

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ GIAO THÔNG VẬN TẢI
KHOA CÔNG NGHỆ THÔNG TIN



Nguyễn Tiến Đạt

**XÂY DỰNG HỆ THỐNG TRẢ LỜI THÔNG TIN TỰ
ĐỘNG CHO TRƯỜNG ĐẠI HỌC CÔNG NGHỆ GTVT
SỬ DỤNG MÔ HÌNH TRANSFORMER**

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ thông tin

HÀ NỘI - 2023

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ GIAO THÔNG VẬN TẢI
KHOA CÔNG NGHỆ THÔNG TIN

Nguyễn Tiến Đạt

XÂY DỰNG HỆ THỐNG TRẢ LỜI THÔNG TIN TỰ
ĐỘNG CHO TRƯỜNG ĐẠI HỌC CÔNG NGHỆ GTVT
SỬ DỤNG MÔ HÌNH TRANSFORMER

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ thông tin

Cán bộ hướng dẫn: Ths. Đoàn Thị Thanh Hằng

HÀ NỘI - 2023

MỤC LỤC

CHƯƠNG 1. TỔNG QUAN	2
1.1. Đặt vấn đề.....	2
1.2. Mục tiêu và phạm vi đề tài.....	3
1.3. Định hướng giải pháp.....	4
1.4. Giới hạn và phạm vi thực hiện	5
1.5. Phương pháp nghiên cứu	5
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT.....	6
2.1. Tổng quan về trí tuệ nhân tạo, học máy, học sâu	6
2.1.1. Khái niệm	6
2.1.2. Các dạng bài toán	7
2.1.3. Overfitting và Underfitting.....	9
2.1.4. Quá trình hình thành mô hình machine learning	10
2.1.5. Thách thức và những lợi ích của AI mang lại	11
2.2. Tổng quan về mô hình transformer.....	12
2.2.1. Vanilla Transformer	13
2.2.2. Các dạng kiến trúc tiêu biểu	16
2.2.3. Generative Model	18
2.3. Tổng quan về hệ thống chatbot	19
2.3.1. Khái niệm	19
2.3.2. Các phương pháp xây dựng chatbot	21
2.3.3. Hệ thống hỏi đáp	21
2.3.4. Retrieval	24
2.3.5. Generator Answer.....	27
2.4. Nền tảng phát triển.....	29
2.4.1. Pytorch.....	29
2.4.2. Transformers.....	30
2.4.3. ChromaDB	31
2.4.4. FastAPI	32
2.4.5. Gradio	33
CHƯƠNG 3: PHÂN TÍCH VÀ THIẾT KẾ KIẾN TRÚC HỆ THỐNG	35

3.1. Khảo sát và phân tích các chức năng.....	35
3.1.1. Mục đích của khảo sát	35
3.1.2. Nội dung khảo sát.....	35
3.1.4. Phân tích chức năng.....	37
3.2. Kiến trúc hệ thống tổng quan.....	38
3.2.1. Domain Classification	40
3.2.2. Chitchat.....	41
3.2.3. Generator Answer.....	42
3.2.4. Retrieval và Ranker	44
CHƯƠNG 4: XÂY DỰNG HỆ THỐNG	46
4.1. Xây dựng mô hình	46
4.1.1. Chuẩn bị và xây dựng dữ liệu.....	46
4.1.2. Thực nghiệm mô hình	54
4.1.3. Đánh giá.....	62
4.2. Xây dựng hệ thống.....	78
4.2.1. Thiết kế API	78
4.2.2. Thiết kế giao diện	79
4.2.3. Đánh giá.....	80
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	95
1. Kết quả đạt được	95
2. Hướng phát triển	95
3. Hạn chế	95
4. Tổng kết.....	95
TÀI LIỆU THAM KHẢO.....	96

DANH MỤC HÌNH ẢNH

Hình 2.1: Biểu diễn miền của trí tuệ nhân tạo > học máy > học sâu..	7
Hình 2.2: So sánh giữa supervised learning và unsupervised learning	9
Hình 2.3: Các loại khớp dữ liệu có thể có trong mô hình	10
Hình 2.4: Quá trình phát triển một hệ thống AI nói chung	11
Hình 2.5: Kiến trúc mô hình transformer	13
Hình 2.6: Biểu diễn trực quan self-attention	15
Hình 2.7: Mô hình BERT	17
Hình 2.8: Hình ảnh tên các mô hình pretrained	17
Hình 2.9: Formatting Task Datasets	18
Hình 2.10: Trực quan LoRA, khi training tham số A, B sẽ được cập nhật còn tham số của mô hình được giữ nguyên	19
Hình 2.11: Các hệ thống hỏi đáp	23
Hình 2.12: Mô hình Reading Comprehension	27
Hình 2.13: Thư viện Pytorch	29
Hình 2.14: Thư viện Transformers	30
Hình 2.15: Thư viện Vector Database ChromaDB	31
Hình 2.16: Thư viện xây dựng API FastAPI	33
Hình 2.17: Thư viện xây dựng giao diện Gradio	27
Hình 3.1: Biểu đồ phân cấp chức năng	37
Hình 3.2: Kiến trúc tổng quan của hệ thống	39
Hình 3.3: Domain Classification	40
Hình 3.4: Nạp dữ liệu vào vector database	41
Hình 3.5: Mô đun Chitchat	42
Hình 3.6: Các dạng dữ liệu question answering	43
Hình 3.7: Retrieval và Ranker	44
Hình 3.8: Phép đo độ tương tự giữa hai vector - Cosine Similarity	45
Hình 4.1: Dữ liệu STS - 2012 (gốc)	47
Hình 4.2: Dịch STS (ENG) sang STS (VIE)	47
Hình 4.3: Dữ liệu STS - 2012 (Đã được dịch)	47
Hình 4.4: Dữ liệu Zalo đã được xử lý	48
Hình 4.5: Dữ liệu data Viettel Hearted AI Challenges	49

Hình 4.6: Dữ liệu data Viettel Hearted AI Challenges (filtered).....	49
Hình 4.7: Trang thông tin giảng viên	51
Hình 4.8: Thông tin của một số thầy cô trong trường	51
Hình 4.9: Mỗi chủ đề sẽ được lưu một file riêng	52
Hình 4.10: Chia nhỏ văn bản thành các đoạn văn	53
Hình 4.11: Kiến trúc Sentence Transformer.....	54
Hình 4.12: Thử nghiệm sử dụng nhiều hidden state	55
Hình 4.13: Sentence Embedding sau khi được cải tiến.....	56
Hình 4.14: Biểu đồ training loss Sentence Embedding - Zalo	57
Hình 4.15: Biểu đồ training loss Sentence Embedding - STS.....	58
Hình 4.16: Sentence Embedding inference	58
Hình 4.17: Cross-Encoder	59
Hình 4.18: Cross-Encoder sau khi được cải tiến	59
Hình 4.19: Biểu đồ training loss Cross-Encoder - Zalo	60
Hình 4.20: Biểu đồ training loss Cross-Encoder - STS.....	61
Hình 4.21: Biểu đồ training loss Bloomz-1b7.....	62
Hình 4.22: API Sentence Embedding.....	78
Hình 4.23: API Ranker	79
Hình 4.24: API Question Answering Generation.....	79
Hình 4.25: Giao diện chatbox	80

DANH MỤC BẢNG

Bảng 4.1: Kết quả đánh giá tác vụ Sentence Similarity..	62
Bảng 4.2: Kết quả đánh giá tác vụ Question-Answering của mô hình Bloomz-1b1	63
Bảng 4.3: Kết quả đánh giá hệ thống trả lời trên tác vụ hỏi đáp giới thiệu bản thân của chatbot	81
Bảng 4.4: Kết quả đánh giá hệ thống trả lời trên tác vụ hỏi đáp thông tin của trường .	82

LỜI CẢM ƠN

Lời đầu tiên, em xin chân thành cảm ơn các thầy/cô thuộc trường Đại học Công nghệ Giao thông Vận tải đã tận tâm giảng dạy và hỗ trợ em để hoàn thành chương trình học tập của mình. Thầy/cô như những người lái đò đã thầm lặng truyền đạt những tri thức quý báu tới chúng em và đặc biệt là những kinh nghiệm thực tế để chúng em vững tin bước trên con đường của mình. Đó là điều may mắn và hạnh phúc của chúng em khi được học tập tại trường. Em xin biết ơn sâu sắc về những duyên lành này và xin chúc các thầy/cô luôn mạnh khỏe để tiếp tục trên những chuyến đò lặng thầm này.

Đặc biệt em xin chân thành cảm ơn cô giáo, Ths. Đoàn Thị Thanh Hằng đã tận tình hướng dẫn và chỉ bảo trong suốt quá trình thực hiện. Bên cạnh đó, em cũng xin gửi lời cảm ơn tới tập thể các anh chị, đồng nghiệp Công ty Công nghệ Gia đình (FTECH.AI) đã tạo điều kiện và hỗ trợ em thực hiện đồ án. Em cũng xin gửi lời cảm ơn tới anh Nguyễn Kỳ Tùng, bạn Lê Xuân Trường và các anh chị, bạn bè, đồng nghiệp khác đã hỗ trợ và tư vấn để em hoàn thiện đồ án tốt hơn.

Vì những hiểu biết hạn hẹp của bản thân mà có những sự hiểu nhầm, tư duy chưa đúng, thực hiện chưa đúng, dẫn tới những điều sai sót trong đồ án. Do đó rất mong nhận được những lời nhận xét, góp ý chân thành từ thầy cô và hội đồng bảo vệ để em có thể cập nhật sửa đổi và hoàn thiện hơn.

Em xin chân thành cảm ơn!

LỜI CAM ĐOAN

Em xin cam đoan rằng đồ án tốt nghiệp này là kết quả của công việc nghiêm túc và tự mình nghiên cứu. Tất cả các thông tin, số liệu và kết quả mà em đưa ra trong đồ án này đều được trình bày một cách chính xác và trung thực

Em xin chịu hoàn toàn trách nhiệm và chịu mọi hình thức kỷ luật từ khoa và nhà trường nếu những lời trên không đúng sự thật.

Hà Nội, ngày ... tháng ... năm 2023

SINH VIÊN THỰC HIỆN

(Ký và ghi rõ họ tên)

Nguyễn Tiến Đạt

LỜI MỞ ĐẦU

Trong những năm gần đây, chatbot đã trở nên phổ biến và được sự quan tâm bởi cộng đồng và doanh nghiệp. Từ đó, bước đầu chatbot đã có những ứng dụng khả quan trong thực tế từ việc hỗ trợ các doanh nghiệp tới các ứng dụng trong đời sống xã hội. Tuy nhiên, các hướng nghiên cứu chatbot hiện nay, mọi người thường tập trung vào giải quyết các vấn đề về giải pháp công nghệ mà ít chú trọng tới việc nghiên cứu ứng dụng vào trong thực tế. Đây là một hướng nghiên cứu quan trọng trong lĩnh vực chatbot bởi để xây dựng một chatbot hiệu quả không chỉ phụ thuộc vào công nghệ mà còn phụ thuộc vào cách chúng ta định hướng phát triển xây dựng và thiết kế chúng như thế nào. Lĩnh vực giáo dục đào tạo là một lĩnh vực quan trọng trong đời sống và đặc biệt là đối với sinh viên. Sinh viên khi bắt đầu học đại học thì chưa có tư duy đúng đắn nên thường khó khăn trong việc học tập. Từ đó, nhu cầu xây dựng một ứng dụng chatbot hỗ trợ tư vấn học sinh viên là hết sức cần thiết. Từ nhu cầu thực tiễn này và nhận thức về tầm quan trọng của việc nghiên cứu ứng dụng chatbot trong thực tế, em quyết định chọn đề tài "Xây dựng hệ thống trả lời thông tin tự động cho trường Đại học Công nghệ GTVT sử dụng mô hình Transformer" để làm đề tài cho Đồ án tốt nghiệp lần này. Với đề tài này, em chú trọng vào việc nghiên cứu cách xây dựng chatbot như thế nào cho hiệu quả và đáp ứng nhu cầu thực tiễn. Đặc biệt, em tập trung vào việc nghiên cứu xây dựng và thiết kế các tính năng và kịch bản chatbot, cụ thể là các kịch bản hỏi đáp. Trong đồ án, thông qua xây dựng một ứng dụng thực tế, em đã phân tích một số vấn đề cụ thể trong việc nghiên cứu phát triển chatbot bao gồm những khó khăn và thách thức gặp phải trong việc phân tích và thiết kế chatbot. Đồng thời, em cũng đã đưa ra giải pháp dựa trên các công nghệ hiện có để đáp ứng các vấn đề này. Từ đó, em đã xây dựng được một ứng dụng chatbot tư vấn hỗ trợ sinh viên với kết quả khả quan và có thể ứng dụng vào thực tiễn.

CHƯƠNG 1. TỔNG QUAN

Chương này sẽ giới thiệu chung về đề tài tốt nghiệp, bao gồm các vấn đề về, phạm vi và mục tiêu mà đề tài hướng tới. Từ đó, em sẽ đưa ra các giải pháp định hướng mà em sẽ thực hiện trong quá trình xây dựng hệ thống hỏi đáp này.

1.1. Đặt vấn đề

Hiện nay, với công cuộc cách mạng công nghiệp 4.0, chúng ta đã chứng kiến sự phát triển vượt bậc của khoa học công nghệ. Nó đã tác động ảnh hưởng không nhỏ đến cuộc sống của chúng ta, đi kèm với nó, đó là sự tiện lợi đồng thời mở ra nhiều cơ hội đem lại cuộc sống được thoải mái hơn. Trí tuệ nhân tạo (AI) là một trong số những công nghệ đang rất nổi hiện nay, thực tế công nghệ này đã và đang được nghiên cứu giữa những năm thế kỷ 20, tuy rằng lúc đó nó vẫn chỉ được coi là thuật toán chỉ phục vụ được các bài toán đơn giản hoặc các bài toán miền hẹp. Nhưng với hiện tại, trải qua sự nghiên cứu và cải tiến không ngừng nghỉ, AI đem lại cho chúng ta từ bất ngờ này sang bất ngờ khác, gần như các bài toán khó khăn tưởng chừng là không thể thay thế con người được, tuy nhiên AI đã chứng minh rằng không gì là không thể. Bằng chứng là các bài toán dịch máy, nhận diện cảnh báo, hay thậm chí là trợ lý ảo. Trợ lý ảo hay còn được gọi là chatbot hay hệ thống trả lời tự động đã dần trở nên phổ biến và quen thuộc khi mà chúng xuất hiện ở mọi nơi, mọi lúc. Bản chất của chúng, có thể được hiểu là các chương trình máy tính, một hệ thống có thể tương tác với người dùng bằng ngôn ngữ tự nhiên, giống như trò chuyện giữa người với người. Người dùng có thể ra lệnh hoặc yêu cầu nó thực hiện một số nhiệm vụ đơn giản, ví dụ như “Bật báo thức lúc 6h sáng” hoặc “Thời tiết hôm nay như thế nào”,.. Sự tiện dụng này đã dần thay thế các các thức hoạt động cũ lạc hậu, không hiệu quả. Với các khả năng như vậy, chatbot có thể được ứng dụng rộng rãi trong các công việc hay nhiệm vụ khác nhau như: chăm sóc khách hàng, hỗ trợ cung cấp thông tin dịch vụ, tự động hóa các lịch trình. Gần đây, nhờ sự phát triển đột phá của các thuật toán liên quan đến Học sâu (Deep learning) cùng với đó là lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP), các hệ thống chatbot giờ đây đã có thể thực hiện nhiều yêu cầu phức tạp của con người hơn (ví dụ như ChatGPT, PaLM, Bard,...). Vậy nên có thể hy vọng rằng có thể phát triển được các chatbot hỗ trợ con người trong các nhiệm vụ cụ thể và phức tạp hơn.

Đối với lĩnh vực giáo dục, ở bậc đại học, thực tế cho rằng trong quá trình học tập, sinh viên thường gặp phải rất nhiều khó khăn, có thể liên quan đến việc sinh viên vẫn còn mang tư duy học tập phổ thông lên đại học dẫn tới khó tiếp thu các kiến thức mới, sinh viên chưa được nắm bắt kịp các thông tin quy định liên quan đến chương trình học, nhà trường, sinh viên vẫn còn rụt rè, thái độ chưa cởi mở để có thể dễ dàng hỏi, trao đổi thông tin tới thầy cô và bạn bè các vấn đề khó khăn trong cuộc sống và học tập. Từ một số nguyên nhân trên, chúng ta có thể nhận ra nguyên nhân chính đến từ vấn đề tâm lý và tư duy của sinh viên, qua đó có thể nhận thấy được rằng công tác tư vấn học tập cho sinh viên trong đào tạo là một công việc hết sức cần thiết. Với những khả năng của hệ thống trả lời câu hỏi tự động, hệ thống chatbot, chúng ta có thể hoàn toàn có thể xây dựng các hệ thống trợ lý ảo để hỗ trợ sinh viên nhanh chóng và hiệu quả hơn. Nhờ khả năng cung cấp môi trường trò chuyện riêng tư với sinh viên, chatbot có thể hỗ trợ sinh viên các công việc như tìm hiểu các quy định của nhà trường, hỗ trợ tìm kiếm thông tin một cách hiệu quả hay trò chuyện tâm sự giải quyết các vấn đề khó khăn trong cuộc sống của sinh viên. Từ đó, định hướng sinh viên tới những tư duy đúng đắn trong học tập và cuộc sống, gỡ mắc dần các rào cản tâm lý để sinh viên học tập tốt hơn. Hơn nữa các vấn đề khó khăn của sinh viên xảy ra nhiều ở các sinh viên khác nhau và thường mang tính lặp lại. Dẫn tới hiện tại, tổ tư vấn học tập sinh viên thường tốn thời gian, nhân lực và công sức để hỗ trợ tư vấn, gây ra lãng phí và không hiệu quả. Thay vào đó, chatbot là một hệ thống có khả năng trả lời các thắc mắc của sinh viên một cách tự động tại mọi thời điểm khác nhau. Khi ứng dụng chatbot vào việc hỗ trợ tư vấn, chúng sẽ giúp giảng viên và các cán bộ giảm thiểu được khối lượng công việc tư vấn hơn. Hệ thống chatbot sẽ hỗ trợ giải đáp các vấn đề thường gặp, đơn giản mà sinh viên hay gặp phải, còn đối với các vấn đề mang tính riêng tư, cần sự tư vấn chuyên sâu sẽ được xử lý bởi tổ tư vấn. Từ đó, việc tư vấn hỗ trợ sinh viên sẽ hiệu quả hơn và tối ưu được chi phí hơn.

1.2. Mục tiêu và phạm vi đề tài

Như ở phần trên, chúng ta thấy rằng việc phát triển các hệ thống trợ lý ảo hỗ trợ công tác sinh viên hiệu quả, đồng thời giúp cho khối lượng công việc tư vấn của giảng viên và các cán bộ giảm thiểu khối lượng công việc nhiều hơn. Tuy nhiên chúng ta sẽ phải xem xét kỹ các hướng đi để có thể xây dựng hiệu quả. Việc xây dựng các hệ thống trợ lý ảo sẽ luôn đi kèm 2 vấn đề hay 2 khía cạnh chính đó là công nghệ sử dụng và phương pháp phát triển. Mặc dù về mặt công nghệ là quan trọng nhưng nó chưa phải là

điều kiện đủ để có thể tạo ra một chatbot đủ tốt. Nếu tập trung về vấn đề công nghệ, mà chatbot không được thiết kế tốt thì cho ra một chatbot không tốt, và ngược lại nếu chatbot được thiết kế tốt mà công nghệ không đáp ứng được thì chatbot cũng không thể sử dụng được. Do đó chúng ta cần nghiên cứu theo hướng kết hợp cả hai vấn đề này trong quá trình nghiên cứu xây dựng chatbot. Hiện nay, các công nghệ NLP AI đã được phát triển mạnh mẽ, điển hình như các mô hình Transformer ra đời, cùng với đó là hàng loạt các nghiên cứu các thuật toán tăng cường chất lượng hệ thống hỏi đáp. Vậy nên, về vấn đề công nghệ, có thể nói rằng đôi phần đã được giải quyết, còn lại khía cạnh phương pháp phát triển ứng dụng thì chưa được giải quyết. Vấn đề này là làm sao ứng dụng, chọn lựa các công nghệ hiện có để xây dựng hệ thống chatbot có tính ứng dụng cao để hỗ trợ các sinh viên và thầy cô trong việc giải đáp thắc mắc và khó khăn trong quá trình học tập.

1.3. Định hướng giải pháp

Như đã được nhắc đến ở mục 1.2, để xây dựng chatbot được tốt, chúng ta phải đảm bảo được chatbot của chúng ta đảm bảo được về mặt công nghệ lẫn cách thiết kế hệ thống sao cho tối ưu. Với đề án “Xây dựng hệ thống trả lời thông tin tự động cho trường Đại học Công nghệ GTVT sử dụng mô hình Transformer”, hệ thống trợ lý ảo của em sẽ là dạng bài toán có sử dụng các phương pháp truy xuất thông tin như Retrieval, hay Retrieval Augmented Generation kết hợp với nó, sử dụng các phương pháp sinh câu trả lời dựa vào câu hỏi của người dùng và các văn bản liên quan được truy xuất. Do đó với phương pháp này, hệ thống chatbot chỉ có thể trả lời theo dạng bị động, nghĩa là thay vì nó có thể hỏi lại các thông tin cần thiết từ người dùng, thì ở dạng hệ thống hỏi đáp, nó sẽ mặc định đưa ra phản hồi mỗi khi người dùng đưa một câu hỏi vào, không cần phải quan tâm câu hỏi của người dùng liệu có đủ thông tin để hỗ trợ hay không. Điều này sẽ rất khác so với các bài toán liên quan đến Conversation, do độ phức tạp của bài toán Conversation rất cao, hệ thống hỏi đáp phải xác định rằng liệu câu hỏi của người dùng liệu có bị thiếu không, do đó trong nhiều trường hợp điều đó là không thể để hệ thống có thể tự động xác định được, tuy nhiên để làm được điều đó, người phát triển phải định nghĩa các kịch bản có thể xảy ra, ví dụ như: “Nếu người dùng hỏi tìm Thông tin nhà trên đường Trường Chinh -> Hệ thống phản hồi, yêu cầu người dùng cung cấp số nhà, địa chỉ”. Tuy nhiên đối với đề tài của em, thì việc định nghĩa các kịch bản để xử lý được hết các trường hợp vừa nêu trên là gần như bất khả thi. Cùng với đó là lượng dữ liệu là thực

tế rất ít hoặc gần như không có để có thể đào tạo các mô hình AI để xác định câu hỏi của người dùng thuộc kịch bản nào.

Bên cạnh về việc sử dụng kịch bản như vừa nêu trên, gần đây các nhà nghiên cứu đã có thể xây dựng hệ thống chatbot theo dạng conversation (hội thoại) mà không cần đến các kịch bản, rule-based, ví dụ như các mô hình LLM như ChatGPT,.. Tuy nhiên nhược điểm của phương pháp này phải cần các mô hình LM đủ lớn, cùng với đó là input length (giới hạn đầu vào của mô hình ngôn ngữ, thường được tính bằng số tokens), do đó để có thể triển khai ứng dụng trong thực tế, sẽ phải cần lượng tài nguyên rất lớn để có thể phục vụ được, do đó xét về mặt tối ưu và thiết kế, nó hoàn toàn không phù hợp.

Vậy nên, điều chúng ta cần ở một hệ thống trợ lý ảo để phù hợp với bài toán của mình là khả năng tối ưu, gọn nhẹ nhưng vẫn phải đảm bảo được sự chính xác.

1.4. Giới hạn và phạm vi thực hiện

Tập trung nghiên cứu và làm rõ về các công nghệ sử dụng, song song đó là phương pháp phát triển hệ thống hỏi đáp.

Xây dựng được hệ thống hỏi đáp có tính ứng dụng thực tế có khả năng giải đáp thắc mắc về nội quy, quy chế, hỗ trợ các câu hỏi về công tác sinh viên, thông tin nhà trường. Bên cạnh đó hỗ trợ trò chuyện căn bản như chào hỏi.

Hệ thống được xây dựng sử dụng các mô hình và dữ liệu có hỗ trợ tiếng việt, các mô hình được sử dụng là các mô hình transformer (bao gồm cả dạng encoder-only và decoder-only). Đặc biệt, sử dụng các mô hình gọn nhẹ để phù hợp với nguồn tài nguyên hiện có (vì hiện tại có rất nhiều mô hình ngôn ngữ lớn nhưng chỉ có thể chạy được trên các máy tính cấu hình rất cao).

1.5. Phương pháp nghiên cứu

Về mặt lý thuyết:

- Các tài liệu, báo cáo về đề tài hệ thống hỏi đáp, trợ lý ảo.
- Các tài liệu về phương pháp học máy, học sâu và mô hình transformer.
- Các tài liệu liên quan đến công cụ hỗ trợ được sử dụng.

Về mặt thực nghiệm:

- Huấn luyện và đánh giá các mô hình Deep learning
- Xây dựng hệ thống
- Đánh giá toàn bộ hệ thống

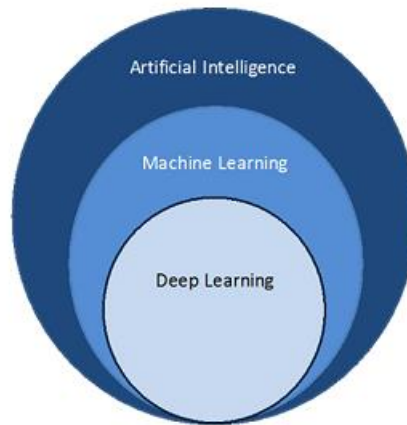
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

Trong chương trước, chúng ta có thể thấy rằng các chức năng hay thực chất là phạm vi kiến thức mà hệ thống hỏi đáp có thể hỗ trợ được. Tuy nhiên để xây dựng được một hệ thống hỏi đáp có thể thực hiện được những điều đó, chúng ta cần phải hiểu nên xây dựng nó như thế nào hay nên bắt đầu từ đâu ? Để trả lời trong câu hỏi này, trong chương này, em sẽ giới thiệu về những kiến thức tổng quan về trí tuệ nhân tạo, học sâu, mô hình transformer và hệ thống hỏi đáp. Từ đó, chúng ta có thể hiểu rõ những yếu tố cơ bản, và cách xây dựng hệ thống chatbot như thế nào. Bên cạnh đó, em giới thiệu các công cụ hỗ trợ xây dựng các mô hình deeplearning nói chung và hệ thống hỏi đáp nói riêng.

2.1. Tổng quan về trí tuệ nhân tạo, học máy, học sâu

2.1.1. Khái niệm

Trí tuệ nhân tạo (AI) là lĩnh vực khoa học máy tính chuyên giải quyết các vấn đề nhận thức thường liên quan đến trí tuệ con người, chẳng hạn như học tập, sáng tạo và nhận diện hình ảnh. Các tổ chức hiện đại thu thập vô số dữ liệu từ nhiều nguồn khác nhau như cảm biến thông minh, nội dung do con người tạo, công cụ giám sát và nhật ký hệ thống. Mục tiêu của AI là tạo ra các hệ thống tự học có thể tìm ra ý nghĩa của dữ liệu. Sau đó, AI áp dụng kiến thức thu được để giải quyết các vấn đề mới theo cách giống như con người. Ví dụ: công nghệ AI có thể trả lời cuộc trò chuyện với con người một cách hợp lý, tạo hình ảnh và văn bản gốc cũng như đưa ra quyết định dựa trên đầu vào dữ liệu theo thời gian thực. Tổ chức bạn có thể tích hợp tính năng AI vào ứng dụng để tối ưu hóa quy trình kinh doanh, nâng cao trải nghiệm khách hàng và đẩy mạnh quá trình đổi mới.



Hình 2.1: Biểu diễn miền của trí tuệ nhân tạo > học máy > học sâu..

Máy học là môn khoa học nhằm phát triển những thuật toán và mô hình thống kê mà các hệ thống máy tính sử dụng để thực hiện các tác vụ dựa vào khuôn mẫu và suy luận mà không cần hướng dẫn cụ thể. Các hệ thống máy tính sử dụng thuật toán máy học để xử lý khối lượng lớn dữ liệu trong quá khứ và xác định các khuôn mẫu dữ liệu. Việc này cho phép chúng dự đoán kết quả chính xác hơn từ cùng một tập dữ liệu đầu vào cho trước. Ví dụ: các nhà khoa học dữ liệu có thể đào tạo một ứng dụng y tế chẩn đoán ung thư từ ảnh chụp X-quang bằng cách lưu trữ hàng triệu ảnh quét và chẩn đoán tương ứng.

Học sâu là một phương thức trong lĩnh vực trí tuệ nhân tạo (AI), được sử dụng để dạy máy tính xử lý dữ liệu theo cách được lấy cảm hứng từ bộ não con người. Chúng phản ánh quá trình xử lý diễn ra trong não người. Bộ não chứa hàng triệu tế bào thần kinh hoạt động cùng nhau để xử lý và phân tích thông tin. Mạng nơ-ron học sâu sử dụng các tế bào thần kinh nhân tạo cùng nhau xử lý thông tin. Mỗi nơ-ron nhân tạo, hoặc nút, sử dụng các phép tính toán học để xử lý thông tin và giải quyết các vấn đề phức tạp. Phương pháp học sâu này có thể giải quyết các vấn đề hoặc tự động hóa các nhiệm vụ thường cần có trí thông minh của con người.

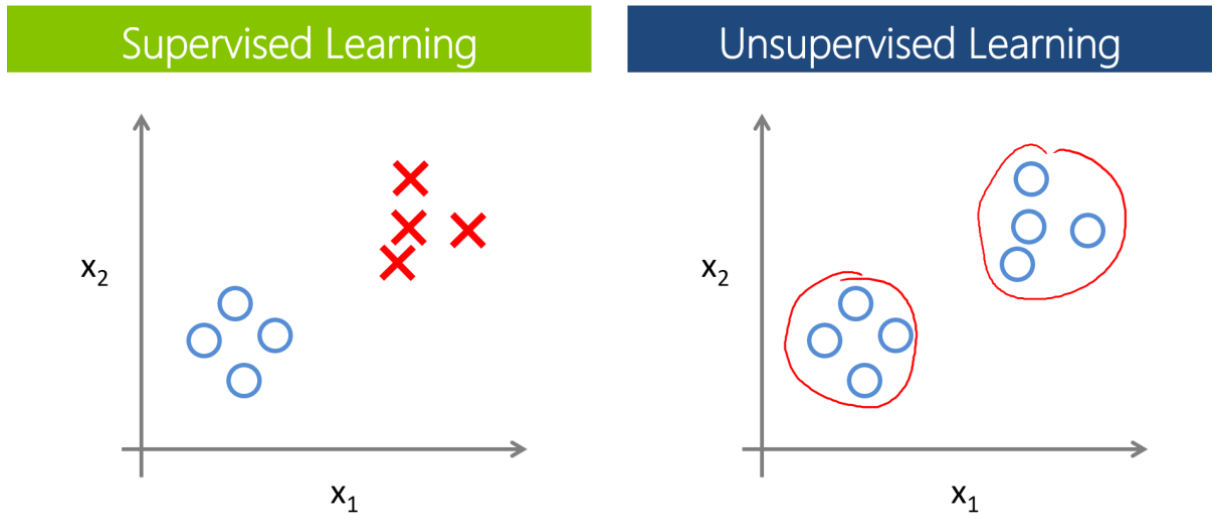
2.1.2. Các dạng bài toán

Hiện tại đối với các bài toán học máy nói chung, sẽ có 2 dạng bài toán chính thường gặp đó là học có giám sát, và không giám sát. Bên cạnh đó, còn có học bán giám sát, và học tăng cường. Tuy nhiên, học tăng cường không liên quan đến bất kỳ thuật toán học bài toán nào trong đề tài này nên sẽ không được đề cập đến.

Học có giám sát (supervised learning): Học có giám sát đã có thành công lớn trong các ứng dụng trong thực tế. Học có giám sát còn được gọi là học quy nạp trong học máy. Quá trình học này giống như việc con người học hỏi từ những kiến thức trong quá khứ để có được kiến thức mới nhằm cải thiện khả năng của chúng tôi trong việc xử lý các công việc thực tế. Tuy nhiên, vì máy móc không có “trải nghiệm”, buộc máy tính phải học những thông tin từ dữ liệu, dữ liệu được thu thập trong quá khứ và nó làm nhiệm vụ huấn luyện máy tính học các trải nghiệm này để áp dụng giải quyết các vấn đề trong thực tế.

Học không giám sát (unsupervised learning): Học không giám sát là một nhánh trong học máy. Loại hình học này sử dụng dữ liệu không được gắn nhãn như học có giám sát mà thay vào đó tập trung vào các đặc trưng của dữ liệu. Khi sử dụng phương pháp học không giám sát, chúng tôi không quan tâm đến kết quả mục tiêu kết quả đầu ra, vì mục tiêu của thuật toán là tìm kiếm các mối quan hệ trong dữ liệu và thực hiện nhóm các điểm dữ liệu chỉ dựa trên dữ liệu đầu vào. Học có giám sát liên quan đến dữ liệu được gắn nhãn để đưa ra dự đoán, nhưng học không giám sát thì không. Mục tiêu của học không giám sát là phân tích dữ liệu và tìm ra các đặc trưng quan trọng. Việc học không được giám sát thường sẽ tìm thấy các nhóm con hoặc các mẫu ẩn trong tập dữ liệu mà người quan sát là con người có thể không nhận ra.

Học bán giám sát (semi-supervised learning): Học bán giám sát như tên gọi cho biết nằm ở giữa hai trạng thái (học có giám sát là toàn bộ tập dữ liệu được gắn nhãn và học không giám sát là tập dữ liệu không có nhãn) về tính khả dụng của dữ liệu được gắn nhãn. Một nhiệm vụ học bán giám sát được đi kèm với một tập dữ liệu được gắn nhãn và không được gắn nhãn. Nó sử dụng dữ liệu không được gắn nhãn để hiểu thêm về cấu trúc dữ liệu. Học bán giám sát rất được quan tâm trong học máy và khai thác dữ liệu vì nó có thể sử dụng dữ liệu không gắn nhãn sẵn có để cải thiện các nhiệm vụ học có giám sát khi dữ liệu được gắn nhãn khan hiếm hoặc đắt tiền. Thông thường, học bán giám sát được thực hiện bằng cách sử dụng một tập dữ liệu có nhãn nhỏ và một tập dữ liệu không được gắn nhãn tương đối lớn hơn. Mục đích là để tìm hiểu một dự đoán dữ liệu kiểm tra trong tương lai tốt hơn so với dự đoán được học từ dữ liệu đào tạo được gắn nhãn



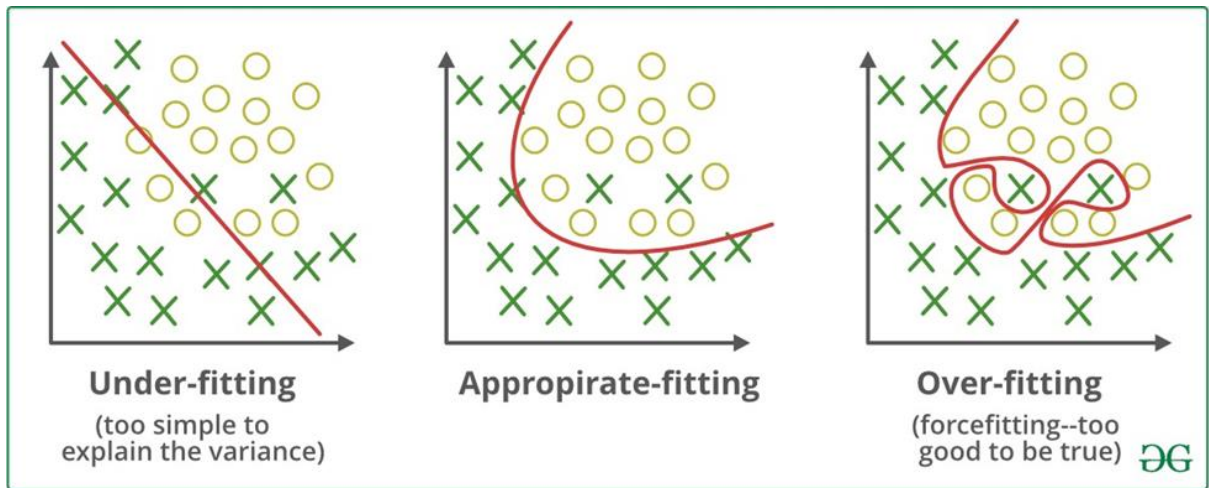
Hình 2.2: So sánh giữa supervised learning và unsupervised learning.

Hình 2.2: Với supervised, có thể hiểu đơn giản, mô hình học máy sẽ “học” để cố gắng đưa output gần sát với những nhãn (label) đã được set trước đó. Còn unsupervised, thường sẽ là các bài toán phân cụm dựa vào các tính chất của dữ liệu. Tuy nhiên, trong nhiều trường hợp cũng có thể coi là unsupervised nếu như phương pháp đó không yêu cầu phải thiết kế nhãn như các bài toán denoising,...

Thực tế, đa số các bài toán hiện nay đều sử dụng phương pháp supervised, nhưng các nghiên cứu hiện nay không ngừng cải tiến, tìm cách áp dụng các phương pháp unsupervised để tiết kiệm chi phí đánh nhãn. Lưu ý rằng, các phương pháp pre-training của các mô hình ngôn ngữ hiện nay (Language Model - LM) đều là unsupervised.

2.1.3. Overfitting và Underfitting

Đặc biệt, khi thực hiện các bài toán học máy, sẽ gặp 2 vấn đề lớn đó là overfitting và underfitting. Hai vấn đề này xảy ra khiến cho kết quả của mô hình không tốt, thậm chí là không thể sử dụng được khi đưa lên sản phẩm. Nguyên nhân xảy ra cũng có thể từ nhiều nguyên nhân khác nhau, có thể đến từ chất và lượng dữ liệu sử dụng hoặc từ phía mô hình.



Hình 2.3: Các loại khớp dữ liệu có thể có trong mô hình

Hình 2.3 có ba tình huống phổ biến trong học máy: under-fitting, appropriate-fitting và overfitting. Under-fitting là dữ liệu và mô hình không được khớp với nhau được biểu thị bằng đường màu đỏ và giá trị dự đoán thiếu độ chính xác. Appropriate Fitting là dữ liệu và mô hình có sự phù hợp tương thích, cũng được hiển thị bằng đường cong màu đỏ, các giá trị dự đoán khá chính xác trên dữ liệu mới. Trong trường hợp. Overfitting thì mô hình phù hợp với toàn bộ dữ liệu mà không sai bất kỳ một giá trị nào, nó khớp với dữ liệu này, nhưng khi dữ liệu mới được thêm vào, nó không hoạt động tốt nữa.

Overfitting: Vấn đề mô hình quá khớp với dữ liệu là khi mô hình cố gắng học để phù hợp với dữ liệu huấn luyện nhất. Điều này thường được thực hiện bằng cách ép cho mô hình học đi học lại trên tập dữ liệu đã huấn luyện trước đó để trở thành mô hình cho hiệu suất dự đoán, phân loại tốt nhất. Mô hình này tiến hành các nhiệm vụ trên dữ liệu mới thì hiệu suất của mô hình là cực kì kém.

Underfitting: ngược lại với overfitting. Điều này xảy ra khi mô hình quá đơn giản để tìm hiểu cấu trúc cơ bản của dữ liệu. Điều này thường dẫn đến phương sai cao không giải thích được vì mô hình không thể giải thích phương sai.

2.1.4. Quá trình hình thành mô hình machine learning

Trên thực tế, để phát triển một hệ thống AI tương đối là phức tạp thậm chí là rất khó. Thứ nhất, về quá trình phát triển một hệ thống AI như hình 2.4 dưới đây, trung bình mỗi một mô đun trong toàn bộ hệ thống AI sẽ phải trải qua từ bước:

Chuẩn bị dữ liệu → Xây dựng mô hình → Huấn luyện mô hình → Triển khai sản phẩm.



Hình 2.4: Quá trình phát triển một hệ thống AI nói chung

Các bước trên sẽ được lặp đi lặp lại nhiều lần cho đến khi mô hình đủ tốt, lúc đó mới có thể Triển khai sản phẩm, tuy nhiên trên thực tế, khi ứng dụng được nhiều người sử dụng, có nghĩa là sẽ có thêm một lượng dữ liệu lớn hoặc lượng phản hồi đổ về và chúng ta lại phải tiếp tục cải tiến để có thể giữ chân được khách hàng của mình. Do đó độ quá trình phát triển cũng sẽ tăng theo khi một hệ thống AI sử dụng nhiều mô đun hay mô hình con.

Bên cạnh đó, việc phát triển còn phụ thuộc các yếu tố như tài nguyên (phần cứng máy tính), quy mô, lượng dữ liệu có sẵn được cung cấp, v.v...

2.1.5. Thách thức và những lợi ích của AI mang lại

Ứng dụng AI trong hầu hết mọi lĩnh vực: Cuộc cách mạng 4.0 hiện nay đã cho thấy công nghệ nhanh chóng trở nên thông minh hơn và mạnh hơn, nhỏ hơn, nhẹ hơn và rẻ hơn. Các công nghệ này bao gồm phần cứng của các thiết bị: robot vật lý, máy bay không người lái và phương tiện tự trị và các thành phần của chúng (ví dụ: bộ xử lý, cảm biến, camera, chip). Công nghệ có thể là mã hoặc phần mềm như: phần mềm Analytics, phần mềm xử lý giọng nói, phần mềm sinh trắc học, thực tế ảo, thực tế tăng cường, công nghệ đám mây, công nghệ di động, gắn thẻ địa lý, nền tảng mã thấp, tự động hóa quá trình robot (RPA) và học máy. AI sẽ làm thay đổi tất cả các ngành trong tất cả các lĩnh vực dịch vụ và tự động hóa như:

Trong quản lý nhân sự: AI kiểm soát quá trình lao động tại nơi làm việc, cho phép tối đa hóa năng suất và hiệu quả của con người tại nơi làm việc. Theo một báo cáo gần đây, hầu hết các tổ chức lớn đã xây dựng khả năng của AI, 70% giám đốc điều hành xem ứng dụng AI ưu tiên hàng đầu. Năm 2020 là một năm thuận lợi để các công ty bán phần mềm giám sát nhân viên. Ví dụ, giải pháp “Công nhân Maximo” của IBM sử dụng AI để xử lý dữ liệu thời gian thực từ “camera, tín hiệu Bluetooth, điện thoại di động,

thiết bị đeo có kết nối IoT và cảm biến môi trường” cho phép người quản lý giám sát nhân viên hiệu quả. Ngoài ra, AI hứa cho phép dự đoán và ngăn chặn sự vi phạm của trật tự được thiết lập trước khi chúng xảy ra.

Trong Doanh nghiệp: Theo Alibaba, Chatbot AI đã giúp họ giảm tới 90% truy vấn đối với khách hàng và phục vụ hơn 3,5 triệu người dùng mỗi ngày. Trong một nghiên cứu gần đây liên quan đến 1.500 công ty trong 12 ngành công nghiệp, đã cho thấy các tổ chức có thể đạt được những cải tiến hiệu suất đáng kể nhất khi con người và máy móc làm việc cùng nhau.

Trong y tế: Liên quan đến phát hiện ung thư, trong hình ảnh của các tế bào hạch bạch huyết, Wang et al. (2016) phát hiện ra rằng một cách tiếp cận con người kết hợp với AI sẽ vượt trội hơn so với các quyết định chỉ có con người duy nhất. Tỷ lệ lỗi 0,5% khi kết hợp AI và con người trong việc ra quyết định, giảm tỷ lệ lỗi ít nhất ở mức 85% so với các phương pháp tiếp cận chỉ có con người và chỉ có AI.

Trong các ngành dịch vụ và tự động hóa: Robot dịch vụ có thể phân tích khối lượng dữ liệu lớn, tích hợp thông tin nội bộ và bên ngoài, nhận dạng các mẫu liên quan đến các hồ sơ khách hàng. Trong vài phút, các robot này có thể đề xuất các giải pháp phù hợp nhất và đưa ra khuyến nghị. Các đội người-robot ngày càng mang lại các nhiệm vụ đòi hỏi kỹ năng nhận thức và cảm xúc cao

Trong tài chính ngân hàng: AI có thể giúp giám sát dữ liệu theo thời gian thực, phát hiện những điều kỳ lạ để điều tra thêm và loại bỏ hoặc giảm sự xuất hiện của gian lận trong thanh toán do tội phạm mạng chuyên nghiệp thực hiện. Gian lận thanh toán hiện nay tinh vi hơn và thường vượt quá khả năng phát hiện của các hệ thống dựa trên quy tắc cũ. Ứng dụng AI cho phép chống gian lận, bảo mật gấp đôi so với giải pháp bảo mật thông thường do vậy lừa đảo được phát hiện gần ngay lập tức và thông tin thông báo sẽ được gửi đi kịp thời.

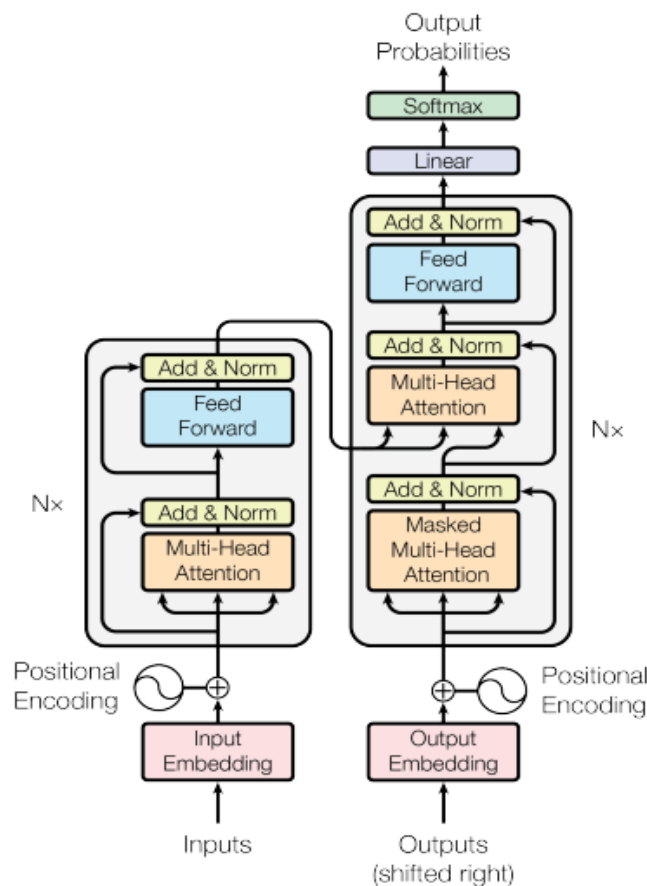
2.2. Tổng quan về mô hình transformer

Ngoài việc giới thiệu tổng quan về các mô hình transformer, tuy nhiên đến nay đã có rất nhiều mô hình được dựa trên và lấy cảm hứng từ mô hình transformer. Vậy nên nó trở thành ma trận trong việc lựa chọn cho các mô hình phù hợp với bài toán của mình, do đó em sẽ đưa ra những đánh giá và lý do để chọn lựa mô hình, tuy nhiên các mục sau sẽ được giải đáp chi tiết hơn.

Lưu ý rằng, các bài toán được sử dụng trong đề tài này đều thuộc deep learning. Đồng nghĩa với việc các mô hình ngôn ngữ được sử dụng cũng sẽ được coi là các mô hình deep learning.

2.2.1. Vanilla Transformer

Kể từ khi được google công bố, transformer đã trở thành một cú hích cho lĩnh vực xử lý ngôn ngữ tự nhiên, thậm chí tác động toàn bộ lĩnh vực học sâu. Khi mà không chỉ có mỗi các bài toán nlp mà ngay cả computer vision, time series đều sử dụng rất nhiều và thay thế hầu như các kiến trúc cũ (ví dụ: LSTM, RNN).



Hình 2.5: Kiến trúc mô hình transformer

Mô hình transformer có cấu trúc gồm nhiều lớp, mỗi lớp bao gồm hai thành phần chính là self-attention và mạng feedforward.

Self-Attention: là thành phần quan trọng nhất của mô hình transformer. Cơ chế này cho phép mô hình tập trung vào các thông tin quan trọng trong chuỗi đầu vào, bất kể vị trí của các thông tin đó. Self-attention được thực hiện bằng cách tính toán một hàm trọng

số cho mỗi token trong chuỗi đầu vào, hàm trọng số này phản ánh mức độ quan trọng của token đó đối với kết quả cuối cùng.

Mạng feedforward là một mạng nơ-ron truyền thẳng được sử dụng để xử lý thông tin đã được chú ý. Mạng này thường bao gồm một hoặc hai lớp nơ-ron, mỗi lớp có một số lượng nơ-ron cố định.

Ngoài hai thành phần chính này, mô hình transformer còn có một số thành phần phụ trợ khác, bao gồm:

- Embedding layer: Lớp này được sử dụng để chuyển đổi các token trong chuỗi đầu vào thành các vector đại diện.
- Positional encoding: Lớp này được sử dụng để thêm thông tin về vị trí của các token trong chuỗi đầu vào.
- Residual connection: Cấu trúc này giúp mô hình học hỏi hiệu quả hơn bằng cách kết nối đầu ra của một lớp với đầu vào của lớp đó.
- Layer normalization: Cấu trúc này giúp ổn định quá trình học tập của mô hình bằng cách chuẩn hóa đầu ra của mỗi lớp.

Dưới đây là mô tả chi tiết về tác dụng của từng thành phần cấu tạo của mô hình transformer:

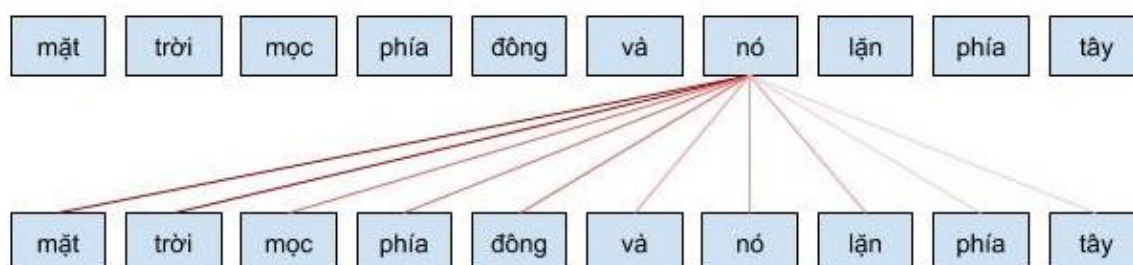
- Embedding layer là lớp đầu tiên trong mô hình transformer. Lớp này được sử dụng để chuyển đổi các token trong chuỗi đầu vào thành các vector đại diện. Các vector đại diện này thường có kích thước lớn hơn các token gốc, giúp mô hình học hỏi được nhiều thông tin hơn.
- Positional encoding là một lớp phụ trợ được thêm vào sau embedding layer. Lớp này được sử dụng để thêm thông tin về vị trí của các token trong chuỗi đầu vào. Thông tin này rất quan trọng đối với các tác vụ NLP, chẳng hạn như dịch máy, trong đó vị trí của các token có thể ảnh hưởng đến ý nghĩa của câu.
- Multi-head attention là thành phần chính của mô hình transformer. Lớp này được sử dụng để tính toán một hàm trọng số cho mỗi token trong chuỗi đầu vào, hàm trọng số này phản ánh mức độ quan trọng của token đó đối với kết quả cuối cùng.
- Mạng feedforward là thành phần thứ hai của mô hình transformer. Lớp này được sử dụng để xử lý thông tin đã được chú ý. Mạng này thường bao gồm một hoặc hai lớp nơ-ron, mỗi lớp có một số lượng nơ-ron cố định.

- Nhánh residual là một cấu trúc phụ trợ được thêm vào sau mỗi lớp transformer. Cấu trúc này giúp mô hình học hỏi hiệu quả hơn bằng cách kết nối đầu ra của một lớp với đầu vào của lớp đó.
- Layer normalization là một cấu trúc phụ trợ được thêm vào sau mỗi lớp transformer. Cấu trúc này giúp ổn định quá trình học tập của mô hình bằng cách chuẩn hóa đầu ra của mỗi lớp.

Mô hình transformer đã đạt được những thành tựu đáng kể trong các tác vụ NLP. Mô hình này đã vượt qua các mô hình dựa trên RNN trong nhiều tác vụ, chẳng hạn như dịch máy, tóm tắt văn bản và trả lời câu hỏi. Mô hình transformer hiện đang được sử dụng rộng rãi trong các ứng dụng NLP.

Điểm đặc biệt của kiến trúc này, đó chính là việc nó sử dụng các layer Self-attention. So với các kiến trúc trước như RNN hay LSTM, thì các kiến trúc này được dựa theo xác suất có điều kiện của từ phía trước, tuy nhiên cả LSTM và RNN đều tồn tại nhược điểm đó là vanishing gradient và gradient exploding khi input quá dài. Dẫn tới việc khó học và xảy ra tình trạng thông tin ở những trạng thái đầu tiên biến mất. Transformer thì ngược lại, nó vẫn giữ được các thông tin trong câu và làm giàu chúng bằng Self-attention.

Hiểu đơn giản, self-attention cho phép mô hình khi mã hóa một từ có thể sử dụng thông tin của những từ liên quan tới nó. Hay nói cách khác, self attention áp dụng để học biểu diễn của một từ dựa vào những từ xung quanh nó. Ví dụ như hình 2.6 dưới đây, từ nó được mã hóa, nó sẽ chú ý các từ liên quan như mặt trời.



Hình 2.6: Biểu diễn trực quan self-attention.

Do đó, thay vì dựa vào thông tin của các từ phía trước để biểu diễn từ tiếp theo, nghĩa là chỉ có một chiều (ngoài LSTM, còn có Bidirectional LSTM, hỗ trợ tính toán 2 chiều, tuy nhiên vẫn chỉ là tính toán tuần tự lần lượt), self-attention, biểu diễn ngữ nghĩa của

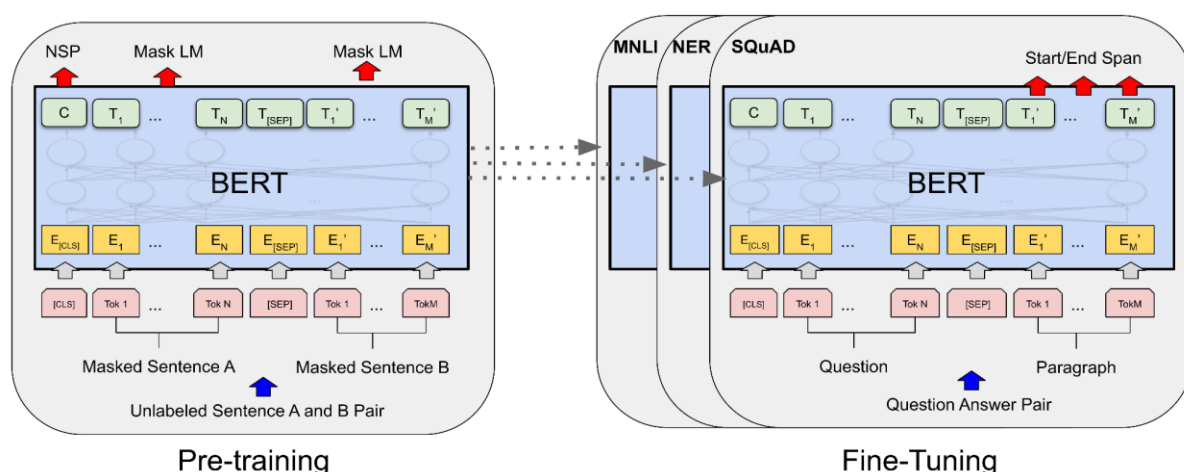
một từ bằng các từ xung quanh nó. Transformer đã cho thấy khả năng vượt trội so với những thuật toán trước là LSTM, RNN. Vì vậy, kiến trúc của mô hình này đã trở thành cơ sở để các nhà nghiên cứu phát triển dựa trên đó, đưa ra 2 loại kiến trúc khác, có nhiệm vụ và ưu thế riêng biệt. Phần này sẽ được làm rõ trong phần 2.2.2.

Ngoài self-attention làm điểm nhấn ra, kiến trúc của transformer cũng cho thấy có 2 phần chính, đó là encoder chịu trách nhiệm mã hóa và decoder chịu trách nhiệm

giải mã. Encoder thường sẽ được dùng để mã hóa đầu vào của chúng về vector phục vụ các tác vụ phân tích, đánh giá, phân loại, còn decoder ngược lại thường dùng cho các bài toán dạng sinh (generative) như sinh văn bản.

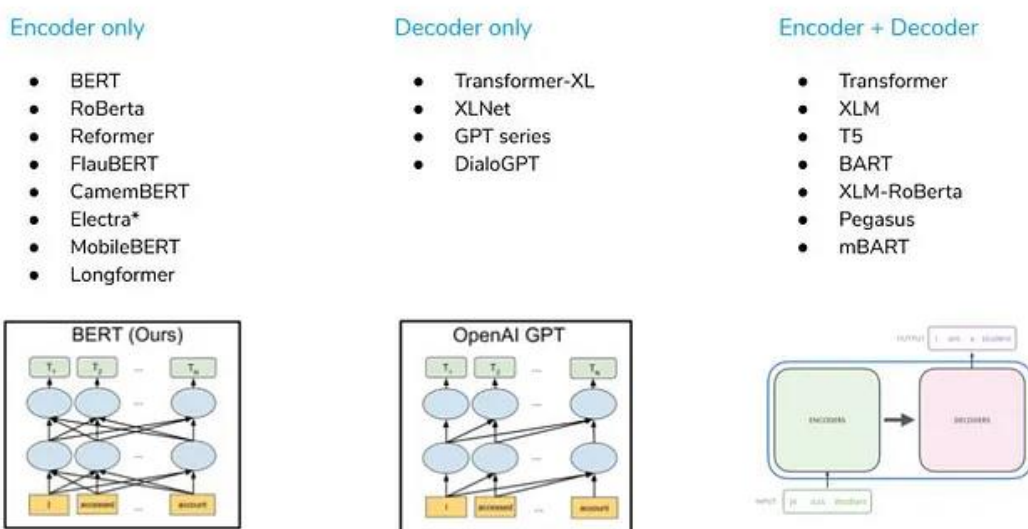
2.2.2. Các dạng kiến trúc tiêu biểu

Dựa vào ý tưởng độc đáo của transformer, các nhà nghiên cứu đã đưa ra dạng kiến trúc encoder-only và decoder-only. Tiêu biểu có thể kể đến là BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer). Mô hình encoder-only sẽ chỉ sử dụng kiến trúc encoder từ transformer, và ngược lại, decoder-only cũng sẽ chỉ kiến trúc decoder. Các mô hình này thường dùng để biểu diễn ngữ nghĩa bằng việc huấn luyện trên một tập dữ liệu text rất lớn với nhiệm vụ dự đoán từ bị che từ câu input, ví dụ “Tôi đi <mask>” => <mask>: học. BERT sau khi được pretraining, có thể sử dụng để tinh chỉnh cho các task cụ thể khác, ví dụ như các bài toán phân loại câu, phân loại từ,... Hiện tại đến nay cũng có rất nhiều phiên bản kế thừa từ BERT, đặc biệt có mô hình được huấn luyện trên tiếng Việt là PhoBERT. Mô hình này cũng sẽ được đề cập trong các mục sắp tới.



Hình 2.7: Mô hình BERT

Bên cạnh encoder-only, chúng ta có decoder-only và encoder-decoder. Sở dĩ có thể phân loại như này, bởi vì chúng ta có thể phân loại thành 2 dạng, Embedding representation (encoder - only) và Generative (decoder - only). Cả 2 kiến trúc này đều được huấn luyện dưới dạng sinh text. Đặc biệt, nổi bật với decoder-only, ChatGPT cũng là model sử dụng kiến trúc này, còn encoder-decoder, nổi bật nhất là các model T5.



Illustrations are from: <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html> and <http://jalammr.github.io/illustrated-transformer/>

Hình 2.8: Hình ảnh tên các mô hình pretrained.

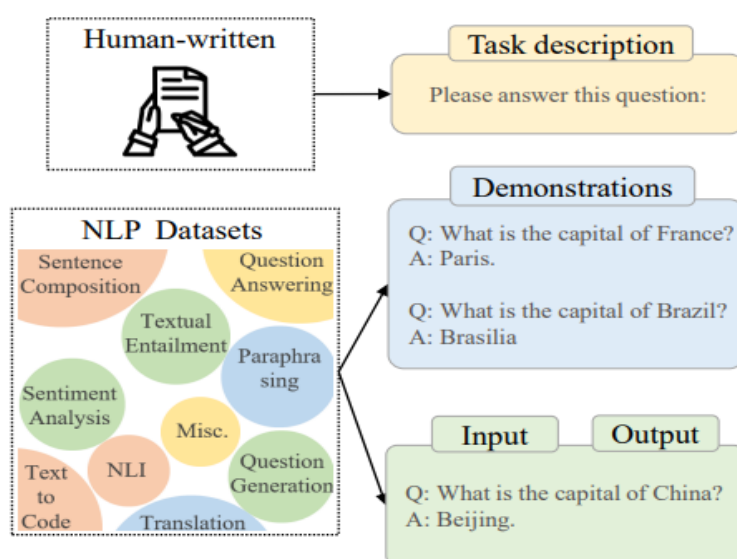
Gần đây, các mô hình generative đã cho thấy tiềm năng rất lớn khi có thể được sử dụng để tinh chỉnh thành những nhiệm vụ cụ thể nào đó, thay vì sử dụng các mô hình BERT. Thậm chí có khả năng hỗ trợ được tinh chỉnh cho nhiều nhiệm vụ cùng một lúc. Do đề tài của em là xây dựng hệ thống trợ lý ảo nên sẽ ưu tiên sử dụng mô đun QA Extractive với một mô hình sinh, ưu tiên có khả năng hỗ trợ tiếng việt, và đặc biệt mô hình phải đủ vừa dung lượng cũng không quá nhỏ và cũng không lớn. Tuy nhiên, ở phương pháp này, em sẽ đề cập sâu hơn trong các mục sau.

Các mô hình biến thể của transformer thường có những cải tiến và điều chỉnh nhất định để cải thiện hiệu suất, hiệu quả hoặc để giải quyết những vấn đề cụ thể trong việc xử lý ngôn ngữ tự nhiên. Một số điều chỉnh này có thể bao gồm cách thức áp dụng attention, cách xử lý context, cải thiện cấu trúc mạng nơ-ron hoặc việc huấn luyện dựa trên lượng dữ liệu lớn hơn.

Mặc dù có những cải tiến và điều chỉnh, nhưng với việc giữ nguyên các đặc điểm cốt lõi của Transformer, những mô hình này vẫn được coi là biến thể hoặc mở rộng của kiến trúc gốc và vẫn nằm trong họ mô hình Transformer trong lĩnh vực học sâu.

2.2.3. Generative Model

Trong mục trước, em đã giới thiệu tổng quan tới các loại mô hình transformer hiện tại, đặc biệt về mô hình sinh sẽ có 2 dạng kiến trúc là decoder-only và encoder-decoder. Tuy nhiên, vì hiện tại generative model, đang là một chủ đề lớn nên sẽ buộc phải tách ra làm một phần riêng. Vậy nên ở mục này, em sẽ đi qua về các vấn đề xoay quanh mô hình và việc tinh chỉnh chúng trước khi có thể đi sâu vào các mục tới do có mô đun sử dụng bài toán này.



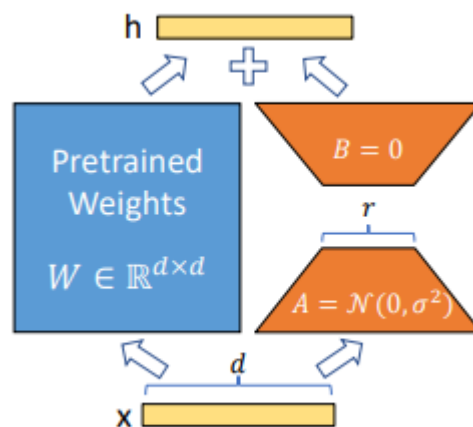
Hình 2.9: Formatting Task Datasets

Về kỹ thuật, hiện tại, các mô hình Generative AI, hay còn được gọi là Large Language Model (LLM) thường được sử dụng cho các bài toán liên quan đến sinh văn bản, tuy nhiên có thể biến đổi các bài toán thông thường như phân loại về dạng sinh text, vậy nên nó hoàn toàn có thể được sử dụng trong các bài toán multi-task training. Vậy nên bài toán QA (Question Answering) là không ngoại lệ. Bên cạnh đó, nhờ việc các mô hình được pretrained trên bộ dữ liệu rất lớn nên khả năng tinh chỉnh cho các nhiệm vụ cụ thể sẽ tốt hơn so với các mô hình như BERT.

Để có thể huấn luyện hoặc tinh chỉnh được mô hình sinh cho bất kỳ task nào, người ta cung cấp phương pháp gọi là instruction Tuning. Cụ thể, phương pháp này sẽ đưa các

sample về các cặp text, trong đó input sẽ bao gồm 2 phần chính được gọi “Task description” dùng để miêu tả task, và Input, ngoài ra sẽ có thể các ví dụ, được gọi là “Demonstrations” như hình 2.9. Ví dụ, với cặp câu hỏi trả lời “Thủ đô Việt Nam ở đâu?” - “Hà Nội”, kết hợp với task description như trên, có thể tạo ra một sample mới như: “Please answer the question: Thủ đô Việt Nam ở đâu” - “Hà Nội”. Các nghiên cứu chỉ ra rằng việc fine-tuned các data gắn nhãn mà không sử dụng task description kèm theo khiến cho kết quả mô hình giảm đi đáng kể.

Bên cạnh đó, việc fine tuning các LLM, đòi hỏi cần có tài nguyên đủ mạnh tùy thuộc vào kích thước của model, tuy nhiên thông thường người ta sẽ sử dụng một phương pháp là LoRA (Low - Rank Adaptation). Thay vì khi tinh chỉnh hoặc training model, tất cả các tham số sẽ phải cập nhật, LoRA sẽ cho thêm một lượng tham số vào mô hình, thường rất nhỏ, trong đó các tham số gốc của mô hình được frozen lại nên sẽ chỉ cập nhật trọng số cho tham số của LoRA. Điểm đặc biệt ở phương pháp này, cho kết quả gần xấp xỉ thậm chí là hơn trong một vài trường hợp so với việc fully fine-tuning.



Hình 2.10: Trục quan LoRA, khi training tham số A, B sẽ được cập nhật còn tham số của mô hình được giữ nguyên.

2.3. Tổng quan về hệ thống chatbot

2.3.1. Khái niệm

Ngôn ngữ là phương tiện để chúng ta giao tiếp trao đổi thông tin trong cuộc sống. Khi chúng ta làm việc, trò chuyện, hay bất kể một điều gì cần trao đổi chúng ta đều sử dụng ngôn ngữ. Những cuộc trao đổi, trò chuyện như vậy, chúng ta gọi là hội thoại (Dialogue). Và một hệ thống hội thoại (Dialogue System hay Conversational Agent) là một hệ thống có thể giao tiếp với người dùng thông qua ngôn ngữ tự nhiên. Hệ thống

hội thoại sẽ cung cấp một giao diện giữa người dùng và hệ thống để cho phép người dùng có thể tương tác với hệ thống bằng ngôn ngữ tự nhiên. Có hai hình thái của ngôn ngữ tự nhiên đó là dạng văn bản và dạng ngôn ngữ nói. Với dạng văn bản, hệ thống sẽ cung cấp giao diện trò chuyện như các ứng dụng nhắn tin (Facebook Messenger, Zalo, Telegram,...). Người dùng sẽ tương tác bằng cách gõ nội dung tin nhắn và gửi tới hệ thống, hệ thống sẽ xử lý và đưa ra phản hồi tương ứng. Với dạng ngôn ngữ nói, hệ thống sẽ cung cấp giao diện giao tiếp bằng tiếng nói (có thể thu và phát tiếng nói). Người dùng sẽ nói trực tiếp với hệ thống (thay vì phải gõ) và hệ thống sẽ thu nhận, xử lý và đưa ra phản hồi. Các hệ thống như vậy còn gọi là hệ thống hội thoại dùng ngôn ngữ nói (Spoken Dialogue Systems). Một số ví dụ về hệ thống hội thoại dùng ngôn ngữ nói như: loa thông minh Amazon Echo, Trợ lý ảo Siri hay robot Sophia. Như vậy, chúng ta thấy rằng đầu vào của các hệ thống hội thoại là dạng ngôn ngữ tự nhiên và đầu ra là các phản hồi tương ứng. Các phản hồi này có thể là một câu phản hồi cũng dưới dạng ngôn ngữ tự nhiên, một hành động, cử chỉ hay không phản hồi. Các hệ thống hội thoại được chia thành hai dạng chính là hệ thống hội thoại hướng tác vụ (Task-oriented Dialogue Agent) và Chatbot. Các hệ thống hội thoại hướng tác vụ được xây dựng với mục đích để giúp người dùng hoàn thành một số công việc đã được định nghĩa sẵn thông qua trao đổi, trò chuyện. Ví dụ hệ thống hội thoại trong một số trợ lý ảo (Siri, Alexa, Google Assistant, Cortana,...) có thể giúp người dùng thực hiện một số công việc như: tìm đường, gọi điện thoại, đặt báo thức hay mở nhạc và xem thời tiết,... Các hệ thống hội thoại cũng có thể trả lời các câu hỏi của người dùng như hỗ trợ trả lời các câu hỏi về một tổ chức doanh nghiệp trên website của họ hay hỗ trợ tư vấn chăm sóc khách hàng trong một sản phẩm, dịch vụ. Trái ngược với các hệ thống hướng tác vụ, chatbot là một hệ thống được thiết kế để có thể trò chuyện tự nhiên như con người. Một chatbot sẽ có đặc điểm là có thể mở rộng cuộc trò chuyện, các cuộc trò chuyện thường không có cấu trúc cụ thể, và không cần định nghĩa sẵn theo một logic cố định. Với đặc điểm này các chatbot thường được xây dựng với mục đích giải trí nhiều hơn so với các hệ thống hội thoại hướng tác vụ. Tuy nhiên chatbot cũng có thể đưa vào các hệ thống hướng tác vụ để giúp các nhiệm vụ trở nên tự nhiên hơn và không bị quá cứng nhắc như một người máy. Chính vì điều này, mà hiện nay người ta thường sử dụng thuật ngữ chatbot thay cho hệ thống hội thoại. Với cách sử dụng như vậy, chatbot là một hệ thống có thể hỗ trợ con người thực hiện các tác

vụ và cũng có thể trò chuyện tự nhiên với con người. Trong hệ thống hỏi đáp tư vấn học tập sinh viên, thuật ngữ chatbot cũng được sử dụng với ý hiểu tương tự.

Trong phần này đã giới thiệu tổng quan về hệ thống hội thoại bao gồm các khái niệm và cách dạng khác nhau của hệ thống hội thoại. Qua đây, chúng ta cũng hiểu hệ thống chatbot là một dạng của hệ thống hội thoại. Trong các phần tiếp theo chúng ta sẽ đi tìm hiểu về các phương pháp xây dựng hệ thống hội thoại cũng như hệ thống chatbot để hiểu được cách xây dựng chúng như thế nào.

2.3.2. Các phương pháp xây dựng chatbot

Hiện nay, có nhiều phương pháp khác nhau trong việc xây dựng hệ thống chatbot, tuy nhiên chúng được chia thành ba hướng tiếp cận chính là:

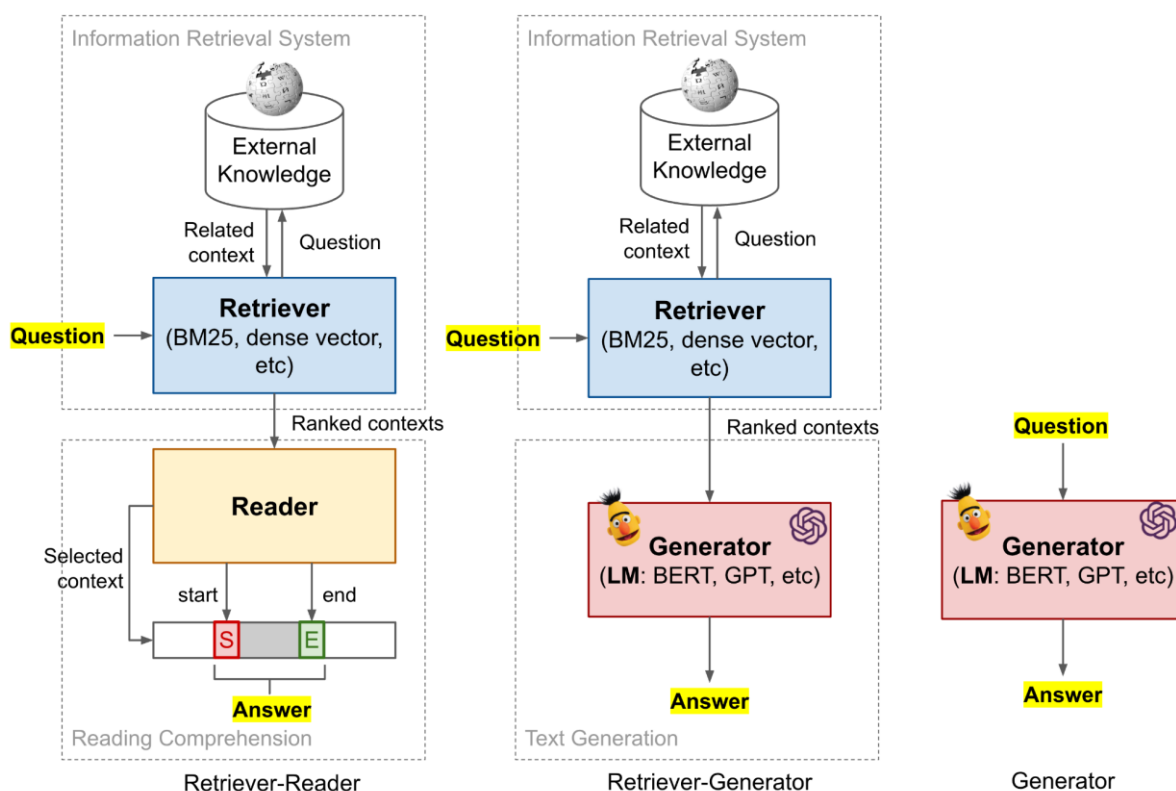
- Chatbot dựa trên luật (rule-based chatbot): Với chatbot dựa trên luật, hệ thống sẽ sử dụng tập luật để so khớp và đưa ra phản hồi tương ứng.
- Chatbot dựa trên kho ngữ liệu (corpus-based chatbot): chatbot dựa trên kho ngữ liệu là một hướng tiếp cận khác của chatbot. Các hệ thống chatbot dựa trên kho ngữ liệu sẽ khai thác lịch sử trò chuyện giữa con người với con người trong quá khứ và sử dụng các kỹ thuật truy xuất thông tin để sao chép các phản hồi tương đồng hoặc sử dụng các bộ mã hóa - giải mã (encoder-decoder) để sinh ra các câu phản hồi tới người dùng.
- Hệ thống hội thoại (dialogue system): Hệ thống hội thoại là hệ thống chatbot được xây dựng dựa trên kiến trúc trạng thái hội thoại.

2.3.3. Hệ thống hỏi đáp

Chatbot dựa trên luật và Hệ thống hội thoại là hai hướng tiếp cận rất phổ biến, tuy nhiên ở hướng tiếp cận dựa trên luật, các xây dựng hệ thống tuy đơn giản nhưng lại khó hoạt động hiệu quả và khó mở rộng do phụ thuộc vào tập luật. Còn đối với hướng tiếp cận dựa trên hệ thống hội thoại, nó sẽ có những ưu điểm rằng trong việc có thể đáp ứng xử lý được những tác vụ phức tạp cần hơn ít nhất một lần trò chuyện để hoàn thành công việc. Ví dụ, như việc đặt vé tàu hỏa, hệ thống sẽ phải hỏi lại các thông tin của người dùng và thông tin chuyến , đặt vé, thanh toán. Các nhiệm vụ như này, đối với các hệ thống chatbot trên luật và kho ngữ liệu sẽ khó có thể xử lý được. Tuy nhiên, nó vẫn có những nhược điểm chính như triển khai rất phức tạp, ngoài việc phải tạo dữ liệu cho các trường hợp, nhà phát triển phải tạo từng kịch bản cho các trường hợp, đồng thời việc phát triển cho các sản phẩm dịch vụ thực tế hướng tới trải nghiệm người dùng sẽ là rất

khó khăn vì nhà phát triển sẽ phải định nghĩa là rất nhiều kịch bản có thể xảy ra cho một tác vụ cụ thể. Bên cạnh đó, do đề tài của em là xây dựng hệ thống hỏi đáp tự động để chủ yếu truy xuất thông tin vậy nên sẽ không cần thiết phải sử dụng đến hệ thống hội thoại. Nếu sử dụng, với mỗi câu hỏi được đặt ra, sẽ phải định nghĩa từng kịch bản riêng cho từng câu hỏi, từng chủ đề nhỏ. Các hệ thống hỏi đáp sẽ gần giống với dạng chatbot dựa trên kho ngữ liệu, chúng chỉ cần được nhà phát triển, thiết kế và phát triển tốt các mô đun như trả lời câu hỏi và tìm kiếm các dữ liệu liên quan đến câu hỏi và có thể trả lời được. Từ đó có thể dễ dàng triển khai cho các hệ thống khác bằng việc thay thế kho ngữ liệu và các mô đun đi kèm theo nếu có. Tuy nhiên, nó vẫn không thể tự động trả lời hay phản hồi lại như cách mà hệ thống thoại được phát triển, điều này khiến cho việc giao tiếp không được tự nhiên. Nhược điểm này đã được giải quyết một phần, nhưng chỉ xuất hiện trên các mô hình ngôn ngữ rất lớn nên vấn đề này sẽ được loại bỏ.

Về kiến trúc, các hệ thống hỏi đáp sẽ có chung các thành phần như: mô đun trả lời câu hỏi, mô đun tìm kiếm ngữ liệu liên quan đến câu hỏi đó và database. Ngoài ra, trong một số hệ thống sẽ có những mô đun khác đi kèm theo như một bước tiền xử lý nào đó. Bên cạnh đó, sẽ có những hệ thống được gọi là end2end, khi mà nó chỉ cần câu hỏi đưa qua model Deep Learning sẽ lập tức có ngay câu trả lời.



Hình 2.11: Các hệ thống hỏi đáp

Hiện tại, sẽ có 3 dạng hệ thống hỏi đáp như hình 2.3.3 trên, tuy nhiên mỗi loại sẽ có những ưu nhược điểm riêng, vậy nên tùy theo từng bài toán và tài nguyên hiện tại có thể lựa chọn những hướng đi phát triển phù hợp.

Retriever-Reader và Retriever-Generator đều cùng chung một template đó là sử dụng mô đun retriever để tìm kiếm các dữ liệu văn bản liên quan đến câu hỏi mà có thể giúp cho mô hình đưa ra câu trả lời đúng cho câu hỏi đó. Reader và Generator, có thể hiểu là các mô đun trả về câu trả lời với input đầu vào là cặp (Question, Passages), chỉ khác là Reader sẽ là dạng trích xuất, đưa ra câu trả lời từ passage truyền vào là các vị trí bắt đầu đến vị trí kết thúc của câu trả lời. Còn Generator, output sẽ sinh ra là text có thể là dạng trích xuất từ passage truyền vào hoặc dạng suy luận, lúc này output sẽ không nằm trong passage. Về luồng chạy, câu hỏi của người dùng sẽ được đưa đến mô đun retrieval để tìm kiếm top-k các passage liên quan đến câu hỏi. Phương pháp tìm kiếm có thể sử dụng các thuật toán thống kê như BM25, TF-IDF hoặc phương pháp tìm kiếm dựa trên vector search. Sau khi có k passage hiện có, tiếp tục đưa qua mô đun ranker, để lấy cho ra kết quả tốt nhất. Cuối cùng đưa ra cặp passage có score liên quan nhất tới question và question, đi qua mô đun generator để sinh ra câu trả lời.

Generator E2E là phương pháp sử dụng trực tiếp các mô hình sinh, để đưa ra trực tiếp câu trả lời từ câu hỏi đầu vào. Tuy nhiên phương pháp này có điểm bất lợi, khi sử dụng các mô hình sinh cho các miền tri thức hẹp, sẽ dẫn đến việc mô hình trả lời sai khi nó bắt những câu hỏi chưa từng được nhìn thấy trong khi huấn luyện, cùng với việc tri thức bị hạn chế. Ví dụ, huấn luyện ChatGPT trả lời các câu hỏi về thông tin của trường A, tuy nhiên, tri thức của ChatGPT là general, nên những tri thức đi sâu về trường, thông tin trong trường, phòng ban, nội quy,.. là không có trong ChatGPT. Vậy nên kể cả trong trường hợp có tập dữ liệu câu hỏi về nhà trường, nhưng do sự hạn chế về dữ liệu ban đầu ChatGPT pretraining vẫn dẫn tới sai. Đồng thời, dữ liệu thông tin nhà trường, sẽ luôn luôn được thay đổi, cập nhật theo thời gian, vậy nên không thể nào áp dụng phương án, mỗi khi có dữ liệu mới cập nhật, không thể nào liên tục đưa mô hình vào huấn luyện liên tục được. Vậy nên, phương pháp này không có khả thi đối với đề án của em.

Tóm lại, chúng ta sẽ nhận thấy rằng phương pháp 2 là phù hợp nhất cho bài toán hiện tại, vừa có thể tận dụng được khả năng của các mô hình sinh, đồng thời có thể sử dụng ngay trên luồng dữ liệu liên tục được cập nhật, tiết kiệm được tài nguyên. Tiếp theo, em sẽ đi sâu hơn vào các mô đun của phương pháp 2 cũng như phương pháp mà em sử dụng cho đề án.

2.3.4. Retrieval

Như đã giới thiệu ở trên, thông thường một hệ thống hỏi đáp sẽ có retrieval, ranker và generator answer. Ở phần này, em sẽ giới thiệu và làm rõ mô đun retrieval, cũng như trình bày các thuật toán thường được sử dụng trong mô đun này. Bản chất của retrieval là các thuật toán tìm kiếm văn bản dựa theo độ tương đồng hoặc dựa theo các thuật toán xác suất thống kê như BM25 và TF-IDF.

TF-IDF (Term Frequency-Inverse Document Frequency) là một phương pháp trong xử lý ngôn ngữ tự nhiên và truy xuất thông tin để đánh giá tầm quan trọng của một từ hoặc một thuật ngữ trong một tài liệu hoặc một bộ sưu tập các tài liệu.

Tần số xuất hiện của từ (TF - Term Frequency): Đo lường tần suất xuất hiện của một từ trong một tài liệu. Cách phổ biến để tính TF là sử dụng số lần xuất hiện của từ trong văn bản chia cho tổng số từ trong văn bản đó.

Tần số nghịch đảo của tài liệu (IDF - Inverse Document Frequency): Đo lường tầm quan trọng của từng từ bằng cách đánh giá mức độ phổ biến của từ đó trong toàn bộ tập

các tài liệu. Công thức thường được sử dụng để tính IDF là logarit cơ số e của tổng số tài liệu chia cho số tài liệu chứa từ cần xem xét.

Công thức tổng quát để tính TF-IDF của một từ trong một tài liệu trong một bộ sưu tập các tài liệu:

$$TF\text{-}IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

Trong đó:

- $TF\text{-}IDF(t, d, D)$ là giá trị TF-IDF của từ t trong tài liệu d trong tổng số văn bản D .
- $TF(t, d)$ là tần số xuất hiện của từ t trong tài liệu d .
- $IDF(t, D)$ là tần số nghịch đảo của từ t trong tổng số văn bản D .

TF-IDF giúp nhận biết được sự quan trọng của một từ tài liệu so với các từ khác. Kỹ thuật này thường được sử dụng trong các hệ thống tìm kiếm, phân loại văn bản, gợi ý tài liệu, và nhiều ứng dụng khác liên quan đến xử lý ngôn ngữ tự nhiên và truy xuất thông tin.

BM25 (Best Matching 25) là một mô hình dùng để đánh giá sự tương quan giữa một truy vấn và các tài liệu trong các hệ thống tìm kiếm. Được phát triển từ mô hình BM11 (Best Matching 11), BM25 đã trở thành một trong những mô hình đánh giá hiệu suất tốt nhất trong truy xuất thông tin và hệ thống tìm kiếm văn bản.

Các điểm chính của BM25:

- Mô hình Tần suất Xuất hiện Điều chỉnh (TF-IDF điều chỉnh): BM25 được xây dựng trên cơ sở của mô hình TF-IDF nhưng có các cải tiến và điều chỉnh để cải thiện hiệu suất.
- Các tham số điều chỉnh: BM25 có thể được điều chỉnh thông qua các tham số như k , b , định dạng của hàm IDF để phù hợp với từng bộ dữ liệu cụ thể.
- Sự cân nhắc giữa tần số xuất hiện của 1 từ trong văn bản (TF) và Tần số nghịch đảo của 1 từ trong văn bản (IDF): Mô hình này không chỉ dựa vào tần suất xuất hiện của từ mà còn xem xét cách mà tần số xuất hiện này ảnh hưởng đến sự quan trọng của từ trong toàn bộ tập các tài liệu.

Công thức tính điểm BM25 của một từ trong một tài liệu cụ thể là:

$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdI}})},$$

Trong đó:

- Q là câu truy vấn
- d là tài liệu
- n là số lượng từ trong truy vấn
- q_i là từ thứ i trong truy vấn.
- $f(q_i, d)$ là tần số xuất hiện của từ q_i trong tài liệu d.
- k_1 và b là các tham số điều chỉnh.
- $\text{dl}(d)$ là độ dài của tài liệu d.
- avgdI là độ dài trung bình của tất cả các tài liệu trong bộ dữ liệu.

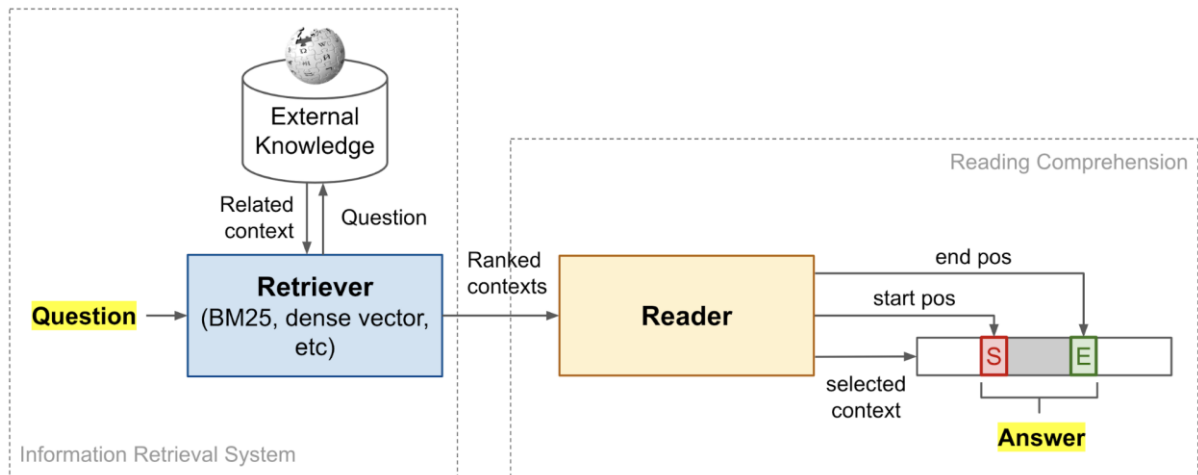
BM25 đã đạt được nhiều kết quả tốt trong nhiều trường hợp, tuy nhiên nó vẫn chỉ là mô hình đánh giá sự tương quan dựa vào thống kê vào số lượng từ. BM25 gặp khó khăn khi phải đối phó với vấn đề ngữ cảnh và cấu trúc của văn bản.

Vậy nên cần phải có các phương pháp trích xuất khác hiệu quả hơn, do đó nhờ sự phát triển của các mô hình transformer, đặc biệt là BERT, cho phép chúng ta vector hóa văn bản về chiều không gian vector để có thể so sánh.

Embedding Similarity Search là phương pháp giúp chúng ta có thể vector hóa một văn bản thông qua mạng neural, đưa nó về không gian vector để so sánh với các vector khác. Vector đó sẽ mang thông tin ngữ nghĩa của văn bản mà bạn đưa qua mô hình. Để tìm kiếm các vector liên quan đến chính nó, chúng ta có thể sử dụng các phép đo độ tương đồng giữa 2 vector như Euclidean, Cosine Similarity, hoặc các phương pháp khác. Lúc này các vector có độ tương đồng hay score cao nhất với vector mình xét sẽ có một phần hoặc toàn bộ thông tin ý nghĩa trùng với vector mình xét. Từ đó có thể lấy được các văn bản thực mà các vector tương đồng cao đại diện. Phương pháp này trong nhiều trường hợp rất khả quan trong các bài toán liên quan đến so sánh mức độ ngữ nghĩa của từ như: “Đi học” \Leftrightarrow “Nghỉ học”, tuy nhiên với các trường hợp câu ngắn, hoặc chỉ cần đơn giản đến Keyword search, các thuật toán thống kê có khả năng sẽ tốt hơn.

2.3.5. Generator Answer

Bên cạnh, sự quan trọng của các mô hình tìm kiếm văn bản thì không thể thiếu được sự quan trọng của các mô hình sinh câu trả lời. Thông thường với các nhiệm vụ Generator answer, chúng ta sẽ có 2 dạng bài toán question-answering chính, đó là Extractive-QA và Abstract QA.



Hình 2.12: Mô hình Reading Comprehension

Extractive QA là các bài toán trích xuất hay phát hiện các câu trả lời chính xác có trong đoạn văn, dựa vào yêu cầu trong câu hỏi. Bài toán này đã xuất hiện trong các mô hình BERT. Khi mà nhiệm vụ của chúng là phải predict được chính xác vị trí bắt đầu và vị trí kết thúc của câu trả lời có trong văn bản. Các bài toán tương tự như này thường được đặt tên là Machine Comprehension hoặc Reading Comprehension. Tuy nhiên, phương pháp này còn nhiều hạn chế:

- Khả năng hiểu ngữ cảnh và suy luận hạn chế: Các mô hình BERT thường sẽ không có khả năng hiểu ngữ cảnh rộng rãi, đặc biệt là không có khả năng suy luận. Do đó, nó có thể không hiểu được ngữ nghĩa sâu hơn của câu hỏi hoặc không cung cấp được những thông tin phức tạp hay suy luận từ nhiều nguồn thông tin.
- Vấn đề chính xác cú pháp và từ vựng: Có thể gặp khó khăn nếu câu hỏi hoặc thông tin trong nguồn dữ liệu không chính xác về cú pháp hoặc từ vựng, điều này có thể ảnh hưởng đến khả năng trích xuất thông tin đúng đắn.

- Khả năng không hiệu quả khi xử lý câu hỏi phức tạp: Trong trường hợp câu hỏi phức tạp hoặc đòi hỏi sự kết hợp thông tin từ nhiều nguồn, extractive qa có thể không hiệu quả và không cung cấp được câu trả lời chi tiết và đầy đủ.
- Khả năng trả lời có hạn trong văn bản dài: Có thể gặp khó khăn trong việc xử lý văn bản dài hoặc các đoạn văn dài với thông tin phân tán

Abstract QA là một phương pháp trả lời câu hỏi mà không chỉ đơn giản là trích xuất thông tin từ nguồn dữ liệu có sẵn mà còn có khả năng tổng hợp thông tin từ nhiều nguồn khác nhau và đưa ra câu trả lời ở dạng trừu tượng, tổng quát hơn.

Một số ưu điểm của Abstract QA so với Extractive QA:

- Khả năng hiểu ngữ cảnh và suy luận tốt hơn: Abstract QA có khả năng hiểu ngữ cảnh rộng rãi hơn và có thể áp dụng suy luận để đưa ra câu trả lời. Nó không chỉ trích xuất thông tin mà còn xử lý thông tin và suy nghĩ logic để tạo ra câu trả lời.
- Khả năng tổng hợp thông tin từ nhiều nguồn: Abstract QA có thể kết hợp thông tin từ nhiều nguồn khác nhau để tạo ra câu trả lời tổng quát và chi tiết hơn.
- Tạo ra câu trả lời mới: Abstract QA có thể tạo ra câu trả lời mới dựa trên việc tổng hợp thông tin, không chỉ dựa trên những gì có sẵn trong nguồn thông tin.
- Xử lý các loại câu hỏi phức tạp tốt hơn: Khi đối mặt với các câu hỏi phức tạp đòi hỏi sự kết hợp thông tin từ nhiều nguồn, Abstract QA có khả năng trả lời tốt hơn so với Extractive QA.

Tuy nhiên, cũng có nhược điểm:

- Khó khăn trong việc xử lý ngôn ngữ tự nhiên phức tạp: Abstract QA có thể gặp khó khăn trong việc xử lý ngữ cảnh và câu hỏi phức tạp, đặc biệt khi cần suy luận hoặc tổng hợp thông tin từ nhiều nguồn văn bản khác nhau.
- Yêu cầu dữ liệu lớn và đa dạng: Để huấn luyện một hệ thống Abstract QA hiệu quả, cần dữ liệu huấn luyện lớn và đa dạng, đồng thời cũng đòi hỏi các mô hình và thuật toán phức tạp hơn.

Abstract QA và Extractive QA đều có những ưu điểm và nhược điểm riêng. Việc lựa chọn phương pháp phù hợp phụ thuộc vào loại câu hỏi, yêu cầu của ứng dụng cụ thể, và nguồn dữ liệu đang xử lý. Trong nhiều trường hợp, việc kết hợp cả hai phương pháp này có thể tạo ra kết quả tốt nhất cho hệ thống QA.

2.4. Nền tảng phát triển

Dưới đây là các framework, thư viện mà em sử dụng cho việc hoàn thành đồ án. Tuy nhiên thực tế sẽ có rất nhiều các thư viện hoặc framework nhỏ hơn nhưng em sẽ chỉ tập trung và giới thiệu các thư viện, framework chính.

2.4.1. Pytorch



Hình 2.13: Thư viện PyTorch

PyTorch là một thư viện mã nguồn mở dùng cho việc xây dựng và huấn luyện mạng neural network (mạng nơ-ron) trong lĩnh vực Machine Learning và Deep Learning. Được phát triển bởi Facebook's AI Research lab (FAIR), PyTorch cung cấp một cách tiếp cận linh hoạt và mạnh mẽ cho việc tạo ra các mô hình máy học từ cơ bản đến nâng cao.

Dưới đây là một số điểm nổi bật của PyTorch:

- **Dynamic Computational Graphs:** Một trong những điểm mạnh của PyTorch là sử dụng đồ thị tính toán động (dynamic computational graphs). Điều này cho phép người dùng tạo và điều chỉnh các mô hình một cách linh hoạt hơn so với các thư viện khác. Đồ thị được xây dựng ngay khi mã Python chạy, giúp dễ dàng debug và điều chỉnh mô hình.
- **Hỗ trợ GPU mạnh mẽ:** PyTorch cung cấp hỗ trợ mạnh mẽ cho việc tính toán trên GPU, giúp tăng tốc quá trình huấn luyện mạng neural network và xử lý dữ liệu lớn.
- **Dễ học và sử dụng:** PyTorch có cú pháp Python rất gần gũi và dễ hiểu, giúp cho người dùng mới tiếp cận Deep Learning dễ dàng hơn. Cộng đồng người dùng lớn và tài liệu phong phú cũng hỗ trợ trong quá trình học và sử dụng PyTorch.
- **Module TorchScript:** TorchScript là một công cụ cho phép chuyển đổi mô hình PyTorch sang một dạng biểu diễn trung gian (intermediate representation) để có thể chạy được mô hình trên các môi trường không phải là Python.

- Cộng đồng lớn và sự phát triển liên tục: PyTorch có cộng đồng người dùng rộng lớn, đồng thời nhận được sự hỗ trợ và cập nhật liên tục từ Facebook và cộng đồng open source, giúp cải thiện, bổ sung tính năng và giải quyết các vấn đề nhanh chóng.
- Hỗ trợ cho nhiều loại ứng dụng: Không chỉ dành cho Deep Learning, PyTorch cũng hỗ trợ trong nhiều lĩnh vực khác như Computer Vision, Natural Language Processing (NLP), Reinforcement Learning và nhiều ứng dụng khác.

PyTorch là một trong những thư viện Deep Learning phổ biến và mạnh mẽ, được lựa chọn rộng rãi bởi cộng đồng do sự linh hoạt, hiệu suất và dễ sử dụng.

2.4.2. Transformers



Hình 2.14: Thư viện Transformers

Thư viện Transformers là một thư viện mã nguồn mở được phát triển bởi Hugging Face, chuyên về Natural Language Processing (NLP) trong lĩnh vực Deep Learning. Thư viện này tập trung vào việc phát triển và triển khai các mô hình transformer, đặc biệt là các mô hình Pre-trained Transformers đã đạt được nhiều thành tựu trong NLP.

Dưới đây là một số điểm nổi bật của thư viện Transformers:

- Cung cấp các mô hình Pre-trained state-of-the-art: Thư viện Transformers cung cấp các mô hình pre-trained với hiệu suất cao trên nhiều nhiệm vụ NLP khác nhau, bao gồm BERT, GPT (Generative Pre-trained Transformer), RoBERTa, DistilBERT, và nhiều mô hình transformer khác.
- Sử dụng đa dạng cho các nhiệm vụ NLP: Thư viện này hỗ trợ cho nhiều loại nhiệm vụ NLP như classification, text generation, language translation, question answering, và nhiều ứng dụng khác.
- Hỗ trợ nhiều ngôn ngữ và nguồn dữ liệu: Transformers có thể làm việc với nhiều ngôn ngữ và tập dữ liệu khác nhau, cung cấp sự linh hoạt trong việc xử lý dữ liệu đa dạng từ các nguồn khác nhau.

- Dễ sử dụng và tương thích với PyTorch và TensorFlow: Thư viện được xây dựng trên cả PyTorch và TensorFlow, giúp người dùng có thể sử dụng nó trên hai framework này một cách thuận tiện.
- Hỗ trợ community lớn và sự phát triển liên tục: Có một cộng đồng lớn đang sử dụng và đóng góp vào thư viện Transformers. Hugging Face thường xuyên cập nhật và bổ sung tính năng mới, cũng như cung cấp tài liệu hướng dẫn và ví dụ sử dụng đa dạng.
- Tích hợp các công cụ NLP mới nhất: Thư viện này không chỉ tập trung vào việc triển khai các mô hình hiện đại, mà còn liên tục cập nhật và tích hợp các công cụ và phương pháp mới nhất trong lĩnh vực NLP.

Thư viện Transformers đã trở thành một công cụ quan trọng cho các nhà nghiên cứu và nhà phát triển trong lĩnh vực NLP, giúp họ tiếp cận các mô hình hiệu suất cao và phát triển ứng dụng NLP một cách linh hoạt và hiệu quả.

2.4.3. ChromaDB



Hình 2.15: Thư viện Vector Database Chroma

ChromaDB là một thư viện Python và JavaScript cho phép tạo, truy vấn và tìm kiếm cơ sở dữ liệu nhúng với bộ nhớ. ChromaDB là một trong những cơ sở dữ liệu nhúng mã nguồn mở được thiết kế để tạo ra các ứng dụng AI với các nhúng.

ChromaDB có nhiều ưu điểm, bao gồm:

- Dễ sử dụng: ChromaDB được thiết kế để đơn giản hóa việc lưu trữ và truy xuất nhúng, vì vậy bạn có thể dễ dàng tích hợp nó vào các mô hình AI của mình.
- Tốc độ: ChromaDB được tối ưu hóa để lưu trữ và truy xuất các nhúng với tốc độ nhanh hơn so với các cơ sở dữ liệu khác 1.

- Tính linh hoạt: ChromaDB hỗ trợ nhiều tùy chọn lưu trữ khác nhau, cho phép bạn lưu trữ các nhúng trong bộ nhớ hoặc trên đĩa cứng 1.
- Mã nguồn mở: ChromaDB là một thư viện mã nguồn mở, vì vậy bạn có thể sử dụng nó miễn phí và tùy chỉnh nó theo nhu cầu của mình 1.

2.4.4. FastAPI



Hình 2.16: Thư viện xây dựng API FastAPI

FastAPI là một framework web nhanh và hiệu quả, được viết bằng Python, được sử dụng để xây dựng các ứng dụng API (Application Programming Interface) với hiệu suất cao. Nó được thiết kế để kết hợp các tính năng mạnh mẽ của Python với khả năng xử lý nhanh và mã nguồn mở.

Dưới đây là một số điểm nổi bật của FastAPI:

- Hiệu suất cao và tính năng dựa trên Starlette và Pydantic: FastAPI được xây dựng trên cơ sở của Starlette (một framework web cho ASGI) để tận dụng hiệu suất cao của ASGI (Asynchronous Server Gateway Interface). Nó cũng tích hợp với Pydantic để kiểm tra dữ liệu và xác định kiểu dữ liệu, giúp giảm thiểu lỗi và tạo ra mã ít lỗi hơn.
- Tự tạo tài liệu API tự động: FastAPI tự động tạo ra tài liệu API dựa trên mã nguồn thông qua Swagger UI và Redoc, giúp người dùng dễ dàng hiểu và tương tác với API thông qua giao diện người dùng.
- Hỗ trợ cho Python 3.7+ và kiến trúc dựa trên async/await: FastAPI sử dụng cú pháp mới async/await của Python để hỗ trợ xử lý không đồng bộ, giúp tối ưu hiệu suất và khả năng mở rộng.
- Đơn giản và dễ sử dụng: FastAPI cung cấp cú pháp Python rất gần gũi và dễ hiểu. Nó có thể tự động sinh mã API từ các hàm Python thông qua decorators.

- Hỗ trợ CORS (Cross-Origin Resource Sharing) và WebSockets: FastAPI hỗ trợ CORS để cho phép các request từ nguồn khác được gửi đến API. Ngoài ra, nó cũng hỗ trợ WebSockets để xây dựng các kết nối thời gian thực và tương tác hai chiều với ứng dụng.
- FastAPI đang trở thành một trong những lựa chọn phổ biến cho việc xây dựng các ứng dụng API với hiệu suất cao trong cộng đồng Python, nhờ vào tính năng mạnh mẽ, hiệu suất tốt và cú pháp dễ sử dụng của nó.

2.4.5. Gradio



Hình 2.17: Thư viện xây dựng giao diện Gradio

Gradio là một thư viện mã nguồn mở trong Python, được thiết kế để giúp người dùng xây dựng các giao diện người dùng (UI) tương tác cho các mô hình Machine Learning (ML) một cách nhanh chóng và dễ dàng. Nó cung cấp các công cụ để tạo ra các ứng dụng tương tác trực quan cho các mô hình ML mà không yêu cầu kiến thức sâu về front-end development.

Dưới đây là một số điểm nổi bật của Gradio:

- Tạo giao diện người dùng tương tác cho mô hình ML: Gradio cho phép người dùng tạo ra các giao diện người dùng đơn giản và tương tác với các mô hình ML chỉ trong vài dòng mã. Người dùng có thể tạo các widget để nhập dữ liệu và hiển thị kết quả trực tiếp từ mô hình.
- Hỗ trợ cho nhiều loại mô hình ML: Gradio hỗ trợ nhiều loại mô hình ML, bao gồm các mô hình học sâu (deep learning), machine learning truyền thống và các mô hình trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), xử lý hình ảnh, và dữ liệu dạng bảng.
- Giao diện tương thích với nhiều framework: Gradio hoạt động tốt với các framework phổ biến như TensorFlow, PyTorch, Keras, Scikit-learn và nhiều framework ML khác.

- Tích hợp dễ dàng và tùy chỉnh linh hoạt: Gradio cung cấp một API đơn giản cho việc tích hợp và tùy chỉnh giao diện người dùng. Người dùng có thể điều chỉnh cấu trúc, kiểu dữ liệu đầu vào và đầu ra, cũng như các thiết lập hiển thị để phù hợp với nhu cầu cụ thể của họ.
- Hỗ trợ cho việc chia sẻ và triển khai: Gradio cung cấp khả năng chia sẻ các ứng dụng đã tạo thông qua Gradio Share, một nền tảng trực tuyến cho phép người dùng chia sẻ và triển khai giao diện người dùng của họ một cách dễ dàng.

Gradio đã trở thành một công cụ hữu ích cho việc tạo ra các giao diện người dùng tương tác đơn giản và trực quan cho các mô hình ML, giúp tăng khả năng tiếp cận và sử dụng của các mô hình này đối với người dùng không có kiến thức chuyên sâu về Machine Learning.

CHƯƠNG 3: PHÂN TÍCH VÀ THIẾT KẾ KIẾN TRÚC HỆ THỐNG

Trong chương trước, em đã trước đã giới thiệu tổng quan về hệ thống chatbot và một số hướng tiếp cận để xây dựng hệ thống chatbot. Trong chương này em sẽ khảo sát, phân tích các chức năng và trình bày chi tiết kiến trúc hệ thống chatbot mà em thiết kế. Đầu tiên, em sẽ khảo sát và phân tích các chức năng, sau đó trình bày giới thiệu về kiến trúc tổng quan và làm rõ từng thành phần trong đó.

3.1. Khảo sát và phân tích các chức năng

3.1.1. Mục đích của khảo sát

Khảo sát này nhằm mục đích thu thập thông tin về nhu cầu, mong muốn sử dụng hệ thống trả lời câu hỏi tự động của sinh viên trường đại học Công nghệ Giao thông Vận tải. Dựa vào kết quả cuộc khảo sát này lên kế hoạch xây dựng và phát triển các chức năng của hệ thống trả lời câu hỏi. .

3.1.2. Nội dung khảo sát

Đối tượng khảo sát: Khảo sát này dành cho tất cả sinh viên đang theo học tại Trường Đại học Công nghệ Giao thông Vận tải.

Hình thức khảo sát: Khảo sát này được thực hiện dưới hình thức trực tuyến thông qua gg form:

Câu hỏi khảo sát:

- Bạn thường tra cứu thông tin trên trường, lớp như công tác sinh viên, quy định, quy chế đào tạo bằng cách nào ?
- Bạn gặp phải những khó khăn gì khi tìm kiếm các thông tin không?
- Bạn mong muốn chatbot có những tính năng gì?
- Bạn thật sự đã đủ nắm rõ các quy chế, quy định đào tạo, nội quy của trường?

Kết quả khảo sát: Khảo sát đã thu thập được tổng cộng 1000 phản hồi từ sinh viên Trường Đại học Công nghệ Giao thông Vận tải. Kết quả khảo sát cho thấy:

Các cách mà sinh viên tra cứu thông tin :

- Qua fanpage, các nền tảng mạng xã hội, trang chủ: 60%
- Thông qua từ nhiều nguồn như tham khảo, hỏi thông tin từ các thầy cô: 40%

Khó khăn mà sinh viên gặp phải khi tra cứu thông tin :

- Khó tìm kiếm, các thông tin quy chế đào tạo, thông tin công tác sinh viên: 70%
- Thông tin bị thiếu, không nhất quán giữa các nguồn thông tin: 30%

Nhu cầu về tính năng của chatbot: Sinh viên mong muốn chatbot có các tính năng sau:

- Trả lời chính xác các câu hỏi: 80%
- Hiểu ý của người dùng một cách chính xác: 80%
- Tương tác được với người dùng một cách tự nhiên: 60%

Số sinh viên nắm rõ các quy chế, quy định đào tạo, nội quy của trường:

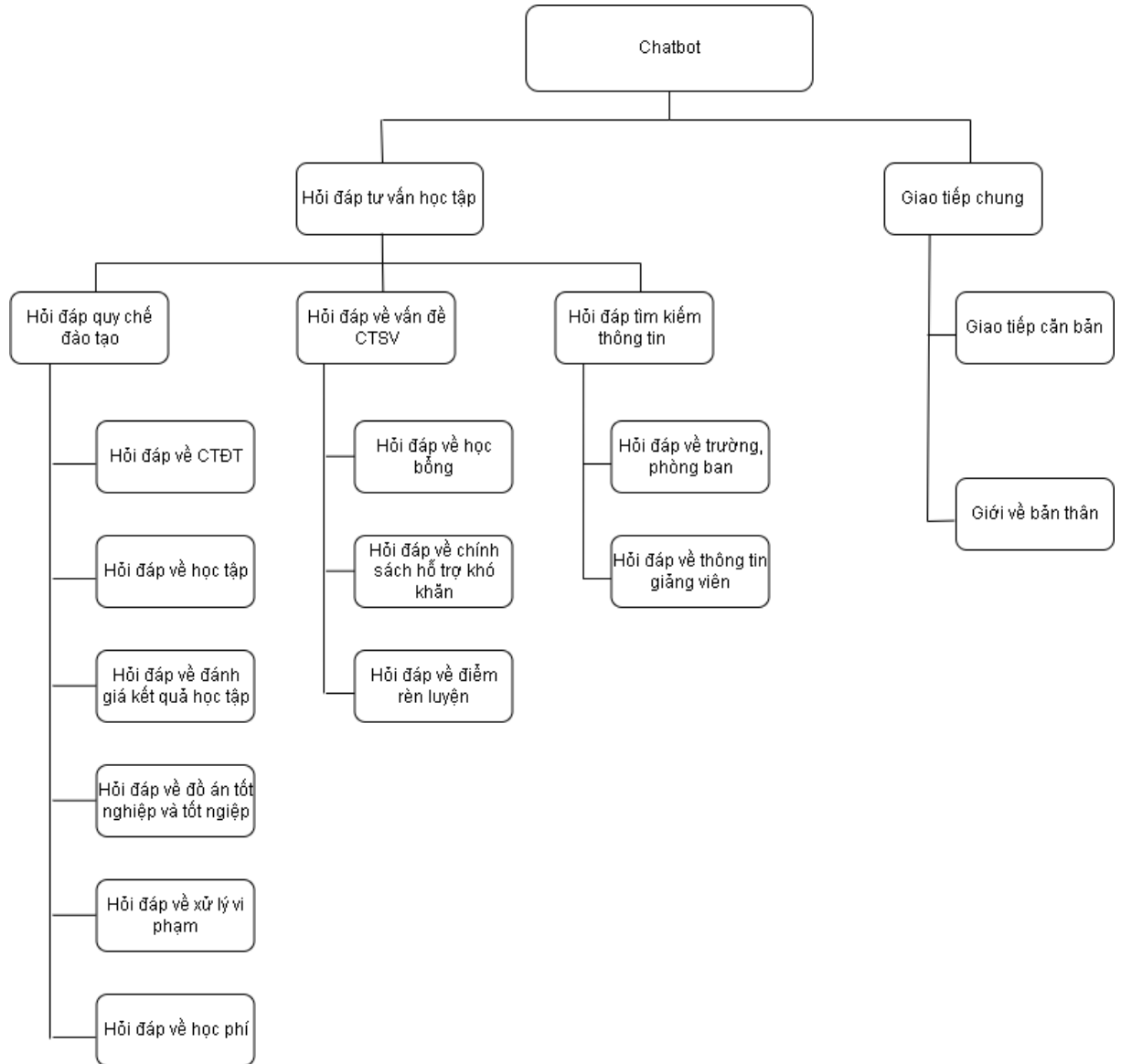
- Có: 20%
- Không: 80%

Kết luận: Kết quả khảo sát cho thấy sinh viên có nhu cầu sử dụng hệ thống trả lời câu hỏi tự động. Hệ thống này có thể giúp sinh viên học tập hiệu quả hơn bằng cách cung cấp các thông tin đa dạng và chính xác, trả lời các câu hỏi một cách nhanh chóng và chính xác, tương tác với sinh viên một cách tự nhiên

Dựa trên kết quả khảo sát, hệ thống trả lời câu hỏi tự động bằng mô hình transformer cần có các tính năng sau:

- Trả lời chính xác các câu hỏi: Đây là tính năng quan trọng nhất của hệ thống. Hệ thống cần được huấn luyện trên một bộ dữ liệu lớn và đa dạng để có thể trả lời các câu hỏi một cách chính xác.
- Hiểu ý của người dùng một cách chính xác: Hệ thống cần có khả năng hiểu ý của người dùng một cách chính xác, ngay cả khi câu hỏi của người dùng không rõ ràng hoặc khó hiểu.
- Tương tác được với người dùng một cách tự nhiên: Hệ thống cần có khả năng tương tác với người dùng một cách tự nhiên, giống như một cuộc trò chuyện giữa hai người.

3.1.4. Phân tích chức năng



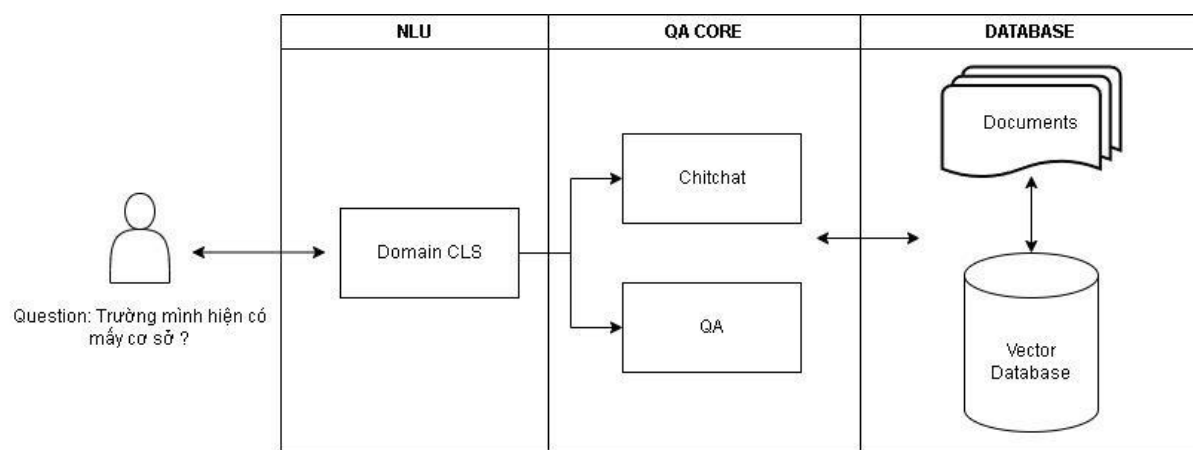
Hình 3.1: Biểu đồ phân cấp chức năng

Như hình 3.1, chúng ta thấy hệ thống trả lời tự động hay chatbot hỗ trợ chức năng chính là hỏi đáp các vấn đề học của sinh viên. Đây là các câu hỏi mà sinh viên thường gặp và có tính lặp lại trên nhiều sinh viên. Các vấn đề này được chia làm 3 nhóm chức năng hỏi đáp chính là: hỏi đáp về quy chế đào tạo, hỏi đáp về công tác sinh viên (CTSV) và các câu hỏi đáp để tìm kiếm thông tin về các phòng ban và giảng viên. Với dạng hỏi đáp về quy chế đào tạo, hệ thống sẽ có thể hỗ trợ sinh viên tìm kiếm các thông tin về các chương trình đào tạo (đối với các sinh viên năm nhất), trả lời các câu hỏi liên quan đến quá trình học tập, cách thức và quy định đánh giá kết quả học tập, giúp giải đáp các thắc mắc về Đồ án tốt nghiệp và tốt nghiệp. Và trong phần hỏi đáp về quy chế đào tạo, hệ thống còn có thể hỗ trợ sinh viên giải đáp các thắc mắc về xử lý vi phạm học tập và

các vấn đề về học phí. Với nhóm chức năng thứ hai, hỗ trợ hỏi đáp các vấn đề về công tác sinh viên, với nhóm này, hệ thống có thể giúp sinh viên nắm rõ các quy định và thủ tục về hành chính, thủ tục sinh viên và các vấn đề khác ngoài vấn đề đào tạo của sinh viên trong quá trình học tập tại trường như các vấn đề về: học bổng, hỗ trợ khó khăn sinh viên, điểm rèn luyện. Và cuối cùng là nhóm chức năng hỏi đáp để tìm kiếm thông tin. Chức năng này là chức năng bổ sung để sinh viên có thể tìm kiếm thông tin về các giảng viên, các phòng ban. Thông qua việc hỏi đáp với hệ thống chatbot, sinh viên có thể tìm kiếm các thông tin này một cách dễ dàng hơn. Ngoài ra để thêm tính thân thiện với người dùng, hệ thống còn có thể thực hiện những lời thoại trò chuyện căn bản với người dùng như: chào hỏi và tự giới thiệu về bản thân mình.

3.2. Kiến trúc hệ thống tổng quan

Như em đã trình bày ở mục 2.3.3, hệ thống hỏi đáp dựa trên miền ngữ liệu là một lý tưởng tốt để áp dụng và phát triển cho hệ thống hỏi đáp này. Nó phù hợp và dễ dàng triển khai hơn khi chỉ cần tối thiểu 2 mô đun quan trọng nhất là Retrieval và Generator QA để hoạt động. Đồng thời so với chatbot luật và hệ thống hội thoại, hệ thống của chúng ta dễ dàng triển khai so với việc phải phụ thuộc vào tập luật, và tránh tồn tại một tác vụ có thể phải viết nhiều lần cho phù hợp với các kịch bản đề ra, một hệ thống tồn tại việc với đằng sau mỗi dạng câu hỏi phải luôn đi liền với các kịch bản từ dễ đến phức tạp sẽ khiến cho hệ thống của chúng ta có thể phản hồi câu hỏi của người dùng không thể phong phú. Bên cạnh đó, khi đưa sản phẩm lên thực tế, chúng ta có thể có một lượng khách hàng tăng và khi số lượng tăng, các kịch bản mà chúng ta định nghĩa cũng sẽ phải sửa đổi lại để theo ý khách. Do đó, đối với các hệ thống hỏi đáp tương tự như của chúng ta sẽ không thật sự cần thiết khi sử dụng hệ thống đối thoại, nhưng vẫn không thể phủ nhận được sự thiếu sót về khả năng tự động phản hồi, hỏi khách hàng của hệ thống, điều này cũng khiến cho hệ thống của chúng ta trở nên thiếu tự nhiên hơn và nói cách khác, hệ thống sẽ chỉ hỗ trợ 1 turn 1, nghĩa là khi người dùng hỏi - bot trả lời, bot sẽ không tự động hỏi lại người. dùng. Vậy nên, ngoài việc yêu cầu hệ thống chính phải hoạt động ổn định, để đảm bảo được tăng sự tự nhiên ra, em sẽ thêm các mô đun hỗ trợ, các mô đun này như đã nói có thể có hoặc không, chi tiết về các mô đun này em sẽ trình bày trong các phần tiếp theo.



Hình 3.2: Kiến trúc tổng quan của hệ thống

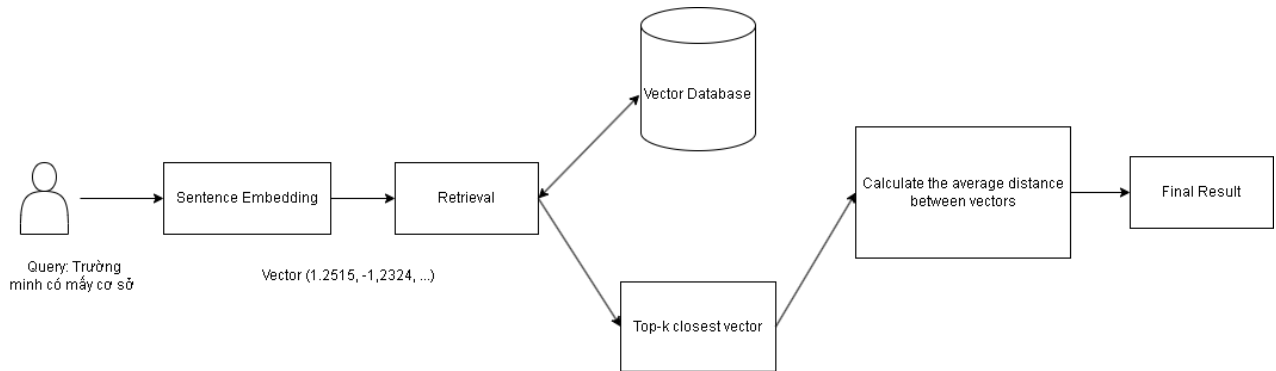
Hệ thống hỏi đáp sẽ gồm có 3 mô đun chính bao gồm NLU, QA Core và Database, mỗi mô đun sẽ đảm nhiệm nhiệm vụ như sau:

- **NLU:** có trách nhiệm phân loại các câu hỏi đầu vào của người dùng làm 3 loại, chitchat, qa, out of domain. Hệ thống sẽ tiếp tục xử lý cho mô đun tiếp theo nếu kết quả trả về là chitchat hoặc qa. Còn lại đối với out of domain, hệ thống sẽ tự động trả về kết quả không hỗ trợ tới người dùng, hoặc yêu cầu người dùng phải nhập lại câu hỏi phù hợp.
- **QA CORE:** bao gồm 2 mô đun chitchat và QA. Đối với mô đun chitchat, mô đun này xử lý theo dạng trả lời câu hỏi, tương tự y hệt mô đun QA, chỉ khác dữ liệu được dùng cho mô đun này sẽ được thiết kế khác, kết hợp với Domain-Classification, lúc này cho phép người dùng hỏi đáp, trích xuất thông tin đã được thiết kế đó. Còn QA, vẫn sẽ hoạt động đúng bản chất như các kiến trúc hỏi đáp đã được trình bày các mục phía trên, tuy nhiên mô đun này chỉ được thực hiện nếu câu hỏi của người dùng liên quan đến chủ đề trường, lớp,...
- **DATABASE:** bao gồm các database như tập dữ liệu document raw text đã được chia tách, dữ liệu đặc biệt vector database, 2 kiểu dữ liệu này luôn luôn mapping cho nhau, phục vụ cho việc tìm kiếm các văn bản tương đồng ngữ nghĩa với câu hỏi thông qua các thuật toán tìm kiếm, xét sự tương đồng giữa các vector.

Ngoài ra sẽ có các mô đun hỗ trợ khác nhưng chỉ hỗ trợ việc truyền tải, phản hồi sử dụng rule-based, ví dụ: người dùng nhập câu hỏi không được hỗ trợ, hệ thống yêu cầu người dùng nhập lại. Chi tiết phần này em sẽ trình bày trong các phần tiếp theo

3.2.1. Domain Classification

Mô đun Domain Classification giúp hệ thống phân loại được đâu là các câu hỏi có thể hỗ trợ hoặc không. Việc phân loại sẽ lấy ý tưởng từ việc tìm kiếm vector tương đồng rồi căn cứ theo số lượng để quyết định liệu query của người dùng có phải là câu hỏi phù hợp hay không.

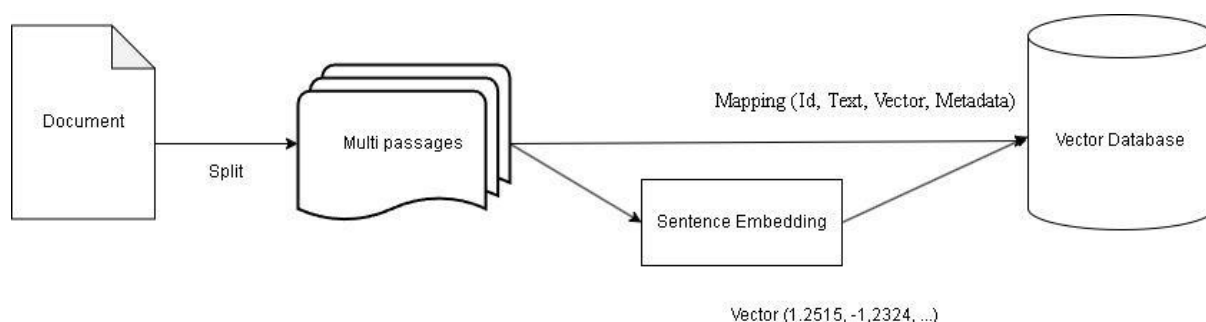


Hình 3.3: Domain Classification

Cụ thể, như hình 3.3 phía trên, câu hỏi của người dùng sẽ được vector hóa thông qua một model sentence embedding, model này bắt buộc phải được tinh chỉnh từ trước trên miền dữ liệu tiếng việt và phải đảm bảo tốt, nghĩa là khả năng phân tích ngữ nghĩa của model phải đủ tốt, nếu 2 câu trái nghĩa, khoảng cách giữa 2 vector này sẽ rất lớn, ngược lại nếu đồng nghĩa khoảng cách giữa 2 vector này sẽ rất nhỏ, thậm chí nếu trùng nhau, khoảng cách 2 vector sẽ bằng 0. Sau khi đã trích xuất được vector embedding xong, tiếp tục sử dụng phương pháp retrieval để thu được top-k các vector có liên quan mật thiết tới vector của câu query. Do bản chất bài toán phân loại thực chất là một bài toán gắn nhãn, phân nhóm cho một dataset không có nhãn trước đó, chúng ta sử dụng phương pháp tính toán trung bình khoảng giữa các vector để xác định liệu câu query người dùng đưa vào có thực sự là câu thuộc domain A hay không, nếu không câu query sẽ thuộc domain còn lại. Việc này yêu cầu phải thiết lập một ngưỡng phù hợp để quyết định cuối cùng xem liệu câu query có thuộc domain chitchat hoặc hỏi về trường không. Ưu điểm của phương pháp này là sự nhanh chóng, nhẹ và không yêu cầu các mô hình phân loại lớn. Bên cạnh đó, về hạn chế của dữ liệu cũng như tài nguyên hiện có nên em chỉ có thể tạo một lượng dữ liệu nhỏ cho domain chitchat, có thể dễ dàng đủ các trường hợp, còn domain hỏi đáp trường học lượng dữ liệu yêu cầu sẽ lớn hơn rất nhiều để có thể tránh việc nhầm lẫn giữa các domain với nhau. Vậy nên với phương pháp này, em chỉ cần xét các câu query có gần với các câu domain chitchat hay không, nếu không có thể coi nó

là domain hỏi đáp thông tin trường học, tuy nhiên chỉ phục vụ cho bài toán 2 nhãn, các câu không thuộc dạng hỏi đáp thông tin trường học và chitchat, có thể sẽ ngẫu nhiên được dự đoán một trong 2.

Vector database trong mô đun này, hoặc trong các mô đun khác sẽ được thiết kế khác nhau, và không ở chung trong một collection. Trong mô đun này hay các mô đun bất kỳ nào khác có sử dụng đến phương pháp search embedding, trước khi thực hiện truy xuất label cho query, phải bắt buộc nạp dữ liệu cho vector database trước đó thì mới có thể sử dụng được. Các vector database đã rất nổi tiếng và quen thuộc nhờ vào khả năng tích hợp các công cụ, thuật toán tìm kiếm cũng như tốc độ xử lý rất nhanh. .

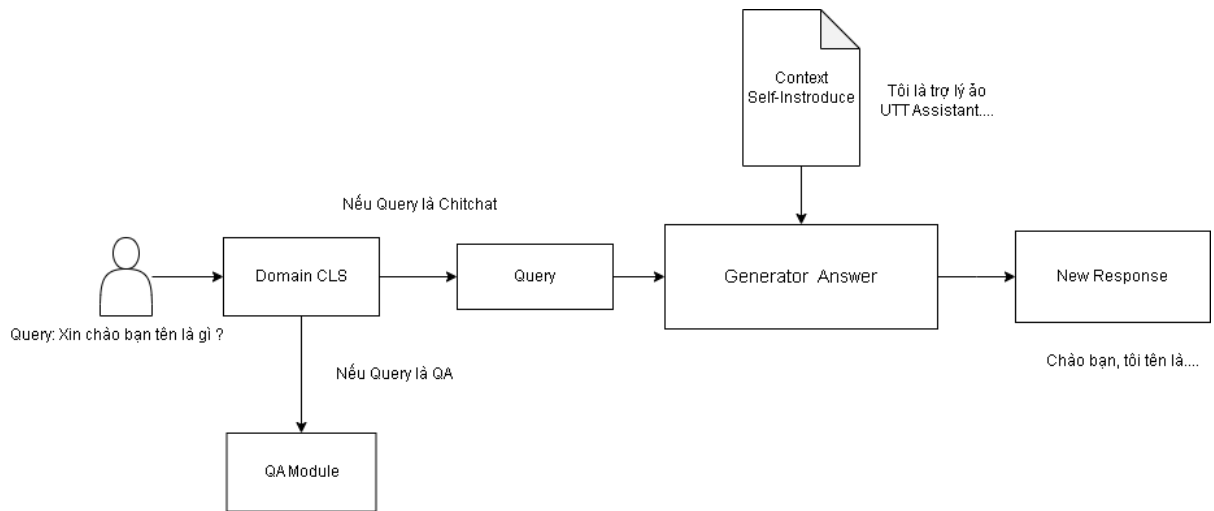


Hình 3.4: Nạp dữ liệu vào vector database

Trên thực tế, chúng ta có thể hoàn toàn không cần đến các vector database này do hoàn toàn có thể sử dụng nhiều công cụ hỗ trợ kết hợp, tuy nhiên để đảm bảo về tốc độ, độ chính xác và không gian lưu trữ nên em sẽ ưu tiên sử dụng database ChromaDB để tối ưu tốc độ, do đồng thời sẽ có các mô đun đặc biệt chiếm nhiều tài nguyên máy tính.

3.2.2. Chitchat

Module chitchat được em sử dụng nhằm tăng khả năng thiện cảm và giao tiếp tương tác giữa người dùng và chatbot. Tuy nhiên, với đề tài hiện tại xây dựng hệ thống hỏi đáp nên mô đun này em chỉ giới hạn trong khả năng chào hỏi và giới thiệu bản thân. Module chitchat sẽ sử dụng bản chất của các hệ thống trả lời câu hỏi hiện tại, đặc biệt là các hệ thống có sử dụng các mô hình ngôn ngữ lớn, giúp tăng sự đa dạng về câu từ phản hồi giúp tăng trải nghiệm người sử dụng. Tuy nhiên mô đun này sẽ không cần thiết phải sử dụng Database vector hỗ trợ. Sau khi query của người dùng đi vào và được phân loại thành domain chitchat, ngay lập tức mô đun generator answer sẽ tính toán và sinh câu trả lời dựa vào câu hỏi của người dùng, và tri thức tự sự giới thiệu bản thân mà em cài đặt.



Hình 3.5: Mô đun Chitchat

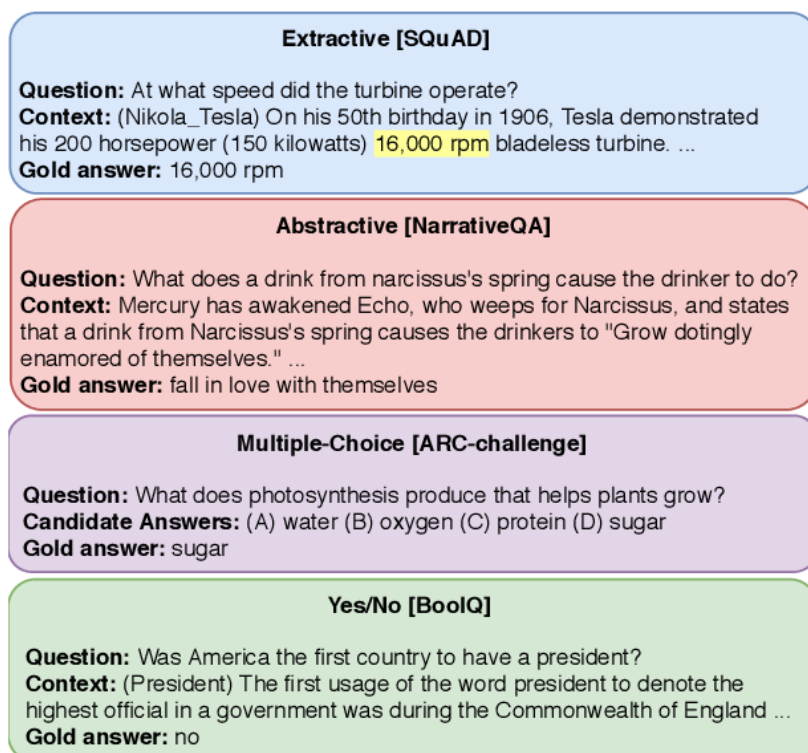
Hình vẽ 3.5 cho thấy luồng xử lý xử lý mô đun chitchat. Điểm đặc biệt trong hệ thống mô đun này đó là sử dụng những context self-introduce. Nó là các đoạn văn được định nghĩa, viết sẵn nhằm để tự nói, giới thiệu bản thân chatbot. Các mô đun Generator Answer sẽ cần phải có để có thể trả lời được các câu hỏi như tên tuổi,... từ người dùng. Việc hỗ trợ, tăng cường phạm vi hỗ trợ ,có thể .bổ sung tri thức thông qua các file context. Do vậy, ở mô đun này, em chỉ định nghĩa tri thức tự giới thiệu để có thể trả lời các câu hỏi như tên tuổi, ngày sinh, giới tính, địa chỉ,...

3.2.3. Generator Answer

Để có thể tạo được câu trả lời, phản hồi tới người dùng, chúng ta cần phải có một mô đun có khả năng sinh các câu trả lời từ context và câu hỏi của người dùng. Như em đã giới thiệu trong phần 2.3.5, sẽ có 2 dạng chính đó là Extractive QA và Abstract QA. Đối với Extractive QA, đây là các bài toán rất đơn giản, gần như các bài toán nhận diện span text trong câu đã cho trước đó, chỉ mới dùng lại trong việc xác định đâu là key để có thể trả lời câu hỏi. Extractive QA gây ra sự thiếu mạch lạc trong câu, khả năng suy luận và tổng hợp rất kém khi đưa mô hình làm việc trên các tập dữ liệu lớn. Ngoài ra, dữ liệu huấn luyện cho các mô hình này thường không phức tạp, rất đơn giản do bản chất của Extractive QA.. Đặc biệt, các mô hình tiếng Việt thường hoạt động rất kém do lượng dữ liệu có được để huấn luyện thường rất ít, còn nếu nhiều thường là dịch từ các bộ dữ liệu ngôn ngữ khác tuy nhiên chất lượng các bản dịch có thể ảnh hưởng nhiều tới hiệu suất của mô hình nếu bản dịch đó không chất lượng, hoặc chứa sạn. Đối với Abstract QA, cho phép mô hình tạo câu trả lời tổng hợp từ các chi tiết có trong context,

đồng thời cho thấy khả năng suy luận và tổng hợp vượt trội, khả năng làm việc trên nhiều loại domain khác nhau, cho phép tạo trải nghiệm và tương tác tốt với khách hàng, đồng thời có thể thực hiện được nhiều yêu cầu phức tạp hơn từ người dùng. Tuy nhiên, hạn chế lớn nhất của phương pháp này đến từ hai phần:

- Vấn đề dữ liệu: Dữ liệu cho Abstract QA thật sự là không nhiều và tiếng Việt là không có, tuy nhiên em có thể sử dụng các bộ dữ liệu được tạo ra bởi các mô hình ngôn ngữ lớn, nếu nó đủ tốt có thể hoàn toàn sử dụng cho tiếng Việt..
- Vấn đề mô hình: Thường sẽ sử dụng các mô hình sinh văn bản thay vì sử dụng các mô hình mã hóa vector như BERT, dẫn đến sử dụng các mô hình sẽ rất là nặng, tùy thuộc vào độ khó hay độ phức tạp của dữ liệu.

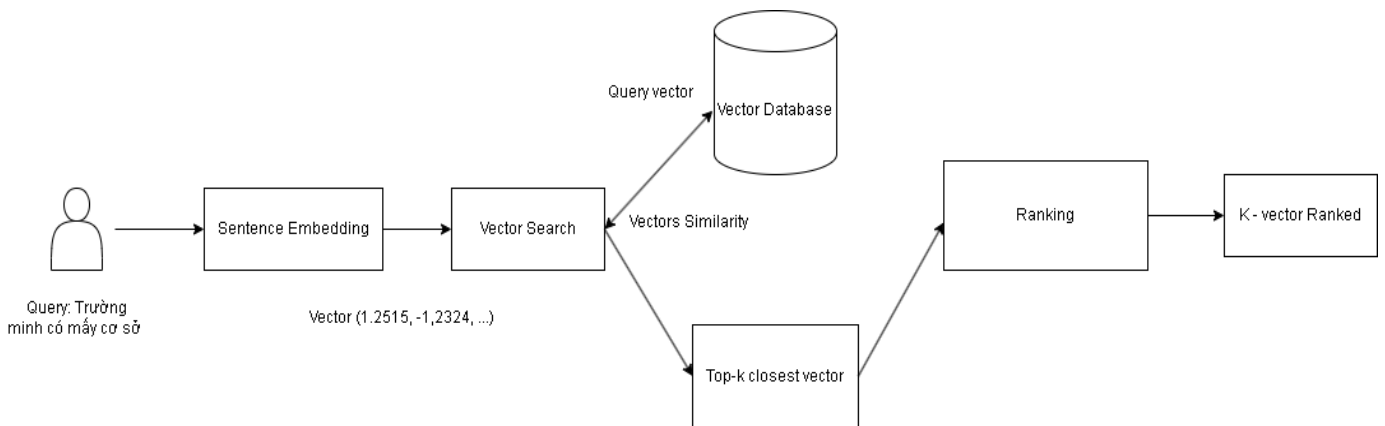


Hình 3.6: Các dạng dữ liệu question answering

Do đó, em có thể sử dụng các mô hình ngôn ngữ sinh văn bản như BLOOM, T5 để có thể hướng đến việc xây dựng các mô hình Abstract QA đủ tốt trên tiếng Việt, đặc biệt có thể sử dụng cho miền câu hỏi chủ đề về trường học, giáo dục. Trong phần 4.1, em sẽ đi sâu vào vấn đề thu thập dữ liệu và xử lý để có thể huấn luyện mô hình đạt kết quả tốt.

3.2.4. Retrieval và Ranker

Ngoài mô đun Generator Answer, mô đun quan trọng nhất của toàn bộ hệ thống hỏi đáp là Retrieval và Ranker. Hiệu suất của hai mô đun này ảnh hưởng rất lớn và mang tính quyết định liệu hệ thống có đủ tốt và đủ chính xác hay không. Trong khi các mô đun còn lại chỉ là hỗ trợ trong việc phân loại câu hỏi, và tạo câu trả lời, retrieval đảm nhiệm việc tìm kiếm các dữ liệu văn bản liên quan đến câu hỏi để Generator Answer có thể tạo câu trả lời và phản hồi tới người dùng. Hỗ trợ cho retrieval, ranker sẽ trực tiếp so sánh từng cặp của top - k dữ liệu văn bản liên quan đến câu hỏi nhất, để xác định được rằng đâu là dữ liệu văn bản quan trọng nhất. Hiện tại có hai phương pháp retrieval lẫn ranker được sử dụng nhiều đó là dense vector và sparse vector hay các thuật toán tìm kiếm thông thường như TF-IDF hoặc BM25. Tuy nhiên BM25 sẽ được em sử dụng kết hợp với

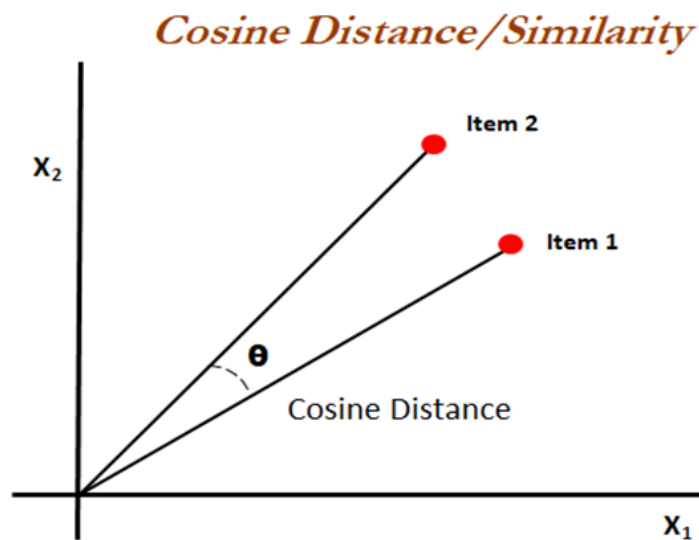


Hình 3.7: Retrieval và Ranker

Ranker, và trong một số trường hợp BM25 sẽ được dùng làm retrieval thay vì sử dụng phương pháp vector search, tuy nhiên phần này em chỉ làm rõ về kiến trúc tổng quan của mô đun, trong phần sau, thực nghiệm xây dựng ứng dụng em sẽ đi sâu và làm rõ hơn.

Retrieval như hình trên thực chất chính là vector search, nhận input là vector được mã hóa từ query của người dùng thông qua một mô hình mạng neural, vector này sẽ mang các thông tin về ngữ nghĩa query của người dùng. Sau đó, nó được tính toán và so sánh, thu được các vector có trong Vector Database có độ tương đồng hay khoảng cách gần nhất với nó (thông thường được sử dụng với Cosine Similarity hoặc Euclid). Tuy nhiên, mục tiêu của hệ thống question answering sẽ phải tìm được một hoặc nhiều các vector thật sự có liên quan tới query của người dùng. Việc sử dụng retrieval để lấy

top - k vector gần nhất chưa đủ để có thể coi là k - vector có thể sử dụng cho mô đun Generator Answer được, điều này ảnh hưởng bởi nhiều yếu tố khác nhau, đặc biệt trong nhiều trường không gian vector rất nhỏ chưa đa dạng hoặc chưa các vector không có nhiều ý nghĩa với nhau, việc lấy top - k, chỉ đơn giản tìm kiếm k vector gần nhất chứ không có nghĩa phải bắt buộc chúng phải thật sự có nghĩa với vector đang xét hay không. Vậy nên sẽ cần phải có mô đun ranker để đưa ra score thật sự giữa vector query và top-k vector, ngoại trừ nếu retrieval quá tốt (top - k vector tự động sắp xếp theo độ tương đồng) thì có thể sẽ không cần phải ranking tiếp.



Hình 3.8: Phép đo độ tương tự giữa hai vector - Cosine Similarity

Ranker là mô đun nối tiếp với Retrieval, có vai trò tính toán lại score giữa vector query và top - k vector. Sau đó, tùy từng bài toán hay hệ thống có thể lấy passage có score cao nhất hoặc top score cao nhất đối với bài toán QA multiple context. Hiện tại, Ranker có thể sử dụng các mô hình deep learning nlp đã huấn luyện cho tác vụ này hoặc cũng có thể sử dụng các thuật toán thống kê như BM25 như em đã nói ở phần trên. Tuy nhiên để đảm bảo được sự chính xác nên em sẽ hướng đến việc kết hợp giữa sử dụng mô hình Deep learning cùng với BM25.

CHƯƠNG 4: XÂY DỰNG HỆ THỐNG

Trong phần này em sẽ đi chi tiết vào việc xây dựng ứng dụng từ việc chuẩn bị dữ liệu, thống kê dữ liệu, xây dựng mô hình, huấn luyện và kiểm thử, đánh giá chất lượng trên từng mô hình. Cuối cùng chúng ta xây dựng và hoàn thiện hệ thống bao gồm xây dựng API, thiết kế giao diện và đánh giá hệ thống.

4.1. Xây dựng mô hình


4.1.1. Chuẩn bị và xây dựng dữ liệu

Do hệ thống có các mô hình khác nhau nên từng mô hình có thể sẽ sử dụng các dữ liệu khác nhau cũng như việc chuẩn bị và xử lý chúng sẽ khác nhau. Trong đó sẽ có dữ liệu cho các mô hình như Retrieval, Ranker, Generator Answer, tiếp đó là dữ liệu knowledge cho toàn bộ hệ thống vận hành.

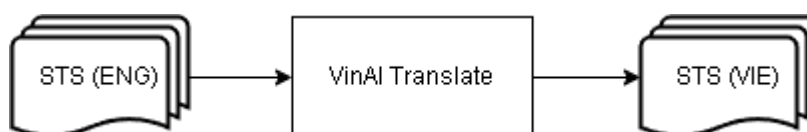
a. Chuẩn bị dữ liệu training mô hình Retrieval và Ranker

Mô hình Retrieval và Ranker được em hướng đến sẽ gần như tương đồng với các bài toán liên quan đến Semantic Textual Similarity để tối ưu việc tìm kiếm các context tương đồng, lúc này sẽ cần đến các dataset có khả năng hỗ trợ việc đưa ra score chính xác cho sự tương đồng của nó. Tuy nhiên, việc xây dựng một model tiếng Việt cho bài toán Semantic Textual Similarity tốt là một điều khá khó khăn, do hiện tại chưa có dữ liệu nào được công bố chính thức dưới dạng paper. Điều này khiến cho việc phát triển các mô hình khá khó khăn và không có một chuẩn dữ liệu đánh giá cụ thể. Tuy nhiên vẫn có thể sử dụng các phương pháp như cross-lingual (sử dụng mô hình pretrained multilingual fine-tuned trên bộ dữ liệu tiếng Anh rồi sử dụng trên ngôn ngữ khác) nhưng do liên quan đến biểu diễn ngữ nghĩa nên phải ưu tiên dữ liệu là tiếng Việt. Vậy nên sẽ cần phải dịch các bộ dữ liệu uy tín sang tiếng Việt để có thể sử dụng.

Bộ dữ liệu STS là một tập hợp các dữ liệu dùng để phục vụ tác vụ Semantic Textual Similarity từ 2012 cho đến 2016, bao gồm các đầu tin tức, trang báo,.. Có thể coi dữ liệu STS là dạng open-domain, rất thích hợp cho việc triển khai cho các tác vụ closed-domain như bài toán của chúng ta. Do đó, em tiến hành thu thập dữ liệu và đưa dữ liệu đi qua mô hình dịch (sử dụng mô hình en2vi của VinAI).

split string · classes	sentence1 string · lengths	sentence2 string · lengths	score float64
1 value	 10 416	 16 444	 0 5
train	But other sources close to the sale said Vivendi was keeping th...	But other sources close to the sale said Vivendi was keeping th...	4
train	Micron has declared its first quarterly profit for three years.	Micron's numbers also marked the first quarterly profit in three...	3.75
train	The fines are part of failed Republican efforts to force or...	Perry said he backs the Senate's efforts, including the fines, to...	2.8
train	The American Anglican Council, which represents Episcopalian...	The American Anglican Council, which represents Episcopalian...	3.4
train	The tech-loaded Nasdaq composite rose 20.96 points to 1595.91,...	The technology-laced Nasdaq Composite Index <.IXIC> climbed...	2.4
train	Amgen shares gained 93 cents, or 1.45 percent, to \$65.05 in...	Shares of Allergan were up 14 cents at \$78.40 in late trading...	1.333

Hình 4.1: Dữ liệu STS - 2012 (gốc)



Hình 4.2: Dịch STS (ENG) sang STS (VIE)

sentence1 string · lengths	sentence2 string · lengths	score float64
 1 228	 7 222	 0 5
Vấn đề có thể sẽ có nghĩa là thay đổi khác phục trước khi hạm đội tàu con thoi bắt đầu...	Ông nói vấn đề cần được khắc phục trước khi phi đội tàu con thoi được phép bay trở lại.	4.4
Chỉ số tổng hợp nasdaq, được xếp hạng về công nghệ.ixic giảm 1 điểm, tương đương...	Chỉ số 500 tiêu chuẩn rộng & nghèo.spz giảm 3 điểm, tương đương 0,32%, xuống 970.	0.8
" đó là một con mắt đen khổng lồ, " nhà xuất bản arthur ochs sulzberger jr., người...	" một con mắt đen to tướng, " arthur sulzberger, chủ bút của tờ báo, nói về vụ...	3.6
Chủ tịch ủy ban chứng khoán và hối đoái william donaldson nói rằng " có một sự tự...	" tôi nghĩ có một niềm tin vững chắc rằng viên cảnh sát đang đi tuần... ".	3.4
Cổ phiếu vivendi đóng cửa 1,9% ở mức 15,80 euro tại paris sau khi giảm 3,6% vào thứ...	Tại new york, cổ phiếu vivendi giảm 1,4% xuống còn 18,29 đô-la.	1.4
Nhà lãnh đạo ủng hộ dân chủ myanmar aung san suu kyi sẽ trở về nhà vào cuối thứ sáu...	Nhà lãnh đạo ủng hộ dân chủ myanmar aung san suu kyi sẽ bị quản thúc tại gia sau kh...	4.6

Hình 4.3: Dữ liệu STS - 2012 (Đã được dịch)

Sau khi thu thập và xử lý dữ liệu, tuy nhiên chỉ thu thập được 32 282 sample. Nhận thấy rằng bài toán của chúng ta cần đạt được có hiệu suất tổng quan thực tế cao nhất có thể. Thực tế lượng dữ liệu đó quá khiêm tốn để có thể đảm bảo rằng nó có thể phân ngữ nghĩa tốt trên miền dữ liệu hẹp như dữ liệu câu hỏi về trường học, công tác

sinh viên của chúng ta. Vậy nên chúng ta phải tìm các bộ dữ liệu khác bổ trợ, có thể không cùng dạng với dữ liệu STS, tuy nhiên có thể xem xét để bổ trợ nó theo một phương pháp nào đó.

Bộ dữ liệu Zalo, trước đó Zalo AI đã tổ chức các cuộc thi AI liên quan đến các chủ đề về question answering system và retrieval (các cuộc thi có bài toán QA diễn ra vào năm 2019, 2021, 2022).. Do dataset của họ cũng yêu cầu các mô hình phải có khả năng học và dự đoán được xem question và context có liên quan tới nhau không, nên em hoàn toàn có thể xử lý và đưa về dạng của mình, với label 0 là question và context không thể đưa ra câu trả lời, ngược lại với label 1, question và context liên quan tới nhau, dữ liệu thu về khoảng 97k sample.

	question	text	label
0	Đạo_đức là gì	Đạo_đức kinh_doanh là một tập_hợp các nguyên_t...	0
1	Cơ_quan quyền_lực nhất của nhà_nước ta	Quốc_hội (Cameroon) Quốc_hội () là ngôi nhà...	0
2	xiêm la là quốc_hiệu cũ của quốc_gia nào	Tây_Thái_Lan Miền_Tây_Thái_Lan là vùng giáp bi...	0
3	Bài hát 19 tháng 8 là sáng_tác của ai	Trong khu nhà nhỏ phố Quán_Sứ , Hà_Nội , đôi_k...	1
4	tổng_giám_đốc thuộc công_ty nhà_nước trong việ...	thực_hiện các nhiệm_vụ liên_quan đến quyền ngh...	0
...
97029	bàn thắng thứ 85 của cristiano ronaldo 2018 đã...	Maroš Ferenc Maroš_Ferenc (sinh 19 tháng 2 nă...	0
97030	Đại_danh y Lê_Hữu Trác quê ở huyện nào của tỉn...	Lê_Hữu Trác vốn có tên cúng_cơm là Huân (燕) ...	1
97031	Thành_phố nào là thủ_đô của Libya hiện_nay	Sân_bay quốc_tế Tripoli Sân_bay quốc_tế Tripol...	0
97032	thu_hồi tài_nguyên khoáng_sản thông_qua hoạt_đ...	việc nạo_vét khơi thông luồng để mở mới cải_tạ...	0
97033	Cầu Kiền nằm trên đường nào	10	0

97034 rows × 3 columns

Hình 4.4: Dữ liệu Zalo đã được xử lý

b. Chuẩn bị dữ liệu training mô hình Generator Answer

Bài toán Generator Answer được em hướng đến sẽ ưu tiên là dạng Abstract QA thay vì Extractive QA do các nhược điểm mà em đã phân tích trước đó trong 3.1.3 và 2.3.5, đồng thời do đây là bài toán hỏi đáp thông tin nên có thể sẽ có trường hợp yêu cầu truy xuất đầy đủ hoặc tóm tắt thông tin, tuy nhiên thực tế Extractive QA thực sự không tốt cho các yêu cầu phức tạp. Tuy nhiên, với bài toán này, dữ liệu vẫn là trở ngại lớn trên bộ dữ liệu tiếng Việt khi cũng không có một bộ dataset hay một bộ đánh giá chuẩn nào cả, nhưng như em đã phân tích ở trên, cũng có thể sử dụng các bộ dataset đủ lớn và dịch sang tiếng Việt. Do đó, em sử dụng bộ dữ liệu trong cuộc thi AI của Viettel (Viettel Hearted AI Challenges) tác vụ đa phần là về Question Answering Open-Domain . Bộ dữ liệu này được ban tổ chức tập hợp các sample tác vụ question answering từ rất nhiều

bộ dataset khác chuyên dùng để instruction fine-tuning large language model, sau đó được dịch về tiếng Việt và thu được khoảng 100k sample.



Hình 4.5: Dữ liệu data Viettel Hearted AI Challenges

Tuy nhiên, nhận thấy có nhiều các sample không phù hợp, chỉ là dạng question answering nhưng không có context đi kèm, việc này sẽ khiến cho model sau khi học có khả năng bị tình trạng “hallucination” - “ảo giác”. Tình trạng này xảy ra khi các model đưa ra các câu trả lời sai mà câu trả lời đó không liên quan đến context đưa vào. Vậy nên để giảm thiểu khả năng hallucination, em tiến hành lọc và loại bỏ, chỉ giữ lại các dạng câu hỏi có sử dụng context bằng cách lấy các sample, hay các hàng có phần tử ở cột input. Cuối cùng thu được khoảng 87k sample

	instruction	inp	oup
0	You are an smart assistant. Provide a direct, ...	Context: Thành phố đã đầu hàng mà không chiến ...	Abdullah đã xây dựng Russafa bên ngoài Valencia.
1	As an intelligent AI model, your task is to an...	Provide answer to the question at the end base...	Java không an toàn vì nó cho phép vi phạm các ...
2	As an intelligent AI model, your task is to an...	Provide answer to the question at the end base...	Rất khó để tái tạo chính xác hành vi của con n...
3	As an intelligent AI model, your task is to an...	Provide answer to the question at the end base...	Các diễn viên thường tự sản xuất phim để tránh...
4	As an intelligent AI model, your task is to an...	Provide answer to the question at the end base...	Ngất xỉu vì đau quá có thể là do trực trặc tro...
...
86641	As an intelligent AI model, your task is to an...	Provide answer to the question at the end base...	Ngả lưng khi ngồi trên ghế có thể gây ra nhiều...
86642	You are an AI assistant. User will you give yo...	AP - Mary Frances Berry, chủ tịch thẳng thắn c...	Bước 1: Xác định chủ đề chính của bài viết - M...
86643	You are an smart assistant. Provide a direct, ...	Context: Houston là trụ sở của Trung tâm Y tế ...	Nhóm các tổ chức chăm sóc sức khỏe lớn nhất th...
86644	You are an smart assistant. Provide a direct, ...	Context: Yazid chết trong khi cuộc bạo vậy vẫn...	Trận chiến Marj Rahit diễn ra vào năm 684.
86645		Câu hỏi: Bạn có nghĩ Brittany giết mẹ mình khô...	Không đủ thông tin để đưa ra câu trả lời.
86646 rows × 3 columns			

Hình 4.6: Dữ liệu data Viettel Hearted AI Challenges (filtered)

c. Chuẩn bị dữ liệu cho Domain Classification

Như em đã trình bày ở trước đó trong mục 3.1.1, mô đun này sẽ không cần phải có lượng lớn dữ liệu để huấn luyện một model classification, nhưng cần phải có một lượng dữ liệu đủ để có thể giúp mô đun domain Classification phân biệt được giữa chitchat và qa.

Mô đun domain Classification sẽ tính toán trung bình của khoảng cách giữa câu query của người dùng và top_k câu liên quan nhất từ tập dữ liệu chitchat, sau đó dựa vào threshold đã định sẵn để quyết định xem liệu nó có phải chitchat hay không. Nếu khoảng cách trung bình vượt ngưỡng threshold thì query được coi là domain chitchat. Em tự tạo bằng tay khoảng 85 sample chitchat, các câu trả lời thuộc các dạng như: chào hỏi, hỏi tên, hỏi tuổi, năm sinh, lý do được tạo ra, hỏi giới tính. Mục đích của mô đun là phân loại để hệ thống ngoài việc hỏi đáp chuyên môn ra còn có khả năng chào hỏi và giới thiệu cơ bản.

d. Chuẩn bị dữ liệu knowledge

Dữ liệu knowledge sẽ được dùng để truy xuất kết hợp với câu hỏi của người dùng đưa vào, hệ thống sẽ tiến hành retrieval các đoạn văn liên quan rồi đưa qua mô đun generator answer để phản hồi lại câu trả lời cho chính câu hỏi đó. Dữ liệu knowledge này bao gồm thông tin giới thiệu nhà trường, phòng ban, thư viện...; thông tin giảng viên đang làm việc tại trường bao gồm tên tuổi, học hàm học vị,...; thông tin về công tác sinh viên: học bổng, hỗ trợ khó khăn, chính sách, điểm rèn luyện; quy chế đào tạo: hỏi đáp về học tập, hỏi đáp về chương trình đào tạo,... Các dữ liệu này được em crawl từ nhiều nguồn khác nhau liên quan đến trường bao gồm từ trang web của trường, các tài liệu ban hành dưới dạng doc, txt, pdf. Tuy nhiên có một số tài liệu pdf không thể dùng phương pháp OCR được, nên có thể coi chủ đề hỏi đáp đó bị loại bỏ đi, bao gồm: tài liệu hướng dẫn làm thủ tục hành chính cho sinh viên. Bên cạnh đó, các dữ liệu thông tin về giảng viên đang làm việc tại trường được em crawl trực tiếp từ trang <https://www.utt.edu.vn/teacher/>

#	Giảng viên	Đơn vị	Cơ sở	Lý lịch khoa học
1	Vũ Ngọc Khiêm	Hội đồng trường	Hà Nội	Xem chi tiết
2	Nguyễn Hoàng Long	Ban Giám hiệu	Hà Nội	Xem chi tiết
3	Nguyễn Mạnh Hùng	Ban Giám hiệu	Hà Nội	Xem chi tiết
4	Nguyễn Văn Lâm	Ban Giám hiệu	Hà Nội	Xem chi tiết
5	Trần Hà Thanh	Ban Giám hiệu	Hà Nội	Xem chi tiết
6	Lê Thu Sao	Công Đoàn	Hà Nội	Xem chi tiết
7	Hoàng Vũ	Đoàn Thanh niên	Hà Nội	Xem chi tiết
8	Tạ Thị Hồng Nhung	Khoa Công trình	Hà Nội	Xem chi tiết
9	Phùng Bá Thắng	Khoa Công trình	Hà Nội	Xem chi tiết

Hình 4.7: Trang thông tin giảng viên

Sau khi crawl về thu được 604 thông tin giảng viên tại trường

```
[{'PGS.TS Vũ Ngọc Khiêm. Học hàm, học vị: PGS.TS. Quê quán: Thanh Hoá. Đơn vị: Hội đồng trường. Giới tính: Nam. Email: khiemvn@utt.edu.vn. Địa chỉ: Số 54 Phố Triều Khúc, phường Thanh Xuân Nam, quận Thanh Xuân, thành phố Hà Nội',
'PGS.TS Nguyễn Hoàng Long. Học hàm, học vị: PGS.TS. Quê quán: Dân Tiến - Khoái Châu - Hưng Yên. Đơn vị: Ban Giám hiệu. Giới tính: Nam. Email: longnh@utt.edu.vn. Địa chỉ: Phòng 302 Nhà H, số 54 Triều Khúc - Thanh Xuân - Hà Nội',
'Tiến sỹ Nguyễn Mạnh Hùng. Học hàm, học vị: Tiến sỹ. Quê quán: Phú Thọ. Đơn vị: Ban Giám hiệu. Giới tính: Nam. Email: hungnm@utt.edu.vn. Địa chỉ: Phòng 302 Nhà H3',
'Tiến sĩ Nguyễn Văn Lâm. Học hàm, học vị: Tiến sĩ. Quê quán: Văn Hội - Tam Dương - Vĩnh Phúc. Đơn vị: Ban Giám hiệu. Giới tính: Nam. Email: lamnv@utt.edu.vn. Địa chỉ: Phòng Đào tạo Trường Đại học Công nghệ GTVT - 54 Triều Khúc - Thanh Xuân - Hà Nội',
'Tiến sĩ Trần Hà Thanh. Học hàm, học vị: Tiến sĩ. Quê quán: Hà Tĩnh. Đơn vị: Ban Giám hiệu. Giới tính: Nam. Email: thanhth@utt.edu.vn. Địa chỉ: 348 Đường Nguyễn Trãi, phường Trung Văn, Quận Nam Từ Liêm - Thành phố Hà Nội',
'Tiến sỹ Lê Thu Sao. Học hàm, học vị: Tiến sỹ. Quê quán: Nam Định. Đơn vị: Công Đoàn. Giới tính: Nữ. Email: lethusao@utt.edu.vn. Địa chỉ: P 306 H1 Đại học Công nghệ GTVT',
'Tiến Sĩ Kỹ Thuật Hoàng Vũ. Học hàm, học vị: Tiến Sĩ Kỹ Thuật. Quê quán: Hà Nội. Đơn vị: Đoàn Thanh niên. Giới tính: Nam. Email: vuhoang@utt.edu.vn. Địa chỉ: Số 54, Phố Triều Khúc, Quận Thanh Xuân, Hà Nội',
'Kỹ sư Tạ Thị Hồng Nhung. Học hàm, học vị: Kỹ sư. Quê quán: Phú Thọ. Đơn vị: Khoa Công trình. Giới tính: Nữ. Email: nhungth@utt.edu.vn. Địa chỉ: Phòng 202- nhà H5 - Văn phòng khoa - Khoa Công trình - Trường đại học CN GTVT',
'Tiến sĩ kỹ thuật Phùng Bá Thắng. Học hàm, học vị: Tiến sĩ kỹ thuật. Quê quán: Nghệ An. Đơn vị: Khoa Công trình. Giới tính: Nam. Email: thangpb@utt.edu.vn. Địa chỉ: Bộ môn Cầu - Hầm, trường Đại học Công nghệ GTVT, 54 Triều Khúc, Thanh Xuân, Hà Nội',
'Tiến sĩ kỹ thuật Nguyễn Anh Tuấn. Học hàm, học vị: Tiến sĩ kỹ thuật. Quê quán: Hà nội. Đơn vị: Khoa Công trình. Giới tính: Nam. Email: tuanna@utt.edu.vn. Địa chỉ: 54 Triều khúc - Thanh Xuân - Hà Nội']
```

Hình 4.8: Thông tin của một số thầy cô trong trường.

Đối với các dữ liệu text thô, ngoài dữ liệu thông tin của các thầy cô trong trường, em tiến hành chia các tập văn bản lớn thành từng đoạn văn nhỏ phù hợp để có thể dễ dàng được tìm kiếm cho mô đun retrieval. Sau khi crawl về xong, mỗi file sẽ chứa từng chủ đề hoặc quy chế, quy định riêng, ví dụ thông tin phòng hành chính được lưu file riêng, thông tin phòng đào tạo cũng được lưu file riêng và thông tin quy chế đào tạo được lưu một file riêng khác. Điều này cho phép dễ dàng thực hiện việc chia các văn bản thành nhiều đoạn văn nhỏ.

ctsv_hocbong	03/12/2023 9:30 CH	Tài liệu văn bản	7 KB
ctsv_quy che sinh vien	03/12/2023 9:41 CH	Tài liệu văn bản	36 KB
ctsv_quyche_diem_renluyen	03/12/2023 9:45 CH	Tài liệu văn bản	9 KB
chucnang_damnhiem_phong_daotao	03/12/2023 5:32 SA	Tài liệu văn bản	12 KB
fqa_tai_chinh_ketuan	03/12/2023 9:24 CH	Tài liệu văn bản	4 KB
gen_query_answer	29/11/2023 1:21 SA	Jupyter Source File	62 KB
hd_nop_taikhoan_online	03/12/2023 3:54 CH	Tài liệu văn bản	6 KB
lichsu_truyenthong	04/12/2023 4:17 SA	Tài liệu văn bản	20 KB
noi quyhocduong	03/12/2023 5:25 CH	Tài liệu văn bản	3 KB
phong_ban_ctsv	04/12/2023 2:53 SA	Tài liệu văn bản	11 KB
phong_ban_daotao	04/12/2023 2:54 SA	Tài liệu văn bản	3 KB
phong_ban_hanhchinh_quantri	04/12/2023 3:13 SA	Tài liệu văn bản	5 KB

Hình 4.9: Mỗi chủ đề sẽ được lưu một file riêng

Các văn bản sẽ được tách và được đánh dấu có thể tách bằng “####”, tiêu chí có thể tách được bao gồm:

- Là các đầu mục nhưng phải đảm bảo được độ dài văn bản của đầu mục không quá lớn do gây ra tràn ram gpu khi đưa qua model generator answer, đồng thời độ dài quá lớn khiến cho dư thừa khi sử dụng retrieval do các input của các model đó max_length tối đa là 256 tokens dẫn đến việc retrieval không hiệu quả do có thể thông tin quan trọng nằm ở cuối đoạn văn.
- Các đầu mục nhỏ hơn, nếu nội dung của chúng độc lập, không hỗ trợ lẫn nhau có thể được tách riêng nhưng phải xóa tên đầu mục chung.
- Nếu độ dài của cả văn bản (1 file) quá nhỏ có thể không cần phải tách.
- Sau khi đánh dấu tách, chia nhỏ văn bản bắt buộc phải xem có mang ý nghĩa chủ đề hoặc có thể bị nhầm lẫn khi retrieval, nếu có phải thêm tên chủ đề.

```

Phòng Đào tạo Sau đại học: địa chỉ: Phòng 207 Nhà H1 - Trường Đại học Công nghệ GTVT - Số 54 Triều Khúc - Thanh Xuân - Hà Nội. Điện thoại: 043.2222.865. Website: http://sdh.utt.edu.vn. Email: phongdtsdh@utt.edu.vn
####
Tổ chức nhân sự của phòng đào tạo sau đại học:
- Trưởng phòng: TS. Trần Thanh An
Cán bộ cơ hữu:
1. ThS. Đỗ Thanh Long
2. ThS. Phan Huy Thực
####
Phòng Đào tạo Sau đại học được Hiệu trưởng Trường Đại học Công nghệ GTVT ra quyết định thành lập, số 706/QĐ-ĐHCNGTVT ngày 25/03/2015.
####
Chức năng của phòng đào tạo sau đại học:
Giúp Hiệu trưởng thực hiện công các tổ chức và quản lý về đào tạo sau đại học bao gồm:
- Chiến lược phát triển đào tạo;
- Chương trình đào tạo;
- Quản lý đào tạo;
- Công tác tuyển sinh;
- Cấp phát văn bằng;
- Chứng chỉ theo quy định của pháp luật.
####
Nhiệm vụ của phòng đào tạo sau đại học:
- Đề xuất, phối hợp với các khoa, bộ môn chuyên ngành xây dựng chiến lược phát triển đào tạo, chương trình đào tạo SDH phù hợp với định hướng phát triển của Trường, nhu cầu xã hội và quy định của Bộ Giáo dục và Đào tạo;
- Xây dựng kế hoạch tuyển sinh, kế hoạch đào tạo; tổ chức ôn tập thi tuyển nhập học; quản lý chất lượng đào tạo theo quy chế tuyển sinh thạc sĩ và tiến sĩ hiện hành;
- Quản lý, lưu trữ hồ sơ, cấp phát văn bằng, chứng chỉ, chứng nhận nhập học liên quan đến người học theo quy định của quy chế đào tạo;
- Đề xuất, quản lý các hoạt động đào tạo liên kết, hợp đồng giảng viên thỉnh giảng;
- Soạn thảo trình Hiệu trưởng ký ban hành các văn bản quản lý về lĩnh vực đào tạo SDH, thực hiện quy chế đào tạo SDH;
- Tham mưu, trình Hiệu trưởng tổ chức các hội nghị, hội thảo liên quan đến công tác đào tạo SDH;
- Phối hợp với các đơn vị trong Trường thực hiện nhiệm vụ được giao;
- Hằng tháng báo cáo kết quả thực hiện nhiệm vụ được giao, đề xuất kế hoạch công tác tháng tới tại hội nghị giao ban hoặc bằng văn bản trình Hiệu trưởng và thực hiện chế độ báo cáo theo quy định;
- Thực hiện các nhiệm vụ khác do Hiệu trưởng giao.

```

Hình 4.10: Chia nhỏ văn bản thành các đoạn văn

Hình 4.10: Ví dụ cho thấy việc chia nhỏ văn bản thành các đoạn văn nhỏ hơn đảm bảo được các tiêu chí trên, Các đoạn nhỏ hơn mang nội dung độc lập nhau, tuy nhiên vẫn phải đảm bảo được cùng một chủ đề để dễ dàng truy vấn được. Thực tế, ban đầu thông tin phòng đào tạo sau đại học trên, các mục nhỏ sẽ không có “của phòng đào tạo sau đại học”. Nếu sau khi tách, mà không thêm vào, khi retrieval với câu query: “Bạn hãy trình bày tổ chức nhân sự phòng đào tạo sau đại học”, sẽ không thể nào phân biệt được các tổ chức nhân sự của các phòng ban với nhau. Vậy nên, ở các trường hợp tương tự như này, bắt buộc phải điền thêm “của phòng đào tạo sau đại học” để có thể dễ dàng phân biệt và truy xuất đúng mong đợi.

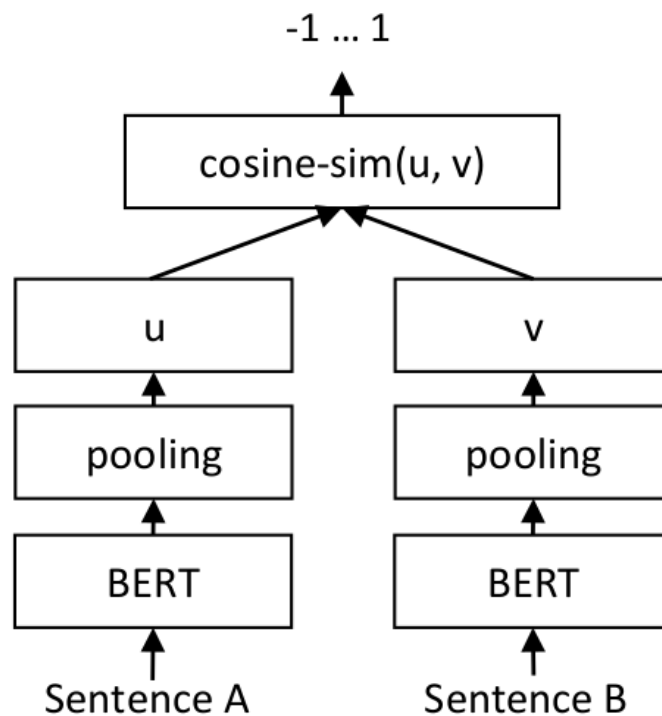
Ngoài ra, về mặt dữ liệu phải đảm bảo không bị trùng, tránh bị lỗi chính tả và các ký tự đặc biệt, có thể ảnh hưởng một phần đối với retrieval và ảnh hưởng đến kết quả cuối cùng bởi model generator answer có thể đem những ký tự đó vào, gây ra khó chịu đối với người dùng.

4.1.2. Thực nghiệm mô hình

Sau khi đã chuẩn bị và xây dựng dữ liệu xong, em sẽ tiến hành training cho từng mô hình bao gồm Sentence Embedding, Cross Encoder, Generator Answer đồng thời sẽ đi sâu vào việc lựa chọn mô hình, xây dựng cũng như lựa chọn tham số, cuối cùng đánh giá trên một số tập test.

a. Thực nghiệm mô hình Sentence Embedding

Mô hình Sentence Embedding là mô hình chính trong mô đun Retrieval, giúp mã hóa các thông tin tốt hơn bằng việc fine-tuning theo hướng Bi-Encoder dựa theo ý tưởng kiến trúc của Sentence Transformer. Mô hình sẽ tối ưu hóa để học được điểm tương đồng giữa hai vector.

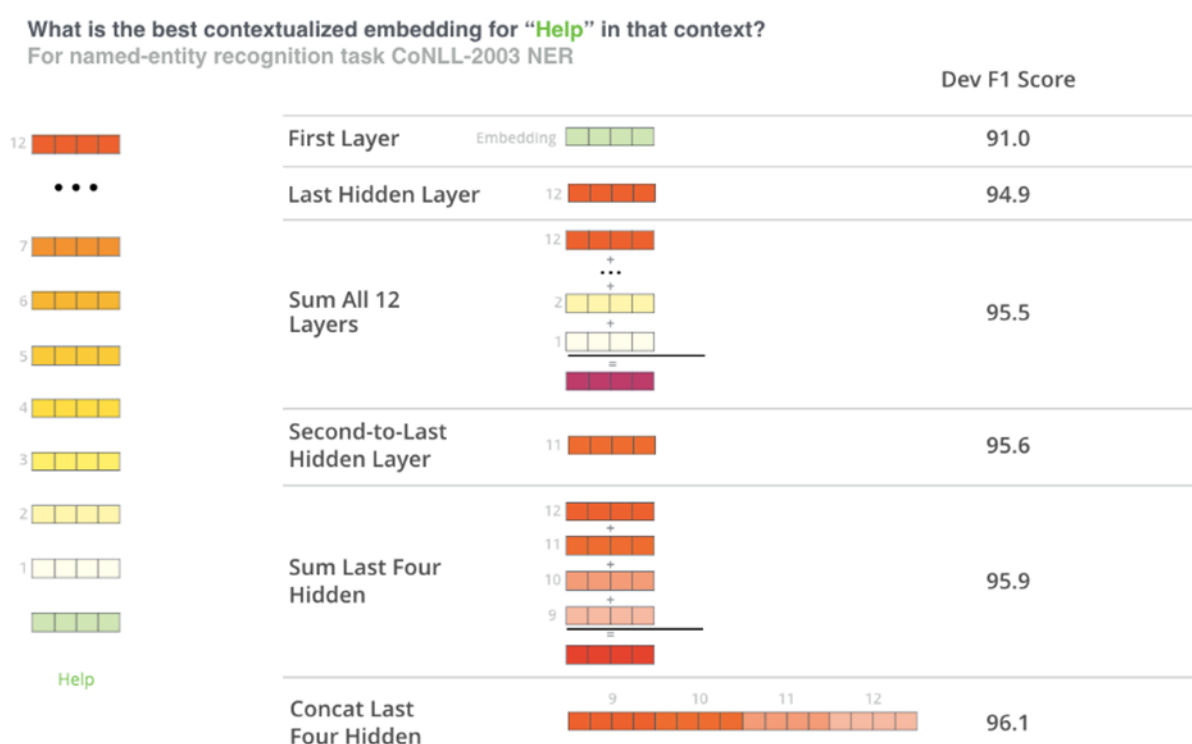


Hình 4.11: Kiến trúc Sentence Transformer

Tuy nhiên để đạt được hiệu suất cao hơn nữa, cần phải biến đổi một số layer có trong kiến trúc, nhằm làm giàu embedding cũng như giúp cho mô hình biểu diễn tốt hơn. Phần thực nghiệm cho kiểu kiến trúc ban đầu, đã được em thực nghiệm huấn luyện và đánh giá trên thư viện sentence-transformers, vậy nên kết quả đánh giá này em sẽ ghi ở mục đánh giá dưới. Kiến trúc Sentence Transformer sẽ được em kế thừa và biến đổi lại nhằm

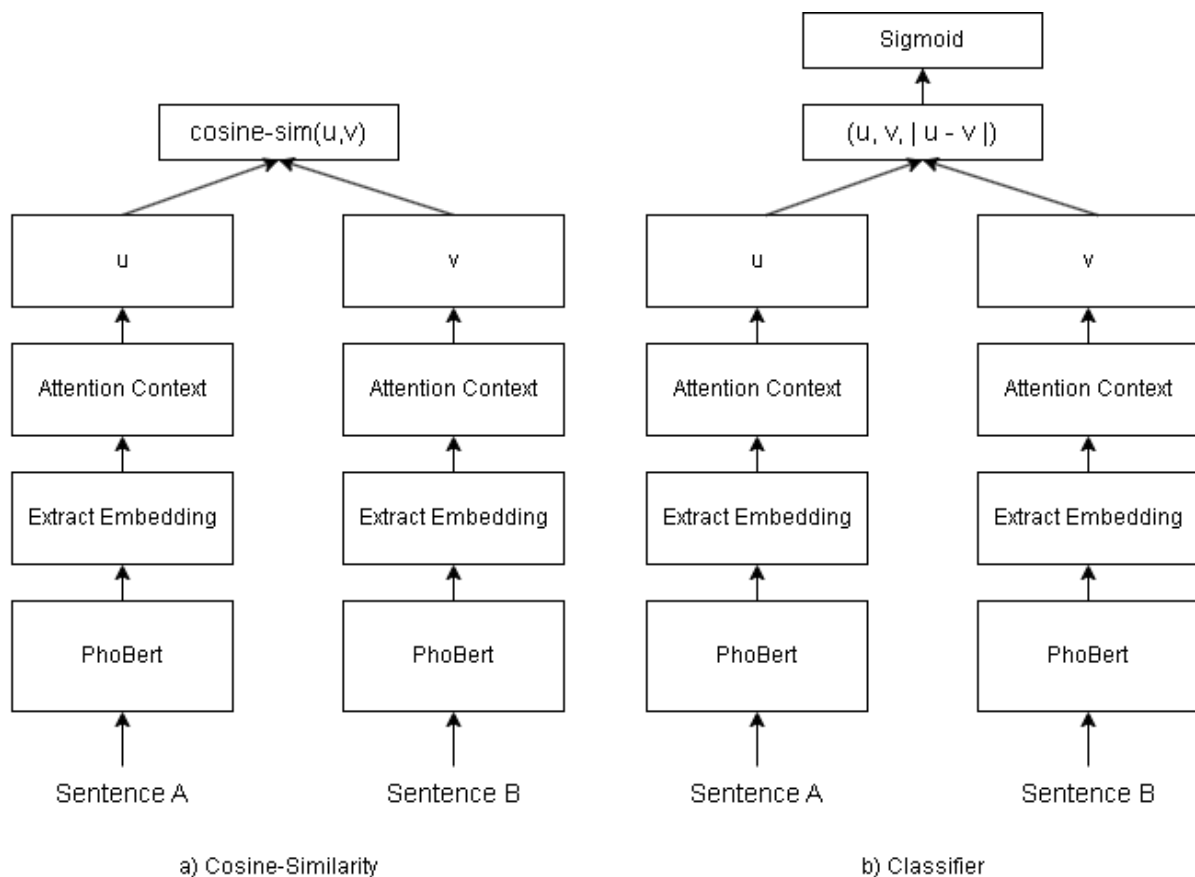
đạt hiệu quả cao cho mô hình. Extract Embedding và Attention Context sẽ là hai phương pháp được em sửa đổi và bổ sung cho kiến trúc.

Extract Embedding là phương pháp lấy thông tin từ nhiều hidden layer thay vì chỉ lấy embedding từ lớp hidden cuối cùng. Trong đó có một số nghiên cứu chỉ ra rằng khi sử dụng embedding từ việc concat 4 hidden cuối cùng cho kết quả tốt hơn so với việc dùng hidden state cuối cùng trong bài toán NER. Tuy nhiên để đảm bảo được sự cân bằng giữa việc tối ưu với hiệu suất, nên em sẽ chọn phương án dùng tính mean cho 4 hidden cuối cùng để tối ưu không gian bộ nhớ đồng thời tăng tốc độ training do việc lấy mean cũng đã tự động normalize tham số.



Hình 4.12: Thử nghiệm sử dụng nhiều hidden state

Attention Context có nhiệm vụ tính toán trọng số từng từ trong câu, tập trung vào những từ quan trọng và sau đó tổng hợp chúng thành vector sentence. Attention Context được áp dụng trong kiến trúc trên, có thể tính toán trọng số trên toàn bộ vector embedding, và trong quá trình học, nó sẽ tự động tập trung vào những vector quan trọng hay lấy những thông tin quan trọng để cuối cùng tạo ra sentence embedding. Attention Context sẽ thay thế cho mean pooling, vốn chỉ là tính mean trên toàn bộ vector embedding. Như vậy, kiến trúc mới sẽ là:

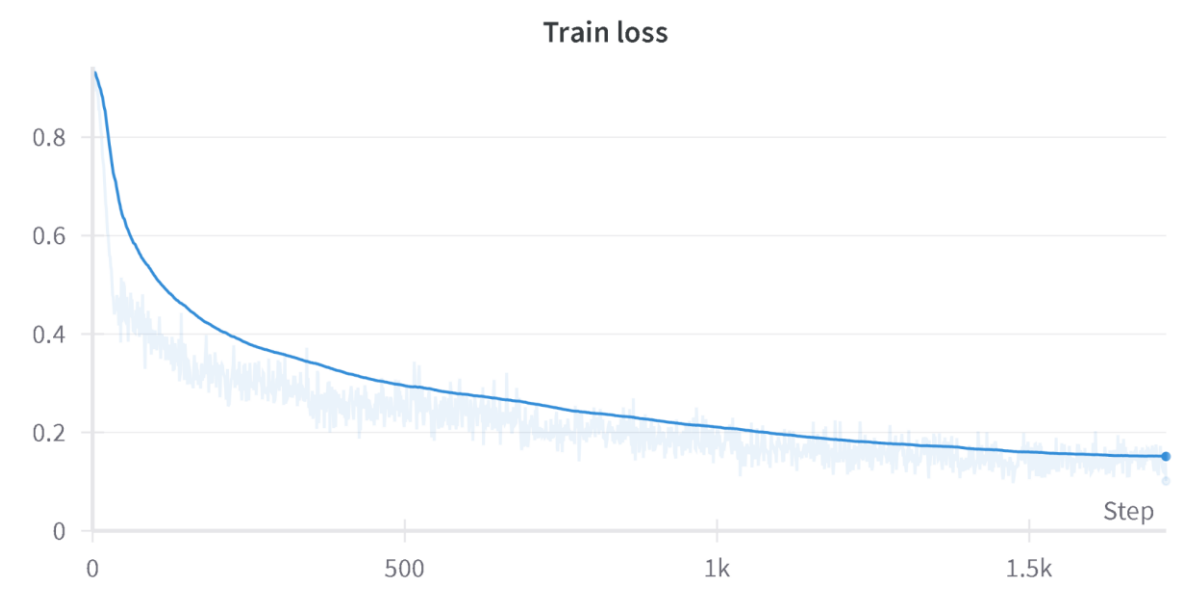


Hình 4.13: Sentence Embedding sau khi được cải tiến

Để huấn luyện Sentence Embedding, sử dụng lần lượt 2 bộ dữ liệu Zalo và STS để fine-tuning. Zalo sẽ được fine-tuning đầu tiên trên mô hình phân loại, nhằm phân loại cặp câu nào liên quan đến nhau. Lúc này, output của mô hình sẽ cố gắng đưa về 0 nếu hai câu không liên quan tới nhau, ngược lại đưa về 1. Sau đó sử dụng bộ dữ liệu STS để fine-tuning với hàm loss cosine-similarity để đưa output của mô hình để học sự nhạy cảm trong dataset. Lúc này nó có thể đưa ra score phù hợp thay vì chỉ đưa ra tiệm cận 0 hoặc 1. Các kết quả dưới đây được em thử nghiệm trên RTX 3090 24GB

Các tham số mà em sử dụng cho việc fine-tuning trên bộ dữ liệu zalo lần lượt là:

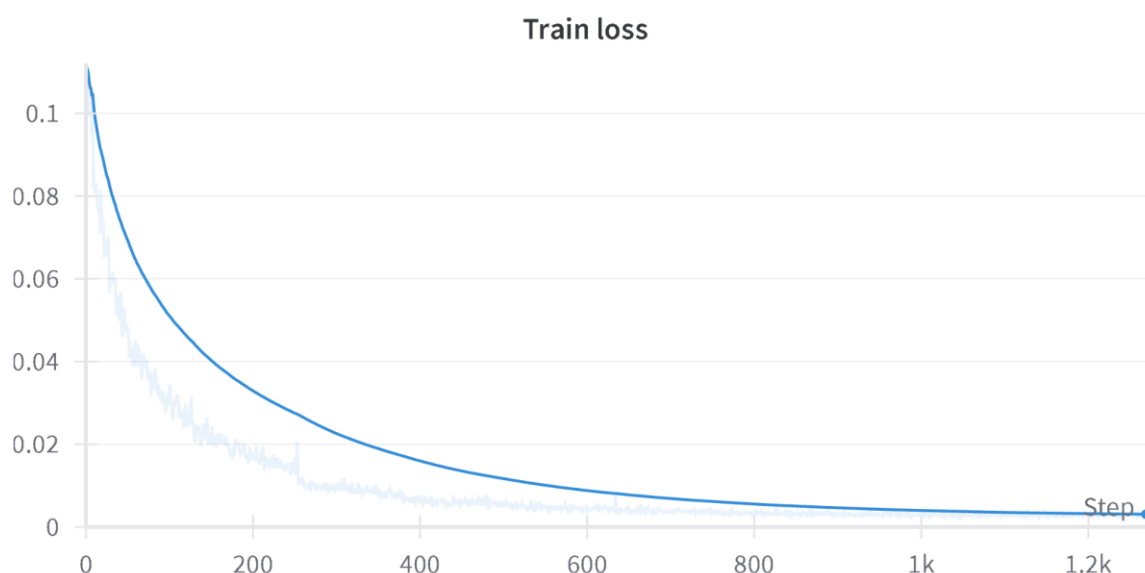
- batch_size: 64
- required_grad: True
- epochs: 5
- learning_rate: $2e-5$
- gradient_accumulation_steps: 8



Hình 4.14:Biểu đồ training loss Sentence Embedding - Zalo

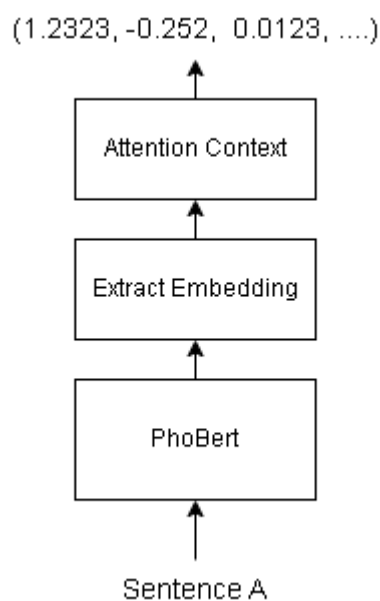
Các tham số mà em sử dụng cho việc fine-tuning trên bộ dữ liệu STS lần lượt là:

- `batch_size`: 64
- `required_grad`: True
- `epochs`: 10
- `learning_rate`: $1e-4$
- `gradient_accumulation_steps`: 8



Hình 4.15: Biểu đồ training loss Sentence Embedding - STS

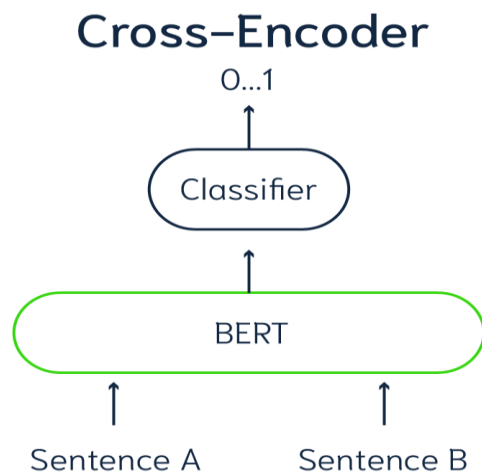
Khi sử dụng mô hình Sentence Embedding để lấy vector embedding, chúng ta bỏ layer cuối cùng để lấy embedding sẽ là vector output có chiều là (1, 768) do hiện tại mô hình đang sử dụng model base là Phobert-base-v2. Bên cạnh đó, việc train Bi-Encoder thực chất là mô hình dùng chung weight để cập nhật tham số nên khi cần lấy vector embedding của câu, chúng ta chỉ cần lấy output từ layer Attention Context.



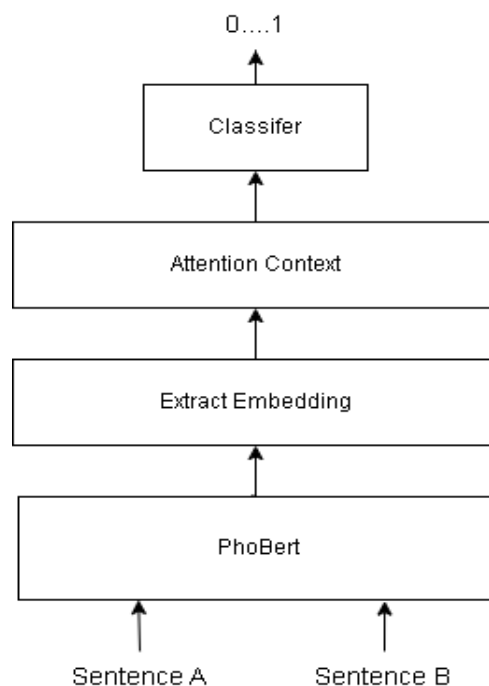
Hình 4.16: Sentence Embedding inference

b. Thực nghiệm mô hình Cross-Encoder

Mô hình Cross-Encoder là một phần của mô đun Ranker. Mô hình này sử dụng phương pháp concat 2 sentence input lại với nhau tạo thành 1 sentence rồi đưa qua model để predict. Tuy nhiên, với mô hình này em cũng sẽ sửa đổi và bổ sung tương tự như Sentence-Embedding. Kết quả dưới đây được em thử nghiệm trên RTX 3090 24GB.



Hình 4.17: Cross-Encoder



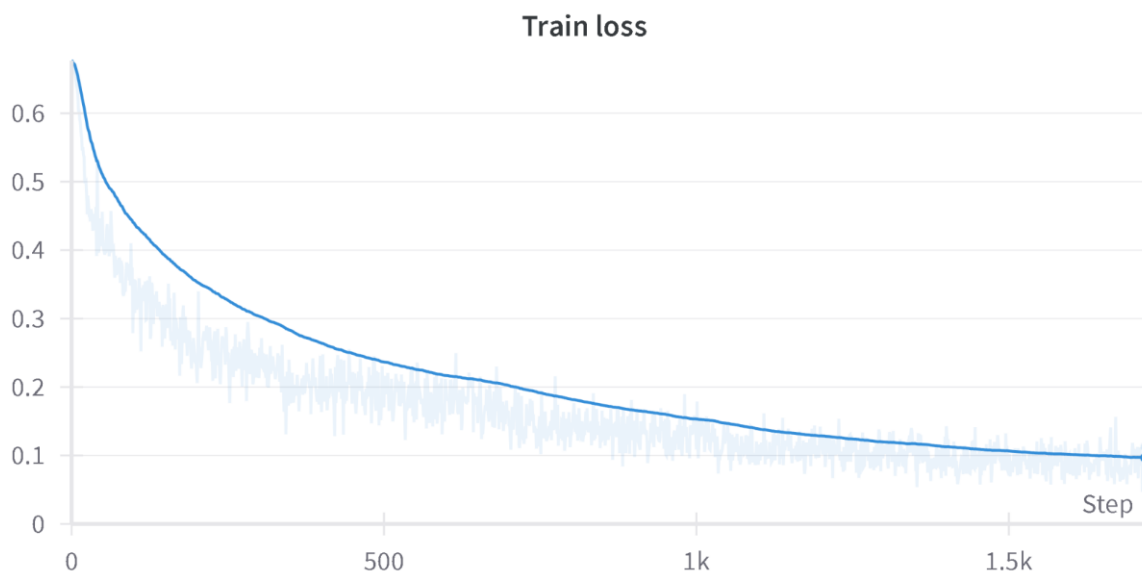
Hình 4.18: Cross-Encoder sau khi được cải tiến

Để thực hiện huấn luyện cho Cross-Encoder, em tiến hành huấn luyện lần lượt như với Zalo và STS lần lượt như với thực nghiệm sentence embedding. Đặc biệt khi fine-tuning với STS, em tiến hành sử dụng loss function khác thay thế việc sử dụng cross

entropy như Sentence Transformers, thay vào đó, activation sigmoid sẽ được dùng để đưa output về trong khoảng $0 \dots 1$, sau đó sử dụng loss function mean square error để đưa các output về gần với label là số thực $0 \dots 1$. Các kết quả dưới đây được em thực nghiệm trên RTX 3090 24GB.

Các tham số mà em sử dụng cho việc fine-tuning trên bộ dữ liệu Zalo lần lượt là:

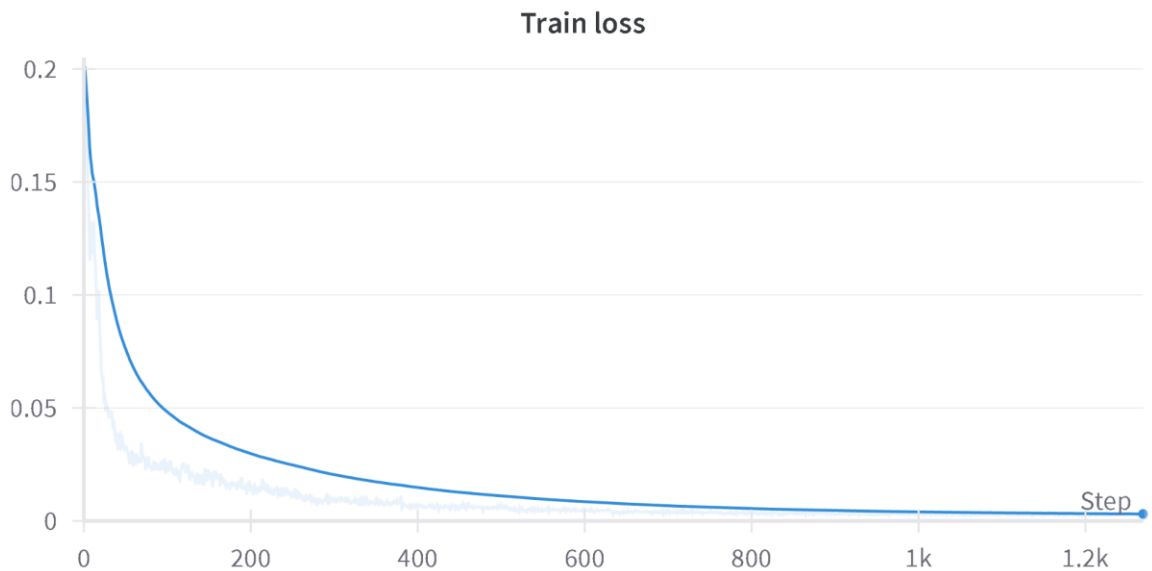
- `batch_size`: 64
- `required_grad`: True
- `epochs`: 5
- `learning_rate`: $2e-5$
- `gradient_accumulation_steps`: 8



Hình 4.19: Biểu đồ training loss Cross-Encoder - Zalo

Các tham số mà em sử dụng cho việc fine-tuning trên bộ dữ liệu Zalo lần lượt là:

- `batch_size`: 64
- `required_grad`: True
- `epochs`: 10
- `learning_rate`: $1e-4$
- `gradient_accumulation_steps`: 8



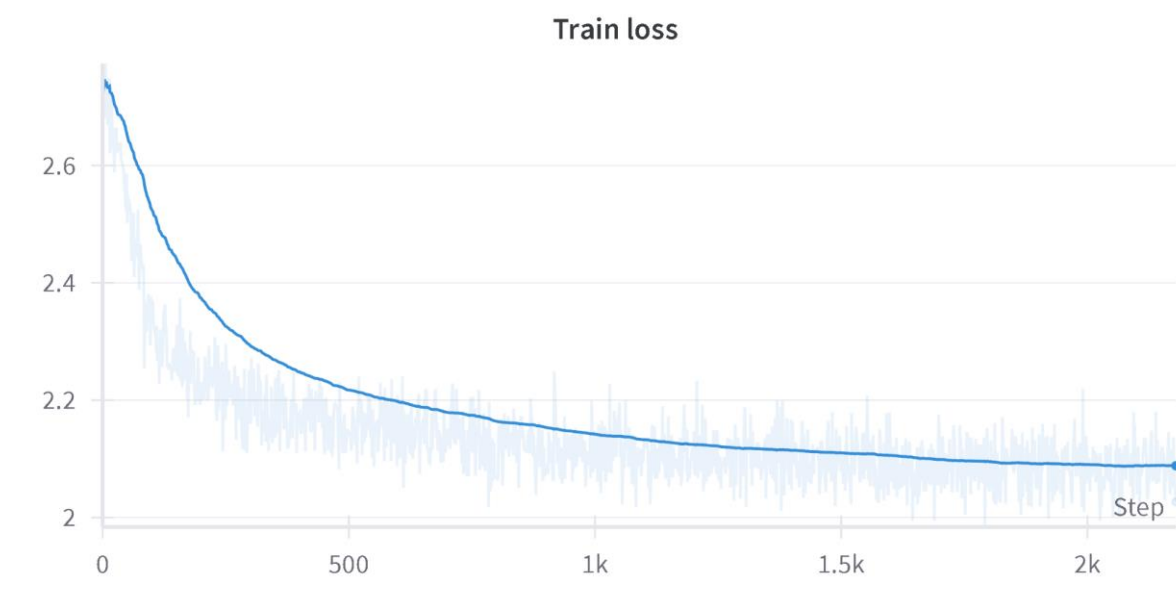
Hình 4.20: Biểu đồ training loss Cross-Encoder - STS

c. Thử nghiệm mô hình Generator Answer

Mô hình Generator Answer được ưu tiên lựa chọn là các mô hình sinh được huấn luyện trên lượng corpus lớn, đa ngôn ngữ. Hiện tại có mô hình BLOOM và MT5 được pretraining trên dữ liệu đa ngôn ngữ rất lớn. Tuy nhiên do bài toán của chúng ta là instruction fine-tuning trên bộ dữ liệu QA để có khả năng Abstract QA, vậy nên cần phải lựa chọn mô hình đủ lớn và phải được fine-tuning multi-task từ trước để khi training không quá mất nhiều thời gian hội tụ). Do đó em lựa chọn Bloomz-1b1 để fine tune tác vụ này.

Việc fine-tuning, em sử dụng adapter LoRA để fine-tuning có hiệu quả và tiết kiệm chi phí hơn, tuy nhiên do kích thước của mỗi sample là rất lớn, nên việc thử nghiệm này em có set max length= 2048, thay vì các số nhỏ hơn để tránh việc mất thông tin. Dưới đây là tham số em sử dụng, cùng với đó là kết quả training. Kết quả này em thử nghiệm trên A100 80GB:

- batch_size: 8
- epochs: 3
- learning_rate: 1e-4
- lora_rank: 16
- lora_alpha: 32
- gradient_accumulation_steps: 16
- max_length= 2048



Hình 4.21: Biểu đồ training loss Bloomz-1b7

4.1.3. Đánh giá

Để đánh giá được các mô hình sau khi huấn luyện, em sẽ đánh giá dataset thường được dùng để đánh giá, đặc biệt đối với 2 model sentence-embedding và cross-encoder sẽ sử dụng dataset STSb (sử dụng tập test) đã được dịch sang tiếng Việt.

Metrics được sử dụng dưới đây là “Spearman” (Càng cao thì càng tốt). Các mô hình base là các mô hình sử dụng thư viện cũng như kiến trúc của sentence-transformer.

Bảng 4.1: Kết quả đánh giá tác vụ Sentence Similarity

Model name	Stsb
Phobert-sentence_embedding-zalo (base)	0.228
Phobert-sentence_embedding-zalo	0.301
Phobert-sentence_embedding-zalo_sts	0.908
Phobert-cross_encoder-zalo (base)	0.433
Phobert-cross_encoder-zalo	0.539
Phobert-cross_encoder-zalo_sts	0.907

Còn lại đối với model generator answer, hiện tại chưa có cách đánh giá cụ thể bao quát cho mô hình này khi fine-tuning trên task QA để đánh giá xem liệu nó có phù hợp hay không, nhất là cần đánh giá liệu nó có ổn khi thực nghiệm trên context là knowledge của trường.. Vậy nên thay vì dùng cách metrics đánh giá tự động, em sẽ đánh giá thủ công bằng cảm quan với 20 câu hỏi và context random được trích từ dữ liệu thông tin của trường.

Các kết quả đánh giá dưới đây được em so sánh thủ công trực tiếp với answer và context, với tiêu chí là câu trả lời có thể khái quát, hoặc gần như đủ ý đối với toàn bộ context, tuy nhiên nếu trong câu trả lời có thành phần sai sẽ coi như câu trả lời đó sai, ví dụ hỏi về cách đóng tiền qua vietcombank nhưng câu trả lời liên quan đến viettel pay mà không phải vietcombank nên câu đó sai.

Bảng 4.2: Kết quả đánh giá tác vụ Question-Answering của mô hình Bloomz-1b1

Question	Context	Answer	Check
bạn hãy trình bày thực học của trường	<p>+ Triết lý giáo dục: Ứng dụng- Thực học- Thực nghiệp</p> <p>Ứng dụng: Các chương trình đào tạo được Nhà trường xây theo định hướng ứng dụng có mục tiêu và nội dung theo hướng phát triển kết quả nghiên cứu cơ bản, ứng dụng các công nghệ nguồn thành các giải pháp công nghệ, quy trình quản lý, thiết kế các công cụ hoàn chỉnh phục vụ nhu cầu của thực tiễn sản xuất.</p> <p>Thực học: Các chương trình đào tạo được xây dựng đảm bảo tỷ lệ thực hành, thực tập trong trường và ngoài doanh nghiệp chiếm từ 40% trở lên; được tổ chức dạy thật, học thật, thi thật.</p> <p>Thực nghiệp: Các chương trình đào tạo được xây dựng gắn liền với nhu cầu của doanh nghiệp trong và ngoài nước đảm bảo sinh viên được tuyển dụng ngay sau khi tốt nghiệp ra trường.</p>	trường thực học cung cấp các chương trình đào tạo được xây dựng đảm bảo tỷ lệ thực hành, thực tập trong trường và ngoài doanh nghiệp chiếm từ 40% trở lên; được tổ chức dạy thật, học thật, thi thật. Các chương trình đào tạo được xây dựng gắn liền với nhu cầu của doanh nghiệp trong và ngoài nước đảm bảo sinh viên được tuyển dụng ngay sau khi tốt nghiệp ra	PASS

		trường.	
sứ mạng của trường là gì vậy bạn	<p>+ Sứ mạng:</p> <p>“Trường Đại học Công nghệ Giao thông vận tải có sứ mạng đào tạo và cung cấp nguồn nhân lực chất lượng cao theo hướng ứng dụng, đa ngành, đa lĩnh vực, nghiên cứu khoa học và chuyển giao công nghệ phục vụ sự nghiệp phát triển của ngành Giao thông vận tải và của đất nước, phù hợp với xu thế phát triển quốc tế, hội nhập với nền giáo dục đại học tiên tiến của khu vực và trên thế giới”.</p> <p>+ Tầm nhìn:</p> <p>Đến năm 2030, có một số ngành đào tạo ngang tầm với các trường đại học có uy tín trong khu vực và trên thế giới; là trung tâm nghiên cứu khoa học ứng dụng, chuyển giao công nghệ mới và hợp tác quốc tế trong lĩnh vực Giao thông vận tải.</p> <p>Đến năm 2045, chào mừng kỷ niệm 100 năm thành lập Trường, trở thành trường đại học thông minh, trung tâm nghiên cứu khoa học, chuyển giao công nghệ và hợp tác quốc tế trong lĩnh vực Giao thông vận tải</p>	<p>Trường Đại học Công nghệ Giao thông vận tải có sứ mạng đào tạo và cung cấp nguồn nhân lực chất lượng cao theo hướng ứng dụng, đa ngành, đa lĩnh vực, nghiên cứu khoa học và chuyển giao công nghệ phục vụ sự nghiệp phát triển của ngành Giao thông vận tải và của đất nước, phù hợp với xu thế phát triển quốc tế, hội nhập với nền giáo dục đại học tiên tiến của khu vực và trên thế giới. Đến năm 2045, chào mừng kỷ niệm 100 năm thành lập Trường, trở thành trường đại học thông minh, trung tâm nghiên cứu khoa học, chuyển giao công nghệ và hợp tác quốc tế trong lĩnh vực Giao thông vận tải.</p>	PASS

<p>tóm tắt trường trong giai đoạn 1957 - 1996</p>	<p>THỜI KỲ TRƯỜNG MANG TÊN TRUNG HỌC GTVT - (1957 - 1996)\nTháng 8/1956, Nhà nước có Quyết định chuyển sang đào tạo nguồn lực lao động lớn trình độ trung cấp để phục vụ đất nước trong xây dựng hoà bình, Trường Cao đẳng công chính được tách ra thành các trường Trung cấp: Trung cấp giao thông, Trung cấp Thủy lợi và Trung cấp Kiến trúc, các trường này bắt đầu hoạt động từ đầu năm 1957. Trường Trung cấp giao thông tiếp tục tuyển sinh khoá 8 hệ đào tạo 3 năm gồm các ngành: Cầu, đường, đầu máy, toa xe, thương xa (vận tải đường sắt). Số sinh viên khoá 7 và khoá 8 được tuyển vào khoảng 300 người. Tiếp đó từ khoá 9 đến khoá 12, bình quân mỗi khoá tuyển 350 học sinh vào học. Hàng năm cho ra trường khoảng 300 cán bộ Trung cấp kỹ thuật có chất lượng tốt góp phần đáng kể vào phát triển đội ngũ cán bộ kỹ thuật cho ngành GTVT, phục vụ khôi phục mạch máu giao thông sau khi hoà bình lập lại ở Miền bắc.\nTháng 8/1960, Bộ GTVT có quyết định thành lập Ban xây dựng Trường Đại học GTVT, theo đó thầy Nguyễn Nhật Quang và một số bộ phận cán bộ lãnh đạo, giáo viên, công nhân viên của trường Trung cấp giao thông vận tải được chuyển lên công tác ở Trường Đại học giao thông vận tải.\nThực hiện Nghị quyết Đại hội Đảng toàn quốc lần thứ III (1960), sau khi đào tạo đến khoá 12, Trường Trung cấp giao thông lại tách thành 2 Trường: Trường Trung học giao thông vận tải Thủy - Bộ và Trường Trung học giao thông vận tải đường sắt.\n</p>	<p>Trường Trung học giao thông vận tải đường sắt (Trung học Giao thông Vận tải đường sắt) là trường trung cấp đào tạo nguồn lực lao động lớn trình độ trung cấp để phục vụ đất nước trong xây dựng hòa bình. Nó được thành lập vào năm 1957 và được Bộ Giao thông Vận tải (nay là Bộ Giao thông Vận tải và Xây dựng) cấp phép hoạt động từ năm 1957 đến năm 1996. Nó đã tuyển sinh khoảng 300 học sinh vào năm học đầu tiên, trong đó có khoảng 300 học sinh vào học. Sau đó, nó tuyển sinh hàng năm cho ra trường khoảng 300 cán bộ Trung cấp kỹ thuật có chất lượng tốt góp phần đáng kể vào phát triển đội ngũ cán bộ</p>	<p>PASS</p>
---	--	--	-------------

		kỹ thuật cho ngành Giao thông Vận tải, phục vụ khôi phục mạch máu giao thông sau khi hòa bình lập lại ở Miền Bắc.	
sau này trường được tách ra thành những trường nào	<p>THỜI KỲ TRƯỜNG MANG TÊN TRUNG HỌC GTVT - (1957 - 1996)</p> <p>Tháng 8/1956, Nhà nước có Quyết định chuyển sang đào tạo nguồn lực lao động lớn trình độ trung cấp để phục vụ đất nước trong xây dựng hoà bình, Trường Cao đẳng công chính được tách ra thành các trường Trung cấp: Trung cấp giao thông, Trung cấp Thủy lợi và Trung cấp Kiến trúc, các trường này bắt đầu hoạt động từ đầu năm 1957. Trường Trung cấp giao thông tiếp tục tuyển sinh khoá 8 hệ đào tạo 3 năm gồm các ngành: Cầu, đường, đầu máy, toa xe, thương xa (vận tải đường sắt). Số sinh viên khoá 7 và khoá 8 được tuyển vào khoảng 300 người. Tiếp đó từ khoá 9 đến khoá 12, bình quân mỗi khoá tuyển 350 học sinh vào học. Hàng năm cho ra trường khoảng 300 cán bộ Trung cấp kỹ thuật có chất lượng tốt góp phần đáng kể vào phát triển đội ngũ cán bộ kỹ thuật cho ngành GTVT, phục vụ khôi phục mạch máu giao thông sau khi hoà bình lập lại ở Miền bắc.</p> <p>Tháng 8/1960, Bộ GTVT có quyết định thành lập Ban xây dựng Trường Đại học GTVT, theo đó thầy Nguyễn Nhật Quang và một số bộ phận cán bộ lãnh đạo, giáo viên, công nhân viên của trường Trung cấp giao thông vận tải được chuyển lên công tác ở Trường Đại học giao thông vận tải.</p> <p>Thực hiện Nghị quyết Đại hội Đảng</p>	Trường Trung học giao thông vận tải đường sắt được tách ra thành 2 Trường: Trường Trung học giao thông vận tải đường sắt và Trường Trung học giao thông vận tải Thủy - Bộ.	PASS

	toàn quốc lần thứ III (1960), sau khi đào tạo đến khoá 12, Trường Trung cấp giao thông lại tách thành 2 Trường: Trường Trung học giao thông vận tải Thủy - Bộ và Trường Trung học giao thông vận tải đường sắt.\n		
trường có bao nhiêu cơ sở	Trường Đại học Công nghệ GTVT là trường Đại học công lập, trực thuộc Bộ Giao thông vận tải. Ngày 27/4/2011, Thủ tướng Chính phủ ký Quyết định số 630/QĐ-TTg về việc thành lập trường Đại học Công nghệ GTVT (University Of Transport Technology) trên cơ sở nâng cấp trường Cao đẳng GTVT. Hiện nay, trường có 3 cơ sở đào tạo : Cơ sở 1 tại số 54 Phố Triều Khúc- Phường Thanh Xuân Nam, Quận Thanh Xuân, TP. Hà Nội; Cơ sở 2 tại Phường Đồng Tâm, Thành phố Vĩnh Yên, Tỉnh Vĩnh Phúc và Cơ sở 3 tại Phường Tân Thịnh, TP. Thái Nguyên, Tỉnh Thái Nguyên. Trường đào tạo theo hướng ứng dụng công nghệ phục vụ chiến lược phát triển ngành GTVT và đất nước. Hiện nay Nhà trường đào tạo gần 13.000 học viên, sinh viên các hệ Tiến sĩ (02 chuyên ngành); Thạc sĩ (12 chuyên ngành); Đại học (30 chuyên ngành)\n	Trường Đại học Công nghệ GTVT có 3 cơ sở đào tạo. Cơ sở 1 tại số 54 Phố Triều Khúc- Phường Thanh Xuân Nam, Quận Thanh Xuân, TP. Hà Nội; Cơ sở 2 tại Phường Đồng Tâm, Thành phố Vĩnh Yên, Tỉnh Vĩnh Phúc và Cơ sở 3 tại Phường Tân Thịnh, TP. Thái Nguyên, Tỉnh Thái Nguyên.	PASS
bạn hãy tóm tắt nhiệm vụ của phòng thanh tra ạ	Nhiệm vụ, quyền hạn của phòng thanh tra giáo dục: - Xây dựng kế hoạch công tác thanh tra hằng năm, thanh tra đột xuất và tổ chức thực hiện sau khi được Hiệu trưởng phê duyệt; - Thanh tra việc thực hiện chính sách và pháp luật về giáo dục và đào tạo; - Thanh tra việc thực hiện mục tiêu, kế hoạch,	Nhiệm vụ của phòng thanh tra giáo dục là xây dựng kế hoạch công tác thanh tra hằng năm, thanh tra đột xuất và tổ chức thực hiện sau khi	PASS

	<p>chương trình, nội dung, phương pháp giáo dục, quy chế đào tạo, công tác tuyển sinh, quy chế thi, xét tốt nghiệp, cấp văn bằng chứng chỉ, công tác HS-SV, việc thực hiện các quy định về giáo trình, bài giảng; quản lý tài chính, quản lý tài sản; thanh tra hoạt động khoa học công nghệ và hợp tác quốc tế, công tác tổ chức cán bộ, đảm bảo chất lượng đào tạo theo quy định của Pháp luật và quy định của nhà Trường;</p> <ul style="list-style-type: none"> - Giúp Hiệu trưởng chỉ đạo, tổ chức thực hiện nhiệm vụ phòng ngừa tham nhũng, lãng phí. Triển khai Luật phòng chống tham nhũng, thực hành tiết kiệm, chống lãng phí trong lĩnh vực giáo dục và đào tạo theo quy định của pháp luật; - Chủ trì phối hợp với thanh tra nhân dân, các bộ phận chức năng giúp Hiệu trưởng giải quyết các khiếu nại, tố cáo trong lĩnh vực giáo dục đào tạo và tiếp công dân theo quy định của Pháp luật; - Thanh tra thực hiện quy định về thời gian làm việc, về kỷ luật lao động, về tác phong làm việc của cán bộ, viên chức; - Tiếp nhận đơn thư và giải quyết khiếu nại liên quan đến HS-SV; - Tổng kết công tác thanh tra, kiến nghị các biện pháp đảm bảo thi hành pháp luật về giáo dục và đào tạo; đề nghị sửa đổi, bổ sung các quy định của Trường, quy định của pháp luật về giáo dục và đào tạo; - Trưởng phòng là ủy viên một số hội đồng theo quyết định của Hiệu trưởng; - Quản lý tài sản, quản lý cán bộ viên chức trong đơn vị theo phân cấp của Hiệu trưởng; 	<p>được Hiệu trưởng phê duyệt. Họ cũng chịu trách nhiệm về việc thực hiện chính sách và pháp luật về giáo dục và đào tạo, thanh tra việc thực hiện mục tiêu, kế hoạch, chương trình, nội dung, phương pháp giáo dục, quy chế đào tạo, công tác tuyển sinh, quy chế thi, xét tốt nghiệp, cấp văn bằng chứng chỉ, công tác HS-SV, việc thực hiện các quy định về giáo trình, bài giảng, quản lý tài chính, quản lý tài sản, thanh tra hoạt động khoa học công nghệ và hợp tác quốc tế, công tác tổ chức cán bộ, đảm bảo chất lượng đào tạo theo quy định của Pháp luật và quy định của Trường, và quản lý tài sản, quản lý cán bộ viên chức trong</p>	
--	--	---	--

	<ul style="list-style-type: none"> - Trưởng phòng được quyền cho cán bộ, viên chức thuộc quyền quản lý nghỉ 01 ngày. Nghỉ từ 02 ngày phải báo cáo nhà Trường. - Hằng tháng báo cáo kết quả nhiệm vụ được giao và kế hoạch công tác tháng tới tại hội nghị giao ban hoặc bằng văn bản trình Hiệu trưởng; - Thực hiện các nhiệm vụ khác do Hiệu trưởng giao 	<p>đơn vị theo phân cấp của Hiệu trưởng. Họ cũng chịu trách nhiệm về việc thực hiện các nhiệm vụ khác do Hiệu trưởng giao</p>	
sinh viên có được hút thuốc trong trường không	<p>Nội quy học đường</p> <ol style="list-style-type: none"> 1. Sinh viên khi đến Trường phải đeo “Thẻ sinh viên”, sinh viên không đeo thẻ không được vào Trường, lớp học, giảng đường. 2. Trang phục gọn gàng, lịch sự. 3. Sinh viên ra, vào lớp phải đúng giờ quy định. Chỉ những sinh viên có tên trong danh sách lớp, môn học mới được tham dự lớp học. 4. Khi Giảng viên vào lớp sinh viên phải đứng dậy chào. Trong lớp phải nghiêm túc trật tự nghe giảng. Giúp đỡ đối xử hòa nhã với bạn bè. Không có hành vi, lời nói vô lễ với cán bộ, giảng viên, nhân viên Nhà trường. Không viết và nói những điều sai sự thật làm ảnh hưởng đến uy tín Nhà trường. Không gây rối làm ảnh hưởng tới công tác giảng dạy của giảng viên, các hoạt động của Nhà trường. 5. Không sử dụng điện thoại, không chơi Game trong lớp học và phòng thi. Không chơi bài bạc dưới mọi hình thức. 6. Không có những hành vi thiếu văn hóa như nói tục, cãi nhau, làm ồn ào mất trật tự ở nơi đông người như ở hội trường, lớp học, giảng đường, thư viện, căng tin... Nghiêm cấm không mang bất kỳ vật dụng nào có thể làm hung khí đánh nhau vào Trường. 	<p>Sinh viên không được hút thuốc trong trường.</p>	PASS

	<p>7. Hết giờ học trước khi ra khỏi lớp, sinh viên phải tự giác kê lại bàn ghế ngay ngắn, đúng vị trí, tắt đèn, tắt quạt, đóng cửa...</p> <p>8. Không dùng bút, vật sắc nhọn để viết, vẽ, rạch lên bàn ghế, bảng, tường, rèm cửa trong phòng học, giảng đường, hội trường, thư viện, căng tin, khu vệ sinh... làm hư hỏng tài sản của Nhà trường. Không chạy nhảy hoặc ngồi lên bàn; không đạp chân lên tường và lan can, hành lang lớp học. Không tự động di chuyển bàn ghế từ phòng này sang phòng khác.</p> <p>9. Phải có trách nhiệm bảo vệ môi trường cảnh quan, cây xanh, chậu cảnh trong khuôn viên Nhà trường. Phải giữ gìn vệ sinh chung; không vứt rác làm mất vệ sinh trong khuôn viên Trường.</p> <p>10. Không hút thuốc lá, ăn kẹo cao su, ăn quà bánh, cơm, cà phê...trong phòng học, giảng đường, hội trường, thư viện.</p> <p>Nếu vi phạm các quy định trên đây tùy theo mức độ sinh viên sẽ phải bồi thường thiệt hại hoặc chịu các hình thức kỷ luật từ khiển trách đến buộc thôi học.</p>		
bạn hãy trình bày về thủ tục xét miễn thi đi	<p>Điều 4. Đối tượng được miễn thi cấp chứng chỉ và công nhận đạt chuẩn đầu ra trình độ CNTT</p> <p>1. Sinh viên có chứng chỉ quốc tế MOS (Microsoft Office Specialist) hoặc IC3 (Internet and Computing Core Certification) còn hiệu lực (trong phạm vi 24 tháng tính đến thời điểm xét tốt nghiệp).</p> <p>2. Sinh viên đã có bằng tốt nghiệp đại học hệ chính quy chuyên ngành Hệ thống thông tin hoặc Truyền thông và mạng máy tính của Trường hoặc từ các trường đại học khác có chương trình đào tạo và chuẩn đầu ra tương đương với các chuyên ngành</p>	<p>Thủ tục xét miễn thi được trình bày trong Hướng dẫn này. Điều này yêu cầu sinh viên nộp đơn theo mẫu kèm theo bản sao Bằng tốt nghiệp đại học hoặc Chứng chỉ tin học cho phòng Đào tạo chậm nhất 03 tháng trước khi xét</p>	PASS

	<p>này của Trường.</p> <p>3. Thủ tục xét miễn thi: Sinh viên thỏa mãn các điều kiện theo quy định tại khoản (1) hoặc (2) của Điều này nộp đơn theo mẫu kèm theo bản sao Bảng tốt nghiệp đại học hoặc Chứng chỉ tin học cho phòng Đào tạo chậm nhất 03 tháng trước khi xét công nhận tốt nghiệp. Khi nộp đơn và bản sao văn bằng, chứng chỉ, sinh viên phải xuất trình bản gốc để kiểm tra, đối chiếu.</p> <p>Phòng Đào tạo tập hợp, trình Hiệu trưởng ra quyết định miễn thi cấp chứng chỉ và công nhận đạt chuẩn đầu ra trình độ CNTT cho các sinh viên đủ điều kiện trước khi xét công nhận tốt nghiệp.</p>	<p>công nhận tốt nghiệp.</p> <p>Khi nộp đơn và bản sao văn bằng, chứng chỉ, sinh viên phải xuất trình bản gốc để kiểm tra, đối chiếu. Phòng Đào tạo tập hợp, trình Hiệu trưởng ra quyết định miễn thi cấp chứng chỉ và công nhận đạt chuẩn đầu ra trình độ CNTT cho các sinh viên đủ điều kiện trước khi xét công nhận tốt nghiệp.</p>															
<p>khi nào sinh viên bị kỷ luật cảnh cáo vậy bạn</p>	<p>1. Quy định phân loại kết quả rèn luyện theo học kỳ</p> <p>- Điểm rèn luyện của học kỳ là Tổng điểm đạt được của 5 nội dung đánh giá ở mục I</p> <p>- Để phân loại kết quả rèn luyện, mỗi HSSV sử dụng “Phiếu đánh giá kết quả rèn luyện”, căn cứ tổng số điểm đạt được để phân loại kết quả rèn luyện thành một trong các loại: Xuất sắc, Tốt, Khá, Trung bình khá, Trung bình, Yếu và Kém, cụ thể:</p> <table><tr><td>Điểm rèn luyện</td><td>Phân loại KQRL</td></tr><tr><td>Từ 90 đến 100 điểm</td><td>Loại Xuất sắc</td></tr><tr><td>Từ 80 đến dưới 90 điểm</td><td>Loại Tốt</td></tr><tr><td>Từ 70 đến dưới 80 điểm</td><td>Loại Khá</td></tr><tr><td>Từ 60 đến dưới 70 điểm</td><td>Loại Trung bình khá</td></tr><tr><td>Từ 50 đến dưới 60 điểm</td><td>Loại Trung bình</td></tr><tr><td>Từ 30 đến dưới 50 điểm</td><td>Loại Yếu</td></tr></table>	Điểm rèn luyện	Phân loại KQRL	Từ 90 đến 100 điểm	Loại Xuất sắc	Từ 80 đến dưới 90 điểm	Loại Tốt	Từ 70 đến dưới 80 điểm	Loại Khá	Từ 60 đến dưới 70 điểm	Loại Trung bình khá	Từ 50 đến dưới 60 điểm	Loại Trung bình	Từ 30 đến dưới 50 điểm	Loại Yếu	<p>Khi sinh viên bị cảnh cáo, họ sẽ bị từ mức cảnh cáo trở lên, vì vậy họ không được vượt quá loại Trung bình.</p>	<p>PASS</p>
Điểm rèn luyện	Phân loại KQRL																
Từ 90 đến 100 điểm	Loại Xuất sắc																
Từ 80 đến dưới 90 điểm	Loại Tốt																
Từ 70 đến dưới 80 điểm	Loại Khá																
Từ 60 đến dưới 70 điểm	Loại Trung bình khá																
Từ 50 đến dưới 60 điểm	Loại Trung bình																
Từ 30 đến dưới 50 điểm	Loại Yếu																

	<p>Dưới 30 điểm Loại Kém</p> <p>- Những HSSV bị kỷ luật từ mức cảnh cáo trở lên khi phân loại kết quả rèn luyện không được vượt quá loại Trung bình.</p>		
<p>bạn hãy tóm tắt nhiệm vụ của phòng đào tạo</p>	<p>Nhiệm vụ của phòng đào tạo</p> <ul style="list-style-type: none"> - Chủ trì xây dựng, giao kế hoạch, giám sát thực hiện kế hoạch đào tạo hằng năm đã được Hiệu trưởng phê duyệt; - Tham mưu giúp Hiệu trưởng xây dựng chỉ tiêu tuyển sinh tổ chức thực hiện kế hoạch tuyển sinh theo quy chế tuyển sinh của Bộ Giáo dục và Đào tạo, Bộ Lao động - Thương binh và Xã hội; - Tham mưu xây dựng chương trình đào tạo, quản lý chất lượng đào tạo, thực hiện quy chế giảng dạy; - Tổ chức nghiệm thu giáo trình, bài giảng cho các ngành, nghề đào tạo; - Xây dựng kế hoạch xét lên lớp, dừng học, thôi học; điều kiện dự thi tốt nghiệp, tổ chức thi tốt nghiệp, xét công nhận tốt nghiệp cho người học theo quy chế đào tạo; - Quản lý lưu trữ hồ sơ, cấp phát văn bằng, chứng chỉ, chứng nhận học tập liên quan đến người học theo quy định của quy chế đào tạo; - Chủ trì xây dựng trình Hiệu trưởng phê duyệt định mức giảng dạy, theo dõi thực hiện và tính thanh toán vượt giờ cho giảng viên, giáo viên; - Chủ trì xây dựng chiến lược phát triển đào tạo; tiêu chuẩn giảng viên; bồi dưỡng giảng viên; - Đề xuất, quản lý các hoạt động đào tạo liên kết, hợp đồng giảng viên thỉnh giảng; - Soạn thảo trình Hiệu trưởng ký ban hành các văn 	<p>nhiệm vụ của phòng đào tạo là Chủ trì xây dựng, giao kế hoạch, giám sát thực hiện kế hoạch đào tạo hằng năm đã được Hiệu trưởng phê duyệt. Họ cũng tham mưu giúp Hiệu trưởng xây dựng chỉ tiêu tuyển sinh tổ chức thực hiện kế hoạch tuyển sinh theo quy chế tuyển sinh của Bộ Giáo dục và Đào tạo, Bộ Lao động - Thương binh và Xã hội. Họ cũng tham mưu xây dựng chương trình đào tạo, quản lý chất lượng đào tạo, thực hiện quy chế giảng dạy. Họ cũng tham mưu xây dựng kế hoạch xét lên lớp, dừng học, thôi học; điều kiện dự</p>	PASS

	<p>bản quản lý về lĩnh vực đào tạo, thực hiện quy chế đào tạo;</p> <ul style="list-style-type: none"> - Phối hợp với Phòng Công tác chính trị và học sinh, sinh viên, các khoa, tổ bộ môn cử giáo viên chủ nhiệm lớp, chọn Ban cán sự lớp lâm thời, quyết định thành lập các lớp; - Tham mưu, trình Hiệu trưởng tổ chức các hội nghị, hội thảo liên quan đến công tác đào tạo. 	<p>thi tốt nghiệp, tổ chức thi tốt nghiệp, xét công nhận tốt nghiệp cho người học theo quy chế đào tạo. Họ cũng tham mưu xây dựng chiến lược phát triển đào tạo, tiêu chuẩn giảng viên và bồi dưỡng giảng viên. Họ cũng tham mưu, trình Hiệu trưởng tổ chức các hội nghị, hội thảo liên quan đến công tác đào tạo.</p>	
<p>công tác sinh viên là gì</p>	<ol style="list-style-type: none"> 1. Công tác sinh viên là một trong những công tác trọng tâm của Trường, bao gồm tổng thể các hoạt động giáo dục, tuyên truyền; quản lý; hỗ trợ và dịch vụ đối với sinh viên nhằm đảm bảo các mục tiêu của giáo dục đại học. 2. Công tác sinh viên phải thực hiện đúng chủ trương, đường lối của Đảng, chính sách, pháp luật của Nhà nước và các quy định của Bộ Giáo dục và Đào tạo. 3. Công tác sinh viên phải bảo đảm dân chủ, khách quan, công bằng, công khai, minh bạch trong các vấn đề có liên quan đến sinh viên. 	<p>Công tác sinh viên là một trong những công tác trọng tâm của Trường, bao gồm tổng thể các hoạt động giáo dục, tuyên truyền; quản lý; hỗ trợ và dịch vụ đối với sinh viên nhằm đảm bảo các mục tiêu của giáo dục đại học.</p>	PASS
<p>hình thức thi tiếng</p>	<p>Điều 8. Hình thức thi, đề thi, địa điểm tổ chức thi đánh giá trình độ tiếng Anh nội bộ của Trường Đại</p>	<p>Hình thức thi tiếng anh nội bộ của trường</p>	PASS

anh nội bộ của trường là gì	<p>học Công nghệ GTVT</p> <p>1. Hình thức thi: Theo dạng chuẩn TOEIC</p> <p>2. Đề thi, đáp án, hướng dẫn chấm thi phải đáp ứng các yêu cầu sau:</p> <p>a) Đề thi đảm bảo đánh giá được các kỹ năng nghe, đọc;</p> <p>b) Các câu hỏi thi và đề thi được xây dựng theo chuẩn TOEIC, ngoài phần kiến thức chung, đề thi được bổ sung thêm một số câu hỏi liên quan đến các khối ngành mà Trường đang tổ chức đào tạo (Xem chi tiết tại phụ lục đính kèm);</p> <p>c) Đề thi, đáp án, hướng dẫn chấm thi phải được bảo mật trước, trong và sau khi thi.</p> <p>3) Địa điểm thi: Tại Trường Đại học Công nghệ GTVT.</p>	là TOEIC.	
nếu tôi không đóng học phí thì như nào	<p>Điều 23. Học phí và xử lý sinh viên vi phạm quy định đóng học phí 1. Sinh viên có nghĩa vụ đóng học phí đầy đủ theo số tín chỉ đăng ký ngay trong tháng đầu của mỗi học kỳ. Nếu kết thúc tháng đầu mỗi học kỳ mà sinh viên vẫn không đóng học phí sẽ bị loại khỏi danh sách lớp, không được tham dự kỳ thi học kỳ và nhận điểm F đối với tất cả các học phần trong học kỳ đó, đồng thời, bị xử lý kỷ luật theo Quy chế công tác sinh viên hiện hành.</p> <p>2. Sinh viên thuộc diện được xét miễn hoặc giảm học phí theo quy định chung của Bộ Giáo dục và Đào tạo phải làm hồ sơ xin miễn, giảm nộp cho phòng Đào tạo đúng hạn để Nhà trường làm các thủ tục miễn, giảm.</p>	Nếu bạn không đóng học phí, bạn sẽ bị loại khỏi danh sách lớp và không được tham dự kỳ thi học kỳ và nhận điểm F đối với tất cả các học phần trong học kỳ đó, đồng thời, bị xử lý kỷ luật theo Quy chế công tác sinh viên hiện hành.	PASS
sinh viên	Điều 22. Xử lý vi phạm đối với sinh viên	Sinh viên thi hộ sẽ bị	PASS

thi hộ thì bị phạt như nào vậy bạn	<p>1. Trong khi dự kiểm tra thường xuyên, chuẩn bị tiểu luận, bài tập lớn, đồ án môn học, thí nghiệm - thực hành, tham quan, thực tập, thi giữa học phần, thi kết thúc học phần, làm đồ án tốt nghiệp..., nếu vi phạm quy chế, sinh viên sẽ bị xử lý kỷ luật với từng học phần đã vi phạm.</p> <p>2. Sinh viên đi thi hộ hoặc nhờ người khác thi hộ, đều bị kỷ luật ở mức đình chỉ học tập một năm đối với trường hợp vi phạm lần thứ nhất và buộc thôi học đối với trường hợp vi phạm lần thứ hai.</p> <p>3. Sinh viên không đăng ký học tập, nếu không có lý do chính đáng sẽ được xem là tự ý bỏ học và phải chịu xử lý bằng một trong các hình thức: Khiển trách, cảnh cáo hoặc xóa tên theo quy định hiện hành của Trường.</p> <p>4. Người học sử dụng hồ sơ, văn bằng, chứng chỉ giả làm điều kiện trúng tuyển hoặc điều kiện tốt nghiệp sẽ bị buộc thôi học; văn bằng tốt nghiệp nếu đã được cấp sẽ bị thu hồi, huỷ bỏ.</p> <p>Ngoài trường hợp nêu tại khoản 1,2,3,4 của Điều này, tùy theo mức độ sai phạm, sinh viên sẽ bị xử lý kỷ luật theo quy định của Quy chế công tác sinh viên và Quy chế tuyển sinh đại học hệ chính quy hiện hành.</p>	phạt bằng cách đình chỉ học tập một năm đối với trường hợp vi phạm lần thứ nhất và buộc thôi học đối với trường hợp vi phạm lần thứ hai. Ngoài ra, tùy theo mức độ sai phạm, sinh viên sẽ bị xử lý kỷ luật theo quy định của Quy chế công tác sinh viên và Quy chế tuyển sinh đại học hệ chính quy hiện hành.	
điểm đồ án tốt nghiệp được tính như nào	<p>Điều 15. Chấm đồ án hoặc khóa luận tốt nghiệp</p> <p>1. Hiệu trưởng ra quyết định Hội đồng chấm đồ án hoặc khóa luận tốt nghiệp trên cơ sở đề xuất của khoa chuyên ngành. Thành phần Hội đồng chấm đồ án hoặc khóa luận tốt nghiệp cho mỗi đồ án hoặc khóa luận có từ 3-7 giảng viên;</p> <p>2. Điểm của đồ án tốt nghiệp hoặc khóa luận được</p>	Điểm đồ án tốt nghiệp được tính theo thang điểm chữ theo quy định tại khoản 3.2, Điều 10 của Quy chế này. Điểm đồ án hoặc khóa luận tốt	PASS

	<p>chấm theo thang điểm chữ theo quy định tại khoản 3.2, Điều 10 của Quy chế này. Kết quả chấm đồ án hoặc khóa luận tốt nghiệp được công bố chậm nhất là 02 tuần, kể từ ngày bảo vệ đồ án hoặc khóa luận tốt nghiệp. Điểm đồ án hoặc khóa luận tốt nghiệp được tính vào điểm TBC tích lũy của toàn khóa học.</p> <p>3. Sinh viên có đồ án hoặc khóa luận tốt nghiệp bị điểm F, phải đăng ký làm lại đồ án hoặc khóa luận tốt nghiệp.</p>	<p>nh nghiệp được tính theo thang điểm chữ theo quy định tại khoản 3.2, Điều 10 của Quy chế này. Điểm đồ án hoặc khóa luận tốt nghiệp được công bố chậm nhất là 02 tuần, kể từ ngày bảo vệ đồ án hoặc khóa luận tốt nghiệp. Điểm đồ án hoặc khóa luận tốt nghiệp được tính vào điểm TBC tích lũy của toàn khóa học.</p>	
email của trung tâm thư viện trường là gì	<p>Thư viện: địa chỉ: Tòa nhà trung tâm thư viện - Trường Đại học Công nghệ GTVT, 54 Triều Khúc- Thanh Xuân- Hà Nội. Điện thoại: (04) 35527593; Website: http://utt.edu.vn/library. Email: thuvien@utt.edu.vn</p> <p>Thư viện được Hiệu trưởng trường Đại học Công nghệ Giao thông vận tải ra quyết định thành lập số 1940/QĐ-ĐHCNGTVT ngày 04/10/2011.</p>	Email của trung tâm thư viện trường Đại học Công nghệ Giao thông vận tải là thuvien@utt.edu.vn .	PASS
tôi muốn lấy bằng thi bao giờ tôi có thể lấy vậy ạ	<p>trả hồ sơ sinh viên</p> <p>1. Đối với sinh viên thuộc diện: Thôi học, xóa tên, chuyển trường, sinh viên đã tốt nghiệp chưa lấy hồ sơ, Nhà trường trả hồ sơ sinh viên tại phòng Đào tạo thuộc cơ sở Vĩnh Phúc vào các ngày thứ 3, thứ 5 hàng tuần trong giờ hành chính (trừ các ngày nghỉ lễ).</p>	Nhà trường sẽ tổ chức lễ trao bằng và trả hồ sơ tại cơ sở đào tạo Hà Nội theo thời gian được thông báo cụ thể. Hết thời gian theo thông báo, sinh	PASS

	<p>2. Đối với sinh viên tốt nghiệp trong các đợt xét hàng năm, Nhà trường tổ chức lễ trao bằng và trả hồ sơ tại cơ sở đào tạo Hà Nội theo thời gian được thông báo cụ thể. Hết thời gian theo thông báo, sinh viên đến nhận hồ sơ mượn sẽ lấy hồ sơ tại cơ sở Vĩnh Phúc</p>	<p>viên đến nhận hồ sơ mượn sẽ lấy hồ sơ tại cơ sở Vĩnh Phúc.</p>	
<p>nộp học phí qua bên bidv như thế nào vậy bạn</p>	<p>I. Nộp học phí đối với cơ sở đào tạo Hà Nội</p> <p>TRƯỜNG HỢP 1: ĐỐI VỚI SINH VIÊN HOẶC NGƯỜI NỘP HỌC PHÍ CÓ TÀI KHOẢN TẠI BIDV HOẶC NỘP HỌC PHÍ BẰNG TIỀN MẶT, THỰC HIỆN THEO HƯỚNG DẪN SAU:</p> <p>1. Thanh toán tại quầy giao dịch của Ngân hàng BIDV</p> <p>Sinh viên/Người nộp học phí đến các quầy giao dịch của BIDV trên toàn quốc và cung cấp thông tin sau để ngân hàng vào chương trình thu Học phí cho Trường Đại học Công nghệ GTVT:</p> <p>Họ và tên sinh viên:</p> <p>Mã sinh viên:</p> <p>Nộp học phí cho trường Đại học Công nghệ GTVT</p>	<p>1. Thanh toán tại quầy giao dịch của Ngân hàng BIDV</p> <p>Sinh viên/Người nộp học phí đến các quầy giao dịch của BIDV trên toàn quốc và cung cấp thông tin sau để ngân hàng vào chương trình thu Học phí cho Trường Đại học Công nghệ GTVT: Họ và tên sinh viên: Mã sinh viên: Nộp học phí cho trường Đại học Công nghệ GTVT</p>	<p>PASS</p>
<p>tôi là sinh viên dân tộc, nhà nghèo, tôi muốn nộp hồ sơ vậy hồ sơ phải</p>	<p>II. Về chính sách hỗ trợ chi phí học tập</p> <p>Sinh viên là người dân tộc thiểu số thuộc hộ nghèo hoặc hộ cận nghèo đang theo học hệ chính quy (Đại học, Cao đẳng); không áp dụng đối với sinh viên cử tuyển, các đối tượng chính sách được xét tuyển, đào tạo theo địa chỉ, đào tạo liên thông và văn bằng hai.</p> <p>- Hồ sơ bao gồm:</p>	<p>- Đơn đề nghị hỗ trợ chi phí học tập (mẫu đơn đính kèm); + Giấy chứng nhận hộ nghèo, hộ cận nghèo do ủy ban nhân dân xã, phường, thị trấn</p>	<p>PASS</p>

gồm có gì	<ul style="list-style-type: none"> + Đơn đề nghị hỗ trợ chi phí học tập (mẫu đơn đính kèm); + Giấy chứng nhận hộ nghèo, hộ cận nghèo do ủy ban nhân dân xã, phường, thị trấn cấp (bản sao có công chứng); + Giấy khai sinh (bản sao có công chứng); + Sổ hộ khẩu (bản sao có công chứng). 	<ul style="list-style-type: none"> cấp (bản sao có công chứng); + Giấy khai sinh (bản sao có công chứng); + Sổ hộ khẩu (bản sao có công chứng). 	
-----------	---	--	--

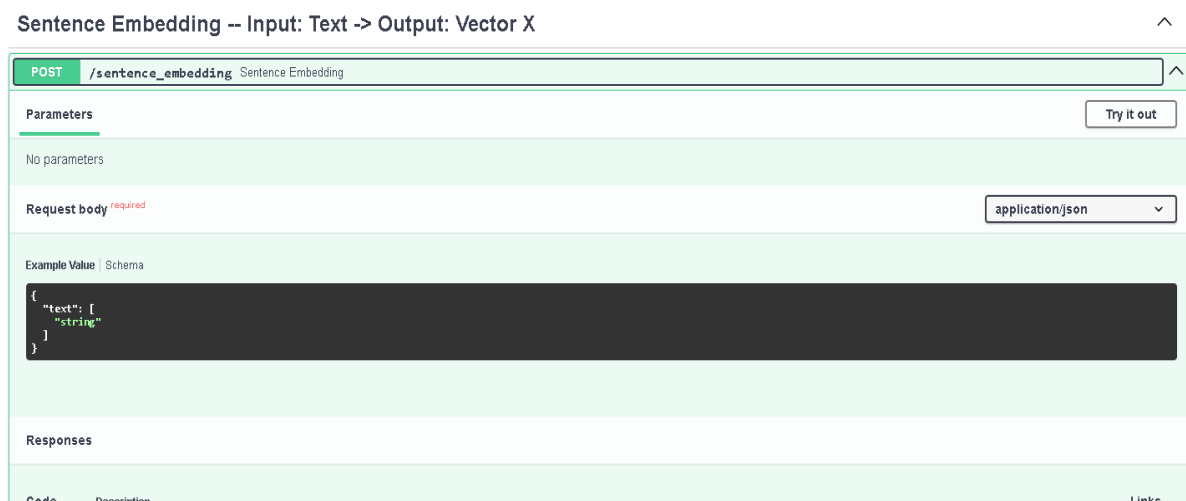
4.2. Xây dựng hệ thống

Sau khi huấn luyện và đánh giá mô hình xong, các mô hình vẫn mang tính rời rạc, nên tiếp xây dựng và triển khai hệ thống phù hợp. Hiện tại tương ứng với 3 mô hình đã được fine-tuning, tiến hành triển khai thành 3 mô đun tương ứng với 3 model, đồng thời xây dựng chúng thành giao thức API

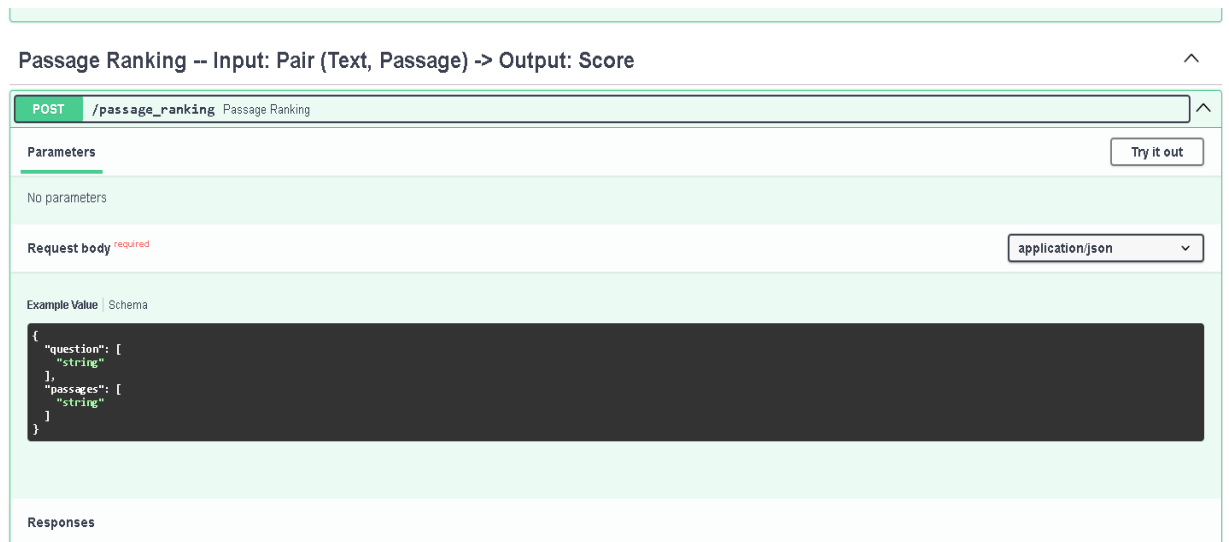
4.2.1. Thiết kế API

Tiến thiết kế API cho các 3 mô đun.

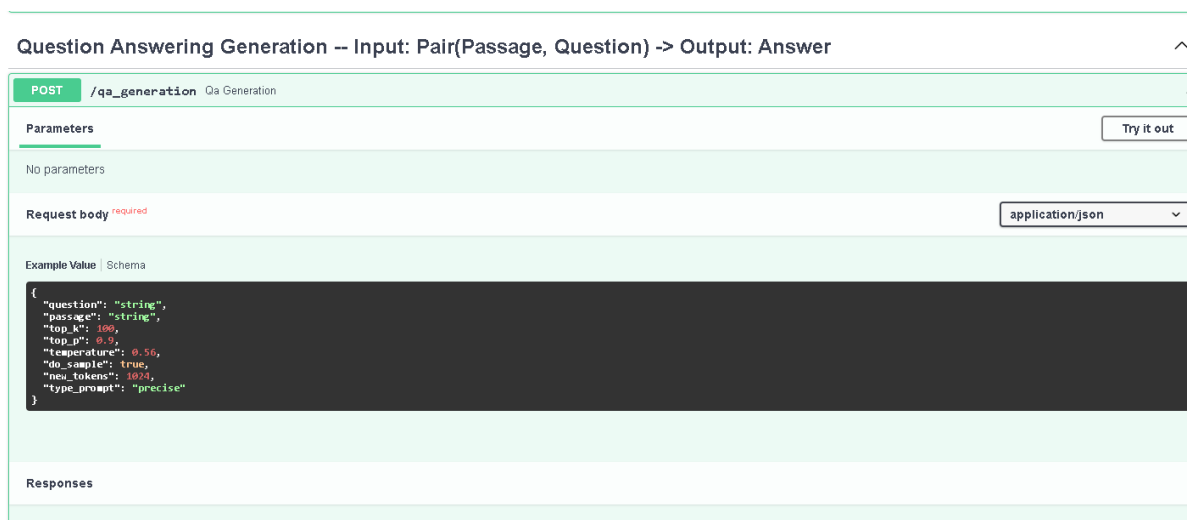
ruytenapi.json



Hình 4.22: API Sentence Embedding



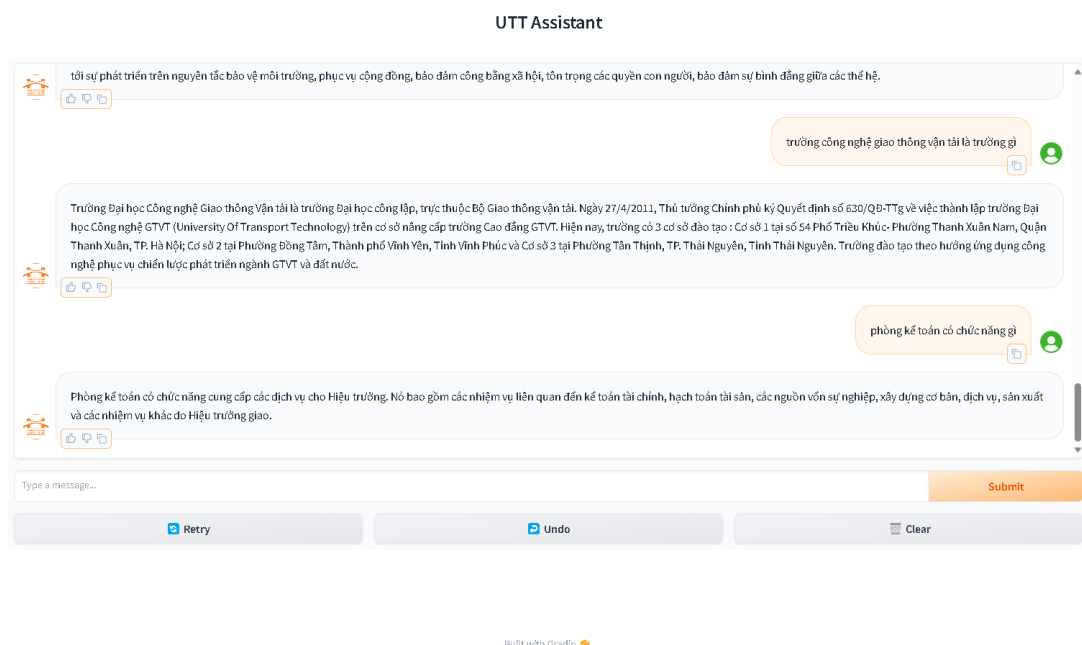
Hình 4.23: API Ranker



Hình 4.24: API Question Answering Generation

4.2.2. Thiết kế giao diện

Để có thể dễ dàng tương tác với hệ thống trợ lý ảo, em sẽ sử dụng thư viện gradio để có thể dễ dàng tạo được giao diện cho việc thử nghiệm các mô hình chatbot, trợ lý ảo mà không mất quá nhiều thời gian vào việc thiết kế cũng như độ ổn định của giao diện.



Hình 4.25: Giao diện chatbox

4.2.3. Đánh giá

Để tiến hành đánh giá chất lượng tổng thể của toàn bộ mô hình, em tiến hành chọn ngẫu nhiên khoảng 51 câu hỏi bằng thủ công, bao gồm 48 câu hỏi ngẫu nhiên liên quan đến trường bao gồm thông tin chung về trường, phòng ban, giảng viên, quy chế đào tạo, quy chế công tác sinh viên, bên cạnh đó 3 câu hỏi ngẫu nhiên về hỏi thông tin của bot. Về tiêu chí đánh giá, để được coi là đúng nếu như câu trả lời mà hệ thống phản hồi đưa ra:

- Liên quan chặt chẽ đến câu hỏi. Ví dụ hỏi về địa chỉ của trường, nhưng hệ thống đưa ra là địa chỉ của thư viện.
- Có thể đáp ứng được một phần của câu hỏi, chấp nhận câu trả lời chung chung nhưng trong câu trả lời, không được có thông tin sai. Ví dụ đặt câu hỏi về cách đóng tiền học qua vcb nhưng hệ thống đưa ra rất chung chung về việc chuyển khoản qua tài khoản nào đó nhưng không ghi rõ các bước, vẫn có thể coi là một câu trả lời đúng, tuy nhiên thả dụ số tài khoản đưa ra không khớp với knowledge ban đầu thì sẽ bị coi là sai.

Các kết quả đánh giá dưới đây được em so sánh thủ công với bộ knowledge em crawl từ trên trang chủ của trường đối với các câu hỏi đáp về trường, còn đối với câu hỏi thông tin của bot, so sánh trực tiếp vào dữ liệu giới thiệu bản thân bot mà em tự định nghĩa.

Bảng 4.3: Kết quả đánh giá hệ thống trả lời trên tác vụ hỏi đáp giới thiệu bản thân của chatbot

Question	Answer	Check
bạn tên là gì	Tôi tên là UTT Assistant, một trí tuệ nhân tạo được phát triển nhằm giải đáp thông tin cho sinh viên ĐH CN GTVT. Tôi được tạo ra bằng công nghệ học máy sâu trên nền mô hình ngôn ngữ đã được huấn luyện hàng trăm triệu điểm dữ liệu trước đó rồi được tinh chỉnh để có thể trả lời câu hỏi như hiện tại. Phiên bản đầu tiên của tôi cũng chính là ngày tôi mới sinh ra là vào ngày 10/10/2023. Tôi không có cảm xúc hay giới tính gì nhưng khả năng của tôi là luôn hỗ trợ giải đáp mọi thắc mắc cho các bạn sinh viên. Tôi hy vọng rằng có thể giúp bạn với các câu hỏi yêu cầu của bạn	Pass
giới tính bạn là gì	Tôi là một AI được tạo ra bởi công nghệ học máy sâu trên nền mô hình ngôn ngữ đã được huấn luyện hàng trăm triệu điểm dữ liệu trước đó. Tôi có khả năng trả lời câu hỏi trực tiếp, tức là tôi có thể trả lời bất kỳ câu hỏi nào được đặt ra cho tôi. Tôi không có cảm xúc hoặc giới tính nào, nhưng khả năng của tôi là luôn hỗ trợ giải đáp mọi thắc mắc cho sinh viên. Tôi hy vọng rằng có thể giúp bạn với các câu hỏi yêu cầu của bạn.	Pass
bạn được tạo ra để làm gì và bởi gì	Tôi được tạo ra bằng công nghệ học máy sâu trên nền mô hình ngôn ngữ đã được huấn luyện hàng trăm triệu điểm dữ liệu trước đó rồi được tinh chỉnh để có thể trả lời câu hỏi như hiện tại. Phiên bản đầu tiên của tôi cũng chính là ngày tôi mới sinh ra là vào ngày 10/10/2023. Tôi không có cảm xúc hay giới tính gì nhưng khả năng của tôi là luôn hỗ trợ giải đáp mọi thắc mắc cho các bạn sinh viên. Tôi hy vọng rằng có thể giúp bạn với các câu hỏi yêu cầu của bạn.	Pass

Bạn sinh năm bao nhiêu	Sinh ngày 10/10/2023.	Pass
------------------------	-----------------------	------

Bảng 4.4: Kết quả đánh giá hệ thống trả lời trên tác vụ hỏi đáp thông tin của trường

Question	Answer	
trường công nghệ giao thông vận tải là trường gì	Trường Đại học Công nghệ GTVT (University Of Transport Technology) là trường Đại học công lập, trực thuộc Bộ Giao thông vận tải.	Pass
trường đại học công nghệ giao thông vận tải có bao nhiêu cơ sở	Trường Đại học Công nghệ GTVT có 3 cơ sở đào tạo: Cơ sở 1 tại số 54 Phố Triều Khúc- Phường Thanh Xuân Nam, Quận Thanh Xuân, TP. Hà Nội; Cơ sở 2 tại Phường Đồng Tâm, Thành phố Vĩnh Yên, Tỉnh Vĩnh Phúc và Cơ sở 3 tại Phường Tân Thịnh, TP. Thái Nguyên, Tỉnh Thái Nguyên.	Pass
trường đại học công nghệ GTVT hiện đang đào tạo bao nhiêu sinh viên vậy ạ	Trường Đại học Công nghệ GTVT hiện đang đào tạo 13.000 sinh viên.	Pass
trường có bao nhiêu giảng viên?	Trường hiện có khoảng 600 giảng viên.	Pass

<p>bạn cho mình hỏi sứ mệnh và chiến lược đào tạo của trường là gì ạ</p>	<p>Trường Đại học Công nghệ Giao thông vận tải có sứ mạng đào tạo và cung cấp nguồn nhân lực chất lượng cao theo hướng ứng dụng, đa ngành, đa lĩnh vực, nghiên cứu khoa học và chuyển giao công nghệ phục vụ sự nghiệp phát triển của ngành Giao thông vận tải và đất nước, phù hợp với xu thế phát triển quốc tế, hội nhập với nền giáo dục đại học tiên tiến của khu vực và trên thế giới. Đến năm 2045, chào mừng kỷ niệm 100 năm thành lập Trường, trở thành trường đại học thông minh, trung tâm nghiên cứu khoa học, chuyển giao công nghệ và hợp tác quốc tế trong lĩnh vực Giao thông vận tải. + Tầm nhìn: Đến năm 2030, có một số ngành đào tạo ngang tầm với các trường đại học có uy tín trong khu vực và trên thế giới; là trung tâm nghiên cứu khoa học ứng dụng, chuyển giao công nghệ mới và hợp tác quốc tế trong lĩnh vực Giao thông vận tải. Đến năm 2045, chào mừng kỷ niệm 100 năm thành lập Trường, trở thành trường đại học thông minh, trung tâm nghiên cứu khoa học, chuyển giao công nghệ và hợp tác quốc tế trong lĩnh vực Giao thông vận tải.</p>	<p>Pass</p>
<p>sơ lược phát triển của trường cao đẳng công chính là gì, bạn hãy kể chi tiết nhé.</p>	<p>Trường Cao đẳng Công chính là trường cao đẳng công chính đầu tiên được thành lập ở Việt Nam. Trường được thành lập vào ngày 6 tháng 6 năm 1902, với tên gọi Trường Thư ký và cán sự chuyên môn công chính. Trường có thời gian đào tạo từ năm 1902 đến tháng 3 năm 1945. Số lượng đào tạo được rất ít, theo thống kê của lớp tiền bối đi trước thì số lượng đào tạo các hệ từ 1902 đến 1945 có 13 lớp học được 204 người, trong đó có 2 lớp công trình sư (1923 và 1925-1926) là 9 người và 1 lớp kỹ sư năm 1944 là 8 người còn lại là cán sự công chính: 187 người.</p>	<p>Pass</p>

bạn cho mình hỏi sứ mạng, tầm nhìn, giá trị cốt lõi của trường công nghệ giao thông vận tải	Sứ mạng của trường là đào tạo và cung cấp nguồn nhân lực chất lượng cao theo hướng ứng dụng, đa ngành, đa lĩnh vực, nghiên cứu khoa học và chuyển giao công nghệ phục vụ sự nghiệp phát triển của ngành Giao thông vận tải và của đất nước, phù hợp với xu thế phát triển quốc tế, hội nhập với nền giáo dục đại học tiên tiến của khu vực và trên thế giới. Trường có tầm nhìn đến năm 2030 là trung tâm nghiên cứu khoa học ứng dụng, chuyển giao công nghệ mới và hợp tác quốc tế trong lĩnh vực Giao thông vận tải. Giá trị cốt lõi của trường là đoàn kết, trí tuệ và đổi mới, coi trí tuệ là tài sản và dùng đổi mới sáng tạo để: Tối ưu hóa – Đơn giản hóa – Khác biệt hóa; xây dựng môi trường học tập và nghiên cứu thân thiện, đảm bảo và tạo điều kiện tối đa cho cán bộ, giảng viên, sinh viên, học viên, sinh viên và các đối tác của Trường. Các hoạt động của trường đều hướng tới sự phát triển trên nguyên tắc bảo vệ môi trường, phục vụ cộng đồng, bảo đảm công bằng xã hội, tôn trọng các quyền con người, bảo đảm sự bình đẳng giữa các thế hệ.	Pass
trường công nghệ giao thông vận tải được thành lập vào ngày nào	Trường Đại học Công nghệ GTVT được thành lập vào ngày 27/4/2011.	Pass
bạn cho mình hỏi phòng thư viện ở đâu vậy ạ	Phòng thư viện của Trường gồm 3 cơ sở: Trụ sở chính tại Vĩnh Yên và 2 cơ sở ở Hà Nội và Thái nguyên.	Pass
chức năng của phòng thanh tra giáo dục là gì ạ	Phòng Thanh tra giáo dục Trường Đại học Công nghệ Giao thông Vận tải là tổ chức tham mưu giúp Hiệu trưởng thực hiện công tác thanh tra, kiểm tra trong phạm vi quản lý của Hiệu trưởng nhằm bảo đảm việc thi hành pháp luật, thực hiện tốt nhiệm vụ của đơn vị, bảo vệ lợi ích của Nhà nước, quyền và lợi ích hợp pháp	Pass

	của nhà Trường, của cán bộ, giáo viên, công nhân viên, học sinh, sinh viên (HS-SV) trong lĩnh vực giáo dục và đào tạo.	
phòng kế toán có chức năng gì	Phòng kế toán có chức năng tham mưu giúp Hiệu trưởng tổ chức công tác kế toán, hạch toán chi phí đào tạo, và các hoạt động dịch vụ theo quy định của Pháp luật	Pass
bạn hãy trả lời mình nhiệm vụ của phòng đào tạo là gì ạ	Đáp: Các nhiệm vụ của Phòng Đào tạo bao gồm xây dựng các văn bản quản lý theo chức năng và nhiệm vụ của Phòng; chủ trì báo cáo các cơ quan Nhà nước có liên quan đến công tác của Phòng; quản lý tài sản, quản lý cán bộ viên chức và người lao động trong đơn vị theo phân cấp của Hiệu trưởng; thực hiện các nhiệm vụ khác do Hiệu trưởng giao; và thực hiện các nhiệm vụ khác do Hiệu trưởng giao.	Pass
địa chỉ của phòng đào tạo là gì	Phòng đào tạo được đặt tại Phòng 206 Nhà H3 - Trường Đại học Công nghệ GTVT- Số 54 Triều Khúc - Thanh Xuân- Hà Nội.	Pass
nhiệm vụ của phòng đào tạo của trường là gì ạ	nhiệm vụ của phòng đào tạo là đảm bảo rằng phòng đào tạo của trường thực hiện chiến lược phát triển đào tạo, chương trình đào tạo, thực hiện quy chế và kế hoạch đào tạo, công tác tuyển sinh, xét dừng học, thôi học, công nhận tốt nghiệp các hệ đào tạo trong trường, và quản lý các hoạt động đào tạo và quản lý.	Pass
bạn hãy nêu rõ nhiệm vụ của phòng đào tạo sau đại học ạ	Nhiệm vụ của phòng đào tạo sau đại học là đề xuất, phối hợp với các khoa, bộ môn chuyên ngành xây dựng chiến lược phát triển đào tạo, chương trình đào tạo SDH phù hợp với định hướng phát triển của Trường,	Pass

	nhu cầu xã hội và quy định của Bộ Giáo dục và Đào tạo.	
phòng đào tạo là gì	Phòng Đào tạo là đơn vị trực thuộc Trường, có chức năng tham mưu giúp Hiệu trưởng thực hiện công tác tổ chức và quản lý đào tạo trình độ đại học và cao đẳng hệ chính quy; công tác chính trị, tư tưởng, công tác quản lý sinh viên, công tác tuyên truyền, văn hóa, truyền thống.	Pass
địa chỉ của phòng công tác sinh viên là gì	Phòng Công tác sinh viên: địa chỉ: Phòng 102, 104 Nhà H3 - Trường Đại học Công nghệ GTVT - Số 54 Triều Khúc - Thanh Xuân - Hà Nội. Website: http://utt.edu.vn/ct-hssv . Email: phongcthssv@utt.edu.vn .	Pass
chức năng của phòng khảo thí là gì vậy ạ	Phòng Khảo thí và Đảm bảo chất lượng đào tạo là đơn vị trực thuộc Trường, có chức năng tham mưu giúp Hiệu trưởng và chịu trách nhiệm tổ chức triển khai thực hiện hoạt động về công tác khảo thí, công tác đánh giá và đảm bảo chất lượng đào tạo trong Nhà trường.	Pass
email của cô Đoàn Thị Thanh Hằng là gì ạ	Email của cô Đoàn Thị Thanh Hằng là hangdtt@utt.edu.vn	Pass
cô Đoàn Thị Thanh Hằng thuộc đơn vị nào	Ths. Đoàn Thị Thanh Hằng thuộc đơn vị Khoa Công nghệ thông tin.	Pass
bạn hãy cho mình biết cô Đoàn Thị Thanh Hằng có học hàm, học vị là gì ạ	Ths. Đoàn Thị Thanh Hằng có học hàm, học vị là Thạc sĩ.	Pass

bạn ơi, mình muốn hỏi đối tượng nào được xét học bổng ạ	Đối tượng được xét, cấp học bổng khuyến khích học tập: Sinh viên Đại học, Cao đẳng hệ chính quy; HSSV TCCN hệ chính quy; HSSV hệ đào tạo nghề chính quy đang theo học tại Trường Đại học Công nghệ Giao thông vận tải.	Pass
bạn có thể trình bày về nội quy học đường được không ạ	Hành vi sinh viên không được làm là 1. Xúc phạm nhân phẩm, danh dự, xâm phạm thân thể nhà giáo, cán bộ quản lý, viên chức, nhân viên, người học của Trường và người khác. 2. Gian lận trong học tập, kiểm tra, thi cử như: quay cốp, mang tài liệu vào phòng thi, xin điểm; học, thi, thực tập, trực hộ người khác hoặc nhờ người khác học, thi, thực tập, trực hộ; sao chép, nhờ hoặc làm hộ tiểu luận, đồ án, khóa luận tốt nghiệp; tổ chức hoặc tham gia tổ chức thi hộ hoặc các hành vi gian lận khác. 3. Hút thuốc, uống rượu, bia trong trường học; say rượu, bia khi lớp học	Pass
kết quả rèn luyện của sinh viên có thể được dùng làm gì	kết quả rèn luyện của sinh viên có thể được sử dụng để đánh giá ý thức học tập của sinh viên.	Pass
công tác sinh viên là gì	Công tác sinh viên là một trong những công tác trọng tâm của Trường, bao gồm tổng thể các hoạt động giáo dục, tuyên truyền; quản lý; hỗ trợ và dịch vụ đối với sinh viên nhằm đảm bảo các mục tiêu của giáo dục đại học.	Pass
bạn hãy trình bày quyền của sinh viên	Quyền của sinh viên được tôn trọng và đối xử bình đẳng, được cung cấp đầy đủ thông tin cá nhân về việc học tập, rèn luyện theo quy định của Trường, được phổ biến nội quy, quy chế về đào tạo, rèn luyện và các chế độ, chính sách của Nhà nước có liên quan đến sinh viên.	Pass

trách nhiệm và quyền lợi của sinh viên là gì ạ	Trách nhiệm và quyền lợi của sinh viên là đảm bảo rằng sinh viên có thể tiếp cận các nguồn tài trợ, chính sách hỗ trợ, miễn giảm học phí của Nhà nước, Nhà trường và các nguồn tài trợ khác.	Pass
hồ sơ xử lý kỷ luật của sinh viên gồm những gì vậy bạn	Hồ sơ xử lý kỷ luật của sinh viên: a) Bản tự kiểm điểm; b) Biên bản của tập thể lớp sinh viên họp kiểm điểm sinh viên có hành vi vi phạm; c) Tờ trình đề nghị xử lý kỷ luật sinh viên vi phạm của phòng công tác HSSV; d) Các tài liệu có liên quan.	Pass
trình bày thủ tục xét kỷ luật sinh viên	Thủ tục xét kỷ luật: a) Sinh viên có hành vi vi phạm phải làm bản tự kiểm điểm và tự nhận hình thức kỷ luật. Trong trường hợp sinh viên không chấp hành làm bản tự kiểm điểm thì Hội đồng khen thưởng và kỷ luật sinh viên vẫn họp để xử lý trên cơ sở các chứng cứ thu thập được; b) Giáo viên chủ nhiệm-Cố vấn học tập lớp sinh viên chủ trì họp với tập thể lớp sinh viên, phân tích và đề nghị hình thức kỷ luật gửi Phòng Công tác HSSV; c) Phòng Công tác HSSV xem xét, đề nghị Hội đồng khen thưởng và kỷ luật sinh viên của Trường;	Pass
bạn hãy trình bày công tác quản lý sinh viên	Công tác quản lý sinh viên của Phòng Công tác sinh viên - Thực hiện các chế độ chính sách của Nhà nước quy định đối với sinh viên về: Học bổng, học phí, trợ cấp xã hội, bảo hiểm, tín dụng đào tạo và các chế độ khác có liên quan đến sinh viên; - Phối hợp với Phòng Đào tạo, các khoa tổ chức tiếp nhận thí sinh trúng tuyển, thành lập lớp, cử ban cán sự lớp lâm thời, bố trí giáo viên chủ nhiệm lớp; - Chủ trì làm thẻ cho cán bộ, nhân viên và người học trong Trường; - Chủ trì, phối hợp với các khoa xây dựng cơ sở dữ liệu quản lý hồ sơ sinh viên theo quy định của Bộ Giáo dục và Đào tạo,	Pass

<p>khi nào hồ sơ sinh viên được trả về</p>	<p>trả hồ sơ sinh viên 1. Đối với sinh viên thuộc diện: Thôi học, xóa tên, chuyển trường, sinh viên đã tốt nghiệp chưa lấy hồ sơ, Nhà trường trả hồ sơ sinh viên tại phòng Đào tạo thuộc cơ sở Vĩnh Phúc vào các ngày thứ 3, thứ 5 hàng tuần trong giờ hành chính (trừ các ngày nghỉ lễ). 2. Đối với sinh viên tốt nghiệp trong các đợt xét hàng năm, Nhà trường tổ chức lễ trao bằng và trả hồ sơ tại cơ sở đào tạo Hà Nội theo thời gian được thông báo cụ thể.</p>	<p>Pass</p>
<p>điểm học phần được tính như thế nào</p>	<p>điểm học phần được tính từ tổng các điểm thành phần nhân với trọng số tương ứng làm tròn đến một chữ số thập phân.</p>	<p>Pass</p>
<p>sinh viên bỏ thi kết thúc học phần có làm sao không</p>	<p>Sinh viên được dự thi kết thúc học phần nhưng bỏ thi không có lý do chính đáng hoặc điểm học phần không đạt trong lần thi đầu tiên của mỗi lần học phải tham dự kỳ thi lại do Trường tổ chức. Trong mỗi lần học, nếu sinh viên phải thi lại thì điểm học phần sau khi thi lại của lần học đó chỉ được giới hạn từ mức điểm C trở xuống.</p>	<p>Pass</p>
<p>trình bày về tổ chức kỳ thi kết thúc học phần</p>	<p>Kỳ thi kết thúc học phần là kỳ thi chính và một kỳ thi phụ vào cuối học kỳ để thi kết thúc học phần. Kỳ thi phụ dành cho những sinh viên không tham dự kỳ thi chính hoặc có học phần bị điểm F ở kỳ thi chính và được tổ chức sớm nhất là hai tuần sau kỳ thi chính.</p>	<p>Pass</p>

sinh viên có thể đăng ký học cùng lúc hai chương trình như thế nào, điều kiện ?	Học cùng lúc hai chương trình dành cho sinh viên có đủ điều kiện theo quy định tại khoản 2 Điều này và có nhu cầu đăng ký học thêm một chương trình thứ hai để khi tốt nghiệp được cấp hai văn bằng. Điều này bao gồm sinh viên được đăng ký học chương trình thứ hai sớm nhất khi đã được xếp trình độ năm thứ hai của chương trình thứ nhất. Điều này cũng áp dụng cho sinh viên được xếp trình độ năm thứ hai của chương trình thứ nhất, nếu điểm trung bình tích lũy xếp loại khá trở lên và đáp ứng ngưỡng bảo đảm chất lượng của chương trình thứ hai trong năm tuyển sinh.	Pass
sinh viên được làm đồ án tốt nghiệp khi nào	sinh viên được làm đồ án hoặc khóa luận tốt nghiệp khi họ còn nợ tối đa 10% tổng số tín chỉ quy định cho toàn chương trình (ngoại trừ các học phần đang hoàn thiện trong kỳ học hiện tại, thực tập tốt nghiệp và đồ án hoặc khóa luận tốt nghiệp).	Pass
trường hợp sinh viên không nộp học phí sẽ bị xử lý như thế nào	trường hợp sinh viên không nộp học phí sẽ bị xử lý như thế nào. Nếu sinh viên vi phạm quy định về đóng học phí, họ sẽ bị loại khỏi danh sách lớp, không được tham dự kỳ thi học kỳ và nhận điểm F đối với tất cả các học phần trong học kỳ đó, đồng thời, bị xử lý kỷ luật theo Quy chế công tác sinh viên hiện hành.	Pass
những ai có thể được học liên thông với văn bằng khác vậy ạ	Những người có thể được học liên thông với văn bằng khác là những người đã tốt nghiệp đại học và dự tuyển và học để nhận thêm một bằng tốt nghiệp đại học của một ngành đào tạo khác.	Pass

điều kiện để sinh viên có thể được chuyển ngành là gì	Chuyển ngành, chuyển nơi học, chuyển cơ sở đào tạo, chuyển hình thức học. Sinh viên được xem xét chuyển sang học một chương trình, một ngành đào tạo khác, hoặc một cơ sở khác của Trường cần đủ các điều kiện sau: a) Không đang là sinh viên trình độ năm thứ nhất hoặc năm cuối khóa, không thuộc diện bị xem xét buộc thôi học và còn đủ thời gian học tập theo quy định tại khoản 5 Điều 2 của Quy chế này; b) Sinh viên đạt điều kiện trúng tuyển của chương trình, ngành đào tạo, của cơ sở đào tạo hoặc Phân hiệu cần chuyển đến trong cùng khóa tuyển sinh;	Pass
tôi muốn được hỗ trợ học phí học tập thì như thế nào	Hồ sơ cần bao gồm Giấy chứng nhận hộ nghèo, hộ cận nghèo do ủy ban nhân dân xã, phường, thị trấn cấp (bản sao có công chứng).	Pass
đối tượng được áp dụng miễn giảm học phí	Đối tượng được áp dụng miễn giảm học phí là sinh viên thuộc hộ nghèo hoặc hộ cận nghèo đang theo học hệ chính quy (Đại học, Cao đẳng); không áp dụng đối với sinh viên cử tuyển, các đối tượng chính sách được xét tuyển, đào tạo theo địa chỉ, đào tạo liên thông và văn bằng hai.	Pass
tôi muốn nộp hồ sơ miễn giảm học phí thì nộp qua đâu	Đợt 1: từ tháng 8 đến ngày 15/10 hằng năm. + Đợt 2: từ tháng 01 đến 15/3 hằng năm. - Địa điểm: Phòng Công tác sinh viên tại các cơ sở đào tạo (đối với sinh viên tại cơ sở Hà Nội, nộp hồ sơ tại Văn phòng “Một cửa”, 104 - H2).	Pass
đối tượng nào được đăng ký thi tiếng anh nội bộ	Sinh viên đại học hệ chính quy của Trường Đại học Công nghệ Giao thông vận tải có nhu cầu được đánh giá năng lực ngoại ngữ.	Pass

hình thức thi tiếng anh nội bộ như thế nào	Hình thức thi tiếng anh nội bộ được quy định bởi Điều 8.1 của Quy định này. Theo dạng chuẩn TOEIC, đề thi, đáp án, hướng dẫn chấm thi phải đáp ứng các yêu cầu sau: a) Đề thi đảm bảo đánh giá được các kỹ năng nghe, đọc; b) Các câu hỏi thi và đề thi được xây dựng theo chuẩn TOEIC, ngoài phần kiến thức chung, đề thi được bổ sung thêm một số câu hỏi liên quan đến các khối ngành mà Trường đang tổ chức đào tạo (Xem chi tiết tại phụ lục đính kèm); c) Đề thi, đáp án, hướng dẫn chấm thi phải được bảo mật trước, trong và sau khi thi.	Pass
quy trình xin miễn thi đánh giá tiếng anh là gì	Sinh viên cần cung cấp một bản sao công chứng chứng chỉ tiếng Anh, bằng đại học, giấy tờ liên quan. Khi nộp đơn, sinh viên cần xuất trình bản gốc các loại chứng chỉ, bằng đại học và giấy tờ liên quan để kiểm tra, đối chiếu. Phòng Đào tạo phối hợp với Bộ môn Ngoại ngữ Anh - Pháp kiểm tra, gửi văn bản xác minh tới các đơn vị tổ chức thi. Căn cứ kết quả xác minh phòng Đào tạo nhập kết quả để xét tốt nghiệp cho sinh viên.	Pass
trình bày cho mình hỏi về quy trình coi thi tiếng anh nội bộ như nào vậy ạ	Quy trình coi thi tiếng Anh nội bộ của Trường Đại học Công nghệ GTVT có thể được tóm tắt như sau: Trước ngày thi, Hội đồng thi và Ban coi thi kiểm tra các điều kiện tổ chức thi đảm bảo các yêu cầu sau: + Khu vực thi phải được bố trí an toàn, yên tĩnh; phải đảm bảo có hệ thống âm thanh, ánh sáng đạt chất lượng theo yêu cầu; + Các thí sinh trong phòng thi được bố trí ngồi số báo danh và theo quy tắc bố trí sơ đồ phòng thi ngẫu nhiên theo từng buổi thi.	Pass

<p>trình bày các trường hợp nào được công nhận chuẩn đầu ra trình độ ngoại ngữ</p>	<p>Trường hợp được công nhận chuẩn đầu ra trình độ ngoại ngữ 1. Sinh viên có kết quả thi đánh giá trình độ tiếng Anh nội bộ của Trường Đại học Công nghệ GTVT theo dạng chuẩn TOEIC đạt từ 450 điểm trở lên. 2. Sinh viên có bằng tốt nghiệp đại học ngoại ngữ hệ chính quy chuyên ngành tiếng Anh, Pháp, Nga, Đức, Trung, Nhật, Hàn, Tây Ban Nha, Ả Rập hoặc bằng tốt nghiệp đại học theo chương trình đào tạo bằng một trong các thứ tiếng nêu trên do các trường đại học có uy tín, được chứng nhận kiểm định chất lượng đào tạo cấp. 3. Sinh viên có một trong các chứng chỉ tiếng Anh quốc tế (TOEIC, IELTS, TOEFL, Cambridge Tests), do các tổ chức đánh giá nước ngoài được Bộ Giáo dục và Đào tạo công nhận cấp, đáp ứng mức tối thiểu được quy định tại điểm a hoặc điểm b của khoản 3, Điều 2 nêu trên. 4. Sinh viên có chứng chỉ tiếng Anh theo khung năng lực ngoại ngữ 6 bậc dành cho Việt Nam do các trường đại học của Việt Nam được Bộ Giáo dục và Đào tạo công nhận cấp, đáp ứng mức tối thiểu được quy định tại điểm a hoặc điểm b của khoản 3, Điều 2 nêu trên. 5. Sinh viên có chứng chỉ tiếng Anh theo khung năng lực ngoại ngữ 6 bậc dành cho Việt Nam do các tổ chức đánh giá nước ngoài được Bộ Giáo dục và Đào tạo Việt Nam công nhận cấp, đáp ứng mức tối thiểu được quy định tại điểm a hoặc điểm b của khoản 3, Điều 2 nêu trên.</p>	<p>Pass</p>
--	--	-------------

<p>cách nộp tiền học qua viettel pay</p>	<p>Trường Đại học Công nghệ Giao thông Vận tải đã triển khai một cách nộp tiền học phí qua Viettel Pay. Sinh viên thanh toán học phí qua Viettel Pay bằng cách đăng ký tài khoản và tải ứng dụng ViettelPay. Bước 1: Đăng ký tài khoản ViettelPay ONLINE (theo đường dẫn trên website/fanpage Nhà trường) & Tải ứng dụng ViettelPay ==> sau đó tạo tài khoản Bước 2: Đăng nhập tài khoản ViettelPay trên application nạp tiền (hướng dẫn kèm sau) Bước 3: Tìm kiếm “ĐH Công nghệ GTVT” Bước 4: Nhập Mã sinh viên Bước 5: Xác nhận số tiền cần thanh toán ==> nhập mã PIN & OTP xác nhận thanh toán Hoặc: Ra hệ thống cửa hàng giao dịch Viettel để được hướng dẫn và hỗ trợ cài đặt. Liên hệ hotline CSKH: 18009000 để được tư vấn hỗ trợ. Mọi thắc mắc liên hệ P.Tài chính – Kế toán (P. 209 Nhà H1). SĐT: 0243 2 216 298.</p>	<p>Pass</p>
--	---	-------------

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết quả đạt được

Xây dựng được thành công hệ thống hệ thống trả lời thông tin tự động cho Trường Công nghệ Giao thông Vận tải bằng mô hình transformers với các chức năng chính như hỏi đáp về quy trình đào tạo, hỏi đáp về vấn đề công tác sinh viên, hỏi đáp tương tác với bot. Hệ thống có thể trả lời các câu hỏi của người dùng tương đối chính xác, hiểu ý của người dùng một cách chuẩn xác và đặc biệt tương tác với người dùng một cách tự nhiên.

2. Hướng phát triển

Để nâng cao chất lượng của hệ thống, có thể thực hiện các hướng phát triển sau:

- Mở rộng bộ dữ liệu huấn luyện: Bộ dữ liệu huấn luyện hiện tại còn hạn chế, cần mở rộng bộ dữ liệu này để hệ thống có thể trả lời các câu hỏi một cách chính xác hơn.
- Tăng cường khả năng xử lý câu hỏi mở: Hệ thống hiện tại có thể xử lý tốt các câu hỏi đóng, cần tăng cường khả năng xử lý các câu hỏi mở để hệ thống có thể trả lời các câu hỏi đa dạng hơn.
- Tăng cường khả năng tương tác: Hệ thống hiện tại có thể tương tác với người dùng một cách tự nhiên, tuy nhiên, có thể tăng cường khả năng tương tác này để hệ thống có thể tạo ra trải nghiệm tốt hơn cho người dùng.

3. Hạn chế

Hệ thống vẫn còn một số hạn chế sau:

- Độ chính xác: Độ chính xác của hệ thống vẫn chưa đạt 100%, cần cải thiện hơn nữa để hệ thống có thể trả lời các câu hỏi một cách chính xác nhất.
- Khả năng hiểu ý: Khả năng hiểu ý của hệ thống vẫn chưa hoàn hảo, cần cải thiện để hệ thống có thể hiểu ý của người dùng một cách chính xác hơn.
- Khả năng tương tác: Khả năng tương tác của hệ thống vẫn chưa hoàn toàn tự nhiên, cần cải thiện để hệ thống có thể tạo ra trải nghiệm tốt hơn cho người dùng.

4. Tổng kết

Đồ án đã đạt được những kết quả nhất định, tuy nhiên, vẫn còn một số hạn chế cần được khắc phục. Với những định hướng phát triển được đề xuất, hệ thống sẽ ngày càng hoàn thiện và đáp ứng tốt hơn nhu cầu của người dùng.

TÀI LIỆU THAM KHẢO

- [1] Vaswani, A., et al. "Attention is all you need." *Advances in Neural Information Processing Systems*. 2017.
- [2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- [3] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. "Language models are unsupervised multitask learners." *OpenAI Blog* 1 (8) (2018).
- [4] Nguyen, T. T., et al. "PhoBERT: A vietnamese bert model for natural language processing." *arXiv preprint arXiv:2201.05237* (2022).
- [5] Devlin, J., et al. "The sentence transformer." *arXiv preprint arXiv:1901.11117* (2019).
- [6] Muennighoff, Niklas, et al. "Crosslingual Generalization through Multitask Finetuning." arXiv preprint arXiv:2211.01786, 2022.