Kỷ yếu Hội nghị Khoa học công nghệ Quốc gia lần thứ XVI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Đà Nẵng, ngày 28-29/9/2023
DOI: 10.15625/vap.2023.0083

NGHIÊN CỨU MỘT SỐ MÔ HÌNH HỎI ĐÁP TỰ ĐỘNG TRÊN MIỀN DỮ LIỆU DU LỊCH

Trần Thanh Phước[1], Nguyễn Duy Khanh[2], Trần Thanh Trâm[3]

[1]Khoa Công nghệ Thông tin, Trường Đại học Tôn Đức Thắng
[2]Phòng Glarus, Công ty Axon Active Việt Nam
[3]Phòng Khoa học Công nghệ, Trường Đại học Công Thương Thành phố Hồ Chí Minh

tranthanhphuoc@tdtu.edu.vn, khanh.nguyenduy@axonactive.com, tramtt@hufi.edu.vn

TÓM TẮT: *Hỏi đáp tự động (QA: Question Answering) là một trong những bài toán thú vị và được nhiều nhà nghiên cứu Khoa học máy tính triển khai nghiên cứu, thực nghiệm trong thời gian gần đây. Có nhiều hướng tiếp cận phổ biến cho bài toán QA như: hướng tiếp cận dựa trên luật, hướng tiếp cận Information Retrieval, hướng tiếp cận Generative,…. Mỗi hướng tiếp cận đều có ưu và khuyết điểm khác nhau, tuỳ theo bài toán mà chúng ta chọn lựa hướng tiếp cận phù hợp. Trong bài báo này, chúng tôi sử dụng hướng tiếp cận Information Retrieval cho bài toán QA trên miền dữ liệu du lịch tiếng Việt. Chúng tôi sử dụng 03 biến thể của mô hình BERT để thực nghiệm, bao gồm: Mô hình RoBERTA-Base, RoBERTa-Large và mô hình PhoBERT. Dữ liệu được dùng để thử nghiệm cho 03 mô hình này được thu thập bán tự động, nội dung dữ liệu đề cập đến các khía cạnh du lịch ở Việt Nam. Kết quả thử nghiệm cho thấy mô hình PhoBERT và mô hình RoBERTa-large cho kết quả tốt gần như nhau.*

Từ khóa: *Vietnamese tourism QA dataset, BERT, ROBERTA-Base, ROBERTA-Lagre, PhoBERT, QA system.*

I. GIỚI THIỆU

Hỏi đáp tự động (QA: Question Answering) hiện đang được cộng đồng khoa học máy tính nói chung và xử lý ngôn ngữ nói riêng tập trung nghiên cứu. Các ứng dụng QA ngày càng phổ biến trong thời gian gần đây, đặc biệt là sự ra đời của ứng dụng ChatGPT (https://chat.openai.com/). Về khía cạnh học thuật, để tạo ra một ứng dụng QA chúng ta cần sử dụng một mô hình theo hướng tiếp cận nào đó. Hiện có một số hướng tiếp cận phổ biến như: Hướng tiếp cận dựa trên luật (Rule-based), hướng tiếp cận dựa trên truy hồi thông tin (Information Retrieval-based) và hướng tiếp cận dựa trên tạo sinh (Generative-based).

Hướng tiếp cận dựa trên luật:

Hướng tiếp cận dựa trên luật sử dụng quy tắc heuristic để tìm kiếm từ vựng và ngữ nghĩa trong câu hỏi và trong tài liệu [1]. Đây là hướng tiếp cận QA được xem là ra đời sớm nhất trong các hướng tiếp cận. Nó được phát triển dựa trên các quy tắc hoặc tập các lệnh được xác định trước. Các hệ thống QA dựa trên hướng tiếp cận này tuân theo một bộ quy tắc được định nghĩa trước để tạo ra câu trả lời cho câu hỏi đầu vào của người dùng. Chúng sử dụng một loạt câu lệnh có điều kiện để kiểm tra từ khóa hoặc cụm từ trong câu hỏi đầu vào của người dùng và đưa ra câu trả lời tương ứng dựa trên các điều kiện này.

Hướng tiếp cận Rule-based có thể hiệu quả trong một số trường hợp nhất định nhưng lại tồn tại một số hạn chế như:

- Khả năng hiểu ngôn ngữ tự nhiên hạn chế: Do các hệ thống QA dạng này được xây dựng dựa trên các quy tắc định nghĩa trước nên chúng sẽ gặp khó khăn khi gặp phải các câu hỏi phức tạp, không giống với các mẫu được định nghĩa trước.
- Thiếu ngữ cảnh: Các hệ thống QA dạng này không thể hiểu ngữ cảnh của cuộc trò chuyện. Chúng không thể hiểu được ý định của người dùng ngoài bộ quy tắc mà chúng được xác định trước đó.
- Không có khả năng học hỏi và thích ứng: Các hệ thống QA dạng này không có khả năng học hỏi hỏi thêm để thích nghi. Chúng không có khả cải thiện phản hồi của mình theo thời gian ngay cả khi các phản hồi này đã lỗi thời, không còn hiệu quả.

Hướng tiếp cận dựa trên truy hồi thông tin:

Như phần trên đã đề cập, hướng tiếp cận dựa trên luật đòi hỏi phải có bộ quy tắc có cấu trúc được định nghĩa trước. Để làm được việc này, chúng ta cần có nguồn lực để tạo các quy tắc có cấu trúc này. Hướng tiếp cận dựa trên truy hồi thông tin không yêu cầu một quy tắc được định nghĩa trước, các hệ thống dựa trên hướng tiếp cận này chỉ cần các tài liệu không cấu trúc và rút trích thông tin từ chúng một cách tự động. Ngày nay, để có được các thông tin không cấu trúc như các tài liệu văn bản, các trang HTML là đơn giản, không quá phức tạp. Từ các tài liệu không cấu trúc này, tập hợp các ngữ cảnh, các câu hỏi - câu trả lời được tạo ra một cách bán tự động.

Hệ thống QA dựa trên truy hồi thông tin được xây dựng dựa trên hai quá trình chính gồm truy xuất tài liệu và rút trích câu trả lời [2]. Trong quá trình truy xuất tài liệu, hệ thống phải chọn lọc ra các tài liệu từ tập dữ liệu lớn liên quan đến câu hỏi. Các tài liệu được chọn lọc sẽ được quá trình rút trích câu trả lời xử lý và trả về câu trả lời phù hợp nhất. Có nhiều phương pháp trong học máy đã được áp dụng hiệu quả cho hai quy trình trên.

Do kho dữ liệu hội thoại được định nghĩa trước nên các phương pháp truy hồi thông tin không mắc lỗi ngữ pháp. Tuy nhiên, chúng không có khả năng xử lý các trường hợp chưa được huấn luyện trước.

Hướng tiếp cận dựa trên tạo sinh:

Khác với hai hướng tiếp cận trước, hướng tiếp cận dựa trên tạo sinh [3] không dựa vào các câu trả lời được xác định trước mà chúng sẽ tạo ra các câu trả lời hoàn toàn mới. Chúng sử mô hình ngôn ngữ lớn (large language model) được huấn luyện trên kho ngữ liệu vô cùng lớn với hàng tỷ từ, cụm từ và câu. Hướng tiếp cận này kết hợp các thuật toán học sâu và và các kỹ thuật xử lý ngôn ngữ tự nhiên để hiểu sự phức tạp của ngôn ngữ và tạo ra phản hồi giống với con người hơn. Ngoài ra, hướng tiếp cận này còn có khả năng học hỏi và cải thiện hiệu suất hỏi đáp theo thời gian.

Các mô hình tạo sinh thông minh hơn so với hai mô hình trước, chúng có thể tham khảo lại các thực thể trước đó. Tuy nhiên, các mô hình theo hướng tiếp cận này khá phức tạp, chúng thường bị sai ngữ pháp khi gặp các câu hỏi dài. Ngoài ra, chúng cần một kho dữ liệu huấn luyện rất lớn.

Mỗi hướng tiếp cận đều có ưu và khuyết điểm khác nhau, tuỳ điều kiện cụ thể, các nhà nghiên cứu sẽ chọn cho mình mô hình phù hợp nhất. Trong bài báo này, chúng tôi chọn hướng tiếp cận Information Retrieval, cụ thể là các biến thể của mô hình BERT (Bidirectional Encoder Representation from Transformer) để thực nghiệm trên miền dữ liệu du lịch Việt Nam. Chúng tôi đã khởi tạo bán tự động kho dữ liệu này và sử dụng ba biến thể của mô hình BERT để thực nghiệm, đó là ROBERTA-Base, ROBERTA-Lagre, PhoBERT để thực nghiệm. Chúng tôi sẽ sử dụng mô hình cho kết quả thực nghiệm tốt nhất để phục vụ cho ứng dụng hỏi đáp tự động về du lịch trong tương lai gần.

Cấu trúc còn lại của bài báo được trình bày theo thứ tự như sau: Ở Mục II, chúng tôi sẽ trình bày các kiến thức nền tảng của bài báo như: Mô hình học chuyển đổi, các mô hình BERT: ROBERTA-Base, ROBERTA-Large, PhoBERT. Nội dung chi tiết về bài toán hỏi đáp tự động trên miền dữ liệu du lịch sẽ được trình bày ở Mục III, bao gồm các nội dung như: Lưu đồ tổng quát và quá trình khởi tạo dữ liệu. Mục IV sẽ đề cập đến các kết quả thực nghiệm cũng như một số thảo luận liên quan. Phần kết luận sẽ được trình bày ở Mục V.

II. KIẾN THỨC NỀN TẢNG

A. Học chuyển đổi

Học chuyển đổi (Transfer Learning) [4] là mô hình học cho phép các mô hình có thể truyền đạt năng lực riêng của chúng cho nhau. Trong đó, một mô hình được học trên một nhiệm vụ cụ thể được gọi là mô hình tiền huấn luyện (Pretrained model). Mô hình này được sử dụng để chuyển đổi (transfer) tri thức được học cho mô hình khác (transefered model). Transfered model được học trên nhiệm vụ đích (target task). Thông thường, Pretraned model được học trên kho dữ liệu lớn, tổng quát; Transfered model được học trên dữ liệu nhỏ hơn thuộc về một lĩnh vực nào đó. Cụ thể trong bài báo nào, Pretrained model được học trên nguồn dữ liệu tiếng Việt tổng quát và Transfered model được học trên nguồn dữ liệu nhỏ du lịch tiếng Việt.

Transfer learning giúp tiết kiệm được rất nhiều thời gian và tài nguyên khi huấn luyện mô hình mới. Tuy nhiên, nếu xảy ra sai sót trong quá trình transfer thì độ chính xác sẽ rất thấp. Ngoài ra, các nhiệm vụ giữa pretrained model và transfered model phải có liên quan trực tiếp với nhau, nếu nhiệm vụ không liên quan, hệ thống sẽ cho hiệu suất rất thấp. Ví dụ, chúng ta không thể dùng mô hình Pre-train QA cho tiếng Anh để transfer cho QA tiếng Việt.

B. BERT

BERT (Bidirectional Encoder Representations from Transformers) [5] là một kỹ thuật biểu diễn từ được Devlin và đồng sự giới thiệu năm 2018. Nhóm tác giả đã sử dụng kiến trúc Transformer [6] để huấn luyện mô hình ngôn ngữ lớn. Hiện tại, BERT được sử dụng như là một mô hình Pre-train cho rất nhiều tác vụ, đặc biệt là các tác vụ trong lĩnh vực xử lý ngôn ngữ tự nhiên, ví dụ như: phân loại văn bản, phân tích cảm xúc người dùng, và đặc biệt là được sử dụng trong tác vụ hỏi đáp tự động. Điểm đặc biệt của mô hình BERT là nó có khả năng tạo ra các vector biểu diễn văn bản trong ngữ cảnh hai chiều trái và phải. Các vector này được tinh chỉnh cho các tác vụ đích khác như: hỏi đáp tự động, phân tích cảm xúc,…

Trong bài báo này, chúng tôi cũng sử dụng các biến thể của BERT như là mô hình pre-trained để tinh chỉnh cho tác vụ hỏi đáp tự động trên dữ liệu du lịch Việt Nam.

C. XLM-RoBERTa

RoBERTa [7] được giới thiệu bởi Facebook kế thừa các kiến trúc và thuật toán của BERT. RoBERTa được thực hiện trên Pytorch, với khả năng thay đổi một số siêu tham số chính trong BERT. RoBERTa lặp lại các thủ tục huấn luyện từ BERT và có một số thay đổi: huấn luyện mô hình lâu hơn, huấn luyện theo nhóm nhỏ và tốc độ học, dữ liệu huấn luyện lớn hơn, không sử dụng cơ chế dự đoán câu tiếp theo mà chỉ sử dụng kỹ thuật mặt nạ động (nó sẽ ẩn đi một số từ ở câu đầu ra bằng token <mask>), tiếp theo các token mặt nạ sẽ thay đổi trong quá trình huấn luyện, RoBERTa sử dụng kích thước bacth lớn hơn.

Một điểm khác biệt chính giữa RoBERTa và BERT là RoBERTa được đào tạo trên tập dữ liệu lớn hơn nhiều lần và sử dụng quy trình đào tạo hiệu quả hơn. RoBERTa hoạt động tốt hơn BERT trong nhiều tác vụ xử lý ngôn ngữ tự nhiên như dịch máy, phân loại văn bản và trả lời tự động.

RoBERTa có 2 biến thể là:

- RoBERTa-Base: bao gồm 250 triệu tham số.
- RoBERTa-Large: bao gồm 560 triệu tham số.

Cả hai biến thế này đều được huấn luyện trên 2.5 terabyte dữ liệu của hơn 100 ngôn ngữ khác nhau, đặc biệt có tiếng Việt trong số các ngôn ngữ này. Mô hình RoBERTa cho kết quả tốt trên nhiều kho dữ liệu chuẩn như MLQA [8].

D. PhoBERT

Năm 2020, Dat Q.N và đồng sự đã công bố mô hình Pre-trained PhoBERT [9]. Có thể xem đây là biến thể của RoBERTa cho phiên bản tiếng Việt. PhoBERT cũng có hai biến thể là PhoBERT-base và PhoBERT-Large; trong đó, PhoBERT-Base có 12 block Transformers, PhoBERT-Large có 24 block Transformers.

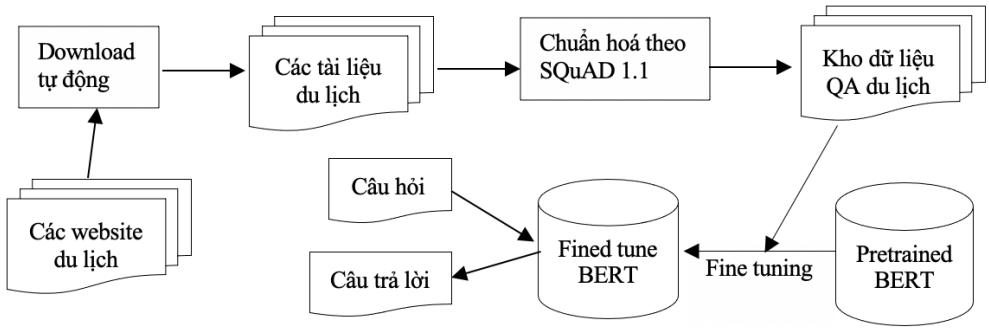
PhoBERT cho hiệu suất vượt trội so với các mô hình theo hướng tiếp cận đơn ngữ và đa ngữ trước đó, đặc biệt ở bốn tác vụ xử lý ngôn ngữ tự nhiên như: gán nhãn từ loại, phân tích cú pháp phụ thuộc, nhận dạng thực thể có tên, và suy luận ngôn ngữ tự nhiên.

III. HỎI ĐÁP TỰ ĐỘNG TRÊN MIỀN DỮ LIỆU DU LỊCH

A. Lưu đồ

Hình 1 thể hiện lưu đồ thực hiện của bài toán hỏi đáp trên miền dữ liệu du lịch, bao gồm các bước chính như:

- Khởi tạo dữ liệu:
  - Thu thập các văn bản du lịch Việt Nam.
  - Tiền xử lý dữ liệu.
  - Khởi tạo dữ liệu QA theo chuẩn SQuAD 1.1.
- Sử dụng các mô hình BERT để fine-tune dữ liệu.
- Kiểm thử dữ liệu.



Hình 1. Nguồn lấy dữ liệu từ các web trực tuyến về du lịch

B. Khởi tạo dữ liệu

1. Nguồn dữ liệu

Nguồn dữ liệu để chúng tôi khai thác là các mục du lịch ở các trang web tiếng Việt nổi tiếng như: Vnexpress (https://vnexpress.net/du-lich/cam-nang), traveloka (https://www.traveloka.com/) và ivivu (https://www.ivivu.com/hoi-dap). Bộ dữ liệu thu được gồm các đoạn văn bản về lĩnh vực Du lịch - Ẩm thực của các khu vực như: Đà Lạt, Đà Nẵng, Phú Quốc, Vũng Tàu,… Hình 2 minh hoạ nguồn dữ liệu được lấy từ website du lịch Vivu.com.

2. Phương pháp thu thập dữ liệu

Chúng tôi sử dụng thư viện BeautifulSoup 4 để thu thập tự động nội dung văn bản của các trang web du lịch. Sau đó, các dữ liệu văn bản này sẽ được làm sạch dựa trên một số thuật toán cơ bản như: Xoá các nội dung rác: Các liên kết, các ký tự đặc biệt, hiệu chỉnh các từ sai chính tả. Dữ liệu văn bản này được định dạng theo chuẩn của kho dữ liệu nổi tiếng SQuAD 1.1 của Đại học Stanford [10]. Hình 3 thể hiện cấu trúc tập tin json của SQuAD 1.1.

Cấu trúc cặp câu hỏi – trả lời theo sQuAD Format:

Mỗi một cặp câu như vậy chứa hai thuộc tính, “Context” và “qas”.

- Context: Đoạn văn hoặc văn bản mà câu hỏi được đặt ra.
- qas: Một danh sách các câu hỏi và câu trả lời.

Các câu hỏi và câu trả lời được trình bày dưới dạng list json. Mỗi cặp câu trong qas có định dạng sau.

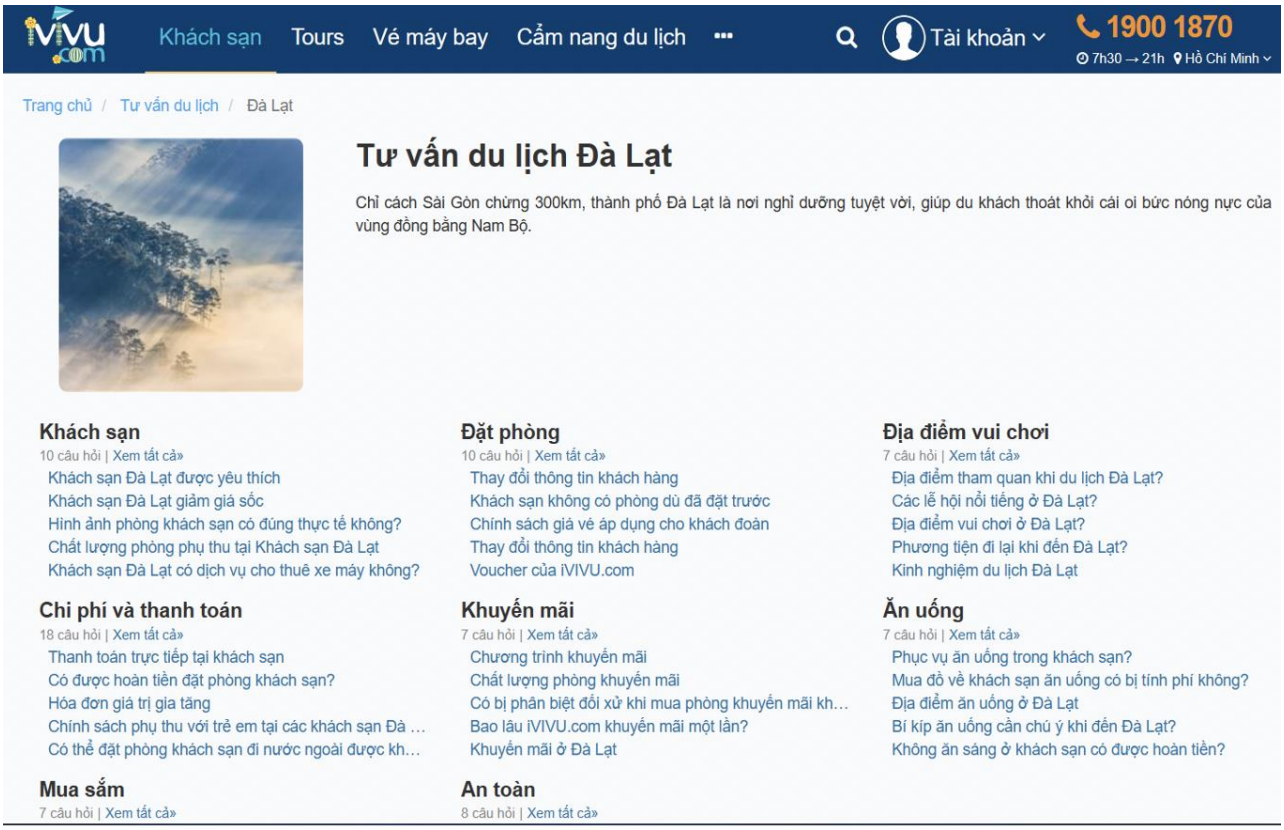
- id: ID duy nhất cho câu hỏi. Phải là duy nhất trên toàn bộ tập dữ liệu.
- question: Câu hỏi.
- answers: Danh sách các câu trả lời đúng cho câu hỏi.

Một câu trả lời được có các các thuộc tính sau:

- Text: Câu trả lời cho câu hỏi. Phải là một chuỗi con của context.
- answer_start: vị trí bắt đầu của câu trả lời trong context.

Để tạo được cặp câu hỏi - trả lời (Q-A) cho kho dữ liệu thử nghiệm, chúng tôi sử dụng 2 phương pháp song song, đó là phương pháp phát sinh câu hỏi tự động từ câu trả lời có sẵn (1) và phương pháp tạo câu hỏi thủ công (2). Phương pháp (1) được chúng tôi sử dụng dựa theo công trình [11]. Chúng tôi sử dụng công cụ cdQA-annotator để khởi tạo thủ công cặp Q-A. Source code của công cụ này được đặt tại "git clone https://github.com/cdqa-suite/cdQA-annotator".

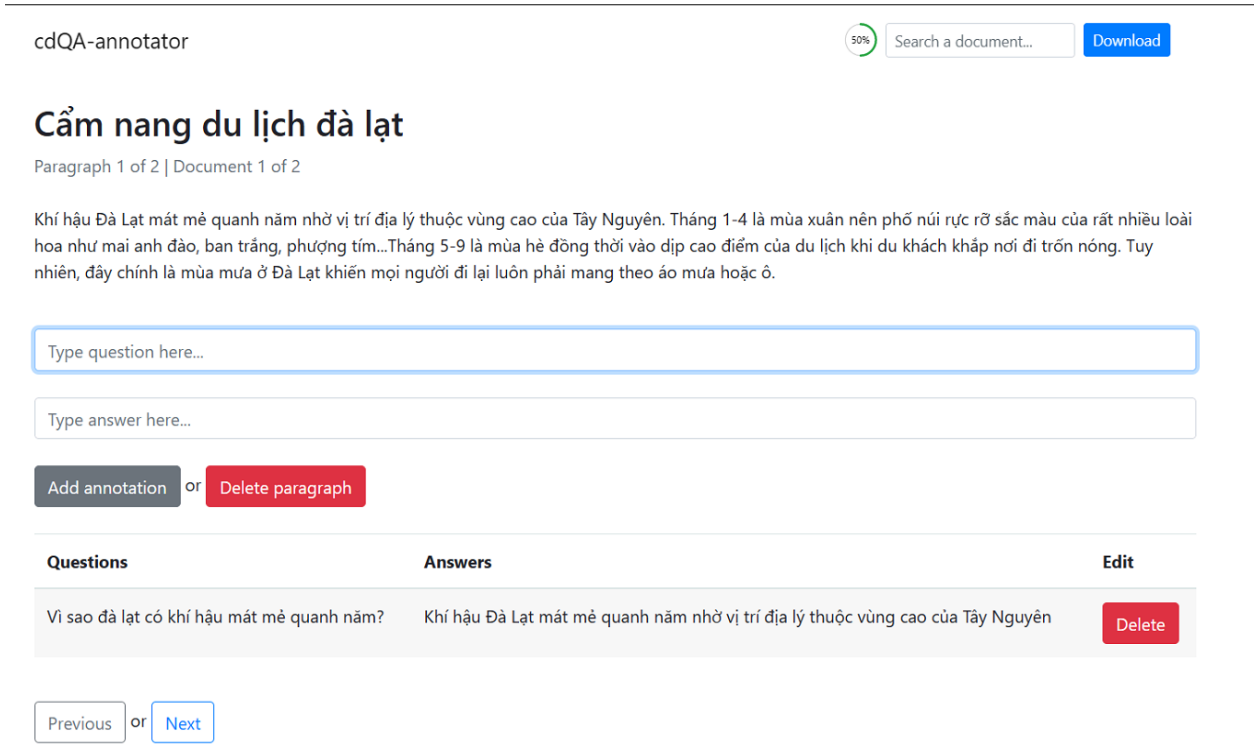
Hình 4 minh hoạ giao diện của cdQA-annotator. Chúng tôi thu thập được tổng cộng gần 5.000 cặp Q-A về du lịch Việt Nam, kho dữ liệu này được đặt tên là NLP-ViQA-Tourism v1.0. **Hình 5** minh hoạ kho dữ liệu này theo định dạng của SQuAD 1.1.



Hình 2. Nguồn lấy dữ liệu từ các web trực tuyến về du lịch


```
"context": "Dư dả thời gian hơn, du khách có thể đi tàu hỏa để
"qas": [
  {
    "question": "Giá vé tàu hỏa đến Đà nẵng",
    "id": "401783b5-d402-4599-bc9e-a4e0b2ed7b72",
    "answers": [
      {
        "answer_start": 145,
        "text": "Vé tàu từ Hà Nội hoặc TP HCM đến Đà Nẵng có g
      }
    ]
  },
```


Hình 3. Cấu trúc tập tin json



Hình 4. Giao diện web công cụ cdQA-annotator


```
{
  "title": "Đà nẵng",
  "paragraphs": [
    {
      "context": "Thành phố Đà Nẵng nằm ở miền Trung, chia đều khoảng cách giữa thủ đô Hà Nội và TP HCM. Đà Nẵng còn là trung tâm của 3 di sản văn hóa thế giới là Cố đô Huế, phố cổ Hội An và thánh địa Mỹ Sơn. Phía bắc Đà Nẵng giáp tỉnh Thừa Thiên – Huế, phía tây và nam giáp tỉnh Quảng Nam, phía đông giáp biển Đông.",
      "qas": [
        {
          "question": "Thành phố Đà Nẵng nằm ở đâu",
          "id": "7b070bd3-215f-47d1-9216-1dc24cc04536",
          "answers": [
            {
              "answer_start": 22,
              "text": "ở miền Trung"
            }
          ]
        },
        {
          "question": "Đà nẵng giáp với các tỉnh nào",
          "id": "76111917-60ce-47bc-bece-4c723f5e1136",
          "answers": [
            {
              "answer_start": 191,
              "text": "Phía bắc Đà Nẵng giáp tỉnh Thừa Thiên – Huế, phía tây và nam giáp tỉnh Quảng Nam, phía đông giáp biển Đông."
            }
          ]
        },
```


Hình 5. Định dạng dữ liệu sau khi xuất từ công cụ cdQA-annotator

IV. THỰC NGHIỆM

A. Dữ liệu thực nghiệm

03 mô hình tiền huấn luyện (Pre-training) được sử dụng trong thực nghiệm là: ROBERTA-Base, ROBERTA-Larger và PhoBERT. Kho dữ liệu được được chuyển đổi (transfering) là kho dữ liệu NLP-ViQA-Tourist v1.0 được thu thập. Kho dữ liệu này bao gồm 4.822 cặp câu Q-A và được chia thành 2 tập dữ liệu như sau: Kho dữ liệu huấn luyện gồm 4.388 cặp Q-A (training), kho dữ liệu kiểm tra (testing) gồm 434 cặp Q-A.

B. Công cụ đánh giá thực nghiệm

Chúng tôi sử dụng một độ đo EM (Exac match) và F1 để đo chất lượng của hệ thống:

- EM: Độ đo này tương đối đơn giản. Đối với một cặp câu QA, nếu các từ trong câu dự đoán khớp chính xác với từ trong câu trả lời tham chiếu thì EM = 1, ngược lại EM = 0.

- F1: Độ đo này khá phổ biến cho các bài toán phân lớp và được sử dụng rộng rãi trong QA, được thể hiện theo công thức (1). Trong đó:
    - P là độ chính xác (Precision)
    - R là độ phủ (Recall)
    - TP (True Positives): số lượng mà bộ phân loại dự đoán đúng lớp
    - FP (False Positives): số lượng mà bộ phân loại dự đoán sai so với lớp đúng thực tế
    - FN (False Negatives): số lượng mà bộ phân loại dự đoán sai qua lớp khác

$$F_1 = \frac{2PR}{P+R} = \frac{2TP}{2TP + FP + FN} \quad (1)$$

C. Kết quả thực nghiệm

Bảng 1 trình bày kết quả thực nghiệm tương ứng với 3 mô hình:

Bảng 1. Kết quả chạy thực nghiệm kho dữ liệu Du lịch General

Models	**Exact match**	**F1**
Phobert-Large	40.7834%	**76.6854%**
XLM-Roberta-Base	36.4055%	73.5173%
XLM-Roberta-Large	**41.7050%**	76.3390%

D. Thảo luận

Từ Bảng 1, chúng ta nhận thấy rằng mô hình PreTrain PhoBERT-Large và XLM-RoBERTa-Large cho kết quả với chất lượng gần tương đồng nhau, trong khi đó mô hình ROBERTA-Base cho kết quả thấp nhất. Có thể giải thích rằng ROBERTA-Base cho kết quả thấp là vì dữ liệu train cũng như các tham số cho mô hình huấn luyện của nó thấp hơn so với ROBERTA-Large và PhoBERT-Large. Bảng 2 thể hiện 6 ví dụ từ dữ liệu kiểm tra của 3 mô hình.

Bảng 2. Một số ví dụ về kết quả thực nghiệm của 3 mô hình.

STT	**Câu hỏi**	**PhoBERT-Large**	**XLM-RoBERTa-Base**	**XLM-RoBERTa-Large**
1	Bến Nhà Rồng ở Sài Gòn còn có tên gọi khác là gì?	Bến Nhà Rồng hay còn có tên gọi khác là Bảo tàng Hồ Chí Minh	Bến Nhà Rồng hay còn có tên gọi khác là Bảo tàng Hồ Chí Minh	Bến Nhà Rồng hay còn có tên gọi khác là Bảo tàng Hồ Chí Minh
2	Bến Nhà Rồng ở Sài Gòn là nơi như thế nào?	đây chính là nơi Bác Hồ ra đi tìm đường cứu nước vào năm 1911	đây chính là nơi Bác Hồ ra đi tìm đường cứu nước vào năm 1911	đây chính là nơi Bác Hồ ra đi tìm đường cứu nước vào năm 1911
3	Giờ mở cửa Suối Tiên ở Sài Gòn ngày thường là gì?	**8h00 - 17h30 (Thứ 2 - Thứ 6), 8h00 - 18h00 (Thứ 7 và Chủ Nhật)**	Giờ mở cửa: 8h00 - 17h30 (Thứ 2 - Thứ 6)	8h00 - 17h30 (Thứ 2 - Thứ 6)
4	Tòa nhà cao ốc cao thứ 2 Việt Nam là gì?	sau Kangnam Hanoi Landmark Tower)	**Bitexco**	**Bitexco**
5	Giờ mở cửa Suối Tiên ở Sài Gòn ngày lễ là gì?	**6h30 - 22h00 (Lễ Tết)**	00 (Lễ Tết)	**6h30 - 22h00 (Lễ Tết)**
6	Diện tích công viên Sun World ở Hạ Long?	**với diện tích 214 ha**	tích 214 ha	**với diện tích 214 ha**

Context cho ví dụ 1 và 2:

Bến Nhà Rồng hay còn có tên gọi khác là Bảo tàng Hồ Chí Minh, đây chính là nơi Bác Hồ ra đi tìm đường cứu nước vào năm 1911. Điểm đặc biệt trong kiến trúc Bến Nhà Rồng nằm ở nóc nhà gắn hình rồng, ở giữa mang chiếc phù hiệu "Đầu ngựa và chiếc mỏ neo". Đặc biệt, dạo bước ra sân phía ngoài du khách còn có thể ngắm sông Sài Gòn với tầm nhìn thoáng đãng, vào ban đêm còn lung linh ánh đèn soi bóng xuống dòng sông.

Context cho ví dụ 3 và 5:

Giờ mở cửa: 8h00 - 17h30 (Thứ 2 - Thứ 6), 8h00 - 18h00 (Thứ 7 và Chủ Nhật), 6h30 - 22h00 (Lễ Tết). Vé vào cổng: 60.000 VND/người lớn, 30.000 VND/trẻ em; Vé các trò chơi: 5.000 - 40.000 VND/lượt.Địa chỉ: 120 Xa Lộ Hà Nội, Phường Tân Phú, Quận 9 (cách trung tâm thành phố 19 km).

- Context cho ví dụ 4:

Công viên Sun World là một trong những khu vui chơi lớn nhất nước với diện tích 214 ha. Ngoài tắm biển, du khách có thể vui chơi tại công viên nước vịnh Lốc Xoáy, công viên Rồng, công viên nước Typhoon Water Park, check-

in trên chiếc cầu Koi nổi tiếng, chiêm bái quần thể tâm linh Bảo Hải Linh Thông Tự hay tham quan khu vườn Nhật Bản, thăm bảo tàng tượng sáp, ngôi nhà lộn ngược. Công viên gồm hai tổ hợp ven biển và trên đỉnh Ba Đèo. Trong đó, các điểm nổi bật như cáp treo nữ hoàng, đồi huyền bí có giá vé 250.000 - 350.000 VND/người. Công viên Dragon Park bán vé 200.000 - 300.000 VND/người không giới hạn số lần chơi các trò giải trí. Công viên nước Typhoon Water Park bán vé 350.000 VND/người.

Dựa trên bộ văn bản kiểm tra, chúng ta đã tìm thấy nhiều trường hợp mà hệ thống XLM-Roberta-Large và Phobert-Large đã cho kết quả tốt hơn so với XLM-RoBERTa-Base. Bảng 2 thể hiện 6 ví dụ trong bộ văn bản kiểm tra của ba mô hình: XLM-RoBERTa-Large, XLM-RoBERT-Base, PhoBert-Large. Trong đó hai ví dụ đầu tiên đưa ra kết quả gần như giống nhau cho cả ba hệ thống. Ví dụ thứ 3 cho ta thấy PhoBERTa-Large tốt hơn 2 mô hình còn lại. Ví dụ 4 cho ta thấy PhoBert-Large sai và XLM-RoBERTa đúng. Hai ví dụ cuối cùng thể hiện rằng PhoBert-Large và XLM-RoBERTa-Large đúng còn XLM-RoBERTa-Base sai. Rõ ràng chúng ta thấy PhoBERT-Large và XLM-RoBERT-Large sẽ tốt hơn XLM-RoBERTa-Base. Tất nhiên sẽ có sai số khi tập train khá là ít nhưng không đáng kể.

V. KẾT LUẬN

Trong bài báo này, chúng tôi đã xây dựng được kho dữ liệu cho bài toán hỏi đáp tự động trên lĩnh vực du lịch. Đây là một trong những lĩnh vực năng động nhất của xã hội phát triển. Kho dữ liệu này cũng được chúng tôi sử dụng để phục vụ cho các mô hình học chuyển đổi; trong đó, mô hình tiền huấn luyện gồm: ROBERTA-Base, ROBERTA-Large, và PhoBERT. Ba mô hình tiền huấn luyện này sẽ chuyển đổi cho mô hình học trên miền dữ liệu đóng, nhỏ hơn: Dữ liệu du lịch Việt Nam. Kết quả thực nghiệm cho thấy mô hình tiền huấn luyện PhoBERT và RoBERTa-Larger cho kết quả tốt gần như nhau. Chúng tôi dự định sẽ chọn một trong hai mô hình này để huấn luyện dữ liệu cho ứng dụng trả lời tự động về lĩnh vực du lịch.

Dữ liệu du lịch hiện tại tương đối nhỏ và mang tính chất chung nhất cho dữ liệu Việt Nam. Sắp tới, chúng tôi sẽ tiếp tục khởi tạo cho nguồn dữ liệu du lịch này thêm phong phú, bao phủ. Ngoài ra, chúng tôi cũng dự định sẽ khởi tạo thêm dữ liệu về du lịch mang tính đặc thù cho từng địa phương. Có được nguồn dữ liệu đặc thù này, hệ thống trả lời tự động về du lịch sẽ chính xác hơn.

LỜI CẢM ƠN

Công trình này được thực hiện bằng nguồn kinh phí hỗ trợ từ Quỹ Phát triển Khoa học Công nghệ Trường Đại học Tôn Đức Thắng, theo Hợp đồng số "FOSTECT.2022.12".

TÀI LIỆU THAM KHẢO

[1] Ellen Riloff and Michael Thelen, "A Rule-based Question Answering System for Reading Comprehension Tests," in *Proceeding of ACL*, 2000, pp. 13-19.

[2] Chao-Chun Hsu, Eric Lind, Luca Soldaini, Alessandro Moschitti, "Answer Generation for Retrieval-based Question Answering Systems," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4276-4282 August 1-6, 2021. ©2021 Association for Computational Linguistics.

[3] Sumit Pandey, Srishti Sharma, "A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning," https://www.sciencedirect.com/science/article/pii/S2772442523000655?via%3Dihub, Elsevier, 2023.

[4] Sewon Min, Minjoon Seo, Hannaneh Hajishirzi, "Question Answering through Transfer Learning from Large Fine-grained Supervision Data," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, (vol. 2: Short Papers), 2017, pp. 510-517

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*, pp. 4171-4186, 2019.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998-6008, 2017.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[8] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk, "MLQA: Evaluating cross-lingual extractive question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7315-7330, 2020.

[9] Dat Quoc Nguyen and Anh Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," *Findings of the Association for Computational Linguistics: EMNLP*, 2020, pp. 1037-1042.

[10] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, (https://arxiv.org/pdf/1606.05250.pdf).

[11] Phuoc Tran, Duy Khanh Nguyen, Tram Tran, and Bay Vo, "Using Syntax and Shallow Semantic Analysis for Vietnamese Question Generation," *KSII transactions on internet and information systems*, (accepted) 2023.

A STUDY OF QA MODELS ON TOURISM DATASET

Tran Thanh Phuoc, Nguyen Duy Khanh, Tran Thanh Tram

ABSTRACT: *Automatic Question Answering (QA) is one of the interesting problems and has been studied by many Computer Science researchers in recent times. There are many popular approaches to QA problems including rule-based approach, Information Retrieval approach, Generative approach, etc. Each approach has different advantages and disadvantages, depending on the specific problem that we choose the appropriate approach. In this paper, we used Information Retrieval approach for QA problem on Vietnamese tourism dataset. We used three variants of BERT model for experiment, including RoBERTA-Base model, RoBERTa-Large model and PhoBERT model. The dataset used for experimenting these three models is semi-automatically collected. Experimental results show that the PhoBERT model and the RoBERTa-large model give almost equally good results.*