XÂY DỰNG HỆ THỐNG TỰ ĐỘNG GIẢI ĐÁP THẮC MẮC VỀ QUY ĐỊNH HỌC TẬP TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP BẰNG KỸ THUẬT HỌC SÂU

ĐẶNG THỊ PHÚC*, NGUYỄN THANH LONG, ĐẶNG VĂN NGHIÊM, TRẦN THỊ MINH KHOA
Khoa Công nghệ thông tin, trường Đại học Công nghiệp thành phố Hồ Chí Minh
**Tác giả liên hệ: phucdt@iuh.edu.vn*
Dois: https://doi.org/10.46242/jstiuh.v61i07.4725

Tóm tắt. Hiện nay, đối với trường đại học có quy mô lớn như Đại học Công nghiệp TPHCM, số lượng quy định, quy chế, thông báo rất lớn và cập nhật thường xuyên dẫn đến việc tìm hiểu và nắm bắt nội dung trở nên khó khăn. Trong bài báo, chúng tôi xây dựng hệ thống tự động trả lời câu hỏi dựa trên nội dung của các file văn bản bằng kỹ thuật deep learning. Hệ thống trích chọn thông tin từ câu hỏi, đưa vào là các từ khoá và trả về đoạn văn bản liên quan bằng thuật toán BM25. Ứng với đoạn văn bản có độ liên quan cao nhất, mô hình deep learning được huấn luyện để trích xuất ra câu trả lời tương ứng. Mô hình được huấn luyện dựa trên bộ dữ liệu huấn luyện với 10000 và bộ dữ liệu test 1600 cặp câu hỏi và câu trả lời tương ứng từ các đoạn văn bản được lấy từ các thông báo, quy định, quy chế của nhà trường. Chúng tôi tinh chỉnh các mô hình deep learning để huấn luyện và đánh giá, dựa trên hiệu quả và độ chính xác để lựa chọn mô hình tối ưu nhất. Kết quả độ chính xác đạt được theo F1-score của mô hình BERT là 73.93%, RoBERTa là 75.59% PhoBERT là 45.13% và DistilBERT là 72.95%. Mô hình RoBERTa được lựa chọn với tốc độ huấn luyện và độ chính xác cao nhất và được triển khai lên hệ thống để đánh giá kết quả.
Từ khóa. Hệ thống trả lời câu hỏi, xử lý ngôn ngữ tự nhiên, học sâu, BM25, BERT

1 GIỚI THIỆU

Trường Đại học Công Nghiệp TP.HCM là một trong những cơ sở giáo dục Đại học, kỹ thuật lớn nhất khu vực phía Nam. Hiện nay nhà trường có 34.000 sinh viên bậc Sau đại học, Đại học, Cao đẳng đang theo học tại trường. Nhằm nâng cao chất lượng giảng dạy và kết nối giữa sinh viên và nhà trường, nhà trường đã đầu tư nhiều kênh thông tin giúp sinh viên tra cứu giải đáp thắc mắc liên quan đến quá trình học tập. Tuy nhiên việc giải đáp cho toàn bộ sinh viên gặp nhiều khó khăn do hiện tại các bộ phận trả lời thông tin nằm ở các phòng ban, cơ sở, các trang web hoặc facebook, gây nên việc trả lời không kịp thời hoặc không thoả đáng. Hiện nay đã có nhiều hệ thống trả lời tự động từ các công ty lớn như Microsoft, Google, Facebook, Samsung. Các hệ thống này phụ thuộc nhiều vào việc xây dựng dữ liệu. Nếu nguồn dữ liệu không đủ chi tiết, hệ thống sẽ không đáp ứng được các câu hỏi đa dạng. Các hệ thống chatbot này phục vụ cho nhu cầu này bằng cách tạo ra các bộ câu hỏi và câu trả lời tương ứng. Tuy nhiên, khi dữ liệu lớn và thay đổi liên tục, việc tổ chức dữ liệu theo yêu cầu của hệ thống trở nên cồng kềnh, tốn kém thời gian, hơn nữa hệ thống trở nên cứng nhắc và phụ thuộc vào số lượng câu hỏi, câu trả lời đã huấn luyện trước đó. Để đảm bảo cho việc sử dụng lâu dài và đáp ứng được các loại câu hỏi đa dạng, chúng tôi hướng đến việc phát triển hệ thống tự động trả lời câu hỏi từ các tài liệu, văn bản bằng các mô hình deep learning. Hướng phát triển này đã được nghiên cứu trong các bài báo [1,2] đem lại hiệu quả khả quan cho bài toán trả lời câu hỏi. Các mô hình deep learning được áp dụng trong bài toán như sequence-to-sequence dựa trên RNN [3], kết hợp Attention [4], Transformer [5] và một số mô hình hiện đại được áp dụng thành công trong nhiều bài toán như BERT [6].

Trong bài báo, chúng tôi xây dựng hệ thống trả lời câu hỏi dựa trên các tài liệu văn bản được cung cấp. Hệ thống hỏi đáp tự động sẽ phân tích câu hỏi, trích chọn các từ khoá có trọng số cao trong câu hỏi, dựa vào từ khoá này để tìm đoạn văn bản có độ liên quan cao nhất (có chứa các từ khoá đó) trong tập tài liệu được cung cấp. Áp dụng vào mô hình deep learning để trích xuất ra câu trả lời tương ứng trong đoạn văn bản đó.

2 CÁC THUẬT TOÁN LIÊN QUAN

2.1 Các thuật toán liên quan

Những năm gần đây, với công nghệ phát triển của deep learning, các bài toán dữ liệu lớn đã được cải thiện đáng kể, một bước phát triển vượt trội là các mô hình mạng nơ ron tích chập, mạng nơ ron hồi quy kết hợp với cơ chế Attention mạng lại hiệu quả cao khi giải quyết các bài toán dữ liệu lớn. Trong đó mạng nơ ron

hồi quy mang lại hiệu quả đáng kể cho bài toán xử lý ngôn ngữ tự nhiên. Trong đó nổi bật hiệu quả đối với các bài toán NLP là mô hình Transformer, BERT.

2.1.1 Mô hình Transformer

Trong mô hình transformer [7] trích xuất đặc trưng của đối tượng được đưa qua cơ chế seft-attention để mã hoá các thông tin ngữ cảnh. Kiến trúc mạng của mô hình transformer gồm 2 hai bộ phận mã hoá (encoder) bên trái và bộ giải mã (decoder) bên phải, encoder thực hiện đọc dữ liệu đầu vào và decoder đưa ra dự đoán. Bộ mã hoá bao gồm nhiều lớp. Lớp con đầu tiên lớp multi-head attention. Lớp con thứ hai là các lớp đầy đủ chuyển tiếp (fully-connected feed-forward). Bộ giải mã cũng là tổng hợp của nhiều lớp xếp chồng nhau, kiến trúc tương tự như các lớp con của bộ mã hóa ngoại trừ thêm một lớp con thể hiện phân phối chú ý ở vị trí đầu tiên. Ngoài ra, mô hình có một bước cộng thêm mã hóa vị trí (Positional Encoding) vào các đầu vào của bộ mã hóa và bộ giải mã nhằm thêm yếu tố thời gian vào mô hình để tăng độ chính xác. Đây đơn giản là phép cộng véc-tơ mã hóa vị trí của từ trong câu với véc-tơ biểu diễn từ.

Việc sử dụng multi-head attention cho phép cơ chế attention được tính toán độc lập, song song với nhau và được kết nối tuyến tính với nhau để lưu giữ các thông tin attention của các phần khác nhau trong chuỗi. Cụ thể, với mỗi bộ truy vấn, khoá, giá trị Q, K, V chúng tôi biến đổi thành h bộ truy vấn phụ, khoá phụ, giá trị phụ và tính giá trị head bằng việc sử dụng dot- product attention để kết hợp các giá trị với nhau. Sau đó chúng tôi kết nối các head với ma trận trọng số cuối W^0.Công thức Multi- head attention như sau [7]:

$$MultiHead(Q,K,V) = Concat_{i=1,2\dots h}(head_i)W^0 \quad (1)$$

Trong đó

$$\mathrm{head_i(Q,K,V)} = \mathrm{Attention}\left(\mathrm{QW_i^Q}, \mathrm{KW_i^K}, \mathrm{VW_i^V}\right) \quad (2)$$

Ở tầng feed-forward, chúng tôi sử dụng feed-forward neural network trích xuất đặc trưng, bằng cách này, việc lấy cập nhật các gradient ở mỗi thời điểm sẽ độc lập, do đó việc học các phụ thuộc tiềm ẩn ở xa dễ dàng hơn. Feed-forward neural network là một mạng nơ ron tuyến tính sử dụng hàm kích hoạt ReLU được định nghĩa như sau [7]:

$$\mathrm{FFN(X)} = \mathrm{W_2ReLU(X,W_1)} \quad (3)$$

Trong đó W_2, W_1 – là các ma trận trọng số.

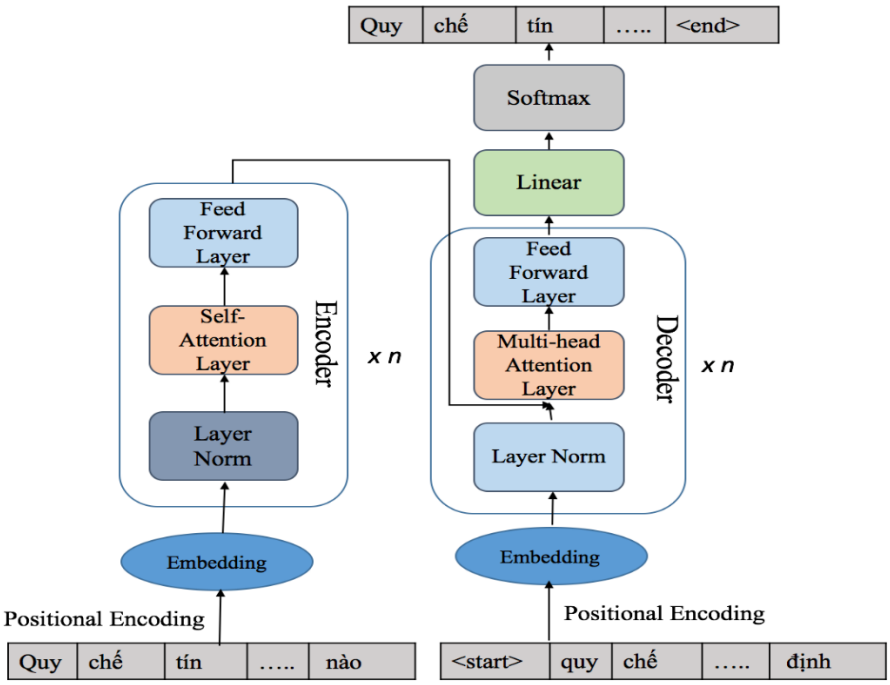
Trong mô hình encoder và decoder chỉ sử dụng multi-head attention và feed-forward neural network mà không sử dụng thêm các lớp convolutional hay recurrent cell, do đó để lưu trữ các thông tin nối tiếp, chúng tôi sử dụng thêm một phương pháp mã hoá là positional encoding, phương pháp này được định nghĩa như sau [7]:

$$\mathrm{PE(pos,2i)} = \sin\left(\mathrm{pos}/10000^{2\mathrm{i}/\mathrm{d_{model}}}\right) \quad (4)$$

$$\mathrm{PE(pos,2i+1)} = \cos\left(\mathrm{pos}/10000^{2\mathrm{i}/\mathrm{d_{model}}}\right), \quad (5)$$

Trong đó pos là vị trí của frame ảnh trong chuỗi ảnh của video hoặc vị trí của từ trong câu, i - kích thước của vector embedding.

Như vậy, với kiến trúc mô hình transformer có thể hỗ trợ việc tính toán song song giữa các từ hoặc các frame ảnh thay vì phải xử lý tuần tự như kiến trúc mạng LSTM trong mô hình sequence to sequence truyền thống, điều này giúp cho hiệu quả tính toán mô hình lên rất nhiều.



Hình 1. Kiến trúc mô hình Transformer

2.1.2 Mô hình BERT

BERT là viết tắt của cụm từ Bidirectional Encoder Representation from Transformer là mô hình biểu diễn từ theo 2 chiều ứng dụng kỹ thuật Transformer. BERT sử dụng Transformer là một mô hình attention học mối tương quan giữa các từ (hoặc 1 phần của từ) trong một văn bản. Transformer gồm có 2 phần chính: Encoder và Decoder, trong khi đó BERT chỉ sử dụng Encoder. Mô hình BERT là mô hình mã hóa sử dụng nhiều lớp Transformer hai chiều [8].

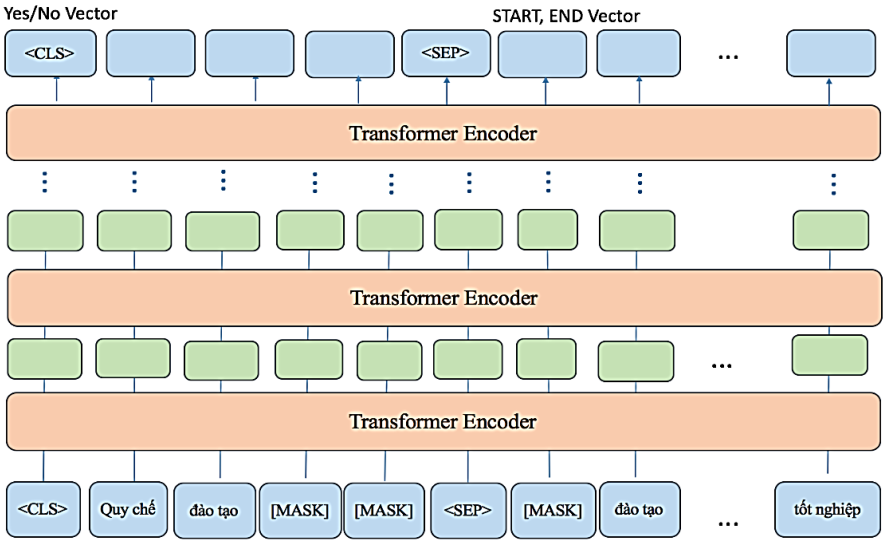
Hiện nay có nhiều phiên bản khác nhau của mô hình BERT, các phiên bản thay đổi dựa trên việc thay đổi kiến trúc của Transformer ở các tham số: L - số lượng lớp, khối trong transformer, H - kích thước của các lớp ẩn là (hay còn gọi là kích thước của embedding vector) A - số lượng head trong multi-head layer. Tên gọi của hai kiến trúc BERT bao gồm:

- $\text{BERT}_{\text{BASE}}(\text{L} = 12, \text{H} = 768, \text{A} = 12)$ - Tổng tham số là 110 triệu.
- $\text{BERT}_{\text{LARGE}}(\text{L} = 24, \text{H} = 1024, \text{A} = 16)$ - Tổng tham số là 340 triệu.

BERT có khả năng huấn luyện dữ liệu theo cả hai chiều, qua đó mô hình có thể học được ngữ cảnh (context) của từ tốt hơn bằng cách sử dụng những từ xung quanh nó (phải & trái). Khi huấn luyện các mô hình truyền thống, một nhược điểm hay gặp phải đó là giới hạn khi học ngữ cảnh của từ. Để khắc phục nhược điểm này, Mô hình BERT được tiền huấn luyện bởi hai tác vụ đó là Mô hình ngôn ngữ với mặt nạ (Masked Language Model) và Dự đoán câu kế tiếp (Next Sentence Prediction.)

Masked Language Model là mô hình mà bối cảnh của từ được học từ cả hai phía bên trái và bên phải cùng một lúc từ những bộ dữ liệu văn bản không giám sát. Dữ liệu đầu vào sẽ được che dấu (thay bằng một token [MASK]) một cách ngẫu nhiên với tỷ lệ thấp. Huấn luyện mô hình dự báo được từ đã được che dấu dựa trên bối cảnh xung quanh là những từ không được che dấu nhằm tìm ra cách biểu diễn của từ.

Next Sentence Prediction là mô hình phân loại với 2 nhãn được huấn luyện song song với mô hình ngôn ngữ mặt nạ. Đầu vào sử dụng một cặp câu và đầu ra dự đoán câu thứ 2 là câu tiếp theo của câu thứ 1 hay không.

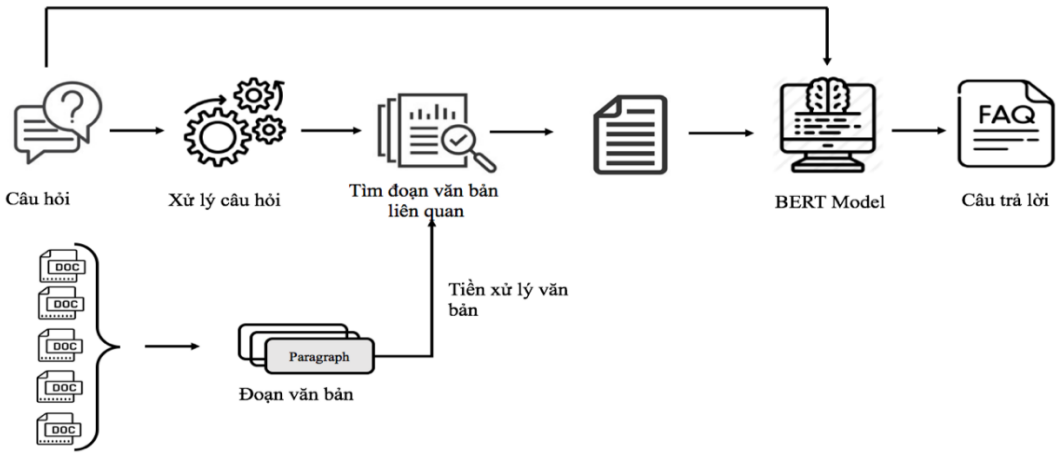


Hình 2: Kiến trúc mạng của BERT

Mô hình Bert sẽ được mô tả chi tiết trong quá trình xây dựng hệ thống trả lời câu hỏi.
Trong bài báo này ngoài mô hình BERT, chúng tôi đề xuất thêm các mô hình cải tiến của BERT như:
Mô hình RoBERTa cải tiến hơn bằng việc loại bỏ Dự đoán câu kế tiếp (Next Sentence Prediction) trong quá trình huấn luyện và đưa ra mặt nạ thay đổi theo thời gian huấn luyện (dynamic masking), thời gian huấn luyện lâu hơn với các kích thước lô (batch size) lớn hơn. Mô hình RoBERTa được đề xuất để cải tiến độ chính xác của mô hình BERT [9].
Mô hình PhoBERT là mô hình cải tiến, được huấn luyện trên 20Gb dữ liệu tiếng Việt nhằm giải quyết các bài toán cho tiếng Việt. PhoBERT cũng có cách tiếp cận tương tự RoBERTa là loại bỏ bước Next Sentence Prediction và chỉ sử dụng Masked Language Model để huấn luyện. Mô hình PhoBERT được đề xuất để đánh giá hiệu quả xử lý tiếng Việt của bài toán [10].
Mô hình DistilBERT: sử dụng kỹ thuật chắt lọc (distillation) bằng cách sử dụng thuật toán xấp xỉ trong thống kê Bayes là Kulback Leiber để xấp xỉ các kiến trúc mô hình mạng nơ-ron lớn bằng các các mạng có kiến trúc nhỏ hơn. DistilBERTcó kiến trúc giảm đi so với BERT 40%. Mô hình DistilBERT được đề xuất để tăng tốc độ tính toán mà vẫn giữ được độ chính xác và hiệu quả của mô hình [11].

2.2 KIẾN TRÚC HỆ THỐNG TRẢ LỜI CÂU HỎI

Kiến trúc hệ thống được mô tả ở hình 3. Mô hình bao gồm các bước: Phân tích câu hỏi, truy xuất văn bản liên quan từ tập dữ liệu và trích xuất câu trả lời từ đoạn văn bản liên quan đó.



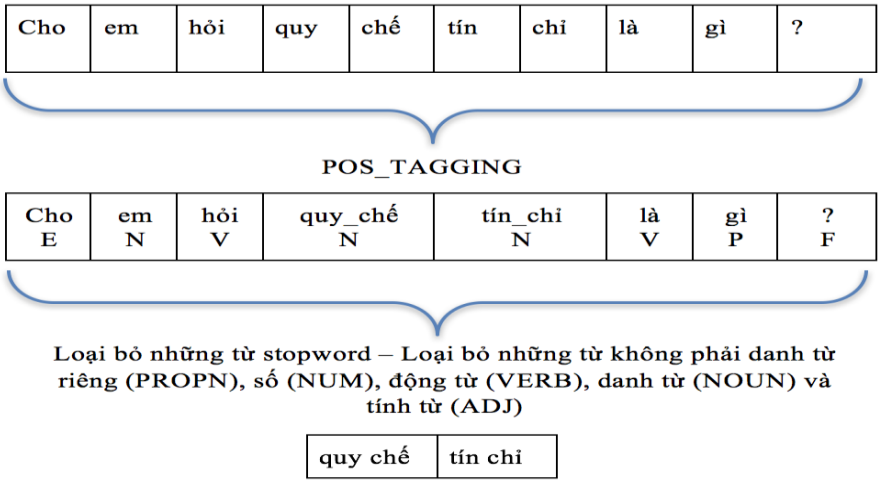
Hình 3: Kiến trúc hệ thống mô hình trả lời câu hỏi từ file, văn bản.

2.2.1 Phân tích câu hỏi

Các bước truy vấn câu hỏi, chúng tôi sử dụng các bước sau để phân tích câu hỏi:

- Đầu tiên câu hỏi được phân tích cú pháp và gắn thẻ bằng các tags part-of-speech.
- Loại bỏ những từ phổ biến trong câu mà không có ý nghĩa liên quan đến nội dung dùng stopwords.

Xoá các từ dựa trên trong tags part-of-speech, những từ không phải danh từ riêng (PROPN), số (NUM), động từ (VERB), danh từ (NOUN) và tính từ (ADJ) bị loại bỏ. Chúng tôi sử dụng thư viện pyvi để phân tích câu hỏi. Ở hình 4, đối với câu hỏi "Cho em hỏi quy chế tín chỉ là gì?" chúng tôi sử dụng thư viện pyvi để gán nhãn cho các từ tương ứng và sau khi loại bỏ các từ stopword, những từ không phải là danh từ riêng, số, danh từ, tính từ, chúng ta thu được 2 từ khoá "quy chế" và "tín chỉ".



Hình 4: Trích xuất từ khoá trong câu hỏi

2.2.2 Truy xuất đoạn văn bản liên quan

Câu hỏi sau khi đã được xử lý truy vấn sẽ được đưa vào trích xuất câu trả lời và đoạn văn tương ứng, tìm kiếm các đoạn văn bản liên quan đến câu hỏi. Khi đưa vào một file văn bản, văn bản trong file được chia nhỏ thành các đoạn văn bản, và các đoạn văn có khả năng chứa câu trả lời sẽ được trích xuất. Để chọn những đoạn văn bản liên quan với câu hỏi, chúng tôi sử dụng thuật BM25 để tạo véc tơ của câu hỏi và đoạn văn. Đối với các keywords cần tìm kiếm, chúng ta tính mức độ quan trọng của từ khoá bằng giá trị TF-IDF [12]:

$$idf(t) = \log\left(1 + \frac{docCount - docFreq + 0.5}{docFreq + 0.5}\right) \quad (6)$$

Trong đó:

docCount là số lượng document.; docFreq là số lượng document chứa term; Log là logarit cơ số tự nhiên (cơ số e).

Mức độ liên quan của từ đối với văn bản được tính bằng công thức BM25 [12]:

$$(BM25)\mathrm{Wt,D} = \frac{IDF * (freq * (k_1 + 1))}{\left(freq * k_1 * \left(1 - b + b * \left(\frac{fieldLength}{avgFieldLength}\right)\right)\right)}$$

$$= \log\left(1 + \frac{docCount - docFreq + 0.5}{docFreq + 0.5}\right) * \frac{(k+1) * freq}{k * (1.0 - b + b * L) + freq} \quad (7)$$

Trong đó:

k - hằng số (thường là 1 và 2); freq: frequency của term trong document; b = 0.75 (mặc định) (b càng về 0 thì độ ảnh hưởng của document length càng nhỏ, và ngược lại, b càng lớn thì độ ảnh hưởng của document Length càng lớn); L = (fieldLength / avgFieldLength): tỉ lệ độ dài của document so với độ dài trung bình của tất cả các đoạn khác.

Trong bảng 1, chúng tôi trích xuất mức độ liên quan của từ khoá "Quy chế" và "tín chỉ" trong tập dữ liệu huấn luyện bằng BM25.

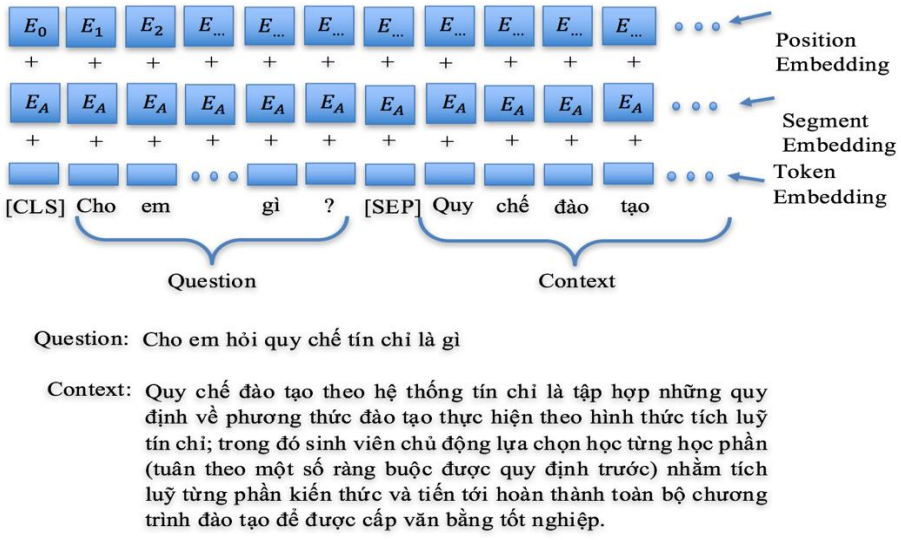
Bảng 1: Mức độ liên quan của từ "quy chế" và "tín chỉ" trong tập dữ liệu huấn luyện.

Đoạn văn bản	Score
Quy chế đào tạo theo hệ thống tín chỉ là tập hợp những quy định về phương thức đào tạo thực hiện theo hình thức tích lũy tín chỉ; trong đó sinh viên chủ động lựa chọn học từng học phần (tuân theo một số ràng buộc được quy định trước) nhằm tích lũy từng phần kiến thức và tiến tới hoàn thành toàn bộ chương trình đào tạo để được cấp văn bằng tốt nghiệp.	4.996
Trên cơ sở chương trình đào tạo, quy chế đào tạo theo hệ thống tín chỉ tạo điều kiện tối đa cho sinh viên phát huy tích cực, chủ động, sáng tạo trong việc sắp xếp lịch học, đăng ký khối lượng kiến thức sẽ tích lũy trong từng học kỳ, tích lũy các học phần. kể cả sắp xếp thời gian học ở trường. thời gian tốt nghiệp, ra trường. quy chế này cũng tạo điều kiện để sinh viên tích cực, chủ động thích ứng với quy trình đào tạo để đạt được những kết quả tốt nhất trong học tập, rèn luyện.	4.982
Khối lượng kiến thức tốt nghiệp đối với bậc đào tạo đại học được phân bổ như sau: khối công nghệ, kỹ thuật (cấp bằng kỹ sư): 13 tín chỉ, trong đó thực tập doanh nghiệp 5 tín chỉ và khóa luận tốt nghiệp 8 tín chỉ. đối với các khối ngành khác (cấp bằng cử nhân): 10 tín chỉ, trong đó thực tập doanh nghiệp 5 tín chỉ và khóa luận tốt nghiệp 5 tín chỉ.	4.482
Sau mỗi học kỳ, căn cứ vào khối lượng kiến thức tích lũy, sinh viên được xếp hạng năm đào tạo như sau: sinh viên năm thứ nhất nếu khối lượng kiến thức tích lũy dưới 35 tín chỉ. sinh viên năm thứ hai nếu khối lượng kiến thức tích lũy từ 35 tín chỉ đến dưới 70 tín chỉ. sinh viên năm thứ ba nếu khối lượng kiến thức tích lũy từ 70 tín chỉ đến dưới 105 tín chỉ. sinh viên năm thứ tư nếu khối lượng kiến thức tích lũy từ 105 tín chỉ trở lên.	3.958

2.2.3 Trích xuất câu trả lời

Đoạn văn bản có độ liên quan cao nhất ở bước truy xuất sẽ làm đầu vào cho việc trích xuất nội dung tương ứng cho đầu ra. Câu trả lời trích xuất được trích xuất từ đoạn văn bản ứng với câu hỏi đó. Ở đây, câu hỏi và đoạn văn bản được đưa vào mô hình học sâu trích xuất câu trả lời và mô hình sẽ đưa ra kết quả với độ chính xác ứng với mỗi câu trả lời. Chúng tôi sử dụng mô hình BERT để trích xuất câu trả lời.

Đầu vào của mô hình BERT bao gồm 1 câu hỏi và đoạn văn bản tương ứng. Các từ trích xuất đặc trưng bằng Token Embedding, đồng thời nhúng thêm segment embeddings để phân biệt câu hỏi với đoạn văn bản và Position Embedding để chỉ định vị trí của các từ trong chuỗi như hình 5.



Hình 5: Tổ chức dữ liệu đầu vào của mô hình BERT

Với đầu vào là câu hỏi và đoạn văn bản, mô hình BERT trích xuất véc tơ Start token và End token được tính bằng xác suất của mỗi token là điểm bắt đầu và điểm kết thúc trong khoảng câu trả lời. Với mục tiêu là xác định điểm bắt đầu và điểm kết thúc cho câu trả lời trong đoạn văn, mô hình sẽ xác định hai véc tơ S

và T (sẽ được học trong quá trình tinh chỉnh) Sau đó, lấy tích vô hướng $e^{S*T} = \Sigma_i e^{S*Tj}$ của các véc tơ với véc tơ đầu ra của câu thứ hai (bắt đầu từ SEP) và trả về scores (điểm xác suất). Sau đó, áp dụng hàm kích hoạt Softmax trên các scores để tính xác suất. Mục tiêu của việc huấn luyện là tổng của các log-likelihoods (khả năng xảy s) của các vị trí bắt đầu và vị trí kết thúc đúng nhất. Chúng ta có một véc vơ trọng số (bộ weights) riêng biệt đó.

Đối với véc tơ xác xuất cho vị trí bắt đầu (S)):

$$P_i = \frac{e^{S*Ti}}{\Sigma_i e^{S*Tj}} \quad (8)$$

Trong đó:

T là từ cần tập trung vào.

Đối với véc tơ xác xuất cho vị trí kết thúc (E)):

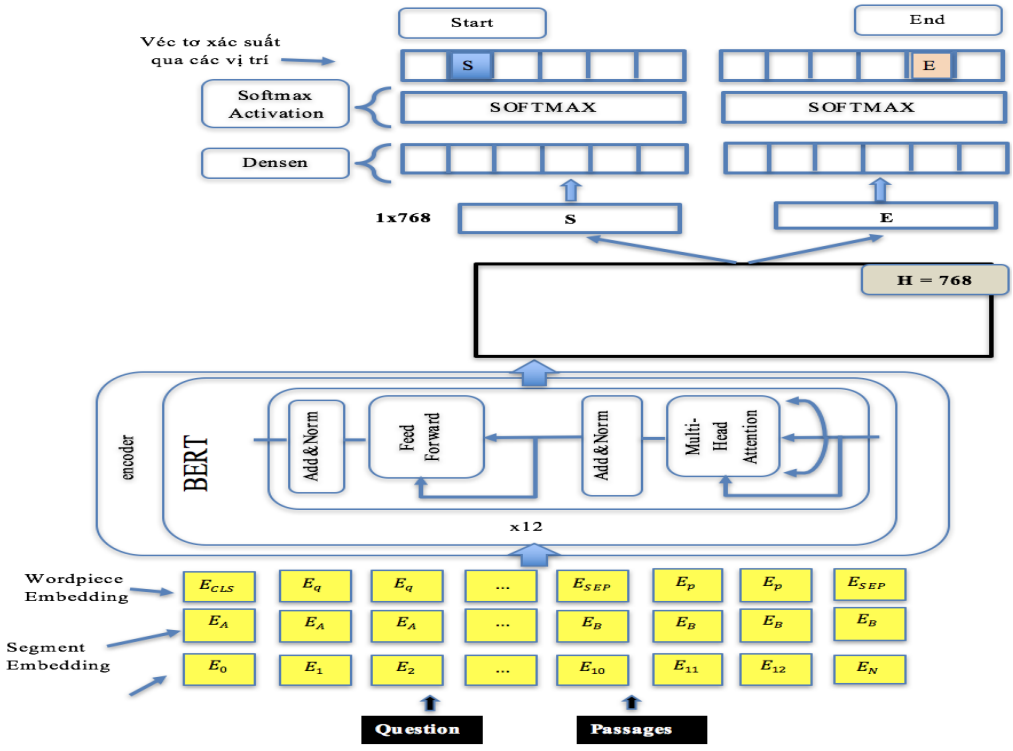
$$P_i = \frac{e^{E*Ti}}{\Sigma_i e^{E*Tj}} \quad (9)$$

Trong đó:

T là từ cần tập trung vào.

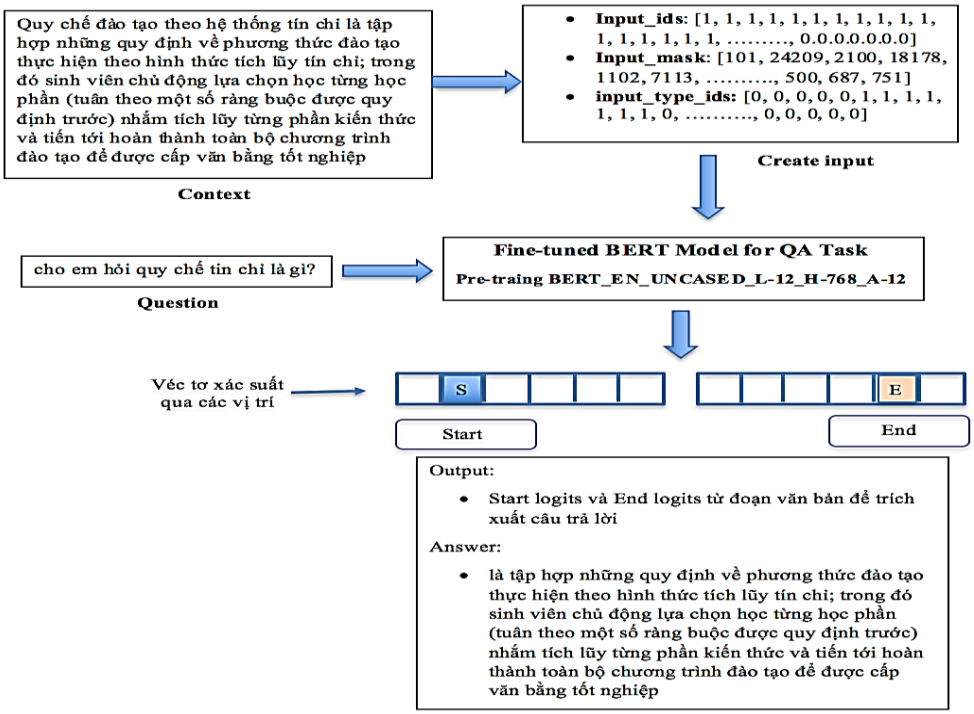
Để dự đoán đúng khoảng câu trả lời, chúng ta lấy tất cả các điển S*T và E*T và lấy khoảng có Score tốt nhất của $(S * T + E * T)$

Chúng tôi sử dụng các mô hình pre-train BERT để huấn luyện. Quy trình huấn luyện ứng với đầu vào và đầu ra tương ứng được mô tả ở hình 6.



Hình 6: Quá trình huấn luyện dữ liệu trên mô hình BERT

Quá trình chuẩn bị dữ liệu đầu vào để huấn luyện mô hình BERT để đạt được đầu ra được mô tả cụ thể ở hình 7.



Hình 7: Sơ đồ huấn luyện mô hình.

3 KẾT QỦA THỰC NGHIỆM

3.1 Bộ dữ liệu

Bộ dữ liệu các câu hỏi được định dạng theo dữ liệu của Stanford Answering Dataset [13], gồm 10.000 câu hỏi và ứng với 10.000 câu trả lời. Nguồn dữ liệu được xây dựng từ các file văn bản của trường Đại học Công nghiệp TP HCM bao gồm: quy chế tín chỉ, quy định xét học bổng, đánh giá rèn luyện, công tác sinh viên và tuyển sinh IUH.

Dữ liệu được tạo bằng cách thêm câu hỏi, câu trả lời và đánh chỉ số bắt đầu của câu trả lời. Sử dụng công cụ cdQA-annotator để tạo dữ liệu câu hỏi và câu trả lời cho bài toán. Dữ liệu huấn luyện được tạo như hình 1, trong đó, trung bình mỗi file được tách thành 10 đến 40 đoạn văn bản khác nhau tuỳ vào độ dài của văn bản. Ứng với mỗi đoạn văn bản được tạo từ 50 đến 100 câu hỏi và câu trả lời khác nhau. Trung bình mỗi file sẽ có 1500 đến 2500 câu hỏi và câu trả lời khác nhau tuỳ vào độ dài của văn bản. Bộ dữ liệu được mô tả ở hình 8.

Tiền xử lý dữ liệu: Đối với bộ dữ liệu, loại bỏ những ký tự (“1.”, ”2.”, “a)”, “b)”, những ký tự đặc biệt, chuẩn hoá chữ hoa, chữ thường, ... để các đoạn văn bản được gọn gàng nhất.

Phân tích câu hỏi: Đối với các câu hỏi, chúng tôi sử dụng thư viện pyvi để gán nhãn từ loại. Từ câu hỏi chúng tôi lấy ra những từ là danh từ riêng (PROPN), số (NUM), động từ (VERB), danh từ (NOUN), tính từ (ADJ) và loại bỏ những từ không thuộc các nhãn này.

Truy xuất đoạn vản bản liên quan: Ứng với mỗi câu hỏi, chúng tôi sử dụng thuật toán BM25 để trích xuất đoạn văn bản có độ liên quan cao nhất.

	title	context	question	id	answers.answer_start	answers.text
0	Phân loại để đánh giá	Sinh viên chuyển trường được sự đồng ý của Hiệ...	Em chào thầy cô, thầy cô cho em hỏi là kết quả...	28471ba0-cc26-48c9-9d50-65e0f139a3e0	[0]	Sinh viên chuyển trường được sự đồng ý của Hiệ...
1	Phân loại để đánh giá	Sinh viên chuyển trường được sự đồng ý của Hiệ...	Các thầy cô cho em hỏi chút kết quả rèn luyện ...	01d6add1-2694-42ff-baa4-3df5c8b2d63a	[0]	Sinh viên chuyển trường được sự đồng ý của Hiệ...
2	Phân loại để đánh giá	Sinh viên chuyển trường được sự đồng ý của Hiệ...	Dạ cho em hỏi là kết quả rèn luyện của đại học...	3fd88e18-36e7-4065-b2ab-589ed5273053	[0]	Sinh viên chuyển trường được sự đồng ý của Hiệ...
3	Phân loại để đánh giá	Sinh viên chuyển trường được sự đồng ý của Hiệ...	Em chào thầy/cô ạ, cho em hỏi kết quả rèn luyệ...	d9846285-07f9-45d4-a62e-75ba97da77c0	[0]	Sinh viên chuyển trường được sự đồng ý của Hiệ...
4	Phân loại để đánh giá	Sinh viên chuyển trường được sự đồng ý của Hiệ...	Em chào thầy/cô ạ, cho em hỏi kết quả rèn luyệ...	567b7b56-1da9-413a-99ac-b0a5d71bd686	[0]	Sinh viên chuyển trường được sự đồng ý của Hiệ...
...	...	...	...	...	...	...

Hình 8: Bộ dữ liệu câu hỏi và trả lời được xây dựng từ văn bản

3.2 Kết quả thực nghiệm

Các mô hình học sâu được huấn luyện trên bộ dữ liệu IUH question answering với số lượng tập train hơn 10000 dữ liệu và tập test 1600 dữ liệu. Mô hình được xây dựng trên ngôn ngữ Python và được thực hiện trên nền tảng Google Colab Pro, GPU Tesla P100-PCIE-16GB, RAM 12.6 GB. Kiến trúc mô hình được tinh chỉnh với thông số như bảng 2.

Bảng 2: Thông số kiến trúc mạng của các mô hình

Mô hình	Kích thước của các lớp ẩn (H)	Số lượng head trong multi-head layer (A)	Số lượng khối trong transformer (L)
BERT	256	4	4
RoBERTa	256	8	8
PhoBERT	768	12	12
DistilBERT	256	8	8

Kết quả huấn luyện trên các mô hình khác nhau và đánh giá độ chính xác như hình được mô tả ở bảng 3. Chúng tôi đánh giá hiệu quả của mô hình thông qua thời gian huấn luyện và số lượng vòng lặp để đạt được độ chính xác. Độ chính xác của mô hình được đánh giá dựa trên 2 giá trị là F1-score và EM. Giá trị EM được tính theo quy tắc: đối với mỗi cặp câu hỏi, câu trả lời, nếu các ký tự trong câu dự đoán của mô hình khớp chính xác với các ký tự của một trongcác câu trả lời đúng thì EM = 1, nếu không thì EM = 0. Giá trị F1-score được định nghĩa bởi công thức [14]:

Precision: tỉ lệ số lượng da ung thư đúng được phân loại đúng trong số lượng ảnh được phân loại là da ung thư đó.

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

Recall: tỉ lệ số lượng da ung thư đúng được phân loại đúng trong số lượng ảnh thực sự đúng là da ung thư đó.

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

Trong đó:

TP - True Positives khi nhãn dự đoán và nhãn thực tế đều là 1; FP - False Positives khi nhãn dự đoán là 1 và nhãn thực tế là 0; FN - False Negatives khi nhãn dự đoán là 0 và nhãn thực tế là 1; TN - True Negatives khi nhãn dự đoán và nhãn thực tế đều là 0.

F1-score: là điểm trung bình hài hòa của precision và recall được tính bằng công thức:

$$F1 = 2 * (\frac{precision*recall}{precision+recall}) \tag{12}$$

Kết quả cho thấy mô hình RoBERTa đạt kết quả tốt nhất cả về tốc độ thực hiện cũng như độ chính xác.

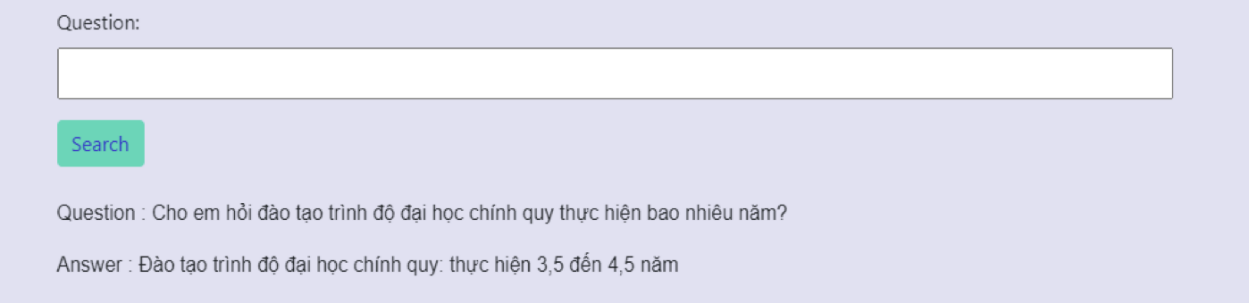
Bảng 3: Hiệu quả và độ chính xác của mô hình

Mô hình	Thời gian thực hiên trên 1 Epoch	Số lượng Epoch	Batch size	F1-Score (%)	EM (%)
BERT	327s	10	4	73.93	71.95
RoBERTa	568s	20	16	75.59	74.2
PhoBERT	939s	20	16	45.13	37.1
DistilBERT	1204s	20	8	72.95	71.5

Chúng tôi lựa chọn mô hình RoBERTa để triển khai lên web, và một số kết quả thu được:

Với câu hỏi đầu vào như hình "*Cho em hỏi đào tạo trình độ đại học chính quy thực hiện bao nhiêu năm?*". Đoạn văn bản được trích xuất từ file có giá trị BM25 cao nhất và từ đó trích xuất câu trả lời được hiện thị trên web như hình 3.

"*Trường Đại học Công nghiệp Thành phố Hồ Chí Minh đào tạo đa bậc, đa hệ, đa ngành với đầu vào của các bậc, các hệ, các ngành khác nhau và thời gian đào tạo của các bậc, các hệ, các ngành cũng khác nhau, cụ thể như sau:* ***Đào tạo trình độ đại học chính quy: thực hiện 3,5 đến 4,5 năm****. Đào tạo trình độ đại học liên thông từ trung cấp: thực hiện 2,5 đến 3,5 năm. Đào tạo trình độ đại học liên thông từ cao đẳng: thực hiện 1,5 – 2 năm. Đào tạo trình độ đại học vừa làm vừa học: thực hiện 4 đến 5 năm. Đào tạo văn bằng 2 trình độ đại học: thực hiện 2,5 đến 3 năm. Đào tạo trình độ cao đẳng: thực hiện 2,5 năm*". Câu trả lời được hiển thị trên web như hình 9. Kết quả cho thấy mô hình trích xuất đúng đoạn văn bản và trích xuất đúng câu trả lời trong văn bản đó.



Hình 9: Hệ thống trích xuất câu trả lời tương ứng trên web

Với câu hỏi: "Sinh viên đi thi hộ bị xử lý như thế nào?". Đoạn văn bản có có độ liên quan cao nhất là: "*Trong khi kiểm tra thường xuyên, chuẩn bị tiểu luận, bài tập lớn, thi giữa học phần, thi kết thúc học phần, chuẩn bị đồ án, khóa luận tốt nghiệp, nếu vi phạm quy chế, sinh viên sẽ bị xử lý kỷ luật với từng học phần đã vi phạm.* ***Sinh viên đi thi hộ hoặc nhờ người khác thi hộ, đều bị kỷ luật bằng hình thức buộc thôi học.*** *Trừ trường hợp được quy định tại khoản 2 điều này, mức độ sai phạm và khung xử lý kỷ luật đối với sinh viên vi phạm được thực hiện theo quy định của quy chế tuyển sinh đại học, cao đẳng hệ chính quy"*. Kết quả cho thấy mô hình trích xuất đúng đoạn văn bản và trích xuất đúng câu trả lời trong văn bản đó.
Với câu hỏi: "Em đã đi học ngày đầu, nhưng không muốn học nữa thì có rút lại được học phí không?". Đoạn văn bản có có độ liên quan cao nhất là: *"Sinh viên chỉ được rút bớt hay huỷ học phần đã đăng ký trong thời gian Phòng Đào tạo chưa khóa lớp học phần. Sau khi đã hết thời hạn cho phép rút hay huỷ bớt các học phần, sinh viên đã được chấp nhận đăng ký các học phần phải đóng học phí cho những học phần đã được chấp nhận. Nếu không đóng học phí đúng thời hạn quy định, phần mềm sẽ tự động hủy đăng ký tất cả các học phần mà sinh viên chưa đóng phí. Những học phần sinh viên đã đăng ký và đóng học phí mà không học thì được xem như tự ý bỏ học và phải nhận điểm F học phần đó.* ***Khi học phần đã được triển khai giảng dạy, Nhà trường không chấp thuận cho sinh viên rút bớt hay huỷ các học phần****"*. Kết quả cho thấy mô hình trích xuất được đoạn văn bản liên quan, tuy nhiên câu trả lời chưa thực sự theo sát với ý của câu hỏi.
Nhận xét: đối với những câu hỏi không quá dài mô hình có thể tìm được các từ quan trọng, đối với những câu hỏi quá phức tạp mô hình trích xuất bị dư thừa hoặc bỏ qua những từ quan trọng dẫn đến câu trả lời không chính xác. Trong hướng phát triển sắp tới, chúng tôi sẽ phân tích thêm ở bước phân tích câu hỏi như phân loại câu hỏi, phân tích đặc trưng ngữ cảnh, phân tích pos tag để tăng độ chính xác trong việc đánh trọng số của từ trong câu.

4 KẾT LUẬN

Trong bài báo, chúng tôi xây dựng mô hình tự động trả lời câu hỏi được trích xuất từ văn bản. Mô hình được huấn luyện trên tập dữ liệu được thu thập từ các quy định, quy chế, thông báo của trường Đại học Công nghiệp TPHCM. Chúng tôi xây dựng 10000 bộ câu hỏi và câu trả lời trong đoạn văn bản liên quan cho tập huấn luyện và 1600 cho tập dữ liệu test. Hệ thống trả lời câu hỏi dựa trên quá trình phân tích câu hỏi để tìm ra các từ khoá có trọng số cao, dựa vào các từ khoá này để tìm đoạn văn bản trong tập dữ liệu các quy định, quy chế có liên quan đến câu hỏi bằng thuật toán BM25. Ứng với đoạn văn bản có độ liên quan cao nhất được trích xuất câu trả lời bằng các mô hình BERT, RoBERTa, PhoBERT và DistilBERT. Kết quả so sánh cho thấy mô hình RoBERTa có độ chính xác trong việc dự đoán vị trí câu trả lời là cao

nhất, đạt 75.59%. Kết quả đánh giá cho thấy, đối với những câu hỏi rõ ràng, hệ thống cho được câu trả lời chính xác. Tuy nhiên đối với những câu trả lời phức tạp, bao gồm nhiều ý để hỏi thì mô hình chưa thể trích xuất được. Với hướng phát triển sắp tới là chúng tôi sẽ bổ sung thêm dữ liệu với nhiều loại câu hỏi đa dạng, đồng thời cải tiến các thuật toán liên quan để mô hình đạt độ chính xác cao hơn.

TÀI LIỆU THAM KHẢO

[1] L. -Q. Cai, M. Wei, S. -T. Zhou and X. Yan, "Intelligent Question Answering in Restricted Domains Using Deep Learning and Question Pair Matching," *IEEE Access*, vol. 8, pp. 32922-32934, 2020, doi: 10.1109/ACCESS.2020.2973728.

[2] D. Singh, K. R. Suraksha and S. J. Nirmala, "Question Answering Chatbot using Deep Learning with NLP," in *2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2021, pp. 1-6, doi: 10.1109/CONECCT52877.2021.9622709.

[3] E. Karimi, B. Majidi and M. T. Manzuri, "Relevant Question Answering in Community Based Networks Using Deep LSTM Neural Networks," in *2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, 2019, pp. 1-5, doi: 10.1109/CFIS.2019.8692168.

[4] A. Sharma and C. Harithas, "Inner Attention Based bi-LSTMs with Indexing for non-Factoid Question Answering," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 1-7, doi: 10.1109/ICMLA.2018.00009.

[5] D. A. Navastara, Ihdiannaja and A. Z. Arifin, "Bilingual Question Answering System Using Bidirectional Encoder Representations from Transformers and Best Matching Method," in *13th International Conference on Information & Communication Technology and System (ICTS)*, 2021, pp. 360-364, doi: 10.1109/ICTS52701.2021.9608905.

[6] A. Saha, M. I. Noor, S. Fahim, S. Sarker, F. Badal and S. Das, "An Approach to Extractive Bangla Question Answering Based On BERT-Bangla And BQuAD," in *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, 2021, pp. 1-6, doi: 10.1109/ACMI53878.2021.9528178.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998-6008.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, Volume 1, pp. 4171-4186.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019, arxiv:1907.11692.

[10] Dat Quoc Nguyen and Anh Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP*, 2020, pp. 1037-1042.

[11] S. Victor, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," 2019, arXiv:1910.01108.

[12] A. Trotman, A. Puurula and B. Burgess, "Improvements to BM25 and Language Models Examined," in *Proceedings of the 2014 Australasian Document Computing Symposium (ADCS '14)*, 2014, pp.58–65, doi: 10.1145/2682862.2682863.

[13] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Lian, "SQuAD: 100000+ questions for machine comprehension of text," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2016, arXiv:1606.05250.

[14] Y. Sasaki, "The truth of the F-measure," in *Teach Tutor mater*, vol. 1, no. 5, 2007, pp. 1-5.

APPLYING DEEP LEARNING FOR AUTOMATIC REGULATION QUESTION ANSWERING SYSTEM AT INDUSTRIAL UNIVERSITY OF HO CHI MINH CITY

ĐANG THI PHUC*, NGUYEN THANH LONG, ĐANG VAN NGHIEM, TRAN THI MINH KHOA

Faculty of Information Technology, Industrial University of Ho Chi Minh City

**Corresponding author: phucdt@iuh.edu.vn*

Abstract. Currently, for a large-scale university like the Industrial University of Ho Chi Minh City, the number of regulations, announcements is very large and frequently updated, making it difficult to understand and grasp the content. In this paper, we build a system to automatically answer questions based on the content of text files using deep learning techniques. The system extracts information from the question, enters the keywords and returns the relevant text using the BM25 algorithm. Given the text with the highest relevance, the deep learning model is trained to extract the corresponding answer. The model is trained on a training data set of 10000 and a test dataset of 1600 pairs of questions and corresponding

answers from texts taken from announcements, regulations of the university. We refine deep learning models for training and evaluation, based on efficiency and accuracy to select the most optimal model. The resulting accuracy obtained according to the F1-score of the BERT model is 73.93%, RoBERTa is 75.59% PhoBERT is 45.13% and DistilBERT is 72.95%. The RoBERTa model was selected with the highest training speed and accuracy and deployed to the system to evaluate the results.
Keywords. Question answering system, natural language processing, deep learning, BM25, BERT

Ngày gửi bài: 15/04/2022

Ngày chấp nhận đăng: 29/06/2022