Incorporating Relevance Feedback for Information-Seeking Retrieval using Few-Shot Document Re-Ranking
Under Review @ EMNLP

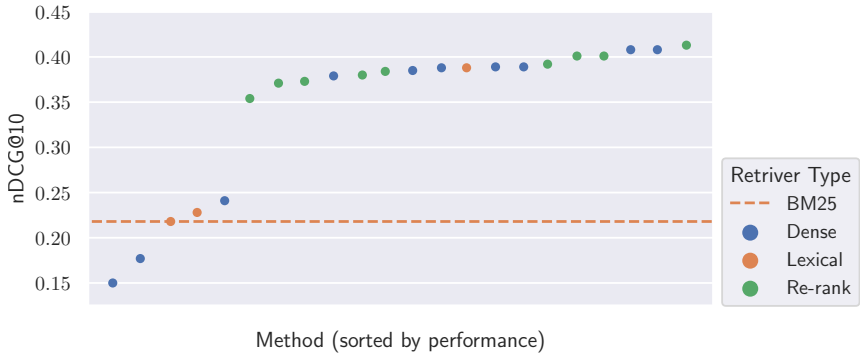
TECHNISCHE UNIVERSITÄT DARMSTADT

Tim Baumgärtner[1], Leonardo F. R. Ribeiro[1], Nils Reimers[2], Iryna Gurevych[1]
[1]Ubiquitous Knowledge Processing Lab, [2]Hugging Face

UKP

Motivation

Lexical vs. Neural Retrieval



nDCG@10 performance on MS MARCO Passage Retrieval Nguyen et al., 2016. Results from BEIR Thakur et al., 2021.

Motivation

Query Types Broder, 2002

TECHNISCHE UNIVERSITÄT DARMSTADT

	Navigational / Transactional	Information-Seeking
Query	”arxiv sentence bert”, ”acl paper submission”	”origin coronavirus”, ”neural sentence representation”
# Relevant	few	many
Scenarios	Navigation Know-Item Retrieval	Scientific Literature Review News Background Retrieval Case Law Retrieval Factoid QA

Motivation

Query Types Broder, 2002

TECHNISCHE UNIVERSITÄT DARMSTADT

	Navigational / Transactional	**Information-Seeking**
Query	”arxiv sentence bert”, ”acl paper submission”	”origin coronavirus”, ”neural sentence representation”
# Relevant	few	many
Scenarios	Navigation Know-Item Retrieval	Scientific Literature Review News Background Retrieval Case Law Retrieval Factoid QA

Challenges: Unknown Search Domain & Query Formulation

Background

Relevance Feedback in Lexical Retrieval

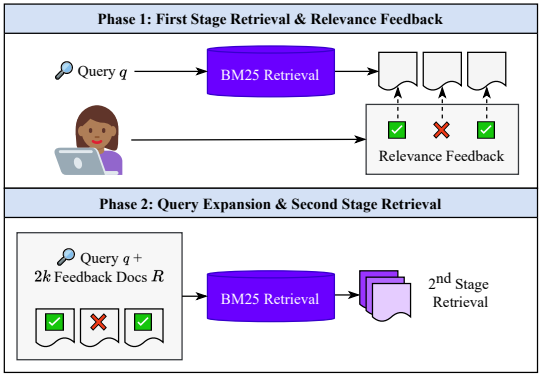
TECHNISCHE UNIVERSITÄT DARMSTADT

Phase 1

1. Retrieve documents using the query
2. Obtain pseudo/implicit/explicit relevance feedback on retrieved Documents

Phase 2

1. Extract additional ”expansion terms” from relevance documents
2. Retrieve documents with query + terms



UKP

Background

Relevance Feedback in Lexical Retrieval

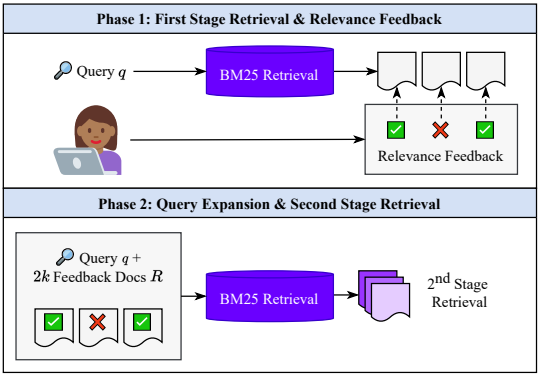
TECHNISCHE UNIVERSITÄT DARMSTADT

Phase 1

1. Retrieve documents using the query
2. Obtain pseudo/implicit/explicit relevance feedback on retrieved Documents

Phase 2

1. Extract additional ”expansion terms” from relevance documents
2. Retrieve documents with query + terms



How to integrate relevance feedback in neural retrieval?

UKP

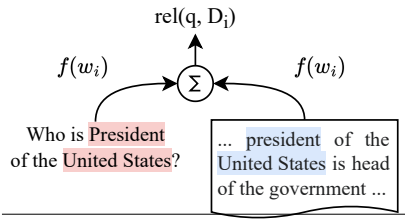
Background

Lexical & Neural Retrieval Methods

TECHNISCHE UNIVERSITÄT DARMSTADT

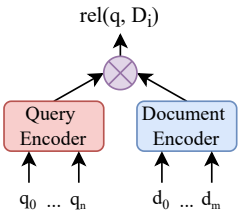
Lexical Retrieval: BM25 Robertson and Zaragoza, 2009

Treat query & documents as Bag of Words and compute document scores via sum of weighted lexical matches



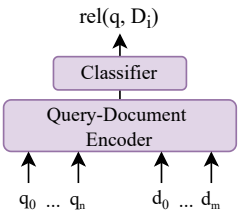
Dense Retrieval Karpukhin et al., 2020

Encode query & documents separately and compute scores between representations



Neural Re-Ranking Nogueira and Cho, 2019

Conduct initial retrieval to get document candidates and re-score by encoding query and document jointly



UKP

Setup

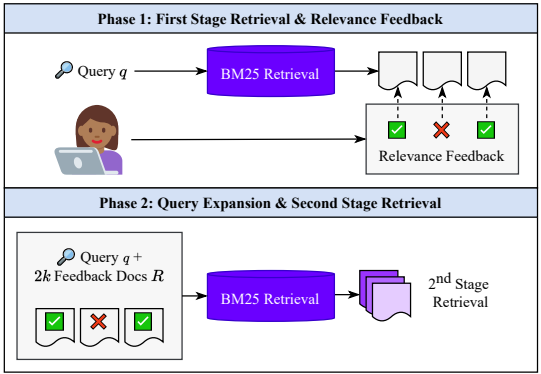
Goal: Re-rank documents from 2nd stage retrieval incorporating relevance feedback.

Given

- Query *q*
- *k* relevant & *k* non-relevant feedback documents, where $k \in \{2, 4, 8\}$
- $n = 1000$ documents obtained from BM25 Query Expansion

Evaluation

- Ranking: nDCG@20 Järvelin and Kekäläinen, 2000 with varying *k*
- Latency: [ms]

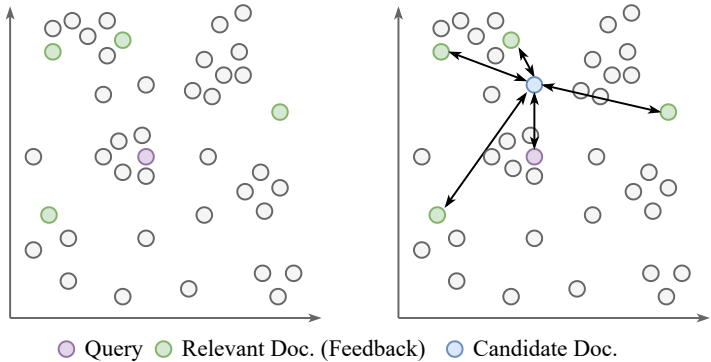


Method

Re-Ranking with kNN

- Compute document representations $d_i \in D$
- Score documents by summing the similarity between query representation q and relevant feedback documents $d_j \in R^+$

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



Model: sentence-transformers/all-MiniLM-L6-v2

Method

Re-Ranking with Cross-Encoder

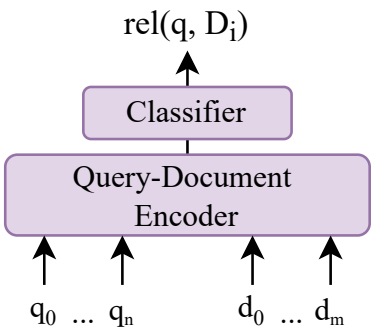
TECHNISCHE UNIVERSITÄT DARMSTADT

Zero-Shot

- Use model without any fine-tuning

Few-Shot

- **CE Query-FT**: Fine-Tune all bias layers per query on $2k$ Feedback Documents
- **CE MAML + Query-FT**:
  1. Fine-Tune bias layers on in-domain annotations with Meta-Learning to obtain "fast parameters"
  2. Fine-Tune per query on on $2k$ Feedback



Model: cross-encoder/ms-marco-MiniLM-L-6-v2

UKP

Method

Reciprocal Rank Fusion Cormack et al., 2009

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Idea**: Merge rankings from different methods
- **Problem**: Methods produce different scores, simple adding is biased
- **Solution**: Use ranks instead of raw scores

$$s_i = \sum_{g \in G} \frac{1}{c + g(d_i)}$$

$c = 60$ (constant),

$g =$ ranking function of a method

UKP

Method

Reciprocal Rank Fusion Cormack et al., 2009

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Idea**: Merge rankings from different methods
- **Problem**: Methods produce different scores, simple adding is biased
- **Solution**: Use ranks instead of raw scores

$$s_i = \sum_{g \in G} \frac{1}{c + g(d_i)}$$

$c = 60$ (constant),

$g =$ ranking function of a method

$\Rightarrow$ The better the rank ($g(d_i)$ is small), the higher the score. Smoothed by c.

Datasets

TECHNISCHE UNIVERSITÄT DARMSTADT

Dataset	Domain	Docs	Queries	Judgments
Robust04 E. M. Voorhees et al., 2004	News	528k	148	1287.14 (±501)
TREC-Covid E. Voorhees et al., 2021	Biomedical	191k	50	1370.36 (±323)
TREC-News Soboroff et al., 2018	News	595k	34	258.85 (±82)
Webis-Touché Bondarenko et al., 2021	Debates	383k	49	49.76 (±7)

Includes only queries with at least 32 relevant documents.

Results

Re-ranking Performance

TECHNISCHE UNIVERSITÄT DARMSTADT

Method	Robust	Covid	News	Touché	Avg.
BM25-QE	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 performance. Results averaged over $k \in \{2, 4, 8\}$ feedback documents

Results

Re-ranking Performance

TECHNISCHE UNIVERSITÄT DARMSTADT

Method	Robust	Covid	News	Touché	Avg.
BM25-QE	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**

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

Results

Re-ranking Performance

TECHNISCHE UNIVERSITÄT DARMSTADT

Method	Robust	Covid	News	Touché	Avg.
BM25-QE	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**

$\Rightarrow$ Fine-Tuning only on $2k$ datapoints works, and MAML additionally helps.

Results

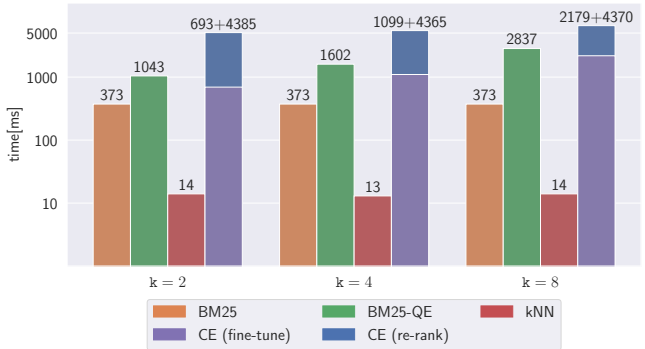
Re-ranking Performance

TECHNISCHE UNIVERSITÄT DARMSTADT

Method	Robust	Covid	News	Touché	Avg.
BM25-QE	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**

⇒ Rank-Fusion is highly effective and complementary.

Results
Re-ranking Latency



⇒ kNN is extremely fast, everything can be precomputed.

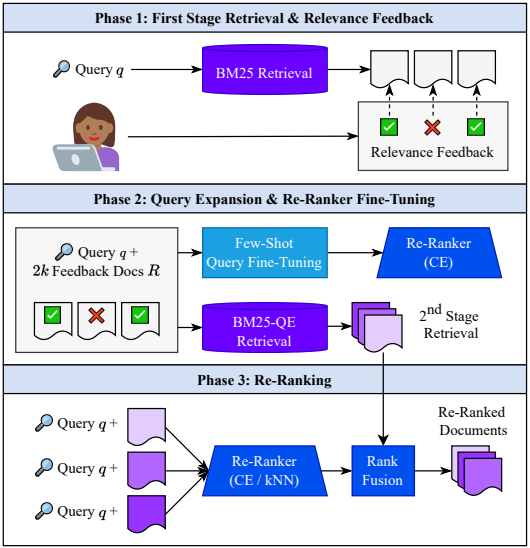
⇒ Fine-Tuning CE only takes a fraction of the time compared with re-ranking.

Results

Ablations

TECHNISCHE UNIVERSITÄT DARMSTADT

Method	Robust	Covid	News	Touché	Avg.
BM25-QE	0.496	0.610	0.392	0.271	0.442
BM25 (w/o Feedback Docs.)	0.045	0.161	0.055	0.105	0.091
kNN	0.443	0.686	0.365	0.174	0.417
kNN (query only)	0.362	0.665	0.254	0.165	0.361
CE Query FT (bias)	0.484	0.723	0.335	0.198	0.435
CE Query FT (full)	0.520	0.722	0.341	0.189	0.443
CE MAML + Query-FT	0.506	0.735	0.314	0.223	0.445
CE supervised + Query-FT	0.493	0.725	0.325	0.217	0.440


Phase 1: First Stage Retrieval & Relevance Feedback
Query q
BM25 Retrieval
Relevance Feedback
Phase 2: Query Expansion & Re-Ranker Fine-Tuning
Query $q+$
$2k$ Feedback Docs R
Few-Shot
Query Fine-Tuning
Re-Ranker
(CE)
BM25-QE
Retrieval
2nd Stage
Retrieval
Phase 3: Re-Ranking
Query $q+$
Query $q+$
Query $q+$
Re-Ranker
(CE / kNN)
Rank
Fusion
Re-Ranked
Documents

References I

Bondarenko, A., Gienapp, L., Fröbe, M., Beloucif, M., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., & Hagen, M. (2021). Overview of touché 2021: Argument retrieval. In K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, & N. Ferro (Eds.), *Experimental IR meets multilinguality, multimodality, and interaction - 12th international conference of the CLEF association, CLEF 2021, virtual event, september 21-24, 2021, proceedings* (pp. 450–467, Vol. 12880). Springer. https://doi.org/10.1007/978-3-030-85251-1_28

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, *36*(2), 3–10. https://doi.org/10.1145/792550.792552

References II

Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 758–759. https://doi.org/10.1145/1571941.1572114

Järvelin, K., & Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 41–48.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

References III

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). Ms marco: A human generated machine reading comprehension dataset. https://www.microsoft.com/en-us/research/publication/ms-marco-human-generated-machine-reading-comprehension-dataset/

Nogueira, R., & Cho, K. (2019). Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Soboroff, I., Huang, S., & Harman, D. (2018). Trec 2018 news track overview.. *TREC*.

References IV

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In J. Vanschoren & S. Yeung (Eds.), *Proceedings of the neural information processing systems track on datasets and benchmarks* (Vol. 1). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/65b9eea6e1cc6bb9f0cd2a47751a186f-Paper-round2.pdf

Voorhees, E., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W. R., Lo, K., Roberts, K., Soboroff, I., & Wang, L. L. (2021). Trec-covid: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, *54*(1). https://doi.org/10.1145/3451964.3451965

Voorhees, E. M., et al. (2004). Overview of trec 2004.. *Trec*.