MỘT PHƯƠNG PHÁP TRÍCH XUẤT KẾT HỢP THỰC THỂ VÀ QUAN HỆ THAM CHIẾU TRONG VĂN BẢN PHÁP QUY

Nguyễn Thị Thanh Thủy, Nguyễn Ngọc Điệp
Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Để có thể xây dựng được các hệ thống xử lý văn bản pháp quy tự động như tìm kiếm, tra cứu, phân tích, hay truy vấn nội dung, thì việc trích xuất ra được những thông tin cần thiết trong các văn bản pháp quy, bao gồm thực thể tham chiếu và quan hệ tham chiếu, là một trong những công việc quan trọng cần phải được thực hiện trước tiên. Các nghiên cứu trước đây khi có yêu cầu trích xuất cả hai loại thông tin thực thể tham chiếu và quan hệ tham chiếu, hoặc khi chỉ có yêu cầu trích xuất quan hệ tham chiếu, sẽ thường thực hiện theo cách làm lần lượt, đầu tiên là trích xuất thực thể, và sau đó là trích xuất quan hệ. Như vậy, độ chính xác của việc trích xuất quan hệ tham chiếu sẽ phụ thuộc vào việc có trích xuất được đúng hay không các thực thể tham chiếu. Trong bài báo này, chúng tôi trình bày một phương pháp cải tiến hơn để giải quyết bài toán trích xuất thông tin trong văn bản pháp quy, đó là phương pháp trích xuất kết hợp thực thể và quan hệ tham chiếu cùng lúc, sử dụng mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer. Kết quả cho thấy mô hình đề xuất có độ chính xác khá cao, với độ đo F_1 lên tới 96,8% cho kết quả trích xuất kết hợp cả hai thông tin. Kết quả trích xuất riêng cũng vượt trội so với các nghiên cứu trước: trích xuất thực thể tham chiếu đạt độ đo F_1 là 98,4%, trích xuất quan hệ tham chiếu đạt độ đo F_1 là 97,7%, trên tập dữ liệu 5031 văn bản pháp quy tiếng Việt.

Từ khóa: văn bản pháp quy, trích xuất thực thể tham chiếu, trích xuất quan hệ tham chiếu, trích xuất thực thể và quan hệ kết hợp.

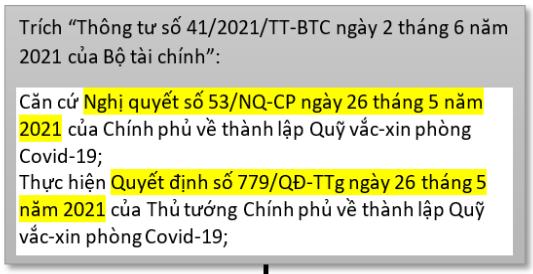
I. GIỚI THIỆU

Xử lý văn bản pháp quy tự động bao gồm các hoạt động như tìm kiếm, tra cứu, truy vấn văn bản pháp luật là một việc khó khăn nhưng rất cần thiết trong các hệ thống xử lý văn bản quy phạm pháp luật (hay còn gọi là văn bản pháp quy). Việc này không những hỗ trợ được cho người dùng bình thường trong cuộc sống hàng ngày, mà còn hỗ trợ được cho cả các chuyên gia về luật, luật sư, do mỗi Quốc gia thường có số lượng rất lớn văn bản pháp quy và các văn bản này vẫn được gia tăng, cập nhật hàng năm. Để có thể xây dựng được các hệ thống xử lý văn bản pháp quy tự động, việc trích xuất ra được những thông tin cần thiết trong các văn bản quy phạm pháp luật là một trong những công việc quan trọng cần phải được thực hiện trước tiên.

Một điều rất dễ nhận thấy là trong nội dung của một văn bản pháp quy thường nhắc đến nhiều các văn bản pháp quy khác có liên quan đến văn bản đang được đọc (hay xem xét), ví dụ như văn bản đang đọc có *căn cứ* từ một văn bản khác, hoặc văn bản đang đọc là *sửa đổi/bổ sung* của một văn bản trước đó, hoặc văn bản đang đọc là văn bản được *thay thế* cho một văn bản trước đó,… Trong các trường hợp này, người dùng thường có nhu cầu tìm kiếm thông tin của các văn bản được nhắc đến trong văn bản đang đọc để tìm hiểu sâu hơn về những nội dung liên quan họ đang cần. Nếu người dùng chỉ sử dụng các công cụ tìm kiếm thông thường như trên các trang văn bản pháp luật hiện có thì sẽ khá tốn kém thời gian và công sức.

Hiện đã có một số nghiên cứu liên quan đến việc trích xuất thông tin trong văn bản pháp quy, với hai nhiệm vụ cụ thể được quan tâm đó là trích xuất thực thể trong văn bản pháp quy và trích xuất quan hệ trong văn bản pháp quy [4, 5, 7]. Nhiệm vụ thứ nhất giải quyết nhu cầu đầu tiên được nhắc đến ở đoạn trên là nhận biết ra được tên của văn bản cũng như những tham chiếu đến văn bản được nhắc đến trong văn bản đang đọc, được gọi chung là thực thể tham chiếu. Đây chính là một loại thực thể đặc trưng cho văn bản pháp quy. Nhiệm vụ thứ hai giải quyết nhu cầu tiếp theo là nhận biết ra được mối liên hệ của văn bản đang đọc với văn bản được nhắc tên/tham chiếu đến trong nội dung của văn bản đang đọc. Đây chính là mối quan hệ tham chiếu giữa các văn bản pháp quy đã được xác định. Việc nhận biết mối quan hệ tham chiếu giữa các văn bản pháp quy thường được thực hiện bằng cách quy về bài toán phân loại quan hệ văn bản. Hình 1 trình bày một ví dụ về trích xuất thực thể tham chiếu và quan hệ tham chiếu trong sử người dùng đang xem xét văn bản quy phạm pháp luật là "*Thông tư số 41/2021/TT-BTC ngày 2 tháng 6 năm 2021 của Bộ tài chính*" có trích đoạn nội dung như trong phần nửa phía trên của Hình 1. Trong đoạn văn bản này, có hai văn bản được nhắc đến là "*Nghị quyết số 53/NQ-CP ngày 26 tháng 5 năm 2021*", và "*Quyết định số 779/QĐ-TTg ngày 26 tháng 5 năm 2021*". Trong đó, ngữ nghĩa ở đây là: "*Thông tư số 41/2021/TT-BTC ngày 2 tháng 6 năm 2021*" căn cứ theo "*Nghị quyết số 53/NQ-CP ngày 26 tháng 5 năm 2021*" và căn cứ theo "*Quyết định số 779/QĐ-TTg ngày 26 tháng 5 năm 2021*". Như vậy, hai nhiệm vụ trích xuất thực thể và

Tác giả liên hệ: Nguyễn Thị Thanh Thủy,
Email: thuyr205@gmail.com
Đến tòa soạn: 9/2021, chỉnh sửa: 10/2021, chấp nhận đăng: 10/2021.



Thực thể 1		Mối quan hệ	Thực thể 2	
Tên thực thể	Loại thực thể		Tên thực thể	Loại thực thể
Thông tư số 41/2021/TT-BTC ngày 2 tháng 6 năm 2021	Thông_tư	Căn_cứ	Nghị quyết số 53/NQ-CP ngày 26 tháng 5 năm 2021	Nghị_quyết
Thông tư số 41/2021/TT-BTC ngày 2 tháng 6 năm 2021	Thông_tư	Căn_cứ	Quyết định số 779/QĐ-TTg ngày 26 tháng 5 năm 2021	Quyết_định

Hình 1. Ví dụ về trích xuất thực thể và quan hệ trong một đoạn văn bản pháp quy

mối quan hệ giữa các thực thể trong văn bản pháp quy bao gồm (phần nửa phía dưới của Hình 1):

1. Trích xuất hai thực thể tham chiếu: "*Nghị quyết số 53/NQ-CP ngày 26 tháng 5 năm 2021*", và "*Quyết định số 779/QĐ-TTg ngày 26 tháng 5 năm 2021*".

2. Trích xuất quan hệ giữa thực thể văn bản đang xem xét với hai thực thể đã trích xuất được ở trên, bao gồm: Quan hệ ***Căn_cứ*** giữa "*Thông tư số 41/2021/TT-BTC ngày 2 tháng 6 năm 2021*" và "*Nghị quyết số 53/NQ-CP ngày 26 tháng 5 năm 2021*"; Quan hệ ***Căn_cứ*** giữa "*Thông tư số 41/2021/TT-BTC ngày 2 tháng 6 năm 2021*" và "*Quyết định số 779/QĐ-TTg ngày 26 tháng 5 năm 2021*".

Theo khảo sát về các nghiên cứu trước đây của chúng tôi, hai nhiệm vụ trích xuất thực thể và trích xuất quan hệ giữa các thực thể trong văn bản pháp quy được thực hiện một cách rời rạc với nhau, nghĩa là thực hiện riêng biệt từng nhiệm vụ [4, 5, 7]. Như vậy, trong trường hợp có yêu cầu trích xuất đồng thời cả hai thông tin về thực thể và quan hệ tham chiếu, hoặc chỉ trích xuất quan hệ tham chiếu, thì bài toán sẽ được giải quyết theo cách: bước đầu tiên sẽ thực hiện trích xuất thực thể văn bản pháp quy, và sau đó bước thứ hai mới thực hiện phân loại quan hệ tham chiếu giữa các thực thể trong văn bản pháp quy. Với cách thực hiện này, độ chính xác của việc trích xuất quan hệ tham chiếu (trong bước thứ hai) sẽ phụ thuộc vào việc có trích xuất được đúng hay không các thực thể tham chiếu (trong bước thứ nhất).

Trong nghiên cứu này, chúng tôi trình bày một phương pháp để giải quyết cùng lúc hai nhiệm vụ: trích xuất thực thể văn bản pháp quy và phân loại quan hệ văn bản pháp quy. Đây là phương pháp trích xuất kết hợp thực thể và quan hệ đồng thời sử dụng mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer. Trong mô hình này, quá trình trích xuất bộ thông tin thực thể, quan hệ pháp quy gồm hai bước: ban đầu, mô hình xác định tất cả các thực thể có thể có trong một câu; sau đó, với mỗi thực thể mô hình sẽ tìm các quan hệ có thể có cho thực thể đó. Mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer đề xuất được cài đặt bao gồm: PhoBERT (một mô-đun mã hóa dựa trên BERT được huấn luyện sẵn cho dữ liệu tiếng Việt), một mô-đun xác định thực thể và một mô-đun xác định loại quan hệ của thực thể đó với văn bản pháp quy đang xem xét. Kết quả thực nghiệm cho thấy mô hình đề xuất có hiệu quả cao trong nhiệm vụ trích xuất kết hợp thực thể và quan hệ trong các văn bản pháp quy tiếng Việt. Ngoài ra, trên các các nhiệm vụ riêng lẻ, kết quả thực nghiệm cho thấy mô hình đề xuất cũng vượt trội so với các mô hình sử dụng để giải quyết từng bài toán, trong đó nhiệm vụ trích xuất thực thể đạt 98,4% so với 95,3% trong mô hình đưa ra của nghiên cứu [5] và nhiệm vụ xác định quan hệ tham chiếu đạt F_1 97,7% so với 95,5% trong mô hình đưa ra của nghiên cứu [7].

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 kết hợp đồng thời thực thể và quan hệ giữa các văn bản pháp quy. Kết quả và những phân tích thực nghiệm được trình bày trong phần Phần IV. Cuối cùng, Phần V là kết luận bài báo và định hướng nghiên cứu.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Một loại thực thể đặc trưng riêng trong lĩnh vực văn bản pháp quy là thực thể tham chiếu, chính là tên và các tham chiếu đến văn bản pháp quy trong nội dung của văn bản đang xem xét. Thực thể tham chiếu trong văn bản pháp quy đã được nghiên cứu trích xuất với một số ngôn ngữ khác nhau [1, 2, 3, 4, 5]. Có thể phân chia các phương pháp trích xuất thực thể tham chiếu thành hai nhóm chính, bao gồm các phương pháp dựa trên luật [1, 2, 3] và các phương pháp dựa trên học máy [4, 5]. Trong hai nhóm phương pháp này, các phương pháp dựa trên học máy có ưu thế về độ chính xác cao hơn, như trong nghiên cứu [4] đã báo cáo kết quả trích xuất tham chiếu có độ chính xác (accuracy) là 85,61%, độ đo F_1 là 80,06% trên một tập dữ liệu văn bản luật của Nhật bản. Đặc biệt, trong nghiên cứu [5] các tác giả đã xây dựng được mô hình trích xuất tiên tiến, kết hợp Bi-LSTM và CRF cho kết quả độ đo F_1 là 95,35% trên bộ dữ liệu 11.000 câu từ các tài liệu văn bản tiếng Việt.

Nhiệm vụ trích xuất quan hệ của tham chiếu trong văn bản pháp quy hơi khác với bài toán trích xuất quan hệ thực thể thông thường trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Lý do là quan hệ trong bài toán này là quan hệ của thực thể tham chiếu với văn bản pháp quy đang xem xét, trong khi với bài toán xử lý ngôn ngữ tự nhiên thông thường thì quan hệ cần trích xuất là quan hệ giữa các thực thể trong câu đầu vào. Theo khảo sát của nhóm tác giả thì chưa có nghiên cứu nào trước đây thực hiện việc trích xuất quan hệ tham chiếu, ngoài một nghiên cứu đầu tiên về trích xuất quan hệ tham chiếu giữa các thực thể trong văn bản pháp quy tiếng Việt do nhóm chúng tôi thực hiện trong [7]. Nghiên cứu này đã đề xuất phương pháp và xây dựng các thử nghiệm với bộ dữ liệu văn bản pháp quy tiếng Việt gồm 5031 văn bản pháp quy, sử dụng các phương pháp học máy có giám sát cho kết quả bước đầu rất khả quan, đạt độ đo F_1 là 95,57% (với bộ phân loại Máy véc-tơ hỗ trợ - SVM).

Về trích xuất quan hệ nói chung hiện nay phần lớn được tiếp cận theo các phương pháp dựa trên học máy thống kê, với dữ liệu là các văn bản đã được chú thích sẵn thực thể. Với cách tiếp cận này, nhiệm vụ trích xuất quan hệ thường được chuyển thành nhiệm vụ phân loại quan hệ [6, 7, 9, 10]. Như vậy, khi có yêu cầu trích xuất cả hai loại thông tin thực thể và mối quan hệ, thì hầu hết các cách tiếp cận hiện nay là sử dụng phương pháp trích xuất tuần tự, ban đầu là trích xuất thực thể và sau đó là phân loại mối quan hệ giữa các thực thể [11, 12, 13]. Cách thực hiện này có nhược điểm là không thể thực hiện trích xuất đồng thời cả hai loại thông tin, đồng thời việc nhận diện ra được đúng mối quan hệ cũng phụ thuộc vào việc có nhận diện ra được hay không thực thể từ trước. Trong nghiên cứu này, chúng tôi đề xuất một mô hình trích xuất thực thể tham chiếu và quan hệ tham chiếu trong văn bản pháp quy nhằm khắc phục nhược điểm kể trên, sử dụng trích xuất thông tin kết hợp dựa trên cơ chế gán nhãn phân tầng với bộ mã hóa Transformer.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Phần này trình bày mô tả bài toán và đề xuất phương pháp trích xuất kết hợp thực thể và quan hệ giữa các thực thể trong văn bản pháp quy tiếng Việt sử dụng mô hình gán nhãn phân tầng dựa trên kiến trúc bộ mã hóa Transformer. Việc trích xuất các thông tin là thực thể tham chiếu và quan hệ tham chiếu giữa các thực thể trong văn bản pháp quy được gọi chung là trích xuất các đối tượng trong văn bản pháp quy.

A. *Mô tả bài toán*

Giả sử cho một tập văn bản pháp quy ***L***, với mỗi văn bản *T* trong tập văn bản ***L***, thông tin cần trích xuất bao gồm:

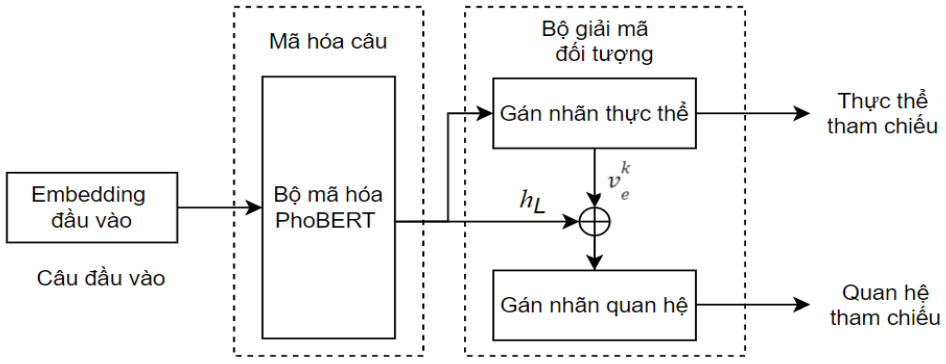
1. Trích xuất thực thể tham chiếu trong văn bản *T*, ký hiệu là *e*. Trong một văn bản *T* có thể có nhiều thực thể tham chiếu *e*.

2. Trích xuất quan hệ giữa (thực thể) văn bản *T* với các thực thể tham chiếu *e* đã trích xuất được ở trên.

Xét mỗi câu *S* trong văn bản *T*. *S* sẽ được sử dụng làm đầu vào cho bài toán. Mỗi câu đầu vào *S* trong một văn bản pháp quy được biểu diễn thành một chuỗi các từ (token) $S=w_1w_2...w_n$, với *n* là số các từ có trong câu. Từ mỗi câu đầu vào *S*, cần trích xuất các đối tượng $O_j=(e,r)$, trong đó *e* là một thực thể tham chiếu trong câu *S* và *r* là mối quan hệ của thực thể tham chiếu *e* với văn bản pháp quy *T* đang xem xét. Về ngữ nghĩa của bài toán trích xuất các đối tượng trong văn bản pháp quy ở đây, thực thể tham chiếu là một chuỗi con của các từ liên tiếp đề cập đến một văn bản quy phạm pháp luật khác, chẳng hạn như *luật*, *nghị định* hoặc *thông tư*. Thực thể tham chiếu *e* trong văn bản pháp quy thường có độ dài lớn hơn nhiều so với các thực thể trong các bài toán trích xuất thực thể NLP thông thường. Trong câu *S*, mỗi tham chiếu *e* chỉ có một quan hệ *r*. Quan hệ này có thể thuộc một trong các loại như "*căn cứ*", "*dẫn chiếu*", "*bị thay thế*", "*hết hiệu lực*", "*được sửa đổi hoặc bổ sung*",...

B. *Mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer*

Mô hình đề xuất cho việc trích xuất kết hợp thực thể và quan hệ của thực thể với văn bản pháp quy đang xem xét là mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer. Chúng tôi mô hình hóa các đối tượng và thiết kế hàm mục tiêu huấn luyện để trích xuất mỗi đối tượng này. Trong kiến trúc này, bộ mã hóa BERT được sử dụng để mã hóa câu đầu vào và bộ giải mã gồm 2 mô-đun: mô-đun gán nhãn thực thể xác định thực thể và mô-đun gán nhãn quan hệ xác định quan hệ của các thực thể. Mô-đun gán nhãn thực thể gồm hai bộ phân lớp nhị phân tương tự nhau với hàm kích hoạt là *sigmoid* để phát hiện vị trí bắt đầu và kết thúc của thực thể tham chiếu, với dữ liệu đầu



Hình 2. Kiến trúc mô hình gán nhãn phân tầng

vào là véc-tơ ẩn của tầng cuối từ bộ mã hóa BERT. Mô-đun gán nhãn quan hệ sử dụng bộ phân lớp *softmax* để xác định kiểu của quan hệ tương ứng với mỗi thực thể tham chiếu đã xác định được, với đầu vào là kết hợp của véc-tơ ẩn của tầng cuối từ bộ mã hóa BERT và đặc trưng tham chiếu. Quá trình huấn luyện mô hình được thực hiện với việc cực đại hóa tổng của hàm *log likelihood* của từng phần gán nhãn thực thể và gán nhãn quan hệ, với Adam stochastic gradient descent [16] qua các *mini-batch* được xáo trộn. Hàm mất mát sử dụng trong mô hình là *cross entropy loss*.

Kiến trúc của mô hình gán nhãn phân tầng dựa trên bộ mã hóa Transformer đề xuất cho bài toán trích xuất thông tin trong văn bản pháp quy tiếng Việt được trình bày trong Hình 2, bao gồm 2 thành phần chính là Bộ mã hóa câu (sử dụng PhoBERT) và Bộ giải mã đối tượng để xác định thực thể và quan hệ tham chiếu. Phần sau sẽ mô tả hai thành phần này.

1) ***Bộ mã hóa***

Bộ mã hóa thực hiện chuyển các câu đầu vào thành các véc-tơ ngữ nghĩa. Ở đây, chúng tôi sử dụng PhoBERT [14] để mã hóa các thông tin ngữ cảnh cho bài toán xử lý văn bản pháp quy tiếng Việt. Kiến trúc của mô hình huấn luyện trước (pretrain model) này tương tự với mô hình BERT, là bộ mã hóa transformer hai chiều L tầng (*L-layer bidirectional Transformer encoder*) [15]. Các véc-tơ ẩn của tầng cuối từ PhoBERT được sử dụng làm biểu diễn chung của mỗi từ (token) trong câu đầu vào S, được ký hiệu là h_L .

2) ***Bộ giải mã***

Bộ giải mã thực hiện dự đoán cặp thực thể tham chiếu và quan hệ tham chiếu từ câu đầu vào. Ý tưởng cơ bản là trích xuất cặp tham chiếu và quan hệ tham chiếu thông qua hai bước liên tiếp. Đầu tiên, các thực thể tham chiếu được xác định từ câu đầu vào. Sau đó với mỗi ứng viên thực thể tham chiếu đã xác định trước đó, tiếp tục xác định quan hệ của tham chiếu đó với câu đầu vào. Các tham chiếu được phát hiện thông qua việc giải mã trực tiếp véc-tơ h_L sinh ra từ bộ giải mã *L-layer* PhoBERT.

Cụ thể, chúng tôi sử dụng hai bộ phân lớp nhị phân tương tự nhau với hàm kích hoạt là *sigmoid* để phát hiện vị trí bắt đầu và kết thúc của thực thể tham chiếu, tương ứng với từ (token) bắt đầu và từ (token) kết thúc được gán bằng 1, và bằng 0 nếu không phải. Giả sử $h_L=(h_1,h_2,...,h_N)$, h_i là biểu diễn mã hóa của từ thứ i trong chuỗi đầu vào. Công thức cụ thể như sau:

$$p_i^{e-start} = \sigma(W_{start} h_i) \quad (1)$$

$$p_i^{e-end} = \sigma(W_{end} h_i) \quad (2)$$

Trong đó, $p_i^{e-start}$ và p_i^{e-end} lần lượt biểu diễn xác suất phát hiện từ thứ i của chuỗi đầu vào và tương ứng chính là vị trí bắt đầu và kết thúc của thực thể. $W(.)$ là ma trận có khả năng học. Từ tương ứng sẽ được gán là nhãn 1 nếu xác suất vượt ngưỡng, ngược lại được gán nhãn là 0.

Tiếp theo là xác định kiểu của quan hệ tham chiếu bằng một bộ phân lớp cho quan hệ của tham chiếu đã trích xuất được ở phần trên. Khác với bộ trích xuất tham chiếu phía trên trực tiếp giải mã véc-tơ h_L , bộ phân loại quan hệ tham chiếu sử dụng thêm cả đặc trưng tham chiếu. Cách xác định như sau:

$$p_i^r = softmax(W_r(h_i + v_e^i)) \quad (3)$$

Trong công thức (3), p_i^r xác định kiểu quan hệ tham chiếu của thực thể, v_e^i thể hiện véc-tơ biểu diễn mã hóa của thực thể thứ i đã xác định trong mô-đun trước đó.

Với mỗi câu đầu vào S, với trong tập huấn luyện $\boldsymbol{L}$ và tập các cặp $\boldsymbol{O}$=*{(e,r)}* trong S, cần cực đại hóa hàm mục tiêu log-likelihood :

$$J(\Theta) = \sum_{e \in T} \log p_\alpha(e|S) + \sum_{r \in T} \log p_\beta(r|e,S) \quad (4)$$

Mô hình được huấn luyện bằng cách cực đại hóa hàm J(Θ) với Adam stochastic gradient descent [16] qua các *mini-batch* được xáo trộn.

IV. TẬP DỮ LIỆU

Chúng tôi sử dụng tập dữ liệu ban đầu đã được xây dựng để thực hiện các thực nghiệm trong chuỗi nghiên cứu của nhóm về văn bản pháp quy [5, 7] để tiếp tục với các thực nghiệm trích xuất kết hợp trong nghiên cứu này. Dữ liệu này được xem xét lại và hiệu chỉnh thêm một lần trước khi thực hiện thực nghiệm trong nghiên cứu này. Phần sau sẽ trình bày tóm tắt lại quá trình xây dựng cũng như các thống kê cụ thể về tập dữ liệu.

Nguồn dữ liệu được thu thập từ Cổng thông tin văn bản quy phạm pháp luật của Nhà nước (http://vbpl.vn), với ba loại văn bản pháp quy phổ biến nhất, bao gồm luật, nghị định và thông tư, sau đó, chọn ngẫu nhiên một tập hợp con trong nguồn này để xây dựng tập dữ liệu.

Bước 1. Tiền xử lý dữ liệu: 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 sử dụng Pyvi (https://github.com/trungtv/pyvi.). Kết quả sau khi tiền xử lý thu được tập dữ liệu gồm 5031 văn bản pháp quy.

Bước 2. Gán nhãn thực thể tham chiếu: bao gồm 2 công đoạn [5]: gán nhãn tự động và gán nhãn thủ công. Việc *gán nhãn tự động* nhằm mục đích làm tăng tốc độ gán nhãn bằng cách sử các biểu thức chính quy, dựa theo một số quan sát trong các văn bản quy phạm pháp luật, ví dụ như: Tham chiếu của văn bản pháp quy thường bắt đầu bằng một từ khóa về loại văn bản pháp quy. Như vậy, có thể xây dựng một từ điển các từ khóa về loại văn bản pháp quy như 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; Mã số văn bản pháp quy (ví dụ 41/2021/TT-BTC);... Loại thực thể được

xác định là từ khóa đầu tiên của tham chiếu văn bản pháp quy. Việc *gán nhãn thủ công* nhằm mục đích kiểm tra và sửa lỗi thủ công các thực thể tham chiếu và loại thực thể đã được gán nhãn ở bước gán nhãn tự động. Kết quả thu được tập dữ liệu đã được gán nhãn thực thể, gồm 61.446 thực thể với 9 loại: Hiến pháp, Bộ luật, Luật, Nghị định, Thông tư, Thông tư liên tịch, Quyết định, Pháp lệnh, Nghị quyết như trong Bảng I. Các loại thực thể tham chiếu có số lượng nhiều nhất là *"luật"* (21.157), *"nghị định"* (22.917), và *"thông tư"* (7.033).

Bảng I. Thống kê số lượng thực thể trong tập dữ liệu

STT	LOẠI THỰC THỂ	SỐ LƯỢNG
1	Hiến pháp	103
2	Bộ luật	960
3	Luật	21.157
4	Nghị định	22.917
5	Thông tư	7.033
6	Thông tư liên tịch	424
7	Quyết định	4.036
8	Pháp lệnh	3.926
9	Nghị quyết	890
Tổng		**61.446**

Bước 3: Gán nhãn quan hệ tham chiếu [7]**.** 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 tham chiếu không có quan hệ với thực thể văn bản đang xét được gán nhãn là "*none*" (được coi là loại quan hệ thứ 7). Tổng cộng có 61.446 quan hệ được gán nhãn cho 7 loại, trong đó hai loại quan hệ có số lượng nhiều nhất là "*dẫn chiếu*" (27.783) và "*căn cứ*" (18.540), như trình bày trong Bảng II.

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

STT	LOẠI QUAN HỆ	NHÃN	SỐ LƯỢNG
1	Căn cứ	CC	18.540
2	Dẫn chiếu	DaC	27.783
3	Hết hiệu lực	HHL	1.618
4	Bị thay thế	BTT	1.765
5	Được sửa đổi hoặc bổ sung	DSD	1.203
6	Được hướng dẫn	DHD	320
7	Không có quan hệ	none	10.217
Tổng			**61.446**

Hình 3 trình bày ví dụ một đoạn văn bản trong "Thông tư 41/2021/TT-BTC" sau khi được gán nhãn thực thể tham chiếu và quan hệ tham chiếu. Các cặp thẻ chứa thực thể tham chiếu: Nghị quyết (*<NQ>,</NQ>*), Quyết định (*<QĐ>,</QĐ>*),…; thuộc tính "*rel*" xác định loại quan hệ: căn cứ "*CC*", dẫn chiếu "*DaC*",… của văn bản đang xem xét (là "Thông tư 41/2021/TT-BTC") với hai thực thể văn bản được tham chiếu trong nội dung (là "Nghị quyết số 53/NQ-CP", và "Quyết định số 779/QĐ-TTg").

Trích "Thông tư số 41/2021/TT-BTC ngày 2 tháng 6 năm 2021 của Bộ tài chính":

Căn cứ **<NQ rel="CC">** Nghị quyết số 53/NQ-CP ngày 26 tháng 5 năm 2021 **</NQ>** của Chính phủ về thành lập Quỹ vắc-xin phòng Covid-19;
Thực hiện **<QĐ rel="CC">** Quyết định số 779/QĐ-TTg ngày 26 tháng 5 năm 2021 **</QĐ>** của Thủ tướng Chính phủ về thành lập Quỹ vắc-xin phòng Covid-19;

Hình 3. Văn bản pháp quy được gán nhãn thực thể tham chiếu và quan hệ tham chiếu

V. CÁC THỰC NGHIỆM VÀ KẾT QUẢ

A. *Thiết lập thực nghiệm*

Hiệu năng của mô hình trích xuất các đối tượng trong văn bản pháp quy được đo bằng độ chính xác (precision), độ bao phủ (recall) và độ đo trung bình điều hòa F_1, theo các công thức như sau:

$$Precision = \frac{|A \cap B|}{|A|} \quad (5)$$

$$Recall = \frac{|A \cap B|}{|B|} \quad (6)$$

và

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

1) *Với các mô hình trích xuất riêng thực thể tham chiếu và quan hệ tham chiếu*, các tham số A và B ở công thức trên được giải thích như sau:

- Với mô hình trích xuất riêng thực thể tham chiếu thì A và B tương ứng là tập các tham chiếu được nhận ra và tập hợp các tham chiếu đúng (được gán nhãn bởi người gán nhãn), của một loại tham chiếu cụ thể (ví dụ như *luật*). Các độ đo độ chính xác, độ bao phủ và F_1 được tính toán cho từng loại tham chiếu và cho tất cả các loại tham chiếu.
- Với mô hình trích xuất quan hệ tham chiếu thì A và B là tập các quan hệ được xác định bởi mô hình và tập các quan hệ đúng (được gán nhãn bởi người gán nhãn) cho một loại quan hệ cụ thể (ví dụ quan hệ "*căn cứ*"). Các độ đo độ chính xác, độ bao phủ và F_1 được tính toán cho từng loại quan hệ.

2) *Với mô hình trích xuất kết hợp thực thể tham chiếu và quan hệ tham chiếu*, chúng tôi thực hiện thử nghiệm như sau: Dữ liệu được chia theo tỷ lệ 3:1:1 cho tập huấn luyện (training set), tập kiểm chứng (validation set) và tập kiểm

tra (test set). Các tham số *A* và *B* ở công thức tương ứng là tập (thực thể, quan hệ) được nhận ra và tập (thực thể, quan hệ) đúng (được gán nhãn bởi người gán nhãn). Ví dụ: ((*Nghị quyết số 53/NQ-CP ngày 26 tháng 5 năm 2021, Nghị quyết), Căn cứ*)

Chúng tôi áp dụng cơ chế *mini-batch* để huấn luyện mô hình đề xuất, trong đó: *batch size* là 6; *learning rate* là $1e^{-5}$; các siêu tham số được xác định trên tập dữ liệu kiểm chứng. Chúng tôi cũng áp dụng cơ chế dừng sớm để ngăn mô hình bị tình trạng quá khớp. Cụ thể, quá trình huấn luyện sẽ dừng khi hiệu suất trên tập dữ liệu kiểm chứng không được cải thiện nào trong ít nhất 5 *epoch* liên tiếp. Số *Transformer block* là 12, với kích thước của véc-tơ trạng thái ẩn h_L là 768. Mô hình BERT đã huấn luyện trước sử dụng cho tiếng Việt trong nghiên cứu này là PhoBERT (base, số lượng tham số là 110M) [14]. PhoBERT được huấn luyện trên khoảng 20GB dữ liệu bao gồm khoảng 1GB Vietnamese Wikipedia corpus và 19GB còn lại lấy từ Vietnamese news corpus. Lượng dữ liệu này là đủ tốt để huấn luyện một mô hình như BERT. PhoBERT sử dụng RDRSegmenter của VNCoreNLP để tách từ tiếng Việt cho dữ liệu đầu vào trước khi qua bộ mã hóa BPE. Độ dài tối đa của câu đầu vào cho mô hình đề xuất được đặt là 60 từ. Ngưỡng xác định cho vị trí bắt đầu và kết thúc của thực thể là không xác định do thực thể tham chiếu có thể ở bất kỳ vị trí nào.

B. *Kết quả thực nghiệm*

Mục đích xây dựng các thực nghiệm:

- Trích xuất riêng các đối tượng trong văn bản pháp quy.
- So sánh kết quả của phương pháp trích xuất kết hợp với kết quả của các phương pháp trích xuất riêng thực thể tham chiếu và quan hệ tham chiếu (đã được thực hiện ở các nghiên cứu trước).
- Trích xuất kết hợp các đối tượng trong văn bản pháp quy

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

1) *Kết quả trích xuất riêng các đối tượng trong văn bản pháp quy*

Thực nghiệm trích xuất riêng các đối tượng trong văn bản pháp quy được thực hiện sử dụng mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer trên tập dữ liệu văn bản pháp quy tiếng Việt. Với thử nghiệm này, chúng tôi trích xuất kết quả đầu ra riêng cho từng đối tượng cần trích xuất, thực thể tham chiếu và quan hệ tham chiếu, để xem xét hiệu quả của mô hình.

Bảng III. Kết quả trích xuất riêng thực thể tham chiếu và quan hệ tham chiếu

Đối tượng trích xuất	Pre. (%)	Rec. (%)	F_1 (%)
Thực thể tham chiếu	98,3	98,7	98,4
Quan hệ tham chiếu	97,3	98,2	97,7

Kết quả đạt được tốt nhất của mô hình đề xuất đối với các nhiệm vụ trích xuất riêng thực thể tham chiếu và quan hệ tham chiếu thể hiện trong Bảng III. Kết quả cho thấy, các độ đo đều đạt trên 96% khi thực hiện trích xuất các đối tượng trong văn bản pháp quy. Cụ thể, khi thực hiện trích xuất riêng thực thể tham chiếu, độ chính xác và độ bao phủ tương đối cao, tương ứng đạt 98,3% và 98,7%. Kết quả cũng tốt tương tự với trích xuất riêng quan hệ tham chiếu, với độ chính xác đạt 97,3% và độ bao phủ đạt 98,2%. Kết quả tính chung theo độ đo F_1 khi thực hiện trích xuất riêng thực thể tham chiếu đạt 98,4%, trích xuất riêng quan hệ tham chiếu đạt 97,7%.

2) *So sánh kết quả của mô hình đề xuất so với kết quả của các nghiên cứu trước*

Bảng IV trình bày kết quả trích xuất thực thể tham chiếu của mô hình đề xuất sử dụng mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer và kết quả trích xuất thực thể tham chiếu đã được thực hiện trong nghiên cứu [5]. Nghiên cứu [5] báo cáo kết quả đạt tốt nhất khi sử dụng mô hình BiLSTM-CRF, là mô hình kết hợp sử dụng mạng bộ nhớ ngắn hạn hai chiều BiLSTM (để học cách biểu diễn từ và câu) kết hợp trường ngẫu nhiên có điều kiện CRF ở lớp suy diễn thay vì sử dụng hàm *softmax* ở lớp này.

Kết quả cho thấy hiệu năng của mô hình trích xuất thực thể tham chiếu đề xuất đạt kết quả tốt hơn khá nhiều so với mô hình đã thực hiện trong nghiên cứu [5]. Cụ thể, mô hình đề xuất có độ chính xác cao hơn 3,2%, độ bao phủ cao hơn 3,1%, và độ đo F_1 cao hơn 3%, so với mô hình sử dụng BiLSTM-CRF.

Bảng IV. Kết quả trích xuất thực thể tham chiếu

BiLSTM-CRF [5]			Mô hình đề xuất		
Pre. (%)	Rec. (%)	F_1 (%)	Pre. (%)	Rec. (%)	F_1 (%)
95,1	95,6	95,4	98,3	98,7	98,4

Bảng V trình bày kết quả trích xuất quan hệ tham chiếu của mô hình đề xuất sử dụng mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer và kết quả trích xuất quan hệ tham chiếu đã được thực hiện trong nghiên cứu [7]. Nghiên cứu [7] báo cáo kết quả đạt tốt nhất khi sử dụng mô hình phân loại Máy véc-tơ hỗ trợ SVM, với các đặc trưng *n*-gram và *TF-IDF*.

Bảng V. Kết quả trích xuất quan hệ tham chiếu

SVM [7]			Mô hình đề xuất		
Pre. (%)	Rec. (%)	F_1 (%)	Pre. (%)	Rec. (%)	F_1 (%)
95,7	95,7	95,6	97,3	98,2	97,7

Kết quả cho thấy hiệu năng của mô hình trích xuất quan hệ tham chiếu đề xuất đạt kết quả tốt hơn so với mô hình đã thực hiện trong nghiên cứu [7]. Cụ thể, mô hình đề xuất có

độ chính xác cao hơn 1,6%, độ bao phủ cao hơn 2,5%, và độ đo F_1 cao hơn 2,1%, so với mô hình sử dụng SVM.

Các kết quả trích xuất thực thể tham chiếu và quan hệ tham chiếu của mô hình đề xuất tốt hơn so với các nghiên cứu trước [5, 7] có thể giải thích được là do mô hình kết hợp tận dụng được mối tương quan khi học đặc trưng, bao gồm cả các đặc trưng về thực thể và các đặc trưng về mối quan hệ. Hơn nữa, việc nhận dạng mối quan hệ tham chiếu không phụ thuộc vào việc có nhận ra được đúng hay không thực thể tham chiếu như khi thực hiện theo phương pháp tuần tự.

3) *Trích xuất kết hợp các đối tượng trong văn bản pháp quy*

Bảng VI. Kết quả trích xuất kết hợp thực thể tham chiếu và quan hệ tham chiếu

Đối tượng trích xuất	Pre. (%)	Rec. (%)	F_1 (%)
Thực thể tham chiếu và quan hệ tham chiếu	96,4	97,1	96,8

Thực nghiệm trích xuất kết hợp các đối tượng trong văn bản pháp quy được thực hiện sử dụng mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer trên tập dữ liệu văn bản pháp quy tiếng Việt. Với thử nghiệm này, chúng tôi trích xuất kết quả đầu ra đồng thời với cả hai đối tượng cần trích xuất, thực thể tham chiếu và quan hệ tham chiếu, để xem xét hiệu quả của mô hình (Bảng VI).

Khi thực hiện trích xuất kết hợp cho ra kết quả đồng thời cả hai thông tin thực thể tham chiếu và quan hệ tham chiếu, kết quả rất khả quan với độ chính xác đạt 96,4% và độ bao phủ đạt 97,1%. Kết quả tính chung theo độ đo F_1 khi trích xuất kết hợp cả hai thông tin thực thể tham chiếu và quan hệ tham chiếu đạt 96,8%. Kết quả này cũng tốt hơn so với kết quả trích xuất riêng từng đối tượng trong các nghiên cứu [5] và [7]. Điều này có thể giải thích là do mô hình trích xuất kết hợp học được thêm thông tin qua sự tương tác giữa nhận dạng thực thể và phân loại mối quan hệ, khiến cho khả năng tối ưu hóa cho nhiệm vụ kết hợp hiệu quả hơn khi so sánh với việc thực hiện các nhiệm vụ riêng biệt.

C. *Phân tích lỗi*

Kết quả trích xuất thực thể tham chiếu và quan hệ tham chiếu sử dụng mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer rất khả quan với các độ đo hiệu năng đều đạt trên 96,4%, trong cả trường hợp trích xuất thông tin riêng và trường hợp trích xuất thông tin kết hợp. Phần sau sẽ trình bày một số lỗi còn tồn tại trong quá trình thực hiện thực nghiệm mô hình, bao gồm các lỗi như sau đây.

1) *Trích xuất sai hoặc không trích xuất ra được thực thể tham chiếu*

Trích xuất thực thể thiếu thông tin ngày tháng. Ví dụ: *"Thông_tư này quy_định chi_tiết và hướng_dẫn thi_hành một_số điều của Nghị_định số 30/2012/NĐ-CP ngày 12 tháng 4 năm 2012 của Chính_phủ về tổ_chức , hoạt_động của quỹ xã_hội , quỹ từ_thiện (sau đây gọi là Nghị_định số 30/2012/NĐ-CP"*, thực thể tham chiếu xác định được là "*Nghị_định số 30/2012/NĐ-CP"*.

Trích xuất thiếu thực thể tham chiếu xảy ra trong trong một số ít câu kiểm tra, trong các câu có 3 thực thể tham chiếu trở lên.

2) *Trích xuất sai quan hệ tham chiếu*

Một số mẫu bị nhận nhầm lẫn giữa các nhãn loại quan hệ tham chiếu: none (*không có quan hệ*), DHD (*được hướng dẫn*), DaC (*dẫn chiếu*), DSD (*được sửa đổi hoặc bổ sung*). Ví dụ: *"Nghị_định này quy_định chi_tiết thi_hành một_số điều của Luật sửa_đổi , bổ_sung một_số điều của Luật Kinh_doanh bảo_hiểm và sửa_đổi , bổ_sung một_số điều của Nghị_định số 45/2007/NĐ-CP ngày 27 tháng 3 năm 2007 của Chính_phủ quy_định chi_tiết thi_hành một_số điều của Luật Kinh_doanh bảo_hiểm (sau đây gọi tắt là Nghị_định 45/2007/NĐ-CP"*. Quan hệ tham chiếu đúng (đã được gán nhãn) là DSD, nhưng mô hình lại xác định quan hệ tham chiếu là DaC. Tuy nhiên, trong thực tế con người cũng có thể bị mắc phải các nhầm lẫn này do các mối quan hệ được hướng dẫn, dẫn chiếu và được sửa đổi hoặc bổ sung dễ lẫn sang nhau nếu không chú ý cẩn thận.

3) *Trích xuất sai thực thể tham chiếu dẫn đến trích xuất sai quan hệ tham chiếu*

Trong câu dài hoặc có nhiều thực thể tham chiếu, thì mô hình có thể trích xuất sai hoặc thiếu thực thể tham chiếu, dẫn đến trích xuất sai thông tin quan hệ tham chiếu. Ví dụ: *"Đối_với dự_án đầu_tư mà tại một trong ba loại giấy_tờ sau đây : Giấy chứng_nhận đầu_tư (Giấy_phép đầu_tư) hoặc Quyết_định cho thuê đất hoặc Hợp_đồng thuê đất được cấp (được ký_kết) theo quy_định của Luật_Đầu_tư nước_ngoài , Luật_Đầu_tư trong nước và pháp_luật có liên_quan có quy_định đơn_giá thuê đất , thuê mặt_nước và nguyên_tắc điều_chỉnh đơn_giá thuê theo các quy_định về đơn_giá cho thuê đất , thuê mặt_nước của Bộ Tài_chính (Quyết_định số 210A-TC/VP ngày 01 tháng 4 năm 1990 , Quyết_định số 1417TC/TCĐN ngày 30 tháng 12 năm 1994 , Quyết_định số 179/1998/QĐ-BTC ngày 24 tháng 02 năm 1998 , Quyết_định số 189/2000/QĐ-BTC ngày 24 tháng 11 năm 2000 , Quyết_định số 1357TC/QĐ-TCT ngày 30 tháng 12 năm 1995) thì được :"*. Đa phần các lỗi này xảy ra trong các câu có số lượng thực thể tham chiếu nhiều hơn hoặc bằng 5, các thực thể được mô tả liên tiếp nhau và được biểu diễn dưới nhiều định dạng khác nhau.

VI. KẾT LUẬN

Bài báo đã trình bày một đề xuất nghiên cứu cải tiến cho bài toán trích xuất đồng thời thực thể tham chiếu và quan hệ tham chiếu trong văn bản pháp quy sử dụng mô hình gán nhãn phân tầng dựa trên kiến trúc của bộ mã hóa Transformer. Mô hình đề xuất không những có thể thực hiện trích xuất riêng thực thể tham chiếu và quan hệ tham chiếu đạt hiệu suất cao, mà còn có thể trích xuất kết hợp đồng thời cả hai thông tin này. Các thực nghiệm được tiến hành trên tập dữ liệu 5031 văn bản pháp quy tiếng Việt đã được gán nhãn cho kết quả trích xuất thông tin rất tốt, với trích xuất riêng thực thể tham chiếu đạt 98,4%, trích xuất riêng quan

hệ tham chiếu đạt 97,7%, trích xuất kết hợp cả hai thông tin đạt 96,8%, tính theo độ đo F_1.

Trong thời gian tới, chúng tôi dự định nghiên cứu giải quyết bài toán này dựa trên cơ chế attention nhằm xây dựng các biểu diễn câu cụ thể cho từng quan hệ, từ đó nâng cao độ chính xác khi trích xuất kết hợp cặp thực thể tham chiếu và quan hệ tham chiếu trong câu.

LỜI CẢM ƠN

Nghiên cứu sinh Nguyễn Thị Thanh Thủy được tài trợ bởi Tập đoàn Vingroup – Công ty CP và hỗ trợ bởi chương trình học bổng đào tạo thạc sĩ, tiến sĩ trong nước của Quỹ Đổi mới sáng tạo Vingroup (VINIF), Viện Nghiên cứu Dữ liệu lớn (VinBigdata), mã số VINIF.2020.TS.94.

TÀI LIỆU THAM KHẢO

[1] De Maat, Emile, Radboud Winkels, and Tom Van Engers. "Automated Detection of Reference." In Legal Knowledge and Information Systems: JURIX 2006: the Nineteenth Annual Conference, vol. 152, p. 41. IOS Press, 2006.

[2] Martínez-González, Mercedes, Pablo de la Fuente, and Dámaso-Javier Vicente. "Reference extraction and resolution for legal texts." In International Conference on Pattern Recognition and Machine Intelligence, pp. 218-221. Springer, Berlin, Heidelberg, 2005.

[3] Palmirani, Monica, Raffaella Brighi, and Matteo Massini. "Automated extraction of normative references in legal texts." In Proceedings of the 9th international conference on Artificial intelligence and law, pp. 105-106. 2003.

[4] Tran, Oanh Thi, Bach Xuan Ngo, Minh Le Nguyen, and Akira Shimazu. "Automated reference resolution in legal texts." Artificial intelligence and law 22, no. 1 (2014): 29-60.

[5] Ngo Xuan Bach, Nguyen Thi Thanh Thuy, Dang Bao Chien, Trieu Khuong Duy, To Minh Hien, and Tu Minh Phuong. "Reference extraction from Vietnamese legal documents." In Proceedings of the Tenth International Symposium on Information and Communication Technology, pp. 486-493. 2019.

[6] Kambhatla, Nanda. "Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction." In Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 178-181. 2004.

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

[8] Wei, Zhepei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. "A novel cascade binary tagging framework for relational triple extraction." arXiv preprint arXiv:1909.03227. 2019.

[9] Zeng, Daojian, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun 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.

[10] Xu, Yan, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. "Classifying relations via long short term memory networks along shortest dependency paths." In Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 1785-1794. 2015.

[11] Chan, Yee Seng, and Dan Roth. "Exploiting syntactico-semantic structures for relation extraction." In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 551-560. 2011.

[12] Zhang, Meishan, Yue Zhang, and Guohong Fu. "End-to-end neural relation extraction with global optimization." In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1730-1740. 2017.

[13] Miwa, Makoto, and Mohit Bansal. "End-to-end relation extraction using lstms on sequences and tree structures." arXiv preprint arXiv:1601.00770. 2016.

[14] Nguyen, Dat Quoc, and Anh Tuan Nguyen. "PhoBERT: Pre-trained language models for Vietnamese." arXiv preprint arXiv:2003.00744. 2020.

[15] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In Advances in neural information processing systems, pp. 5998-6008. 2017.

[16] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980. 2014.

A JOINT REFERENCE ENTITY AND RELATION EXTRACTION METHOD FOR LEGAL DOCUMENT

Abstract: One of the most important tasks that need to be done first to build automated legal document processing systems, such as searching, querying, analyzing, or question-answering, is extracting the necessary information in legal documents, including reference entities and reference relations. When there is a need of extracting both information of reference entity and reference relation, or extract only reference relations, previous studies would usually do this work in the sequential way, first extracting entities, and then extracting relations. Thus, the accuracy of reference relation extraction often depends on whether the reference entities are correctly extracted or not. In this paper, we present an novel method to solve the problem of extracting information in legal documents, that is joint reference entity and relation extraction for legal document using cascade tagging model based on Transformer encoder architecture. The results show that the proposed model achieves quite high accuracy, with the F_1 score up to 96.8% for the joint reference entity and relation extraction. The individual extraction results are outperform compared to previous studies: the reference entity extraction achieved the F_1 score of 98.4%, and the reference relation extraction achieved the F_1 score of 97.7%, on a dataset of 5031 Vietnamese legal documents.

Keywords: Legal document, reference entity extraction, reference relation extraction, joint entity and relation extraction.

Nguyễn Thị Thanh Thủy. Nhận học vị Thạc sĩ năm 2009. 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.

Nguyễn Ngọc Điệp. Nhận học vị Tiến sĩ năm 2017. 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, an toàn thông tin, xử lý ngôn ngữ tự nhiên.