Structured representation for Information Retrieval

Yuxuan Zong[1,*], Benjamin Piwowarski[1]

[1]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France

Abstract

Generative information retrieval represents documents as a sequence of identifier tokens. Some of the works propose to use arbitrary identifiers while others propose to use meta-data (text, URL, title). Both approaches have limits: the former exhibit generalization problems while the latter might be limited by the correctness of the meta-data. In this work, we propose a new generative approach, named REFERENTIAL, that combines these two methods. Namely, we use prefix-biased identifiers but do not require a one-to-one relationship between an identifier and a document. In this paper, we briefly expose the model and the conducted experiments.

Keywords

information retrieval, hierarchical structure, generative retrieval,

1. Background and introduction

In 2022, Tay et al. [1] introduce the idea of generative retrieval-based methods where instead of representing a document as an embedding and building indices for retrieval, the model *directly predicts the identifier of relevant documents*. Said otherwise, the information about documents is stored directly into the model parameters. Within this research area, we can distinguish two different directions. Some of the works [1, 2, 3, 4, 5, 6] propose to use arbitrary identifiers while others [1, 7, 8, 9, 10, 11, 12, 13] propose to use meta-data (text, URL, title) as identifiers. On the one hand, for arbitrary identifiers, the most promising works [1, 2, 3, 4, 5] rely on identifiers that exhibit some structure: namely, the sequences that share the same prefix are closer than others that do not, e.g. documents identified by the sequences (1,5) and (1,6) have more in common than a document identified by the sequence (2,5). In these works, there is a one-to-one mapping between identifiers and documents that allows using the model to perform retrieval by generating the document identifiers directly. However, as these works try to maximizing the probability that a document generates the corresponding identifier, generalizing to new documents is difficult. The alternative, relying on meta-data, uses more interpretable document identifiers but the lack of one-to-one mapping (in the case of titles) and clear semantics results in heuristics to define a document's score concerning a query, as well as potential generalization problems (although less than for arbitrary IDs).

To overcome these limitations, we propose a new generative approach, named REFERENTIAL, that combines these two lines of work: using prefix-biased identifiers and removing the one-to-one relationship between an identifier and a document. For each document, we model it as a

CORIA 2024: COnférence en Recherche d'Information et Applications, 3 et 4 avril 2024, La Rochelle, France

*Corresponding author.

yuxuan.zong@isir.upmc.fr (Y. Zong); benjamin.piwowarski@isir.upmc.fr (B. Piwowarski)

0009-0002-0376-1369 (Y. Zong); 0000-0001-6792-3262 (B. Piwowarski)

combination of several identifiers. To allow efficient retrieval, we introduce a structure within the identifiers. However, different from the work [10] which uses document meta-data, the arbitrary token sequences are not easy to train. We have explored several ways to accelerate and stabilize the training procedure including dynamic hard negatives, and progressive training, which we describe next.

2. REFERENTIAL IR model

In this work, we use an encoder-decoder to generate sequences of arbitrary identifiers given the input texts. To enforce a structure on identifier sequences, we propose to model the relevance information given the generated identifier sequence of the query (q), positive document (α), negative document (β) by "α is better than β given the q," denoted as $\alpha \succ_q \beta$, by supposing that the common prefix between the relevant document and a query should be longer than the one between the non-relevant document and the query. For instance, assuming we have $\alpha = (5, 6, 7)$, $\beta = (5, 3, 1)$ and $q = (5, 6, 4)$: we don't have $\alpha = q$, but they share a prefix of length 2. While for β and q, the common prefix is of length 1. Different from the previous works [1, 2, 7, 8, 9, 14, 12, 13], which has a clear one-to-one mapping of a sequence to a unique document, our representation associate a document/query with a sparse probability distribution over "identifier" sequences.

For the learning procedure, most of the previous works propose to maximize the likelihood of generating the document identifier given the document/query text, which makes learning closer to a generative task than to an IR task. We instead propose to maximize the probability that the relevant document α is more relevant than the non-relevant one β for a given query q, i.e. we maximize $P(\alpha \succ_q \beta)$.

To compute its gradient, we use a recursive process based on a score function optimization. For each generation step, we sample the tokens based on the conditional probability distributions from α, q, β respectively: If α and q share the same tokens but not the β, we can stop the recursion by ensuring the $\alpha \succ_q \beta$ based on the prefix. If α, β, q all share the same prefix, we cannot conclude anything so we have to continue the recursion. In all the other cases, we can also stop the recursion by concluding that $\alpha \succ_q \beta$ is false given the prefix.

During inference, we cannot compute $P(\alpha \succ_q \beta)$ for every pair of documents. Instead, we propose to use the following score function:

$$rsv(q, d) = P(\alpha \succ_q D)$$

where D is a document whose distribution over sequences is uniform. This score can be interpreted as "the probability of the document is better than a random document based on the query to be evaluated." Further, different from the loss where we sample one token at each depth to make the training more efficient, we propose to use a beam search during the inference stage to cover more sequences according to the probability distribution of the query.

3. Training REFERENTIAL

To make our model work better, we have tried various strategies from the literature:

Query augmentation : According to [15], inside MSMARCO Passage, the annotated query-document pairs only cover 10% of the total documents, which means that if we use only the manually annotated relevance information, the model can easily overfit. Following [7, 8], we propose to use Doc2Query [16] to generate several queries for each document and consider them to be relevant. We use these synthetic queries in a pre-training stage and then switch to a fine-tuning stage by using the human-annotated dataset.

Dynamic negative sampling The negative sampling strategy is quite important when training IR models, especially for generative ones [17, 3, 18]. If the non-relevant documents (negatives) are too easy to distinguish from relevant ones, our model can distinguish using only a short prefix, and we cannot learn long and meaningful identifier sequences. To sample harder negatives, during training, we can cache which sequence maps to which documents. This information can, in turn, be used by sampling negative documents whose identifier sequences match the most probable ones for the query or the relevant document.

KL-Divergence regularization and bias logit initialization In our model, documents can be associated with sequences of different lengths. However, shorter sequences have a higher probability of being generated. To cope with that, we need to ensure that the probability of two sequences is the same on average (at initialization) by introducing a factor that downscales the probability of short sequences. During training, we use a KL loss to ensure the model does not converge towards generating short sequences.

Progressive training As the model is initialized random and the tokens are arbitrary, quite often when training, the loss drops into a local minimum. We propose a progressive training strategy by progressively increasing the limit on the length of generated sequences.

4. Conclusion and Future work

While the above learning strategies did improve the training process, learning properly REFERENTIAL is still problematic and leads to sub-optimal performance. We are currently considering the following directions, which may make our model perform better.

Smarter initialization : With a random initialization, we observed that the model could still be stuck in a local minimum. We hypothesize that this might be due to an improper pre-conditioning. To cope with this issue, we propose to follow the work of [1, 2] and leverage a hierarchical k-means clustering of documents (based on some embeddings). This clustering can be used to associate to each document a unique sequence. We can then warm up the model by training it in a teacher-forcing mode, i.e., by maximizing the probability that a document generates the sequence associated with its cluster.

Depthwise embedding table : Identifier sequences are quite different from sequences in natural language. More precisely, the semantics of a token depend much more on the prefix that for natural language, especially since the number of tokens is much lower (4 to 32 in our experiments). Approaches like [3, 4] use different codebooks (embedding

matrices) for the decoder at different depths, but they do not depend on the prefix; we could rather use a different embedding matrix for each possible prefix.

References

[1] Y. Tay, V. Tran, M. Dehghani, J. Ni, D. Bahri, H. Mehta, Z. Qin, K. Hui, Z. Zhao, J. Gupta, T. Schuster, W. W. Cohen, D. Metzler, Transformer memory as a differentiable search index, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 21831–21843. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/892840a6123b5ec99ebaab8be1530fba-Paper-Conference.pdf.

[2] Y. Wang, Y. Hou, H. Wang, Z. Miao, S. Wu, Q. Chen, Y. Xia, C. Chi, G. Zhao, Z. Liu, et al., A neural corpus indexer for document retrieval, Advances in Neural Information Processing Systems 35 (2022) 25600–25614.

[3] H. Zeng, C. Luo, B. Jin, S. M. Sarwar, T. Wei, H. Zamani, Scalable and effective generative information retrieval (2023). URL: http://arxiv.org/abs/2311.09134, arXiv:2311.09134 [cs].

[4] W. Sun, L. Yan, Z. Chen, S. Wang, H. Zhu, P. Ren, Z. Chen, D. Yin, M. Rijke, Z. Ren, Learning to tokenize for generative retrieval, in: A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 46345–46361. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/91228b942a4528cdae031c1b68b127e8-Paper-Conference.pdf.

[5] T. Yang, M. Song, Z. Zhang, H. Huang, W. Deng, F. Sun, Q. Zhang, Auto search indexer for end-to-end document retrieval, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 6955–6970. URL: https://aclanthology.org/2023.findings-emnlp.464. doi:10.18653/v1/2023.findings-emnlp.464.

[6] Y. Zhou, J. Yao, Z. Dou, L. Wu, J.-R. Wen, Dynamicretriever: A pre-training model-based ir system with neither sparse nor dense index (2022). URL: http://arxiv.org/abs/2203.00537, arXiv:2203.00537 [cs].

[7] S. Zhuang, H. Ren, L. Shou, J. Pei, M. Gong, G. Zuccon, D. Jiang, Bridging the gap between indexing and retrieval for differentiable search index with query generation (2022). URL: http://arxiv.org/abs/2206.10128, arXiv:2206.10128 [cs].

[8] X. Chen, Y. Liu, B. He, L. Sun, Y. Sun, Understanding differential search index for text retrieval, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10701–10717. URL: https://aclanthology.org/2023.findings-acl.681. doi:10.18653/v1/2023.findings-acl.681.

[9] Y. Li, N. Yang, L. Wang, F. Wei, W. Li, Multiview identifiers enhanced generative retrieval, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 6636–6648. URL: https://aclanthology.org/2023.acl-long.366. doi:10.18653/v1/2023.acl-long.366.

[10] Z. Wang, Y. Zhou, Y. Tu, Z. Dou, Novo: Learnable and interpretable document identifiers for

model-based ir, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, ACM, Birmingham United Kingdom, 2023, p. 2656–2665. URL: https://dl.acm.org/doi/10.1145/3583780.3614993. doi:10.1145/3583780.3614993.

[11] Y. Tang, R. Zhang, J. Guo, J. Chen, Z. Zhu, S. Wang, D. Yin, X. Cheng, Semantic-enhanced differentiable search index inspired by learning strategies, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 4904–4913. URL: https://doi.org/10.1145/3580305.3599903. doi:10.1145/3580305.3599903.

[12] R. Ren, W. X. Zhao, J. Liu, H. Wu, J.-R. Wen, H. Wang, TOME: A two-stage approach for model-based retrieval, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 6102–6114. URL: https://aclanthology.org/2023.acl-long.336. doi:10.18653/v1/2023.acl-long.336.

[13] Y. Zhou, J. Yao, Z. Dou, L. Wu, P. Zhang, J.-R. Wen, Ultron: An ultimate retriever on corpus with a model-based indexer (2022). URL: http://arxiv.org/abs/2208.09257, arXiv:2208.09257 [cs].

[14] Y. Li, N. Yang, L. Wang, F. Wei, W. Li, Learning to rank in generative retrieval, in: AAAI 2024, 2023. URL: https://www.microsoft.com/en-us/research/publication/learning-to-rank-in-generative-retrieval/.

[15] R. Pradeep, K. Hui, J. Gupta, A. Lelkes, H. Zhuang, J. Lin, D. Metzler, V. Tran, How does generative retrieval scale to millions of passages?, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 1305–1321. URL: https://aclanthology.org/2023.emnlp-main.83. doi:10.18653/v1/2023.emnlp-main.83.

[16] R. Nogueira, W. Yang, J. Lin, K. Cho, Document expansion by query prediction (2019). URL: http://arxiv.org/abs/1904.08375, arXiv:1904.08375 [cs].

[17] S. Lee, M. Choi, J. Lee, GLEN: Generative retrieval via lexical index learning, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 7693–7704. URL: https://aclanthology.org/2023.emnlp-main.477. doi:10.18653/v1/2023.emnlp-main.477.

[18] H. Lee, J. Kim, H. Chang, H. Oh, S. Yang, V. Karpukhin, Y. Lu, M. Seo, Nonparametric decoding for generative retrieval (2023). URL: http://arxiv.org/abs/2210.02068, arXiv:2210.02068 [cs].