On the Limitations of Query Performance Prediction for Neural IR[*]

Discussion paper

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, France*

Abstract

The evaluation of Information Retrieval (IR) relies on human-made relevance assessments whose collection is time-consuming and expensive. To alleviate this limitation, Query Performance Prediction (QPP) models have been developed to estimate system performance without relying on human-made relevance judgements. QPP models have been applied to traditional IR methods with varying success. The shift towards semantic signals thanks to Neural IR (NIR) models has changed the retrieval paradigm. In this study, we investigate the ability of current QPP models to predict the performance of NIR systems. We evaluate seven traditional IR systems and seven NIR (BERT-based) approaches, as well as nineteen QPPs, on two collections: Deep Learning '19 and Robust '04. Our results highlight that QPPs perform significantly worse on NIR systems. When semantic signals are prevalent, such as in passage retrieval, their performance on neural models decreases by up to 10% compared to bag-of-words approaches.

1. Introduction

The advent of Neural IR (NIR) and Pre-trained Language Models (PLM) induced considerable changes in several central Information Retrieval (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 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 [2]. At the current time, no large-scale work has been devoted to assessing whether traditional QPP models can be used for NIR systems [3]. To address such a gap, 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. As contributions:

- 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 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 a decrease in QPP performance is particularly prominent on queries where TIR and NIR performances differ the most.

The remainder of this paper is organized as follows: Section 2 outlines the main related endeavours. Section 3 details our methodology and experimental setting. Empirical results are reported in Section 4. Section 5 draws the conclusions.

2. Related Work

Large PLMs like BERT [2] have given birth to a new generation of NIR systems. Indeed, dense representations based on contextualized embeddings, combined with approximate nearest neighbors algorithms, have proven to be effective and efficient first-stage retrieval approaches [4, 5, 6, 7, 8, 9]. In the meantime, another research branch brought lexical models by learning contextualized term weights [10, 11, 12, 13], query or document expansion [14], or both mechanisms jointly [15, 16]. This new wave of NIR systems demonstrate state-of-the-art results on several datasets [17, 18, 19].

A well-known problem linked to IR evaluation is the variation in performance achieved by different IR systems, even on a single query [20, 21]. To 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 [22]. For example, SCQ [23], VAR [23] and IDF [24, 25] 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 [20]. Among them, Clarity [26] compares the language model of the first k retrieved documents with the one of the entire corpus. NQC [27], WIG [28] and SMV [29] exploit the retrieval scores distribution for the top-ranked documents to compute their predictive score. Finally, Utility Estimation Framework (UEF) [30] serves as a general framework that can be instantiated with many of the mentioned predictors, pre-retrieval ones included. We further divide QPP models into traditional and neural approaches. Among neural predictors, one of the first approaches is NeuralQPP [31] which computes its predictions by combining semantic and lexical signals using a feed-forward neural network. A similar approach for Question Answering is NQA-QPP [32], which also relies on three neural components but, unlike NeuralQPP, exploits BERT [2] to embed tokens semantics. Similarly, BERT-QPP [33] encodes semantics via BERT, but directly *fine-tunes* it to predict query performance based on the first retrieved document. Only a little work has been done to apply traditional QPP on NIR models [32, 34]. Similarly, neural QPP methods – which model the semantic interactions between query and document terms – have been mostly *designed for* and *evaluated on* TIR 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.

3. Experimental Methodology and Setup

To assess the effect induced by NIR systems on QPP performance, we employ the following ANalysis Of VAriance (ANOVA) models, using sARE [35, 36] as a performance measure. 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) [37]

is assessed using ω^2. 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.

Our analyses focus on Robust '04 [38], and TREC Deep Learning 2019 Track (Deep Learning '19) [39] collections. 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 [40].Deep Learning '19 concerns passage retrieval from natural questions making the retrieval harder for TIR approaches, while NIR systems tend to have an edge in retrieving relevant documents.

As reference points, we consider seven TIR methods: Language Model with Dirichlet (LMD) and Jelinek–Mercer (LMJM) smoothing [41], BM25, vector space model [42] (TFIDF), InExpB2 [43] (InEB2), Axiomatic F1-EXP [44] (AxF1e), and Divergence From Independence (DFI) [45]. 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. We consider *dense* models, *i*) a "standard" bi-encoder (bi) trained with negative log-likelihood, *ii*) TAS-B [7] (bi-tasb) whose training relies on topic-sampling and knowledge distillation *iii*) and finally CoCondenser [8] (bi-cc) and Contriever [9] (bi-ct) which are based on contrastive pre-training. We also consider two models from the *sparse* family: SPLADE [15] (sp) with default training strategy, and its improved version SPLADE++ [46, 16] (sp++) based on distillation, hard-negative mining and pre-training. We finally consider the *late-interaction* ColBERTv2 [19] (colb2). Models are fine-tuned on the MS MARCO passage dataset and applied in a zero-shot manner on Robust '04 [18, 19].We focus our analyses on Normalized Discounted Cumulated Gain (nDCG) with cutoff 10, as it is employed across NIR benchmarks consistently.

Concerning QPP models, we consider 9 pre-retrieval models: Simplified query Clarity Score (SCS) [47], Similarity Collection-Query (SCQ) [23], VAR [23], IDF and Inverse Collection Term Frequency (ICTF) [24, 25]. 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 [26], Normalized Query Commitment (NQC) [27], Score Magnitude and Variance (SMV) [29], Weighted Information Gain (WIG) [28] and their UEF [30] counterparts. Among post-retrieval predictors, we also include a supervised approach, BERT-QPP [33], using both bi-encoder (*bi*) and cross-encoder (*ce*) formulations. We train BERT-QPP[1] for each IR system on the MS MARCO training set, as proposed in [33]. Similarly to what is done for NIR models, we apply BERT-QPP models on Robust '04 queries in a zero-shot manner.

4. Experimental Results

Figures 1a and 1b refer, respectively, to Robust '04 and Deep Learning '19 collections and report the Pearson's r correlation between prediction scores and nDCG@10, for both TIR and NIR runs. For Robust '04, we notice that pre-retrieval (top) predictors (mean correlation: 15.9%) tend to perform 52.3% worse than post-retrieval ones (bottom) (mean correlation: 30.2%). 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 we notice that post-retrieval predictors tend to perform similarly

---

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

(a) Robust '04

	axF1e	BM25	LMD	bi	bi-cc	bi-ct	bi-tasb	sp	sp++	colbv2
SCSsum	0.27	0.27	0.34	-0.1	0.01	0.02	0.03	0.06	0.08	0.02
SCQavg	0.17	0.19	0.25	-0.00	0.08	0.08	0.13	0.12	0.10	0.09
SCQmax	0.2	0.21	0.29	0.1	0.13	0.10	0.14	0.12	0.10	0.1
ICTFavg	0.28	0.27	0.34	-0.06	0.02	0.02	0.02	0.06	0.08	0.03
CTFmax	0.27	0.26	0.34	0.02	0.03	0.04	0.03	0.06	0.09	0.03
IDFavg	0.29	0.28	0.35	-0.04	0.02	0.04	0.03	0.08	0.09	0.04
IDFmax	0.28	0.27	0.36	0.04	0.04	0.05	0.05	0.07	0.10	0.04
VARavg	0.15	0.13	0.18	0.04	0.05	0.07	0.06	0.11	0.11	0.06
VARmax	0.2	0.21	0.29	0.1	0.13	0.10	0.14	0.12	0.10	0.1
Clarity	0.24	0.29	0.43	-0.32	0.22	0.1	0.24	0.3	0.25	0.41
NQC	0.29	0.43	0.46	-0.25	0.27	0.28	0.32	0.41	0.37	0.31
SMV	0.24	0.45	0.46	0.6	0.32	0.33	0.36	0.47	0.44	0.41
WIG	0.35	0.34	0.44	0.63	0.39	0.26	0.34	0.42	0.33	0.51
UEFClarity	0.32	0.41	0.48	0.02	0.43	0.24	0.39	0.44	0.4	0.51
UEFNQC	0.23	0.42	0.48	-0.23	0.38	0.3	0.4	0.5	0.39	0.44
UEFSMV	0.19	0.4	0.49	0.61	0.4	0.32	0.41	0.5	0.44	0.5
UEFWIG	0.37	0.43	0.5	0.63	0.48	0.28	0.42	0.48	0.42	0.52
RTQPPce	0.22	0.24	0.25	0.51	0.28	0.22	0.15	0.36	0.08	0.34
RTQPPbi	0.19	0.13	0.13	0.43	0.02	0.03	0.08	0.16	0.01	0.23

(b) Deep Learning '19

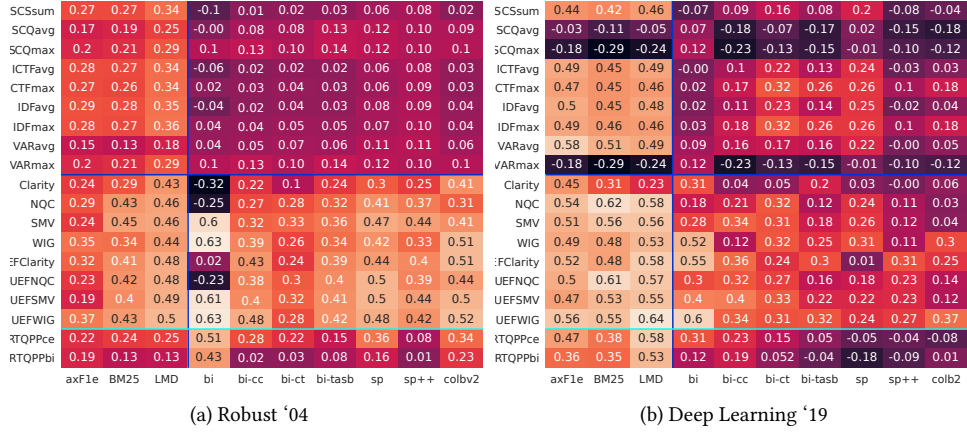|	axF1e	BM25	LMD	bi	bi-cc	bi-ct	bi-tasb	sp	sp++	colb2
SCSsum	0.44	0.42	0.46	-0.07	0.09	0.16	0.08	0.2	-0.08	-0.04
SCQavg	-0.03	-0.11	-0.05	0.07	-0.18	-0.07	-0.17	0.02	-0.15	-0.18
SCQmax	-0.18	-0.29	-0.24	0.12	-0.23	-0.13	-0.15	-0.01	-0.10	-0.12
ICTFavg	0.49	0.45	0.49	-0.00	0.1	0.22	0.13	0.24	-0.03	0.03
CTFmax	0.47	0.45	0.46	0.02	0.17	0.32	0.26	0.26	0.1	0.18
IDFavg	0.5	0.45	0.48	0.02	0.11	0.23	0.14	0.25	-0.02	0.04
IDFmax	0.49	0.46	0.46	0.03	0.18	0.32	0.26	0.26	0.1	0.18
VARavg	0.58	0.51	0.49	0.09	0.16	0.17	0.16	0.22	0.07	0.03
VARmax	-0.18	-0.29	-0.24	0.12	-0.23	-0.13	-0.15	-0.01	-0.10	-0.12
Clarity	0.45	0.31	0.23	0.31	0.04	0.05	0.2	0.03	-0.00	0.06
NQC	0.54	0.62	0.58	0.18	0.21	0.32	0.12	0.24	0.11	0.03
SMV	0.51	0.56	0.56	0.28	0.34	0.31	0.18	0.26	0.12	0.04
WIG	0.49	0.48	0.53	0.52	0.12	0.32	0.25	0.31	0.11	0.3
UEFClarity	0.52	0.48	0.58	0.55	0.36	0.24	0.3	0.01	0.31	0.25
UEFNQC	0.5	0.61	0.57	0.3	0.32	0.27	0.16	0.18	0.23	0.14
UEFSMV	0.47	0.53	0.55	0.4	0.4	0.33	0.22	0.22	0.23	0.12
UEFWIG	0.56	0.55	0.64	0.6	0.34	0.31	0.32	0.24	0.27	0.37
RTQPPce	0.47	0.38	0.58	0.31	0.23	0.15	0.05	-0.05	-0.04	-0.08
RTQPPbi	0.36	0.35	0.53	0.12	0.19	0.052	-0.04	-0.18	-0.09	0.01

Figure 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.

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. Conversely, on NIR the performance is overall lower (13.1%) and relatively more uniform between pre- (5.4%) and post-retrieval (19.9%) models. The maximum correlation achieved by pre-retrieval predictors for NIR on Deep Learning '19 is much higher than the one achieved on Robust '04, 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. 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, 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. 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 [33]. 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 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. To further statistically quantify the phenomena observed in the previous subsection, we apply MD1 to our data, considering both collections at once. From a quantitative standpoint, we notice that all the
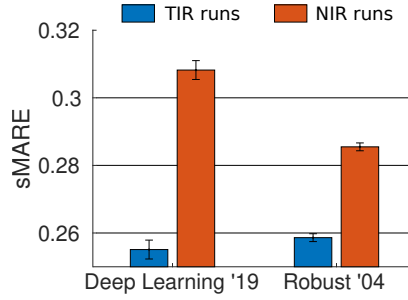
Figure 2: Comparison between the mean sARE (sMARE) achieved over TIR or NIR on different corpora.

Table 1
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%

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 [48, 21], where most of the SOA range between medium to large. Nevertheless, it is in line with what was observed by Faggioli et al. [35] for the performance QPP methods, who showed that all the factors besides the topic are small to medium.

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. The performance achieved by predictors on NIR models is *on average* worse than the one 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. 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 when considering NIR models: QPPs approaches applied on Deep Learning '19 perform by far worse than on the former Robust '04. While *on average* we will be less satisfied by QPP predictors applied to NIR regardless of the type of collection, there might be good performing predictors also for NIR systems. To verify this hypothesis, we apply MD2 to
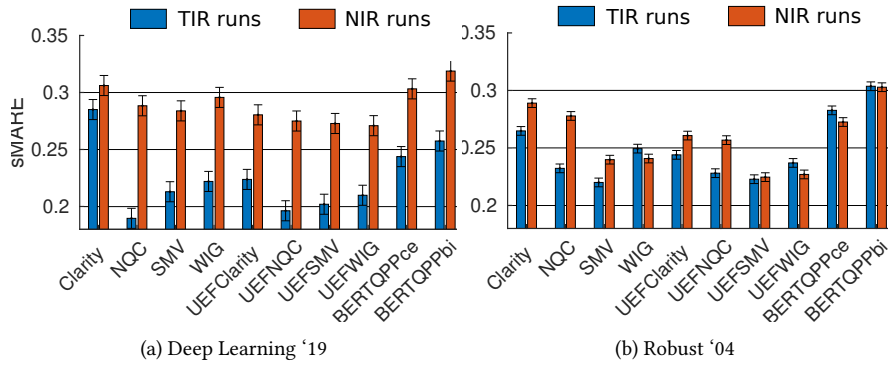
Figure 3: sMARE observed for different predictors on Deep Learning '19 (left) and Robust '04 (right).

each collection separately, and measure what happens to each predictor individually[2]. Table 1 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 close 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. This indicates that whether a model will be better than another depends on the topic considered. An almost identical pattern was observed also in [35]. Therefore, to improve QPP's generalizability, it is important to address challenges caused by differences in NIR and TIR as well as to take account of the large variance introduced by topics.

If we consider Robust '04, 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 1 – 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 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.

Figure 4a shows the performance of topics that maximize the difference between TIR and NIR and can be considered as more "semantically defined" [49]. If we consider the results of

---

[2]We focus on post-retrieval predictors – similar observations hold for pre-retrieval ones.

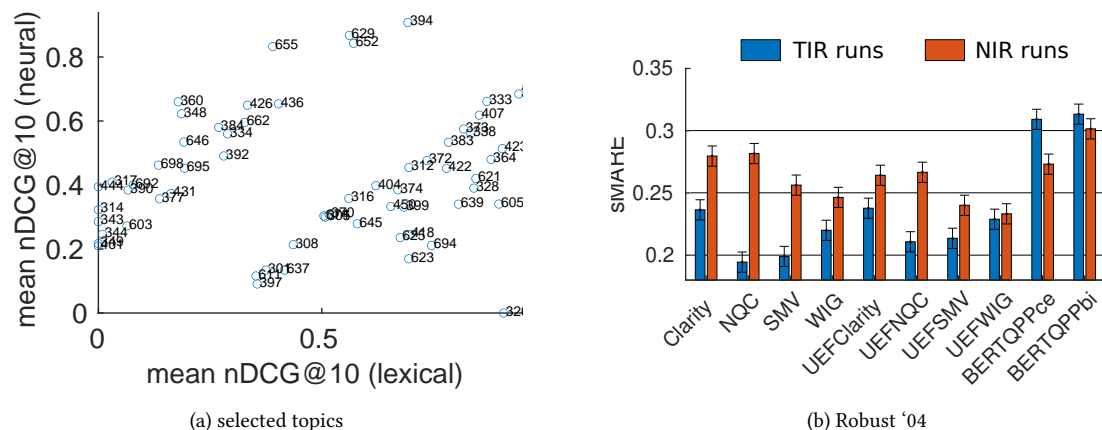(a) selected topics

(b) Robust '04

Figure 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.

applying MD2 on this set of topics, we notice that compared to Robust '04 (Table 1, 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.

5. Conclusion and Future Work

This study examined the applicability of current QPPs on first-stage NIR models based on PLMs. The study evaluated 19 diverse QPP models on seven TIR and seven first-stage NIR methods based on BERT, applied to the Robust '04 and Deep Learning '19 collections. We observe that indeed QPPs are effective in predicting TIR systems' performance but fail in dealing with NIR ones. Moreover, the study found that QPPs tend to fail on those topics where NIR and TIR models differ the most, which impairs the possibility of using QPP models to choose between NIR and TIR approaches where it is most needed. Furthermore, semantic QPP approaches such as BERT-QPP do not solve the problem and work properly only on lexical IR systems. These results highlight the need for QPPs specifically tailored to neural IR.

Future work will consider query variations to understand the impact of changing how a topic is formulated on QPP, and the development of QPP methods explicitly designed for NIR models that take into consideration the large variance introduced by topics.

References

[1] G. Faggioli, T. Formal, S. Marchesin, S. Clinchant, N. Ferro, B. Piwowarski, Query Performance Prediction for Neural IR: Are We There Yet?, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, 2023, pp. 232–248. URL: https://doi.org/10.1007/978-3-031-28244-7_15. doi:10.1007/978-3-031-28244-7_15.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186.

[3] G. Faggioli, N. Ferro, J. Mothe, F. Raiber, QPP++ 2023: Query-Performance Prediction and Its Evaluation in New Tasks, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, 2023, pp. 388–391.

[4] N. Reimers, I. Gurevych, 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, 2019, pp. 3980–3990.

[5] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense Passage Retrieval for Open-Domain Question Answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781.

[6] L. Xiong, C. Xiong, Y. Li, K. Tang, J. Liu, P. N. Bennett, J. Ahmed, A. Overwijk, 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.

[7] S. Hofstätter, S. Lin, J. Yang, J. Lin, A. Hanbury, 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, 2021, pp. 113–122.

[8] L. Gao, J. Callan, 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, 2022, pp. 2843–2853.

[9] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Towards Unsupervised Dense Information Retrieval with Contrastive Learning, CoRR abs/2112.09118 (2021).

[10] Z. Dai, J. Callan, 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, 2020, pp. 1533–1536.

[11] A. Mallia, O. Khattab, T. Suel, N. Tonellotto, 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, 2021, pp. 1723–1727.

[12] S. Zhuang, G. Zuccon, 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, 2021, pp. 1483–1492.

[13] J. Lin, X. Ma, A Few Brief Notes on DeepImpact, COIL, and a Conceptual Framework for Information Retrieval Techniques, CoRR abs/2106.14807 (2021).

[14] R. F. Nogueira, W. Yang, J. Lin, K. Cho, Document Expansion by Query Prediction, CoRR abs/1904.08375 (2019).

[15] T. Formal, B. Piwowarski, S. Clinchant, 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, 2021, pp. 2288–2292.

[16] T. Formal, C. Lassance, B. Piwowarski, S. Clinchant, 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, 2022, pp. 2353–2359.

[17] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, 2016.

[18] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, 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.

[19] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, M. Zaharia, 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, 2022, pp. 3715–3734.

[20] D. Carmel, E. Yom-Tov, Estimating the Query Difficulty for Information Retrieval, Morgan & Claypool Publishers, 2010.

[21] J. S. Culpepper, G. Faggioli, N. Ferro, O. Kurland, Topic Difficulty: Collection and Query Formulation Effects, ACM Trans. Inf. Syst. 40 (2022) 19:1–19:36.

[22] C. Hauff, D. Hiemstra, F. de Jong, 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, 2008, pp. 1419–1420.

[23] Y. Zhao, F. Scholer, Y. Tsegay, 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, volume 4956, 2008, pp. 52–64.

[24] S. Cronen-Townsend, Y. Zhou, W. B. Croft, A Language Modeling Framework for Selective Query Expansion, Technical Report, CIIR, UMass, 2004.

[25] F. Scholer, H. E. Williams, A. Turpin, Query association surrogates for Web search, J. Assoc. Inf. Sci. Technol. 55 (2004) 637–650.

[26] S. Cronen-Townsend, Y. Zhou, W. B. Croft, 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, 2002, pp. 299–306.

[27] A. Shtok, O. Kurland, D. Carmel, 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, volume 5766, 2009, pp. 305–312.

[28] Y. Zhou, W. B. Croft, 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, 2007, pp. 543–550.

[29] Y. Tao, S. Wu, 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, 2014, pp. 1891–1894.

[30] A. Shtok, O. Kurland, D. Carmel, 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, 2010, pp. 259–266.

[31] H. Zamani, W. B. Croft, J. S. Culpepper, 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, 2018, pp. 105–114.

[32] H. Hashemi, H. Zamani, W. B. Croft, 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, 2019, pp. 55–58.

[33] N. Arabzadeh, M. Khodabakhsh, E. Bagheri, 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, 2021, pp. 2857–2861.

[34] S. Datta, D. Ganguly, M. Mitra, D. Greene, A Relative Information Gain-Based Query Performance Prediction Framework with Generated Query Variants, ACM Trans. Inf. Syst. (2022) 1–31.

[35] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, 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, volume 12656, 2021, pp. 115–129.

[36] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, sMARE: a new paradigm to evaluate and understand query performance prediction methods, Inf. Retr. J. 25 (2022) 94–122.

[37] A. Rutherford, ANOVA and ANCOVA: a GLM approach, John Wiley & Sons, 2011.

[38] E. M. Voorhees, The TREC robust retrieval track, SIGIR Forum 39 (2005) 11–20.

[39] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees, I. Soboroff, 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, 2021, pp. 2369–2375.

[40] E. M. Voorhees, I. Soboroff, J. Lin, Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models?, CoRR abs/2201.11086 (2022).

[41] C. Zhai, Statistical Language Models for Information Retrieval: A Critical Review, Found. Trends Inf. Retr. 2 (2008) 137–213.

[42] G. Salton, C. Buckley, Term-Weighting Approaches in Automatic Text Retrieval, Inf. Process. Manag. 24 (1988) 513–523.

[43] G. Amati, C. J. van Rijsbergen, Probabilistic Models of Information Retrieval based on measuring the Divergence From Randomness, ACM Trans. Inf. Syst 20 (2002) 357–389.

[44] H. Fang, C. Zhai, 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, 2005, pp. 480–487.

[45] I. Kocabas, B. T. Dinçer, B. Karaoglan, A nonparametric term weighting method for information retrieval based on measuring the divergence from independence, Inf. Retr. 17 (2014) 153–176.

[46] T. Formal, C. Lassance, B. Piwowarski, S. Clinchant, SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval, CoRR abs/2109.10086 (2021).

[47] J. He, M. A. Larson, M. de Rijke, 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, volume 4956, 2008, pp. 689–694.

[48] N. Ferro, G. Silvello, Toward an anatomy of IR system component performances, J. Assoc. Inf. Sci. Technol. 69 (2018) 187–200.

[49] G. Faggioli, S. Marchesin, 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, volume 2950 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 61–69. URL: http://ceur-ws.org/Vol-2950/paper-06. pdf.