Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 1277–1279

DOI: 10.15439/2023F3900
ISSN 2300-5963 ACSIS, Vol. 35

Multi-index Retrieve and Rerank with Sequence-to-Sequence Model

Konrad Wojtasik
Wrocław University of Science and Technology
Email: konrad.wojtasik@pwr.edu.pl

***Abstract*—This paper presents the solution to PolEval 2022 Task 3: Passage Retrieval. The main goal of the task, was to retrieve relevant text passages for the query. There were three different domains of passages: wikipedia passages, allegro faq and legal documents. The proposed solution incorporated both dense and lexical indexes, as well as, reranking model and reached 67.44 NDCG@10 score in official evaluation.**

I. Introduction

INFORMATION retrieval is a task that aims at finding relevant information from a collection of documents. It involves searching through the whole collection based on a query provided by a user. The query is usually a question, and the system has to provide the documents or text fragments that contain the answer to the question.

The landscape of existing information retrieval datasets is predominantly populated by English language resources. This presents a challenge for the development and evaluation of models designed to handle a variety of languages. In response to this, the BEIR benchmark[11] was developed, a unique benchmark that focuses on the evaluation of information retrieval models in a zero-shot setting. The primary training dataset utilized in this benchmark is MS MARCO[8], a large-scale dataset in English, while a diverse array of other datasets are employed for zero-shot evaluation.

Recognizing the need for multilingual resources, a multilingual version of MS MARCO[1] was created. This version was translated into various languages using state-of-the-art automated machine translation techniques, expanding the reach of the dataset beyond English. However, it was observed that the Polish language was conspicuously absent from this collection.

In an effort to fill this linguistic gap and foster the development of Polish language information retrieval models, a Polish version of the dataset was introduced in the BEIR-PL benchmark[13], as well as Massive Automatically-created Polish Question Answering Dataset[10], which is a large collection of question and passage pairs.

The main goal of PolEval 2022 Information Retrieval task was to propose a cross-domain question-answering retrieval system in Polish language. The task encompassed three distinct domains of queries and documents. The training set was exclusively composed of data related to the trivia domain. The other domains, namely legal and customer support, were approached from a zero-shot perspective.

II. Related Work

The most common approach to the information retrieval systems is incorporating the two-step retrieval process with reranking. With this two-step process, the information retrieval system can provide more accurate and relevant results, enhancing the overall effectiveness.

A. Retrieval

In the retrieval phase of information retrieval, the system matches a user's query with the indexed collection of documents to identify the initial set of relevant items. It should be fast and efficient, as the collections may contain millions of documents.

One way to perform retrieval is lexical matching. Usually, it utilizes the Best Matching 25 (BM25), which compares the frequency of terms in the query and the document. BM25 is a ranking function used by search engines to estimate the relevance of a document to a given search query based on the terms it contains.

Elasticsearch[1], an open-source search engine, is a popular implementation of this approach. It uses BM25 as its default scoring function. This method has become a standard baseline approach for most retrieval benchmarks due to its good performance, effectiveness and lack of training.

Recent trends indicate that neural retrievers are capable of surpassing the performance of lexical term matching[6]. The neural network, based on pre-trained transformer model, encode both the query and the document into a low-dimensional space. The encodings are compared using inner product or cosine similarity. By pre-encoding the corpus into the index, retrieval can become very efficient and run online with millisecond level latency with libraries that support similarity search of dense vectors, such as FAISS[2]. Neural retrievers are trained as bi-encoders with contrastive loss, which makes the representations of passages and queries with the same information similar.

B. Reranking

The reranking phase is a subsequent process that follows the initial retrieval. It involves reordering the retrieved documents based on more complex models or additional features to

[1] https://www.elastic.co/

[2] https://github.com/facebookresearch/faiss

improve the ranking and elevate the most relevant documents to the top positions in the ranked list. The reranking problem can be formulated as a classification problem, where the query and the document are passed jointly to the model and the result is a binary classification, if the document is relevant for the query or not. Pretrained transformer models, like BERT model, can be effectively utilized for this task. The BERT model[9], trained using cross-entropy loss, is capable of performing the classification based on the representation of the [CLS] token. Additionally, a single-layer neural network is employed to compute the probability of a passage's relevance to a given query. This approach leverages the power of transformer architectures to capture complex semantic relationships in the data, thereby enhancing the accuracy of the classification task. Encoder-decoder models, such as T5, have also been employed for the reranking task. In the context of classification tasks, the tokenizer is augmented with two unique tokens. The first token, representing a 'true' value, is generated when the text passage demonstrates relevance to the query in question. On the other hand, the second token, denoting a 'false' value, is generated in instances where the passage does not exhibit relevance to the query.

III. Dataset

Data domains in PolEval 2022 Information Retrieval task:

- trivia domain - knowledge based general questions from a popular quizzes and passages from Wikipedia pages. The corpus contains 7M passages.
- customer support domain - FAQ based customer questions from the allegro.pl platform. The corpus includes 921 passages.
- legal domain - dataset was constructed based on legal documents, with questions formulated around the content of these documents. The corpus comprises a total of 26287 passages.

The training set contained only data related to the trivia domain, other domains were treated as zero-shot approach.

IV. Solution

The final solution incorporates three dense indexes, where documents are embedded with different neural encoders and one lexical BM25 index. The reranking is performed by the multilingual T5 model.

A. Experiment Setup

Most of the experiments were performed on NVIDIA RTX 3090 with 24GB GPU memory, except the final submission with mT5-13B model, which was run on A100 with 80GB GPU memory, due to the model size and computational complexity.

B. Experiments

The system was constructed from a combination of various dense indexes, each created with different models, as well as a lexical index created using the BM25 algorithm. From each index, the top documents were retrieved, and the collective set of all these documents was then forwarded to the reranking model for further refinement.

TABLE I
Retrievers results on TEST A without reranking. Combined* represents a score of all taken set of all returned passages from all retrievers at top k.

Retriever	NDCG@10	Recall@10	Recall@100	Recall@1000
BM25	50.77	37.55	45.29	51.65
mContriever	**58.43**	**56.03**	**70.05**	**77.93**
LaBSE	29.84	32.01	50.59	62.87
mDPR	31.42	33.95	51.32	66.09
Combined*	-	*63.84*	*75.15*	*81.34*

Experimental results shown in the table I demonstrated that, the best performance achieve mContriever dense retriever. Other dense retrievers got lower NDCG@10 and Recall@10 metrics than BM25, but they achieve better recall score at higher number of retrieved passages. Combined* results show recall when all top passages are combined from all four retrievers. In practice, due to duplicates, there are on average 3 times of the k passages. So for Recall@10, there were on average 30 passages taken into account for each query.

In the final solution, the following three dense retriever models were employed:

- mContriever-base-msmarco [3] - mBERT[7] based retriever trained in unsupervised manner with contrastive loss on multilingual data and afterward fine-tuned on English MS MARCO dataset[5].
- LaBSE [4] - is a language-agnostic BERT model for sentence embedding[3].
- mDPR-question-nq [5] and mDPR-passage-nq [6] - mDPR is a multilingual Dense Passage Retriever[15], a bi-encoder network where query and passage are encoded with different encoders trained in contrastive manner.

For reranking stage, different rerankers were taken into account. Initial testing were performed with following rerankers fine-tunned on MS MARCO dataset to reranking task:

- mMiniLMv2 [7] - multilingual MiniLMv2 model[12].
- mDeBERTa - improved multilingual DeBERTa model[4].
- mBERT[8] - multilingual BERT model[7].
- plT5[9] - T5-based language model train on Polish corpora[2]. Fine-tunned on Polish MS MARCO from BEIR-PL.
- mT5[10] - multilingual text-to-text transformer[14].

As shown in the table II, the bigger model, the better result. That is why, in the final solution, the mT5 model[11] with 13 billion parameters was employed. This model is very

[3] https://huggingface.co/nthakur/mcontriever-base-msmarco
[4] https://huggingface.co/sentence-transformers/LaBSE
[5] https://huggingface.co/castorini/mdpr-question-nq
[6] https://huggingface.co/castorini/mdpr-passage-nq
[7] https://huggingface.co/cross-encoder/mmarco-mMiniLMv2-L12-H384-v1
[8] https://huggingface.co/amberoad/bert-multilingual-passage-reranking-msmarco
[9] https://huggingface.co/clarin-knext/plt5-base-msmarco
[10] https://huggingface.co/unicamp-dl/mt5-3B-mmarco-en-pt
[11] https://huggingface.co/unicamp-dl/mt5-13b-mmarco-100k

TABLE II
RERANKER RESULTS ON TEST A RERANKING TOP 1000 BM25 RESULTS.

Reranker	NDCG@10 TEST A
plt5-base	66.02
plt5-large	69.09
mMini-lm	67.87
mDeberta	68.67
mBert	59.78
mT5-3B	**70.28**

TABLE III
FINAL RESULT ON TEST A AND TEST B.

	NDCG@10 TEST A	NDCG@10 TEST B
Final solution	**74.28**	**67.44**

computationally expensive to run, that is why there are no additional experiments performed on this model.

The final solution was achieved by retrieving the top 100 passages for each query from each dense retriever, and the top 100 reranked passages using the plT5-large model from the top 1000 retrieved from the BM25 index. The use of plT5 was dictated by the computational complexity of the mT5-13B model. As shown in table I, combined retrievers achieve already very high 75.15 recall at top 100 passages from each retriever.

Subsequently, the set of all top 100 passages from all sources was reranked using the mT5-13B model. The total number of passages to rerank was approximately 310 instead of 400, as some passages were duplicated. The final results are shown in the table III.

V. CONCLUSION

The optimal strategy to enhance the results of information retrieval involves the utilization of various dense retrievers, in conjunction with lexical BM25 matching. This approach amplifies the overall recall of the system, ensuring that passages not deemed relevant by one model may be identified as such by another. Another key insight is the importance of employing an effective reranker. The performance of the reranker is often correlated with the size of the model, with larger models typically achieving superior scores compared to their smaller counterparts. However, this advantage is accompanied by an increase in computational cost, which must be taken into account. Furthermore, reranking a larger number of top retrieved passages enhances the likelihood that a relevant passage is included in the set of reranked passages. This strategy, while potentially more computationally intensive, can significantly improve the precision of the retrieval system at the top ranks, which is often a critical requirement in many information retrieval applications.

ACKNOWLEDGMENT

The work was financed by the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN - Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19.

REFERENCES

[1] Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of MS MARCO passage ranking dataset. *CoRR*, abs/2108.13897, 2021.

[2] Aleksandra Chrabrowa, Łukasz Dragan, Karol Grzegorczyk, Dariusz Kajtoch, Mikołaj Koszowski, Robert Mroczkowski, and Piotr Rybak. Evaluation of transfer learning for polish with a text-to-text model. *arXiv preprint arXiv:2205.08808*, 2022.

[3] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[4] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543, 2021.

[5] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2021.

[6] 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 2020. Association for Computational Linguistics.

[7] Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. How language-neutral is multilingual bert?, 2019.

[8] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. November 2016.

[9] Rodrigo Frassetto Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *CoRR*, abs/1901.04085, 2019.

[10] Piotr Rybak. MAUPQA: Massive automatically-created Polish question answering dataset. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 11–16, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[11] 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.

[12] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *CoRR*, abs/2012.15828, 2020.

[13] Konrad Wojtasik, Vadim Shishkin, Kacper Wołowiec, Arkadiusz Janz, and Maciej Piasecki. Beir-pl: Zero shot information retrieval benchmark for the polish language, 2023.

[14] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934, 2020.

[15] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. Towards best practices for training multilingual dense retrieval models, 2022.