HAL
open science

Query Performance Prediction for Neural IR: Are We There Yet?

Guglielmo Faggioli, Thibault Formal, Stefano Marchesin, Stéphane Clinchant, Nicola Ferro, Benjamin Piwowarski

To cite this version:

Guglielmo Faggioli, Thibault Formal, Stefano Marchesin, Stéphane Clinchant, Nicola Ferro, et al.. Query Performance Prediction for Neural IR: Are We There Yet?. European Conference on Information Retrieval, Apr 2023, Dublin (Ireland), Ireland. pp.232-248, 10.1007/978-3-031-28244-7_15 . hal-04290230

HAL Id: hal-04290230

https://hal.sorbonne-universite.fr/hal-04290230

Submitted on 16 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Query Performance Prediction for Neural IR: Are We There Yet?

Guglielmo Faggioli[1], Thibault Formal[2,3], Stefano Marchesin[1], Stéphane Clinchant[2], Nicola Ferro[1], and Benjamin Piwowarski[3,4]

[1] University of Padova, Padova, Italy
[2] Naver Labs Europe, Meylan, France
[3] Sorbonne Université, ISIR Paris, France
[4] CNRS

Abstract. Evaluation in Information Retrieval (IR) relies on post-hoc empirical procedures, which are time-consuming and expensive operations. To alleviate this, Query Performance Prediction (QPP) models have been developed to estimate the performance of a system without the need for human-made relevance judgements. Such models, usually relying on lexical features from queries and corpora, have been applied to traditional sparse IR methods – with various degrees of success. With the advent of neural IR and large Pre-trained Language Models, the retrieval paradigm has significantly shifted towards more semantic signals. In this work, we study and analyze to what extent current QPP models can predict the performance of such systems. Our experiments consider seven traditional bag-of-words and seven BERT-based IR approaches, as well as nineteen state-of-the-art QPPs evaluated on two collections, Deep Learning '19 and Robust '04. Our findings show that QPPs perform statistically significantly worse on neural IR systems. In settings where semantic signals are prominent (e.g., passage retrieval), their performance on neural models drops by as much as 10% compared to bag-of-words approaches. On top of that, in lexical-oriented scenarios, QPPs fail to predict performance for neural IR systems on those queries where they differ from traditional approaches the most.

1 Introduction

The advent of Neural IR (NIR) and Pre-trained Language Models (PLM) induced considerable changes in several central IR research and application areas, with implications that are yet to be fully tamed by the research community. Query Performance Prediction (QPP) is defined as the prediction of the performance of an IR system without human-crafted relevance judgements and is one of the areas the most interested by advancements in NIR and PLM domains. In fact, *i)* PLM can help developing better QPP models, and *ii)* it is not fully clear yet whether current QPP techniques can be successfully applied to NIR. With this paper, we aim to explore the connection between PLM-based first-stage retrieval techniques and the available QPP models. We are interested in

investigating to what extent QPP techniques can be applied to such IR systems, given *i*) their fundamentally different underpinnings compared to traditional lexical IR approaches, *ii*) that they hold the promise to replace – or at least complement – them in multi-stage ranking pipelines. In return, QPP advantages are multi-fold: it can be used to select the best-performing system for a given query, help users in reformulating their needs, or identify pathological queries that require manual intervention from the system administrators. Said otherwise, the need for QPP still holds for NIR methods. Among the plethora of available QPP methods, most of them rely on lexical aspects of the query and the collection. Such approaches have been devised, tested, and evaluated in predicting the performance of lexical bag-of-words IR systems – from now on referred to as Traditional IR (TIR) – with various degrees of success. Recent advances in Natural Language Processing (NLP) led to the advent of PLM-based IR systems, which shifted the retrieval paradigm from traditional approaches based on lexical matching to exploiting contextualized semantic signals – thus alleviating the semantic gap problem. To ease the readability throughout the rest of the manuscript, with an abuse of notation, we use the more general term NIR to explicitly refer to first-stage IR systems based on BERT [13].

At the current time, no large-scale work has been devoted to assessing whether traditional QPP models can be used for NIR systems – which is the goal of this study. We compare the performance of nineteen QPP methods applied to seven traditional TIR systems, with those achieved on seven state-of-the-art first-stage NIR approaches based on PLM. We consider both pre- and post-retrieval QPPs, and include in our analyses post-retrieval QPP models that exploit lexical or semantic signals to compute their predictions. To instantiate our analyses on different scenarios we consider two widely adopted experimental collections: Robust '04 and Deep Learning '19. Our contributions are as follows:

- we apply and evaluate several state-of-the-art QPP approaches to multiple NIR retrievers based on BERT, on Robust '04 and Deep Learning '19;
- we observe a correlation between QPPs performance and how different NIR architectures perform lexical match;
- we show that currently available QPPs perform reasonably well when applied to TIR systems, while they fail to properly predict the performance for NIR systems, even on NIR oriented collections;
- we highlight how such decrease in QPP performance is particularly prominent on queries where TIR and NIR performances differ the most – which are those queries where QPPs would be most beneficial.

The remainder of this paper is organized as follows: Section 2 outlines the main related endeavours. Section 3 details our methodology, while Section 4 contains the experimental setting. Empirical results are reported in Section 5. Section 6 summarizes the main conclusions and future research directions.

2 Related Work

The rise of large PLM like BERT [13] has given birth to a new generation of NIR systems. Initially employed as re-rankers in a standard learning-to-rank framework [35], a real paradigm shift occurred when the first PLM-based retrievers outperformed standard TIR models as candidate generators in a multi-stage ranking setting. For such a task, dense representations, based on a simple pooling of contextualized embeddings, combined with approximate nearest neighbors algorithms, have proven to be both highly effective and efficient [37, 30, 49, 28, 22, 29]. ColBERT [31, 41] avoids this pooling mechanism, and directly models semantic matching at the token level – allowing it to capture finer-grained relevance signals. In the meantime, another research branch brought lexical models up to date, by taking advantage of BERT and the proven efficiency of inverted indices in various manners. Such sparse approaches for instance learn contextualized term weights [10, 34, 55, 33], query or document expansion [36], or both mechanisms jointly [21, 20]. This new wave of NIR systems, which substantially differ from lexical ones – and from each other – demonstrate state-of-the-art results on several datasets, from MS MARCO [3] on which models are usually trained, to zero-shot settings such as the BEIR [46] or LoTTE [41] benchmarks.

A well-known problem linked to IR evaluation is the variation in performance achieved by different IR systems, even on a single query [4, 9]. To partially account for it, a large body of work has focused on predicting the performance that a system would achieve for a given query, using QPP models. Such models are typically divided into pre- and post-retrieval predictors. Traditional pre-retrieval QPPs leverage statistics on the query terms occurrences [26]. For example, SCQ [53], VAR [53] and IDF [8, 42] combine query tokens' occurrence indicators, such as Collection Frequency (CF) and Inverse Document Frequency (IDF), to compute their performance prediction score. Post-retrieval QPPs exploit the results of IR models for the given query [4]. Among them, Clarity [7] compares the language model of the first k retrieved documents with the one of the entire corpus. NQC [43], WIG [54] and SMV [45] exploit the retrieval scores distribution for the top-ranked documents to compute their predictive score. Finally, Utility Estimation Framework (UEF) [44] serves as a general framework that can be instantiated with many of the mentioned predictors, pre-retrieval ones included. Post-retrieval predictors are based on lexical signals – SMV, NQC and WIG rely on the Language Model scores estimated from top-retrieved documents, while Clarity and UEF exploit the language models of the top-k documents.

We further divide QPP models into traditional and neural approaches. Among neural predictors, one of the first approaches is NeuralQPP [50] which computes its predictions by combining semantic and lexical signals using a feed-forward neural network. Notice that NeuralQPP is explicitly designed for TIR and is hence not expected to work better with NIR [50]. A similar approach for Question Answering is NQA-QPP [24], which also relies on three neural components but, unlike NeuralQPP, exploits BERT [13] to embed tokens semantics. Similarly, BERT-QPP [2] encodes semantics via BERT, but directly *fine-tunes* it to predict query performance based on the first retrieved document. Subsequent

approaches extend BERT-QPP by employing a groupwise predictor to jointly learn from multiple queries and documents [5] or by transforming its pointwise regression into a classification task [12]. Since we did not consider multiple formulations, we did not experiment with such approach in our empirical evaluation.

Although traditional QPP methods have been widely used over the years, only few works have been done to apply them on NIR models. Similarly, neural QPP methods – which model the semantic interactions between query and document terms – have been mostly *designed for* and *evaluated on* TIR models. Two noteworthy exceptions concerning the tested IR models are [24] who evaluate the devised QPP on pre-BERT approaches for Question Answering (QA), while [11] assess the performance of their approach on DRMM [23] (pre-BERT) and ColBERT [31] (BERT-based) as NIR models. Hence, there is an urgent need to deepen the evaluation of QPP on state-of-the-art NIR models to understand where we are, what are the challenges, and which directions are more promising.

A third category that can be considered a hybrid between the groups of predictors mentioned above is passage retrieval QPP [38]. In [38], authors exploit lexical signals obtained from passages' language models to devise a predictor meant to better deal with passage retrieval prediction.

3 Methodology

Evaluating Query Performance Predictors. QPP models compute a score for each query, that is expected to correlate with the quality of the retrieval for such query. Traditional evaluation of QPP models relies on measuring the correlation between the predicted QPP scores and the observed performance measured with a traditional IR measure. Typical correlation coefficients include Kendall's τ, Spearman's ρ and the Pearson's r. This evaluation procedure has the drawback of summarizing, through the correlation score, the performance of a QPP model into a single observation for each system and collection [15, 16]. Therefore, Faggioli et al. [15] propose a novel evaluation approach based on the scaled Absolute Rank Error (sARE) measure that, given a query q, is defined as $\text{sARE}(q) = \frac{|R_q^e - R_q^p|}{|Q|}$, where R_q^e and R_q^p are the ranks of the query q induced by the IR measure and the QPP score respectively, over the entire set of queries of size $|Q|$. With "rank" we refer to the ordinal position of the query if we sort all the queries of the collection either by IR performance or prediction score. By switching from a single-point estimation to a distribution of performance, sARE has the advantage of allowing conducting more powerful statistical analyses and carrying out failure analyses on queries where the predictors are particularly bad. To be comparable with previous literature, we report in Sec. 5.1 the performance of the analyzed predictors using the traditional Pearson's r correlation-based evaluation. On the other hand, we use sARE as the evaluation measure for the statistical analyses, to exploit its additional advantages. Such analyses, whose results are reported in Sec. 5.2, are described in the remainder of this section.

ANOVA. To assess the effect induced by NIR systems on QPP performance, we employ the following ANalysis Of VAriance (ANOVA) models. The first model, dubbed MD1, aims at explaining the sARE performance given the predictor, the type of IR model and the collection. Therefore, we define it as follows:

$$sARE_{ijpqr} = \mu + \pi_p + \eta_i + \chi_j + (\eta\chi)_{ij} + \epsilon_{ijpqr}, \tag{MD1}$$

where μ is the grand mean, π_p is the effect of the p-th predictor, η_i represents the type of IR model (either TIR or NIR), χ_j stands for the effect of the j-th collection on QPP's performance, and $(\eta\chi)_{ij}$ describes how much the type of run and the collection interact and ϵ is the associated error.

Secondly, since we are interested in determining the effect of different predictors in interaction with each query, we define a second model, dubbed MD2, that also includes the interaction factor and is formulated as follows:

$$sARE_{ipqr} = \mu + \pi_p + \tau_q + \eta_i + (\pi\tau)_{qp} + (\pi\eta)_{pi} + (\tau\eta)_{iq} + \epsilon_{ipqr}, \tag{MD2}$$

Differently from MD1, we apply MD2 to each collection separately. Therefore, having a single collection, we replace the effect of the collection with τ_q, the effect for the q-th topic. Furthermore, the model includes also all the first-order interactions.

The Strength of Association (SOA) [39] is assessed using ω^2 measure computed as:

$$\omega^2_{<fact>} = \frac{\mathrm{df}_{<fact>} * F_{<fact>}}{\mathrm{df}_{<fact>} * (F_{<fact>} - 1) * N},$$

where N is the number of experimental data-points, $\mathrm{df}_{<fact>}$ is the factor's number of Degrees of Freedom (DF), and $F_{<fact>}$ it the F statistics computed by ANOVA. As a rule-of-thumb, $\omega^2 < 6\%$ indicates a small SOA, $6\% \leq \omega^2 < 14\%$ is a medium-sized effect, while $\omega^2 \geq 14\%$ represent a large-sized effect.

ANOVA Models have been fitted using `anovan` function from the stats MATLAB package. In terms of sample size, depending on the model and collection at hand, we considered 19 predictors, 249 topics in the case of Robust '04 and 43 for Deep Learning '19 and 14 different IR systems for a total of 66234 and 11438 observations for Robust '04 and Deep Learning '19 respectively.

4 Experimental Setup

Our analyses focus on two distinct collections: Robust '04 [47], and TREC Deep Learning 2019 Track (Deep Learning '19) [6]. The collections have respectively 249 and 43 topics each and are based on TIPSTER and MS MARCO passages corpora. Robust '04 is one of the most used collections to test lexical approaches, while providing a reliable benchmark for NIR models [48] – even though they struggle to perform well on this collection, especially when evaluated in a zero-shot setting [46]. Deep Learning '19 concerns passage retrieval from natural questions – the formulation of queries and the nature of the documents (passages)

make the retrieval harder for TIR approaches, while NIR systems tend to have an edge in retrieving relevant documents.

Our main objective is to assess whether existing QPPs are effective in predicting the performance of different state-of-the-art NIR models. As reference points, we consider seven TIR methods: Language Model with Dirichlet (`LMD`) and Jelinek–Mercer (`LMJM`) smoothing [52], BM25, vector space model [40] (`TFIDF`), InExpB2 [1] (`InEB2`), Axiomatic F1-EXP [17] (`AxF1e`), and Divergence From Independence (DFI) [32]. TIR runs have been computed using Lucene. For the NIR methods, we focus on BERT-based first-stage models. We consider state-of-the-art models from the three main families of NIR models, which exhibit different behavior, and thus might respond to QPPs differently. We consider *dense* models, *i*) a "standard" bi-encoder (`bi`) trained with negative log-likelihood, *ii*) TAS-B [28] (`bi-tasb`) whose training relies on topic-sampling and knowledge distillation *iii*) and finally CoCondenser [22] (`bi-cc`) and Contriever [29] (`bi-ct`) which are based on contrastive pre-training. We also consider two models from the *sparse* family: SPLADE [21] (`sp`) with default training strategy, and its improved version SPLADE++ [19, 20] (`sp++`) based on distillation, hard-negative mining and pre-training. We finally consider the *late-interaction* ColBERTv2 [41] (`colb2`). Models are fine-tuned on the MS MARCO passage dataset; given the absence of training queries in Robust '04, they are evaluated in a zero-shot manner, similarly to previous settings [46, 41]. Besides the bi-encoder we trained on our own, we rely on open-source weights available for every model. The advantage of considering multiple TIR and NIR models is that *i)* we achieve more generalizable results: different models, either TIR or NIR perform the best in different scenarios and therefore our conclusions should be as generalizable as possible; *ii)* it allows to achieve more statistical power in the experimental evaluation. We focus our analyses on Normalized Discounted Cumulated Gain (nDCG) with cutoff 10, as it is employed across NIR benchmarks consistently. This is not the typical setting for evaluating traditional QPP – which usually considers Average Precision (AP)@1000. Nevertheless, given our objective – determining how QPP performs on settings where NIR models can be used successfully – we are also interested in selecting the most appropriate measure.

Concerning QPP models, we select the most popular state-of-the-art approaches. In details, we consider 9 pre-retrieval models: Simplified query Clarity Score (SCS) [27], Similarity Collection-Query (SCQ) [53], VAR [53], IDF and Inverse Collection Term Frequency (ICTF) [8, 42]. For SCS, we use the sum aggregation, while for others we use max and mean, which empirically produce the best results. In terms of post-retrieval QPP models, our experiments are based on Clarity [7], Normalized Query Commitment (NQC) [43], Score Magnitude and Variance (SMV) [45], Weighted Information Gain (WIG) [54] and their UEF [44] counterparts. Among post-retrieval predictors, we also include a supervised approach, BERT-QPP [2], using both bi-encoder (*bi*) and cross-encoder (*ce*) formulations. We train BERT-QPP[5] for each IR system on the MS MARCO

[5] We use the implementation provided at https://github.com/Narabzad/BERTQPP

Table 1: nDCG@10 for the selected TIR and NIR systems. NIR outperform traditional approaches on Deep Learning '19, and have comparable performance on Robust '04.

	axF1e	BM25	LMD	LMJM	TFIDF	DFI	InEB2	bi	bi-tasb	bi-cc	bi-ct	sp	sp++	colbv2
Deep Learning '19	0.45	0.48	0.45	0.48	0.37	0.47	0.49	0.64	0.72	0.72	0.67	0.71	0.73	0.75
Robust '04	0.39	0.44	0.43	0.40	0.31	0.44	0.44	0.23	0.45	0.30	0.46	0.39	0.45	0.47

training set, as proposed in [2]. Similarly to what is done for NIR models, we apply BERT-QPP models on Robust '04 queries in a zero-shot manner.

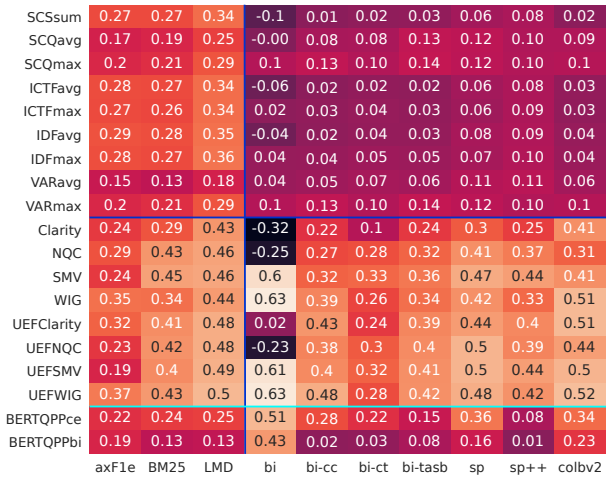
5 Experimental Results

5.1 QPP models performance

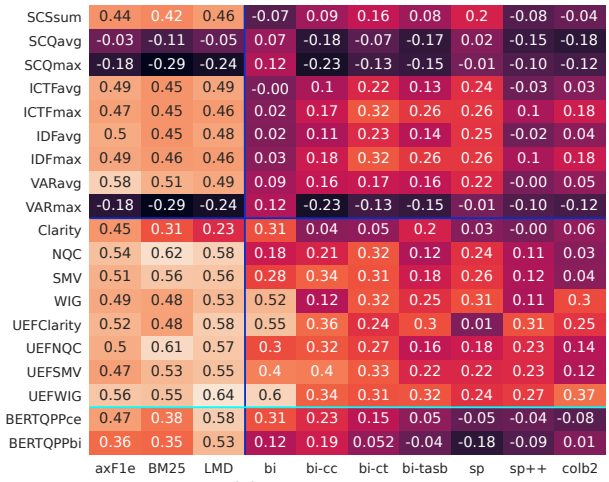
Figures 1a and 1b refer, respectively, to Robust '04 and Deep Learning '19 collections and report the Pearson's r correlation between the scores predicted by the chosen predictors and the nDCG@10, for both TIR and NIR runs[6]. The presence of negative values indicates that some predictors fail in specific contexts and has been observed before in the QPP setting [25]. For Robust '04, we notice that – following previous literature – pre-retrieval (top) predictors (mean correlation: 15.9%) tend to perform 52.3% worse than post-retrieval ones (bottom) (mean correlation: 30.2%). Pre-retrieval results are in line with previous literature [51]. The phenomenon is more evident (darker colors) for NIR runs (right) than TIR ones (left). Pre-retrieval predictors fail in predicting the performance of NIR systems (mean correlation 6.2% vs 25.6% for TIR), while in general, to our surprise, we notice that post-retrieval predictors tend to perform similarly on TIR and NIR (34.5% vs 32.3%) – with some exceptions. For instance, for `bi`, post-retrieval predictors either perform extremely well or completely fail. This happens particularly on Clarity, NQC, and their UEF counterparts. Note that `bi` is the worst performing approach on Robust '04, with 23% of nDCG@10 – the second worst is `bi-cc` which achieves 30% nDCG@10.

The patterns observed for Robust '04 hold only partially on Deep Learning '19. For example, we notice again that pre-retrieval predictors (mean correlation: 14.7%) perform 58.3% worse than post-retrieval ones (mean correlation: 35.3%). On the contrary, the difference in performance is far more evident between NIR and TIR. On TIR runs, almost all predictors perform particularly well (mean correlation: 38.1%) – even better than on Robust '04 collection. The only three exceptions are SCQ (both in avg and max formulations) and VAR using max formulation. Conversely, on NIR the performance is overall lower (13.1%) and relatively more uniform between pre- (5.4%) and post-retrieval (19.9%) models. In absolute value, maximum correlation achieved by pre-retrieval predictors for NIR on Deep Learning '19 is much higher than the one achieved on Robust '04,

[6] Additional IR measures and correlations, as well as full ANOVA tables are available at: https://github.com/guglielmof/ECIR2023-QPP



(a) Robust '04



(b) Deep Learning '19

Fig. 1: Pearson's r correlation observed for different pre (top) and post (bottom) retrieval predictors on lexical (left) and neural (right) runs. To avoid cluttering, we report the results for the 3 main TIR models, other models achieve highly similar results.

especially for `bi-ct`, `sp`, and `bi-tasb` runs. On the other hand, post-retrieval predictors, perform worse than on the Robust '04. The only exception to this pattern is again represented by `bi`, on which some post-retrieval predictors, namely WIG, UEFWIG, and UEFClarity work surprisingly well. The supervised BERT-QPP shows a trend similar to other post-retrieval predictors on Deep Learning '19 (42.3% mean correlation against 52.9% respectively) for what concerns TIR, with performance in line with the one reported in [2]. This is exactly the setting where BERT-QPP has been devised and tested. If we focus on Deep Learning '19 and NIR systems, its performance (mean correlation: 4.5%) is far lower than

Table 2: Pearson's r QPP performance for three versions of `sp++` applied on Robust '04, with varying degree of sparsity ($\texttt{sp++}_2 \succ \texttt{sp++}_1 \succ \texttt{sp++}_0$ in terms of sparsity). The more "lexical" are the models, the better QPP performs. $\mathbf{d}_l$ and $\mathbf{q}_l$ represent respectively the average document/query sizes (i.e. non-zero dimensions in SPLADE) on Robust '04.

	$\mathbf{d_l}/\mathbf{q_l}$	Clarity	NQC	SMV	WIG	UEFClarity	UEFNQC	UEFSMV	UEFWIG
$\texttt{sp++}_2$	55/22	0.26	0.31	0.46	0.42	0.44	0.4	0.48	0.5
$\texttt{sp++}_1$	79/29	0.2	0.34	0.47	0.35	0.38	0.4	0.46	0.43
$\texttt{sp++}_0$	204/45	0.25	0.37	0.44	0.33	0.4	0.39	0.44	0.42

those of other post-retrieval predictors (mean correlation without BERT-QPP: 23.8%). Finally, its performance on Robust '04 – applied in zero-shot – is considerably lower compared to other post-retrieval approaches.

Interestingly, on Robust '04, post-retrieval QPPs achieve, on average, top performance on the late interaction model (`colb2`), followed by sparse approaches (`sp` and `sp++`). Finally, excluding `bi`, where predictors achieve extremely inconsistent performance, dense approaches are those where QPP perform the worst. In this sense, the performance that QPP methods achieve on NIR systems seems to correlate with the importance these systems give to lexical signals. In this regard, Formal et al. [20] observed how late-interaction and sparse architectures tend to rely more on lexical signals, compared to dense ones. To further corroborate this observation, we apply the predictors to three versions of SPLADE++ with various levels of sparsit as controlled by the regularization hyperparameter. Increasing the sparsity of representations leads to models that cannot rely as much on expansion – emphasizing the importance given to lexical signals in defining the document ranking. Therefore, as a first approximation, we can deem sparser methods to be also more lexical. Given the low performance achieved by pre-retrieval QPPs, we focus this analysis on post-retrieval methods only. Table 2 shows the Pearson's r for the considered predictors and different SPLADE++ versions. Interestingly, in the majority of the cases, QPPs perform the best for the sparser version ($\texttt{sp++}_2$), followed by $\texttt{sp++}_1$ and $\texttt{sp++}_0$ – which is the one used in Fig. 1. There are a few switches, often associated with very close correlation values (SMV and UEFClarity). Only one predictor, NQC, completely reverses the order. This goes in favour of our hypothesis that indeed QPP performance tends to correlate with the degree of lexicality of the NIR approaches. Although not directly comparable, following this line of thought, `sp`, being handled better by QPPs (cfr. Fig. 1a), is more lexical than all the `sp++` versions considered: this is reasonable, given the different training methodology. Finally, `colb2`, being the method where QPPs achieve the best performance, might be the one that, at least for what concerns the Robust '04 collection, gives the highest importance to lexical signals – in line with what was observed in [21].

5.2 ANOVA Analysis

To further statistically quantify the phenomena observed in the previous subsection, we apply MD1 to our data, considering both collections at once. From a

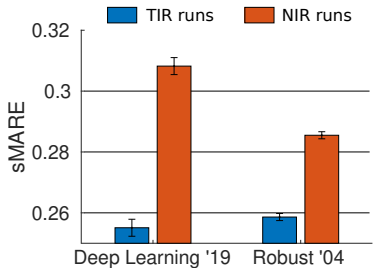


Fig. 2: Comparison between the mean sARE (sMARE) achieved over TIR or NIR when changing the corpus. Observe the large distance between results on NIR – especially for Deep Learning '19 – compared to the one on TIR runs.

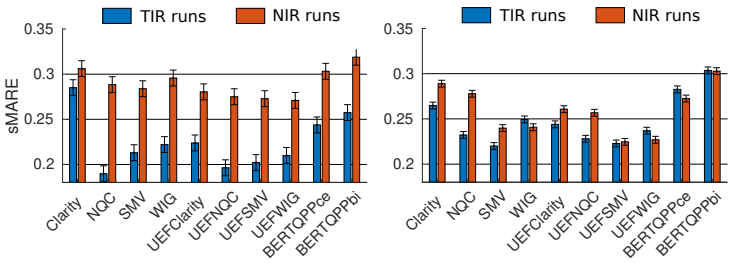
quantitative standpoint, we notice that all the factors included in the model are statistically significant (p-value $< 10^{-4}$). In terms of SOA, the collection factor has a small effect (0.02%). The run type, on the other hand, impacts for $\omega^2 = 0.48\%$. Finally, the interaction between the collection and run type, although statistically significant, has a small impact on the performance ($\omega^2 = 0.05\%$): in both collections QPPs perform better on TIR models. All factors are significant but have small-size effects. This is in contrast with what was observed for the performance of IR systems [18, 9], where most of the SOA range between medium to large. Nevertheless, it is in line with what was observed by Faggioli et al. [15] for the performance QPP methods, who showed that all the factors besides the topic are small to medium. A second observation is that it is likely that the small SOAs are due to a model unable to accrue for all the aspects of the problem – more factors should be considered. Model MD2, introducing also the topic effect, allows for further investigation of this hypothesis.

We are now interested in breaking down the performance of the predictors according to the collection and type of run. Figure 2 reports the average performance (measured with sMARE, the lower the better) for QPPs applied on NIR or TIR runs over different collections, with their confidence intervals as computed using ANOVA. Interestingly, regardless of the type of collection, the performance achieved by predictors on NIR models will *on average* be worse than those achieved on TIR runs. QPP models perform better on TIR than NIR on both collections: this explains the small interaction effect between collections and run types. Secondly, there is no statistical difference between the performance achieved by QPPs applied to TIR models when considering Deep Learning '19 and Robust '04– the confidence intervals are overlapping. This goes in contrast with what happens on Robust '04 and Deep Learning '19 when considering NIR models: QPPs approaches applied on the latter dataset perform by far worse than on the former.

While *on average* we will be less satisfied by QPP predictors applied to NIR regardless of the type of collection, there might be some noticeable exceptions of good performing predictors also for NIR systems. To verify this hypothesis,

Table 3: p-values and ω^2 SOA using MD2 on each collection

	Deep Learning '19		Robust '04	
	p-value	ω^2	p-value	ω^2
topic	$< 10^{-4}$	22.5%	$< 10^{-4}$	24.0%
qpp	$< 10^{-4}$	1.65%	$< 10^{-4}$	2.21%
run type	$< 10^{-4}$	4.35%	$< 10^{-4}$	0.11%
topic*qpp	$< 10^{-4}$	22.7%	$< 10^{-4}$	17.2%
topic*run type	$< 10^{-4}$	15.2%	$< 10^{-4}$	10.0%
qpp*run type	0.0012	0.23%	$< 10^{-4}$	0.30%

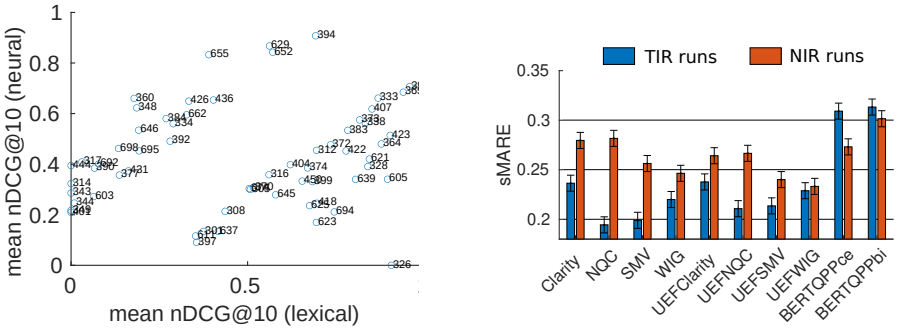


(a) Deep Learning '19 (b) Robust '04

Fig. 3: sMARE observed for different predictors on Deep Learning '19 (left) and Robust '04 (right). On Deep Learning '19, predictors behave differently on TIR and NIR runs, while they are more uniform on Robust '04.

we apply MD2 to each collection separately, and measure what happens to each predictor individually[7]. Table 3 reports the p-values and ω^2 SOA for the factors included in MD2, while Figure 3 depicts the phenomena visually. We observe that, concerning Deep Learning '19, the run type (TIR or NIR) is significant, while the interaction between the predictor and the run type is small: indeed predictors always perform better on TIR runs than on NIR ones. The only model that behaves slightly differently is Clarity, with far closer performance for both classes of runs – this can be explained by the fact that Clarity is overall the worst-performing predictor. Notice that, the best predictor on TIR runs – NQC – performs almost 10% worse on NIR ones. Finally, we notice a large-size interaction between topics and QPP models – even bigger than the topic or QPP themselves. This indicates that whether a model will be better than another strongly depends on the topic considered. An almost identical pattern was observed also in [15]. Therefore, to improve QPP's generalizability, it is important not only to address challenges caused by differences in NIR and TIR but also to take into consideration the large variance introduced by topics. We analyze more in detail this variance later, where we consider only "semantically defined" queries.

[7] To avoid cluttering, we report the subsequent analyses only for post-retrieval predictors – similar observations hold for pre-retrieval ones.



(a) selected topics

(b) Robust '04

Fig. 4: left: topics selected to maximize the difference between lexical and neural models; right: results of MD2 applied on Robust '04 considering only the selected topics.

If we consider Robust '04, the behaviour changes deeply: Figure 3 shows that predictors performances are much more similar for TIR and NIR runs compared to Deep Learning '19. This is further highlighted by the far smaller ω^2 for run type on Robust '04 in Table 3 – 4.35% against 0.11%. The widely different pattern between Deep Learning '19 and Robust '04 suggests that current QPPs are doomed to fail when used to predict the performance of IR approaches that learned the semantics of a collection – which is the case for Deep Learning '19 that was used to fine-tune the models. Current QPPs evaluate better IR approaches that rely on lexical clues. Such approaches include both TIR models and NIR models applied in a zero-shot fashion, as it is the case for Robust '04. Thus, QPP models are expected to fail where NIR models behave differently from the TIR ones. This poses at stake one of the major opportunities provided by QPP: if we fail in predicting the performance of NIR models where they behave differently from TIR ones, then a QPP cannot be safely used to carry out model selection. To further investigate this aspect, we carry out the following analysis: we select from Robust '04 25% of the queries that are mostly "semantically defined" and rerun MD2 on the new set of topics. We call "semantically defined" those queries where NIR behave, on average, oppositely w.r.t. the TIR, either failing or succeeding at retrieving documents. In other terms, we select queries in the top quartile for the absolute difference in performance (nDCG), averaged over all TIR or NIR models.

Figure 4a shows the performance of topics that maximize the difference between TIR and NIR and can be considered as more "semantically defined" [14]. There are 62 topics selected (25% of the 249 topics available on Robust '04). Of these, 35 topics are better handled by TIR models, while 27 obtain better nDCG if dealt with NIR rankers. If we consider the results of applying MD2 on this set of topics, we notice that compared to Robust '04 (Table 3, last column) the effect of the different QPPs increases to 2.29%: on these topics, there is more difference between different predictors. The interaction between predictors and run types

grows from 0.30% to 0.91%. Furthermore, the effect of the run type grows from 0.11% to 0.67% – 6 times bigger. On the selected topics, arguably those where a QPP is the most useful to help select the right model, using NIR systems has a negative impact (6 times bigger) on the performance of QPPs. Figure 4b, compared to Figure 3b, is more similar to Figure 3a – using only topics that are highly semantically defined, we get similar patterns as those observed for Deep Learning '19 on Figure 3a. The only methods that behave differently are BERT-QPP approaches, whose performance is better on NIR runs than on TIR ones, but are the worst approaches in terms of predictive capabilities for both run types. In this sense, even though the contribution of the semantic signals appears to highly important to define new models with improved performance in the NIR setting, it does not suffice to compensate for current QPPs limitations.

6 Conclusion and Future Work

With this work, we assessed to what extent current QPPs are applicable to the recent family of first-stage NIR models based on PLM. To verify that, we evaluated 19 diverse QPP models, used on seven traditional bag-of-words lexical models (TIR) and seven first-stage NIR methods based on BERT, applied to the Robust '04 and Deep Learning '19 collections. We observed that if we consider a collection where NIR systems had the chance to learn the semantics – i.e., Deep Learning '19 – QPPs are effective in predicting TIR systems performance, but fail in dealing with NIR ones. Secondly, we considered Robust '04. In this collection, NIR models were applied in a zero-shot fashion, and thus behave similarly to TIR models. In this case, we observed that QPPs tend to work better on NIR models than in the previous scenario, but they fail on those topics where NIR and TIR models differ the most. This, in turn, impairs the possibility of using QPP models to choose between NIR and TIR approaches where it is most needed. On the other hand, semantic QPP approaches such as BERT-QPP do not solve the problem: being devised and tested on lexical IR systems, they work properly on such category of approaches but fail on neural systems. These results highlight the need for QPPs specifically tailored to Neural IR.

As future work, we plan to extend our analysis by considering other factors, such as the query variations to understand the impact that changing how a topic is formulated has on QPP. Furthermore, we plan to devise QPP methods explicitly designed to synergise with NIR models, but that also take into consideration the large variance introduced by topics.

Acknowledgements The work was partially supported by University of Padova Strategic Research Infrastructure Grant 2017: "CAPRI: Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione", ExaMode project, as part of the EU H2020 program under Grant Agreement no. 825292.

References

1. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval based on measuring the Divergence From Randomness. ACM Trans. Inf. Syst **20**(4), 357–389 (2002)
2. Arabzadeh, N., Khodabakhsh, M., Bagheri, E.: BERT-QPP: Contextualized Pre-trained transformers for Query Performance Prediction. In: CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021. pp. 2857–2861 (2021)
3. Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., Wang, T.: MS MARCO: A Human Generated MAchine Reading COmprehension Dataset (2016)
4. Carmel, D., Yom-Tov, E.: Estimating the Query Difficulty for Information Retrieval. Morgan & Claypool Publishers (2010)
5. Chen, X., He, B., Sun, L.: Groupwise Query Performance Prediction with BERT. In: Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II. Lecture Notes in Computer Science, vol. 13186, pp. 64–74 (2022)
6. Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M., Soboroff, I.: TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime. In: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 2369–2375 (2021)
7. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland. pp. 299–306 (2002)
8. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: A Language Modeling Framework for Selective Query Expansion. Tech. rep., CIIR, UMass (2004)
9. Culpepper, J.S., Faggioli, G., Ferro, N., Kurland, O.: Topic Difficulty: Collection and Query Formulation Effects. ACM Trans. Inf. Syst. **40**(1), 19:1–19:36 (2022)
10. Dai, Z., Callan, J.: Context-Aware Term Weighting For First Stage Passage Retrieval. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020. pp. 1533–1536 (2020)
11. Datta, S., Ganguly, D., Mitra, M., Greene, D.: A Relative Information Gain-Based Query Performance Prediction Framework with Generated Query Variants. ACM Trans. Inf. Syst. pp. 1–31 (2022)
12. Datta, S., MacAvaney, S., Ganguly, D., Greene, D.: A 'Pointwise-Query, Listwise-Document' based Query Performance Prediction Approach. In: SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022. pp. 2148–2153 (2022)
13. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: 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). pp. 4171–4186 (2019)
14. Faggioli, G., Marchesin, S.: What makes a query semantically hard? In: Proceedings of the Second International Conference on Design of Experimental Search

& Information REtrieval Systems, Padova, Italy, September 15-18, 2021. CEUR Workshop Proceedings, vol. 2950, pp. 61–69. CEUR-WS.org (2021), http://ceur-ws.org/Vol-2950/paper-06.pdf

15. Faggioli, G., Zendel, O., Culpepper, J.S., Ferro, N., Scholer, F.: An Enhanced Evaluation Framework for Query Performance Prediction. In: Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I. vol. 12656, pp. 115–129 (2021)
16. Faggioli, G., Zendel, O., Culpepper, J.S., Ferro, N., Scholer, F.: sMARE: a new paradigm to evaluate and understand query performance prediction methods. Inf. Retr. J. **25**(2), 94–122 (2022)
17. Fang, H., Zhai, C.: An exploration of axiomatic approaches to information retrieval. In: SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005. pp. 480–487 (2005)
18. Ferro, N., Silvello, G.: Toward an anatomy of IR system component performances. J. Assoc. Inf. Sci. Technol. **69**(2), 187–200 (2018)
19. Formal, T., Lassance, C., Piwowarski, B., Clinchant, S.: SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval. CoRR **abs/2109.10086** (2021)
20. Formal, T., Lassance, C., Piwowarski, B., Clinchant, S.: From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective. In: SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022. pp. 2353–2359 (2022)
21. Formal, T., Piwowarski, B., Clinchant, S.: SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 2288–2292 (2021)
22. Gao, L., Callan, J.: Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. pp. 2843–2853 (2022)
23. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A Deep Relevance Matching Model for Ad-hoc Retrieval. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016. pp. 55–64 (2016)
24. Hashemi, H., Zamani, H., Croft, W.B.: Performance Prediction for Non-Factoid Question Answering. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2019, Santa Clara, CA, USA, October 2-5, 2019. pp. 55–58 (2019)
25. Hauff, C.: Predicting the effectiveness of queries and retrieval systems. SIGIR Forum **44**(1), 88 (2010)
26. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008. pp. 1419–1420 (2008)
27. He, J., Larson, M.A., de Rijke, M.: Using Coherence-Based Measures to Predict Query Difficulty. In: Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings. vol. 4956, pp. 689–694 (2008)

28. Hofstätter, S., Lin, S., Yang, J., Lin, J., Hanbury, A.: Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 113–122 (2021)
29. Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E.: Towards Unsupervised Dense Information Retrieval with Contrastive Learning. CoRR **abs/2112.09118** (2021)
30. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense Passage Retrieval for Open-Domain Question Answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781 (2020)
31. Khattab, O., Zaharia, M.: ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020. pp. 39–48. ACM (2020)
32. Kocabas, I., Dinçer, B.T., Karaoglan, B.: A nonparametric term weighting method for information retrieval based on measuring the divergence from independence. Inf. Retr. **17**(2), 153–176 (2014)
33. Lin, J., Ma, X.: A Few Brief Notes on DeepImpact, COIL, and a Conceptual Framework for Information Retrieval Techniques. CoRR **abs/2106.14807** (2021)
34. Mallia, A., Khattab, O., Suel, T., Tonellotto, N.: Learning Passage Impacts for Inverted Indexes. In: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 1723–1727 (2021)
35. Nogueira, R.F., Cho, K.: Passage re-ranking with BERT. CoRR **abs/1901.04085** (2019)
36. Nogueira, R.F., Yang, W., Lin, J., Cho, K.: Document Expansion by Query Prediction. CoRR **abs/1904.08375** (2019)
37. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 3980–3990 (2019)
38. Roitman, H.: An extended query performance prediction framework utilizing passage-level information. In: Song, D., Liu, T., Sun, L., Bruza, P., Melucci, M., Sebastiani, F., Yang, G.H. (eds.) Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2018, Tianjin, China, September 14-17, 2018. pp. 35–42. ACM (2018). https://doi.org/10.1145/3234944.3234946, https://doi.org/10.1145/3234944.3234946
39. Rutherford, A.: ANOVA and ANCOVA: a GLM approach. John Wiley & Sons (2011)
40. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988)
41. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., Zaharia, M.: ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022. pp. 3715–3734 (2022)

42. Scholer, F., Williams, H.E., Turpin, A.: Query association surrogates for Web search. J. Assoc. Inf. Sci. Technol. **55**(7), 637–650 (2004)
43. Shtok, A., Kurland, O., Carmel, D.: Predicting Query Performance by Query-Drift Estimation. In: Advances in Information Retrieval Theory, Second International Conference on the Theory of Information Retrieval, ICTIR 2009, Cambridge, UK, September 10-12, 2009, Proceedings. vol. 5766, pp. 305–312 (2009)
44. Shtok, A., Kurland, O., Carmel, D.: Using statistical decision theory and relevance models for query-performance prediction. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010. pp. 259–266 (2010)
45. Tao, Y., Wu, S.: Query Performance Prediction By Considering Score Magnitude and Variance Together. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014. pp. 1891–1894 (2014)
46. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual (2021)
47. Voorhees, E.M.: The TREC robust retrieval track. SIGIR Forum **39**(1), 11–20 (2005)
48. Voorhees, E.M., Soboroff, I., Lin, J.: Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models? CoRR **abs/2201.11086** (2022)
49. Xiong, L., Xiong, C., Li, Y., Tang, K., Liu, J., Bennett, P.N., Ahmed, J., Overwijk, A.: Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (2021)
50. Zamani, H., Croft, W.B., Culpepper, J.S.: Neural Query Performance Prediction using Weak Supervision from Multiple Signals. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018. pp. 105–114 (2018)
51. Zendel, O., Shtok, A., Raiber, F., Kurland, O., Culpepper, J.S.: Information needs, queries, and query performance prediction. In: Piwowarski, B., Chevalier, M., Gaussier, É., Maarek, Y., Nie, J., Scholer, F. (eds.) Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019. pp. 395–404. ACM (2019). https://doi.org/10.1145/3331184.3331253, https://doi.org/10.1145/3331184.3331253
52. Zhai, C.: Statistical Language Models for Information Retrieval: A Critical Review. Found. Trends Inf. Retr. **2**(3), 137–213 (2008)
53. Zhao, Y., Scholer, F., Tsegay, Y.: Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. In: Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings. vol. 4956, pp. 52–64 (2008)
54. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007. pp. 543–550 (2007)
55. Zhuang, S., Zuccon, G.: TILDE: Term Independent Likelihood moDEl for Passage Re-ranking. In: SIGIR '21: The 44th International ACM SIGIR Conference on

Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 1483–1492 (2021)