HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

PHÍ MẠNH KIÊN

TÌM KIẾM VĂN BẢN PHÁP QUY SỬ DỤNG KỸ THUẬT HỌC SÂU

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

HÀ NỘI - 2020

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: GS. TS. TỪ MINH PHƯƠNG

Phản biện 1: ……

Phản biện 2: ……

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ....giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Ngày nay, trong kỉ nguyên kỹ thuật số, với sự bùng nổ của thông tin, số lượng các tài liệu điện tử do con người tạo ra ngày càng khổng lồ. Trong quá trình học tập, nghiên cứu hay làm việc, chúng ta cần tìm kiếm và đọc rất nhiều tài liệu để tìm được thông tin ta mong muốn. Việc này đôi khi mất nhiều thời gian, điển hình là trong lĩnh vực pháp luật. Một văn bản pháp luật thường có thể dài tới 15-20 trang hoặc thậm chí nhiều hơn. Một vụ việc có thể liên quan đến nhiều văn bản khác nhau. Các luật sư, nhân viên pháp lý... phải đọc rất nhiều văn bản và so sánh các điều, khoản trong đó với trường hợp đang xử lý. Theo một khảo sát năm 2013 tại Mỹ [19], trung bình, gần 47,3% số người được hỏi dành 15% thời gian, 36.6% số người dành 15-50% thời gian, 10.3% số người dành từ 50% thời gian trở lên mỗi tuần cho việc tìm kiếm và nghiên cứu văn bản pháp luật. Đây là một vấn đề thực tiễn, mang lại giá trị mà chúng ta cần giải quyết.

Bài toán tìm kiếm thông tin ra đời chính là để xử lý vấn đề trên. Nhiệm vụ chính của bài toán tìm kiếm thông tin là tìm kiếm các thông tin thoả mãn nhu cầu thông tin của người dùng. Người sử dụng của một hệ thống tìm kiếm thông tin không chỉ muốn tìm những văn bản có chứa những từ khóa trong câu truy vấn mà còn quan tâm tới việc thu nhận được những văn bản mang lại thông tin phù hợp với mục đích tìm kiếm.

Các hệ thống tìm kiếm thông tin thường biểu diễn văn bản và câu truy vấn dưới dạng các véc-tơ. Chất lượng biểu diễn văn bản và so sánh các véc-tơ biểu diễn có ảnh hưởng quan trọng tới kết quả. Gần đây, các kỹ thuật sử dụng học sâu cho thấy khả năng biểu diễn văn bản rất tốt trong xử lý ngôn ngữ tự nhiên nói chung và tìm kiếm thông tin văn bản nói riêng. Vì vậy, tôi chọn đề tài **"Tìm kiếm văn bản pháp quy sử dụng kỹ thuật học sâu"** cho luận văn của mình. Mục tiêu của luận văn là tìm hiểu các phương pháp biểu diễn văn bản và đề xuất mô hình sử dụng kỹ thuật học sâu ứng dụng trong tìm kiếm văn bản pháp quy tiếng Việt. Đầu vào của hệ thống là một câu hỏi về pháp luật. Đầu ra của hệ thống là văn bản pháp quy có liên quan, trả lời

được cho câu hỏi đó, cụ thể đến mức điều. Ví dụ, với câu hỏi "*Vợ chồng ly hôn tài sản chung được phân chia như thế nào?*" hệ thống sẽ trả về kết quả là: *Điều 59 Luật Hôn nhân và gia đình, Điều 7 Thông tư liên tịch hướng dẫn một số quy định của Luật Hôn nhân và gia đình*.

Nội dung luận văn được chia thành 3 chương như sau:

- **CHƯƠNG 1:** Bài toán tìm kiếm thông tin và các phương pháp biểu diễn văn bản: Trình bày tổng quan về bài toán tìm kiếm thông tin và các phương pháp biểu diễn văn bản phục vụ tìm kiếm, tìm kiếm thông tin.
- **CHƯƠNG 2:** Ứng dụng biểu diễn văn bản bằng mạng nơ-ron sâu trong tìm kiếm văn bản pháp quy: Giới thiệu về bài toán tìm kiếm văn bản pháp quy, trình bày phương pháp biểu diễn văn bản sử dụng mạng nơ-ron sâu.
- **CHƯƠNG 3:** Thử nghiệm và đánh giá: Mô tả quá trình xây dựng bộ dữ liệu và so sánh, đánh giá hiệu quả của mô hình đề xuất so với các phương pháp khác.

Các kết quả của luận văn đã được chấp nhận công bố tại hội nghị COLING 2020, hội nghị hạng A về xử lý ngôn ngữ tự nhiên.

CHƯƠNG 1. BÀI TOÁN TÌM KIẾM THÔNG TIN VÀ CÁC PHƯƠNG PHÁP BIỂU DIỄN VĂN BẢN

Chương này sẽ trình bày tổng quan về bài toán tìm kiếm thông tin nói chung và bài toán tìm kiếm văn bản pháp quy nói riêng, bao gồm khái niệm, kiến trúc hệ thống và mô hình tìm kiếm thông tin, cùng với các phương pháp biểu diễn văn bản phục vụ tìm kiếm.

1.1. Bài toán tìm kiếm thông tin

1.1.1. Tìm kiếm văn bản quy phạm pháp luật

Bài toán tìm kiếm thông tin

Input:

- Một tập tài *liệu lớn, ổn định.*
- Một nhu cầu thông tin thể hiện dưới dạng câu truy vấn (các từ khoá hoặc câu hỏi).

Output:

- Tìm tất cả tài liệu có *liên quan* đến câu truy vấn.

Những vấn đề cần giải quyết của bài toán tìm kiếm thông tin

- Biểu diễn tập tài liệu như thế nào?
- Biểu diễn nhu cầu thông tin của người dùng như thế nào?
- Bằng cách nào hệ thống có thể trả về những tài liệu có liên quan đến nhu cầu thông tin một cách có hiệu quả?
- Kết quả trả về được trình bày như thế nào?

Bài toán tìm kiếm văn bản pháp quy

- Đầu vào: Truy vấn của người dùng dưới dạng một câu hỏi.
- Đầu ra: Các điều khoản có liên quan, giúp trả lời được cho câu hỏi của người dùng.

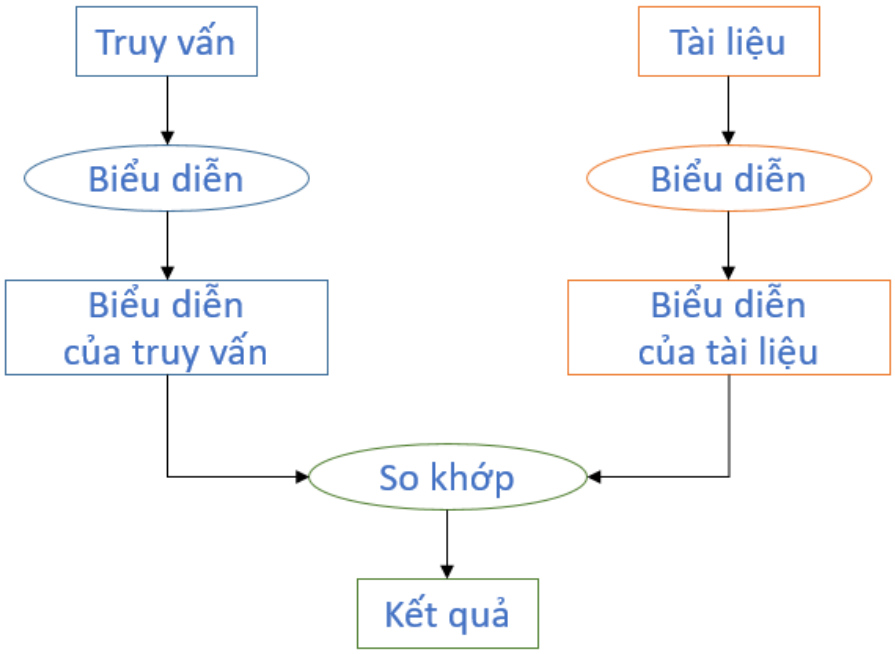
Ví dụ minh họa đầu vào và đầu ra của bài toán được mô tả bằng bảng bên dưới:

Bảng 1.1. Ví dụ minh họa bài toán tìm kiếm văn bản pháp quy.

Câu hỏi đầu vào	Con riêng có quyền hưởng thừa kế của bố đã mất không di chúc không?
Đầu ra	Điều 651 Bộ luật dân sự 2015
Nội dung điều luật	**Điều 651. Người thừa kế theo pháp luật** 1. Những người thừa kế theo pháp luật được quy định theo thứ tự sau đây: a) Hàng thừa kế thứ nhất gồm: vợ, chồng, cha đẻ, mẹ đẻ, cha nuôi, mẹ nuôi, con đẻ, con nuôi của người chết; b) Hàng thừa kế thứ hai gồm: ông nội, bà nội, ông ngoại, bà ngoại, anh ruột, chị ruột, em ruột của người chết; cháu ruột của người chết mà người chết là ông nội, bà nội, ông ngoại, bà ngoại; c) Hàng thừa kế thứ ba gồm: cụ nội, cụ ngoại của người chết; bác ruột, chú ruột, cậu ruột, cô ruột, dì ruột của người chết; cháu ruột của người chết mà người chết là bác ruột, chú ruột, cậu ruột, cô ruột, dì ruột; chắt ruột của người chết mà người chết là cụ nội, cụ ngoại. 2. Những người thừa kế cùng hàng được hưởng phần di sản bằng nhau. 3. Những người ở hàng thừa kế sau chỉ được hưởng thừa kế, nếu không còn ai ở hàng thừa kế trước do đã chết, không có quyền hưởng di sản, bị truất quyền hưởng di sản hoặc từ chối nhận di sản.

1.1.2. Hệ thống tìm kiếm và tìm kiếm thông tin

Hoạt động của một hệ thống tìm kiếm thông tin được mô tả trong Hình 1.1, bao gồm ba bước chính: biểu diễn văn bản, biểu diễn truy vấn và so khớp – đánh giá độ liên quan giữa văn bản và truy vấn.



Hình 1.1. Kiến trúc tổng quan của hệ thống tìm kiếm thông tin.

1.2. Biểu diễn văn bản sử dụng từ khóa

1.2.1. TF-IDF

Term Frequency – Inverse Document Frequency (TF-IDF), là một thống kê số học phản ánh tầm quan trong của một từ (word) với một văn bản (document) trong tập các văn bản (corpus). Nó thường được dùng để làm trọng số trong việc thu thập thông tin và khai phá văn bản.

Các tham số trong TF-IDF:

- Term frequency: Tần số xuất hiện
- Inverse document frequency: Tần số nghịch đảo văn bản
- Document Length: Độ dài văn bản

$$tf - idf(t,d) = tf(t) \times idf(t,d) \times norm(d)$$

1.2.2. BM25

Term frequency trong BM25

Đối với TF-IDF, giá trị của nó sẽ tăng vô hạn khi TF tăng lên. Để giảm tác động của TF thì BM25 đã chỉnh sửa công thức của TF lại.

Độ dài văn bản trong BM25

Công thức của TF-IDF chưa thực sự hoàn chỉnh, nó đúng với những văn bản có độ dài trung bình trong toàn bộ tập dữ liệu. Nếu độ dài văn bản quá ngắn hoặc quá dài so với độ dài trung bình, thì công thức trên sẽ cho kết quả thiếu chính xác.

Bởi vậy, người ta thêm vào trong công thức trên 2 tham số, một hằng số b và một giá trị độ dài L, công thức sẽ trở thành:

$$\frac{(k+1) \times tf}{k \times (1.0 - b + b \times L) + tf}$$

Inverse Document Frequency trong BM25

Biểu đồ Hình 1.3 cho thấy IDF trong BM25 khá giống IDF trong TF-IDF. Tuy nhiên BM25 đã chỉnh sửa công thức tính lại để thêm khả năng đưa ra điểm âm khi tần suất xuất hiện của từ trên toàn bộ tập văn bản rất cao.

$$idf_t = log \frac{1 + (D - d + 0.5)}{d + 0.5}$$

Trong đó:

- D: tổng số văn bản
- d: số lượng văn bản chứa từ t

1.3. Biểu diễn văn bản sử dụng chủ đề ẩn

1.3.1. Khái niệm mô hình Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation [4] [5] là một trong những mô hình phát hiện chủ đề ẩn thành công nhất hiện nay được phát triển bởi David Blei, Andrew Ng và Michael Jordan.

LDA là một mô hình tự sinh xác suất cho các bộ dữ liệu rời rạc như bộ văn bản ký tự. Bản chất của LDA là một mô hình Bayes phân cấp với 3 mức, trong đó

mỗi một phần tử của bộ dữ liệu là một tập hợp hữu hạn của một tập các chủ đề nằm ẩn bên trong nội dung nhìn thấy được của thành phần đó.

1.3.2. Tổng quan về mô hình sinh trong LDA

Với một tập các văn bản (*corpus*) gồm M văn bản ký hiệu là $D = \{d_1, d_2, \dots, d_M\}$, trong đó văn bản thứ m trong tập văn bản sẽ có N_m từ, các từ trong văn bản sẽ được lấy từ tập từ vựng của các thuật ngữ (term) $= \{t_1, t_2, \dots, t_V\}$,. Mục đích của LDA là tìm ra cấu trúc ẩn của các chủ đề (topic) hay các lĩnh vực (concept) trong văn bản.

Quá trình sinh trong LDA được mô tả như sau: LDA sinh ra một luồng các từ quan sát được $w_{m,n}$ (là các từ có trong nội dung văn bản), được phân chia thành các văn bản. Với mỗi văn bản, một tỷ lệ chủ đề $\overrightarrow{\vartheta_m}$ sẽ được đưa ra, và từ đó, các từ đặc tả chủ đề được tạo ra. Nghĩa là, với mỗi từ, một chỉ số chỉ thi chủ đề $z_{m,n}$ được lấy mẫu theo các văn bản – tỷ lệ trộn cụ thể, và sau đó phân phối chủ đề tương ứng $\vec{\varphi}_{z_{m,n}}$ được sử dụng để sinh ra các từ. Các chủ đề $\vec{\varphi}_k$ sẽ được lấy mẫu một lần cho mọi văn bản trong tập văn bản D.

1.3.3. Suy luận

Với một mô hình LDA đã cho, có thể thực hiện suy luận ra các chủ đề có trong một văn bản mới chưa có trong tập văn bản huấn huyện bằng một tiến trình lấy mẫu tương tự.

Nhiệm vụ cụ thể của việc suy luận này là từ một văn bản mới $\tilde{\boldsymbol{m}}$, được biểu diễn bởi một véc-tơ các từ $\overrightarrow{\boldsymbol{w}}$, chúng ta phải đi ước lượng các xác suất hậu nghiệm của các chủ đề $\tilde{\vec{\boldsymbol{z}}}$ cho bởi véc-tơ các từ của câu truy vấn $\overrightarrow{\boldsymbol{w}}$ và mô hình LDA đã cho trước $\boldsymbol{L}(\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{\Phi}})$.

1.4. Biểu diễn văn bản sử dụng véc-tơ từ

1.4.1. Giới thiệu

Phương pháp biểu diễn văn bản bằng véc-tơ từ, hay biểu diễn bằng từ khóa phân tán, biểu diễn các từ dưới dạng véc-tơ có số chiều cố định và nhỏ hơn nhiều so với kích thước từ vựng. Giá trị của mỗi thành phần trong véc-tơ biểu diễn đều là số thực và có giá trị và thường khác 0 (không chỉ là 0 hay 1 như one-hot), do vậy cách biểu diễn này còn được gọi là biểu diễn đặc (dense) khác với biểu diễn thưa (sparse) kiểu one-hot.

Mô hình này hướng đến việc phân tích ngữ nghĩa của từ và biểu diễn quan hệ giữa các từ thông qua véc-tơ biểu diễn của chúng.

1.4.2. Các bước thực hiện

Cách biểu diễn của từ trong phương pháp này thu được thông qua tiến hành học máy (không giám sát) trên các mô hình ngôn ngữ mạng nơ-ron nhân tạo [21] hoặc các mô hình giảm số chiều khác [24]. Người ta đưa vào mạng nơ-ron một tập dữ liệu huấn luyện lớn có độ bao quát rộng để xác định trọng số thích hợp nhất của các nơ-ron trong mạng. Cuối quá trình huấn luyện, sau khi đã xác định trọng số người ta đưa từng từ vào đầu vào của mạng và lấy kết quả là biểu diễn dạng véc-tơ của từ ở đầu ra.

Có 2 thuật toán học máy thường dùng trong việc học các biểu diễn từ của máy là CBOW (continuous bag of words) và Skip-gram.

Ngoại trừ hai thuật toán nói trên, gần đây hơn (2014) nhóm nghiên cứu của đại học Stanford cũng giới thiệu thuật toán học máy GloVe (Global Vector) [24] cho phép đạt được véc-tơ từ với độ chính xác tốt hơn.

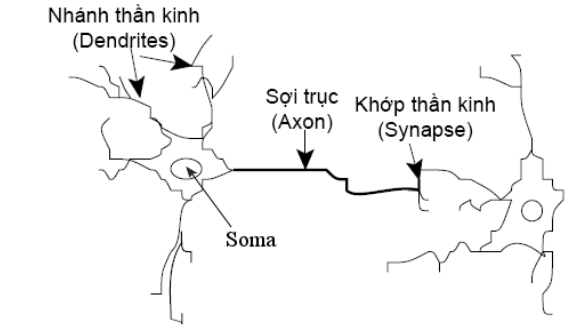
1.5. Biểu diễn văn bản sử dụng mạng nơ-ron sâu

1.5.1. Giới thiệu về mạng nơ-ron nhân tạo

Mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) là mô hình xử lý thông tin được mô phỏng dựa trên hoạt động của hệ thống thần kinh của sinh vật, bao gồm số lượng lớn các nơ-ron được gắn kết để xử lý thông tin. ANN giống như bộ não con người, được học bởi kinh nghiệm (thông qua huấn luyện), có khả năng lưu giữ

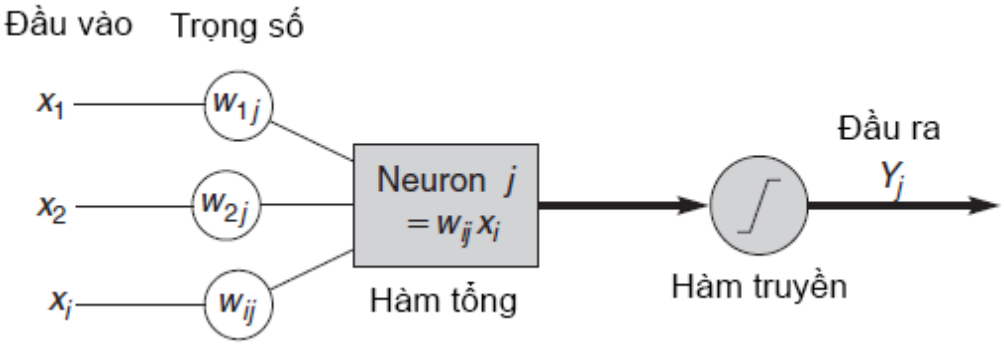
những kinh nghiệm hiểu biết (tri thức) và sử dụng những tri thức đó trong việc dự đoán các dữ liệu chưa biết (unseen data).

1.5.2. Cấu trúc và mô hình của một nơ-ron nhân tạo



Hình 1.2. Mô hình một nơ-ron sinh học.

Mạng nơ-ron nhân tạo được lấy cảm hứng từ cách làm việc của bộ não con người. Các nơ-ron nhân tạo mô phỏng lại hoạt động của nơ-ron sinh học



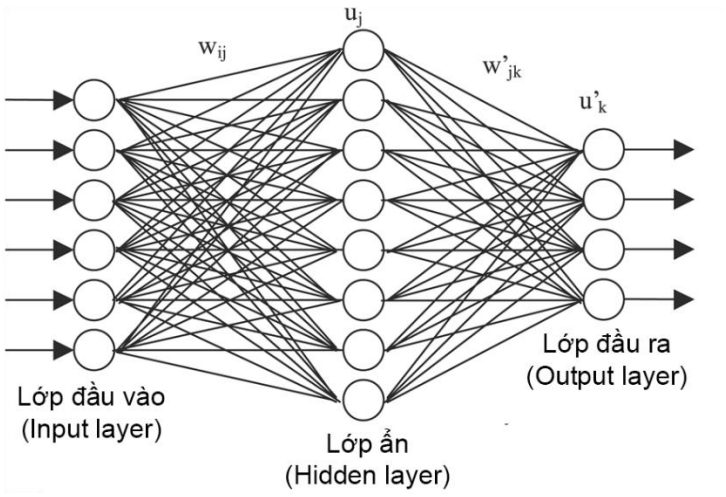
Hình 1.3. Mô hình một nơ-ron nhân tạo.

Tương tự như nơ-ron sinh học, nơ-ron nhân tạo cũng nhận các tín hiệu đầu vào, xử lý (nhân các tín hiệu này với trọng số liên kết, tính tổng các tích thu được rồi gửi kết quả đến hàm truyền) và cho một tín hiệu đầu ra (là kết quả của hàm lan truyền).

1.5.3. Cấu tạo và phương thức làm việc của mạng nơ-ron

Khi liên kết các đầu vào, đầu ra của nhiều nơ-ron với nhau, ta sẽ thu được một mạng nơ-ron. Việc ghép nối các nơ-ron trong mạng với nhau có thể theo nguyên tắc bất kỳ.

Nguyên lý cấu tạo chung của mạng nơ-ron gồm nhiều lớp, mỗi lớp bao gồm nhiều nơ-ron có cùng chức năng trong mạng. Thông thường một mạng nơ-ron sẽ bao gồm: lớp đầu vào (input layer), lớp ẩn (hidden layer) và lớp đầu ra (output layer). Trong đó có thể có nhiều lớp ẩn.



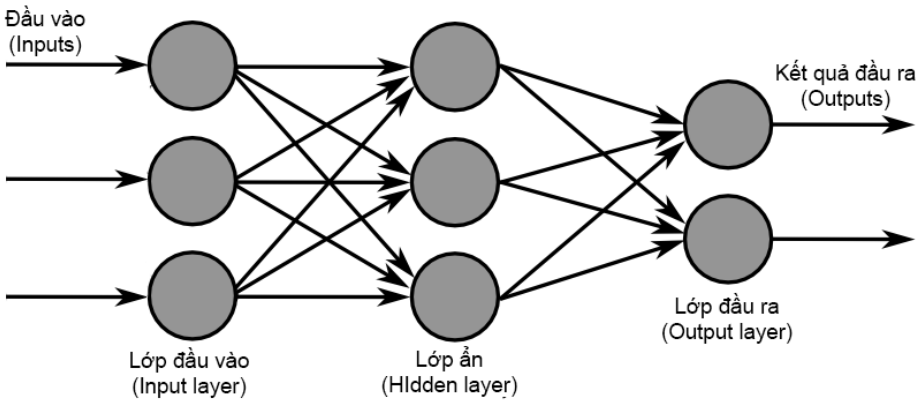
Hình 1.4. Mô hình cấu tạo của một mạng nơ-ron cơ bản.

Khi mới được hình thành thì mạng nơ-ron chưa có tri thức, tri thức của mạng sẽ được hình thành dần dần sau một quá trình học.

1.5.4. Phân loại mạng nơ-ron

1.5.4.1. *Mạng nơ-ron truyền thẳng (Feed-forward Neural Network - FNN)*

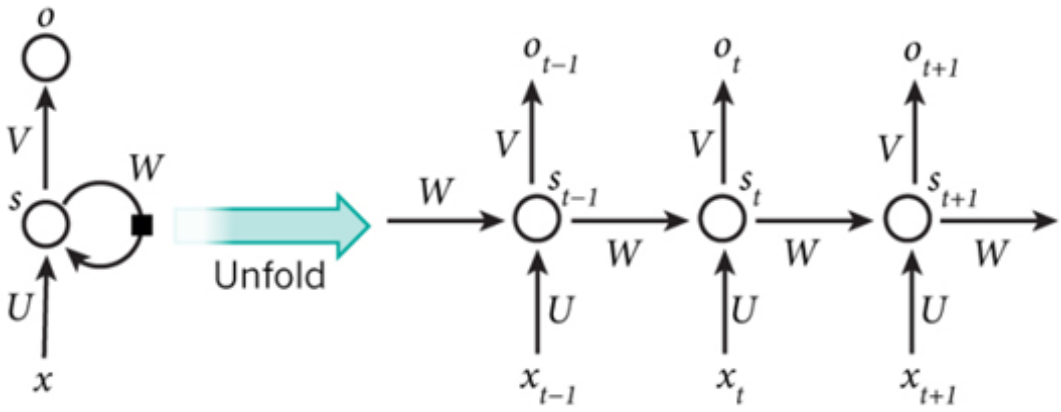
Mạng nơ-ron truyền thẳng là kiến trúc mạng nơ-ron được sử dụng phổ biến. Đúng như tên của nó, các giá trị sẽ đi thẳng từ lớp đầu vào tới lớp đầu ra chứ không có chiều quay ngược lại (khác với mạng nơ-ron hồi quy được trình bày ở phần sau).



Hình 1.5. Mô hình mạng nơ-ro truyền thẳng.

1.5.4.2. *Mạng nơ-ron hồi quy (Recurrent Neural Networks – RNN)*

Ý tưởng về mạng nơ-ron hồi quy xuất phát từ mục đích muốn chuyển hóa một chuỗi các đầu vào thành chuỗi kết quả đầu ra, trong đó các thành phần trong chuỗi đều ảnh hưởng tới nhau. Ví dụ đối với bài toán chat bot, đầu vào là một câu (gồm nhiều từ và rõ ràng các từ phải liên quan tới nhau), mỗi từ được biểu diễn bằng một véc-tơ và ta mong muốn sử dụng mạng nơ-ron để ghi nhớ ngữ nghĩa của câu đó. Mạng nơ-ron truyền thẳng - FNN đã đề cập ở trên không thể làm được điều này vì đầu vào của FNN chỉ là một bản ghi và các bản ghi khác nhau hoàn toàn không ảnh hưởng lẫn nhau. Nhưng mạng nơ-ron hồi quy có thể làm được điều này.



Hình 1.6. Mô hình mạng nơ-ron hồi quy.

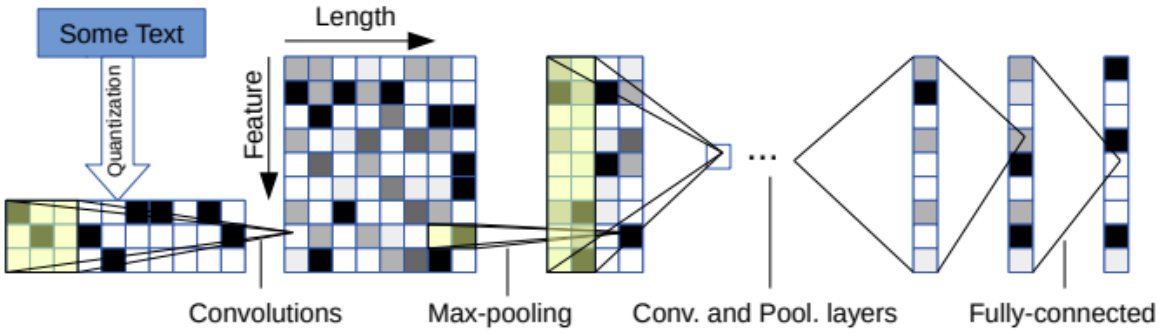
1.5.5. Các mạng nơ-ron sâu

Mạng nơ-ron sâu Deep Neural Networks - DNN) là một mạng nơ-ron nhân tạo với nhiều lớp ẩn ẩn giữa lớp đầu vào và lớp đầu ra. Các mạng nơ-ron sâu có thể mô hình mối quan hệ phi tuyến tính phức tạp.

Mạng nơ-ron nhân chập

Mạng nơ-ron nhân chập là một dạng đặc biệt của mạng nơ-ron nhiều lớp. Trong mạng các lớp nhân chập (convolution layer) kết hợp với các hàm kích hoạt phi tuyến (nonlinear activation function) như ReLU hay tanh để tạo ra thông tin trừu tượng hơn cho các lớp tiếp theo [2] [23].

CNN được áp dụng trong các tác vụ như phân loại câu [14] [13] [31], phân tích cảm xúc, quan điểm [22], tìm kiếm theo ngữ nghĩa [10] [25], nhận dạng tiếng nói [1].



Hình 1.7. Mô hình CNN trong nghiên cứu [31].

1.5.6. Biểu diễn văn bản sử dụng mạng nơ-ron

Nhiều nghiên cứu đã sử dụng mạng nơ-ron để biểu diễn văn bản và thu được kết quả khả quan.

1.6. Kết luận chương

Trong phần đầu của chương này, luận văn đã trình bày tổng quan về bài toán tìm kiếm thông tin nói chung và bài toán tìm kiếm văn bản pháp quy nói riêng, bao gồm khái, kiến trúc hệ thống và mô hình tìm kiếm thông tin.

Chương này cũng đã trình bày về các phương pháp biểu diễn văn bản bao gồm: biểu diễn sử dụng từ khóa, biểu diễn sử dụng chủ đề ẩn, biểu diễn sử dụng véc-tơ từ, biểu diễn sử dụng mạng nơ-ron sâu. Trong đó, phương pháp biểu diễn sử dụng từ khóa còn nhiều hạn chế, chưa biểu diễn được tốt về ngữ nghĩa, phương pháp biểu sử dụng mạng nơ-ron sâu đang cho thấy hiệu quả cao trong các nghiên cứu gần đây.

CHƯƠNG 2. ỨNG DỤNG BIỂU DIỄN VĂN BẢN BẰNG MẠNG NƠ-RON SÂU TRONG TÌM KIẾM VĂN BẢN PHÁP QUY

Chương này sẽ đề xuất phương pháp biểu diễn văn bản sử dụng mạng nơ-ron nhân chập kết hợp với cơ chế Attention áp dụng cho bài toán tìm kiếm văn bản pháp quy.

2.1. Ý tưởng

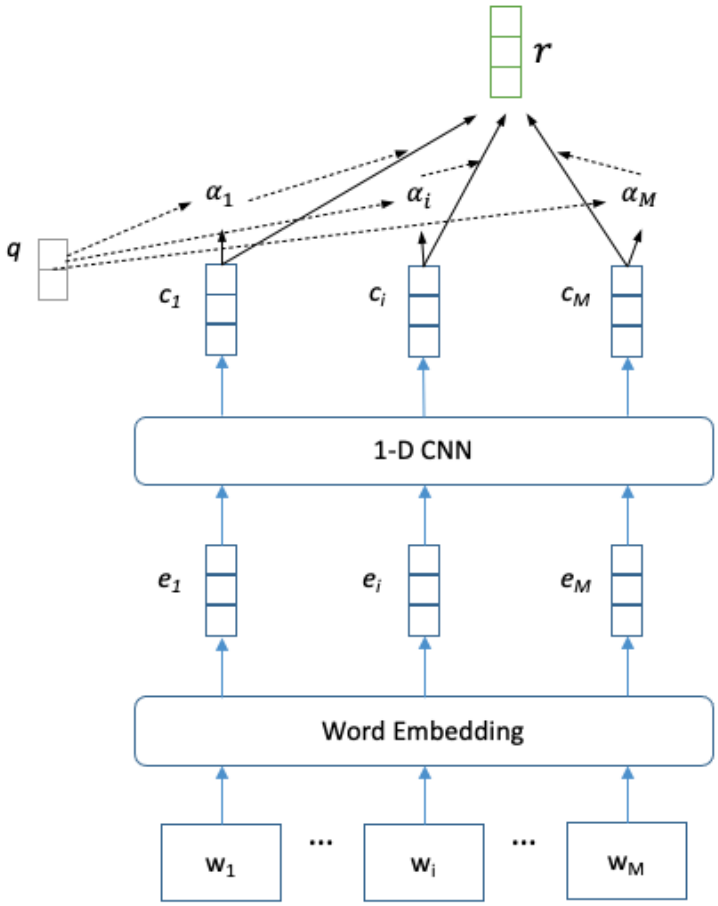
Mỗi điều luật được coi như một văn bản cần tìm kiếm. Tổng quan ý tưởng của phương pháp hai khâu chính. Đầu tiên là biểu điều luật và truy vấn dưới dạng véc-tơ. Sau đó dùng hàm tích vô hướng để so khớp, ước tính độ liên quan giữa chúng.

Mô hình gồm hai mô-đun chính là Mô-đun Biểu diễn truy vấn (Query Encoder) và Mô-đun Biểu diễn điều luật (Article Encoder). Hai mô-đun này sẽ được mô tả chi tiết hơn ở các mục phía sau trong chương này.

Trong mỗi mô-đun, mạng nơ-ron nhân chập sẽ được dùng để ghi nhận các thông tin ngữ cảnh. Sau đó cơ chế Attention sẽ được áp dụng để tính toán các biểu diễn của truy vấn hoặc điều luật.

2.2. Mô-đun Biểu diễn truy vấn

Mô-đun này biến đổi truy vấn thành véc-tơ biểu diễn. Kiến trúc của nó được mô tả ở Hình 2.2, bao gồm ba lớp: word embedding, lớp nhân chập (Convolutional Neural Network - CNN) và attention.

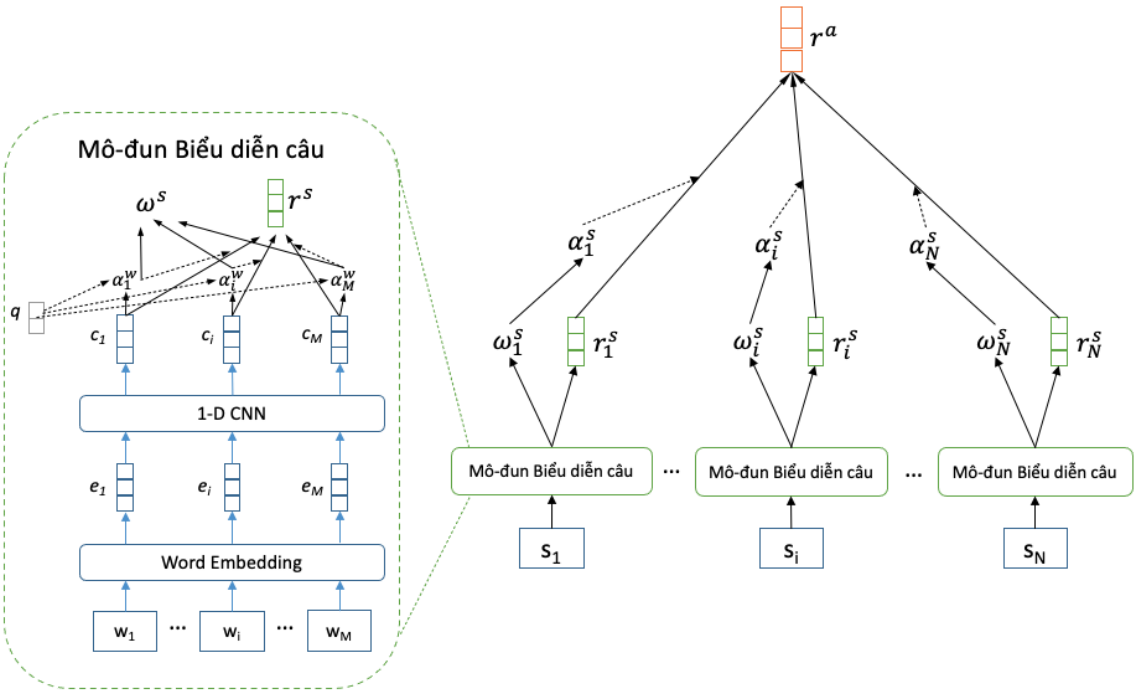


Hình 2.1. Kiến trúc của Mô-đun Biểu diễn truy vấn.

2.3. Mô-đun Biểu diễn điều luật

Mô-đun này biến đổi điều luật dưới dạng một đoạn văn thành một véc-tơ. Kiến trúc của nó được mô tả ở Hình 2.3.

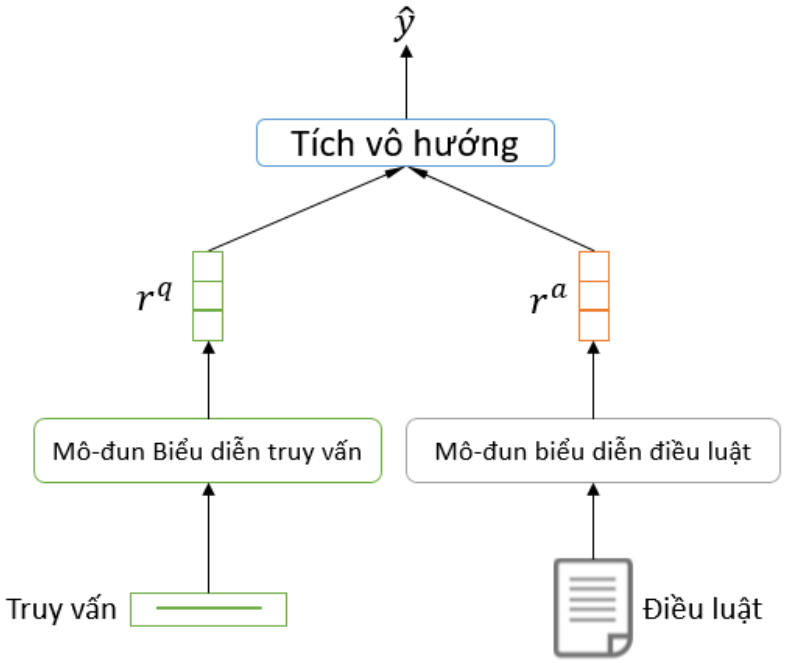
Thay vì xem như một chuỗi dài, mô hình xem điều luật như một đoạn văn tạo thành bởi các câu và sẽ được biểu diễn bằng kiến trúc phân cấp.



Hình 2.2. Kiến trúc của Mô-đun Biểu diễn điều luật.

2.4. So khớp, tính độ liên quan

Hình 2.4 mô tả cách hệ thống tính độ liên quan giữa một điều luật và một truy vấn.



Hình 2.3. Tính độ liên quan giữa một điều luật và một truy vấn.

Độ liên quan giữa một điều luật và một truy vấn được tính bằng tích vô hướng giữa hai véc-tơ biểu diễn của chúng.

Hệ thống được huấn luyện bằng kỹ thuật "*negative sampling*". Hệ thống gán nhãn các điều luật liên quan tới một truy vấn là "*positive*", các điều luật không liên quan là "*negative*". Với mỗi điều luật liên quan, hệ thống chọn mẫu K điều luật không liên quan. Hệ thống sẽ học để phân loại K + 1 điều luật này là liên quan tới truy vấn hay không.

2.5. Kết luận chương

Chương này đã đề xuất phương pháp biểu diễn văn bản sử dụng mạng nơ-ron nhân chập kết hợp với cơ chế Attention áp dụng cho bài toán tìm kiếm văn bản pháp quy.

Chương tiếp theo sẽ trình bày quá trình thu thập, xây dựng dữ liệu, hệ thống và thử nghiệm, đánh giá phương pháp đã đề xuất.

CHƯƠNG 3. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Chương này sẽ trình bày quá trình xây dựng tập dữ liệu văn bản quy, câu hỏi về lĩnh vực pháp luật, việc ứng dụng phương pháp biểu diễn văn bản sử dụng mạng nơ-ron nhân chập kết hợp với cơ chế Attention để xây dựng hệ thống tìm văn bản pháp quy. Cuối cùng là phần thực nghiệm, so sánh với các phương pháp khác.

3.1. Xây dựng tập dữ liệu văn bản pháp quy và câu hỏi

3.1.1. Xây dựng tập dữ liệu văn bản pháp quy tiếng Việt

Tập dữ liệu văn bản pháp quy sử dụng trong luận văn được thu thập từ trang vbpl.vn. Các văn bản được thu thập là các văn bản còn hiệu lực và thuộc các loại sau: bộ luật, luật, nghị định, thông tư, thông tư liên tịch.

Tổng cộng đã thu thập được **8586** văn bản, chia thành **117545** điều.

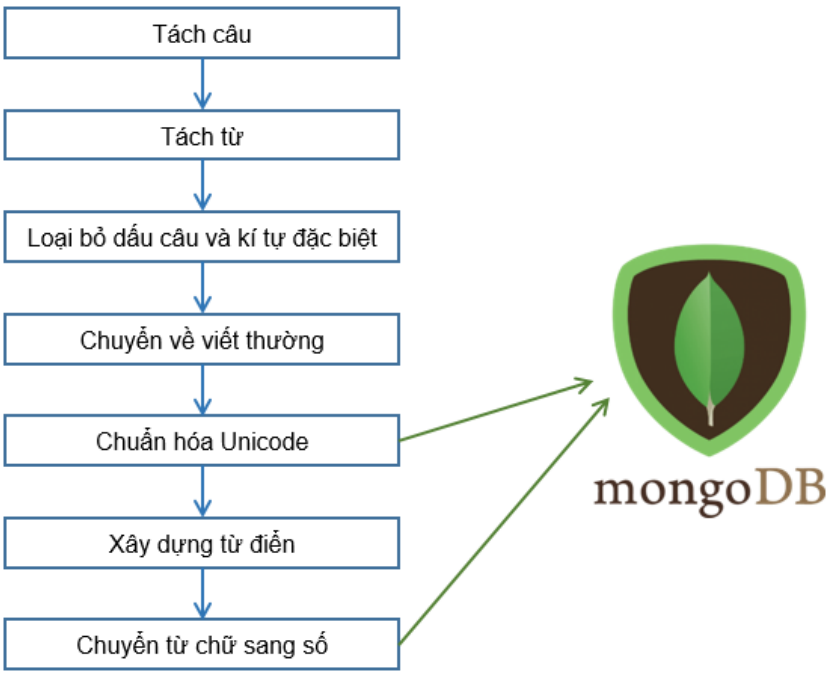
3.1.2. Xây dựng tập câu hỏi và câu trả lời chuẩn

Các câu hỏi được thu thập từ các trang hỏi đáp pháp luật [32][33][34]. Mỗi câu hỏi ban đầu gồm có tiêu đề câu hỏi, chi tiết câu hỏi và câu trả lời.

Tập câu hỏi cuối cùng gồm 2925 câu, mỗi câu hỏi có thể có một hoặc nhiều cách hỏi khác nhau sau đây được gọi là truy vấn. Tổng cộng có **5922** truy vấn.

3.2. Xây dựng hệ thống

3.2.1. Tiền xử lý dữ liệu



Hình 3.1. Các bước tiền xử lý dữ liệu.

3.2.2. Xây dựng hệ thống tìm kiếm sử dụng phương pháp TF-IDF và BM25

Hệ thống này được xây dựng sử dụng Elasticsearch. Elasticsearch cho phép lưu trữ dữ liệu và tạo chỉ mục theo phương pháp biểu diễn TF-IDF và BM25. Mỗi điều sau khi tiền xử lý được lưu thành một bản ghi trong Elasticsearch

Khi nhận được truy vấn, hệ thống sẽ tiền xử lý rồi sử dụng API của Elasticsearch để tìm kiếm theo phương pháp tương ứng.

3.2.3. Xây dựng hệ thống tìm kiếm sử dụng phương pháp biểu diễn văn bản bằng mạng CNN kết hợp với cơ chế Attention

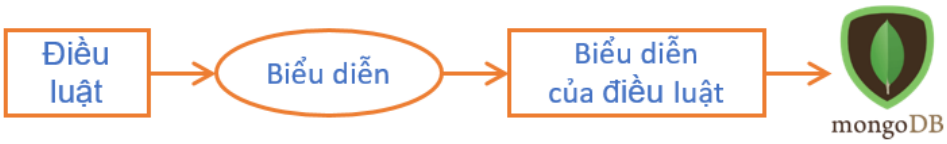
3.2.3.1. *Huấn luyện*

Hệ thống này sẽ lấy kết quả tìm kiếm bằng phương pháp BM25 dựa trên Elasticsearch làm đầu vào rồi dùng mô hình mạng nơ-ron để xếp hạng lại kết quả.

Mô hình được huấn luyện dựa trên kỹ thuật negative sampling.

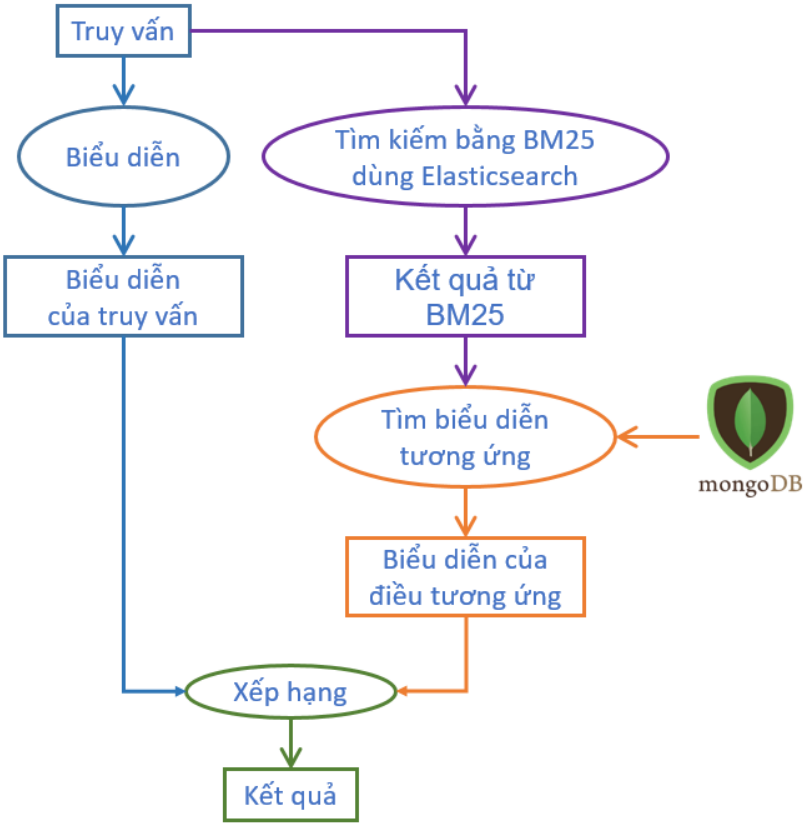
3.2.3.2. *Tìm kiếm*

Các điều luật trong tập dữ liệu văn bản pháp quy sẽ được tính toán biểu diễn sử dụng mô-đun Biểu diễn điều luật đã được huấn luyện. Sau đó, véc-tơ biểu diễn sẽ được lưu kèm với số hiệu văn bản và tên điều để có thể sử dụng được dễ dàng.



Hình 3.2. Lưu trữ biểu diễn của các điều luật.

Khi nhận một truy vấn, hệ thống sẽ biểu diễn nó thành véc-tơ bằng mô-đun biểu diễn truy vấn. Sau đó, hệ thống thực hiện tìm kiếm bằng phương pháp BM25 trong Elasticsearch để thu được một tập kết quả. Hệ thống sẽ tìm kiếm biểu diễn đã được lưu sẵn của các điều tương ứng trong tập kết quả đó. Tiếp theo, độ tương đồng giữa biểu diễn của câu truy vấn và từng điều sẽ được tính rồi dùng để xếp hạng và cho ra kết quả cuối cùng.



Hình 3.3. Quá trình tìm kiếm khi nhận một truy vấn.

3.3. Phương pháp đánh giá

3.3.1. Recall

3.3.2. NDCG

3.4. Kết quả thực nghiệm

Tập câu truy vấn được chia thành 2 phần: 90% dùng để huấn luyện mô hình mạng nơ-ron và 10% dùng để đánh giá các phương pháp.

Đầu tiên là thử nghiệm so sánh hiệu quả của phương pháp sử dụng mạng nơ-ron nhân chập kết hợp với cơ chế Attention, sau đây sẽ được gọi là NATR (Neural Attentive Text Representation), với phương pháp dùng TF-IDF và BM25. Tiếp theo sẽ là thực nghiệm so sánh hiệu quả khi thay đổi các tham số trong NATR. Cuối cùng là thực nghiệm kết hợp điểm của phương pháp BM25 và NATR khi xếp hạng.

3.4.1. Thực nghiệm so sánh hiệu quả của các phương pháp

Trong thực nghiệm này, hệ thống NATR được huấn luyện với 1 điều positive đi kèm 15 điều negative từ Elasticsearch và 15 điều negative. Khi tìm kiếm, hệ thống NATR ấy 1000 kết quả trả về từ Elasticsearch để xếp hạng lại. Kết quả được cho bởi bảng sau:

Bảng 3.1. So sánh hiệu quả các phương pháp.

Phương pháp	**Recall@20**	**NDCG@20**
TF-IDF	0.4716	0.3537
BM25	0.5593	0.3755
NATR	**0.7261**	**0.4642**

3.4.2. Thực nghiệm hiệu quả khi thay đổi các tham số

Trong thực nghiệm này, các tham số được thay đổi để đánh giá tác động lên hiệu quả của hệ thống NATR. Các tham số được thực nghiệm bao gồm:

- K: Số điều negative trong dữ liệu huấn luyện, một nửa lấy từ kết quả trả về của Elasticsearch, một nửa được lấy ngẫu nhiên
- N: Số kết quả trả về từ Elasticsearch dùng để xếp hạng lại khi tìm kiếm.

Kết quả thay đổi tham số K khi huấn luyện và cố định tham số N = 1000 khi tìm kiếm được cho bởi bảng sau:

Bảng 3.2. Kết quả khi thay đổi tham số K

K	Recall@20	NDCG@20	Thời gian huấn luyện
30	0.7261	0.4642	3 giờ 24 phút
60	0.7785	0.5305	6 giờ 20 phút
80	0.7842	0.5452	8 giờ 49 phút
100	**0.8115**	**0.5849**	10 giờ 50 phút
120	0.8103	0.5766	13 giờ 39 phút

Kết quả khi cố định K = 100 khi huấn luyện và thay đổi tham số N khi tìm kiếm được cho bởi bảng sau:

Bảng 3.3. Kết quả khi thay đổi tham số N

N	Recall@20	NDCG@20
300	0.8049	**0.6269**
400	0.8084	0.6147
500	0.8051	0.6063
1000	**0.8115**	0.5849
1500	0.7917	0.5569

3.4.3. Thực nghiệm kết hợp điểm của BM25 và NATR

Trong thực nghiệm này, điểm của phương pháp BM25 và NATR sẽ được kết hợp với nhau để xếp hạng lại các điều trả về từ Elasticsearch. Điểm kết hợp sẽ được tính theo công thức:

$$score = w \times BM25\_score + (1 - w) \times NATR\_score$$

Kết quả thực nghiệm khi cố định K = 100, N = 1000 và thay đổi tham số w được cho bởi bảng sau:

Bảng 3.4. Kết quả khi thay đổi tham số w.

w	**Recall@20**	**NDCG@20**
0.0	0.8155	0.5849
0.1	**0.8245**	**0.6882**
0.2	0.8122	0.6821
0.3	0.7970	0.6741
0.4	0.7954	0.6682
0.5	0.7852	0.6547

3.4.4. Hình ảnh hóa trọng số Attention

3.5. Kết luận chương

Chương này đã trình bày quá trình xây dựng bộ dữ liệu văn bản và câu hỏi pháp quy. Tiếp theo đó là trình bày quá trình áp dụng các phương pháp biểu diễn văn bản để xây dựng hệ thống tìm kiếm văn bản pháp quy và thực so sánh hiệu quả dựa trên bộ dữ liệu đã xây dựng.

Quá trình thực nghiệm đã cho thấy phương pháp biểu diễn văn bản sử dụng mạng nơ-ron nhân chập kết hợp với cơ chế Attention được đề xuất đã cho kết quả tốt hơn các phương pháp hiện có như TF-IDF, BM25.

KẾT LUẬN

Luận văn tập trung nghiên cứu các phương pháp biểu diễn văn bản phục vụ truy xuất, tìm kiếm thông tin và đã đạt được một số kết quả sau:

- Trình bày các phương pháp biểu diễn văn bản
- Đề xuất phương pháp biểu diễn văn bản sử dụng mạng nơ-ron nhân chập và cơ chế Attention.
- Xây dựng bộ dữ liệu văn bản và câu hỏi pháp quy, áp dụng một số phương pháp biểu diễn văn bản để xây dựng hệ thông tìm kiếm thông tin, thử nghiệm và đánh giá các phương pháp đó.
- Kết quả của luận văn đã được chấp nhận công bố tại hội nghị COLING 2020.

Trong tương lai, luận văn có thể tiếp tục được nghiên cứu theo hướng ứng dụng xây dựng hệ thống truy xuất văn bản trong một chủ đề xác định.