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](#_Toc1197373473)

[DANH MỤC BẢNG BIỂU viii](#_Toc72753690)

[DANH MỤC CÁC CHỮ VIẾT TẮT ix](#_Toc273501090)

[CHƯƠNG 1. MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI 1](#_Toc308688145)

[1.1 Lý do chọn đề tài 1](#_Toc1596410667)

[1.2 Mục tiêu thực hiện đề tài 1](#_Toc203633156)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 2](#_Toc1859134618)

[2.1 Hệ thống hỏi đáp mở (ODQA) 2](#_Toc1391475391)

[2.1.1 Hệ thống hỏi đáp mở là gì? 2](#_Toc89627787)

[2.1.2 Các loại hệ thống Q&A mở trong ODQA 3](#_Toc6085625)

[2.2 Open-book QA: Retriever-Reader 3](#_Toc196645934)

[2.2.1 Tổng quan về quy trình thực hiện 4](#_Toc2120497140)

[2.2.2 Mô hình Retriever 5](#_Toc1710779058)

[2.3 Open-book QA: Retriever-Generative 5](#_Toc886694100)

[CHƯƠNG 3. MÔ HÌNH ĐỀ XUẤT 6](#_Toc2114985694)

[3.1 Mô hình classicIR(Retriever) 6](#_Toc1991186671)

[3.1.1 Giới thiệu mô hình classicIR 6](#_Toc528924564)

[3.1.2 Phương pháp tìm kiếm thông tin 6](#_Toc1797788998)

[3.2 Mô hình BiLSTM(Reader) 6](#_Toc2111574929)

[3.2.1 Giới thiệu mô hình BiLSTM 6](#_Toc475149601)

[3.2.2 Quá trình đọc và trích xuất thông tin 6](#_Toc1014985609)

[3.3 Mô hình PhoBERT (Generative) 6](#_Toc786167079)

[3.3.1 Giới thiệu mô hình PhoBERT 6](#_Toc337615184)

[3.3.2 Tạo câu trả lời tự nhiên 6](#_Toc1891933431)

[CHƯƠNG 4. THỰC NGHIỆM 6](#_Toc308950404)

[4.1 Dữ liệu thực nghiệm 6](#_Toc1103927606)

[4.2 Cài đặt thực nghiệm 7](#_Toc361094310)

[CHƯƠNG 5. KẾT LUẬN 8](#_Toc985126430)

[5.1 Kết luận 8](#_Toc1953554310)

[5.2 Hướng phát triển 8](#_Toc1000189501)

[TÀI LIỆU THAM KHẢO 9](#_Toc2120527696)

# DANH MỤC HÌNH VẼ

# DANH MỤC BẢNG BIỂU

[Bảng 4.1 : Thống kê kiểu thực thể trong tập VLSP 2016 7](#_Toc1135401266)

# 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.

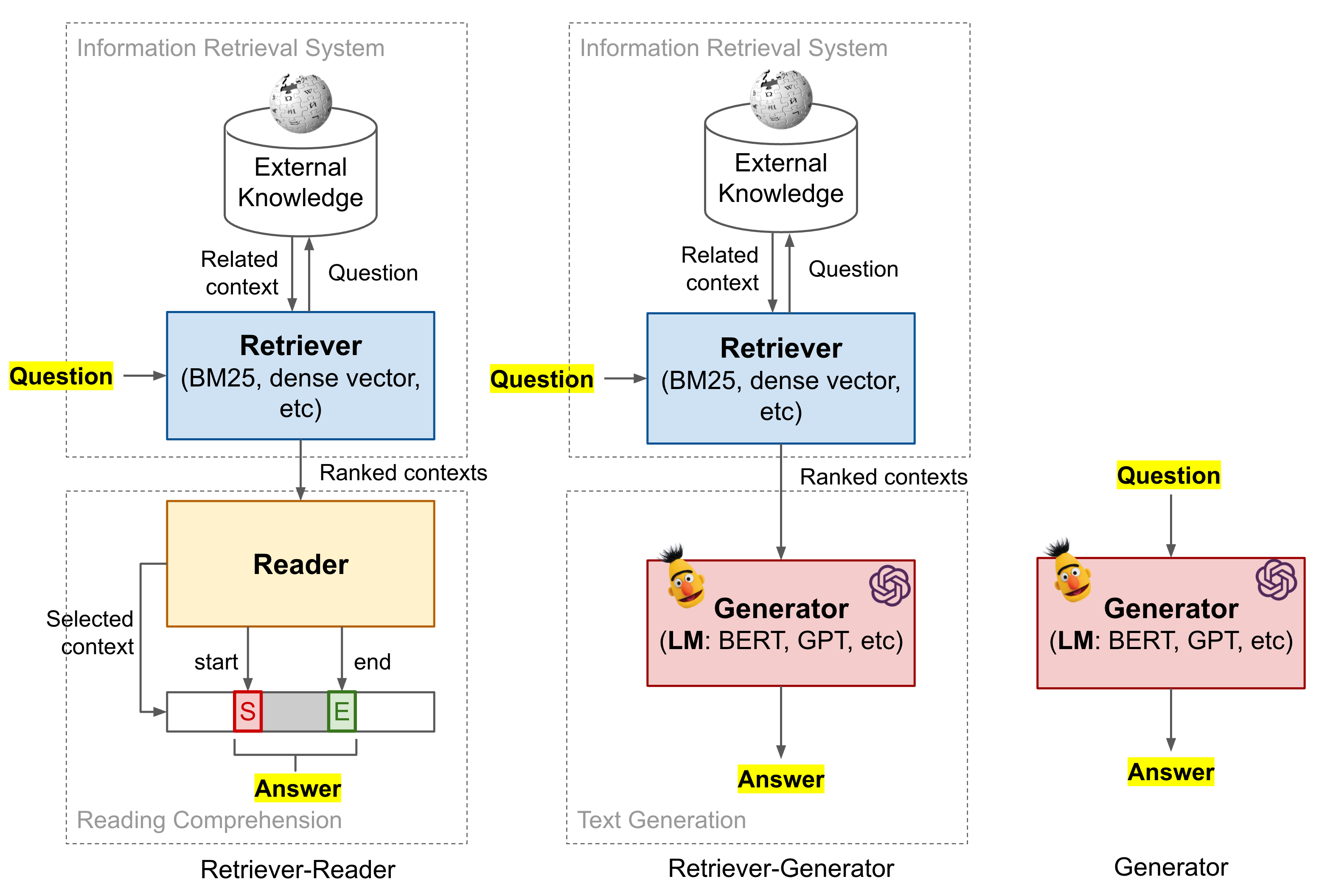
Phần "open-domain" đề cập đến sự thiếu hụt của ngữ cảnh liên quan đối với bất kỳ câu hỏi thực tế nào được đặt ra một cách tùy ý. Trong ví dụ trên, mô hình chỉ nhận câu hỏi làm đầu vào mà không có bất kỳ thông tin nào về "tại sao quyết định đăng ký xét tuyển lại có thời gian như vậy" được cung cấp. Trong 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 Q&A mở trong ODQA

Khi xem xét các loại Q&A 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 Q&A trong open-domain

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

## Open-book QA: Retriever-Reader

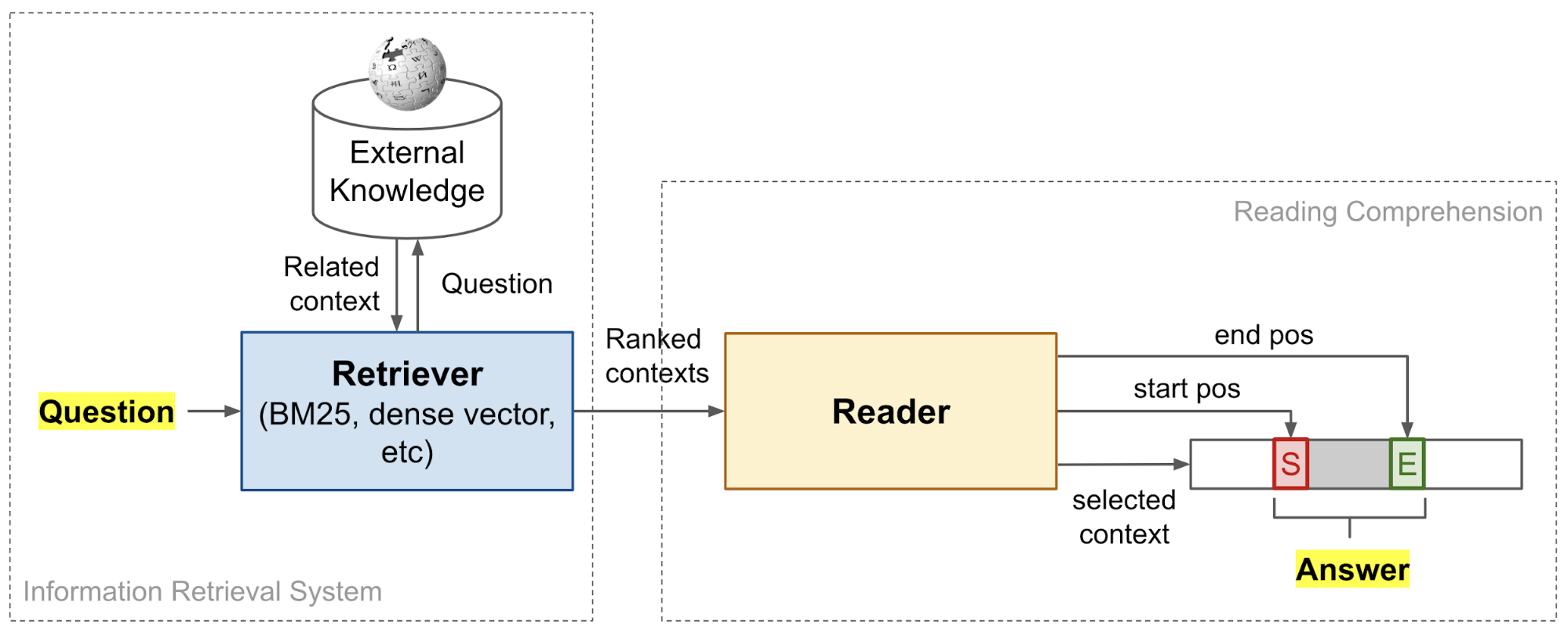
### Tổng quan về quy trình thực hiện

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 tìm kiếm câu trả lời cho các câu hỏi cho trước thành hai giai đoạn:

* Tìm kiếm ngữ cảnh liên quan trong một kho dữ liệu kiến thức bên ngoài:
* 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:
* 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 rõ 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/))

### Mô hình Retriever

Có hai phương pháp phổ biến để triển khai hệ thống thu thập thông tin (retriever) trong hệ thống trả lời câu hỏi mở (ODQA): sử dụng hệ thống thu thập thông tin (IR) dựa trên các đặc trưng TF-IDF cổ điển không dựa trên học ("Classic IR") hoặc (2) các vector dense của văn bản được tạo ra bởi các mạng nơ-ron ("neural IR").

## Open-book QA: Retriever-Generative

# MÔ HÌNH ĐỀ XUẤT

## Mô hình classicIR(Retriever)

### Giới thiệu mô hình classicIR

### Phương pháp tìm kiếm thông tin

## Mô hình BiLSTM(Reader)

### Giới thiệu mô hình BiLSTM

### 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 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)