TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

KHÓA LUẬN TỐT NGHIỆP

CHUYÊN NGÀNH KHOA HỌC MÁY TÍNH

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

Sinh viên: Trần Mỹ Linh 18066361, Đặng Văn Nghiêm 18056331
Giảng viên hướng dẫn: Ts. Đặng Thị Phúc

1.Đặt vấn đề

1. Hiện nay, với xu hướng số hóa tài liệu, đa số tài liệu văn bản được cập nhật lên mạng internet toàn cầu với số lượng lớn.

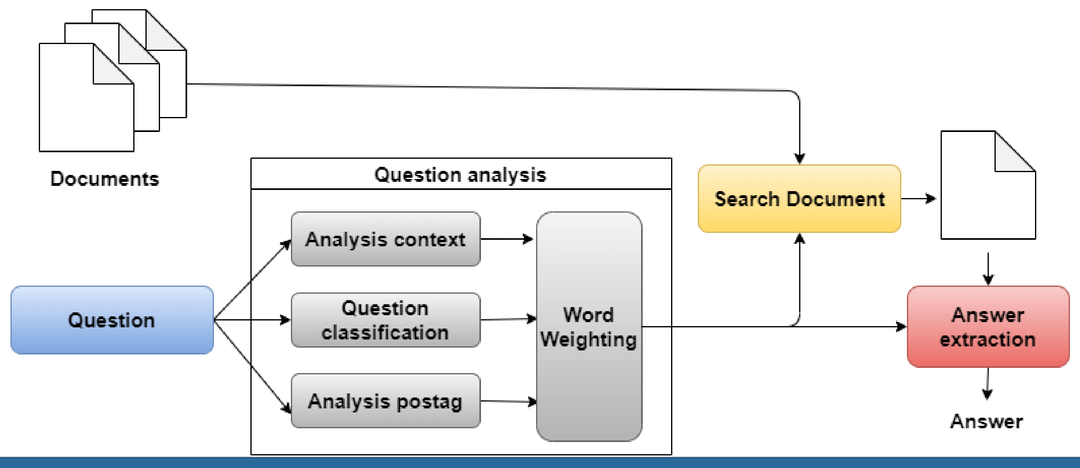
2. Phân tích câu hỏi là yếu tố đầu tiên trong kiến trúc chung của một hệ thống hỏi đáp, nó có nhiệm vụ tìm ra các thông tin cần thiết làm đầu vào cho quá trình xử lý của các quá trình sau

2. Mục tiêu đề tài

1. Phân tích câu hỏi: đánh trọng số cho các từ trong câu hỏi nhằm tìm ra những từ quan trọng trong câu hỏi.

2. Áp dụng trọng số vào các tác vụ phía sau: tìm kiếm đoạn văn chứa câu trả lời, trích xuất câu trả lời từ đoạn văn.

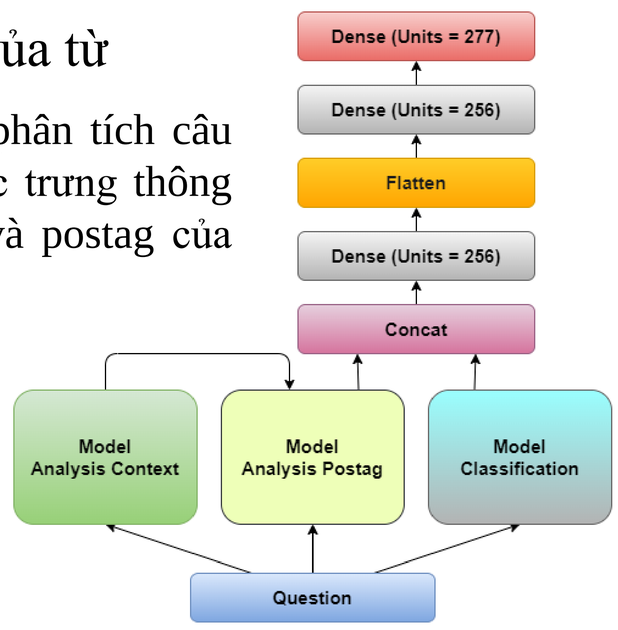
3. Phương hướng giải quyết



4. Các mô hình

Mô hình tìm trọng số của từ

Mô hình tìm trọng số sẽ phân tích câu hỏi bao gồm phân tích đặc trưng thông tin câu hỏi, loại câu hỏi và postag của các từ trong câu hỏi.



Ứng dụng trọng số vào thuật toán BM25

BM25 là thuật toán tìm kiếm tài liệu liên quan dựa vào câu hỏi

Công thức BM25:

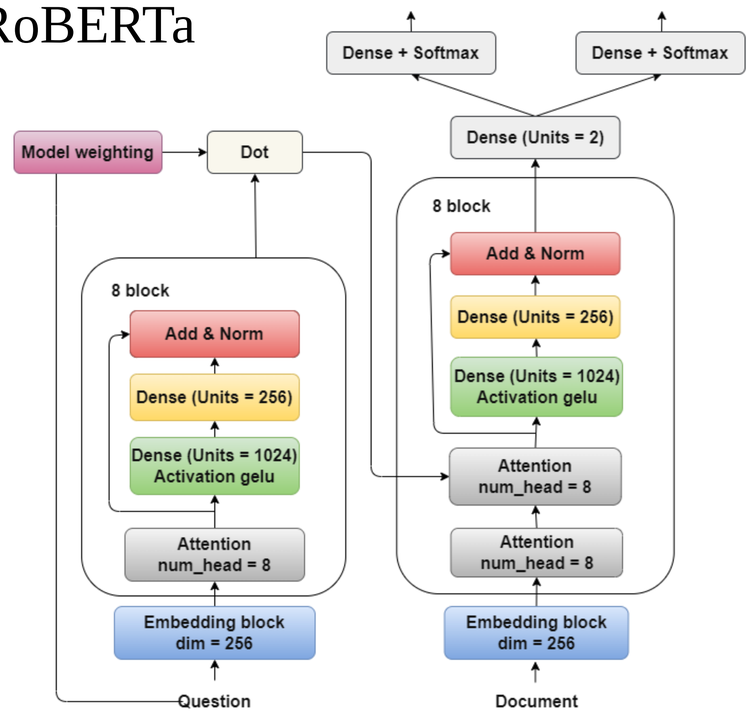
$$BM25 = \sum_{i}^{n} IDF(q_i) \frac{f(q_i, D) \times (k1 + 1 + w_i)}{f(q_i, D) + k1 \times \left(1 - b + b \times \frac{fieldLen}{avgFieldLen}\right)}$$

Công thức của IDF(q_i):

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

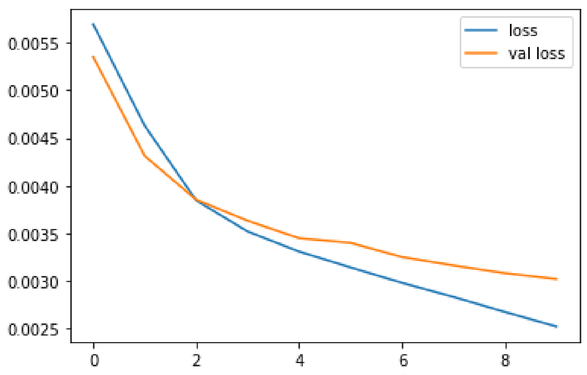
Tinh chỉnh mô hình RoBERTa

Mô hình sẽ được tinh chỉnh thêm decoder để tách thành 2 đầu vào cho câu hỏi và đoạn văn và áp dụng mô hình phân tích trọng số



5. Kết quả

Mô hình tìm trọng số của từ



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

Ứng dụng trọng số và thuật toán BM25

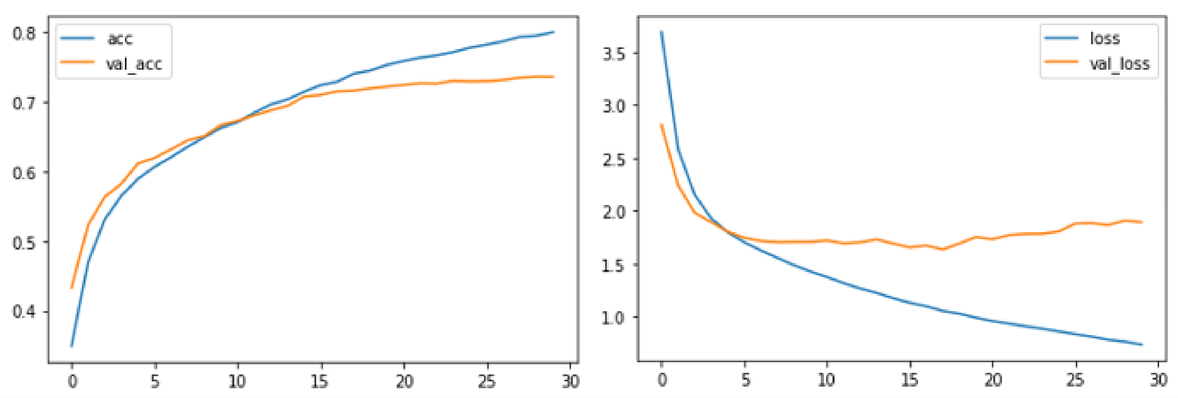
	Accuracy (%)	F1-score (%)
BM25	83.34	90.92
BM25 cải tiến	85.87	92.40

Mô hình RoBERTa

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

$$loss = \frac{loss\_start + loss\_end}{2}$$

Đánh giá trên tập test:
F1-score: 73%
EM: 73%



6. Hướng phát triển

Thêm nhiều dữ liệu hơn cho bài toán

Cải thiện mô hình để đảm bảo tính ứng dụng bởi vì mô hình hiện tại khá nặng.Thử nghiệm trên nhiều mô hình mới hơn nữa để tìm mô hình phù hợp với bài toán

Thông tin liên hệ:
Trần Mỹ Linh. Email: tranmylinh26042000@gmail.com Đặng Văn Nghiêm. Email: vannghiem848@gmail.com