung tương tự như trong thực nghiệm sử dụng mô hình sửa đổi của SPERT, trong đó một thực thể tham chiếu cố định được bổ sung vào cuối mỗi câu đầu vào và quan hệ của tham chiếu gốc trong câu với văn bản đang xem xét được sử dụng như quan hệ của tham chiếu gốc và tham chiếu thêm vào. Bộ ba đầu ra sẽ bao gồm một tham chiếu gốc, tham chiếu thêm vào trong câu và quan hệ cần trích xuất.

Độ chính xác được tính toán lại theo thực thể tham chiếu và quan hệ gốc cần trích xuất, tương tự như các mô hình khác theo yêu cầu của thực nghiệm. Các tham số của mô hình được sử dụng theo mặc định trong nghiên cứu [132].

d) SPN [108]

Mô hình trích xuất kết hợp thực thể và quan hệ với các mạng dự đoán theo tập hợp *(set prediction networks)*. Mô hình này là cơ sở của mô hình đề xuất trong nghiên cứu, chỉ khác biệt tại đầu vào của Bộ giải mã và sửa đổi cho phù hợp với việc gán nhãn dữ liệu đầu vào, đầu ra (thay vì bộ ba với 2 thực thể và quan hệ giữa chúng, mô hình có đầu ra chỉ có 1 thực thể tham chiếu và quan hệ giữa tham chiếu này với văn bản pháp quy đang xử lý). SPN sử dụng một véc-tơ đầu vào chung cho tất cả các

trường hợp dữ liệu đầu vào, mà không sử dụng Bộ tăng cường đầu vào như nghiên cứu đề xuất trong luận án.

Tương tự như các mô hình đã trình bày ở trên, thay vì sử dụng BERT, mô hình SPN sửa đổi sử dụng trong thực nghiệm sử dụng PhoBERT nhằm có được biểu diễn tốt hơn cho các câu tiếng Việt. Các tham số và cấu trúc tương tự như mô hình đề xuất trong luận án.

4.3.3. Huấn luyện mạng

Mô hình đề xuất được triển khai trong Pytorch sử dụng HuggingFace (*https://huggingface.co/*) và sử dụng PhoBERT [82] làm mô hình ngôn ngữ huấn luyện trước. Trong tất cả các thử nghiệm, độ dài của chuỗi được đặt tối đa là 512, số lượng khối transformer hai chiều xếp chồng (K) là 12 và kích thước nhúng (d) được đặt là giá trị mặc định trong PhoBERT, là 768. Tham số (N) được đặt giá trị là 15 trong mô hình đề xuất. Đây là số tham chiếu đầu ra tối đa dự kiến, nhiều hơn số tham chiếu nhiều nhất trong một câu có trong tập dữ liệu. Mô hình được huấn luyện bằng bộ tối ưu hóa AdamW [68] với epsilon và phân rã trọng số (*weight decay*) được đặt là giá trị mặc định trong PyTorch, là 1e-8. Số lớp giải mã, tốc độ học và kích thước lô (*batch*) được thiết lập tương ứng là {1, 2, 3, 4, 5}, {1e-4, 2e-4, 3e-4, 4e-4, 5e-4} và {4, 8, 16, 32}. Để giảm thiểu *overfitting*, chúng tôi đã sử dụng *dropout rate* là 0,1 cho mỗi lớp ẩn. Đối với bộ chọn cụm từ, độ dài của ngữ cảnh trái/phải (q) được đặt là 3. Kích thước lô, độ dài của cụm từ (m) và tốc độ học được thiết lập tương ứng là {8, 16, 32, 64}, {1,2,3,4} và {1e-3, 2e-3, 3e-3}. Thuật toán *grid search* được sử dụng để tìm ra tập tham số tối ưu cho mô hình. Trong mỗi thử nghiệm, mô hình được huấn luyện 100 *epochs* và tính toán độ chính xác, độ phủ và độ đo F_1 sau mỗi *epoch* trên tập kiểm chứng (*validation set*). Kết quả trích xuất kết hợp có độ đo F_1 cao nhất được chọn để áp dụng cho tập kiểm tra. Các siêu tham số của mô hình được tóm tắt trong Bảng 4.1.

Bảng 4.1. Các siêu tham số của mô hình

Mô hình	**Siêu tham số**	**Giá trị**
Mô hình trích xuất kết hợp	Số lớp giải mã	3
	Kích thước lô (Batch size)	8
	Tốc độ học	2e-4
	Phân rã trọng số (Weight decay)	1e-8
	Dropout rate	0.1
Bộ tăng cường đầu vào	Kích thước lô (Batch size)	32
	Độ dài cụm từ (m)	2
	Độ dài của ngữ cảnh trái/phải (q)	3
	Tốc độ học	1e-3

Để đảm bảo việc so sánh các phương pháp là tương đương với nhau, chúng tôi sử dụng PhoBert-base [82] làm bộ mã hóa BERT cho tất cả các phương pháp, ngoại trừ JointER do không phụ thuộc BERT để mã hóa. Việc đánh giá hiệu năng của các mô hình sẽ được thực hiện trên cùng một tập dữ liệu với phương pháp đánh giá giống nhau. Cấu hình nền tảng thử nghiệm nhất quán với máy chủ sử dụng hệ điều hành Linux, bộ vi xử lý Intel E5v4, 64GB RAM và hai card đồ họa NVIDIA GTX 2080 Ti.

4.3.4. Kết quả thực nghiệm

Thử nghiệm đầu tiên được thực hiện nhằm so sánh mô hình đề xuất với các mô hình cơ sở. Bảng 4.2 thể hiện kết quả thực nghiệm của các mô hình trích xuất thực thể tham chiếu và quan hệ. Mô hình đề xuất đạt kết quả vượt trội hơn tất cả các mô hình cơ sở trong cả hai trường hợp, chỉ trích xuất thực thể tham chiếu và trích xuất kết hợp cả thực thể tham chiếu và quan hệ. Với trường hợp chỉ trích xuất thực

thể tham chiếu, mô hình đề xuất đạt độ đo F_1 là 99,7%, cải thiện 0,4% (giảm tỷ lệ lỗi 57%) so với mô hình SPN (là mô hình đạt độ đo F_1 tốt nhất trong nhóm các mô hình cơ sở đang xem xét). Với trường hợp trích xuất kết hợp cả thực thể tham chiếu và quan hệ, mô hình đề xuất đạt độ đo F_1 là 99,4%, cải thiện 1,1% (giảm tỷ lệ lỗi 65%) so với mô hình SPN.

Có thể thấy với việc trích xuất riêng thực thể tham chiếu cũng như trích xuất kết hợp thực thể tham chiếu và quan hệ đều cho kết quả vượt trội so với các mô hình tốt nhất đã được đề xuất trong nghiên cứu Chương 3 (F_1 = 96,62% cho trích xuất thực thể tham chiếu và F_1 = 97,03% cho trích xuất quan hệ).

Bảng 4.2. Kết quả thực nghiệm của các mô hình trích xuất thực thể tham chiếu và quan hệ

Mô hình	Chỉ trích xuất thực thể tham chiếu			Trích xuất kết hợp thực thể tham chiếu và quan hệ		
	Độ chính xác (%)	Độ phủ (%)	Độ đo F_1 (%)	Độ chính xác (%)	Độ phủ (%)	Độ đo F_1 (%)
CasRel	98,7	91,3	94,8	94,8	87,7	91,1
SPERT	98,3	91,7	94,8	96,9	90,4	93,5
JointER	98,4	98,0	98,2	97,2	96,9	97,1
SPN	99,8	98,7	99,3	98,8	97,7	98,3
Mô hình đề xuất	**99,8**	**99,6**	**99,7**	**99,5**	**99,3**	**99,4**

Độ phức tạp của mô hình đề xuất được đánh giá và so với các mô hình trích xuất khác, dựa trên số lượng tham số thời gian huấn luyện của mỗi mô hình. Kết quả trong Bảng 4.3 cho thấy mô hình JointER có số lượng tham số ít nhất và thời gian huấn luyện ngắn nhất. Mô hình đề xuất có số lượng tham số và thời gian huấn luyện tương đương với các mô hình khác còn lại, nhưng mang lại kết quả tốt nhất. Sự khác

biệt về số tham số và thời gian huấn luyện giữa các mô hình là chấp nhận được, tuy nhiên, với thời gian huấn luyện tương tự, độ chính xác vẫn là yếu tố quan trọng hơn.

Bảng 4.3. Số lượng tham số và thời gian huấn luyện của các mô hình trích xuất thực thể tham chiếu và quan hệ

Mô hình	Số lượng tham số (triệu)	Thời gian huấn luyện
CasRel	135,10	4h20p
SPERT	135,62	4h30p
JointER	13,17	33p
SPN	165,77	4h45p
Mô hình đề xuất	165,80	4h50p

Thử nghiệm tiếp theo được thực hiện nhằm đo hiệu năng của các mô hình trích xuất thực thể tham chiếu và quan hệ giữa các thực thể theo độ phức tạp của các câu trong văn bản pháp quy đầu vào. Tập dữ liệu kiểm tra được chia thành 5 tập con và tiến hành đo F_1 trên mỗi tập con một cách riêng biệt. Tập con i ($1 \leq i \leq 4$) chứa các câu có chính xác i thực thể tham chiếu trong tập nhãn gốc. Tập con 5 chứa các câu khác với ít nhất 5 thực thể tham chiếu trong tập nhãn gốc. Ví dụ một câu trong Tập con 5 (đã được gán nhãn thực thể và quan hệ) như sau: *"2 . Thông_tư này thay_thế* ***<QĐ rel="BTT">*** *Quyết_định số 0808/2001/QĐ-BTM ngày 31 tháng 7 năm 2001* ***</QĐ>*** *của Bộ_trưởng Bộ Thương_mại quy_định về quản_lý và sử_dụng Thẻ kiểm_tra thị_trường , phù_hiệu , cờ hiệu , cấp_hiệu , biển_hiệu Quản_lý thị_trường ;* ***<QĐ rel="BTT">*** *Quyết_định số 1152/2001/QĐ-BTM ngày 29 tháng 10 năm 2001* ***</QĐ>*** *và* ***<QĐ rel="BTT">*** *Quyết_định số 1211/2004/QĐ-BTM ngày 25 tháng 8 năm 2004* ***</QĐ>*** *của Bộ_trưởng Bộ Thương_mại về việc bổ_sung một_số điều quy_định về quản_lý và sử_dụng Thẻ kiểm_tra thị_trường , phù_hiệu , cờ hiệu , cấp_hiệu , biển_hiệu Quản_lý thị_trường ;* ***<TT rel="BTT">*** *Thông_tư số 30/2012/TT-BCT ngày 10 tháng 10 năm 2012* ***</TT>*** *của Bộ_trưởng Bộ*

Công_Thương quy_định việc cấp , quản_lý và sử_dụng Thẻ kiểm_tra thị_trường đối_với công_chức của Cục Quản_lý thị_trường ; khoản 1 Điều 6 ***<TT rel="BTT">*** *Thông_tư số 09/2013/TT-BCT ngày 02 tháng 5 năm 2013* ***</TT>*** *của Bộ_trưởng Bộ Công_Thương quy_định về hoạt_động kiểm_tra và xử_phạt vi_phạm hành_chính của Quản_lý thị_trường .”*

Bảng 4.4. Hiệu năng của các mô hình trích xuất thực thể tham chiếu và quan hệ theo độ phức tạp của các câu văn bản pháp quy đầu vào tính theo độ đo F_1 (%)

Mô hình	Tập con 1	Tập con 2	Tập con 3	Tập con 4	Tập con 5
CasRel	97,7	90,1	80,6	75,1	56,0
SPERT	96,4	94,3	90,3	87,5	82,3
JointER	99,8	97,4	91,1	90,3	90,0
SPN	99,4	98,7	98,2	96,4	91,9
Mô hình đề xuất	**99,9**	**99,6**	**99,1**	**98,9**	**96,8**

Bảng 4.4 trình bày hiệu năng của các mô hình trích xuất trên 5 tập con. Có một số quan sát như sau: 1) hiệu năng của tất cả các mô hình giảm theo độ phức tạp của các câu đầu vào; 2) mô hình đề xuất trong nghiên cứu này đạt được kết quả tốt hơn tất cả các mô hình cơ sở trên 5 tập con; 3) giá trị độ lệch về hiệu năng giữa mô hình đề xuất với mô hình tốt nhất trong 4 mô hình cơ sở tăng lên theo độ phức tạp của các câu đầu vào. Điều này cho thấy tính hiệu quả của mô hình đề xuất trong các trường hợp phức tạp.

Thử nghiệm thứ ba được tiến hành nhằm đánh giá tác động của bộ tăng cường đầu vào đối với hiệu năng của mô hình trích xuất thực thể tham chiếu và quan hệ. Nghiên cứu thực hiện xem xét các biến thể sau của mô hình với các bộ tăng cường đầu vào khác nhau:

- Không sử dụng: Không sử dụng bộ tăng cường đầu vào.

- Dựa trên từ điển: Sử dụng bộ tăng cường đầu vào dựa trên từ điển.
- MLP dựa trên phân loại: Mô hình đề xuất sử dụng bộ tăng cường đầu vào dựa trên phân loại với perceptron đa lớp (multi-layer perceptron).
- CNN dựa trên phân loại: Mô hình đề xuất sử dụng bộ tăng cường đầu vào dựa trên phân loại với mạng nơ-ron tích chập (*convolutional neural network, CNN*) để phân loại. Thử nghiệm sử dụng một kiến trúc CNN đơn giản: hai lớp tích chập 1D, một lớp kết nối đầy đủ ẩn và một lớp đầu ra với hàm kích hoạt sigmoid. Tương tự như biến thể MLP, thử nghiệm cũng sử dụng hàm mất mát entropy chéo nhị phân và bộ tối ưu hóa Adam. Số lượng bộ lọc và kích thước bộ lọc được thiết lập tương ứng là {64, 128, 256} và {3, 4, 5} bằng cách sử dụng tập xác nhận *(validation test)*.

Bảng 4.5. Tác dụng của bộ tăng cường đầu vào

Mô hình (Các biến thể)	**Chỉ trích xuất thực thể tham chiếu**			**Trích xuất kết hợp thực thể tham chiếu và quan hệ**		
	Độ chính xác (%)	Độ phủ (%)	Độ đo F_1 (%)	Độ chính xác (%)	Độ phủ (%)	Độ đo F_1 (%)
Không sử dụng	99,5	98,7	99,1	98,7	97,6	98,2
Dựa trên từ điển	99,5	99,1	99,3	98,6	98,1	98,4
Phân loại dựa trên MLP	99,8	99,6	99,7	99,5	99,3	99,4
Phân loại dựa trên CNN	99,9	99,7	99,8	99,5	99,3	99,4

Như được trình bày trong Bảng 4.5, ba biến thể sau sử dụng bộ tăng cường đầu vào hoạt động tốt hơn so với biến thể đầu không sử dụng bộ tăng cường đầu vào. Điều này khẳng định tính hiệu quả của phương pháp tăng cường đầu vào bộ giải mã đã đề xuất. Kết quả thực nghiệm cũng chỉ ra rằng cách tiếp cận dựa trên phân loại vượt trội so với cách tiếp cận dựa trên từ điển đơn giản. Hơn nữa, hai biến thể dựa

trên phân loại cho kết quả tương tự, cho thấy tính ổn định của phương pháp tăng cường được đề xuất.

Bảng 4.6. Ảnh hưởng của số lớp giải mã tới hiệu quả của mô hình đề xuất

Số lớp giải mã	**Độ chính xác (%)**	**Độ phủ (%)**	**Độ đo F_1 (%)**
1	99,2	99,0	99,1
2	99,3	99,2	99,3
3	99,5	99,3	99,4
4	99,4	99,3	99,4
5	99,4	99,3	99,4

Một thử nghiệm khác nữa cũng được tiến hành nhằm đánh giá về việc số lượng lớp giải mã ảnh hưởng đến hiệu năng của mô hình, bằng cách thử nghiệm mô hình đề xuất với số lượng lớp giải mã khác nhau. Kết quả trong Bảng 4.6 cho thấy việc tăng số lượng lớp giải mã có thể dẫn đến kết quả tốt hơn, nhưng hiệu năng cao nhất được đạt khi sử dụng 3 lớp giải mã. Kết quả tốt nhất đạt được khi số lượng lớp giải mã được chọn là 3, với giá trị F_1 là 99,4%. Khi số lượng lớp giải mã tăng lên 4 hoặc 5, kết quả không được cải thiện đáng kể. Lý do chính có thể là khi lớp giải mã sâu hơn, sẽ có nhiều mô-đun multi-head self-attention và inter-attention hơn, làm cho phép tích hợp thông tin của câu vào các truy vấn một cách toàn diện hơn. Tuy nhiên, khi số lượng lớp giải mã tăng nhiều hơn nữa, ưu thế này thể hiện kém rõ ràng hơn.

4.4. KẾT LUẬN CHƯƠNG 4

Nội dung Chương 4 đã trình bày một nghiên cứu về trích xuất kết hợp đồng thời thực thể tham chiếu và quan hệ giữa các thực thể trong văn bản pháp quy dựa trên Transformer. Nghiên cứu đã chứng minh tính hiệu quả của mô hình đề xuất bằng cách tiến hành các thử nghiệm trên một tập dữ liệu đã được xây dựng để trích xuất thực thể tham chiếu và quan hệ giữa các thực thể riêng (như trình bày trong Chương

3), so sánh kết quả của các thử nghiệm này với một số mô hình cơ sở đã được đánh giá có kết quả tốt (trong các nghiên cứu trước đây). Kết quả thực nghiệm cho thấy: 1) Mô hình đề xuất có thể trích xuất chính xác đồng thời cả thực thể tham chiếu và quan hệ đạt độ đo F_1 là 99,4%; 2) Mô hình đề xuất cho kết quả tốt hơn so với các mô hình cơ sở khác, đặc biệt là đối với các câu văn bản pháp quy phức tạp có nhiều thực thể tham chiếu; 3) Bộ chọn cụm từ hoạt động ổn định, nghĩa là nó không bị ảnh hưởng nhiều bởi thuật toán học máy.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết luận

Sau quá trình nghiên cứu, đề tài luận án "Nghiên cứu các phương pháp học máy cho trích xuất thông tin tự động từ văn bản" đã đạt được những kết quả nhất định như sau:

1) *Đề xuất giải pháp nâng cao hiệu quả cho trích xuất khía cạnh và phân loại quan điểm trong ngôn ngữ tiếng Việt bằng cách khai thác nguồn dữ liệu đã được gán nhãn sẵn từ ngôn ngữ khác*. Việc xác định các loại khía cạnh và phân loại quan điểm được thực hiện theo từng câu thay vì toàn bộ bài đánh giá sẽ thực tế hơn và có thể áp dụng trong nhiều ứng dụng thế giới thực. Phương pháp đề xuất khá tổng quát và linh hoạt do không phụ thuộc vào ngôn ngữ và các thuật toán học máy, giúp giải quyết khó khăn do việc thiếu tài nguyên dữ liệu huấn luyện trong một số ngôn ngữ có ít tài nguyên cho bài toán này (như tiếng Việt). Các kết quả thực nghiệm trên tập dữ liệu, gồm 575 bài đánh giá tiếng Việt và 440 bài đánh giá tiếng Anh, cho thấy với việc sử dụng thêm dữ liệu (đã được gán nhãn) dịch từ ngôn ngữ khác (như tiếng Anh), phương pháp đề xuất đã cải thiện hiệu năng trong cả hai nhiệm vụ trích xuất khía cạnh và phân loại quan điểm tiếng Việt. Nội dung đề xuất này được trình bày chi tiết trong Chương 2, với kết quả được tổng hợp từ các công trình nghiên cứu đã công bố [4, 6] (Theo danh mục các công trình công bố).
2) *Đề xuất phương pháp trích xuất thông tin sử dụng học máy truyền thống và học sâu cho văn bản pháp quy tiếng Việt. Các thông tin được trích xuất bao gồm thực thể tham chiếu và mối quan hệ giữa các thực thể văn bản pháp quy*. Ngoài việc sử dụng các phương pháp học máy truyền thống, nghiên cứu còn sử dụng các phương pháp học sâu, đồng thời kết hợp lợi thế của mô hình học sâu và các đặc trưng được thiết kế thủ công (theo phương pháp học máy truyền thống). Kết quả thực nghiệm trên tập dữ liệu tự xây dựng, gồm 5.031 văn bản

pháp quy tiếng Việt, cho thấy phương pháp đề xuất có kết quả khả quan với cả hai nhiệm vụ trích xuất thực thể tham chiếu và phân loại quan hệ, với độ đo F_1 đều đạt trên 95%. Nội dung đề xuất này được trình bày chi tiết trong Chương 3, với kết quả được tổng hợp từ các công trình nghiên cứu đã công bố [1, 5] (Theo danh mục các công trình công bố).

3) *Đề xuất phương pháp trích xuất kết hợp thực thể và quan hệ trong văn bản pháp quy tiếng Việt sử dụng mô hình dựa trên học sâu*. Mô hình trích xuất kết hợp sử dụng kiến trúc bộ mã hóa-giải mã dựa trên Transformer với cơ chế giải mã song song không tự hồi quy *(non-autoregressive decoding mechanism)* để trích xuất đồng thời các thực thể tham chiếu và quan hệ trong văn bản pháp quy (khác với nghiên cứu trong Chương 3 thực hiện trích xuất các thông tin này theo cách tuần tự). Kết quả thực nghiệm trên tập dữ liệu văn bản pháp quy đã được xây dựng cho thấy mô hình đề xuất có thể trích xuất chính xác đồng thời cả thực thể tham chiếu và quan hệ (đạt độ đo F_1 là 99,4%), đồng thời mô hình đề xuất cho kết quả tốt hơn so với các mô hình cơ sở khác, đặc biệt là đối với các câu pháp quy phức tạp có nhiều thực thể tham chiếu. Nội dung đề xuất này được trình bày chi tiết trong Chương 4, với kết quả được tổng hợp từ các công trình nghiên cứu đã công bố [2, 3] (Theo danh mục các công trình công bố).

Hướng phát triển

Với những kết quả nghiên cứu có được, một số định hướng phát triển nghiên cứu tiếp theo của đề tài luận án như sau:

- Nghiên cứu các phương pháp tiên tiến sử dụng dữ liệu có sẵn từ các ngôn ngữ khác nhằm cải thiện các nhiệm vụ NLP cho các ngôn ngữ ít tài nguyên (như tiếng Việt). Ví dụ, có thể thực hiện: (1) tận dụng nhiều bộ dữ liệu của cùng một nhiệm vụ NLP từ các ngôn ngữ khác nhau để giải quyết nhiệm vụ đó cho tiếng Việt; và (2) tận dụng nhiều bộ dữ liệu của các nhiệm vụ NLP khác nhau

từ một ngôn ngữ khác để giải quyết các nhiệm vụ đó cho tiếng Việt (nghĩa là học đa tác vụ).

- Với lĩnh vực xử lý văn bản pháp quy: (1) Có thể sử dụng kết quả trích xuất thực thể tham chiếu và phân loại quan hệ giữa các thực thể văn bản pháp quy cho các tác vụ xử lý văn bản pháp quy khác như truy xuất thông tin pháp quy, tóm tắt văn bản pháp quy và trả lời câu hỏi trong lĩnh vực pháp quy; (2) Có thể xây dựng các ứng dụng xử lý văn bản pháp quy nhằm giúp người dùng đọc, hiểu và truy xuất thông tin cần thiết từ các văn bản pháp. Những việc này rất có ý nghĩa trong thực tiễn.

DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ

TẠP CHÍ KHOA HỌC

[1] **Nguyễn Thị Thanh Thủy**, Đặng Bảo Chiến, Triệu Khương Duy, Ngô Xuân Bách, Từ Minh Phương, Phân loại quan hệ tham chiếu trong văn bản pháp quy, *Vol 1 No 3 (2020): Journal of Science and Technology on Information and Communications (ISSN: 2525-2224)*, pp.69-78, 2020.

[2] **Nguyễn Thị Thanh Thủy**, Nguyễn Ngọc Điệp, Một phương pháp trích xuất kết hợp thực thể và quan hệ tham chiếu trong văn bản pháp quy, *Vol 1 No 3 (2021): Journal of Science and Technology on Information and Communications (ISSN: 2525-2224)*, pp.100-108, 2021.

[3] **Nguyen Thi Thanh Thuy**, Nguyen Ngoc Diep, Ngo Xuan Bach, Tu Minh Phuong, Joint Reference and Relation Extraction from Legal Documents with Enhanced Decoder Input, *Vol 23 No 2 (2023): Cybernetics and Information Technologies (ISSN: 1314-4081*), pp.72-86, 2023. (**Scopus, Q2**).

HỘI NGHỊ KHOA HỌC

[4] **Nguyen Thi Thanh Thuy**, Ngo Xuan Bach, Tu Minh Phuong, Cross-Language Aspect Extraction for Opinion Mining, *Proceedings of the 10th International Conference on Knowledge and Systems Engineering (KSE 2018)*, pp. 67-72, 2018.

[5] Ngo Xuan Bach, **Nguyen Thi Thanh Thuy**, Dang Bao Chien, Trieu Khuong Duy, To Minh Hien, and Tu Minh Phuong, Reference Extraction from Vietnamese Legal Documents, *Proceedings of* the *Tenth International Symposium on Information and Communication Technology (SoICT 2019)*, pp. 486-493, 2019.

[6] **Nguyen Thi Thanh Thuy,** Ngo Xuan Bach, Tu Minh Phuong, Leveraging Foreign Language Labeled Data for Aspect-Based Opinion Mining, *International Conference* on *Computing and Communication Technologies (RIVF 2020)*, pp. 1-6, 2020.

TÀI LIỆU THAM KHẢO

[1] Alvarez-López, T., Juncal-Mart\'\inez, J., Fernández-Gavilanes, M., Costa-Montenegro, E. and González-Castano, F.J. (2016), Gti at semeval-2016 task 5: Svm and crf for aspect detection and unsupervised aspect-based sentiment analysis, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 306–11.

[2] Andrew, J.J. (2018), Automatic extraction of entities and relation from legal documents, *Proceedings of the Seventh Named Entities Workshop*, p. 1–8.

[3] Bach, N.X., Hai, V.T. and Phuong, T.M. (2016), Cross-domain sentiment classification with word embeddings and canonical correlation analysis, *Proceedings of the Seventh Symposium on Information and Communication Technology*, p. 159–66.

[4] Bach, N.X., Le Minh, N., Oanh, T.T. and Shimazu, A. (2013), A two-phase framework for learning logical structures of paragraphs in legal articles. 2013, *ACM Transactions on Asian Language Information Processing (TALIP)*, **12**(1), p. 3.

[5] Bach, N.X. and Phuong, T.M. (2015), Leveraging user ratings for resource-poor sentiment classification, *Procedia Computer Science*, Elsevier. **60**, p. 322–31.

[6] Bach, N.X., Thien, T.H.N., Phuong, T.M. and others. (2017), Question analysis for vietnamese legal question answering, *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, p. 154–9.

[7] Bahdanau, D., Cho, K. and Bengio, Y. (2014), Neural machine translation by jointly learning to align and translate, *ArXiv Preprint ArXiv:14090473*,.

[8] Borkar, V., Deshmukh, K. and Sarawagi, S. (2001), Automatic segmentation of text into structured records, *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, p. 175–86.

[9] Broder, A., Fontoura, M., Josifovski, V. and Riedel, L. (2007), A semantic approach to contextual advertising, *Proceedings of the 30th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, p. 559–66.

[10] Bui, T.D. and Ho, Q.B. (2014), An approach for automatically structuring vietnamese legal text, *2014 International Conference on Asian Language Processing (IALP)*, p. 187–90.

[11] Bui, T.D., Nguyen, S.T. and Ho, Q.B. (2015), Towards a conceptual search for Vietnamese legal text, *IFIP International Conference on Computer Information Systems and Industrial Management*, p. 175–85.

[12] Bunescu, R., Ge, R., Kate, R.J., Marcotte, E.M., Mooney, R.J., Ramani, A.K. et al. (2005), Comparative experiments on learning information extractors for proteins and their interactions, *Artificial Intelligence in Medicine*, Elsevier. **33**(2), p. 139–55.

[13] Bunescu, R. and Mooney, R. (2005), A shortest path dependency kernel for relation extraction, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 724–31.

[14] Çetinda\ug, C., Yaz\ic\io\uglu, B. and Koç, A. (2022), Named-entity recognition in Turkish legal texts, *Natural Language Engineering*, Cambridge University Press. p. 1–28.

[15] Chakaravarthy, V.T., Gupta, H., Roy, P. and Mohania, M. (2006), Efficiently linking text documents with relevant structured information, *Proceedings of the 32nd International Conference on Very Large Data Bases*, p. 667–78.

[16] Chakrabarti, S. (2002), Mining the Web: Discovering knowledge from hypertext data, Morgan Kaufmann.

[17] Chakrabarti, S., Mirchandani, J. and Nandi, A. (2005), Spin: searching personal information networks, *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 674.

[18] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N. and Androutsopoulos, I. (2020), LEGAL-BERT:" Preparing the Muppets for Court’"., *EMNLP (Findings)*, p. 2898–904.

[19] Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. et al. (2022), {L}ex{GLUE}: A Benchmark Dataset for Legal Language Understanding

in {E}nglish, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland. p. 4310–30. http://dx.doi.org/10.18653/v1/2022.acl-long.297

[20] Chalkidis, I. and Kampas, D. (2019), Deep learning in law: early adaptation and legal word embeddings trained on large corpora, *Artificial Intelligence and Law*, Springer. **27**(2), p. 171–98.

[21] Chan, Y.S. and Roth, D. (2011), Exploiting syntactico-semantic structures for relation extraction, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, p. 551–60.

[22] Chang, C.-C. and Lin, C.-J. (2011), LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology (TIST)*, Acm New York, NY, USA. **2**(3), p. 1–27.

[23] Chau, C.-N., Nguyen, T.-S. and Nguyen, L.-M. (2020), Vnlawbert: A vietnamese legal answer selection approach using bert language model, *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, p. 298–301.

[24] Chen, H., Wu, L., Chen, J., Lu, W. and Ding, J. (2022), A comparative study of automated legal text classification using random forests and deep learning, *Information Processing \& Management*, Elsevier. **59**(2), p. 102798.

[25] Chen, Y., Sun, Y., Yang, Z. and Lin, H. (2020), Joint entity and relation extraction for legal documents with legal feature enhancement, *Proceedings of the 28th International Conference on Computational Linguistics*, p. 1561–71.

[26] Cheng, T.T., Cua, J.L., Tan, M.D., Yao, K.G. and Roxas, R.E. (2009), Information extraction from legal documents, *2009 Eighth International Symposium on Natural Language Processing*, p. 157–62.

[27] Cohen, J. (1960), A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, Sage Publications Sage CA: Thousand Oaks, CA. **20**(1), p. 37–46.

[28] Correia, F.A., Almeida, A.A.A., Nunes, J.L., Santos, K.G., Hartmann, I.A., Silva,

F.A. et al. (2022), Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court, *Information Processing \& Management*, Elsevier. **59**(1), p. 102794.

[29] Cutrell, E. and Dumais, S.T. (2006), Exploring personal information, *Communications of the ACM*, ACM New York, NY, USA. **49**(4), p. 50–1.

[30] Van Dang, T., Nguyen, V.D., Van Kiet, N. and Ngan, N.L.T. (2018), A transformation method for aspect-based sentiment analysis, *Journal of Computer Science and Cybernetics*, **34**(4), p. 323–33.

[31] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 4171–86.

[32] Doorenbos, R.B., Etzioni, O. and Weld, D.S. (1997), A scalable comparison-shopping agent for the world-wide web, *Proceedings of the First International Conference on Autonomous Agents*, p. 39–48.

[33] Duyen, N.T., Bach, N.X. and Phuong, T.M. (2014), An empirical study on sentiment analysis for Vietnamese, *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*, p. 309–14.

[34] Eberts, M. and Ulges, A. (2020), Span-based Joint Entity and Relation Extraction with Transformer Pre-training, *Proceedings of the 24th European Conference on Artificial Intelligence*,.

[35] Elman, J.L. (1990), Finding structure in time, *Cognitive Science*, Wiley Online Library. **14**(2), p. 179–211.

[36] Filtz, E., Kirrane, S. and Polleres, A. (2021), The linked legal data landscape: linking legal data across different countries, *Artificial Intelligence and Law*, Springer. **29**(4), p. 485–539.

[37] Goldberg, Y. (2017), Neural network methods for natural language processing, *Synthesis Lectures on Human Language Technologies*, Morgan \& Claypool Publishers. **10**(1), p. 1–309.

[38] Graves, A. (2012), Long short-term memory, *Supervised Sequence Labelling with*

Recurrent Neural Networks, Springer. p. 37–45.

[39] Graves, A. and Schmidhuber, J. (2005), Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Networks*, Elsevier. **18**(5–6), p. 602–10.

[40] Grishman, R. (2012), Information extraction: Capabilities and challenges, *Notes Prepared for The,*.

[41] Grishman, R., Huttunen, S. and Yangarber, R. (2002), Information extraction for enhanced access to disease outbreak reports, *Journal of Biomedical Informatics*, Elsevier. **35**(4), p. 236–46.

[42] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002), Gene selection for cancer classification using support vector machines, *Machine Learning*, Springer. **46**(1), p. 389–422.

[43] Ha, Q.-V., Nguyen-Hoang, B.-D. and Nghiem, M.-Q. (2016), Lifelong Learning for Cross-Domain Vietnamese Sentiment Classification, *International Conference on Computational Social Networks*, p. 298–308.

[44] Hasegawa, T., Sekine, S. and Grishman, R. (2004), Discovering relations among named entities from large corpora, *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 415–22.

[45] He, B., Patel, M., Zhang, Z. and Chang, K.C.-C. (2007), Accessing the deep web, *Communications of the ACM*, ACM New York, NY, USA. **50**(5), p. 94–101.

[46] He, R., Lee, W.S., Ng, H.T. and Dahlmeier, D. (2017), An unsupervised neural attention model for aspect extraction, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 388–97.

[47] Hercig, T., Brychc\'\in, T., Svoboda, L. and Konkol, M. (2016), Uwb at semeval-2016 task 5: Aspect based sentiment analysis, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 342–9.

[48] Hui, Y., Wang, J., Cheng, N., Yu, F., Wu, T. and Xiao, J. (2021), Joint Intent Detection and Slot Filling Based on Continual Learning Model, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p.

7643–7.

[49] Ireson, N., Ciravegna, F., Califf, M.E., Freitag, D., Kushmerick, N. and Lavelli, A. (2005), Evaluating machine learning for information extraction, *Proceedings of the 22nd International Conference on Machine Learning*, p. 345–52.

[50] Ji, D., Gao, J., Fei, H., Teng, C. and Ren, Y. (2020), A deep neural network model for speakers coreference resolution in legal texts, *Information Processing \& Management*, **57**(6), p. 102365.

[51] Jiang, X., Wang, Q., Li, P. and Wang, B. (2016), Relation extraction with multi-instance multi-label convolutional neural networks, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, p. 1471–80.

[52] Jihan, N., Senarath, Y., Tennekoon, D., Wickramarathne, M. and Ranathunga, S. (2017), Multi-domain aspect extraction using support vector machines, *Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017)*, p. 308–22.

[53] Judith Jeyafreeda Andrew, X.T. (2018), Automatic Extraction of Entities and Relation from Legal Documents, *Proceedings of the Seventh Named Entities Workshop, ACL*, p. 1–8.

[54] Kambhatla, N. (2004), Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction, *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, p. 178–81.

[55] Kanapala, A., Pal, S. and Pamula, R. (2019), Text summarization from legal documents: a survey, *Artificial Intelligence Review*, **51**, p. 371–402.

[56] Kien, P.M., Nguyen, H.-T., Bach, N.X., Tran, V., Nguyen, M. Le and Phuong, T.M. (2020), Answering Legal Questions by Learning Neural Attentive Text Representation, *Proceedings of the 28th International Conference on Computational Linguistics*, p. 988–98.

[57] Kieu, B.T. and Pham, S.B. (2010), Sentiment analysis for Vietnamese, *2010 Second International Conference on Knowledge and Systems Engineering*, p. 152–7.

[58] Kudo, T., Yamamoto, K. and Matsumoto, Y. (2004), Applying conditional random fields to Japanese morphological analysis, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, p. 230–7.

[59] Kuhn, H.W. (1955), The Hungarian Method for the assignment problem, *Naval Research Logistics Quarterly*, **2**, p. 83–97.

[60] Lafferty, J., McCallum, A. and Pereira, F.C.N. (2001), Conditional random fields: Probabilistic models for segmenting and labeling sequence data,

[61] Lawrence, S., Giles, C.L. and Bollacker, K. (1999), Digital libraries and autonomous citation indexing, *Computer*, IEEE. **32**(6), p. 67–71.

[62] Le, H.S., Van Le, T. and Pham, T.V. (2015), Aspect analysis for opinion mining of Vietnamese text, *2015 International Conference on Advanced Computing and Applications (ACOMP)*, p. 118–23.

[63] Leitner, E., Rehm, G. and Moreno-Schneider, J. (2019), Fine-grained named entity recognition in legal documents, *Semantic Systems The Power of AI and Knowledge Graphs: 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9--12, 2019, Proceedings*, p. 272–87.

[64] Li, J., Sun, A., Han, J. and Li, C. (2020), A survey on deep learning for named entity recognition, *IEEE Transactions on Knowledge and Data Engineering*, IEEE. **34**(1), p. 50–70.

[65] Li, Q. and Ji, H. (2014), Incremental Joint Extraction of Entity Mentions and Relations, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 402–12.

[66] Lin, Y., Shen, S., Liu, Z., Luan, H. and Sun, M. (2016), Neural relation extraction with selective attention over instances, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 2124–33.

[67] Liu, B. (2012), Sentiment analysis and opinion mining, *Synthesis Lectures on Human Language Technologies*, Morgan \& Claypool Publishers. **5**(1), p. 1–167.

[68] Loshchilov, I. and Hutter, F. (2019), Decoupled Weight Decay Regularization, *International Conference on Learning Representations*,.

[69] De Maat, E., Winkels, R. and Van Engers, T. (2006), Automated Detection of Reference, *Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference*, p. 41.

[70] Mandal, A., Ghosh, K., Ghosh, S. and Mandal, S. (2021), A sequence labeling model for catchphrase identification from legal case documents, *Artificial Intelligence and Law*, Springer. p. 1–34.

[71] Mart\'\inez-González, M., la Fuente, P. de and Vicente, D.-J. (2005), Reference extraction and resolution for legal texts, *International Conference on Pattern Recognition and Machine Intelligence*, p. 218–21.

[72] Martins, P.H., Marinho, Z. and Martins, A.F.T. (2019), Joint Learning of Named Entity Recognition and Entity Linking, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 190–6.

[73] McCallum, A., Freitag, D. and Pereira, F.C.N. (2000), Maximum entropy Markov models for information extraction and segmentation., *Icml*, p. 591–8.

[74] McCallum, A., Nigam, K., Reed, J., Rennie, J. and Seymore, K. (2000), Cora: Computer science research paper search engine,

[75] Michelson, M. and Knoblock, C.A. (2005), Semantic annotation of unstructured and ungrammatical text, *IJCAI*, p. 1091–8.

[76] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *Advances in Neural Information Processing Systems*, **26**.

[77] Miwa, M. and Bansal, M. (2016), End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1105–16.

[78] Mukherjee, A. and Liu, B. (2012), Aspect extraction through semi-supervised modeling, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 339–48.

[79] Muslea, I., Minton, S. and Knoblock, C.A. (2001), Hierarchical wrapper induction for semistructured information sources, *Autonomous Agents and Multi-Agent Systems*,

Springer. **4**(1), p. 93–114.

[80] Nayak, T. and Ng, H.T. (2020), Effective Modeling of Encoder-Decoder Architecture for Joint Entity and Relation Extraction, *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence*, p. 8528–35.

[81] Nebhi, K. (2013), A rule-based relation extraction system using DBpedia and syntactic parsing, *Proceedings of the NLP-DBPEDIA-2013 Workshop Co-Located with the 12th International Semantic Web Conference (ISWC 2013)*,.

[82] Nguyen, D.Q. and Nguyen, A.T. (2020), PhoBERT: Pre-trained language models for Vietnamese, *ArXiv Preprint ArXiv:200300744*,.

[83] Nguyen, H.-T., Nguyen, V.-H. and Vu, V.-A. (2017), A knowledge representation for Vietnamese legal document system, *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, p. 30–5.

[84] Nguyen, H.T. and Le Nguyen, M. (2018), Effective attention networks for aspect-level sentiment classification, *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, p. 25–30.

[85] Nguyen, H.T.M., Nguyen, H. V, Ngo, Q.T., Vu, L.X., Tran, V.M., Ngo, B.X. et al. (2018), VLSP shared task: sentiment analysis, *Journal of Computer Science and Cybernetics*, **34**(4), p. 295–310.

[86] Nguyen, L.T. and Nguyen, D.Q. (2021), PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, {NAACL-HLT} 2021, Online, June 6-11, 2021*, p. 1–7.

[87] Nguyen, T.H. and Grishman, R. (2014), Employing word representations and regularization for domain adaptation of relation extraction, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 68–74.

[88] Nguyen, T.H. and Grishman, R. (2015), Relation extraction: Perspective from convolutional neural networks, *Proceedings of the 1st Workshop on Vector Space*

Modeling for Natural Language Processing, p. 39–48.

[89] Nguyen, T.H. and Shirai, K. (2015), Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 2509–14.

[90] Palmirani, M., Brighi, R. and Massini, M. (2003), Automated extraction of normative references in legal texts, *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, p. 105–6.

[91] Pang, B., Lee, L. and others. (2008), Opinion mining and sentiment analysis, *Foundations and Trends® in Information Retrieval*, Now Publishers, Inc. **2**(1--2), p. 1–135.

[92] Phu, V.N., Chau, V.T.N., Tran, V.T.N., Duy, D.N. and Duy, K.L.D. (2019), A valence-totaling model for Vietnamese sentiment classification, *Evolving Systems*, Springer. **10**(3), p. 453–99.

[93] Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J. and Leser, U. (2006), AliBaba: PubMed as a graph, *Bioinformatics*, Oxford University Press. **22**(19), p. 2444–5.

[94] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M. et al. (2016), Semeval-2016 task 5: Aspect based sentiment analysis, *International Workshop on Semantic Evaluation*, p. 19–30.

[95] Popescu, A.-M. and Etzioni, O. (2007), Extracting product features and opinions from reviews, *Natural Language Processing and Text Mining*, Springer. p. 9–28.

[96] Qin, L., Liu, T., Che, W., Kang, B., Zhao, S. and Liu, T. (2021), A co-interactive transformer for joint slot filling and intent detection, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 8193–7.

[97] Quaresma, P. and Gonçalves, T. (2010), Using linguistic information and machine learning techniques to identify entities from juridical documents, *Semantic Processing of Legal Texts*, Springer. p. 44–59.

[98] Quinlan, J.R. (1986), Induction of decision trees, *Machine Learning*, Springer. **1**(1), p. 81–106.

[99] Rabiner, L.R. (1989), A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, Ieee. **77**(2), p. 257–86.

[100] Rish, I. and others. (2001), An empirical study of the naive Bayes classifier, *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, p. 41–6.

[101] Sarawagi, S. (2008), Information Extraction, *Foundations and Trends in Databases*, **1**(3), p. 261–377. http://dx.doi.org/10.1561/1500000003

[102] Sarawagi, S. and Bhamidipaty, A. (2002), Interactive deduplication using active learning, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 269–78.

[103] Seymore, K., McCallum, A., Rosenfeld, R. and others. (1999), Learning hidden Markov model structure for information extraction, *AAAI-99 Workshop on Machine Learning for Information Extraction*, p. 37–42.

[104] Sha, F. and Pereira, F. (2003), Shallow parsing with conditional random fields, *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, p. 213–20.

[105] Soentpiet, R. and others. (1999), Advances in kernel methods: support vector learning, MIT press.

[106] Son, N.T., Duyen, N.T.P., Quoc, H.B. and Le Minh, N. (2015), Recognizing logical parts in vietnamese legal texts using conditional random fields, *The 2015 IEEE RIVF International Conference on Computing \& Communication Technologies-Research, Innovation, and Vision for Future (RIVF)*, p. 1–6.

[107] Song, D., Vold, A., Madan, K. and Schilder, F. (2022), Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training, *Information Systems*, Elsevier. **106**, p. 101718.

[108] Sui, D., Chen, Y., Liu, K., Zhao, J., Zeng, X. and Liu, S. (2020), Joint Entity and Relation Extraction with Set Prediction Networks, *CoRR*,.

[109] Sun, A., Grishman, R. and Sekine, S. (2011), Semi-supervised relation extraction with large-scale word clustering, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, p. 521–9.

[110] Sun, C., Lv, L., Liu, T. and Li, T. (2021), A joint model based on interactive gate mechanism for spoken language understanding, *Applied Intelligence*, p. 1–8.

[111] Sutskever, I., Vinyals, O. and Le, Q. V. (2014), Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems*, **27**.

[112] Tran, O.T., Ngo, B.X., Nguyen, M. Le and Shimazu, A. (2014), Automated reference resolution in legal texts, *Artificial Intelligence and Law*, Springer. **22**(1), p. 29–60.

[113] Turian, J., Ratinov, L. and Bengio, Y. (2010), Word representations: a simple and general method for semi-supervised learning, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 384–94.

[114] Turmo, J., Ageno, A. and Catala, N. (2006), Adaptive information extraction, *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA. **38**(2), p. 4-- es.

[115] Tuyet, H.N.T., Hanh, T. and Cong, T.H. (2015), Extracting semantic relations between vietnamese legislative documents, *2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*, p. 191–6.

[116] Vapnik, V.N. (1999), An overview of statistical learning theory, *IEEE Transactions on Neural Networks*, IEEE. **10**(5), p. 988–99.

[117] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. et al. (2017), Attention is all you need, *Advances in Neural Information Processing Systems*, p. 5998–6008.

[118] Vu, T.-T., Pham, H.-T., Luu, C.-T. and Ha, Q.-T. (2011), A feature-based opinion mining model on product reviews in Vietnamese, *Semantic Methods for Knowledge Management and Communication*, Springer. p. 23–33.

[119] Vu, T., Nguyen, D.Q., Nguyen, D.Q., Dras, M. and Johnson, M. (2018), VnCoreNLP: A Vietnamese natural language processing toolkit, *ArXiv Preprint ArXiv:180101331*,.

[120] Walter, S. (2008), Linguistic Description and Automatic Extraction of Definitions from German Court Decisions., *LREC*,.

[121] Wang, J. and Lu, W. (2020), Two are Better than One: Joint Entity and Relation

Extraction with Table-Sequence Encoders, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics. p. 1706–21.

[122] Wang, W., Pan, S.J., Dahlmeier, D. and Xiao, X. (2016), Recursive neural conditional random fields for aspect-based sentiment analysis, *ArXiv Preprint ArXiv:160306679*,.

[123] Wang, Y., Huang, M., Zhu, X. and Zhao, L. (2016), Attention-based LSTM for aspect-level sentiment classification, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 606–15.

[124] Wang, Y., Sun, C., Wu, Y., Zhou, H., Li, L. and Yan, J. (2021), UniRE: A Unified Label Space for Entity Relation Extraction, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 220–31.

[125] Wei, Z., Su, J., Wang, Y., Tian, Y. and Chang, Y. (2020), A Novel Cascade Binary Tagging Framework for Relational Triple Extraction, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1476–88.

[126] Xenos, D., Theodorakakos, P., Pavlopoulos, J., Malakasiotis, P. and Androutsopoulos, I. (2016), AUEB-ABSA at SemEval-2016 Task 5: Ensembles of Classifiers and Embeddings for Aspect Based Sentiment Analysis., *SemEval@ NAACL-HLT*, p. 312–7.

[127] Xiao, C., Hu, X., Liu, Z., Tu, C. and Sun, M. (2021), Lawformer: A pre-trained language model for chinese legal long documents, *AI Open*, Elsevier. **2**, p. 79–84.

[128] Xuan Bach, N., Khuong Duy, T. and Minh Phuong, T. (2019), A POS tagging model for vietnamese social media text using BiLSTM-CRF with rich features, *Pacific Rim International Conference on Artificial Intelligence*, p. 206–19.

[129] Xue, W. and Li, T. (2018), Aspect based sentiment analysis with gated convolutional networks, *ArXiv Preprint ArXiv:180507043*,.

[130] Yang, J. and Zhang, Y. (2018), Ncrf++: An open-source neural sequence labeling toolkit, *ArXiv Preprint ArXiv:180605626*,.

[131] Young, T., Hazarika, D., Poria, S. and Cambria, E. (2018), Recent trends in deep

learning based natural language processing, *Ieee Computational IntelligenCe Magazine*, IEEE. **13**(3), p. 55–75.

[132] Yu, B., Zhang, Z., Shu, X., Liu, T., Wang, Y., Wang, B. et al. (2020), Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy, *Proceedings of the 24th European Conference on Artificial Intelligence - ECAI*,.

[133] Zeng, D., Liu, K., Lai, S., Zhou, G. and Zhao, J. (2014), Relation classification via convolutional deep neural network, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, p. 2335–44.

[134] Zeng, D., Zhang, H. and Liu, Q. (2020), CopyMTL: Copy Mechanism for Joint Extraction of Entities and Relations with Multi-Task Learning, *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, p. 9507–14.

[135] Zeng, X., Zeng, D., He, S., Liu, K. and Zhao, J. (2018), Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 506–14.

[136] Zhang, C., Zhang, X., Jiang, W., Shen, Q. and Zhang, S. (2009), Rule-based extraction of spatial relations in natural language text, *2009 International Conference on Computational Intelligence and Software Engineering*, p. 1–4.

[137] Zheng, H., Wen, R., Chen, X., Yang, Y., Zhang, Y., Zhang, Z. et al. (2021), PRGC: Potential Relation and Global Correspondence Based Joint Relational Triple Extraction, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 6225–35.

[138] Zheng, L., Guha, N., Anderson, B.R., Henderson, P. and Ho, D.E. (2021), When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, p. 159–68.

[139] Zheng, S., Hao, Y., Lu, D., Bao, H., Xu, J., Hao, H. et al. (2017), Joint entity and relation extraction based on a hybrid neural network, *Neurocomputing*, p. 59–66.