

← Go to **ACL ARR 2024 October** homepage (/group?id=aclweb.org/ACL/ARR/2024/October)

Knowledge-Aware Query Expansion with Large Language Models for Textual and Relational Retrieval



Yu Xia (/profile?id=~Yu_Xia9), *Junda Wu* (/profile?id=~Junda_Wu1),
Sungchul Kim (/profile?id=~Sungchul_Kim1), *Tong Yu* (/profile?id=~Tong_Yu3),
Ryan A. Rossi (/profile?id=~Ryan_A._Rossi2),
Haoliang Wang (/profile?id=~Haoliang_Wang1),
Julian McAuley (/profile?id=~Julian_McAuley1)

16 Oct 2024 (modified: 20 Dec 2024) ACL ARR 2024 October Submission October, Senior Area Chairs, Area Chairs, Reviewers, Authors, Commitment Readers Revisions (/revisions?id=e2q6SZGWTd) CC BY 4.0
 (https://creativecommons.org/licenses/by/4.0/)

Abstract:

Large language models (LLMs) have been used to generate query expansions augmenting original queries for improving information search. Recent studies also explore providing LLMs with initial retrieval results to generate query expansions more grounded to document corpus. However, these methods mostly focus on enhancing textual similarities between search queries and target documents, overlooking document relations. For queries like "Find me a highly rated camera for wildlife photography compatible with my Nikon F-Mount lenses", existing methods may generate expansions that are semantically similar but structurally unrelated to user intents. To handle such semi-structured queries with both textual and relational requirements, in this paper we propose a knowledge-aware query expansion framework, augmenting LLMs with structured document relations from knowledge graph (KG). To further address the limitation of entity-based scoring in existing KG-based methods, we leverage document texts as rich KG node representations and use document-based relation filtering for our Knowledge-Aware Retrieval (KAR). Extensive experiments on three datasets of diverse domains show the advantages of our method compared against state-of-the-art baselines on textual and relational semi-structured retrieval.

Paper Type: Long

Research Area: Information Retrieval and Text Mining

Research Area Keywords: query expansion, document retrieval

Contribution Types: Model analysis & interpretability, NLP engineering experiment

Languages Studied: English

Reassignment Request Action Editor: This is not a resubmission

Reassignment Request Reviewers: This is not a resubmission

A1 Limitations Section: This paper has a limitations section.

A2 Potential Risks: N/A

B Use Or Create Scientific Artifacts: No

B1 Cite Creators Of Artifacts: N/A

B2 Discuss The License For Artifacts: N/A

B3 Artifact Use Consistent With Intended Use: N/A

B4 Data Contains Personally Identifying Info Or Offensive Content: N/A

B5 Documentation Of Artifacts: N/A

B6 Statistics For Data: N/A

C Computational Experiments: Yes

C1 Model Size And Budget: Yes

C1 Elaboration: Section 5

C2 Experimental Setup And Hyperparameters: Yes

C2 Elaboration: Section 5

C3 Descriptive Statistics: Yes

C3 Elaboration: Section 5

C4 Parameters For Packages: Yes

C4 Elaboration: Section 5

D Human Subjects Including Annotators: No

D1 Instructions Given To Participants: N/A

D2 Recruitment And Payment: N/A

D3 Data Consent: N/A

D4 Ethics Review Board Approval: N/A

D5 Characteristics Of Annotators: N/A

E Ai Assistants In Research Or Writing: No

E1 Information About Use Of Ai Assistants: N/A

Reviewing Volunteers: 👁 Junda Wu (/profile?id=~Junda_Wu1)

Reviewing No Volunteers Reason: 👁 N/A - An author was provided in the previous question.

Reviewing Volunteers For Emergency Reviewing: 👁 The volunteers listed above are willing to serve either as regular reviewers or as emergency reviewers.

Preprint: 👁 no

Preprint Status: 👁 We plan to release a non-anonymous preprint in the next two months (i.e., during the reviewing process).

Consent To Share Data: 👁 yes

Consent To Share Submission Details: 👁 On behalf of all authors, we agree to the terms above to share our submission details.

Author Submission Checklist: 👁 I confirm that the paper is anonymous and that all links to data/code repositories in the paper are anonymous., I confirm that the paper has proper length (Short papers: 4 content pages maximum, Long papers: 8 content pages maximum, Ethical considerations and Limitations do not count toward this limit), I confirm that the paper is properly formatted (Templates for *ACL conferences can be found here: <https://github.com/acl-org/acl-style-files> (<https://github.com/acl-org/acl-style-files>).

Association For Computational Linguistics - Blind Submission License Agreement: 👁 On behalf of all authors, I agree

Submission Number: 2488

Discussion (/forum?id=e2q6SZGWTd#discussion)

Filter by reply type... ▼

Filter by author... ▼

Search keywords...

Sort: Newest First

☰

☰

☰

-

=

☰

🔗

👁

Everyone

Submission2488...

Submission2488 Area...

Submission2488 Authors

14 / 15 replies shown

Submission2488...

Program Chairs

Submission2488...

Submission2488...

Submission2488...

Submission2488...

✕

Add: Author-Editor Confidential Comment Withdrawal

Meta Review of Submission2488 by Area

Chair hVop

Meta Review by Area Chair hVop 📅 04 Dec 2024, 09:22 (modified: 20 Dec 2024, 12:17)

👁️ Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

📄 Revisions (/revisions?id=6muYwaeBtu)

Metareview:

This paper proposes an approach for query expansion that considers relations among retrieved documents, rather than relying on only textual information. The proposed approach performs well in comparison with reasonable baselines, and the proposed document filtering approach outperforms entity-based filtering from prior work.

Summary Of Reasons To Publish:

This is an interesting angle on LLM-powered query expansion, which improves upon reasonable baselines by considering both textual information and relations among retrieved documents. Reviewers appreciated this perspective, noted that the proposed approach performs well in experiments, and all agree that the paper is sound.

Summary Of Suggested Revisions:

There were several minor clarifications in the author discussion that could be incorporated into the paper, such as the prompt used and comparisons with additional methods. Given the question raised by one reviewer about how the method extends prior work, it might help to clarify this as done in the author response (i.e., "Response to Weakness 2").

Overall Assessment: 4 = There are minor points that may be revised

Suggested Venues: NAACL

Best Paper Ae: No

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Author Identity Guess: 4 = From an allowed pre-existing preprint or workshop paper, I know/can guess at least one author's name.

Reported Issues: Yes, and I took them into account in my meta-review

Note To Authors: The reported issues were taken into account.

Add: **Author-Editor Confidential Comment**

Official Review of Submission2488 by Reviewer tbxG

Official Review by Reviewer tbxG 📅 13 Nov 2024, 15:15 (modified: 20 Dec 2024, 12:17)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer tbxG, Commitment Readers

📄 Revisions (/revisions?id=5dEKVCzhI6)

Paper Summary:

This paper proposes the knowledge-aware retrieval for handling semi-structured queries, thereby handling both textual and relational representation and boosting the retrieval performance.

Summary Of Strengths:

- Experimental results shows that proposed method boost the retrieval performance.
- The paper is easy to read.

Summary Of Weaknesses:

- Maybe explanation is not sufficient to fully understand. Because, for example in Section 4 — Entity Parsing by LLM, author referenced Gao et al. (2023) and mentioned followed the idea. But, it seems there is no prompt example (even in appendix). Did you use prompt in Table 8 for Entity Parsing?
- Could you please insert the experimental results of [1], [2]?
- Inserting the experiment result (e.g., GritLM-7b) from [3] — Table 6, 7 would be better for confirming the effectiveness of the proposed paper.

- I recommend moving the Section 5.3 (Implementation Details) to Appendix.
- I have no idea to confirm the reproducibility because author doesn't upload the any software data. But, I believe that it can be reproduced with difficulty.

[1]. **Query2doc: Query Expansion with Large Language Models, Wang et al, EMNLP 23'**

[2]. **Query Expansion by Prompting Large Language Models, Jagerman et al, arxiv preprint 23'**

[3]. **STaRK: Benchmarking LLM Retrieval on Textual and Relational Knowledge Bases, Wu et al, NeurIPS 2024'**

Comments Suggestions And Typos:

See above.

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Soundness: 3.5

Overall Assessment: 3.5

Best Paper: No

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**



Reply to Reviewer tbxG

Official Comment

by Authors (Yu Xia (/profile?id=~Yu_Xia9), Tong Yu (/profile?id=~Tong_Yu3), Julian McAuley (/profile?id=~Julian_McAuley1), Haoliang Wang (/profile?id=~Haoliang_Wang1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission2488/Authors))

25 Nov 2024, 13:24 (modified: 02 Jan 2025, 08:15)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer tbxG, Commitment Readers

Revisions (/revisions?id=uVvinyxNL0)

Comment:

Thank you for your time and efforts in reviewing our paper. We provide our point-by-point responses as follows:

Response to Weakness 1:

Thank you for your feedback. Yes, we directly use the prompt in Table 8 (Parse) of the Appendix to parse entities mentioned in the query, where we instruct the LLM to output these entities in document formats as *{document type: {document attributes}}* so that they can be directly used for subsequent entity document retrieval.

For example, the query in Figure 2 *"Find me a paper about high-resolution image recovery written by Andrew Stokes in 2010 and citing the paper Multi-aperture coherent imaging"* is first parsed by the LLM using the prompt in Table 8 into the following two entities:

- *{author: {author name: Andrew Stokes}}*
- *{paper: {paper title: Multi-aperture coherent imaging}}*

Similarly to Gao et al. (2023), where they consider the query itself as a pseudo-document in query expansion, we include the query itself as a pseudo-entity representing the target entity document to be retrieved, as discussed in Line 229 and illustrated in Figure 2. We will include additional details and explanations to improve the clarity of each module in our method.

Response to Weakness 2:

Thank you for your suggestion. As discussed in Line 128 of Section 2, Query2Doc [1] can be considered as a few-shot variant of HyDE [4], where the LLM is provided with a few query expansion examples before generation. In our experiments, we follow [5] to consider a zero-shot setting and thus mainly compare our method with zero-shot query expansion baselines, e.g., [4][5].

As it is indeed helpful to have more comprehensive comparisons with existing methods, we report and will include the results of Q2D (Query2Doc [1]) and CoT (Chain-of-Thought as Query Expansion [2]) on human-generated queries as follows:

	AMAZON				MAG				PRIME			
	Hit@1	Hit@5	R@20	MRR	Hit@1	Hit@5	R@20	MRR	Hit@1	Hit@5	R@20	MRR
Q2D	38.27	61.73	36.90	51.13	32.14	42.86	37.48	37.08	28.44	53.21	53.04	38.63
CoT	49.38	71.60	38.75	60.09	22.62	36.90	30.92	29.63	26.61	46.79	51.13	36.84
KAR (Ours)	61.73	72.84	40.62	66.32	51.20	58.33	46.60	54.52	44.95	60.55	59.90	51.85

Response to Weakness 3:

Thank you for your suggestion. Since the updated version of the STaRK benchmark [3] with new experimental results (e.g., GritLM-7b) was released after the ARR October submission deadline, our current submission only includes the benchmark results of the earlier version. We will include the updated benchmark results as suggested for more comprehensive comparisons.

Response to Weakness 4:

Thank you for your suggestion. We will improve the organization of the main body of our paper.

Response to Weakness 5:

Thank you for your feedback. We are planning to release our source code upon acceptance and approval to facilitate reproducibility and further studies.

[1] Query2doc: Query Expansion with Large Language Models, Wang et al, EMNLP 23'

[2] Query Expansion by Prompting Large Language Models, Jagerman et al, arxiv preprint 23'

[3] STaRK: Benchmarking LLM Retrieval on Textual and Relational Knowledge Bases, Wu et al, NeurIPS 2024'

[4] Gao, Luyu, et al. Precise Zero-Shot Dense Retrieval without Relevance Labels. ACL 2023.

[5] Chen, Xinran, et al. Analyze, generate and refine: Query expansion with LLMs for zero-shot open-domain QA. ACL 2024 Findings.

Add: Author-Editor Confidential Comment



Response by Reviewer

Official Comment by Reviewer tbxG 📅 26 Nov 2024, 01:39 (modified: 02 Jan 2025, 08:15)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer tbxG, Commitment Readers

📄 Revisions (/revisions?id=KwHZm0Urbf)

Comment:

Thank you for your response. I have carefully read your response, I believe that my concerns have mostly resolved.

- Concern 1. According to the **Response to Weakness 1**, authors have clearly clarified the use of prompt template and confirmed that will be included additional details and explanations to improve the clarity of each module in paper.
- Concern 2. According to the **Response to Weakness 2**, author have provided us experimental results which I asked. Based on the experimental results, KAR have confirmed the boosting improvements.
- Concern 3. According to the **Response to Weakness 3**, author have clarified that the result what I have asked is updated after the ARR October submission deadline. Even my concern is not resolved, I will not ask more clarification. This weakness I have mentioned will not be reflected to my score.
- Concern 4. According to the **Response to Weakness 4**, author have mentioned that authors will improve the organization of the main body of our paper.
- Concern 5. I still have no idea to confirm the reproducibility. However, author have mentioned that they are planning to release the source code. If released, it will be helpful for NLP community.

In overall, I believe that proposed method is succinct and efficient to handle the query expansion for improving the performance of retrieval. The novelty of proposed method may be somewhat limited, however I believe that this is sufficiently over the bar. Therefore, I raised my score from 3 to 3.5
(Overall Assessment)

Add: **Author-Editor Confidential Comment**



➔ *Replying to Response by Reviewer*

Reply to Reviewer tbxG

Official Comment

by Authors (Yu Xia (/profile?id=~Yu_Xia9), Tong Yu (/profile?id=~Tong_Yu3), Julian McAuley (/profile?id=~Julian_McAuley1), Haoliang Wang (/profile?id=~Haoliang_Wang1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission2488/Authors))

26 Nov 2024, 09:27 (modified: 02 Jan 2025, 08:15)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer tbxG, Commitment Readers

Revisions (/revisions?id=OBHrsGrgBD)

Comment:

Thank you for your recognition of our work and your valuable time in reviewing our paper.

Add: **Author-Editor Confidential Comment**

Official Review of Submission2488 by Reviewer jrGA

Official Review by Reviewer jrGA 10 Nov 2024, 20:36 (modified: 20 Dec 2024, 12:17)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer jrGA, Commitment Readers

Revisions (/revisions?id=VoQmtBNaor)

Paper Summary:

The paper proposes Knowledge-Aware Retrieval (KAR), a method to enhance query expansion by combining LLMs with KG relations. This approach improves retrieval accuracy for queries that need both textual and relational information. Evaluated on diverse datasets, KAR demonstrates strong effectiveness for semi-structured retrieval tasks.

Summary Of Strengths:

- S1: How to utilize the relations in knowledge graphs is an interesting and practically important research problem.
- S2: The proposed method achieves strong empirical performance across datasets.
- S3: The authors conducted comprehensive analyses on the proposed method. This will provide rich insights for practitioners when they use the proposed method.

Summary Of Weaknesses:

- W1: Since the proposed method uses document details, it is unclear whether the method is scalable for documents that have long details. It might be better if the authors provide a scalability test (e.g., latency vs detail length).
- W2: The proposed method seems to have limited novelty. As the author said in Section 4, the main novelty is the document-based relation filtering (DRF) component while all other components are prior work. In addition to that, this DRF component (i.e., utilizing document details) is rather simple. It is more like a feature engineering trick than a novel method.
- W3: Table 3 shows that KAR without DRF is significantly better than KAR on Amazon in terms of Hit@5, and Table 4 shows that even the simple method HyDE sometimes outperforms KAR. These results seem to suggest that DRF can be unhelpful in some cases. However, the authors did not analyze these results. It might be better if the authors discuss the underlying assumptions that DRF relies on and empirically analyze why these assumptions are violated here.

Comments Suggestions And Typos:

- C1: The authors did not justify the evaluation metrics they chose. It is unclear why they chose Hit@1, Hit@5, and Recall@20 but did not choose other metrics like Hit@20 or Recall@5.
- C2: The authors did not discuss the limitations of the proposed method.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Soundness: 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.

Overall Assessment: 2.5

Best Paper: No

Best Paper Justification:

NA

Limitations And Societal Impact:

The authors did not discuss limitations or societal impact of this work. See weaknesses and comments above.

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**



Kind Reminder of Author-Reviewer Discussion Period

Official Comment

by Authors (👁️ Yu Xia (/profile?id=~Yu_Xia9), Tong Yu (/profile?id=~Tong_Yu3), Julian McAuley (/profile?id=~Julian_McAuley1), Haoliang Wang (/profile?id=~Haoliang_Wang1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission2488/Authors))

📅 26 Nov 2024, 14:21 (modified: 02 Jan 2025, 08:15)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer jrGA, Commitment Readers

📄 Revisions (/revisions?id=Gtu0hBZADH)

Comment:

Dear Reviewer jrGA,

As the author-reviewer discussion period is ending soon (in about 13 hours), we would like to follow up on whether our responses have fully addressed your concerns. We are also happy to answer any further questions you might have.

Best,

Authors

Add: **Author-Editor Confidential Comment**



Reply to Reviewer jrGA

Official Comment

by Authors (👁️ Yu Xia (/profile?id=~Yu_Xia9), Tong Yu (/profile?id=~Tong_Yu3), Julian McAuley (/profile?id=~Julian_McAuley1), Haoliang Wang (/profile?id=~Haoliang_Wang1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission2488/Authors))

📅 25 Nov 2024, 13:35 (modified: 02 Jan 2025, 08:15)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer jrGA, Commitment Readers

📄 Revisions (/revisions?id=3KxmvpdBJU)

Comment:

Thank you for your time and efforts in reviewing our paper. We provide our point-by-point responses as follows:

Response to Weakness 1:

Thank you for your suggestion. With longer document details, the latency for LLM-based query expansion methods indeed increases as input token sequences grow longer for both LLM inference and dense embedding generation. This can be further observed from the dataset statistics in Table 1 and the latency comparison in Figure 5.

Specifically, from Table 1, we can observe that AMAZON (avg. 571.7 tokens per document) has longer document details than MAG (avg. 113.5 tokens per document). Despite MAG having a larger retrieval candidate pool than AMAZON, which might increase latency, all compared methods in Figure 5 show relatively lower overall latency on MAG than on AMAZON (KAR -28.0%, AGR -33.9%, RAR -27.3%, HyDE -41.8%). While in our experiments, the latency of API calls might also be affected by varying server load, we will include additional discussion regarding these observations about the scalability to different document lengths.

Response to Weakness 2:

Thank you for your feedback. Unlike prior works [1][2][3][4] that simply leverage either texts or KGs, we propose a novel knowledge-aware query expansion framework that jointly utilizes semantically rich textual descriptions and structured KG relations for LLM-enhanced document retrieval. As texts and KGs are among the most

common types of knowledge sources, we believe our method can inspire further research and studies on integrating them for more accurate and grounded LLM outputs. In addition, the strong empirical effectiveness of our method also demonstrates its applicability for handling complex user queries with various requirements in practice.

Response to Weakness 3:

Thank you for your insightful observation and suggestion. As discussed in Line 078 and Line 260, the motivation and underlying assumption of applying DRF is that user queries often involve textual details of target entities that cannot be simply reflected by entity names or titles. For example, users may search for a product with high-rated reviews or for a paper published in 2010—information about reviews and dates is unlikely to appear in entity titles. With DRF, our method utilizes document details to capture more fine-grained relevance between user queries and entity documents for more accurate retrieval.

Regarding your observation, while in Table 1 our KAR method shows a clear advantage over KAR_w/o_DRF on AMAZON in Hit@1, it is outperformed by KAR_w/o_DRF in Hit@5. This indeed implies that DRF can be unhelpful in some cases as you suggested. By examining specific cases on AMAZON where KAR_w/o_DRF performed better than KAR, we find that while document details used in DRF help identify the target product satisfying all user query requirements (favoring Hit@1 metrics), the textual documents for AMAZON products sometimes include lengthy and irrelevant Q&A and reviews, introducing noise when calculating the relevance between the query and document during DRF.

For example, the target entity document for the query *“Could you help me find shooting sticks that have an aluminum alloy tipped foot designed for all kinds of uneven terrain?”* includes not only a detailed description of alloy tipped foot for shooting sticks but also customer reviews complaining about the service *“... Called and spoke to ... just before they closed for the day. Tech support basically said they couldn't believe my product was defective as they make 'great' stuff...”*. On the other hand, the entity-based filtering in KAR_w/o_DRF retains entities that are directly relevant to the entity names mentioned in the query, e.g., “shooting sticks,” which can favor the Hit@5 metric over first-document accuracy.

We will include additional analysis of these results as you suggested and also discuss possible further improvements, e.g., document chunking.

Response to Comment 1:

Thank you for your comment. As described in Line 356, we use the same evaluation metrics reported in the STaRK benchmark [5], i.e., Hit@1, Hit@5, Recall@20, and MRR. These metrics capture top-ranked accuracy, broader coverage of relevant items, and ranking quality, enabling direct comparisons with benchmark results.

Response to Comment 2:

Thank you for your comment. As discussed in the Limitations section (Line 595–Line 609), one main limitation of our proposed method is the extra retrieval latency introduced by additional LLM inference and KG usage compared to direct retrieval. We will also include further discussion of our method on handling lengthy document details as you suggested above.

Add: **Author-Editor Confidential Comment**




Official Comment by Authors

Official Comment

by Authors (👁️ Yu Xia (/profile?id=~Yu_Xia9), Tong Yu (/profile?id=~Tong_Yu3), Julian McAuley (/profile?id=~Julian_McAuley1), Haoliang Wang (/profile?id=~Haoliang_Wang1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission2488/Authors))

📅 25 Nov 2024, 13:36 (modified: 02 Jan 2025, 08:15)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer jrGA, Commitment Readers

 Revisions (/revisions?id=RdTCKrppvO)

Comment:

[1] Gao, Luyu, et al. Precise Zero-Shot Dense Retrieval without Relevance Labels. ACL 2023.

[2] Chen, Xinran, et al. Analyze, generate and refine: Query expansion with LLMs for zero-shot open-domain QA. ACL 2024 Findings.

[3] Yasunaga, Michihiro, et al. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. NAACL 2021.

[4] Zhang, Xikun, et al. GreaseLM: Graph REASONing Enhanced Language Models. ICLR 2022

[5] Wu, Shirley, et al. STaRK: Benchmarking LLM Retrieval on Textual and Relational Knowledge Bases. NeurIPS 2024 Benchmark and Datasets

Add: **Author-Editor Confidential Comment**



Official Review of Submission2488 by Reviewer YLJQ

Official Review by Reviewer YLJQ  10 Nov 2024, 18:19 (modified: 20 Dec 2024, 12:17)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer YLJQ, Commitment Readers

 Revisions (/revisions?id=FizaGCRrtG)

Paper Summary:

The paper proposes combining domain knowledge from a knowledge graph with the general intrinsic knowledge of a large language model (LLM) for query expansion in information retrieval settings. The approach first uses an LLM to identify a set of seed entities from the original query. By following links (relations) in the knowledge graph, these seed entities are expanded into a small query-focused graph representing the relational context of the query. Since the initially constructed query graph could be noisy, the authors introduce a filtering step based on the semantic similarity between the original query and the entities' descriptions. Finally, the query-focused graph is used to prompt an LLM to generate an expanded query, which is then combined with the original query for retrieval. The paper conducts multiple experiments and ablation studies to evaluate the effectiveness and efficiency of the proposed approach. The reported results seem to support the authors' claims.

Summary Of Strengths:

- The paper presents an interesting approach with good results, demonstrating the complementary benefits of using knowledge graphs and language models for query expansion. This KG-augmented approach provides a complementary perspective to the recent retrieval-augmented generation approaches for query expansion.
- An extensive set of experiments is conducted to assess the effectiveness of the components introduced in the paper (e.g., adding a knowledge graph, document-based filtering). Ablation studies are included to evaluate the sensitivity of the model to its hyperparameters.
- The paper is generally well written and logically organized.

Summary Of Weaknesses:

- The proposed method is not technically different from previous works (which the authors cite) that use knowledge graphs for question answering (QA) tasks. The newly introduced technique—using descriptions for filtering instead of entity names—is interesting, but there is no direct comparison between the two variants.
- The main results of the paper are obtained using commercial LLMs (OpenAI APIs) for both query expansion and dense retrieval, which could raise concerns about reproducibility. While the authors also experimented with LLaMA for query expansion, it would be useful to see how an open-source dense embedding model performs with the proposed expansion methods. The results with BM25 are a good example of this, but the paper could benefit from exploring recent neural dense/sparse models.

- It is unclear from the paper how the relevance labels for retrieval tasks and the KG relations are constructed. If KG relations are harvested from the same original raw datasets used to construct the retrieval test results, there could be label contamination. The authors should clarify this and confirm whether it is the case.

Comments Suggestions And Typos:

- There might be a typo in Equation 6. Should the correct expression be (v_i, r_{ij}, v_j) ?

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.

Overall Assessment: 3.5

Best Paper: No

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**



Kind Reminder of Author-Reviewer Discussion Period

Official Comment

by Authors (Yu Xia (/profile?id=~Yu_Xia9), Tong Yu (/profile?id=~Tong_Yu3), Julian McAuley (/profile?id=~Julian_McAuley1), Haoliang Wang (/profile?id=~Haoliang_Wang1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission2488/Authors))

26 Nov 2024, 14:22 (modified: 02 Jan 2025, 08:15)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer YLJQ, Commitment Readers

Revisions (/revisions?id=6rh5yIiEPA)

Comment:

Dear Reviewer YLJQ,

As the author-reviewer discussion period is ending soon (in about 13 hours), we would like to follow up on whether our responses have fully addressed your concerns. We are also happy to answer any further questions you might have.

Best,

Authors

Add: **Author-Editor Confidential Comment**



Reply to Reviewer

YLJQ

Official Comment

by Authors ([Yu Xia](/profile?id=~Yu_Xia9) (/profile?id=~Yu_Xia9), [Tong Yu](/profile?id=~Tong_Yu3) (/profile?id=~Tong_Yu3), [Julian McAuley](/profile?id=~Julian_McAuley1) (/profile?id=~Julian_McAuley1), [Haoliang Wang](/profile?id=~Haoliang_Wang1) (/profile?id=~Haoliang_Wang1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission2488/Authors))

25 Nov 2024, 13:41 (modified: 02 Jan 2025, 08:15)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer YLJQ, Commitment Readers

Revisions (/revisions?id=rWy0yw5ihN)

Comment:

Thank you for your time and efforts in reviewing our paper. We provide our point-by-point responses as follows:

Response to Weakness 1:

Thank you for your feedback. Previous works on knowledge graph question answering (e.g., [1][2][3]) primarily utilize entity names for relation filtering. Our proposed method differs in that we leverage textual document descriptions as rich entity representations for more accurate and query-focused relation filtering.

We have compared these two approaches in our experiments by:

1. Including QAGNN [1] as a representative knowledge graph question answering baseline (detailed in Line 392).
2. Including an ablated variant of our method, KAR_wo_DRF, which performs entity-based relation filtering as in [1][2] instead of our proposed document-based relation filtering (detailed in Line 385).

We believe the results in Tables 1 and 2 effectively validate the advantages of our proposed method compared to previous entity-based methods.

Response to Weakness 2:

Thank you for your suggestion. As shown in Table 4 (BM25 as retriever), Table 5 (LLaMA as backbone LLM), our proposed method can also perform well with open-source models.

To further support this, we report the results below on human-generated queries using LLM-Embedder [4], a recent open-source embedding model for dense retrieval. The results consistently validate the advantages of our method.

	AMAZON				MAG				PRIME			
	Hit@1	Hit@5	R@20	MRR	Hit@1	Hit@5	R@20	MRR	Hit@1	Hit@5	R@20	MRR
Base	34.57	60.49	31.01	46.33	28.57	40.78	31.11	33.75	25.69	40.37	38.78	32.09
PRF	40.12	61.73	30.34	50.01	27.98	39.29	32.17	33.36	23.31	37.61	37.82	30.51
HyDE	39.91	62.55	31.48	50.21	28.57	39.68	32.03	33.79	24.77	37.92	38.13	30.76
RAR	42.90	62.96	31.79	52.37	30.95	40.77	32.91	35.58	31.01	44.40	45.67	37.21
AGR	42.80	60.91	29.74	51.33	32.14	42.86	34.34	37.66	31.31	43.73	45.75	37.33
KAR (Ours)	45.43	63.95	32.21	54.28	34.76	44.29	35.51	39.19	35.32	48.62	47.34	41.71

Response to Weakness 3:

Thank you for your suggestion. As mentioned in Section 5.1 of our paper and detailed in Section 2.2 of the STaRK benchmark paper [5], the semi-structured knowledge bases are constructed using textual document descriptions and KG relations obtained from different sources with complementary information, rather than being harvested from the same original raw datasets. Thus, we believe the relevance labels are not contaminated. We will include additional dataset information to clarify this.

Response to Comment:

Thank you for your feedback. As discussed in Line 292, unlike traditional entity-based knowledge triples (v_i, r_{ij}, v_j) , we construct document-based knowledge triples (d_i, r_{ij}, d_j) as in Equation 6 where d_i represents the textual document associated with entity v_i , which provide the LLM with both structured document relations and rich textual details.

[1] Yasunaga, Michihiro, et al. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. NAACL 2021.

[2] Zhang, Xikun, et al. GreaseLM: Graph REASONing Enhanced Language Models. ICLR 2022

[3] Taunk, Dhaval, et al. GrapeQA: Graph augmentation and pruning to enhance question-answering. Companion Proceedings of the ACM Web Conference 2023.

[4] Zhang, Peitian, et al. Retrieve anything to augment large language models. arXiv preprint arXiv:2310.07554 (2023).

[5] Wu, Shirley, et al. STaRK: Benchmarking LLM Retrieval on Textual and Relational Knowledge Bases. NeurIPS 2024 Benchmark and Datasets

Add: **Author-Editor Confidential Comment**



Reponse to authors

Official Comment by Reviewer YLJQ 📅 26 Nov 2024, 14:45 (modified: 02 Jan 2025, 08:15)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer YLJQ, Commitment Readers

📄 Revisions (/revisions?id=ohPjVVv43c)

Comment:

Thank you for your response. After reviewing the information provided, I feel that most of my concerns have been addressed. I believe the work is an interesting addition to the community, and as a result, I have decided to increase the soundness score to 4.

Add: **Author-Editor Confidential Comment**



➔ *Replying to Reponse to authors*

Reply to Reviewer YLJQ

Official Comment

by Authors (👁 Yu Xia (/profile?id=~Yu_Xia9), Tong Yu (/profile?id=~Tong_Yu3), Julian McAuley (/profile?id=~Julian_McAuley1), Haoliang Wang (/profile?id=~Haoliang_Wang1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission2488/Authors))

📅 26 Nov 2024, 14:50 (modified: 02 Jan 2025, 08:15)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer YLJQ, Commitment Readers

📄 Revisions (/revisions?id=nTw49o5z60)

Comment:

Thank you for your recognition of our work and your valuable time in reviewing our paper.

Add: **Author-Editor Confidential Comment**

[About OpenReview \(/about\)](#)

[Hosting a Venue \(/group?id=OpenReview.net/Support\)](#)

[All Venues \(/venues\)](#)

[Sponsors \(/sponsors\)](#)

[Frequently Asked Questions](#)

[\(https://docs.openreview.net/getting-started/frequently-asked-questions\)](https://docs.openreview.net/getting-started/frequently-asked-questions)

[Contact \(/contact\)](#)

[Feedback](#)

[Terms of Use \(/legal/terms\)](#)

[Privacy Policy \(/legal/privacy\)](#)

[OpenReview \(/about\)](#) is a long-term project to advance science through improved peer review, with legal nonprofit status through [Code for Science & Society \(https://codeforscience.org/\)](https://codeforscience.org/). We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](#). © 2025 OpenReview