ADAM MICKIEWICZ UNIVERSITY, POZNAŃ

Passage Retrieval of Polish Texts Using OKAPI BM25 and an Ensemble of Cross Encoders

Jakub Pokrywka

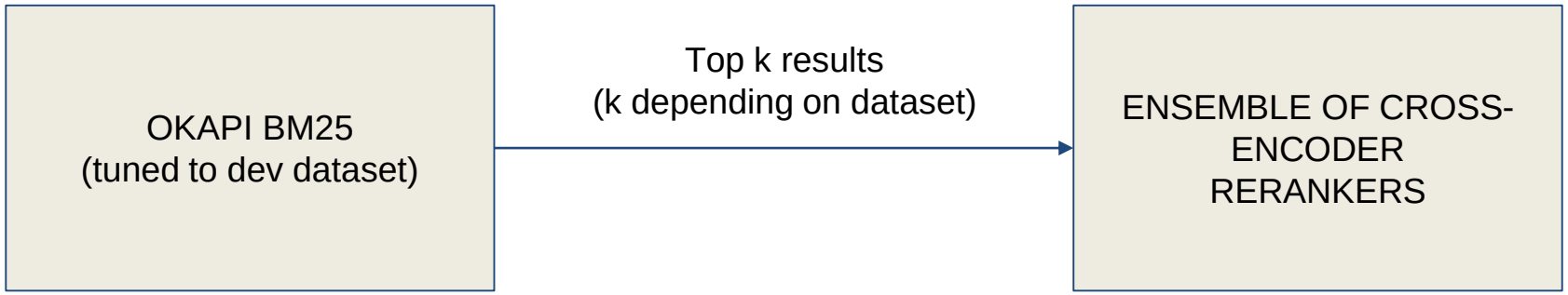
Competition dataset

-	wiki-trivia	legal-questions	allegro-faq
train questions	4401	0	0
dev questions	599	0	0
test-A questions	400	400	400
mean test-A rel. passages	3.46	1.97	1.09
test-B questions	891	318	500
mean test-B rel. passages	3.39	2.03	1.05
passages	7097322	26287	921
mean word per passage	44.6	155.1	50.0

mMARCO dataset and models

- MS MARCO (English only) - over half a million questions from Bing's search query logs
- mMARCO - translation into 14 languages (but not including Polish)
- Some multilingual models trained on mMARCO

My final method



Retriever

Okapi BM25 tuned to wiki dataset:

- k_1 = 1.2, b = 0.75, epsilon=0.25
- Word tokenization with NLTK
- Lowercase normalization
- Stemming with pystempel with Polimorf
- Removal of Polish stopwords

Rerankers

wiki-trivia- reranking of 3000 positions (super slow even on A100):

- no-ft unicamp-dl/mt5-13b-mmarco-100k
- ft cross-encoder/mmarco-mMiniLMv2-L12-H384-v1
- ft cross-encoder/mmarco-mdeberta-v3-base-5negs-v1 (available during the competition, but withdrawn from huggingface now)

legal questions (1500 positions), allegro-faq (all 921 positions)

- no-ft unicamp-dl/mt5-13b-mmarco-100k
- no-ft unicamp-dl/mt5-3B-mmarco-en-pt

model	test-B	wiki-trivia	legal-questions	allegro-faq
final ensemble	69.36	55.13	86.39	83.88
OKAPI BM25	42.55	23.48	81.31	51.87
mmarco-mMiniLMv2-L12-H384-v1 no-ft 10	48.85	28.45	83.00	63.47
mt5-3B-mmarco no-ft 10	50.31	29.47	84.35	65.81
mt5-13B-mmarco no-ft 10	50.36	29.63	83.59	66.15
mmarco-mMiniLMv2-L12-H384-v1 no-ft 50	56.18	35.88	85.26	73.84
mt5-3B-mmarco no-ft 50	59.04	38.06	86.75	78.80
mt5-13B-mmarco no-ft 50	59.79	39.30	85.30	80.08
mmarco-mMiniLMv2-L12-H384-v1 no-ft 100	57.76	38.22	85.54	74.91
mt5-3B-mmarco no-ft 100	61.42	41.24	87.06	81.09
mt5-13B-mmarco no-ft 100	62.65	43.17	85.63	82.75
mmarco-mMiniLMv2-L12-H384-v1 no-ft 500	58.52	39.86	85.61	74.56
mt5-3B-mmarco no-ft 500	63.48	44.41	86.67	82.70
mt5-13B-mmarco no-ft 500	65.04	47.21	85.42	83.86
mmarco-mMiniLMv2-L12-H384-v1 no-ft 1000	58.91	40.49	85.66	74.72
mt5-3B-mmarco no-ft 1000	64.12	45.48	86.64	83.01
mt5-13B-mmarco no-ft 1000	65.59	48.13	85.22	84.21
mmarco-mMiniLMv2-L12-H384-v1 no-ft 1500	58.99	40.70	85.51	74.72
mmarco-mMiniLMv2-L12-H384-v1 ft 1500	-	47.64	-	-
mmarco-mdeberta-v3-base-5negs-v1 no-ft 1500	-	45.30	-	-
mmarco-mdeberta-v3-base-5negs-v1 ft 1500	-	51.73	-	-
mt5-3B-mmarco no-ft 1500	64.46	46.17	86.55	83.01
mt5-13B-mmarco no-ft 1500	65.99	48.96	85.04	84.21

Other experiments

- Translating Polish into English and using English Cross-Encoder
- Bi Encoder models (but did not try Contriever)
- Translating MS MARCO into Polish and training HerBERT

Conclusions

- The OKAPI BM25 is a highly effective retrieval system, particularly when properly optimized.
- Increasing the number of reranked items improves NDCG, but beyond 500 positions, the improvement is minimal.
- The above conclusion contradicts my other research: Reranking for a Polish Medical Search Engine
- Finetuning the model for a specific domain improves performance slightly within that domain but reduces performance in other domains (~3-6 NDCG@10).

- Trying out Contriever and comparing it to the well-tuned OKAPI BM25 baseline as a retrieval model in the future would be beneficial.

Thanks for your attention

jakub.pokrywka@amu.edu.pl