PHÂN LOẠI QUAN HỆ THAM CHIẾU TRONG VĂN BẢN PHÁP QUY

Nguyễn Thị Thanh Thủy, Đặng Bảo Chiến, Triệu Khương Duy, Ngô Xuân Bách, Từ Minh Phương
Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Xác định quan hệ tham chiếu trong văn bản quy phạm pháp luật là bước quan trọng trong các hệ thống xử lý văn bản pháp quy tự động. Quan hệ tham chiếu giúp người dùng thuận tiện trong việc tìm kiếm, tra cứu, phân tích, hay truy vấn nội dung văn bản quy phạm pháp luật. Đây chính là bài toán trích xuất và phân loại quan hệ giữa các thực thể, trong đó một thực thể là tham chiếu được đề cập đến trong nội dung và thực thể còn lại là văn bản pháp quy đang xem xét. Hướng tiếp cận đề xuất giải quyết bài toán này là sử dụng học máy có giám sát, là phương pháp phổ biến và đạt được độ chính xác cao trong các nghiên cứu về trích xuất quan hệ. Để trích xuất đặc trưng, ngoài thông tin về các thực thể, bài báo đề xuất sử dụng các thông tin ngữ cảnh liên quan đến các thực thể nhằm cải thiện độ chính xác trích xuất quan hệ. Bài báo cũng giới thiệu một tập dữ liệu gồm 5031 văn bản pháp quy được gán nhãn thực thể và mối quan hệ giữa các thực thể, được trích xuất từ cổng thông tin văn bản quy phạm pháp luật của Việt Nam. Các thử nghiệm trích xuất quan hệ trên tập dữ liệu này với ba thuật toán học máy Phân loại Bayes đơn giản, Cây quyết định (C4.5) và Máy véc-tơ tựa (SVM) cho kết quả khả quan, trong đó SVM đạt giá trị F_1 95,57%.

Từ khóa: trích xuất quan hệ, văn bản pháp quy, tham chiếu, học có giám sát.

I. GIỚI THIỆU

Văn bản quy phạm pháp luật (văn bản pháp quy) như hiến pháp, luật, nghị định, thông tư là 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 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.

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 Luật Ngân sách nhà nước…*". 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*", "*bị thay thế*", "*hết hiệu lực*", "*được 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 mối quan hệ giữa các văn bản là một phần công việc quan trọng.

Bài báo trình bày phương pháp trích xuất tự động quan hệ tham chiếu từ văn bản pháp quy. Bài toán này bao gồm hai bước: (1) *trích xuất tham chiếu từ văn bản pháp quy*, và (2) *phân loại quan hệ giữa các tham chiếu và văn bản pháp quy đang xem xét thành các loại như "căn cứ", "dẫn chiếu", "bị thay thế", "hết hiệu lực", "được sửa đổi hoặc bổ sung"*,... Bước (1) đã được đề cập đến trong một nghiên cứu trước [1], trong đó tham chiếu là văn bản pháp quy được đề cập đến trong nội dung của văn bản đang xem xét. Trong nghiên cứu này, chúng tôi tập trung giải quyết bước (2), tức là xác định quan hệ giữa các thực thể, trong đó một thực thể là tham chiếu được đề cập đến trong nội dung và thực thể còn lại là văn bản pháp quy đang xem xét (sau đây sẽ gọi tắt là *trích xuất quan hệ giữa các thực thể*).

Hình 1 trình bày ví dụ kết quả trích xuất thực thể tham chiếu và xác định quan hệ giữa các thực thể từ một đoạn văn bản trong "Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004" (ví dụ được nêu ở phần trên). Có ba 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**) "Thông tư số 59/TT-BTC ngày 23/6/2003", và (**3**) "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 (**2**), và không có quan hệ với thực thể tham chiếu (**3**) (trong Hình

Tác giả liên hệ: Nguyễn Thị Thanh Thủy
Email: thuyntt@ptit.edu.vn
Đến tòa soạn: 9/2020, chỉnh sửa:10/2020 , chấp nhận đăng: 10/2020

1 giá trị quan hệ là "***none***").

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, [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 1. Ví dụ tham chiếu và mối quan hệ giữa các tham chiếu với văn bản pháp quy

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 văn bản "Thông tư số 96/2004/TT-BTC ngày 13 tháng 10 năm 2004 của Bộ Tài chính" (Hình 1). Về xác định thực thể, ví dụ với thực thể thứ nhất, tham chiếu 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*". Do vậy, để trích xuất được thực thể cần phải có quy định về định dạng nhận diện thực thể. Về 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 (**2**) được đề cập trong nội dung. Tuy nhiên, có thể xác định quan hệ theo cách khác là hai thực thể tham chiếu (**1**) và (**2**) được xác định trong nội dung có quan hệ ***"dẫn chiếu"*** với thực thể văn bản đang xem xét. Thêm nữa, thực thể tham chiếu (**2**) 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**) trong cùng một câu.

Có hai hướng tiếp cận chính để giải quyết bài toán trích xuất quan hệ trong văn bản nói chung, bao gồm hướng tiếp cận dựa trên luật [2, 3, 4], và hướng tiếp cận dựa trên học máy [5, 6, 7, 8, 9]. Hướng tiếp cận dựa trên luật cần có chuyên gia xử lý và sinh luật theo từng lĩnh vực riêng. Hướng tiếp cận dựa trên học máy thống kê được nghiên cứu và phát triển nhiều hơn do không phụ thuộc vào tri thức chuyên gia, đồng thời được đánh giá là có độ chính xác cao. Gần đây, cũng có một số nghiên cứu tiếp cận giải quyết bài toán dựa trên các mô hình học sâu [10, 11, 12], tuy nhiên yêu cầu cần phải có lượng dữ liệu huấn luyện đủ lớn, và các mô hình này cũng có hạn chế về tốc độ xử lý. Do vậy, trong nghiên cứu này, chúng tôi tập trung vào hướng tiếp cận dựa trên học máy thống kê để giải quyết bài toán trích xuất quan hệ giữa các thực thể trong văn bản pháp quy.

Đóng góp của nghiên cứu gồm hai phần. Thứ nhất, nghiên cứu đề xuất phương pháp giải quyết bài toán phân loại quan hệ giữa các tham chiếu và văn bản pháp quy sử dụng học máy có giám sát. Cụ thể, chúng tôi sử dụng học có giám sát với các đặc trưng văn bản phù hợp cho bài toán đang xét. Để trích xuất đặc trưng, ngoài thông tin về thực thể, chúng tôi sử dụng các thông tin ngữ cảnh liên quan được trích chọn từ đoạn văn bản chứa thực thể tham chiếu nhằm cải thiện độ chính xác trích xuất quan hệ. Thứ hai, để kiểm tra tính hiệu quả của phương pháp đề xuất, chúng tôi xây dựng tập dữ liệu gồm 5031 văn bản pháp quy tiếng Việt được gán nhãn thực thể và quan hệ giữa các thực thể, và tiến hành thực nghiệm trên tập dữ liệu này. Các thử nghiệm trích xuất quan hệ trên tập dữ liệu cho kết quả khả quan với độ chính xác tốt nhất của hầu hết các quan hệ đều đạt độ đo F_1 trên 83%, độ đo F_1 tối đa đạt 95.57%.

Phần còn lại của bài báo được tổ chức như sau. Phần II mô tả các nghiên cứu liên quan. Phần III trình bày đề xuất phương pháp thực hiện trích xuất quan hệ trong văn bản pháp quy tiếng Việt. Việc xây dựng bộ dữ liệu và các thực nghiệm được trình bày trong phần Phần IV và Phần V. Cuối cùng, Phần VI là kết luận bài báo và định hướng nghiên cứu.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Phần này trình bày các nghiên cứu liên quan đến trích xuất quan hệ và trích xuất thông tin trong văn bản pháp quy.

A. Trích xuất quan hệ

Các nghiên cứu trước đây về trích xuất quan hệ thường sử dụng phương pháp tiếp cận dựa trên luật, ví dụ như [2, 3, 4]. 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 sắc về nền tảng và đặc điểm của lĩnh vực xử lý. Do vậy, nhược điểm chính của cách tiếp cận này là cần phải có sự tham gia của chuyên gia và khó chuyển đổi giữa các lĩnh vực khác nhau.

Một cách tiếp cận phổ biến hiện nay 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 [5, 6]. 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 Kambhatla [7] 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 [8] 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. Nghiên cứu [9] sử dụng một nhân dạng cây mới để trích xuất quan hệ đã được đề xuất, bằng cách chú thích mỗi nút trên cây với một tập các đặc trưng phân biệt để tinh chỉnh biểu diễn cho cây cú pháp.

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 [10, 11, 12] 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 thống kê là tốc độ, cùng với yêu cầu phải có tập dữ liệu huấn luyện đủ lớn.

B. Trích xuất thông tin trong văn bản pháp quy

Walter [13] 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 từ văn bản pháp quy tiếng Đức. Nghiên cứu [14] mô tả hệ thống Legal TRUTHS, nhằm trích xuất các thông tin quan trọng cho các vụ án hình sự, như tội phạm, thời gian, ủy ban, nguyên đơn và hình phạt được xác định từ một bộ tài liệu mẫu. Nghiên cứu [15] sử dụng cách tiếp cận kết hợp học máy và đặc trưng về ngôn ngữ để trích xuất thông tin và kết quả đạt được độ chính xác tương đối cao. Nghiên cứu đề xuất sử dụng bộ phân loại SVM để liên kết các khái niệm với tài liệu pháp lý và bộ phân tích cú pháp ngôn ngữ tự nhiên để xác định các thực thể, gồm vị trí, tổ chức, ngày tháng và tham chiếu đến các tài liệu khác. Andrew [16] sử dụng kết hợp cả luật dựa trên biểu thức chính quy và CRF (Conditional Random Field) để trích xuất thông tin thực thể gồm tên người, tổ chức, vai trò và chức năng của người cùng với các quan hệ giữa các thực thể trong văn bản luật và cũng thu được độ chính xác khá cao. Nghiên cứu [1] sử dụng cả CRF và mô hình học sâu để trích xuất thực thể pháp luật tham chiếu trong văn bản luật Việt Nam. Kết quả tốt nhất thu được F_1 lớn hơn 95% với mô hình học sâu BiLSTM (Bidirectional Long-Short Term Memory) kết hợp CRF.

Các khảo sát trên cho thấy bài toán trích xuất thông tin trong văn bản luật khá phổ biến và đã đạt được nhiều kết quả đáng chú ý. Tuy nhiên, với hiểu biết của nhóm nghiên cứu, cho đến nay chưa thấy nghiên cứu nào đề cập đến bài toán trích xuất quan hệ giữa thực thể tham chiếu với văn bản pháp quy hiện tại đang xem xét, như được đề xuất trong nghiên cứu này.

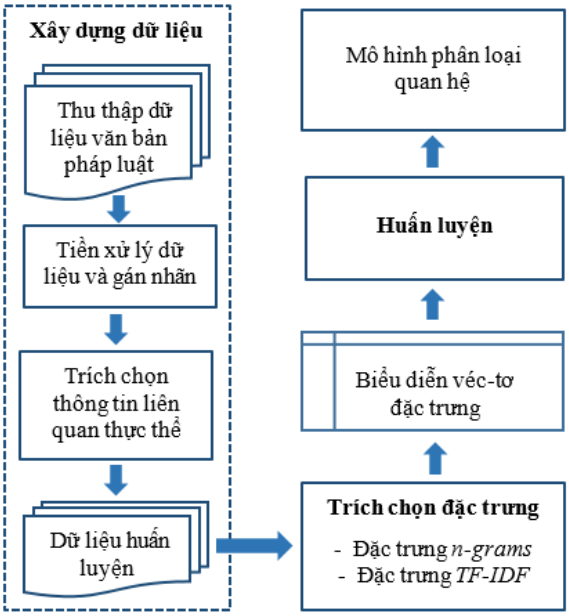
III. PHƯƠNG PHÁP ĐỀ XUẤT

Phần này trình bày đề xuất phương pháp phân loại quan hệ tham chiếu trong văn bản pháp quy có chứa thực thể tham chiếu. Các loại quan hệ được xác định bao gồm: căn cứ, dẫn chiếu, được hướng dẫn, được sửa đổi hoặc bổ sung, bị thay thế,…

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

Với mỗi tham chiếu B_k, xét đoạn văn bản chứa 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 tham chiếu B_k.

Hình 2 trình bày sơ đồ các bước đề xuất giải quyết bài toán phân loại quan hệ tham chiếu 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.



Hình 2. Sơ đồ các bước đề xuất giải quyết bài toán phân loại quan hệ tham chiếu trong văn bản pháp quy

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, chúng tôi 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:

1) Thực thể tham chiếu B_k,
2) Phần văn bản ở phía trước thực thể tham chiếu B_k (trong cùng câu với B_k),
3) Phần văn bản ở phía sau thực thể tham chiếu B_k (trong cùng câu với B_k),
4) Tên của thực thể văn bản A,
5) 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 trích xuất 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*. 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 bài toán.

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-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:

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

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

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

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

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

Giá trị *TF-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-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-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). Việc kết hợp đặc trưng *n-grams* với đặc trưng *TF-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* $v_{tf\text{-}idf}(d_i)$ 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 $v_{tf\text{-}idf}(d_i)$ tạo thành véc-tơ $v(d_i)$ (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 luật đã được xác định.

Để huấn luyện mô hình, chúng tôi sử dụng ba thuật toán học máy khác nhau là Phân loại Bayes đơn giản (Naïve Bayes) [17], Cây quyết định [18, 19] và Máy véc-tơ tựa [20], đại diện cho ba nhóm thuật toán khác nhau: dựa trên trên mô hình xác suất, dựa trên cây và dựa trên hàm nhân. Đây là các thuật toán đã được chứng minh là hiệu quả cho các bài toán phân loại. Phần sau sẽ giới thiệu tóm tắt ba thuật toán này.

1) ***Phân loại Bayes đơn giản*** (Naïve Bayes).

Phân loại Bayes đơn giản [17] là thuật toán phân loại dựa trên định lý Bayes và có giả thiết về tính độc lập giữa các thuộc tính. Cho một ví dụ mẫu $x = (x_1, x_2, \ldots, x_n)$, phương pháp dựa trên thuật toán Naïve Bayes sẽ tìm lớp y_{NB} phù hợp nhất với mẫu *x* như sau:

$$y_{NB} = argmax_{y \in Y} p(x|y)p(y)$$

Trong đó *Y* là tập tất cả các lớp. Do giả thiết về tính độc lập giữa các thuộc tính nên:

$$p(x|y) = \prod_{i=1}^{n} p(x_i|y)$$

Xác suất $p(y)$ và $p(x_i|y)$ có thể được tính toán đơn giản dựa trên dữ liệu huấn luyện.

2) ***Cây quyết định (C4.5)***.

Cây quyết định [18] là một mô hình phân loại dưới dạng cấu trúc cây. Mô hình này chia một tập dữ liệu ban đầu thành các tập con nhỏ hơn theo kiểu đệ quy, và đồng thời một cây quyết định được phát triển dần dần. Kết quả cuối cùng là một cây, với mỗi nút bên trong đại diện cho một thuộc tính, mỗi nhánh đại diện cho một quyết định và mỗi nút lá đại diện cho một nhãn lớp. Quyết định được thực hiện sau khi tính toán tất cả các thuộc tính. Các đường dẫn từ gốc đến nút lá đại diện cho các quy tắc phân loại. C4.5 [19] là một mô hình cây quyết định dựa trên khái niệm entropy. Tại mỗi nút trên cây, C4.5 chọn ra thuộc tính tốt nhất để chia dữ liệu vào các nút con một cách hiệu quả nhất. Thuộc tính được chọn là thuộc tính có độ lợi thông tin sau chuẩn hóa cao nhất.

3) ***Máy véc-tơ tựa***.

Máy véc-tơ tựa [20] (Support Vector Machine) là thuật toán phân loại rất hiệu quả đối với nhiều bài toán phân loại khác nhau trong xử lý ngôn ngữ tự nhiên [21, 22]. SVM 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 Gaussian.

IV. TẬP DỮ LIỆU

Phần này sẽ mô tả về việc xây dựng tập dữ liệu để sử dụng cho các thực nghiệm.

A. *Thu thập và tiền xử lý dữ liệu*

Nguồn dữ liệu được thu thập từ Cổng thông tin văn bản quy phạm 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 5031 văn bản pháp quy. Tập dữ liệu này sẽ được sử dụng cho bước tiếp theo là gán nhãn dữ liệu.

B. *Gán nhãn dữ liệu*

Có ba người thực hiện việc gán nhãn dữ liệu cho từng văn bản sau khi đã được tiền xử lý. Hai người gán nhãn đầu là sinh viên ngành Công nghệ thông tin và người gán nhãn thứ ba là Cử nhân ngành Luật.

Việc gán nhãn được thực hiện bao gồm 2 công đoạn như sau:

1) ***Gán nhãn thực thể*** là tham chiếu của văn bản được đề cập (văn bản *B*) trong nội dung của văn bản đang xét (văn bản *A*). Quy trình gán nhãn thực thể tham chiếu được thực hiện theo hướng dẫn trong nghiên cứu [1], bao gồm 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 các biểu thức chính quy. Có một số quan sát và thảo luận như sau:

- Tham chiếu của văn bản pháp quy thường bắt đầu bằng một từ khóa về loại văn bản pháp quy. Do vậy, chúng tôi 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,…
- Tham chiếu của văn bản pháp quy thường kết thúc theo một trong các dạng sau:
  - Ngày tháng năm (có các dạng: năm *yyyy*, tháng *mm* năm *yyyy* hoặc ngày *dd* tháng *mm* năm *yyyy*).
  - Mã số văn bản pháp quy (ví dụ như *85/2015/QH13*)
  - Một từ có xác suất cao là từ cuối cùng trong tên văn bản pháp quy. Danh sách các từ này được tạo ra bằng cách thực hiện thống kê tên của tất cả các tài liệu/văn bản pháp quy được thu thập.

Loại thực thể được xác định là từ khóa đầu tiên của tham chiếu văn bản pháp quy.

Gán nhãn thủ công. Trong bước này thực thể tham chiếu và loại thực thể đã được gán nhãn ở bước gán nhãn tự động sẽ được kiểm tra và sửa lỗi thủ công bởi hai người gán nhãn độc lập, là hai sinh viên ngành Công nghệ thông tin. Người gán nhãn thứ ba, là cử nhân ngành Luật, sẽ kiểm tra và đưa ra quyết định cuối cùng khi có sự bất đồng giữa hai người gán nhãn đầu.

Kết quả thu được tập dữ liệu đã được gán nhãn thực thể, với 9 loại thực thể, bao gồm: Hiến pháp, Bộ luật, Luật, Pháp lệnh, Nghị định, Nghị quyết, Quyết định, Thông tư, Thông tư liên tịch.

2) ***Gán nhãn mối quan hệ*** giữa thực thể văn bản *A* với thực thể văn bản *B*. Sau khi khảo sát nguồn dữ liệu văn bản pháp quy, chúng tôi xác định 6 loại quan hệ được gán nhãn bao gồm: căn cứ, dẫn chiếu, hết hiệu lực, bị thay thế, được sửa đổi hoặc bổ sung và được hướng dẫn. Thực thể không có quan hệ với thực thể văn bản đang xét được gán nhãn là "*none*" (được coi là loại quan hệ thứ 7).

Ban đầu, hai sinh viên ngành Công nghệ thông tin thực hiện việc gán nhãn quan hệ độc lập với nhau. Sau đó, người gán nhãn thứ ba là cử nhân Luật sẽ kiểm tra lại. Nếu có ý kiến bất đồng giữa hai người gán nhãn đầu thì người thứ ba sẽ đưa ra quyết định cuối cùng.

Hình 3 trình bày ví dụ một đoạn văn bản pháp quy được gán nhãn thực thể tham chiếu và mối quan hệ. 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 văn bản đang xem xét với thực thể văn bản được tham chiếu trong nội dung.

```
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
Luật Ngân sách nhà nước, <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. Văn bản pháp quy được gán nhãn quan hệ với văn bản tham chiếu trong nội dung

Bảng I trình bày chi tiết thống kê số lượng quan hệ có trong tập dữ liệu. Tổng cộng có 60.688 quan hệ được gán nhãn cho 7 loại, trong đó hai loại quan hệ có số lượng nhiều nhất là "*dẫn chiếu*" (27.502) và "*căn cứ*" (18.377).

Bảng I. Thống kê số lượng quan hệ trong tập dữ liệu

STT	LOẠI QUAN HỆ	NHÃN	SỐ LƯỢNG
1	Căn cứ	CC	18.377
2	Dẫn chiếu	DaC	27.502
3	Hết hiệu lực	HHL	1.473
4	Bị thay thế	BTT	1.751
5	Được sửa đổi hoặc bổ sung	DSD	1.359
6	Được hướng dẫn	DHD	368
7	Không có quan hệ	none	9.858
Tổng			**60.688**

V. CÁC THỰC NGHIỆM VÀ KẾT QUẢ

A. *Thiết lập thực nghiệm*

Dữ liệu được chia ngẫu nhiên thành 5 phần để thực hiện kiểm tra chéo. Hiệu năng của mô hình trích xuất quan hệ được đo bằng:

1) ***Độ chính xác (accuracy)***: số quan hệ được trích xuất chính xác trên tổng số quan hệ cần được trích xuất.

$$acc = \frac{Số\ quan\ hệ\ được\ trích\ xuất\ chính\ xác}{Tổng\ số\ quan\ hệ\ cần\ được\ trích\ xuất}$$

2) ***Độ chính xác (precision), độ bao phủ (recall) và độ đo F_1*** cho từng loại quan hệ. Lấy ví dụ với loại quan hệ "*căn cứ*". Giả sử A ký hiệu cho tập các quan hệ được xác định bởi mô hình, và B ký hiệu cho tập các quan hệ được gán nhãn bởi người gán nhãn, thì độ chính xác, độ bao phủ và độ đo F_1 cho quan hệ "*căn cứ*" được tính như sau (tương tự cho các loại quan hệ khác):

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

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

và

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

B. *Kết quả thực nghiệm*

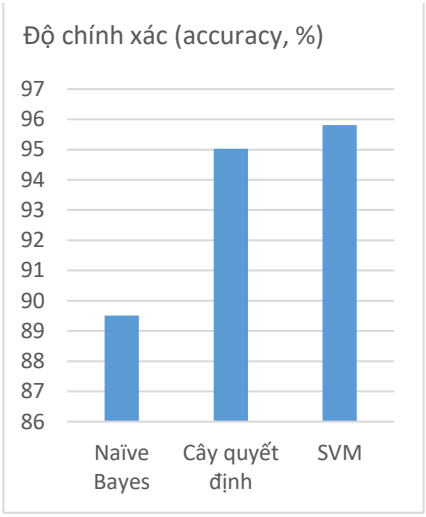
Mục đích xây dựng các thực nghiệm:

- Giải quyết bài toán trích xuất quan hệ giữa các thực thể văn bản luật bằng các phương pháp học máy khác nhau và so sánh hiệu năng của các bộ phân loại.
- So sánh các phương pháp trích chọn thông tin liên quan đến các thực thể để xây dựng dữ liệu huấn luyện.
- So sánh các phương pháp trích chọn đặc trưng để xây dựng mô hình trích xuất quan hệ.

Phần sau sẽ mô tả các thực nghiệm và kết quả.

1) ***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 SVM. Với mỗi phương pháp, chúng tôi 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 4. So sánh các bộ phân loại khác nhau

Hình 4 trình bày kết quả tốt nhất thực nghiệm được với ba bộ phân loại đề xuất. Nhìn chung, cả ba đều có kết quả trích xuất quan hệ tương đối tốt, với độ chính xác (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, chúng tôi sẽ thực hiện thực nghiệm sử dụng bộ phân loại tốt nhất là SVM.

2) ***So sánh các phương pháp trích chọn thông tin liên quan thực thể***

Bảng II. Ví dụ trích chọn thông tin liên quan thực thể

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

Để trích xuất 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 II trình bày ví dụ về các thông tin được trích chọn trong một đoạn văn bản luật 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*".

Chúng tôi đề xuất ba phương pháp trích chọn thông tin liên quan thực thể được sử dụng để 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 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 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 III trình bày tóm tắt các phương pháp trích chọn thông tin liên quan thực thể.

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

THỬ NGHIỆM	PHƯƠNG PHÁP TRÍCH CHỌN
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*

Để so sánh kết quả giữa các phương pháp trích chọn thông tin liên quan thực thể, chúng tôi chỉ sử dụng một loại đặc trưng đơn giản là *n-grams*. Mô hình huấn luyện sử dụng SVM tuyến tính với các tham số mô hình được tinh chỉnh dựa trên hàm *gridsearch* (dùng thư viện *sklearn* của Python).

Bảng IV trình bày kết quả trích xuất 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.

Bảng IV. Kết quả trích xuất 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**

Kết quả trong Bảng IV cho thấy độ chính xác của trích xuất cho từng loại quan hệ tương đối cao. 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 "*được 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 368 lần (trên tổng số 60.688 quan hệ, xem Bảng I). 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 18.377 lần (*căn cứ*) và 27.502 (*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 95,33%. 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.

3) ***So sánh các phương pháp trích chọn đặc trưng***

Để thực nghiệm với các phương pháp trích chọn đặc trưng khác nhau, chúng tôi 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 phần mô tả trên (phần 2). 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. Chúng tôi 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 luật.

Bảng V 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* (precision), *độ bảo phủ* (recall) và độ đo F_1.

Bảng V. Kết quả trích xuất quan hệ với các phương pháp trích chọn đặc trưng

QUAN HỆ	n-grams + TF-IDF			n-grams (F_1)
	Pre.	Rec.	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**

Có thể thấy, việc kết hợp đặc trưng *n-grams* và *TF-IDF* cho kết quả trích xuất quan hệ giữa các thực thể văn bản luật 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 (precision) là 95,68%, độ bao phủ (recall) 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.

C. *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 VI 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. Do trong Bảng VI, FNR khá thấp nên chúng ta 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 VI. Phân tích lỗi trích xuất 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 27.502, 18.377 và 9.858 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 *none* bị nhận nhầm thành BTT. *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 trích xuất 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 VI). 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 hay 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.

VI. KẾT LUẬN

Bài báo đã trình bày một nghiên cứu thực nghiệm về bài toán trích xuất quan hệ giữa các thực thể là tham chiếu với thực thể là văn bản pháp quy hiện tại đang xem xét. Phương pháp đề xuất sử dụng SVM và các đặc trưng được trích chọn dựa trên sự kết hợp của các thông tin về các thực thể cùng các thông tin ngữ cảnh liên quan giúp làm tăng độ chính xác trích xuất quan hệ. Các thực nghiệm được hành trên tập dữ liệu hơn 5000 văn bản pháp quy Việt Nam, với các thực thể và mối quan hệ giữa các thực thể được gán nhãn thủ công. Kết quả thực nghiệm cho thấy phương pháp đề xuất có độ chính xác khả quan, với hầu hết các quan hệ đều đạt trên 83% tính theo độ đo F_1. 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, và ngược lại.

Trong thời gian tới, chúng tôi dự định nghiên cứu giải quyết bài toán này dựa trên các kỹ thuật học sâu ứng dụng cho các bài toán có tập dữ liệu nhỏ. Đây là một hướng nghiên cứu thú vị, hứa hẹn với khả năng tăng tính hiệu quả cho việc trích xuất các quan hệ giữa các thực thể tham chiếu với văn bản pháp quy.

LỜI CẢM ƠN

Nghiên cứu sinh được hỗ trợ bởi chương trình học bổng đào tạo tiến sĩ trong nước của Quỹ Đổi mới sáng tạo Vingroup, mã số VINIF.2019.TS.65.

TÀI LIỆU THAM KHẢO

[1] N. X. Bach, N. T. T. Thuy, D. B. Chien, T. K. Duy, T. M. Hien, and T.M Phuong. "Reference Extraction from Vietnamese Legal Documents". In Proceedings of the Tenth International Symposium on Information and Communication Technology, pp. 486-493. 2019.

[2] T.M. Phuong, D. Lee and K.H. Lee. "Learning rules to extract protein interactions from biomedical text". In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 148-158. Springer, Berlin, Heidelberg. 2003.

[3] C. Zhang, X. Zhang, W. Jiang, Q. Shen and S. Zhang. "Rule-based extraction of spatial relations in natural language text". In 2009 International Conference on Computational Intelligence and Software Engineering, pp. 1-4. IEEE. 2009.

[4] K. Nebhi. "A rule-based relation extraction system using DBpedia and syntactic parsing". In Proceedings of the NLP-DBPEDIA-2013 Workshop co-located with the 12th International Semantic Web Conference (ISWC 2013). 2013.

[5] T. Hasegawa, S. Sekine, and R. Grishman, R. "Discovering relations among named entities from large corpora". In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pp. 415-422. 2004.

[6] A. Sun, R. Grishman, and S. Sekine. "Semi-supervised relation extraction with large-scale word clustering". In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp. 521-529. 2011.

[7] N. Kambhatla. "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations". In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, pp. 22-es. 2004.

[8] R. Bunescu and R. Mooney. "A shortest path dependency kernel for relation extraction". In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 724-731. 2005.

[9] L. Sun, and X. Han. "A feature-enriched tree kernel for relation extraction". In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, pp. 61-67. 2014.

[10] X. Jiang, Q. Wang, P. Li and B. Wang. "Relation extraction with multi-instance multi-label convolutional neural networks". In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1471-1480. 2016.

[11] Y. Lin, S. Shen, Z. Liu, H. Luan and M. Sun. "Neural relation extraction with selective attention over instances". In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, pp. 2124-2133. 2016.

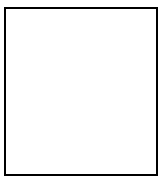
[12] D. Zeng, K. Liu, S. Lai, G. Zhou and J. Zhao. "Relation classification via convolutional deep neural network". In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 2335-2344. 2014.

[13] S. Walter. "Linguistic Description and Automatic Extraction of Definitions from German Court Decisions". In LREC. 2008.
[14] T. T. Cheng, J. L. Cua, M. D. Tan, K. G. Yao and R. E. Roxas. "Information extraction from legal documents". In 2009 eighth international symposium on natural language processing, pp. 157-162. IEEE. 2009.
[15] P. Quaresma and T. Gonçalves. "Using linguistic information and machine learning techniques to identify entities from juridical documents". In Semantic Processing of Legal Texts, pp. 44-59. Springer, Berlin, Heidelberg. 2010.
[16] J. J. Andrew. "Automatic extraction of entities and relation from legal documents". In Proceedings of the Seventh Named Entities Workshop, pp. 1-8. 2018.
[17] I. Rish. "An Empirical Study of the Naive Bayes classifier". In Proceedings of IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. 2001.
[18] J. R. Quinlan. "Induction of decision trees". Machine learning, 1(1), 81-106. 1986.
[19] I. H. Witten and E. Frank. "Data mining: practical machine learning tools and techniques with Java implementations". ACM Sigmod Record, 31(1), 76-77. 2002.
[20] V.N. Vapnik. "Statistical Learning Theory". Wiley-Interscience, 1998.
[21] N. Jihan, Y. Senarath, D. Tennekoon, M. Wickramarathne, and S. Ranathunga. "Multi-Domain Aspect Extraction using Support Vector Machines". In Proceedings of the Conference on Computational Linguistics and Speech Processing (ROCLING), pp. 308–322. 2017.
[22] M. Pontiki et al. "SemEval-2016 Task 5: Aspect Based Sentiment Analysis". In Proceedings of SemEval–2016, pp. 19–30, 2016.

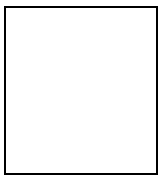
REFERENCE RELATIONS CLASSIFICATION IN LEGAL DOCUMENTS

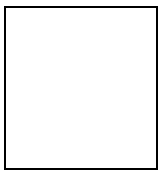
Abstract: Identifying reference relations in legal documents is an important step in automated legal document processing systems. Using reference relations helps users to conveniently search, consult, analyze, or query the content of legal documents. This is the problem of extracting and classifying relations between entities, in which one entity is the reference mentioned in the text and the other is the legal document under consideration. The proposed approach to solving this problem is to use supervised machine learning, which is a popular method and achieves high accuracy in relation extraction works. For feature extraction, contextual information related to the entities is proposed to use in combination with entity information in order to improve relation extraction accuracy. We also introduces an annotated dataset of 5031 legal documents extracted from Vietnam's legal document portal in which entities and relations among entities are labelled. Experiments are conducted on this dataset with three machine learning algorithms including Naïve Bayes, Decision Tree (C4.5) and SVM, yielding positive results with F_1-score of 95.57% (SVM).

Keywords: relation extraction, legal document, reference, supervised learning.

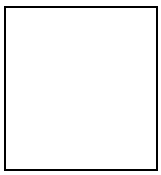


Nguyễn Thị Thanh Thủy. Nhận học vị Thạc sĩ năm 2009 tại Hàn Quốc. Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, xử lý ngôn ngữ tự nhiên.

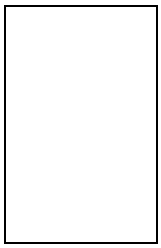




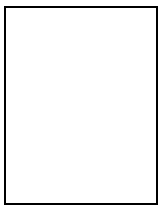
Đặng Bảo Chiến. Nhận bằng Kỹ sư Công nghệ thông tin năm 2019. Hiện đang làm nghiên cứu tại Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, xử lý ngôn ngữ tự nhiên.



Triệu Khương Duy. Nhận bằng Kỹ sư Công nghệ thông tin năm 2019. Hiện đang làm nghiên cứu tại Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, xử lý ngôn ngữ tự nhiên.



Ngô Xuân Bách. Nhận học vị Tiến sĩ năm 2014 tại Viện Khoa học và Công nghệ tiên tiến Nhật Bản (JAIST). Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: xử lý ngôn ngữ tự nhiên, học máy, hệ khuyến nghị.



Từ Minh Phương. Nhận học vị Tiến sĩ năm 1995. Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, hệ khuyến nghị, xử lý ngôn ngữ tự nhiên.