Evaluating Temporal Persistence Using Replicability Measures

Notebook for the LongEval Lab at CLEF 2023

Jüri Keller[1,*], Timo Breuer[1] and Philipp Schaer[1]

[1]TH Köln (University of Applied Sciences), Claudiusstr. 1, Cologne, 50678, Germany

Abstract

In real-world Information Retrieval (IR) experiments, the Evaluation Environment (EE) is exposed to constant change. Documents are added, removed, or updated, and the information need and the search behavior of users is evolving. Simultaneously, IR systems are expected to retain a consistent quality. The LongEval Lab seeks to investigate the longitudinal persistence of IR systems, and in this work, we describe our participation. We submitted runs of five advanced retrieval systems, namely a Reciprocal Rank Fusion (RRF) approach, ColBERT, monoT5, Doc2Query, and E5, to both sub-tasks. Further, we cast the longitudinal evaluation as a replicability study to better understand the temporal change observed. As a result, we quantify the persistence of the submitted runs and see great potential in this evaluation method.

Keywords

web search, longitudinal evaluation, continuous evaluation, replicability

1. Introduction

This paper describes our contribution to the CLEF 2023 LongEval Lab [1].[1] The lab seeks to investigate the temporal persistence of retrieval systems. It, therefore, provides a first-of-its-kind web retrieval collection with three sub-collections from different points in time [2]. We participated in the retrieval task by providing runs of five systems to both sub-task.

A retrieval system's Evaluation Environment (EE) is under constant change. Not only but especially web retrieval systems are exposed to this due to the dynamic nature of the web. Documents, i.e., websites, get created, updated, or created [3, 4]. But besides the evolving collection, all other aspects of an EE underlay change as well, from the information need and search behavior of the users [5] all the way to the evolving language itself [6]. These changes raise questions about the persistence and generalizability of IR system effectiveness evaluations.

By requiring a temporarily reliable system to perform consistently over time, evaluating this can be understood as a replicability task. Oriented at the ACM definition of replicability[2], the

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

jueri.keller@smail.th-koeln.de (J. Keller); timo.breuer@th-koeln.de (T. Breuer); philipp.schaer@th-koeln.de (P. Schaer)

0000-0002-9392-8646 (J. Keller); 0000-0002-1765-2449 (T. Breuer); 0000-0002-8817-4632 (P. Schaer)

[1]https://clef-longeval.github.io

[2]https://www.acm.org/publications/policies/artifact-review-and-badging-current

goal is to achieve the same measurements in a different experimental setup, in this case, at a proceeded point in time.

To investigate temporal persistence, we submitted runs of five advanced retrieval systems to both sub-tasks of the LongEval Lab. The systems are not specifically adapted to changes in the LongEval dataset to validate the temporal reliability of system-oriented IR evaluations following the Cranfield paradigm. Further, as a proof of concept, we use the replicability measures Delta Relative Improvement (Δ RI) and the Effect Ratio (ER) [7] to investigate the temporal persistence. In short, the contributions of this work are:

- Descriptions of **five state-of-the-art systems** submitted to both retrieval sub-tasks,
- an **extensive evaluation** of retrieval effectiveness,
- an **adaptation of replicability measures** to evaluate temporal persistence,
- an **open-source release** of the experimental setup.

The remainder of this paper is structured as follows. Section 2 contains an analysis of the LongEval dataset. The five retrieval systems are described in Section 3. Further, Section 4 provides the results on the train slice and a preliminary evaluation of the results. In Section 5, we describe the replicability efforts. This paper concludes with a short discussion and some future work in Section 6. The code is publicly available on GitHub.[3]

2. LongEval Dataset

To our knowledge, the LongEval dataset [2] is the first dataset specifically designed to investigate temporal changes in IR. On a high level, the collection consists of three sub-collections from different points in time. Each collection contains topics and qrels. The documents as well as the topics and qrels originate from the French, privacy-focused search engine Qwant.[4] For this work, we entirely rely on the English automatic translations of the dataset. The documents contain the cleaned content of websites. They are filtered for adult and spam content, but no further processing was done, sometimes leaving unconnected phrases, keywords, or code artifacts in the documents.

The topics are selected according to *"popularity, stability, generality, and diversity"* [2]. For these topics, queries are selected from the Qwant search engine logs if they contain the topic as a sub-string. The qrels for the shared task are simulated based on the Cascade Click Model [8, 9]. Documents are assessed as not relevant, relevant, and highly relevant. Further, human-assessed gold labels are announced for September (2023). More details can be found in the original publication [2].

The sub-collections are sequential snapshots of an evolving search environment for temporal comparison. The topics are constructed once, but the queries are partially changing across sub-collections. The documents, i.e., the websites identified by the URL, are also mainly static across sub-collections but the content of the documents changes.

The collections are organized into a WT, ST, and LT sub-collection. The WT (within time) sub-collection was created in June 2022. The ST (short-term) sub-collection was created in July 2022,

[3] https://github.com/irgroup/CLEF2023-LongEval-IRC
[4] https://www.qwant.com/

Table 1
LongEval subcollection statistics. The length of documents and queries are measured in tokens, split by white spaces. The query WT q062213307 and ST q072211861 is excluded as outlier since it only contains the token *leg* 108 and 110 times.

	WT	ST	LT	Intersection
Timeframe	June 2022	July 2022	September 2022	
Number documents	1,570,734	1,593,376	1,081,334	1,011,613
Mean document length	794.11	793.96	807.28	
Min document length	0	0	1	
Max document length	7065	12210	7255	
Number queries	753	860	910	124
Mean query length	2.73	2.71	2.52	
Min query length	1	1	1	
Max query length	6	11	9	

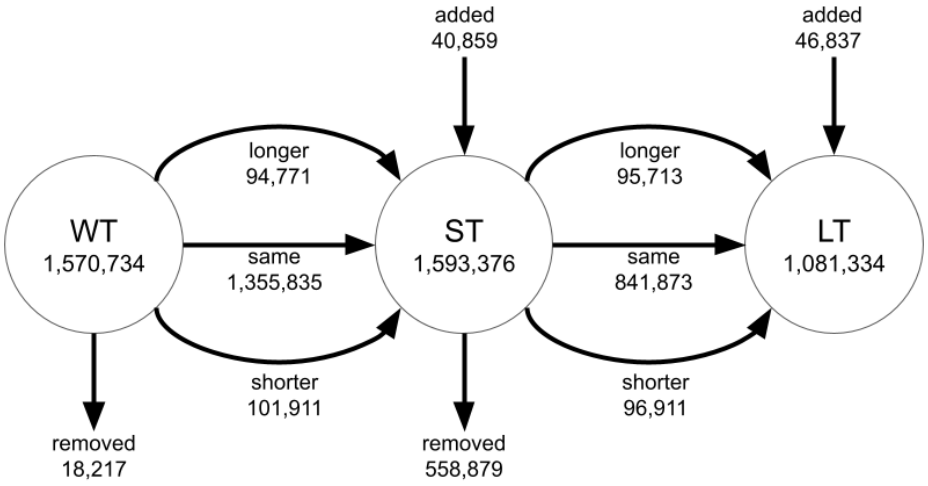

Figure 1: The evolution of the LongEval dataset documents across the three sub-collections. Transitioning from one sub-collection to the next, documents are added, removed, or updated. All documents were harmonized by their URLs.

immediately after the WT collection. The third sub-collection, LT (long term), contains more distant data as it was created with a two-month gap from ST in September 2022. Table 1 gives an overview of the sub-collections.

The LongEval dataset contains over 1.5 million documents. Not every document is present in every sub-collection, but most documents are. The core document collection contains 1,011,613 documents, as identified by matching their URLs. Versions of these documents are present in

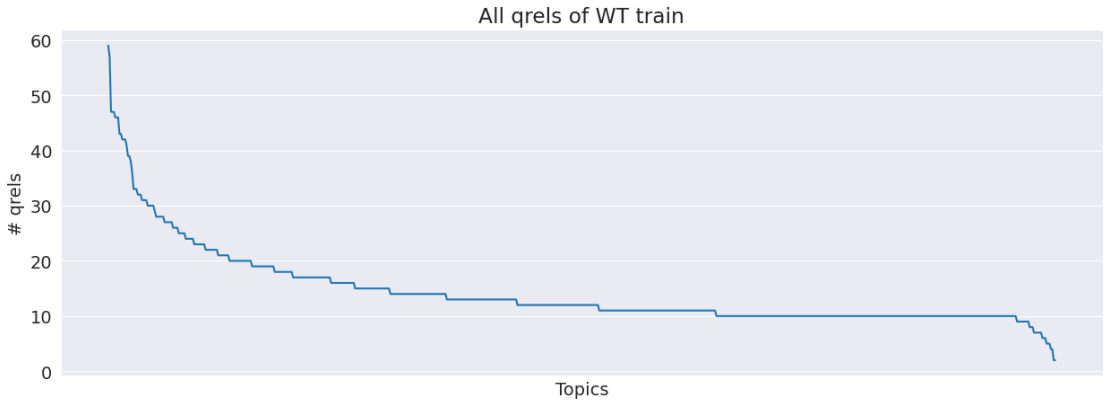

Figure 2: Distribution of qrels per query for the 672 WT train sub-collection queries.

every sub-collection but do not necessarily contain exactly the same content. The documents evolve over time, meaning that the content of one website might change over time. To capture this change on a general level, Figure 1 shows how many documents increase or decrease in character length and how many documents are added, deleted, or stay the same in length. We note that between ST and LT considerably more documents are removed from the collections than between WT and ST.

Like the documents, the queries change over time as well. However, relatively fewer core queries that appear in all sub-collections exist. In total, only 124 unique query strings appear in all collections. However, the overlap of query IDs is larger due to duplicate queries that are probably caused by automatic translations.

The relevance judgments (qrels) classify the documents' relevance on a three-graded scale, including *not relevant*, *relevant*, and *highly relevant* labels. In general, the dataset has few assessed documents per topic. While the mean number of qrels is 14 per topic, the absolute number fluctuates between 2 and 59. Figure 2 shows the distribution of all qrels per query. Most of the documents are marked as not relevant, and the distribution of relevant and highly relevant qrels is skewed as well. Especially the highly relevant qrels are rare, with a maximum of only four and a mean of only one highly relevant document per topic. In the evaluations, these single documents heavily influence the final outcome as their position in the ranking especially impacts the score of rank-based measures like nDCG. While relevant qrels are generally rare, 16 queries do not have a single relevant document.

3. Approaches and Implementations

We compared different ranking functions and multi-stage retrieval systems on the WT train slice of the LongEval dataset. The systems were chosen as they represent state-of-the-art, off-the-shelf methods that are used in many recent IR experiments. Therefore, it is especially interesting how these systems behave over time without being specifically adapted to a changing environment.

3.1. Statistical Ranking Functions

Different ranking functions were used as baselines in their default configurations. Special attention was given to the BM25 [10] ranking function as it is a robust, efficient, and often hard-to-beat baseline. We use this run to compare advanced systems to it. Since we use the PyTerrier [11] framework for experiments, the default parameters $k_1 = 1.2$ and $b = 0.75$ were kept. Further we included PL2 [12], TF-IDF and DFR χ^2 [13].

To further improve these ranking functions, two query expansion methods are employed. Namely, RM3 [14] and Bo1 [12] are used to extend the queries through pseudo-relevance feedback. The default PyTerrier parameters are also kept here; three feedback documents were used to gather ten feedback terms.

3.2. Rank Fusion

Multiple runs were combined into a single ranking to profit from the diversity of multiple ranking functions. First, BM25, DFR χ^2 and PL2 are fused through Reciprocal Rank Fusion (RRF) [15] with the `ranx` Python library [16]. Further runs are created by using the pseudo-relevance-feedback methods on top of BM25. The default parameters $min_k = 10$, $max_k = 100$ and $step = 10$ were used for the RRF.

3.3. ColBERT

ColBERT [17] applies the BERT [18] Language Model (LM) to overcome the lexical gap [19] by creating semantic representations of queries and documents as embeddings. In contrast to traditional BERT-based approaches like cross-encoders, the interaction mechanism used to calculate the similarity between a document and a query is detached from the embedding creation process. However, in contrast to bi-encoder systems, nuanced similarities can be calculated. To do so, semantic representations for a query or a document are calculated as a set of token embeddings. The relevance score between a query and a document is then calculated as the sum of the max of the cosine similarity or the L2 distance between all embeddings for the query and the document.

By separating the scoring from the embedding process, the efficiency at run time can be greatly improved as all document embeddings can be calculated beforehand offline. ColBERT can also be used in a later retrieval stage as a reranker. The PyTerrier version of ColBERT [5] was used in a zero-shot fashion. Besides using ColBERT as a first-stage retriever, where the whole corpus is converted to embeddings, ColBERT was also used to rerank the top 1000 BM25 results.

3.4. monoT5

The potential of sequence-to-sequence models can be fostered for the ranking task by providing a query and a document as input and asking the model to decide if the document is relevant for this query by generating "true" or "false." The softmax of the generated token probability is then used as confidence for the predicted class to compute the final relevance of the document [20]. The T5 [21] model was fine-tuned in this fashion on the MS Marco passage retrieval dataset [22] as

[5] https://github.com/terrierteam/pyterrier_colbert

monoT5 by Pradeep et al. [23]. This model is then used in a second stage to rerank BM25 rankings and achieves great results, even as a pre-trained model on other datasets and domains [23].

The T5 model supports 512 sub-word tokens, and the LongEval dataset consists of documents with an average length of around 800 tokens. To avoid arbitrary truncation, the document retrieval task is formulated as a passage retrieval task, and the top 1000 BM25 results are split into (still arbitrary but shorter) passages with an overlap half the size of the passage. By that, the whole document texts are reranked by monoT5. Further, the maximum relevance score of all passages from one document is used as the relevance score of the document for the final ranking.

For comparison and to avoid arbitrary sequences, the full documents are used instead as well. This approach seems reasonable since not too much text is cut off from the average document, and the title and introductions with high-level terms, similar to the query terms, are often located at the beginning of a document and are therefore captured by the model.

3.5. Doc2Query

Instead of applying a language model at the reranking stage, Doc2Query [24] uses the T5 model to generate likely queries that a document could answer. These additional queries are then indexed along the document itself. By that, natural language queries can result in exact matches using traditional ranking functions, and alleged relevant terms are boosted. This results in an advanced index that can be efficiently searched independent of methods.

The effectiveness is highly dependent on the number of queries that are added to the documents during indexing since this determines how much content is added. For this experiment, we used three and ten queries. While Rodrigo and Lin [24] used up to 80 queries, a maximum of ten queries were chosen to match the available resources. Three queries are the default of the implementation and were used as a lower bound to test the effect.

3.6. E5

Recently Wang et al. [25] achieved superior performance with the E5 model family. It is the first model that outperforms BM25 in a zero-shot retrieval setting on the BEIR [26] benchmark. The performance is attributed to the large and high-quality dataset, the contrastive pre-training and the advanced fine-tuning process. The new paired dataset CCPairs [26] of query passage pairs was used for training. It contains 1.3 billion query document pairs from Reddit, Wikipedia, SemanticScoolar, CommonCrawl, Stack Exchange, and news websites.

The models $E5_{small}$ and $E5_{base}$ are used in a zero-shot fashion to create embeddings for all queries and documents. The documents are truncated at 512 sub-word tokens to fit in the model and not split into passages for efficiency. A Faiss[6] flat index was created from all embeddings, and L2 was used to score the query document similarity.

Table 2
Results on the train slice of the WT sub-collection. The best results per group are highlighted in **bold**, and significant differences with Bonferroni correction to the BM25 baseline are denoted by an asterisk (∗).

System	MAP	Bpref	RR	P@20	nDCG	nDCG@20
BM25	0.1452	0.3245	0.2604	0.0654	0.2884	0.2087
PL2	0.1408	**0.3352**	0.2572	0.0650	0.2884	0.2064
TF-IDF	**0.1467**	0.3259	**0.2637**	0.0660	**0.2907**	**0.2109**
DFR χ^2	0.1428	0.3265	0.2629	**0.0633**	0.2871	0.2042
BM25+Bo1	**0.1470**	**0.3341**	**0.2534**	**0.0661**	**0.2922**	**0.2075**
BM25+RM3	0.1426	0.3295	0.2408	0.0658	0.2867	0.2035
RRF(BM25, DFR χ^2, PL2)	0.1462	0.3380*	0.2646	0.0656	0.2967*	0.2101
RRF(BM25+Bo1, DFR χ^2, PL2)	**0.1511**	0.3466*	**0.2686**	0.0673	**0.3040***	**0.2156**
RRF(BM25+RM3, DFR χ^2, PL2)	0.1472	**0.3472***	0.2589	**0.0676**	0.3008*	0.2125
BM25+passages+monoT5	0.1540	0.3369	0.2743	0.0708*	0.2969	0.2196
BM25+monoT5	**0.1809***	**0.3494***	**0.3216***	**0.0768***	**0.3208***	**0.249***
d2q(3)>BM25	0.1578	**0.3411**	0.2630	**0.0752***	0.2940	0.2284*
d2q(10)>BM25	**0.1638***	0.3382	**0.2862***	0.0707*	**0.3070***	**0.2287***
colBERT	0.1652	0.3435	0.3045*	0.0689	0.2989	0.2290
BM25+colBERT	**0.1682***	**0.3447**	**0.3046***	**0.0692**	**0.3082***	**0.231***
E5_small	0.1437	0.3265	0.2705	0.0619	0.2762	0.2039
E5_base	**0.1545**	**0.3483**	**0.2826**	**0.0634**	**0.2910**	**0.2128**

4. Evaluation

In the following, results for the initial experiments on the train slice of the WT sub-collection are reported, and the submitted systems are analyzed. Then, the runs and results on the full dataset are described.

4.1. System Selection

Table 2 gives an extensive overview of the initial experiments. BM25 appeared to be a strong baseline, outperformed only by some systems and most often not statistically significant on all measures. The best runs of the different types were chosen for submission, also with the goal in mind to provide a diverse set of runs for the planned pooled gold annotation [2].

For the official ranking, we submitted to both sub-tasks the five systems:

1. RRF(BM25+Bo1, DFR χ^2, L2) as **`IRC_RRF(BM25+Bo1-XSqrA_M-PL2)`**
2. BM25+colBERT as **`IRC_BM25+colBERT`**
3. BM25+monoT5 as **`IRC_BM25+monoT5`**
4. d2q(10)>BM25 as **`IRC_d2q(10)>BM25`**
5. $E5_{base}$ as **`IRC_E5_base`**

[6] https://faiss.ai/

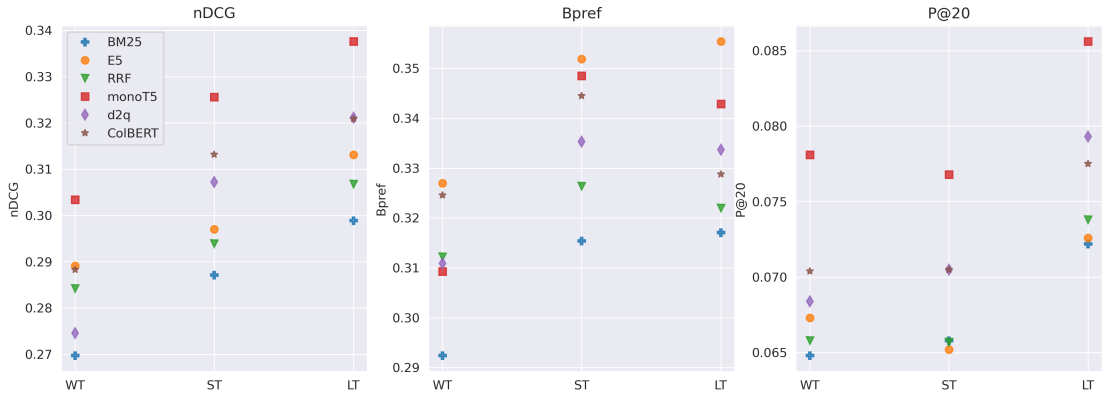

Figure 3: The ARP of nDCG (left), Bpref (center), and Reciprocal Rank (right) from the submitted systems at WT, ST, and LT.

The BM25 baseline achieved an nDCG of 0.2884 on the WT train sub-collection slice. A MAP of 0.1452 is reported, but as initially shown in the data analysis in Section 2, only a few qrels per query are available; we relied on the bpref [27] measure instead. Here, a score of 0.3245 is achieved. Notably, compared to BM25, TF-IDF outperforms BM25 slightly but is not statistically significant. Regarding the runs with additional pseudo-relevance feedback, no significant improvements are made as well.

The RRF runs show the first significant improvements. The fusion run of the three runs BM25+Bo1, DFR χ^2, and PL2 significantly outperform the BM25 baseline on MAP and nDCG to some extent. Larger improvements and the overall best results are achieved with BM25+monoT5. This run is significantly better on all measures and archives a 0.0324 higher nDCG. The passage retrieval version of the run performs considerably worse, similar to the baseline. The gap between the BM25 results on the two Doc2Query extended indexes is similar. While the results on the version with three additional queries per document make statistically no difference to the baseline, the results on the ten queries indexes are almost as good as the ones with BM25+monoT5 on all measures, except for P@20, which is even better. BM25+ColBERT performs slightly worse overall. Focusing on P@20, the system differs not from the baseline. Employing ColBERT as a first-stage ranker impairs the performance further. The results achieved with the E5 models as first-stage rankers are not significantly different from the baseline. Still, the base version outperforms the baseline in all measures, and the small version does on Bpref and RR.

4.2. Results

For the evaluation of the result, the main goal is not a high but rather persistent performance. The underlying assumption is that the system would continuously achieve the same performance. To evaluate this, the Result Delta ($\mathcal{R}_e\Delta$) between the averaged retrieval performances at two different points in time is measured as proposed by Sáez et al. [28]. The results are presented in Table 3 and visualized in Figure 3.

`IRC_RRF(BM25+Bo1-XSqrA_M-PL2)`: The fused run contains at least 1000 results for all topics in the WT sub-collection. For the ST sub-collection the system could not find any docu-

Table 3
Results on the three (test) sub-collections as well as the deltas between them. The best system per measure and group is highlighted in **bold**, and significant differences from the BM25 baseline are denoted with an asterisk*.

		ARP			$\mathcal{R}_e\Delta$	
		WT	ST	LT	WT, ST	WT, LT
Bpref	BM25	0.2924	0.3154	0.3171	-0.0230	-0.0247
	RRF	0.3122	0.3264*	0.3220	**-0.0142**	-0.0098
	ColBERT	0.3246	0.3445*	0.3288	-0.0392	-0.0336
	monoT5	0.3093	0.3485*	0.3429*	-0.0244	-0.0228
	d2q	0.3109	0.3353*	0.3337*	-0.0199	**-0.0042**
	E5	**0.3270**	**0.3519***	**0.3554***	-0.0249	-0.0284
P@20	BM25	0.0648	0.0658	0.0722	-0.0010	-0.0074
	RRF	0.0658	0.0657	0.0738	**0.0001**	-0.0080
	ColBERT	0.0704	0.0705*	0.0775*	0.0013	-0.0075
	monoT5	**0.0781***	**0.0768***	**0.0856***	-0.0021	-0.0109
	d2q	0.0684	0.0705*	0.0793*	**-0.0001**	-0.0071
	E5	0.0673	0.0652	0.0726	0.0021	**-0.0053**
nDCG	BM25	0.2697	0.2871	0.2989	-0.0174	-0.0292
	RRF	0.2842*	0.2939*	0.3068*	-0.0097	**-0.0226**
	ColBERT	0.2883	0.3132*	0.3209*	-0.0222	-0.0342
	monoT5	**0.3034**	**0.3256***	**0.3376***	-0.0326	-0.0465
	d2q	0.2746	0.3072*	0.3211*	-0.0249	-0.0326
	E5	0.2891	0.2970	0.3131	**-0.0079**	-0.0240

ments for four queries. Namely the queries *to*, *a*, *the* and *the*[7] resulted in empty rankings. These queries consist only of stopwords, which leave an empty query string after query processing. These queries are most likely bad translations from the terms *verseau, argentique, nanterre* and *falloir*, mostly containing named entities. For the two LT sub-collection topics *cadreemploi* and *a*[8], no BM25 first stage ranking could be created. While *a* is again just a stopword, for the term *cadreemploi* no results were found, which could possibly be explained by a spelling error of the French job exchange website *cadremploi*. Similarly, the topic *cadreemploi* is also present in the French queries.

The Average Retrieval Performance (ARP) — defined by the mean retrieval performance over multiple topics — improves slightly over time. In general, the measured differences between the sub-collections are fairly small. The Δ nDCG between WT and ST is only -0.0097 and between WT and LT -0.0226.

`IRC_BM25+colBERT`: Based on the WT sub-collection for the topic *ducielalaterre*[9] no documents were found, and for all other topics, at least 1000 documents could be retrieved. Since ColBERT was employed as a reranker on top of BM25, the four topics *to*, *a*, *the* and *the*[10] still

[7]LongEval ST qid: q072214697, q072222604, q072224942, q072212314
[8]LongEval LT qid: q0922511 and q092219105
[9]LongEval WT held out qid: q062216851
[10]LongEval ST qid: q072214697, q072222604, q072224942, and q072212314

remain empty. For 28 other topics, only less than 1000 documents, ranging between three and 663, could be found. Like before, the LT sub-collection topics *cadreemploi* and the topic *a*[11] remain empty. For further 22 topics, less than 1000 results were found. For example, the fewest results were found for the topic *the audeau*.[12]

The ARP is increasing over time, as already observed for the RRF system. However, the differences are larger for this system. Between WT and ST the Δ nDCG is -0.0249, and between WT and LT -0.0326.

IRC_BM25+monoT5: The composition of the runs stayed mostly the same for these runs. Since they also use BM25 as the first-stage ranking, the issue of empty or short rankings remains.

As already observed on the train slice of the WT sub-collection, the ARP is the highest achieved on all measures and sub-collections compared to the other submitted systems, with small exceptions. One strong exception is the Bpref of only 0.3093 on the WT sub-collection, the smallest score achieved overall. However, the results are inconsistent; the deltas are higher, especially for Bpref.

IRC_d2q(10)>BM25: Through the document expansion with Doc2Query, at least 37 documents were found for the previously empty WT sub-collection topic *ducielalaterre*.[13] However, for the other sub-collections, the results stayed similar. Doc2Query performed weaker than initially on the train slice before, especially in comparison to monoT5. The result deltas between WT and ST and WT and LT are among the highest for nDCG and P@20.

IRC_E5_base: Since the E5 model is based on k-NN and no stopwords were removed, for every topic, 1000 results were found. Compared to the train slice of the WT sub-collection, the system performed better. It achieved the highest Bpref on all three sub-collections and a high overall nDCG. The results are especially consistent between sub-collections with a Δ nDCG of 0.0079 between WT and ST and -0.0240 between WT and LT.

5. Temporal Persistence as Replicability

Building upon the result delta evaluation as introduced by Sáez et al. [28], we propose to use replicability measures to investigate the environment effect on the systems further. As described and implemented by Breuer et al. [7, 29], the ARP may hide differences between the topic score distributions. For example, the RRF system achieved a high nDCG (0.28) at WT and is relatively stable considering the $\mathcal{R}_e\Delta(WT, ST)$ of 0.001. However, the per-topic results fluctuate between -0.4 and 0.8, as shown in Figure 4. For some topics, the retrieval performance improves, while the changes in the EE harm retrieval performance for other topics. We note that these circumstances require a more in-depth evaluation.

For a more detailed analysis of how the topic score distributions change, we cast the temporal comparison into a replication task, i.e., we evaluate the same set of systems on different data. Naturally, a direct comparison based on different sub-collections is difficult since it remains unclear if the observed effects should be attributed to the system or the changing EE. To overcome this problem, a pivot system similar to that described by Sáez et al. [28] is used, and

[11] LongEval LT qid: q0922511, q092219105
[12] LongEval LT qid: q092220802
[13] LongEval WT held out qid: q062216851

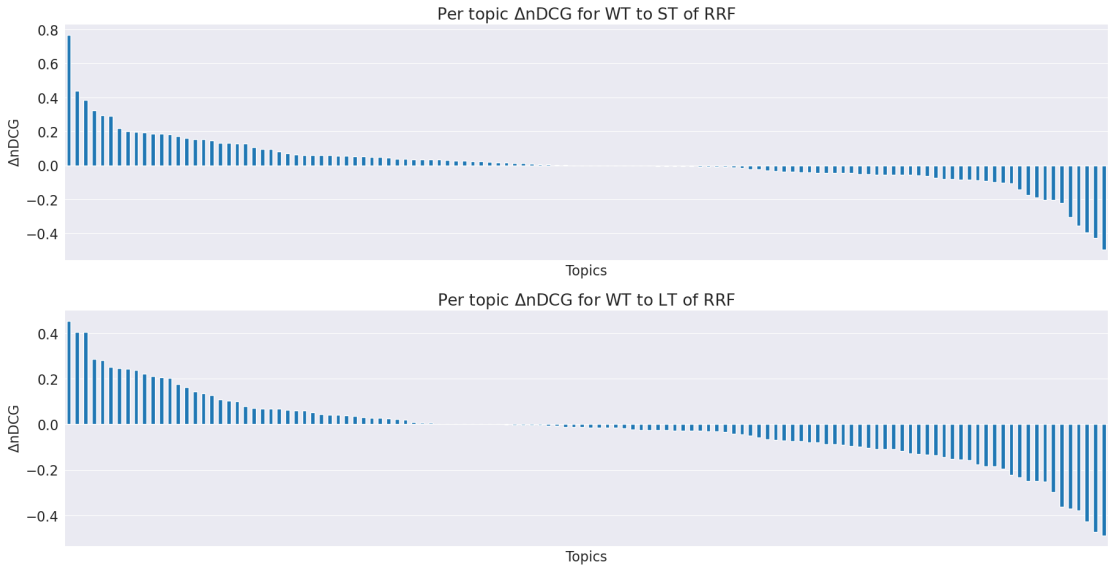

Figure 4: RRF $\Delta nDCG$ results per topic for WT to ST (top) and WT to LT (bottom). The topics are ordered according to the delta.

likewise, the experimental system is kept fixed in both EE. Effects are measured in comparison to this pivot system on one sub-collection and then compared to the same setup on a later sub-collection. To align the terminology, the pivot system is a baseline run, BM25 for simplicity in this example, and the advanced run is the experimental system investigated.

In addition to the $\mathcal{R}_e\Delta$, as reported earlier in Table 3, we report the Effect Ratio (ER) and the Delta Relative Improvement (Δ RI). The ER [7] is originally defined by the ratio between relative improvements of an advanced run over a baseline run. The relative improvements are based on the per-topic improvements, which are adapted for changing EEs as follows:

$$\Delta M_j^{EE_1} = M_j^{EE_1}(S) - M_j^{EE_1}(P), \Delta' M_j^{EE_2} = M_j^{EE_2}(S) - M_j^{EE_2}(P) \tag{1}$$

where $\Delta M_j^{EE_1}$ denotes the difference in terms of a measure M between the pivot system P and the experimental system S for the j-th topic of the evaluation environment EE_1. Correspondingly, $\Delta' M_j^{EE_2}$ denotes the topic-wise improvement in the evaluation environment EE_2. The ER is then defined as:

$$\mathrm{ER}\left(\Delta' M^{EE_2}, \Delta M^{EE_1}\right) = \frac{\overline{\Delta' M^{EE_2}}}{\overline{\Delta M^{EE_1}}} = \frac{\frac{1}{n_{EE_2}} \sum_{j=1}^{n_{EE_2}} \Delta' M_j^{EE_2}}{\frac{1}{n_{EE_1}} \sum_{j=1}^{n_{EE_1}} \Delta M_j^{EE_1}}. \tag{2}$$

More specifically, the mean improvement per topic between the pivot and experimental system on one sub-collection (of EE_1) in comparison to the effect on the other sub-collection (of EE_2) is measured. Thereby, the ER is sensitive to the effect size. If the effect size is completely replicated in the second sub-collection, the ER is 1, i.e., the retrieval system is robust. If the ER is between 0 and 1, the effect is smaller, indicating a less robust system with performance drops.

If the ER is larger than 1, the effect is larger, indicating performance gains caused by the change of the EE. Additionally, we include the Δ RI [7], based on the relative improvements (RI) that are adapted to the LongEval definitions as follows:

$$\mathrm{RI} = \frac{\overline{M^{EE_1}(S) - M^{EE_1}(P)}}{\overline{M^{EE_1}(P)}}, \qquad \mathrm{RI}' = \frac{\overline{M^{EE_2}(S) - M^{EE_2}(P)}}{\overline{M^{EE_2}(P)}} \tag{3}$$

where M^{EE} denotes the score of a measure M determined with EE, and S and P denote the experimental and pivot system, respectively. The Δ RI is then defined as:

$$\Delta\mathrm{RI} = \mathrm{RI} - \mathrm{RI}'. \tag{4}$$

Therefore, a comparison between different sub-collections is straightforward. The ideal Δ RI of 0 is achieved if the RI is the same between both sub-collections, indicating a robust system. The more Δ RI deviates from 0, the less robust is the system, whereas negative scores indicate a more effective experimental system S in the evaluation environment EE_2, and higher scores correspond to a less effective experimental systems than in the evaluation environment EE_1. All of the replicability measures were implemented with the help of `repro_eval` [29], which is a dedicated reproducibility and replicability evaluation toolkit.

Even though the replicability measures do not necessarily require the same topics for each sub-collection, we harmonized the topics. Therefore, we only rely on the core queries that are shared between the sub-collections in this analysis. Given this methodology, the extended results are presented in Table 4. For all systems, the ARP decreases slightly at first (WT to ST) but increases in the long run (WT to LT) — a circumstance that is also reflected by the lower $\mathcal{R}_e\Delta$ scores for WT to ST compared to WT to LT.

The ER and Δ RI complement $\mathcal{R}_e\Delta$. For instance, monoT5 achieved similar P@20 scores on WT and ST, resulting in a $\mathcal{R}_e\Delta$ score of 0, which indicates perfect robustness in terms of $\mathcal{R}_e\Delta$. However, when comparing ER and also Δ RI, more granular analysis is possible. In this case, the scores are close to but different from the perfect scores of 1 and 0, respectively, which would indicate perfect robustness. In general, the $\mathcal{R}_e\Delta$ scores do not always agree on the most robust system with ER and Δ RI. By these findings, we conclude that the replicability measures provide another perspective of the robustness, and we emphasize once again that it is also important to consider the topical variance over time.

Furthermore, we see that it is not enough to consider the differences of a single retrieval measure like nDCG. Depending on the evaluation measure, different systems perform best in terms of robustness. For instance, $\mathcal{R}_e\Delta$ of nDCG is lower for ColBERT and Doc2Query than that of monoT5, while $\mathcal{R}_e\Delta$ of P@20 is lower for monoT5. Similarly, the replicability measures should be instantiated with different retrieval measures to get a more comprehensive understanding of robustness. While our RRF-based submissions achieve the best $\mathrm{ER}_{\mathrm{nDCG}}$ on both tasks, monoT5 is the most robust system in terms of $\mathrm{ER}_{\mathrm{P@20}}$. Likewise, ER and Δ RI identify different systems as the most robust for the same measures and tasks, which shows that it is insightful to evaluate both replicability measures.

In addition, we also included the p-values of unpaired tests based on the topic score distributions from different EE that were determined with the same experimental system as proposed in [7]. The general idea of these evaluations proposes to determine the quality of replicability

Table 4
Extended results on the core queries, including the replicability measures.

		ARP			$\mathcal{R}_e\Delta$		ER		Δ RI		p-val	
	System	WT	ST	LT	WT, ST	WT, LT	WT, ST	WT, LT	WT, ST	WT, LT	WT, ST	WT, LT
P@20	BM25	0.070	0.067	0.085	0.002	-0.015	1.000	1.000	0.000	0.000	1.000	1.000
	RRF	0.075	0.069	0.088	0.006	**-0.013**	0.311	0.544	0.051	0.041	0.591	0.269
	colBERT	0.072	0.071	0.087	0.002	-0.015	1.244	0.933	**-0.011**	**0.009**	0.875	0.190
	monoT5	**0.081**	**0.081**	**0.096**	**0.000**	-0.014	**1.191**	**0.953**	-0.039	0.037	**0.998**	0.229
	d2q	0.079	0.072	0.091	0.007	**-0.013**	0.499	0.726	0.062	0.051	0.547	**0.303**
	E5	0.071	0.066	0.088	0.005	-0.017	-1.452	2.903	0.040	-0.022	0.616	0.125
nDCG	BM25	0.269	0.272	0.306	-0.003	-0.037	1.000	1.000	0.000	0.000	1.000	1.000
	RRF	0.285	0.282	0.314	0.003	-0.030	**0.925**	**0.786**	**0.003**	0.013	0.945	0.227
	colBERT	0.276	0.275	0.297	**0.001**	-0.021	0.441	-1.198	0.015	0.053	**0.967**	0.412
	monoT5	**0.295**	**0.302**	0.311	-0.007	**-0.015**	1.146	0.187	-0.013	0.083	0.817	**0.580**
	d2q	0.285	0.287	**0.327**	**-0.001**	-0.042	0.916	1.317	0.006	**-0.010**	0.960	0.150
	E5	0.290	0.300	0.313	-0.010	-0.023	1.333	0.362	-0.025	0.054	0.720	0.382
Bpref	BM25	0.314	0.314	0.324	**-0.000**	-0.010	1.000	1.000	0.000	0.000	1.000	1.000
	RRF	0.346	0.328	0.347	0.019	-0.001	0.574	**1.007**	0.032	**0.002**	0.784	0.756
	colBERT	0.324	0.317	0.338	0.007	-0.013	0.286	1.278	0.024	-0.008	0.826	0.668
	monoT5	0.337	0.344	0.337	-0.007	**0.000**	1.261	0.553	-0.019	0.034	0.850	**0.997**
	d2q	0.335	0.331	0.368	0.004	-0.033	**0.779**	2.034	**0.015**	-0.067	**0.894**	0.300
	E5	**0.368**	**0.354**	**0.371**	0.014	-0.003	0.738	0.863	0.045	0.028	0.692	0.931

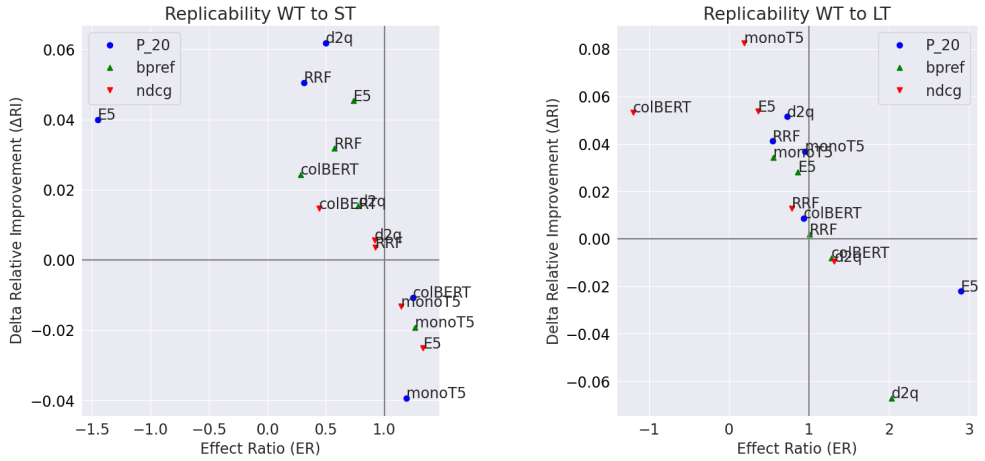

Figure 5: The ER plotted against the Δ RI for the replication WT to ST (left) and WT to LT (right).

(in our case, robustness) by the p-values. It follows the assumption that lower p-values give a higher probability of failed replications or systems that are not robust. As can be seen, the highest p-values are achieved for the monoT5, ColBERT, or d2q, which generally agrees with our earlier observations.

The full potential of the ER and Δ RI can be seen if plotted against each other as in Figure 5. The closer the systems are located to the point (1, 0), the more persistent they are, with the preferable regions bottom right and top left. For the comparison of WT to ST, the monoT5 system performs well on all three measures. However, the effect and the absolute scores are larger. The

E5 system completely fails to replicate the absolute P@20 score and shows a generally larger difference. The RRF system, like most others, shows smaller absolute scores according to the Δ RI and a slightly decreased effect ratio. The plot regarding WT to LT shows more outliers with larger effect sizes for P@20 for the E5 system and Bpref for the d2q system. The systems are shifted to the top right of the plot, a trend similar to the increased $\mathcal{R}_e\Delta$ for WT to LT.

6. Conclusion and Outlook

In this work, we described our participation in the LongEval Lab at CLEF 2023. As the core contribution, we applied five advanced retrieval systems to the LongEval dataset and submitted the runs to both sub-tasks. As this is a new challenge, the interpretation of the results is difficult. The results for the different systems are very similar. The measured differences are statistically significant but appear small as compared to the same methods on different datasets as listed on the IR experiment platform [30].[14] Interestingly, an increasing ARP over time was observed for most systems and measures. Still, the performance difference, measured by $\mathcal{R}_e\Delta$, is smaller for WT to ST compared to WT to LT, which complies with the natural assumption that persistence deteriorates over time.

Further, we report preliminary results applying replicability measures to quantify temporal persistence, an extension on common practices of these measures and their interpretation [31]. It was shown that the results based on different measures and likewise for different topics do not necessarily agree with each other. Therefore, we see great potential in using replicability measures to gain further insights into robustness and also saw similarities to the measured result deltas. All in all, a strong environment effect on the systems was shown and could be analyzed.

Future work will be regarding the selection of the pivot system and qualitative core queries. Also, further harmonizing the dataset by unifying the document IDs would allow us to cast the problem as a reproducibility task and investigate persistence on an even more specific level with reproducibility measures.

References

[1] R. Alkhalifa, I. Bilal, H. Borkakoty, J. Camacho-Collados, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, G. Gonzalez-Saez, P. G. and Lorraine Goeuriot, E. Kochkina, M. Liakata, D. Loureiro, H. T. Madabushi, P. Mulhem, F. Piroi, M. Popel, C. Servan, A. Zubiaga, Overview of the clef-2023 longeval lab on longitudinal evaluation of model performance, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science (LNCS), Springer, Thessaliniki, Greece, 2023.

[2] P. Galuscáková, R. Deveaud, G. G. Sáez, P. Mulhem, L. Goeuriot, F. Piroi, M. Popel, Longeval-retrieval: French-english dynamic test collection for continuous web search evaluation, CoRR abs/2303.03229 (2023). doi:10.48550/arXiv.2303.03229.

[14] https://www.tira.io/task/ir-benchmarks

[3] J. Bar-Ilan, Criteria for evaluating information retrieval systems in highly dynamic environments, in: M. Levene, A. Poulovassilis (Eds.), Proceedings of the Second International Workshop on Web Dynamics, WebDyn@WWW 2002, Honululu, HW, USA, May 7, 2002, volume 702 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2002, pp. 70–77. URL: https://ceur-ws.org/Vol-702/paper7.pdf.

[4] S. T. Dumais, Temporal dynamics and information retrieval, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, Association for Computing Machinery, 2010, pp. 7–8. doi:10.1145/1871437.1871442.

[5] E. Adar, J. Teevan, S. T. Dumais, J. L. Elsas, The web changes everything: Understanding the dynamics of web content, in: R. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto, B. B. Cambazoglu (Eds.), Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009, ACM, 2009, pp. 282–291. doi:10.1145/1498759.1498837.

[6] A. Jatowt, K. Tanaka, Large scale analysis of changes in english vocabulary over recent time, in: X.-w. Chen, G. Lebanon, H. Wang, M. J. Zaki (Eds.), 21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012, ACM, 2012, pp. 2523–2526. doi:10.1145/2396761.2398682.

[7] T. Breuer, N. Ferro, N. Fuhr, M. Maistro, T. Sakai, P. Schaer, I. Soboroff, How to measure the reproducibility of system-oriented IR experiments, in: J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J.-R. Wen, Y. Liu (Eds.), Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, ACM, 2020, pp. 349–358. doi:10.1145/3397271.3401036.

[8] O. Chapelle, Y. Zhang, A dynamic bayesian network click model for web search ranking, in: J. Quemada, G. León, Y. S. Maarek, W. Nejdl (Eds.), Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009, ACM, 2009, pp. 1–10. doi:10.1145/1526709.1526711.

[9] N. Craswell, O. Zoeter, M. J. Taylor, B. Ramsey, An experimental comparison of click position-bias models, in: M. Najork, A. Z. Broder, S. Chakrabarti (Eds.), Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008, ACM, 2008, pp. 87–94. doi:10.1145/1341531.1341545.

[10] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3., 1994.

[11] C. Macdonald, N. Tonellotto, Declarative experimentation in information retrieval using PyTerrier, in: K. Balog, V. Setty, C. Lioma, Y. Liu, M. Zhang, K. Berberich (Eds.), ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020, ACM, 2020, pp. 161–168. doi:10.1145/3409256.3409829.

[12] G. Amati, Probability Models for Information Retrieval Based on Divergence from Randomness, 2003. URL: http://theses.gla.ac.uk/1570/.

[13] G. Amati, Frequentist and bayesian approach to information retrieval, in: M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, A. Yavlinsky (Eds.), Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK,

April 10-12, 2006, Proceedings, volume 3936 of *Lecture Notes in Computer Science*, Springer, 2006, pp. 13–24. doi:10.1007/11735106_3.

[14] N. A. Jaleel, J. Allan, W. B. Croft, F. Diaz, L. S. Larkey, X. Li, M. D. Smucker, C. Wade, Umass at TREC 2004: Novelty and HARD, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, volume 500-261 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2004. URL: http://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf.

[15] G. V. Cormack, C. L. A. Clarke, S. Büttcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, J. Zobel (Eds.), Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, ACM, 2009, pp. 758–759. doi:10.1145/1571941.1572114.

[16] E. Bassani, L. Romelli, Ranx.fuse: A python library for metasearch, in: M. A. Hasan, L. Xiong (Eds.), Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, ACM, 2022, pp. 4808–4812. doi:10.1145/3511808.3557207.

[17] O. Khattab, M. Zaharia, ColBERT: Efficient and effective passage search via contextualized late interaction over BERT, in: J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J.-R. Wen, Y. Liu (Eds.), Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, ACM, 2020, pp. 39–48. doi:10.1145/3397271.3401075.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.

[19] G. W. Furnas, T. K. Landauer, L. M. Gomez, S. T. Dumais, The Vocabulary Problem in Human-System Communication, Commun. ACM 30 (1987) 964–971.

[20] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document Ranking with a Pretrained Sequence-to-Sequence Model, in: Document Ranking with a Pretrained Sequence-to-Sequence Model, Association for Computational Linguistics, 2020, pp. 708–718. doi:10.18653/v1/2020.findings-emnlp.63.

[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[22] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated machine reading comprehension dataset, in: T. R. Besold, A. Bordes, A. S. d'Avila Garcez, G. Wayne (Eds.), Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016, volume 1773 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016. URL: https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

[23] R. Pradeep, R. F. Nogueira, J. Lin, The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models, CoRR abs/2101.05667 (2021). URL: https://arxiv.org/abs/2101.05667. arXiv:2101.05667.
[24] R. Nogueira, J. Lin, From doc2query to docTTTTTquery, 2019.
[25] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, CoRR abs/2212.03533 (2022). doi:10.48550/arXiv.2212.03533.
[26] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, 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 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.
[27] C. Buckley, E. M. Voorhees, Retrieval evaluation with incomplete information, in: M. Sanderson, K. Järvelin, J. Allan, P. Bruza (Eds.), SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004, ACM, 2004, pp. 25–32. doi:10.1145/1008992.1009000.
[28] G. N. G. Sáez, P. Mulhem, L. Goeuriot, Towards the evaluation of information retrieval systems on evolving datasets with pivot systems, 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, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 91–102. doi:10.1007/978-3-030-85251-1_8.
[29] T. Breuer, N. Ferro, M. Maistro, P. Schaer, Repro_eval: A python interface to reproducibility measures of system-oriented IR experiments, in: D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II, volume 12657 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 481–486. doi:10.1007/978-3-030-72240-1_51.
[30] M. Fröbe, J. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The Information Retrieval Experiment Platform, in: 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), ACM, 2023.
[31] M. Maistro, T. Breuer, P. Schaer, N. Ferro, An in-depth investigation on the behavior of measures to quantify reproducibility, Inf. Process. Manag. 60 (2023) 103332. doi:10.1016/j.ipm.2023.103332.