

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP. HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN



TRẦN MỸ LINH - 18066361

ĐẶNG VĂN NGHIÊM - 18056331

**ỨNG DỤNG HỌC SÂU ĐỂ CẢI TIẾN MÔ HÌNH
PHÂN TÍCH CÂU HỎI TRONG BÀI TOÁN TRẢ LỜI
CÂU HỎI**

Ngành: Khoa học máy tính

Giảng viên hướng dẫn: TS. Đặng Thị Phúc

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 06 NĂM 2022

**BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP. HCM
KHOA CÔNG NGHỆ THÔNG TIN**



TRẦN MỸ LINH - 18066361

ĐẶNG VĂN NGHIÊM - 18056331

**ỨNG DỤNG HỌC SÂU ĐỂ CẢI TIẾN MÔ HÌNH
PHÂN TÍCH CÂU HỎI TRONG BÀI TOÁN TRẢ LỜI
CÂU HỎI**

Ngành: Khoa học máy tính

Giảng viên hướng dẫn: TS. Đặng Thị Phúc

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 06 NĂM 2022

INDUSTRIAL UNIVERSITY OF HO CHI MINH CITY
FACULTY OF INFORMATION TECHNOLOGY



TRAN MY LINH - 18066361

DANG VAN NGHIEM - 18056331

**DEEP LEARNING APPLICATION TO IMPROVE
THE QUESTION ANALYSIS MODEL IN THE
QUESTION ANSWER PROBLEM**

Major: Computer Science

Supervisor: Ph.D Dang Thi Phuc

HO CHI MINH CITY, 2022

DEEP LEARNING APPLICATION TO IMPROVE THE QUESTION ANALYSIS MODEL IN THE QUESTION ANSWER PROBLEM

ABSTRACT

Recently, open domain question answering (QA) has been combined with machine understanding models to find answers in a large of knowledge. Since QA requires retrieving relevant documents from a dataset to answer questions (what, who, when, how, where yes/no, ...), its performance largely depends on the processing. Therefore, the question analysis stage plays an important role, directly affecting the accuracy of the model. If the question analysis is not good, it will not be possible to find the right question. In this article, we will introduce how to analyze the question more specifically, the passages find better answers and less noise.

Keywords: question answering, word weighting, bm25, bert

LỜI CẢM ƠN

Trong suốt quá trình học tập và thực hiện khóa luận tốt nghiệp chúng em luôn được sự quan tâm, hướng dẫn và giúp đỡ tận tình của các thầy, cô giáo trong khoa Công nghệ thông tin cùng với sự động viên giúp đỡ của bạn bè.

Lời đầu tiên, nhóm chúng em xin được bày tỏ lòng biết ơn sâu sắc đến Ban giám hiệu Trường Đại học Công Nghiệp Thành Phố Hồ Chí Minh, ban chủ nhiệm khoa Công nghệ thông tin đã tận tình giúp đỡ cho chúng em suốt thời gian học tại trường.

Đặc biệt, nhóm chúng em xin bày tỏ lòng biết ơn chân thành sâu sắc đến cô TS. Đặng Thị Phúc - người đã hướng dẫn và chỉ bảo tận tình cho chúng em trong suốt quá trình học tập. Đặc biệt hơn thì nhóm chúng em còn được cô hướng dẫn trong suốt quá trình thực hiện khóa luận khóa luận tốt nghiệp.

Có lẽ kiến thức là vô hạn mà sự tiếp nhận kiến thức của nhóm thì có những hạn chế nhất định. Do đó, trong quá trình hoàn thành bài tiểu luận, chắc chắn sẽ không thoát khỏi những thiếu sót. Do đó, rất mong có sự đóng góp của quý thầy cô để đề tài của nhóm được hoàn thiện hơn. Nhóm xin chân thành cảm ơn!

Người thực hiện đề tài

Trần Mỹ Linh

Đặng Văn Nghiêm

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày.... Tháng năm....

CHỮ KÝ CỦA GIẢNG VIÊN

NHẬN XÉT CỦA GIẢNG VIÊN PHẢN BIỆN 1

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày.... Tháng năm....

CHỮ KÝ CỦA GIẢNG VIÊN

NHẬN XÉT CỦA GIẢNG VIÊN PHẢN BIỆN 2

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày.... Tháng năm....

CHỮ KÝ CỦA GIẢNG VIÊN

MỤC LỤC

DANH LỤC HÌNH ẢNH.....	1
DANH MỤC BẢNG BIỂU	4
DANH MỤC CÁC THUẬT NGỮ VIẾT TẮT	5
CHƯƠNG 1 TỔNG QUAN	7
1.1 Đặt vấn đề.....	7
1.2 Phương hướng giải quyết	7
CHƯƠNG 2 CƠ SỞ LÝ THUYẾT.....	9
Các phương pháp đã có	9
2.1 Mạng nơ-ron hồi quy (Recurrent Neural Network - RNN)	11
2.2 Attention	14
2.2.1 Scale Dot-Product Attention	15
2.2.2 Multi-head Attention.....	18
2.3 TF-IDF.....	19
2.4 BM25 (Best matching)	21
CHƯƠNG 3 MÔ HÌNH	23
3.1 Mô hình phân loại câu hỏi	23
3.2 Mô hình tìm trọng số của từ	24
3.2.1 Trích xuất đặc trưng ngữ cảnh	26
3.2.2 Phân tích pos-tags cho câu hỏi.....	27
3.2.3 Phân tích loại câu hỏi	28

3.2.4	Kết hợp các mô hình và đưa ra trọng số	29
3.3	Ứng dụng trọng số vào thuật toán BM25	29
3.4	Mô hình transformer.....	30
3.5	Mô hình BERT	33
3.6	Tinh chỉnh mô hình RoBERTa.....	36
3.6.1	Khởi tạo vector nhúng.....	37
3.6.2	Attention.....	39
3.6.3	Encoder (Bộ mã hóa)	41
3.6.4	Decoder (Bộ giải mã).....	42
CHƯƠNG 4	THỰC NGHIỆM VÀ ĐÁNH GIÁ.....	43
4.1	Kết quả mô hình phân loại câu hỏi.....	43
4.2	Kết quả mô hình tìm trọng số của từ	46
4.3	Ứng dụng trọng số vào thuật toán BM25	50
4.4	Kết quả mô hình RoBERTa	53
CHƯƠNG 5	KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....	78
5.1	Kết luận	78
5.2	Hướng phát triển.....	78
TÀI LIỆU THAM KHẢO.....		79

DANH LỤC HÌNH ẢNH

Hình 1. 1 Phương hướng phát triển	7
Hình 2. 1 Kiến trúc RNN chứa một tầng ẩn	11
Hình 2. 2 Kiến trúc LSTM	12
Hình 2. 3 Ví dụ self-attention	14
Hình 2. 4 Scaled Dot-Product Attention	15
Hình 2. 5 Tạo ma trận Q, K, V	16
Hình 2. 6 Score matrix	16
Hình 2. 7 Scale score	17
Hình 2. 8 Ví dụ tìm đặc trưng của từ "Tôi"	17
Hình 2. 9 Multi-head Attention	18
Hình 2. 10 Mô hình hóa tính toán multi head attention	19
Hình 3. 1 Cấu trúc mô hình sử dụng Multi-head Attention	23
Hình 3. 2 Cấu trúc mô hình đánh trọng số của từ	25
Hình 3. 3 Trích xuất đặc trưng ngữ cảnh và feedforward	26
Hình 3. 4 Phân tích postag.....	27

Hình 3. 5 Phân tích câu hỏi	28
Hình 3. 6 Output của mô hình tìm trọng số.....	29
Hình 3. 7 Sơ đồ kiến trúc transformers	31
Hình 3. 8 Tính mã hóa vị trí	32
Hình 3. 9 Masked language model.....	34
Hình 3. 10 Next Sentence Prediction	35
Hình 3. 11 Mô hình RoBERTa Encoder Decoder.....	36
Hình 3. 12 Embedding block.....	37
Hình 3. 13 Ảnh minh họa layer embedding hoạt động	38
Hình 3. 14 Ví dụ token type cho một câu.....	38
Hình 3. 15 Ví dụ token type cho 2 câu.....	39
Hình 3. 16 Attention	39
Hình 3. 17 Các bước tính attention	40
Hình 3. 18 Encoder trong Bert	41
Hình 3. 19 Decoder trong Bert	42
Hình 4. 1 Các bước giải quyết bài toán	43
Hình 4. 2 Dữ liệu phân loại câu hỏi	43
Hình 4. 3 Biểu đồ phân bố các lớp câu hỏi	44
Hình 4. 4 Kết quả mô hình phân loại sử dụng Multi-head Attention.....	45

Hình 4. 5 Kết quả mô hình phân loại sử dụng Bi_LSTM	45
Hình 4. 6 Dữ liệu mô hình trọng số.....	46
Hình 4. 7 Ví dụ minh họa huấn luyện và tính loss cho mô hình	47
Hình 4. 8 Loss mô hình đánh trọng số của từ.....	48
Hình 4. 9 Kết quả test model đánh trọng số của từ	48
Hình 4. 10 Ảnh dữ liệu cho mô hình trả lời câu hỏi.....	53
Hình 4. 11 Quá trình embedding block	54
Hình 4. 12 Quá trình train encoder	54
Hình 4. 13 Áp dụng trọng số vào encoder.....	55
Hình 4. 14 Quá trình train decoder.....	55
Hình 4. 15 Output mô hình.....	56
Hình 4. 16 Độ chính xác và loss của mô hình RoBERTa	56

DANH MỤC BẢNG BIỂU

Bảng 4.1 So sánh mô hình Bi_LSTM và Multi-head Attention

Bảng 4.2 So sánh các mô hình BERT

DANH MỤC CÁC THUẬT NGỮ VIẾT TẮT

Từ viết tắt	Từ đầy đủ	Nghĩa
NLP	Natural language processing	Xử lý ngôn ngữ tự nhiên
QA	Question answering	Trả lời câu hỏi
LSTM	Long short-term memory	Bộ nhớ đệm dài và ngắn
Bi_LSTM	Bidirectional long short-term memory	Bộ nhớ đệm dài và ngắn 2 chiều
RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
MHA	Multi head attention	Trích xuất các đặc trưng của câu hỏi
TF	Term frequency	Tần suất xuất hiện của từ
IDF	Inverse document frequency	Tần xuất nghịch đảo văn bản
POS tags	Part-of-speech	Gán nhãn từ loại

ỨNG DỤNG HỌC SÂU ĐỂ CẢI TIẾN MÔ HÌNH PHÂN TÍCH CÂU HỎI TRONG BÀI TOÁN TRẢ LỜI CÂU HỎI

Tóm tắt

Gần đây, trả lời câu hỏi (QA) miễn mở đã được kết hợp với các mô hình hiểu máy để tìm câu trả lời trong một lượng lớn kiến thức. Vì QA yêu cầu truy xuất các tài liệu liên quan từ tập dữ liệu để trả lời các câu hỏi (cái gì, ai, khi nào, như thế nào, ở đâu, có / không, ...), hiệu suất của nó phần lớn phụ thuộc vào quá trình xử lý. Vì vậy, khâu phân tích câu hỏi đóng vai trò quan trọng, ảnh hưởng trực tiếp đến độ chính xác của mô hình. Nếu phân tích câu hỏi không tốt sẽ không thể tìm ra câu hỏi đúng. Trong bài này, chúng tôi sẽ giới thiệu cách phân tích câu hỏi cụ thể hơn, các đoạn văn tìm ra câu trả lời hay hơn và ít nhiễu hơn.

Từ khóa: trả lời câu hỏi, trọng số, bm25, bert

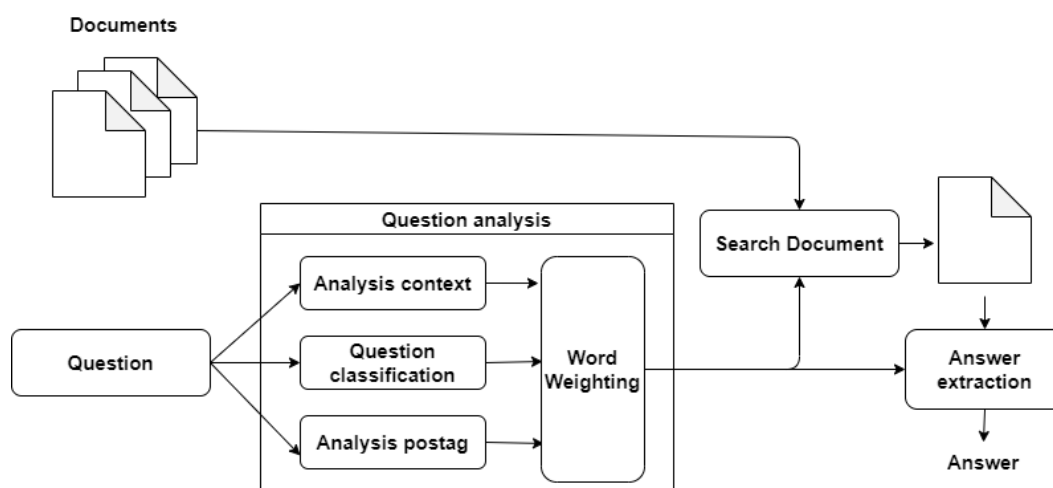
CHƯƠNG 1 TỔNG QUAN

1.1 Đặt vấn đề

Hiện nay, với xu hướng số hóa tài liệu, đa số tài liệu văn bản được cập nhật lên mạng toàn cầu với số lượng lớn. Việc tìm kiếm tài liệu theo câu hỏi trở nên khó khăn. Do đó, chúng ta cần một công cụ tìm kiếm câu trả lời nhanh nhất và sát nhất với câu hỏi được đặt ra.

Phân tích câu hỏi là yếu tố đầu tiên trong kiến trúc chung của một hệ thống hỏi đáp, nó có nhiệm vụ tìm ra các thông tin cần thiết làm đầu vào cho quá trình xử lý của các quá trình sau (trích chọn tài liệu, trích xuất ra câu trả lời, ...). Vì vậy, phân tích câu hỏi có vai trò hết sức quan trọng và ảnh hưởng trực tiếp đến hoạt động của toàn bộ hệ thống. Nếu phân tích câu hỏi không đúng thì sẽ không tìm ra được câu trả lời phù hợp. Trong quá trình thực hiện, nhóm em đã thử nghiệm nhiều phương pháp khác nhau. Tuy nhiên, với phạm vi và trình độ có hạn nên kết quả chưa được tốt.

1.2 Phương hướng giải quyết



Hình 1. 1 Phương hướng phát triển

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Vấn đề cần giải quyết ở đây là việc phân tích câu hỏi để nâng cao độ chính xác cho việc tìm đoạn văn có khả năng chứa câu trả lời từ nguồn tri thức khổng lồ trên các website và nhiều nguồn tài liệu khác nhau.

Sau quá trình phân tích câu hỏi chúng ta cần phải tìm kiếm lại tài liệu, tìm ra tài liệu có khả năng chứa câu trả lời dựa vào kết quả của việc phân tích câu hỏi.

Và cuối cùng sau khi có đoạn văn chứa câu trả lời việc cần làm hiện tại là trích xuất câu trả lời từ đoạn văn đó.

Tóm lại các công việc cần phải làm:

Phân tích câu hỏi: phân tích loại câu hỏi, phân tích từ loại trong câu hỏi.

Tìm kiếm đoạn văn chứa câu trả lời.

Trích xuất câu trả lời.

CHƯƠNG 2 CƠ SỞ LÝ THUYẾT

Các phương pháp đã có

Việc phân tích câu hỏi hết sức quan trọng để tìm kiếm được câu trả lời hợp lý. Phân tích câu hỏi nhận đầu vào là câu hỏi dưới dạng ngôn ngữ tự nhiên của người dùng, đưa ra câu truy vấn cho bước trích chọn tài liệu liên quan và các thông tin cần thiết cho bước trích xuất câu trả lời.

Phân tích câu hỏi bao gồm phân tích loại câu hỏi và phân tích từ loại trong câu hỏi.

Thông thường có hai hướng tiếp cận được sử dụng rộng rãi trong việc phân lớp câu hỏi đó là hướng tiếp cận dựa trên luật và hướng tiếp cận dựa trên xác suất thống kê.

Hướng tiếp cận dựa trên luật: hướng tiếp cận này yêu cầu phải có các chuyên gia ngôn ngữ cung cấp các luật, các biểu thức chính quy, các từ khóa cho từng câu hỏi ... để hệ thống hoạt động. Có các hạn chế của hướng tiếp cận này được chỉ ra như sau:

Xây dựng mô hình cho phương pháp này rất tốn thời gian và công sức, cần có sự cộng tác của những chuyên gia trong lĩnh vực ngôn ngữ học khi xây dựng các mẫu câu hỏi và văn phạm cho từng loại câu hỏi đó.

Các luật ngữ pháp viết tay và văn phạm của từng loại câu hỏi rất cứng nhắc, không linh hoạt. Khi một dạng câu hỏi mới xuất hiện, mô hình theo hướng này không thể xử lý. Muốn xử lý được mô hình cần phải được cung cấp theo những luật mới.

Các vấn đề nhập của các văn phạm ngữ pháp rất khó xử lý, kiểm soát và phụ thuộc và đặc điểm của từ ngôn ngữ.

Khi bộ dữ liệu câu hỏi được mở rộng hoặc có sự thay đổi kéo theo việc phải viết loại hoàn toàn các luật trước đó nên hệ thống rất khó mở rộng.

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Hướng tiếp cận dựa trên xác suất thông kê: được tổng hợp lại gồm hai cách tiếp cận chính đó là:

Phương pháp học máy: sử dụng tập dữ liệu câu hỏi đủ lớn đã được qua bước xử lý gán nhãn để huấn luyện mô hình có thể tự động nắm bắt được các mẫu có ích trong việc phân lớp câu hỏi. Các thuật toán thường được dùng là Support Vector Machines (SVM), láng giềng gần nhất (K-Nearest Neighbors – kNN), Naïve Bayes (NB), ...

Phương pháp sử dụng mô hình ngôn ngữ: Xây dựng một mô hình ngôn ngữ thông kê để ước lượng được phân phối của ngôn ngữ tự nhiên chính xác nhất có thể. Có thể nói bài toán phân lớp câu hỏi là việc ước lượng xác suất có điều kiện.

Ngoài ra, hiện nay có một phương pháp trích xuất các từ khóa, các từ quan trọng đó là phương pháp TF-IDF [1].

Sau khi phân tích câu hỏi chúng ta sẽ tìm được các từ quan trọng. Chúng ta sẽ sử dụng một thuật toán nào đó để tìm các đoạn văn bản liên quan. Sau đó sẽ áp dụng các mô hình học máy để trích xuất câu trả lời.

Hướng tiếp cận dựa trên học máy hiện đang được rất nhiều nhà nghiên cứu quan tâm vì nó không chỉ tốn ít công sức của con người mà tính khả chuyển cao, dễ dàng áp dụng cho nhiều miền ứng dụng khác nhau. Tuy nhiên cũng sẽ có khó khăn khi số lượng câu hỏi lớn.

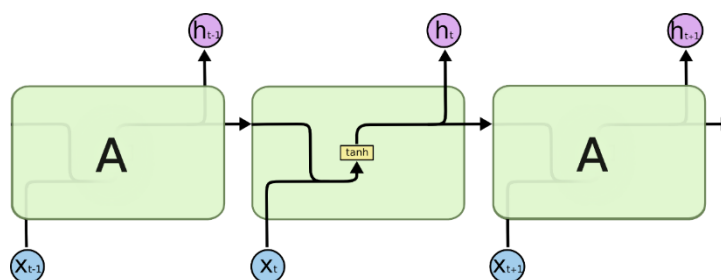
Hướng tiếp cận dựa trên mô hình học máy cho các tác vụ xử lý ngôn ngữ tự nhiên thường áp dụng các kiến trúc của mạng nơ-ron hồi quy [2] (Recurrent Neural Network – RNN). Mạng nơ-ron hồi quy đạt được nhiều kết quả tốt. Bên cạnh đó nó có những khuyết điểm, nhưng dưới sự xuất hiện của attention một phần nào góp phần khắc phục những điểm yếu đó.

Hiện nay, sự xuất hiện của attention dần dần cũng đã thay thế RNN, các mô hình gần đây hầu như đều loại bỏ RNN trong kiến trúc của mình. Nổi bật hiện nay là các mô hình Transformer [3], BERT [4].

Dưới đây chúng tôi sẽ trình bày chi tiết về phương pháp đánh trọng số cũ là TF-IDF, thuật toán BM25 [5], ưu nhược điểm của mạng nơ-ron hồi quy và lý do vì sao các mô hình hiện tại đa phần đều sử dụng attention.

2.1 Mạng nơ-ron hồi quy (Recurrent Neural Network - RNN)

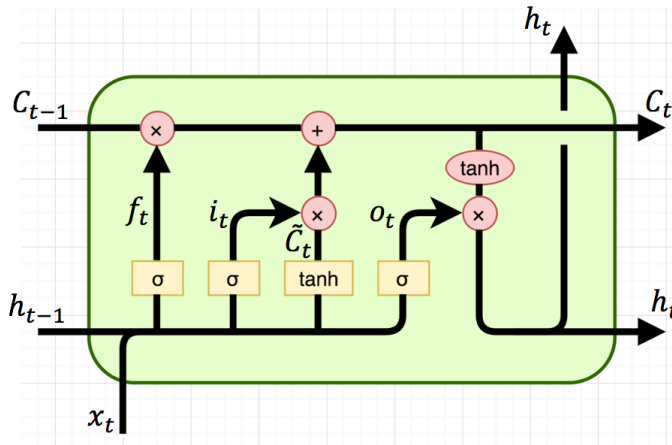
Trong xử lý ngôn ngữ tự nhiên, dữ liệu sẽ là một câu văn. Về cấu tạo của một câu thì câu được tạo thành bằng cách kết hợp các từ với nhau. Nếu xét từng từ riêng lẻ thì nó sẽ không mang ý nghĩa, từ phía sau sẽ cần phải có từ trước đó mới có thể hiểu được. Do đó, cần phải có một kiến trúc mô hình nào đó để tạo sự liên kết giữa các từ và từ đầu đến cuối câu. Mạng RNN được thiết kế để giải quyết vấn đề này.



Hình 2. 1 Kiến trúc RNN chứa một tầng ẩn¹

Cấu trúc của RNN còn khá đơn giản nên khả năng kết nối các thông tin xa phía trước không được tốt. Nguyên nhân chính được giải thích là do sự triệt tiêu đạo hàm của hàm chi phí khi trải qua một chuỗi các tính toán hồi quy trên các dữ liệu dài hạn. Một phiên bản của RNN là LSTM (Long short term memory) ra đời.

¹ https://phamdinhhkhanh.github.io/2019/04/22/Ly_thuyet_ve_mang_LSTM.html



Hình 2. 2 Kiến trúc LSTM²

Đầu vào: C_{t-1} , h_{t-1} , x_t . Trong đó x_t là đầu vào ở trạng thái thứ t , C_{t-1} , h_{t-1} là đầu ra của lớp trước đó.

Đầu ra: C_t , h_t gọi C là cell state, h là hidden state.

Như chúng ta thấy được trên hình kiến trúc của LSTM có một đường thẳng ký hiệu là C nó được gọi là ô trạng thái (cell state) ô này sẽ lưu trạng thái của các lần trước đó và qua mỗi time step nó sẽ được cập nhật loại bỏ hoặc giữ lại các thông tin cần thiết.

Bước đầu tiên LSTM sẽ quyết định thông tin nào được phép đi qua ô trạng thái tức là cell state. Ở đây nó sử dụng hàm sigmoid ở tầng quên để chuẩn hóa các giá trị vào khoảng 0 và 1. Ở đây nó nhận vào 2 giá trị là h_{t-1} và x_t . Nó chuẩn hóa giá trị trước khi đưa vào ô trạng thái với 0 là thông tin không quan trọng, 1 là thông tin quan trọng cần giữ lại và được nhân vào cell state để loại bỏ hoặc giữ lại thông tin cần thiết.

$$f_t = \sigma(x_t U^f + h_{t-1} W^f) \quad (1)$$

Bước tiếp theo chúng ta sẽ quyết định thông tin nào được lưu thêm vào thanh trạng thái. Ở bước này gồm 2 tầng ẩn với 2 hàm kích hoạt là hàm sigmoid và hàm tanh.

² <https://medium.com/@saurabh.rathor092/simple-rnn-vs-gru-vs-lstm-difference-lies-in-more-flexible-control-5f33e07b1e57>

Với tầng tanh thì sẽ tạo ra thông tin để có thể thêm vào trạng thái. Đối với hàm sigmoid sẽ nhân với đầu ra của hàm tanh để quyết định xem thông tin nào nên đưa vào trạng thái.

$$i_t = \sigma(x_t U^i + h_{t-1} W^i) \quad (2)$$

$$\bar{C}_t = \tanh(x_t U^g + h_{t-1} W^g) \quad (3)$$

$$C_t = \sigma(f_t * C_{t-1} + i_t * \bar{C}_t) \quad (4)$$

Cuối cùng chúng ta cần phải chọn những thông tin cần thiết ở trạng thái để trả ra kết quả. Chúng ta cũng sẽ có 2 tầng ẩn với 2 hàm kích hoạt là hàm sigmoid và hàm tanh. Hàm sigmoid sẽ quyết định lấy thông tin nào. Trạng thái được qua hàm tanh để chuyển về khoảng -1 và 1. Sau nó nhân với đầu ra tại hàm sigmoid, do đó sẽ thu được những phần cần thiết.

$$o_t = \sigma(x_t U^o + h_{t-1} W^o) \quad (5)$$

$$h_t = \tanh(C_t) * o_t \quad (6)$$

Nhược điểm của RNN [6]:

Vanishing gradient: hiện tượng gradient sẽ bị nhỏ lại tới mức gần như biến mất ở những hidden state cuối của một đầu vào dài như đoạn văn.

RNN có nhiệm vụ đưa thông tin kịp thời. Nhưng việc truyền tất cả các thông tin này là một việc khá khó khăn khi bước thời gian quá dài. Khi một mạng có quá nhiều lớp sâu, nó sẽ trở nên không thể kiểm tra được.

RNN không thể hệ thống dài dòng nếu việc sử dụng tanh hoặc relu như một tính năng kích hoạt.

RNN học tuần tự từ trái sang phải nên mất khá nhiều thời gian và bộ nhớ cho mô hình RNN.

2.2 Attention

Attention được dịch sang tiếng Việt là “sự chú ý”. Đúng với tên của nó, Attention giúp mô hình chú ý vào những thông tin quan trọng. Đối với tác vụ xử lý ngôn ngữ tự nhiên nó giúp chú ý vào những từ quan trọng, còn với tác vụ xử lý ảnh nó giúp chú ý vào các điểm ảnh quan trọng.

Attention giúp giải quyết được các nhược điểm của các mô hình RNN. Attention giúp đánh lại trọng số lại các thông tin đặc trưng tránh gây mất thông tin đối với câu dài.

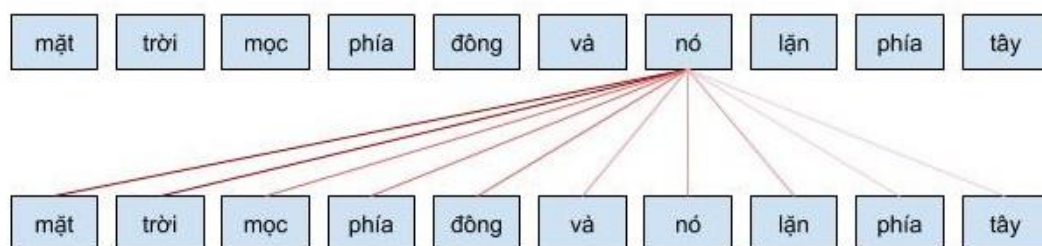
Sự xuất hiện của attention dần dần phát sinh ra các loại attention có thể thay thế các mô hình RNN dùng để trích xuất đặc trưng. Hiện nay nổi bật là self-attention (tự chú ý), xuất hiện trong bài báo Attention is all you need. Tức là từ nay chúng ta không còn phải phụ thuộc vào RNN mà thứ cần thiết là attention giải quyết tất cả tác vụ mà RNN làm được thậm chí còn tốt hơn hẳn.

Trong bài báo tác giả cũng đã đề cập tới 3 lý do sử dụng self-attention [7]:

Độ phức tạp tính toán.

Khối lượng tính toán được song song hóa.

Khả năng học phụ thuộc xa.



Hình 2. 3 Ví dụ self-attention³

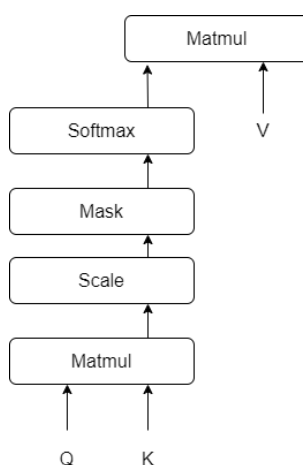
³ <https://pbcquoc.github.io/transformer/>

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Sau này self-attention được tinh chỉnh cho nhiều bài toán và đạt được kết quả cao điển hình như BERT. Về cơ bản self-attention ở các mô hình được tính toán gần như là giống nhau.

2.2.1 Scale Dot-Product Attention

Đây chính là một cơ chế self-attention khi mỗi từ có thể điều chỉnh trọng số của nó cho các từ khác trong câu sao cho từ ở vị trí càng gần nó nhất thì trọng số càng lớn và càng xa thì càng nhỏ dần.



Hình 2. 4 Scaled Dot-Product Attention

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [3] \quad (7)$$

Trong đó:

Q: ma trận query có số chiều là dk .

K: ma trận key có số chiều là dk .

V: ma trận value có số chiều là dk .

Các ma trận Q, K, V được tạo bằng cách:

$$Q = XW_Q$$

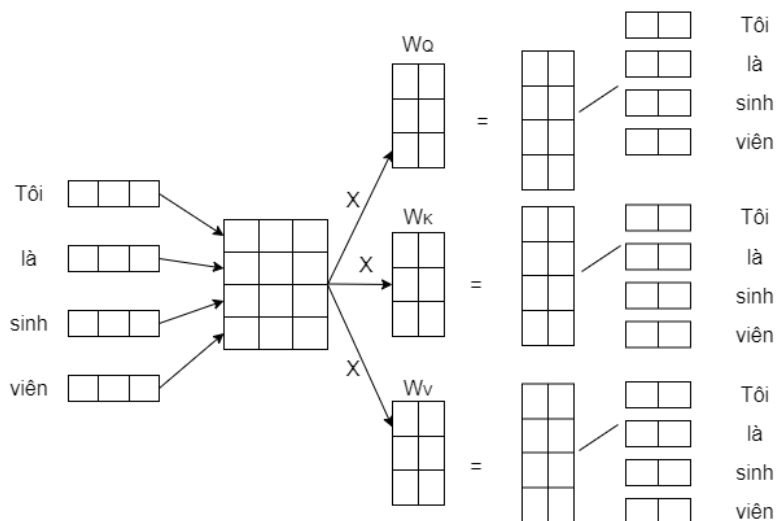
$$K = XW_K$$

$$V = XW_V$$

Trong đó:

X : là ma trận đầu vào.

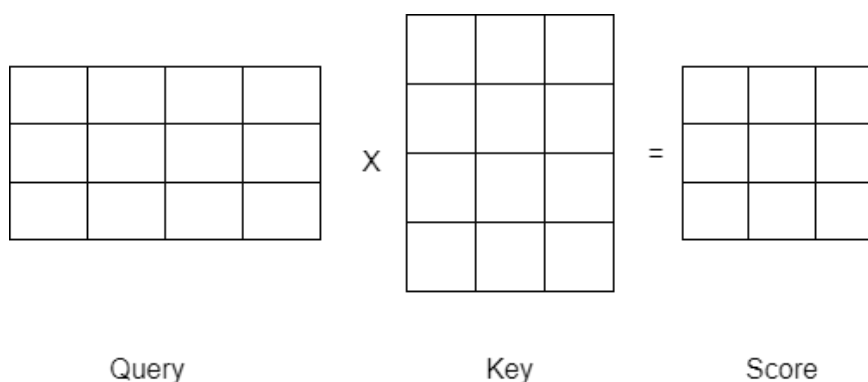
W_Q, W_K, W_V : là các ma trận trọng số được khởi tạo và được cập nhật trong quá trình huấn luyện.



Hình 2. 5 Tạo ma trận Q, K, V

Giải thích query, key, value: ví dụ tìm kiếm một nội dung trên google như “NLP” thì query ở đây là “NLP” thì google sẽ so sánh với key (tiêu đề các bài viết, nội dung, ...) sau đó sẽ trả lời về các kết quả thì đó là value.

Đầu tiên, thực hiện nhân ma tra trận query và ma trận key lại với nhau. Phép tính này nhằm tìm ra mối quan hệ trọng số giữa các cặp từ.



Hình 2. 6 Score matrix

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

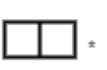




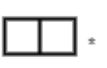



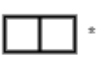



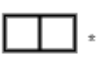



Score matrix là ma trận thể hiện sự tương quan giữa query và key.

Sau đó, score matrix được scale lại bằng cách chia cho căn bậc 2 của chiều query hoặc key. Điều này giúp gradient ổn định hơn.

$$\frac{\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}}{\sqrt{d_k}} = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \text{ Scaled Scores}$$

Hình 2. 7 Scale score

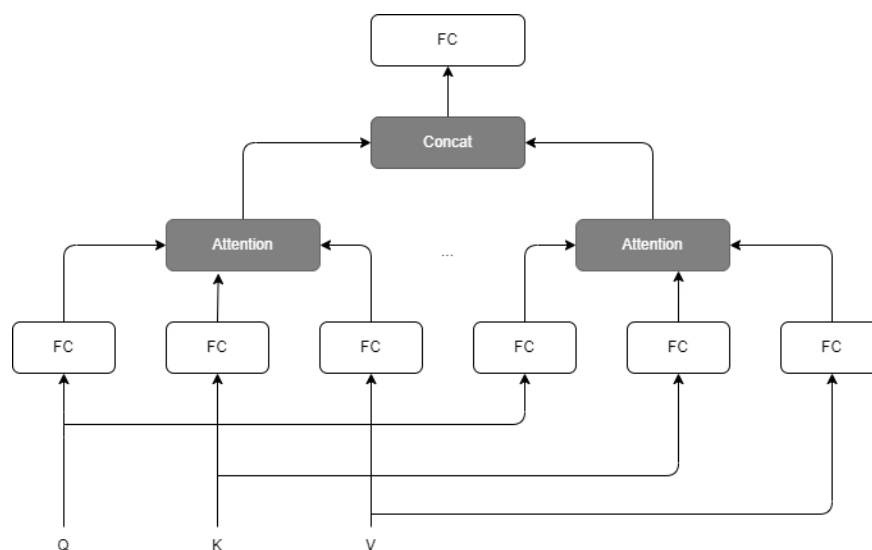
Tuy nhiên, điểm số sau khi nhân lại với nhau chưa được chuẩn hóa. Do đó, chúng ta cần chuẩn hóa bằng một hàm softmax để đưa về một phân phối xác suất mà độ lớn của nó sẽ đại diện cho mức độ chú ý của từ query tới key.

Tôi	Query * Key.T		Softmax	Value	Softmax * Value	Sum
Tôi * tôi	 * 	= 100	0.7			
Tôi * là	 * 	= 90	0.2			
Tôi * sinh	 * 	= 70	0.08			
Tôi * viên	 * 	= 5	0.02			

Hình 2. 8 Ví dụ tìm đặc trưng của từ "Tôi"

Sau khi qua hàm softmax ta được ma trận trọng số. Tiếp theo ta nhân ma trận trọng số với ma trận value ta sẽ được một ma trận mới có được sự chú ý vào các phần quan trọng.

2.2.2 Multi-head Attention



Hình 2. 9 Multi-head Attention

Như vậy sau quá trình scale dot product chúng ta sẽ thu được một ma trận attention. Các tham số mô hình cần tinh chỉnh là các ma trận \mathbf{W} . Mỗi quá trình như vậy gọi là một head của attention.

Trên thực tế, với cùng một tập hợp các query, key và value, chúng ta muốn mô hình của mình kết hợp kiến thức từ các hành vi khác nhau của cùng một cơ chế chú ý, chẳng hạn như nắm bắt sự phụ thuộc của các phạm vi khác nhau (ví dụ: phạm vi ngắn hơn so với phạm vi dài hơn) trong một chuỗi.

Do đó, có thể có lợi khi cho phép cơ chế chú ý của chúng ta sử dụng các không gian con biểu diễn khác nhau của các query, key và value. Với mỗi giá trị query, key và value sẽ học được các đặc trưng khác nhau.

Vì vậy, thay vì thực hiện một tập hợp sự chú ý duy nhất, các query, key và value có thể được chuyển đổi với \mathbf{h} các phép chiếu tuyến tính đã học độc lập. Sau đó, các query, key và value dự kiến này được đưa vào nhóm sự chú ý song song.

Trước khi cung cấp việc triển khai sự chú ý từ nhiều phía, chúng ta hãy chính thức hóa mô hình này về mặt toán học.

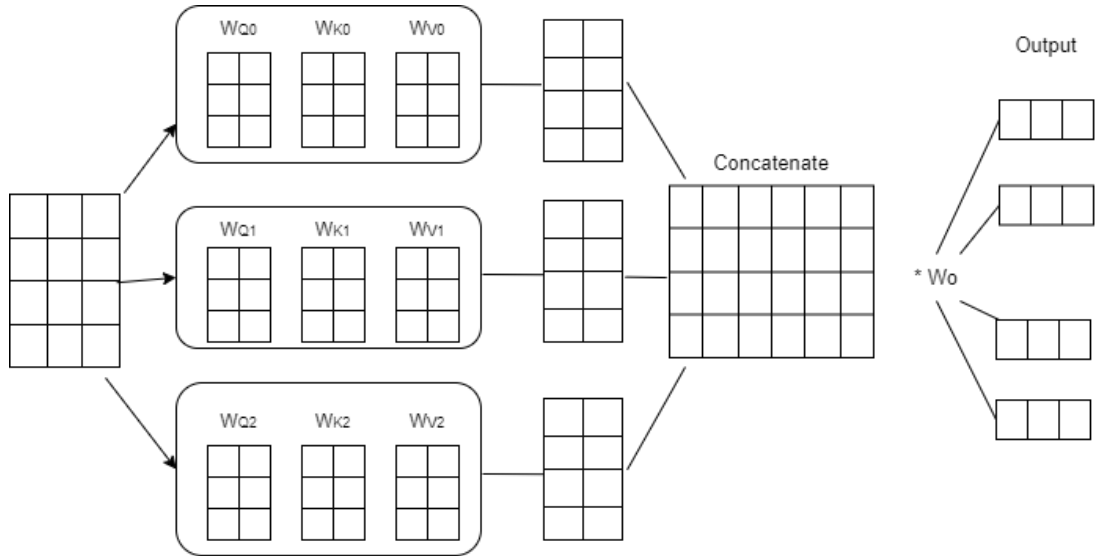
Đưa ra một query $q \in \mathbb{R}^{d_q}$, một key $k \in \mathbb{R}^{d_k}$, một value $v \in \mathbb{R}^{d_v}$, mỗi đầu chú ý h_i ($i = 1, \dots, h$) được tính là:

$$h_i = f(W_i^{(q)}q, W_i^{(k)}k, W_i^{(v)}v) \in \mathbb{R}^{d_v}, \quad (8)$$

nơi các thông số có thể học được $W_i^{(q)} \in \mathbb{R}^{p_q \times d_q}$, $W_i^{(k)} \in \mathbb{R}^{p_k \times d_k}$ và

$W_i^{(v)} \in \mathbb{R}^{p_v \times d_v}$, và f là tổng hợp sự chú ý. Đầu ra sự chú ý nhiều đầu là một phép biến đổi tuyến tính khác thông qua các tham số có thể học được $W_o \in \mathbb{R}^{p_o \times hp_v}$ của sự nối các đầu h .

Cuối cùng, các đầu ra tổng hợp sự chú ý được nối và biến đổi với một phép chiếu tuyến tính đã học khác để tạo ra đầu ra cuối cùng. $W_o \begin{bmatrix} h_1 \\ \dots \\ h_h \end{bmatrix} \in \mathbb{R}^{p_o}$



Hình 2. 10 Mô hình hóa tính toán multi head attention

2.3 TF-IDF

Trọng số từ (TF – term frequency) góp phần quan trọng trong nâng cao độ chính xác của phân lớp văn bản dài, trong khi với câu hỏi các từ hầu như chỉ xuất hiện một lần duy nhất. Do đó việc biểu diễn trọng số của mỗi từ trong câu hỏi tại những ngữ cảnh

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

khác nhau sẽ hết sức khó khăn cho nên việc sử dụng TF không có ý nghĩa cho câu hỏi.

Mô hình đánh trọng số từ nhằm tìm ra từ quan trọng trong câu hỏi hỗ trợ cho việc tìm kiếm đoạn văn có chứa câu trả lời.

Phương pháp TF-IDF: là một thống kê số học nhằm phản ánh tầm quan trọng của từ bên trong tài liệu.

Công thức TF:

$$TF(t, d) = \frac{f(t, d)}{\max \{f(w, d) : w \in d\}} \quad (9)$$

Trong đó:

$TF(t, d)$: tần suất xuất hiện của từ t trong văn bản d .

$f(t, d)$: số lần xuất hiện của từ t trong văn bản d .

$\max(\{f(w, d) : w \in d\})$: số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d .

Công thức IDF:

$$IDF(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (10)$$

Trong đó:

$IDF(t, D)$: giá trị idf của từ t trong tập văn bản.

$|D|$: Tổng số văn bản trong tập D .

$|\{d \in D : t \in d\}|$: thể hiện số văn bản trong tập D có chứa từ t .

Cụ thể, chúng ta sẽ có công thức tính TF-IDF hoàn chỉnh là:

$$TF - IDF = TF(t, d) * IDF(t, D) \quad (11)$$

Nhưng đối với câu hỏi thì TF-IDF không được hiệu quả cao bởi vì tần xuất hiện của các từ trong câu hỏi gần như tương tự nhau. Ví dụ như câu hỏi: “Bạn tên gì?” thì đối

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

với câu hỏi ngắn như thế này thì TF và IDF của từng từ sẽ bằng nhau và lần lượt là 1.0 và 0.0 cho nên TF-IDF của từng từ sẽ là 0.0. Cho nên, nó sẽ không phản ánh được tầm quan trọng của các từ bên trong câu hỏi.

Mô hình tìm trọng số sẽ phân tích câu hỏi bao gồm loại câu hỏi, pos tags của các từ trong câu hỏi và trích xuất đặc trưng về ngữ cảnh của câu hỏi.

Pos tags (part-of-speech tagging): gán nhãn từ loại. Nó là quá trình đánh dấu một từ trong văn bản tương ứng với một từ loại nào đó vào bối cảnh văn phạm của từ đó. Công việc gán nhãn cho một văn bản là xác định từ loại của mỗi từ trong phạm vi văn bản đó, tức là phân loại các từ vào các lớp từ loại của ngôn ngữ đó.

Ví dụ: “Thomas Edison phát minh bóng đèn” thì Thomas Edison sẽ là danh từ, phát minh là động từ, bóng đèn là danh từ.

Mô hình tìm trọng số của các từ bên trong câu hỏi sẽ được chúng tôi trình bày bên dưới tại mục 3.2.

2.4 BM25 (Best matching)

Tìm kiếm tài liệu là quá trình trích chọn những tài liệu có khả năng chứa câu trả lời trong số nhiều tài liệu được cung cấp.

Quá trình này giúp giảm thiểu được đầu vào của mô hình trả lời câu hỏi, loại bỏ các đoạn văn không cần thiết. Nó làm cho mô hình trả lời câu hỏi được chính xác hơn, giảm thiểu thời gian tìm câu trả lời.

Trong tìm kiếm thông tin, Okapi BM25 [5] là hàm tính thứ hạng được các công cụ tìm kiếm sử dụng để xếp hạng các văn bản theo độ phù hợp với truy vấn nhất định. Hàm xếp hạng này dựa trên mô hình xác suất, được phát minh vào những năm 1970 – 1980. Phương pháp tìm kiếm này có tên là BM25 (BM – Best Matching), nhưng thường được gọi là Okapi BM25 vì được sử dụng lâu đầu tiên trong hệ thống tìm kiếm Okapi. BM25 là một hệ thống xếp hạng dựa trên TF-IDF. Hàm này tìm kiếm trên một tổ hợp từ và xếp hạng các tập tài liệu dựa trên từ truy vấn trong tài liệu mà

không quan tâm đến quan hệ của từ đó bên trong văn bản. Đây cũng chính là nhược điểm mà BM25 gặp phải khi trong truy vấn và trong tài liệu của những từ cùng nghĩa nhưng khác cách viết.

Công thức BM25:

$$\sum_i^n IDF(q_i) \frac{f(q_i, D) \times (k_1 + 1)}{f(q_i, D) + k_1 \times (1 - b + b \times \frac{fieldLen}{avgFieldLen})} \quad (12)$$

Trong đó:

$IDF(q_i)$ là nghịch đảo tần suất của tài liệu thứ i trong q (câu truy vấn).

q là viết tắt của từ query tức là câu truy vấn.

q_i là từ thứ i trong câu truy vấn.

D là tài liệu.

b là một hằng số ($b = 0.75$).

k_1 là một hằng số. $k_1 \in [1.2, 2.0]$.

$fieldLen$ là độ dài của tài liệu.

f là tần suất xuất hiện của từ q_i trong tài liệu D .

$avgFieldLen$ là độ dài trung bình của tài liệu.

Công thức của $IDF(q_i)$:

$$IDF(q_i) = \ln \left(1 + \frac{(docCount - f(q_i) + 0.5)}{f(q_i) + 0.5} \right) \quad (13)$$

Trong đó:

$docCount$ là tổng số lượng tài liệu.

$f(q_i)$ là số lượng tài liệu chứa q_i .

CHƯƠNG 3 MÔ HÌNH

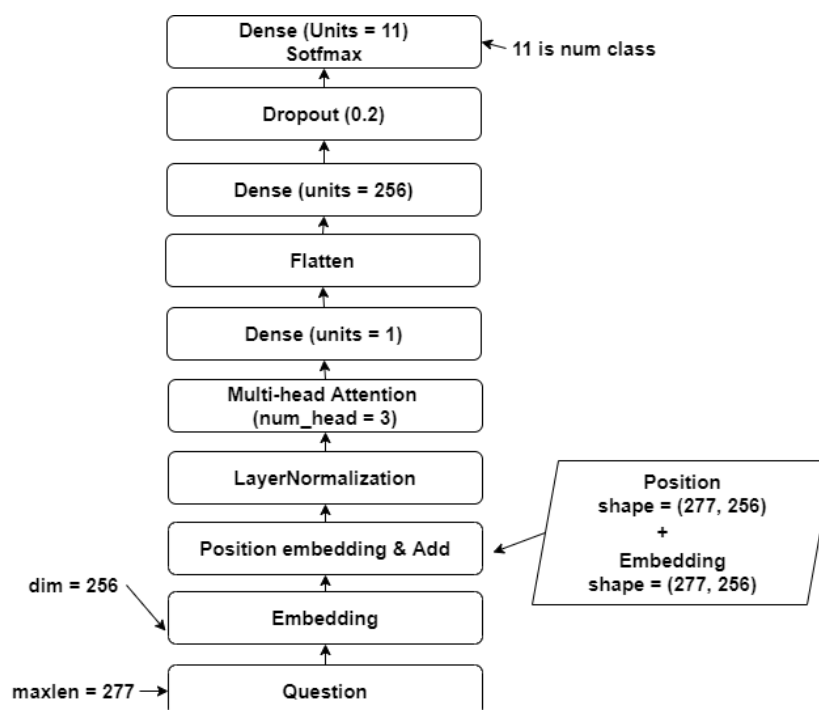
3.1 Mô hình phân loại câu hỏi

Phân loại câu hỏi có ý nghĩa quan trọng trong việc phân tích câu hỏi. Nó giúp xác định loại ngữ nghĩa của câu trả lời. Ví dụ câu hỏi “Ai phát minh ra bóng đèn?” thì câu trả lời là “Thomas Edison” hoặc “một nhà phát minh người Mỹ”. Trong khi câu hỏi “Tên người phát minh ra bóng đèn điện?” thì chỉ chấp nhận câu trả lời “Thomas Edison”. Điều này cho thấy việc quan trọng trong tổ chức câu hỏi.

Có nhiều cách để xác định loại câu hỏi như: xây dựng mô hình học máy thống kê, các kỹ thuật trong xử lý ngôn ngữ tự nhiên, xác định bằng so khớp các quan hệ của các từ có sẵn ...

Ở chúng tôi sử dụng mô hình học máy trong xử lý ngôn ngữ tự nhiên.

Phân loại các nhãn câu hỏi như: Ai, cái gì, con vật, như thế nào, con số, tại sao, điểm thời gian, khoảng thời gian, thực vật, yes/no, vị trí.



Hình 3. 1 Cấu trúc mô hình sử dụng Multi-head Attention

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Đầu vào được tách từ bằng thư viện `vncorenlp`⁴ và chuyển thành số với vị trí (gọi là các tokens) trong từ điển để đưa vào huấn luyện mô hình.

Giống như các mô hình xử lý ngôn ngữ tự nhiên khác chúng ta cần chuyển các tokens của đầu vào thành các vector có d chiều nhờ các mô hình ngôn ngữ như `word2vec` hay `fasttext` ở đây mô hình sử dụng lớp nhúng. Lớp này sẽ tạo một ma trận có kích thước (`vocab_size` x d) ma trận trọng số này sẽ được cập nhật trong quá trình huấn luyện mỗi từ trong từ điển sẽ ứng với mỗi dòng trong ma trận và với mỗi câu tạo thành ma trận có kích thước (`maxlen` x d). Với mô hình sử dụng multi-head attention thì câu được đưa vào tính toán song song nên cần một cơ chế để đánh dấu vị trí cho mỗi tokens đầu vào. Để giải quyết vấn đề vị trí của từ, mô hình này sẽ sử dụng một lớp nhúng khác để học bộ trọng số để làm sao có thể phân biệt được vị trí tokens đầu vào. Bộ trọng số của mã hóa vị trí này sẽ có kích thước (`maxlen` x d). Vector nhúng sẽ được cộng với trọng số này để đánh dấu vị trí cho từ.

Đầu ra sau khi qua lớp nhúng thì sẽ đi qua lần lượt các lớp: Normalization, Multi-head Attention (`heads` = 3), fully connected với `unit` = 1 để bỏ đi chiều cuối tức là còn lại (`batch_size` x `maxlen`).

Cuối cùng ta cho qua các lớp: fully connected (`units` = d), dropout (0.2), fully connected (`units` = 11) với hàm kích hoạt là softmax cho ra xác suất của đối tượng phân loại.

3.2 Mô hình tìm trọng số của từ

Mô hình tìm trọng số sẽ phân tích câu hỏi bao gồm loại câu hỏi và postags của các từ trong câu hỏi.

Pos tags (part-of-speech tagging): gán nhãn từ loại. Nó là quá trình đánh dấu một từ trong văn bản tương ứng với một từ loại nào đó vào bối cảnh văn phạm của từ đó.

⁴ <https://github.com/vncorenlp/VnCoreNLP>

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

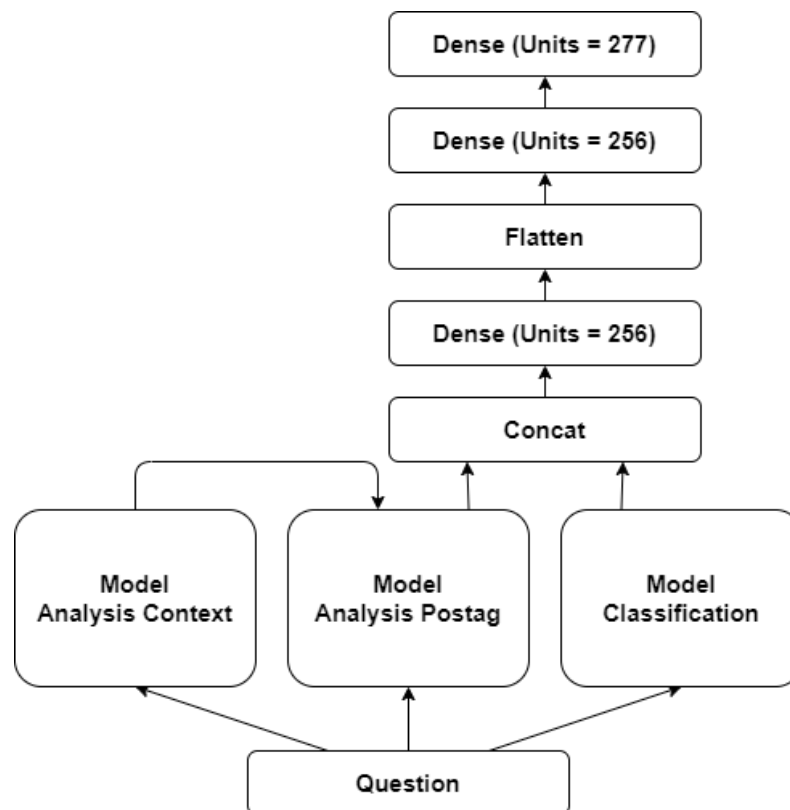
Công việc gán nhãn cho một văn bản là xác định từ loại của mỗi từ trong phạm vi văn bản đó, tức là phân loại các từ vào các lớp từ loại của ngôn ngữ đó.

Đầu vào sẽ chia thành 3 hướng cho 3 tác vụ để phân tích:

Trích xuất đặc trưng ngữ cảnh cho câu hỏi.

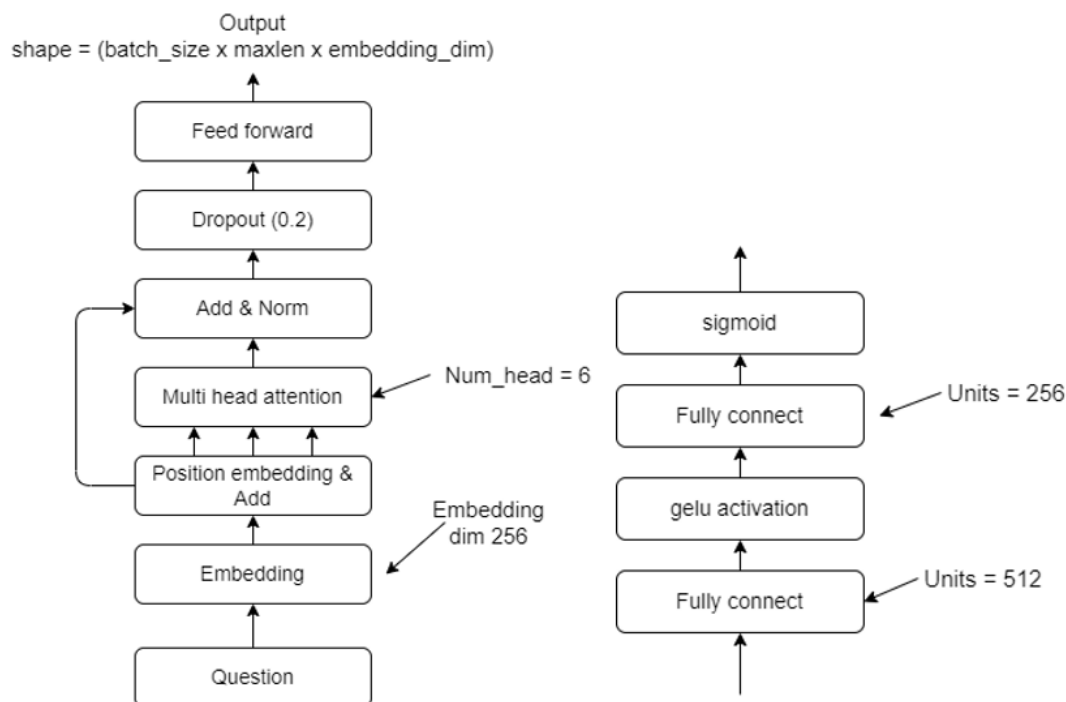
Phân tích pos tags cho câu hỏi.

Phân tích loại câu hỏi: hỏi về cái gì, người hay con vật, ...



Hình 3. 2 Cấu trúc mô hình đánh trọng số của từ

3.2.1 Trích xuất đặc trưng ngữ cảnh



Hình 3. 3 Trích xuất đặc trưng ngữ cảnh và feedforward

Mô hình sẽ tái sử dụng lại 2 lớp của mô hình phân tích câu hỏi là lớp nhúng và lớp đánh dấu vị trí để giảm kích thước mô hình và tránh việc huấn luyện không cần thiết.

Lớp tiếp theo là multi-head attention với heads = 6 để tìm vector ngữ cảnh cho câu hỏi.

Sẽ có một kết nối từ nhúng vị trí đến đầu ra của multi-head attention để bổ sung thêm thông tin trước đó để tránh mất mát thông tin trong lúc tính attention. Tiếp theo là 1 lớp normalization.

Sau đó qua một lớp dropout với tỉ lệ 20%.

Lớp cuối cùng là 1 lớp feedforward với 2 lớp fully connected liên tiếp và hàm kích hoạt là gelu [8] và sigmoid ở lớp cuối để loại bỏ đi những đặc trưng không quan trọng khi sử dụng kết nối từ lớp đầu tiên (nó chứa quá nhiều thông tin không cần thiết).

Công thức:

$$O_1 = (x \times W_1 + b_1) \times P(X \leq x) \text{ với } P(x) \sim N(0,1) \quad (14)$$

$$O_2 = \frac{1}{1 + e^{-(o_1 \times W_2 + b_2)}} \quad (15)$$

3.2.2 Phân tích pos-tags cho câu hỏi

Đầu vào của postags là một ma trận với mỗi câu sẽ là có kích thước ($\text{maxlen} \times \text{num_pos_tag}$).

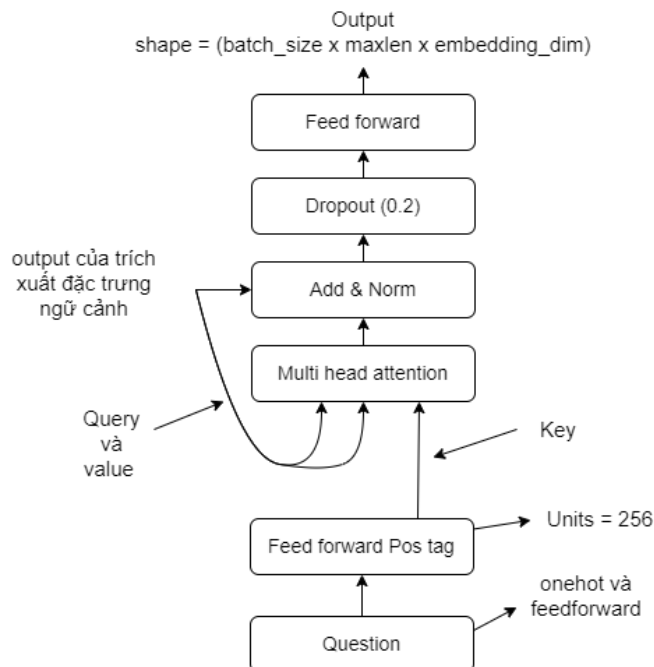
Trong đó:

maxlen : là độ dài tối đa của câu hỏi được đệm với giá trị 0.

num_pos_tag : là số lượng loại pos tags bao gồm 25 loại.

Mỗi từ one hot thành vector có giá trị 0 và 1, với 1 là vị trí postag của câu hỏi trong danh sách postags.

Tiếp theo là multi-head attention với value là postag, query và key là kết quả của việc phân tích đặc trưng ngữ cảnh. Để tìm mối quan hệ giữa ngữ cảnh và postag của các từ bên trong.



Hình 3. 4 Phân tích postag

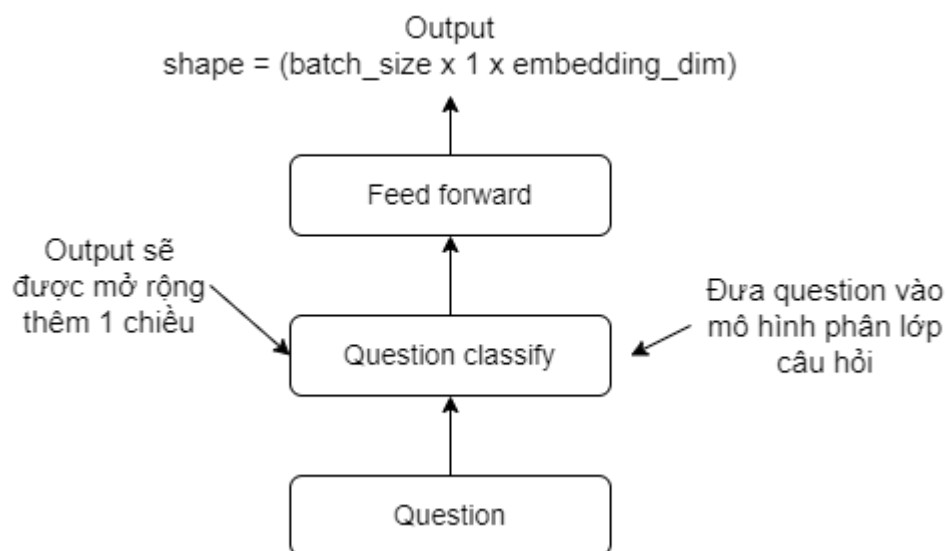
Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Output sẽ được cộng với output của phân tích ngữ cảnh sau đó qua dropout 20% và feedforward.

3.2.3 Phân tích loại câu hỏi

Khó khăn của việc phân tích câu hỏi là câu hỏi đầu vào của hệ thống hỏi đáp tự động là câu hỏi dưới dạng ngôn ngữ tự nhiên của người dùng. Vì vậy, việc phân tích câu hỏi cũng gặp những khó khăn của xử lý ngôn ngữ tự nhiên.

Mô hình phân tích câu hỏi bên trên sẽ được dùng lại thành một lớp trong mô hình đánh giá trọng số này để góp phần thêm sự chú ý vào các đối tượng bên trong câu hỏi. Ví dụ: “Ai phát minh ra bóng đèn điện”, thì loại câu hỏi sẽ là con người từ cần chú ý ở đây sẽ là “Ai”.

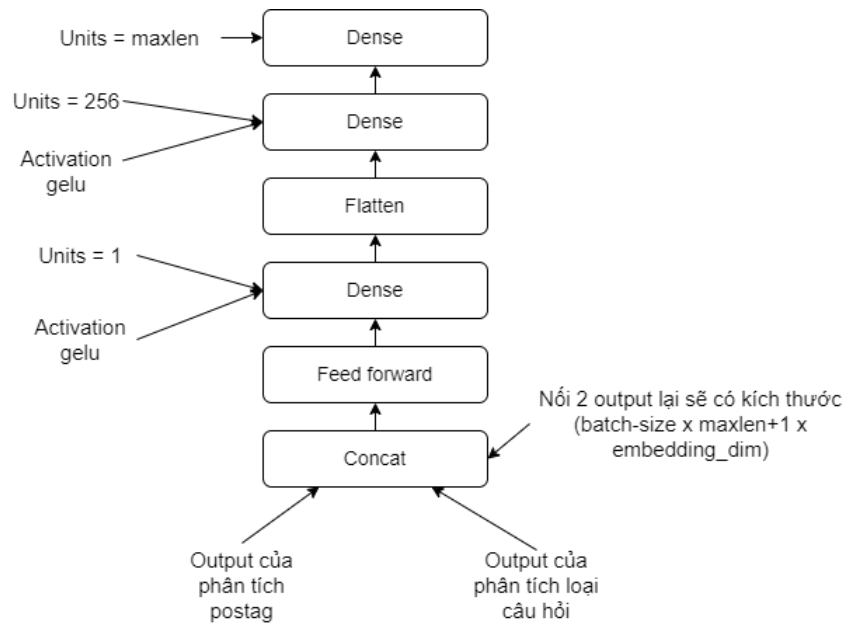


Hình 3. 5 Phân tích câu hỏi

Câu hỏi được đưa vào mô hình phân loại thì đầu ra của nó sẽ là xác suất của các nhãn với kích thước (batch size x 11). Tiếp theo cho qua feedforward.

Đầu ra của phân tích postag và phân tích loại câu hỏi sẽ được nối lại với nhau tạo thành ma trận (batch size x maxlen + 1 x d).

3.2.4 Kết hợp các mô hình và đưa ra trọng số



Hình 3. 6 Output của mô hình tìm trọng số

Đầu ra cuối cùng sẽ có kích thước (batch size x maxlen) với từng giá trị sẽ là trọng số của từ tương ứng.

3.3 Ứng dụng trọng số vào thuật toán BM25

Để giải quyết vấn đề của BM25 thì mô hình học sâu được sinh ra để giải quyết các nhược điểm bên trên bằng việc học các đặc trưng của từ, mối quan hệ của các từ trong văn bản và giữa các từ trong văn bản với nhau. Nó được kết hợp mô hình trọng số nhằm xác định được mức độ quan trọng của các từ trong câu truy vấn.

Đối với một truy vấn sẽ mang một trọng số nhất định. Trọng số càng cao thì mức độ quan trọng của từ đó càng cao, tức là từ đó ý chính của câu hỏi. Sử dụng trọng số này để bổ sung tầm quan trọng cho từ đó. Mỗi từ sẽ được cộng với một trọng số nhất định. Từ quan trọng sẽ bổ sung trọng số cao hơn hẳn các từ còn lại làm nổi bật từ trọng tâm. Cụ thể công thức sẽ được bổ sung trọng số như bên dưới.

Công thức của $IDF(q_i)$:

$$IDF(q_i) = \ln \left(1 + \frac{(docCount - f(q_i) + 0.5 + w_i)}{f(q_i) + 0.5} \right) \quad (16)$$

Trong đó:

docCount là tổng số lượng tài liệu.

$f(q_i)$ là số lượng tài liệu chứa q_i .

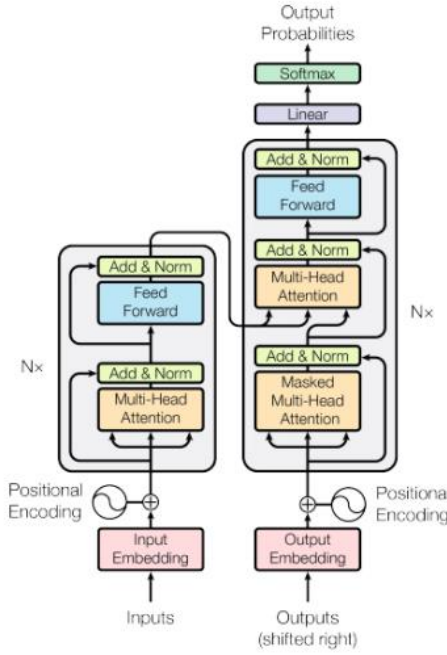
w_i là trọng số của từ q_i .

Công thức BM25:

$$BM25 = \sum_i^n IDF(q_i) \frac{f(q_i, D) \times (k_1 + 1 + w_i)}{f(q_i, D) + k_1 \times \left(1 - b + b \times \frac{fieldLen}{avgFieldLen} \right)} \quad (17)$$

3.4 Mô hình transformer

Trong bài báo Attention is all you need [3], tác giả đã sử dụng mô hình chú ý đã được trình bày tại mục [2.2](#). Giống như cái tên của bài báo, kiến trúc của mô hình đã loại bỏ toàn bộ kiến trúc RNN và thay vào đó là các lớp attention. Transformer cũng là một mô hình dạng Sequence to Sequence tức là đầu vào là một câu và đầu ra cũng là một câu cho nên quá trình tính toán cũng sẽ gồm 2 phần chính là bộ mã hóa (encoder) và bộ giải mã (decoder). Transformer không xử lý các thành phần trong chuỗi một cách tuần tự mà sẽ tính toán song song. Do tính năng này, Transformer có thể giảm đáng kể thời gian tính toán.



Hình 3. 7 Sơ đồ kiến trúc transformers⁵

Mã hóa vị trí (Position encoding): Đầu tiên, Input của mô hình được biểu diễn bằng một ma trận embedding với một từ là một vector. Nhưng Transformer xử lý các từ song song, do đó thì ma trận embedding không thể hiện được vị trí các từ. Do đó, sinh ra cơ chế mã hóa vị trí để giải quyết vấn đề vị trí của từ khi cho vào mô hình. Vị trí của các từ được mã hóa bằng một vector bằng với vector từ và được cộng vào vector từ. Cụ thể vị trí đó sẽ được tính theo công thức:

$$PE_{(pos, 2i)} = \sin\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right) \quad (18)$$

$$PE_{(pos, 2i+1)} = \cos\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right) \quad (19)$$

Trong đó:

PE: Position encoding.

pos: vị trí của từ trong chuỗi.

⁵ <https://arxiv.org/abs/1706.03762>

i : thứ tự chiều embedding.

d_{model} : chiều dài vector embedding.

	0	1	2	3	4	5
0	$\sin(\frac{1}{1000^{0/512}} \times 0)$	$\cos(\frac{1}{1000^{0/512}} \times 0)$	$\sin(\frac{1}{1000^{2/512}} \times 0)$	$\cos(\frac{1}{1000^{2/512}} \times 0)$	$\sin(\frac{1}{1000^{4/512}} \times 0)$	$\cos(\frac{1}{1000^{4/512}} \times 0)$
1	$\sin(\frac{1}{1000^{0/512}} \times 1)$	$\cos(\frac{1}{1000^{0/512}} \times 1)$	$\sin(\frac{1}{1000^{2/512}} \times 1)$	$\cos(\frac{1}{1000^{2/512}} \times 1)$	$\sin(\frac{1}{1000^{4/512}} \times 1)$	$\cos(\frac{1}{1000^{4/512}} \times 1)$
2	$\sin(\frac{1}{1000^{0/512}} \times 2)$	$\cos(\frac{1}{1000^{0/512}} \times 2)$	$\sin(\frac{1}{1000^{2/512}} \times 2)$	$\cos(\frac{1}{1000^{2/512}} \times 2)$	$\sin(\frac{1}{1000^{4/512}} \times 2)$	$\cos(\frac{1}{1000^{4/512}} \times 2)$

Hình 3. 8 Tính mã hóa vị trí⁶

Trong hình bên trên minh họa cho cách tính mã hóa vị trí với vector nhúng có 6 chiều và mỗi dòng ứng với một từ.

Encoder (mã hóa): bao gồm một loạt 6 lớp giống nhau được xếp chồng lên nhau. Mỗi lớp mã hóa sẽ được chia thành 2 phần nhỏ. Đầu tiên là cơ chế multi-head attention đã được đề cập ở mục 2.2 và thành phần thứ 2 là một mạng nơ-ron truyền thẳng. Ngoài ra, còn có skip connection và normalization layer. Ở đây áp dụng skip residual connection (kết nối giá trị đầu ra của lớp hiện tại và giá trị của lớp trước đó để tiếp tục tính toán) điều này giúp tránh việc mất mát thông tin khi truyền qua nhiều lớp trong mạng. Sau khi áp dụng residual connection thì kết quả sẽ được qua một lớp chuẩn hóa để cải thiện tốc độ huấn luyện. Phần mạng nơ-ron truyền thẳng có 3 lớp linear + relu + linear. Giá trị tại normalization layer sẽ được tính như sau:

$$Norm(x) = \frac{x - avg(X)}{\sqrt{\sigma^2 + 0.001}} \quad (20)$$

Trong đó:

Norm(x): giá trị của x sau khi normalize.

X: tập các giá trị cần normalize.

avg(X): trung bình của các giá trị đang xét.

⁶ <https://pbcquoc.github.io/transformer/>

σ : Độ lệch chuẩn.

Tham số 0.001 được cộng vào để tránh hiện tượng chia cho 0.

Decoder (giải mã): Cũng như bộ mã hóa bộ giải mã cũng mang 6 lớp nối tiếp nhau. Ngoài 2 thành phần giống bộ mã hóa, bộ giải mã chèn thêm phần tử thứ ba, nó sẽ thực hiện tính multi-head attention trên đầu ra của bộ mã hóa. Các thành phần của bộ giải mã được tính toán và áp dụng kỹ thuật residual connection tương tự như bộ mã hóa. Nhưng ở đây có sự chỉnh sửa multi-head attention đầu tiên để nhằm ngăn chặn việc tham gia giải mã của các từ không cần thiết. Việc che đi để phù hợp với việc dự đoán sau này (sử dụng các từ phía trước để dự đoán từ phía sau).

3.5 Mô hình BERT

BERT[4] là viết tắt của cụm từ Bidirectional Encoder Representation from Transformer, có nghĩa là mô hình biểu diễn từ theo 2 chiều ứng dụng kỹ thuật transformer. Là một kiến trúc mới cho bài toán Language Representation được Google, được tạo và xuất bản vào năm 2018 bởi Jacob Devlin và các đồng nghiệp của ông.

Điểm đặc biệt của BERT đó là có thể cân bằng bối cảnh ngữ nghĩa theo 2 chiều trái phải. Cơ chế của BERT là truyền toàn bộ các từ của câu văn đồng thời vào mô hình mà không cần quan tâm đến thứ tự của các từ đầu vào. Do đó, BERT được coi là huấn luyện 2 chiều (bidirectional) [9].

Non-context (không bối cảnh): Là các thuật toán không tồn tại bối cảnh trong biểu diễn từ. Đó là các thuật toán xử lý ngôn ngữ tự nhiên đời đầu như: word2vec, fasttext... Chúng ta chỉ có duy nhất một biểu diễn vector cho mỗi từ tại mọi bối cảnh khác nhau.

Uni-directional (một chiều): Là các thuật toán bắt đầu xuất hiện bối cảnh của từ. Các phương pháp dựa trên RNN là những phương pháp nhúng từ một chiều. Các vector nhúng được tạo ra nhờ bối cảnh một chiều từ trái sang phải.

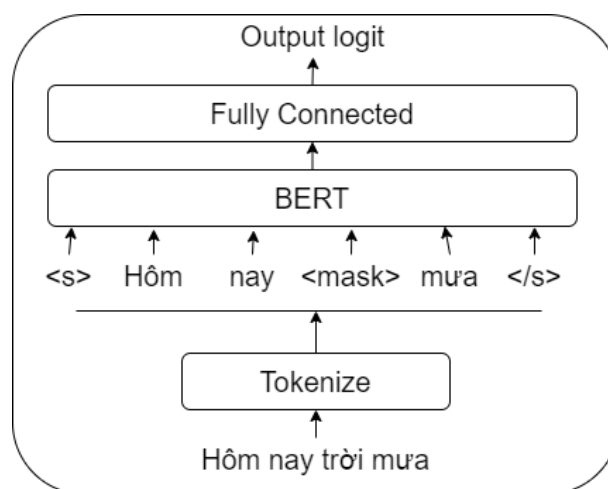
Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Bi-directional (hai chiều): ngữ nghĩa của từ không chỉ được biểu diễn ở từ liền trước mà còn được biểu diễn dựa vào các từ xung quanh. Đại diện cho các mô hình sử dụng kỹ thuật transformer.

Bert còn được coi như là đột phá lớn trong machine learning bởi vì khả năng ứng dụng của nó vào nhiều bài toán xử lý ngôn ngữ tự nhiên khác nhau: question answering, natural language inference,... với kết quả rất tốt.

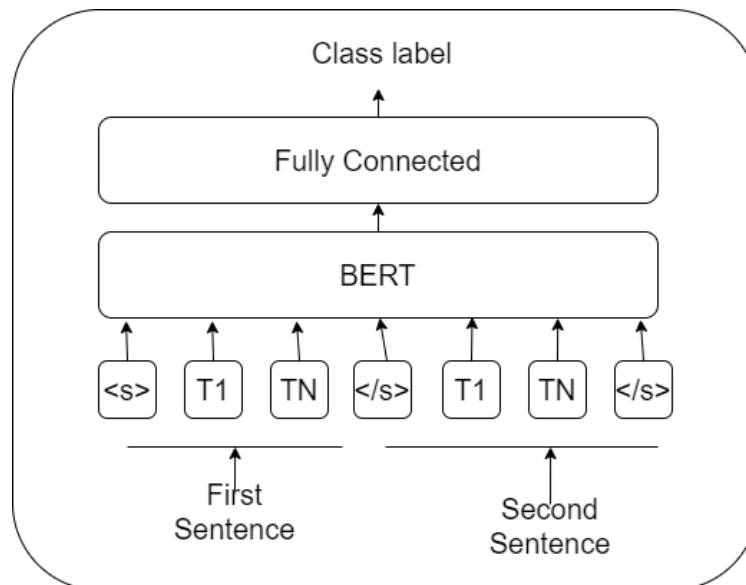
Pre-training BERT (mô hình tiền huấn luyện): Mô hình BERT được huấn luyện bởi hai tác vụ: Mô hình ngôn ngữ với mặt nạ (Masked language model) và dự đoán câu tiếp theo (Next sentence prediction).

Mô hình ngôn ngữ với mặt nạ là một tác vụ cho phép chúng ta fine-tuning lại các biểu diễn từ trên các bộ dữ liệu unsupervised bất kỳ. Thông thường khoảng 15% các token của câu sẽ được thay thế bằng mã mặt nạ trước khi truyền vào mô hình. mã mặt nạ sẽ đại diện cho từ bị che đi. Mô hình sẽ dựa trên các từ không được che xung quanh mã mặt nạ để dự đoán ra từ bị che đi mất. Đồng thời, các từ xung quanh mã mặt nạ cũng sẽ tạo nên bối cảnh cho từ bị che. Để tính toán xác suất cho từ đầu ra, chúng ta thêm một fully connected layer ngay sau layer cuối cùng. Hàm softmax có tác dụng tính toán phân phối xác suất. Số lượng units của fully connected sẽ bằng với kích thước của từ điển.



Hình 3. 9 Masked language model

Dự đoán câu tiếp theo là một bài toán phân loại học có giám sát với 2 nhãn. Đầu vào của mô hình sẽ là một cặp câu sao cho 50% câu thứ 2 được lựa chọn là câu tiếp theo của câu thứ nhất và 50% được lựa chọn một cách ngẫu nhiên từ bộ văn bản mà không có mối liên hệ gì với câu thứ nhất. Nhãn của mô hình sẽ tương ứng với isNext tức là câu tiếp theo và notNext tức là cặp câu không liên tiếp.



Hình 3. 10 Next Sentence Prediction

Mô hình BERT định nghĩa số lớp ẩn là L , kích thước lớp ẩn là H (hay còn gọi là kích thước của vector embedding) và số head ở attention là A . BERT chủ yếu có 2 kiến trúc:

BERT_{base}: $L = 12$, $H = 768$, $A = 12$

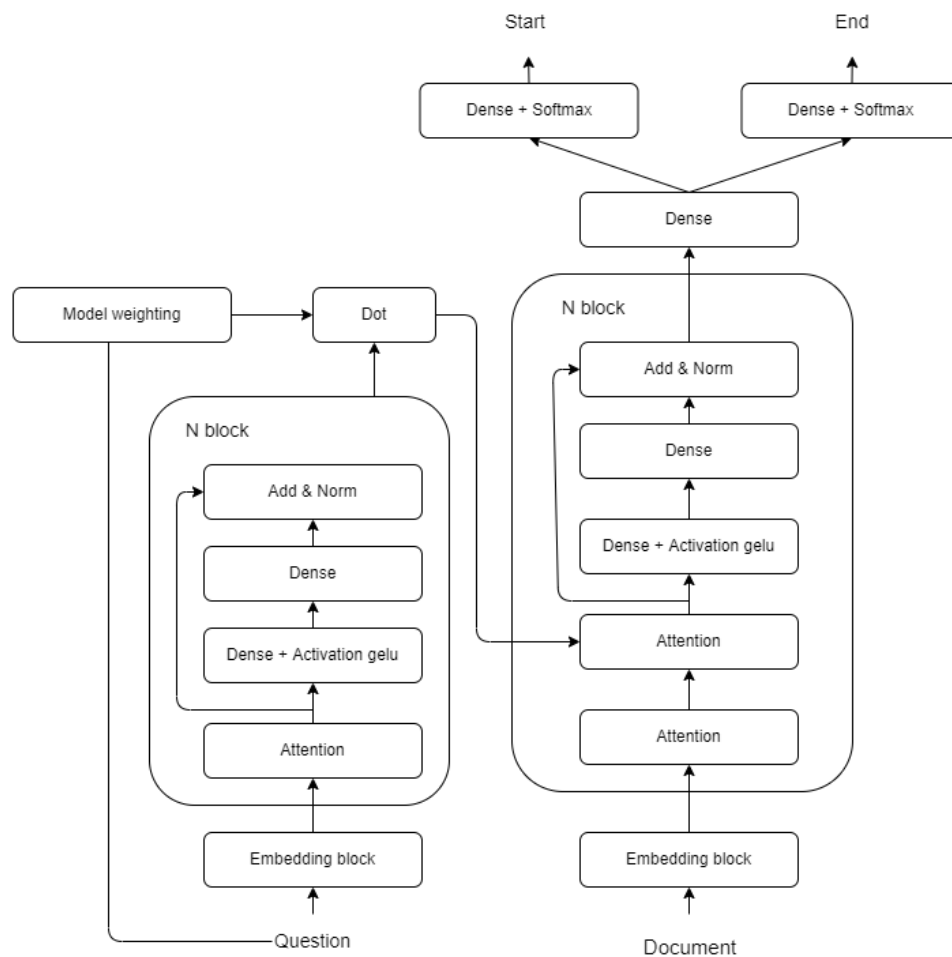
BERT_{large}: $L = 24$, $H = 1024$, $A = 16$

Mô hình RoBERTa [10]: 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 batchsize lớn hơn. Mô hình RoBERTa được đề xuất để cải tiến độ chính xác của mô hình BERT.

Mô hình DistilBERT [11]: 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. DistilBERT có 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.

3.6 Tinh chỉnh mô hình RoBERTa

Mô hình sẽ được tinh chỉnh thêm decoder để tách thành 2 đầu vào cho câu hỏi và đoạn văn khác với BERT thông thường là gộp 2 thành phần vào 1 để đưa vào mô hình. Chúng tôi có sử dụng mô hình trọng số ở cuối tầng mã hóa để tăng sự chú ý cho từ quan trọng.



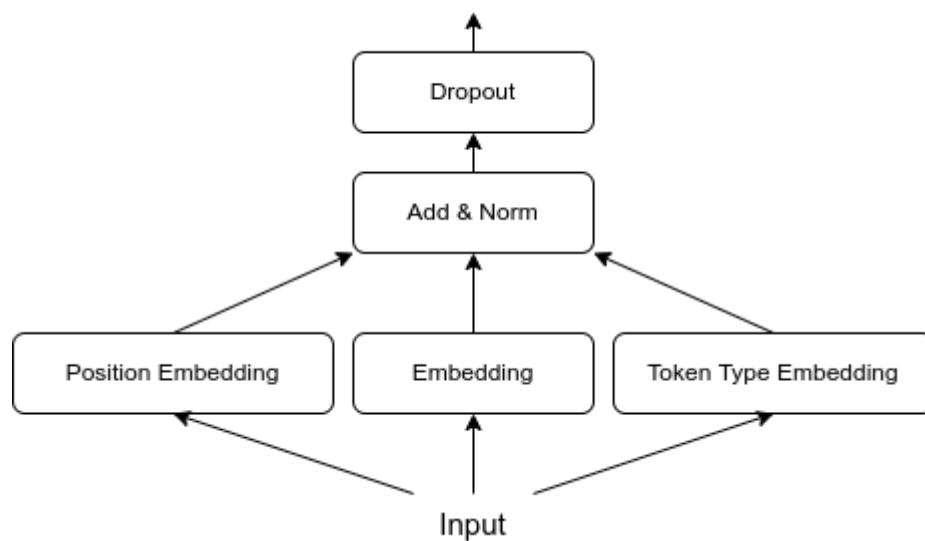
Hình 3. 11 Mô hình RoBERTa Encoder Decoder

3.6.1 *Khởi tạo vector nhúng*

Đầu vào là một câu với các từ sẽ được đánh dấu bằng vị trí của từ đó trong bộ từ điển gọi là tokens.

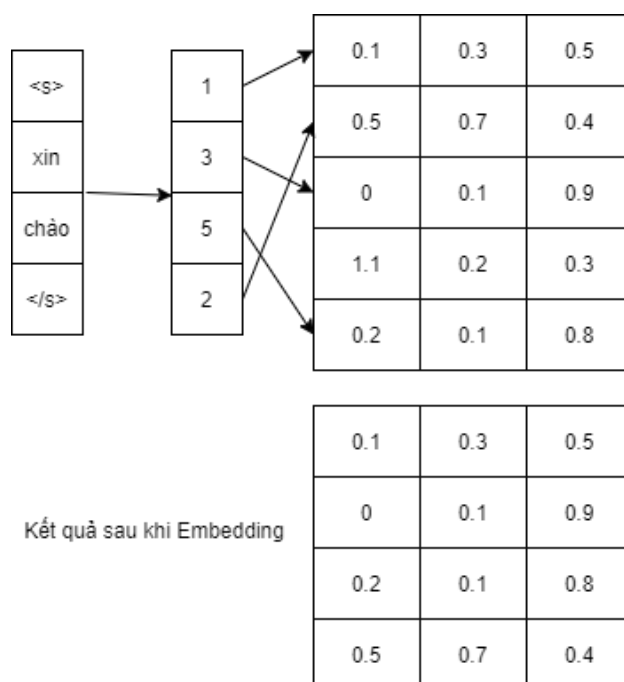
Khởi tạo vector nhúng có nhiệm vụ chuyển các tokens này thành vector nhúng có D chiều.

Đầu vào sẽ là các tokens cho qua các layer lần lượt là: embedding, position embedding, token type embedding có cấu trúc như hình bên dưới.



Hình 3. 12 *Embedding block*

Cả 3-layer đều tạo ra các ma trận trọng số hay còn gọi là ma trận nhúng với mỗi từ trong từ điển sẽ ứng với mỗi dòng trong ma trận.

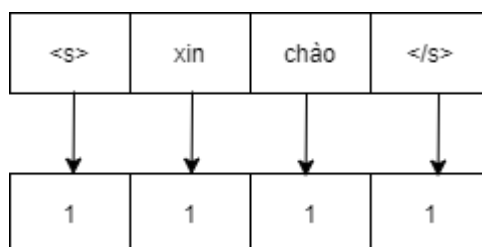


Hình 3. 13 Ảnh minh họa layer embedding hoạt động

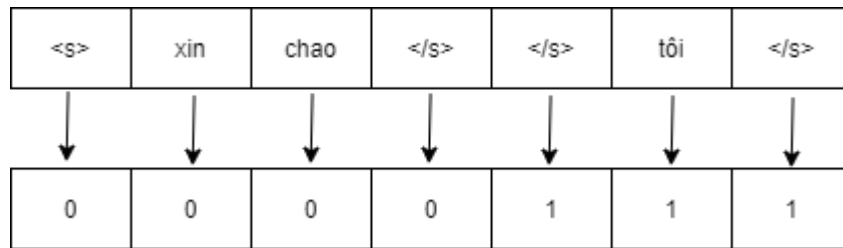
Với layer embedding sẽ được huấn luyện để tạo ra ma trận nhúng có kích thước (vocab size \times D). Với mỗi giá trị vị trí trong từ điển sẽ được ánh xạ vào ma trận nhúng để lấy ra vector nhúng. Ví dụ từ “anh” có vị trí trong từ điển là 3 thì sẽ ánh xạ đến dòng thứ 3 trong ma trận nhúng lấy ra vector có D chiều.

Position embedding sẽ đánh dấu vị trí cho từng từ. Đối với mô hình Bert dữ liệu đầu vào sẽ là một câu nên cần một cơ chế đánh dấu vị trí cho từng từ đầu vào.

Token type embedding dùng để đánh dấu các thành phần của đầu vào là một câu hoặc 1 cặp câu.



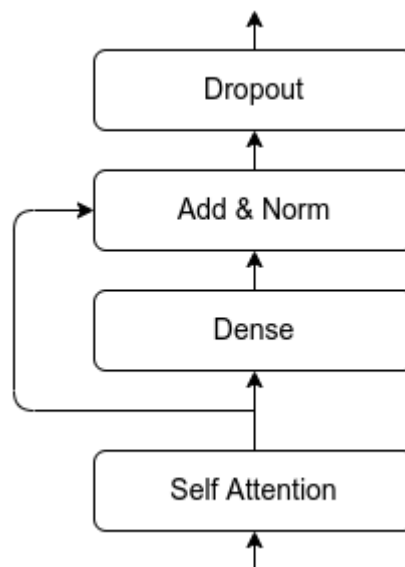
Hình 3. 14 Ví dụ token type cho một câu



Hình 3. 15 Ví dụ token type cho 2 câu

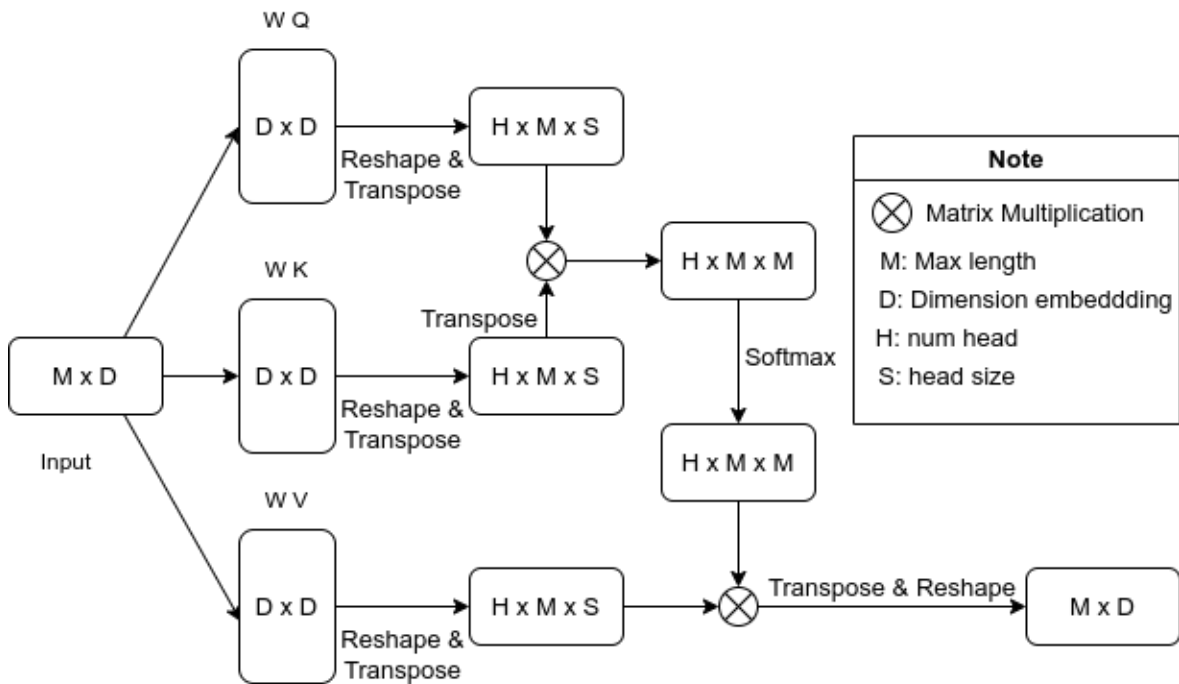
3.6.2 Attention

Đầu vào đi lần lượt đi qua các lớp self attention, densen và sẽ tạo một kết nối từ đầu ra tại self attention và đầu ra của dense sau đó là lớp normalization.



Hình 3. 16 Attention

Self attention cho phép mô hình khi mã hóa một từ thì có thể tìm kiếm sự liên quan của từ đó với cái từ khác trong câu. Cơ thể attention giống như cơ chế tìm kiếm. Với một từ cho trước, cơ chế này cho phép mô hình tìm các từ còn lại và thông tin sẽ được mã hóa dựa trên các từ còn lại.



Hình 3. 17 Các bước tính attention

Các embedding vector sau khi tính toán sẽ được truyền vào kiến trúc multi-head attention với nhiều khối.

Đầu tiên ma trận biểu diễn từ có kích thước ($M \times D$), sẽ được nhân với các ma trận WQ , WK , WV có kích thước ($D \times D$) tạo ra các ma trận Q , K , V . các ma trận WQ , WK , WV sẽ được cập nhật trong quá trình huấn luyện.

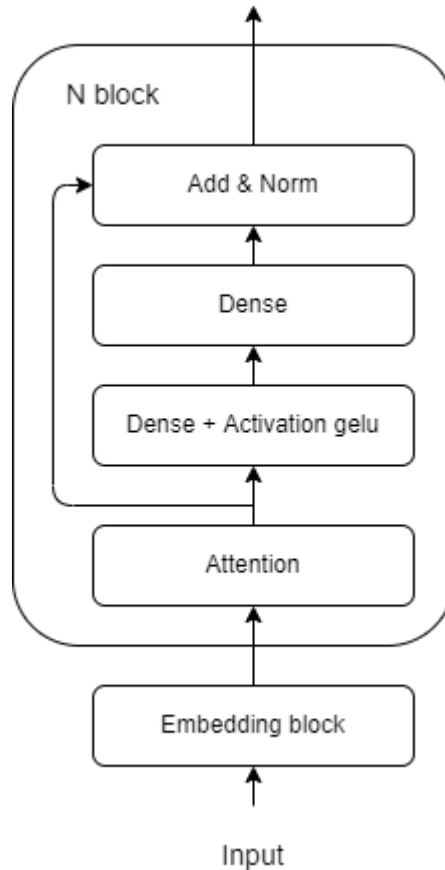
Mỗi ma trận Q , K , V sẽ được thay đổi kích thước và chuyển vị tại thành ma trận có kích thước ($H \times M \times S$). Với chiều cuối D sẽ được tách thành 2 phần là H và S tức là số lượng head H và kích thước mỗi head là S cho nên D phải chia hết cho H . Việc phân chia thành nhiều head khác nhau để mong đợi với mỗi head sẽ học được các đặc trưng khác nhau.

Để tính mối tương quan, chúng ta đơn giản chỉ cần nhân ma trận Q với ma trận chuyển vị của K . Tạo ra ma trận có kích thước ($H \times M \times M$). Sau đó chuẩn hóa bằng căn bậc 2 của chính nó và sau khi qua hàm softmax ta được ma trận chỉ số tương quan trong đoạn từ 0 đến 1.

$$score = softmax\left(Q \times \frac{K}{\sqrt{Q \times K}}\right) \quad (21)$$

Cuối cùng nhân điểm số này với ma trận V, chuyển vị và thay đổi kích thước về vị trí ban đầu.

3.6.3 Encoder (Bộ mã hóa)



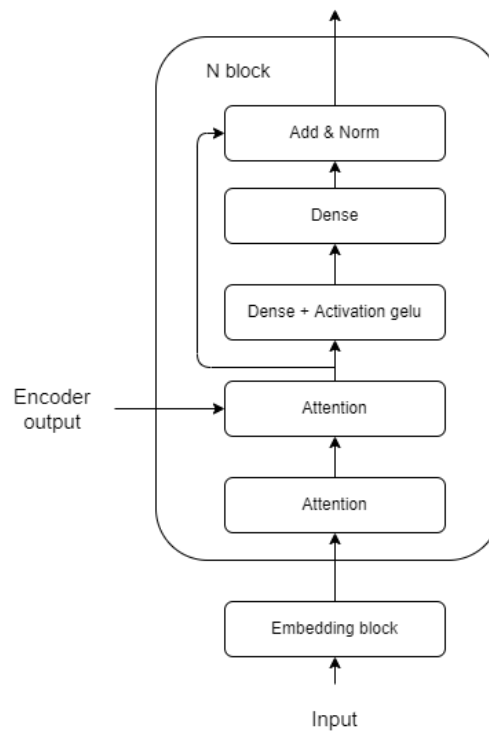
Hình 3. 18 Encoder trong Bert

Encoder sẽ bao gồm nhiều nhiều block xếp chồng lên nhau thường sẽ là 6 block, 12 block hoặc 24 block.

Tại mỗi layer encoder sẽ bao gồm một attention và các layer Dense, LayerNormalization.

Kết thúc quá trình mã hóa ta thu được đặc trưng của câu hỏi và sau đó ta sẽ nhân nó với trọng số của từng từ và chuyển sang decoder để tính encoder-decoder attention để tìm sự chú ý của câu hỏi vào đoạn văn.

3.6.4 Decoder (Bộ giải mã)

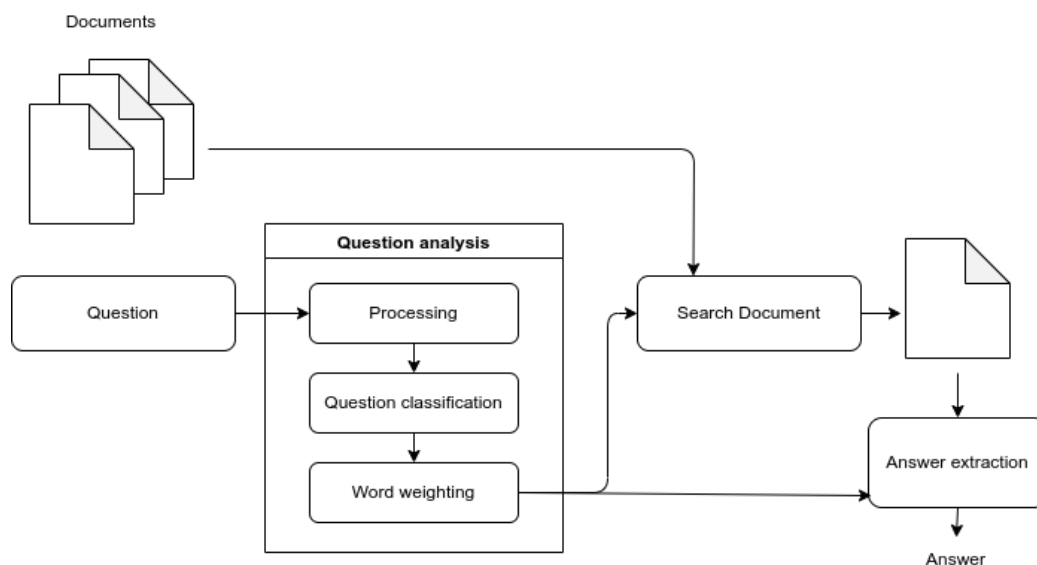


Hình 3. 19 Decoder trong Bert

Kiến trúc cũng bao gồm các lớp liên tiếp nhau. Bộ giải mã cũng bao gồm các lớp tương tự như bộ mã hóa nhưng khác ở chỗ bộ giải mã sẽ có thêm một attention gọi là encoder-decoder attention.

Encoder-decoder attention sẽ tìm sự chú ý của câu hỏi vào trong đoạn văn. Câu hỏi sau khi qua bộ mã hóa và nhân với bộ trọng số sẽ được đưa vào attention để tính key và value, còn query sẽ là attention trước đó trích xuất đặc trưng từ đoạn văn.

CHƯƠNG 4 THỰC NGHIỆM VÀ ĐÁNH GIÁ



Hình 4. 1 Các bước giải quyết bài toán

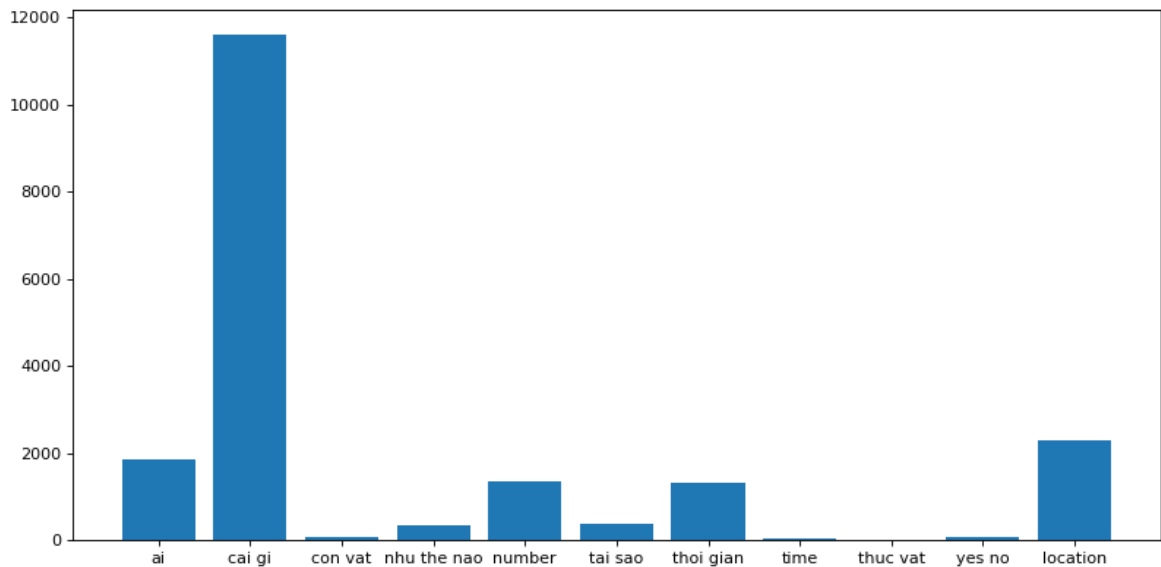
4.1 Kết quả mô hình phân loại câu hỏi

Bộ dữ liệu cho mô hình phân loại câu hỏi bao gồm nhiều chủ đề khác nhau và được gán nhãn gồm 19339 câu hỏi.

Dữ liệu bao gồm 11 nhãn, cụ thể là: Ai, cái gì, con vật, như thế nào, con số, tại sao, điểm thời gian, khoảng thời gian, thực vật, yes/no, vị trí.

	A	B
1	Question	Type
2	Hiện nay lăng mộ thừa tướng Lữ Gia nằm ở đâu	khu vuc
3	Thừa tướng Lữ Gia quê ở đâu	khu vuc
4	Tỉnh nào là nơi đầu tiên Sông Hồng đổ vào đất Việt	khu vuc
5	Hiện nay, nước ta có bao nhiêu sông lớn nhỏ	number
6	Sông bắt nguồn từ Tây Nguyên, được xem là có khởi nguồn nội địa, dài nhất nước ta	cai gi
7	Dòng sông huyền thoại nào ở Tây Sơn gắn liền sự phát tích của vương triều Tây Sơn	cai gi
8	Dòng sông Bạch Đằng nổi tiếng đổ ra biển ở tỉnh nào của nước ta	khu vuc
9	Loại đồ ăn nổi tiếng nào trong đêm Giáng sinh bắt nguồn từ Anh	cai gi
10	Thánh địa lớn nhất của đạo Cao Đài tên là gì	cai gi
11	Quần đảo Trường Sa được Việt Nam chia làm mấy cụm	number

Hình 4. 2 Dữ liệu phân loại câu hỏi



Hình 4. 3 Biểu đồ phân bố các lớp câu hỏi

Mô hình được huấn luyện trên colab có cấu hình: Device: Colab Pro, GPU Tesla P100- PCIE-16GB, RAM 12.6GB.

Hàm tối ưu Adam với: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$, $\text{learning_rate} = 0.001$.

Hàm mất mát: Categorical cross entropy

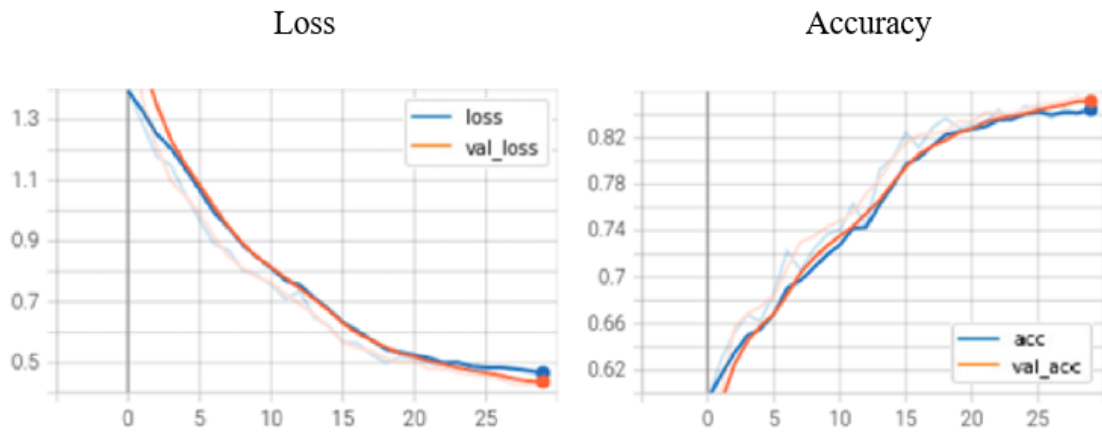
$$\text{Loss} = - \sum_{i=1}^{\text{output size}} y_i \times \log \hat{y}_i \quad (22)$$

Trong đó:

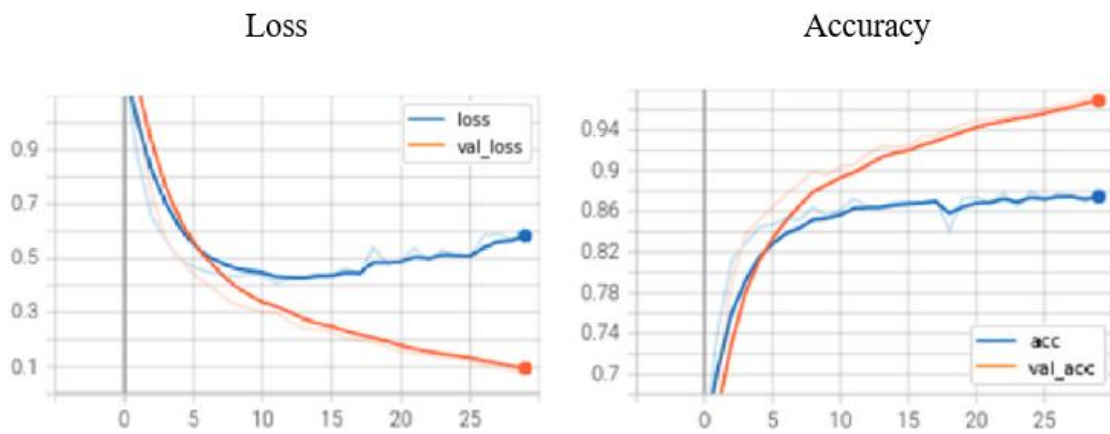
output size: số lượng nhãn.

y_i : là xác suất của đối tượng thứ i.

\hat{y}_i : là xác suất dự đoán của đối tượng thứ i.



Hình 4. 4 Kết quả mô hình phân loại sử dụng Multi-head Attention



Hình 4. 5 Kết quả mô hình phân loại sử dụng Bi_LSTM

Tiêu chí	Bi-LSTM	Multi-head Attention
Param	16.004.991	13.784.267
Learning rate	0.001	0.001
Batch size	64	64
Evaluate loss	0.42	0.46
Evaluate accuracy	0.88	0.84
F1-score	0.85	0.83
Epochs	30	30

Bảng 4.1 So sánh mô hình Bi-LSTM và Multi-head Attention

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

4.2 Kết quả mô hình tìm trọng số của từ

Dữ liệu bao gồm: 21073 câu hỏi với nhiều chủ đề khác nhau.

Dữ liệu được gán nhãn một cách thủ công. Theo như dữ liệu trên hình cột đầu tiên là câu hỏi đã được tách từ, cột thứ 2 là các từ được cho là quan trọng ý chính trong câu nó được chúng tôi đọc và trích ra một cách thủ công.

	A	B
1	hiện_nay_lăng_mộ_thừa_tướng_lữ_gia_nằm_ở_đâu , hôm_nay_trời_đẹp_quá	lăng_mộ_thừa_tướng_lữ_gia_nằm
2	thừa_tướng_lữ_gia_quê_ở_đâu	thừa_tướng_lữ_gia_quê
3	theo_mình_được_biết_sông_hồng_là_sông_lớn_ở_miền_bắc_tỉnh_nào_là_nơi_đầu_tiên_sông_hồ	theo_biết_sông_sông_miền_bắc_tỉnh_nơi_sông_hồng_đổ_đất_việt
4	hiện_nay , nước_ta_có_bao_nhiều_sông_lớn_nhỏ , nước_ta_có_rất_nhiều_sông_lớn_nhỏ	nước_sông_nước_sông
5	sông_bắt_nguồn_từ_tây_nguyên , được_xem_là_có_khởi_nguồn_nội_địa , dài_nhất_nước_ta	sông_bắt_nguồn_tây_nguyên_xem_khởi_nguồn_nội_địa_nước
6	dòng_sông_huyền_thoại_nào_ở_tây_sơn_gắn_liền_sự_phát_tích_của_vương_triều_tây_sơn	dòng_sông_huyền_thoại_tây_sơn_gắn_sự_phát_tích_vương_triều_tây_s
7	sông_bạch_đăng_gắn_liền_với_lịch_sử_chống_giặc_ngoại_xâm . dòng_sông_bạch_đăng_nổi_t	sông_đăng_gắn_lịch_sử_chống_giặc_ngoại_xâm_dòng_sông_bạch_đăng_d
8	loại_đồ_ăn_nổi_tiếng_nào_trong_đêm_giáng_sinh_bắt_nguồn_từ_anh_có_nhiều_món_ăn_nổi	đồ_ăn_đêm_giáng_sinh_bắt_nguồn_anh_món_ăn
9	đạo_cao_đài_là_một_đạo_lớn_ở_miền_nam_thì_cho_mình_hỏi_thánh_địa_lớn_nhất_của_đạo_cac	đạo_đài_đạo_miền_nam_hỏi_thánh_địa_đạo_cao_đài_tên
10	hôm_nay_mình_có_diệp_đi_ra_ngoài_biển_quần_đảo_trường_sa_được_việt_nam_chia_làm_m	hôm_nay_diệp_đi_ra_biển_quần_đảo_trường_sa_việt_nam_chia_cụm
11	einstein_được_trao_giải_thưởng_nobel_cho_lĩnh_vực_khoa_học_nào	einstein_trao_giải_thưởng_nobel_lĩnh_vực_khoa_học
12	albert_einstein_được_trao_giải_nobel_vào_năm_nào	albert_einstein_trao_giải_nobel_năm
13	tỉnh_bình_thuận_có_nhiều_đặc_sản_nổi_tiếng_thì_món_nào_nổi_tiếng_với_loại_cây_đặc_sản	tỉnh_bình_thuận_nhiều_đặc_sản_món_cây_đặc_sản
14	facebook_messenger_là_một_mạng_xã_hội_lớn_nhất_thế_giới_thì_phát_hành_lần_đầu_khi	messenger_mạng_xã_hội_thế_giới_phát_hành_lần_đầu_khi
15	người_dân_đi_biển_thường_dùng_từ_nào_để_nói_về_cái_chết_của_cá_voi	người_dân_đi_biển_dùng_nói_chết_cá_voi
16	hiện_nay_có_rất_nhiều_mạng_xã_hội_thì_cho_mình_hỏi_mạng_xã_hội_nào_tại_việt_nam_ra_đ	mạng_xã_hội_hỏi_mạng_xã_hội_việt_nam_ra_đời_năm
17	tỉnh_khánh_hoà_thuộc_khu_vực_nào_của_nước_ta	tỉnh_khánh_hoà_thuộc_khu_vực_nước
18	mã_độc_tấn_công_toàn_cầu_vào_ngày_12/5_có_tên_là_gì	mã_tấn_công_toàn_cầu_ngày_tên
19	ông_donald_trump_là_một_tỉ_phủ_tài_năng_và_tranh_cử_tổng_thống_và_cho_mình_hỏi_ông	donald_trump_tỉ_phủ_tài_năng_tranh_cử_tổng_thống_hỏi_ông_nhậm_ư
20	chỉ_pu_là_một_ca_sĩ_trẻ_xinh_đẹp_thì_cho_mình_hỏi_cô_ấy_có_tên_thật_là_gì	chỉ_pu_ca_sĩ_hỏi_cô_tên
21	suboi_một_rapper_tài_năng_đã_rap_cho_tổng_thống_obama_và_cô_có_tên_thật_là	suboi_tài_năng_rap_tổng_thống_obama_cô_tên
22	đầu_là_tên_thật_của_ca_sĩ_bảo_thy	tên_ca_sĩ_bảo_thy
23	phạm_lưu_tuấn_tài_là_tên_của_thành_viên_nào_trong_nhóm_365	phạm_lưu_tuấn_tài_tên_thành_viên_nhóm
24	ca_sĩ_miu_lê_tên_thật_là_gì	ca_sĩ_miu_lê_tên

Hình 4. 6 Dữ liệu mô hình trọng số

Sau đó dữ liệu sẽ được chuyển thành số để đưa vào mô hình huấn luyện.

Với mỗi câu hỏi sẽ được chuyển thành số chính là vị trí của từ đó trong bộ từ điển.

Còn đối với đầu ra thì chúng ta sẽ dựa vào cột dữ liệu thứ 2, đối với những từ quan trọng chúng ta sẽ gán nhãn ngẫu nhiên từ 0.8 đến 0.95, những từ không quan trọng sẽ được gán từ 0.1 đến 0.3.

Câu hỏi	Từ quan trọng	Nhãn
Hiện nay		0.12
Lăng mộ	X	0.84
Thừa tướng	X	0.9
Lữ Gia	X	0.93
Năm	X	0.88

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

ở		0.2
đâu		0.22
Hôm nay		0.18
Trời		0.21
Đẹp		0.3
Quá		0.23

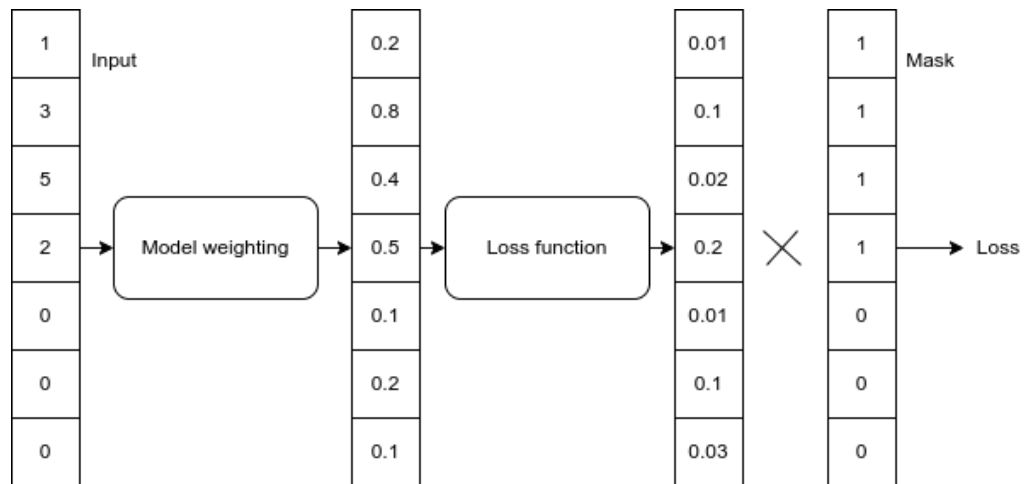
Sau đó mỗi câu sẽ được thêm các giá trị 0 vào cuối để dữ liệu đầu vào được bằng nhau và bằng câu dài nhất 277 từ.

Hàm kích hoạt cho mô hình: hàm tối ưu Adam với: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-7$, $\text{learning_rate} = 0.0001$.

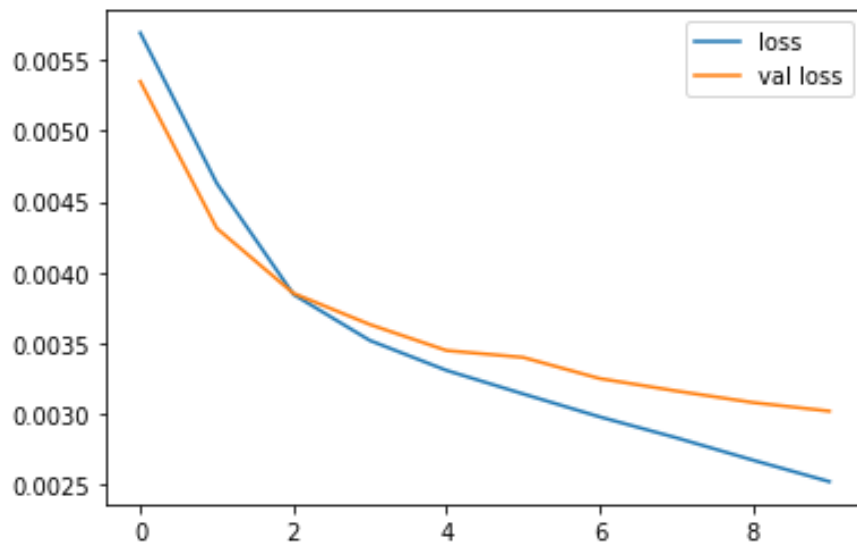
Hàm loss sẽ là hàm bình phương lỗi (Mean Squared Error - MSE).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 * mask \quad (23)$$

Trong đó: mask là ma trận có kích thước (batch size x maxlen) gồm 2 giá trị 0 và 1. 0 tương ứng với giá trị padding, 1 không padding.



Hình 4. 7 Ví dụ minh họa huấn luyện và tính loss cho mô hình



Hình 4. 8 Loss mô hình đánh trọng số của từ

```

không ==> 0.21
biết ==> 0.20
trách_nhiệm ==> 0.94
của ==> 0.23
phòng ==> 0.91
công_tác ==> 0.87
sinh_viên ==> 0.93
là ==> 0.84
làm ==> 0.74
gì ==> 0.26
ạ ==> 0.44
Em ==> 0.32
cám_ơn ==> 0.69

quy_chế ==> 0.89
đào_tạo ==> 0.87
theo ==> 0.89
hệ_thống ==> 0.84
tín_chỉ ==> 0.84
là ==> 0.27
gì ==> 0.28

chào ==> 0.15
admin ==> 0.08
em ==> 0.07
muốn ==> 0.11
hỏi ==> 0.57
niên_giám ==> 0.47
là ==> 0.44
gì ==> 0.27
ạ ==> 0.32
    
```

Hình 4. 9 Kết quả test model đánh trọng số của từ

Chúng ta sẽ so sánh kết quả với các thuật toán cũ trước đó. Các thuật toán cũ trước đó sẽ phân tích từ loại sau đó sẽ chỉ giữ lại danh từ làm từ quan trọng. Chúng ta sẽ sử dụng thư viện pyvi phân tích và so sánh với 5 từ có trọng số cao của mô hình.

Ví dụ 1: quy chế đào tạo theo hệ thống tín chỉ là gì.

	Trọng số từ mô hình	Từ loại pyvi
Quy chế	0.89	Danh từ

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Đào tạo	0.87	Động từ
Theo	0.89	Giới từ
Hệ thống	0.84	Danh từ
Tín chỉ	0.84	Danh từ
Là	0.27	Động từ
Gì	0.28	Danh từ

Nhận xét: Ý chính trong câu là hỏi về quy chế đào tạo về tín chỉ nếu chỉ giữ lại danh từ thì sẽ bỏ đi các từ quan trọng như “đào tạo”.

Ví dụ 2: không biết trách nhiệm của phòng công tác sinh viên là làm gì ạ em cảm ơn.

	Trọng số từ mô hình	Từ loại pyvi
Không	0.22	Trạng từ
Biết	0.17	Động từ
Trách nhiệm	0.91	Danh từ
Của	0.21	Giới từ
Phòng	0.92	Danh từ
Công tác	0.93	Động từ
Sinh viên	0.97	Danh từ
Là	0.85	Động từ
Làm	0.82	Động từ
Gì	0.14	Đại từ
ạ	0.25	Từ bổ trợ, phương thức
Em	0.16	Danh từ
Cảm ơn	0.64	Động từ

Nhận xét: Ý chính trong câu là hỏi về phòng công tác sinh viên nếu chỉ giữ lại danh từ thì sẽ bị mất đi ý nghĩa.

Ví dụ 3: chào admin em muốn hỏi niên giám là gì ạ.

	Trọng số từ mô hình	Từ loại pyvi
Chào	0.15	Động từ
Admin	0.08	Danh từ
Em	0.07	Danh từ
Muốn	0.11	Động từ
Hỏi	0.57	Động từ
Niên giám	0.47	Trạng từ
Là	0.44	Động từ
Gì	0.27	Đại từ
ạ	0.32	Tính từ

Nhận xét: Ý chính trong câu là hỏi về niên giám nếu chỉ giữ lại danh từ thì sẽ bị mất đi ý nghĩa.

4.3 Ứng dụng trọng số vào thuật toán BM25

Trong quá trình thử nghiệm thì nhóm đã áp dụng bộ trọng số vào 2 công thức là:

Đánh giá trên tập dữ liệu được lấy từ <https://github.com/maillong25/bert-vietnamese-question-answering>

Bộ dữ liệu được định dạng json như sau:

```
[
  {
    "id": "u7-1570446247_1",
    "question": "Quang Hải giành được chức vô địch U21 quốc gia năm bao nhiêu tuổi",
    "title": "Nguyễn Quang Hải (sinh 1997)",
    "text": "Năm 2013, Nguyễn Quang Hải giành chức vô địch U21 quốc gia 2013 cùng với đội trẻ Hà Nội T&T và tạo nên cú sốc khi trở thành cầu thủ 16 tuổi đầu tiên giành được danh hiệu vô địch U21 quốc gia.",
    "label": true
  },
  {
    "id": "u7-1570446247_2",
    "question": "Quang Hải giành được chức vô địch U21 quốc gia năm bao nhiêu tuổi",
    "title": "Nguyễn Quang Hải (sinh 1997)",
    "text": "Sau chức vô địch U-21 quốc gia 2013, Nguyễn Quang Hải mới 16 tuổi lập tức được HLV Phan Thanh Hùng điền vào danh sách của đội bóng thủ đô tham dự V-League 2014.",
    "label": true
  },
  ...
]
```

Kết quả đánh giá trên tập dữ liệu bên trên:

	Accuracy (%)	F1-score (%)
BM25	83.34	90.92

BM25 + weight	85.87	92.40
---------------	-------	-------

Ví dụ: BM25 sai nhưng BM25 + weight lại ra kết quả đúng:

Câu hỏi: Iran nằm ở khu_vực nào.

Đoạn 1	Iran (tiếng Ba Tư : ایران\u200e Irān [ʔiːˈɾɒːn] (nghe)) , gọi chính_thức là nước Cộng_hoà Hồi_giáo Iran (tiếng Ba Tư : جمهوری_اسلامی ایران\u200e Jomhuri-ye Eslāmi-ye Irān Về âm_thanh nàyphát âm (trợ_giúp · thông_tin)) , [14] là một quốc_gia có chủ_quyền tại Tây_Á . [15] [16] Iran có biên_giới về phía tây bắc với Armenia , Azerbaijan , và Cộng_hoà Artsakh tự_xung ; phía bắc giáp biển Caspi ; phía đông bắc giáp Turkmenistan ; phía đông giáp Afghanistan và Pakistan ; phía nam giáp vịnh Ba Tư và vịnh Oman ; còn phía tây giáp Thổ_Nhĩ_Kỳ và Iraq (Khu_vực Kurdistan)
Đoạn 2	'Châu_Nam_Cực là lục_địa nằm xa nhất về phía nam của Trái_Đất , chứa cực Nam địa_lý và nằm trong vùng Nam_Cực của Nam_bán_cầu , gần như hoàn_toàn ở trong vòng Nam_Cực và được bao quanh bởi Nam_Băng_Dương . Với diện_tích 14 triệu km2 (5,4 triệu dặ m2) , châu Nam_Cực là lục_địa lớn thứ năm về diện_tích sau châu_Á , châu_Phì , Bắc_Mỹ , và Nam_Mỹ . Khoảng 98% châu Nam_Cực bị bao_phủ bởi một lớp băng có bề dày trung_bình 1,9 km (1,2 dặm) . [3] Băng trải rộng ra khắp mọi phía , xa nhất lên phía bắc tới điểm cực Bắc của bán_đảo Nam_Cực .' ,
Đoạn 3	'Vịnh Hạ_Long 1994 (tái công_nhận : 2000 , 2011) ; Vịnh Hạ_Long nằm trong Vịnh Bắc_Bộ là một quần_thể gồm hơn 1.600 đảo lớn_nhỏ , tạo nên một phong_cảnh tuyệt đẹp giữa biển với những cột đá_vôi nhô lên . Hầu_hết những hòn đảo đều không có người và không có sự tác_động của con người do

	đặc_tính dốc của chúng . Ngoài vẻ đẹp kỳ_diệu , vịnh Hạ_Long còn sở_hữu hệ sinh thái đặc_sắc . [7]',
Đoạn 4	'Núi Chứa_Chan hay còn gọi núi Gia_Lào , cách TP. Hồ_Chí_Minh khoảng 110km , núi có độ cao 800m so với mực nước_biển , thuộc huyện Xuân_Lộc , tỉnh Đồng_Nai . Đây là ngọn núi cao thứ hai khu_vực Nam_Bộ .',
Đoạn 5	'Đảo Ilha da Queimada_Grande rộng 45 ha , nằm lẻ_loi ở Nam_Đại_Tây_Dương , thuộc quyền_sở_hữu của Brazil , cách bờ biển Sao Paulo khoảng 35 km .',
Đoạn 6	'Quốc triểu hình_luật trong cuốn sách A. 341 có 13 chương , ghi_chép trong 6 quyển (5 quyển có 2 chương / quyển và 1 quyển có 3 chương) , gồm 722 điều .',
Đoạn 7	'Lưỡng quốc_trạng nguyên Nguyễn Đẳng_Đạo',

	BM25	BM25 cải tiến
Đoạn 1	2.18	4.73
Đoạn 2	1.75	2.34
Đoạn 3	0.73	1.21
Đoạn 4	1.92	2.13
Đoạn 5	2.56	3.26
Đoạn 6	0.0	0.0
Đoạn 7	0.0	0.0

Nhận xét: theo như chúng ta thấy đoạn văn chứa câu trả lời sẽ nằm trong đoạn văn thứ nhất nhưng đối với thuật toán BM25 cũ thì lại bị sai theo đó đoạn văn có điểm cao nhất là đoạn 5 và cũng gần bằng đoạn 1 không tạo được sự khác biệt. Thuật toán BM25 + weight điểm số của đoạn 1 sẽ cao hơn hẳn các đoạn còn lại.

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

4.4 Kết quả mô hình RoBERTa

Bộ dữ liệu sẽ gồm 3 phần: đoạn văn, câu hỏi và câu trả lời được trích từ đoạn văn. Tổng cộng bao gồm 21073 câu hỏi về những quy định của Trường Đại học Công Nghiệp Thành phố Hồ Chí Minh.

Bộ dữ liệu được định dạng như sau:

	A	B	C
1	Document	Question	Answer
2	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	không biết căn_cứ xét tuyển dựa vào điểm thi trun	căn_cứ xét tuyển xét tổng điểm của
3	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	thưa thầy_cô , em có câu hỏi ạ không biết rằng căn	căn_cứ xét tuyển xét tổng điểm của
4	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	thầy_cô có thể nào cho em biết rằng căn_cứ xét tu	căn_cứ xét tuyển xét tổng điểm của
5	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	cho em hỏi rằng căn_cứ xét tuyển dựa vào điểm th	căn_cứ xét tuyển xét tổng điểm của
6	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	cho em hỏi căn_cứ xét tuyển dựa vào điểm thi trur	căn_cứ xét tuyển xét tổng điểm của
7	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	thầy_cô cho em hỏi căn_cứ xét tuyển dựa vào điểm	căn_cứ xét tuyển xét tổng điểm của
8	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	xin chào thầy_cô em muốn hỏi căn_cứ xét tuyển d	căn_cứ xét tuyển xét tổng điểm của
9	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	em chào các thầy cô ạ thầy_cô cho em hỏi là căn_c	căn_cứ xét tuyển xét tổng điểm của
10	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	ad cho em hỏi căn_cứ xét tuyển dựa vào điểm thi t	căn_cứ xét tuyển xét tổng điểm của
11	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	cô ơi cho em hỏi căn_cứ xét tuyển dựa vào điểm t	căn_cứ xét tuyển xét tổng điểm của
12	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	thầy_cô cho em hỏi căn_cứ xét tuyển dựa vào điểm	căn_cứ xét tuyển xét tổng điểm của
13	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	em xin hỏi về căn_cứ xét tuyển dựa vào điểm thi t	căn_cứ xét tuyển xét tổng điểm của
14	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	em có thắc_mắc muốn được giải_đáp căn_cứ xét tu	căn_cứ xét tuyển xét tổng điểm của
15	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	cho em hỏi trường mình căn_cứ xét tuyển dựa vào	căn_cứ xét tuyển xét tổng điểm của
16	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	hiện_tại em đang muốn hỏi là căn_cứ xét tuyển d	căn_cứ xét tuyển xét tổng điểm của
17	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	cho e hỏi căn_cứ xét tuyển dựa vào điểm thi trung	căn_cứ xét tuyển xét tổng điểm của
18	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	cho mình hỏi , căn_cứ xét tuyển dựa vào điểm thi t	căn_cứ xét tuyển xét tổng điểm của
19	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	thưa thầy_cô căn_cứ xét tuyển dựa vào điểm thi tr	căn_cứ xét tuyển xét tổng điểm của
20	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	em xin hỏi căn_cứ xét tuyển dựa vào điểm thi trun	căn_cứ xét tuyển xét tổng điểm của
21	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	cho em hỏi về căn_cứ xét tuyển dựa vào điểm thi t	căn_cứ xét tuyển xét tổng điểm của
22	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	nhà_trường cho em hỏi căn_cứ xét tuyển dựa vào	căn_cứ xét tuyển xét tổng điểm của
23	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	em chào thầy_cô căn_cứ xét tuyển dựa vào điểm t	căn_cứ xét tuyển xét tổng điểm của
24	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	chào_admin em muốn hỏi căn_cứ xét tuyển dựa v	căn_cứ xét tuyển xét tổng điểm của
25	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	cho em xin hỏi là căn_cứ xét tuyển dựa vào điểm t	căn_cứ xét tuyển xét tổng điểm của
26	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	em chào các thầy_cô , em là sinh_viên năm nhất , c	căn_cứ xét tuyển xét tổng điểm của
27	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	chào mọi người em là sinh_viên năm nhất , cho e	căn_cứ xét tuyển xét tổng điểm của
28	căn_cứ xét tuyển xét tổng điểm của điểm thi trung_học_p	xin phép thầy_cô cho em hỏi căn_cứ xét tuyển d	căn_cứ xét tuyển xét tổng điểm của

Hình 4. 10 Ảnh dữ liệu cho mô hình trả lời câu hỏi

Các bước tiền xử lý dữ liệu:

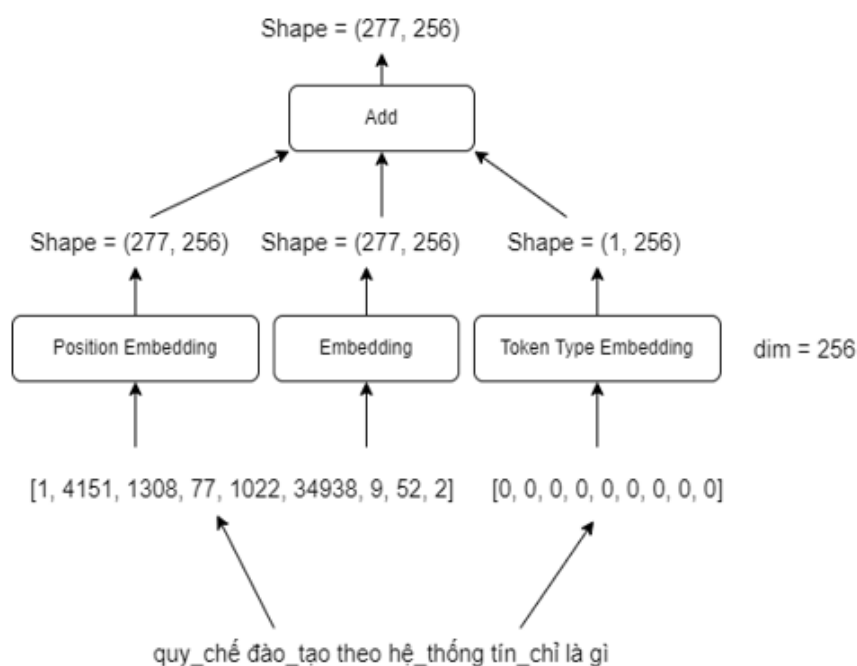
Khi xử lý văn bản tiếng Việt, thường sẽ thực hiện một số thao tác như là chuyển chữ in hoa sang chữ thường (ví dụ: “Tôi” và “tôi” là hai từ cùng nghĩa nhưng lại khác nhau nếu không chuyển về cùng một dạng thì sau này mô hình sẽ xem như hai từ là khác nhau).

Loại bỏ các ký tự đặc biệt như: @, #, %, &, ...

Tách từ: sử dụng thư viện vncorenlp.

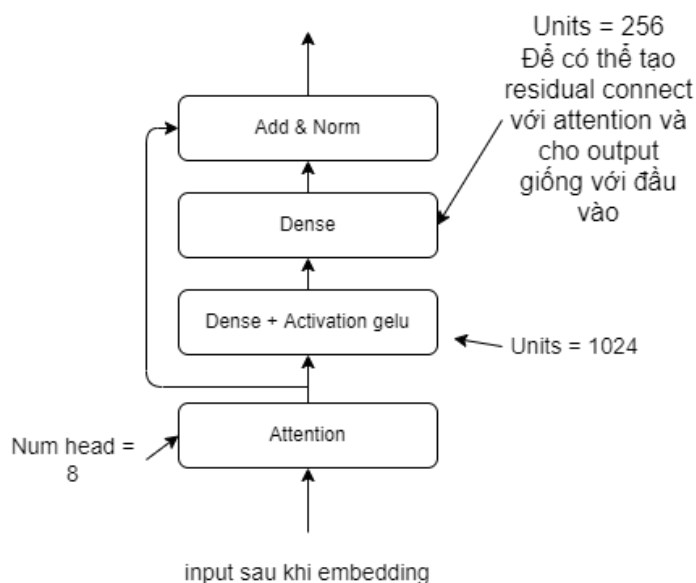
Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Sau khi tiền xử lý dữ liệu sẽ giống như hình trên. Dữ liệu sẽ được đưa và huấn luyện với các tham số trong các hình bên dưới.



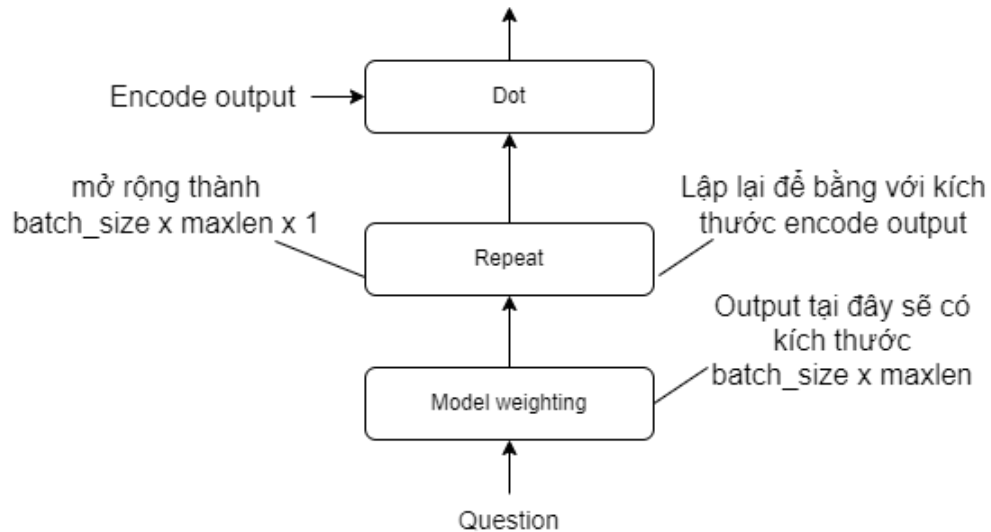
Hình 4. 11 Quá trình embedding block

Input sau khi chuyển thành các tokens sẽ được đưa vào các lớp nhúng để chuyển thành các vector đặc trưng có chiều là 256.



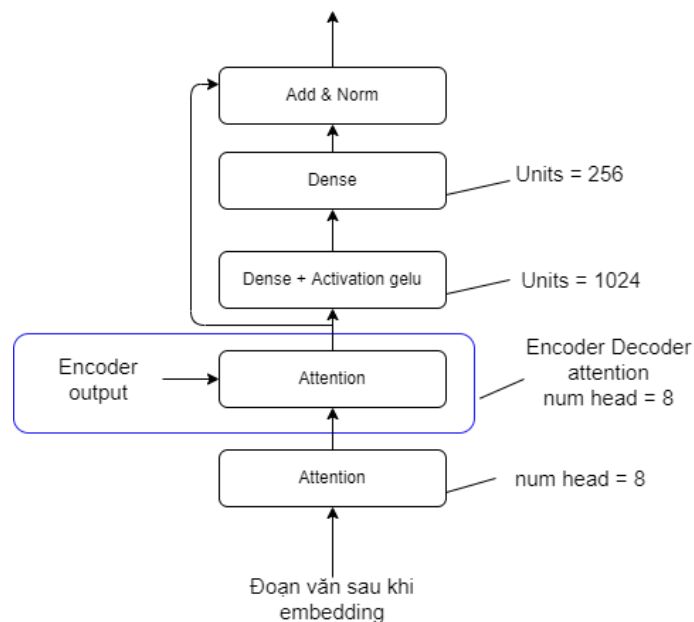
Hình 4. 12 Quá trình train encoder

Câu hỏi sau khi nhúng vector sẽ đưa vào các khối mã hóa. Hình trên thể hiện 1 khối mã hóa. Ở đây sẽ sử dụng 8 khối nối như nhau nối tiếp nhau.



Hình 4. 13 Áp dụng trọng số vào encoder

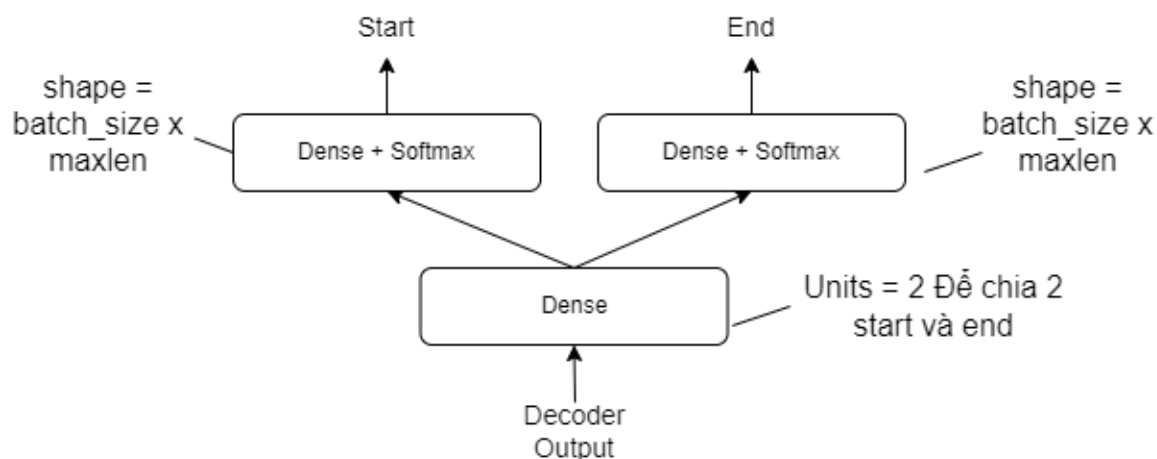
Câu hỏi sẽ được tính trọng số để đánh dấu những từ quan trọng. Output tại mô hình trọng số có kích thước (batch_size x maxlen), để có thể nhàn vào đầu ra của bộ mã hóa thì chúng ta sẽ mở rộng một chiều cuối là lặp lại để trở thành kích thước (batch_size x maxlen x embedding_dim).



Hình 4. 14 Quá trình train decoder

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Bộ giải mã sẽ bao gồm 2 đầu vào: 1 là đoạn văn sau khi nhúng, 1 còn lại là đầu ra của bộ giải mã sau khi nhân với trọng số.



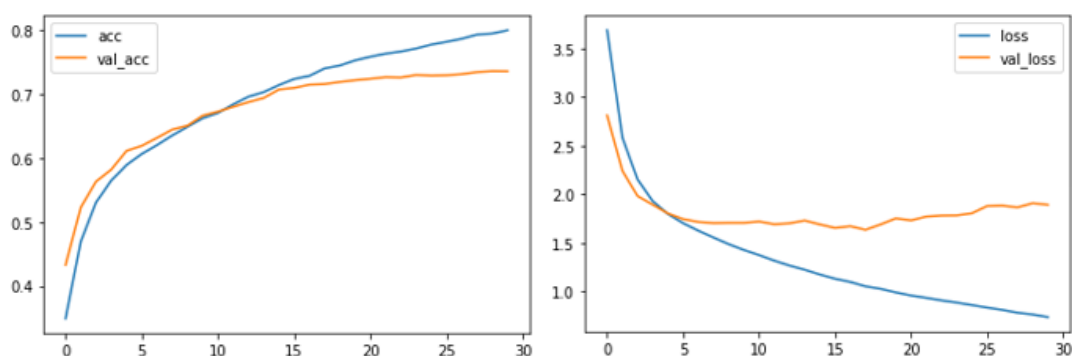
Hình 4. 15 Output mô hình

Mô hình có 2 đầu ra với mục đích xác định điểm bắt đầu và kết thúc của câu trả lời bên trong đoạn văn. Đầu ra sẽ là một vector xác suất có kích thước là (1x800).

Do đó hàm loss của mô hình sẽ có dạng:

$$loss_{start/end} = - \sum_{i=1}^{output\ size} y_i \times \log \hat{y}_i \quad (24)$$

$$loss = \frac{loss_{start} + loss_{end}}{2} \quad (25)$$



Hình 4. 16 Độ chính xác và loss của mô hình RoBERTa

Model	L	H	A	time per epoch	epochs	f1-score	EM
BERT	8	256	8	812s	20	0.65	0.61
BERT + LSTM	8	256	8	784s	20	0.59	0.53
RoBERTa	8	256	8	845	20	0.74	0.74
DistilBERT	8	256	8	892	20	0.72	0.71
RoBERTa+weight	8	256	8	922	20	0.73	0.73
BERT +weight	8	256	8	782	20	0.69	0.65
DistilBERT+weight	8	256	8	1027	20	0.62	0.58

Bảng 4. 2 So sánh các mô hình BERT

Dưới đây là quá trình thử nghiệm các mô hình bằng các qua các ví dụ của thể:

Ví dụ 1:

Câu hỏi	Quy chế đào tạo theo hệ thống tín chỉ là gì?
Đoạn văn 1	Sinh viên chuyển trường được sự đồng ý của hiệu trưởng hai cơ sở giáo dục đại học thì được bảo lưu kết quả rèn luyện của cơ sở giáo dục cũ khi học tại cơ sở giáo dục đại học mới và được tiếp tục đánh giá kết quả rèn luyện ở các kỳ học tiếp theo.
Đoạn văn 2	Sinh viên bị kỷ luật mức buộc thôi học không được đánh giá kết quả rèn luyện.
Đoạn văn 3	Sinh viên không tự đánh giá kết quả rèn luyện theo quy định học kỳ nào sẽ bị nhận điểm 0 trong học kỳ đó.
Đoạn văn 4	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 tự chủ động lựa chọn học từng học phần (tuân theo một số ràng buộc 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.

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Đoạn văn 5	Sinh viên nghỉ học tạm thời được bảo lưu kết quả rèn luyện, sẽ được đánh giá kết quả rèn luyện khi tiếp tục trở lại học tập theo quy định
Đoạn văn 6	Kết quả rèn luyện được phân thành các loại: Xuất sắc, tốt, khá, trung bình, yếu và kém. Phân loại kết quả rèn luyện: từ 90 đến 100 điểm đạt loại xuất sắc; từ 80 đến dưới 90 đạt loại tốt; từ 65 đến dưới 80 đạt loại khá; từ 50 đến dưới 65 đạt loại trung bình; từ 35 đến dưới 50 đạt loại yếu; dưới 35 đạt loại kém.

Đưa vào mô hình tìm trọng số của từ

Sau khi tiền xử lý câu hỏi sẽ được chuyển thành số chính là vị trí của từ đó bên trong từ điển.

Trọng số của các từ như sau:

Quy chế	0.89
Đào tạo	0.87
Theo	0.89
Hệ thống	0.84
Tín chỉ	0.84
Là	0.27
Gì	0.28

Tìm đoạn văn chứa câu trả lời

Sau khi tính trọng số chúng ta sẽ áp dụng nó vào thuật toán BM25 để tìm đoạn văn chứa câu trả lời.

Đoạn văn	Điểm số
Đoạn 1	0.00

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Đoạn 2	0.00
Đoạn 3	1.40
Đoạn 4	13.13
Đoạn 5	1.38
Đoạn 6	0.00

Trích xuất câu trả lời từ đoạn văn

Sau khi tìm được đoạn văn chứa câu trả lời bằng thuật toán BM25 cải tiến thì ta đưa câu hỏi và đoạn văn vào mô hình BERT để tìm vị trí của câu trả lời bên trong đoạn văn.

Câu trả lời: 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 tự_chủ_động lựa_chọn học từng học_phần tuân theo một_số ràng_buộc quy_định

Nhận xét:

Trích xuất đúng văn bản.

Trích xuất câu trả lời chính xác.

Ví dụ 2:

Câu hỏi	Không biết trách nhiệm của phòng công tác sinh viên là làm gì ạ? Em cảm ơn ạ
Đoạn văn 1	Yêu cầu của công tác sinh viên. 1. Sinh viên là nhân vật trung tâm trong Nhà trường, được Nhà trường bảo đảm điều kiện thực hiện đầy đủ nhiệm vụ và quyền trong quá trình học tập và rèn luyện tại Trường. 2, Công tác sinh viên phải thực hiện đúng đường lối, chính sách của Đảng, pháp luật của Nhà nước và các quy chế, quy định của Bộ Giáo dục và Đào tạo. Công tác sinh viên phải bảo đảm dân chủ, khách quan, công bằng, công khai, minh bạch trong các vấn đề có liên quan đến sinh viên.

Đoạn văn 2	<p>Quy định về khung xử lý kỷ luật sinh viên vi phạm quy chế (Ban hành kèm theo Quyết định số 1034/QĐ-ĐHCN ngày 28 tháng 05 năm 2019 của Hiệu trưởng Trường Đại học Công nghiệp Thành phố Hồ Chí Minh).</p> <p>1. Đến muộn giờ học, giờ thực tập không có lý do chính đáng, nghỉ học không phép hoặc nghỉ quá phép bị kỷ luật Cấm thi nếu nghỉ quá 20% số tiết trong môn học; Buộc thôi học nếu tự ý nghỉ 02 học kỳ liên tiếp.</p> <p>2. Mất trật tự làm việc riêng trong giờ học, giờ thực tập và tự học bị kỷ luật Kiểm điểm, trừ điểm rèn luyện theo Quy định đánh giá kết quả rèn luyện.</p> <p>3. Vô lễ với Thầy, Cô giáo và cán bộ viên chức Nhà trường bị kỷ luật Tùy mức độ, xử lý từ khiển trách đến buộc thôi học.</p> <p>4. Học hộ hoặc nhờ người khác học hộ bị kỷ luật Cảnh cáo đối với lần 1, đình chỉ học tập 1 năm nếu tái phạm lần 2 và buộc thôi học nếu vi phạm lần thứ 3.</p> <p>5. Thi, kiểm tra hộ hoặc nhờ thi, kiểm tra hộ; làm hộ, nhờ làm hoặc sao chép tiểu luận, đồ án, khoá luận tốt nghiệp bị kỷ luật Cảnh cáo đối với lần 1 và buộc thôi học nếu vi phạm lần 2.</p> <p>6. Tổ chức học, thi, kiểm tra, hộ; tổ chức làm hộ tiểu luận, đồ án, khoá luận tốt nghiệp bị kỷ luật buộc thôi học nếu vi phạm lần 1; Tùy theo mức độ có thể giao cho cơ quan chức năng xử lý theo quy định của pháp luật.</p> <p>7. Mang tài liệu vào phòng thi, viết hoặc vẽ bậy vào bài thi bị kỷ luật Đình chỉ thi và cho bài thi điểm không. nếu nội dung viết, vẽ vào bài thi vi phạm pháp luật, đạo đức xã hội, tùy mức độ xử lý kỷ luật từ khiển trách đến buộc thôi học.</p> <p>8. Đưa đề thi ra ngoài nhờ làm hộ, ném tài liệu vào phòng thi và các hình thức gian lận khác trong học tập, thi, kiểm tra bị kỷ luật xử lý theo quy chế đào tạo Nhà trường.</p> <p>9. Cố tình không nộp bảo hiểm y tế theo quy định của Nhà trường mà không có lý do chính đáng bị kỷ luật tùy mức độ, xử lý từ nhắc nhở, khiển trách đến buộc thôi học.</p> <p>10. Xả rác bừa bãi, bôi xóa, viết vẽ lên bàn, tường trong phòng học và trong khuôn viên Trường, làm hư hỏng tài sản của Nhà</p>
------------	---

	<p>trường bị kỷ luật tùy thuộc mức độ xử lý từ khiển trách đến buộc thôi học và phải bồi thường thiệt hại. 11. Uống rượu, bia, trong giờ học; say rượu, bia khi đến lớp bị kỷ luật khiển trách lần 1, cảnh cáo nếu tái phạm lần thứ 2, đình chỉ học tập 1 năm nếu tái phạm lần thứ 3 và buộc thôi học nếu tái phạm lần thứ 4. 12. Hút thuốc là trong giờ học, phòng họp, phòng thí nghiệm và nơi cấm hút thuốc theo quy định bị kỷ luật từ lần 3 trở lên, xử lý từ khiển trách đến cảnh cáo. 13. Đánh bạc dưới mọi hình thức bị kỷ luật khiển trách nếu vi phạm lần thứ 1, cảnh cáo nếu vi phạm lần thứ 2, đình chỉ học tập 1 năm nếu vi phạm lần thứ 3 và buộc thôi học nếu vi phạm lần thứ 4. 14. Tàng trữ lưu hành, truy cập, sử dụng sản phẩm văn hoá đồi trụy hoặc tham gia các hoạt động mê tín dị đoan, hoạt động tôn giáo trái phép bị kỷ luật khiển trách nếu vi phạm lần thứ 1, cảnh cáo nếu vi phạm lần thứ 2, đình chỉ học tập 1 năm nếu vi phạm lần thứ 3 và buộc thôi học nếu vi phạm lần thứ 4. 15. Buôn bán, vận chuyển, tàng trữ, sử dụng lời kéo người khác sử dụng ma túy bị kỷ luật buộc thôi học nếu vi phạm lần đầu và giao cho các cơ quan chức năng xử lý theo quy định của pháp luật. 16. Sử dụng ma túy bị xử lý theo quyết định số 48/2006/QĐ-BGDĐT ngày 25/10/2006 của Bộ Giáo dục và Đào tạo. 17. Hoạt động mại dâm bị xử lý theo quy định hiện hành của Bộ GDĐT. 18. Chứa chấp, môi giới hoạt động mại dâm bị kỷ luật buộc thôi học nếu vi phạm lần 1 và giao cho cơ quan chức năng xử lý theo quy định của pháp luật. 19. Lừa đảo, lấy cắp tài sản, chứa chấp, tiêu thụ tài sản do lấy cắp mà có bị xử lý tùy theo mức độ xử lý từ cảnh cáo đến buộc thôi học. Nếu nghiêm trọng giao cho cơ quan chức năng xử lý theo quy định của pháp luật. 20. Chứa chấp buôn bán vũ khí, chất dễ cháy và các hàng cấm theo quy định của nhà nước bị kỷ luật buộc thôi học và giao cho cơ quan chức năng xử lý theo quy định của pháp luật.</p>
--	---

Đoạn văn 3	<p>Trách nhiệm của Phòng Công tác sinh viên. 1. Quản lý hồ sơ nhập học, hồ sơ miễn giảm học phí của sinh viên. 2. Theo dõi, tổng hợp kết quả rèn luyện sinh viên theo từng học kỳ, năm học và toàn khóa học. 3. Tham mưu cho Hội đồng khen thưởng và kỷ luật sinh viên để khen thưởng cá nhân và tập thể có thành tích cao trong học tập, rèn luyện và các hoạt động khác hoặc xử lý khi vi phạm quy chế, quy định, nội quy Nhà trường. 4. Xác nhận kết quả rèn luyện, cấp giấy chứng nhận tham gia Tuần sinh hoạt công dân sinh viên và các giấy tờ khác cho sinh viên thuộc thẩm quyền của đơn vị. 5. Tổ chức triển khai công tác giáo dục tư tưởng chính trị, đạo đức, lối sống, nhân cách cho sinh viên; phối hợp tổ chức cho sinh viên tham gia các hoạt động chính trị xã hội, văn hóa văn nghệ, thể dục thể thao và các hoạt động khác ngoài giờ lên lớp ở cấp trường. 6. Phối hợp với các ngành, các cấp chính quyền địa phương trên địa bàn nơi trường trú đóng, xây dựng kế hoạch đảm bảo an ninh chính trị, trật tự và an toàn cho sinh viên; giải quyết kịp thời các vụ việc liên quan đến sinh viên. 7. Tổ chức tuyên truyền, giáo dục cho sinh viên các kiến thức cơ bản về kỹ năng sống, tư vấn học đường, giáo dục an toàn giao thông, phòng chống tội phạm và các tệ nạn xã hội; hướng dẫn sinh viên chấp hành các quy định của pháp luật và nội quy, quy định, quy chế của Nhà trường. 8. Phối hợp với Trung tâm Thư viện làm và cấp phát thẻ sinh viên. 9. Phối hợp tổ chức đối thoại định kỳ giữa Hiệu trưởng với sinh viên. 10. Tổ chức và phối hợp thực hiện các chế độ chính sách của Nhà nước quy định đối với sinh viên về học bổng, học phí, trợ cấp xã hội và các chế độ khác có liên quan đến sinh viên. 11. Tổ chức triển khai thực hiện công tác quản lý sinh viên ngoại trú theo quy định của Nhà trường.</p>
------------	--

Đoạn văn 4	Hồ sơ xét khen thưởng. Hồ sơ xét khen thưởng bao gồm: Biên bản họp lớp; báo cáo tổng kết; bảng tổng hợp kết quả rèn luyện chuyên cần và danh sách đề nghị khen thưởng.
Đoạn văn 5	<p>Các hành vi sinh viên không được làm. 1. Xúc phạm nhân phẩm, danh dự, xâm phạm thân thể nhà giáo, cán bộ quản lý, viên chức, nhân viên, người học của Nhà trường và người khác. 2. Gian lận trong học tập, kiểm tra, thi cử như: quay cốp, mang tài liệu vào phòng thi, xin điểm; học, thi, thực tập, trực hộ người khác hoặc nhờ người khác học, thi, thực tập; sao chép, nhờ hoặc làm hộ tiểu luận, đồ án, khóa luận tốt nghiệp; tổ chức hoặc tham gia tổ chức thi hộ hoặc các hành vi gian lận khác. 3. Hút thuốc, uống rượu, bia trong khuôn viên Trường; say rượu, bia khi đến lớp học. 4. Xả rác bừa bãi, bôi xóa, viết vẽ lên bàn, tường trong phòng học và trong khuôn viên của Nhà trường: làm hư hại các tài sản, trang thiết bị của Nhà trường. 5. Tổ chức hoặc tham gia tụ tập đông người, biểu tình, khiếu kiện trái pháp luật; tham gia tệ nạn xã hội, gây rối an ninh, trật tự an toàn trong Nhà trường hoặc ngoài xã hội. 6. Tổ chức hoặc tham gia đua xe, cò vũ đua xe trái phép. 7. Tổ chức hoặc tham gia đánh bạc dưới mọi hình thức. 8. Sản xuất, buôn bán, vận chuyển, phát tán, tàng trữ, sử dụng hoặc lôi kéo người khác sử dụng vũ khí, chất nổ, các chất ma túy, các loại dược phẩm, hóa chất cấm sử dụng; các tài liệu, ấn phẩm, thông tin phản động, đồi trụy và các tài liệu cấm khác theo quy định của Nhà nước; tổ chức, tham gia, truyền bá các hoạt động mê tín dị đoan, các hoạt động tôn giáo trong Nhà trường và các hành vi vi phạm đạo đức khác. 9. Thành lập, tham gia các hoạt động mang tính chất chính trị trái pháp luật; tổ chức, tham gia các hoạt động tập thể mang danh nghĩa Nhà trường khi chưa được Hiệu trưởng cho phép. 10. Đăng tải, bình luận, chia sẻ bài viết, hình ảnh có nội dung dung tục, bạo lực, đồi trụy, xâm phạm an ninh quốc gia, chống phá</p>

	Đảng và Nhà nước, xuyên tạc, vu khống, xúc phạm uy tín của tổ chức, danh dự và nhân phẩm của cá nhân trên mạng internet. 11. Tổ chức hoặc tham gia các hoạt động vi phạm pháp luật khác.
--	--

Không	0.20
Biết	0.18
Trách nhiệm	0.93
Của	0.22
Phòng	0.92
Công tác	0.93
Sinh viên	0.96
Là	0.87
Làm	0.83
Gì	0.16
ạ	0.41
Em	0.13
Cám ơn	0.68

Đoạn văn	Điểm số
Đoạn 1	6.66
Đoạn 2	5.90
Đoạn 3	8.38
Đoạn 4	0.00
Đoạn 5	4.90

Câu trả lời: trách_nhiệm của phòng công_tác sinh_viên 1 quản_lý hồ_sơ nhập_học hồ_sơ miễn_giảm học_phí của sinh_viên 2 theo_dõi tổng_hợp kết_quả rèn_luyện sinh_viên theo từng học_kỳ năm_học và toàn_khoá học 3 tham_mưu cho hội_đồng khen_thưởng và kỷ_luật sinh_viên để khen_thưởng cá_nhân và tập_thể có thành_tích cao trong học_tập rèn_luyện và các hoạt_động khác hoặc xử_lý khi vi_phạm quy_chế quy_định nội_quy nhà_trường 4 xác_nhận kết_quả rèn_luyện cấp giấy_chứng_nhận tham_gia tuần sinh_hoạt công_dân sinh_viên và các giấy_tờ khác cho sinh_viên thuộc thẩm_quyền của đơn_vị 5 tổ_chức triển_khai công_tác giáo_dục tư_tưởng

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

chính_trị đạo_đức lối sống nhân_cách cho sinh_viên phối_hợp tổ_chức cho sinh_viên tham_gia các hoạt_động chính_trị xã_hội văn_hoá văn_nghệ thể_dục thể_thao và các hoạt_động khác ngoài giờ lên_lớp ở cấp trường 6 phối_hợp với các ngành các cấp chính_quyền địa_phương trên địa_bàn nơi trường trú đóng xây_dựng kế_hoạch đảm_bảo an_ninh chính_trị trật_tự.

Nhận xét:

Đánh trọng số hợp lý. Trọng số cao nằm ở những từ như là “trách nhiệm”, “phòng”, “công tác”, “sinh viên”, “là”, “làm”.

Ví dụ 3:

Câu hỏi	Không biết trách nhiệm của phòng tổ chức hành chính là gì ạ? Em cảm ơn ạ.
Đoạn văn 1	Trách nhiệm của Phòng Tổ chức Hành chính. 1. Giải quyết các công việc hành chính có liên quan cho sinh viên; sao y các văn bằng, chứng chỉ của Trường. 2. Tiếp nhận và phân phối bưu phẩm, thư cho sinh viên.
Đoạn văn 2	Quy định về khung xử lý kỷ luật sinh viên vi phạm quy chế (Ban hành kèm theo Quyết định số 1034/QĐ-ĐHCN ngày 28 tháng 05 năm 2019 của Hiệu trưởng Trường Đại học Công nghiệp Thành phố Hồ Chí Minh). 1. Đến muộn giờ học, giờ thực tập không có lý do chính đáng, nghỉ học không phép hoặc nghỉ quá phép bị kỷ luật Cấm thi nếu nghỉ quá 20% số tiết trong môn học; Buộc thôi học nếu tự ý nghỉ 02 học kỳ liên tiếp. 2. Mất trật tự làm việc riêng trong giờ học, giờ thực tập và tự học bị kỷ luật Kiểm điểm, trừ điểm rèn luyện theo Quy định đánh giá kết quả rèn luyện. 3. Vô lễ với Thầy, Cô giáo và cán bộ viên chức Nhà trường bị kỷ luật Tuỳ mức độ, xử lý từ khiển trách đến buộc thôi học. 4. Học hộ hoặc nhờ người khác học hộ bị kỷ luật Cảnh cáo đối với lần 1, đình chỉ học tập 1 năm nếu tái phạm lần 2 và buộc thôi học nếu vi phạm lần thứ 3. 5. Thi, kiểm tra hộ hoặc nhờ thi, kiểm tra hộ; làm hộ, nhờ làm hoặc sao chép tiểu luận,

	<p>đề án, khoá luận tốt nghiệp bị kỷ luật Cảnh cáo đối với lần 1 và buộc thôi học nếu vi phạm lần 2. 6. Tổ chức học, thi, kiểm tra, hộ; tổ chức làm hộ tiểu luận, đề án, khoá luận tốt nghiệp bị kỷ luật buộc thôi học nếu vi phạm lần 1; Tuỳ theo mức độ có thể giao cho cơ quan chức năng xử lý theo quy định của pháp luật. 7. Mang tài liệu vào phòng thi, viết hoặc vẽ bậy vào bài thi bị kỷ luật Đình chỉ thi và cho bài thi điểm không. nếu nội dung viết, vẽ vào bài thi vi phạm pháp luật, đạo đức xã hội, tuỳ mức độ xử lý kỷ luật từ khiển trách đến buộc thôi học. 8. Đưa đề thi ra ngoài nhờ làm hộ, ném tài liệu vào phòng thi và các hình thức gian lận khác trong học tập, thi, kiểm tra bị kỷ luật xử lý theo quy chế đào tạo Nhà trường. 9. Cố tình không nộp bảo hiểm y tế theo quy định của Nhà trường mà không có lý do chính đáng bị kỷ luật tuỳ mức độ, xử lý từ nhắc nhở, khiển trách đến buộc thôi học. 10. Xả rác bừa bãi, bôi xoá, viết vẽ lên bàn, tường trong phòng học và trong khuôn viên Trường, làm hư hỏng tài sản của Nhà trường bị kỷ luật tuỳ thuộc mức độ xử lý từ khiển trách đến buộc thôi học và phải bồi thường thiệt hại. 11. Uống rượu, bia, trong giờ học; say rượu, bia khi đến lớp bị kỷ luật khiển trách lần 1, cảnh cáo nếu tái phạm lần thứ 2, đình chỉ học tập 1 năm nếu tái phạm lần thứ 3 và buộc thôi học nếu tái phạm lần thứ 4. 12. Hút thuốc là trong giờ học, phòng họp, phòng thí nghiệm và nơi cấm hút thuốc theo quy định bị kỷ luật từ lần 3 trở lên, xử lý từ khiển trách đến cảnh cáo. 13. Đánh bạc dưới mọi hình thức bị kỷ luật khiển trách nếu vi phạm lần thứ 1, cảnh cáo nếu vi phạm lần thứ 2, đình chỉ học tập 1 năm nếu vi phạm lần thứ 3 và buộc thôi học nếu vi phạm lần thứ 4. 14. Tàng trữ lưu hành, truy cập, sử dụng sản phẩm văn hoá đồi trụy hoặc tham gia các hoạt động mê tín dị đoan, hoạt động tôn giáo trái phép bị kỷ luật khiển trách nếu vi phạm lần thứ 1, cảnh cáo nếu vi phạm lần thứ 2, đình chỉ học tập 1 năm nếu vi phạm lần thứ 3 và buộc thôi học</p>
--	---

	<p>nếu vi phạm lần thứ 4. 15. Buôn bán, vận chuyển, tàng trữ, sử dụng lôi kéo người khác sử dụng ma túy bị kỷ luật buộc thôi học nếu vi phạm lần đầu và giao cho các cơ quan chức năng xử lý theo quy định của pháp luật. 16. Sử dụng ma túy bị xử lý theo quyết định số 48/2006/QĐ-BGDĐT ngày 25/10/2006 của Bộ Giáo dục và Đào tạo. 17. Hoạt động mại dâm bị xử lý theo quy định hiện hành của Bộ GDĐT. 18. Chứa chấp, môi giới hoạt động mại dâm bị kỷ luật buộc thôi học nếu vi phạm lần 1 và giao cho cơ quan chức năng xử lý theo quy định của pháp luật. 19. Lừa đảo, lấy cắp tài sản, chứa chấp, tiêu thụ tài sản do lấy cắp mà có bị xử lý tùy theo mức độ xử lý từ cảnh cáo đến buộc thôi học. Nếu nghiêm trọng giao cho cơ quan chức năng xử lý theo quy định của pháp luật. 20. Chứa chấp buôn bán vũ khí, chất dễ cháy và các hàng cấm theo quy định của nhà nước bị kỷ luật buộc thôi học và giao cho cơ quan chức năng xử lý theo quy định của pháp luật.</p>
Đoạn văn 3	<p>Trách nhiệm của Phòng Công tác sinh viên. 1. Quản lý hồ sơ nhập học, hồ sơ miễn giảm học phí của sinh viên. 2. Theo dõi, tổng hợp kết quả rèn luyện sinh viên theo từng học kỳ, năm học và toàn khóa học. 3. Tham mưu cho Hội đồng khen thưởng và kỷ luật sinh viên để khen thưởng cá nhân và tập thể có thành tích cao trong học tập, rèn luyện và các hoạt động khác hoặc xử lý khi vi phạm quy chế, quy định, nội quy Nhà trường. 4. Xác nhận kết quả rèn luyện, cấp giấy chứng nhận tham gia Tuần sinh hoạt công dân sinh viên và các giấy tờ khác cho sinh viên thuộc thẩm quyền của đơn vị. 5. Tổ chức triển khai công tác giáo dục tư tưởng chính trị, đạo đức, lối sống, nhân cách cho sinh viên; phối hợp tổ chức cho sinh viên tham gia các hoạt động chính trị xã hội, văn hóa văn nghệ, thể dục thể thao và các hoạt động khác ngoài giờ lên lớp ở cấp trường. 6. Phối hợp với các ngành, các cấp chính quyền địa phương trên địa bàn nơi trường</p>

	<p>trú đóng, xây dựng kế hoạch đảm bảo an ninh chính trị, trật tự và an toàn cho sinh viên; giải quyết kịp thời các vụ việc liên quan đến sinh viên. 7. Tổ chức tuyên truyền, giáo dục cho sinh viên các kiến thức cơ bản về kỹ năng sống, tư vấn học đường, giáo dục an toàn giao thông, phòng chống tội phạm và các tệ nạn xã hội; hướng dẫn sinh viên chấp hành các quy định của pháp luật và nội quy, quy định, quy chế của Nhà trường. 8. Phối hợp với Trung tâm Thư viện làm và cấp phát thẻ sinh viên. 9. Phối hợp tổ chức đối thoại định kỳ giữa Hiệu trưởng với sinh viên. 10. Tổ chức và phối hợp thực hiện các chế độ chính sách của Nhà nước quy định đối với sinh viên về học bổng, học phí, trợ cấp xã hội và các chế độ khác có liên quan đến sinh viên. 11. Tổ chức triển khai thực hiện công tác quản lý sinh viên ngoại trú theo quy định của Nhà trường.</p>
Đoạn văn 4	<p>Quyền của sinh viên. 1. Được nhận vào học đúng ngành nghề đã đăng ký dự tuyển nếu đủ các điều kiện trúng tuyển theo quy định của Nhà trường. 2. Được tôn trọng và đối xử bình đẳng: được cung cấp đầy đủ thông tin cá nhân về việc học tập, rèn luyện theo quy định của Nhà trường: được phổ biến nội quy, quy chế về đào tạo, rèn luyện và các chế độ, chính sách của Nhà nước có liên quan đến sinh viên. 3. Được tạo điều kiện trong học tập, nghiên cứu khoa học và rèn luyện, bao gồm: Sử dụng hệ thống thư viện, các trang thiết bị và phương tiện phục vụ các hoạt động học tập, nghiên cứu khoa học, văn hóa, văn nghệ, thể dục, thể thao; Tham gia nghiên cứu khoa học, thi sinh viên giỏi, thi Olympic các môn học, thi sáng tạo khoa học, kỹ thuật; Chăm sóc, bảo vệ sức khỏe theo quy định hiện hành của Nhà nước; Đăng ký dự tuyển đi học, tham gia các hoạt động giao lưu, trao đổi sinh viên ở nước ngoài; học chuyển tiếp ở các trình độ đào tạo cao hơn theo quy định hiện hành; Tham gia hoạt động trong tổ chức Đảng Cộng sản Việt Nam, Đoàn Thanh niên Cộng sản</p>

	<p>(TNCS) Hồ Chí Minh, Hội Sinh viên Việt Nam; tham gia các tổ chức tự quản của sinh viên, các hoạt động xã hội có liên quan ở trong và ngoài trường học theo quy định của pháp luật; các hoạt động văn hóa, văn nghệ, thể thao lành mạnh, phù hợp với mục tiêu đào tạo của Nhà trường; Sử dụng các dịch vụ công tác xã hội hiện có của Nhà trường (bao gồm các dịch vụ về hướng nghiệp, tư vấn việc làm, tư vấn sức khỏe, tâm lý, hỗ trợ sinh viên có hoàn cảnh đặc biệt,...); Nghỉ học tạm thời, tạm ngừng học, học theo tiến độ chậm, tiến độ nhanh, học cùng lúc hai chương trình, chuyển trường theo quy định, quy chế về đào tạo của Bộ Giáo dục và Đào tạo; được nghỉ hè, nghỉ tết, nghỉ lễ theo quy định. 4. Được hưởng các chế độ, chính sách, được xét nhận học bổng khuyến khích học tập, học bổng do các tổ chức, cá nhân trong và ngoài nước tài trợ theo quy định hiện hành; được miễn giảm phí khi sử dụng các dịch vụ công cộng về giao thông, giải trí, tham quan bảo tàng, di tích lịch sử, công trình văn hóa theo quy định của Nhà nước. 5. Được góp ý kiến, tham gia giám sát hoạt động giáo dục và các điều kiện đảm bảo chất lượng giáo dục; trực tiếp hoặc thông qua đại diện hợp pháp của mình kiến nghị các giải pháp góp phần xây dựng và phát triển Nhà trường; đề bạt nguyện vọng và khiếu nại lên Hiệu trưởng giải quyết các vấn đề có liên quan đến quyền, lợi ích chính đáng của sinh viên. 6. Được xét tiếp nhận vào ở ký túc xá và ưu tiên theo quy định. 7. Sinh viên đủ điều kiện công nhận tốt nghiệp được cấp bằng tốt nghiệp, chứng chỉ, bảng điểm học tập và rèn luyện, các giấy tờ liên quan và giải quyết các thủ tục hành chính khác.</p>
Đoạn văn 5	<p>Các hành vi sinh viên không được làm. 1. Xúc phạm nhân phẩm, danh dự, xâm phạm thân thể nhà giáo, cán bộ quản lý, viên chức, nhân viên, người học của Nhà trường và người khác. 2. Gian lận trong học tập, kiểm tra, thi cử như: quay cốp, mang tài liệu vào</p>

	phòng thi, xin điểm; học, thi, thực tập, trực hộ người khác hoặc nhờ người khác học, thi, thực tập; sao chép, nhờ hoặc làm hộ tiểu luận, đồ án, khóa luận tốt nghiệp; tổ chức hoặc tham gia tổ chức thi hộ hoặc các hành vi gian lận khác. 3. Hút thuốc, uống rượu, bia trong khuôn viên Trường; say rượu, bia khi đến lớp học. 4. Xả rác bừa bãi, bôi xóa, viết vẽ lên bàn, tường trong phòng học và trong khuôn viên của Nhà trường: làm hư hại các tài sản, trang thiết bị của Nhà trường. 5. Tổ chức hoặc tham gia tụ tập đông người, biểu tình, khiếu kiện trái pháp luật; tham gia tệ nạn xã hội, gây rối an ninh, trật tự an toàn trong Nhà trường hoặc ngoài xã hội. 6. Tổ chức hoặc tham gia đua xe, cò vũ đua xe trái phép. 7. Tổ chức hoặc tham gia đánh bạc dưới mọi hình thức. 8. Sản xuất, buôn bán, vận chuyển, phát tán, tàng trữ, sử dụng hoặc lôi kéo người khác sử dụng vũ khí, chất nổ, các chất ma túy, các loại dược phẩm, hóa chất cấm sử dụng; các tài liệu, ấn phẩm, thông tin phản động, đồi trụy và các tài liệu cấm khác theo quy định của Nhà nước; tổ chức, tham gia, truyền bá các hoạt động mê tín dị đoan, các hoạt động tôn giáo trong Nhà trường và các hành vi vi phạm đạo đức khác. 9. Thành lập, tham gia các hoạt động mang tính chất chính trị trái pháp luật; tổ chức, tham gia các hoạt động tập thể mang danh nghĩa Nhà trường khi chưa được Hiệu trưởng cho phép. 10. Đăng tải, bình luận, chia sẻ bài viết, hình ảnh có nội dung dung tục, bạo lực, đồi trụy, xâm phạm an ninh quốc gia, chống phá Đảng và Nhà nước, xuyên tạc, vu khống, xúc phạm uy tín của tổ chức, danh dự và nhân phẩm của cá nhân trên mạng internet. 11. Tổ chức hoặc tham gia các hoạt động vi phạm pháp luật khác.
--	--

Không	0.25
Biết	0.16
Trách nhiệm	0.95
Của	0.22
Phòng	0.96

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Tổ chức	0.94
Hành chính	0.97
Là	0.49
Gì	0.18
ạ	0.13
Em	0.16
Cám ơn	0.64

Đoạn văn	Điểm số
Đoạn 1	6.82
Đoạn 2	4.30
Đoạn 3	3.19
Đoạn 4	2.15
Đoạn 5	2.86

Câu trả lời: trách_nhiệm của phòng tổ_chức hành_chính 1 giải_quyết các công_việc hành_chính có liên_quan cho sinh_viên sao y.

Nhận xét:

Trọng số đánh hợp lý cho nên kết quả tìm đoạn văn theo như chúng ta có thể thấy khoảng cách số điểm của đoạn văn chứa câu trả lời cao hơn các đoạn còn lại. Kết quả không tốt nên chỉ đưa ra đáp án gần đúng.

Ví dụ 4:

Câu hỏi	Chào Admin em muốn hỏi niên giám là gì ạ?
Đoạn văn 1	Lớp học phần là lớp học được tổ chức theo từng học phần dựa trên kết quả đăng ký khối lượng học tập của SV ở từng học kỳ, trong cùng thời điểm. Mỗi lớp học phần được gán một mã số riêng. Số lượng sinh viên của một lớp học phần được giới hạn bởi sức chứa của phòng học hoặc phòng thí nghiệm. Nếu số lượng sinh viên đăng ký thấp hơn chuẩn tối thiểu thì lớp học sẽ không được tổ chức và sinh viên phải đăng ký chuyển sang học phần khác có lớp. Số lượng sinh viên tối thiểu để xem xét mở lớp học phần trong học kỳ đối với

	các học phần lý thuyết là: - Ít nhất 50 sinh viên cho các học phần cơ bản, cơ sở của các nhóm ngành lớn và 30 sinh viên đăng ký cho các học phần khác; - Trong trường hợp các học phần chuyên ngành, nếu có đề nghị của khoa quản lý ngành, trường sẽ xem xét để mở các lớp có sĩ số chưa đủ điều kiện mở lớp; - Số lượng sinh viên để xem xét mở lớp học phần trong học kỳ đối với các học phần thí nghiệm, thực hành được xác định bởi đơn vị đào tạo, tối thiểu 20 sinh viên.
Đoạn văn 2	Có hai hình thức tổ chức lớp: lớp học phần và lớp sinh viên.
Đoạn văn 3	Đối với sinh viên hệ chính quy, khối lượng học tập được thể hiện trong Niên giám mà Trường đã cung cấp khi sinh viên làm thủ tục nhập học, trong đó thể hiện lịch trình học dự kiến cho từng chương trình, ngành học, từng học kỳ và từng năm học, bao gồm: danh sách các học phần bắt buộc và tự chọn dự kiến sẽ dạy, điều kiện tiên quyết để được đăng ký học cho từng học phần. Đối với các hệ bậc đào tạo khác, sinh viên liên hệ ĐVĐT để được cung cấp kế hoạch học tập. Hằng năm, vào cuối năm học, Nhà trường sẽ công bố tiến độ đào tạo cho năm học tiếp theo, bao gồm: kế hoạch giảng dạy, kế hoạch kiểm tra và thi cuối kỳ. ĐVĐT sẽ chịu trách nhiệm cung cấp hình thức kiểm tra và thi đối với từng học phần trong buổi học đầu tiên của học phần. Sinh viên tùy theo khả năng và điều kiện học tập của mình, trước mỗi học kỳ đăng ký các lớp học phần cho học kỳ đó và các học kỳ còn lại của năm học (đầu mỗi học kỳ tiếp theo, sinh viên có quyền thay đổi các học phần đã đăng ký trước khi Phòng Đào tạo chấp nhận mở lớp).
Đoạn 4	A. Tổ chức lớp học phần. - Sinh viên bậc đại học hệ vừa làm vừa học, hệ liên thông, hệ chính quy đại trà được phép đăng ký học chung với nhau trong các lớp học phần của cùng một môn học và cùng số tín chỉ. Sinh viên đại học hệ chất lượng cao chỉ được đăng ký học chung với các hệ đào tạo đại học khác đối với các môn đã có điểm

	tổng kết học phần. - Khi có ít nhất 2 sinh viên làm đơn xin học một môn học thì đơn vị đào tạo phải mở lớp học phần.
Đoạn 5	Đầu mỗi học kỳ, sinh viên phải theo dõi thông báo của Trường để nắm thông tin về các học phần sẽ mở trong học kỳ và đăng ký các học phần qua cổng thông tin sinh viên. Kết quả đăng ký học tập của mỗi sinh viên được thông báo trong tài khoản sinh viên và trên website Trường. Kết quả đăng ký học phần ghi rõ tên học phần, mã học phần, mã lớp, số tín chỉ của mỗi học phần, lịch học của các học phần.
Đoạn 6	Lớp sinh viên là lớp được tổ chức cho sinh viên cùng khóa để duy trì các sinh hoạt chính trị, xã hội của tập thể sinh viên ngoài giờ lên lớp theo lớp học phần như: xét khen thưởng, kỷ luật, bình xét điểm rèn luyện.. Lớp sinh viên cũng là cơ sở để hình thành các chỉ đoàn thanh niên, chỉ hội sinh viên. Sau 4 năm đối với đại học, sau 3 năm đối với cao đẳng chuyên nghiệp chính quy sau khi tốt nghiệp thì lớp sinh viên tự giải thể. Số sinh viên chưa được tốt nghiệp do tích lũy chưa đủ số tín chỉ sẽ chuyển về khoa chuyên ngành quản lý (gọi là sinh viên vắng lai).

Chào_admin	0.14
Em	0.10
Muốn	0.08
Hỏi	0.11
Niên_giám	0.57
Là	0.47
Gì	0.44
ạ	0.28

Đoạn văn	Điểm số
Đoạn 1	1.53
Đoạn 2	0.00

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.

Đoạn 3	1.59
Đoạn 4	0.00
Đoạn 5	0.00
Đoạn 6	2.11

Câu trả lời: lớp sinh_viên là lớp được tổ_chức cho sinh_viên cùng khoá để duy_trì các sinh_hoạt chính_trị xã_hội của tập_thể sinh_viên ngoài giờ lên_lớp theo lớp_học phần như xét_khen_thưởng kỷ_luật bình_xét điểm rèn_luyện lớp sinh_viên cũng là cơ_sở để hình_thành các chỉ đoàn thanh_niên chỉ hội sinh_viên sau 4 năm đối_với đại_học sau 3 năm đối_với cao_đẳng chuyên_nghiep chính_quy sau khi tốt_nghiep thì lớp sinh_viên tự giải_thể số sinh_viên.

Nhận xét:

Tính điểm số của các đoạn văn bị sai cho nên tìm câu trả lời bị sai.

Ví dụ 5:

Câu hỏi	không biết trách nhiệm của trạm y tế trường là gì ạ? em cảm ơn.
Đoạn văn 1	"Trách nhiệm của Phòng Công tác sinh viên. 1. Quản lý hồ sơ nhập học, hồ sơ miễn giảm học phí của sinh viên. 2. Theo dõi, tổng hợp kết quả rèn luyện sinh viên theo từng học kỳ, năm học và toàn khóa học. 3. Tham mưu cho Hội đồng khen thưởng và kỷ luật sinh viên để khen thưởng cá nhân và tập thể có thành tích cao trong học tập, rèn luyện và các hoạt động khác hoặc xử lý khi vi phạm quy chế, quy định, nội quy Nhà trường. 4. Xác nhận kết quả rèn luyện, cấp giấy chứng nhận tham gia Tuần sinh hoạt công dân sinh viên và các giấy tờ khác cho sinh viên thuộc thẩm quyền của đơn vị. 5. Tổ chức triển khai công tác giáo dục tư tưởng chính trị, đạo đức, lối sống, nhân cách cho sinh viên; phối hợp tổ chức cho sinh viên tham gia các hoạt động chính trị xã hội, văn hóa văn nghệ, thể dục thể thao và các hoạt động khác ngoài giờ lên lớp ở cấp trường. 6. Phối hợp với

	<p>các ngành, các cấp chính quyền địa phương trên địa bàn nơi trường trú đóng, xây dựng kế hoạch đảm bảo an ninh chính trị, trật tự và an toàn cho sinh viên; giải quyết kịp thời các vụ việc liên quan đến sinh viên. 7. Tổ chức tuyên truyền, giáo dục cho sinh viên các kiến thức cơ bản về kỹ năng sống, tư vấn học đường, giáo dục an toàn giao thông, phòng chống tội phạm và các tệ nạn xã hội; hướng dẫn sinh viên chấp hành các quy định của pháp luật và nội quy, quy định, quy chế của Nhà trường. 8. Phối hợp với Trung tâm Thư viện làm và cấp phát thẻ sinh viên. 9. Phối hợp tổ chức đối thoại định kỳ giữa Hiệu trưởng với sinh viên. 10. Tổ chức và phối hợp thực hiện các chế độ chính sách của Nhà nước quy định đối với sinh viên về học bổng, học phí, trợ cấp xã hội và các chế độ khác có liên quan đến sinh viên. 11. Tổ chức triển khai thực hiện công tác quản lý sinh viên ngoại trú theo quy định của Nhà trường.</p>
Đoạn văn 2	<p>Trách nhiệm của Phòng Tổ chức Hành chính. 1. Giải quyết các công việc hành chính có liên quan cho sinh viên; sao y các văn bằng, chứng chỉ của Trường. 2. Tiếp nhận và phân phối bưu phẩm, thư cho sinh viên.</p>
Đoạn văn 3	<p>Trách nhiệm của Trạm y tế. 1. Tổ chức thực hiện công tác y tế trường học; tổ chức khám sức khỏe cho sinh viên khi vào nhập học; chăm sóc, phòng chống dịch, bệnh, khám sức khỏe định kỳ và giải quyết các trường hợp sơ cấp cứu ban đầu cho sinh viên trong thời gian học tập tại Trường; Báo cáo những trường hợp không đủ tiêu chuẩn sức khỏe để học tập. 2. Triển khai công tác bảo hiểm y tế, bảo hiểm tai nạn cho sinh viên, phối hợp với cơ quan bảo hiểm, các phòng ban có liên quan giải quyết các trường hợp sinh viên bị ốm đau, tai nạn rủi ro. 3. Phổ biến, tuyên truyền nâng cao nhận thức cho sinh viên về ăn uống đảm bảo dinh dưỡng, vệ sinh an toàn thực phẩm, sinh hoạt điều độ, không lạm dụng rượu, bia, sử dụng chất kích thích, gây nghiện;</p>

	kiến thức và kỹ năng chăm sóc sức khỏe, phòng chống dịch, bệnh, tai nạn thương tích.
Đoạn văn 4	<p>Trách nhiệm Ban Cán sự lớp sinh viên. 1. Lớp sinh viên: Bao gồm những sinh viên cùng ngành, cùng khóa học. Lớp sinh viên được duy trì ổn định trong cả khóa học, là nơi để Nhà trường tổ chức, quản lý về thực hiện các nhiệm vụ học tập, rèn luyện, các hoạt động đoàn thể, các hoạt động xã hội, thi đua, khen thưởng, kỷ luật. 2. Ban Cán sự lớp sinh viên gồm: Lớp trưởng và các lớp phó do tập thể sinh viên trong lớp bầu, được Hiệu trưởng Nhà trường công nhận. Nhiệm kỳ Ban Cán sự lớp sinh viên theo năm học. Nhiệm vụ của Ban Cán sự lớp sinh viên: Tổ chức thực hiện các nhiệm vụ học tập, rèn luyện, các hoạt động sinh hoạt, đời sống và các hoạt động xã hội theo kế hoạch của Nhà trường; Đôn đốc sinh viên trong lớp chấp hành nghiêm chỉnh nội quy, quy chế về học tập, rèn luyện; xây dựng nề nếp tự quản trong lớp; Tổ chức, động viên giúp đỡ những sinh viên gặp khó khăn trong học tập, rèn luyện; thay mặt cho sinh viên của lớp liên hệ với giáo viên chủ nhiệm và các giáo viên bộ môn; đề nghị các khoa, đơn vị phụ trách công tác sinh viên và Ban Giám hiệu giải quyết những vấn đề có liên quan đến nhiệm vụ và quyền hạn của sinh viên trong lớp; Phối hợp chặt chẽ và thường xuyên với tổ chức Đoàn TNCS Hồ Chí Minh và Hội Sinh viên trong hoạt động của lớp; Báo cáo đầy đủ, chính xác tình hình học tập, rèn luyện theo học kỳ, năm học và những việc đột xuất của lớp với khoa hoặc đơn vị phụ trách công tác sinh viên. Quyền lợi của Ban Cán sự lớp sinh viên: Được ưu tiên cộng điểm rèn luyện và các chế độ khác theo quy định của Nhà trường.</p>

Không	0.22
Biết	0.17
Trách nhiệm	0.90

Của	0.23
Trạm	0.93
Y tế	0.86
Trường	0.91
là	0.34
gì	0.26
ạ	0.59
em	0.30
Cảm ơn	0.87

Đoạn văn	Điểm số
Đoạn 1	1.46
Đoạn 2	1.97
Đoạn 3	8.53
Đoạn 4	1.86

Câu trả lời: trách_nhiệm của trạm y_tế 1 tổ_chức thực_hiện công_tác y_tế trường_học tổ_chức khám sức_khoẻ cho sinh_viên khi vào nhập_học chăm_sóc phòng_chống dịch bệnh khám sức_khoẻ định_kỳ và giải_quyết các trường_hợp sơ_cấp cứu ban_đầu cho sinh_viên trong thời_gian học_tập tại trường báo_cáo những trường_hợp không đủ tiêu_chuẩn sức_khoẻ để học_tập 2 triển_khai công_tác bảo_hiểm_y_tế bảo_hiểm tai_nạn cho sinh_viên phối_hợp với cơ_quan bảo_hiểm.

Nhận xét:

 Tìm kiếm đoạn văn và câu trả lời chính xác.

CHƯƠNG 5 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi hiện nay rất được chú trọng đặc biệt bởi các nhà nghiên cứu và doanh nghiệp trong ngành công nghệ thông tin. Ở đây thì giai đoạn phân tích câu hỏi rất quan trọng trong ứng dụng. Trong khóa luận đã khảo sát và thử nghiệm qua các model, phân tích và đưa ra các phương pháp phù hợp cho ứng dụng này. Mô hình tìm trọng số của từ đã góp phần cải thiện các tác vụ như: tìm kiếm đoạn văn chứa câu trả lời và mô hình trích xuất câu trả lời từ đoạn văn bản.

Bên cạnh đó, do kiến thức còn hạn hẹp nên độ chính xác của mô hình chưa cao. Mô hình còn khá nặng mặc dù đã tái sử dụng đầu ra của các mô hình khác mà không cần phải tính toán lại.

5.2 Hướng phát triển

Để chương trình hoạt động hiệu quả hơn thì có rất nhiều hướng có thể thực hiện trong tương lai khi có nhiều thời gian hơn nữa chúng tôi xin đề xuất một vài giải pháp để cải thiện độ chính xác như sau:

Thêm nhiều dữ liệu hơn cho bài toán.

Cải thiện mô hình hơn nữa để đảm bảo chất lượng lẫn hiệu năng được tốt hơn.

Thử nghiệm trên nhiều mô hình mới để tìm ra được mô hình phù hợp với bài toán.

TÀI LIỆU THAM KHẢO

- [1] "TF-IDF," 11 2 2022. [Online]. Available: <https://en.wikipedia.org/wiki/Tf%E2%80%93idf>.
- [2] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network," 9 8 2018. [Online]. Available: <https://arxiv.org/abs/1808.03314>.
- [3] A. Vaswani, "Attention is all you need," 06 12 2017. [Online]. Available: <https://arxiv.org/abs/1706.03762>.
- [4] M.-W. C. K. L. K. T. Jacob Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 11 10 2018. [Online]. Available: <https://arxiv.org/abs/1810.04805>.
- [5] "BM25," 19 4 2018. [Online]. Available: <https://www.elastic.co/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables>.
- [6] S. A. Afshine Amidi, "Mạng neural hồi quy cheatsheet," [Online]. Available: <https://stanford.edu/~shervine/1/vi/teaching/cs-230/cheatsheet-recurrent-neural-networks>. [Accessed 14 2 2022].
- [7] N. S. N. P. J. U. L. J. A. N. G. L. K. I. P. Ashish Vaswani, "Attention Is All You Need," p. 6, 12 7 2017.
- [8] K. G. Dan Hendrycks, "Gaussian Error Linear Units (GELUs)," p. 2, 27 7 2016.

- [9] phamdinhkhanh, "Bài 36 - BERT model," 23 5 2020. [Online]. Available:
<https://phamdinhkhanh.github.io/2020/05/23/BERTModel.html>.
[Accessed 3 2022].
- [10] Y. Liu, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 26 7 2019. [Online]. Available:
<https://arxiv.org/abs/1907.11692>.
- [11] L. D. J. C. T. W. Victor Sanh, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," 2 10 2019.

Ứng dụng học sâu để cải tiến mô hình phân tích câu hỏi trong bài toán trả lời câu hỏi.
