Hybrid retrievers with generative re-rankers

Marek Kozlowski

National Information Processing Institute (OPI PIB)

Introduction

- **Information retrieval (IR)** is focused on obtaining relevant information from a collection of diverse data resources (mainly textual ones).
- Traditionally, lexical methods like TF-IDF+Cosine or BM25 were often applied in the retrieval systems. They are **robust, interpretable, and unsupervised**.
- Recently, neural information retrieval (NIR) has surpassed those above lexical methods by fine-tuning the pre-trained language models like Bert or T5. **They beat the classical methods according to the quality efficiency, but there are also some drawbacks e.g. need for a high quality, relevant training dataset.**
- Existing neural information retrieval (NIR) models have often been analysed **in single domain narrowed area, where training sets are scarce in languages other than English**.
- **Poleval 2022 third task is dedicated to Passage retrieval in polish language area,** a specialized type of IR application that retrieves relevant polish passages (usually paragraphs) for sentence like queries.

Data, that I used

- Poleval dev – 599 wiki queries

wiki-trivia	**Jak nazywa się dowolny odcinek łączący dwa punkty okręgu?**
wiki-trivia	W którym państwie leży Bombaj?
wiki-trivia	Co budował w Egipcie inżynier Tarkowski, ojciec Stasia z powieści „W pustyni i w puszczy”?
wiki-trivia	Kwartet – to ilu wykonawców?
wiki-trivia	Jak nazywa się trucizna przyrządzona z korzenia szaleju?
wiki-trivia	Do którego państwa należą wyspy Rodos i Korfu?
wiki-trivia	W którym filmie Krzysztofa Kieślowskiego zagrali Olaf Lubaszenko i Grażyna Szapołowska?

- Poleval train - 4401 wiki quries, with assigned positive passages

1	**wiki-trivia**	**Jak nazywa się pierwsza litera alfabetu greckiego?**
2	**wiki-trivia**	Czy w państwach starożytnych powoływani byli posłowie i poselstwa?
3	**wiki-trivia**	W jakim zespole występowała Hanka w filmie „Żona dla Australijczyka”?
4	**wiki-trivia**	Który numer boczny nosi czołg Rudy z „Czterech pancernych”?

1	**19291-0**	**127-27**	**1091819-0**	**464044-0**	**5137974-0**
2	**3742-17**	85719-4	85719-5		
3	**643566-1**	588756-39			
4	**12381-0**	11959-15	1874562-1	447358-1	509400-73

- MS MARCO - MS MARCO (MicroSoft MAchine Reading COmprehension) is a large-scale dataset focused on machine reading comprehension. There are also available translations of this English corpora called MMARCO (a multilingual version of the MS MARCO passage ranking dataset comprising 13 languages that was created using machine translation). MS MARCO was translated into polish by us.

Classical IR approaches

- BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of their proximity within the document.
- **Okapi BM25** is a retrieval **model based on the probabilistic retrieval framework**
- BM25 often achieves better performance compared to TF-IDF
- Simple TF-IDF rewards term frequency and penalizes document frequency. **BM25 goes beyond this to account for document length and term frequency saturation.**

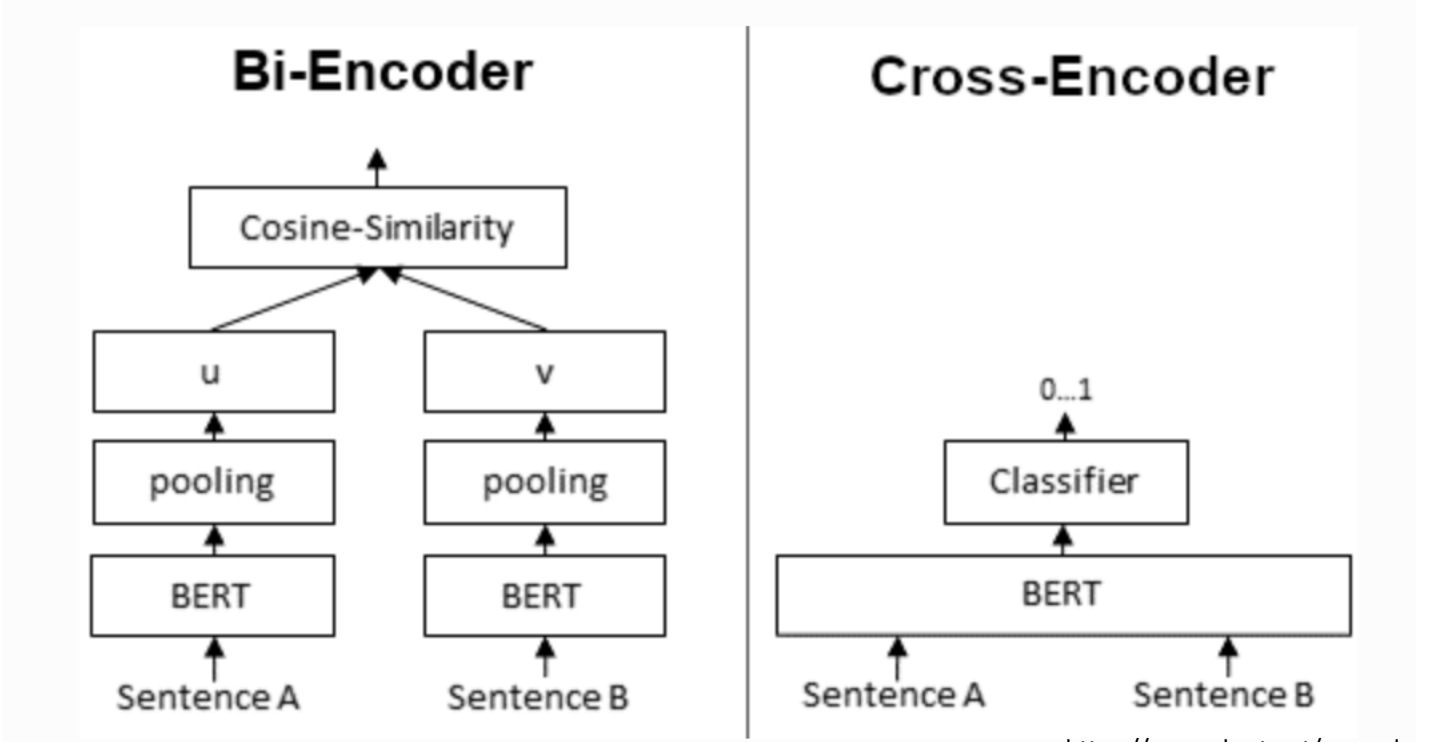
Given a query Q, containing keywords $q_1, \ldots, q_n$, the BM25 score of a document D is:

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

Deep learning IR approaches

Bi-encoders produce an embedding vectors for a given pair of two sentences. We pass to a BERT independently the sentences A and B, which result in the sentence embeddings u and v. These sentence embedding can then be compared using cosine similarity.

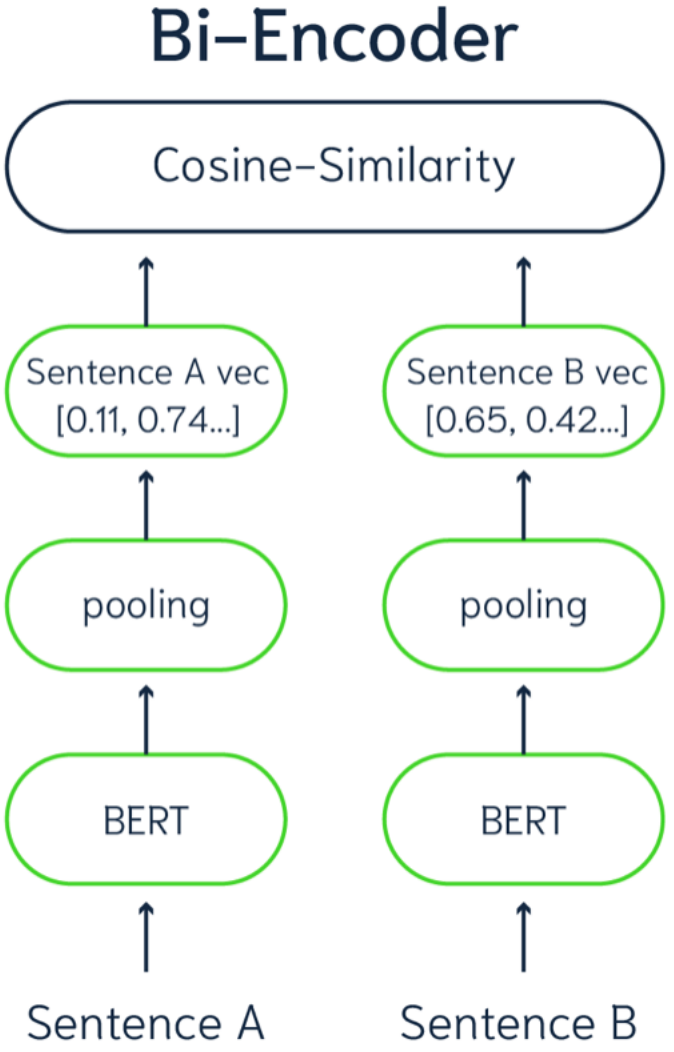
A Cross-Encoder does not produce a sentence embedding. We pass both sentences simultaneously to the Transformer network. It produces than an output value between 0 and 1 indicating the similarity of the input sentence pair.



https://www.sbert.net/examples/applications/cross-encoder/README.html

Bi-encoder

- Bi-Encoders is a two-branch type model architecture.
- The queries and passages are passed independently to the transformer network to produce fixed sized embeddings.
- We store passage embeddings in FAISS indexes (FAISS is a library for the efficient similarity search and clustering of dense vectors).
- Training phase consumes triplets in the format: (query, positive_passage, negative_passage).
- Negative passage are hard negative examples, that where retrieved by lexical search. We use Elasticsearch to get (hard_negatives_max = 10) hard negative examples given a positive passage.
- Bert type model used RoBERTa base and large (https://github.com/sdadas/polish-Roberta)
- For training, we use such settings:
  - Contrastive Learning, namely MultipleNegativesRankingLoss
  - Batch size = 32 (limited by the GPU V100 RAM capacity)

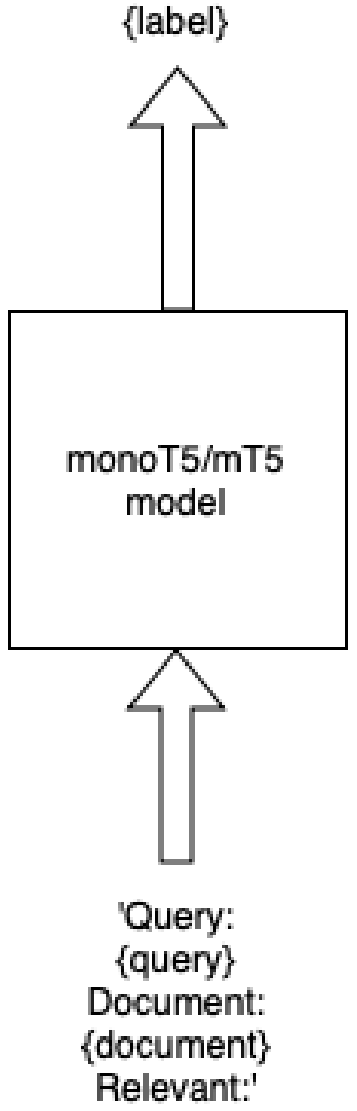


https://weaviate.io/blog/2022/08/Using-Cross-Encoders-as-reranker-in-multistage-vector-search.html

T5 as a cross-encoder

- In typical cross-encoders we have language model as BERT/RoBERTa/miniLM where we pass both sentences simultaneously to the network. It produces than an output value between 0 and 1 indicating the similarity of the input sentence pair.
- T5 model can be used in a cross-encoder typical way, where the target text is a label, and we get also its probability scores.
- In 1.01 version (June 22) https://github.com/beir-cellar/beir/releases/tag/v1.0.1

"Added support for the T5 reranking model: monoT5 reranker. We added the support of the monoT5 reranking model within BEIR. These are stronger (but complex) rerankers that can be used to attain the best reranking performances currently on the BEIR benchmark."



{label} can be e.g 'yes'/'no', we are also interested in logprobs

Re-rankers evaluation

TABLE I
EVALUATION OF DIFFERENT RE-RANKERS THAT HAVE THE SAME CLASSICAL, LEXICAL RETRIEVER: BM25, AND TOP-K=100 RESULTS RETRIEVED FROM BM25. NDCG METRICS WERE CALCULATED ON THE POLEVAL VALIDATION DATASET.

Method name	Retriever	Re-ranker	NDCG@10(%)
baseline	BM25 (default)	None	21.05
baseline@plugged	BM25 (morfologik)	None	27.67
bi-enc@base	BM25 (morfologik)	Bi-encoder (RoBERTa-base)	38.03
gpt3@curie	BM25 (morfologik)	GPT3 (curie)	40.06
bi-enc@large	BM25 (morfologik)	Bi-encoder (RoBERTa-large)	41.19
mT5@base	BM25 (morfologik)	mT5-base-mmarco	42.87
mT5@xxl	**BM25 (morfologik)**	**mT5-13b-mmarco**	**45.88**

My inspiration for the final solution

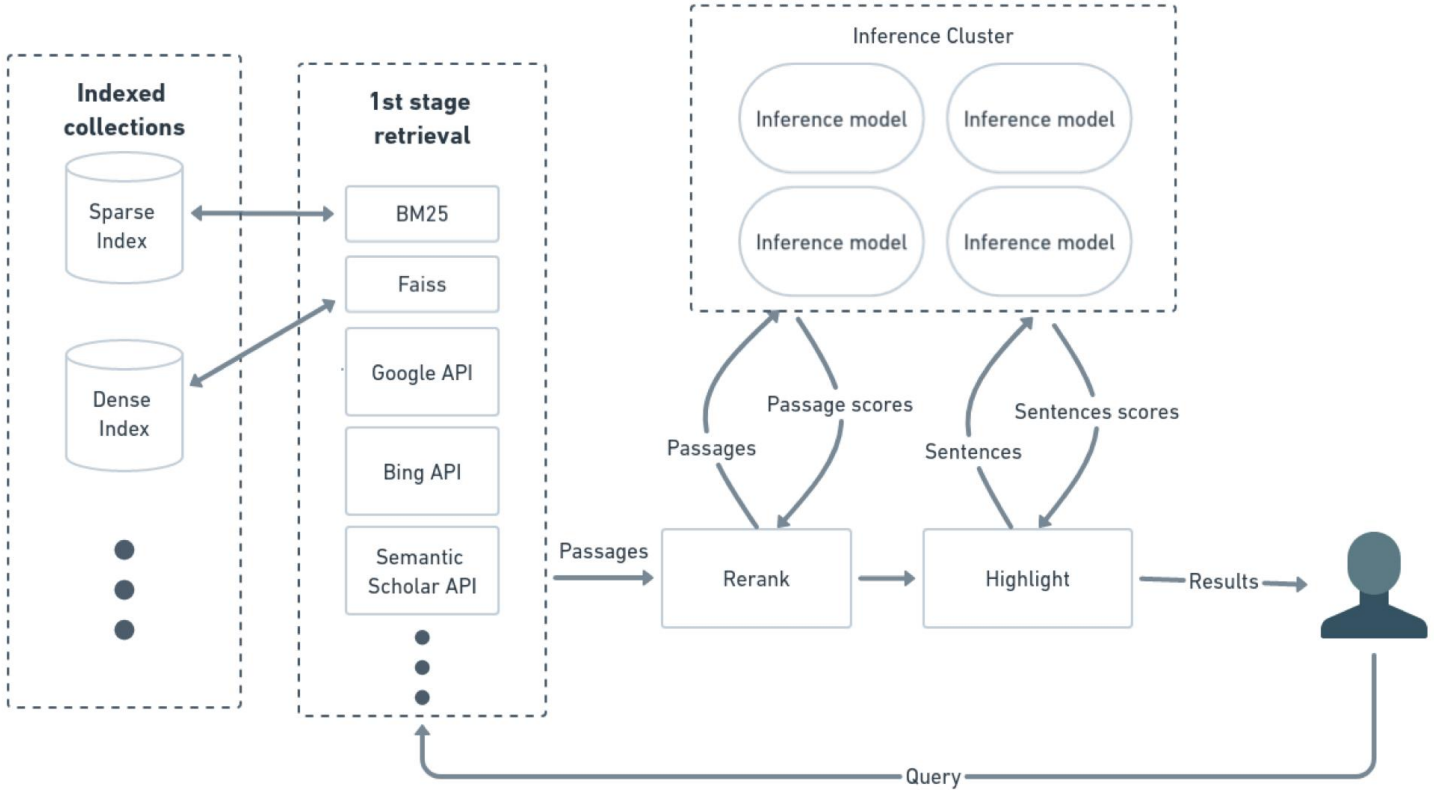


Figure 1: Retrieval architecture of NeuralSearchX. Note that reranking and highlighting are performed by the same neural model, which decreases hardware idle time.

Almeida, Thales Sales, et al. "NeuralSearchX: Serving a Multi-billion-parameter Reranker for Multilingual Metasearch at a Low Cost." *arXiv preprint arXiv:2210.14837* (2022).

The submitted solution

- The goal of the task was to create a system for cross-domain passage retrieval so I developed **one generic solution (including three partial models) for all mentioned domains**
- Our solution to this task is a **multi-stage neural information retrieval system**.
- The **first stage consists of a candidate passage retrieval step** in which passages are retrieved using federated search over sparse (BM25) and dense indexes (two FAISS indexes built using bi-encoder type retrievers based on polish RoBERTa models).
- We trained two types of dense retrievers:
  - using RoBERTa-base-v2 as a transformer model, and fine-tuning one epoch on 500 000 triplets (135 000 Poleval ones and 370 000 Polish MS Marco ones randomly selected);
  - using RoBERTa-large-v2 as a transformer model, and training 10 epochs on a few million triplets (several million Polish MS Marco ones randomly selected), then fine-tuned for one epoch on all 135 000 PolEval triplets.
- After the candidate passage retrieval, **the next step is to merge all the candidates in a single list, and then re-ranking** them with mt5-13b-mmarco model. The model scores each passage by its relevance to a given query.

Retrievers' settings evaluation of final solution

TABLE II

EVALUATION OF DIFFERENT RETRIEVERS' SETTINGS (SPARSE—BM25, AND TWO DENSE RETRIEVERS BASED ON BI-ENCODER ARCHITECTURE, AND TWO TYPES OF ROBERTA MODELS) OF OUR SOLUTION. THE NDCG METRICS WERE CALCULATED ON THE POLEVAL TEST-A DATASET, THE FIRST RELEASED DATASET THAT WAS USED FOR PUBLIC LEADERBOARD PURPOSES.

Top@K sparse retriever – BM25	**Top@K dense retriever – bi-encoder (RoBERTa-base)**	**Top@K dense retriever – bi-encoder (RoBERTa-large)**	**NDCG@10(%)**
1	28	28	73.71
7	50	0	74.61
7	0	50	74.71
7	**25**	**25**	**75.32**
7	45	45	74.83

Conclusions

- Our solution was a multi-stage neural information retrieval system:
  - a candidate passage retrieval step is a federated search over sparse and dense indices
  - re-ranking step of the previously-selected passages with mt5-13b-mmarco.
- Hybrid retrievers outperform baseline single retriever (BM25)
- Reranker model as mT5-13b trained on mmarco outperforms significantly other type of rerankers
- High computational costs and high time consumption of mt5-xxl model application - even in the inference phase, so I have to:
  - use at least two V100 gpus to perform inference with mt5-xxl
  - use maximum several hundred candidate passages for re-ranking
- There is a huge need for further research in the both directions – retrieval and rerank on polish custom datasets

Future plans

Spaces | sdadas/pirb | like 0 | Running | App | Files | Community

Polish Information Retrieval Benchmark

Polish Information Retrieval Benchmark (PIRB) covers 41 Polish multidomain information retrieval tasks. Its purpose is to evaluate Polish and multilingual information retrieval methods on a wide range of problems with different characteristics, thus testing the generalization ability of the models and their zero-shot performance. The benchmark includes pre-existing datasets such as MaupQA, BEIR-PL and PolEval-2022. We have also added new, previously unpublished datasets. The "Web Datasets" group contains real questions and answers from Polish web services.

Above each task group there is a panel which contains the following actions: + expands the group into individual tasks, ? displays a tooltip providing additional information about the group, 🔗 includes a link to the paper or the official website, × closes the group and removes it from the evaluation.

Evaluation metric: NDCG@10 MRR@10 Recall@100 Accuracy@1

Filter models

Model	Tasks won	Average (41 tasks)	PolEval-2022 Average (7 tasks)	Web Datasets Average (9 tasks)	BEIR-PL Average (11 tasks)	MaupQA Average (12 tasks)	Other Average (2 tasks)
intfloat/multilingual-e5-large	24	57.29	65.86	64.34	48.99	50.53	81.81
ipipan/silver-retriever-base-v1	14	53.32	60.87	61.89	37.18	52.64	81.18
intfloat/multilingual-e5-base	0	53.11	60.16	59.06	44.01	48.38	80.18
intfloat/multilingual-e5-small	1	50.65	57.84	53.80	42.45	46.77	79.72

https://huggingface.co/spaces/sdadas/pirb

Acknowledgments

- Many thanks to Sławek Dadas for the translation of MSMARCO into polish, and expertise support
- Many thanks to Microsoft for getting me limited access preview to the Azure OpenAI Service as GPT3 fine tuning endpoints etc.

OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Thank you

Ośrodek Przetwarzania Informacji-
Państwowy Instytut Badawczy
al. Niepodległości 188 B
00-608 Warszawa

tel.: +48 22 570 14 00
faks: +48 22 825 33 19
e-mail: opi@opi.org.pl
www.opi.org.pl