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

TÓM TẮT LUẬN ÁN TIẾN SĨ KỸ THUẬT

Hà Nội - 2023

Công trình được hoàn thành tại:

Học viện Công nghệ Bưu chính Viễn thông

Người hướng dẫn khoa học:

1. GS.TS. Từ Minh Phương
2. PGS.TS. Ngô Xuân Bách

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án được bảo vệ trước Hội đồng chấm luận án cấp Học viện họp tại:

Học viện Công nghệ Bưu chính Viễn thông

Vào hồi ……. ngày …… tháng …… năm …….

Có thể tìm hiểu luận án tại:

Thư viện Học viện Công nghệ Bưu chính Viễn thông

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. 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. *Trích xuất thông tin* thực hiện trích xuất tự động những thông tin có cấu trúc như các thực thể, mối quan hệ giữa các thực thể, các ý kiến/quan điểm mô tả 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 hơn để có thể tiếp tục xử lý, nhằm hỗ trợ tốt hơn cho người dùng.

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 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: (1) trích xuất các loại khía cạnh và (2) phân loại quan điểm cho khía cạnh (đã được trích xuất). 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 cho toàn bộ văn bản đầu vào).

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 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. Văn bản pháp quy là những văn bản do cơ quan Nhà nước ban hành để điều tiết hoạt động của Nhà nước và xã hội, có số lượng lớn và được gia tăng, cập nhật theo thời gian. 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 pháp quy tự động, như tìm kiếm, tra cứu, phân tích, truy vấn, nhằm hỗ trợ tốt hơn cho người dùng.

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.

3. Các đóng góp của luận án

Đóng góp thứ nhất 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.

Đóng góp thứ hai 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.

Đóng góp thứ ba là đề xuất phương pháp trích xuất kết hợp đồng thời thực thể và quan hệ trong văn bản pháp quy tiếng Việt sử dụng mô hình dựa trên học sâu.

4. Bố cục của luận án

Nội dung luận án được tổ chức thành bốn chương. **Chương 1.** Tổng quan về trích xuất thông tin tự động từ văn bản. **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 [4, 6]. **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 [1, 5]. **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 [2, 3]. 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

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,…). Có thể chia thành bốn nhóm bài toán trích xuất thông tin: 1) Trích xuất thực thể có tên; 2) Trích xuất ý kiến/quan điểm mô tả thực thể; 3) Trích xuất quan hệ; 4) Trích xuất sự kiện và kịch bản.

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).

1.2. Các phương pháp tiếp cận dựa trên học máy để giải quyết các bài toán trích xuất thông tin

1) *Phương pháp tiếp cận dựa trên phân loại:* quy bài toán trích xuất thông tin về bài toán phân loại sử dụng các phương pháp học có giám sát. 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, Cây quyết định, Máy véc-tơ tựa (SVM). Trong đó, SVM được đánh giá là một kỹ

thuật phân lớp có độ chính xác cao đối với nhiều bài toán phân loại khác nhau trong xử lý ngôn ngữ tự nhiên.

2) *Phương pháp tiếp cận dựa trên gán nhãn chuỗi:* coi bài toán trích xuất thông tin như là một nhiệm vụ gán nhãn chuỗi. Một số mô hình gán nhãn chuỗi được sử dụng rộng rãi bao gồm: mô hình Markov ẩn, Mô hình Markov cực đại hóa Entropy và Trường ngẫu nhiên có điều kiện (CRF). Trong đó, CRF là phương pháp được sử dụng phổ biến nhất và rất hiệu quả trong nhiều bài toán gán nhãn chuỗi.

3) Phương pháp tiếp cận sử dụng 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. Một số phương pháp học sâu được sử dụng cho trích xuất thông tin: Kỹ thuật nhúng từ, Mạng nơ-ron hồi quy, LSTM (Long Short-Term Memory), Mô hình Seq2Seq, Cơ chế Attention, Transformer.

1.3. Phương pháp thực nghiệm và đánh giá kết quả

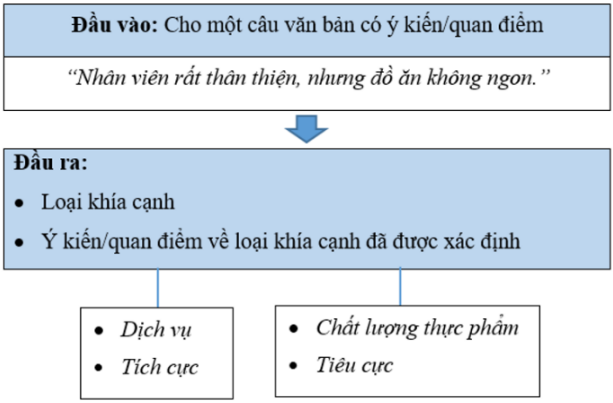
Các bước thực hiện thực nghiệm như sau: thu thập và gán nhãn dữ liệu, trích chọn đặc trưng, huấn luyện mô hình học máy, kiểm tra mô hình với các mẫu dữ liệu mới, và đánh giá kết quả. Để đánh giá kết quả, thực nghiệm sẽ được tiến hành nhiều lần trên tập dữ liệu, theo phương pháp kiểm tra chéo. Kết quả 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.

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

Nội dung Chương 2 trình bày 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).

2.1. Trích xuất khía cạnh và phân loại quan điểm

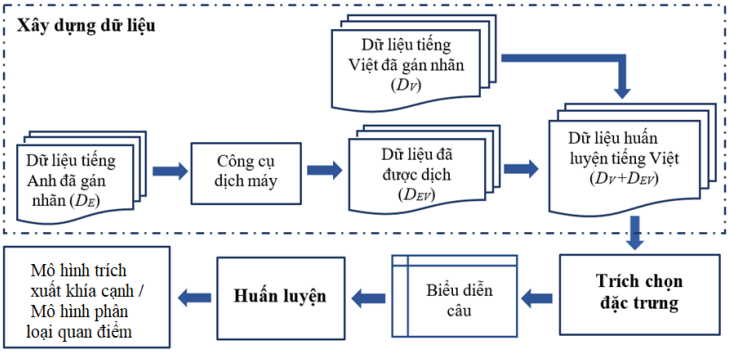
Trích xuất khía cạnh và phân loại quan điểm là hai nhiệm vụ trong bài toán khai phá quan điểm dựa trên khía cạnh, trong đó: (1) Trích xuất các loại khía cạnh, nghĩa là thực hiện xác định danh mục khía cạnh (cặp thực thể và thuộc tính), mà có một ý kiến/quan điểm được thể hiện trong văn bản; và (2) Phân loại quan điểm, nghĩa là thực hiện gán nhãn quan điểm cho từng loại khía cạnh đã được xác định trong nhiệm vụ (1). (Hình 2.1).



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

2.2. Đề xuất phương pháp trích xuất khía cạnh và phân loại quan điểm cho tiếng Việt

Phương pháp tổng thể đề xuất để giải quyết cả hai nhiệm vụ trích xuất khía cạnh và phân loại quan điểm bao gồm ba bước chính (Hình 2.2): (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.



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

1) *Xây dựng dữ liệu:* Sự khác biệt của phương pháp đề xuất là 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). Dữ liệu tiếng Anh được dịch sang tiếng Việt bằng một công cụ dịch tự động (Google Translate).
2) *Trích chọn đặc trưng:* Với nhiệm vụ *trích xuất khía cạnh*, hai loại đặc trưng được sử dụng là đặc trưng cơ bản (*n*-grams tiếng Việt) và nhúng từ. Với nhiệm vụ *phân loại*

quan điểm, ba loại đặc trưng được sử dụng là: từ quan trọng, nhúng từ, và đặc trưng loại khía cạnh.

3) *Mô hình huấn luyện:* Cho *N* là số lượng các loại khía cạnh muốn trích xuất, nghiên cứu thực hiện huấn luyện *N* bộ phân loại cho *N* loại khía cạnh và một bộ phân loại để xác định loại quan điểm. Thuật toán học có giám sát được sử dụng là Máy véc-tơ tựa.

2.3. Xây dựng tập dữ liệu

Tập dữ liệu tiếng Việt được thu thập từ trang web Foody (có tại: https://www.foody.vn/). Tập dữ liệu tiếng Anh được trích xuất từ nhiệm vụ 5 trong SemEval-2016. Dữ liệu tiếng Việt được thu thập, tiền xử lý và gán nhãn dữ liệu với các nhãn loại khía cạnh và loại quan điểm. (Bảng 2.1).

Bảng 2.1. Loại khía cạnh và quan điểm 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	**1357**	886	235	41	**1162**
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	**3062**	**881**	**422**	**4365**	**2268**	**953**	**145**	**3366**

2.4. Kết quả thực nghiệm

1) Các mô hình thực nghiệm: Mô hình *Cơ sở* (*baseline*), *CRL* (Cross-Language), và *WEmb* (Word Embedding) (Bảng 2.2).

Bảng 2.2. 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) Kết quả thử nghiệm:

Kết quả trích xuất khía cạnh (Bảng 2.3)**:** So với mô hình cơ sở, mô hình CRL đạt được 9/12 loại khía cạnh cao hơn, cho thấy hiệu quả của việc sử dụng dữ liệu dịch bổ sung cho trích xuất khía cạnh. 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. 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.

Kết quả phân loại quan điểm (Bảng 2.4)**:** Với tất cả các mô hình, độ đo F_1 của nhãn tích cực cao hơn nhiều so với nhãn tiêu cực: 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). Có hai lý do chính: 1) số lượng các mẫu tích cực trong các tập dữ liệu đều cao hơn nhiều so với số lượng các mẫu tiêu cực; và 2) quan điểm tích cực thường được nêu trực tiếp và rõ ràng, trong khi quan điểm tiêu cực thường ở dạng tiềm ẩn. Ví dụ câu có quan điểm tiêu cực "*Chúng tôi phải đợi thức ăn khoảng nửa tiếng*.", hay "*Kim chi không cay mà lại hơi ngọt.*".

Bảng 2.3. 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**

Bảng 2.4. Kết quả 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**

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

Nội dung Chương 3 trình bày 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.

3.1. Trích xuất thông tin trong văn bản pháp quy

Trích xuất thông tin trong văn bản pháp quy tiếng Việt được nghiên cứu trong Chương 3 bao gồm 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 (Hình 3.1 trình bày một ví dụ). *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.

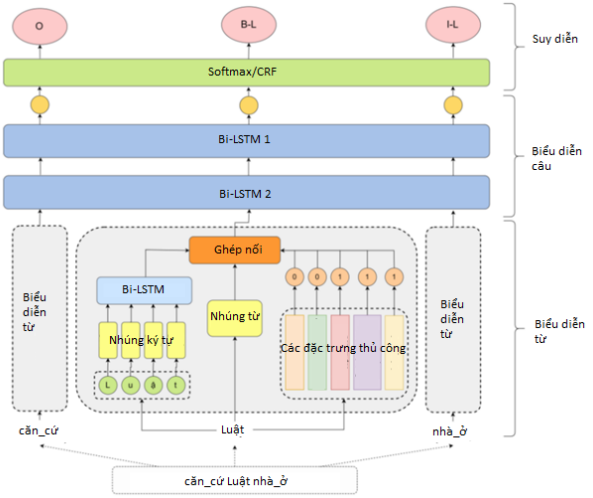
"Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004 của Bộ Tài chính"

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

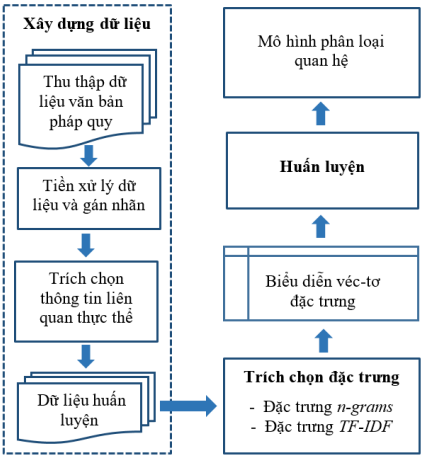
3.2. Đề xuất phương pháp trích xuất thực thể và quan hệ

1) Trích xuất thực thể tham chiếu: Đề xuất 2 mô hình: (1) Mô hình dựa trên CRF, và (2) Mô hình BiLSTM và BiLSTM-CRF. Các mô hình BiLSTM và BiLSTM-CRF bao gồm ba lớp: biểu diễn từ, biểu diễn câu và suy diễn (Hình 3.2).

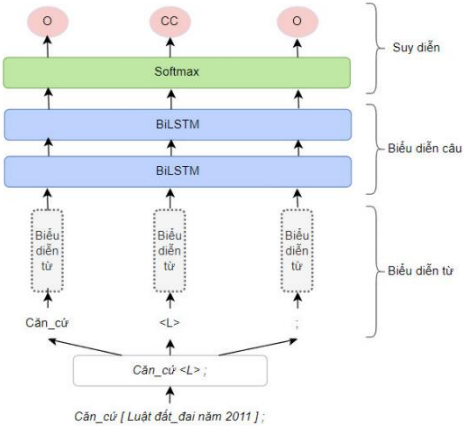


Hình 3.2. Các mô hình BiLSTM và BiLSTM-CRF cho trích xuất thực thể tham chiếu

2) Phân loại quan hệ giữa các thực thể văn bản pháp quy: Được thực hiện với cả hai phương pháp học máy truyền thống (Hình 3.3) và học sâu (Hình 3.4).



Hình 3.3. Phân loại quan hệ giữa các thực thể trong văn bản pháp quy sử dụng học máy truyền thống



Hình 3.4. Mô hình BiLSTM cho phân loại quan hệ giữa các thực thể văn bản pháp quy

3.3. Xây dựng tập 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. Dữ liệu được thu thập, tiền xử lý và gán nhãn dữ liệu với hai nhãn loại thực thể tham chiếu và loại quan hệ. Các thông tin thống kê dữ liệu được trình bày trong Bảng 3.1, 3.2.

Bảng 3.1 Thông tin thống kê về các loại thực thể tham chiếu

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**

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

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. Kết quả thực nghiệm

1) Trích xuất thực thể tham chiếu

Kết quả trong Bảng 3.3 cho thấy: 1) Tất cả các mô hình đều có kết quả khá cao (từ 95,78% đến 96,62% tính theo độ đo F_1); 2) 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), 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. 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**

Hiệu năng của mô hình BiLSTM-CRF trên từng loại thực thể tham chiếu thu được tương đối tốt trên hầu hết các loại thực thể tham chiếu (Bảng 3.4), 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" 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%), 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 3.4. Hiệu năng của mô hình BiLSTM-CRF trên từng loại thực thể tham chiếu

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

2) Phân loại quan hệ giữa các thực thể văn bản pháp quy

Phương pháp sử dụng học máy truyền thống (Bảng 3.5): phương pháp kết hợp đặc trưng *n*-grams và TF-IDF cho kết quả tốt hơn, đạt được độ chính xác là 95,68%, độ phủ là 95,67% và độ đo F_1 là 95,57%.

Phương pháp sử dụng học sâu (Bảng 3.6): Kết quả thực nghiệm phân loại quan hệ giữa các thực thể văn bản pháp quy với mô hình BiLSTM đề xuất cho kết quả tốt hơn so với phương pháp học máy truyền thống tốt nhất (SVM). 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%.

Bảng 3.5. Kết quả phân loại quan hệ sử dụng SVM (%)

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 3.6. 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**

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

Nội dung Chương 4 trình bày đề xuất phương pháp trích xuất kết hợp đồng thời thực thể tham chiếu và quan hệ giữa các thực thể trong văn bản pháp quy tiếng Việt sử dụng kiến trúc bộ mã hóa-giải mã dựa trên Transformer với cơ chế giải mã song song không tự hồi quy.

4.1. Đặt vấn đề

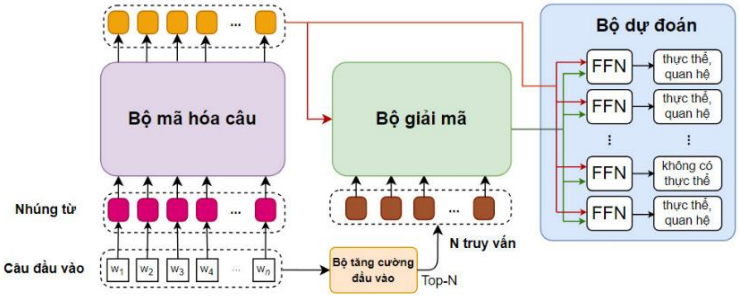
Nghiên cứu trong Chương 3 đề xuất phương pháp trích xuất các thông tin về thực thể tham chiếu và quan hệ giữa các thực thể văn bản pháp quy theo cách tuần tự, đầu tiên (1) trích xuất thực thể tham chiếu, và sau đó (2) phân loại quan hệ giữa thực thể tham chiếu đã được trích xuất và thực thể văn bản đang 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.

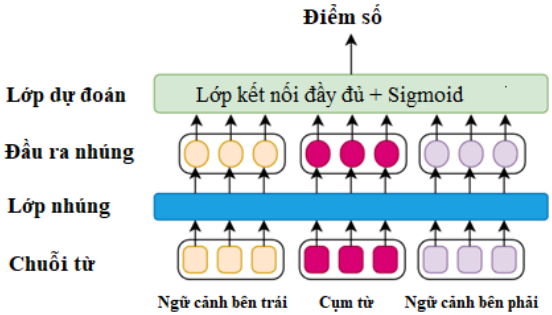
4.2. Đề xuất mô hình trích xuất kết hợp thực thể và quan hệ

Mô hình đề xuất sẽ thực hiện xử lý theo từng câu s (được biểu diễn dưới dạng một chuỗi n từ $s = t_1t_2...t_n$) 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 4.1. Minh họa kiến trúc của mô hình đề xuất

Kiến trúc tổng thể của mô hình bao gồm bốn thành phần chính (Hình 4.1): bộ mã hóa câu, bộ tăng cường đầu vào, bộ giải mã và bộ dự đoán.



Hình 4.2. Bộ tăng cường đầu vào

Nghiên cứu ở đây khác các nghiên cứu trước là đề xuất sử dụng phương pháp tăng cường đầu vào bộ giải mã với các thông tin đầu mối quan trọng của văn bản tham chiếu nhằm cải thiện hiệu năng của mô hình trích xuất kết hợp. Bộ tăng cường đầu vào được huấn luyện độc lập với mô hình trích xuất kết hợp (Hình 4.2).

4.3. Kết quả thực nghiệm

1) Các mô hình thực nghiệm: 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 (đã được đánh giá tốt): CasRel, SPERT, JointER và SPN. Các thực nghiệm được thực hiện trên tập dữ liệu đã được xây dựng trong Chương 3 của luận án.

2) Kết quả thử nghiệm

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 (Bảng 4.1). 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% 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.

Bảng 4.1. Kết quả so sánh các mô hình trích xuất

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**

Kết quả sử dụng bộ tăng cường đầu vào (Bảng 4.2): 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 4.2. 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

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 đóng góp 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*. Phương pháp đề xuất 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). Kết quả có trong các công trình [4, 6].
2) *Nghiên cứu 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*. Kết quả có trong các công trình [1, 5].
3) *Nghiên cứu 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 (khác với nghiên cứu trong đóng góp thứ hai thực hiện trích xuất các thông tin này theo cách tuần tự). Kết quả có trong các công trình [2, 3].

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, *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, *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, *RIVF 2020*, pp. 1-6, 2020.