MMLF: Multi-query Multi-passage Late Fusion Retrieval

Yuan-Ching Kuo[1] , Yi Yu[1,2], Chih-Ming Chen[1] , Chuan-Ju Wang[1]
[1]Academia Sinica, [2]The Ohio State University

Abstract

Leveraging large language models (LLMs) for query expansion has proven highly effective across diverse tasks and languages. Yet, challenges remain in optimizing query formatting and prompting, often with less focus on handling retrieval results. In this paper, we introduce **Multi-query Multi-passage Late Fusion (MMLF)**, a straightforward yet potent pipeline that generates sub-queries, expands them into pseudo-documents, retrieves them individually, and aggregates results using reciprocal rank fusion. Our experiments demonstrate that MMLF exhibits superior performance across five BEIR benchmark datasets, achieving an **average improvement of 4%** and a **maximum gain of up to 8% in both Recall@1k and nDCG@10** compared to state of the art across BEIR information retrieval datasets.

Methodology

Multi-query Generation
- Generate multiple sub-queries (q_1, q_2, ..., q_n) from an original query q using an LLM with the MQR prompt.
- Captures diverse interpretations of user intent, broadening the retrieval scope.

Query-to-passage Expansion
- Expand each sub-query into a passage (pseudo-document) using an LLM with the CQE prompt.
- Provides richer context, enhancing retrieval accuracy.

Ranked List Fusion
- Retrieve documents separately using the original query and expanded passages.
- Apply Reciprocal Rank Fusion (RRF) to merge ranked lists, prioritizing consistently relevant documents.
- Unlike simple concatenation, RRF mitigates dilution of relevance.

Experiments

- LLM: Llama-3-70B-Instruct (temperature = 1, top-p = 1)
- Encoder: e5-small-v2 (384-dimensional embeddings)
- Sub-queries fixed at 3 for consistency

MMLF consistently outperforms all baselines in Recall@1k and nDCG@10 across five datasets. Specifically, our approach achieves an average improvement of 4% in Recall@1k over the closest competitor, MILL, demonstrating a substantial gain, particularly given the high performance of existing methods. (Table 1)

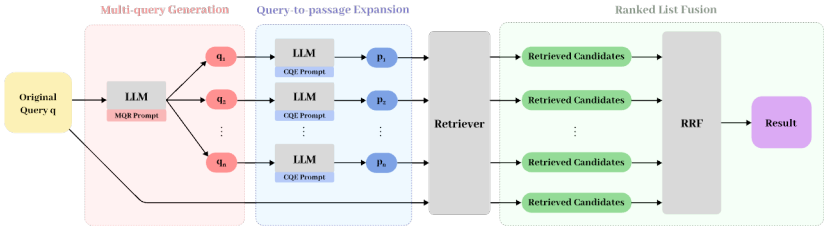

Figure 1. The MMLF pipeline3

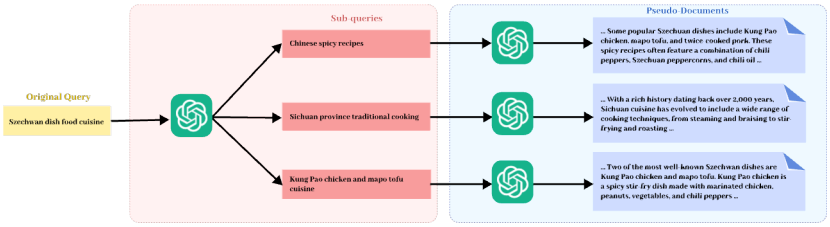

Figure 2. Illustration of Multi-Query Generation and Query-to-Passage Expansion on the DBPedia query.

Ablation Study

Fusion Method Comparison (Figure 3)

✓ **RRF:** Rank-based late fusion. Aggregates retrieval results based on document ranks instead of similarity scores. The final score for each document d is:

$$score_{RRF}(d) = \sum_{i=0}^{3} \frac{1}{k + rank_i(d)}$$

- **CombSUM:** Score-based late fusion. Aggregates similarity scores from retrieval results of the query and individual passages. The final score for each document d is:

$$score_{CombSUM}(d) = \sum_{i=0}^{3} score_i(d)$$

- **Concatenation:** Early fusion. Concatenates the original query and passages into a single sequence before retrieval:

$$R(concat(q, [SEP], p_1, [SEP], p_2, [SEP], p_3))$$

Role of the Original Query (Figure 4)
- **RRF w/o q:** Uses only expanded passages.
- **RRF w/ q concatenated:** Concatenates the original query with each passage.
- ✓ **RRF w/ q included:** Retrieves documents separately using the original query and passages.
- **RRF w/ q included + concatenated:** Combines both strategies for retrieval.

Query Reformulation Pipeline (Figure 5)
- **RawQuery:** Uses query without reformulation.
- **MQ:** Uses sub-queries directly for retrieval.
- **MP:** Expands the original query into passages without sub-queries.
- ✓ **MQ2MP:** Generates sub-queries first, then expands them into passages.

The check mark (✓) denotes the best-performing method in each category, also used for MMLF.

	DBPEDIA		FIQA-2018		NFCORPUS		TREC-COVID		TOUCHE-2020	
	Recall@1k	nDCG@10	Recall@1k	nDCG@10	Recall@1k	nDCG@10	Recall@1k	nDCG@10	Recall@1k	nDCG@10
RawQuery	73.76	36.06	86.74	35.50	60.72	31.81	40.49	52.61	70.16	13.23
Query2Doc	73.36	39.93	90.05	36.20	65.42	32.47	45.89	73.92	79.47	28.24
CoT	71.78	37.57	88.08	35.25	64.65	30.53	43.19	74.17	78.71	28.19
LC-MQR w/RRF	74.52	34.6	89.49	34.34	65.11	31.03	42.17	61.24	73.30	18.91
MILL (w/o PRF & MV)	73.48	40.21	88.56	35.05	64.86	31.57	45.13	75.86	80.84	27.73
MMLF	**79.17**	**42.96**	**91.02**	**37.86**	**67.03**	**34.09**	**48.82**	**77.27**	**81.44**	**28.60**

Table 1: Main Results

Conclusions

We introduced MMLF, a robust and efficient information retrieval pipeline that significantly enhances performance across multiple datasets without requiring model fine-tuning. By uniquely integrating query decomposition and passage generation, MMLF offers a scalable and adaptable solution for improving search effectiveness across diverse domains.

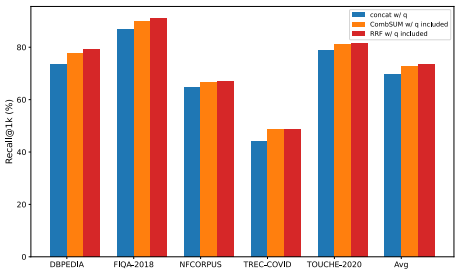

Figure 3. Fusion methods comparison

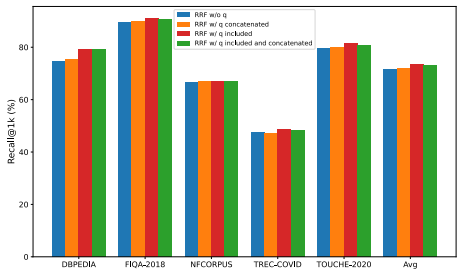

Figure 4. Impact of including the original query

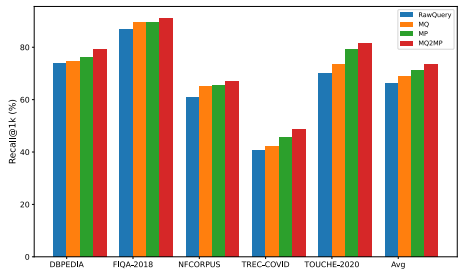

Figure 5. Impact of generating passages in two stages