Recent Advances in Generative Information Retrieval

Yubao Tang[1,2][0009−0003−8010−3404], Ruqing Zhang[1,2][0000−0003−4294−2541], Zhaochun Ren[3][0000−0002−9076−6565], Jiafeng Guo[1,2][0000−0002−9509−8674], and Maarten de Rijke[4][0000−0002−1086−0202]

[1] CAS Key Lab of Network Data Science and Technology, ICT, CAS
[2] University of Chinese Academy of Sciences
{tangyubao21b, zhangruqing,guojiafeng}@ict.ac.cn
[3] Leiden University
z.ren@liacs.leidenuniv.nl
[4] University of Amsterdam
m.derijke@uva.nl

Abstract. Generative retrieval (GR) has become a highly active area of information retrieval that has witnessed significant growth recently. Compared to the traditional "index-retrieve-then-rank" pipeline, the GR paradigm aims to consolidate all information within a corpus into a single model. Typically, a sequence-to-sequence model is trained to directly map a query to its relevant document identifiers (i.e., docids). This tutorial offers an introduction to the core concepts of the novel GR paradigm and a comprehensive overview of recent advances in its foundations and applications. We start by providing preliminary information covering foundational aspects and problem formulations of GR. Then, our focus shifts towards recent progress in docid design, training approaches, inference strategies, and applications of GR. We end by outlining remaining challenges and issuing a call for future GR research. This tutorial is intended to be beneficial to both researchers and industry practitioners interested in developing novel GR solutions or applying them in real-world scenarios.

Keywords: Generative retrieval

1 General information

Information retrieval (IR) is a core task in a wide range of real-world applications, such as web search [21,24] and question answering [9,10]. It aims to retrieve information from a large repository that is relevant to an information need. Most existing IR methods follow a common pipeline paradigm of "index-retrieve-then-rank," which includes (i) building an index for each document in the corpus [14]; (ii) retrieving an initial set of candidate documents for a query [17]; and (iii) determining the relevance degree of each candidate [14]. Despite its wide usage, this paradigm has limitations: (i) during training, heterogeneous

modules with different optimization objectives may lead to sub-optimal performance, and capturing fine-grained relationships between queries and documents is challenging; and (ii) during inference, a large document index is needed to search over the corpus, which may come with substantial memory and computational requirements.

Recently, a fundamentally different paradigm, known as *generative retrieval* (GR) [16], has garnered attention to replace the long-standing pipeline paradigm. The key idea of the GR paradigm is to parameterize the indexing, retrieval, and ranking components of traditional IR systems into a single consolidated model. Specifically, a sequence-to-sequence (Seq2Seq) model is trained to directly map queries to their relevant document identifiers (docids). Such a single-step generative model dramatically simplifies the search process, could be optimized in an end-to-end manner, and could better leverage the capabilities of large language models (LLMs). Based on [19], there are two families of GR, namely *closed-book* GR and *open-book* GR. Closed-book GR refers to the scenario where the language model that is used for directly generating relevant information resources for an information need, and the model is the only source of knowledge leveraged during generation. Open-book GR, on the other hand, allows the language model to draw on external memory prior to, during, and after generation. Here, our main focus is on closed-book GR.

Many publications have emerged in reputable conferences, e.g., SIGIR [3,6], CIKM [4,5,29], KDD [26], NeurIPS [2,25,27,28], ICLR [8], and ACL [7,11,13,23], in Gen-IR@SIGIR2023 [15,20,22,33], in journals [31], and on arXiv [12,18,30,32]. At SIGIR'23, Marc Najork, serving as the keynote speaker, provided a comprehensive summary of existing GR systems and discussed many open challenges in this emerging field [19]. The first workshop on generative information retrieval at SIGIR'23 (Gen-IR@SIGIR2023) [1] welcomed many submissions and attendees, underscoring the IR community's current keen interest in GR.

The time is right to offer a tutorial on the topic of GR. Therefore, we have organized and presented a tutorial dedicated to GR at the 1st International ACM SIGIR Conference on Information Retrieval in the Asia Pacific (SIGIR-AP 2023) on November 26, 2023, in Beijing, China. At ECIR'24 we offer a new edition of the tutorial that has been revised based on the feedback received and incorporates coverage of new relevant work. We hope this tutorial will generate the interest of more researchers and help them gain a better understanding of this novel field.

2 Tutorial information

2.1 Format and length

This is a 3-hour and lecture-style tutorial.

2.2 Tutorial outline

1. **Introduction** (15 minutes)

- An overview of the tutorial
- Why generative retrieval?

2. **Preliminaries** (15 minutes)
   - Retrieval task formulation: generative models vs. discriminative models
   - Basic concepts in generative retrieval
3. **Generative retrieval: Docid design** (30 minutes)
   - Pre-defined static docids
     - Single docids: number-based and word-based docids
     - Multiple docids
   - Learnable docids: jointly with retrieval tasks
4. **Generative retrieval: Training approaches** (40 minutes)
   - Static corpora: supervised learning with labeled data, and pre-training with unlabeled data
   - Dynamic corpora: continual learning
5. **Generative retrieval: Inference strategies** (25 minutes)
   - For a single docid: constrained beam search, constrained greedy search and FM-index
   - For multiple docids: heuristic scoring functions
6. **Generative retrieval: Applications** (35 minutes)
   - Offline application: e.g., entity retrieval, fact checking, recommender systems, multi-hop retrieval and code generation
   - Industry applications
7. **Conclusions and future directions** (20 minutes)

3 Target audience and prerequisites

The tutorial will be accessible to anyone who has a basic knowledge of IR and NLP. The topic will be of interest to both IR and NLP researchers in academia and practitioners in the industry.

4 Presenters

Yubao Tang is a Ph.D. student at the Institute of Computing Technology, Chinese Academy of Sciences. She obtained her M.Sc. degree from the Institute of Information Engineering, Chinese Academy of Sciences, and her B.Eng. from Sichuan University. Her research focuses on information retrieval, and she is the first author of a full paper on generative retrieval at KDD'23 [26].

Ruqing Zhang is an Associate Researcher at the Institute of Computing Technology, Chinese Academy of Sciences. Her recent research focuses on information retrieval, with a particular emphasis on generative information retrieval, the robustness of neural ranking models, and trustworthy retrieval through the lens of causality. She has authored several papers in the field of generative retrieval [3,4,5,6,15,26]. Additionally, Ruqing co-organized the first workshop on generative information retrieval at SIGIR'23 (Gen-IR@SIGIR23) to foster discussions and innovations in GR.

Zhaochun Ren is an Associate Professor at Leiden University. His research interests focus on research problems at the interface of information retrieval and natural language processing, with an emphasis on generative retrieval, recommender systems, and conversational information seeking. He aims to develop intelligent systems that can address complex user requests and solve core challenges in both information retrieval and natural language processing towards that goal. He has been working on various topics related to generative retrieval research. In addition to his academic experience, he worked on e-commerce search and recommendation at JD.com for 2+ years. He has been invited to give tutorials at SIGIR'18 and NLPCC'22.

Jiafeng Guo is a Researcher at the Institute of Computing Technology, Chinese Academy of Sciences (CAS) and a Professor at the University of Chinese Academy of Sciences. He is the director of the CAS key lab of network data science and technology. He has worked on a number of topics related to web search and data mining, with a current focus on neural models for information retrieval and natural language understanding. He has received multiple best paper (runner-up) awards at leading conferences (CIKM'11, SIGIR'12, CIKM'17, WSDM'22). He has been (co)chair for many conferences, e.g., reproducibility track co-chair of SIGIR'23, workshop co-chair of SIGIR'21 and short paper co-chair of SIGIR'20. He serves as an associate editor for ACM Transactions on Information Systems and Information Retrieval Journal. Jiafeng has previously taught tutorials at ACML, CCIR and CIPS ATT.

Maarten de Rijke is a Distinguished University Professor of Artificial Intelligence and Information Retrieval at the University of Amsterdam. His research is focused on designing and evaluating trustworthy technology to connect people to information, particularly search engines, recommender systems, and conversational assistants. He is the scientific director of the Innovation Center for Artificial Intelligence and a former editor-in-chief of ACM Transactions on Information Systems and of Foundations and Trends in Information Retrieval, and a current co-editor-in-chief of Springer's Information Retrieval book series, (associate) editor for various journals and book series. He has been general (co)chair or program (co)chair for CIKM, ECIR, ICTIR, SIGIR, WSDM, WWW, and has previously taught tutorials at these same venues and AAAI.

5 Tutorial materials

We plan to share the following materials on this website:[1] (i) Slides: All slides are made publicly available. (ii) Annotated bibliography: An annotated compilation of references that lists all works discussed in the tutorial and provides a good basis for further study. (iii) Code: An annotated list of pointers to open source code bases and datasets for the work discussed in the tutorial. (iv) Videos: A video recording of the presentation will be made available.

[1] https://ecir2024-generative-ir.github.io

References

1. Bénédict, G., Zhang, R., Metzler, D.: Gen-IR@SIGIR 2023: The first workshop on generative information retrieval. In: SIGIR. pp. 3460–3463 (2023)
2. Bevilacqua, M., Ottaviano, G., Lewis, P., Yih, W.t., Riedel, S., Petroni, F.: Autoregressive search engines: Generating substrings as document identifiers. In: NeurIPS. pp. 31668–31683 (2022)
3. Chen, J., Zhang, R., Guo, J., Fan, Y., Cheng, X.: GERE: Generative evidence retrieval for fact verification. In: SIGIR. pp. 2184–2189 (2022)
4. Chen, J., Zhang, R., Guo, J., Liu, Y., Fan, Y., Cheng, X.: CorpusBrain: Pre-train a generative retrieval model for knowledge-intensive language tasks. In: CIKM. pp. 191–200 (2022)
5. Chen, J., Zhang, R., Guo, J., de Rijke, M., Chen, W., Fan, Y., Cheng, X.: Continual learning for generative retrieval over dynamic corpora. In: CIKM. p. 306–315 (2023)
6. Chen, J., Zhang, R., Guo, J., de Rijke, M., Liu, Y., Fan, Y., Cheng, X.: A unified generative retriever for knowledge-intensive language tasks via prompt learning. In: SIGIR. pp. 1448–1457 (2023)
7. Chen, X., Liu, Y., He, B., Sun, L., Sun, Y.: Understanding differential search index for text retrieval. In: Findings of ACL. p. 10701–10717 (2023)
8. De Cao, N., Izacard, G., Riedel, S., Petroni, F.: Autoregressive entity retrieval. In: ICLR (2021)
9. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: ICML. pp. 3929–3938 (2020)
10. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.W., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics **7**, 452–466 (2019)
11. Lee, H., Kim, J., Chang, H., Oh, H., Yang, S., Karpukhin, V., Lu, Y., Seo, M.: Nonparametric decoding for generative retrieval. In: Findings of the ACL 2023. pp. 12642–12661 (2023)
12. Li, Y., Yang, N., Wang, L., Wei, F., Li, W.: Learning to rank in generative retrieval. arXiv preprint arXiv:2306.15222 (2023)
13. Li, Y., Yang, N., Wang, L., Wei, F., Li, W.: Multiview identifiers enhanced generative retrieval. In: ACL. pp. 6636–6648 (2023)
14. Liu, S., Xiao, F., Ou, W., Si, L.: Cascade ranking for operational e-commerce search. In: KDD. pp. 1557–1565 (2017)
15. Liu, Y.A., Zhang, R., Guo, J., Chen, W., Cheng, X.: On the robustness of generative retrieval models: An out-of-distribution perspective. In: Gen-IR@SIGIR (2023)
16. Metzler, D., Tay, Y., Bahri, D., Najork, M.: Rethinking search: Making domain experts out of dilettantes. SIGIR Forum **55**(1), 1–27 (2021)
17. Mitra, B., Nalisnick, E., Craswell, N., Caruana, R.: A dual embedding space model for document ranking. arXiv preprint arXiv:1602.01137 (2016)
18. Nadeem, U., Ziems, N., Wu, S.: CodeDSI: Differentiable code search. arXiv preprint arXiv:2210.00328 (2022)
19. Najork, M.: Generative information retrieval. In: SIGIR. pp. 1–1 (2023)
20. Nguyen, T., Yates, A.: Generative retrieval as dense retrieval. In: Gen-IR@SIGIR (2023)
21. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset.

In: Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (2016)
22. Pradeep, R., Hui, K., Gupta, J., Lelkes, A.D., Zhuang, H., Lin, J., Metzler, D., Tran, V.Q.: How does generative retrieval scale to millions of passages? In: Gen-IR@SIGIR (2023)
23. Ren, R., Zhao, W.X., Liu, J., Wu, H., Wen, J.R., Wang, H.: TOME: A two-stage approach for model-based retrieval. In: ACL. pp. 6102–6114 (2023)
24. Rose, D.E., Levinson, D.: Understanding user goals in web search. In: WWW. pp. 13–19 (2004)
25. Sun, W., Yan, L., Chen, Z., Wang, S., Zhu, H., Ren, P., Chen, Z., Yin, D., de Rijke, M., Ren, Z.: Learning to tokenize for generative retrieval. In: NeurIPS (2023)
26. Tang, Y., Zhang, R., Guo, J., Chen, J., Zhu, Z., Wang, S., Yin, D., Cheng, X.: Semantic-enhanced differentiable search index inspired by learning strategies. In: KDD. pp. 4904–4913 (2023)
27. Tay, Y., Tran, V.Q., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., Schuster, T., Cohen, W.W., Metzler, D.: Transformer memory as a differentiable search index. In: NeurIPS. vol. 35, pp. 21831–21843 (2022)
28. Wang, Y., Hou, Y., Wang, H., Miao, Z., Wu, S., Sun, H., Chen, Q., Xia, Y., Chi, C., Zhao, G., Liu, Z., Xie, X., Sun, H., Deng, W., Zhang, Q., Yang, M.: A neural corpus indexer for document retrieval. In: NeurIPS. vol. 35, pp. 25600–25614 (2022)
29. Wang, Z., Zhou, Y., Tu, Y., Dou, Z.: NOVO: Learnable and interpretable document identifiers for model-based ir. In: CIKM. pp. 2656–2665 (2023)
30. Zhang, P., Liu, Z., Zhou, Y., Dou, Z., Cao, Z.: Term-sets can be strong document identifiers for auto-regressive search engines. arXiv preprint arXiv:2305.13859 (2023)
31. Zhou, Y.J., Yao, J., Dou, Z.C., Wu, L., Wen, J.R.: DynamicRetriever: A pre-trained model-based ir system without an explicit index. Machine Intelligence Research **20**(2), 276–288 (2023)
32. Zhou, Y., Yao, J., Dou, Z., Wu, L., Zhang, P., Wen, J.R.: Ultron: An ultimate retriever on corpus with a model-based indexer. arXiv preprint arXiv:2208.09257 (2022)
33. Zhuang, S., Ren, H., Shou, L., Pei, J., Gong, M., Zuccon, G., Jiang, D.: Bridging the gap between indexing and retrieval for differentiable search index with query generation. In: Gen-IR@SIGIR (2023)