ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Ngày nhận hồ sơ	
(Do CQ quản lý ghi)	

THUYẾT MINH
ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ CẤP SINH VIÊN 2022

A. THÔNG TIN CHUNG

A1. Tên đề tài

- Tên tiếng Việt (IN HOA): UIT-VICOVIDQA: HỎI ĐÁP DỰA VÀO CỘNG ĐỒNG VỀ COVID-19 TRÊN DỮ LIỆU TIẾNG VIỆT.

- Tên tiếng Anh (IN HOA): UIT-VICOVIDQA: A NOVEL DATASET FOR COVID-19 COMMUNITY-BASED QUESTION ANSWERING ON VIETNAMESE LANGUAGE.

A2. Thời gian thực hiện

..06.. tháng (kể từ khi được duyệt).

A3. Tổng kinh phí

(Lưu ý tính nhất quán giữa mục này và mục B8. Tổng hợp kinh phí đề nghị cấp)

Tổng kinh phí: …**6**.. triệu đồng, gồm

- Kinh phí từ Trường Đại học Công nghệ Thông tin: ..**6**.. triệu đồng

A4. Chủ nhiệm

Họ và tên: **Thái Minh Triết** .

Ngày, tháng, năm sinh: 25/07/2000 . Giới tính (Nam/Nữ): Nam.

Số CMND: 092200003803 ; Ngày cấp: 15/06/2016 ; Nơi cấp: Cục Cảnh sát quản lý hành chính về trật tự xã hội .

Mã số sinh viên: 19522397 .

Số điện thoại liên lạc: 0868075398 .

Đơn vị (Khoa): Khoa học và Kỹ thuật thông tin .

Số tài khoản: 0111000362124 Ngân hàng: Vietcombank

A5. Thành viên đề tài

TT	Họ tên	MSSV	Khoa
1	Thái Minh Triết	19522397	KH&KTTT
2	Chu Hà Thảo Ngân	19521882	KH&KTTT
3	Võ Tuấn Anh	19521226	KH&KTTT

B. MÔ TẢ NGHIÊN CỨU

B1. Giới thiệu về đề tài

Tình hình nghiên cứu liên quan đến đề tài:

Trên thế giới đã có nhiều bộ dữ liệu hỏi đáp về lĩnh vực y khoa được công bố trong những năm gần đây. Có thể nhắc đến bộ dữ liệu COVIDRead [1], với cấu trúc tương tự với bộ dữ liệu SquAD của đại học Stanford, gồm hơn 100 nghìn cặp câu hỏi đáp tiếng Anh về COVID-19 dựa trên văn bản ngữ cảnh. Công trình nghiên cứu của Ben Abacha và Demner-Fushman năm 2019 [2] cho thấy một hướng tiếp cận mới mẻ trong lĩnh vực hỏi đáp y khoa khi xây dựng bộ dữ liệu MedQuAD gồm 47,457 cặp câu hỏi đáp y khoa thực tế cùng các phương pháp RQE kết hợp với các mô hình trích xuất thông tin. Một nghiên cứu khác của tác giả Kacupaj và các đồng sự dựa trên bộ dữ liệu ParaQA [3] tuy không liên quan đến chủ đề y khoa nhưng hướng tiếp cận của họ sử dụng kỹ thuật diễn giải câu trả lời trên các mô hình Sequence-to-Sequence đã cho thấy sự hiệu quả trên bài toán hỏi đáp trong hội thoại theo lượt một (Single-turn Conversational Question Answering).

Lí do thực hiện đề tài:

Tính từ đợt bùng phát dịch lần thứ 4 đến nay, tình hình dịch bệnh COVID-19 ở Việt Nam vẫn chưa được kiểm soát hoàn toàn, khi mỗi ngày nước ta ghi nhận hàng chục nghìn ca nhiễm mới, trong khi số ca tử vong vẫn chưa dừng lại. Bên cạnh đó, sự lan nhanh của những thông tin sai sự thật khiến một bộ phận người dân gặp nhiều khó khăn, mất phương hướng trong việc tiếp cận được những nguồn tin chính thống để bảo vệ sức khỏe cho chính mình. Đặc biệt trong bối cảnh nước ta đang bước vào trạng thái "bình thường mới", thì ý thức phòng bệnh là một trong những yếu tố không chỉ quyết định đến tình hình dịch bệnh mà còn quyết định đến hoạt động giáo dục, y tế, lao động, sản xuất, phát triển kinh tế, du lịch,,.. của toàn thể xã hội.

Nhận thấy sự cần thiết phải xây dựng một bộ dữ liệu tiếng Việt phục vụ cho các hệ thống hỏi đáp tự động về chủ đề COVID-19, chúng tôi quyết định lựa chọn và thực hiện nghiên cứu đề tài "UIT-ViCovidQA: HỎI ĐÁP DỰA VÀO CỘNG ĐỒNG VỀ COVID-19 TRÊN DỮ LIỆU TIẾNG VIỆT". Mục tiêu chính của đề tài nhằm cung cấp cho cộng đồng nghiên cứu trong nước một bộ dữ liệu hỏi đáp dựa trên cộng đồng đầu tiên về chủ đề COVID-19 cho tiếng Việt - một ngôn ngữ low-resource, cũng như thực nghiệm các mô hình học sâu trên bộ dữ liệu để đưa ra kết quả benchmark nhằm đánh giá khả năng giải đáp các câu hỏi về dịch bệnh COVID-19.

Các thách thức:

Hiện tại chưa có bộ dữ liệu chính thức cho tác vụ hỏi đáp dựa trên cộng đồng về COVID-19 cho tiếng Việt. Để tiến hành nghiên cứu, chúng tôi cần xây dựng một bộ dữ liệu hoàn chỉnh,

đáng tin cậy, có kích thước đủ lớn và phải được biên soạn trên ngôn ngữ tiếng Việt. Yêu cầu đó đã đặt ra những thách thức nhất định cho đề tài nghiên cứu này.

Thứ nhất, các nguồn hỏi đáp chính thống về COVID-19 cho Tiếng Việt vẫn còn nhiều hạn chế so với những nguồn hỏi đáp trên tiếng Anh, ngoài ra để có các câu hỏi đáp chất lượng đòi hỏi chúng phải được biên soạn và cập nhật thường xuyên bởi đội ngũ y, bác sĩ có trình độ chuyên môn cao thuộc các tổ chức y tế uy tín tại Việt Nam và nước ngoài. Đó là chưa kể đến sự đa dạng về nền tảng đăng tải của mỗi tổ chức, khiến cho việc chọn lọc và thu thập dữ liệu hỏi đáp COVID-19 cho tiếng Việt gặp nhiều khó khăn.

Thứ hai, bộ dữ liệu phục vụ bài toán hỏi đáp dựa trên cộng đồng thường chứa nội dung hỏi đáp dài, có thể là một tập hợp nhiều câu văn hoặc nhiều đoạn văn bản cho mỗi điểm dữ liệu. Điều này gây hao tổn chi phí và tài nguyên tính toán, ảnh hưởng lớn đến thời gian huấn luyện và hiệu suất của các mô hình học sâu.

Thách thức cuối cùng là về mặt đa dạng ngữ nghĩa của ngôn ngữ tiếng Việt và việc thiếu hụt một độ đo tiêu chuẩn đặc thù cho tác vụ hỏi đáp trên ngôn ngữ này khiến cho việc đánh giá mô hình chưa thể đạt được sự khách quan và toàn diện để có thể xem xét khả năng áp dụng kết quả vào thực tế.

B2. Mục tiêu, nội dung, kế hoạch nghiên cứu

B2.1 Mục tiêu

Mục tiêu nghiên cứu của chủng tôi trước tiên là xây dựng một bộ dữ liệu hỏi đáp dựa trên cộng đồng đầu tiên về COVID-19 cho tiếng Việt. Sau quá trình thu thập, chúng tôi mở rộng bộ dữ liệu bằng cách diễn giải (paraphrase) thêm các câu trả lời dựa trên câu trả lời gốc, nhằm đánh giá khả năng cải thiện hiệu suất cho các mô hình baseline. Bộ dữ liệu hoàn chỉnh phải đáp ứng được tính chính xác, kịp thời và phù hợp với tình hình dịch bệnh, bản sắc văn hóa, lối sống của người dân Việt Nam.

Sau khi đã có bộ dữ liệu hoàn chỉnh, chúng tôi sẽ tiến hành thực nghiệm các mô hình baseline theo kiến trúc Encoder-Decoder và đánh giá mô hình thông qua các độ đo tiêu chuẩn trên tác vụ này là BLEU, METEOR và ROUGE-L. Từ những kết quả đó, chúng tôi có thể thẩm định chất lượng của bộ dữ liệu; phân tích, so sánh và đánh giá hiệu quả của quá trình diễn giải câu trả lời đến hiệu suất các mô hình và đưa ra kết quả benchmark cho những nghiên cứu tiếp theo trong tương lai.

B2.2 Nội dung và phương pháp nghiên cứu

Nội dung 1: Xây dựng UIT-ViCovidQA, bộ dữ liệu hỏi đáp dựa trên cộng đồng đầu tiên về COVID-19 cho tiếng Việt:

- Phương pháp:

+ Thu thập các cặp câu hỏi-đáp tiếng Việt từ các trang thông tin hỏi đáp và tư vấn trực tuyến về COVID-19 của các tổ chức y tế uy tín trong và ngoài nước, như Trung tâm kiểm soát và phòng ngừa dịch bệnh Hoa Kỳ (CDC), Quỹ Nhi đồng Liên Hợp Quốc (Unicef), Cổng thông tin điện tử chính phủ Việt Nam, Học viện Quân y, Hệ thống y tế VinMec, Báo điện tử VnExpress,...

+ Nội dung hỏi đáp xoay quanh các chủ đề về COVID-19 và những chính sách liên quan bao gồm: nguồn gốc và tên gọi của dịch bệnh; cách thức lây nhiễm; triệu chứng thường gặp; cách phòng ngừa, điều trị bệnh và đảm bảo dinh dưỡng; các biến thể và ảnh hưởng của chúng; vaccine, đối tượng tiêm chủng và quá trình tiêm chủng; nhập cảnh, du lịch và di chuyển giữa các địa phương; cách ly y tế và giãn cách xã hội; quy định, chỉ thị và chế tài xử lý vi phạm; các mô hình điều trị bệnh nhân COVID-19; chính sách hỗ trợ người lao động gặp khó khăn do COVID-19; di chứng hậu COVID-19 và cách điều trị; COVID-19 ở trẻ em. Bộ dữ liệu này không bao gồm hỏi đáp về số liệu COVID-19 tại Việt nam và trên thế giới.

+ Thực hiện tiền xử lý dữ liệu bằng cách loại bỏ các câu hỏi đáp trùng lặp và các câu hỏi đáp không liên quan COVID-19, không đảm báo tính chính xác, không phù hợp với văn hóa lối sống và tình hình dịch bệnh COVID-19 tại Việt Nam hiện nay.

+ Tạo thêm tối đa ba câu trả lời cho câu hỏi bằng phương pháp diễn giải (paraphrase) dựa trên câu trả lời gốc. Các kỹ thuật diễn giải gồm có:

- Chuyển đổi vị trí các từ, hoặc sắp xếp lại các câu với nhau sao cho tạo thành câu trả lời khác mà vẫn giữ nguyên ý nghĩa của câu trả lời gốc;
- Tự đặt câu trả lời mới, cắt giảm nội dung ngắn gọn hoặc giải thích một cách chi tiết hơn;
- Đặt câu trả lời mới tương tự sử dụng các từ ngữ đồng nghĩa với nhau hoặc diễn đạt từ ngữ theo các cách khác nhau.

-Kết quả dự kiến: Bộ dữ liệu UIT-ViCovidQA hoàn chỉnh dự kiến bao gồm từ 4,500 đến 5,000 cặp câu hỏi-đáp COVID-19 tiếng Việt, trong đó mỗi câu hỏi sẽ có tối thiểu một câu trả lời và có tối đa bốn câu trả lời, các câu trả lời mới được diễn giải theo các cách khác nhau từ câu trả lời ban đầu.

Nội dung 2: Thực nghiệm các mô hình Sequence-to-Sequence

- Phương pháp:

+ Thiết lập các mô hình Sequence-to-Sequence sử dụng kiến trúc Encoder-Decoder gồm: Mô hình Recurrent Neural Netwok (RNN) sử dụng cơ chế Attention gồm Badahnau Attention [4] và Luong Attention [5]; mô hình mạng Transformer [6] và mô hình mạng Convolutional [7].

+ Để đảm bảo sự khách quan khi so sánh hiệu suất các mô hình, chúng tôi dự kiến thiết lập cùng các siêu tham số như sau: hai mô hình RNN đều sử dụng nhân là GRU; cấu trúc mỗi mô hình gồm 2 lớp với số chiều embedding là 512 cùng chỉ số drop out được thiết lập 0.5. Huấn luyện các mô hình với batch size = 8 trên 30 epoch. Ở mỗi mô hình, chúng tôi dự định sử dụng thuật toán tối ưu Adam với learning_rate = 0.001 cùng hàm loss là Cross Entropy Loss.

+ Lần lượt huấn luyện từng mô hình trên từng điều kiện thực ngiệm bao gồm sử dụng tối đa: một câu trả lời, hai câu trả lời, ba câu trả lời và bốn câu trả lời.

-Kết quả dự kiến: Sau khi kết thúc mỗi thực nghiệm, các kết quả sau cùng gồm: tham số mô hình, giá trị loss của quá trình huấn luyện, kết quả độ đo BLEU, METEOR, ROUGE-L cùng câu trả lời dự đoán trên tập kiểm tra sẽ được lưu lại để tiến hành bước so sánh, phân tích và đánh giá kết quả.

Nội dung 3: Đánh giá hiệu suất mô hình, phân tích lỗi và đưa ra hướng phát triển

- Phương pháp:

+ Lập bảng tổng hợp kết quả các độ đo BLEU, METEOR và ROUGE-L cho mỗi mô hình theo từng điều kiện thực nghiệm cụ thể (số lượng câu trả lời tối đa được sử dụng cho mỗi câu hỏi).

+ Trên mỗi mô hình, thống kê kết quả tốt nhất trong các điều kiện thực nghiệm. So sánh các câu trả lời dự đoán ở mỗi điều kiện nhằm phát hiện và phân tích những mặt hạn chế của từng điều kiện.

+ Trên mỗi điều kiện thực nghiệm, thống kê kết quả tốt nhất trong các mô hình. So sánh các câu trả lời dự đoán của mỗi mô hình nhằm phát hiện và phân tích những mặt hạn chế của từng mô hình.

-Kết quả dự kiến:

+ Đưa ra kết quả benchmark trên bộ dữ liệu UIT-ViCovidQA làm cơ sở so sánh cho các nghiên cứu trong tương lai.

+ Đưa ra kết luận về mô hình tốt nhất và những mặt hạn chế của từng mô hình ảnh hưởng đến việc dự đoán câu trả lời.

+ Đưa ra kết luận về điều kiện thực nghiệm tốt nhất và những mặt hạn chế của từng điều kiện thực nghiệm ảnh hưởng đến việc dự đoán câu trả lời.

+ Đưa những thách thức và hướng phát triển của bộ dữ liệu.

B2.3 Kế hoạch nghiên cứu.

Thời gian thực hiện	Nội dung thực hiện	Phân công thực hiện	Ghi chú
01/04/2022 – 20/06/2022	Thu thập dữ liệu hỏi đáp tiếng Việt về COVID-19	Thái Minh Triết Chu Hà Thảo Ngân Võ Tuấn Anh	
21/06/2022 – 30/06/2022	Tiền xử lý dữ liệu đã thu thập	Thái Minh Triết	
01/07/2022 – 31/08/2022	Diễn giải thêm các câu trả lời từ câu trả lời gốc	Thái Minh Triết Chu Hà Thảo Ngân Võ Tuấn Anh	
01/09/2022 – 10/09/2022	Thực nghiệm các mô hình baseline trên bộ dữ liệu	Thái Minh Triết Chu Hà Thảo Ngân Võ Tuấn Anh	
11/09/2022 – 30/09/2022	So sánh kết quả và phân tích lỗi của các mô hình. Đưa ra những thách thức và hướng phát triển cho bộ dữ liệu	Thái Minh Triết Chu Hà Thảo Ngân Võ Tuấn Anh	

B3. Kết quả dự kiến

Kết quả nghiên cứu dự kiến cung cấp cho cộng đồng bộ dữ liệu UIT-ViCovidQA phục vụ tác vụ hỏi đáp về COVID-19 cho tiếng Việt. Bộ dữ liệu gồm 4,500 cặp câu hỏi đáp đảm bảo tính tin cậy, chính xác và phù hợp với mục tiêu nghiên cứu của đề tài.

Thông qua quá trình thực nghiệm, chúng tôi dự kiến đưa ra kết quả benchmark làm cơ sở so sánh cho các nghiên cứu sau này. Chúng tôi cũng kỳ vọng vào khả năng của các mô hình Sequence-to-Sequence trên tác vụ hỏi đáp trên tiếng Việt, đặc biệt là mô hình Transformer, vốn là kiến trúc đạt những kết quả nổi bật ở nhiều nghiên cứu trong lĩnh vực Xử lý ngôn ngữ tự nhiên những năm gần đây.

Ngoài ra, nghiên cứu của chúng tôi cho thấy được trên tác vụ hỏi đáp, việc sử dụng nhiều câu trả lời được diễn giải trong quá trình thực nghiệm là có cơ sở, giúp cho các mô hình Sequence-

to-Sequence đạt hiệu suất tốt hơn, hạn chế được tình trạng overfitting và hiện tượng thoái hóa văn bản.

B4. Tài liệu tham khảo

[1]. “COVIDRead: A Large-scale Question Answering Dataset on COVID-19”; Saikh, Tanik and Sahoo, Sovan Kumar and Ekbal, Asif and Bhattacharyya, Pushpak; EuropePMC, 2021.

[2]. "A Question-Entailment Approach to Question Answering"; Asma Ben Abacha and Dina Demner-Fushman; BMC Bioinformatics, 2019.

[3]. “ParaQA: A Question Answering Dataset with Paraphrase Responses for Single-Turn Conversation”; Kacupaj Endri, Banerjee Barshana, Singh Kuldeep and Lehmann Jens; European Semantic Web Conference, 2021.

[4]. “Neural Machine Translation by Jointly Learning to Align and Translate”; Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio; International Conference on Learning Representations (ICLR), 2015.

[5]. “Effective Approaches to Attention-based Neural Machine Translation”; Minh-Thang Luong, Hieu Pham, Christopher D. Manning; Empirical Methods in Natural Language Processing (EMNLP), 2015.

[6]. “Attention is All you Need”; Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin; Neural Information Processing Systems (NeurIPS), 2017.

[7]. “Convolutional Sequence to Sequence Learning”; Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin; International Conference on Machine Learning (ICML), 2017.

*Ngày __ tháng __ năm 20_*	*Ngày __ tháng __ năm 20_*
Giảng viên hướng dẫn	**Chủ nhiệm đề tài**
(Ký và ghi rõ họ tên)	(Ký và ghi rõ họ tên)