MULTI-INDEX RETRIEVE AND RERANK WITH SEQUENCE-TO-SEQUENCE MODEL

Konrad Wojtasik, Wrocław University of Science and Technology, Clarin-PL

FedCSIS 2023, PolEval Task3: Passage Retrieval

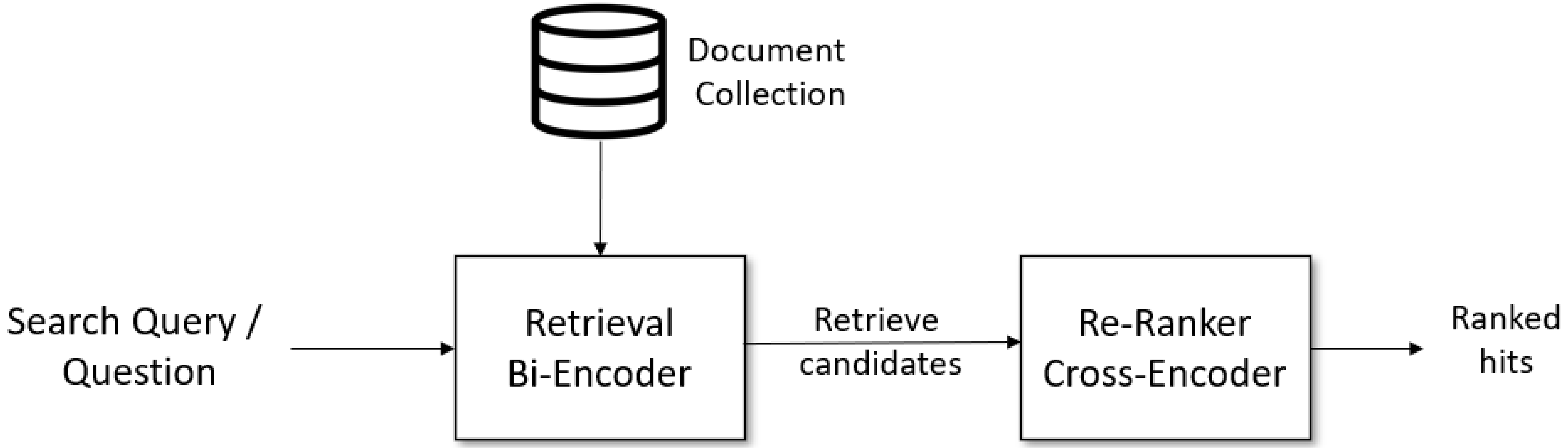
POLEVAL TASK3: PASSAGE RETRIEVAL

Goal of the task was to **retrieve** relevant **passages** corresponding to the **query** from the **passage collection**. Queries were questions and passages were text fragments containing the answer.

Solution: Create effective Information Retrieval pipeline

INFORMATION RETRIEVAL

Information Retrieval (IR) is a process of obtaining relevant information form a collection of documents. User input is a query and the output of the system is the relevant documents, passages.



https://www.sbert.net/examples/applications/information-retrieval/README.html

DATASETS

3 domains: Wikipedia based questions, allegro customer questions, legal questions

No training data for allegro and legal questions domain.

Testsets:

- wiki-trivia: 7M passages, 400 queries in Test A, 891 queries in Test B
- allegro-faq: 921 passages, 400 queries in Test A, 500 queries in Test B
- legal-questions: 26287 passages, 400 queries in Test A, 318 queries in Test B

AVAILABLE DATASETS

- MSMARCO – large dataset created by Microsoft from Bing search questions,
- BEIR benchmark – contains diverse range of IR datasets

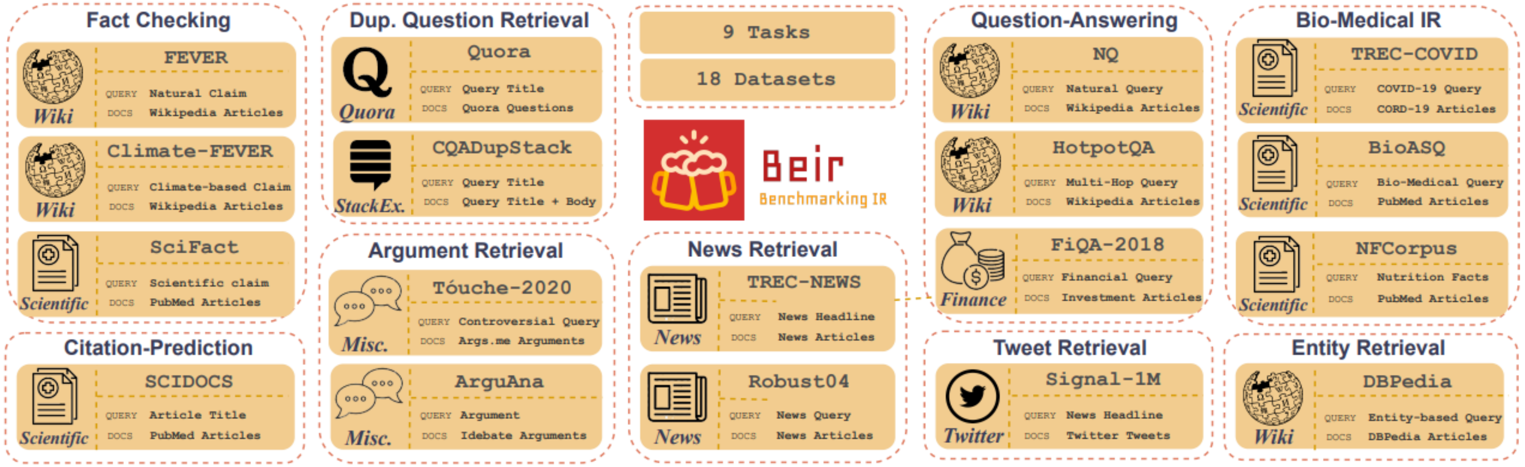
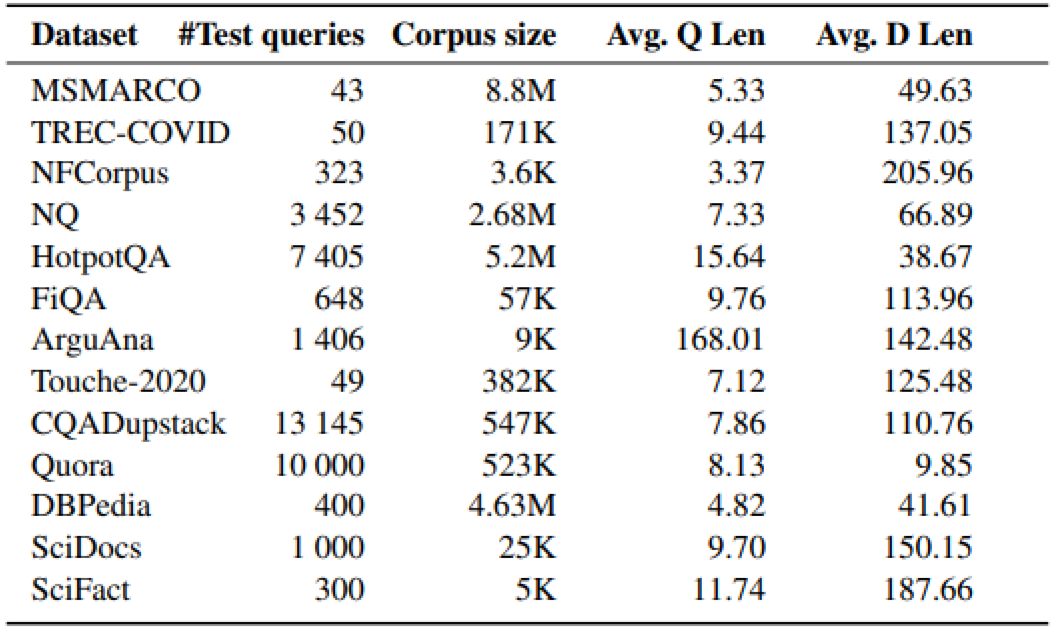


Figure 1: An overview of the diverse tasks and datasets in BEIR benchmark.

BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models

AVAILABLE DATASETS

Number of queries and size of corpuses in BEIR and **BEIR-PL** benchmarks.



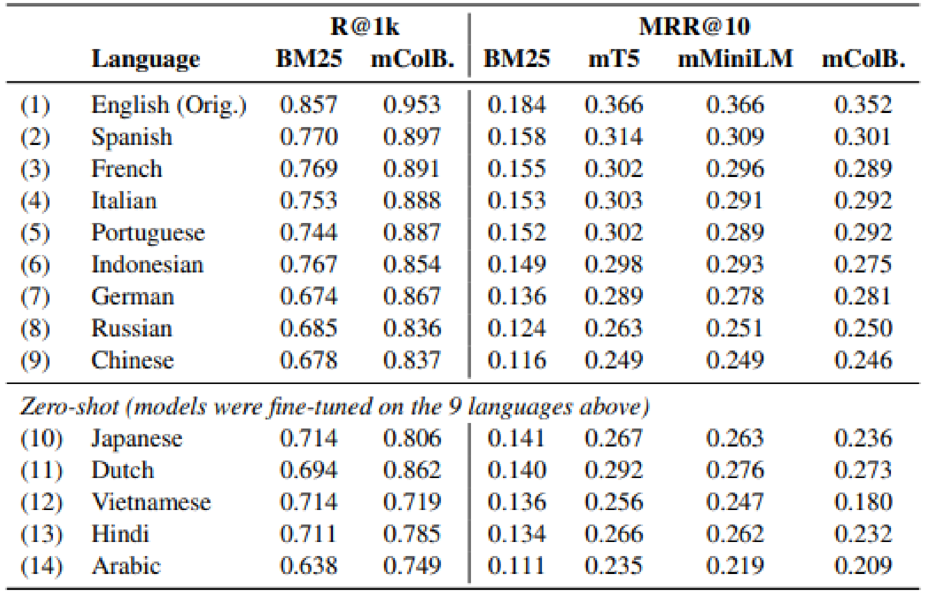
Dataset	#Test queries	Corpus size	Avg. Q Len	Avg. D Len
MSMARCO	43	8.8M	5.33	49.63
TREC-COVID	50	171K	9.44	137.05
NFCorpus	323	3.6K	3.37	205.96
NQ	3 452	2.68M	7.33	66.89
HotpotQA	7 405	5.2M	15.64	38.67
FiQA	648	57K	9.76	113.96
ArguAna	1 406	9K	168.01	142.48
Touche-2020	49	382K	7.12	125.48
CQADupstack	13 145	547K	7.86	110.76
Quora	10 000	523K	8.13	9.85
DBPedia	400	4.63M	4.82	41.61
SciDocs	1 000	25K	9.70	150.15
SciFact	300	5K	11.74	187.66

BEIR-PL: Zero Shot Information Retrieval Benchmark for the Polish Language

AVAILABLE DATASETS

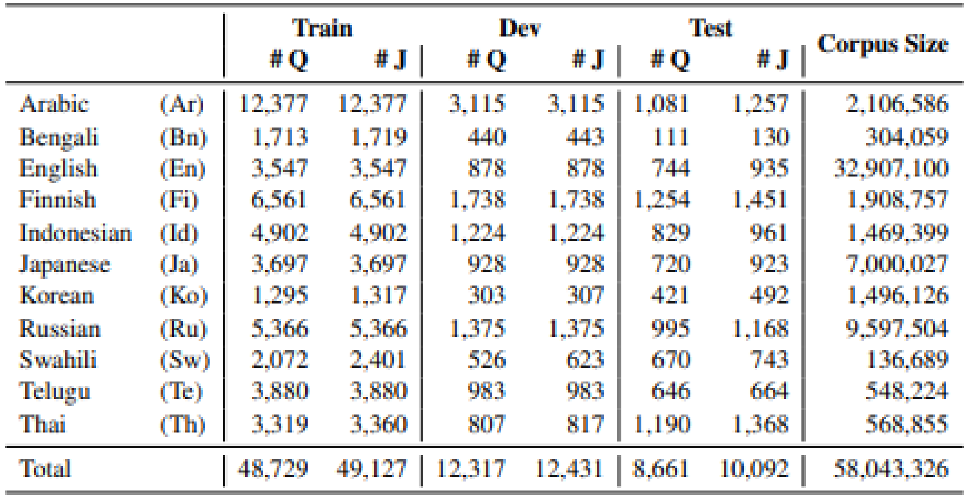
Multilingual:

- mMARCO - Automatic translation of original MS MARCO dataset to 13 languages
- Mr. TyDi – Dataset constructed from TyDI QA dataset



		R@1k		MRR@10			
	Language	BM25	mColB.	BM25	mT5	mMiniLM	mColB.
(1)	English (Orig.)	0.857	0.953	0.184	0.366	0.366	0.352
(2)	Spanish	0.770	0.897	0.158	0.314	0.309	0.301
(3)	French	0.769	0.891	0.155	0.302	0.296	0.289
(4)	Italian	0.753	0.888	0.153	0.303	0.291	0.292
(5)	Portuguese	0.744	0.887	0.152	0.302	0.289	0.292
(6)	Indonesian	0.767	0.854	0.149	0.298	0.293	0.275
(7)	German	0.674	0.867	0.136	0.289	0.278	0.281
(8)	Russian	0.685	0.836	0.124	0.263	0.251	0.250
(9)	Chinese	0.678	0.837	0.116	0.249	0.249	0.246
Zero-shot (models were fine-tuned on the 9 languages above)							
(10)	Japanese	0.714	0.806	0.141	0.267	0.263	0.236
(11)	Dutch	0.694	0.862	0.140	0.292	0.276	0.273
(12)	Vietnamese	0.714	0.719	0.136	0.256	0.247	0.180
(13)	Hindi	0.711	0.785	0.134	0.266	0.262	0.232
(14)	Arabic	0.638	0.749	0.111	0.235	0.219	0.209

mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset



		Train		Dev		Test		Corpus Size
		# Q	# J	# Q	# J	# Q	# J	
Arabic	(Ar)	12,377	12,377	3,115	3,115	1,081	1,257	2,106,586
Bengali	(Bn)	1,713	1,719	440	443	111	130	304,059
English	(En)	3,547	3,547	878	878	744	935	32,907,100
Finnish	(Fi)	6,561	6,561	1,738	1,738	1,254	1,451	1,908,757
Indonesian	(Id)	4,902	4,902	1,224	1,224	829	961	1,469,399
Japanese	(Ja)	3,697	3,697	928	928	720	923	7,000,027
Korean	(Ko)	1,295	1,317	303	307	421	492	1,496,126
Russian	(Ru)	5,366	5,366	1,375	1,375	995	1,168	9,597,504
Swahili	(Sw)	2,072	2,401	526	623	670	743	136,689
Telugu	(Te)	3,880	3,880	983	983	646	664	548,224
Thai	(Th)	3,319	3,360	807	817	1,190	1,368	568,855
Total		48,729	49,127	12,317	12,431	8,661	10,092	58,043,326

Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval

METRICS

Recall (Recall@k) cut off at the k ranking position. The recall@k informs how many relevant documents from the collection were classified to @k ranking.

$$recall = \frac{|relevant| \cap |retrieved|}{|relevant|}$$

METRICS

Normalised Cumulative Discount Gain(nDCG@k) – reported in the original BEIR benchmark. NCDG@k measures the quality of ranking considering all relevant passages and its position in @k retrieved documents,

$$NDCG@k = \frac{\sum_{i=1}^{k(rank\_order)} \frac{Gain}{log_2(i+1)}}{\sum_{i=1}^{k(real\_order)} \frac{Gain}{log_2(i+1)}},$$

RETRIEVERS – BM25 BASELINE

BM25 is a ranking function used in information retrieval systems to estimate the relevance of a document to a given query. It uses the relevance score of a document based on the query terms and the document's content.

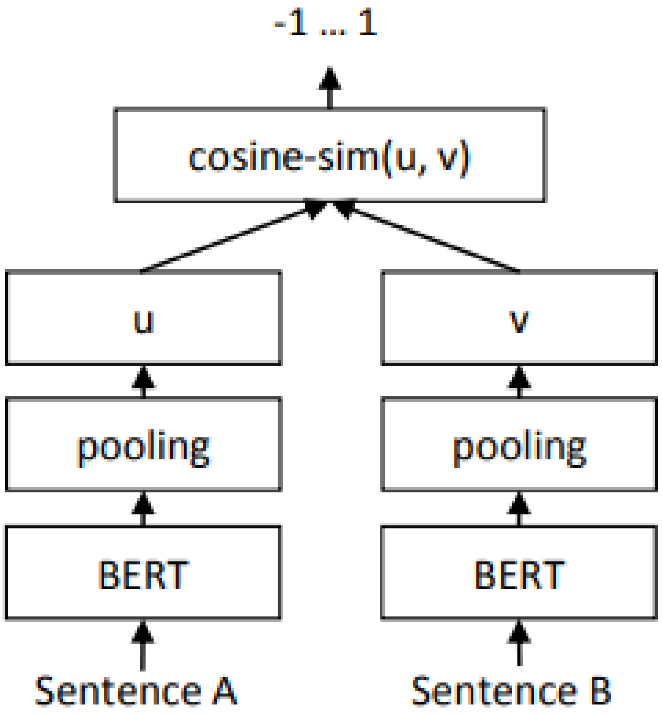
$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

RETRIEVERS - BIENCODER ARCHITECTURE

Get representation of the sentences from the model and calculate similarity with dot product or cosine similarity.

We have to get the representation of the whole sentence and we want to store it in one vector, that is why pooling is needed.

Pooling: CLS or MEAN or MAX(not used right now)



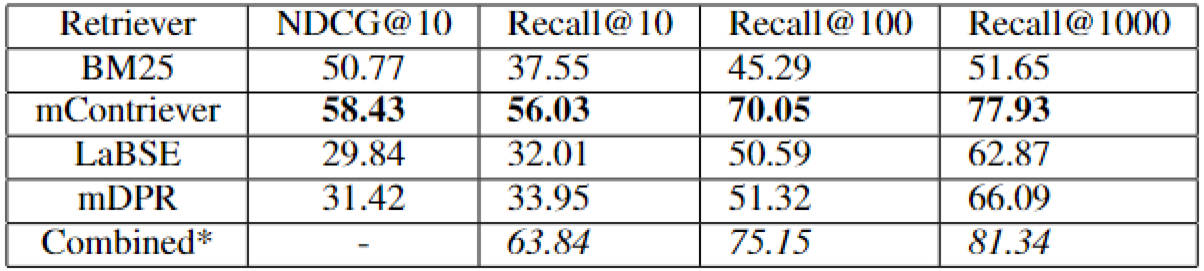
Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

RETRIEVERS

Retrievers:

- Elasticsearch BM25
- LaBSE - language-agnostic BERT model for sentence embedding
- mDPR-question-nq and mDPR-passage-nq - bi-encoder where query and passage are encoded with different encoders trained in contrastive manner
- mContriever-base-msmarco - mBERT based retriever trained in unsupervised manner with contrastive loss on multilingual data and afterward fine-tuned on English MS MARCO dataset.

RETRIEVERS



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*

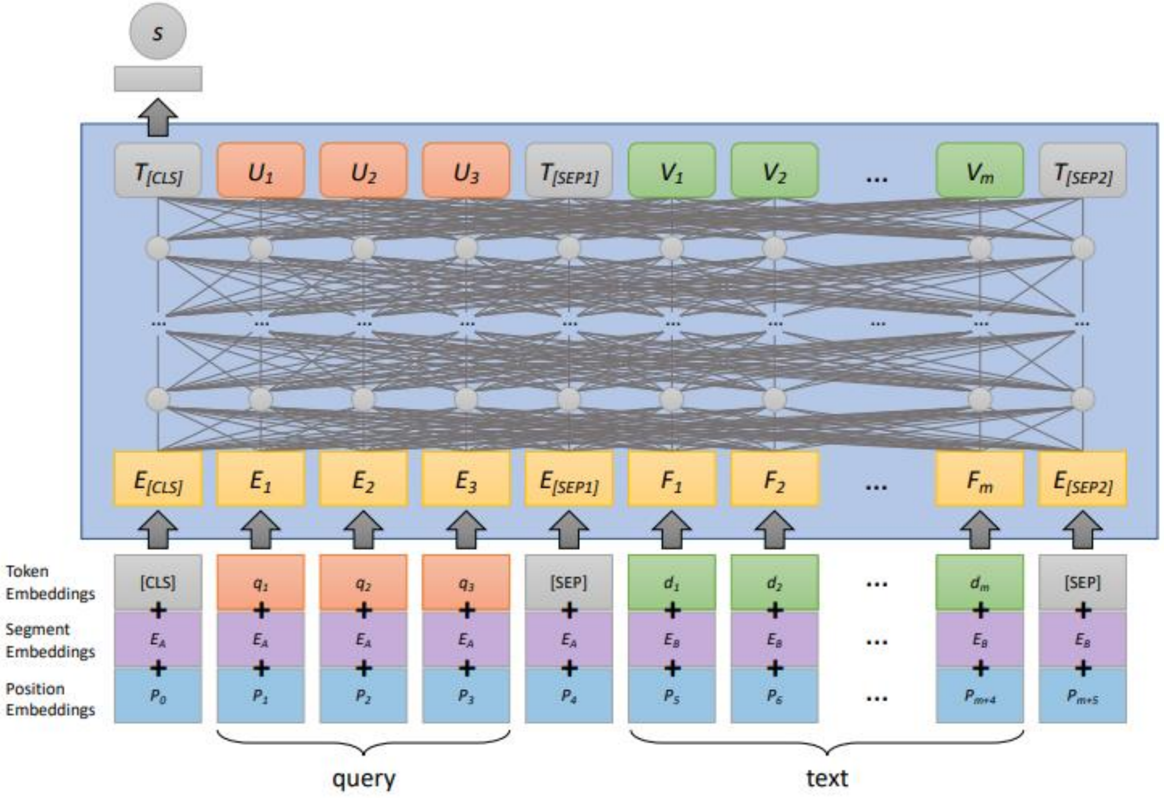
Combined* is a set of all results from all retrievers.

RERANKING – BERT MODEL

The **monoBERT** ranking model adapts BERT for sequence classification task.

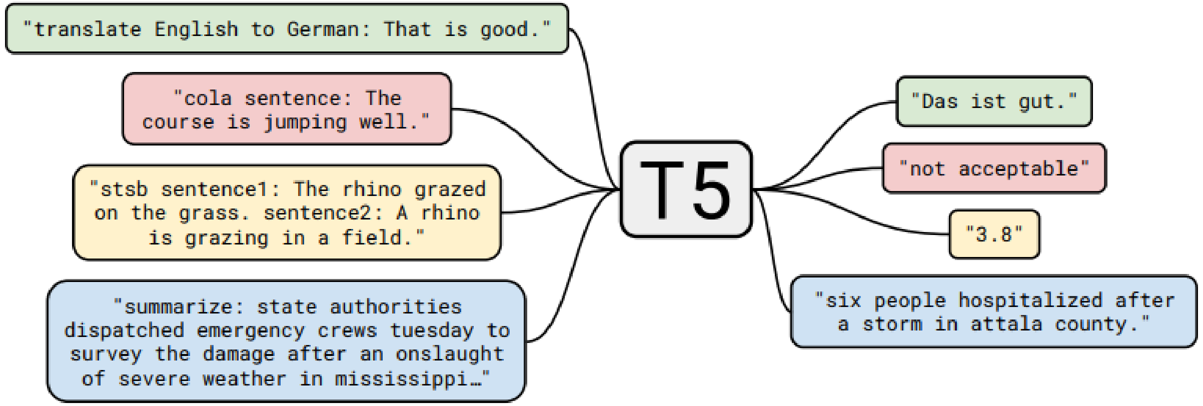
The input is a query and a candidate text separated with [SEP] token.

The final representation of the [CLS] token is fed to a fully-connected layer that produces the relevance score of the text with respect to the query.



Pretrained Transformers for Text Ranking: BERT and Beyond

RERANKING – T5 MODEL


"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

We can also use sequence-to-sequence models for reranking.

monoT5 – extra tokens are added. One represents **true**, when Query and Document are relevant, or **false** they are not. Model is trained to predict the token based on input.

Example input: Query: [q] Document: [d] Relevant:

Example of new tokens: '_true' , '_false'

Document Ranking with a Pretrained Sequence-to-Sequence Model

RERANKERS

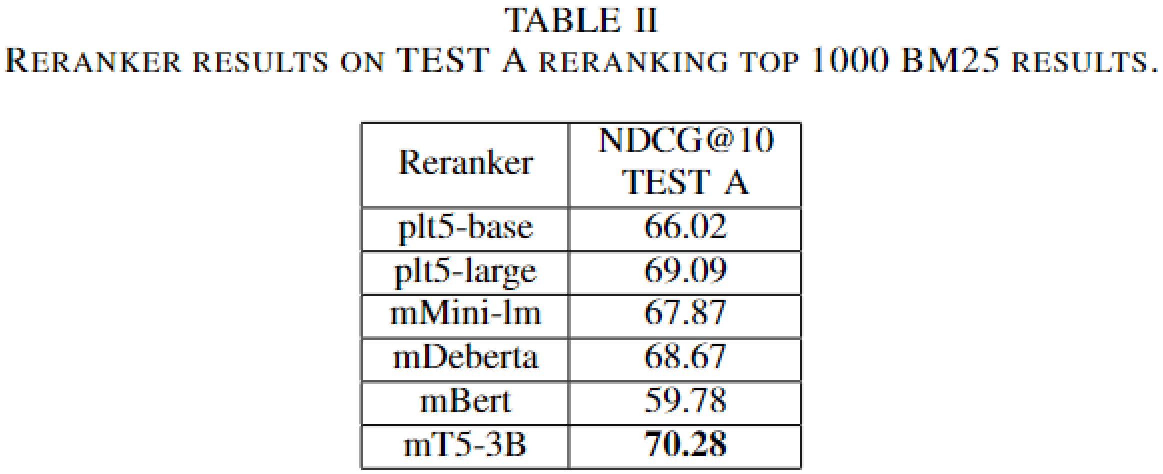


TABLE II
RERANKER RESULTS ON TEST A RERANKING TOP 1000 BM25 RESULTS.

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**

Tested reranker models:

- mMiniLMv2 – mMiniLMv2 trained on multilingual MS Marco
  - nreimers/mmarco-mMiniLMv2-L12-H384-v1)
- mDeBERTa – mDeBERTa trained on multilingual MS Marco (not on huggingface anymore)
- mBERT – mBERT trained as reranker on MS Marco
  - amberoad/bert-multilingual-passage-reranking-msmarco
- plT5 – polish T5 model trained on MS Marco translated to polish from BEIR-PL
  - clarin-knext/plt5-large-msmarco
- mT5 – multilingual T5 model trained on multilingual MS Marco dataset
  - unicamp-dl/mt5-13b-mmarco-100k
  - unicamp-dl/mt5-3B-mmarco-en-pt

FINAL SOLUTION

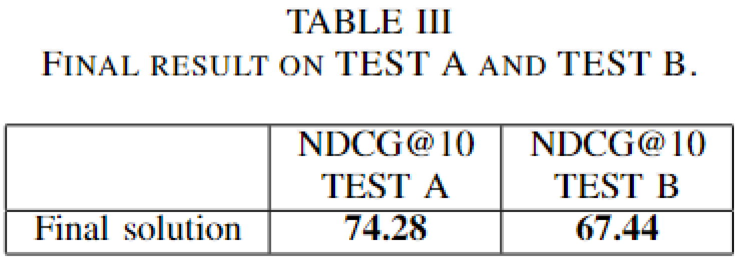


TABLE III
FINAL RESULT ON TEST A AND TEST B.

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

The final solution:

- retrieve the top 100 passages for each query from each dense retriever
- retrieve top 1000 passages from BM25 index and rerank with plT5-large model
- Combine top 100 passages from each dense retriever and top 100 reranked results from BM25 index and create a set of passages
- reranked obtained set with 13B mT5 model

NOTES&IMPROVEMENTS

- Retrieval and reranking is highly computationally intensive task
- Larger rerankers got better performance, but it is hard to fine-tune larger models (mT5-13B model)
- Retrieval part has to have a high recall, so correct passages are not omitted

Improvements:

- Retrieve more passages for reranking, especially from wiki-trivia domain
- Use additional reranking with smaller model on larger number of retrieved documents

THANK YOU FOR YOUR ATTENTION!

Any questions?

REFERENCES

1. Document Ranking with a Pretrained Sequence-to-Sequence Model
2. Pretrained Transformers for Text Ranking: BERT and Beyond
3. Unsupervised Dense Information Retrieval with Contrastive Learning
4. Dense Passage Retrieval for Open-Domain Question Answering
5. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
6. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
7. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval
8. mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset
9. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models
10. BEIR-PL: Zero Shot Information Retrieval Benchmark for the Polish Language