A Geometric Framework for Query Performance Prediction in Conversational Search

Guglielmo Faggioli
guglielmo.faggioli@unipd.it
University of Padova
Padova, Italy

Nicola Ferro
ferro@dei.unipd.it
University of Padova
Padova, Italy

Cristina Ioana Muntean
cristina.muntean@isti.cnr.it
ISTI-CNR
Pisa, Italy

Raffaele Perego
raffaele.perego@isti.cnr.it
ISTI-CNR
Pisa, Italy

Nicola Tonellotto
nicola.tonellotto@unipi.it
University of Pisa
Pisa, Italy

ABSTRACT

Thanks to recent advances in IR and NLP, the way users interact with search engines is evolving rapidly, with multi-turn conversations replacing traditional one-shot textual queries. Given its interactive nature, Conversational Search (CS) is one of the scenarios that can benefit the most from Query Performance Prediction (QPP) techniques. QPP for the CS domain is a relatively new field and lacks a proper framing. In this study, we address this gap by proposing a framework for the application of QPP in the CS domain and use it to evaluate the performance of predictors. We characterize what it means to predict the performance in the CS scenario, where information needs are not independent queries but a series of closely related utterances. We identify three main ways to use QPP models in the CS domain: as a diagnostic tool, as a way to adjust the system's behaviour during a conversation, or as a way to predict the system's performance on the next utterance. Due to the lack of established evaluation procedures for QPP in the CS domain, we propose a protocol to evaluate QPPs for each of the use cases. Additionally, we introduce a set of spatial-based QPP models designed to work the best in the conversational search domain, where dense neural retrieval models are the most common approaches and query cutoffs are typically small. We show how the proposed QPP approaches improve significantly the predictive performance over the state-of-the-art in different scenarios and collections.

CCS CONCEPTS

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

KEYWORDS

Conversational Search, QPP, Dense Representation

ACM Reference Format:
Guglielmo Faggioli, Nicola Ferro, Cristina Ioana Muntean, Raffaele Perego, and Nicola Tonellotto. 2023. A Geometric Framework for Query Performance Prediction in Conversational Search. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591625

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591625

1 INTRODUCTION

Conversational Search (CS) is the Information Retrieval (IR) paradigm where users converse with an automatic agent to satisfy their information needs. CS allows for an intuitive human-machine interaction since the user interrogates the machine using natural language. Rhetorical figures – e.g., anaphoras, ellipses, coreferences – and complex speech constructs in users' utterances make CS challenging. Nevertheless, thanks to recent advances in Natural Language Processing (NLP) and Neural Information Retrieval (NIR) and the advent of Large Language Models (LLMs), it is becoming increasingly popular and ubiquitously adopted. CS can benefit from employing Query Performance Prediction (QPP) techniques in tasks such as determining the utterance rewriting approach to adopt, identifying the topic shifts, or determining if the system needs to ask the user clarifying questions.

QPP is the task of estimating the performance of an IR system in the absence of human-assessed relevance judgements [5]. It has been successfully employed in many tasks, such as query suggestion [47], adaptive model selection [36, 42, 47, 48], and pathological queries discovering [5]. We argue that the QPP for CS cannot be addressed using the traditional strategies due to the profound differences between CS and classical adhoc-ish IR. In a similar fashion to what was pointed out by Hashemi et al. [18] about Question Answering (QA), we can observe that CS involves *i)* highly precision-oriented metrics and small cutoff retrieval, while traditional QPP techniques have been often devised and tested to predict Average Precision (AP) at large cutoffs; *ii)* retrieving passages or short documents, while classical QPP techniques are often designed for long documents; *iii)* heavy usage of NIR techniques which have not been yet explored extensively in the QPP domain; *iv)* in the CS domain, utterances are correlated and grouped into conversations, and therefore this characteristic should be taken into consideration, at least when evaluating the QPP models. While a good share of effort has been devoted to both the CS and QPP tasks alone, at the current time only a few works studied the application of QPP techniques to CS. Most of these works rely on the use of well-established classical QPP methods to choose how the system should interact with the

user [1] or to determine if the answer provided to the user contains the relevant information [39], without taking into consideration all the peculiarities of the CS domain described above. In this work, we aim at address this gap by proposing a set of predictors explicitly designed to synergize the best with CS models. We start by considering that most of the modern CS approaches rely on NIR techniques. Thus, we focus on CS models that exploit documents' and queries' dense representations and propose QPP methodologies relying on measuring how close retrieved documents' representations are to the query. We devise two predictors that measure the volume of the hypercube encompassing the top k retrieved documents in response to a given query and show that such quantity effectively correlates with the actual performance achieved. While we stress the importance of QPP models for the conversational scenario, we want to point out that, at the current time, it is yet missing a framework describing how to correctly address QPP evaluation in the CS scenario. Therefore, we detail a set of use cases and devise practical evaluation protocols to assess the QPP models in each of them.

Our research contributions are the following:

- **RC1** Identify possible applications of QPP in the CS settings and relevant figures that practitioners might be interested in predicting.
- **RC2** Define how the QPP evaluation should be adapted to correctly determine and compare the performance achieved by QPP models for CS.
- **RC3** Devise a QPP model that relies on the specific characteristics of the CS task, namely the heavy usage of dense representations and precision-oriented measures.

To deliver these research contributions, we first define three main possible use cases for QPP models in the CS scenario: *i)* QPP as a diagnostic tool to evaluate CS systems in a post-hoc fashion after the system has been deployed; *ii)* QPPs as a means to assess the system's behavior within the conversation; *iii)* QPP to predict "on-line" how well the next user's utterance will perform.

For each use case we propose an evaluation methodology capable of measuring the performance of a QPP model according to its purpose. For the first use case, we define a collection-wise evaluation, that mimics the current QPP evaluation approach and is suited for scenarios where each utterance is considered an independent event. For the second use case, we define a conversation-wise evaluation approach that allows measuring the coherence within the conversation, to understand whether there are specific conversation topics where the IR system is bound to fail. For the third use case, we define an utterance-wise methodology, that describes how the single utterance is expected to perform. Finally, we propose a set of QPP models explicitly designed to work the best in the CS scenario that exploit morphological characteristics of the embedding space used to represent conversational utterances and documents. We show that, depending on the scenario, the proposed QPP techniques can overcome the current state of the art up to 8.9%.

The remainder of this manuscript is organized as follows: Section 2 surveys the main efforts in CS, QPP and QPP applied to CS. Section 3 introduces our definition of QPP applied to CS and how to evaluate them. Section 4 formalizes the proposed predictors while Section 5 details the empirical evaluation. Finally, Section 6 concludes and outlines our future work.

2 RELATED WORK

2.1 Conversational Search

CS implies a dialogue made of natural language utterances between a user and a conversational agent. The main challenge in this setting is keeping track of the conversation context [35]. The users might in fact shift topics throughout the dialogue or make references to previously mentioned topics [1, 30], asking for details or clarifications. Differently from ad-hoc search, CS systems focus on the reply, might ask clarifying questions and offer the user a single answer rather than several sources.

Several works focus on rewriting the utterances by reusing context from the dialogue [28–30, 50, 54] so as to build self-explanatory queries suitable for the search engine. A slightly different line of research addresses CS systems based on dense retrieval models [16, 57]. Yu et al. propose a few-shot generative approach to conversational query rewriting [56]. The authors develop two methods, based on rules and self-supervised learning, to generate weak supervision data using large amounts of ad-hoc search sessions. These data are used to fine-tune GPT-2 to rewrite conversational queries. Their experiments show that GPT-2 effectively learns to capture context dependencies, even for hard cases involving long-turn dependencies. Yu et al. also propose ConvDR [57], a query rewriting model for conversational dense retrieval. It initially uses the ANCE model [52] to encode both documents and queries with dense representations and then, using a teacher-student model, uses the context and query to learn an enriched representation similar to the one of the manually rewritten query.

2.2 Query Performance Prediction

QPP consists in estimating search effectiveness in the absence of human relevance judgments [5]. While its definition suggests that the primary function for QPP models is being a diagnostic tool to evaluate IR models with a reduced cost, they also proved useful in a number of interactive IR tasks. Examples of such tasks include selecting the best model given the user query [5, 47] or identifying the best query rewritings [12, 42, 47]. Other usages in which QPPs are particularly effective are as an external signal in rank fusion algorithms [36] and as a tool to diagnose pathological queries that require the system administrator's intervention to provide additional relevant documents or labeling [5].

A common classification for QPP models consists in dividing them into pre- and post-retrieval predictors [5, 20, 21]. In particular, pre-retrieval predictors rely on features that are available prior to the retrieval phase, such as the query terms collection frequency [61] or linguistic features, such as query terms polysemy and synonymy degrees [31]. Post-retrieval predictors on the other hand use the similarity between the documents and the query and so require one (or more) retrieval phases to compute the prediction.

A recently developed line of research in the QPP scenario, involves the usage of dense terms representations to devise QPP models. Roy et al. [40] experiment with pre-retrieval predictors in the NIR domain. In particular, they show how the distribution in the space of word vectors with respect to query vectors correlates with the performance of the system. They noticed that, also in this scenario, pre-retrieval QPP alone are not capable of achieving satisfactory results, and therefore combined it with post-retrieval

methods, showing performance improvement. Similarly, Arabzadeh et al. [4] propose, and extend in [3], a set of measures based on neural embeddings aimed at measuring the specificity of each term – specific query terms are likely to allow for better identification of relevant documents. Such measures behave as pre-retrieval predictors and are shown to correlate with the system's performance. Differently from the approaches proposed in this manuscript, both Roy et al. [40] and Arabzadeh et al. [4] consider pre-retrieval techniques.

2.3 QPP for Conversational Search

The research community has already recognized some of the advantages that properly applied QPPs can provide to CS [38, 58]. Traditionally, the QPP task in the conversational search domain has been declined into three distinct lines of research: *i)* QPP models for QA and passage retrieval; *ii)* information elicitation; and *iii)* Query Rewriting. A contiguous research line concerns the prediction of user satisfaction in interacting with a conversational agent [24, 33]. Notice that, these works focus on predicting aspects related to human-computer interaction and, therefore, differ from traditional QPP which focuses mostly on offline experimentation.

Passage Retrieval and QA. One of the first QPP models that switches from traditional document-based retrieval to passage-retrieval is [27]. More in detail, Krikon et al. [27] devise a post-retrieval predictor that employs named entities to determine if a passage contains the answer to the user's question. Similarly, Roitman [37] breaks down documents into passages and exploits information in such passages – namely the one with the maximum retrieval score for the query – to devise a new QPP technique. One of the seminal approaches of QPP in the CS domain is represented by Roitman et al. [39]. In particular, [39] propose a method to filter answers of a conversational system, based on the predicted probability that such answers respond to the user information need.More in detail, multiple classifiers are trained using several sets of features (traditional LETOR features, pre- and post-retrieval QPP scores, passage-based calibration scores) to the task considered, achieving good performance. The same task is also tackled by Tan et al. [45] and Hasanain and Elsayed [17]. This task has some commonalities with the traditional QPP and is a possible application of QPP techniques, but it is intrinsically different from the performance prediction that we tackle in this manuscript. Hashemi et al. [18] devise a BERT-based approach to carry out QPP in the QA scenario that outperforms traditional ad-hoc retrieval QPP, showing the importance of QPP techniques specifically designed for a given IR task. Even though strictly related, QPP models designed for passage retrieval and QA do not take into consideration idiosyncrasies of the CS setting, such as the natural inter-correlation between utterances: with this work, we aim at filling this gap.

QPP models for Mixed-Initiative Conversation Agents. QPP models have proved to be an effective tool to help in deciding if and what information the conversational agent should elicit from the user to better understand the context.

Pal and Ganguly [32] propose an approach based on preexisting QPP models to identify which concepts and entities named in a conversation (either between two humans or in the utterances issued by the user) need further context in order to be understood.Arabzadeh et al. [2] use QPP to predict whether the system needs to ask a clarifying question to properly understand the user's query. Their QPP is based on constructing a coherency network with a LLM and computing some centrality measures on it.

Aliannejadi et al. [1] exploit the power of QPP in the conversational domain, to decide which question the system should issue next to understand what the user is looking for, based on the user's previous answers. They use the post-retrieval QPP predictor proposed by Pérez-Iglesias and Araujo [34]. Similarly, Hashemi et al. [19] use the predictor in [34] to decide which one, among a set of possible clarifying questions, to submit to the user.

[1, 2, 19, 32] adopt QPP techniques devised for the traditional ad-hoc IR, we argue that QPP models explicitly devised for the conversational scenario can further improve the performance of information elicitation systems.

Query Rewriting. A recent line of research explored the possibility of applying QPP to decide which is the best approach to query expansion. For example, Lin et al. [29] use a degenerated QPP system (i.e., the score produced by BM25) to decide whether to expand or not the utterance with additional tokens – if BM25 scores are low, then there is the need for expansion, vice-versa, the utterance can be processed without further modifications.

Similar to what was observed for the mixed-initiative scenario, we consider the development of QPP models explicitly designed for the CS task beneficial also to adaptive query rewriters.

3 QPP IN THE CS DOMAIN

While multiple sources [1, 38, 58] recognize the advantages that QPP brings to CS (See Subsection 2.3), defining what it means to predict performance in the conversational domain is still an open issue. Furthermore, when switching from single independent queries to highly correlated utterances grouped into conversations, it is necessary to determine how to properly assess the performance of a QPP model for a CS system.

In the remainder of this section, we introduce a categorization of the possible objectives and types of predictions that we could carry out in the CS domain and define a unified framework (Subsection 3.1). Subsequently, in Subsection 3.2 we discuss possible evaluation methodologies usable to assess the performance of a QPP model with respect to the prediction task.

3.1 RC1: QPP Use Cases in CS

Defining what it means to predict the performance of a CS system is related to how we measure its performance. For example, we might consider a system to be the best if it is capable of optimizing the performance of each utterance, regardless of the conversation it is taken from. Alternatively, we might be interested in predicting the performance achieved over the entire conversation, to understand more in detail how the system behaves when queried on a specific topic. Finally, we might be interested in determining how each utterance will behave with respect to other utterances within the same conversation, e.g. to counterbalance topic shifts. With this categorization in mind, we define the following possible use cases for a QPP model in the conversational search domain.

3.1.1 Collection-wise prediction. we evaluate the performance over the entire collection, once multiple conversations already happened.
Which signals it employs: past, current and "future" signals – to predict the performance we are allowed to use signals from subsequent utterances.
Effective for: evaluating a system without relevance judgements. QPP can be used to assess how well the system performed in a purely offline setting.
Not effective for: adapting the system to a conversation as it happens: since it exploits "future" signals that are not available on-line, it is not suited to adapt the system while it runs.

3.1.2 Conversation-wise prediction. We assume that the user has just issued a query and received a response. We aim at measuring the performance of the system up to the current moment in the conversation, either considering the average performance accrued up to the current point in the conversation, or for the latest utterance.
Which signals it employs: past and current signals, user's feedback on the current utterance, if available or elicited.
Effective for: evaluating a system without relevance judgements and understanding how the system is performing; if we predict that the current utterance fails, we could also adapt the system online.
Not effective for: devising general trends in the collection.

3.1.3 Utterance-wise prediction. We aim at predicting the performance of the next utterance before the ranked list is presented to the user.
Which signals it employs: past and current signals that are available prior to the response being provided to the user.
Effective for: model selection, query suggestion, topic shift detection, mixed-initiative interactions.
Not effective for: While allowing for a punctual analysis of the single utterance, it might fail in grasping general patterns that arise within conversations or are due to specific features of the collection.

3.2 RC2: Evaluation Procedure

3.2.1 Collection-wise evaluation (sota approach). This approach is devised for the use case in 3.1.1. It corresponds exactly to the state-of-the-art evaluation procedure applied in QPP. In detail, the most common strategy to assess whether a QPP is performing well consists in computing a prediction score for each utterance (query, in the ad-hoc IR case), which we refer to as $\hat{p}_i$, and the actual performance achieved on the utterance, indicated as p_i. Then, the correlation $\rho(p, \hat{p})$ between the two lists $\hat{p} = [\hat{p}_1, ..., \hat{p}_n]$ and $p = [p_1, ..., p_n]$ is computed. A large correlation suggests a well-performing QPP. Examples of possible correlations commonly adopted in the traditional QPP scenario include Pearson, Spearman, Kendall, and RBO [51]. A possible drawback of adopting this evaluation strategy is linked to the fact that it does not allow encompassing the natural correlation between utterances derived from the same conversation. It is well-known [13] that utterances from the same conversation tend to naturally have more similar performance than those achieved by utterances for different conversations. When applying this evaluation methodology straightforwardly, we are disregarding this aspect and treating each utterance as an independent query.

3.2.2 Conversation-wise evaluation (proposed approach). This approach is devised for the use case in 3.1.2. A natural approach to extending the classical QPP evaluation to the CS scenario consists in treating each conversation as a single evaluation instance. Therefore, similarly to the previous case, we can compute $p_{c,i}$ and $\hat{p}_{c,i}$ that describe respectively the performance and the prediction score of the i-th utterance within the c-th conversation. Once the performance and prediction scores lists p_c and $\hat{p}_c$ have been computed, it is possible to measure the correlation between the two and use this value as a performance indicator for the QPP on the c-th conversation. This has the advantage of allowing for a pointwise evaluation: we can determine on which conversations our predictor works properly, and carry out failure analysis. Conversely, it also presents a weakness in terms of "accuracy": the decreased number of utterances considered to compute the correlation leads to greater performance variability within the single conversation. This new approach allows revising the way in which we can carry out the statistical comparison of our approach with baselines. Having a performance score for each conversation allows for carrying out proper statistical analysis, keeping into consideration the effect of each conversation (as done for the topics in IR), something that could have not been done with the traditional evaluation based on correlations.

3.2.3 Utterance-wise evaluation (proposed approach). This approach is devised for the use case in 3.1.3. In particular, it is possible to rely on an utterance-based procedure that allows further breaking down the QPP performance over different factors, such as the type of utterance. Moving from a list-wise correlation-based strategy to an utterance-wise one requires also to change the performance indicator used to evaluate the QPP. In particular, we cannot rely anymore on the correlation measures, but we need to switch to the scaled Absolute Ranked Error (sARE) evaluation [14, 15]. sARE is a point-wise evaluation measure that allows determining the performance achieved by the QPP model on a single query. Given a query q and its true performance and prediction scores p_q and $\hat{p}_q$, sARE is defined as $sARE(q) = \frac{|r_q^e - r_q^p|}{|Q|}$, where r_q^e and r_q^p are respectively the ranks of p_q and $\hat{p}_q$, the ordinal positions of p_q and $\hat{p}_q$ if we sorted the list of performances and predicted scores for all the queries and $|Q|$ is the number of queries. Notice that, akin to the previous cases, sARE fits both a conversation and collection-oriented evaluation procedure. In fact, r_q^e and r_q^p can be computed either with respect to the entire collection or only by considering utterances within the specific conversation. It is also possible to devise a global counterpart of the sARE measure, by computing its average over all queries – this aggregation is called scaled Mean Absolute Ranked Error (sMARE). Being an error, the lower the sARE, the better performs the evaluated QPP.

The main advantage deriving from this last evaluation methodology is that it allows for a very precise breakdown of the performances according to different factors. For example, we can treat separately the performance for the first utterances of each conversation, which are likely to be easier than others from the IR model perspective, and utterances referring to previous ones.

4 RC3: HYPER-VOLUME BASED PREDICTORS

One contribution of the proposed framework is that it relies on the geometric properties of the vector space in which queries and documents are represented. We do not only focus on similarity aspects, such as the angle between the query's and documents' representations – as typically done in the retrieval phase – but we focus on topological properties and how such vectors are distributed in the multidimensional space. In Subsection 4.1 we describe how dense representations are typically constructed in the CS setting, focusing on the models that we are going to use in the experimental evaluation: STAR and ConvDR. Later on, in Subsection 4.2, we present two predictors that exploit the geometric properties of the dense representation space to compute a prediction score. To the best of our knowledge, we are among the first to consider a geometric-driven post-retrieval predictor based on geometric and topological characteristics of dense vector representations.

4.1 Dense representations for CS

While IR-related tasks were previously dominated by lexical signals, the introduction of neural models based on LLMs changed drastically the way in which queries and documents are represented and matched. We focus in this work on single-representation dense retrieval models where documents and queries are encoded with low-dimension vectors within the same embedding space, capturing the semantic relations between documents and queries. In this setting, document embeddings are stored for efficient access in a specialized metric index, such as that provided by the FAISS toolkit [23]. Given a query, embedded in the same multi-dimensional space of the collection documents, online ranking is performed by means of a top-k nearest neighbor similarity search based on a distance function, e.g., L2 or Inner Product. The top-ranked documents are the ones closer, and thus more similar, to the issued query. We consider different state-of-the-art single-representation models such as DPR [25], ANCE [53], and STAR [60], but eventually select STAR for our experiments because the model uses hard negative sampling during fine-tuning, rather than random sampling, obtaining better representations in terms of effectiveness w.r.t. ANCE and DPR. Our CS system uses STAR to encode CAsT queries (original and manually rewritten) and documents as embeddings with 768 dimensions. On the other hand, for testing with state-of-the-art automatic query rewriting techniques, we use CondDR [57] to obtain dense query representations including the conversation context. ConvDR learns contextualized embeddings for multi-turn conversational queries by using the current query together with the previous utterances in order to build up the context. A few-shot teacher-student learner tries to close the gap between the obtained dense representation and the golden-standard one corresponding to the manually rewritten query. ConvDR, akin to STAR, generates 768-dimensional vectors.

4.2 Performance Prediction

Concerning the notation, we call v_q the d-dimensional vector representation of the utterance, and D_i the vector representation of the i-th document retrieved in response to the information need. As commonly done for most post-retrieval predictors, only the top-k documents retrieved are considered in computing the prediction – we will refer to the set of the top-k retrieved documents as $\mathcal{D}@k$.

4.2.1 Hyper-cube definition. Given a query q, we consider the top-k documents retrieved to answer it. Given the multi-dimension representation of the query v_q and documents $D_1, ..., D_k$, we would like to compute how densely such documents distribute around the query. If the hyperspace around the query is densely occupied by retrieved documents, we assume that the model correctly characterizes the query in a semantic way and understands which documents are strongly correlated with it. Vice-versa, if the space is loosely occupied, we can assume that retrieved documents are not meaningful or semantically close to the query itself. Nevertheless, given the absence of a fixed defined reference space, we cannot compute a density. Therefore, we consider the volume that encompasses the query and all the top-k retrieved documents. If such a volume is small, we can expect a high semantic correlation between the query and the documents. Contrarily, a large volume might indicate documents poorly coherent with the query. To define a convex hull around the vectors, we would need as many points as dimensions considered – Notice that our embeddings lie in hyperspace with $d = 768$ dimensions, and such a high number of documents required to construct the convex hull is unlikely to be informative in the QPP domain. A second alternative relies on computing the volume of the hyper-cube containing all the documents. To do this, we consider each dimension h of the learned representation and determine the length of the hyper-cube's edge laying on h as $l_h = |\max(\{D_i(h), \forall i \in [1,k]\} \cup \{v_q(h)\}) - \min\{\{D_i(h), \forall i \in [1,k]\} \cup \{v_q(h)\}\}|$ where k is the ranked list cutoff, $v_q(h)$ and $D_i(h)$ are respectively the values of the h-th dimension for the query and i-th document. Finally, the volume v_q^k of the hyper-cube constructed around the top-k documents for query q is computed as: $v_q^k = \prod_{h=1}^{d} l_h$. Notice that, while no specific bound is present on l_h, it is likely that such values are small, thus it is numerically more stable to compute the log sum of such value. We define the first predictor, dubbed Reciprocal Volume (RV), as:

$$RV_k(q) = -\frac{1}{\sum_{h=1}^{d} \log(l_h)}.$$

Assuming that each dimension represents a latent aspect of the query, having a smaller hyper-cube on a certain dimension suggests that all the retrieved documents are closely related to that query's latent aspect. Vice versa, if the cube is particularly big on that dimension, it is likely that the retrieved documents treat the latent aspect in a very different way from the query.

Discounted Matryoshka. The reference measure used most often in conversational search [9, 10] is normalize Discounted Cumulative Gain (nDCG) [22]. Such a measure is based on the model of a user browsing the ranked list of retrieved documents and accruing a certain amount of utility proportional to the relevance of the document and inversely proportional to its position [6]. Inspired by a similar rationale, we propose a second predictor, dubbed Discounted Matryoshka (DM). The DM predictor is defined as follows:

$$DM_k(q) = \sum_{j=1}^{k} \frac{RV_j(q)}{\log(j+1)}.$$

Ideally, starting from the first document retrieved, we construct the hyper-cube containing the document(s) and the query and determine its volume. Each hyper-cube constructed by adding a new

document contains (or is equal to) the previous one $DM_j(q) \leq DM_{j+1}(q)$ – and therefore they can be seen as Matryoshka dolls. If moving from document to document such volume remains limited – all Matryoshkas are similar and small – we assume that all top retrieved documents are consistent with the query in all its dimensions and therefore we could assume a successful retrieval. Vice versa, if either the hyper-volume is large or we observe a quick change in the hyper-volume, then we can assume that the dense representation does not characterize well that part of the space, and therefore retrieved documents are likely to be not particularly coherent with the query, which suggests a failed retrieval. Notice that, the hyper-volume of each hyper-cube used to compute $DM_k(q)$ is discounted by a discount factor proportional to the number of points in the space used to construct it. Such a discount factor is useful for two reasons. First, in CS, the most common evaluation scenario consists of computing measures at small cutoffs (e.g., nDCG@3). The discount factor allows us to take into consideration this aspect – by discounting less the hyper-cube volumes constructed using top-most documents, we enforce that the prediction remains coherent with the most common use case, where just a few documents are taken into consideration. Secondly, by adding new points (documents) hyper-volumes are inherently bounded to grow: the discount factor allows for this growth to be slower.

5 EXPERIMENTAL EVALUATION

This section briefly introduces the QPP state-of-the-art approaches taken into consideration as baseline, the experimental settings and our experimental findings.

5.1 State-of-the-art Approaches

We report here a brief description of the state-of-the-art QPP approaches used as the baselines.

Clarity [7]. It is one of the first proposed post-retrieval predictors. It relies on calculating $\theta_{\mathcal{D}@k}$, the language model of the first k retrieved documents, and comparing it to θ_C, the language model of the entire corpus, using Kullback–Leibler (KL) Divergence:

$$Clarity(q) = \sum_{w \in V} p(w|\theta_{\mathcal{D}@k}) \frac{p(w|\theta_{\mathcal{D}@k})}{p(w|\theta_C)},$$

where V is the vocabulary and $p(w|\theta)$ is the probability of observing the token w according to θ the language model.

Weighted Information Gain (WIG) [62]. This predictor calculates the difference between the scores of the retrieved documents and the score that the entire corpus would achieve in response to q:

$$WIG(q) = \frac{1}{k\sqrt{|q|}} \sum_{d \in \mathcal{D}@k} \left(s(q,d) - s(q,C)\right).$$

Normalized Query Commitment (NQC) [44]. It computes the prediction score by analyzing the variance of the scores for the top k retrieved documents.

$$NQC(q) = \frac{\sqrt{\frac{1}{k}\sum_{d \in \mathcal{D}@k}(s(q,d) - \hat{\mu}_{\mathcal{D}@k})^2}}{s(q,C)},$$

where $\hat{\mu}_{\mathcal{D}@k} = \frac{1}{k} \cdot \sum_{d \in \mathcal{D}@k} s(q,d)$.

Score Magnitude and Variance (SMV) [46]. Builds on NQC and WIG by considering both the magnitude of the scores (WIG) and their variance (NQC)

$$SMV(q) = \frac{\frac{1}{k}\sum_{d \in \mathcal{D}@k}\left(s(q,d) \cdot \left|\ln \frac{s(q,d)}{\hat{\mu}_{\mathcal{D}@k}}\right|\right)}{s(q,C)}.$$

Utility Estimation Framework (UEF) [43]. This framework can be instantiated using any of the previously mentioned QPP models. It consists in reweighting the prediction scores by the correlation between the ranked list obtained using the query and the one obtained by considering the language model of the top k documents, in a Pseudo-Relevance Feedback (PRF) fashion. Called θ_q the language model constructed using the query, $\pi(\theta, \mathcal{D})$ the retrieval scores for the documents in $\mathcal{D}$ computed using the language model θ, and $\mathcal{M}(q)$ the prediction score for the query q using a given QPP model $\mathcal{M}$, UEF approach is defined as follows.

$$UEF_{\mathcal{M}}(q) = \rho(\pi(\theta_q, \mathcal{D}@k), \pi(\theta_{\mathcal{D}@k}, \mathcal{D}@k)) \cdot \mathcal{M}(q),$$

where ρ is a similarity function between two lists – the most commonly used approach sets ρ to be Pearson's correlation.

5.2 Experimental Setting

Our experiments are based on 2019, 2020, and 2021 TREC Conversational Assistant Track (CAsT)[1] datasets. The CAsT 2019 [10] dataset consists of 20 human-assessed test conversations, while CAsT 2020 [9] and CAsT 2021 include 25 and 26 conversations respectively, with an average of 10 turns per conversation. The CAsT 2019 and 2020 include relevance judgments at passage level, whereas for CAsT 2021 the relevance judgments are provided at the document level. The judgments have a three-point graded scale and refer to passages of the TREC Complex Answer Retrieval (CAR), and MS-MARCO (MAchine Reading COmprehension) collections for CAsT 2019 and 2020, and to documents of MS-MARCO, KILT Wikipedia, and Washington Post 2020 for CAsT 2021. In our experiments we try to predict nDCG@3, the most commonly used measure in CS [9, 10]. In our experiments, we use dense representations of original, automatically rewritten, and manually rewritten queries, where missing keywords or references to previous topics are resolved by human assessors. Original and manually rewritten queries are encoded using the STAR model, while the automatically rewritten ones are obtained by using the ConvDR model. In all the cases, the dense representations of documents and queries is made of 768-dimensional vectors. As a reference line, our ConvDR runs achieve nDCG@3 equal to 0.46 and 0.37 on CAsT 2019 and CAsT 2020 respectively. Conversely, STAR achieves nDCG@3 of 0.38, 0.34, and 0.34 on CAsT 2019, CAsT 2020, and CAsT 2021 respectively. For ConvDR we used publicly available weights[2]. Notice that we do not report ConvDR results for the CAsT 2021 since, at the current time, there are no publicly available weights for this dataset. For all models employing it, the cutoff hyperparameter k has been selected from the set {3, 5, 10, 50, 100, 500}. In particular, all QPP models have been fine-tuned using the traditional two-fold repeated sampling [11, 44, 58, 59] with 30 repetitions to select hyperparameters. For those evaluation procedures based on a correlation measure

[1] Conversational Assistant Track, https://www.treccast.ai/
[2] https://github.com/thunlp/ConvDR

(i.e., full-collection and conversation-wise evaluations) we employ Pearson's correlation, Kendall's correlation and Rank-Biased Overlap (RBO) [51]. Pearson's correlation allows us to take into account the magnitudes of the predictions and performance observed. On the contrary, Kendall's correlation takes into consideration only the ordering of the different queries. Finally, RBO considers only the ordering, it is top-heavy, awarding more systems capable of sorting better systems on the upper part of the ranking.

5.3 Results

In paragraph 5.3.1 we begin our empirical analysis by considering a very simple yet effective baseline to set a common reference point for the subsequent analyses. After that, we evaluate the proposed QPP models according to the three use cases and evaluation methodologies devised in Section 3: collection-wise, conversation-wise and utterance-wise use cases and evaluation protocols.

5.3.1 A baseline predictor. As a first analysis, we are interested in determining a baseline to understand how challenging the predictive task is in the conversational domain. In particular, we adopt the utterance classification proposed by Mele et al. [30] as a QPP to predict the performance of the STAR model when applied straightforwardly to original utterances (without any further rewriting). According to Mele et al. [30], utterances can be classified into two main groups: Self-Explanatory (SE) and non Self-Explanatory (non-SE) utterances. The former consist in utterances that do not require any further context to be answered, since they do not contain anaphoras, ellipses or coreferences. Vice versa, the latter, to be converted into effective queries, requires some form of rewriting, at least by replacing the pronouns with the entity they refer to. We use the manual utterance labeling provided by Mele et al. [30] for CAsT 2019 and CAsT 2020, while we manually annotate utterances for CAsT 2021, since such annotation is not publicly available. Notice that, while we used manually annotated labels, Mele et al. [30] propose a strategy to automatically classify utterances into SE and non-SE that achieves 91% of F-measure, making the task approachable automatically with high effectiveness. To devise a basic predictor from the utterance labels, we apply the following procedure: we assign a prediction score of 1 to each utterance labeled as SE and a prediction score of 0 to non-SE utterances. This procedure follows the simple rationale that, if the utterances contain enough context to be answered, they are likely to be "simpler" and thus more effective. Vice versa - utterances that do not contain all the required information (non-SE utterances), if used without any further expansion, are doomed to fail.

Figure 1 reports a visual depiction of the performance, measured using sARE, of such a trivial predictor to predict the performance of STAR as a retrieval model, without any further query expansion on CAsT 2019, CAsT 2020, and CAsT 2021, thus using the plain original queries. It is interesting to see that, even though the QPP model is extremely simple, it is also highly effective, with mean sARE of 0.150, 0.136, and 0.131 on CAsT 2019, CAsT 2020, and CAsT 2021 respectively (Cfr. Table 3 to see the performance of other predictors in a more realistic scenario). Figure 1 suggests that the approach is more effective on non-SE utterances, with more observations (orange crosses) on the left part of the plots. This is also confirmed numerically considering that, on CAsT 2019, non-SE utterances are predicted with an average sARE of 0.122, against 0.198 for SE. Similarly, on CAsT 2020, the predictor on non-SE utterances achieves an sARE of 0.099, against 0.232 for SE ones. On CAsT 2021, the mean sARE performance is 0.109 against 0.199 for SE and non-SE utterances respectively. This can be explained by the fact that, by assigning a predicted performance of 0 to a non-SE utterance, the predictor is likely to guess correctly. Without any form of additional utterance expansion, non-SE utterances are very likely to fail, making it reasonable to predict 0. Vice versa, the prediction for SE utterances is much more complex since they are likely to achieve a performance in the range [0, 1] – the trivial prediction of 1 is almost always a large upper bound of the performance.

Table 1: Collection-wise performance.

	Pearson			Kendall			RBO		
	CDR-o	CDR	STAR	CDR-o	CDR	STAR	CDR-o	CDR	STAR
CAsT 2019									
Utt. Lbl	0.135	0.135	0.035	0.110	0.110	0.033	**0.535**†	**0.535**†	0.522
Clarity	0.284	0.282	0.296	0.216	0.218	0.224	0.508	0.501	0.505
NQC	0.230	**0.422**†	0.129	0.189	0.271†	0.105	0.507	0.505	0.510
SMV	0.239	0.408	0.157	0.180	0.257	0.124	0.513	0.507	0.516
WIG	0.287	0.283	0.406	0.188	0.181	0.273	0.505	0.517	0.512
UEF$_{Clr}$	0.261	0.259	0.286	0.181	0.180	0.196	0.505	0.506	0.516
UEF$_{NQC}$	0.240	0.379	0.363	0.203	0.259	0.222	0.508	0.497	0.532†
UEF$_{SMV}$	0.254	0.389	0.363	0.208	**0.274**†	0.225	0.514	0.507	**0.539**†
UEF$_{WIG}$	0.306	0.257	0.280	0.244	0.185	0.201	0.519†	0.504	0.508
RV	0.323	0.323	0.410	0.236	0.236	0.239	0.524†	0.524†	0.522†
DM	**0.376**†	0.376	**0.432**†	**0.262**†	0.262	**0.304**†	0.523†	0.523†	0.528†
CAsT 2020									
Utt. Lbl	0.101	0.101	-0.011	0.088	0.088	-0.013	**0.529**†	**0.529**†	**0.524**†
Clarity	0.230	0.230	0.042	0.135	0.136	0.012	0.499	0.497	0.506
NQC	0.397†	**0.478**†	0.236	**0.299**†	**0.350**†	0.193	0.516†	0.526†	0.505
SMV	**0.400**†	0.470	0.246	0.296†	0.342	0.179	0.491	0.523†	0.514†
WIG	0.240	0.444	0.237	0.147	0.323	0.172	0.522†	0.490	0.507†
UEF$_{Clr}$	0.278	0.276	0.228	0.196	0.198	0.179	0.492	0.494	0.495
UEF$_{NQC}$	0.366	0.420	0.304	0.291	0.309	**0.240**†	0.506	0.496	0.499
UEF$_{SMV}$	0.366	0.418	0.307	0.281	0.312	0.238†	0.509	0.501	0.500
UEF$_{WIG}$	0.287	0.342	0.235	0.206	0.253	0.193	0.510	0.502	0.496
RV	0.241	0.241	**0.338**†	0.180	0.180	0.230	0.501	0.501	0.505
DM	0.271	0.271	0.325	0.198	0.198	**0.240**†	0.507	0.507	0.487
CAsT 2021									
Utt. Lbl	–	–	0.154	–	–	0.138	–	–	0.510
Clarity	–	–	0.341	–	–	0.244	–	–	0.518†
NQC	–	–	0.422	–	–	0.327	–	–	0.512
SMV	–	–	0.434	–	–	**0.339**†	–	–	0.523†
WIG	–	–	0.468	–	–	0.329	–	–	0.513
UEF$_{Clr}$	–	–	0.035	–	–	0.015	–	–	0.507
UEF$_{NQC}$	–	–	0.200	–	–	0.089	–	–	0.502
UEF$_{SMV}$	–	–	0.199	–	–	0.097	–	–	0.500
UEF$_{WIG}$	–	–	0.126	–	–	0.050	–	–	0.515†
RV	–	–	0.424	–	–	0.299	–	–	0.518†
DM	–	–	**0.480**†	–	–	0.336†	–	–	**0.531**†

5.3.2 Collection-wise evaluation. Switching to a more in-depth analysis of the proposed approaches, we start by considering the evaluation methodology described in 3.2.1. In particular, each utterance is considered an independent event with respect to other utterances in the collection and the QPP performance measure is the correlation between the utterances' observed IR performance and prediction score. To assess the presence of statistically significant differences between considered baselines and the proposed models we employed ANalysis Of the VAriance (ANOVA) [41] with the model described by equation MD1.

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \tag{MD1}$$

where y_{ij} represent the observed performance for the i-th QPP model measured using any of the previously mentioned correlation measures on the j-th fold, μ is the grand mean, α_i is the effect of

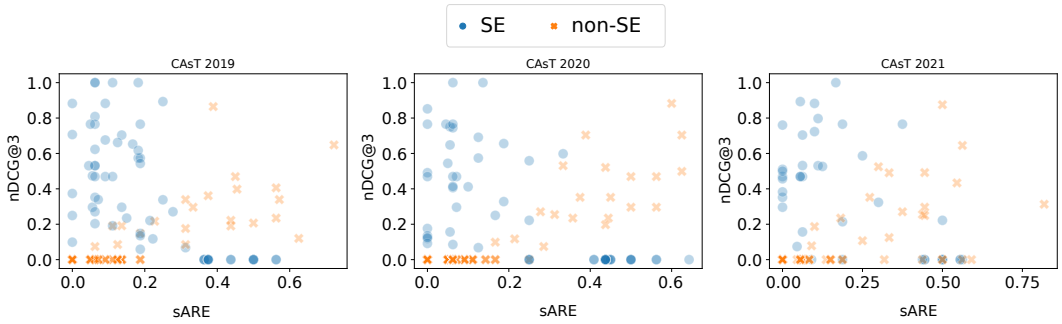

Figure 1: sARE performance for the utterance labeling baseline on a trivial prediction task: predicting the performance of a STAR model that uses straightforwardly original queries without any rewriting/processing.

the i-th model on the performance and ε_{ij} is the unexplained variance (i.e., the ANOVA model error). To carry out post-hoc pairwise comparisons, we employ the Tukey Honestly Significant Difference (HSD) test [49], accounting also for the multiple comparisons problem. The replicates needed to carry out ANOVA are obtained using the classical 2-fold samplings. Table 1 reports our experimental findings. First of all, it is interesting to notice that, when using STAR vectors, for both CAsT 2019 and CAsT 2021, the DM predictor is either the best predictor or not statistically significantly different from the best. This holds for all correlation measures considered. The high RBO for the utterance labeling baseline is due to the fact that, by ranking higher self-explanatory utterances, it is more likely to put in first positions utterances that are in fact "easier" – being top-heavy, RBO awards this behaviour.

On CAsT 2020, on the other hand, the DM predictors beat the baselines only when considering Kendall's and Pearson's correlations, while failing in doing so on the RBO correlation. In general, the RV predictor is always lower compared to DM, with a few exceptions (RBO correlation for CAsT 2019 and Pearson's correlation for CAsT 2020). To predict the performance of ConvDR, we consider two alternatives, the first consists in using the same utterances that ConvDR uses, namely the original ones (indicated with CDR-o), or the manually rewritten utterances (indicated with CDR), to make results more comparable to those observed in STAR. It is important to notice that while traditional predictors are influenced by the usage of either original or rewritten queries, this is not the case for the proposed RV and DM predictors – they rely on the dense representation of the utterance, regardless of its textual content. In terms of retrieval, both ConvDR and ConvDR-o are exactly the same: the difference is the type of utterances used to predict ConvDR performance for the traditional lexical QPP baselines. Notice that, in this sense, the usage of ConvDR and rewritten utterances represent a non-realistic scenario. Let's consider the performance of the predictors of ConvDR with rewritten utterances. We notice that the proposed predictors tend to fail compared to the baselines in the majority of the cases with the exception of CAsT 2019 using RBO as correlation, where the performance is statistically not diverse from the best method (WIG). In particular, the overall best method to predict convDR performance is NQC, which is the best or not statistically worse than the best in 5 evaluation settings out of 6 (the only exception is CAsT 2019 using RBO as correlation measure). If we consider the most conversational and realistic scenario,

Table 2: Conversation-wise performance.

	Pearson			Kendall			RBO		
	CDR-o	CDR	STAR	CDR-o	CDR	STAR	CDR-o	CDR	STAR
CAsT 2019									
Utt. Lbl	0.209†	0.209†	0.050†	0.200†	0.200†	0.080†	0.595†	0.595†	0.595†
Clarity	0.226†	0.226†	0.169†	0.175†	0.175†	0.081†	0.604†	0.605†	0.598†
NQC	0.293†	**0.455**†	0.158†	0.245†	0.338†	0.159†	0.608†	0.610†	0.583†
SMV	0.284†	0.437†	0.188†	0.247†	**0.362**†	0.174†	0.619†	0.593†	0.589†
WIG	0.248†	0.308†	**0.381**†	0.236†	0.218†	0.241†	0.598†	0.583†	0.591†
UEF$_{Clt}$	0.253†	0.253†	0.135†	0.135†	0.135†	0.085†	0.583†	0.570†	0.618†
UEF$_{NQC}$	0.363†	0.433†	0.223†	0.292†	0.319†	0.144†	0.618†	0.594†	0.619†
UEF$_{SMV}$	**0.372**†	0.445†	0.222†	**0.300**†	0.324†	0.142†	**0.626**†	0.599†	0.610†
UEF$_{WIG}$	0.321†	0.282†	0.165†	0.225†	0.160†	0.114†	0.574†	0.539†	0.601†
RV	0.316†	0.316†	0.347†	0.242†	0.242†	0.271†	0.621†	**0.621**†	**0.640**†
DM	0.354†	0.354†	0.365†	0.224†	0.224†	**0.276**†	0.594†	0.594†	0.614†
CAsT 2020									
Utt. Lbl	0.111	0.111	0.048†	0.074	0.074	0.044†	0.600†	0.600†	0.596†
Clarity	0.143†	0.140	0.018†	0.121†	0.114	-0.021	**0.622**†	0.621†	**0.634**†
NQC	0.362†	0.408†	0.282†	0.277†	0.329†	0.220†	0.592†	0.614†	0.620†
SMV	0.354†	0.402†	0.276†	0.289†	**0.337**†	0.220†	0.611†	0.624†	0.621†
WIG	0.145†	0.382†	0.234†	0.113†	0.277†	0.147†	0.597†	**0.632**†	0.617†
UEF$_{Clt}$	0.267†	0.267†	0.166†	0.225†	0.225†	0.118†	0.589†	0.589†	0.588†
UEF$_{NQC}$	0.372†	0.405†	0.318†	0.291†	0.329†	0.229†	0.602†	0.600†	0.580†
UEF$_{SMV}$	**0.380**†	**0.409**†	0.310†	**0.300**†	0.321†	0.215†	0.603†	0.604†	0.574†
UEF$_{WIG}$	0.240†	0.325†	0.178†	0.169†	0.248†	0.102†	0.602†	0.575†	0.564†
RV	0.269†	0.269†	**0.335**†	0.238†	0.238†	**0.255**†	0.586†	0.586†	0.609†
DM	0.311†	0.311†	0.333†	0.258†	0.258†	0.240†	0.588†	0.588†	0.609†
CAsT 2021									
Utt. Lbl	–	–	0.194	–	–	0.079	–	–	0.498
Clarity	–	–	0.311†	–	–	0.244†	–	–	0.586†
NQC	–	–	0.537†	–	–	0.426†	–	–	0.648†
SMV	–	–	**0.544**†	–	–	**0.431**†	–	–	**0.654**†
WIG	–	–	0.411†	–	–	0.335†	–	–	0.639†
UEF$_{Clt}$	–	–	0.053	–	–	0.034	–	–	0.575†
UEF$_{NQC}$	–	–	0.204	–	–	0.115	–	–	0.564†
UEF$_{SMV}$	–	–	0.206	–	–	0.123	–	–	0.568†
UEF$_{WIG}$	–	–	0.095	–	–	0.086	–	–	0.558†
RV	–	–	0.364†	–	–	0.350†	–	–	0.597†
DM	–	–	0.392†	–	–	0.378†	–	–	0.635†

ConvDR with original utterances and predictors based on original utterances, We notice that DM is always the best method (with the only exception of the RBO measure, where it ranks third, behind RV and the trivial utterance labeling predictor wins, but statistically they are equivalent). It is interesting to notice that the Utterance labeling trivial predictor performs always better on ConvDR than on STAR (See also Table 2). This is because ConvDR, by using original utterances, is still influenced by the class of the utterances. This suggests that, when a system relies on original utterances to carry out retrieval, the utterance labeling predictor can be used in combination with other predictors as an orthogonal signal.

5.3.3 Conversation-wise evaluation. We apply now the evaluation methodology described in 3.2.2. While it would be still possible to

Table 3: Utterance-wise evaluation using sMARE.

	CAsT 2019			CAsT 2020			CAsT 2021
	CDR-o	CDR	STAR	CDR-o	CDR	STAR	STAR
Utt. Label	0.259†	0.259†	0.274†	0.272†	0.272†	0.282†	0.319
Clarity	0.276†	0.276†	0.306†	0.290†	0.292	0.324	0.258†
NQC	0.253†	**0.232**†	0.271†	0.243†	**0.235**†	0.263†	0.219†
SMV	0.252†	0.233†	0.265†	**0.242**†	0.238†	0.265†	**0.216**†
WIG	0.275†	0.274†	0.264†	0.294†	0.246†	0.271†	0.228†
UEFClarity	**0.248**†	0.302	0.302†	0.259†	0.259†	0.293†	0.314
UEFNQC	0.252†	0.242†	0.295†	0.249†	0.241†	0.264†	0.296
UEFSMV	0.249†	0.238†	0.299†	0.244†	0.241†	0.267†	0.294
UEFWIG	0.267†	0.295	0.297†	0.271†	0.253†	0.293†	0.304
RV	**0.248**†	0.248†	0.257†	0.259†	0.259†	0.256†	0.238†
DM	0.256†	0.256†	**0.248**†	0.258†	0.258†	**0.256**†	0.233†

adopt MD1 in this context, to improve the expressiveness of the evaluation model, we extend it into MD2 by also considering the effect that different conversations might have on the performance:

$$y_{ik} = \mu + \alpha_i + \chi_k + \varepsilon_{ik}, \tag{MD2}$$

where, with respect to MD1, y_{ik} is the correlation measured on the k-th conversation, while χ_k represents the effect of the k-th conversation. Table 2 reports the results for this second analysis. It is possible to observe that the results are less stable, with no clear winner in all scenarios. Furthermore, all the results appear to be statistically not significantly different: this highlights strong variance of the correlation observed for different conversations. This behaviour is somehow expected: having only 20 to 26 conversations, it is reasonable that our tests are underpowered – similar phenomena are observable also in the traditional IR scenario when a small number of topics is considered. This urges to expand the collections with more conversations. In terms of numerical results, we observe that, for CAsT 2019 and using STAR as the ranking function, the DM predictor ranks second if the Pearson correlation is used, and ranks first if Kendall's correlation is used. Vice Versa, if we consider CAsT 2020 and STAR, the best-performing method when Kendall's and Pearson's correlations are used is RV. For CAsT 2021, the method that systematically performs the best is the SMV baseline. Notice that, the UEF counterpart of SMV is also one of the best methods for both CAsT 2019 and CAsT 2020 if we consider the predictions for ConvDR, both using original and rewritten queries.

5.3.4 Utterance-wise evaluation. Finally, we report the results according to the evaluation methodology described in 3.2.3. Also in this case we could consider to use MD1, but we switch to a more expressive model that includes the individual effect of each query.

$$y_{il} = \mu + \alpha_i + \upsilon_l + \varepsilon_{il}, \tag{MD3}$$

where, with respect to MD1, y_{il} is the sARE measured on the l-th utterance, while υ_l represents the effect of the l-th utterance. Table 3 reports the results of the utterance-wise analysis. Notice that, given that we report the mean sARE, which is an error, the lower the figure in the table, the better the performance. For CAsT 2019 and CAsT 2020 we observe similar patterns to those highlighted by Table 1, with the proposed approaches overcoming all the other baselines for the STAR model. For the ConvDR model, on the other hand, the behaviour is less stable, with RV being the best model in CAsT 2019 if the original utterances are considered. Akin to what was observed in Table 2, for the ConvDR system, the best models in the utterance-wise evaluation scenario seem to be SMV and NQC.

Following what was observed in Table 2, we notice that most of the comparisons are not statistically significant, stressing again the need for more conversations in a collection. Even though theoretically more suited to the conversational task, both conversation- and utterance-wise evaluation protocols show that there are no evident statistical differences between the baselines and there is no clear winner. This should raise concern within the community that, without a proper evaluation framework, we are in fact comparing with weak baselines. The non-negligible risk is that new methods are deemed significantly better than the baselines due to the wrong evaluation methodology used, as already observed in neighbouring areas, including IR [26, 55] and Recommender Systems [8].

6 CONCLUSION AND FUTURE WORK

In this study, we explore the potential of a geometric framework for performance prediction in the CS domain. The lack of a clear definition for QPP in conversational settings is addressed, and three relevant use cases for QPP in CS are identified: as a post-hoc evaluation technique, to diagnose anomalies within the conversation, and to predict the performance of the next utterance. We define an evaluation procedure for each use case, including a collection-wise evaluation procedure that mimics current QPP evaluation, as well as conversation- and utterance-wise evaluation procedures. We propose two geometric post-retrieval predictors, which measure the proximity of retrieved documents to the query encapsulating them within a hypercube. The predictors are applied to two conversational models, ConvDR and STAR, on three established conversational collections. The results demonstrate that our proposed methodology outperforms QPP baselines on CAsT 2019 and CAsT 2021 at the collection- and the utterance-wise level. However, the low statistical power of conversation- and utterance-wise evaluations highlight the need for larger conversational collections and revisiting the evaluation procedure used to devise the baselines. In conclusion, the significance of QPP in the CS domain is emphasized, and our proposed models show promising results in improving QPP for conversational search. In future research, we plan to investigate how to incorporate in the predictors signals from previous utterances and their linguistic content.

ACKNOWLEDGMENTS

This work is supported, in part, by the spoke "FutureHPC & BigData" of the ICSC – Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, the Spoke "Human-centered AI" of the M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research", funded by European Union – NextGenerationEU, the FoReLab project (Departments of Excellence), and by the University of Padova Strategic Research Infrastructure Grant 2017: "CAPRI: Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione", European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu), the scheme "World Leading Data and Computing Technologies 2022", Grant Agreement n. 101093026, "EFRA: Extreme Food Risk Analytics".

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. 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, 475–484. https://doi.org/10.1145/3331184.3331265

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

[6] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft (Eds.). ACM, 903–912. https://doi.org/10.1145/2009916.2010037

[7] 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

[8] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 101–109. https://doi.org/10.1145/3298689.3347058

[9] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2020: The Conversational Assistance Track Overview. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020 (NIST Special Publication, Vol. 1266)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST), 1–10. https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.C.pdf

[10] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview. *CoRR* abs/2003.13624 (2020), 1–10. arXiv:2003.13624 https://arxiv.org/abs/2003.13624

[11] 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* 41, 2 (jun 2022), 1–31.

[12] 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), 1–15. https://www.mdpi.com/2076-3417/11/19/9075

[13] Guglielmo Faggioli, Marco Ferrante, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. 2021. Hierarchical Dependence-aware Evaluation Measures for Conversational Search. 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, 1935–1939. https://doi.org/10.1145/3404835.3463090

[14] 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

[15] 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

[16] Ophir Frieder, Ida Mele, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Nicola Tonellotto. 2022. Caching Historical Embeddings in Conversational Search. *ACM Trans. Web* (dec 2022), 1–18. https://doi.org/10.1145/3578519

[17] Maram Hasanain and Tamer Elsayed. 2017. Query performance prediction for microblog search. *Inf. Process. Manag.* 53, 6 (2017), 1320–1341. https://doi.org/10.1016/j.ipm.2017.08.002

[18] 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

[19] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2020. Guided Transformer: Leveraging Multiple External Sources for Representation Learning in Conversational Search. 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, 1131–1140. https://doi.org/10.1145/3397271.3401061

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

[21] 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

[22] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446. https://doi.org/10.1145/582415.582418

[23] J. Johnson, M. Douze, and H. Jegou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* 7, 03 (2021), 535–547.

[24] Mohammad Kachuee, Hao Yuan, Young-Bum Kim, and Sungjin Lee. 2021. Self-Supervised Contrastive Learning for Efficient User Satisfaction Prediction in Conversational Agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 4053–4064. https://doi.org/10.18653/v1/2021.naacl-main.319

[25] 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 *Proc. EMNLP*. 6769–6781.

[26] Sadegh Kharazmi, Falk Scholer, David Vallet, and Mark Sanderson. 2016. Examining Additivity and Weak Baselines. *ACM Trans. Inf. Syst.* 34, 4 (2016), 23:1–23:18. https://doi.org/10.1145/2882782

[27] Eyal Krikon, David Carmel, and Oren Kurland. 2012. Predicting the performance of passage retrieval for question answering. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, Xue-wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki (Eds.). ACM, 2451–2454. https://doi.org/10.1145/2396761.2398664

[28] Yongqi Li, Wenjie Li, and Liqiang Nie. 2022. Dynamic Graph Reasoning for Conversational Open-Domain Question Answering. *ACM Trans. Inf. Syst.* 40, 4, Article 82 (jan 2022), 24 pages. https://doi.org/10.1145/3498557

[29] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021. Multi-Stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting. *ACM Trans. Inf. Syst.* 39, 4 (2021), 48:1–48:29. https://doi.org/10.1145/3446426

[30] Ida Mele, Cristina Ioana Muntean, Franco Maria Nardini, R. Perego, Nicola Tonellotto, and Ophir Frieder. 2021. Adaptive utterance rewriting for conversational search. *Inf. Process. Manag.* 58 (2021), 102682.

[31] 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*. 7–10.

[32] Dipasree Pal and Debasis Ganguly. 2021. Effective Query Formulation in Conversation Contextualization: A Query Specificity-based Approach. In *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, Faegheh Hasibi, Yi Fang, and Akiko Aizawa (Eds.). ACM, 177–183. https://doi.org/10.1145/3471158.3472237

[33] Dookun Park, Hao Yuan, Dongmin Kim, Yinglei Zhang, Spyros Matsoukas, Young-Bum Kim, Ruhi Sarikaya, Edward Guo, Yuan Ling, Kevin Quinn, Pham Hung, Benjamin Yao, and Sungjin Lee. 2020. Large-scale Hybrid Approach for Predicting User Satisfaction with Conversational Agents. *CoRR* abs/2006.07113 (2020). arXiv:2006.07113 https://arxiv.org/abs/2006.07113

[34] Joaquín Pérez-Iglesias and Lourdes Araujo. 2010. Standard Deviation as a Query Hardness Estimator. In *String Processing and Information Retrieval - 17th International Symposium, SPIRE 2010, Los Cabos, Mexico, October 11-13, 2010. Proceedings (Lecture Notes in Computer Science, Vol. 6393)*, Edgar Chávez and Stefano Lonardi (Eds.). Springer, 207–212. https://doi.org/10.1007/978-3-642-16321-0_21

[35] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proc. CHIIR*. ACM, New York, NY, USA, 117–126.

[36] 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

[37] Haggai Roitman. 2018. Query Performance Prediction using Passage Information. 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, 893–896. https://doi.org/10.1145/3209978.3210070

[38] Haggai Roitman. 2020. ICTIR Tutorial: Modern Query Performance Prediction: Theory and Practice. In *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*, Krisztian Balog, Vinay Setty, Christina Lioma, Yiqun Liu, Min Zhang, and Klaus Berberich (Eds.). ACM, 195–196. https://doi.org/10.1145/3409256.3409813

[39] Haggai Roitman, Shai Erera, and Guy Feigenblat. 2019. A Study of Query Performance Prediction for Answer Quality Determination. 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, 43–46. https://doi.org/10.1145/3341981.3344219

[40] 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

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

[42] 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

[43] 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

[44] 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

[45] Chuanqi Tan, Furu Wei, Qingyu Zhou, Nan Yang, Weifeng Lv, and Ming Zhou. 2018. I Know There Is No Answer: Modeling Answer Validation for Machine Reading Comprehension. In *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 11108)*, Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan (Eds.). Springer, 85–97. https://doi.org/10.1007/978-3-319-99495-6_8

[46] 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

[47] 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

[48] Nicola Tonellotto and Craig Macdonald. 2020. Using an Inverted Index Synopsis for Query Latency and Performance Prediction. *ACM Trans. Inf. Syst.* 38, 3, Article 29 (may 2020), 33 pages. https://doi.org/10.1145/3389795

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

[50] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. *Query Resolution for Conversational Search with Limited Supervision*. Association for Computing Machinery, New York, NY, USA, 921–930. https://doi.org/10.1145/3397271.3401130

[51] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* 28, 4 (2010), 20:1–20:38. https://doi.org/10.1145/1852102.1852106

[52] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *CoRR* abs/2007.00808 (2020). arXiv:2007.00808 https://arxiv.org/abs/2007.00808

[53] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *Proc. ICLR*.

[54] Jheng-Hong Yang, Sheng-Chieh Lin, Chuan-Ju Wang, Jimmy J. Lin, and Ming-Feng Tsai. 2019. Query and Answer Expansion from Conversation History. In *TREC*.

[55] Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. 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, 1129–1132. https://doi.org/10.1145/3331184.3331340

[56] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-Shot Generative Conversational Query Rewriting. In *Proc. SIGIR*. ACM, New York, NY, USA, 1933–1936.

[57] Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-Shot Conversational Dense Retrieval. 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, 829–838. https://doi.org/10.1145/3404835.3462856

[58] 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

[59] 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

[60] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. 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, 1503–1512. https://doi.org/10.1145/3404835.3462880

[61] 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

[62] 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