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

Đặng Anh Đạt - 250101009

Tóm tắt

- Lớp: CS2205.CH201
- Link Github: danganhdat/CS2205.CH201
- Link YouTube video: CS2205.CH201 - Đăng Anh Đat - 250101009

Đặng Anh Đạt
250101009

Giới thiệu

- **Xu hướng** chuyển dịch từ tìm kiếm từ khóa sang tìm kiếm ngữ nghĩa thông qua Vector Embeddings.
- **Hạn chế của mô hình giám sát (Supervised):**
  - Phụ thuộc vào tập dữ liệu gán nhãn khổng lồ (Ví dụ: MS-MARCO)
  - Chi phí thu thập dữ liệu cao, giới hạn quyền sử dụng thương mại
- **Sự lệch pha (mismatch)** về đặc điểm phân phối giữa Câu truy vấn (ngắn) và Tài liệu (dài) trong không gian vector khi không có nhãn giám sát.

→ Một hệ thống **Zero-shot Dense Retrieval** hoạt động hiệu quả mà không cần dữ liệu huấn luyện đặc thù.

Giới thiệu

- Thay đổi từ đối sánh **"Truy vấn - Tài liệu"** sang **"Tài liệu - Tài liệu"**
- Chuyển đổi bài toán mô hình hóa sự liên quan phức tạp sang tác vụ tạo văn bản. Tận dụng LLM trong việc hiểu chỉ dẫn (*Instructions*) để mô phỏng hình dáng của tài liệu mục tiêu

1. Input: Query q + Instruction (Chỉ dẫn tác vụ)
2. LLM Step: Tạo tài liệu giả định $\hat{d}$ (Phản ánh cấu trúc, dù có thể có thông tin ảo giác)
3. Encoder Step: Chuyển $\hat{d}$ thành Vector nhúng $\hat{v}$ (Đóng vai trò bộ lọc nén, giữ lại cốt lõi ngữ nghĩa)
4. Output:
    - Vector truy vấn đại diện.
    - Danh sách tài liệu thực tế thông qua tìm kiếm láng giềng gần nhất (MIPS).

Mục tiêu

1. **Thiết lập Quy trình HyDE:**
   - Chuyển đổi đối sánh Truy vấn-Tài liệu → Tài liệu-Tài liệu
   - Sử dụng LLM tạo văn bản giả định (Instruction Prompting)
   - Loại bỏ hoàn toàn sự phụ thuộc vào dữ liệu gán nhãn (Relevance labels)
2. **Đánh giá Hiệu suất Đa tác vụ:**
   - Đo lường chỉ số *nDCG@10*, *MAP* trên các Benchmark: TREC DL, SciFact, FiQA
   - So sánh trực tiếp với BM25 (Truyền thống) và Contriever (Tiên tiến)
3. **Kiểm chứng Khả năng Khái quát hóa:**
   - Đánh giá tính thích ứng **Đa ngôn ngữ Zero-shot** (Hàn, Nhật, Swahili...).
   - Phân tích ảnh hưởng của các dòng LLM khác nhau (*InstructGPT, Flan-T5*).

Nội dung và Phương pháp (1)

Nội dung 1: Xây dựng khung phương pháp HyDE để phân rã quy trình truy xuất và loại bỏ sự phụ thuộc vào dữ liệu gán nhãn

Chuyển đổi đối sánh *{Query → Document}* sang *{Document → Document}*

Phân rã tác vụ (Task Factorization):

1. **NLG (LLM):** Dự đoán hình thái tài liệu liên quan $\hat{d} = g(q, INST)$
2. **NLU (Encoder):** Ánh xạ ngữ nghĩa vào không gian vector dày đặc

Mô hình hóa toán học:

- Vector đại diện: $\hat{v}_q = \frac{1}{N} \sum_{k=1}^{N} f(\hat{d}_k)$ (Trung bình cộng *N* tài liệu giả định)
- Hiệu chỉnh ổn định: $\hat{v}_q = \frac{1}{N+1}[\sum f(\hat{d}_k) + f(q)]$.

Cơ chế Lossy Compressor: Encoder lọc bỏ thông tin ảo giác (hallucination), giữ lại đặc trưng ngữ nghĩa cốt lõi để định vị tài liệu thực.

Nội dung và Phương pháp (2)

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

Kỳ vọng trên Tìm kiếm Web (TREC DL19/20)

- Mục tiêu *nDCG@10*: Tiệm cận mức 61.3
- Vượt xa Unsupervised: BM25 (~50.6) và Contriever (~44.5).
- Cạnh tranh Supervised: Tiệm cận Contriever-FT (~62.1).

Tính bền vững trên Benchmarks đa miền (BEIR)

- Đột phá tại SciFact: Kỳ vọng đạt 69.1 (Vượt xa DPR: 31.8 và ANCE: 50.7).
- Độ ổn định cao: Duy trì sai số cực thấp (< 0.5) so với BM25 trên các kịch bản khó như TREC-Covid.

Giả thuyết thực nghiệm kết luận: Chứng minh khả năng ủy thác mô hình hóa sự liên quan cho LLM. Khẳng định hệ thống Zero-shot có thể đạt năng lực tương đương mô hình huấn luyện trên hàng triệu cặp dữ liệu gán nhãn.

Nội dung và Phương pháp (3)

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

Thích ứng Đa ngôn ngữ (Multilingual Zero-shot)

- Phạm vi: Thử nghiệm trên tập dữ liệu **Mr.TyDi** (Hàn, Nhật, Swahili, Bengali)
- Kỳ vọng (Tiếng Hàn): Chỉ số *MRR@100* tăng từ 22.3 (mContriever) → 30.6 (HyDE)
- Giả thuyết: Tài liệu giả định từ LLM cung cấp tín hiệu ngữ nghĩa tốt hơn các mô hình Transfer Learning có giám sát (mDPR)

Hiệu ứng quy mô LLM (Scaling Laws)

- Backbone thử nghiệm: *Flan-T5 (11B), Cohere (52B), InstructGPT (175B)*
- Xu hướng bậc thang (*nDCG@10*): Tăng trưởng tỷ lệ thuận với quy mô tham số (11B: ~48.9 → 175B: ~61.3)

Tối ưu hóa Chỉ dẫn (Instruction Engineering):

- Giải pháp: Thiết kế *Task-specific Instructions* (Ví dụ: phong cách y sinh, tài chính)
- Giá trị thực tiễn: Giải quyết bài toán *Cold-start*, không cần tái huấn luyện mô hình

Kết quả dự kiến

1. **Đột phá về Hiệu năng và Cơ chế Lọc nhiễu**
   - Mô hình hóa sự liên quan: LLM thu hẹp khoảng cách ngữ nghĩa giữa truy vấn (ngắn) và tài liệu thực (dài).
   - Cơ chế "Bộ lọc đặc trưng": Chứng minh Encoder có thể lọc bỏ sai lệch từ LLM để trích xuất ngữ nghĩa cốt lõi.
2. **Hiệu suất Tiệm cận mô hình Giám sát (Supervised)**
   - Benchmark (TREC DL, BEIR): Chỉ số *nDCG@10* và *MAP* dự kiến vượt xa BM25, Contriever; tương đương với các mô hình tinh chỉnh trên MS-MARCO.
   - Tính ổn định: Hoạt động tốt trên các miền dữ liệu chuyên biệt (Y sinh, Tài chính).
3. **Khả năng Khái quát hóa và Sản phẩm**
   - Đa ngôn ngữ Zero-shot: Hiệu quả trên các ngôn ngữ nguồn lực thấp (Hàn, Nhật, Swahili...).
   - Triển khai "Cold-start": Giải pháp lý tưởng cho hệ thống mới khi chưa có dữ liệu gán nhãn.
   - Đóng góp: Bộ mã nguồn hoàn chỉnh và hướng dẫn thiết kế Instruction Prompting.

Tài liệu tham khảo

[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).