SQL QUERIES GENERATING FROM VIETNAMESE LANGUAGE TEXT APPLICATION USING NATURAL LANGUAGE PROCESSING TECHNIQUE

Nguyễn Phạm Minh Quân - 230201052

Tóm tắt

- Lớp: CS2205.MAR2024
- Link Github: https://github.com/Biscottezi/CS2205.MAR2024.git
- Link YouTube video: https://youtu.be/2XJ_2MC2zlo
- Tên: Nguyễn Phạm Minh Quân

Giới thiệu

- 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
- Để thành thạo SQL cần rất nhiều thời gian và nỗ lực.

=> Một ứng dụng giúp chuyển đổi văn bản tiếng Việt thành câu truy vấn SQL.

Giới thiệu

- Mô hình T5 đã được chứng minh là có thể được ứng dụng để chuyển đổi ngôn ngữ tự nhiên thành câu truy vấn SQL.
- T5 không thể hiểu sâu sắc ngữ nghĩa và các sắc thái trong tiếng Việt.

=> Kết hợp T5 với PhoBERT nhằm tăng độ chính xác khi chuyển đổi ngôn ngữ tiếng Việt thành câu truy vấn SQL.

Giới thiệu

- **Input:** Câu hỏi bằng tiếng Việt.

  VD: Lấy tất cả dữ liệu trong bảng HocVien
- **Output:** Câu truy vấn SQL.

  VD: SELECT * FROM HocVien

Mục tiêu

- Nghiên cứu độ hiệu quả của việc kết hợp mô hình PhoBERT và T5 để tăng độ chính xác trong việc tạo câu truy vấn SQL từ văn bản tiếng Việt.
- Tạo ứng dụng chatbot giúp tạo câu truy vấn SQL từ văn bản tiếng Việt.

Nội dung và 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 chia input của người dùng thành các segment phù hợp trước khi đưa vào mô hình PhoBERT.
- Xây dựng cơ chế biến đổi output của mô hình PhoBERT thành input phù hợp với 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.

Kết quả dự kiến

- 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.

Tài liệu tham khảo

[1]. Dat Quoc Nguyen, Anh Tuan Nguyen:
PhoBERT: Pre-trained language models for Vietnamese. CoRR abs/2003.00744 (2020)
[2]. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu:
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. CoRR abs/1910.10683 (2019)
[3]. Albert Wong, Lien Pham, Young Lee, Shek Chan, Razel Sadaya, Youry Khmelevsky, Mathias Clement, Florence Wing Yau Cheng, Joe Mahony, Michael Ferri:
Translating Natural Language Queries to SQL Using the T5 Model. CoRR abs/2312.12414 (2023)
[4]. Anh Tuan Nguyen, Mai Hoang Dao, Dat Quoc Nguyen:
A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese. EMNLP (Findings) 2020: 4079-4085