ĐẠI HỌC QUỐC GIA HÀ NỘI
ĐẠI HỌC CÔNG NGHỆ

VƯƠNG THỊ HẢI YẾN

MÔ HÌNH HÓA VÀ HỌC CÁC MỐI QUAN HỆ VĂN BẢN VÀ CẤU TRÚC CHO VIỆC TRUY HỒI THÔNG TIN PHÁP LUẬT HỌC SÂU

Hanoi, 2024

Tóm tắt

Với sự tiến bộ gần đây trong số hóa và chuyển đổi số, các chuyên gia pháp lý hiện có thể dễ dàng truy cập một lượng lớn tài liệu pháp lý trực tuyến. Điều này cực kỳ quan trọng vì thường xuyên các chuyên gia pháp lý cần tìm thông tin pháp lý liên quan khi làm việc với một vụ án mới, thực hiện nghiên cứu pháp lý, phân tích vụ án, chuẩn bị tài liệu trước phiên tòa, cung cấp lời khuyên pháp lý cho khách hàng, hoặc ra quyết định về một vụ án hiện tại. Tuy nhiên, cơ sở dữ liệu pháp lý vàng lớn càng khó để họ tìm được tài liệu liên quan theo cách thủ công. Ngoài ra, các tài liệu pháp lý như luật hành văn, án lệ hoặc hợp đồng thường dài và phức tạp, bao gồm nhiều phần, chương, mục, điều, khoản, và vân vân. Do đó, xây dựng một hệ thống truy hồi thông tin pháp lý (information retrieval-IR) thông minh và tự động là điều quan trọng để cải thiện và tăng tốc quy trình và luồng công việc của họ. Tổng thể, luận án này nhằm đề xuất các phương pháp và giải pháp IR pháp lý khác nhau dựa trên sự hiểu biết sâu rộng về tính chất và đặc điểm của dữ liệu pháp lý cũng như sự phức tạp của các vấn đề IR pháp lý.

Do đó, hai vấn đề chính mà chúng tôi cần cân nhắc kỹ lưỡng trong nghiên cứu này là tài liệu pháp lý và các vấn đề IR pháp lý. Tài liệu pháp lý là đa dạng, bao gồm nhiều loại tài liệu khác nhau như hiến pháp, pháp luật, quy định, quyết định, án lệ, tài liệu tòa án, hợp đồng, thông báo pháp lý, bằng sáng chế,v.v. Trong số đó, chúng tôi tập trung vào hai loại văn bản pháp lý chính – luật hành văn và án lệ – vì làm việc với tất cả các loại tài liệu pháp lý là quá rộng lớn và vượt ra ngoài phạm vi của luận án. Liên quan đến các vấn đề IR pháp lý, nghiên cứu này tập trung vào ba nhiệm vụ IR chính: (i) truy hồi án lệ; (ii) truy hồi luật – vụ án; và (iii) trả lời câu hỏi pháp luật dựa trên IR. Nhiệm vụ đầu tiên tìm kiếm và trả về tài liệu án lệ từ một cơ sở dữ liệu án lệ liên quan và hỗ trợ quyết định của một vụ án pháp lý đầu vào. Nhiệm vụ thứ hai truy hồi pháp luật từ cơ sở dữ liệu pháp luật mà liên quan đến một vụ án truy vấn. Và nhiệm vụ thứ ba tìm kiếm và trả về các điều luật có khả năng chứa đựng câu trả lời cho một câu hỏi pháp lý cụ thể.

Ba vấn đề IR pháp lý đã nêu ở trên thách thức nhiều hơn so với IR truyền thống cho các văn bản trong lĩnh vực chung. Khái niệm về sự liên quan trong những nhiệm vụ

này không còn đơn giản là về việc phù hợp từ khóa hoặc chủ đề. Sự tương đồng giữa các văn bản pháp lý đòi hỏi sự hiểu biết về các luận điểm pháp lý và lập luận logic xa vời hơn so với so sánh từ vựng hoặc chủ đề. Ngoài ra, trong khi làm việc với dữ liệu pháp lý, chúng tôi nhận ra rằng ngôn ngữ pháp lý thường chặt chẽ và phức tạp. Các tài liệu pháp lý thường dài và phụ thuộc nặng nề vào các thuật ngữ, biệt ngữ và sự tinh tế ngôn ngữ cụ thể của lĩnh vực. Hơn nữa, có một cấu trúc đồ thị phức tạp được ẩn trong bất kỳ tập dữ liệu pháp lý nào do sự thường xuyên được đề cập, trích dẫn, tham khảo trong và giữa các tài liệu pháp lý. Ngoài ra, phong cách và nội dung của các tài liệu pháp lý phụ thuộc nhiều vào lĩnh vực và hệ thống pháp luật của mỗi quốc gia. Và một vấn đề quan trọng khác nữa là dữ liệu được có nhãn hạn chế vì việc gán nhãn cho dữ liệu pháp lý đòi hỏi nhiều thời gian, công sức và chuyên môn về lĩnh vực pháp lý. Tất cả những lý do này đều là thách thức cũng như động lực trong nghiên cứu này.

Mục tiêu chính của luận án này là nâng cao hiệu suất và độ chính xác của ba vấn đề IR pháp lý bằng cách tận dụng mối quan hệ và cấu trúc trong dữ liệu pháp lý. Đầu tiên, chúng tôi đề xuất một mô hình hỗ trợ mã hóa cả mối quan hệ từ vựng và pháp lý ở các mức độ khác nhau của độ chi tiết để xử lý vấn đề truy hồi án lệ. Ngoài ra, chúng tôi giới thiệu một phương pháp để tự động tạo ra một tập dữ liệu nhãn yếu lớn để vượt qua hạn chế của dữ liệu được gán nhãn. Thứ hai, một biểu đồ tri thức pháp lý không đồng nhất đã được đề xuất và xây dựng để tận dụng các mối quan hệ vụ án pháp luật và luật pháp trong việc truy hồi luật – án lệ. Thứ ba, luận án trình bày một phương pháp mới mà xây dựng một mạng tham khảo điều luật để khám phá cả mối quan hệ cục bộ và xa giữa các điều luật để nâng cao hiệu suất của việc trả lời câu hỏi pháp lý dựa trên IR. Hơn nữa, trong suốt luận án, chúng tôi đề xuất các kiến trúc học sâu phù hợp để mã hóa các đặc điểm văn bản và cấu trúc của dữ liệu pháp lý và kết hợp chúng với các mô hình ngôn ngữ được đào tạo trước mạnh mẽ để nâng cao hiệu suất tổng thể của ba vấn đề IR. Ngoài các đóng góp kỹ thuật, việc đánh giá, phân tích và thảo luận trong suốt luận án này sẽ cung cấp hiểu biết sâu và rõ ràng hơn về bản chất và các hạn chế trong xử lý ngôn ngữ tự nhiên pháp lý nói chung và trong IR pháp lý nói riêng. Điều này cũng có thể là một tài liệu tham khảo tiềm năng cho các nghiên cứu trong tương lai trong lĩnh vực này, đặc biệt là đối với các ngôn ngữ tài nguyên thấp như tiếng Việt.

Từ khóa: luật hành văn, án lệ, vụ án pháp lý, truy hồi thông tin pháp lý sâu, trả lời câu hỏi pháp lý, truy hồi án lệ, truy hồi luật, trả lời câu hỏi pháp luật dựa trên IR, mô hình hỗ trợ, dữ liệu nhãn yếu, liên quan, mối quan hệ văn bản, mối quan hệ cấu trúc, biểu đồ tri thức pháp lý, mạng tham chiếu, mô hình ngôn ngữ huấn luyện trước.

Chương 1

Giới thiệu

1.1 Các vấn đề truy hồi thông tin pháp lý sâu

IR và QA cho văn bản pháp lý là các nhiệm vụ liên quan đến việc truy hồi thông tin liên quan đến một truy vấn đầu vào hoặc tìm ra một câu trả lời chính xác cho một câu hỏi đầu vào. Cũng có nhiều loại tài liệu pháp lý. Do đó, chúng ta có thể có các cách khác nhau để định nghĩa các nhiệm vụ IR và QA. Tuy nhiên, như đã đề cập trong phần trước, trong phạm vi của nghiên cứu này, chúng tôi chỉ làm việc với hai loại tài liệu pháp lý chính: pháp luật và án lệ. Do đó, chúng tôi sẽ giới hạn ba vấn đề IR và QA chính cho hai loại tài liệu pháp lý này trong luận án này. Các vấn đề đó là:

(i) Truy hồi án lệ;

(ii) Truy hồi luật – vụ án;

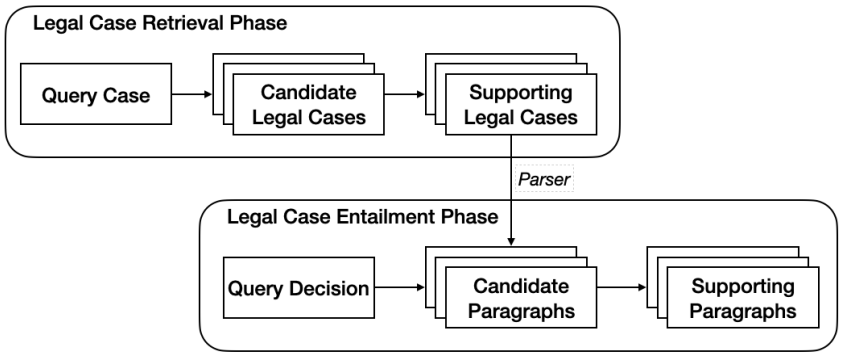
(iii) Hỏi đáp pháp luật dựa trên IR.

Tất cả các ý tưởng và phương pháp mà chúng tôi đề xuất cũng như các đóng góp kỹ thuật của chúng tôi trong luận án này đều xoay quanh ba vấn đề IR và QA này. Chúng tôi sẽ giải thích chi tiết về những vấn đề này ở đây để hiểu rõ hơn về chúng, vì chúng ta sẽ gặp chúng thường xuyên trong toàn bộ luận án này.

Truy hồi án lệ

Với sự tiến bộ gần đây trong số hóa và biến đổi số, các thẩm phán và luật sư hiện có thể dễ dàng truy cập vào một lượng lớn tài liệu pháp lý trực tuyến. Tuy nhiên, càng

nhiều tài liệu pháp lý thì việc tìm kiếm các án lệ liên quan nhất trở nên khó khăn hơn. Do đó, việc phát triển một hệ thống tự động truy hồi án lệ sẽ tăng tốc và cải thiện hiệu suất của quy trình làm việc của thẩm phán và luật sư. Vấn đề truy hồi án lệ được định nghĩa để đáp ứng nhu cầu này. Vấn đề này bao gồm hai giai đoạn (hoặc hai phần con) là Nhiệm vụ 1 (Truy hồi án lệ) và Nhiệm vụ 2 (Kế thừa án lệ) của cuộc thi COLIEE, tương ứng. Hình 1.1 cho thấy hai giai đoạn và luồng logic của vấn đề truy hồi án lệ.



Hình 1.1: Luồng logic của vấn đề truy hồi án lệ

Giai đoạn truy hồi án lệ: Đặt C là không gian của tất cả các vụ án pháp lý và các luật pháp và đặt $C \subset \mathrm{C}$ là một tập hợp các án lệ. Cho một vụ án truy vấn đầu vào $c_q \in \mathrm{C}$. Truy vấn c_q thường là một vụ án pháp lý mới mà một thẩm phán hoặc luật sư đang làm việc. Mục tiêu của giai đoạn này là tìm kiếm và truy hồi một tập hợp tất cả các án lệ có liên quan $C^r = c_1^r, c_2^r, \ldots, c_k^r \subset C$ hỗ trợ cho quyết định của c_q. Trong lĩnh vực pháp lý, những vụ án hỗ trợ này $c_1^r, c_2^r, \ldots, c_k^r$ cũng được gọi là "vụ án đã được chú ý". Giai đoạn truy hồi án lệ này có thể được biểu diễn như một ánh xạ như sau:

$$f_{case\_retrieval}(c_q, C) \rightarrow C^r \tag{1.1}$$

Giai đoạn kế thừa án lệ: Cho một bộ ba bao gồm vụ án truy vấn đầu vào c_q, một quyết định d_q của vụ án truy vấn c_q, và danh sách tất cả các vụ án hỗ trợ C^r được trả về từ giai đoạn trước. Đặt P^r là tập hợp tất cả các đoạn văn được phân đoạn từ một vụ án hỗ trợ cụ thể $c^r \in C^r$. Mục tiêu của giai đoạn này là xác định một tập hợp các đoạn văn hỗ trợ $P^e = p_1^e, p_2^e, \ldots, p_l^e \subset P^r$ được kế thừa để hỗ trợ cho quyết định d_q của vụ án truy vấn c_q. Giai đoạn truy hồi này có thể được biểu diễn như một ánh xạ như sau:

$$f_{case\_entailment}(c_q, d_q, P^r) \rightarrow P^e \tag{1.2}$$

Mối quan hệ kế thừa giữa hai đoạn văn bản pháp lý tương tự như khái niệm của kế thừa văn bản trong hiểu biết và suy luận ngôn ngữ tự nhiên. Đó là mối quan hệ giữa hai

đoạn văn nơi một đoạn (được gọi là *giả thiết*) có thể được suy luận hoặc ngụ ý bởi đoạn văn kia (được gọi là *văn bản* hoặc *tiền đề*). Nói cách khác, nếu văn bản là đúng, thì giả thiết cũng có thể là đúng. Cả việc hỗ trợ trong giai đoạn đầu tiên và việc kế thừa trong giai đoạn thứ hai là mối quan hệ phức tạp dựa trên lý luận pháp lý và logic. Chúng sâu sắc hơn và vượt ra ngoài khái niệm thông thường về sự tương quan trong IR truyền thống chỉ dựa trên sự gần gũi về từ vựng và chủ đề. Đó là lý do tại sao chúng tôi gọi và xem xét những nhiệm vụ này như là các vấn đề **truy hồi thông tin pháp lý sâu**.

Truy hồi luật – vụ án

Truy hồi luật – vụ án nhằm tìm kiếm các văn bản pháp luật có liên quan tới vụ án pháp luật. Mục tiêu của nó là định vị cả các văn bản pháp luật và vụ án pháp luật liên quan đến một truy vấn pháp lý cụ thể, cung cấp một nguồn tài nguyên pháp lý toàn diện bao gồm cả khung pháp luật và tiền lệ tư pháp.

Giả sử S là một bộ luật thành văn (tức là, một cơ sở dữ liệu về luật thành văn). Cho một vụ án truy vấn đầu vào c_q (thông thường, c_q là vụ án pháp lý mới mà các thẩm phán và luật sư đang làm việc), mục tiêu của vấn đề này là định vị và truy hồi tất cả các luật thành văn $S^r = s_1^r, s_2^r, \ldots, s_k^r$ từ bộ luật S mà liên quan nhất đến vụ án truy vấn c_q. Điều này có thể được biểu diễn dưới dạng ánh xạ sau:

$$f_{law\_retrieval}(c_q, S) \rightarrow S^r \tag{1.3}$$

Hỏi đáp pháp luật dựa trên IR

Giả sử A là một danh sách điều luật (tức là, một cơ sở dữ liệu) các điều luật. Cho một câu hỏi q về bất kỳ vấn đề pháp lý nào có thể được bao quát bởi bộ luật A, mục tiêu của vấn đề này là tìm kiếm các điều luật có liên quan nhất $A^r = a_1^r, a_2^r, \ldots, a_k^r$ từ bộ luật A có khả năng chứa các câu trả lời cho câu hỏi đầu vào q. Điều này có thể được biểu diễn dưới dạng ánh xạ sau:

$$f_{statute\_retrieval}(q, A) \rightarrow A^r \tag{1.4}$$

1.2 Câu hỏi nghiên cứu và mục tiêu nghiên cứu

Dựa trên những thách thức và động lực nghiên cứu, cũng như những gì đã được thực hiện trong các nghiên cứu trước đó và những vấn đề còn chưa giải quyết, luận án này đề cập đến các câu hỏi nghiên cứu sau:

- **Q1**: Làm thế nào để xử lý và biểu diễn các văn bản pháp lý phức tạp và dài dòng? Làm thế nào để đề xuất và học các mối quan hệ văn bản pháp lý và sự tương đồng giữa các văn bản pháp lý ở các mức độ chi tiết khác nhau (vụ án, đoạn văn, quyết định . . .) để nâng cao tính liên quan và độ chính xác cho các vấn đề IR và QA?
- **Q2**: Làm thế nào để vượt qua hạn chế của dữ liệu đã được gắn nhãn trong các vấn đề IR và QA pháp lý? Làm thế nào để có được nhiều dữ liệu được gắn nhãn hơn trong lĩnh vực này để cải thiện hiệu suất truy hồi?
- **Q3**: Làm thế nào chúng ta có thể biểu diễn và học các mối quan hệ cấu trúc là các liên kết đồ thị giữa các văn bản pháp lý (ví dụ, tham chiếu cục bộ và xa) và các liên kết giữa các thực thể pháp lý (ví dụ, tòa án, các vụ án, luật pháp, lĩnh vực) để giúp cải thiện hiệu suất của các vấn đề IR và QA?
- **Q4**: Làm thế nào để tích hợp và tận dụng các đặc điểm văn bản và cấu trúc pháp lý với các mô hình học sâu mạnh mẽ (bao gồm các mô hình ngôn ngữ được huấn luyện trước) để cải thiện hiệu suất của các vấn đề IR và QA?

Mục tiêu tổng quan của luận án này là nâng cao hiệu suất và hiệu quả của các vấn đề IR và QA pháp lý theo nhiều cách khác nhau. Chúng tôi có ba mục tiêu cụ thể như sau:

- **O1**: Đề xuất các phương pháp và mô hình mới để nâng cao hiệu quả của các vấn đề IR và QA pháp lý.
- **O2**: Tận dụng và khai thác tối đa bản chất và đặc điểm của dữ liệu pháp lý (tức là, cả các mối quan hệ pháp lý và cấu trúc) để tăng cường hiệu suất của ba vấn đề IR pháp lý được nêu trong Phần 1.1: truy hồi án lệ, truy hồi luật – vụ án và trả lời câu hỏi pháp lý dựa trên IR.
- **O3**: Đề xuất các phương pháp phù hợp để kết hợp và tích hợp các đặc điểm văn bản và cấu trúc của dữ liệu pháp lý với các mô hình học sâu mạnh mẽ (bao gồm các mô

hình ngôn ngữ được đào tạo trước) để cải thiện thêm hiệu quả của các nhiệm vụ IR và QA pháp lý.

1.3 Contributions

Luận án này đóng góp đáng kể ở các khía cạnh khác nhau: học biểu diễn các đặc điểm pháp lý, tăng cường dữ liệu, định nghĩa và tạo ra đồ thị tri thức pháp lý, khám phá và sử dụng các mối quan hệ đồ thị, và tích hợp mô hình học sâu lấy cảm hứng từ đồ thị. Đầu tiên, luận án tập trung vào khám phá và biểu diễn các mối quan hệ pháp lý giữa các văn bản ở các mức độ chi tiết khác nhau để xử lý các tài liệu dài cũng như tận dụng cả các mối quan hệ từ vựng và logic phức tạp vào một mô hình được gọi là *mô hình hỗ trợ* để giải quyết nhiệm vụ truy hồi án lệ. Thứ hai, chúng tôi đề xuất một chiến lược gán nhãn yếu để vượt qua việc thiếu dữ liệu đã được gắn nhãn và cải thiện hiệu suất truy hồi. Thứ ba, chúng tôi định nghĩa và tạo ra *một đồ thị tri thức không đồng nhất* của các loại thực thể pháp lý khác nhau để tăng cường hiệu suất của vấn đề truy hồi luật – vụ án. Chúng tôi cũng định nghĩa và xây dựng *một mạng tham chiếu* để nắm bắt và sử dụng các kết nối hoặc mối quan hệ đồ thị giữa các văn bản pháp lý để cải thiện hiệu suất của nhiệm vụ trả lời câu hỏi. Hơn nữa, suốt toàn bộ luận án này, chúng tôi đề xuất các kiến trúc mô hình sâu để tích hợp mượt mà cả các đặc điểm văn bản và cấu trúc pháp lý của dữ liệu pháp lý để cải thiện hiệu suất của các mô hình IR và QA. Các kiến trúc mô hình được giới thiệu trong phần Phương pháp nghiên cứu thí nghiệm và thực hiện các thí nghiệm để xác nhận tính chính xác và hiệu quả của các mô hình được đề xuất trong luận án đã thể hiện hiệu suất tốt hơn so với các tiêu chuẩn hiện tại, với một số kết quả vượt trội trên các bộ dữ liệu được thiết lập. Việc cải thiện hiệu suất được thể hiện qua các thí nghiệm, phân tích, đánh giá làm sáng tỏ tính hiệu quả của các phương pháp và phương pháp được đề xuất. Cuối cùng, phân tích và thảo luận trong suốt công việc này sẽ giúp cung cấp hiểu biết sâu sắc về các văn bản pháp lý và các vấn đề xử lý, trình bày các tiến bộ và hạn chế còn lại của NLP pháp lý nói chung và IR và QA pháp lý cụ thể; và cũng sẽ đề xuất hướng nghiên cứu IR và QA pháp lý trong tương lai, đặc biệt là cho các ngôn ngữ có tài nguyên thấp như tiếng Việt.

Luận án đưa ra ba đóng góp chính:

- Chúng tôi nghiên cứu mối quan hệ hỗ trợ trong các văn bản pháp luật và đề xuất một phương pháp gọi là mô hình hỗ trợ có thể xử lý cả giai đoạn truy hồi và làm rõ trong nhiệm vụ truy hồi án lệ. Ý tưởng cơ bản là các mối quan hệ hỗ trợ giữa

các vụ án, giữa các đoạn văn và giữa các quyết định và đoạn văn để tăng cường tính liên quan cho việc truy hồi văn bản pháp luật. Ngoài ra, dựa trên mối quan hệ hỗ trợ, chúng tôi cũng đề xuất một phương pháp để tự động tạo ra một tập dữ liệu nhãn yếu lớn để vượt qua sự thiếu hụt dữ liệu được gán nhãn.

- Chúng tôi đề xuất và xây dựng một đồ thị tri thức không đồng nhất bao gồm các loại thực thể pháp luật khác nhau (án lệ, tòa án, pháp luật, và miền pháp luật) để cải thiện tổ chức và truy hồi thông tin pháp luật trong nhiệm vụ truy hồi pháp luật - vụ án.

- Chúng tôi nghiên cứu các mối quan hệ trích dẫn, tham chiếu giữa các điều luật và đề xuất một phương pháp mạng tham chiếu để tăng cường hiệu suất của nhiệm vụ trả lời câu hỏi văn bản pháp luật. Nhúng và mã hóa các tham chiếu cục bộ và các toàn cục (xa) giữa các bài viết pháp luật vào các mô hình ngôn ngữ đào tạo trước giúp mô hình QA trở nên mạnh mẽ và chính xác hơn. Ngoài ra, bằng cách phát hiện các kết nối ẩn giữa các luật, phương pháp của chúng tôi có thể hỗ trợ trong việc xác định những không nhất quán và lỗ hổng trong hệ thống pháp luật, từ đó cải thiện tính hiệu quả và đáng tin cậy của nó.

Luận án tiến sĩ này đóng góp cho cả lĩnh vực khoa học và thực tiễn. Luận án trình bày một cái nhìn toàn diện về NLP miền pháp luật cho truy hồi và trả lời câu hỏi văn bản pháp luật. Nó cũng cung cấp cái nhìn sâu sắc về các đặc điểm của các văn bản pháp luật và mối quan hệ giữa chúng. Ngoài ra, các phương pháp biểu diễn, thiết kế kiến trúc của các mô hình, và các bước thực hiện cho việc huấn luyện và đánh giá các mô hình này được mô tả chi tiết trong luận án này.

Chương 2

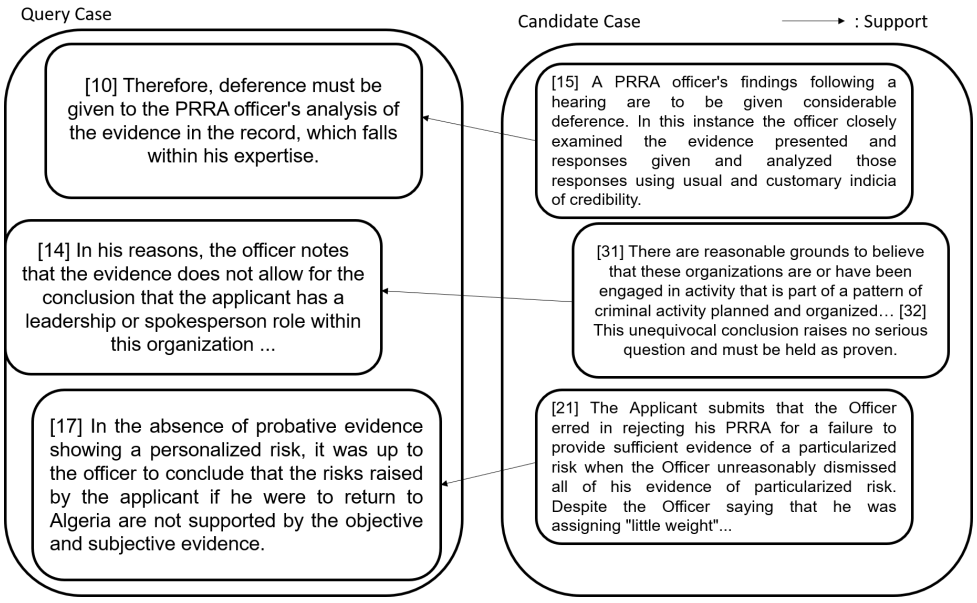
Mô hình hỗ trợ trong truy hồi án lệ

Truy hồi án lệ là nhiệm vụ định vị các văn bản pháp luật thực sự liên quan đến án lệ truy vấn đầu vào. Không giống như việc truy hồi thông tin cho văn bản tổng quát, nhiệm vụ này bao gồm hai giai đoạn (*truy hồi văn án lệ* và *kế thừa án lệ*) và khó khăn hơn nhiều do một số lý do. Thứ nhất, cả truy vấn và các án lệ ứng cử là các tài liệu dài gồm nhiều đoạn văn. Điều này khiến việc mô hình chúng với học biểu diễn thường có hạn chế về độ dài đầu vào trở nên khó khăn. Thứ hai, khái niệm *liên quan* trong lĩnh vực này được định nghĩa dựa trên mối quan hệ pháp lý vượt ra ngoài sự liên quan từ vựng hoặc theo chủ đề. Điều này là một thách thức thực sự vì việc so khớp văn bản thông thường sẽ không hoạt động. Thứ ba, việc xây dựng một bộ dữ liệu văn bản pháp luật lớn và chính xác đòi hỏi rất nhiều công sức và chuyên môn. Điều này rõ ràng là một trở ngại đối với việc tạo ra đủ dữ liệu để huấn luyện các mô hình truy hồi sâu. Trong chương này, chúng tôi đề xuất một phương pháp mới gọi là *mô hình hỗ trợ* có thể xử lý cả hai giai đoạn. Ý tưởng cơ bản là mối quan hệ hỗ trợ giữa các vụ án cũng như chiến lược so khớp giữa các đoạn văn và quyết định-đoạn văn. Ngoài ra, chúng tôi đề xuất một phương pháp để tự động tạo ra một bộ dữ liệu gắn nhãn yếu lớn để vượt qua thiếu hụt dữ liệu. Các thí nghiệm đã cho thấy rằng giải pháp của chúng tôi đã đạt được kết quả tiên tiến nhất cho cả hai giai đoạn truy hồi án lệ và kế thừa án lệ.

2.1 Mối quan hệ hỗ trợ án lệ

Mối quan hệ hỗ trợ giữa các án lệ không chỉ liên quan đến các tình huống tương tự. Những vụ án hỗ trợ có thể được đề cập và trích dẫn để hỗ trợ án lệ truy vấn. Theo quan sát của chúng tôi, một văn án lệ s là một án lệ được nhận biết của một án lệ truy vấn qc,

điều này không có nghĩa là tất cả các phần của s đều hỗ trợ cho qc. Nói cách khác, nếu chỉ có một số đoạn văn trong s hỗ trợ một số quyết định trong qc, chúng ta có thể kết luận rằng s ủng hộ qc. Do đó, chúng tôi giới thiệu một khái niệm vụ án hỗ trợ dựa trên thành phần hỗ trợ. Văn bản pháp lý dài được chia thành các thành phần giống đoạn văn và mối quan hệ hỗ trợ ở cấp độ thành phần thay vì tập trung vào mối quan hệ hỗ trợ với đơn vị án lệ như trong các nghiên cứu trước đó. Hình 2.1 minh họa một ví dụ về thành phần hỗ trợ của chúng tôi. Đây là một phần của đồ thị mối quan hệ hỗ trợ giữa một vụ án truy vấn và một vụ án ứng cử.

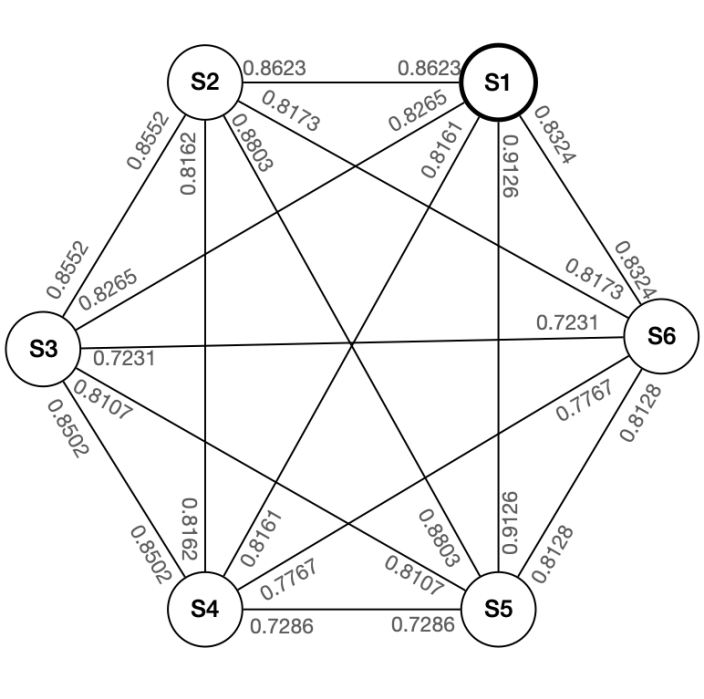


Hình 2.1: Ví dụ về mối quan hệ hỗ trợ giữa trường hợp truy vấn và trường hợp ứng cử viên

Tương tự, các đoạn văn trong án lệ thường được cấu trúc theo hình thức lập luận, trình bày các lập luận pháp lý một cách rõ ràng, chính xác và có sự nhất quán logic. Mỗi đoạn văn tập trung vào một vấn đề pháp lý cụ thể hoặc một điểm pháp lý cụ thể; sử dụng logic, bằng chứng và các trích dẫn cụ thể để làm sáng tỏ vấn đề hoặc quan điểm được trình bày. Có sự thống nhất chủ đề trong mỗi đoạn văn pháp lý, đảm bảo rằng câu chuyện được trình bày một cách rõ ràng. Hình 2.2 cho thấy một ví dụ về đồ thị mối quan hệ hỗ trợ giữa các câu trong đoạn văn pháp luật.

2.2 Mối quan hệ hỗ trợ trong việc truy hồi án lệ

Với sự tiến bộ gần đây trong quá trình số hóa và biến đổi số, luật sư hiện có thể dễ dàng truy cập một lượng lớn tài liệu pháp luật trực tuyến. Tuy nhiên, càng nhiều tài liệu



Case IMM-2683-96, paragraph 4:
S1: The applicant alleges that the Board used standard form "boiler-plate" reasons and therefore denied the applicant a fair hearing.
S2: Key passages in the Board's reasons are identical or virtually identical to two other decisions, Jafari v. Minister of Citizenship & Immigration , Board
S3: In all three cases involving Iranian refugee claimants, Mr. Jack Davis was the presiding Board member.
S4: Jafari was heard three days after the case at bar.
S5: The respondent in turn argues that the Board clearly made an independent decision.
S6: The identical passages are nothing more than digests of the law on such legal questions as credibility and the documentary evidence

Hình 2.2: Một ví dụ về đồ thị mối quan hệ hỗ trợ giữa các câu trong đoạn văn án lệ, mỗi câu trong đoạn văn được đại diện bằng một đỉnh, các cạnh là sự tương đồng ý nghĩa giữa các câu. S1 là câu đề đề cập đến chủ đề trong ví dụ này.

pháp luật, việc tìm kiếm các văn bản pháp luật có liên quan nhất để hỗ trợ việc chuẩn bị tòa của luật sư càng trở nên khó khăn hơn. Do đó, việc phát triển một hệ thống truy hồi án lệ tự động là rất quan trọng để tăng tốc quy trình làm việc của luật sư.

Cuộc thi truy hồi thông tin pháp lý (COLIEE) là một cuộc thi hàng năm dành cho các nhà nghiên cứu để giải quyết các vấn đề về truy hồi thông tin, trích xuất và lập luận trong lĩnh vực pháp luật. Một trong những thách thức chính trong cuộc thi là nhiệm vụ về án lệ. Dữ liệu cho nhiệm vụ này dựa trên các văn bản vụ án của Tòa án Liên bang Canada được cung cấp bởi vLex Canada.

Một án lệ thường là một tập hợp các kết luận pháp lý trước đó được viết bởi tòa án. Một luật sư có thể tìm các văn bản án lệ có liên quan và sử dụng các kết luận thích hợp để ủng hộ quyết định trong vụ án hiện tại. Văn bản án lệ có thể khác nhau về cấu trúc, các thành phần có thể không giống nhau trong tất cả các trường hợp, điều này đòi hỏi nỗ lực đáng kể trong quá trình xử lý. Thậm chí, việc đọc, quét và tìm kiếm các văn bản án lệ thực sự liên quan từ một cơ sở dữ liệu án lệ lớn cũng là rất khó khăn đối với các luật sư đã được đào tạo. Do đó, việc truy hồi văn bản án lệ là một nhiệm vụ phức tạp có một số thách thức như sau:

Thách thức 1: Cả truy vấn lẫn các vụ án hỗ trợ đều là các văn bản cực kỳ dài, với

trung bình khoảng 3000 từ.

Trong nhiệm vụ truy hồi, truy vấn dài là một thách thức. Cả hai phương pháp học biểu diễn và học so khớp đều có nhược điểm trong việc xử lý các văn bản dài. Việc học biểu diễn cho văn bản dài trong không gian vector có hạn là một thách thức. Xây dựng và tổng hợp các văn bản dài trong học so khớp cũng là một vấn đề khó khăn.

Thách thức 2: Định nghĩa về sự liên quan trong lĩnh vực pháp luật khá khác biệt so với định nghĩa chung về sự liên quan về chủ đề.

Trong tình huống pháp lý, các vụ án có liên quan là những vụ án có thể hỗ trợ quyết định của một vụ án mới, thường có tình huống tương tự và các quy định phù hợp. Việc xác định mối quan hệ hỗ trợ giữa các văn bản pháp luật là rất quan trọng. Mối quan hệ này vượt xa sự liên quan về chủ đề và từ vựng. Việc so khớp giữa vụ án truy vấn và các vụ án ứng cử, giữa quyết định truy vấn và các vụ án hỗ trợ trở nên khó khăn hơn nhiều so với việc truy hồi văn bản chung.

Thách thức 3: Việc tạo ra một bộ dữ liệu lớn và chính xác cho nhiệm vụ về pháp luật đòi hỏi nhiều nỗ lực và kiến thức chuyên môn trong lĩnh vực pháp luật. Sự thiếu hụt dữ liệu được gán nhãn là một rào cản đối với việc huấn luyện và đánh giá các mô hình mạng nơ-ron.

Trong nghiên cứu này, chúng tôi đề xuất một phương pháp học sâu với một mô hình hỗ trợ cho việc truy hồi văn bản án lệ gọi là SM-BERT-CR để giải quyết những thách thức trên. Chúng tôi đề xuất một khái niệm về vụ án hỗ trợ cho giai đoạn truy hồi văn bản án lệ dựa trên thành phần hỗ trợ của chúng tôi trong mối quan hệ hỗ trợ văn bản pháp luật (**Thách thức 1 và 2**). Mối quan hệ giữa các đoạn văn hỗ trợ và một quyết định đã cho trong giai đoạn kế thừa tương tự như mối quan hệ giữa các đoạn văn trong các vụ án hỗ trợ và một vụ án truy vấn trong giai đoạn truy hồi.

Mối quan hệ hỗ trợ $support(a, b)$ (a supports b), các nhiệm vụ truy hồi và kế thừa văn bản pháp luật được hình thành như sau:

Đặt $\mathbf{C}$ là không gian của tất cả các vụ án pháp lý và các luật pháp và đặt $C \subset \mathbf{C}$ là một tập hợp các án lệ. Cho một vụ án truy vấn đầu vào $c_q \in \mathbf{C}$. Truy vấn c_q thường là một vụ án pháp lý mới mà một thẩm phán hoặc luật sư đang làm việc. Mục tiêu của giai đoạn này là tìm kiếm và truy hồi một tập hợp tất cả các án lệ có liên quan $C^r = c_1^r, c_2^r, \ldots, c_k^r \subset C$ hỗ trợ cho quyết định của c_q. Chúng tôi giả định rằng một án lệ ứng cử C_i^r hỗ trợ cho vụ án truy vấn c_q nếu và chỉ nếu có một hoặc nhiều đoạn văn trong s hỗ trợ một quyết định trong c_q:

$$support(c_i^r, c_q) \iff \exists p_j \in c_i^r \wedge \exists p_k \in c_q : support(p_j, p_k)$$

Giai đoạn kế thừa án lệ: Cho một bộ ba bao gồm vụ án truy vấn đầu vào c_q, một quyết định d_q của vụ án truy vấn c_q, và danh sách tất cả các vụ án hỗ trợ C^r được trả về từ giai đoạn trước. Đặt P^r là tập hợp tất cả các đoạn văn được phân đoạn từ một vụ án hỗ trợ cụ thể $c^r \in C^r$. Nhiệm vụ là xác định một tập hợp các đoạn văn đưa ra quyết định $P^e = p_i^e \mid p_i^e \in P^r \wedge support(p_i^e, d_q)$.

Các nghiên cứu trước đó thường giải quyết việc tìm mối quan hệ hỗ trợ giữa vụ án truy vấn/quyết định và vụ án ứng viên/đoạn ẩn gián tiếp thông qua các phương pháp đo độ tương đồng. Không giống như các nghiên cứu trước đó, chúng tôi xây dựng một mô hình hỗ trợ để dự đoán trực tiếp mối quan hệ hỗ trợ (**Thách thức 2**). Lấy cảm hứng từ sự thành công của mô hình ngôn ngữ tiền huấn luyện BERT trên nhiều nhiệm vụ xử lý ngôn ngữ tự nhiên, chúng tôi điều chỉnh mô hình BERT để xây dựng mô hình hỗ trợ cho các nhiệm vụ văn bản pháp luật.

Ngoài mô hình hỗ trợ, chúng tôi cũng khai thác nhiều độ đo tương đồng như sự tương đồng từ ngữ (so sánh từ khóa) và sự tương đồng ngữ cảnh (so sánh bối cảnh). Mặc dù sự tương đồng từ ngữ và sự tương đồng ngữ cảnh khá khác biệt, chúng có thể được kết hợp và bổ sung cho nhau. Sự tương đồng từ ngữ có thể được đạt được bằng cách so sánh từng từ với một số biến thể như stemming, loại bỏ từ dừng, lemmatization, v.v. Điểm số cao trong sự tương đồng từ ngữ có thể chỉ ra sự phù hợp cao giữa hai tài liệu, nhưng với sự tương đồng từ ngữ thấp, điều đó không có nghĩa là những tài liệu này không có bất kỳ mối quan hệ nào. Do đó, chúng tôi kết hợp mô hình hỗ trợ với mô hình từ ngữ trong hệ thống truy hồi văn bản án lệ của chúng tôi.

Để giải quyết thách thức của việc thiếu dữ liệu được gán nhãn, chúng tôi sử dụng một số phương pháp heuristics để tự động xây dựng tập dữ liệu huấn luyện về mối quan hệ hỗ trợ trong vụ án pháp luật được gọi là một tập dữ liệu hỗ trợ nhãn yếu (**Thách thức 3**). Tập dữ liệu này được xây dựng dựa trên đồ thị mối quan hệ hỗ trợ trong đoạn văn pháp luật mà một đoạn văn chứa một câu quyết định và các câu còn lại hỗ trợ câu quyết định này. Hơn nữa, chúng tôi giả định rằng câu quyết định là câu chủ đề trong đoạn văn ứng cử. Để xác định câu quyết định trong đoạn văn ứng cử, chúng tôi áp dụng thuật toán TextRank - một mô hình xếp hạng dựa trên đồ thị cho việc trích xuất câu tự động. Việc giới thiệu tập dữ liệu này có thể giảm sự phụ thuộc của các mô hình mạng thần kinh vào dữ liệu được gán nhãn.

Chương 3

Đồ thị tri thức cho truy hồi luật – vụ án

Trong chương này, chúng tôi phát triển một biểu đồ kiến thức bao gồm các tài liệu về vụ án pháp luật và các văn bản pháp luật liên quan để cải thiện tổ chức và truy hồi thông tin pháp lý. Phương pháp của chúng tôi bao gồm việc thu thập dữ liệu, trích xuất thực thể và xây dựng đồ thị. Biểu đồ không đồng nhất được xây dựng kết nối các tòa án, các vụ án, các lĩnh vực và luật pháp, làm phong phú thông tin được cung cấp bởi các hệ thống truy hồi. Phương pháp của chúng tôi thể hiện tiềm năng trong phân tích vụ án, đề xuất pháp lý và hỗ trợ quyết định, cung cấp những hiểu biết và tài nguyên quý giá cho lĩnh vực pháp lý.

3.1 Đồ thị tri thức pháp lý

Một đồ thị tri thức pháp lý biểu diễn thông tin pháp lý được tổ chức theo dạng đồ thị không đồng nhất, ghi lại các mối quan hệ giữa các thực thể pháp lý như luật, quy định, vụ án và khái niệm. Sự biểu diễn dựa trên đồ thị này cho phép hiểu rõ hơn về các lĩnh vực pháp lý bằng cách tổ chức và kết nối các điểm dữ liệu pháp lý phân tán. Bằng cách mô hình hóa kiến thức pháp lý như các nút và cạnh kết nối, các biểu đồ kiến thức pháp lý hỗ trợ nhiều nhiệm vụ khác nhau, bao gồm nghiên cứu pháp lý, truy hồi thông tin và các hệ thống hỗ trợ quyết định.

Trong việc trình bày về pháp luật và vụ án, các đồ thị tri thức cung cấp các khung được tổ chức cho việc phân tích thông tin pháp lý được xây dựng từ các vụ án và các văn bản pháp luật, một cách cụ thể. Sử dụng lý thuyết đồ thị, các biểu đồ kiến thức này biểu diễn các thực thể pháp lý, chẳng hạn như các vụ án, luật hành văn, quy định, các lĩnh vực pháp luật, và các mối quan hệ, dưới dạng các nút và cạnh.

Phương pháp xây dựng một đồ thị tri thức đóng vai trò là một công cụ thích hợp để xác định và biểu diễn các mối quan hệ giữa các vụ án và các luật liên quan. Các đồ thị tri thức có thể mô tả hiệu quả lượng lớn kiến thức với ý nghĩa ngữ nghĩa, giúp truy cập và truy vấn cấu trúc dễ dàng. Những đồ thị tri thức này được thiết kế một cách thân thiện với người dùng, phục vụ cho những người không chuyên như luật sư, thẩm phán, học giả, v.v., giúp họ dễ dàng sử dụng và khám phá thông tin. Hơn nữa, các đồ thị tri thức có thể được áp dụng để cải thiện các nhiệm vụ phụ thuộc xuôi dòng khác nhau trong lĩnh vực pháp lý như truy hồi thông tin, trả lời câu hỏi, phân loại, và nhiều hơn nữa.

3.2 Định nghĩa đồ thị tri thức án lệ

Một vụ án được lưu trữ trên trang web của Tòa án nhân dân Tối cao Việt Nam bao gồm hai phần: dữ liệu siêu dữ liệu và tài liệu vụ án. Hình 3.1 minh họa cấu trúc và nội dung của một vụ án. Dữ liệu siêu dữ liệu chứa thông tin cơ bản về vụ án, bao gồm số vụ án, tên vụ án, loại vụ án, v.v. Phần thân của tài liệu vụ án bao gồm bốn phần: giới thiệu, nội dung của vụ án, nhận định của tòa án, và quyết định của tòa án. Mô tả của mỗi phần được hiển thị trong Bảng 3.1.

Bảng 3.1: The description of a law document.

Part	Description
Giới thiệu	thông tin chi tiết về vụ việc, tòa án, bị đơn, nguyên đơn, các bên liên quan (ví dụ: họ tên, ngày sinh, địa chỉ của các bên)
Nội dung của vụ án	ý kiến của vụ án, tòa án, bị đơn, nguyên đơn, các bên liên quan
Nhận định của tòa án	Ý kiến, phân tích của tòa án
Quyết định của tòa án	Quyết định của tòa án dựa trên các phần trên

Chúng tôi xây dựng đồ thị tri thức về vụ án pháp lý Việt Nam dựa trên một đồ thị không đồng nhất, có thể có các nút và cạnh thuộc các loại khác nhau. Một đồ thị không đồng nhất $G = (V, E)$ chứa một tập hợp thực thể V và một tập hợp quan hệ E với một hàm ánh xạ loại thực thể $f : V \rightarrow A$ và một hàm ánh xạ loại quan hệ $g : E \rightarrow R$. A và R chỉ ra các tập hợp loại thực thể và loại quan hệ, trong đó $|A| + |R| > 2$.

Bản án số: 577/2022/HC-PT ngày 28/07/2022 **a**
Tên bản án: Phạm Đăng M kiện UBND TP PR-TC
Đối tượng khởi kiện: QĐ hành chính, hành vi hành chính về quản lý đất đai […]
Cấp xét xử: Phúc thẩm
Loại án: Hành chính
Tòa án xét xử: TAND cấp cao tại TP Hồ Chí Minh
Áp dụng án lệ: Không
Đính chính: 0
Thông tin về vụ án: Không chấp nhận yêu cầu kháng cáo của người khởi kiện ông Phạm Đăng M [...]

1. Mở đầu: **b**
- Thành phần Hội đồng xét xử phúc thẩm gồm có [...]
- Thư ký phiên tòa [...]
[...]
2. Nội dung vụ án:
Theo đơn khởi kiện, biên bản đối thoại và tại phiên tòa người đại diện theo ủy quyền của người khởi kiện ông Lê Văn H trình bày: [...]
3. Nhận định của tòa án:
Sau khi nghiên cứu các tài liệu có trong hồ sơ vụ án đã được thẩm tra tại phiên tòa và căn cứ vào kết quả tranh tụng, ý kiến của đại diện Viện kiểm sát, các quy định pháp luật, Hội đồng xét xử nhận định: [...]
4. Quyết định:
Căn cứ khoản 1 Điều 241 của Luật tố tụng Hành chính năm 2015 [...]

Hình 3.1: The structure of a case law (a: meta-data, b: case content)

Đặc biệt, chúng tôi xác định 4 loại thực thể dựa trên đặc điểm của vụ án pháp lý Việt Nam. Thực thể Vụ án, chứa thông tin về mỗi quyết định/tranh tụng hiện đang có hiệu lực. Thực thể Lĩnh vực, chứa thông tin về tội phạm, loại tranh chấp và quyết định. Thực thể Tòa án, chứa thông tin về tên của mỗi tòa án và cấp bậc trong hệ thống tư pháp. Thực thể Luật, chứa tên của các luật cụ thể hoặc mã luật.

Tổng cộng có 3 loại quan hệ giữa các thực thể. Quan hệ *Xét xử* giữa các tòa án và các vụ án, chỉ ra mối quan hệ của một tòa án cụ thể xem xét vụ án. Quan hệ *Thuộc* về giữa các vụ án và các lĩnh vực, chỉ ra mối quan hệ của một lĩnh vực cụ thể và các lĩnh vực phụ thuộc mà vụ án thuộc về. Quan hệ *Dựa trên* giữa các vụ án và các luật, chỉ ra mối quan hệ của một quyết định hoặc phán quyết cụ thể đã tham khảo một tập hợp các luật/mã luật để hỗ trợ quyết định của mình.

Trong một đồ thị không đồng nhất, hai thực thể có thể được kết nối thông qua các

đường dẫn khác nhau. Một cách chính thức, các đường dẫn này được gọi là meta-path. Một meta-path P được định nghĩa dưới dạng $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_k} A_{k+1}$, biểu diễn một mối quan hệ tổng hợp $R = R_1 \circ R_2 \circ \cdots \circ R_k$ giữa A_1 và A_{k+1}, trong đó $\circ$ biểu thị toán tử hợp thành trên các mối quan hệ. Hai vụ án có thể được kết nối thông qua các meta-path khác nhau, ví dụ: Case-Court-Case (CCC) hoặc Case-Domain-Case (CDC). Các meta-path khác nhau mô tả các mối quan hệ ngữ nghĩa trong các góc độ khác nhau. Ví dụ, đường dẫn CCC có nghĩa là các vụ án này được tòa án cùng một bên xét xử, trong khi đường dẫn CDC chỉ ra rằng chúng thuộc cùng một lĩnh vực.

3.3 Statutory – Case Law Retrieval Model

Cùng với sự phát triển của công nghệ, khối lượng tài liệu kỹ thuật số đã tăng đáng kể, đặc biệt là trong lĩnh vực pháp lý. Sự tiến bộ này đã làm cho việc tìm kiếm và truy cập thông tin pháp lý hiệu quả hơn. Tài liệu pháp lý thường dài và có cấu trúc, được trình bày theo một phong cách viết cụ thể. Việc tận dụng hiệu quả dữ liệu này chủ yếu phụ thuộc vào cách nó được tổ chức và chuẩn hóa. Trong lĩnh vực pháp lý, đặc biệt là trong các tài liệu về quy phạm tư pháp, người có thể tìm thấy thông tin về các vụ án, quyết định của tòa án và luật pháp liên quan đến những vụ án đó. Mặc dù thông tin có sẵn, việc truy hồi thông tin pháp lý có thể phức tạp, đặc biệt là khi xử lý các vụ án cụ thể hoặc điều tra một vụ án cụ thể như một chuyên gia pháp lý. Thông tin mong muốn có thể cần phải được tìm kiếm từ nhiều nguồn và tiếp cận theo nhiều cách khác nhau.

Về việc truy hồi thông tin, chúng tôi đã xác định các thực thể và mối liên kết trong các văn bản pháp lý không cấu trúc để tạo ra một biểu đồ đa dạng. Ngoài việc giúp cải thiện hiệu suất của nhiệm vụ truy hồi luật – vụ án, phương pháp này cũng hỗ trợ một loạt các ứng dụng khác trong lĩnh vực pháp luật, bao gồm phân tích vụ án, hướng dẫn pháp lý và hỗ trợ ra quyết định. Mô hình cơ sở, sử dụng các kỹ thuật học không giám sát và biểu đồ kiến thức, đã cho thấy kết quả tích cực trong việc nhận diện các luật pháp liên quan cho một vụ án cụ thể. Công việc nghiên cứu tương lai có thể tập trung vào tinh chỉnh việc trích hồi thông tin, kết hợp các phương pháp học dựa trên biểu đồ tiên tiến, và mở rộng phạm vi của biểu đồ kiến thức để tăng hiệu suất và khả năng sử dụng rộng rãi hơn.

Chương 4

Mạng tham chiếu điều luật cho hỏi đáp dựa trên IR

Sự phức tạp ngày càng tăng của luật pháp đã dẫn đến sự tăng cường nhu cầu về các phương pháp truy hồi hiệu quả. Chương này trình bày một phương pháp mới cho việc truy hồi pháp luật sử dụng các mạng tham chiếu để khám phá các kết nối giữa các luật. Bằng cách trình bày các điều luật như một mạng lưới các tham chiếu, phương pháp của chúng tôi cho phép người dùng nhanh chóng xác định các điều luật liên quan và điều hướng trong mạng phức tạp của các tài liệu pháp lý. Điểm chính là mạng tham chiếu có thể mã hóa cả các mối quan hệ pháp lý nội bộ và bên ngoài, giúp tích hợp cả tính liên quan cục bộ và sự phụ thuộc xa vào mô hình truy hồi cuối cùng. Chúng tôi đánh giá hiệu suất của phương pháp của mình bằng cách sử dụng một tập lớn các tài liệu luật pháp và chứng minh rằng nó vượt trội hơn so với các phương pháp truy hồi hiện có. Phương pháp của chúng tôi có thể góp phần vào việc phát triển các công cụ nghiên cứu pháp lý hỗ trợ trí tuệ nhân tạo, giúp cho các chuyên gia pháp lý dễ dàng tìm ra các luật và tiền lệ liên quan. Hơn nữa, bằng cách khám phá các kết nối ẩn giữa các điều luật, phương pháp của chúng tôi có thể giúp trong việc xác định những không nhất quán trong hệ thống pháp luật, từ đó cải thiện hiệu quả và đáng tin cậy của nó.

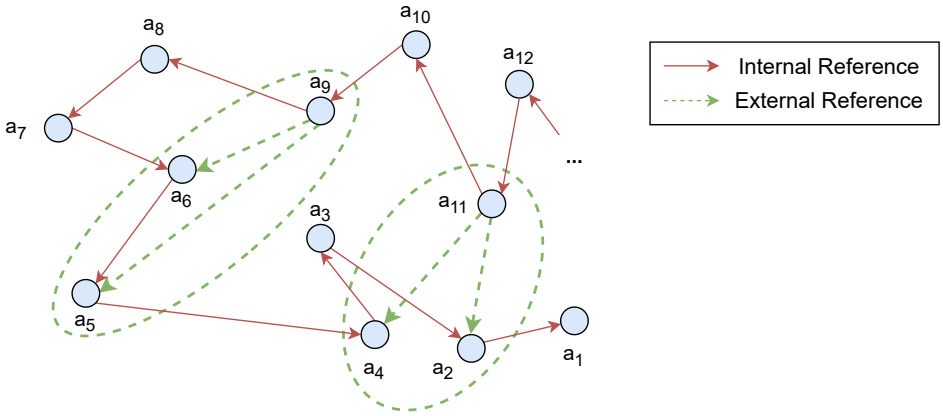
Ngoài ra, chương này tổng hợp các mô hình đại diện cho các mối quan hệ trong lĩnh vực pháp lý để giải quyết vấn đề trả lời câu hỏi về pháp luật tiếng Việt.

4.1 The Article Reference Relation Network

Các tài liệu pháp lý được đặc trưng bởi độ dài đáng kể và một cấu trúc tổ chức nghiêm ngặt, thường được chia thành các cấp độ phân cấp như phần, chương, mục, điều, và khoản với mức độ điều lệ là cấp độ phổ biến và được sử dụng rộng rãi nhất.

Internal reference: Trong ngữ cảnh của các tài liệu pháp lý, các điều liên tiếp trong một chương thường có một mối quan hệ gần gũi về nội dung hoặc thông qua các tham chiếu trực tiếp bằng các thuật ngữ liên hệ, chúng tôi gọi đó là *internal reference*.

External reference: Ngoài ra, thường xuyên các điều luật có thể tham chiếu đến các điều luật trước đó trong cùng một hoặc khác nhau tài liệu pháp lý, chúng tôi đặt tên cho nó là *external reference*.



Hình 4.1: Minh họa mối quan hệ tham chiếu giữa các điều luật

Trong thực tế, tài liệu pháp lý bao gồm một lượng lớn các mối quan hệ tham chiếu, và việc bỏ qua những mối quan hệ này sẽ dẫn đến mất mát thông tin đáng kể. Trong nghiên cứu này, chúng tôi đề xuất xây dựng một biểu đồ tri thức có thể ghi lại các mối quan hệ tham chiếu trong các tài liệu pháp lý. Biểu đồ mối quan hệ tham chiếu pháp lý được xây dựng dựa trên một biểu đồ không đồng nhất $\mathcal{G} = \mathcal{V}, \mathcal{E}$ như được minh họa trong Hình 4.1. Các nút trong biểu đồ là các điều pháp lý $\mathcal{V} = a_1, a_2, ..., a_N$. Tổng số loại mối quan hệ của thực thể là hai, bao gồm các cạnh *internal reference* từ a_i là $\mathcal{E}_{a_i}^{IN} = \left\{ in_{a_{i-1}}^{a_i} | a_i, a_{i-1} \in \mathcal{V} : InSameChapter(a_i, a_{i-1}) = 1 \right\}$ và các cạnh *external reference* từ a_i là $\mathcal{E}_{a_i}^{EX} = ex_1^{a_i}, ex^{a_i}2, \ldots, ex^{a_i}n_i$.

4.2 Mạng tham chiếu cho hỏi đáp dựa trên IR

Trong một thế giới ngày càng phức tạp và chuyên sâu, các chuyên gia pháp lý cần phải điều hướng qua một loạt các luật lệ pháp luật ngày càng phát triển và tăng lên về quy mô. Việc xác định các điều luật liên quan từ một nguồn tài liệu lớn không chỉ là công việc đòi hỏi nhiều công sức mà còn là rất quan trọng cho lập luận pháp lý, soạn thảo pháp luật, kiện tụng và nghiên cứu pháp lý. Sự chuyển đổi hướng tới tài liệu pháp lý số hóa đã thúc đẩy nhu cầu về các hệ thống truy hồi pháp luật hiệu quả có thể hỗ trợ các chuyên gia pháp lý trong công việc này. Các phương pháp nghiên cứu pháp lý truyền thống, chủ yếu dựa vào tìm kiếm thủ công hoặc tìm kiếm dựa trên từ khóa đơn giản, thường không đủ để đối phó với tính phức tạp của văn bản pháp lý. Cấu trúc của các tài liệu pháp lý được đặc trưng bởi một mạng lưới các tham chiếu, trong đó các điều luật trích dẫn các điều luật khác, tạo ra một mạng lưới các luật phụ thuộc lẫn nhau. Hiểu và điều hướng qua những phụ thuộc này là rất quan trọng cho việc phân tích pháp luật toàn diện. Tuy nhiên, do số lượng lớn và ngôn ngữ phức tạp của văn bản pháp lý, việc theo dõi các tham chiếu này thủ công có thể là một công việc dễ mắc lỗi. Sự xuất hiện của công nghệ truy hồi thông tin (IR), kết hợp với những tiến bộ gần đây trong xử lý ngôn ngữ tự nhiên (NLP) và học máy, đã thúc đẩy sự phát triển của các hệ thống IR có thể xử lý các lượng văn bản lớn để tìm thông tin liên quan. Tuy nhiên, các thách thức cụ thể mà văn bản pháp lý đặt ra, như ngôn ngữ cụ thể cho lĩnh vực, sự cần thiết của độ chính xác cao và tầm quan trọng của ngữ cảnh và các tham chiếu giữa các tài liệu, đòi hỏi các giải pháp được tinh chỉnh và phức tạp hơn.

Trong chương này, chúng tôi giới thiệu một phương pháp truy hồi luật mới sử dụng khái niệm của mạng lưới tham chiếu để tăng cường quá trình truy hồi. Phương pháp của chúng tôi tận dụng quan sát rằng các điều luật pháp luật không phải là các thực thể độc lập; thay vào đó, chúng hoạt động trong một mạng lưới tham chiếu, với các luật thường trích dẫn các luật khác. Bằng cách xem xét các luật như các nút trong một mạng lưới tham chiếu, chúng tôi có thể khám phá các kết nối trực tiếp và gián tiếp giữa các điều luật, từ đó tạo điều kiện cho việc xác định các luật liên quan đến câu hỏi đầu vào một cách hiệu quả hơn. Chúng tôi đề xuất một kiến trúc kết hợp thông tin từ các điều luật được trích dẫn để làm giàu biểu diễn của một điều luật cụ thể, từ đó bắt lấy cả nội dung và ngữ cảnh của các tham chiếu. Phương pháp này đại diện cho một bước đi quan trọng so với các kỹ thuật truy hồi tài liệu truyền thống thường chỉ dựa vào độ tương tự về nội dung mà thôi. Bằng cách xem xét mạng lưới tham chiếu, hệ thống của chúng tôi được trang bị tốt hơn để hiểu ngữ cảnh và sự liên quan pháp lý của tài liệu, từ đó giúp tạo ra

các kết quả truy hồi chính xác hơn.

Các nghiên cứu gần đây về việc truy hồi pháp luật đã sử dụng nhiều kỹ thuật mạng nơ-ron khác nhau, như CNNs, LSTMs, cơ chế chú ý và mạng nơ-ron đồ thị, để đạt được các kết quả đáng chú ý trong lĩnh vực pháp lý. Những phương pháp nghiên cứu trước đó chủ yếu tập trung vào độ tương tự dựa trên nội dung và có thể không hoàn toàn nắm bắt được mạng lưới phức tạp của các tham chiếu trong tài liệu pháp lý. Ngược lại, phương pháp của chúng tôi nhằm giải quyết vấn đề này bằng cách tận dụng sức mạnh của các mạng tham chiếu, đặc biệt là khai thác tối đa cả tính phù hợp nội bộ (tức là, cục bộ) và sự phụ thuộc ngoại lai (tức là, xa) để tăng cường mô hình truy hồi cuối cùng. Thông qua các đánh giá toàn diện của một tập dữ liệu lớn các tài liệu về pháp luật, chúng tôi chứng minh rằng phương pháp của chúng tôi vượt qua các phương pháp truy hồi hiện có sẵn về mặt tính phù hợp và hiệu quả. Ngoài ra, chúng tôi thảo luận về những đóng góp tiềm năng của mô hình của chúng tôi cho việc phát triển các công cụ nghiên cứu pháp lý hỗ trợ trí tuệ nhân tạo, giúp tối ưu hóa quy trình khám phá pháp lý.

Chúng tôi mô tả phương pháp được sử dụng để tận dụng các mối tương quan giữa các điều luật để xây dựng một biểu diễn dữ liệu nhằm tăng cường kết quả của nhiệm vụ truy hồi pháp luật. Đầu tiên, chúng tôi trình bày một cái nhìn tổng quan toàn diện về vấn đề truy hồi điều luật. Tiếp theo, chúng tôi giới thiệu các ký hiệu khác nhau, cấu trúc biểu đồ tri thức và phương pháp được sử dụng trong quá trình xây dựng chúng. Cuối cùng, chúng tôi làm rõ kiến trúc và quá trình huấn luyện của một mô hình tích hợp biểu diễn đồ thị các mối quan hệ pháp lý với các mô hình ngôn ngữ được huấn luyện trước.

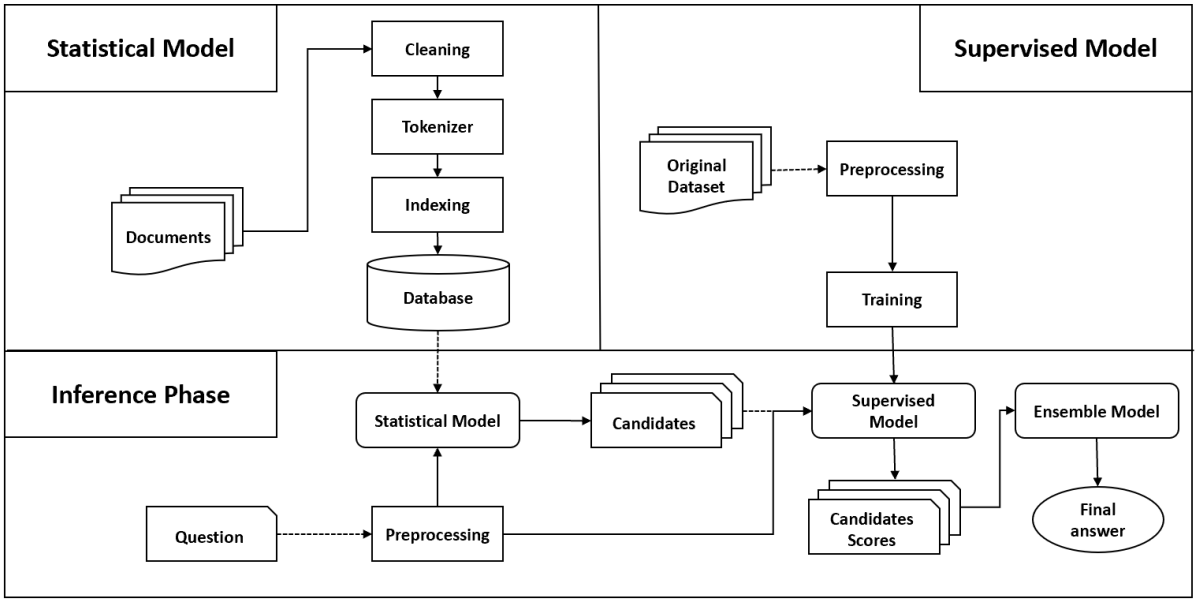
Truy hồi điều luật là một trong những nhiệm vụ truyền thống và phổ biến nhất trong lĩnh vực xử lý văn bản pháp lý. Giả sử A là một tập hợp các điều luật (tức là, một cơ sở dữ liệu) về luật pháp. Cho một câu hỏi q về bất kỳ vấn đề pháp lý nào có thể được bao gồm trong tập hợp A, hệ thống nhằm truy hồi một tập con $A^r \subset A$ sao cho mỗi điều luật $a_i^r \in A^r$ có liên quan ngữ nghĩa hoặc hỗ trợ cho một câu hỏi cụ thể q (câu hỏi hoặc tuyên bố pháp lý). Vấn đề có thể được mô tả như sau:

$$\textit{Relevance}(q, a_i^r) = \begin{cases} 1 & \text{if } a_i^r \text{ is semantically related to } q \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

$$A^r = \{a_i^r \in A : \textit{Relevance}(q, a_i^r) = 1\} \tag{4.2}$$

4.3 Hệ hỏi đáp tiếng Việt

Kiến trúc hệ trả lời câu hỏi dựa trên truy hồi điều luật mà chúng tôi đề xuất được minh họa trong Hình 4.2. Bao gồm ba giai đoạn chính: giai đoạn tiền xử lý, huấn luyện, và suy luận, các giai đoạn này hoạt động cùng nhau để cung cấp các câu trả lời chính xác và hiệu quả cho các câu hỏi của người dùng.



Hình 4.2: Quy trình của hệ thống QA dựa trên truy hồi bài viết đầu cuối

Giai đoạn tiền xử lý Một cơ sở dữ liệu bao gồm các điều luật được tạo ra bằng cách xử lý các tài liệu pháp luật dân sự Việt Nam. Cơ sở dữ liệu cấp điều luật kết quả cho phép truy cập và truy hồi dễ dàng các thông tin cụ thể được chứa trong các tài liệu.

Gia đoạn huấn luyện Một mô hình máy học giám sát được phát triển để xếp hạng các điều luật liên quan đến câu hỏi đầu vào. Mô hình này sử dụng dữ liệu huấn luyện để học các mẫu và mối quan hệ trong các bài viết và áp dụng kiến thức này để cung cấp các xếp hạng chính xác cho các điều luật liên quan.

Pha suy luận Giai đoạn suy luận đề cập đến quá trình tạo ra một phản hồi cho một câu hỏi đầu vào mới. Giai đoạn này thường bao gồm việc áp dụng một mô hình máy học đã được huấn luyện vào câu hỏi đầu vào và chọn lựa phản hồi phù hợp nhất từ một tập hợp các câu trả lời tiềm năng.

Các phương pháp được đề xuất nhằm cải thiện hiệu suất cho nhiệm vụ trả lời câu hỏi pháp lý bằng tiếng Việt bằng cách sử dụng các mô hình ngôn ngữ thông qua việc gán nhãn yếu và mạng tham chiếu. Bằng cách thực nghiệm phương pháp này, chúng tôi đã

xác minh giả thuyết rằng cải thiện chất lượng và số lượng dữ liệu là phương pháp đúng đắn cho vấn đề này, đặc biệt là trong các ngôn ngữ nguồn lực thấp như tiếng Việt. Kết quả của công việc của chúng tôi có thể cung cấp những cái nhìn quý giá và phục vụ như một tài liệu tham khảo cho các nỗ lực trong tương lai để giải quyết các thách thức tương tự trong việc trả lời câu hỏi pháp lý cho ngôn ngữ nguồn lực thấp.

Kết luận

Tóm tắt kết quả và đóng góp

Luận án đã tiến hành một nghiên cứu có hệ thống và kỹ lưỡng về các nhiệm vụ truy hồi và trả lời câu hỏi pháp luật, đó là hai trong những vấn đề quan trọng và khó khăn nhất trong lĩnh vực NLP pháp lý. Theo các thách thức nghiên cứu, động lực và mục tiêu được đề cập trong Chương 1, luận án đã trình bày phát biểu vấn đề, công thức hóa, và đề xuất việc sử dụng các loại đặc điểm pháp lý khác nhau cũng như giới thiệu một số kiến trúc mô hình sâu để tích hợp những đặc điểm đó nhằm tăng cường hiệu suất của ba nhiệm vụ IR và QA. Tóm lại, luận án có những kết quả và đóng góp quan trọng sau đây:

• Để tận dụng và tận dụng tối đa bản chất và đặc điểm của dữ liệu pháp lý và tăng hiệu suất của ba nhiệm vụ IR và QA chính được đề cập trong luận án này (mục tiêu nghiên cứu - O2), chúng tôi đã giới thiệu mô hình hỗ trợ (trong Chương 2) giúp tích hợp các mối quan hệ hỗ trợ ở các cấp độ khác nhau (trường hợp-trường hợp, đoạn-đoạn, và quyết định-đoạn) cho vấn đề truy hồi án lệ. Ngoài các đặc điểm văn bản pháp lý, các đặc điểm cấu trúc hoặc dựa trên đồ thị cũng rất hữu ích cho các nhiệm vụ IR và QA pháp lý. Do đó, chúng tôi đã định nghĩa và xây dựng một biểu đồ tri thức không đồng nhất bao gồm các văn bản vụ án pháp lý và luật ahnhf văn liên quan để cải thiện việc tổ chức thông tin pháp lý và nhiệm vụ truy hồi luật – vụ án (trong Chương 3). Biểu đồ tri thức liên kết các vụ án, tòa án, lĩnh vực và luật pháp để làm phong phú các đặc điểm dựa trên đồ thị và do đó giúp cải thiện hiệu suất truy hồi một cách đáng kể. Trong Chương 4, chúng tôi đề xuất việc sử dụng một mạng tham chiếu để cải thiện hiệu suất của vấn đề trả lời câu hỏi pháp luật. Mạng tham chiếu nắm bắt cả các trích dẫn cục bộ và xa giữa các điều luật nhằm khám phá các liên kết tiềm năng giúp xác định và truy hồi các điều luật thực sự liên quan mà không thể tìm thấy bằng cách kết hợp truy hồi từ điển truyền thống, bằng cách sử dụng từ đồng nghĩa, hoặc thậm chí là bằng các phương pháp nhúng văn bản.

• Ngoài việc khám phá và sử dụng các đặc điểm pháp lý, luận án này đã cố gắng

giới thiệu các kiến trúc hợp lý dựa trên học sâu cho các nhiệm vụ IR và QA (các mục tiêu nghiên cứu O1 và O3). Các kiến trúc mô hình này giúp (i) tận dụng các nguồn lực, phương pháp và mô hình hiện có (ví dụ: các mô hình ngôn ngữ tiền huấn mạnh mẽ) và (ii) học các biểu diễn và tích hợp tốt hơn của các đặc điểm văn bản và cấu trúc pháp lý. Những cải tiến này dẫn đến việc tăng cường hiệu suất và hiệu quả của các nhiệm vụ. Kỹ thuật, Chương 2 đã đề xuất kiến trúc SM-BERT-CR, một mô hình hỗ trợ cho nhiệm vụ truy hồi án lệ. Trong Chương 3, việc xây dựng biểu đồ không đồng nhất liên quan đến việc thu thập dữ liệu, trích xuất thực thể và xây dựng đồ thị bằng các kỹ thuật NLP. Phương pháp của chúng tôi đã thể hiện tiềm năng của nó trong nhiệm vụ truy hồi luật – vụ án và các nhiệm vụ phụ thuộc khác như phân tích vụ án, đề xuất pháp lý và hỗ trợ ra quyết định, cung cấp thông tin và tài nguyên quý báu cho lĩnh vực pháp lý. Chương 4 đề xuất mô hình mạng tham chiếu để giải quyết nhiệm vụ trả lời câu hỏi văn bản pháp lý. Các liên kết tham chiếu cục bộ và toàn cầu trong mạng được nhúng bằng các mô hình tiền huấn mạnh mẽ và sau đó được tích hợp vào mô hình trả lời câu hỏi cuối cùng để cải thiện hiệu suất.

• Hơn nữa, các kết quả thực nghiệm trong luận án này cạnh tranh với các kết quả tốt nhất hiện nay. Trong Chương 2, mô hình hỗ trợ, tức SM-BERT-CR, đạt được điểm F_1 là 0.6060 và 0.6528 cho giai đoạn truy hồi án lệ trên tập dữ liệu COLIEE 2019 và 2020. Những kết quả này chỉ thấp hơn kết quả tốt nhất hiện nay 2 điểm phần trăm mặc dù chúng tôi không sử dụng dữ liệu huấn luyện cho giai đoạn này. Trong giai đoạn thứ hai (kế thừa án lệ), mô hình này đã đạt được kết quả rất cao với điểm F_1 là 0.7253 và 0.6753 trên tập dữ liệu COLIEE 2019 và 2020. Những kết quả này cao hơn đáng kể (khoảng 6 điểm phần trăm) so với đội đứng thứ hai. Trong nhiệm vụ truy hồi luật –vụ án (Chương 3), phương pháp của chúng tôi dựa trên biểu đồ kiến thức đạt được điểm F_1 là 0.503, cao hơn nhiều so với mô hình cơ sở ($F_1 = 0.288$) không sử dụng biểu đồ kiến thức. Trong Chương 4, phương pháp dựa trên mạng tham chiếu đã đem lại kết quả đáng kể trên cả hai tập dữ liệu COLIEE 2019 và 2020 với điểm F_2 là 0.7648 và 0.8266. Ngoài ra, chúng tôi cũng xây dựng một hệ thống từ đầu đến cuối cho nhiệm vụ trả lời câu hỏi văn bản pháp lý tiếng Việt và đạt được kết quả cao nhất trên một số tập dữ liệu tiếng Việt.

• Bên cạnh những đóng góp kỹ thuật, việc phân tích và thảo luận suốt cuộc luận án này sẽ giúp cung cấp một hiểu biết tốt hơn về văn bản pháp lý và các vấn đề xử lý, trình bày những tiến bộ và thách thức còn lại của NLP pháp lý nói chung và IR và QA pháp lý nói riêng. Nghiên cứu này cũng sẽ đề xuất các hướng nghiên cứu tương lai về IR và QA pháp lý, đặc biệt kích thích các nghiên cứu tiếp theo trong NLP pháp lý cho các ngôn ngữ ít tài nguyên như tiếng Việt.

Hạn chế của luận án

Mặc dù luận án đã cố gắng tận dụng các đặc điểm pháp lý khác nhau và giới thiệu các kiến trúc dựa trên học sâu khác nhau để tăng hiệu suất và hiệu suất của ba nhiệm vụ IR và QA, nhưng vẫn còn một số hạn chế và vấn đề còn lại có thể làm tốt hơn.

Trong giai đoạn đầu tiên của nhiệm vụ truy hồi vụ án pháp lý, mô hình SM-BERT-CR đã xác định và truy hồi án lệ hỗ trợ cho một truy vấn cụ thể từ toàn bộ bộ sưu tập văn bản pháp lý dựa trên cả gần gũi văn bản và mối quan hệ pháp lý. Tuy nhiên, trong lĩnh vực pháp lý, một án lệ hỗ trợ thực sự được gọi là "án lệ được nhận biết"được giả định là có liên quan đến án lệ truy vấn bởi luật sư. Điều này thường gây ra sự không nhất quán trong dữ liệu. Kết quả là, giai đoạn đầu tiên của nhiệm vụ thường truy hồi ra nhiều vụ án liên quan hơn cần thiết. Điều này vẫn là một vấn đề có thể được cải thiện hơn trong các nghiên cứu tiếp theo. Thứ hai, thông tin trong biểu đồ kiến thức pháp lý được xây dựng trong Chương 3 vẫn chưa được khai thác hết. Điều này phần là do biểu đồ kiến thức đã được xác định và xây dựng để bao gồm nhiều nhiệm vụ khác trong NLP pháp lý. Cuối cùng, các mô hình được đề xuất trong luận án này vẫn cần có hệ thống máy tính có sức mạnh cao để huấn luyện và suy luận do cả phức tạp của các mô hình cũng như việc sử dụng các mô hình ngôn ngữ được huấn luyện trước. Điều này vẫn cần được cải thiện cho các ứng dụng thực tế.

Hướng nghiên cứu tiếp theo

Các nghiên cứu trong tương lai sẽ cải thiện phương pháp được đề xuất theo nhiều hướng khác nhau. Đầu tiên, tiếp tục cải thiện các phương pháp để giải quyết các vấn đề liên quan đến độ dài và độ phức tạp của các tài liệu pháp lý. Thứ hai, hiệu quả của việc tích hợp các mối quan hệ pháp lý vào các nhiệm vụ IR và QA của tài liệu pháp lý cho thấy chúng ta có thể mở rộng các phương pháp của mình với sự trình bày kiến thức pháp lý lớn hơn và phức tạp hơn, tức là, về cả quy mô và đa dạng. Mở rộng nghiên cứu về biểu diễn logic trong các tài liệu pháp lý để cải thiện độ chính xác cho các nhiệm vụ truy hồi nói riêng và NLP pháp lý nói chung. Ngoài ra, chúng ta có thể thử các mô hình ngôn ngữ được huấn luyện trước lớn hơn, đặc biệt là các mô hình chuyên sâu cho mỗi ngôn ngữ cụ thể. Cuối cùng, phát triển các giải pháp và mô hình cho IR và QA pháp lý từ nhiều góc độ khác nhau để phục vụ nhiều loại người dùng khác nhau bao gồm các nhà lập pháp, thẩm phán, nguyên đơn, bị đơn và người dùng không chuyên.

List of Publications

[VTHY1] **Yen Thi-Hai Vuong**, Quan Minh Bui, Ha-Thanh Nguyen, Thi-Thu-Trang Nguyen, Vu Tran, Xuan-Hieu Phan, Ken Satoh, and Le-Minh Nguyen. "SM-BERT-CR: a deep learning approach for case law retrieval with supporting model."*Artificial Intelligence and Law* 31, no. 3 (2023): 601-628. (SCIE, ISI Q1)

[VTHY2] **Thi-Hai-Yen Vuong**, Hai-Long Nguyen, Tan-Minh Nguyen, Hoang-Trung Nguyen, Thai-Binh Nguyen, and Ha-Thanh Nguyen. "NOWJ at COLIEE 2023: Multi-task and Ensemble Approaches in Legal Information Processing."*The Review of Socionetwork Strategies* (2024): 1-21. (ESCI, WoS)

[VTHY3] **Thi-Hai-Yen Vuong**, Hoang Minh-Quan, Tan-Minh Nguyen, Hoang-Trung Nguyen, and Ha-Thanh Nguyen. "Constructing a Knowledge Graph for Vietnamese Legal Cases with Heterogeneous Graphs."*In 2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 1-6. IEEE, 2023. (Scopus)

[VTHY4] **Thi-Hai-Yen Vuong**, Ha-Thanh Nguyen, Quang-Huy Nguyen, Le-Minh Nguyen, Xuan-Hieu Phan. "Improving Vietnamese Legal Question-Answering System based on Automatic Data Enrichment". *In JSAI-isAI 2023. Lecture Notes in Computer Science*. Springer, Cham. (In press, Scopus)

[VTHY5] Hai-Long Nguyen, Thai-Binh Nguyen, Tan-Minh Nguyen, Ha-Thanh Nguyen, and **Hai-Yen Thi Vuong**. "Vlh team at alqac 2022: Retrieving legal document and extracting answer with bert-based model."*In 2022 14th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 1-6. IEEE, 2022. (Scopus)

[VTHY6] Nguyen, Hai-Long, Dieu-Quynh Nguyen, Hoang-Trung Nguyen, Thu-Trang Pham, Huu-Dong Nguyen, Thach-Anh Nguyen, **Thi-Hai-Yen Vuong** and Ha-Thanh Nguyen. "NeCo@ ALQAC 2023: Legal Domain Knowledge Acquisition for Low-Resource Languages through Data Enrichment."*In 2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 1-6. IEEE, 2023. (Scopus)