LREC-COLING 2024

UC San Diego

IR2: Information Regularization for Information Retrieval

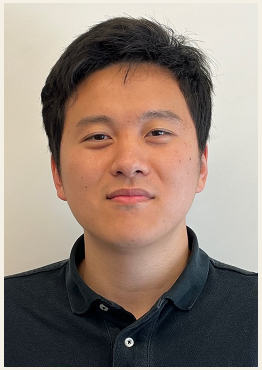
Jianyou Wang ✦ Kaicheng Wang ✦ Xiaoyue Wang ✦ Weili Cao
Ramamohan Paturi ❄ Leon Bergen ❄

Laboratory for Emerging Intelligence | University of California, San Diego | La Jolla, CA, 92093

LREC-COLING 2024| ID: 1078

LEI

Andre Jianyou Wang



Kaicheng Wang

Xiaoyue Wang

Weili Cao

Ramamohan Paturi

Leon Bergen

LEI

UC San Diego

Multifaceted and Complex-Query Datasets

DORIS-MAE[1]

Retrieve scientific abstract for complex research question

60 queries
~110 candidates per query

ArguAna[2]

Retrieve counter-arguments that refute a query argument

1,406 queries
8,674 candidates

WhatsThatBook[3]

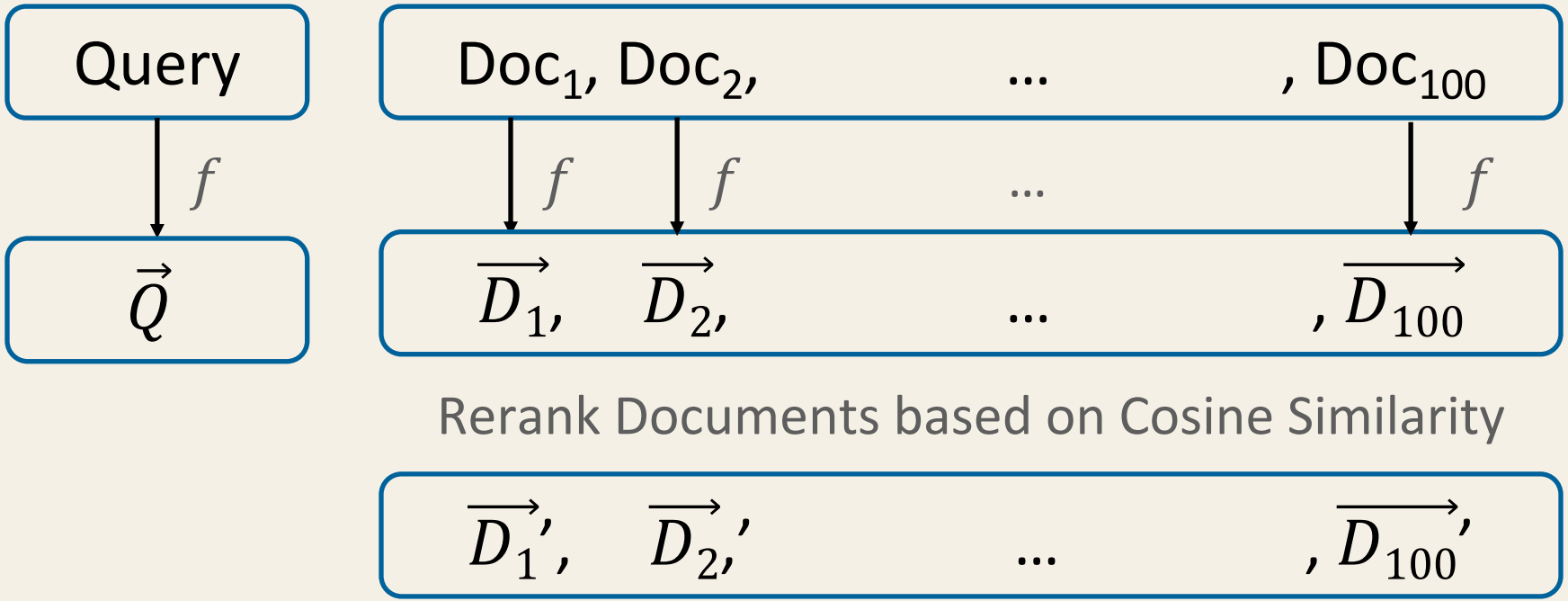
Retrieve book that match a tip-of-the-tongue query

1,445 queries
4,000 candidates

[1] Wang, Jianyou Andre, et al. "Scientific document retrieval using multi-level aspect-based queries." NeurIPS Dataset & Benchmark (2024).
[2] Wachsmuth, et al. "Retrieval of the best counterargument without prior topic knowledge." ACL 2018.
[3] Lin, Kevin, et al. "Decomposing Complex Queries for Tip-of-the-tongue Retrieval." EMNLP 2023.

UC San Diego

Information Retrieval Framework



Fine-tuning Embedding Models

Challenge: not enough queries for training

- Pretrained embedding models need to be fine-tuned to understand these specialized IR tasks.
- Not enough queries can be used for training.
- Synthetic query generation via LLM. (gpt-4-0613)
- But there are issues with the quality of the synthetic queries.

Baseline Synthetic Data Generation

8-shot Generation

- Promptagator [4] sees 8 pairs of example (document, query).
- Using LLM's ability of in-context learning, it learns the transformation from document to query.
- After seeing the 9^{th} document, Promptagator generates the synthetic 9^{th} query.

- D1, Q1, D2, Q2, …, D8, Q8, D9 => Synthetic Q9

[4] Dai, Zhuyun, et al. "Promptagator: Few-shot Dense Retrieval From 8 Examples." ICLR 2022.

Superficial Lexical Overlaps

Synthetic Query from Promptagator[4]:
(Previous state-of-the-art Synthetic Query Generation Method)

As a medical researcher, I am looking for a tool that can help me efficiently explore biomedical literature, especially in the context of the COVID-19 pandemic. I need a tool that can not only retrieve relevant articles but also filter them based on clinically-relevant concepts and their relations. This tool should be able to decrease the proportion of unjudged documents and increase the precision of the search results, ensuring that I am exposed to a larger number of relevant documents. I am also interested in understanding how this concept-based literature exploration compares to traditional text-based retrieval. I would like to see both quantitative and qualitative insights into the characteristics of this approach.

Original Abstract:
(A paper abstract from which Synthetic Queries are generated)

The COVID-19 pandemic has driven ever greater demand for tools which enable efficient exploration of biomedical literature. Although semi-structured information resulting from concept recognition and detection of the defining elements of clinical trials (e.g.PICO criteria) has been commonly used to support literature search, the contributions of this abstraction remain poorly understood, especially in relation to text-based retrieval. In this study, we compare the results retrieved by a standard search engine with those filtered using clinically relevant concepts and their relations. With analysis based on the annotations from the TREC-COVID shared task, we obtain quantitative as well as qualitative insights into characteristics of relational and concept-based literature exploration. Most importantly, we find that the relational concept selection filters the original retrieved collection in a way that decreases the proportion of unjudged documents and increases the precision, which means that the user is likely to be exposed to a larger number of relevant documents.

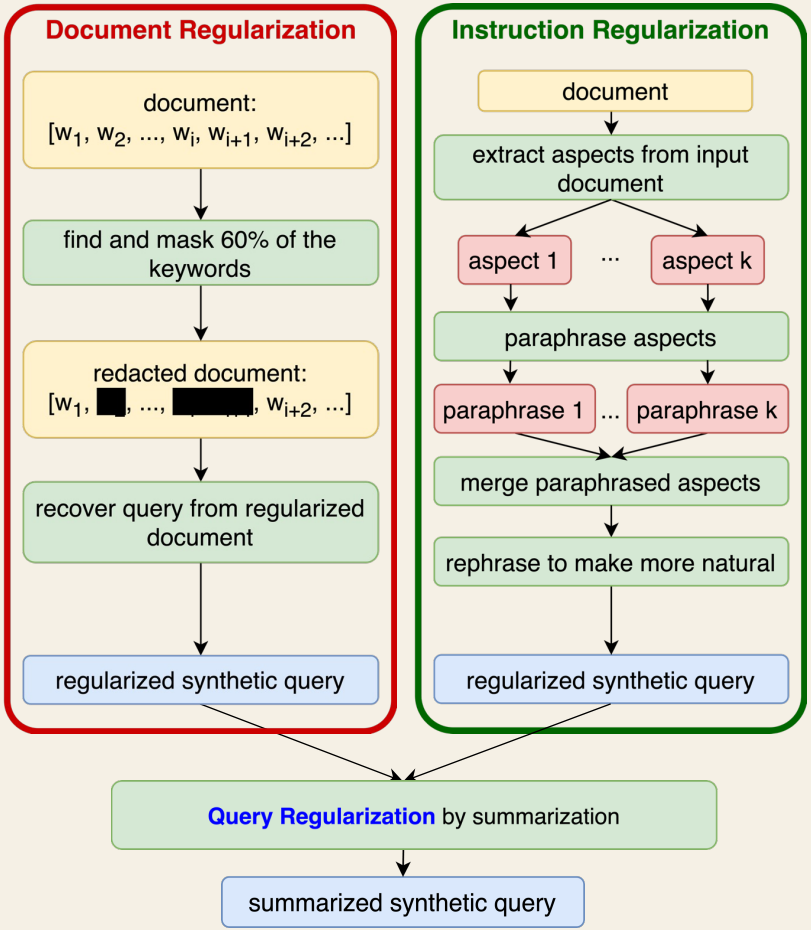
[4] Dai, Zhuyun, et al. "Promptagator: Few-shot Dense Retrieval From 8 Examples." ICLR 2022.

Real Query

I want to introduce an operation planning system for transportation authorities and truck companies. This system will coordinate between different truck companies and transportation authorities … Therefore, currently I am thinking about using data mining approaches to learn patterns in truck routes…Overall, the system should also reduce total fuel consumption to achieve the goal of energy savings.

Real Abstract

Freight transportation …energy consumption and the environment…In this paper, we review how modern information and communication technology supports a cyber-physical transportation system architecture with an integrated logistic system coordinating fleets of trucks traveling together in vehicle platoons. From the reduced air drag, platooning trucks traveling close together can save about 10% of their fuel consumption. … A realistic case study with 200 heavy-duty vehicles performing transportation tasks in Sweden is described…



Information Regularization Methods

We generate the synthetic queries with the following regularization methods:

- $\text{Doc}_{\text{60\% reg}}$
- $\text{Query}_{\text{reg}} \circ \text{Doc}_{\text{60\% reg}}$
- $\text{Instr}_{\text{reg}}$
- $\text{Query}_{\text{reg}} \circ \text{Instr}_{\text{reg}}$

Synthetic Query shares too many similarities with Abstract

Synthetic Query from Promptagator[4]:
(Previous state-of-the-art Synthetic Query Generation Method)

As a medical researcher, I am looking for a tool that can help me efficiently explore biomedical literature, especially in the context of the COVID-19 pandemic. I need a tool that can not only retrieve relevant articles but also filter them based on clinically-relevant concepts and their relations. This tool should be able to decrease the proportion of unjudged documents and increase the precision of the search results, ensuring that I am exposed to a larger number of relevant documents. I am also interested in understanding how this concept-based literature exploration compares to traditional text-based retrieval. I would like to see both quantitative and qualitative insights into the characteristics of this approach.

Original Abstract:
(A paper abstract from which Synthetic Queries are generated)

The COVID-19 pandemic has driven ever greater demand for tools which enable efficient exploration of biomedical literature. Although semi-structured information resulting from concept recognition and detection of the defining elements of clinical trials (e.g.PICO criteria) has been commonly used to support literature search, the contributions of this abstraction remain poorly understood, especially in relation to text-based retrieval. In this study, we compare the results retrieved by a standard search engine with those filtered using clinically relevant concepts and their relations. With analysis based on the annotations from the TREC-COVID shared task, we obtain quantitative as well as qualitative insights into characteristics of relational and concept-based literature exploration. Most importantly, we find that the relational concept selection filters the original retrieved collection in a way that decreases the proportion of unjudged documents and increases the precision, which means that the user is likely to be exposed to a larger number of relevant documents.

Synthetic Query with Regularization:
(Our method)

I am a researcher studying the impact of COVID-19 on various sectors. I am particularly interested in the development of tools that can efficiently manage the challenges brought about by the pandemic. I am considering using a method that involves the detection of defining elements of certain trials, which has been commonly used to support related studies. However, I am unsure of the contributions of this method, especially in relation to specific aspects of the pandemic. I am planning to compare the results retrieved by this method with those filtered using a different approach. I am also interested in understanding the implications of the selection filters used in the original method. I am particularly concerned about the decrease in the number of unique elements and the increase in other factors. Ultimately, I want to determine if the original method is likely to be applicable to a larger number of pandemic-related studies.

[4] Dai, Zhuyun, et al. "Promptagator: Few-shot Dense Retrieval From 8 Examples." ICLR 2022.

Experiment Setup

- Synthetic Data Generation Methods with regularization, and Baseline.
- 3 datasets (DORIS-MAE, ArguAna, WhatsThatBook)
  - 4000 real training documents are used to generate 4000 synthetic queries per dataset per method.
  - No training document is used in any way during testing.
- 4 pre-trained embedding models
  - E5-v2 (0.3B), RoBERTa (0.3B), SimCSE-supervised (0.3B), SPECTER-v2 (0.1B)
- Fine-tuning by contrastive learning using NT-Xent loss.
  - 4 NVIDIA A100 GPU, 40GB each. (Batch size 80)
  - One epoch per configuration, 20 runs with random seeds, average results.
  - All hyperparameters use default value.

DORIS-MAE Experiment Result

	E5-Large-v2[5]	RoBERTa[6]	SimCSE[7]	SPECTER-v2[8]
Method	NDCG@10	NDCG@10	NDCG@10	NDCG@10
Pretrained	71.98	66.86	70.81	71.46
Promptagator	73.95	72.23	73.97	71.55
$Doc_{60\%\ reg}$	74.98 ↑ 1.03	72.94 ↑ 0.71	74.70 ↑ 0.73	72.81 ↑ 1.26
$Query_{reg} \circ Doc_{60\%\ reg}$	74.90 ↑ 0.95	**74.61** ↑ 2.38	74.65 ↑ 0.68	72.19 ↑ 0.64
$Instr_{reg}$	74.42 ↑ 0.47	73.53 ↑ 1.30	**75.60** ↑ 1.63	**73.06** ↑ 1.51
$Query_{reg} \circ Instr_{reg}$	**76.02** ↑ 2.07	74.08 ↑ 1.85	75.43 ↑ 1.46	72.46 ↑ 0.91

Bonferroni adjusted p-value

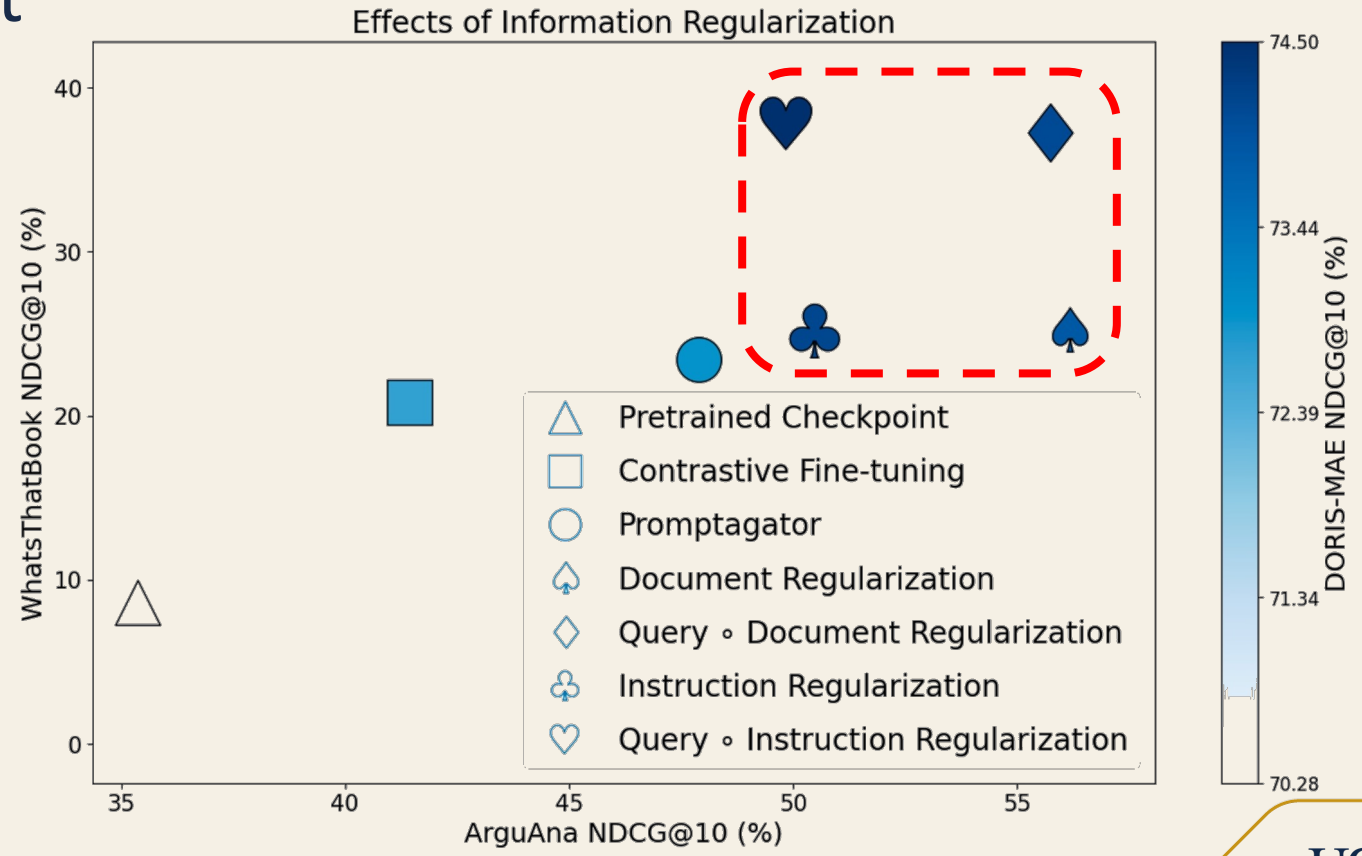
WhatsThatBook Results

	E5-Large-v2[5]	RoBERTa[6]	SimCSE[7]	SPECTER-v2[8]
Method	NDCG@10	NDCG@10	NDCG@10	NDCG@10
Pretrained	15.25	2.17	14.58	2.43
Promptagator	37.07	23.69	27.15	5.72
$\text{Doc}_{60\%\ \text{reg}}$	39.64 ↑ 2.57	25.14 ↑ 1.45	28.87 ↑ 1.72	7.46 ↑ 1.74
$\text{Query}_{\text{reg}} \circ \text{Doc}_{60\%\ \text{reg}}$	**41.58** ↑ 4.51	**30.15** ↑ 6.46	30.77 ↑ 3.62	6.55 ↑ 0.83
$\text{Instr}_{\text{reg}}$	37.33	24.51 ↑ 0.82	29.74 ↑ 2.59	9.30 ↑ 3.58
$\text{Query}_{\text{reg}} \circ \text{Instr}_{\text{reg}}$	39.29 ↑ 2.22	29.28 ↑ 5.59	**33.18** ↑ 6.03	**9.59** ↑ 3.87

ArguAna Results

	E5-Large-v2[5]	RoBERTa[6]	SimCSE[7]	SPECTER-v2[8]
Method	NDCG@10	NDCG@10	NDCG@10	NDCG@10
Pretrained	47.75	23.56	39.23	30.91
Promptagator	55.14	49.80	56.38	30.24
$\text{Doc}_{60\%\ \text{reg}}$	**64.41** ↑ 9.27	**54.85** ↑ 5.05	58.46 ↑ 2.08	**46.55** ↑ 16.31
$\text{Query}_{\text{reg}} \circ \text{Doc}_{60\%\ \text{reg}}$	64.10 ↑ 8.96	54.41 ↑ 4.61	**60.94** ↑ 4.56	42.90 ↑ 12.66
$\text{Instr}_{\text{reg}}$	57.76 ↑ 2.62	51.49 ↑ 1.69	57.41	34.80 ↑ 4.56
$\text{Query}_{\text{reg}} \circ \text{Instr}_{\text{reg}}$	56.89 ↑ 1.75	52.25 ↑ 2.45	58.16 ↑ 1.78	31.72 ↑ 1.48

Result



UC San Diego

Summary

- Information Retrieval (IR) tasks with multifaceted and complex queries.
  - These IR tasks are much harder for embedding models. Not enough training queries.
  - Synthetic data (query) generation methods via LLM.
  - The effectiveness of synthetic data for fine-tuning IR embedding models.
- The concept of information regularization during synthetic data generation.
  - Information regularization reduces superficial lexical correspondence between synthetic query and real document.
  - Propose several regularized generation methods that outperform baseline generation method.

Reference

[1] Wang, Jianyou Andre, et al. "Scientific document retrieval using multi-level aspect-based queries." Advances in Neural Information Processing Systems 36 (2024).

[2] Wachsmuth, Henning, Shahbaz Syed, and Benno Stein. "Retrieval of the best counterargument without prior topic knowledge." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.

[3] Lin, Kevin, et al. "Decomposing Complex Queries for Tip-of-the-tongue Retrieval." Findings of the Association for Computational Linguistics: EMNLP 2023. 2023.

[4] Dai, Zhuyun, et al. "Promptagator: Few-shot Dense Retrieval From 8 Examples." The Eleventh International Conference on Learning Representations. 2022.

[5] Wang, Liang, et al. "Text embeddings by weakly-supervised contrastive pre-training." arXiv preprint arXiv:2212.03533 (2022).

[6] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[7] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple Contrastive Learning of Sentence Embeddings." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

[8] Cohan, Arman, et al. "SPECTER: Document-level Representation Learning using Citation-informed Transformers." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

Thank you!