|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 2021

# THÔNG TIN CHUNG

## A1. Tên đề tài

* Tên tiếng Việt: NGHIÊN CỨU ĐỌC HIỂU TỰ ĐỘNG TRÊN VĂN BẢN TIN TỨC TIẾNG VIỆT
* Tên tiếng Anh: MACHINE READING COMPREHENSION FOR VIETNAMESE NEWS

## A2. Loại hình nghiên cứu

*(Tham khảo tiêu chuẩn đề tài đối với từng loại hình NC, chọn 01 trong 03 loại hình)*

 Nghiên cứu cơ bản

 Nghiên cứu ứng dụng

 Nghiên cứu triển khai

## A3. Thời gian thực hiện

**..06..** tháng (kể từ khi được duyệt).

## A4. 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

## A5. Chủ nhiệm

Họ và tên: **Nguyễn Thị Hồng Nhung**.

Ngày, tháng, năm sinh: 28/05/2000 Giới tính (Nam/Nữ): Nữ

Số CMND: 281230339 Ngày cấp: 11/06/2015 Nơi cấp: Bình Dương

Mã số sinh viên: 18521218.

Số điện thoại liên lạc: 0327004473.

Đơn vị: Khoa Khoa học và Kỹ thuật Thông tin.

## A6. Nhân lực nghiên cứu

|  |  |  |  |
| --- | --- | --- | --- |
| **TT** | **Họ tên** | **MSSV** | **Khoa/ Bộ Môn** |
| 1 | Nguyễn Thị Hồng Nhung | 18521218 | Khoa KH&KTTT |
| 2 | Hà Phan Diệu Phương | 18521268 | Khoa KH&KTTT |

# MÔ TẢ NGHIÊN CỨU

## B1. Giới thiệu về đề tài

*(Ghi các ý về tổng quan tình hình nghiên cứu liên quan đến đề tài, lí do thực hiện đề tài, các thách thức)*

### B1.1. Lý do chọn đề tài

Question Answering System (Tạm dịch: “Hệ thống trả lời câu hỏi”) được định nghĩa là một hệ thống mô hình máy tính tự động trả lời các câu hỏi do con người đặt ra bằng ngôn ngữ tự nhiên. Hệ thống trả lời câu hỏi hiện đại có hai thành phần chính [1], trong đó thành phần đầu tiên để truy xuất thông tin chọn các đoạn văn bản liên quan đến các câu hỏi từ corpus (retrieval component) và thành phần thứ hai đọc hiểu tự động (MRC component) để trích xuất các câu trả lời trong đoạn văn đã tìm và trả lời cho người dùng. Đầu ra của hệ thống này là một câu trả lời được truy vấn ở một nguồn tri thức nào đó như một đoạn văn bản hay là từ một trang web có chứa lượng thông tin đầy đủ và luôn được cập nhật như Wikipedia. Hiện nay, có hai hệ thống trả lời câu hỏi dựa trên nguồn tri thức thông tin mà nó sử dụng để đưa ra câu trả lời. Một là, nguồn thông tin mở (open domain): Hệ thống có thể trả lời gần như mọi câu hỏi về mọi chủ đề. Hai là, nguồn thông tin giới hạn (closed domain): Hệ thống sẽ tìm ra câu trả lời dựa trên thông tin từ một lĩnh vực cụ thể.

Hệ thống đọc hiểu tự động là một lĩnh vực đang nhận được nhiều sự quan tâm của cộng đồng AI trên thế giới, do đó có không ít các công trình nghiên cứu về đọc hiểu văn bản tự động đã và đang đạt được nhiều thành tựu đáng kể. Lĩnh vực này có tính ứng dụng vô cùng đối lớn đối với doanh nghiệp. Bởi vì các hệ thống đọc hiểu được giao nhiệm vụ phân tích dữ liệu và hỗ trợ cho doanh nghiệp đưa ra các quyết định một cách đúng đắng. Ngoài ra, việc phát triển lĩnh vực này được triển khai vào các ứng dụng trí tuệ nhân tạo khác nhau như công cụ tìm kiếm, công nghệ AI, Chatbot và Robot.

Từ đó, chúng tôi nắm bắt được tầm quan trọng của các hệ thống trả lời câu hỏi tự động đối với con người. Tuy nhiên, lĩnh vực này có khá mới mẻ ở Việt Nam vì vậy cộng đồng xử lý ngôn ngữ tự nhiên vẫn chưa đạt được nhiều công trình nghiên cứu. Việc thực hiện đề tài này mang lại rất nhiều thách thức. Chúng tôi quyết định thực hiện đề tài với mong muốn xây dựng được một hệ thống trả lời câu hỏi liên quan đến những vấn đề liên quan về Sức khỏe. Bộ dữ liệu của đề tài là closed domain về lĩnh vực Sức khỏe.

### B1.2. Các công trình liên quan

B1.2.1. Các công trình nghiên cứu trên thế giới:

Với lợi ích mà các hệ thống đọc hiểu tự động mang lại, nó đã trở thành một chủ đề nóng và có sức hấp dẫn lớn đối với các cộng đồng nghiên cứu hiện nay, do đó có không ít các công trình nghiên cứu về đọc hiểu văn bản tự động đã và đang đạt được nhiều thành tựu đáng kể, đáng chú ý đầu tiên phải kể đến là những công trình nghiên cứu trên bộ dữ liệu TriviaQA [2]. Một bộ dữ liệu tiếng Anh gồm 95.956 cặp câu hỏi và câu trả lời được thu thập từ 14 trang website đố vui. Số lượng tài liệu bằng chứng (evidence documents) là 662.659 được tổng hợp từ Wikipedia (chứa câu trả lời cho 79.7% trên tổng số câu hỏi) và tìm kiếm trên trình duyệt Bing (chứa câu trả lời cho 75.4% trên tổng số câu hỏi). Những nghiên cứu trên bộ dữ liệu SQuAD [3] của Đại học Standford – một bộ dữ liệu tiếng Anh gồm 107.785 cặp câu hỏi và trả lời được tổng hợp từ 536 bài báo trên Wikipedia. Ngoài ra, còn có nhiều công trình nghiên cứu thành công cho bài toán đọc hiểu văn bản tự động (thành phần thứ hai của hệ thống trả lời câu hỏi) trên nhiều bộ dữ liệu với ngôn ngữ khác. Chẳng hạn như bộ dữ liệu tiếng Trung CMRC [4], bộ dữ liệu tiếng Hàn KorQuAD [5], bộ dữ liệu tiếng Pháp FquAD [6] và bộ dữ liệu tiếng Nga SberQuAD [6].

Về chủ đề hỏi đáp sức khỏe, một số bộ dữ liệu tiếng Anh nổi bật như bộ dữ liệu CliCR [7] gồm 100.000 câu truy vấn dạng điền vào chỗ trống dựa trên báo cáo trường hợp lâm sàng, bộ dữ liệu MedQA [8] có kích cỡ 270.000 với câu trả lời đầu ra dạng trắc nghiệm, hay là bộ dữ liệu PubMedQA [9] có kích cỡ 273.000 với câu trả lời cho câu hỏi đầu vào là đúng hoặc sai. Những bộ dữ liệu này đều hướng tới giải quyết bài toán đọc hiểu văn bản tự động đơn giản.

B1.2.2. Các công trình nghiên cứu trong nước:

Hiện nay tại Việt Nam vẫn chưa có nhiều công trình nghiên cứu cho chủ đề trả lời câu hỏi tự động thuộc ngôn ngữ tiếng Việt. Ba bộ dữ liệu tiếng Việt có thể kể tới là bộ dữ liệu giải quyết bài toán đọc hiểu văn bản (MRC): ViMMRC [10], UIT-ViQuAD [11] và ViNewsQA [12]. Đáng chú ý nhất là bộ dữ liệu ViNewsQA bao gồm 22.057 cặp câu hỏi và trả lời. Những người tạo dữ liệu dựa trên nội dung tri thức của 4.416 bài viết Sức khỏe trên VNEXPRESS. Hiện tại, mô hình tốt nhất đã chạy trên bộ dữ liệu này là ALBERT [12] với độ đo EM là 65.26% và ở độ đo F1 là 84.89%. Trong đề tài của chúng tôi, chúng tôi cũng sẽ kế thừa bộ dữ liệu cũng như những mô hình đã chạy trên bộ dữ liệu này.

### B1.3. Vấn đề thách thức

Bước chiếm hầu hết thời gian trong một bài toán AI chắc chắn là bước xây dựng dữ liệu. Việc xây dựng một bộ dữ liệu chất lượng để phục vụ cho mô hình đạt kết quả cao thì điều không thể thiếu đó là một bộ dữ liệu đủ lớn về số lượng cũng như về chất lượng. Bên cạnh đó, nguồn nhân lực để gán nhãn dữ liệu còn hạn chế, đòi hỏi những người có sự hiểu biết nhất định về bài toán để có thể xác định đúng đắn đầu ra.

Bài toán xây dựng hệ thống đọc hiểu tự đồng còn mới mẻ, manh nha đối với tiếng Việt. Vì vậy mà các công trình nghiên cứu, tài liệu tham khảo cũng ít hơn so với tiếng Anh hay các ngôn ngữ khác.

Tiếng Việt bản thân nó cũng mang những đặc điểm, độ khó nhất định (ví dụ như từ ghép, đa nghĩa) không giống với các ngôn ngữ khác. Việc xử lý ngôn ngữ tiếng Việt cũng là một thách thức lớn đòi hỏi chúng tôi không ngừng tìm tòi, nghiên cứu.

Hơn nữa, vấn đề về sức khỏe không chỉ bằng một vài câu chữ là có thể diễn tả, giải đáp được cho người thắc mắc, vì vậy những bài viết chứa câu trả lời dài xuất hiện nhiều và việc rút trích ra câu trả lời chính xác, đúng trọng tâm trong đoạn văn câu trả lời đòi hỏi người gán nhãn cần có khả năng đọc hiểu, khả năng truy xuất và rút trích thông tin đúng đắn, uy tín mới có thể mang lại độ chính xác cao cho quá trình đào tạo mô hình sau này.

Trên đây là một vài những thách thức, khó khăn mà chúng tôi nhận thấy ở thời điểm hiện tại mà chúng tôi cần đối mặt. Trong quá trình thực hiện đề tài, chúng tôi sẽ cố gắng giải quyết tốt các vấn đề đã được nêu như trên.

Một thách thức lớn nhất đối với chúng tôi là mất nhiều thời gian để hiện thực bộ dữ liệu bởi vì muốn hệ thống đọc hiểu đạt độ chính xác cao thì yêu cầu bộ dữ liệu khá lớn trong khi đó nguồn nhân lực còn hạn chế.

Các mô hình của đọc hiểu văn bản tự động khá phức tạp, khó hiểu và khó triển khai. Hơn thế đọc hiểu văn bản tự động là một lĩnh vực khá mới mẻ ở Việt Nam, tài liệu tham khảo cũng như tài nguyên các bộ dữ liệu về lĩnh vực này chủ yếu trên tiếng Anh trong khi đó trên tiếng Việt khá khiêm tốn gây khó khăn trong việc tìm hiểu và nghiên cứu.

Khó khăn của đề tài còn xuất phát từ đặc điểm của ngôn ngữ tiếng Việt, không như tiếng Anh mỗi từ mang nghĩa riêng biệt còn ở tiếng Việt xuất hiện các cụm từ ghép cũng như đa nghĩa ở một từ gây khó khăn trong việc xử lý.

Một thách thức nữa chúng tôi muốn nhắc tới là quá trình tạo câu hỏi và chọn câu trả lời, đòi hỏi người tạo phải có khả năng đọc hiểu văn bản, sử dụng đúng cấu trúc ngữ pháp tiếng Việt… để làm cho bộ dữ liệu có độ chính xác cao tốt cho qua trình đào tạo sau này.

Những bài báo chúng tôi thu thập được từ trang báo điện tử tương đối dài, các câu hỏi có thể không bao quát hết nội dụng của bài báo, ảnh hưởng đến quá trình đào tạo.

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

### B2.1 Xác định bài toán

**Đối tượng và phạm vi nghiên cứu:** tập trung nghiên cứu bộ dữ liệu tiếng Việt phục vụ cho bài toán trả lời câu hỏi tự động tiếng Việt với miền dữ liệu thuộc lĩnh vực Sức khỏe và nghiên cứu thực nghiệm các phương pháp giải quyết bài toán.

**Đầu vào và đầu ra một bài toán:**

* **Đầu vào:** Một câu hỏi tiếng Việt về những vấn đề thuộc lĩnh vực sức khỏe?
* **Đầu ra:** Một đoạn văn bản bằng chứng (evidence document) và câu trả lời dạng span-based được tìm thấy trong văn bản này.

Một ví dụ cụ thể được thể hiện ở bảng 1.

*Bảng 1: Ví dụ cụ thể đầu ra và đầu vào của bài toán.*

|  |  |
| --- | --- |
| **Đầu vào** | **Đầu ra** |
| **Bài báo tiếng Việt được lấy từ trang báo điện tử VinMec[[1]](#footnote-2)**   * **Câu hỏi:** Chào bác sĩ! Em đau bụng (thi thoảng giống kiểu đau bụng đi ngoài) nhưng không phải dạng đi ngoài (1 ngày nhiều nhất chỉ 2 lần hiếm đến 3 lần). Khi đi ngoài thì không ra phân như bình thường, mà ra chất lỏng (ngồi 2-3 phút thì lại không đau bụng nữa và không ra phân)? Thường ngày em ăn khá ít. Bác sĩ cho em hỏi đau bụng, đi ngoài nhiều ra chất lỏng là dấu hiệu của bệnh gì ạ? Rất mong bác sĩ tư vấn, cảm ơn bác sĩ! | * **Câu trả lời**   Chào bạn, bác sĩ xin được giải đáp câu hỏi: “Đau bụng, đi ngoài nhiều ra chất lỏng là dấu hiệu của bệnh gì?” như sau:  Theo mô tả, bạn có thể có vấn đề [bệnh lý ở đường tiêu hóa](https://www.vinmec.com/vi/tin-tuc/thong-tin-suc-khoe/nhung-benh-nao-thuong-gap-o-duong-tieu-hoa/), nhất là các bệnh lý ở đại tràng như [viêm loét đại tràng](https://www.vinmec.com/vi/benh/viem-dai-trang-3061/), u đại tràng... . Bạn nên đăng ký khám tại phòng khám tiêu hóa, hoặc ngoại tiêu hóa, bệnh viện thuộc [Hệ thống Y tế Vinmec](https://www.vinmec.com/vi/danh-sach/ca-nuoc/coso-benh-vien-v-phong-kham) để được thăm khám và tư vấn, bác sĩ có thể đề nghị một số xét nghiệm như [nội soi tiêu hóa](https://www.vinmec.com/vi/tin-tuc/thong-tin-suc-khoe/nhung-kieu-noi-soi-tieu-hoa-khong-dau/), siêu âm bụng...thì mới có chẩn đoán cụ thể được. Trân trọng! |

**Đề tài chúng tôi bao gồm hai bài toán chính:**

* Bài toán thứ nhất là bài toán truy xuất văn bản liên quan nhất đến câu trả lời (Document Retrieval). Đầu vào là câu hỏi và đầu ra là đoạn văn bản có nội dung liên quan đến câu hỏi. Bài toán được đánh giá bằng độ đo Accuracy.
* Bài toán thứ hai là bài toán đọc hiểu văn bản (MRC/Document Reader). Đầu vào là một đoạn văn bản và câu hỏi về nội dung liên quan có trong đoạn văn bản này. Đầu ra là một câu trả lời dạng span-based cho câu hỏi. Bài toán được đánh giá bằng độ đo EM và độ đo F1.

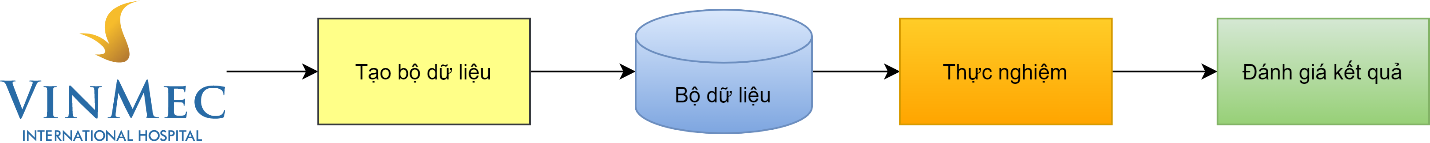
### B2.2 Mục tiêu

Bên cạnh việc sử dụng bộ dữ liệu ViNewsQA – bộ dữ liệu tiếng Việt bao gồm 22.057 cặp câu hỏi và trả lời. Nguồn tri thức của bộ dữ liệu này được xây dựng từ 4.416 bài viết với chủ đề sức khỏe của báo điện tử VNEXPRESS, là một trang tin tức được đánh giá là uy tín và cũng là trang tin tức trực tuyến lớn với số lượng đọc giả cao nhất tính đến thời điểm hiện tại. Chúng tôi xây dựng thêm bộ dữ liệu dự kiến khoảng 5.000 cặp câu hỏi và đoạn văn bản bằng chứng (evidence document) trả lời cho câu hỏi tương ứng. Câu hỏi được hỏi bởi những những người thắc mắc về những vấn đề liên quan sức khỏe được đăng lên những diễn đàn uy tín có những bác sĩ có chuyên môn cao giải đáp.

Nghiên cứu các mô hình, kỹ thuật có liên quan cho hai bài toán của đề tài. Chúng tôi cũng sẽ kế thừa những mô hình đọc hiểu văn bản (MRC) [12] đã tiến hành trên bộ dữ liệu ViNewsQA, trong đó mô hình ALBERT thực nghiệm trên bộ dữ liệu ViNewsQA cho kết quả với độ đo EM là 65.26% và ở độ đo F1 là 84.89% để trích xuất đầu ra câu trả lời dạng span-based (mục tiêu cuối cùng của bài toán). Chúng tôi kỳ vọng sẽ xây dựng được một mô hình tốt nhất đạt trên 50% theo độ đo EM và trên 70% theo độ F1-score để phục vụ cho ứng dụng thực tế như xây dựng ứng dụng hỗ trợ đọc báo, tìm kiếm thông tin, giải đáp thắc mắc về các vấn đề trong y tế nhằm chăm sóc tốt cho sức khỏe người dùng.

Chúng tôi mong muốn có thể đóng góp được bộ dữ liệu và đề xuất những phương pháp tốt giải quyết bài toán trả lời câu hỏi tự động cho lĩnh vực xử lý ngôn ngữ tự nhiên nói riêng và cộng đồng AI nói chung.

### B2.2 Nội dung và phương pháp nghiên cứu



*Hình 1: Quy trình nghiên cứu hỏi đáp văn bản tự động.*

Hình 1 trình bày quy trình nghiên cứu hỏi đáp văn bản tự động của chúng tôi, có hai quy trình chính là việc tạo bộ dữ liệu, nghiên cứu mô hình học sâu và thực nghiệm, đánh giá cho bộ dữ liệu của chúng tôi.

Nội dung 1: Xây dựng bộ dữ liệu

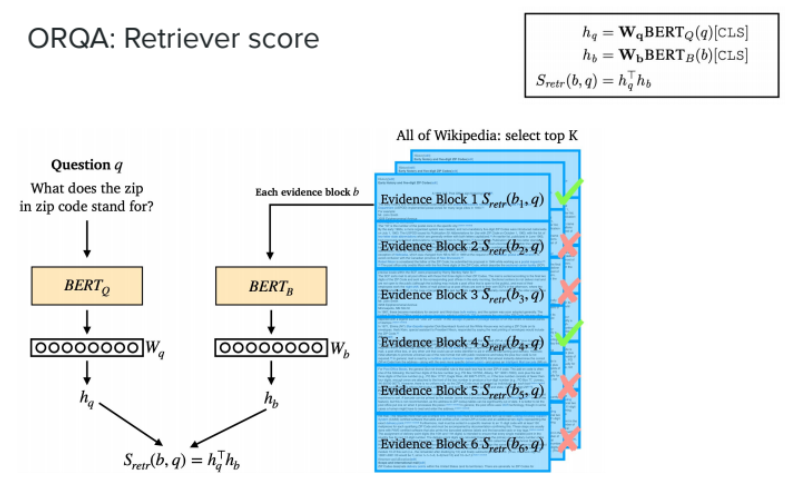
Chúng tôi tiến hành xây dựng thêm một bộ dữ liệu mới dự kiến tối thiểu có khoảng 5.000 cặp câu hỏi và văn bản trả lời cho câu hỏi tương ứng. Câu hỏi được hỏi bởi những những người thắc mắc về những vấn đề liên quan sức khỏe được đăng lên những diễn đàn uy tín trên trang báo điện tử có những bác sĩ có chuyên môn cao giải đáp (như diễn đàn sức khỏe trên VinMec và VNEXPRESS). Bộ dữ liệu mới được xây dựng với mục đích chính là phục vụ giải quyết cho bài toán đầu tiên là truy xuất đoạn văn bản có chứa câu trả lời cho câu hỏi đó. Lý do chúng tôi xây dựng thêm bộ dữ liệu mới vì nhận thấy bộ dữ liệu ViNewsQA được xây dựng bằng phương pháp dựa theo quy trình quy trình tạo ra bộ dữ liệu SQuAD [3]. Thành viên tham gia tạo bộ dữ liệu sẽ đặt câu hỏi dựa trên mỗi bài viết. Vì thế, phần lớn câu hỏi so với đoạn nội dung kèm theo sẽ có mức độ tương đồng giống nhau cao về mặt từ vựng (word matching). Phương pháp xây dựng bộ dữ liệu của chúng tôi sẽ lấy câu hỏi của người dùng thực sự. Và bác sĩ có chuyên môn sẽ trả lời với nội dung tương ứng câu hỏi. Chính vì thế, mức độ tương đồng giống nhau giữa câu hỏi và đoạn văn trả lời sẽ cao thiên về mặt ngữ nghĩa [13]. Hệ thống trả lời câu hỏi sẽ mang tính ứng dụng thực tế hơn nếu câu hỏi và đoạn văn bản đáp án có độ tương đồng nhau về mặt ngữ nghĩa hơn là về mặt từ vựng.

Nội dung 2: Nghiên cứu phương pháp thực nghiệm

* **Bài toán thứ nhất là bài toán truy xuất văn bản liên quan nhất đến câu trả lời (Document Retrieval)**

Hai phương pháp tiếp cận chính:

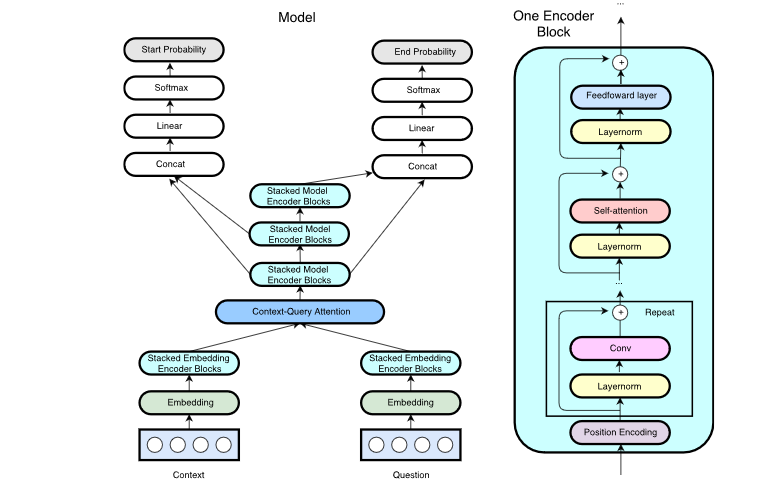
* Truy xuất văn bản dựa trên biểu diễn thưa thớt (sparse representation). Đây là phương pháp truy xuất không huấn luyện. Tiêu biểu là phương pháp BM25 [14], bản chất là công thức TF – IDF.
* Truy xuất văn bản dựa trên biểu diễn dày đặc (dense representation). Cùng với sự phát triển mô hình BERT, việc phát triển encode biểu diễn ngữ nghĩa của dữ liệu văn bản trở nên dễ dàng. Vì thế, chúng tôi muốn tiếp cận phương pháp này. Một trong những mô hình mà chúng tôi tham khảo là mô hình ORQA [13]. Mô hình này sẽ có hai module: Một module encode câu hỏi và một module encode đoạn văn bản chứng minh. Sau đó, sẽ nhân dot product vô hướng giữa vector đặc trưng của câu hỏi và vector đặc trưng của đoạn văn bản. Cuối cùng, sẽ xếp hạng (ranking) top K, đoạn văn bản có nội dung liên quan câu trả lời.



*Hình 2: Kiến trúc mô hình ORQA – Retrieval.*

* **Bài toán thứ hai là bài toán đọc hiểu văn bản (MRC/Document Reader)**
* **QANet:**

Mô hình QANet được đề xuất bởi Yu *et al.* [15]*,* theo tác giả thì mô hình này ra đời với mục đích giúp cho việc đào tạo và inference của mô hình nhanh hơn so với các mô hình trước đây bởi vì họ loại bỏ đi RNNs thay vào đó họ chỉ dùng convolutional và self-attention mechanism cho cả hai embedding và modeling encoders từ đó giúp cho mô hình của họ xử lý song song các tokens đầu vào. Cấu trúc mô hình QANet bao gồm có 5 thành phần: Input Embedding Layer, Embedding Encoder Layer, Context-Query Attention Layer, Model Encoder Layer và Output layer. Chi tiết cấu trúc của mô hình được thể hiện trong hình số 1. Chúng tôi cài đặt thử nghiệm mô hình QANet với một loạt các bộ word-embedding được đào tạo sẵn với số chiều là 300 để có được kích thước nhúng của mỗi từ như W2V\_ner, FastText\_ner, ELMO\_ner, Bert\_Base\_ner, MULTI\_W\_F\_B\_E [16] được đào tạo từ các bài báo trên Wikipedia, baomoi.model.bin [17] cùng với bộ character embedding W2V\_W2C\_ner [16].



*Hình 3: Kiến trúc mô hình QANet.*

* **BERT:**

Được Google công bố vào năm 2018, sự xuất hiện của BERT [18] mang đến một sự đột phá mới trong lĩnh vực xử lý ngôn ngữ tự nhiên. BERT cung cấp một khả năng mạnh mẽ trong việc học ra các vector đại diện theo ngữ cảnh hai chiều của từ cực kỳ chuẩn xác so với các kỹ thuật biểu diễn từ trước đó như Word2vec, Glove. Thay vì phải sử dụng một lượng dữ liệu cực kỳ lớn để có thể khái quát được các quan hệ ngữ nghĩa giữa các từ thì BERT sử dụng một kiến trúc hai chiều để tận dụng tối đa các quan hệ cần thiết với ít dữ liệu hơn. Nó đã cho thấy khả năng của mình khi lần lượt vươn lên đứng đầu và là sự lựa chọn hàng đầu cho các bài toán về Question Answering và Language Inference.

Về kiến trúc, BERT có kiến trúc đa tầng gồm nhiều lớp Transformer [19] hai chiều (**Bidirectional Transformer encoder**). Mỗi kiến trúc BERT có 2 nhiệm vụ con, gồm:

* **Masked LM:** Có nhiệm vụ che đi một số token đầu vào một cách ngẫu nhiên và sau đó dựa vào dữ liệu còn lại để dự đoán các token được giấu đi đó.
* **Next Sentence Prediction:** Có nhiệm vụ dự đoán mối quan hệ giữa 2 câu khác nhau. Nhằm phục vụ cho các bài toán về Question Answering như bài toán của chúng tôi. Với những câu trả lời hoặc câu hỏi có phạm vi ý nghĩa từ 2 câu trở lên thì việc xem xét các câu để hiểu đúng nghĩa là một nhiệm vụ vô cùng quan trọng, nhờ đó có thể hiểu chính xác và chọn được câu trả lời phù hợp.
* **ALBERT:**

Albert [20] được Google đề xuất nhằm hướng tới việc cải tiến độ chính xác và hiệu xuất của mô hình. Về kiến trúc, Albert được thiết kế tương tự như BERT, đều sử dụng kiến trúc encoder của transformer (Vaswani et al., 2017) [19] với hàm kích hoạt Gelu làm kiến trúc chính, sử dụng bidirectional self-attention để hiểu bối cảnh của một từ. Nhưng đối với Albert, có những điểm cải tiến hơn so với BERT, đó là:

* **Embedding có trọng số:** Thay vì ánh xạ các vecto embedding trực tiếp đến các lớp ẩn (tức là E = H) như Bert thì ở Albert chúng được ánh xạ đến một ma trận ít chiều hơn, sau đó mới ánh xạ ma trận này tới các lớp ẩn. Như vậy, số tham số sẽ được giảm đi đáng kể (từ **O(V × H)** đến **O(V × E + E × H)**).
* **Chia sẻ tham số giữa các lớp:** Cụ thể, các tham số được dùng chung là các tham số feed-forward network (FFN) và tham số attention, giúp thông tin lưu chuyển giữa các lớp thông suốt hơn và giúp mạng trở nên ổn định hơn.
* **Thay đổi cách huấn luyện:** Sử dụng SOP (Sentence Order Prediction) thay vì Next Sentence Prediction như Bert để giúp mô hình có năng lực dự đoán thứ tự giữa các câu nhờ vào sự mạch lạc giữa chúng.

### B2.3 Kế hoạch nghiên cứu

Chúng tôi tiến hành thực hiện nghiên cứu đề tài hỏi đáp văn bản tự động trong vòng 6 tháng và kế hoạch thực hiện chi tiết thể hiện trong Bảng 1.

**Bảng 1 - Kế hoạch thực hiện đề tài trong 6 tháng**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Công việc** | **T1** | **T2** | **T3** | **T4** | **T5** | **T6** |
| Tìm hiểu tổng quan |  |  |  |  |  |  |
| Tạo bộ dữ liệu |  |  |  |  |  |  |
| Nghiên cứu phương pháp |  |  |  |  |  |  |
| Báo cáo |  |  |  |  |  |  |

## B3. Kết quả nghiên cứu

Chúng tôi đặt ra những kết quả nghiên cứu của đề tài sau đây:

* Xây dựng bộ dữ liệu cho bài toán hỏi đáp văn bản tự động về sức khỏe cho tiếng Việt để phục vụ cho nghiên cứu và phát triển ứng dụng thực tế.
* Nghiên cứu và ứng dụng các phương pháp QANet, BERT, ALBERT cho bài toán đọc hiểu văn bản tự động trên chính bộ dữ liệu của chúng tôi đã tạo ra.

# TÀI LIỆU THAM KHẢO

[1] D. Chen, A. Fisch, J. Weston, and A. Bordes, ‘Reading wikipedia to answer open-domain questions’, *arXiv preprint arXiv:1704.00051*, 2017.

[2] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, ‘Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension’, *arXiv preprint arXiv:1705.03551*, 2017.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, ‘Squad: 100,000+ questions for machine comprehension of text’, *arXiv preprint arXiv:1606.05250*, 2016.

[4] Y. Cui *et al.*, ‘A span-extraction dataset for chinese machine reading comprehension’, *arXiv preprint arXiv:1810.07366*, 2018.

[5] S. Lim, M. Kim, and J. Lee, ‘Korquad1. 0: Korean qa dataset for machine reading comprehension’, *arXiv preprint arXiv:1909.07005*, 2019.

[6] P. Efimov, A. Chertok, L. Boytsov, and P. Braslavski, ‘SberQuAD–Russian reading comprehension dataset: Description and analysis’, in *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2020, pp. 3–15.

[7] S. Šuster and W. Daelemans, ‘CliCR: a dataset of clinical case reports for machine reading comprehension’, *arXiv preprint arXiv:1803.09720*, 2018.

[8] X. Zhang, J. Wu, Z. He, X. Liu, and Y. Su, ‘Medical exam question answering with large-scale reading comprehension’, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32, no. 1.

[9] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, ‘PubMedQA: A dataset for biomedical research question answering’, *arXiv preprint arXiv:1909.06146*, 2019.

[10] K. Van Nguyen, K. V. Tran, S. T. Luu, A. G.-T. Nguyen, and N. L.-T. Nguyen, ‘Enhancing lexical-based approach with external knowledge for Vietnamese multiple-choice machine reading comprehension’, *IEEE Access*, vol. 8, pp. 201404–201417, 2020.

[11] K. Nguyen, V. Nguyen, A. Nguyen, and N. Nguyen, ‘A Vietnamese Dataset for Evaluating Machine Reading Comprehension’, in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 2595–2605.

[12] K. Van Nguyen, D.-V. Nguyen, A. G.-T. Nguyen, and N. L.-T. Nguyen, ‘New vietnamese corpus for machine readingcomprehension of health news articles’, *arXiv preprint arXiv:2006.11138*, 2020.

[13] K. Lee, M.-W. Chang, and K. Toutanova, ‘Latent retrieval for weakly supervised open domain question answering’, *arXiv preprint arXiv:1906.00300*, 2019.

[14] S. Robertson and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[15] A. W. Yu *et al.*, ‘Qanet: Combining local convolution with global self-attention for reading comprehension’, *arXiv preprint arXiv:1804.09541*, 2018.

[16] X.-S. Vu, T. Vu, S. N. Tran, and L. Jiang, ‘ETNLP: A toolkit for extraction, evaluation and visualization of pre-trained word embeddings’, *arXiv preprint arXiv:1903.04433*, 2019.

[17] ‘word2vecVN’. https://github.com/sonvx/word2vecVN.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘Bert: Pre-training of deep bidirectional transformers for language understanding’, *arXiv preprint arXiv:1810.04805*, 2018.

[19] A. Vaswani *et al.*, ‘Attention is all you need’, in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[20] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, ‘Albert: A lite bert for self-supervised learning of language representations’, *arXiv preprint arXiv:1909.11942*, 2019.

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

1. https://www.vinmec.com/vi/tieu-hoa-gan-mat/tu-van-bac-si/dau-bung-di-ngoai-nhieu-ra-chat-long-la-dau-hieu-cua-benh-gi/ [↑](#footnote-ref-2)