ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Xây dựng và cải tiến mô-đun Retrieval cho hệ thống hỏi đáp về pháp luật Việt Nam

TRẦN QUANG ĐẠI
dai.tq194005@sis.hust.edu.vn

Ngành: Khoa học máy tính

Giảng viên hướng dẫn: PGS. TS. Lê Thanh Hương _____

Chữ kí GVHD

Khoa: Khoa học máy tính

Trường: Công nghệ Thông tin và Truyền thông

HÀ NỘI, 01/2024

LỜI CẢM ƠN

Trước hết, em xin gửi lời cảm ơn sâu sắc nhất tới gia đình em, nguồn động lực không thể thiếu trong suốt quá trình học tập và thực hiện đồ án tốt nghiệp này. Những lời động viên, sự ủng hộ và cả những hi sinh thầm lặng của mọi người đã giúp em vượt qua những thử thách khó khăn.

Em cũng muốn gửi lời cảm ơn tới các thầy cô tại Đại học Bách Khoa Hà Nội, đặc biệt là cô Lê Thanh Hương, người đã truyền đạt kiến thức, hướng dẫn và tạo điều kiện tốt nhất cho em để hoàn thành đồ án này. Em cũng xin cảm ơn thầy Ngô Văn Linh, người đã hỗ trợ em rất nhiều trong quá trình học tập tại trường.

Cuối cùng, em xin gửi lời cảm ơn đến bạn bè, những người đã luôn sát cánh cùng em trong suốt quá trình này, cùng em chia sẻ niềm vui, nỗi buồn và giúp em vượt qua những khó khăn.

Và em cũng muốn cảm ơn chính mình, đã không ngừng nỗ lực, quyết tâm và kiên trì để hoàn thành đồ án tốt nghiệp này. Đây chính là thành quả của sự cố gắng không mệt mỏi.

TÓM TẮT NỘI DUNG ĐỒ ÁN

Đồ án tốt nghiệp của em tập trung vào vấn đề xây dựng và cải thiện mô-đun Retrieval cho hệ thống hỏi đáp về pháp luật Việt Nam. Vấn đề này đặc biệt quan trọng trong bối cảnh ngày nay, khi mà nhu cầu tìm kiếm thông tin pháp luật của người dân ngày càng tăng. Tuy nhiên, việc tìm kiếm thông tin pháp luật chính xác, đầy đủ và nhanh chóng vẫn còn là một thách thức. Các hệ thống hỏi đáp truyền thống thường gặp hạn chế trong việc hiểu ngữ cảnh và nội dung của câu hỏi, dẫn đến việc trả lời không chính xác.

Em đã lựa chọn hướng tiếp cận thông qua việc kết hợp các mô hình truy xuất khắc nhau, từ các mô hình thuần túy dựa trên thống kê cho đến các mô hình học biểu diễn văn bản sâu. Đồng thời, em cũng áp dụng nhiều kiến trúc mô hình tiên tiến được xây dựng cho tác vụ truy xuất văn bản để có độ chính xác cao nhất. Qua đó, em tận dụng được sức mạnh của từng mô hình, từ việc so khớp từ vựng cho đến việc hiểu ngữ nghĩa và ngữ cảnh của câu truy vấn.

Giải pháp của em bao gồm việc sử dụng các mô hình bi-encoder và cross-encoder đã được tiền huấn luyện và tinh chỉnh cẩn thận để học được biểu diễn không gian nhúng hiệu quả. Em khai thác các mẫu âm khó cho việc học biểu diễn tương phản để các mô hình học được biểu diễn mạnh mẽ hơn sử dụng kỹ thuật Grad Cache để tương kích thước batch trên phần cứng hạn chế. Thêm vào đó, em cũng bổ trợ mô hình BM25 để tận dụng được thế mạnh về so khớp từ vựng. Cuối cùng, bằng việc ensemble các mô hình bi-encoder, em đã tận dụng được thế mạnh của các mô hình đơn lẻ để nâng cao chất lượng truy xuất.

Đóng góp chính của đồ án tốt nghiệp này là việc xây dựng và cải thiện mô-đun Retrieval, giúp hệ thống hỏi đáp về pháp luật Việt Nam hoạt động chính xác và hiệu quả hơn. Kết quả đạt được sau cùng là việc tăng độ chính xác trong việc truy xuất các văn bản pháp luật, giúp người dùng tiếp cận thông tin pháp luật một cách nhanh chóng và dễ dàng hơn.

Sinh viên thực hiện

(Ký và ghi rõ họ tên)

ABSTRACT

My thesis focuses on the issue of building and improving the Retrieval module for the Vietnamese legal question-answering system. This issue is particularly important in today's context, when the demand for legal information search is increasing. However, accurately, fully, and quickly searching for legal information remains a challenge. Traditional question-answering systems often face limitations in understanding the context and content of the question, leading to inaccurate answers.

I have chosen an approach that combines various retrieval models, from purely statistical-based models to deep text representation learning models. At the same time, I also apply many advanced model architectures built for text retrieval tasks to achieve the highest accuracy. Through this, I leverage the power of each model, from vocabulary matching to understanding the semantics and context of the query.

My solution includes using bi-encoder and cross-encoder models that have been carefully pre-trained and fine-tuned to learn effective embedding space representations. I exploit hard negative samples for contrastive representation learning so that the models learn stronger representations using the Grad Cache technique to boost batch size on limited hardware. In addition, I also supplement the BM25 model to leverage its strength in vocabulary matching. Finally, by ensembling the bi-encoder models, I have leveraged the strengths of individual models to enhance retrieval quality.

The main contribution of this thesis is to build and improve the Retrieval module, helping the Vietnamese legal question-answering system operate more accurately and effectively. The ultimate result is an increase in the accuracy of legal document retrieval, helping users access legal information more quickly and easily.

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 1

1.1 Đặt vấn đề 1

1.2 Các giải pháp hiện tại và hạn chế 1

1.3 Mục tiêu và định hướng giải pháp 2

1.4 Phạm vị của đồ án 3

1.5 Đóng góp của đồ án 3

1.6 Bố cục đồ án 4

CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT 5

2.1 BM25 5

2.1.1 Hàm xếp hạng BM25 5

2.1.2 Ưu điểm và Hạn chế của BM25 5

2.2 Học biểu diễn tương phản 6

2.2.1 Mục tiêu huấn luyện tương phản 6

2.2.2 Các kỹ thuật chính 7

2.3 Bi-encoder 7

2.3.1 Kiến trúc tiêu chuẩn 8

2.3.2 Kiến trúc Condenser 8

2.3.3 Kỹ thuật Grad Cache 10

2.3.4 Kiến trúc coCondenser 11

2.4 Cross-encoder 12

2.5 Ensemble 13

CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT 14

3.1 Tổng quan giải pháp 14

3.1.1 Mô hình backbone 14

3.1.2 Dữ liệu sử dụng 14
3.1.3 Các thành phần chính 15
3.1.4 Sử dụng bổ sung BM25 16
3.2 Tiền xử lý dữ liệu 16
3.3 Retriever 17
3.3.1 Tiền huấn luyện các mô hình bi-encoder 17
3.3.2 Tinh chỉnh các mô hình bi-encoder 17
3.3.3 Ensemble 18
3.4 Re-ranker 19
CHƯƠNG 4. ĐÁNH GIÁ THỰC NGHIỆM 20
4.1 Các tham số đánh giá 20
4.2 Tập dữ liệu đánh giá 20
4.3 Phương pháp thí nghiệm 20
4.3.1 Các baseline lựa chọn để so sánh 21
4.3.2 Các thiết lập tham số 22
4.3.3 Các kịch bản thí nghiệm 22
4.4 Tinh chỉnh với các hàm mất mát khác nhau 23
4.5 So sánh độ chính xác của các mô hình bi-encoder vòng 1 và vòng 2 24
4.6 So sánh độ chính xác với baseline 25
4.7 Tính cần thiết của các thành phần trong mô-đun 26
4.8 Tốc độ truy xuất 27
CHƯƠNG 5. KẾT LUẬN 28
5.1 Kết luận 28
5.1.1 Các vấn đề đã giải quyết 28
5.1.2 Các vấn đề còn tồn đọng 28
5.2 Hướng phát triển trong tương lai 29

TÀI LIỆU THAM KHẢO..... 31

DANH MỤC HÌNH VẼ

Hình 2.1 Kiến trúc chung của bi-encoder 8
Hình 2.2 Kiến trúc Condenser 9
Hình 2.3 Kiến trúc chung của cross-encoder 12

Hình 3.1 Luồng hoạt động của mô-đun Retrieval 15
Hình 3.2 Quá trình tiền huấn luyện và tinh chỉnh các mô hình trong Retriever 17

DANH MỤC BẢNG BIỂU

Bảng 3.1 Số lượng cặp câu trong bộ dữ liệu truy xuất 15
Bảng 3.2 Số lượng cặp câu trong bộ dữ liệu truy xuất vòng 2 của các mô hình bi-encoder vòng 1 18

Bảng 4.1 Độ chính xác (%) của mô hình PhoBERT-base bi-encoder vòng 1 khi tinh chỉnh với các hàm mất mát khác nhau 23
Bảng 4.2 Độ chính xác (%) của các mô hình bi-encoder vòng 1 và vòng 2 24
Bảng 4.3 Độ chính xác (%) của mô hình BM25, hai mô hình baseline, các mô hình bi-encoder vòng 2 và ensemble 25
Bảng 4.4 Độ chính xác (%) của các mô hình bi-encoder vòng 2 khi sử dụng độc lập, bổ sung mô hình BM25, cross-encoder và toàn bộ mô-đun 26
Bảng 4.5 Tốc độ truy xuất (truy vấn/s) của các mô hình bi-encoder vòng 2, các mô hình bi-encoder vòng 2 bổ sung mô hình cross-encoder và toàn bộ mô-đun Retrieval 27