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 [1,2]

[1] datda.20@grad.uit.edu.vn

[2] Trường Đại học Công nghệ Thông tin, ĐHQG TP. HCM

What ?

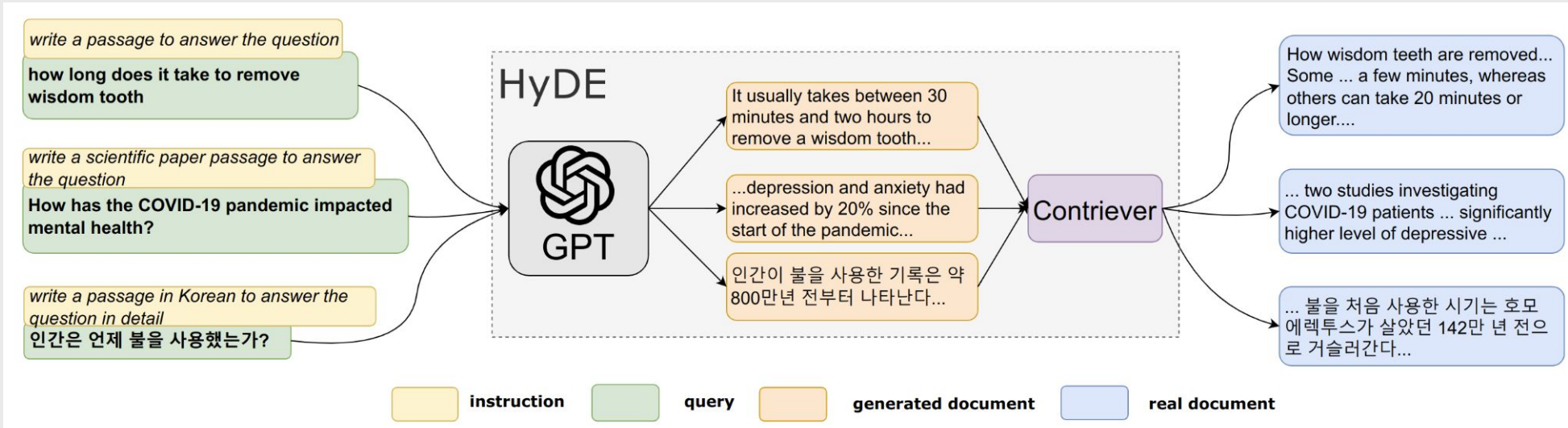
Giới thiệu **HyDE (Hypothetical Document Embeddings)**, một phương pháp truy xuất thông tin dày đặc (dense retrieval) đột phá:

- **Fully Zero-Shot:** Hệ thống hoạt động hoàn toàn không cần nhãn liên quan (relevance labels) và không cần huấn luyện mô hình (training-free).
- **Cơ chế:** Phân tách bài toán truy xuất thành hai bước: (1) Tạo sinh văn bản (Generative) bằng mô hình ngôn ngữ (LLM) và (2) So khớp độ tương đồng (Similarity) bằng bộ mã hóa tương phản không giám sát.
- **Hiệu suất cao:** HyDE cải thiện đáng kể so với các phương pháp không giám sát hiện có (như Contriever) và đạt hiệu suất tương đương với các hệ thống được tinh chỉnh (fine-tuned) trên nhiều tác vụ khác nhau.

Why ?

- **Thách thức của Zero-Shot:** Truy xuất dày đặc thường yêu cầu học không gian embedding chung giữa câu hỏi và tài liệu, điều này rất khó thực hiện nếu không có dữ liệu nhãn để định nghĩa sự "liên quan".
- **Sự khan hiếm dữ liệu:** Các bộ dữ liệu lớn có nhãn như MS-MARCO không phải lúc nào cũng có sẵn hoặc phù hợp cho các miền dữ liệu thực tế (do vấn đề bản quyền hoặc miền đặc thù).
- **Tận dụng LLM:** Các mô hình ngôn ngữ lớn (LLM) có khả năng hiểu chỉ dẫn và nắm bắt các mẫu liên quan (relevance patterns) rất tốt, ngay cả khi chúng tạo ra thông tin sai lệch (hallucinations). HyDE tận dụng điều này để thay thế cho việc học từ nhãn thủ công.

Overview



Description

1. Định nghĩa bài toán

Truy xuất dày đặc (Dense Retrieval) tính độ tương đồng giữa truy vấn $\mathbf{q}$ và tài liệu $\mathbf{d}$ qua tích vô hướng của hai vector:

$$sim(q, d) = \langle enc_q(q), enc_d(d) \rangle$$

Thách thức: Trong zero-shot, chúng ta không có dữ liệu nhãn để học hàm $\mathbf{enc_q}$ ánh xạ truy vấn vào cùng không gian với tài liệu

2. Hypothetical Document Embeddings - HyDE

Thay vì học $\mathbf{enc_q}$, **HyDE** sử dụng mô hình ngôn ngữ tạo sinh (Generative Language Model) $\mathbf{g}$ và bộ mã hóa tài liệu f (Contrastive Encoder) có sẵn.

- **Bước 1: Sinh tài liệu giả định (Hypothetical Generation):** Sử dụng Instruction-following LM để sinh tài liệu $\mathbf{d'}$ dựa trên truy vấn $\mathbf{q}$ và chỉ dẫn $\mathbf{INST}$

$$d' = g(q, INST)$$

- *Mục đích:* $\mathbf{d'}$ đóng vai trò như một mẫu dữ liệu chứa các patterns liên quan (relevance patterns).
- **Bước 2: Mã hóa & Lọc nhiễu (Encoding & Filtering)**: Vector truy vấn $\hat{\mathbf{v}}_\mathbf{q}$ được ước lượng bằng trung bình vector của $\mathbf{N}$ tài liệu giả định được lấy mẫu:

$$\hat{v}_q = \frac{1}{N} \sum_{k=1}^{N} f(\hat{d}_k)$$

- *Cơ chế:* Bộ mã hóa f hoạt động như một bộ nén có tổn thất (lossy compressor), lọc bỏ các chi tiết "ảo giác" (hallucinations) sai lệch và chỉ giữ lại các đặc trưng dày đặc cốt lõi.

3. Thiết lập thực nghiệm

3.1. Mô hình (Models):

- *Backbone:* InstructGPT (text-davinci-003) cho phần tạo sinh và Contriever/mContriever cho phần mã hóa.
- *Baselines:* BM25 (từ vựng), Contriever (unsupervised), DPR và ANCE (fine-tuned trên MS-MARCO).

3.2. Bộ dữ liệu (Datasets): Đánh giá trên 3 nhóm tác vụ chính để kiểm chứng khả năng tổng quát hóa:

- **Tìm kiếm Web (Web Search):** TREC DL19 và DL20 (dựa trên MS-MARCO).
- **Tài nguyên thấp (Low-Resource - BEIR):** Gồm 6 tập dữ liệu đa dạng miền: *SciFact* (Khoa học), *Arguana* (Tranh luận), *TREC-COVID* (Y tế), *FiQA* (Tài chính), *DBPedia* (Thực thể), *TREC-NEWS* (Tin tức).
- **Truy xuất đa ngôn ngữ (Multilingual Retrieval - Mr. TyDi):** Gồm 4 ngôn ngữ typology khác nhau: *Swahili, Korean, Japanese, Bengali*.

3.3. Kết quả chính

- **Web Search:** Trên TREC DL19, HyDE đạt MAP **41.8**, vượt xa Contriever (24.0) và tương đương mô hình có giám sát ContrieverFT (41.7).
- **Low-Resource:** HyDE vượt trội hơn Contriever gốc trên tất cả các dataset BEIR. Ví dụ: Trên Arguana, HyDE đạt nDCG@10 là **46.6** so với 37.9 của Contriever và 41.5 của ANCE.
- **Kết luận:** HyDE chứng minh rằng việc mô hình hóa sự liên quan (relevance modeling) có thể được chuyển giao cho LLM thông qua cơ chế sinh, loại bỏ nhu cầu về dữ liệu nhãn.

NII

Đặng Anh Đạt - Trường Đại học Công nghệ Thông tin ĐHQG TP. HCM

TEL: 0932091822 - Email: datda.20@grad.uit.edu.vn