TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TRẦN VĂN THẢO - 52000805**

**NGUYỄN VÕ CÔNG HUY - 52000765**

**XÂY DỰNG HỆ THỐNG TƯ VẤN SINH VIÊN**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**TÍNH TOÁN THÔNG MINH**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TRẦN VĂN THẢO - 52000805**

**NGUYỄN VÕ CÔNG HUY - 52000765**

**XÂY DỰNG HỆ THỐNG TƯ VẤN SINH VIÊN**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**TÍNH TOÁN THÔNG MINH**

Người hướng dẫn

**TS. TRẦN LƯƠNG QUỐC ĐẠI**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn …………………………………… …………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

*TP. Hồ Chí Minh, ngày 03 tháng03 năm 2024*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS. Trần Lương Quốc Đại. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 03 tháng 03 năm 2024*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

**XÂY DỰNG HỆ THỐNG TƯ VẤN SINH VIÊN**

**TÓM TẮT**

(Time New Romans – 13)

**TITLE**

**ABSTRACT**

(Time New Romans – 13)

**MỤC LỤC**

[DANH MỤC HÌNH VẼ vii](#_Toc1704507187)

[DANH MỤC BẢNG BIỂU viii](#_Toc681540259)

[DANH MỤC CÁC CHỮ VIẾT TẮT ix](#_Toc613665809)

[CHƯƠNG 1. MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI 1](#_Toc517161626)

[1.1 Lý do chọn đề tài 1](#_Toc565157308)

[1.2 Mục tiêu thực hiện đề tài 1](#_Toc83623914)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 2](#_Toc1405889382)

[2.1 Hệ thống hỏi đáp mở (ODQA) 2](#_Toc760865402)

[2.1.1 Hệ thống hỏi đáp mở là gì? 2](#_Toc226976765)

[2.1.2 Các loại hệ thống hỏi đáp mở trong ODQA 3](#_Toc2135893499)

[2.2 Hệ thống hỏi đáp mở: Retriever-Reader 3](#_Toc245593057)

[2.2.1 Tổng quan về quy trình thực hiện 4](#_Toc2142179560)

[2.2.2 Hệ thống thu thập thông tin (Retriever) 5](#_Toc73486709)

[2.2.3 Quá trình đọc và trả lời (Reader) 7](#_Toc698758535)

[2.3 Hệ thống hỏi đáp mở: Retriever-Generative 7](#_Toc1727040831)

[2.3.1 Tổng quan về quy trình thực hiện 8](#_Toc1499419334)

[2.3.2 Quá trình tạo câu trả lời (Generator) 8](#_Toc1072602732)

[CHƯƠNG 3. MÔ HÌNH ĐỀ XUẤT 8](#_Toc1565019267)

[3.1 Mô hình TF-IDF 8](#_Toc2122988554)

[3.1.1 Giới thiệu mô hình TF-IDF 9](#_Toc1917736486)

[3.1.2 Phương pháp tìm kiếm thông tin 9](#_Toc41504787)

[3.2 Mô hình Bi directional LSTM 11](#_Toc1983974638)

[3.2.1 Giới thiệu mô hình Bi directional LSTM 11](#_Toc330538447)

[3.2.2 Quá trình đọc và trích xuất thông tin 11](#_Toc310127272)

[3.3 Mô hình PhoBERT (Generative) 11](#_Toc583565534)

[3.3.1 Giới thiệu mô hình PhoBERT 11](#_Toc1017952354)

[3.3.2 Tạo câu trả lời tự nhiên 11](#_Toc1112052172)

[CHƯƠNG 4. THỰC NGHIỆM 11](#_Toc2021892495)

[4.1 Dữ liệu thực nghiệm 11](#_Toc1692015313)

[4.2 Cài đặt thực nghiệm 12](#_Toc2121163106)

[CHƯƠNG 5. KẾT LUẬN 13](#_Toc805795170)

[5.1 Kết luận 13](#_Toc1249038852)

[5.2 Hướng phát triển 13](#_Toc655219717)

[TÀI LIỆU THAM KHẢO 14](#_Toc1419460979)

# DANH MỤC HÌNH VẼ

# DANH MỤC BẢNG BIỂU

[Bảng 4.1 : Thống kê kiểu thực thể trong tập VLSP 2016 11](#_Toc931386322)

# DANH MỤC CÁC CHỮ VIẾT TẮT

|  |  |
| --- | --- |
| BERT | Bidirectional Encoder Representations from Transformers |
| GEC | Grammatical Error Correction |
| MLM | Masked Language Model |
| NLP | Natural Language Processing |
| NSP | Next Sentence Prediction |

# MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI

## Lý do chọn đề tài

Một trong những lý do quan trọng để chọn đề tài này là đáp ứng nhu cầu ngày càng tăng về thông tin trong quá trình tuyển sinh. Hệ thống tuyển sinh truyền thống thường gặp khó khăn trong việc cung cấp thông tin nhanh chóng và chính xác đối với các học sinh đang tìm hiểu về quy trình tuyển sinh. Sự chậm trễ trong phản hồi có thể tạo ra rủi ro về việc học sinh không nhận được thông tin đầy đủ và kịp thời, ảnh hưởng đến quyết định của họ trong việc chọn lựa trường học.

Hệ thống Q&A có khả năng ngay lập tức cung cấp thông tin chi tiết và rõ ràng, giúp học sinh giải đáp mọi thắc mắc một cách hiệu quả. Bằng cách này, đề tài không chỉ giải quyết vấn đề bất cập trong quá trình tuyển sinh mà còn đáp ứng nhu cầu ngày càng cao của học sinh đối với thông tin nhanh chóng và dễ tiếp cận.

## Mục tiêu thực hiện đề tài

Mục tiêu chính của nghiên cứu là phát triển một hệ thống Q&A trong quá trình tuyển sinh, nhằm giảm thời gian phản hồi, đa dạng hóa khả năng trả lời, tự động cập nhật thông tin, và thực nghiệm ứng dụng trong môi trường thực tế. Mục tiêu này nhằm tối ưu hóa trải nghiệm thông tin cho học sinh, giảm công sức của tổ chức tuyển sinh, và tạo cơ sở cho phát triển tương lai trong ứng dụng trí tuệ nhân tạo trong giáo dục.

# CƠ SỞ LÝ THUYẾT

## Hệ thống hỏi đáp mở (ODQA)

### Hệ thống hỏi đáp mở là gì?

Hệ thống hỏi đáp mở (Open Domain Question Answering) là một loại nhiệm vụ ngôn ngữ tự nhiên, yêu cầu mô hình tạo ra câu trả lời cho các câu hỏi đối với dữ liệu thực tế bằng ngôn ngữ tự nhiên. Câu trả lời đúng là khách quan nên việc đánh giá hiệu quả của mô hình rất đơn giản.

Ví dụ:

Câu hỏi: Khi nào bắt đầu thời gian đăng ký xét tuyển đại học năm nay?

Câu trả lời: Thời gian đăng ký xét tuyển đại học năm nay bắt đầu từ ngày 15 tháng 6 đến hết ngày 30 tháng 6.

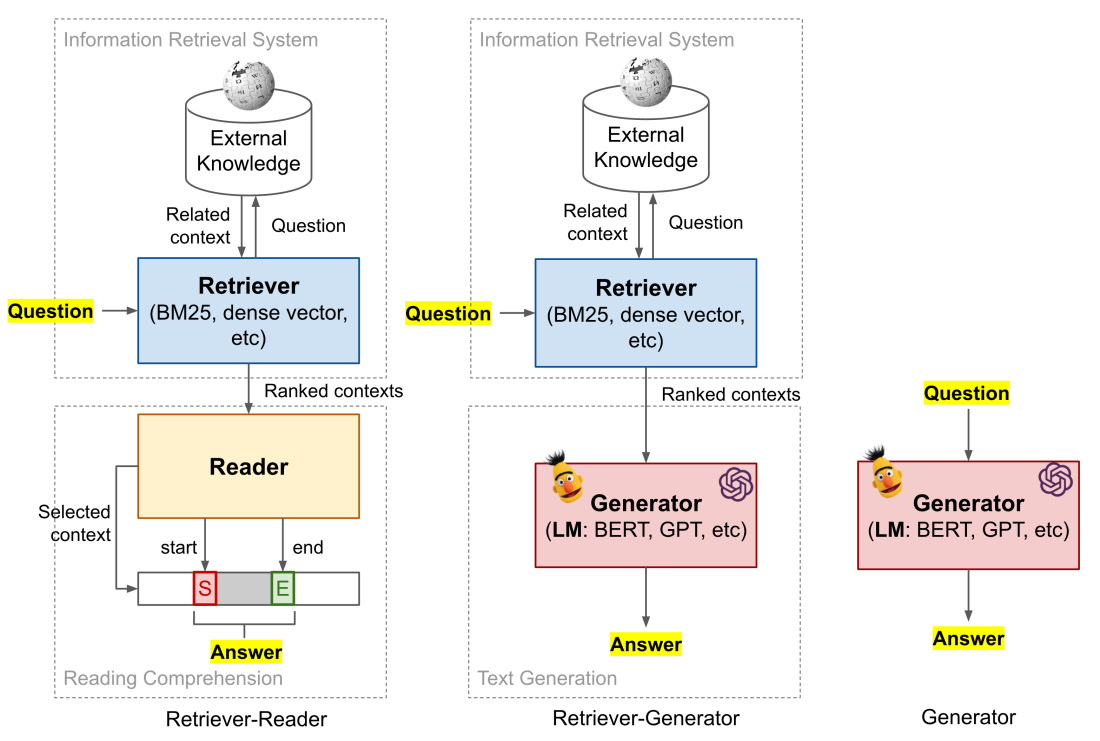
Phần "open-domain" đề cập đến sự thiếu hụt của ngữ cảnh liên quan đối với bất kỳ câu hỏi thực tế nào được đặt ra một cách tùy ý. Trong ví dụ trên, mô hình chỉ nhận câu hỏi làm đầu vào mà không có bất kỳ thông tin nào về "tại sao quyết định đăng ký xét tuyển lại có thời gian như vậy" được cung cấp. Trong trường hợp cả câu hỏi và ngữ cảnh đều được cung cấp, nhiệm vụ này được biết đến là Đọc hiểu (Reading Comprehension - RC).

Một mô hình ODQA có thể hoạt động có hoặc không có quyền truy cập vào nguồn tri thức bên ngoài (ví dụ: trang web của trường đại học), và điều này có thể được gọi là open-book hoặc closed-book.

### Các loại hệ thống hỏi đáp mở trong ODQA

Khi xem xét các loại hỏi đáp trong open-domain, thường được phân loại theo 3 loại có độ khó tăng dần:[[1]](#footnote-0)

* Mô hình có khả năng ghi nhớ đúng và trả lời với câu trả lời của một câu hỏi mà nó đã thấy trong quá trình huấn luyện.
* Mô hình có khả năng trả lời các câu hỏi mới trong quá trình kiểm thử và chọn câu trả lời từ tập các câu trả lời mà nó đã thấy trong quá trình huấn luyện.
* Mô hình có khả năng trả lời các câu hỏi mới mà có câu trả lời không xuất hiện trong tập dữ liệu huấn luyện.



Hình 2. 1 : Tổng quan về ba loại mô hình hỏi đáp mở

(Nguồn: [Lil’Log How to Build an ODQA System?](https://lilianweng.github.io/posts/2020-10-29-odqa/))

## Hệ thống hỏi đáp mở: Retriever-Reader

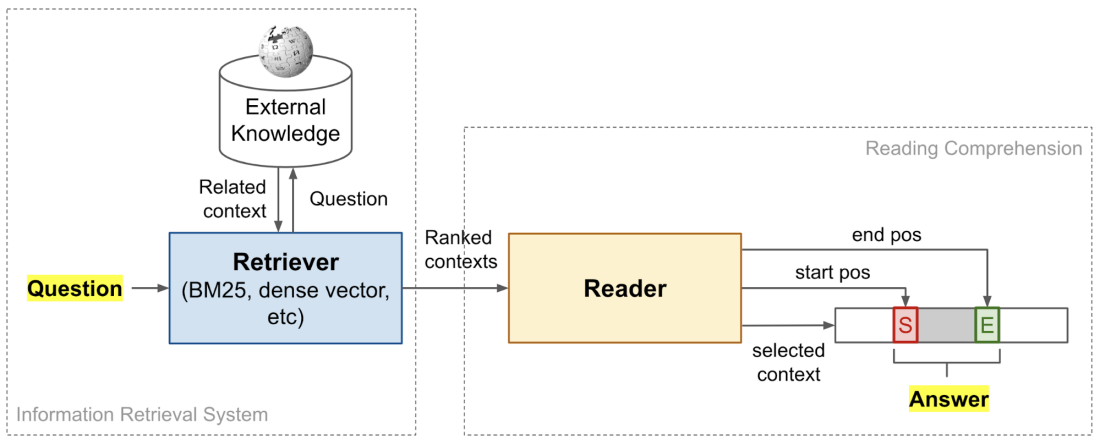
### Tổng quan về quy trình thực hiện

Dựa trên một câu hỏi thông tin cụ thể, nếu một mô hình ngôn ngữ không có ngữ cảnh hoặc không đủ lớn để ghi nhớ ngữ cảnh có sẵn trong bộ dữ liệu huấn luyện, khả năng đoán đúng câu trả lời là khá thấp. Trong kỳ thi cho phép xem tài liệu, sinh viên được phép tham khảo các nguồn tài liệu bên ngoài như ghi chú và sách khi trả lời các câu hỏi kiểm tra. Tương tự, hệ thống Hỏi-Đáp mở (ODQA) có thể được kết hợp với một cơ sở kiến thức phong phú để xác định các tài liệu liên quan làm bằng chứng cho câu trả lời.

Chúng ta có thể phân tích quy trình này thành hai giai đoạn:

* Tìm kiếm ngữ cảnh liên quan trong một kho dữ liệu kiến thức bên ngoài (Retriever):
* Ở giai đoạn này, hệ thống cần xác định các tài liệu, văn bản hoặc nguồn thông tin ngoại vi chứa thông tin liên quan đến câu hỏi.
* Quá trình này có thể thực hiện thông qua một hệ thống hỏi đáp mở (ODQA) được kết hợp với một kho dữ liệu lớn chứa kiến thức đa dạng.
* Xử lý ngữ cảnh đã tìm được để trích xuất câu trả lời (Reader):
* Sau khi có được ngữ cảnh liên quan, hệ thống cần phân tích và xử lý thông tin để trích xuất câu trả lời đáp ứng câu hỏi.
* Các phương pháp xử lý ngôn ngữ tự nhiên và trích xuất thông tin có thể được áp dụng để hiểu và tìm ra câu trả lời trong ngữ cảnh tìm kiếm.

Để hiểu tổng quan về quy trình tìm kiếm và cho ra câu trả lời chúng ta có thể quan sát hình 2.2

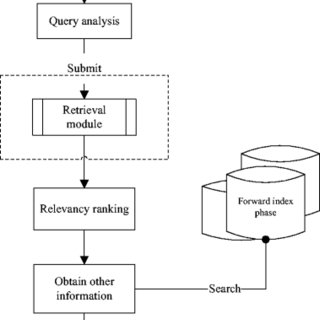


Hình 2. 2 : Quy trình tìm kiếm câu trả lời cho câu hỏi của Retriever-Reader

(Nguồn: [Lil’Log How to Build an ODQA System?](https://lilianweng.github.io/posts/2020-10-29-odqa/))

### Hệ thống thu thập thông tin (Retriever)

Hệ thống thu thập thông tin (Retriever) trong ngữ cảnh truy vấn thông tin (Information Retrieval) là một phần quan trọng của mô hình hệ thống truy vấn thông tin. Nhiệm vụ chính của Retriever là tìm kiếm và trả về các tài liệu phù hợp nhất với truy vấn từ người dùng.



Hình 2. 3 : Sơ đồ quy trình hệ thống thu thập thông tin hoạt động

(Nguồn: [Information retrieval of mass encrypted data over multimedia networking with N-level vector model-based relevancy ranking](https://www.researchgate.net/publication/291184443_Information_retrieval_of_mass_encrypted_data_over_multimedia_networking_with_N-level_vector_model-based_relevancy_ranking))

Có hai phương pháp chính để triển khai hệ thống thu thập thông tin (retriever) trong hệ thống trả lời câu hỏi mở (ODQA):

* Hệ thống thu thập thông tin (IR) dựa trên các đặc trưng TF-IDF cổ điển không dựa trên học máy (Classic IR): Phương pháp này tập trung vào việc sử dụng hệ thống thu thập thông tin (IR) truyền thống, dựa trên đặc trưng TF-IDF để xác định sự tương đồng giữa câu hỏi và các văn bản nguồn. Các truy vấn được biểu diễn bằng các vector dựa trên tần suất xuất hiện của các từ khóa và các văn bản được sắp xếp theo độ tương đồng.
* Hệ thống thu thập thông tin (IR) dựa trên các vector dense của văn bản được tạo ra bởi các mạng nơ-ron (neural IR): Phương pháp này sử dụng các mô hình mạng nơ-ron để tạo ra các biểu diễn vector dense của văn bản, thường thông qua các mô hình như BERT, RoBERTa, hay các biến thể khác của Transformer. Các biểu diễn này mang lại khả năng hiểu ngữ nghĩa sâu rộng và tương đồng ngữ cảnh giữa câu hỏi và nguồn thông tin.

### Quá trình đọc và trả lời (Reader)

Mô hình Reader là một phần quan trọng của hệ thống hỏi đáp mở, chịu trách nhiệm đọc và hiểu nội dung của văn bản để tìm kiếm thông tin cần thiết để trả lời các câu hỏi từ người dùng. Mục tiêu chính của mô hình là hiểu và rút trích thông tin ý nghĩa từ văn bản nguồn để tạo ra câu trả lời chính xác.

Kiến trúc chung của mô hình Reader gồm các phần sau:

* Lớp Embedding: Chuyển đổi từ ngôn ngữ tự nhiên thành các vector số thực để máy tính có thể hiểu được ý nghĩa của từng từ.
* Kiến Trúc Tự Chú Ý (Self-Attention): Các lớp tự chú ý giúp mô hình tập trung vào các phần quan trọng của văn bản, xử lý thông tin từ cấp từ đến cấp câu.
* Trích Xuất Thông Tin: Sau khi đọc và hiểu, mô hình trích xuất thông tin quan trọng để sử dụng trong quá trình tạo câu trả lời.

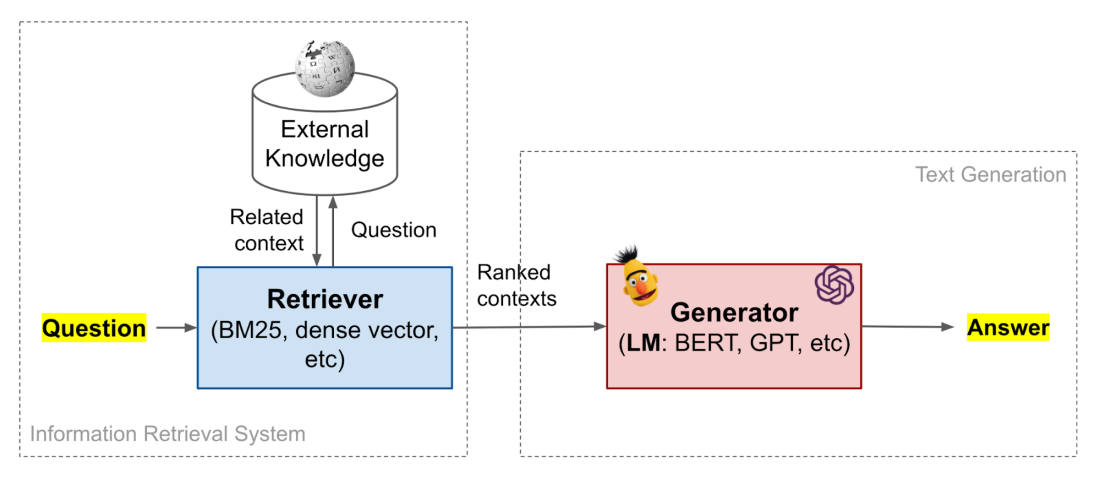
Các mô hình thường sử dụng trong vai trò Reader: Bi directional LSTM, BERT,...

## Hệ thống hỏi đáp mở: Retriever-Generative

### Tổng quan về quy trình thực hiện

So với phương pháp thu thập thông tin và đọc hiểu (retriever-reader), phương pháp thu thập thông tin và tạo văn bản (retriever-generator) cũng gồm 2 giai đoạn, nhưng giai đoạn thứ hai được thiết kế để tạo ra văn bản tự do trực tiếp để trả lời câu hỏi thay vì trích xuất vị trí bắt đầu/kết thúc trong một đoạn văn được thu thập. Một số nghiên cứu cũng gọi phương pháp này là "Generative Question Answering" (Trả lời câu hỏi bằng cách tạo văn bản).

Để hiểu tổng quan về quy trình tìm kiếm và cho ra câu trả lời của retriever - generator chúng ta có thể quan sát hình 2.4



Hình 2. 4 : Quy trình tìm kiếm câu trả lời cho câu hỏi của Retriever-Generator

(Nguồn: [Lil’Log How to Build an ODQA System?](https://lilianweng.github.io/posts/2020-10-29-odqa/))

### Quá trình tạo câu trả lời (Generator)

### Không giống như mô hình Reader, mô hình Generator trong hệ thống hỏi đáp mở là một thành phần quan trọng giúp tạo ra câu trả lời chính xác và có ý nghĩa cho các câu hỏi mà hệ thống không có sẵn câu trả lời trong nguồn dữ liệu cố định.

Kiến trúc chung của mô hình Generator gồm các phần sau:

* Mạng Nơ-ron Tự Sinh (Seq2Seq): Sử dụng mô hình Seq2Seq, trong đó có một mạng nơ-ron Encoder chịu trách nhiệm biểu diễn câu hỏi đầu vào thành một vector ngữ cảnh. Mạng nơ-ron Decoder sẽ sinh ra câu trả lời dựa trên vector ngữ cảnh đã biểu diễn.
* Attention Mechanism: Các mô hình thường sử dụng cơ chế chú ý để tập trung vào các phần quan trọng của câu hỏi và nguồn dữ liệu khi tạo ra câu trả lời. Cơ chế này giúp mô hình tập trung vào các từ quan trọng và mối quan hệ trong ngữ cảnh.
* Transfer Learning và Pre-training: Sử dụng kỹ thuật pre-training để học các biểu diễn ngôn ngữ tự nhiên từ dữ liệu lớn trước khi fine-tuning cho nhiệm vụ ODQA cụ thể. Mô hình có thể sử dụng các kiến thức từ các nguồn lớn như Wikipedia để cải thiện khả năng sinh câu trả lời.

Các mô hình thường sử dụng trong vai trò Generator: T5, GPT, BERT,...

# MÔ HÌNH ĐỀ XUẤT

## Mô hình TF-IDF

### Giới thiệu mô hình TF-IDF

Mô hình TF-IDF (Term Frequency-Inverse Document Frequency), là một trong những phương pháp truy vấn thông tin cổ điển nhất được sử dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên và truy vấn thông tin.

### Phương pháp tìm kiếm thông tin

Dưới đây là mô tả chi tiết về cách mô hình TF-IDF hoạt động tìm kiếm thông tin:

1. \*\*Thu thập dữ liệu:\*\*

- Dữ liệu đầu vào của mô hình là tập văn bản (ví dụ: tài liệu, bài viết, văn bản) được gọi là "tập tài liệu".

- Mỗi tài liệu trong tập được biểu diễn bởi một bộ từ vựng, tức là tập hợp tất cả các từ xuất hiện trong tài liệu đó.

2. \*\*Tạo chỉ số (Indexing):\*\*

- Mỗi từ trong tập hợp từ vựng được gán một số thứ tự (index).

- Dùng index này để tạo ra một ma trận, với mỗi dòng biểu diễn một tài liệu và mỗi cột biểu diễn một từ trong bộ từ vựng.

- Mỗi ô trong ma trận chứa giá trị TF-IDF của từ đó trong tài liệu tương ứng.

3. \*\*Tính giá trị Term Frequency (TF):\*\*

- Term Frequency đo tần suất xuất hiện của mỗi từ trong một tài liệu. Công thức tính TF cho một từ w trong tài liệu d là:

\[TF(w, d) = \frac{\text{số lần xuất hiện của w trong d}}{\text{Tổng số từ trong d}}\]

4. \*\*Tính giá trị Inverse Document Frequency (IDF):\*\*

- Inverse Document Frequency đo mức độ quan trọng của một từ trong toàn bộ tập tài liệu. Công thức tính IDF cho một từ w trong toàn bộ tập tài liệu là:

\[IDF(w) = \log\left(\frac{\text{Tổng số tài liệu}}{\text{Số tài liệu chứa từ w + 1}}\right)\]

(+1 ở mẫu để tránh trường hợp chia cho 0 nếu từ đó không xuất hiện trong bất kỳ tài liệu nào)

5. \*\*Tính giá trị TF-IDF:\*\*

- Giá trị TF-IDF của mỗi từ trong mỗi tài liệu được tính bằng cách nhân giá trị TF với giá trị IDF:

\[TF-IDF(w, d) = TF(w, d) \times IDF(w)\]

6. \*\*Truy vấn và tính điểm tương tự:\*\*

- Khi có một truy vấn, cũng được biểu diễn dưới dạng vector TF-IDF.

- Tính điểm tương tự giữa truy vấn và mỗi tài liệu bằng cách sử dụng các phương pháp đo lường sự tương đồng như cosine similarity.

- Các tài liệu được sắp xếp theo điểm tương tự và trả về kết quả.

Mô hình TF-IDF là một mô hình đơn giản nhưng hiệu quả trong nhiều trường hợp truy vấn thông tin và tìm kiếm văn bản. Tuy nhiên, nó không quan tâm đến cấu trúc ngữ nghĩa của câu hoặc văn bản, điều này là một hạn chế khi so sánh về mặt ngữ nghĩa.

## Mô hình Bi directional LSTM

### Giới thiệu mô hình Bi directional LSTM

### Quá trình đọc và trích xuất thông tin

## Mô hình PhoBERT (Generative)

### Giới thiệu mô hình PhoBERT

### Tạo câu trả lời tự nhiên

# THỰC NGHIỆM

## Dữ liệu thực nghiệm

Dữ liệu là bộ câu hỏi về tuyển sinh đại học trong VLSP-2016 đã được chia sẵn thành hai tập huấn luyện và kiểm thử. Thống kê chi tiết của các kiểu thực thể và các tập dữ liệu được mô tả trong Bảng 4.1**.**

Bảng 4.1: Thống kê kiểu thực thể trong tập VLSP 2016

|  |  |  |
| --- | --- | --- |
| **Label** | **Train** | **Test** |
| **LOC** | 1210 | 1377 |
| **ORG** | 7478 | 274 |
| **PER** | 6230 | 1294 |
| **MISC** | 250 | 47 |
| **Total** | 15168 | 2992 |

## Cài đặt thực nghiệm

…

# KẾT LUẬN

## Kết luận

## Hướng phát triển

# TÀI LIỆU THAM KHẢO

Tiếng Việt

…

Tiếng Anh

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation*, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

1. [Lewis, et al., 2020](https://arxiv.org/abs/2008.02637) [↑](#footnote-ref-0)