Impact of Shallow vs. Deep Relevance Judgments on BERT-based Reranking Models

Gabriel Iturra-Bocaz
Danny Vo
Petra Galuščáková

Information Access and Artificial Intelligence
University of Stavanger

Paper:

Repository:

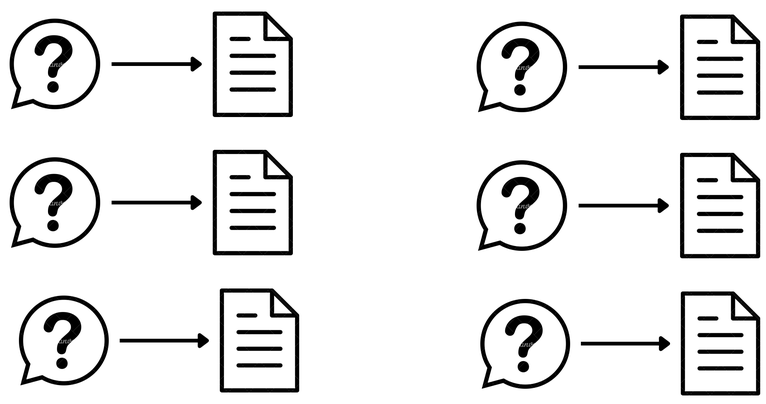
PROBLEM

- This study focus on how training on deep (many judgments per query) versus shallow (many queries, few judgments each) datasets affects neural reranker models performance.
- It examines the impact of negative sampling and the reuse of manual relevance judgments for training.
- **Research question:** When do deep judgments outperform shallow ones for training rerankers?

NEGATIVE SAMPLING

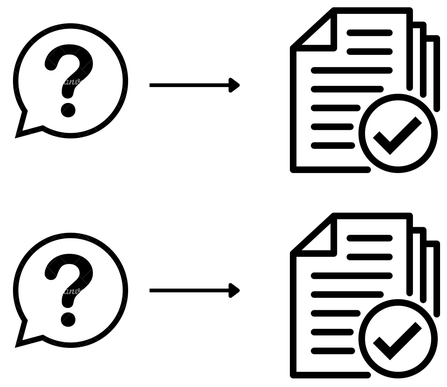
- Positive samples require humans intervention; negatives samples can be automated.
- For each query negative sampling are extracted using BM25 and filtered out the 10 first documents to avoid false positives.
- Negative sampling improves performance in deep-judged datasets with few positive samples.

SHALLOW-JUDGED DATASETS



Datasets with **many queries**, each having only a **few relevance judgments** (qrels).

DEEP-JUDGED DATASETS



Datasets with relatively **few queries**, each having **many relevance judgments** (qrels).

RESULTS

MS MARCO V1 COLLECTION

Model	Type	Inst.	Q/Q Ratio	MAP@10	NDCG@10	MRR@10
BM25	---	0	---	0.1793	**0.2269**	0.1852
BM25 + BERT	---	0	---	0.1553	**0.2068**	0.1582
BM25 + BERT	**Deep**	4,200	70/60	0.1136	**0.1756**	**0.1183**
BM25 + BERT	**Deep**	5,000	50/100	**0.1145**	**0.1748**	0.1173
BM25 + BERT	**Shallow**	4,200	2,100/2	0.2128	**0.2578**	**0.2255**
BM25 + BERT	Shallow	5,000	2,500/2	**0.2201**	**0.2594**	**0.2277**

LONGEVAL SHORT-TERM AND LONG-TERM COLLECTIONS

Test	Model	Type	Inst.	Q/Q Ratio	MAP@10	NDCG@10	MRR@10
Short-term	BM25	---	---	---	0.1189	**0.1746**	0.2430
	BM25 + BERT	---	---	---	0.1183	**0.1749**	0.2474
	BM25 + BERT	Deep	2,250	45/60	0.1034	**0.1607**	**0.2068**
	BM25 + BERT	Deep	3,100	31/100	**0.1155**	**0.1709**	**0.2324**
	BM25 + BERT	Shallow	1,508	754/2	0.1064	**0.1637**	**0.2199**
Long-term	BM25 + BERT	---	---	---	0.1150	**0.1736**	**0.2505**
	BM25 + BERT	---	---	---	**0.0702**	**0.1297**	**0.1337**
	BM25 + BERT	Deep	2,250	45/60	**0.1027**	**0.1618**	0.2208
	BM25 + BERT	Deep	3,100	31/100	0.1103	**0.1681**	**0.2357**
	BM25 + BERT	Shallow	1,508	754/2	0.1070	**0.1654**	**0.2339**

Performance of the fine-tuned BERT-based reranker models and evaluated on the test MS MARCO V1 and test LongEval collection, depending on the number of training instances (Inst.) and their Query/Qrels Ratio (Q/Q Ratio). **Evaluation metrics** are computed on **top 10 documents.**

CONCLUSIONS

- Shallow training sets consistently outperform deep training sets in all our experiments.
- The issue of lack of training data on deep datasets can be partially mitigated by increasing the number of negative training sample.

University of Stavanger, Norway **gabriel.e.iturrabocaz@uis.no**