BioASQ 11B: Integrating Domain Specific Vocabulary to BERT-based Model for Biomedical Document Ranking.

Notebook for the BioASQ Lab at CLEF 2023

Maël Lesavourey[1], Gilles Hubert[1]

[1]*IRIT lab, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9*

Abstract

In this paper we present the results obtained during BioASQ Task 11B Phase A on document ranking. We proposed a two-stage retrieval approach composed of a retriever and a reranker. The former is based on BM25 and developed with Pyserini. The latter is an implementation of a BERT cross-encoder named CEDR. It tackles the issue of input length limitation and takes advantage of the word embeddings in output of the model to compute a relevance score. We investigate a strategy to integrate biomedical thesaurus (MeSH) knowledge into this model.

Keywords

biomedical document ranking, information retrieval, thesaurus-based knowledge, BERT cross-encoder, multi-stage retrieval

1. Introduction

The number of biomedical publications keeps growing year after year. This allows researchers to have access to a vast amount of knowledge but a drawback is that it becomes more and more difficult to get a precise result when querying information sources. The Covid-19 pandemic reminded us that it is necessary to facilitate the navigation of researchers through publication databases in order to help their bibliographic work so they can easily assimilate information.

The BioASQ[1] initiative [1] aims to push forward biomedical information retrieval by running an annual evaluation campaign to solve several tasks. Task B of the challenge [2] especially focuses on question answering (QA) through two phases. Phase A tackles the issue of document retrieval and text snippets extraction while Phase B relates to answer generation and summarization.

TaskB-PhaseA is particularly interesting for us because it provides an evaluation framework on a similar problem we would like to address in the context of the FORUM portal[2]. FORUM is a Semantic Web framework based on a knowledge graph that "provides links between chemical

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
mael.lesavourey@irit.fr (M. Lesavourey); gilles.hubert@irit.fr (G. Hubert)
https://www.irit.fr/~Gilles.Hubert/ (G. Hubert)

[1] http://www.bioasq.org/

[2] https://forum-webapp.semantic-metabolomics.fr

compounds and biomedical concepts supported by literature" [3]. The links are built using articles metadata and combining a thesaurus and different ontologies of the biomedical domain. In addition, it is possible to retrieve the list of articles that support each link.

We intend to extend FORUM by taking advantage of the textual content of the articles and provide suggestion on prior articles for each pair chemical compound – biomedical concept. The whole system will be a multi-stage retrieval approach where the knowledge graph plays the role of a first-stage retriever. The main difference with BioASQ TaskB is that the first-stage retrieval is already completed and the articles to prioritize are necessarily within the retrieved list. Despite this detail, we can still benefit from BioASQ benchmarks to evaluate document ranking approaches.

In this paper we present the system used for Task 11B Phase A on document ranking based on a retrieve and rerank strategy. We applied a first stage retrieval based on BM25 [4] which is a straightforward yet effective approach. In order to rerank the documents we used CEDR-KNRM [5], a BERT-based model for full-length documents that combines the [CLS] token representation and the contextual embeddings of the query and document terms. We propose to integrate thesaurus knowledge to this model via a marking strategy.

2. Materials and Method

2.1. System description

Our method to address this task is composed of two main modules: a retriever and a reranker. This is a common approach that consists in decomposing document ranking into several parts [6]. The retriever aims at creating a smaller candidate list (hundreds of documents) from the whole corpus (generally more than ten million). It is usually based on bag of words (BoW) representations [7] which are less efficient than deep contextualized representations but drastically reduce computational cost. The candidate list is then reranked using a high-performance model. State-of-the-art models to achieve this task are transformer-based methods like BERT [8], especially when using a cross-encoder architecture where the query and candidate document are passed together as input. Computing a relevance score is usually done by adding a single linear layer on top of the model output.

As mentioned in the introduction our goal is to develop a robust and effective reranker. For our retriever, we decided to take advantage of a well-known system rather than proposing a new approach. We used Pyserini [9], a Python library for reproducing information retrieval research. We created our own indices with PubMed articles and used BM25 implementation to retrieve the top 500 articles most likely to respond to the query.

Classical BERT approaches for text ranking take query and candidate document as input with the following sequence: [CLS] query [SEP] document [SEP]. The system is trained to predict if a document is relevant or not by using the [CLS] token embedding. Documents are ranked using their probability of being relevant to a certain query. We identify several challenges of common approaches regarding Task 11B Phase A. BERT produces contextual embeddings for each word given in input along with the [CLS] token representation. Taking advantage of these embeddings to compute the probability of a document being pertinent could extend the classical approach. Another limitation is the maximum number of tokens that BERT can handle at once, i.e., 512. This limit is easily exceeded with scientific publications even if we consider only their title and abstract. Finally, we would like to investigate if the BERT-based model can benefit from the knowledge of biomedical structured vocabularies. We illustrate the pipeline used for our method in Figure 1.

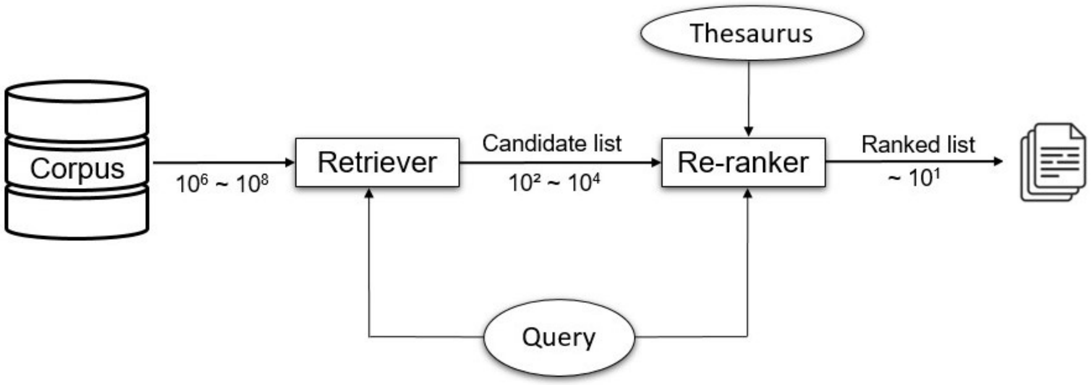


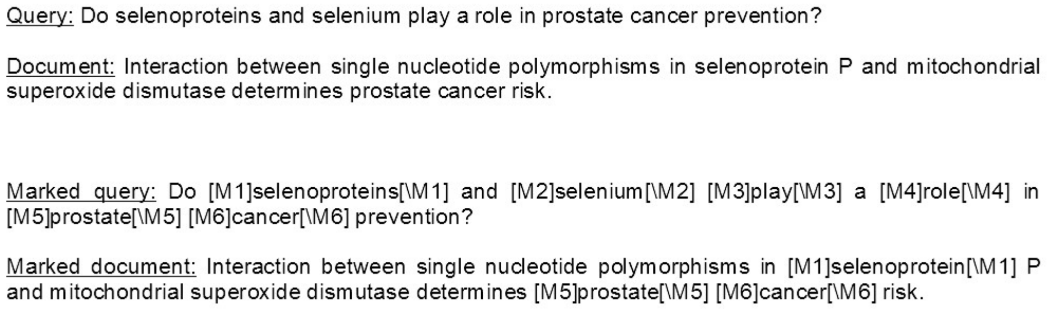
Figure 1: Overview of our proposed two-stage retrieval architecture

We decided to base our method on the Contextualized Embeddings for Document Ranking (CEDR) model as it addresses two challenges: input length limitation and the use of term embeddings. [5] proposed to keep the query as a whole and to divide too long documents in chunks of the same length, then apply a cross-encoder on each query-chunk pairs. The final [CLS] representation is obtained by computing an average pooling on the [CLS] vector of each of these pairs. Moreover, they created a similarity matrix between the query and document by concatenating the outputs of the cross-encoders. They passed the similarity matrix along with the classification token as input of 3 neural ranking models. Best performance were obtained with K-NRM [10], a kernel-based ranking model that takes advantage of word embeddings and soft match features to produce a ranking score.

We implemented a version of the CEDR model using K-NRM algorithm to create a baseline to evaluate our work. We propose to extend this method by taking advantage of the Medical Subject Headings[3] (MeSH) thesaurus, a controlled vocabulary produced by the National Library of Medicine[4] (NLM) to index every citation appearing in MEDLINE/PubMed. Some works [11] highlighted the significant effect of input sequence organisation in the performance of BERT-based systems. [12] proposed to exploit exact term-matches between the query and the document by marking the input of a BERT cross-encoder. We propose to integrate MeSH thesaurus knowledge the same way. Given a query and a document we implemented a marking strategy of biomedical terms assuming that if a word is referenced in the MeSH thesaurus it will carry a more important information. Differing from [12], we limited the marking to the main headings and entry terms referenced in MeSH. We used the same tag for a main heading and its entry terms in order to highlight not only the exact matches but also synonyms. An example of the

[3] https://www.nlm.nih.gov/mesh/meshhome.html
[4] https://www.nlm.nih.gov/

marking strategy is given in Figure 2.



Query: Do selenoproteins and selenium play a role in prostate cancer prevention?

Document: Interaction between single nucleotide polymorphisms in selenoprotein P and mitochondrial superoxide dismutase determines prostate cancer risk.

Marked query: Do [M1]selenoproteins[\M1] and [M2]selenium[\M2] [M3]play[\M3] a [M4]role[\M4] in [M5]prostate[\M5] [M6]cancer[\M6] prevention?

Marked document: Interaction between single nucleotide polymorphisms in [M1]selenoprotein[\M1] P and mitochondrial superoxide dismutase determines [M5]prostate[\M5] [M6]cancer[\M6] risk.

Figure 2: Example of the marking strategy using a query and the title of its most relevant article.

2.2. Data

We worked with the PubMed Annual Baseline[5] for 2023 for which we removed all citations that do not contain an abstract. We also worked with the dataset provided by BioASQ[6] which contains all the questions of the previous editions along with their gold standards (relevant articles).

We built a training set composed of relevant and irrelevant articles for each query. The relevant articles are the ones given by BioASQ. Irrelevant articles were chosen in the top 100 articles retrieved by BM25 for each query. We selected two times more irrelevant articles than relevant ones. Indeed [13] has shown that randomly choosing negative samples leads to worst results than selecting hard negative ones. So we selected negative samples close to the positive ones in terms of BoW representations.

In order to mark the queries and candidate documents we downloaded the latest release of MeSH thesaurus and we kept their main headings and entry terms.

3. Results

Task 11B Phase A was organized on 4 different batches. Each participating system had to return a ranked list of at most 10 relevant articles and/or text snippets. The evaluation metric used for the official scores is the Mean Average Precision (MAP) due to its capability to take into account the order of the submitted items.

In Table 1 we present the results of our runs for batches 2, 3, and 4 of the Task 11B Phase A on document ranking. We submitted predictions with two systems: CEDR and Mark-CEDR. The first is an implementation of the CEDR-KNRM model proposed by [5]. The latter is the same model trained with marked queries and documents as described in section 2.1.

[5] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/
[6] http://participants-area.bioasq.org/datasets/

Table 1
Results of our systems in Task 11B Phase A for document ranking

Batch	Model	MAP	System Rank	Team Rank
2	Mark-CEDR	0.2405	19/33	5
3	Mark-CEDR	0.1332	25/35	6
3	CEDR	0.1020	26/35	6
4	Mark-CEDR	0.1339	17/27	4
4	CEDR	0.1564	16/27	4

The best result for our proposed method was obtained during batch 2 of the challenge where we achieved a 0.2405 MAP score. During batch 3 and batch 4 Mark-CEDR performed the same way while the base model (CEDR) gained more than 0.05 points.

In order to understand these results we led further investigations regarding question type (Table 2) and query length (Table 3). From Table 2 we learn that both models always perform better when treating yes/no questions. There is a huge drop (-0.092) between batch 3 and 4 for list questions treated by Mark-CEDR while the base model maintains its performance. That partially explains why CEDR performed better during the last batch.

Table 2
MAP scores per question type

	Yes/No		**List**		**Summary**		**Factoid**	
Batch	Mark-CEDR	CEDR	Mark-CEDR	CEDR	Mark-CEDR	CEDR	Mark-CEDR	CEDR
3	0.163	0.167	0.139	0.074	0.139	0.092	0.097	0.07
4	0.214	0.23	0.047	0.09	0.11	0.108	0.181	0.208

Table 3 indicates that, from batch 2 to batch 4, the mean number of MeSH terms detected per query is decreasing along with the mean number of tokens in each question. When question length grows we detect more MeSH and it seems that Mark-CEDR achieves better results in that case. On the contrary best score for CEDR was obtained with smaller queries.

Table 3
Queries lengths and MeSH terms detected per query

Batch	Mean number of tokens	Mean number of MeSH terms
2	10.72	1.52
3	10.4	1.3
4	9.1	1.08

4. Conclusion/Discussion

The first version of our approach that consists in integrating external structured knowledge provides average performance results on all batches and, unfortunately, inconsistent ones. The scores obtained in batch 3 were promising because the mark version was performing better by 0.03 points. However it was quite the opposite during the last batch which reveals that the performance of this first implementation is unstable. This observation is in line with the questioning of [11] about the effectiveness of exact matching against strong baselines.

However we identified several points that can be addressed to improve our model. The first one would be to use a stronger retriever. We observed that for many questions some of the expected articles were not in the top 500 retrieved. In that cases the reranker cannot perform properly. In order to improve the reranker we could explore new types of marking strategies to investigate the instability issue. Moreover we could take MeSH structure into account. Indeed we only used MeSH as a controlled vocabulary but we did not take advantage of its hierarchical tree structure. Finally, it would be wise to exploit a BERT model pre-trained on a biomedical corpus like BioLinkBERT [14] or PubMedBERT [15] which achieved state-of-the-art results on the BLURB benchmark[7] [16].

References

[1] A. Nentidis, A. Krithara, G. Paliouras, E. Farre-Maduell, S. Lima-Lopez, M. Krallinger, Bioasq at clef2023: The eleventh edition of the large-scale biomedical semantic indexing and question answering challenge, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2023, p. 577–584. URL: https://doi.org/10.1007/978-3-031-28241-6_66. doi:10.1007/978-3-031-28241-6_66.

[2] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Scientific Data 10 (2023) 170.

[3] M. Delmas, O. Filangi, N. Paulhe, F. Vinson, C. Duperier, W. Garrier, P.-E. Saunier, Y. Pitarch, F. Jourdan, F. Giacomoni, C. Frainay, FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases, Bioinformatics 37 (2021) 3896–3904. URL: https://doi.org/10.1093/bioinformatics/btab627. doi:10.1093/bioinformatics/btab627.

[4] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.

[5] S. MacAvaney, A. Yates, A. Cohan, N. Goharian, Cedr: Contextualized embeddings for document ranking, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1101–1104. URL: https://doi.org/10.1145/3331184.3331317. doi:10.1145/3331184.3331317.

[6] R. F. Nogueira, W. Yang, K. Cho, J. Lin, Multi-stage document ranking with BERT, CoRR abs/1910.14424 (2019). URL: http://arxiv.org/abs/1910.14424. arXiv:1910.14424.

[7] https://microsoft.github.io/BLURB/

[7] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, X. Cheng, Semantic models for the first-stage retrieval: A comprehensive review, ACM Trans. Inf. Syst. 40 (2022). URL: https://doi.org/10.1145/3486250. doi:10.1145/3486250.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[9] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, R. Nogueira, Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2356–2362. URL: https://doi.org/10.1145/3404835.3463238. doi:10.1145/3404835.3463238.

[10] C. Xiong, Z. Dai, J. Callan, Z. Liu, R. Power, End-to-end neural ad-hoc ranking with kernel pooling, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 55–64. URL: https://doi.org/10.1145/3077136.3080809. doi:10.1145/3077136.3080809.

[11] J. Lin, R. F. Nogueira, A. Yates, Pretrained transformers for text ranking: BERT and beyond, CoRR abs/2010.06467 (2020). URL: https://arxiv.org/abs/2010.06467. arXiv:2010.06467.

[12] L. Boualili, J. G. Moreno, M. Boughanem, Markedbert: Integrating traditional ir cues in pre-trained language models for passage retrieval, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1977–1980. URL: https://doi.org/10.1145/3397271.3401194. doi:10.1145/3397271.3401194.

[13] J. Zhan, J. Mao, Y. Liu, J. Guo, M. Zhang, S. Ma, Optimizing dense retrieval model training with hard negatives, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1503–1512.

[14] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, 2022. arXiv:2203.15827.

[15] R. Tinn, H. Cheng, Y. Gu, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Fine-tuning large neural language models for biomedical natural language processing, CoRR abs/2112.07869 (2021). URL: https://arxiv.org/abs/2112.07869. arXiv:2112.07869.

[16] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, CoRR abs/2007.15779 (2020). URL: https://arxiv.org/abs/2007.15779. arXiv:2007.15779.