ỨNG DỤNG TẠO CÂU TRUY VẤN SQL TỪ VĂN BẢN TIẾNG VIỆT SỬ DỤNG KỸ THUẬT XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Nguyễn Phạm Minh Quân

Trường Đại học Công nghệ Thông tin - ĐHQG TP.HCM

Mục tiêu

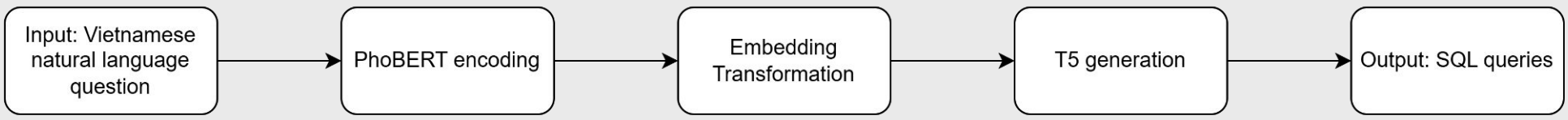
Báo cáo này đưa ra một ứng dụng chuyển đổi văn bản tiếng Việt thành câu truy vấn SQL, trong đó:

- Kết hợp khả năng dịch văn bản của mô hình T5 và khả năng hiểu biết sâu về tiếng Việt của mô hình PhoBERT để tăng độ chính xác cho quá trình chuyển đổi văn bản tiếng Việt thành câu truy vấn SQL.
- Xây dựng pipeline kết hợp hai mô hình trên và ứng dụng chatbot để sử dụng.

Lý do chọn đề tài

- Ngày nay,rất nhiều thông tin được lưu trữ trong các cơ sử dữ liệu quan hệ. Vì thế việc viết câu lệnh truy vấn SQL đã trở thành một điều thiết yếu để có thể truy vấn thông tin từ các cơ sở dữ liệu này. Tuy nhiên, để sử dụng thành thạo ngôn ngữ SQL trong việc truy vấn dữ liệu, cần tới rất nhiều thời gian và nỗ lực để học tập.
- Mô hình T5 ứng dụng trong các tác vụ dịch máy hiện nay không thể nắm bắt đầy đủ ngữ nghĩa và sắc thái trong tiếng Việt.

Tổng quan



Hình 1. Mô hình tích hợp PhoBERT và T5

- **Input:** Câu hỏi bằng tiếng Việt chứa yêu cầu truy vấn dữ liệu mà người dùng nhập vào (VD: Lấy tất cả dữ liệu trong bảng HocVien)
- **Output:** Câu truy vấn SQL dùng để thực hiện truy vấn theo yêu cầu của người dùng (VD: SELECT * FROM HocVien)

- **PhoBERT (Phoneme-based Vietnamese BERT):** Mô hình ngôn ngữ transformer được huấn luyện trên dữ liệu văn bản tiếng Việt.
- **T5 (Text-to-text Transfer Transformer):** Mô hình ngôn ngữ transformer có khả năng thực hiện nhiều nhiệm vụ xử lý ngôn ngữ tự nhiên khác nhau.

Mô tả

1. Nội dung

- Nghiên cứu mô hình T5 nhằm tạo câu truy vấn SQL và mô hình PhoBERT nhằm phân tích ngữ nghĩa và sắc thái của văn bản tiếng Việt đầu vào và tạo input cho T5.
- Fine-tune mô hình T5 bằng bộ dữ liệu ViText2SQL giúp cho mô hình học thêm về ngữ nghĩa và sắc thái trong tiếng Việt để tạo câu truy vấn SQL chính xác hơn
- Xây dựng ứng dụng chatbot nhằm thực hiện các yêu cầu tạo câu truy vấn SQL của người dùng.

2. Phương pháp

- Tìm hiểu cấu trúc mô hình PhoBERT và T5, xây dựng pipeline kết hợp cả hai mô hình này để cho ra câu truy vấn SQL từ văn bản tiếng Việt.
- Fine-tune mô hình T5 bằng bộ dữ liệu ViText2SQL.
- Xây dựng cơ chế segment input của ứng dụng trước khi đưa vào mô hình PhoBERT.
- Xây dựng cơ chế biến đổi embedding sinh ra từ PhoBERT thành input phù hợp cho mô hình T5.
- Xây dựng ứng dụng dựa trên pipeline của hai mô hình PhoBERT và T5 để thực hiện các yêu cầu tạo câu truy vấn SQL của người dùng.

3. Kết quả mong đợi

- Ứng dụng chatbot có thể đưa ra các câu truy vấn SQL từ văn bản tiếng Việt được nhập vào.
- Mô hình kết hợp giữa PhoBERT và T5 có thể tạo câu truy vấn SQL từ văn bản đầu vào bằng tiếng Việt với độ chính xác đạt từ 70% trở lên.

NII

Nguyễn Phạm Minh Quân – Trường Đại học Công nghệ Thông tin - ĐHQG TP.HCM

TEL : +84988536799 Email : quannpm.18@grad.uit.edu.vn