DOI: https://doi.org/10.59294/HIUJS.KHQG.2024.001

PHÁT TRIỂN TRỢ LÝ ẢO THÔNG MINH BẰNG MÔ HÌNH NGÔN NGỮ LỚN HỖ TRỢ GIẢNG DẠY

Trần Ngọc Oanh, Bùi Công Tuấn, Nguyễn Việt Phương, Hồ Nguyễn Ngọc Bảo, Nguyễn Song Thiên Long, Bùi Hoài Thắng và Quản Thành Thơ*

Trường Đại học Bách Khoa, Đại học Quốc gia TP.HCM

TÓM TẮT

Bài báo này trình bày về việc phát triển một Trợ lý Ảo Thông minh cho lĩnh vực giáo dục. Trợ lý ảo này được phát triển dựa trên các kỹ thuật AI tiên tiến nhất hiện nay bao gồm các Mô hình Ngôn ngữ Lớn (LLMs), Đồ thị Tri thức (Knowledge Graph) và các kỹ thuật RAG (Retrieval Augmented Generation). Trước tiên, chúng tôi thảo luận việc xây dựng KG từ dữ liệu học vụ thực tế từ nhiều nguồn của Trường Đại học Bách Khoa – ĐHQG TPHCM (HCMUT). Đặc biệt, chúng tôi nhấn mạnh vào việc sử dụng các kỹ thuật học máy để phát hiện các ý định mở (open intent) từ các trao đổi học vụ của sinh viên và nhân viên của Nhà trường để phát triển một KG có tính thực tế. Tiếp theo, KG này được kết hợp với một LLM thông qua kỹ thuật RAG để hiện thực một trợ lý ảo có khả năng giao tiếp và hướng dẫn sinh viên các vấn đề học thuật. Trợ lý ảo này hứa hẹn sẽ biến đổi các hoạt động giao tiếp giáo dục truyền thống. Bằng cách cung cấp các gợi ý từ các nguồn tài nguyên giáo dục của HCMUT và tham gia vào cuộc trò chuyện một cách tự nhiên, hệ thống này giúp Nhà trường truyền đạt những kiến thức học vụ một cách tương tác và hiệu quả hơn. Hơn nữa, nó thúc đẩy sự tham gia và tự chủ của học sinh bằng cách cung cấp hỗ trợ và phản hồi được tùy chỉnh trong quá trình học. Chúng tôi cũng đã đánh giá hiệu quả của hệ thống bằng các phương pháp học máy thông qua các cơ chế phản hồi của người dùng và các chỉ số hiệu suất của học máy. Nghiên cứu này đóng góp vào việc tiến xa hơn trong các công nghệ giáo dục được điều khiển bởi trí tuệ nhân tạo, mở ra con đường cho những trải nghiệm học tập linh hoạt và cá nhân hóa hơn trong thời đại số.

***Từ khóa**: Trợ lý ảo giáo dục, Mô hình ngôn ngữ lớn, Đồ thị Tri thức*

DEVELOPMENT OF INTELLIGENT VIRTUAL ASSISTANTS USING LARGE LANGUAGE MODELS TO SUPPORT ACADEMIC ACTIVITIES

Tran Ngoc Oanh, Bui Cong Tuan, Nguyen Viet Phuong, Ho Nguyen Ngoc Bao, Nguyen Song Thien Long, Bui Hoai Thang and Quan Thanh Tho

ABSTRACT

This paper presents the development of an Intelligent Virtual Assistant for the education domain. This virtual assistant is developed based on the most advanced AI techniques currently available, including Large Language Models (LLMs), Knowledge Graphs (KG), and Retrieval Augmented Generation (RAG) techniques. First, we discuss building a KG from real academic data from multiple sources at Ho Chi Minh City University of Technology (HCMUT). In particular, we emphasize using machine learning techniques to detect open intents from academic interactions between students and staff to develop a realistic KG. Next, this KG is combined with an LLM via the RAG technique to realize a virtual assistant capable of communicating and guiding students on academic issues. This virtual assistant promises to transform traditional educational communication activities. By

*Tác giả liên hệ: PGS.TS Quản Thành Thơ, Email**:** qttho@hcmut.edu.vn
(Ngày nhận bài: 15/04/2024; Ngày nhận bản sửa: 02/05/2024; Ngày duyệt đăng: 04/05/2024)

providing suggestions from HCMUT's educational resources and engaging in natural conversation, this system helps the university convey academic knowledge more interactively and effectively. Moreover, it promotes student engagement and autonomy by providing customized support and feedback throughout the learning process. We also evaluated the system's effectiveness using machine learning methods through user feedback mechanisms and machine learning performance metrics. This research contributes to further advancements in AI-driven educational technologies, paving the way for more flexible and personalized learning experiences in the digital age.

Keywords: *Educational Virtual Assistant, Large Language Model, Knowledge Graph*

1. ĐẶT VẤN ĐỀ

1.1. Mô hình ngôn ngữ lớn

Một Mô hình Ngôn ngữ Lớn (LLM) là một mô hình ngôn ngữ được đào tạo trên một tập hợp văn bản khổng lồ. Những mô hình này đã thu hút được sự chú ý từ các cộng đồng khác nhau nhờ khả năng thực hiện các nhiệm vụ tạo sinh ngôn ngữ và hiểu được nhiều loại ngôn ngữ. Các ví dụ đáng chú ý về LLM bao gồm GPT của OpenAI [1], BART của Google [2] và LLaMa của Meta [3]. Để đạt được kiến thức tổng quát sâu rộng [4] và cải thiện khả năng sử dụng ngôn từ [5], các LLM tiên tiến ngày nay được đào tạo trên các tập hợp văn bản rất lớn.

Tuy nhiên, những LLM này có thể chứa đựng "ảo giác" [6] do hạn chế trong việc tiếp cận thông tin mới nhất, chính xác hoặc thiếu kiến thức chuyên môn trong các lĩnh vực. Để giải quyết vấn đề này và các hạn chế khác, một số mô hình dung hợp đã cho thấy kết quả ấn tượng bằng cách kết hợp bộ nhớ tham số với bộ nhớ không tham số [7]. Vì cơ sở tri thức của những mô hình này có thể được chỉnh sửa và mở rộng trực tiếp nên kiến thức được truy xuất có thể được con người kiểm tra và diễn giải. Kỹ thuật này, được gọi là Retrieval-Augmented Generation (RAG) [8], lần đầu tiên được nhóm nghiên cứu tại Facebook công bố trong một bài báo nổi tiếng.

1.2. Đồ thị Tri thức

Trong lĩnh vực biểu diễn tri thức, một Đồ thị Tri thức (KGs) [9] là một ngân hàng tri thức được biểu diễn dưới dạng đồ thị, trong đó các khái niệm hoặc thực thể được biểu thị bằng các nút và các mối quan hệ giữa chúng được biểu diễn bằng các cạnh nối các nút. Đồ thị Tri thức cho phép biểu diễn tri thức một cách trực quan và dễ hiểu, đồng thời cũng tạo điều kiện thuận lợi cho việc truy vấn và suy luận trên tri thức đó.

Đồ thị Tri thức đã được ứng dụng trong nhiều lĩnh vực khác nhau, bao gồm tìm kiếm thông tin, hỗ trợ ra quyết định, trí tuệ nhân tạo, phân tích dữ liệu và nhiều lĩnh vực khác. Chúng cung cấp một cách biểu diễn tri thức gọn gàng và dễ hiểu, đồng thời cho phép khai thác và suy luận trên tri thức đó một cách hiệu quả.

1.3. Dữ liệu giáo dục

Trong hầu hết các lĩnh vực thực tế, chẳng hạn như giáo dục, dữ liệu thường xuất phát từ các bộ phận và phòng ban khác nhau trong cùng một tổ chức, dẫn đến nguồn dữ liệu đa dạng với các thuộc tính khác nhau như văn bản có cấu trúc, văn bản không cấu trúc, cơ sở dữ liệu, hình ảnh hoặc dữ liệu truy cập thông qua cơ chế API từ các hệ thống web/ứng dụng hiện có. Ví dụ, tại *Đại học Bách Khoa Thành phố Hồ Chí Minh (HCMUT)*, dữ liệu không cấu trúc xuất phát từ tài liệu pháp lý, câu hỏi thường gặp (FAQs) từ hệ thống hỗ trợ sinh viên BKSI, tin tức từ trang web, dữ liệu từ cơ sở dữ liệu và thông tin được truy xuất từ các hệ thống dựa trên API như hệ thống quản lý giảng dạy LMS. Do tính chất dữ liệu chéo. Việc sử dụng Đồ thị Tri thức như mô tả trên có thể xem như một hình thức bộ nhớ không tham số thích hợp cho biểu diễn kiến thức trong môi trường giáo dục, cho phép triển khai hiệu quả một hệ thống RAG trong ngữ cảnh này. Tuy nhiên, xây dựng một Đồ thị Tri thức từ nhiều nguồn dữ liệu, không được chủ ý thiết kế để tương tác với nhau, không phải là một nhiệm vụ dễ dàng, đặc biệt là khi cần xử lý các ý định mở (Open Intent) thường xuyên xuất hiện trong cuộc trò chuyện thông thường giữa sinh viên và nhân viên của trường đại học. Do đó, dưới hiểu biết hiện tại của chúng tôi,

chưa có một ứng dụng vào thực tiễn nào đáng chú ý của kỹ thuật RAG từ KG cho LLMs trong các tình huống thực tế.

Trong bài báo này, chúng tôi đề xuất một phương pháp tiên tiến tiếp cận RAG dựa trên KG cho Hệ thống trả lời câu hỏi về giáo dục. Chúng tôi đề xuất một kiến trúc cho việc xây dựng Đồ thị Tri thức từ nguồn dữ liệu chéo trong lĩnh vực giáo dục, hiện được triển khai tại HCMUT và sử dụng ngôn ngữ tiếng Việt. Phương pháp này được áp dụng như một thử nghiệm trong một hệ thống dựa trên mô hình ngôn ngữ lớn. Đóng góp của chúng tôi gồm hai phần: (i) Chúng tôi giới thiệu kỹ thuật phát hiện thực thể ý định cho văn bản không cấu trúc trong cuộc trò chuyện phong cách FAQ trên tiếng Việt; và (ii) chúng tôi tiến hành các thực nghiệm, cụ thể là trả lời câu hỏi của sinh viên dựa trên LLM ngay tại HCMUT bằng cách sử dụng kỹ thuật RAG với KG giáo dục.

2. CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

2.1. Phát hiện Ý định Mở

Nhiệm vụ của *Phát hiện Ý định Mở* [10] đặt ra nhiều thách thức và khó khăn trong lĩnh vực *Hiểu Ngôn ngữ Tự nhiên (NLU)* và hệ thống đối thoại. Một thách thức đáng kể là sự không rõ ràng và biến thiên không một khuôn mẫu nào trong cách biểu đạt của người dùng. Khác với việc nhận diện ý định truyền thống, trong đó một tập hợp được định nghĩa trước các danh mục được sử dụng cho phân loại, phát hiện ý định mở bao gồm việc xác định các ý định của người dùng có thể chưa được gặp trong quá trình đào tạo. Điều này mang lại một mức độ không chắc chắn và không thể dự đoán, vì người dùng có thể diễn đạt ý định của họ theo nhiều cách đa dạng và không được dự đoán trước. Một thách thức khác là sự thiếu hụt dữ liệu được gán nhãn đủ cho việc đào tạo các mô hình trong một môi trường thế giới mở. Việc tạo ra các tập dữ liệu được gán nhãn cho mỗi ý định tiềm năng trở nên không thực tế, đặc biệt là khi xử lý với một loạt các lĩnh vực và ứng dụng rộng lớn. Sự khan hiếm của các ví dụ được gán nhãn cho các ý định mới làm cho việc mô hình hóa một cách hiệu quả và chính xác để xác định các ý định mở trở nên khó khăn. Trong những năm gần đây, đã có một sự tăng cường quan tâm đến việc xác định ý định của người dùng từ cả ngôn ngữ viết và ngôn ngữ nói, với sự tập trung vào mô hình hóa và hiểu các tương tác. Các nghiên cứu gần đây, như phương pháp sử dụng mạng LSTM hai chiều và CRF để phát hiện ý định của người dùng một cách hiệu quả [11], phương pháp không giám sát ở hai giai đoạn nhằm khám phá ý định và tạo ra nhãn ý định có ý nghĩa tự động từ các lời nói chưa được gán nhãn [12], và phương pháp khai thác dữ liệu lời nói chưa được gán nhãn để phát hiện ra các ý định phổ biến [13]. Những nghiên cứu này cùng nhấn mạnh tính quan trọng của các phương pháp mạnh mẽ để hiểu ý định của người dùng, mở đường cho việc cải thiện hệ thống đối thoại và trợ lý ảo. Các ý định mở đã được nghiên cứu trong các lĩnh vực như Chăm sóc Khách hàng trong doanh nghiệp [12, 13] và các cơ sở Y tế [14]; tuy nhiên, trong lĩnh vực giáo dục, đặc biệt là trong Chăm sóc Sinh viên, nó vẫn chưa được khám phá một cách toàn diện.

2.2. Đồ thị Tri thức trong lĩnh vực giáo dục

Đồ thị Tri thức (KGs) đã phát triển thành một cách hiệu quả để biểu diễn kiến thức. KGs cung cấp một biểu diễn cấu trúc và tích hợp của các khái niệm, mối quan hệ và thuộc tính trong một lĩnh vực. Đã có nhiều nghiên cứu về KGs trong lĩnh vực giáo dục như đồ thị tri thức cho toán học [15], mô hình Ontology cho chương trình dạy và hồ sơ sinh viên (EducOnto) [16], lược đồ kiến thức (knowledge schema) cho việc dạy ở trường đại học [17], và đồ thị Tri thức để xác định nhu cầu thị trường lao động (giáo dục và việc làm) [18]. Đáng chú ý, nghiên cứu về KGs cho lĩnh vực giáo dục Việt Nam còn rất hạn chế. Huong và Phuc [19] giới thiệu một Kiến trúc để trích xuất các bộ ba (mối quan hệ chủ-định-phụ) từ tiếng Việt. Tuy nhiên, kỹ thuật của họ có hạn chế trong việc xử lý câu phức tạp và chỉ áp dụng cho lĩnh vực du lịch.

2.3. Mô hình ngôn ngữ lớn được tăng cường bởi Đồ thị Tri thức

Một trong những rào cản lớn nhất đối với các LLM là việc mô hình gặp khó khăn trong khả năng ghi nhớ các sự kiện thực tế và thường xuyên xảy ra hiện tượng "ảo giác". Đây là nơi mà Đồ thị Tri thức đóng một vai trò quan trọng. Đồ thị Tri thức cung cấp một biểu diễn cấu trúc của kiến thức, mã hóa các thực thể và mối quan hệ của chúng trong một định dạng có thể đọc hiểu bởi máy. *Mô hình ngôn ngữ lớn được tăng cường bởi Đồ thị Tri thức (KG-augmented LLMs)* nhằm mục tiêu xóa bỏ rào cản này bằng cách tận dụng các ưu điểm của cả hai phương pháp. Nếu thông tin được tạo sinh từ LLMs tương đồng với "sự thật nền tảng" ("ground truth") được biểu diễn bởi các KG, nó có thể được coi là chính xác về mặt thông tin và đã không xảy ra hiện tượng "ảo giác". Phương pháp cải thiện này bao gồm tinh chỉnh quá trình suy luận của LLMs [20], tối ưu hóa cơ chế học của mô hình [21], và thiết lập một cơ chế xác nhận kết quả tạo sinh [22]. Nhiều tiến bộ đáng kể đã đạt được thông qua những nghiên cứu này, làm nổi bật tầm quan trọng của sự đổi mới không ngừng và hỗ trợ cho việc phát triển các Mô hình ngôn ngữ lớn được tăng cường bởi Đồ thị Tri thức tiên tiến hơn.

3. HƯỚNG TỚI XÂY DỰNG ĐỒ THỊ TRI THỨC CHUNG CHO HỆ THỐNG HỎI ĐÁP GIÁO DỤC SỬ DỤNG LLM

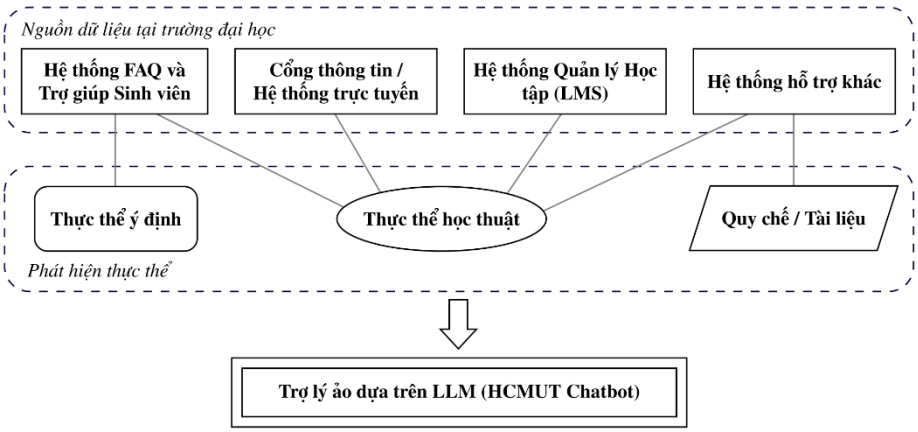
Trong phần này, chúng tôi đi sâu vào phương pháp tiếp cận được sử dụng trong nghiên cứu của chúng tôi để trích xuất những thông tin hữu ích từ môi trường dữ liệu giáo dục đa nguồn để xây dựng Đồ thị Tri thức, là nền tảng của hệ thống Trả lời Câu hỏi Giáo dục được vận hành bởi mô hình ngôn ngữ lớn. Chúng tôi mở đầu bằng việc thảo luận về bản chất của các nguồn dữ liệu trong môi trường đại học của HCMUT và cách chúng đóng góp vào việc hiểu biết toàn diện về hệ sinh thái giáo dục. Tiếp theo, chúng tôi giới thiệu *Kiến trúc E-OED (Khám phá Thực thể Mở Giáo dục)*, một phương pháp mạnh mẽ được thiết kế để khám phá ý định bằng cách sử dụng các phương pháp học không giám sát. Kiến trúc này đặc biệt thích hợp trong việc xử lý những thách thức như các thực thể trùng lặp, ý định mở, và sự kết hợp dữ liệu từ nhiều nguồn. Cuối cùng, chúng tôi trình bày phương pháp của chúng tôi cho Khám phá Mối quan hệ, cho phép chúng tôi khám phá mối quan hệ giữa các loại thực thể khác nhau và xây dựng một Đồ thị Tri thức đại diện cho lĩnh vực giáo dục.

3.1. Dữ liệu giáo dục từ nhiều nguồn tại HCMUT

Trong bối cảnh của một môi trường đại học năng động như HCMUT, dữ liệu có thể đến từ vô số nguồn khác nhau, tạo thành một hệ sinh thái dữ liệu đa nguồn và phức tạp. Mạng lưới các thực thể và liên kết này cung cấp một bức tranh toàn diện về hệ sinh thái một của trường đại học, cho phép hiểu sâu hơn về nhu cầu của người dùng, quy trình giáo dục và môi trường tổng thể của trường đại học bao gồm các nguồn dữ liệu sau. Hình 1 trình bày môi trường dữ liệu tại HCMUT, bao gồm các nguồn dữ liệu sau đây.

Hệ thống FAQ và Trợ giúp Sinh viên: Tại HCMUT, có một hệ thống trực tuyến tương tác cho phép sinh viên nêu ra các vấn đề phát sinh trong quá trình học tập và nhận phản hồi được từ cán bộ hỗ trợ. Kiến thức có thể rút ra từ nguồn này bao gồm các ý định mở, sự kiện, vấn đề của sinh viên, dịch vụ của trường đại học, hệ thống phần mềm của trường đại học.

Cổng thông tin/Hệ thống trực tuyến: Đây là tất cả các trang web mà HCMUT thường xuyên cập nhật về tin tức, học thuật, các chính sách/quy định của trường đại học và thông tin về các khóa học. Kiến thức có thể rút ra từ nguồn này bao gồm các tài liệu hướng dẫn, chương trình đào tạo và mô tả, cùng với các thực thể khác như các tổ chức (trường đại học, khoa, phòng/ban), địa điểm (địa chỉ, thành phố, quận), và tên của cá nhân (sinh viên, giảng viên, nhân viên).

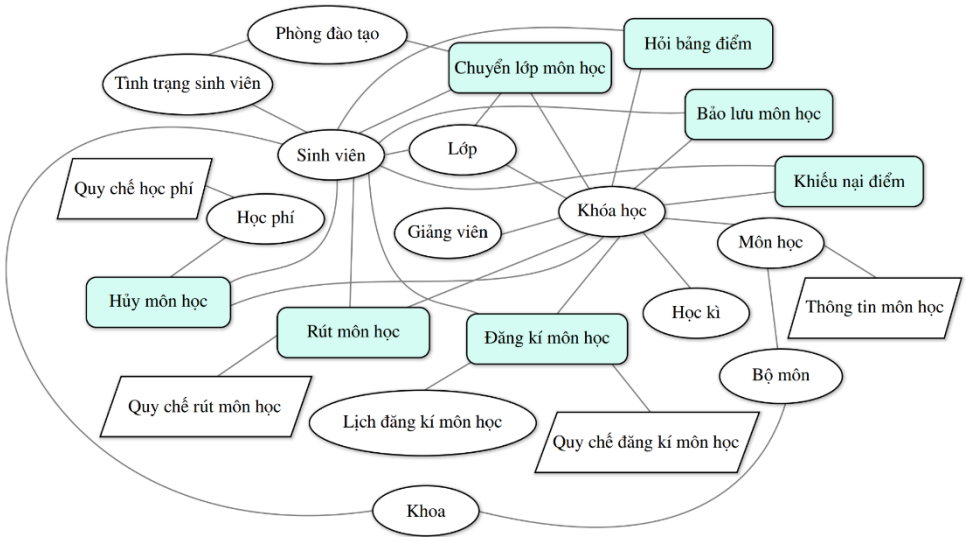


Hình 1. Dữ liệu giáo dục từ nhiều nguồn tại HCMUT

Hệ thống Quản lý Học tập (Learning Management System- LMS) và các hệ thống hỗ trợ khác: Các hệ thống này giúp sinh viên theo dõi các khóa học của mình. Kiến thức có thể rút ra từ nguồn này bao gồm các khóa học, *Từ khóa*, thuật ngữ ngành nghề, lý thuyết ngành nghề, sách, tài liệu, tạp chí và các thành phần của khóa học (sinh viên, giảng viên, trợ giảng).

3.2. Kiến trúc E-OED

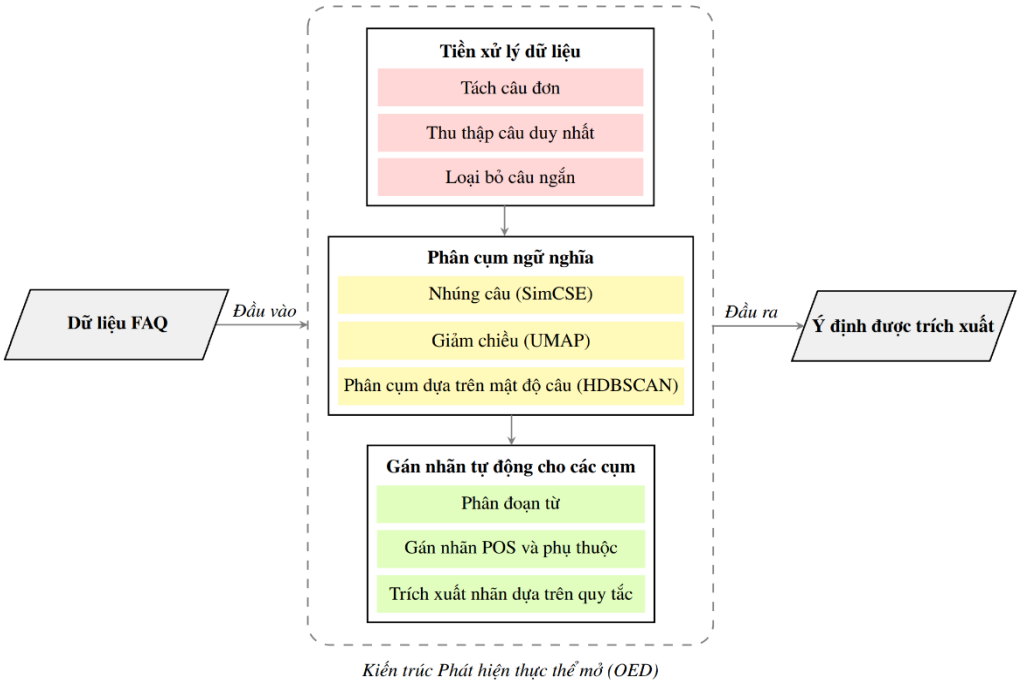
Hình 2 mô tả nhiệm vụ phức tạp của việc trích xuất các mối quan hệ trong lĩnh vực giáo dục và khám phá ý định. Ví dụ, một *Sinh viên* cụ thể có thể có *Tình trạng Sinh viên* của mình, được quản lý bởi *Văn phòng Học vụ*. Sinh viên này cũng thuộc về một *Khoa* cụ thể và đang tham gia vào các *Lớp học* khác nhau. Sinh viên cũng có thể sử dụng một số dịch vụ khi cần như *Rút môn học* hoặc *Tra cứu bảng điểm*. Đồ thị này thể hiện sự phức tạp của việc ánh xạ các ý định vào các thực thể giáo dục, minh họa các kết nối tinh tế tồn tại trong lĩnh vực này.



Hình 2. Các loại thực thể trong môi trường giáo dục

Sự phức tạp này nhấn mạnh sự cần thiết của một phương pháp cấp tiến để giải mã và phân tích các mối quan hệ tương tự một cách chính xác. Nhiệm vụ khó khăn nhất đối với việc xây dựng một Đồ thị Tri thức như vậy có lẽ là khám phá thực thể mở giáo dục. Những thách thức này bao gồm xử lý các thực thể chồng chéo, khám phá các ý định mở (thiếu nghiên cứu hơn so với các nghiên cứu về xử lý các ý định đã được phân loại), và kết hợp dữ liệu từ nhiều nguồn. Hơn nữa, ngôn ngữ tiếng Việt đặt ra những rào cản đặc biệt do là một ngôn ngữ thiếu tài nguyên (low-resource language). Bên cạnh đó, trong môi trường giáo dục, ý định của các cuộc trò chuyện thường là không cố định (dynamic) và liên

tục thay đổi (open) ra thay vì cố định và có giới hạn. Do đó, việc xác định các thực thể mở là rất quan trọng trong lĩnh vực giáo dục.

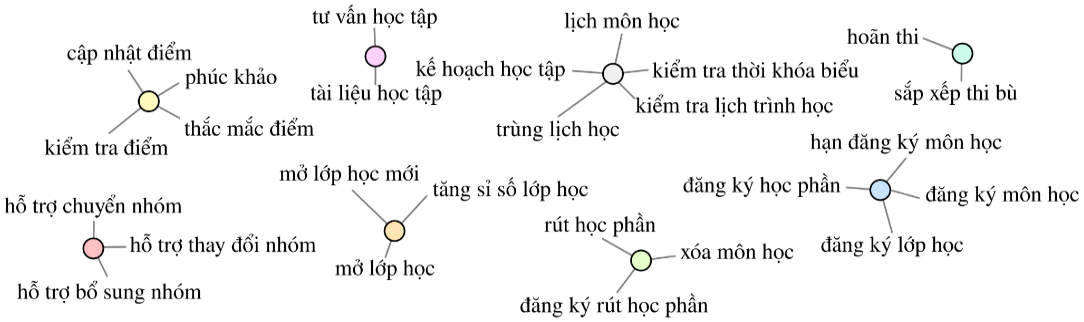


Hình 3. Kiến trúc Phát hiện thực thể mở giáo dục

Tiền xử lý dữ liệu: Trong bước này, chúng tôi chia đoạn văn thành một danh sách các câu đơn, lọc ra các nhiễu, v.v. Bước lọc ban đầu này giúp chúng tôi tập trung vào phần chính của văn bản.

Phân cụm ngữ nghĩa: Để có được sentence embedding, chúng tôi sử dụng *SimCSE* [23] là một mô hình nhúng. Mô hình của chúng tôi sử dụng framework của SimCSE và đã được phát triển sử dụng mô hình cơ bản được đào tạo trước *PhoBERT* [24] nổi tiếng. Trước khi áp dụng các embedding vào thuật toán phân cụm, việc thực hiện các phương pháp giảm chiều là cần thiết. Bước phòng ngừa này được thực hiện để giải quyết "curse of dimensionality", có thể ảnh hưởng tiêu cực đến độ chính xác của các thuật toán phân cụm bằng cách tác động đến các phép đo khoảng cách và xác định các dữ liệu nhiễu không mong muốn. Chúng tôi lựa chọn UMAP thay vì PCA và LDA. Sau khi thực hiện việc giảm chiều, chúng tôi lựa chọn *HDBSCAN* [25] làm thuật toán phân cụm của chúng tôi.

Gán nhãn tự động cho các cụm: Bước cuối cùng trong quy trình này là việc tự động gán nhãn cho các cụm. Để thực hiện công việc này, phương pháp tiếp cận ban đầu của chúng tôi bao gồm việc sử dụng công cụ phân đoạn từ để phân đoạn câu trong mỗi cụm. Sau đó, các câu đã được phân đoạn sử dụng *PhoNLP* [26] để thực hiện cả gán nhãn POS (POS tagging) và gán nhãn phụ thuộc (dependency tagging). Sau khi rút trích các nhãn bởi PhoNLP, một phương pháp dựa trên quy tắc (rule-based method) được thực hiện. Việc làm rõ thêm được đạt được bằng cách xác định các token xuất hiện thường xuyên nhất trong mỗi tập con, sau đó được chỉ định là nhãn của cụm của chúng tôi. Hình 4 trình bày một bộ sưu tập các ý định được khám phá từ tập dữ liệu FAQ như là kết quả của quá trình *Phân cụm ngữ nghĩa* và *Gán nhãn tự động*. Liên quan đến các khóa học và các nhóm lớp học, tồn tại một loạt các yêu cầu bao gồm việc đăng ký khóa học, chuyển lớp học và mở rộng sức chứa, đến các vấn đề về thời gian biểu. Khám phá các ý định và các thực thể liên quan có thể giúp cung cấp các phản hồi tốt hoặc khởi đầu cho các hành động tiếp theo.



Hình 4. Danh sách một số ý định đã được phát hiện

4. THÍ NGHIỆM

4.1. Bộ dữ liệu

Các bộ dữ liệu thử nghiệm bao gồm ba bộ dữ liệu khác nhau *Bank- ing77_eng, Banking77_vni,* và *FAQ_HCMUT_vni* được trình bày trong Bảng 1. Banking77_eng [27] là một bộ dữ liệu tiếng Anh chứa 77 ý định của khách hàng từ hơn 10,000 câu hỏi trong lĩnh vực ngân hàng. Banking77_vni là một bộ dữ liệu tiếng Việt, kết quả đạt được qua việc tự động dịch (sử dụng Google Translate) từ bộ dữ liệu Banking77_eng. FAQ_HCMUT_vni là một bộ dữ liệu tiếng Việt chứa hơn 200,000 câu hỏi thường gặp (FAQs) được thu thập từ hệ thống HelpDesk của HCMUT. Các bộ dữ liệu Banking77_eng và Banking77_vni được sử dụng để đánh giá hiệu suất của framework của chúng tôi. Trong khi bộ dữ liệu FAQ_HCMUT_vni được sử dụng để thể hiện kết quả của framework E-OED. Để nghiên cứu tác động của các vector nhúng khác nhau đối với hiệu suất phân cụm, framework của chúng tôi sử dụng Vietnamese SimCSE cho các bộ dữ liệu tiếng Việt, trong khi BERTopic gốc sử dụng một mô hình sentence-transformers (biến thể của all-miniLM-L6).

4.2. Các Ý định Mở trong Lĩnh vực Giáo dục

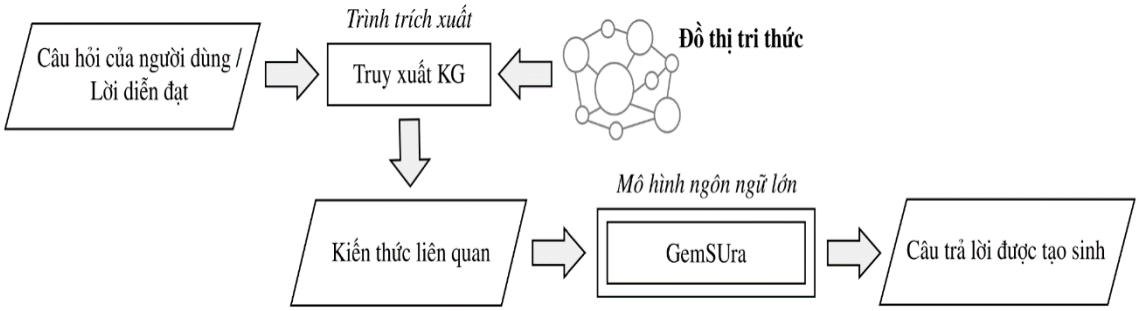
Bảng 1 minh họa kết quả của các thử nghiệm của chúng tôi trong việc khám phá các ý định trên các bộ dữ liệu cụ thể. Để xác minh framework E-OED, chúng tôi thực hiện thử nghiệm framework BERTopic trên hai bộ dữ liệu, Banking77_eng và Banking77_vni. Các kết quả phân cụm trong các Trường hợp 1 và 2 cho thấy hiệu suất ưu việt của framework BERTopic với bộ dữ liệu tiếng Anh, trong đó đã xác định được 73 ý định, gần với 77 danh mục đã xác định trước. Tuy nhiên, trong bộ dữ liệu Banking77_vni, số lượng ý định được xác định giảm đáng kể xuống còn 65. Trong Trường hợp 3, framework E-OED thể hiện hiệu suất tốt hơn với bộ dữ liệu Banking77_vni so với framework BERTopic cho ngôn ngữ tiếng Việt, cho ra 76 ý định được trích xuất. Tiếp theo, chúng tôi áp dụng framework E-OED vào bộ dữ liệu HCMUT_FAQ_vni (Trường hợp 4), khám phá đến tổng cộng 284 cụm với các cụm nhiễu và các cụm trùng lặp đáng kể. Mặc dù vậy, khoảng 372 ý định đã được rút ra từ các cụm này, mặc dù một số lượng lớn ý định vẫn chưa được khám phá. Như đã đề cập trước đó, Hình 4 mô tả một số ý định mở được phát hiện bởi phương pháp của chúng tôi.

Bảng 1. Minh họa kết quả của các thử nghiệm

Trường hợp	Phương pháp	Bộ dữ liệu	Số cụm	Số ý định
1	BERTopic	Banking77_eng	157	73
2	BERTopic	Banking77_vni	147	65
3	E-OED	Banking77_vni	257	**76**
4	E-OED	FAQ_HCMUT_vni	284	**372**

4.3. Phương pháp sử dụng mô hình ngôn ngữ lớn được tương cường bởi Đồ thị Tri thức trong bối cảnh giáo dục

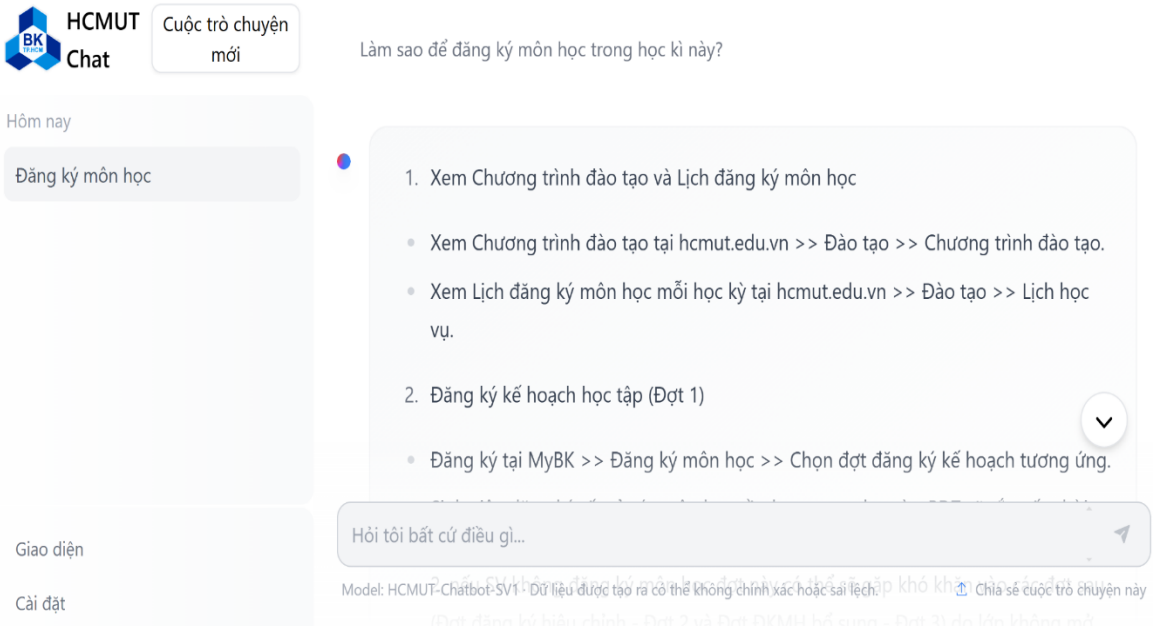
Để minh họa khía cạnh mô hình ngôn ngữ lớn được tương cường bởi Đồ thị Tri thức (KG-augmented LLM) của nghiên cứu và ứng dụng của nó trong lĩnh vực giáo dục tại HCMUT, chúng tôi hiện thực một pipeline tổng quát cho Hệ thống Trả lời Câu hỏi. Quá trình trả lời câu hỏi của người dùng bao gồm việc đưa câu hỏi vào mô hình ngôn ngữ lớn GemSUra do chúng tôi tự phát triển. GemSUra tạo ra một truy vấn có cấu trúc dưới dạng ngôn ngữ Cypher của Neo4j, vì cơ sở dữ liệu Đồ thị Tri thức của chúng tôi được xây dựng trên Neo4j. Truy vấn này tìm kiếm cơ sở dữ liệu Đồ thị Tri thức được tạo bằng kết quả khám phá ý định để truy xuất các đồ thị con và bộ ba tương ứng. Tiếp theo, các bộ ba được chuyển đổi thành định dạng ngôn ngữ tự nhiên và kết hợp với câu hỏi gốc của người dùng để tạo ra một gợi ý cuối cùng cho mô hình ngôn ngữ lớn (LLM). Gợi ý này hướng dẫn LLM trong việc tạo sinh ra một câu trả lời mạch lạc và chính xác.



Hình 5. Phương pháp tiếp cận bổ sung KG cho LLM

5. MỘT SỐ VÍ DỤ MINH HỌA

Trong phần này, chúng tôi trình bày một số ví dụ minh họa về hoạt động của Trợ lý ảo sau khi đã xây dựng và hoạt động thực tế. Trong Hình 6 là hoạt động của Trợ lý ảo khi trả lời câu hỏi về tư vấn tuyển sinh. Hình 7 và Hình 8 minh họa phản hồi của chatbot trong trường hợp sinh viên muốn đăng ký môn học hoặc rút môn học.



Hình 6. Minh họa phản hồi của chatbot cho câu hỏi về tư vấn tuyển sinh

Hình 7. Minh họa phản hồi của chatbot cho câu hỏi về đăng ký môn học

Hình 8. Minh họa phản hồi của chatbot cho câu hỏi về rút môn học

6. KẾT LUẬN

Trong bài viết này, chúng tôi giới thiệu một phương pháp khám phá ý định mở bằng cách sử dụng các phương pháp học không giám sát. Theo kiến thức của chúng tôi, đây là bài báo đầu tiên sử dụng phương pháp này cho việc khám phá ý định mở trong tiếng Việt. Kết quả được đánh giá trên ba bộ dữ liệu: Banking77_eng, Banking77_vni và HCMUT_FAQ_vni. Ngoài ra, chúng tôi thực hiện các thử nghiệm sơ bộ về việc khám phá mối quan hệ giữa các ý định và các thực thể khác để xây dựng một Đồ thị Tri thức cho lĩnh vực giáo dục. Chúng tôi cũng thực hiện một số thử nghiệm về việc áp dụng Đồ thị Tri thức vào mô hình ngôn ngữ lớn để vận hành một Hệ thống Trả lời Câu hỏi tiên tiến.

7. KIẾN NGHỊ

Các thí nghiệm của chúng tôi đã cho thấy một số điểm quan trọng. Đầu tiên, mô hình nhúng gốc, mặc dù không được huấn luyện trên dữ liệu tiếng Việt, nhưng vẫn đạt được hiệu suất hợp lý trên tập dữ liệu banking77_vni. Điều này gợi ý về khả năng chuyển giao do sự tương đồng cấu trúc giữa các ngôn ngữ. Tuy nhiên, việc điều chỉnh một mô hình nhúng riêng biệt dành riêng cho tiếng Việt có thể gây ra những cải tiến tiềm năng hơn. Việc đánh giá chất lượng của mô hình nhúng có thể được thực hiện bằng cách phân tích các câu đại diện trong mỗi cụm. Sự tương đồng ngữ nghĩa lớn hơn trong các cụm cho thấy một mô hình hiệu quả hơn.

Cuối cùng, hệ thống của chúng tôi nhạy cảm với việc lựa chọn siêu tham số. Do sự hiện diện của nhiều siêu tham số cần được điều chỉnh, cần phải xem xét cẩn thận khi áp dụng phương pháp này vào lĩnh vực cụ thể hoặc tập dữ liệu FAQ.

TÀI LIỆU THAM KHẢO

[1] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training", 2018.

[2] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoy-anov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension", 2019.

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models", 2023.

[4] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?", 2019.

[5] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models", 2022.

[6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation", *ACM Computing Surveys*, vol. 55, no. 12, p. 1–38, Mar. 2023. [Online]. Available: http://dx.doi.org/10.1145/3571730

[7] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering", 2020.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks", 2021.

[9] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs", *ACM Computing Surveys*, vol. 54, no. 4, p. 1–37, Jul. 2021. [Online]. Available: http://dx.doi.org/10.1145/3447772

[10] M. Chen, B. Jayakumar, M. Johnston, S. E. Mahmoodi, and D. Pressel, "Intent discovery for enterprise virtual assistants: Applications of utterance embedding and clustering to intent mining", in *North American Chapter of the Association for Computational Linguistics*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250390998

[11] N. Vedula, N. Lipka, P. Maneriker, and S. Parthasarathy, "Towards open intent discovery for conversational text", 2019.

[12] "Open intent extraction from natural language interactions", in *Proceedings of The Web Conference 2020*, ser. WWW '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2009–2020. [Online]. Available: https://doi.org/10.1145/3366423.3380268

[13] P. Liu, Y. Ning, K. K. Wu, K. Li, and H. Meng, "Open intent discovery through unsupervised semantic clustering and dependency parsing", 2021.

[14] A. Mullick, I. Mondal, S. Ray, R. Raghav, G. S. Chaitanya, and P. Goyal, "Intent identification and entity extraction for healthcare queries in indic languages", 2023.

[15] P. Chen, Y. Lu, V. W. Zheng, X. Chen, and B. Yang, "Knowedu: A system to construct knowledge graph for education", *IEEE Access*, vol. 6, pp. 31 553–31 563, 2018.

[16] N. Hubert, A. Brun, and D. Monticolo, "New ontology and knowledge graph for university curriculum recommendation", 2022.

[17] M. Rizun, "Knowledge graph application in education: a literature review", *Acta Universitatis Lodziensis. Folia Oeconomica*, vol. 3, pp. 7–19, 08 2019.

[18] Y. Fettach, M. Ghogho, and B. Benatallah, "Knowledge graphs in education and employability: A survey on applications and techniques", *IEEE Access*, vol. 10, pp. 80 174–80 183, 2022.

[19] H. D. To and P. Do, "Extracting triples from Vietnamese text to create knowledge graph", in *12th International Conference on Knowledge and Systems Engineering, KSE 2020, Can Tho City, Vietnam, November 12-14, 2020*. IEEE, 2020, pp. 219–223. [Online]. Available: https://doi.org/10.1109/KSE50997.2020.9287471

[20] J. Baek, A. F. Aji, and A. Saffari, "Knowledge-augmented language model prompting for zero-shot knowledge graph question answering", in *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, B. Dalvi Mishra, G. Durrett, P. Jansen, D. Neves Ribeiro, and J. Wei, Eds. Toronto, Canada: Association for Computational Linguistics, Jun. 2023, pp. 78–106. [Online]. Available: https://aclanthology.org/2023.nlrse-1.7

[21] S. Kim, S. Joo, D. Kim, J. Jang, S. Ye, J. Shin, and M. Seo, "The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning", in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 685–12 708. [Online]. Available: https://aclanthology.org/2023.emnlp-main.782

[22] M. Kang, J. M. Kwak, J. Baek, and S. J. Hwang, "Knowledge graph-augmented language models for knowledge-grounded dialogue generation", 2023.

[23] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings", 2022.

[24] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1037–1042. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.92

[25] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based cluster-ing", *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017.

[26] L. T. Nguyen and D. Q. Nguyen, "PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition, and dependency parsing", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2021, pp. 1–7.

[27] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić, "Efficient intent detection with dual sentence encoders", 2020.