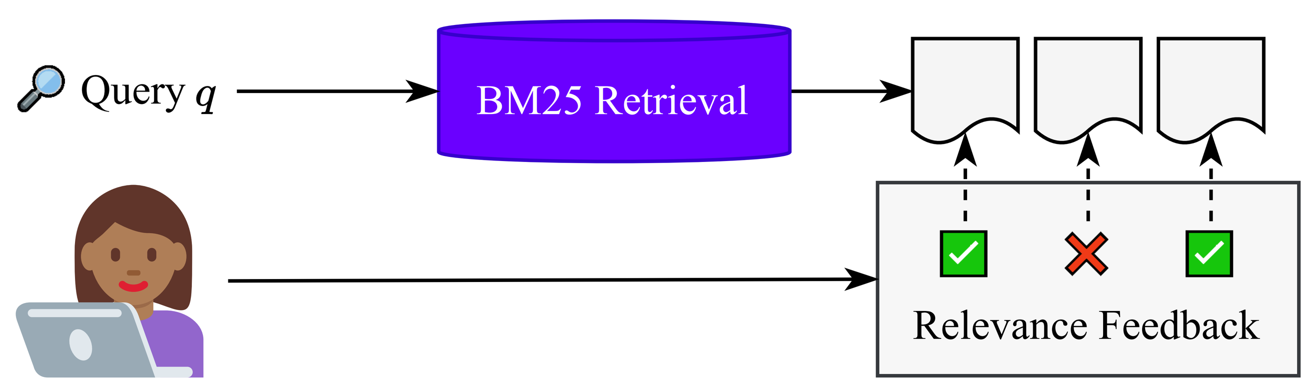
Incorporating Relevance Feedback for Information-Seeking Retrieval using Few-Shot Document Re-Ranking

Tim Baumgärtner,[1] Leonardo F. R. Ribeiro,[1] Nils Reimers,[2] Iryna Gurevych[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab), Technical University of Darmstadt, [2]cohere.ai

For information-seeking retrieval scenarios, we propose to fine-tune a re-ranking model using few-shot learning based on data from relevance feedback. We transform four existing information retrieval datasets into this setup by simulating a user. We evaluate a kNN approach and fine-tune a Cross-Encoder model per query. Despite recent advances, we show that BM25 with query expansion is a tough baseline to beat. Our final model, however, is able to outperform it across all 4 datasets by 5.2% nDCG@20 on average.

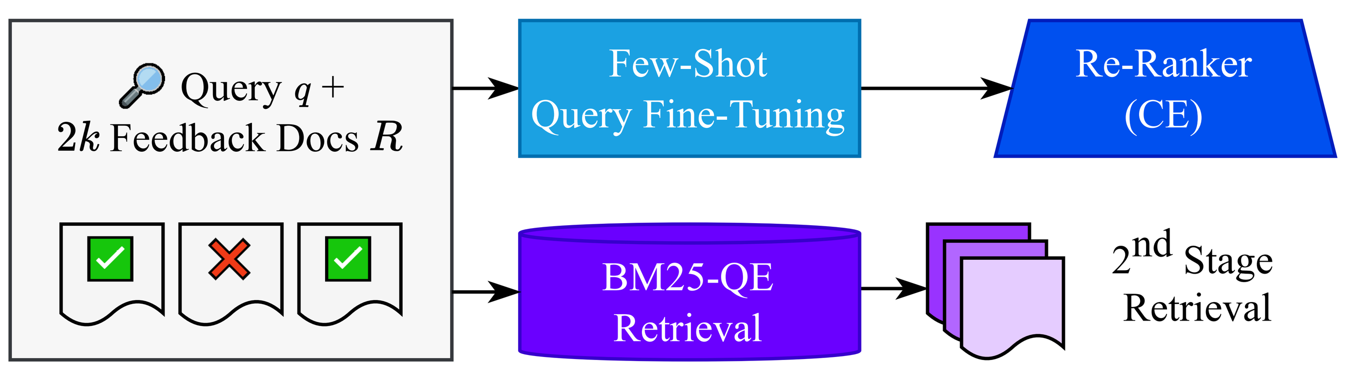
Task Setup

1) First Stage Retrieval & Relevance Feedback



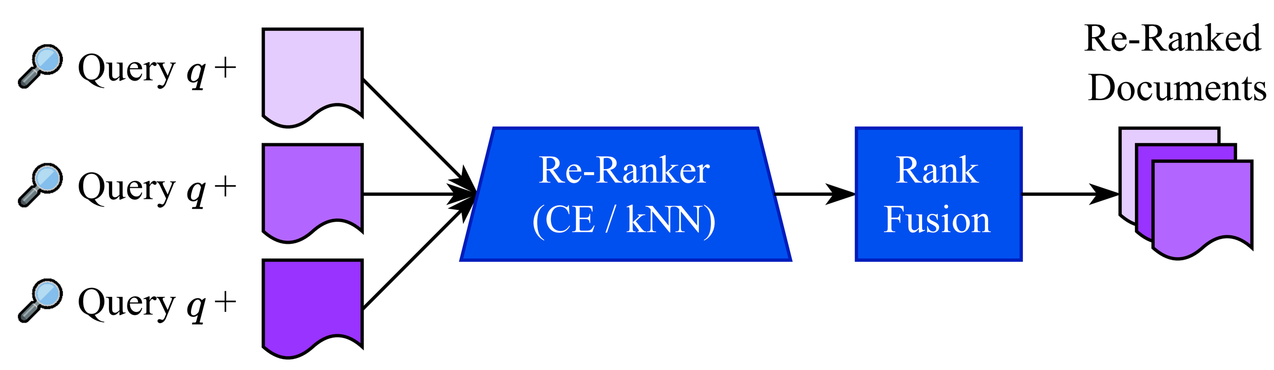
Retrieve first set of documents and annotate the top k relevant and non-relevant documents R. We use annotations from existing datasets to simulate a user.

2) Query Expansion & Re-Ranker Fine-Tuning



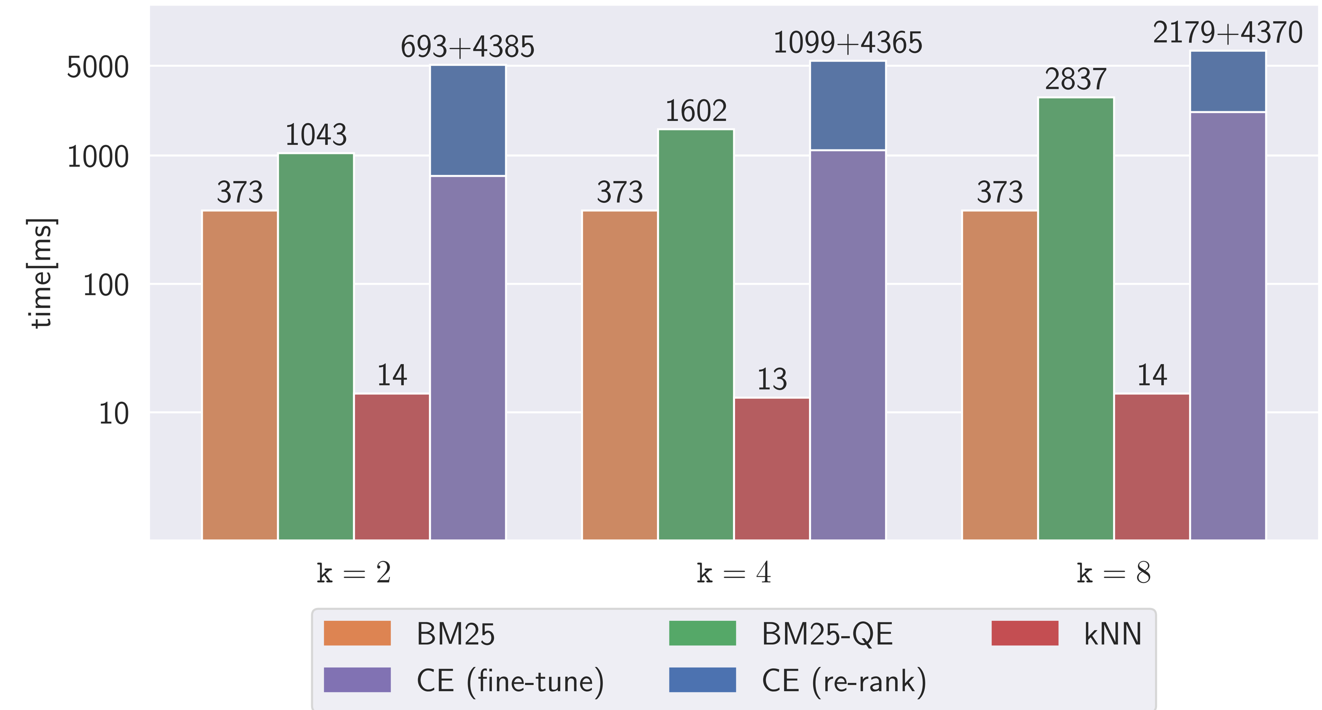
Using the feedback documents R, (1) fine-tune the re-ranking model and (2) expand the query to obtain the 2^{nd} stage retrieval documents.

3) Re-Ranking



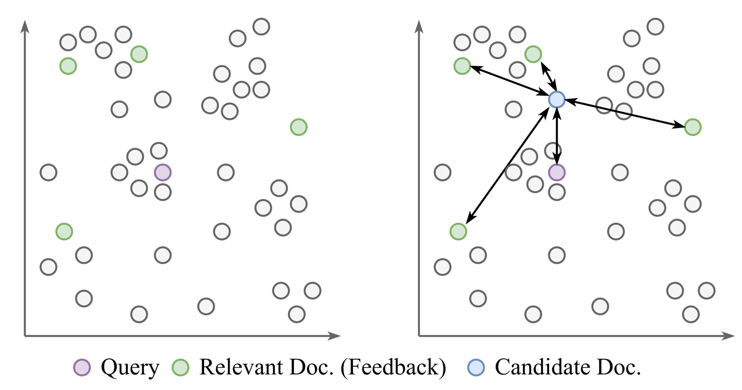
Re-rank 2^{nd} stage retrieval documents using the fine-tuned re-ranker from Phase 2. Optionally fuse ranks between BM25-QE and re-ranker.

Retrieval, Re-Ranking and Fine-tuning Latency



Average time in milliseconds for retrieval and re-ranking 1000 documents .

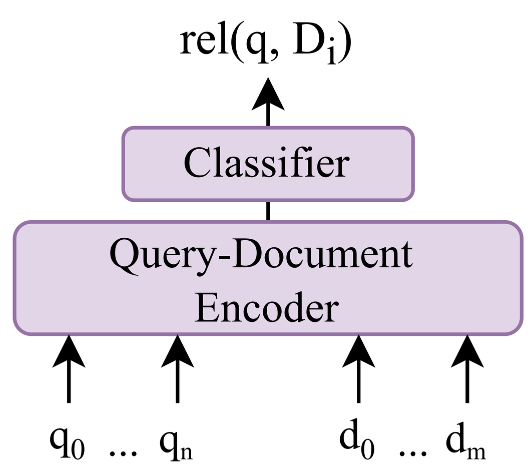
Method: kNN



1. Pre-compute document representations for all $d_i \in D$ (offline)
2. Score documents by summing the similarities between the candidate d_i, query q and relevant documents $d_j \in R^+$

$$s_i = f(d_i, q) + \sum_{d_j \in R^+} f(d_i, d_j)$$

Method: CE Query Fine-Tuning



CE Query-FT
Fine-tune bias layers per query on $2k$ feedback documents

CE MAML + Query-FT

1. Fine-tune bias layers on the training set with MAML to obtain "fast parameters"
2. Fine-tune bias layers per query on the $2k$ feedback documents

Method: Rank Fusion

$$s_i = \sum_{h \in H} \frac{1}{c + h(d_i)}$$

Given rankings from different models, merge their results according to the individual ranks.

Results

Method	Robust	Covid	News	Touché	Avg.
BM25-QE (2^{nd} Stage Retrieval)	0.496	0.610	0.392	0.271	0.442
kNN	0.443	0.686	0.365	0.174	0.417
CE Zero-Shot	0.415	0.702	0.314	0.176	0.402
CE Query FT	0.484	0.723	0.335	0.198	0.435
CE MAML + Query-FT	0.506	0.735	0.314	0.223	0.445
BM25-QE ∩ kNN	0.507	0.707	**0.412**	0.248	0.468
BM25-QE ∩ CE MAML + Query-FT	**0.570**	**0.740**	0.405	**0.272**	**0.497**

nDCG@20 results averaged over $k = \{2, 4, 8\}$.

⇒ kNN and CE Zero-Shot cannot outperform BM25-QE

⇒ Fine-tuning only on 2k datapoints works, and MAML additionally helps

⇒ Rank-fusion is highly effective and complementary

⇒ kNN is fast & query fine-tuning takes only a fraction of the overall time