

Recontextualize, Revise and Retrieve

Yoonsang Lee



Seoul National University

Natural Language Processing, 23 Spring

Problem Formulation

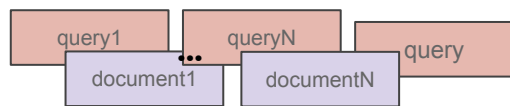


LM-based Retrieval

- Recent studies involve language models in retrieval
 - Applied before training retriever models
 - Query expansion / Data augmentation / Domain Transfer

Query2doc

(Wang et al. 2023)

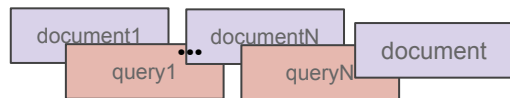


document'

Supervised
(query + document',
document)

Promptagator

(Dai et al. 2022)

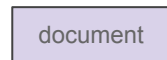


query'

Few-shot
(query', document)

GPL

(Wang et al. 2021)

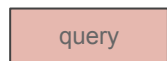


query'

Zero-shot
(query', document)

HyDE

(Gao et al. 2022)



document'

No training

Reliability on generated outputs

- Several papers report the errors in generated outputs of language model

query	who sings monk theme song
LLM generation	The theme song for the television show Monk is entitled " It's a Jungle Out There " and is sung by American singer-songwriter Randy Newman . The song was written specifically for the show, and <i>it has been used as the theme song since the series premiered in 2002</i> . It has been praised by critics and fans alike and is often regarded as one of the best theme songs in television history.
Groundtruth	exists and is an alternate of. The Monk theme song is It's a Jungle Out There by Randy Newman . The Monk theme song is It's a Jungle Out There by Randy Newman .

Type 1 : External error

query	trumbull marriott fax number
LLM generation	The fax number for the Trumbull Marriott Shelton is 203-378-4444 .
Groundtruth	Business name: Trumbull Marriott Merritt Parkway; Address: 180 Hawley Lane Trumbull, Connecticut 06611; Phone number: 203-378-4958; Fax number: 203-378-1400 ; Business hours: 24; Credit cards accepted: Yes; Number of employees: 10-19; Map:

Type 2 : Internal error

- Hyde: The generated document *is not* real, can and is likely to be ungrounded factually. We *only* require it to capture **relevance pattern**.
- Query2doc: Although such errors may appear subtle and difficult to verify, they pose a significant challenge to building **trustworthy systems** using LLMs.



Can't we revise the generated output by **looking the input again**, precisely?

Thorough Examination

- Language Model can capture internal errors and revise

FLAN-T5-xl

Input

T prompt

Revise query if it contains errors based on passage.
passage: Business name: Trumbull Marriott Merritt Parkway; Address: 180 Hawley Lane Trumbull, Connecticut 06611; Phone number: 203-378-4958; Fax number: 203-378-1400; Business hours: 24; Credit cards accepted: Yes; Number of employees: 10-19; Map: query: The fax number for the Trumbull Marriott Shelton is 203-378-4444.

Prompt to send to FLAN-T5.

Output

The fax number for the Trumbull Marriott Merritt Parkway is 203-378-1400.

Generated in 1.70 seconds

Share

Share on Discord

Report

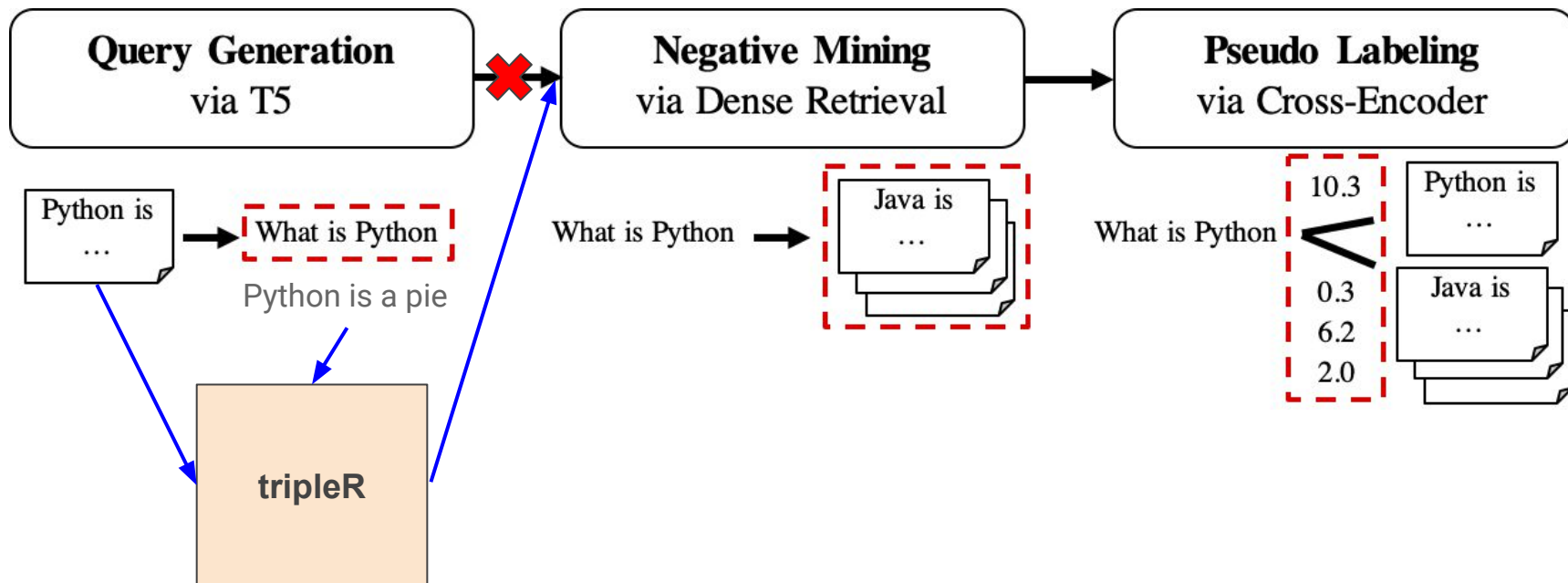
ChatGPT

Revise query if it contains errors based on passage. passage: Business name: Trumbull Marriott Merritt Parkway; Address: 180 Hawley Lane Trumbull, Connecticut 06611; Phone number: 203-378-4958; Fax number: 203-378-1400; Business hours: 24; Credit cards accepted: Yes; Number of employees: 10-19; Map: query: The fax number for the Trumbull Marriott Shelton is 203-378-4444.

Query: The fax number for the Trumbull Marriott Merritt Parkway is 203-378-1400.

Recontextualize, Revise and Retrieve

- To revise the pseudo-document, we need the original document
- GPL as our baseline, since it doesn't involve LLM, which is limited for project



Methods



Training Details

- BeIR : Widely used benchmark containing 18 IR datasets
- We use 3 dataset which have the smallest corpus sizes
 - Arguana : Argument retrieval (argument : counter argument)
 - NFCorpus : Bio-Medical IR (article : query)
 - Scifact : Scientific fact-checking (passage : finding)
- We modify GPL training step from 140K to 70K
- We modify |corpus_size| x qpp from 250K to 100K
- We don't use TSDAE nor TAS-B method
- We report nDCG@10 and MRR@10 as evaluation metric, following Wang et al. 2023
- All codes and results are publicly available (<https://github.com/lilys012/tripleR>)

Proposed Methods (T5)

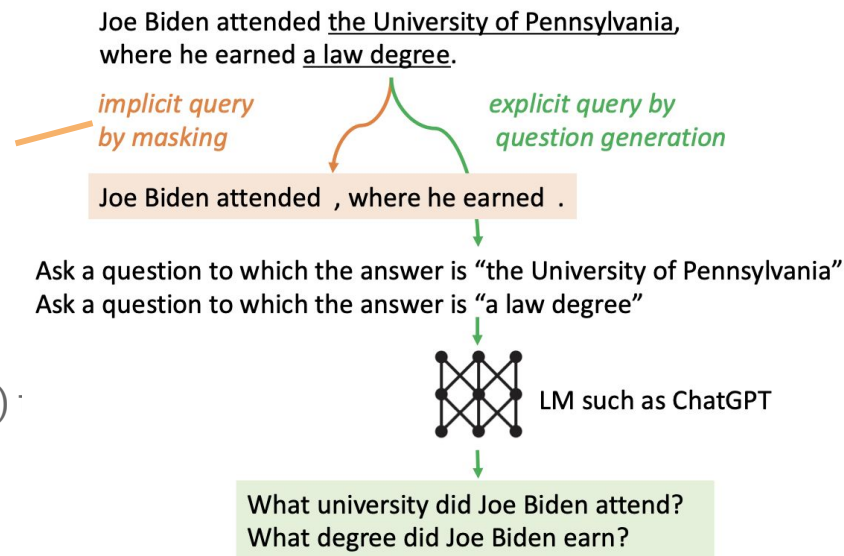
- Query Expansion by Prompting LLMs (Jagerman et al. 2023)
 - **Larger** FLAN-T5 models yield better performance in sparse retrieval; flan-ul2 (20B)
 - v1) Generate with FLAN-T5-xl + task-specific prompt following Dai et al. 2022
- LMCOR (Vernikos et al. 2023)
 - T5-base (250M) can **correct grammatical errors** in outputs of LLM; PaLM (62B)
 - v2) Generate with T5-base, then revise with FLAN-T5-xl

Dataset	Arguana		NFCorpus		Scifact	
Method	nDCG@10	MRR@10	nDCG@10	MRR@10	nDCG@10	MRR@10
GPL	49.6	41.1	32.8	51.8	64.1	60.9
v1	49.7	41.0	32.0	52.1	65.9	63.0
v2	36.0	27.5	-	-	57.2	53.7

- FLAN-T5-xl is not enough to capture context of long passages
- Maybe we need to enlarge our model to 20B~175B, unrealistic 😭

Proposed Methods (BERT)

- FLARE (Jiang et al. 2023)
 - Use **confidence** of output tokens to revise generated texts for text generation
 - v3) mask if confidence < 0.8 and random < 0.4
 - v4) replace random 1 token with [MASK]
 - v5) replace tokens with [MASK]
 - Then feed (pseudo-doc, [SEP], document) · BERT
 - Fill [MASK] to **recontextualize**



Proposed Methods (Ensemble)

- Query2doc (Wang et al. 2023)

- Revise ~~original query by concatenating pseudo document~~

pseudo-documentrecontextualized pseudo-document
- v6) initial pseudo-document + v3 (erase)
- v7) initial pseudo-document + v5 (replace)

Dense Retrieval The new query q^+ is a simple concatenation of the original query q and the pseudo-document d' separated by [SEP]:

$$q^+ = \text{concat}(q, [\text{SEP}], d') \quad (2)$$

- v8) v3 (erase) + v2 (flan-t5-xl revision)

- Only experimented on arguana dataset due to resource issue

Results

- Retrieve results

- Arguana : v8 emphasizes important terms for counter argument
 - why are animals less eco friendly [SEP] less eco friendly
- NFCorpus : specific nouns are critical, so doesn't depend on particular method
 - apobec3g is an innate virus. [SEP] apobec3g innate virus.
- Scifact : transfers msmarco query-like style to finding-like style
 - what mri is used to measure white matter -> diffusion defects in preterm cerebral white matter

Dataset	Arguana		NFCorpus		Scifact	
Method	nDCG@10	MRR@10	nDCG@10	MRR@10	nDCG@10	MRR@10
GPL	47.9	39.4	32.2	51.1	64.1	60.9
v3	48.5	39.9	32.3	51.5	64.5	61.3
v4	49.1	40.5	32.2	52.1	64.2	61.0
v5	48.3	39.7	32.0	51.7	65.2	61.9
v6	51.6	43.2	32.2	51.7	64.2	60.6
v7	50.1	41.5	31.9	51.8	64.6	61.7
v8	53.9	45.2	-	-	-	-

Conclusion

- Our contributions are as follows:
 - We addressed current problems on LM based retrieval
 - We proposed tripleR, novel methods on recontextualizing and revising for dense retrieval
 - We conducted extensive experiments and achieved better performance than the original paper
- Future study
 - Exploit large language models
 - Word-level masking
 - Experiment on other retrieval models or dataset
 - Domain transfer

References

- **Query2doc** (Wang, Liang et al. "Query2doc: Query Expansion with Large Language Models." *ArXivabs/2303.07678* (2023): n. pag.)
- **Promptagator** (Dai, Zhuyun et al. "Promptagator: Few-shot Dense Retrieval From 8 Examples." *ArXivabs/2209.11755* (2022): n. pag.)
- **GPL** (Wang, Kexin et al. "GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval." *ArXiv abs/2112.07577* (2021): n. pag.)
- **HyDE** (Gao, Luyu et al. "Precise Zero-Shot Dense Retrieval without Relevance Labels." *ArXivabs/2212.10496* (2022): n. pag.)
- **Query expansion by Prompting LLMs** (Jagerman, Rolf et al. "Query Expansion by Prompting Large Language Models." *ArXivabs/2305.03653* (2023): n. pag.)
- **LMCor** (Vernikos, Giorgos et al. "Small Language Models Improve Giants by Rewriting Their Outputs." *ArXiv abs/2305.13514* (2023): n. pag.)
- **FLARE** (Jiang, Zhengbao et al. "Active Retrieval Augmented Generation." *ArXiv abs/2305.06983* (2023): n. pag.)
- **BEIR** (Thakur, Nandan et al. "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models." *ArXiv abs/2104.08663* (2021): n. pag.)