Towards Query Performance Prediction for Neural Information Retrieval: Challenges and Opportunities

Guglielmo Faggioli
University of Padua
Padova, Italy

Thibault Formal
Naver Labs Europe
Meylan, France

Simon Lupart
Naver Labs Europe
Meylan, France

Stefano Marchesin
University of Padua
Padova, Italy

Stéphane Clinchant
Naver Labs Europe
Meylan, France

Nicola Ferro
University of Padua
Padua, Italy

Benjamin Piwowarski
Sorbonne Université, ISIR, CNRS
Paris, France

ABSTRACT

In this work, we propose a novel framework to devise features that can be used by Query Performance Prediction (QPP) models for Neural Information Retrieval (NIR). Using the proposed framework as a periodic table of QPP components, practitioners can devise new predictors better suited for NIR. Through the framework, we detail what challenges and opportunities arise for QPPs at different stages of the NIR pipeline. We show the potential of the proposed framework by using it to devise two types of novel predictors. The first one, named MEMory-based QPP (MEM-QPP), exploits the similarity between test and train queries to measure how much a NIR system can memorize. The second adapts traditional QPPs into NIR-oriented ones by computing the query-corpus semantic similarity. By exploiting the inherent nature of NIR systems, the proposed predictors overcome, under various setups, the current State of the Art, highlighting – at the same time – the versatility of the framework in describing different types of QPPs.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**.

KEYWORDS

Query Performance Prediction, Neural IR, QPP Framework

ACM Reference Format:
Guglielmo Faggioli, Thibault Formal, Simon Lupart, Stefano Marchesin, Stéphane Clinchant, Nicola Ferro, and Benjamin Piwowarski. 2023. Towards Query Performance Prediction for Neural Information Retrieval: Challenges and Opportunities. In *Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '23), July 23, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3578337.3605142

ICTIR '23, July 23, 2023, Taipei, Taiwan

ACM ISBN 979-8-4007-0073-6/23/07…$15.00
https://doi.org/10.1145/3578337.3605142

1 INTRODUCTION

Neural Information Retrieval (NIR) encompasses a broad range of Information Retrieval (IR) techniques that rely on Neural Networks [42, 49, 50, 66, 99]. Large Pre-trained Language Models (PLMs) have impacted IR beyond expectation, leading to unprecedented results on various benchmarks [11–13, 18, 68, 92] as well as in IR evaluation [26, 64]. These systems have initially been developed to learning-to-rank and re-rank results from Traditional Information Retrieval (TIR) approaches such as BM25 [68], but have been more recently employed to directly tackle first-stage retrieval [41]. To evaluate the performance of IR systems without human-made relevance judgments, Query Performance Prediction (QPP) [7] has been investigated for decades and used in various tasks, including model selection [7, 93], query rewriting [25, 83, 93], rank fusion [76], query diagnosis [7], and predicting the best pool's cut-off to reduce annotation cost [38].

QPP is of particular importance for NIR[1] for several reasons. *i)* Unlike TIR models, NIR models come in a range of architectures, each with its unique characteristics. Therefore, choosing the best approach for a given query could significantly enhance the overall performance. *ii)* Training NIR models can be time-consuming and expensive. Effective QPP models could identify underperforming queries and guide the practitioner in gathering additional training examples to improve performance on such queries. *iii)* NIR models are frequently employed in a zero-shot fashion [92], and targeted QPPs could aid in determining beforehand if the model trained on the source collection would perform well on the target one.

Nevertheless, applying QPP to NIR poses some critical challenges that have yet to be addressed [27, 29]. In particular, traditional QPPs often depend on measuring the amount of lexical matching between queries and retrieved documents [14, 88, 90, 109], whereas NIR models are explicitly designed to use semantic matching. The misalignment between the signals considered by QPPs and those used by NIR systems hinders the successful prediction of NIR performance. Additionally, NIR systems are trained on labeled data, contrasting with the unsupervised nature of TIR systems. Their performance is therefore tied to what can be learned from the train set,

[1] In this work, we focus on NIR models, although similar observations likely hold for Learning-to-Rank models. In-depth study of the relationship between LtR and QPP is left for future work.

which is particularly critical for systems evaluated in a zero-shot setting [36, 63, 72, 85, 92, 106].

To overcome these limitations, we propose a new QPP framework that goes further from the traditional pre- and post-retrieval dichotomy. The framework considers a generic NIR pipeline as a reference point and uses it as a periodic table to identify potential challenges that arise when applying QPP. The framework foundation lies on the concept of n-stage features. We relate them to features that have been used by pre- and post-retrieval predictors, as well as which signals are specific to NIR models. We show the power of the proposed framework by using it to develop two QPP approaches that outperform the current State of the Art. The first approach, named MEMory-based QPP (MEM-QPP), directly relies on the supervised nature of NIR models, by measuring the similarity between training and test queries to predict performance. By exploiting information obtained during training, MEM-QPP achieves improved results compared to classical pre- and post-retrieval predictors. The second approach, based on post-retrieval information, adapts the regularization term used in many traditional QPPs – e.g., Clarity [14], WIG [109], NQC [88], SMV [90] – to the NIR setting. Our experimental results show that this approach achieves significantly better performance than the current State of the Art when predicting the performance of NIR models.

To summarize, our contributions are:

- We propose a novel QPP framework based on the concept of *stages* that overcome the classical dichotomy between pre- and post-retrieval QPP. We further show how classical predictors can be interpreted within the framework.
- Using our framework, we devise MEM-QPP, a model-agnostic predictor that leverages semantic similarity between training and test queries to predict the performance of NIR systems in the In-Domain scenario.
- We further use the framework to adapt classical predictors to the NIR scenario. In particular, we compare the distributions of retrieval scores for TIR and NIR systems, and show that they present a high level of similarity. Motivated by this, we propose two sets of QPPs, SPLADE- and DenseCentroid-predictors. Both sets of predictors adapt classical QPPs to the NIR scenario, going beyond the current State of the Art.

The remainder of this work is organized as follows: Section 2 reviews the main efforts in NIR and QPP areas. Section 3 describes the proposed QPP framework for NIR. Section 4 illustrates the challenges highlighted by the framework and outlines possible solutions that can guide future researchers. Finally, Section 5 draws conclusions and introduces future directions enabled by our framework.

2 BACKGROUND

With the advent of NIR based on PLMs, there is a growing interest in developing QPP methods that are tailored to such systems. This is especially appealing for systems tackling the retrieval step in multi-stage pipelines, which differ the most from TIR models. We survey here the main advances in NIR and classical QPP domains. Then, we describe three related families of QPPs. The first family exploits neural networks to nonlinearly combine different features and generate predictions, without any notion of semantics. The second type of QPPs explicitly encodes semantic signals to make predictions, but has primarily been designed for TIR systems. The third family regards QPP models that have been used to predict NIR performance, which is the most similar setting to the one examined in this paper.

Neural Information Retrieval. First-stage NIR models have received increasing attention, due to their ability to overcome the inherent limitations of TIR approaches, such as the vocabulary mismatch [37]. They can be divided into two main categories: sparse and dense [41]. Sparse retrieval models represent documents and queries as sparse high-dimensional vectors in the vocabulary space $\mathbb{R}^{|\mathcal{V}|}$, allowing for efficient indexing and inference with standard inverted indexes [5, 10, 19, 34, 35, 57, 59, 65, 69, 103, 107]. Among them, SPLADE [34, 35] unifies term weighting and expansion in an end-to-end fashion and has shown impressive performance on both In- and Out-of-Domain settings. Dense retrieval models move away from the sparse view by representing queries and documents as continuous low-dimensional vectors in a latent semantic space $\mathbb{R}^d$ [39, 48, 51, 53, 61, 71]. Other approaches like ColBERT [54, 82] or COIL [40] lie in between, by generating fine-grained term-level dense representations for queries and documents. All these approaches rely on different inductive biases that are complementary to some extent – as shown for hybrid retrieval models that combine the strengths of sparse and dense retrieval methods to improve the overall effectiveness [8, 58, 62, 97, 100]. Therefore, being able to predict the performance of various systems is of particular interest.

Traditional QPP. Traditionally, QPPs can be divided into pre- and post-retrieval predictors [7, 44, 45]. Pre-retrieval predictors use features available before retrieval, such as the collection frequency of query terms [67, 108], while post-retrieval predictors compute predictions after one or more retrieval phases. Post-retrieval predictors can further be classified as coherency-based, score-based, and robustness-based. Coherency-based predictors estimate the coherence between the query and the retrieved documents (e.g., Clarity [14]), score-based predictors use the scores of retrieved documents (e.g., Weighted Information Gain (WIG) [109]), and robustness-based predictors measure similarity between original ranking and one after perturbations (e.g., Utility Estimation Framework (UEF) [86]).

Deep-Learning-driven QPP. In recent years, the advent of Deep Learning (DL) has fostered the development of QPP approaches based on neural networks. Zamani et al. [102] propose NeuralQPP, one of the first attempts to apply DL to the QPP task. The authors devise a DL approach that combines three distinct signals to formulate the prediction: the query text, the retrieval scores, and signals derived from the terms distribution. On the other hand, Roy et al. [79] experiment with pre-retrieval predictors. In their work, they show how the distribution of terms in relation to query vectors, estimated using Gaussian Mixture Models, correlates with system performance. The study highlights the need to combine the proposed pre-retrieval QPP with post-retrieval predictors to achieve satisfactory results. Similarly, Arabzadeh et al. [4] propose, and later extend [3], a set of measures based on neural embeddings that quantify the specificity of each term. The underlying intuition is that specific query terms can better identify relevant documents. These measures serve as pre-retrieval predictors and have been shown to

correlate with system performance. As for NeuralQPP [102], also the predictors by [3, 4, 79] are evaluated on TIR systems.

NIR for QPP. To exploit semantic signals for QPP, Khodabakhsh and Bagheri [55] propose three neural features based on dense word representations: Neural Matching, Neural Aggregated Matching, and Neural Distance. These features combine the embeddings of query and document tokens to capture the semantic relationships occurring between them. The authors use the matching signals provided by such features to encode semantic aspects within classic predictors such as WIG [109], Normalized Query Commitment (NQC) [88], Score Magnitude and Variance (SMV) [90], and Clarity [14]. Datta et al. [21] present Deep-QPP, a method based on a convolutional neural network that exploits word embeddings and early-stage interaction signals. The method represents a supervised approach to regress query performances, which has been tested on Language Model (LM) [105]. Arabzadeh et al. [1] propose BERT-QPP, one of the first approaches to leverage PLMs for QPP. Specifically, they fine-tune BERT [24] by utilizing the performance of BM25 [75] on each training query and the first retrieved document as a form of supervision. Several works stemmed from BERT-QPP. Datta et al. [23] expand on BERT-QPP by incorporating clusters of queries – rather than considering only one query at a time – to extract more comprehensive signals. Likewise, Chen et al. [9] build upon BERT-QPP and introduce a groupwise approach that enables learning to predict the performance of a query using signals from multiple queries simultaneously, instead of examining only one query at a time, as in BERT-QPP. Arabzadeh et al. [2] also exploit PLMs to devise a predictor for conversational search. The authors leverage BERT to create a graph of retrieved documents and determine if the documents can be categorized into a single cluster. If multiple clusters exist, they identify the user's information need by asking clarifying questions to determine which cluster contains the relevant documents. The approach is then tested on BM25. Although these methods lean towards NIR models, their primary use remains associated with TIR approaches. This creates a misalignment between the query/document representations used during the ranking and prediction phases, with the former being more lexical and the latter more semantic. Faggioli et al. [28] exploit the geometric properties of dense documents and queries representations to predict the performance in the conversational search task.

Since most of these predictors have been devised for and tested on TIR models, they are not the focus of this paper, which instead concerns QPP predictors for NIR models.

QPP for NIR. More closely related to our work, Hashemi et al. [43] propose NQA-QPP, an approach based on three families of signals – i.e., retrieval scores, lexical features of the query, and lexical features of the query and answer – combined with a deep neural network to address the Non-Factoid Question Answering task. [43] is one of the first and few works evaluating the performance of QPP on NIR models. Specifically, the authors consider BM25 and two neural reranking strategies, aNMM [101] and Conv-KNRM [20]. It is interesting to note that they are among the first to notice the large gap between the prediction quality for BM25 and NIR models. The authors attribute this outcome to the different scale and distribution of scores generated by neural models.

Recently, Faggioli et al. [29] investigate to what extent traditional QPP methods can predict the performance of NIR systems. The authors conduct experiments applying QPPs to several TIR and NIR systems, evaluating them on Deep Learning '19 and Robust '04 collections. The results show that current QPPs perform significantly worse on NIR systems. This situation occurs even when BERT-QPP is considered as a predictor for NIR. Datta et al. [22] also note that previous QPP methods are not as effective for NIR as they are for TIR. To address this, the authors propose Weighted Relative Information Gain-based model (WRIG), a statistical method that involves using probabilistic combinations of retrieval scores for multiple formulations of the same query. To show the effectiveness of the devised strategy, they use WRIG to predict the performance of BM25, four variants of the Deep Relevance Matching Model (DRMM) [42], and a first-stage NIR method, ColBERT [54].

3 A FRAMEWORK TO MODEL QPP FOR NIR AND ITS CHALLENGES

Figure 1 reports a visual depiction of the proposed framework, which is organized into *stages*. The figure is composed of three layers corresponding to retrieval, QPP operations, and the features that can be extracted at each stage. The first three stages concern *learned* approaches, either for NIR or QPP. They contain the choice of the training and test corpora (Stage 0), the collection of training queries and annotations (Stage 1), and the learning procedure (Stage 2). Subsequently, we have Stage 3, which consists in collecting test queries. This Stage, as well as Stages 4 and 5, are common to NIR, TIR, and QPP. Stage 4 represents the moment when the representation of queries and documents is computed. Starting from Stage 4, IR and QPP operations might differ. For instance, the IR model might be based on lexical representations, while the QPP could use dense ones. In Stage 5, the similarity between the query and the documents works for both IR and QPP operations. On the other hand, for QPP, Stage 5 might also contain operations over the entire corpus, such as computing its language model or the retrieval score that it would achieve in response to a query. Stage 6 consists in computing the distribution of the scores over the retrieved documents to select the top-k ones most similar to the query. Finally, a performance measure can be computed, which in turn is the ground truth for the QPP.

In terms of QPP features, no model exploits features derived from the first three stages yet. They are linked to the supervised nature of NIR models. Stage 3 and 4 features are those most commonly used by classical pre-retrieval QPPs. Finally, Stages 5 and 6 allow devising features typically used by post-retrieval QPPs. To show the benefit of the framework, we detail each stage – showing what challenges, pitfalls, and opportunities might arise with respect to each of them, as well as what features could be collected.

Stage 0 – Corpora. This first stage mainly concerns the zero-shot scenario, in which training and inference corpora differ.

Challenge Stage 0 – Exploiting the difference between training and inference collections: The main challenge associated with this stage concerns how to measure the difference between train and test corpora to determine whether a model trained on the former achieves satisfactory results on the latter. The importance of addressing this challenge will be highlighted by our

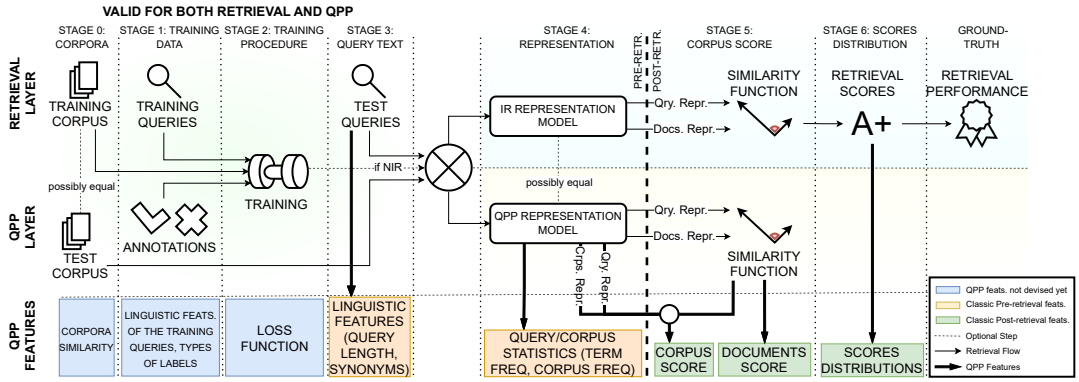

Figure 1: A retrieval pipeline and its correspondence with multi-stage QPPs.

experiments in Section 4, which show how this is one of the most prominent pitfalls in adapting QPP to NIR.

Stage 1 – Training Data. Features derived from this stage are those concerning training queries and annotations. Such features might include the textual content of the training queries or concern aspects linked to how we choose the documents that should be fed to the training model (e.g., hard negatives, number and quality of the annotations).

Challenge Stage 1 – Exploiting training queries information: The community should understand how to take advantage of the information conveyed by training queries to improve QPP in the NIR setup. To demonstrate this possibility, we propose a novel predictor, dubbed MEM-QPP, in Subsection 4.2. This predictor exploits Stage 1 and 3 features, and could therefore be framed as pre-retrieval QPP. MEM-QPP considers, for a given test query, how similar training queries are. The rationale is that the more similar training and test queries are, the better the test queries can be answered – this is directly linked to the fact that NIR models might overfit on training queries [70, 89].

Stage 2 – Training Procedure. A straightforward Stage 2 predictor is the loss function used by NIR approaches. For instance, if the loss function does not decrease during training, it is easy for the practitioner to diagnose a malfunctioning NIR system and predict low performance. The training procedure of the NIR system itself can therefore provide signals on how the system will perform.

Challenge Stage 2 – Lack of Labeled Data: Collections containing a suitable amount of queries to train NIR models (e.g., MS-MARCO [6]) have a shallow pool of annotated documents – often just one. Therefore, we cannot accurately compute IR measures, which are the labels required to train QPP models. Addressing this challenge requires investing, as a community, in obtaining additional annotations and organizing shared tasks.

Stage 3 – Query text. The features that can be computed in this stage are those used by traditional pre-retrieval methods, which are based on linguistic aspects of the query. Such features include, among others, the query length or the presence of synonymous or polysemous words.

Challenge Stage 3 – Query representation: While no major challenges arise with respect to Stage 3, our framework contributes by separating pre-retrieval operations into Stage 3 and Stage 4. This separation provides a clear dichotomy between predictors that use only the query textual content and predictors that also take into account the corpus representation.

Stage 4 – Representation. Together with Stage 3, this stage contains features traditionally considered the most suited for pre-retrieval QPP. Stage 4 corresponds, in an IR pipeline, to when a representation of the queries and the documents has been computed. At this point, it is possible to compute predictors such as Inverse Document Frequency (IDF) and Inverse Collection Term Frequency (ICTF) [15, 84], which require access to the inverted index.

Challenge Stage 4 – Representation misalignment between QPP and NIR models: Representations used by traditional QPPs are often aligned with those by the retrieval model. For example, Clarity is based on a frequentist LM that, although providing a different representation than those by the Vector Space Model [81] or BM25 [75], presents a similar rationale. This is not the case for most of the NIR systems, which rely on dense or sparse learned representations.

Stage 5 – Corpus Score. The features of this stage are derived from the document scores as well as the representation of the entire corpus according to the chosen model. The representation of the corpus is used to *i)* measure the relevance of the corpus – i.e., the retrieval score – in response to the query; *ii)* in the Clarity case, measure how likely it is that the corpus generated the retrieved documents. Such features underlie most classical and theoretically well-grounded post-retrieval approaches. To understand the importance of Stage 5, we consider one of the main efforts in standardizing QPP models by Kurland et al. [56], later expanded by Shtok et al. [87] and Roitman et al. [78]. Kurland et al. [56] notice how post-retrieval predictors can be framed using the following model:

$$Pred(\mathcal{D}@k|q) = p(r|\mathcal{D}@k) \sum_{d \in \mathcal{D}@k} p(d|q,r)p(d|\mathcal{D}@k,r), \quad (1)$$

where q is the query, d is a document, r is the "event of relevance", $\mathcal{D}@k$ the list of top k documents retrieved and $Pred$ is the probability that $\mathcal{D}@k$ contains relevant documents for q (the predictor). The factor $p(r|\mathcal{D}@k)$ describes how likely it is that $\mathcal{D}@k$ contains relevant documents, regardless of the query, and is based on properties of $\mathcal{D}@k$, such as its cohesion or diversity [56, 78, 87]. $p(d|q,r)$ is the document likelihood – usually approximated by the retrieval score $s(q,d)$. Finally, $p(d|\mathcal{D}@k,r)$ represents a *regularization factor* that measures the strength of "association" between d and $\mathcal{D}@k$, assuming the former is relevant. According to [56, 78], this factor can be estimated by measuring how "closer" is d to $\mathcal{D}@k$ than to the entire corpus. Hence, following [56, 78, 87], classical predictors focus on how this regularization factor is computed. For example, Clarity estimates it by comparing the likelihood of generating d from $\mathcal{D}@k$ against the likelihood of generating it from the background model induced by the corpus C [78]. Other QPP models, such as SMV, NQC, or WIG set $p(d|\mathcal{D}@k,r) \propto \frac{1}{s(q,C)}$, where $s(q,C)$ is the relevance of the entire corpus to the query. $s(q,C)$ is traditionally approximated by the retrieval score of the concatenation of all the documents, in response to the query [56, 78, 87]. In turn, this query-sensitive regularization term allows for comparing the predictions across queries [77, 88]. When we consider NIR models, only two of the three components in Equation 1 can be derived directly: $p(r|\mathcal{D}@k)$ and $p(d|q,r)$. Specifically, $p(r|\mathcal{D}@k)$ is model-agnostic, while $p(d|q,r)$ represents the retrieval score as returned by the NIR system. On the other hand, the regularization factor $p(d|\mathcal{D}@k,r)$ requires a notion of distance between the document and the entire corpus that cannot be computed directly by current NIR methods. Therefore, to use classical QPPs on NIR models, at the current time, the only solution consists of applying Eq. 1 using LM to instantiate its different components but considering the top-k documents retrieved by the NIR approach, as done in [29]. The reasons for this impairment are reported below.

Challenge Stage 5 – How to compute $s(q,C)$ for NIR models: In contrast to TIR models, computing the score that the collection would achieve is not feasible for NIR models. This is because, with current neural architectures, it is impossible to input the entire collection into the model to obtain its representation. As a result, how to compute $s(q,C)$ is an open issue. If we could compute $s(q,C)$, adapting the majority of the QPP frameworks – such as those outlined in [56, 78, 87] – to the NIR scenario would be seamless. In Subsection 4.4, we propose two methodologies to approximate $s(q,C)$ for sparse and dense NIR architectures.

Stage 6 – Scores Distribution. This stage concerns the features derived from the distribution of the scores and underlies most of the traditional QPPs, such as WIG, NQC, and SMV. Such predictors are based on statistics – e.g., mean and variance – of the scores for the top-k documents. This requires putting the single document's score (Stage 5 feature) in relation to other documents' scores.

Challenge Stage 6 – Modeling scores distributions for NIR: While TIR models usually have scores that are naturally interpretable and often account for term matching, NIR models do not have any a priori on the distributions of the possible scores they follow. In particular, each model has its own embedding space structure, and scores depend on the model architecture itself. It might be challenging to use traditional QPPs with scores from NIR models, as neural models could exhibit different behavior for each predictor. To investigate this, in Section 4.3, we perform an analysis comparing scores distributions for TIR and NIR models.

3.1 Reinterpreting the State of the Art

We illustrate, using some examples, how the current SotA QPP approaches can be interpreted in light of the proposed framework.

Pre-retrieval QPPs. This class of QPP models focuses mostly on Stage 3 and Stage 4 features. Consider, for example, predictors that exploit syntactic features of the query, such as its length [46], or more articulate linguistic features, such as external knowledge bases or thesauri to identify synonyms or polysemous words [67]. They rely only on the textual content of the query and, therefore, can be framed as Stage 3 predictors. Other QPPs, such as those in [4, 79], exploit latent representations of the query tokens based on word embeddings. It is important to note that the query representation provided by [4] does not depend on the corpus. As a result, the proposed predictors are based solely on Stage 3 features. Finally, other approaches, such as Simplified query Clarity Score (SCS) [47], Similarity Collection-Query (SCQ) [108], VAR [108], IDF and ICTF [15, 84], are based on frequentist aspects of the query and the corpus. Thus, these predictors exploit Stage 4 features. Despite relying on conceptually different query representations, all of these QPP models have traditionally been considered pre-retrieval, as they occur prior to Stage 5.

Classical post-retrieval QPPs. According to [56, 78, 87], most of the classical post-retrieval predictors, such as Clarity [14], WIG [109], NQC [88], and SMV [90], follow the framework described by Equation 1. These approaches re-weigh statistics of the retrieval scores for the top-k retrieved documents (Stage 6 feature) – e.g., the mean or the standard deviation – with the retrieval score the corpus would achieve (Stage 5 feature).

Learned post-retrieval QPPs. Some additional challenges arise if we consider QPPs based on learned representations. As an example, let us consider BERT-QPP [1], which feeds the query text (Stage 3) and the text (Stage 0) of the first k retrieved documents (Stage 6) to a bi-encoder network based on BERT. Thus, BERT-QPP uses both pre- and post-retrieval features. However, the representations used by BERT-QPP differ from those used by the target IR system. Indeed, BERT-QPP was originally developed to predict BM25 performance, as detailed in [1]. A similar reasoning also holds for most of the learned QPPs, such as [43, 102].

4 NIR PERFORMANCE PREDICTION: ADDRESSING SOME CHALLENGES

We now evaluate the framework, showing its capabilities. In Section 4.2, we illustrate how to devise a model-agnostic predictor that exploits features derived from Stages 1 and 3 (the text of the training and test queries) and that relies on the memorization capabilities of NIR models. Following our observations about Stage 6 features, in Section 4.3, we analyze the IR scores distributions for TIR and NIR models. The similarities of the scores distributions of the two IR families motivate us to try adapting traditional QPP models to the NIR scenario. To this end, in Section 4.4, we illustrate how to

obtain the proper regularization term (Stage 5 feature) depending on the architecture of the considered NIR model.

4.1 Experimental Setup

We evaluate QPPs drawn from our framework using three collections: Robust '04 [95], Deep Learning '19 [13], and TREC-COVID [73]. These collections vary in terms of the number of topics, corpus, and availability of training queries. Robust '04 contains 249 topics and is commonly used for ad-hoc document TIR without training queries. Deep Learning '19 has 43 annotated topics and has been designed towards NIR models by providing over $500k$ In-Domain training queries – with a few annotated passages – from the MS-MARCO collection. TREC-COVID is a bio-medical dataset capturing the growth of the COVID-19 literature over time. It is based on the CORD-19 document corpus [96], and contains 50 test queries.

In terms of NIR, we consider the following State-of-the-Art models: four dense models (Bi-Encoder, TAS-B [48], CoCondenser [39], and Contriever [51]), a dense late-interaction model (ColBERT-v2 [82]), and a sparse model (SPLADE [33, 34]). Following standard practices, all models are fine-tuned on the MS-MARCO passage dataset and evaluated in a zero-shot manner on Robust '04 and TREC-COVID – where training queries are not available. To handle longer documents, we truncate them to the models' maximum length. We train the Bi-Encoder from scratch, while for the other models, we rely on open-source weights. For retrieval, we either use internal code – based on FAISS [52] and HuggingFace transformers [98] – or the code from the corresponding open-source repositories – as in the case of SPLADE and ColBERT-v2. As additional baselines, we consider three TIR approaches: the probabilistic model BM25 [74], the Dirichlet Language Modeling approach (LMD) [105], and the axiomatic F1-EXP [32], which together define a diverse set of traditional models. In terms of QPP, we experiment with models that are theoretically grounded in previously defined QPP frameworks [56, 78, 87]. In particular we consider Clarity [14], WIG [109], NQC [88], SMV [90], and their UEF versions [86]. All predictors have been fine-tuned by considering different cutoffs for $\mathcal{D}@k$ with $k \in \{5, 10, 50, 100, 500\}$, following the 2-fold partitioning procedure described in [22, 88, 102, 104] with 30 repetitions. To test for the statistical significance of the results, we apply one-way ANalysis Of the VAriance (ANOVA) [80] with Tukey's Honestly Significant Differences (HSD) post-hoc comparison procedure [94], which also corrects for multiple comparisons. Given its popularity in the NIR scenario, we focus on normalized Discounted Cumulative Gain (nDCG)@10 as the target measure that we wish to predict. To evaluate the performance of the QPPs, we employ Pearson's and Kendall's correlations and scaled Mean Absolute Rank Error (sMARE) [30, 31][2].

4.2 Stage 1 and 3 Features: NIR Models as Memorizers

Following the framework devised in Section 3 we propose a QPP strategy – dubbed MEM-QPP– based on Stage 1 and Stage 3 features, i.e., the textual content of training and test queries. The rationale underneath MEM-QPP is that, if a test query has a "close" training query, then the NIR system might have learned how to retrieve from it. Vice versa, for test queries that are too "far" from the training set, we can assume that the system did not gain enough information on that topic, making it hard to perform well at inference time. It is directly linked to memorization capabilities of PLMs [36, 70, 85, 89], and is inspired from Lupart et al. [63], who propose a similar indicator which correlates with performance drops on zero-shot settings. More formally, it is defined as follows:

$$\text{MEM-QPP}(q_t) \underset{def}{=} \max\{s(q_t, q_r) : q_r \in Q_T\}$$

where q_t is a test query, Q_T is the set of training queries, and $s(q_t, q_r)$ is a similarity function. In other terms, MEM-QPP measures the similarity between test queries and the most similar training query. To embed the *learned* component, we consider as similarity functions s three NIR models: Bi-Encoder, SPLADE and ColBERT-v2[3]. Notice that the representation is not necessarily the same between the predictor and the predicted IR system. For example, we could use Bi-Encoder as the similarity function to instantiate MEM-QPP and predict the performance of a SPLADE run.

MEM-QPP allows understanding of on which topics the training phase was not thorough enough. Once test queries that are likely to fail are identified, the system administrator can expand the training set to include annotations for such queries. Also, note that such an indicator is insufficient to fully characterize models' performance: some queries are intrinsically more difficult than others, regardless of how many times they have been seen at training time. We hypothesize, however, that performance should still be correlated with such indicators – even loosely.

Table 1 reports the empirical evaluation of the three considered variants of MEM-QPP on four NIR models and three different collections. To avoid cluttering, we report the performance of the best – on average – pre-retrieval (i.e., ICTF and IDF) and post-retrieval (i.e., WIG and UEF variants of WIG and NQC) predictors.

MEM-QPP using the Bi-Encoder similarity is the best predictor on Deep Learning '19 for three out of four NIR systems: TAS-B, SPLADE, and ColBERT-v2. It exhibits an improvement as large as 17% in the case of SPLADE. Nevertheless, it fails to beat the baselines when the Bi-Encoder is also used as the ranking function. In Out-of-Domain (OOD) collections, MEM-QPP fails to overcome the considered baselines. This is a first glimpse of how important it would be to be able to devise Stage 0 features. Indeed, MEM-QPP performs well when used as a predictor for In-Domain IR, as for the Deep Learning '19 track. Since the NIR approaches were both trained and tested on MS-MARCO passages, we might assume that the models embed part of the information about the relevance of the documents. Vice versa, MEM-QPP is not as effective when we consider the zero-shot setup. In a nutshell, when we shift the corpus, this information is lost, and MEM-QPP fails. The results show that MEM-QPP is an indicator of how much information from the training set the NIR system can memorize. In real-case scenarios, where training and test documents come from the same distribution, MEM-QPP is a simple yet useful predictive signal. On the other hand, its ineffectiveness for OOD collections is related to "Challenge Stage 0". If we consider different similarity functions besides Bi-Encoder, we see that they fail to achieve satisfactory results (cfr.

[2]To avoid cluttering, we report Kendall's correlation only when it behaves differently than other measures.

[3]In our experiments, we use the same models as the ones used in retrieval.

Table 1: Performance of MEM-QPP predictor, compared to state-of-the-art QPPs on three collections and four NIR.

	Bi-Encoder		TAS-B		SPLADE		ColBERT-v2	
	Prs. ↑	sARE ↓	Prs. ↑	sARE ↓	Prs. ↑	sARE ↓	Prs. ↑	sARE ↓
	Deep Learning '19 (In-Domain)							
ICTF	0.026	0.312	0.243	0.285	0.099	0.325	0.177	0.299
IDF	0.036	0.310	0.246	0.279	0.099	0.321	0.175	0.296
WIG	0.514	0.237	0.298	0.295	0.187	0.315	0.290	0.286
UEF_{NQC}	0.410	0.224	0.204	0.285	0.212	0.287	0.101	0.315
UEF_{WIG}	**0.618**†	**0.207**†	0.315	0.293	0.260	0.304	0.286	0.286
MEM_{SPLADE}	0.161	0.281	0.066	0.321	0.160	0.300	0.118	0.318
$MEM_{ColBERTv2}$	0.144	0.311	0.187	0.298	0.194	0.293	0.167	0.303
$MEM_{biencoder}$	0.451	0.262	**0.369**†	**0.258**†	**0.430**†	**0.264**†	**0.397**†	**0.259**†
	Robust '04 (Out-of-Domain)							
ICTF	0.021	0.324	0.032	0.317	0.088	0.307	0.027	0.326
IDF	0.039	0.322	0.051	0.315	0.100	0.305	0.046	0.322
WIG	0.636	0.168	0.344	0.260	0.333	0.263	0.507	0.223
UEF_{NQC}	0.640	0.164	**0.427**†	**0.241**†	**0.462**†	**0.232**†	0.502	0.213
UEF_{WIG}	**0.646**†	**0.156**†	0.425†	0.242†	0.436	0.244	**0.548**†	**0.201**†
MEM_{SPLADE}	0.055	0.309	0.009	0.325	-0.042	0.339	-0.015	0.339
$MEM_{ColBERTv2}$	-0.067	0.351	0.130	0.309	0.097	0.320	0.003	0.332
$MEM_{biencoder}$	0.067	0.323	0.158	0.298	0.147	0.302	0.053	0.314
	TREC-COVID (Out-of-Domain)							
ICTF	-0.186	0.376	-0.036	0.345	-0.281	0.374	-0.132	0.358
IDF	-0.139	0.369	0.031	0.339	-0.232	0.365	-0.116	0.363
WIG	**0.562**†	**0.205**†	**0.724**†	**0.167**†	**0.555**†	**0.211**†	**0.381**†	0.258†
UEF_{NQC}	0.155	0.297	0.060	0.351	0.384	0.245	0.045	0.301
UEF_{WIG}	0.479	0.227	0.596	0.198	0.497	0.227	0.329	**0.256**†
MEM_{SPLADE}	0.052	0.324	0.286	0.278	0.027	0.312	0.153	0.282
$MEM_{ColBERTv2}$	-0.108	0.363	-0.198	0.394	-0.173	0.352	-0.034	0.343
$MEM_{biencoder}$	0.099	0.312	0.210	0.286	-0.044	0.322	0.220	0.299

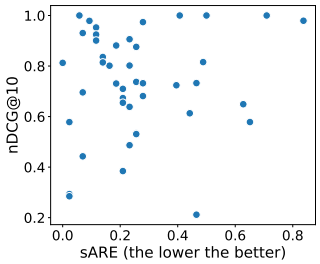

Figure 2: sARE measured for MEM-QPP on predicting the performance of a SPLADE NIR, on Deep Learning '19.

Table 1). To explain this, notice that Bi-Encoder stands apart from methods like SPLADE and ColBERT-v2 in that it uses a bi-encoder design more generally optimized to capture sentence similarity. On the other hand, SPLADE and ColBERT-v2 are more specific architectures dedicated to retrieval. Consequently, when these functions are used to recognize similarities between training and test queries, they exhibit inferior performance. Finally, we are interested in understanding on which queries MEM-QPP performs the best. To this end, we compute the scaled Absolute Rank Error (sARE) [30, 31] between the prediction and the nDCG@10 performance. The presence of observations that lean toward the right on the upper part of Figure 2 indicates that MEM-QPP performs better in predicting when the query is not well handled – lower (sARE) error for lower nDCG. This finding is surprising, given the train-test leakages observed in MS-MARCO [106]. Therefore, a small MEM-QPP indicates that the NIR model did not learn properly the query topic – and the training set could be expanded to account for this.

Table 2: Mean Kolmogorov–Smirnov D-statistic (over queries) on three collections and several TIR and NIR models.

	BM25	LMD	F1EXP	Bi-Encoder	TAS-B	SPLADE	ColBERT-v2
	Deep Learning '19 (In-Domain)						
N_1N_0	0.182	0.223	0.219	0.156	0.149	0.155	0.145
L_1L_0	**0.056**	**0.199**	**0.114**	**0.044**	**0.044**	**0.043**	**0.046**
G_1G_0	0.162	0.265	0.203	0.049	0.059	0.086	0.080
	Robust '04 (Out-of-Domain)						
N_1N_0	0.168	0.175	0.192	0.189	0.148	0.163	0.135
L_1L_0	**0.064**	**0.078**	**0.083**	**0.057**	**0.044**	**0.047**	**0.047**
G_1G_0	0.121	0.101	0.139	0.081	0.064	0.095	0.053
	TREC-COVID (Out-of-Domain)						
N_1N_0	0.144	0.155	0.142	0.127	0.136	0.123	0.107
L_1L_0	**0.055**	**0.087**	**0.069**	0.043	**0.040**	0.039	**0.045**
G_1G_0	0.078	0.101	0.107	**0.040**	**0.040**	**0.038**	**0.045**

4.3 Stage 6 Features: Scores Distributions

Looking at the proposed framework in Figure 1, we observe that Stage 6 features – i.e., the distribution of the scores – are a key aspect of traditional QPP models. If we look back at Equation 1, we see that traditional post-retrieval predictors modeled under such framework include the term $p(d|q, r)$. This term represents the document score in response to the query – assuming its relevance. Therefore, a number of works from the traditional QPP literature focus on modeling the retrieval score distributions [16, 17]. In particular, they aim to determine what are the score distributions for relevant and non-relevant documents and how far apart they are. Cummins [17] shows that, in the TIR scenario, scores for both relevant and non-relevant documents follow a *Log-Normal* distribution. When moving to the NIR setting, we would like to know if considerations for the TIR scenario still hold, or if we should re-consider Stage 6 features. In other terms, we are interested in determining if NIR scores for relevant and non-relevant documents follow the same distributions as TIR ones. If so, this would allow us to adapt classical predictors to neural ranking models. Therefore, we test different distribution pairs to empirically observe which one fits the best to the NIR data. In particular, we experiment with *Normal/Normal*, *Log-Normal/Log-Normal*, and *Gamma/Gamma*, respectively N_1N_0, L_1L_0, and G_1G_0[4]. More specifically, using Maximum Likelihood Estimation, we fit two distributions on the retrieval scores: one for relevant documents $f(s|1)$ and one for non-relevant $f(s|0)$. Then, by referring to λ as the proportion of relevant documents, we define the mixture model of the scores as $f(s) = \lambda\ f(s|1) + (1 - \lambda)\ f(s|0)$. Comparing the fitted mixture model with the observed IR scores gives a measure of goodness-of-fit. Following [17], we then evaluate the fitted mixtures distributions using the Kolmogorov–Smirnov D-statistic. In Table 2, we compare the goodness-of-fit of the mixture to the observed scores for several IR models. In almost all scenarios, regardless of the collection or the IR model considered, the best mixture for the data is the one produced using two *Log-Normal*. It is surprising to note that although BM25, ColBERT-v2, and Bi-Encoder have different score ranges due to their respective architectures, relevant and non-relevant document scores still exhibit a similar distribution – and goodness-of-fit. Thus, the similar probabilistic

[4]We also experimented with other combinations using the *Exponential* distribution, that performed worse and are not reported to avoid cluttering.

distribution between NIR and TIR scores motivates us to use NIR scores in traditional QPPs.

4.4 Stage 5 Features for NIR QPP

We start from what was observed regarding Stage 5 features in Section 3. According to [56, 78, 87], the importance of the regularization term required to compute $p(d|\mathcal{D}@k, r)$ in Equation 1 is key to define a well-grounded QPP model. This regularization allows obtaining comparable predictions across different queries. Depending on the model, the regularization term can be expressed either in the form of the corpus score $s(q, C)$ or as a probability distribution representing the corpus' language model $p(w|\theta_C)$.

When it comes to applying a QPP to a specific TIR model, such regularization can be computed either by *a)* looking at the internal representation of the corpus – e.g., by considering the corpus term frequency – or *b)* actually passing the concatenation of all documents to the model. In the NIR setting, models are trained to score documents, given a query. This prevents us from directly defining the score $s(q, C)$ for NIR models – it is not as straightforward as for TIR to devise indicators that measure the goodness of the entire collection. Secondly, it is computationally unfeasible, at the current time, to compute a representation of the entire corpus by concatenating all the documents: encoders are only able to deal with strings from hundreds to thousands of tokens for the most efficient transformers [91] – while it is reasonable to assume that modern corpora have millions to billions of tokens

The solution we propose consists in approximating the representation of the corpus using the representation of each document. We provide two methodologies, one to compute the corpus representation for sparse models, and the other for dense ones.

4.4.1 Regularization term for sparse NIR models. Sparse models, such as SPLADE, compute high-dimensional sparse representations in the vocabulary space $\mathbb{R}^{|\mathcal{V}|}$ for documents.

Given a document $d_i \in C$ and a term $t_j \in \mathcal{V}$, we define the weight of the term t_j in the document d_i as $d_{ij} = \text{SPLADE}(t_j, d_i)$. If we want to represent the entire corpus, we need to compute a weight c_j for each of its terms. Since the new representation is in the same space as a TIR Bag-of-Words (BoW) model, a straightforward way to represent the weight of each term j in the corpus is:

$$c_j = \frac{\sum_{d_i \in C} d_{ij}}{\sum_{d_i \in C} |d_i|} \tag{2}$$

This weight is unlikely to be the same that the term j would obtain if we would have fed the entire collection to the sparse NIR model. Nevertheless, it can be used to compute the score of the entire corpus given the query (used by WIG, NQC, and SMV) as well as devising a language model to compute the Clarity predictions.

To evaluate the proposed QPPs, we apply them to SPLADE. We first compute the performance of a set of QPPs using the traditional approach based on LM. We then replace it with the SPLADE BoW representations of queries and documents, using the strategy proposed by [60] to use SPLADE together with classical index-based IR libraries. Once the SPLADE BoW inverted index has been computed, we use it and apply classical predictors as they are. We refer to these new predictors as "SPLADE variant of the predictors", and indicate them with "S-".

Table 3: Comparison between traditional SotA predictors and SPLADE versions for a SPLADE run. sMARE, being an error, should be minimized. In bold the best value, † indicates values that are statistically equivalent to the best, while $^{\Delta}$ indicates that the SPLADE version is statistically better than the original one.

	Deep Learning '19 (In-Domain)			Robust '04 (OOD)			TREC-COVID (OOD)		
	Prs. ↑	Knd. ↑	sMARE ↓	Prs. ↑	Knd. ↑	sMARE ↓	Prs. ↑	Knd. ↑	sMARE ↓
Cly	0.014	0.046	0.319	0.263	0.176	0.278	0.079	0.104	0.297
NQC	0.126	0.098	0.309	0.448	0.311	0.236	0.478	0.382	0.218
SMV	0.132	0.094	0.311	0.444	0.306	0.240	0.496	0.403	0.211
WIG	0.187	0.059	0.315	0.333	0.219	0.263	0.555	0.409	0.213
UEF$_{Cly}$	0.252†	0.128	0.293	0.408	0.279	0.244	0.037	0.010	0.320
UEF$_{NQC}$	0.212	0.154	0.287	0.462	0.328	0.232	0.384	0.271	0.245
UEF$_{SMV}$	0.215	0.154	0.287	0.454	0.328	0.232	0.416	0.304	0.236
UEF$_{WIG}$	**0.260**†	0.110	0.304	0.436	0.288	0.244	0.497	0.368	0.227
S-Cly	0.139$^{\Delta}$	0.057$^{\Delta}$	0.312$^{\Delta}$	0.217	0.146	0.288	0.001	0.024	0.337
S-NQC	0.230$^{\Delta}$	0.139$^{\Delta}$	0.294$^{\Delta}$	0.456$^{\Delta}$	0.325$^{\Delta}$	0.232$^{\Delta}$	0.589†	0.433$^{\Delta}$	0.205$^{\Delta}$
S-SMV	0.239†	0.152$^{\Delta}$	0.297$^{\Delta}$	0.449$^{\Delta}$	0.322$^{\Delta}$	0.232$^{\Delta}$	**0.601**†	**0.484**†	**0.189**†
S-WIG	0.234†	0.165$^{\Delta}$	0.297$^{\Delta}$	0.294	0.195	0.272	0.491	0.349	0.229
S-UEF$_{Cly}$	0.231	**0.199**†	**0.270**†	0.423$^{\Delta}$	0.292$^{\Delta}$	0.246	-0.056	-0.055	0.339
S-UEF$_{NQC}$	0.244†	0.157$^{\Delta}$	0.287	**0.474**†	0.335†	0.227†	0.382	0.264	0.261
S-UEF$_{SMV}$	0.238†	0.156	0.293	0.465$^{\Delta}$	**0.336**†	**0.226**†	0.441$^{\Delta}$	0.298	0.243
S-UEF$_{WIG}$	0.236†	0.187†	0.272†	0.439	0.298$^{\Delta}$	0.241	0.287	0.191	0.270

Table 3 reports the empirical results of our analysis. The newly devised predictors are capable of beating the State of the Art consistently on all datasets, and by considering almost every measure. In particular, we observe that, in the case of Deep Learning '19, the SPLADE version of almost all predictors is better than the "original" version, with the exception of UEF-Clarity and UEF-WIG when using Pearson's correlation. Regardless, most of the novel predictors achieve statistically comparable performance to the best-performing system. There is not a unique winner in the case of the Deep Learning '19 collection: UEF-WIG is the best-performing method when Pearson's correlation is used as an evaluation measure, while SPLADE UEF-Clarity is the best when it comes to Kendall's correlation and sMARE. If we consider Robust '04 collection, SPLADE regularization provides an improvement for both NQC and SMV, and for all UEF versions of the predictors, while the performance for Clarity and WIG degrades. This suggests that the variance of the scores is a better indicator of performance when using SPLADE as a retrieval method – NQC and SMV (and thus their UEF variants) being based on such statistics. Akin to what was observed on Deep Learning '19, there is not a unique winning predictor on Robust '04. SPLADE UEF-NQC is the best method in terms of Pearson's correlation, while SPLADE UEF-SMV is the best both with respect to Kendall's correlation and sMARE. Finally, in the case of TREC-COVID, we notice patterns similar to those observed for Robust '04. In particular, only the NQC and SMV actually benefit from the usage of the novel regularization. Nevertheless, the boost obtained by SPLADE-SMV makes it the best performing predictor for the TREC-COVID collection. The improvement observed in Deep Learning '19, not observed on other collections, is similar to what was previously seen for MEM-QPP. This improvement can be attributed to the model used to instantiate the new predictors, which was trained on MS-MARCO. This further highlights the importance of Step 0 features.

4.4.2 Regularization Term for dense NIR models. For dense NIR models, documents and queries are projected into a low-dimensional latent space. Therefore, their representations are d-dimensional vectors, with $d << |\mathcal{V}|$. Such representations are usually obtained by

Table 4: Comparison between traditional SotA predictors and Dense-Centroid versions for four dense NIR runs. sMARE, being an error, should be minimized. In bold the best value, † indicates values that are statistically equivalent to the best, while $^{\Delta}$ indicates that the "DC" version is statistically better than the original one.

	Bi-Encoder		TAS-B		CoCondenser		Contriever	
	Prs. ↑	sMARE ↓	Prs. ↑	sMARE ↓	Prs. ↑	sMARE ↓	Prs. ↑	sMARE ↓
	Deep Learning '19 (In-Domain)							
SMV	0.300	0.246	0.185	0.288	0.347	0.259	0.330	0.250
NQC	0.283	0.258	0.184	0.290	0.338	0.254	0.341	0.248
WIG	0.514	0.237	0.298	0.295	0.102	0.331	0.336	0.282
DCSMV	**0.570**†	**0.202**†	0.321†	0.260†	**0.452**†	0.248†	0.475$^{\Delta}$	0.228†
DCNQC	0.565†	0.213$^{\Delta}$	**0.331**†	**0.259**†	0.444†	0.252$^{\Delta}$	0.496†	**0.224**†
DCWIG	0.445	0.264	0.214	0.284$^{\Delta}$	0.368$^{\Delta}$	**0.244**†	**0.498**†	0.246$^{\Delta}$
	Robust '04 (Out-of-Domain)							
SMV	0.600	0.180	0.366	0.256	0.317	0.272	0.331	0.262
NQC	0.614	0.175	0.376	0.254	0.317	0.264	0.330	0.261
WIG	**0.636**†	0.168†	0.344	0.260	0.390	0.247	0.263	0.278
DCSMV	0.618$^{\Delta}$	0.171$^{\Delta}$	0.437$^{\Delta}$	0.231$^{\Delta}$	0.426$^{\Delta}$	0.245$^{\Delta}$	**0.376**†	0.250†
DCNQC	0.631$^{\Delta}$	**0.167**†	**0.443**†	**0.228**†	**0.434**†	**0.236**†	0.368$^{\Delta}$	**0.249**†
DCWIG	-0.231	0.374	0.346$^{\Delta}$	0.261	0.264	0.270	0.331$^{\Delta}$	0.263$^{\Delta}$
	TREC-COVID (Out-of-Domain)							
SMV	0.241	0.277	0.102	0.326	0.246	0.285	0.022	0.331
NQC	0.211	0.286	0.114	0.329	0.250	0.279	-0.005	0.333
WIG	**0.562**†	**0.205**†	**0.724**†	**0.167**†	**0.677**†	**0.181**†	**0.714**†	**0.168**†
DCSMV	0.441$^{\Delta}$	0.246$^{\Delta}$	0.081	0.336	0.460$^{\Delta}$	0.217$^{\Delta}$	0.097$^{\Delta}$	0.314$^{\Delta}$
DCNQC	0.412$^{\Delta}$	0.250$^{\Delta}$	0.091	0.327$^{\Delta}$	0.483$^{\Delta}$	0.217$^{\Delta}$	0.054$^{\Delta}$	0.314$^{\Delta}$
DCWIG	0.394	0.247	0.044	0.327	0.568	0.209	-0.056	0.338

pooling contextualized term representations of the input sequence – for instance, by averaging or simply considering the `[CLS]` token in the case of BERT. Once all document representations $\mathbf{d_i}$ for $d_i \in C$ have been computed, we can approximate the representation of the entire collection by considering the centroid of all vectors:

$$\mathbf{C} = \frac{\sum_{d_i \in C} \mathbf{d_i}}{|C|} \tag{3}$$

By referring to $\mathbf{q}$ as the representation of the query, the score of the collection is computed as $s(q, C) = \mathbf{q}^T \mathbf{C}$.

Using the novel regularization explicitly designed for dense models, we can instantiate traditional predictors. We call these new predictors Dense-Centroid (DC). Differently from what was observed for sparse models, this collection representation does not allow to instantiate Clarity, since it does not provide a language model.

We report in Table 4 the empirical evaluation of the proposed predictors based on the DC representation of the collection. Table 4 shows that for Deep Learning '19 we are able to improve over the original versions of the models in almost all scenarios. The only exception is WIG for the Bi-Encoder and TAS-B retrievers, where the original model performs the best. For Robust '04 we notice similar patterns, with improved results on almost all scenarios except for the Bi-Encoder, where the original version of WIG remains the best approach in terms of Pearson's correlation, while, for what concerns sMARE the best method is DCNQC. Notice that, for Robust '04 we observe smaller improvements compared to those exhibited on Deep Learning '19. Finally, when it comes to TREC-COVID, the original WIG remains the best-performing solution on all retrieval models. Regardless, the DC versions of both SMV and NQC achieve better results than their original counterparts in the majority of the scenarios, with TAS-B being the only exception.

4.4.3 Discussion. As a first observation, our experiments highlight the importance of Stage 0 features, which at the current time do not exist. Indeed, Tables 1, 3, and 4 show that the improvement is evident for Deep Learning '19 collection, which shares the corpus with the collection on which NIR models have been trained. If we switch to the zero-shot setup, that is TREC-COVID and Robust '04, either the benefit decreases (Tables 3 and 4) or disappears (Table 1). There are also wide differences between target collections. Both Tables 3 and 4 exhibit an advantage in using NIR-derived regularization terms for Robust '04, while the DC procedure fails on the TREC-COVID collection. This allows us to hypothesize that TREC-COVID is more distant from MS-MARCO than Robust '04.

Not all models are equally easy to predict. This emphasizes the need for the community to also focus on Stage 2 features – i.e., those driven by the training procedure. In fact, although dense models may have a similar architecture, their training differs, leading to variations in performance among predictors. For instance, as shown in Table 4, Bi-Encoder does not benefit from MEM-QPP, while TAS-B does, yet Bi-Encoder, in general, tends to be easier to predict.

Not all QPPs benefit from the same features. For example, WIG appears to be the method that benefits the least from the proposed Stage 5 features. This can be explained by considering that WIG is one of the best methods. Therefore, it is likely that the Stage 5 features on which it depends are already expressive enough.

This highlights another advantage of our framework: it allows an understanding of where features come from and which ones should be used in the predictor. Rather than changing features, we can select the best fitting according to the context. Having a complete overview of what features exist, how they can be extracted, and in which setting they work the best, will allow practitioners to combine them into powerful and tailored QPP models.

5 CONCLUSION AND FUTURE WORK

In this work, we propose a novel QPP framework that allows us to interpret and devise features suited for predicting the performance of NIR models. The framework is drawn upon a NIR retrieval pipeline divided into 6 stages, each providing challenges and opportunities for predicting the retrieval performance. We show the benefit of the proposed framework from both descriptive and experimental perspectives. Concerning its descriptive capabilities, we frame traditional QPP models within the proposed framework. We show that most of the classical pre-retrieval predictors exploit Stage 3 and 4 features, while post-retrieval ones are based on Stages 5 and 6. Furthermore, we used the Stage 1 features – the text of the training queries – to define a model-agnostic predictor that exploits the memorization capabilities of NIR models and predicts the performance in the in-domain scenario. Later on, following the framework structure, we proposed a strategy to adapt traditional post-retrieval QPPs. These new predictors modify the regularization term used by classical QPPs to fit sparse and dense architectures and outperform the current state of the art. We argue that our framework can serve as a periodic table for future practitioners to identify unexplored aspects that can further enhance the advantages of QPP systems for the NIR scenario. As future work, we plan to expand the framework to also include re-ranking systems. Furthermore, we plan to adapt the framework to additional tasks where NIR performs best, such as question answering and conversational search.

REFERENCES

[1] Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. 2021. 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*, Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong (Eds.). ACM, 2857–2861. https://doi.org/10.1145/3459637.3482063

[2] Negar Arabzadeh, Mahsa Seifikar, and Charles L. A. Clarke. 2022. Unsupervised Question Clarity Prediction through Retrieved Item Coherency. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, Mohammad Al Hasan and Li Xiong (Eds.). ACM, 3811–3816. https://doi.org/10.1145/3511808.3557719

[3] Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, Feras N. Al-Obeidat, and Ebrahim Bagheri. 2020. Neural embedding-based specificity metrics for pre-retrieval query performance prediction. *Inf. Process. Manag.* 57, 4 (2020), 102248. https://doi.org/10.1016/j.ipm.2020.102248

[4] Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, and Ebrahim Bagheri. 2020. Neural Embedding-Based Metrics for Pre-retrieval Query Performance Prediction. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12036)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 78–85. https://doi.org/10.1007/978-3-030-45442-5_10

[5] Yang Bai, Xiaoguang Li, Gang Wang, Chaoliang Zhang, Lifeng Shang, Jun Xu, Zhaowei Wang, Fangshan Wang, and Qun Liu. 2020. SparTerm: Learning Term-based Sparse Representation for Fast Text Retrieval. https://doi.org/10.48550/ARXIV.2010.00768

[6] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[7] David Carmel and Elad Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00235ED1V01Y201004ICR015

[8] Tao Chen, Mingyang Zhang, Jing Lu, Michael Bendersky, and Marc Najork. 2022. Out-of-Domain Semantics to the Rescue! Zero-Shot Hybrid Retrieval Models. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 95–110. https://doi.org/10.1007/978-3-030-99736-6_7

[9] Xiaoyang Chen, Ben He, and Le Sun. 2022. 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)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 64–74. https://doi.org/10.1007/978-3-030-99739-7_8

[10] Eunseong Choi, Sunkyung Lee, Minijn Choi, Hyeseon Ko, Young-In Song, and Jongwuk Lee. 2022. SpaDE: Improving Sparse Representations Using a Dual Document Encoder for First-Stage Retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 272–282. https://doi.org/10.1145/3511808.3557456

[11] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *CoRR* abs/2102.07662 (2021). arXiv:2102.07662 https://arxiv.org/abs/2102.07662

[12] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2022. Overview of the TREC 2021 deep learning track. In *Text REtrieval Conference (TREC)*. TREC. https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2021-deep-learning-track/

[13] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *CoRR* abs/2003.07820 (2020). arXiv:2003.07820 https://arxiv.org/abs/2003.07820

[14] Stephen Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. 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*, Kalervo Järvelin, Micheline Beaulieu, Ricardo A. Baeza-Yates, and Sung-Hyon Myaeng (Eds.). ACM, 299–306. https://doi.org/10.1145/564376.564429

[15] Stephen Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2004. A framework for selective query expansion. In *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004*, David A. Grossman, Luis Gravano, ChengXiang Zhai, Otthein Herzog, and David A. Evans (Eds.). ACM, 236–237. https://doi.org/10.1145/1031171.1031220

[16] Ronan Cummins. 2012. On the Inference of Average Precision from Score Distributions. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (Maui, Hawaii, USA) *(CIKM '12)*. Association for Computing Machinery, New York, NY, USA, 2435–2438. https://doi.org/10.1145/2396761.2398660

[17] Ronan Cummins. 2014. Document Score Distribution Models for Query Performance Inference and Prediction. *ACM Trans. Inf. Syst.* 32, 1, Article 2 (jan 2014), 28 pages. https://doi.org/10.1145/2559170

[18] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 985–988. https://doi.org/10.1145/3331184.3331303

[19] Zhuyun Dai and Jamie Callan. 2020. 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* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1533–1536. https://doi.org/10.1145/3397271.3401204

[20] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek (Eds.). ACM, 126–134. https://doi.org/10.1145/3159652.3159659

[21] Suchana Datta, Debasis Ganguly, Derek Greene, and Mandar Mitra. 2022. Deep-QPP: A Pairwise Interaction-based Deep Learning Model for Supervised Query Performance Prediction. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). ACM, 201–209. https://doi.org/10.1145/3488560.3498491

[22] Suchana Datta, Debasis Ganguly, Mandar Mitra, and Derek Greene. 2022. A Relative Information Gain-Based Query Performance Prediction Framework with Generated Query Variants. *ACM Transactions on Information Systems* (jun 2022).

[23] Suchana Datta, Sean MacAvaney, Debasis Ganguly, and Derek Greene. 2022. A 'Pointwise-Query, Listwise-Document' Based Query Performance Prediction Approach. In *Proceedings of 45th international ACM SIGIR conference research development in information retrieval*. 2148—-2153. https://doi.org/10.1145/3477495.3531821

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. 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)*. ACL, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[25] Giorgio Maria Di Nunzio and Guglielmo Faggioli. 2021. A Study of a Gain Based Approach for Query Aspects in Recall Oriented Tasks. *Applied Sciences* 11, 19 (2021). https://www.mdpi.com/2076-3417/11/19/9075

[26] Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelo Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Advances in Information Retrieval Theory, 9th International Conference on the Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, July 23, 2023*. ACM. https://doi.org/10.1145/3578337.3605136

[27] Guglielmo Faggioli, Nicola Ferro, Josiane Mothe, and Fiana Raiber. 2023. QPP++ 2023: Query-Performance Prediction and Its Evaluation in New Tasks. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 13982)*. Springer, 388–391. https://doi.org/10.1007/978-3-031-28241-6_42

[28] Guglielmo Faggioli, Nicola Ferro, Cristina Muntean, Raffaele Perego, and Nicola Tonellotto. 2023. A Geometric Framework for Query Performance Prediction in Conversational Search. In *Proceedings of 46th international ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2023 July 23–27, 2023, Taipei, Taiwan*. ACM. https://doi.org/10.1145/3539618.3591625

[29] Guglielmo Faggioli, Thibault Formal, Stefano Marchesin, Stéphane Clinchant, Nicola Ferro, and Benjamin Piwowarski. 2023. Query Performance Prediction for Neural IR: Are We There Yet?. In *Advances in Information Retrieval - 45th European Conference on IR Research, ECIR 2023, Dublin, Ireland, April 2-6, 2023*. 1–18. https://doi.org/10.48550/ARXIV.2302.09947

[30] Guglielmo Faggioli, Oleg Zendel, J. Shane Culpepper, Nicola Ferro, and Falk Scholer. 2021. 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 (Lecture Notes in Computer Science, Vol. 12656)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 115–129. https://doi.org/10.1007/978-3-030-72113-8_8

[31] Guglielmo Faggioli, Oleg Zendel, J. Shane Culpepper, Nicola Ferro, and Falk Scholer. 2022. sMARE: a new paradigm to evaluate and understand query performance prediction methods. *Inf. Retr. J.* 25, 2 (2022), 94–122. https://doi.org/10.1007/s10791-022-09407-w

[32] Hui Fang and ChengXiang Zhai. 2005. An Exploration of Axiomatic Approaches to Information Retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) *(SIGIR '05)*. Association for Computing Machinery, New York, NY, USA, 480–487. https://doi.org/10.1145/1076034.1076116

[33] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval. *CoRR* abs/2109.10086 (2021).

[34] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. 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*. 2353–2359.

[35] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2288–2292. https://doi.org/10.1145/3404835.3463098

[36] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2022. Match Your Words! A Study of Lexical Matching in Neural Information Retrieval. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 120–127.

[37] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. 1987. The Vocabulary Problem in Human-System Communication. *Commun. ACM* 30, 11 (nov 1987), 964–971. https://doi.org/10.1145/32206.32212

[38] Debasis Ganguly and Emine Yilmaz. 2023. Query-specific Variable Depth Pooling via Query Performance Prediction towards Reducing Relevance Assessment Effort. *CoRR* abs/2304.11752 (2023). https://doi.org/10.48550/arXiv.2304.11752 arXiv:2304.11752

[39] Luyu Gao and Jamie Callan. 2022. 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*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 2843–2853. https://doi.org/10.18653/v1/2022.acl-long.203

[40] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 3030–3042. https://doi.org/10.18653/v1/2021.naacl-main.241

[41] Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic Models for the First-Stage Retrieval: A Comprehensive Review. *ACM Trans. Inf. Syst.* 40, 4 (2022), 66:1–66:42. https://doi.org/10.1145/3486250

[42] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. 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*, Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi (Eds.). ACM, 55–64. https://doi.org/10.1145/2983323.2983769

[43] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2019. 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*, Yi Fang, Yi Zhang, James Allan, Krisztian Balog, Ben Carterette, and Jiafeng Guo (Eds.). ACM, 55–58. https://doi.org/10.1145/3341981.3344249

[44] Claudia Hauff. 2010. Predicting the effectiveness of queries and retrieval systems. *SIGIR Forum* 44, 1 (2010), 88. https://doi.org/10.1145/1842890.1842906

[45] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. 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*, James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury (Eds.). ACM, 1419–1420. https://doi.org/10.1145/1458082.1458311

[46] Ben He and Iadh Ounis. 2006. Query performance prediction. *Inf. Syst.* 31, 7 (2006), 585–594. https://doi.org/10.1016/j.is.2005.11.003

[47] Jiyin He, Martha A. Larson, and Maarten de Rijke. 2008. 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 (Lecture Notes in Computer Science, Vol. 4956)*, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (Eds.). Springer, 689–694. https://doi.org/10.1007/978-3-540-78646-7_80

[48] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. 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*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 113–122. https://doi.org/10.1145/3404835.3462891

[49] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. ACM International Conference on Information and Knowledge Management (CIKM). https://www.microsoft.com/en-us/research/publication/learning-deep-structured-semantic-models-for-web-search-using-clickthrough-data/

[50] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. PACRR: A Position-Aware Neural IR Model for Relevance Matching. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 1049–1058. https://doi.org/10.18653/v1/d17-1110

[51] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards Unsupervised Dense Information Retrieval with Contrastive Learning. *CoRR* abs/2112.09118 (2021). arXiv:2112.09118 https://arxiv.org/abs/2112.09118

[52] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* 7, 3 (2021), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[53] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[54] Omar Khattab and Matei Zaharia. 2020. 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*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 39–48. https://doi.org/10.1145/3397271.3401075

[55] Maryam Khodabakhsh and Ebrahim Bagheri. 2021. Semantics-enabled query performance prediction for ad hoc table retrieval. *Inf. Process. Manag.* 58, 1 (2021), 102399. https://doi.org/10.1016/j.ipm.2020.102399

[56] Oren Kurland, Anna Shtok, Shay Hummel, Fiana Raiber, David Carmel, and Ofri Rom. 2012. Back to the roots: a probabilistic framework for query-performance prediction. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, Xuewen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki (Eds.). ACM, 823–832. https://doi.org/10.1145/2396761.2396866

[57] Carlos Lassance and Stéphane Clinchant. 2022. An Efficiency Study for SPLADE Models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2220–2226. https://doi.org/10.1145/3477495.3531833

[58] Hang Li, Shuai Wang, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, Jimmy Lin, and Guido Zuccon. 2022. To Interpolate or Not to Interpolate: PRF, Dense and Sparse Retrievers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2495–2500. https://doi.org/10.1145/3477495.3531884

[59] Jimmy Lin and Xueguang Ma. 2021. A Few Brief Notes on DeepImpact, COIL, and a Conceptual Framework for Information Retrieval Techniques. https://doi.org/10.48550/ARXIV.2106.14807

[60] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2356–2362. https://doi.org/10.1145/3404835.3463238

[61] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. Association for Computational Linguistics, Online, 163–173. https://doi.org/10.18653/v1/2021.repl4nlp-1.17

[62] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. *Trans. Assoc. Comput. Linguistics* 9 (2021), 329–345. https://doi.org/10.1162/tacl_a_00369

[63] Simon Lupart, Thibault Formal, and Stéphane Clinchant. 2022. MS-Shift: An Analysis of MS MARCO Distribution Shifts on Neural Retrieval. https://doi.org/10.48550/ARXIV.2205.02870

[64] Sean MacAvaney and Luca Soldaini. 2023. One-Shot Labeling for Automatic Relevance Estimation. *CoRR* abs/2302.11266 (2023). https://doi.org/10.48550/arXiv.2302.11266 arXiv:2302.11266

[65] Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. 2021. Learning Passage Impacts for Inverted Indexes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1723–1727. https://doi.org/10.1145/3404835.3463030

[66] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to Match using Local and Distributed Representations of Text for Web Search. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 1291–1299. https://doi.org/10.1145/3038912.3052579

[67] Josiane Mothe and Ludovic Tanguy. 2005. Linguistic features to predict query difficulty. In *ACM Conference on research and Development in Information Retrieval, SIGIR, Predicting query difficulty - methods and applications workshop*. Salvador de Bahia, Brazil, 7–10.

[68] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. https://doi.org/10.48550/ARXIV.1901.04085

[69] Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docTTTTTquery.

[70] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. 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)*. Association for Computational Linguistics, Hong Kong, China, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[71] Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2825–2835. https://doi.org/10.18653/v1/2021.emnlp-main.224

[72] Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2022. A Thorough Examination on Zero-shot Dense Retrieval. https://doi.org/10.48550/ARXIV.2204.12755

[73] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen M. Voorhees, Lucy Lu Wang, and William R. Hersh. 2021. Searching for scientific evidence in a pandemic: An overview of TREC-COVID. *J. Biomed. Informatics* 121 (2021), 103865. https://doi.org/10.1016/j.jbi.2021.103865

[74] Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. 0–.

[75] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

[76] Haggai Roitman. 2018. Enhanced Performance Prediction of Fusion-based Retrieval. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2018, Tianjin, China, September 14-17, 2018*, Dawei Song, Tie-Yan Liu, Le Sun, Peter Bruza, Massimo Melucci, Fabrizio Sebastiani, and Grace Hui Yang (Eds.). ACM, 195–198. https://doi.org/10.1145/3234944.3234950

[77] Haggai Roitman. 2019. Normalized Query Commitment Revisited. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 1085–1088. https://doi.org/10.1145/3331184.3331334

[78] Haggai Roitman, Shai Erera, Oren Sar Shalom, and Bar Weiner. 2017. Enhanced Mean Retrieval Score Estimation for Query Performance Prediction. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017*, Jaap Kamps, Evangelos Kanoulas, Maarten de Rijke, Hui Fang, and Emine Yilmaz (Eds.). ACM, 35–42. https://doi.org/10.1145/3121050.3121051

[79] Dwaipayan Roy, Debasis Ganguly, Mandar Mitra, and Gareth J. F. Jones. 2019. Estimating Gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. *Inf. Process. Manag.* 56, 3 (2019), 1026–1045. https://doi.org/10.1016/j.ipm.2018.10.009

[80] Andrew Rutherford. 2011. *ANOVA and ANCOVA: a GLM approach*. John Wiley & Sons.

[81] Gerard Salton and Chris Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* 24, 5 (1988), 513–523.

[82] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. 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*. 3715–3734.

[83] Harrisen Scells, Leif Azzopardi, Guido Zuccon, and Bevan Koopman. 2018. Query Variation Performance Prediction for Systematic Reviews. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 1089–1092. https://doi.org/10.1145/3209978.3210078

[84] Falk Scholer, Hugh E. Williams, and Andrew Turpin. 2004. Query association surrogates for Web search. *J. Assoc. Inf. Sci. Technol.* 55, 7 (2004), 637–650. https://doi.org/10.1002/asi.20011

[85] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple Entity-Centric Questions Challenge Dense Retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6138–6148. https://doi.org/10.18653/v1/2021.emnlp-main.496

[86] Anna Shtok, Oren Kurland, and David Carmel. 2010. 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*, Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy (Eds.). ACM, 259–266. https://doi.org/10.1145/1835449.1835494

[87] Anna Shtok, Oren Kurland, and David Carmel. 2016. Query Performance Prediction Using Reference Lists. *ACM Trans. Inf. Syst.* 34, 4 (2016), 19:1–19:34. https://doi.org/10.1145/2926790

[88] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting Query Performance by Query-Drift Estimation. *ACM Trans. Inf. Syst.* 30, 2 (2012), 11:1–11:35. https://doi.org/10.1145/2180868.2180873

[89] Michael Tänzer, Sebastian Ruder, and Marek Rei. 2022. Memorisation versus Generalisation in Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 7564–7578. https://doi.org/10.18653/v1/2022.acl-long.521

[90] Yongquan Tao and Shengli Wu. 2014. 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*, Jianzhong Li, Xiaoyang Sean Wang, Minos N. Garofalakis, Ian Soboroff, Torsten Suel, and Min Wang (Eds.). ACM, 1891–1894. https://doi.org/10.1145/2661829.2661906

[91] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient Transformers: A Survey. *ACM Comput. Surv.* 55, 6, Article 109 (dec 2022), 28 pages. https://doi.org/10.1145/3530811

[92] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. 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*.

[93] Paul Thomas, Falk Scholer, Peter Bailey, and Alistair Moffat. 2017. Tasks, Queries, and Rankers in Pre-Retrieval Performance Prediction. In *Proceedings of the 22nd Australasian Document Computing Symposium, ADCS 2017, Brisbane, QLD, Australia, December 7-8, 2017*, Bevan Koopman, Guido Zuccon, and Mark James Carman (Eds.). ACM, 11:1–11:4. https://doi.org/10.1145/3166072.3166079

[94] John W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2 (1949), 99–114. http://www.jstor.org/stable/3001913

[95] Ellen Voorhees. 2005. Overview of the TREC 2004 Robust Retrieval Track. https://doi.org/10.6028/NIST.SP.500-261

[96] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online. https://aclanthology.org/2020.nlpcovid19-acl.1

[97] Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. BERT-Based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (Virtual Event, Canada) *(ICTIR '21)*. Association for Computing Machinery, New York, NY, USA, 317–324. https://doi.org/10.1145/3471158.3472233

[98] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online,*

November 16-20, 2020, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[99] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 55–64. https://doi.org/10.1145/3077136.3080809

[100] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. 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*. OpenReview.net. https://openreview.net/forum?id=zeFrfgyZln

[101] Liu Yang, Qingyao Ai, Jiafeng Guo, and W. Bruce Croft. 2016. aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi (Eds.). ACM, 287–296. https://doi.org/10.1145/2983323.2983818

[102] Hamed Zamani, W. Bruce Croft, and J. Shane Culpepper. 2018. 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*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 105–114. https://doi.org/10.1145/3209978.3210041

[103] Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik G. Learned-Miller, and Jaap Kamps. 2018. From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.). ACM, 497–506. https://doi.org/10.1145/3269206.3271800

[104] Oleg Zendel, Anna Shtok, Fiana Raiber, Oren Kurland, and J. Shane Culpepper. 2019. Information Needs, Queries, and Query Performance Prediction. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 395–404. https://doi.org/10.1145/3331184.3331253

[105] Cheng Xiang Zhai. 2008. Statistical Language Models for Information Retrieval: A Critical Review. *Found. Trends Inf. Retr.* 2, 3 (2008), 137–213. https://doi.org/10.1561/1500000008

[106] Jingtao Zhan, Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2022. Evaluating Interpolation and Extrapolation Performance of Neural Retrieval Models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 2486–2496. https://doi.org/10.1145/3511808.3557312

[107] Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. SPARTA: Efficient Open-Domain Question Answering via Sparse Transformer Matching Retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 565–575. https://doi.org/10.18653/v1/2021.naacl-main.47

[108] Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. 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 (Lecture Notes in Computer Science, Vol. 4956)*, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (Eds.). Springer, 52–64. https://doi.org/10.1007/978-3-540-78646-7_8

[109] Yun Zhou and W. Bruce Croft. 2007. 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*, Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando (Eds.). ACM, 543–550. https://doi.org/10.1145/1277741.1277835