THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): CS2205.CH201 - Đăng Anh Đạt - 250101009
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/danganhdat/CS2205.CH201/blob/main/%C4%90%E1%BA%A1t%20%C4%90%E1%BA%B7ng%20Anh%20-%20CS2205.SEP2025.DeCuong.FinalReport.Template.Doc.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Đạng Anh Đạt
- MSSV: 250101009

- Lớp: CS2205.CH201
- Tự đánh giá (điểm tổng kết môn): 8.5/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 0
- Số câu hỏi QT của cả nhóm: 0
- Link Github: https://github.com/mynameuit/CS2205.CS2205.CH201/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

NGHIÊN CỨU PHƯƠNG PHÁP NHÚNG TÀI LIỆU GIẢ ĐỊNH (HYDE) NHẰM TỐI ƯU HÓA HỆ THỐNG TRUY XUẤT DỮ LIỆU DÀY ĐẶC TRONG ĐIỀU KIỆN KHÔNG GIÁM SÁT

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

PRECISE ZERO-SHOT DENSE RETRIEVAL WITHOUT RELEVANCE LABELS

TÓM TẮT *(Tối đa 400 từ)*

Mặc dù tìm kiếm dày đặc (dense retrieval) đã đạt được những tiến bộ vượt bậc, việc triển khai các hệ thống này trong điều kiện hoàn toàn không giám sát (zero-shot) vẫn là một thách thức lớn. Nguyên nhân chủ yếu nằm ở sự thiếu hụt dữ liệu nhãn liên quan (relevance labels) để huấn luyện bộ mã hóa (encoder), dẫn đến việc mô hình khó xác định được mối liên hệ ngữ nghĩa giữa câu truy vấn ngắn và tài liệu dài. Nghiên cứu này đề xuất phương pháp Nhúng tài liệu giả định (Hypothetical Document Embeddings - HyDE) nhằm thay đổi cách thức mô hình hóa sự liên quan mà không cần dữ liệu giám sát.

Phương pháp HyDE đề xuất một quy trình tìm kiếm gồm hai giai đoạn tách biệt: tạo văn bản và mã hóa đối chiếu

Giai đoạn 1: Khi nhận được một câu truy vấn, hệ thống sử dụng một mô hình ngôn ngữ lớn (LLM) đã được huấn luyện theo chỉ dẫn (ví dụ: InstructGPT) để tạo ra một tài liệu giả định. Tài liệu này không nhất thiết phải chính xác về mặt sự thật, nhưng được kỳ vọng sẽ mang các đặc điểm cấu trúc và ngữ nghĩa của một câu trả lời tiềm năng.

Giai đoạn 2: Một bộ mã hóa đối chiếu không giám sát (ví dụ: Contriever) sẽ chuyển đổi tài liệu giả định này thành một vector nhúng. Trong bước này, bộ mã hóa đóng vai trò như một bộ lọc nén, giúp trích xuất các đặc trưng liên quan cốt lõi và loại bỏ các chi tiết hư cấu từ LLM. Vector này sau đó được dùng để truy xuất các tài liệu thực có độ tương đồng cao trong không gian vector.

Kết quả thực nghiệm trên các bộ dữ liệu tiêu chuẩn (web search, QA, fact-checking) chứng minh HyDE vượt trội hơn hẳn mô hình không giám sát và đạt hiệu suất tương đương với các hệ thống đã qua tinh chỉnh (fine-tuned). Phương pháp này cho thấy khả năng thích ứng mạnh mẽ trên nhiều ngôn ngữ (đa ngôn ngữ zero-shot). Nghiên cứu khẳng định bằng cách sử dụng LLM để mô hình hóa sự liên quan, HyDE loại bỏ sự phụ thuộc vào dữ liệu gán nhãn, cung cấp một giải pháp tìm kiếm mạnh mẽ, linh hoạt và hiệu quả cho các hệ thống thông tin ngay từ giai đoạn triển khai ban đầu.

GIỚI THIỆU *(Tối đa 1 trang A4)*

1. Bối cảnh và Nguồn gốc đề tài

Trong những năm gần đây, truy xuất dày đặc (Dense Retrieval) đã trở thành kỹ thuật then chốt trong các hệ thống tìm kiếm hiện đại, cho phép xác định tài liệu dựa trên sự tương đồng ngữ nghĩa trong không gian vector. Mặc dù các mô hình giám sát đã đạt được thành công đáng kể trong các tác vụ như tìm kiếm web hay trả lời câu hỏi, hiệu quả của chúng phụ thuộc rất lớn vào các tập dữ liệu được gán nhãn liên quan (relevance labels) quy mô lớn (như MS-MARCO). Tuy nhiên, trên thực tế, việc thu thập dữ liệu nhãn không phải lúc nào cũng khả thi do rào cản về chi phí và quyền sử dụng thương mại. Điều này đặt ra nhu cầu cấp thiết về việc phát triển các hệ thống truy xuất dày đặc không giám sát (Zero-shot Dense Retrieval) có khả năng hoạt động hiệu quả ngay lập tức mà không cần dữ liệu huấn luyện đặc thù.

2. Lý do chọn đề tài và Mục tiêu nghiên cứu

Thách thức lớn nhất của việc tìm kiếm không giám sát nằm ở sự lệch pha (mismatch) giữa câu truy vấn ngắn của người dùng và các tài liệu dài trong kho lưu trữ, khiến việc mã hóa chúng vào cùng một không gian vector mà không có nhãn giám sát trở nên khó kiểm soát.

Đề tài này nghiên cứu phương pháp Nhúng tài liệu giả định (Hypothetical Document Embeddings - HyDE) nhằm phá vỡ rào cản này. Thay vì so khớp trực tiếp câu truy vấn với kho dữ liệu, HyDE tận dụng khả năng hiểu ngôn ngữ và thực hiện chỉ dẫn của các Mô hình ngôn ngữ lớn (LLM) để tạo ra một "tài liệu giả định" mang các đặc trưng của tài liệu liên quan. Cách tiếp cận này chuyển đổi bài toán mô hình hóa sự liên quan phức tạp từ không gian nhúng sang tác vụ tạo văn bản - một lĩnh vực mà các LLM hiện nay đã đạt tới độ chín muồi.

3. Mô tả hệ thống (Input/Output)

Hệ thống HyDE phân rã quá trình truy xuất thành hai giai đoạn chính với các thành phần cụ thể như sau:

- **Đầu vào (Input):**
  - Câu truy vấn (q): Yêu cầu tìm kiếm từ người dùng (Ví dụ: "Cần bao lâu để nhổ răng khôn?").
  - Chỉ dẫn (Instruction): Văn bản điều hướng mô hình (Ví dụ: "Hãy viết một đoạn văn khoa học để trả lời câu hỏi"), có thể tinh chỉnh tùy theo loại tác vụ (y khoa, xác minh sự thật hoặc đa ngôn ngữ).
- **Quá trình trung gian:**
  1. Tạo tài liệu giả định ($\hat{d}$): LLM (như InstructGPT) nhận q và chỉ dẫn để tạo ra một văn bản phản hồi. Dù tài liệu này có thể chứa thông tin giả (hallucination), nó lại phản ánh chính xác cấu trúc và mẫu hình (pattern) của một câu trả lời thực tế.
  2. Mã hóa đối chiếu: $\hat{d}$ được đưa vào một bộ mã hóa không giám sát (như Contriever). Tại đây, bộ mã hóa đóng vai trò như một bộ lọc nén, lọc bỏ các sai lệch thông tin và chỉ giữ lại các đặc điểm ngữ nghĩa cốt lõi để tạo ra vector nhúng đại diện ($\hat{v}$).
- **Đầu ra (Output):**
  - Vector truy vấn đại diện: Một vector trong không gian dày đặc tích hợp đầy đủ ngữ nghĩa của tài liệu tiềm năng.
  - Danh sách tài liệu thực: Hệ thống thực hiện tìm kiếm láng giềng gần nhất (MIPS) trong kho dữ liệu và trả về các tài liệu thực tế có độ tương đồng cao nhất.

Bằng cách thay đổi quy trình xây dựng vector truy vấn, HyDE không chỉ vượt trội hơn các hệ thống không giám sát tiên tiến nhất mà còn đạt hiệu suất tương đương với các mô hình đã qua tinh chỉnh giám sát trên nhiều ngôn ngữ và tác vụ khác nhau.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. **Thiết kế và hiện thực hóa quy trình truy xuất "Tài liệu giả định" (HyDE)**

- Xây dựng khung nghiên cứu HyDE nhằm chuyển đổi mô hình tìm kiếm từ đối sánh "Truy vấn - Tài liệu" sang đối sánh "Tài liệu - Tài liệu".
- Thiết lập quy trình hai giai đoạn: (1) Ứng dụng mô hình ngôn ngữ lớn (LLM) để chuyển hóa câu truy vấn thành các văn bản giả định thông qua cơ chế

Instruction Prompting; (2) Sử dụng bộ mã hóa đối chiếu (Encoder) để trích xuất đặc trưng ngữ nghĩa cốt lõi.

- Mục tiêu then chốt là chứng minh khả năng loại bỏ hoàn toàn sự phụ thuộc vào dữ liệu gán nhãn (Relevance labels) trong việc xây dựng không gian vector nhúng.

2. **Đánh giá định lượng hiệu suất trên hệ thống Benchmark đa tác vụ**

- Đo lường hiệu quả của phương pháp HyDE thông qua các chỉ số tiêu chuẩn ($nDCG@10, MAP, Recall@k$) trên đa dạng loại hình dữ liệu: Tìm kiếm web (TREC DL), Trả lời câu hỏi (DL Hard) và Xác minh sự thật (SciFact).
- So sánh đối chiếu hiệu suất với các phương pháp truyền thống (BM25), các mô hình không giám sát tiên tiến (Contriever) và các hệ thống đã qua tinh chỉnh giám sát (Fine-tuned models).
- Xác định ranh giới hiệu quả của HyDE trong các kịch bản tìm kiếm chuyên biệt như tài chính (FiQA) hoặc y sinh.

3. **Kiểm chứng tính thích ứng đa ngôn ngữ và sự ảnh hưởng của các dòng LLM**

- Đánh giá khả năng khái quát hóa của hệ thống trên môi trường đa ngôn ngữ (Zero-shot Cross-lingual) với các ngôn ngữ có nguồn lực thấp và trung bình (như tiếng Hàn, Nhật, Swahili) thông qua tập dữ liệu Mr.TyDi.
- Phân tích sức mạnh biểu diễn của các dòng LLM khác nhau (như InstructGPT, Flan-T5) tác động như thế nào đến chất lượng tài liệu giả định và kết quả truy xuất cuối cùng.
- Khẳng định tính linh hoạt và khả năng triển khai tức thời (Plug-and-play) của phương pháp mà không cần tái huấn luyện mô hình.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung 1: Nghiên cứu thiết kế khung phương pháp HyDE và cơ chế phân rã tác vụ truy xuất

- **Thiết lập cơ chế phân rã tác vụ (Task Factorization):** Đề xuất thay đổi căn bản trong quy trình truy xuất thông tin (Information Retrieval - IR) bằng cách chuyển đổi bài toán đối sánh **Query** $\rightarrow$ **Document** sang **Document** $\rightarrow$ **Document**. Quy trình này loại bỏ sự phụ thuộc vào dữ liệu gán nhãn (relevance labels) thông qua việc phân tách thành hai module độc lập:
    - **Generative Module (NLG):** Sử dụng LLM để dự đoán và sinh ra các văn bản giả định (Hypothetical Documents) nhằm nắm bắt các hình thái ngữ nghĩa có thể có của tài liệu đích.

    - **Contrastive Encoder (NLU):** Sử dụng mô hình mã hóa không giám sát để ánh xạ các văn bản giả định này vào không gian vector dày đặc (dense embedding space).
- **Mô hình hóa toán học và Vector đại diện:** Thực hiện tính toán vector đại diện cho truy vấn v_q bằng cách lấy trung bình cộng (mean pooling) của N vector tài liệu giả định. Phương pháp này giúp làm mượt nhiễu và tăng tính ổn định cho biểu diễn ngữ nghĩa.
- **Cơ chế "Lossy Compressor" (Nén có tổn hao):** Tận dụng đặc tính của Encoder như một bộ lọc thông tin. Trong khi LLM có thể sinh ra các chi tiết sai lệch (hallucinations), Encoder sẽ đóng vai trò lọc bỏ các nhiễu thông tin này, chỉ giữ lại và mã hóa các đặc trưng ngữ nghĩa cốt lõi, giúp định vị chính xác tài liệu thực trong không gian vector.

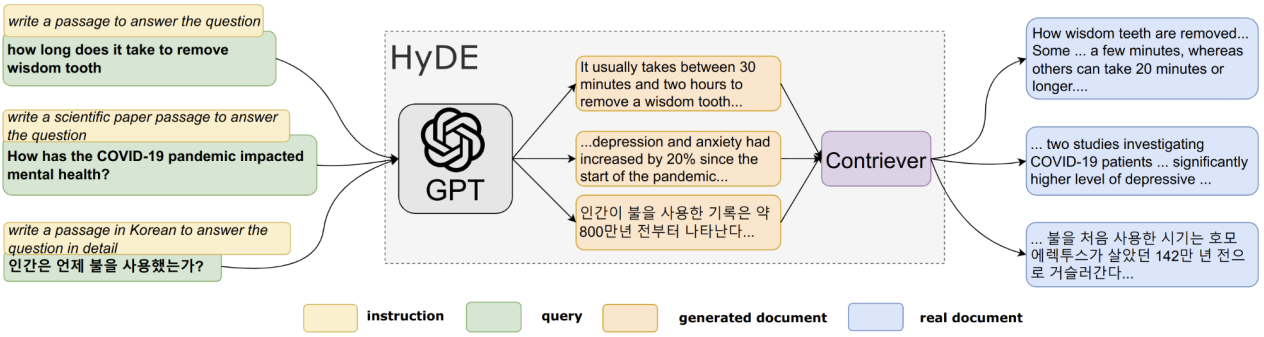

Figure 1: An illustration of the HyDE model. Documents snippets are shown. HyDE serves all types of queries without changing the underlying GPT-3 and Contriever/mContriever models.

Hình 1: Sơ đồ kiến trúc phương pháp HyDE. Quy trình bắt đầu với truy vấn q, được đưa qua LLM (như InstructGPT) để sinh ra k tài liệu giả định. Các tài liệu này được mã hóa bởi Contriever thành các vector, sau đó được tính trung bình để tạo thành vector truy vấn duy nhất v_q. Vector này được dùng để tìm kiếm láng giềng gần nhất (Nearest Neighbor Search) trong kho ngữ liệu.

Nội dung 2: Đánh giá hiệu suất thực nghiệm đa tác vụ và đối sánh hệ thống

- **Kiểm chứng trên tác vụ Tìm kiếm Web (Web Search):** Đánh giá hiệu năng Zero-shot trên bộ dữ liệu chuẩn TREC DL19 và DL20.
    - **Mục tiêu:** Đạt chỉ số $nDCG@10$ tiệm cận mức **61.3**, thu hẹp khoảng cách với các mô hình Fine-tuned.
    - **Đối sánh:** Chứng minh hiệu quả vượt trội so với các phương pháp Unsupervised truyền thống như BM25 (~ 50.6) và Contriever (~ 44.5),

đồng thời cạnh tranh trực tiếp với mô hình Supervised mạnh là Contriever-FT (~ 62.1).

- **Đánh giá tính bền vững trên các miền dữ liệu chuyên sâu (BEIR Benchmark):** Mở rộng đánh giá trên các tập dữ liệu Low-resource và đa dạng miền (Y sinh, Tài chính, Khoa học).
    - **Hiệu quả đột phá:** Tại tập SciFact, hệ thống kỳ vọng đạt điểm số 69.1, vượt xa các baseline mạnh như DPR (31.8) và ANCE (50.7).
    - **Độ ổn định:** Duy trì sai số cực thấp (< 0.5) so với BM25 ngay cả trên các kịch bản khó và biến động như TREC-Covid.
- **Kết luận giả thuyết thực nghiệm:** Các kết quả này khẳng định tính khả thi của việc ủy thác tác vụ mô hình hóa sự liên quan (Relevance Modeling) cho LLM. Hệ thống Zero-shot HyDE chứng minh năng lực tương đương với các mô hình được huấn luyện trên hàng triệu cặp dữ liệu gán nhãn thủ công.

Nội dung 3: Phân tích khả năng khái quát hóa đa ngôn ngữ và ảnh hưởng của quy mô mô hình

- **Nghiên cứu khả năng thích ứng Đa ngôn ngữ (Multilingual Zero-shot):**
    - **Phạm vi:** Thử nghiệm trên tập dữ liệu Mr.TyDi bao gồm các ngôn ngữ có hình thái khác biệt (Tiếng Hàn, Nhật, Swahili, Bengali).
    - **Kết quả kỳ vọng:** Đối với Tiếng Hàn, chỉ số $MRR@100$ tăng đáng kể từ 22.3 (mContriever) lên **30.6** (HyDE).
    - **Phân tích nguyên nhân:** Tài liệu giả định sinh bởi LLM cung cấp tín hiệu ngữ nghĩa (semantic signals) phong phú và chính xác hơn so với việc chỉ dựa vào Transfer Learning từ các mô hình mDPR giám sát.
- **Phân tích hiệu ứng quy mô LLM (Scaling Laws) và Tối ưu hóa chỉ dẫn:**
    - **Thực nghiệm Scaling Laws:** Sử dụng các backbone có kích thước tăng dần: Flan-T5 (11B) $\rightarrow$ Cohere (52B) $\rightarrow$ InstructGPT (175B). Kết quả cho thấy hiệu suất truy xuất ($nDCG@10$) tăng trưởng tuyến tính theo quy mô tham số (từ ~ 48.9 lên ~ 61.3).
    - **Kỹ thuật Instruction Engineering:** Thiết kế các chỉ dẫn đặc thù cho từng tác vụ (Task-specific Instructions) nhằm thay đổi phong cách văn bản giả định (ví dụ: phong cách y sinh hoặc tài chính).
- **Giá trị thực tiễn:** Giải pháp này giải quyết triệt để bài toán **Cold-start** trong các hệ thống tìm kiếm mới, cho phép triển khai ngay lập tức mà không cần chi phí thu thập dữ liệu và tái huấn luyện mô hình.

KẾT QUẢ MONG ĐỢI

- **Chứng minh hiệu năng đột phá của cơ chế "Tài liệu giả định" (HyDE)**
    - Mô hình hóa sự liên quan: Nghiên cứu dự kiến khẳng định LLM có khả năng chuyển đổi các câu truy vấn ngắn, thiếu ngữ cảnh thành các văn bản giả định giàu đặc trưng cấu trúc. Điều này giúp thu hẹp khoảng cách ngữ nghĩa giữa truy vấn và tài liệu thực tế.
    - Cơ chế lọc nhiễu: Chứng minh vai trò của bộ mã hóa đối chiếu (Contrastive Encoder) như một "bộ lọc đặc trưng". Hệ thống dự kiến sẽ cho thấy dù tài liệu giả định chứa các lỗi thực tế (hallucinations), bộ mã hóa vẫn có thể trích xuất các vector nhúng tập trung vào ngữ nghĩa cốt lõi để định vị chính xác tài liệu liên quan trong không gian lân cận (Nearest Neighbors).
- **Hiệu suất vượt trội trên hệ thống Benchmark đa tác vụ**
    - Tìm kiếm Web (TREC DL): HyDE dự kiến đạt các chỉ số $nDCG@10$ và MAP vượt xa các phương pháp không giám sát (unsupervised) như Contriever hay truyền thống như BM25. Kết quả kỳ vọng sẽ tiệm cận hoặc tương đương với các mô hình giám sát (supervised) đã tinh chỉnh trên MS-MARCO (như DPR, ANCE), xóa bỏ rào cản về dữ liệu gán nhãn.
    - Tính bền vững trên miền dữ liệu mới (BEIR): Đối với các tác vụ chuyên biệt (Tài chính - FiQA, Y sinh - SciFact), hệ thống được dự báo sẽ duy trì tính ổn định cao, khắc phục nhược điểm "nhạy cảm với dữ liệu huấn luyện" của các mô hình giám sát thông thường.
- **Khả năng khái quát hóa đa ngôn ngữ và tính tương thích hệ thống**
    - Vượt trội đa ngôn ngữ: Kết quả dự kiến sẽ xác nhận hiệu quả của HyDE trên các ngôn ngữ không phải tiếng Anh (Swahili, Hàn, Nhật, Bengali). Ngay cả với ngôn ngữ nguồn lực thấp, hệ thống vẫn đạt kết quả tốt hơn các mô hình *mContriever* thuần túy nhờ tận dụng tri thức đa ngữ của LLM.
    - Tương quan năng lực LLM: Nghiên cứu dự kiến chỉ ra một quy luật tỷ lệ thuận: Năng lực thực hiện chỉ dẫn của LLM càng mạnh (ví dụ từ *Flan-T5* lên *InstructGPT*), chất lượng vector nhúng và độ chính xác truy xuất càng cao. Điều này mở ra khả năng nâng cấp hệ thống liên tục theo sự phát triển của các dòng LLM mới.
- **Đóng góp về mặt giải pháp và mã nguồn**
    - Giải pháp triển khai tức thời: Khẳng định HyDE là mô hình lý tưởng cho

các hệ thống tìm kiếm mới triển khai (cold-start), nơi dữ liệu người dùng và nhãn liên quan chưa tồn tại.

- Sản phẩm khoa học: Bộ mã nguồn hoàn chỉnh và hướng dẫn thiết kế Instruction Prompting tối ưu cho các loại tác vụ tìm kiếm khác nhau, phục vụ cộng đồng nghiên cứu và ứng dụng thực tiễn.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Luyu Gao, Xueguang Ma, Jimmy Lin, Jamie Callan: Precise Zero-Shot Dense Retrieval without Relevance Labels. CoRR abs/2212.10496 (2022).

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, et al.: Language Models are Few-Shot Learners. NeurIPS 2020.

[3] Gautier Izacard, Mathilde Caron, Lucas Hosseini, et al.: Towards Unsupervised Dense Information Retrieval with Contrastive Learning. CoRR abs/2112.09118 (2021).

[4] Vladimir Karpukhin, Barlas Oguz, Sewon Min, et al.: Dense Passage Retrieval for Open-Domain Question Answering. EMNLP (1) 2020: 6769-6781.

[5] Long Ouyang, Jeff Wu, Xu Jiang, et al.: Training language models to follow instructions with human feedback. CoRR abs/2203.02155 (2022).

[6] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych: BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. CoRR abs/2104.08663 (2021)