MMLF: Multi-query Multi-passage Late Fusion Retrieval

**Yuan-Ching Kuo , Yi Yu, Chih-Ming Chen , Chuan-Ju Wang**

Academia Sinica, 2The Ohio State University

---

**Abstract**

MMLF là một phương pháp truy hồi thông tin mới, khai thác mô hình ngôn ngữ lớn (LLM) để tạo các truy vấn phụ (sub-queries), mở rộng thành các đoạn văn (passages), truy xuất độc lập, sau đó hợp nhất kết quả bằng kỹ thuật Reciprocal Rank Fusion (RRF). Phương pháp này cải thiện rõ rệt độ chính xác trên 5 tập dữ liệu chuẩn BEIR.

**Introduction**

Truy xuất thông tin (IR) nhằm mục đích xác định các tài liệu có liên quan từ các tập đoàn lớn để trả lời các truy vấn của người dùng. Mặc dù đã có nhiều nghiên cứu, việc truy xuất tùy ý vẫn tiếp tục đặt ra những thách thức, đặc biệt là với các truy vấn ngắn hoặc mơ hồ làm phức tạp việc suy luận chính xác ý định của người dùng. Các hệ thống IR truyền thống, chẳng hạn như BM25, dựa vào việc khớp thuật ngữ chính xác, thường gặp khó khăn với sự chồng chéo từ vựng hạn chế giữa các truy vấn và tài liệu.

Để giải quyết vấn đề này, các kỹ thuật mở rộng truy vấn sử dụng cơ sở kiến thức từ vựng, tận dụng các nguồn lực như các thuật ngữ học hoặc mạng ngữ nghĩa để thêm các thuật ngữ có liên quan về mặt ngữ nghĩa, bao gồm cả các từ đồng nghĩa (Bhogal và cộng sự, 2007; Qiu và Frei, 1993; Voorhees, 1994).

Với sự ra đời của các mô hình truy xuất dày đặc, việc cải tiến truy vấn đã chuyển từ khớp từ vựng sang sự tương đồng về ngữ nghĩa. Các mô hình này nhúng các truy vấn và tài liệu vào không gian vectơ được chia sẻ, tạo điều kiện thuận lợi lên đến 8% trong cả Recall@1k và nDCG@10 so với phương pháp tiên tiến, MILL.

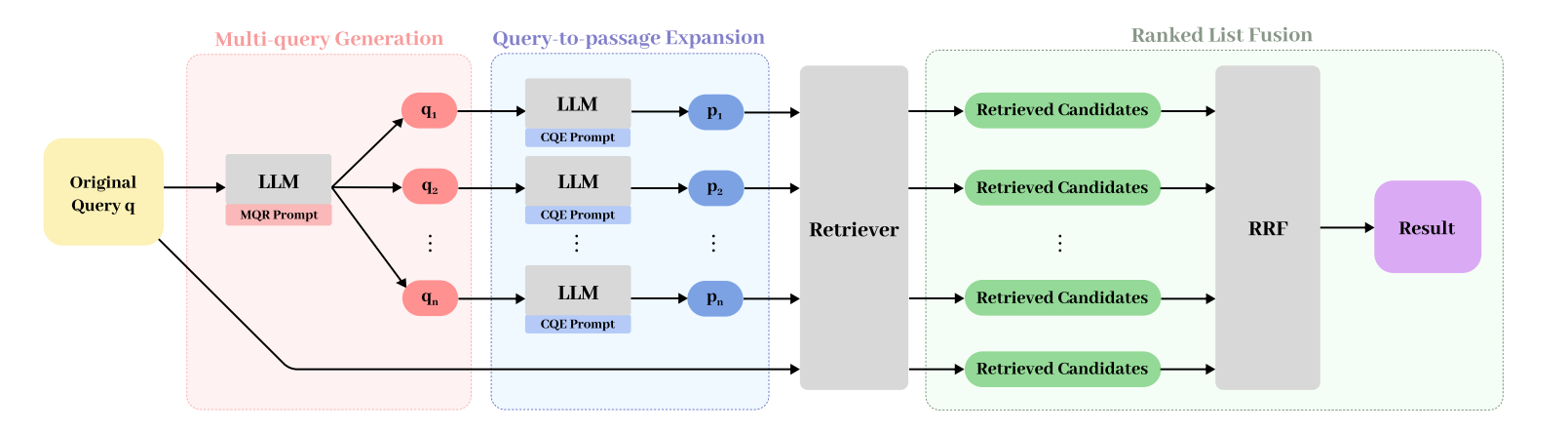
Ngoài ra, chúng tôi tiến hành các nghiên cứu cắt bỏ toàn diện để đánh giá các biến thể khác nhau của MMLF, bao gồm các phương pháp hợp nhất khác nhau, vai trò của

truy vấn ban đầu và các phương pháp cải cách truy vấn thay thế. Các nghiên cứu này cung cấp cái nhìn sâu sắc có giá trị về cơ sở lý luận đằng sau thiết kế của chúng tôi và tiếp tục xác nhận tính mạnh mẽ của MMLF được đề xuất.

**Tóm tắt (pptx)**

1. Bài toán truy hồi thông tin vẫn gặp khó khăn khi truy vấn ngắn hoặc mơ hồ.
2. Truy hồi truyền thống (như BM25) dựa vào từ vựng, không hiệu quả khi không có sự trùng lặp từ.
3. Mô hình dense retrieval và LLM giúp cải thiện bằng cách biểu diễn ngữ nghĩa truy vấn.
4. Các phương pháp hiện tại như Query2Doc, Chain-of-Thought, MILL tuy hiệu quả nhưng vẫn có hạn chế, đặc biệt là trong cách xử lý kết quả truy hồi.
5. MMLF được đề xuất như một pipeline hai giai đoạn nhằm tối ưu hóa quá trình phân tích truy vấn và hợp nhất kết quả.

**Pipeline Đề Xuất: MMLF**



Gồm 3 bước chính:

1. **Tạo truy vấn phụ (Multi-query Generation)**

Sử dụng LLM và prompt MQR để tạo ra 3 truy vấn phụ phản ánh nhiều khía cạnh của truy vấn gốc.

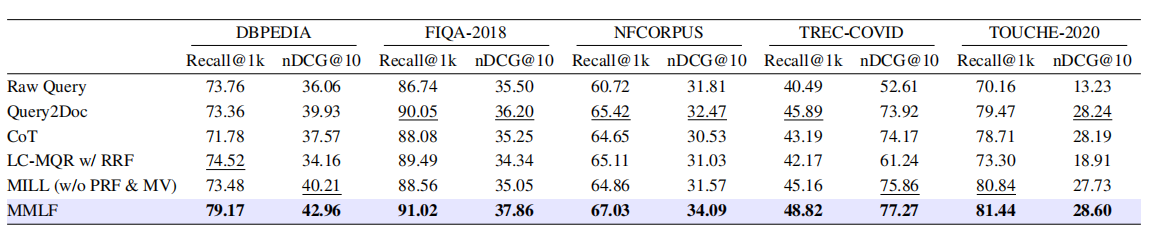
1. **Mở rộng truy vấn thành đoạn văn (Query-to-passage Expansion)**

Mỗi truy vấn phụ được LLM mở rộng thành đoạn văn giàu ngữ cảnh (pseudo-document) bằng prompt CQE.

1. **Hợp nhất danh sách xếp hạng (Ranked List Fusion)**

Truy xuất tài liệu từ mỗi đoạn văn (và truy vấn gốc), sau đó hợp nhất kết quả bằng RRF thay vì nối chuỗi như phương pháp MILL.

**Experiments**



1. **Dữ liệu và đánh giá**

Sử dụng 5 tập dữ liệu BEIR quy mô nhỏ: DBPEDIA, FIQA-2018, NFCORPUS, TREC-COVID, TOUCHE-2020.

Đánh giá bằng Recall@1k và nDCG@10.

1. **Thiết lập thực nghiệm**

Dùng mô hình LLaMA-3-70B-Instruct(top\_p = 1, in-clude temperature = 1) và bộ mã hóa e5-small-v2.

Tạo ra nhúng 384 chiều cho mỗi truy vấn và tài liệu

Tạo 3 truy vấn phụ cố định.

Các thí nghiệm bổ sung sử dụng LLM nhỏ hơn, Llama-3-8B-Instruct và một bộ mã hóa thay thế, Contriever5.

1. **So sánh với các phương pháp khác**

So sánh cách tiếp cận với 5 phương pháp cơ bản:

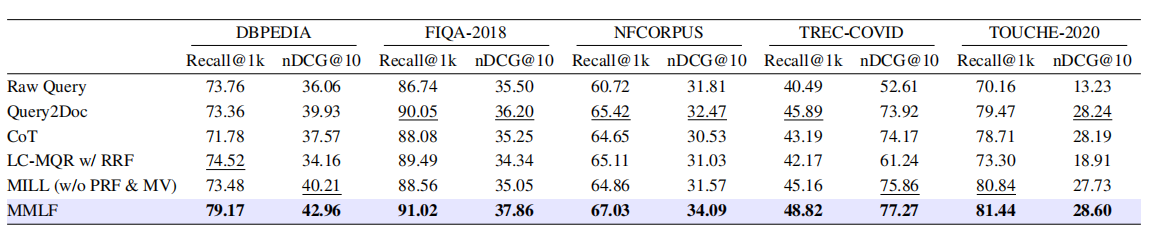
1. Raw Query, sử dụng truy vấn gốc mà không cần định dạng lại;
2. Query2Doc mở rộng truy vấn gốc thành một đoạn trích để truy xuất cùng với truy vấn;
3. Chain of-Thought (CoT) biến truy vấn thành câu trả lời và lý do để truy xuất cùng với truy vấn;
4. LangChain MultiQueryRetriever (LC-MQR) tạo nhiều truy vấn phụ và sử dụng hợp nhất thứ hạng qua lại (RRF) để xếp hạng lại;
5. MILL tạo truy vấn phụ và đoạn trích nhưng bỏ qua các kỹ thuật nâng cao như phản hồi giả liên quan và xác minh lẫn nhau để so sánh trực tiếp.

So sánh với 5 phương pháp: Raw Query, Query2Doc, CoT, LC-MQR (với RRF), MILL.

MMLF vượt trội nhất, tăng trung bình 4% Recall@1k so với MILL.

1. **Kết quả chính**

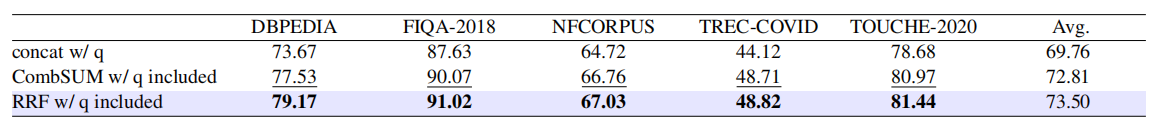
MMLF có kết quả tốt nhất ở cả Recall@1k và nDCG@10 trên tất cả 5 tập dữ liệu.



1. **Nghiên cứu tách biệt (Ablation Study)**

* So sánh kỹ thuật hợp nhất:

RRF vượt trội hơn CombSUM và concat trong Recall@1k, nhưng CombSUM tốt hơn về nDCG@10.



* Vai trò của truy vấn gốc:

Thí nghiệm này kiểm tra tác động của việc đưa truy vấn gốc vào quy trình RRF của đường ống MMLF.

Các cấu hình được đánh giá như sau:

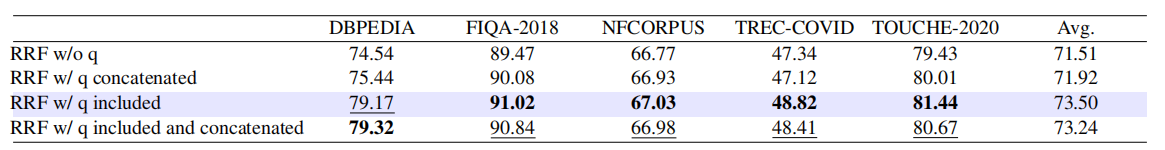
• **RRF w/o q**: Chỉ sử dụng các đoạn trích để truy xuất, không bao gồm truy vấn gốc.

• **RRF w/ q concatenated**: Nối truy vấn gốc với từng đoạn trích trước khi truy xuất.

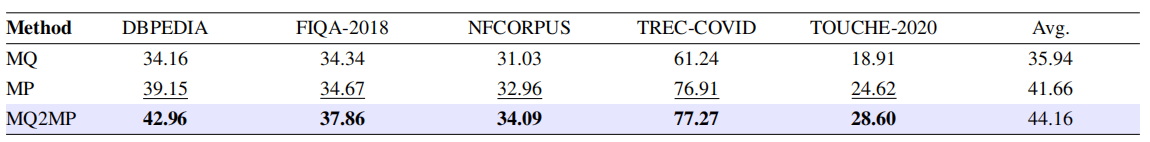
• **RRF w/ q included**: Sử dụng truy vấn gốc và các đoạn trích riêng biệt để truy xuất.

• **RRF w/ q included and concatenated**: Bao gồm truy vấn gốc riêng biệt và nối với từng đoạn trích trước khi truy xuất.

Đối với cả bốn cấu hình được mô tả ở trên, RRF được áp dụng cho kết quả truy xuất thu được.



Kết quả cho thấy việc bao gồm truy vấn gốc cùng với các đoạn văn luôn mang lại hiệu suất tốt hơn trong Recall@1k trên nDCG@10.

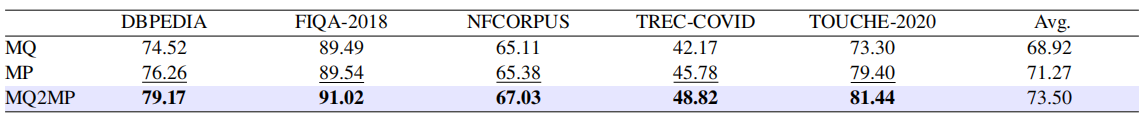


nó không vượt qua phương pháp tiếp cận hai giai đoạn. Phương pháp MQ2MP hai giai đoạn tạo ra các đoạn văn theo ngữ cảnh đa dạng, tương tự như phương pháp tiếp cận MP một giai đoạn. Tuy nhiên, như được chứng minh bằng kết quả Recall@1k và nDCG@10, MQ2MP luôn vượt trội hơn MP, cho thấy bước bổ sung là mở rộng các truy vấn phụ thành các đoạn văn nâng cao hiệu quả truy xuất.

Kết hợp truy vấn gốc với các đoạn văn cải thiện kết quả rõ rệt. “RRF w/ q included” đạt Recall@1k cao nhất.

* So sánh pipeline tạo truy vấn:

MQ2MP (hai bước: tạo sub-query rồi mở rộng) vượt trội hơn các phương án MQ (chỉ tạo sub-query) và MP (chỉ tạo passage).



**Conclusion**

Tác giả giới thiệu MMLF, một pipeline truy xuất thông tin (IR) hiệu quả và mạnh mẽ tận dụng các mô hình ngôn ngữ lớn (LLM) để tạo ra các truy vấn phụ đa dạng và mở rộng chúng thành các tài liệu giả theo ngữ cảnh.

Các thí nghiệm chứng minh những cải tiến đáng kể về hiệu suất truy xuất, đặc biệt là trong **Recall@1k**, trên nhiều tập dữ liệu mà không cần tinh chỉnh mô hình. MMLF trình bày một giải pháp có khả năng mở rộng và thích ứng để cải thiện hiệu suất tìm kiếm trên nhiều miền khác nhau, kết hợp độc đáo việc phân tích truy vấn và tạo đoạn văn để thúc đẩy việc định dạng lại truy vấn cho việc truy xuất dày đặc.

**Limitations**

Mặc dù có hiệu quả, MMLF vẫn có một số thách thức về mặt tính toán. Về mặt tạo,

phương pháp này yêu cầu chạy suy luận với LLM, **có thể chậm đáng kể** do giải mã tự động hồi quy từng mã thông báo. Nhu cầu tạo nhiều truy vấn phụ và mở rộng từng truy vấn thành các đoạn chi tiết làm **tăng đáng kể khối lượng tính toán**.

Mặc dù xử lý song song các đoạn có thể giảm bớt một số chi phí này, nhưng nó vẫn

là một **nút thắt cổ chai**, đặc biệt là trong các hệ thống thời gian thực hoặc quy mô lớn.

Về mặt truy xuất, MMLF xử lý và truy xuất độc lập các tài liệu cho mỗi đoạn mở rộng, làm **tăng khối lượng công việc truy xuất** tổng thể.

Việc áp dụng hợp nhất thứ hạng qua lại (RRF) để kết hợp các kết quả này làm tăng thêm tính phức tạp, đặc biệt là khi mở rộng quy mô lên các tập hợp dữ liệu lớn hơn hoặc khối lượng truy vấn lớn.Mặc dù đã cố định số lượng truy vấn phụ được tạo ra là ba trong các thí nghiệm nhưng không thực hiện nghiên cứu cắt bỏ để đánh giá cách

thay đổi số lượng truy vấn phụ ảnh hưởng đến hiệu suất truy xuất và chi phí tính toán. Đây là một lĩnh vực để điều tra trong tương lai có thể khám phá sự đánh đổi giữa tính đa dạng của truy vấn, hiệu quả truy xuất và nhu cầu tính toán. Những thách thức tính toán này có thể hạn chế khả năng áp dụng của MMLF trong các môi trường bị hạn chế bởi tài nguyên hoặc thời gian.

Nghiên cứu trong tương lai có thể tập trung vào việc tối ưu hóa giai đoạn suy luận LLM và tác giả đang chứng minh hiệu quả của các kỹ thuật truy xuất và hợp nhất

để giảm chi phí tính toán mà không ảnh hưởng đến hiệu quả truy xuất.

**References**