Biomedical Information Retrieval using large scale PubMed References

Study Project Exposé

Duy Le Thanh

November 13, 2023

1 Introduction

Information Retrieval (IR) is the process of retrieving relevant information or documents from a collection of data based on a user's query. It consists of processing the given queries and storing, representing, ranking and finally retrieving the relevant data[Ibrihich et al., 2022]. Domain-specific applications, such as in the biomedical domain, cover a range of tasks, including literature search [Lu, 2011], question answering [Jin et al., 2022], and the recommendation of citations [Jin et al., 2023], related articles [Lin and Wilbur, 2007] and related sentences [Allot et al., 2019]. Information retrieval systems are of particular interest for the biomedical field, due to various reasons. For instance, they play a crucial role in efficiently accessing the vast biomedical literature in databases like PubMed®[1], ensuring that healthcare professionals and researchers can keep up with the rapidly evolving field[Nadkarni, 2002]. PubMed contains more than 36M citations and abstracts from biomedical literature. For approximately 8M of the abstracts and citations, their full-text articles are accessible via PubMed Central®[2] (PMC). From 2021 to 2022 alone, PMC increased by over 1M articles, demonstrating the substantial growth of accessible biomedical literature. Moreover, IR systems should help to alleviate the challenges associated with specific medical vocabulary and synonyms, helping users navigate the complex and heterogeneous terminology used in biomedical research[Sankhavara and Majumder, 2017].

[1] https://pubmed.ncbi.nlm.nih.gov/about/

[2] https://www.ncbi.nlm.nih.gov/pmc/about/intro/

To handle these challenges, a powerful method seems necessary. Previous retrieval models, such as BM25 [Robertson and Zaragoza, 2009], solely capture the lexical features of queries and documents. State-of-the-art systems incorporate transformers[Vaswani et al., 2017] to acquire and use semantic meanings of queries and documents when solving IR tasks[Ni et al., 2021, Neelakantan et al., 2022, Jin et al., 2023].

2 Goals of the Study Project

The main goals of this study project are to train a retriever for citation recommendation using the retriever-part of the MedCPT[Jin et al., 2023] framework, training it on a self-generated dataset from PubMed Central full-text references and to evaluate it on the BEIR[Thakur et al., 2021] dataset for comparison with the original MedCPT model.

3 Background and Related Work

Lexical (Sparse) Retrievers. Sparse retrievers use lexical characteristics of documents to compute relevance scores between queries and documents. An early approach considered term frequencies (TF) in a single document and inverse document frequencies (IDF) in the corpus to determine suitable documents. In this model, terms are weighted higher if they occur frequently in a document and rarely in the corpus[Salton et al., 1975]. Best Matching 25 (BM25) is an extension of the TF-IDF model that further takes into account the saturation of a term in a document and the length of the document[Robertson and Zaragoza, 2009].

Dense Retrievers. Dense retrievers use neural networks to encode and match queries and documents in low-dimensional semantic space, which have been shown to outperform sparse retrievers like BM25 in natural language processing (NLP) tasks, such as question answering[Karpukhin et al., 2020b] and citation recommendation[Nogueira and Cho, 2019, Khattab and Zaharia, 2020, Lin et al., 2020, Jin et al., 2023].

In our work, we follow the *bioMedical Contrastive Pre-trained Transformers* (MedCPT) framework of Jin et al. [2023]. In this approach, a *retriever* efficiently retrieves thousands of candidates from millions of documents and a *re-ranker* further refines the relevance of the candidates. They use a 255M query-article pairs data set generated from PubMed click logs for training. The retriever consists of two 12-layer Transformers (Trm)[Vaswani et al., 2017]: a query encoder $QEnc$ and a document encoder $DEnc$, which are initialized with PubMedBERT[Gu et al., 2020]. The relevance of a query q and a document d is modeled by the dot product of their $[CLS]$ encoder embeddings $E(q) \in \mathbb{R}^h$ and $E(d) \in \mathbb{R}^h$ where $h = 768$. The re-ranker is a 12-layer transformer cross-encoder that is also initialized with PubMedBERT. For this part, the relevance of queries and documents is calculated by passing them into the same cross-encoder. Jin et al. achieved state-of-the-art performances for query-article relevance on the BEIR[Thakur et al., 2021] benchmark dataset, article similarity task on the RELISH[Brown et al., 2019] dataset and sentence similarity task on the BIOSSES[Soganciogl u et al., 2017] without any task-specific training or fine-tuning.

4 Approach

Our model. We use the implementation of the MedCPT retriever[3] as the starting point and follow the framework of Jin et al. [2023]. To train our retriever, we extract query-article pairs from PMC and use the same parameter configuration as described in the original paper. Finally, we evaluate our model on the BEIR data set and compare it against the model of Jin et al. [2023] and their competitors in their original paper. The adapted workflow for the retriever-only framework of Jin et al. [2023] is shown in Figure 1.

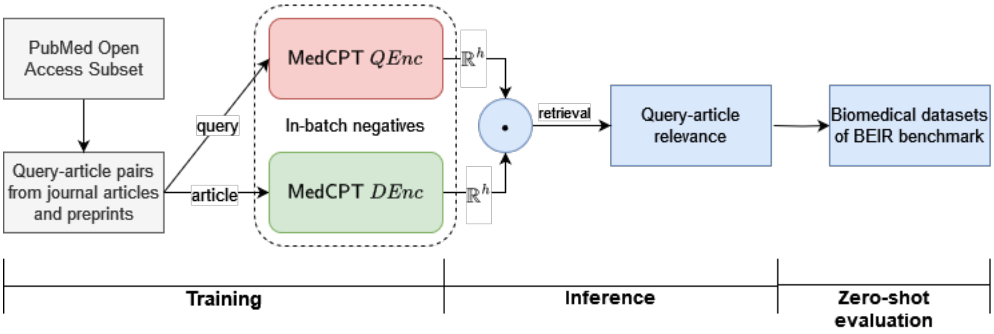

Figure 1: A high-level overview of our retriever-only model. Adapted from Jin et al. [2023].

[3] https://github.com/ncbi/MedCPT/tree/main/retriever

Motivation. We argue that MedCPT's success is largely due to the fact that they have extracted training data from the click logs of knowledgeable PubMed users, thus ensuring high quality of the data. We plan on extracting high quality query-article pairs from existing journal articles and preprints. Sentences within the articles can be viewed as potential queries, with the accompanying citations serving as recommended articles.

PMC Open Access Subset. The PMC Open Access Subset includes more than 3M journal articles and preprints from PubMed Central®. Documents from that subset are made available under Creative Commons or similar licenses to allow a more liberal use and we download them in XML-format using their FTP download service[4].

Extraction of Query-Article Pairs from PMC. For each full-text article in the PMC Open Access Subset, we filter out sentences that contain at least one citation. For each citation in a sentence, we interpret the sentence leading up to that citation as the corresponding query to generate our query-article pairs. E.g., the sentence in Figure 2 results in the query-article pair ("The majority [...] drug targets", "McFadden and Roos 1999"). It is also possible for query-article pairs to share the same query if the citations are in the same group, as can be seen in Figure 3. Here, we generate the two query-article pairs:

> The majority [...] excellent drug targets (McFadden and Roos 1999).

Figure 2: Sentence with one citation extracted from Bozdech et al. [2003].

> Periodicity in [...] human cells (Spellman et al. 1998; Whitfield et al. 2002).

Figure 3: Sentence with two citations extracted from Bozdech et al. [2003].

1. ("Periodicity in [...] human cells", "Spellman et al. 1998") and
2. ("Periodicity in [...] human cells", "Whitfield et al. 2002").

Other citation variants exist, but can be reduced to the above cases. Analogous to Jin et al. [2023], we omit pairs containing articles that do not have a title or abstract in PubMed. We note that we only use the full-texts from PMC to extract the queries. When training the retriever, we only use the title and abstract to compute the corresponding document embeddings.

[4] `https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/#bulk`

References

Olalere A. Abass and Oluremi A. Arowolo. Information retrieval models, techniques and applications. *International Research Journal of Advanced Engineering and Science*, 2:197–202, 2017. ISSN 2455-9024. URL `http://irjaes.com/wp-content/uploads/2020/10/IRJAES-V2N2P214Y17.pdf`.

Alexis Allot, Qingyu Chen, Sun Kim, Roberto Vera Alvarez, Donald C Comeau, W John Wilbur, and Zhiyong Lu. LitSense: making sense of biomedical literature at sentence level. *Nucleic Acids Research*, 47(W1): W594–W599, 04 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz289. URL `https://doi.org/10.1093/nar/gkz289`.

Vera Boteva, Demian Gholipour Ghalandari, Artem Sokolov, and Stefan Riezler. A full-text learning to rank dataset for medical information retrieval. volume 9626, pages 716–722, 03 2016. ISBN 978-3-319-30670-4. doi: 10.1007/978-3-319-30671-1_58.

Zbynek Bozdech, Manuel Llinás, Brian Pulliam, Edith Wong, Jingchun Zhu, and Joseph DeRisi. The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS biology*, 1:E5, 11 2003. doi: 10.1371/journal.pbio.0000005.

Peter Brown, RELISH Consortium, and Yaoqi Zhou. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database*, 2019:baz085, 10 2019. ISSN 1758-0463. doi: 10.1093/database/baz085. URL `https://doi.org/10.1093/database/baz085`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL `https://arxiv.org/abs/2005.14165`.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, pages 2270–2282, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.207. URL https://aclanthology.org/2020.acl-main.207.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Luyu Gao, Zhuyun Dai, and Jamie Callan. Rethink training of BERT rerankers in multi-stage retrieval pipeline. *CoRR*, abs/2101.08751, 2021. URL https://arxiv.org/abs/2101.08751.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779, 2020. URL https://arxiv.org/abs/2007.15779.

S. Ibrihich, A. Oussous, O. Ibrihich, and M. Esghir. A review on recent research in information retrieval. *Procedia Computer Science*, 201:777–782, 2022. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2022.03.106. URL https://www.sciencedirect.com/science/article/pii/S1877050922005191. The 13th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 5th International Conference on Emerging Data and Industry 4.0 (EDI40).

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. Biomedical question answering: A survey of approaches and challenges. *ACM Comput. Surv.*, 55(2), jan 2022. ISSN 0360-0300. doi: 10.1145/3490238. URL https://doi.org/10.1145/3490238.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval, 2023.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Jimmy Lin and W. Wilbur. Pubmed related articles: A probabilistic topic-based model for content similarity. *BMC bioinformatics*, 8:423, 02 2007. doi: 10.1186/1471-2105-8-423.

Jimmy Lin, Rodrigo Frassetto Nogueira, and Andrew Yates. Pretrained transformers for text ranking: BERT and beyond. *CoRR*, abs/2010.06467, 2020. URL https://arxiv.org/abs/2010.06467.

Zhiyong Lu. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036, 01 2011. ISSN 1758-0463. doi: 10.1093/database/baq036. URL https://doi.org/10.1093/database/baq036.

P Nadkarni. An introduction to information retrieval: Applications in genomics. *The pharmacogenomics journal*, 2:96–102, 02 2002. doi: 10.1038/sj.tpj.6500084.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris

Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training. *CoRR*, abs/2201.10005, 2022. URL https://arxiv.org/abs/2201.10005.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers. *CoRR*, abs/2112.07899, 2021. URL https://arxiv.org/abs/2112.07899.

Rodrigo Frassetto Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *CoRR*, abs/1901.04085, 2019. URL http://arxiv.org/abs/1901.04085.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL http://arxiv.org/abs/1910.10683.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389, 01 2009. doi: 10.1561/1500000019.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, nov 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL https://doi.org/10.1145/361219.361220.

Jainisha Sankhavara and Prasenjit Majumder. Biomedical information retrieval. In *Fire*, 2017. URL https://api.semanticscholar.org/CorpusID:3768012.

Gizem Soganciogl u, Hakime Öztürk, and Arzucan Ozgur. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33:i49–i58, 07 2017. doi: 10.1093/bioinformatics/btx238.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663, 2021. URL https://arxiv.org/abs/2104.08663.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, and et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 2015. doi: 10.1186/s12859-015-0564-6. URL https://doi.org/10.1186/s12859-015-0564-6.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. Trec-covid: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1), feb 2021. ISSN 0163-5840. doi: 10.1145/3451964.3451965. URL https://doi.org/10.1145/3451964.3451965.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: the covid-19 open research dataset. *CoRR*, abs/2004.10706, 2020. URL https://arxiv.org/abs/2004.10706.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL `http://arxiv.org/abs/1609.08144`.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *CoRR*, abs/1707.00600, 2017. URL `http://arxiv.org/abs/1707.00600`.