ĐẠI HỌC HUẾ
TRƯỜNG ĐẠI HỌC KHOA HỌC

BẢN THUYẾT MINH SẢN PHẨM DỰ THI

CUỘC THI LẬP TRÌNH DÀNH CHO HỌC SINH TRUNG HỌC PHỔ THÔNG VÀ SẢN PHẨM SÁNG TẠO CÔNG NGHỆ THÔNG TIN DÀNH CHO SINH VIÊN CAO ĐẲNG, ĐẠI HỌC NĂM 2024 (HUE-ICT CHALLENGE-2024)

Tên sản phẩm:

XÂY DỰNG ỨNG DỤNG HỎI ĐÁP CHO TIẾNG VIỆT DỰA TRÊN CÁC MÔ HÌNH TRANSFORMER

Lĩnh vực: **TRÍ TUỆ NHÂN TẠO VÀ PHẦN MỀM**

Tác giả:

1. Nguyễn Luôn Mong Đổ

Thừa Thiên Huế, ngày 07 tháng 06 năm 2024

I NỘI DUNG

1 Tên sản phẩm

Xây dựng ứng dụng hỏi đáp cho tiếng Việt dựa trên các mô hình Transformer

2 Ý tưởng của đề tài

Ngôn ngữ của con người là một hệ thống các tín hiệu/ký hiệu được xây dựng một cách đặc biệt để truyền đạt được thông tin có chủ đích của người viết/người nói. Các tín hiệu/ký hiệu này được con người sử dụng để giao tiếp với nhau. Xử lý ngôn ngữ tự nhiên là một lĩnh vực đặc biệt, đó là sự kết hợp giữa các ngành khoa học máy tính, trí tuệ nhân tạo và ngôn ngữ học. Mục tiêu của việc xử lý ngôn ngữ tự nhiên là làm cho máy tính hiểu và xử lý được ngôn ngữ tự nhiên của con người.

Nghiên cứu về hệ thống hỏi đáp tự động(Q&A) đã thu hút sự quan tâm lớn trên thế giới từ rất lâu. Vào đầu những năm 1960, các hệ thống hỏi đáp đầu tiên sử dụng cơ sở dữ liệu đã được tạo ra, các hệ thống này được xây dựng dựa trên hai mô hình chính bao gồm: mô hình dựa trên trích xuất thông tin và mô hình dựa trên kiến thức. Các hệ thống này được xây dựng nhằm mục đích trả lời các câu hỏi về số liệu thống kê các trận đấu bóng chày và trả lời các sự kiện khoa học [1].

Vào cuối những năm 1990, World Wide Web ra đời và phát triển nhanh chóng tạo ra kho dữ liệu khổng lồ. Cho đến này nay, khi internet đã phát triển vượt trội, lượng thông tin mà con người được cung cấp và tiêu thụ ngày càng lớn. Hệ thống trả lời câu hỏi tự động ra đời nhằm cung cấp cho con người giải

pháp giúp tiết kiệm thời gian đọc và làm giảm khối lượng kiến thức mà con người phải tiếp thu.

Với sự phát triển của các mô hình học sâu đã tạo ra bước độ phá của trí tuệ nhân tạo trong lĩnh vực xử lý ngôn ngữ tự nhiên. Cùng với sự thành công của Chat GPT trong những năm gần đây, việc hiểu ngôn ngữ tự nhiên của máy tính đã trở nên dễ dàng hơn so với trước. Thay vì phải đọc toàn bộ bài viết để hiểu được nội dung, với hệ thống trả lời câu hỏi, người dùng chỉ cần cung cấp nội dung của bài viết và đặt ra các câu hỏi thắc mắc liên quan đến bài viết. Hệ thống giúp người dùng tiết kiệm thời gian đọc và loại bỏ các thông tin gây nhiễu cho người đọc.

Tuy nhiên, việc xử lý ngôn ngữ tự nhiên đối với tiếng Việt vẫn còn nhiều hạn chế. Nguyên nhân đến từ việc tiếng Việt là một ngôn ngữ đơn lập với hệ thống từ ghép và từ láy đa dạng. Các yếu tố đó tạo ra sự "nhập nhằng"trong quá trình xử lý ngôn ngữ tự nhiên. Các mô hình học sâu và các mô hình Transformer được giới thiệu trước đó như Transformer, BERT, RoBERTa, OpenGPT, ELMo... đều được đào tạo dựa trên tập dữ liệu tiếng Anh dẫn đến không thể tối ưu trên tập dữ liệu tiếng Việt.

Với những lợi ích và thách thức mà hệ thống trả lời câu hỏi cho tiếng Việt mang lại, trong công trình nghiên cứu này, tôi xin thực hiện đề tài "Xây dựng ứng dụng hỏi đáp cho tiếng Việt dựa trên các mô hình Transformer".

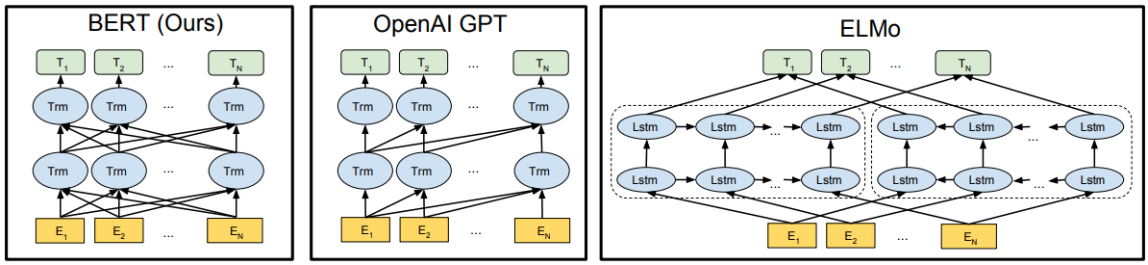
II MÔ TẢ VỀ SẢN PHẨM

1 Tính mới, tính sáng tạo của sản phẩm

Mô hình được đào tạo trong ứng dụng này sử dụng cấu trúc của mô hình BERT.

So với các mô hình mạng Nơ-ron truyền thống (RNNs, LSTM, GRUs), BERT biểu diễn các vector đại diện từ theo ngữ cảnh của câu, do đó cho kết quả tốt hơn trong quá trình huấn luyện.

Ngoài ra, so với các mô hình Transformer khác như ELMo hay OpenAI GPT (hình 1), BERT có những ưu điểm về việc biểu diễn vector theo ngữ cảnh hai chiều. Với OpenAI GPT, nhóm tác giả sử dụng kiến trúc left-to-right, nghĩa là các tokens chỉ phụ thuộc vào các token ở trước đó. Đối với ELMo, mô hình sử dụng hai mạng lưới LSTM tách rời để biểu diễn ngữ cảnh của câu. Trong khi BERT sử dụng một mạng lưới Transformer duy nhất để biểu diễn từ theo vector ngữ cảnh trong câu.



Hình 1: Sự khác nhau giữa BERT, OpenAI GPT, ELMo (Nguồn: [2] figure 3)

Ngoài ra, đối với các mô hình học sâu và các mô hình Transformer được giới thiệu trước đó, các mô hình đều được tiền đào tạo dựa trên tập dữ liệu tiếng Anh, khả năng xử lý tiếng Việt của các mô hình đó còn gặp nhiều hạn chế và hiệu suất chưa được tối ưu. Mô hình mà đề tài nghiên cứu này xây dựng được tiền đào tạo dựa trên tập dữ liệu tiếng Việt - được xây dựng với mục đích sử

dụng riêng cho tiếng Việt.

2 Các nguyên vật liệu làm ra mô hình, sản phẩm

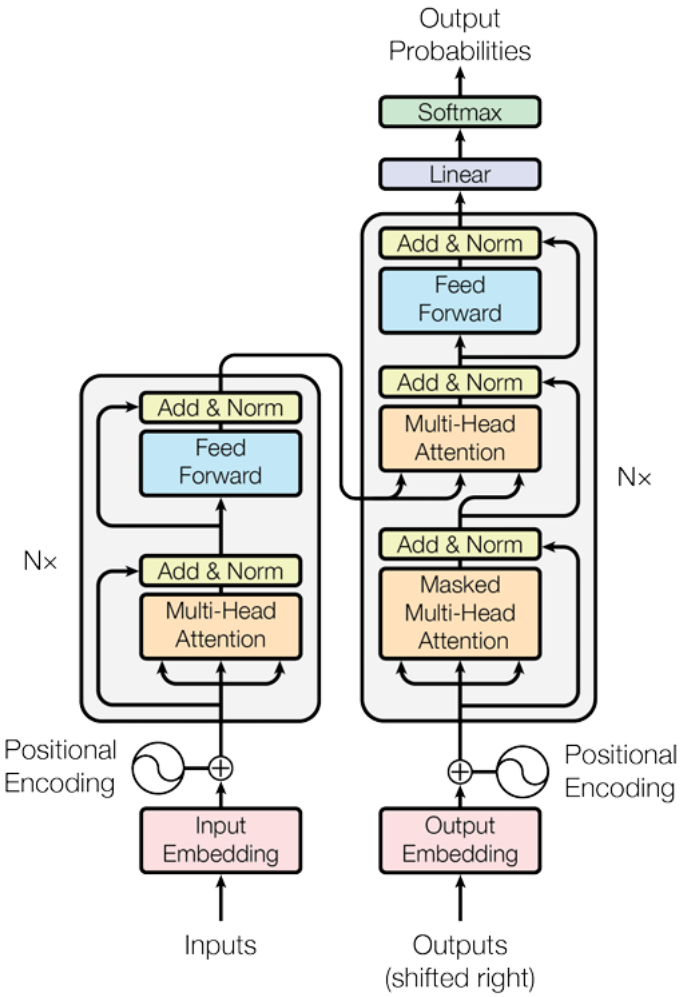
Ứng dụng hỏi đáp sử dụng cho tiếng Việt được xây dựng dựa trên cơ sở lý thuyết sau:

2.2 Mô hình Transformer

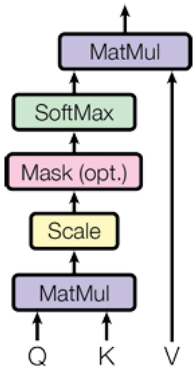
Transformer là một mô hình học sâu được giới thiệu vào năm 2017 trong paper "Attention is all you need" [3] bởi nhóm tác giả đến từ Google. Khác với các mô hình Neural hồi quy (RNN, LSTM...) được sử dụng trong bài toán seq2seq, mô hình Transformer hoàn toàn hoạt động dựa trên cơ chế Attention. Mô hình Transformer có 2 phần bao gồm Bộ mã hóa (Encoder) và Bộ giải mã (Decoder).

Cơ chế chú ý (Attention)

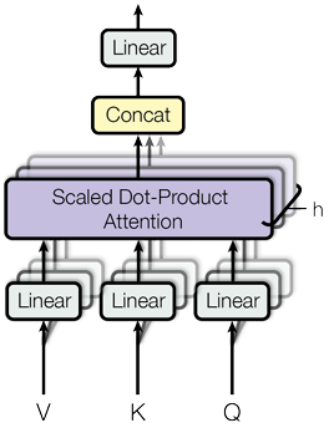
Cơ chế chú ý (Attention) được mô tả như việc ánh xạ một truy vấn (query) vầ một cặp khóa - giá trị (key - value) với đầu ra. Trong đó truy vấn (query), khóa (key), giá trị (value) và đầu ra đều là các vector. Đầu ra được tính toán bằng cách tính tổng có trọng số của các vector giá trị (value), với trọng số được gán cho mỗi vector giá trị (value) được tính bằng một hàm liên quan giữa vector truy vấn (query) và khóa (key) tương ứng.



Hình 2: Cấu trúc của mô hình Transformer (Nguồn: [3])



Hình 3: Scaled Dot-Product (Nguồn: [3])



Hình 4: Multi-Head Attention (Nguồn: [3])

2.3 Mô hình BERT

BERT, viết tắt của Bidirectional Encoder Representations from Transformers. BERT là một kiến trúc mô hình mới cho lớp bài toán Language Representa-

tion được công bố vào tháng 11 năm 2018 bởi nhóm tác giả đến từ Google, bao gồm Jacob Devlin, Ming-Wei Chang, Kenton Lee và Kristina Toutanova [2]. Khác với các mô hình trước đó, BERT được thiết kế để tiền huấn luyện các biểu diễn hai chiều từ dữ liệu văn bản chưa được gán nhãn dựa trên ngữ cảnh hai chiều của chúng. Từ đó, mô hình BERT sau khi được tiền huấn luyện có thể tinh chỉnh với một lớp đầu ra bổ sung phù hợp với từng nhiệm vụ của mô hình, có thể là như trả lời câu hỏi (question and answering) và suy luận ngôn ngữ (language inference), mà không cần thay đổi cấu trúc của cả mô hình cho từng nhiệm vụ cụ thể.

Mô hình BERT gồm cấu trúc đa tầng với nhiều lớp Bidirectional Transformer encoder. Các phiên bản BERT được đào tạo với những thiết lập khác nhau được mô tả trong bảng 1

Bảng 1: Thông tin về các phiên bản mô hình BERT được đào tạo

	H=128	**H=256**	**H=512**	**H=768**
L=2	2/128 BERT-tiny	2/256	2/512	2/768
L=4	4/128	4/256 BERT-mini	4/512 BERT-small	4/768
L=6	6/128	6/256	6/512	6/768
L=8	8/128	8/256	8/512 BERT-medium	8/768
L=10	10/128	10/256	10/512	10/768
L=12	12/128	12/256	12/512	12/768 BERT-base

Có hai phiên bản BERT được giới thiệu trong tài liệu năm 2018 [2] bao gồm:

$$BERT_{\text{BASE}}(L = 12,\ H = 768,\ A = 12,\ \text{Tổng tham số} = 110M)$$

$$BERT_{\text{LARGE}}(L = 24,\ H = 1024,\ A = 16,\ \text{Tổng tham số} = 340M)$$

3 Cách lắp ráp, cài đặt sản phẩm

3.1 Ràng buộc của ứng dụng

- Yêu cầu ngôn ngữ lập trình: Python, JavaScript.

- Các framework được sử dụng: Pytorch, ReactJS

3.2 Cài đặt mô hình thu được từ tinh chỉnh PhoBERT

Cài đặt mã nguồn:

git clone https://github.com/SpiderMan-XiaoDo/KhoaLuanTotNghiep.git

Tải mô hình đã được huấn luyện và lưu nó vào thư mục modelv2 tại đường dẫn:

https://www.kaggle.com/datasets/nguyendolikeyou/pretrain-phobert-base/data

Nguyên nhân: Mô hình được đào tạo có kích thước quá lớn (hơn 525MB) không thể lưu trữ trên môi trường github

Thực hiện triển khai API

uvicorn main:app –reload

3.3 Cài đặt giao diện cho ứng dụng

Cài đặt mã nguồn

git clone https://github.com/SpiderMan-XiaoDo/myapp.git

Cài đặt các gói thư viện

cd myapp

npm install

Triển khai ứng dụng

npm start

4 Nguyên tắc hoạt động, vận hành của sản phẩm dự thi

4.1 Mô tả đầu vào và đầu ra của mô hình

Hệ thống trả lời câu hỏi được tạo ra nhằm mục đích tóm tắt lại cho người dùng nội dung của đoạn văn bản, trả lời các câu hỏi mà người dùng đặt ra liên quan đến đoạn văn bản đã cung cấp.

Hệ thống trả lời câu hỏi yêu cầu người dùng cung cấp tri thức liên quan đến vấn đề thắc mắc (được gọi là Context) và câu hỏi cho hệ thống (được gọi là Question).

Thông qua tri thức được người dùng cung cấp (context), hệ thống sẽ thực hiện truy xuất dữ liệu phù hợp với câu hỏi và đưa ra câu trả lời xuất hiện trong tri thức được cung cấp trước đó đến người dùng. Thông tin về đầu vào và đầu ra của hệ thống trả lời câu hỏi bài báo cáo này sẽ xây dựng được mô trả trong hình 5

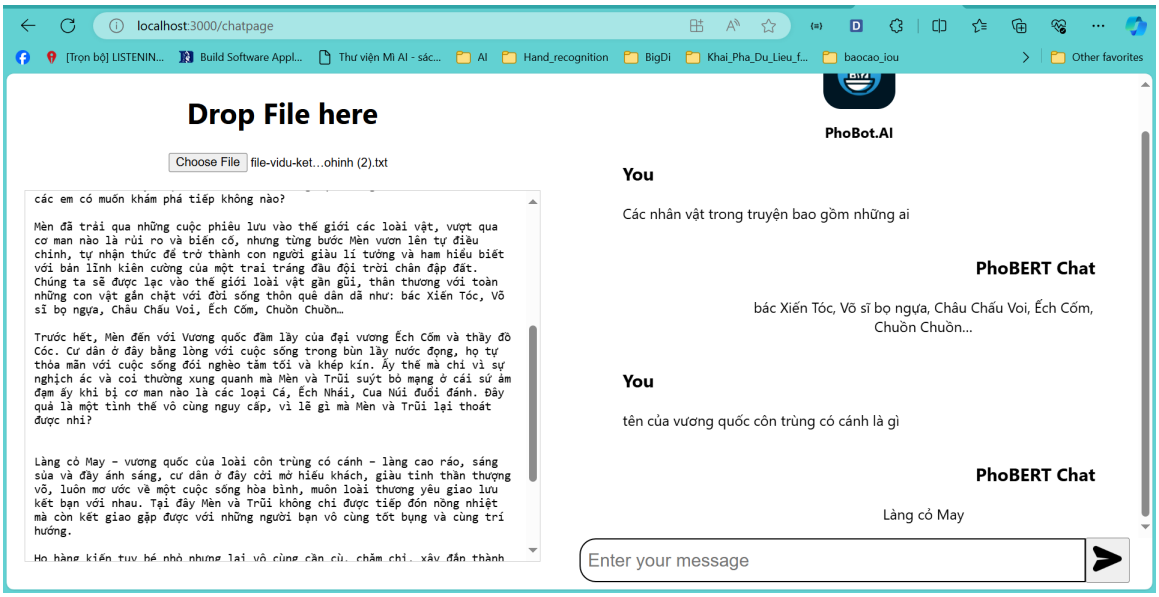
```
Question: Tên khoa học của động vật dưới nước là gì?

Context: Động vật lưỡng cư (danh pháp khoa học: Amphibia) là một lớp động
vật có xương sống máu lạnh. Tất cả các loài lưỡng cư hiện đại đều là phân
nhánh Lissamphibia của nhóm lớn Amphibia này. Động vật lưỡng cư phải trải
qua quá trình biến thái từ ấu trùng sống dưới nước tới dạng trưởng thành
có phổi thở không khí, mặc dù vài loài đã phát triển qua nhiều giai đoạn
khác nhau để bảo vệ hoặc bỏ qua giai đoạn ấu trùng ở trong nước dễ gặp
nguy hiểm. Da được dùng như cơ quan hô hấp phụ, một số loài kỳ giông và
ếch thiếu phổi phụ thuộc hoàn toàn vào da. Động vật lưỡng cư có hình dáng
giống bò sát, nhưng bò sát, cùng với chim và động vật có vú, là các loài
động vật có màng ối và không cần có nước để sinh sản. Trong những thập kỷ
gần đây, đã có sự suy giảm số lượng của nhiều loài lưỡng cư trên toàn cầu.

Answer: Amphibia
```

Hình 5: Hình ảnh mô tả đầu vào và đầu ra của bài toán. Đầu vào của bài toán bao gồm câu hỏi (question) và nội dung có liên quan (context). Đầu ra của bài toán là câu trả lời cho câu hỏi đó.

4.2 Triển khai ứng dụng

Để sử dụng ứng dụng hỏi đáp trên, hãy thực hiện các bước sau:

- **Đầu tiên**, tiến hành cung cấp tệp văn bản chứa thông tin liên quan cần hỏi đáp (dưới dạng tệp .txt) cho hệ thống tại giao diện **Drop File here**.
- **Tiếp theo**, sau khi đã cung cấp tệp thông tin, tiến hành đặt câu hỏi cho hệ thống.
- **Sau đó**, mô hình sẽ tiến hành trả lời câu hỏi đưa ra dựa vào nội dung tệp văn bản vừa được cung cấp và hiển thị ra giao diện.

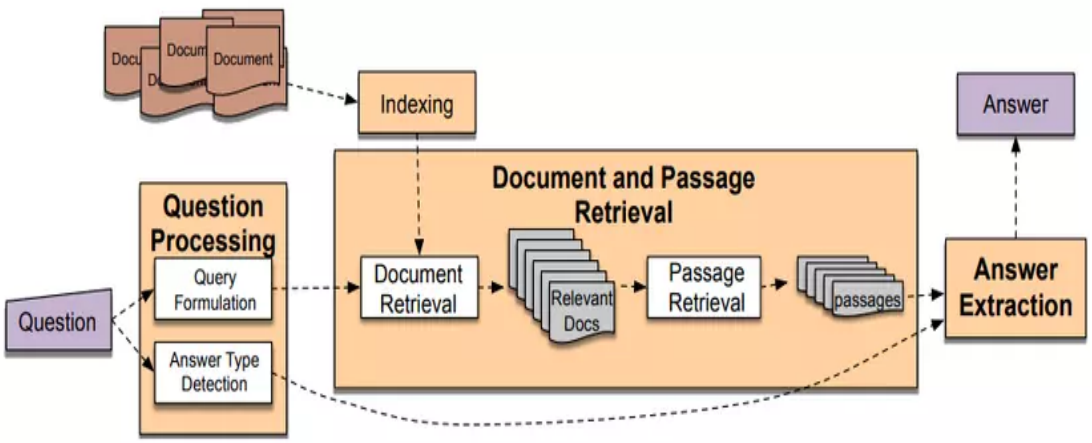


Hình 6: Hình ảnh mô tả kết quả xây dựng của hệ thống hỏi đáp sử dụng cho tiếng Việt.

5 Khả năng ứng dụng của sản phẩm

Hệ thống trả lời câu hỏi (Q&A) được ứng dụng trong nhiều lĩnh vực của đời sống, bao gồm:

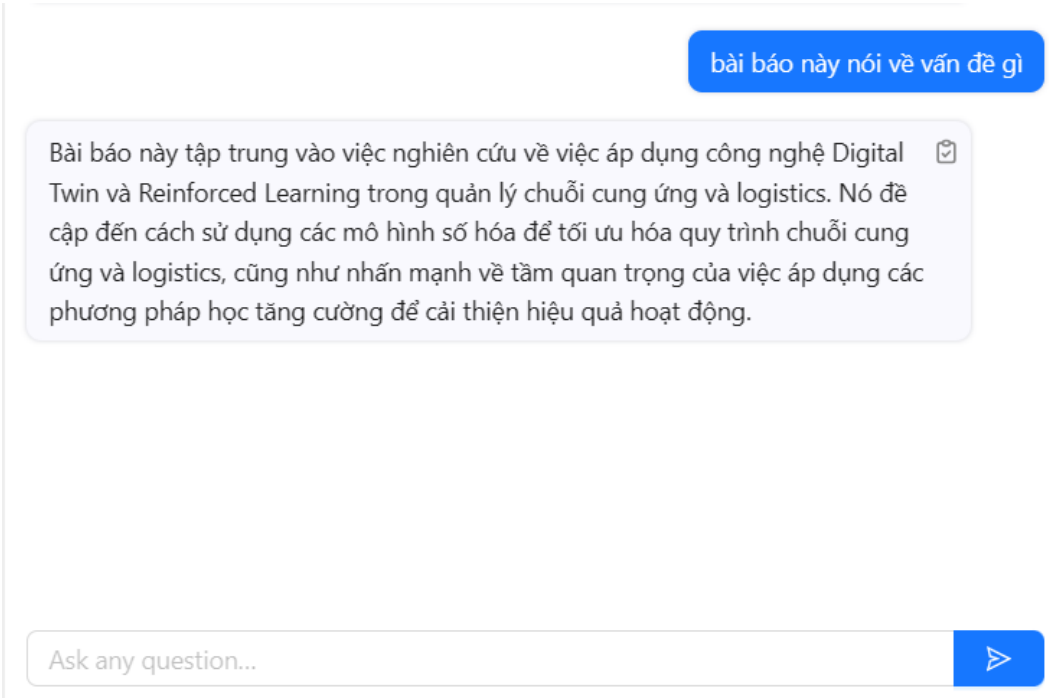
Tìm kiếm thông tin nhanh chóng bằng cách trả lời các câu hỏi dựa trên nội dung có trong các cuốn sách, các bài báo liên quan.



Hình 7: Mô hình mô tả ứng dụng của hệ thống trả lời câu hỏi trong việc tìm kiếm thông tin

Hỗ trợ việc học tập và nghiên cứu: người dùng có thể sử dụng hệ thống trả lời câu hỏi để tìm hiểu về một chủ đề cụ thể, đọc hiểu các bài báo, tóm tắt các ý chính... Chat PDF là một ứng dụng được xây dựng nhằm đáp ứng nhu cầu này, được mô tả tại hình8 và hình 9

Hình 8: Chat PDF yêu cầu người dung cung cấp tập tin chứa thông tin



Hình 9: Sau khi đã cung cấp thông tin, người dùng có thể đặt bất kỳ câu hỏi nào liên quan đến tập tin đó

Ngoài ra, hệ thống trả lời câu hỏi còn được ứng dụng trong các lĩnh vực:

- Y tế: ứng dụng giúp các bệnh nhân tìm hiểu về các triệu chứng bệnh tật và các phương pháp điều trị
- Quản lý giáo dục:ứng dụng giúp giải đáp thắc mắc của sinh viên liên quan đến các điều lệ, quy chế, quy định của nhà trường...
- Phân tích dữ liệu: ứng dụng giúp các doanh nghiệp trích xuất các thông tin có trong hợp đồng, các số liệu báo cáo...

6 Hiệu quả đạt được của sản phẩm

	Ensemble	F1 (%)	EM (%)
Retrospective Reader + XLM-R	x	81.013	71.316
BLANC + XLM-R/SemBERT	x	82.622	73.698
$XLM\text{-}R_{Large}$	x	80.578	70.662
$XLM\text{-}R_{Large}$		79.594	69.092
$PhoBERT_{Large}$+R3F+CS		75.842	63.544
mBERT – baseline		63.031	53.546

Hình 10: Một số kết quả từ cuộc thi VLSP 2021 - Vietnamese Machine Reading Comprehension (Nguồn: [4])

Các tham số thiết lập thí nghiệm với mô hình phobert-base được mô tả trong bảng 2

Bảng 2: Thông tin tham số được thiết lập cho việc fine-tuning lại mô hình PhoBERT

Max length	**Strike**	**Learning rate**	**Batch size**	**Epoch**	**Train Size**	**Valid Size**
256	128	2.10^{-5}	10	20	25270	3187

Bảng 3: Thông tin kết quả tốt nhất mà mô hình thu được với việc thiết lập các tham số tại bảng 2 đối với tập dữ liệu UIT-ViQuAD [5].

Mô hình	**EM**	**F1**
PhoBERT-base	53.8473	77.9264

So sánh kết quả thu được của mô hình thông qua bảng 3 và kết quả thu được từ cuộc thi VLSP 2021 - Vietnamese Machine Reading Comprehension tại hình 10 cho thấy rằng mô hình PhoBERT do tôi đào tạo cho kết quả tốt đối với tiếng Việt. Với kết quả này, mô hình có thể áp dụng vào thực tiễn trong các hệ thống hỏi đáp cho các mục đích cụ thể.

7 Địa chỉ đăng tải sản phẩm

Đường dẫn Github phía BE của ứng dụng: https://github.com/SpiderMan-XiaoDo/KhoaLuanTotNghiep.git

Đường dẫn Github phía FE của ứng dụng: https://github.com/SpiderMan-XiaoDo/myapp.git

8 Cam kết về bản quyền sản phẩm

- Sản phẩm chưa từng được công bố hoặc tham gia trong bất kỳ cuộc thi nào trước đây;

- Sản phẩm đúng bản quyền của sinh viên dự thi, trường hợp sử dụng mã nguồn mở phải tuân thủ các yêu cầu giấy phép mã nguồn mở của các tổ chức, cá nhân phát hành mã nguồn mở.

Tài liệu tham khảo

[1] Alice K. Wolf Carol Chomsky Bert F. Green, Jr. and Kenneth Laughery Lincoln Laboratory. Baseball: an automatic question-answerer. 1961.

[2] Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[3] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Łukasz Kaiser Illia Polosukhin Ashish Vaswani, NoamShazeer. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[4] *VLSP 2021 - Vietnamese Machine Reading Comprehension Result (https://aihub.ml/ competitions/public_submissions/35)*.

[5] Anh Gia Tuan Nguyen Ngan Luu Thuy Nguyen Kiet Van Nguyen, Duc Vu Nguyen. A vietnamese dataset for evaluating machine reading comprehension, 2020.