


# THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):  
[https://youtu.be/2XJ\\_2MC2zlo](https://youtu.be/2XJ_2MC2zlo)
- Link slides (dạng .pdf đặt trên Github):  
<https://github.com/Biscottezi/CS2205.MAR2024/blob/5f0a8d75729f4f133645fcd9cc0a5fdf1c4dcb58/SQL%20QUERIES%20GENERATING%20FROM%20VIETNAMESE%20LANGUAGE%20TEXT%20APPLICATION%20USING%20NATURAL%20LANGUAGE%20PROCESSING%20TECHNIQUE.pdf>

<ul style="list-style-type: none"><li>• Họ và Tên: Nguyễn Phạm Minh Quân</li><li>• MSSV: 230201052</li></ul> 	<ul style="list-style-type: none"><li>• Lớp: CS2205.MAR2024</li><li>• Tự đánh giá (điểm tổng kết môn): 7.0/10</li><li>• Số buổi vắng: 0</li><li>• Số câu hỏi QT cá nhân: 0</li><li>• Link Github: <a href="https://github.com/Biscottezi/CS2205.MAR2024.git">https://github.com/Biscottezi/CS2205.MAR2024.git</a></li></ul>
--	---

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

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

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

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

## TÓM TẮT (Tối đa 400 từ)

Ngày nay, rất nhiều thông tin được lưu trữ trong các cơ sở 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. Do đó, chúng tôi nhận thấy việc chuyển đổi ngôn ngữ tự nhiên thành ngôn ngữ truy vấn SQL có thể hỗ trợ tăng năng suất cho công việc truy vấn thông tin từ các cơ sở dữ liệu.

Đề tài này có mục đích nghiên cứu và cho ra một ứng dụng chatbot, trong đó kết hợp khả năng thực hiện nhiệm vụ dịch máy của mô hình T5 và khả năng hiểu sâu về tiếng Việt của mô hình PhoBERT nhằm giải quyết bài toán tạo câu truy vấn SQL từ văn bản tiếng Việt, với độ chính xác từ 70% trở lên.

## GIỚI THIỆU (Tối đa 1 trang A4)

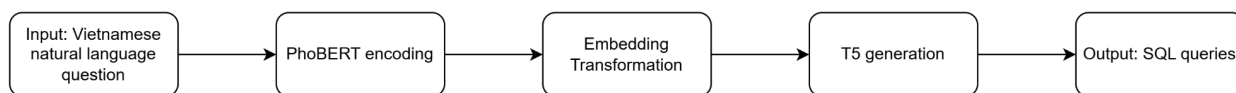
Ngày nay, rất nhiều thông tin được lưu trữ trong các cơ sở dữ liệu quan hệ. Vì thế việc viết câu lệnh truy vấn SQL đã trở thành một điều thiết yếu để có thể truy vấn thông tin từ các cơ sở dữ liệu này. Tuy nhiên, để sử dụng thành thạo ngôn ngữ SQL trong việc truy vấn dữ liệu, cần tới rất nhiều thời gian và nỗ lực để học tập.

Natural Language Processing (NLP), hay Xử lý ngôn ngữ tự nhiên, là một kỹ thuật được sử dụng phổ biến trong việc tương tác giữa con người và máy tính. Đây là một lĩnh vực trong nghiên cứu trí tuệ nhân tạo (AI) tập trung vào việc nghiên cứu và phân tích ngôn ngữ tự nhiên, với mục tiêu tìm hiểu và xử lý thông tin trong ngôn ngữ hàng

ngày của con người. Tóm lại, NLP giúp máy tính hiểu được ngôn ngữ con người và tương tác với con người một cách tự nhiên.

Từ thực tiễn khó khăn của việc phải thành thạo SQL, chúng tôi đề xuất xây dựng một ứng dụng giúp tạo câu truy vấn SQL từ văn bản tiếng Việt nhằm hỗ trợ cho công việc cần phải truy vấn dữ liệu nhiều. Đồng thời ứng dụng có thể đóng vai trò như một nguồn tham khảo cho những người đang học SQL. T5 [2], một mô hình transformer sequence-to-sequence đã được chứng minh là có khả năng chuyển đổi văn bản ngôn ngữ tự nhiên thành câu truy vấn SQL với độ chính xác cao [3]. Tuy nhiên, mô hình T5 không thể nắm bắt đầy đủ ngữ nghĩa và sắc thái trong tiếng Việt. Ngược lại, PhoBERT [1], một mô hình transformer được công bố bởi VinAI Research, được thiết kế đặc biệt để biểu diễn tiếng Việt. Do đó, luận văn này đề xuất kết hợp mô hình PhoBERT và T5 nhằm tăng độ chính xác cho các câu truy vấn được tạo ra từ văn bản tiếng Việt. Cụ thể:

- Input: Câu hỏi bằng tiếng Việt (VD: Lấy tất cả dữ liệu trong bảng HocVien)
- Output: Câu truy vấn SQL (VD: SELECT \* FROM HocVien)



Hình 1. Mô hình tích hợp PhoBERT và T5

## 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

### 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 [4] 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.

#### **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.

#### **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.

#### **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