A Spatial Approach to Predict Performance of Conversational Search Systems*

Discussion Paper

Guglielmo Faggioli[1], Nicola Ferro[1], Cristina Muntean[2], Raffaele Perego[2] and Nicola Tonellotto[3]

[1]University of Padova, Padova, Italy
[2]ISTI-CNR, Pisa, Italy
[3]University of Pisa, Pisa, Italy

Abstract

Recent advancements in Information Retrieval and Natural Language Processing have led to significant developments in the way users interact with search engines, with traditional one-shot textual queries being replaced by multi-turn conversations. As a highly interactive search scenario, Conversational Search (CS) can significantly benefit from Query Performance Prediction (QPP) techniques. However, the application of QPP in the CS domain is a relatively new field and requires proper framing. This study proposes a set of spatial-based QPP models, designed to work effectively in the conversational search domain, where dense neural retrieval models are the most common approach and query cutoffs are small. The proposed QPP approaches are shown to improve the predictive performance over the state-of-the-art in different scenarios and collections, highlighting the utility of QPP in the CS domain.

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 the advent of Large Language Models (LLMs), it is becoming increasingly popular and ubiquitously adopted. CS can benefit from the 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 [2]. It has been successfully employed in many tasks, such as query suggestion [3], adaptive model selection [3, 4, 5, 6], and pathological queries discovering [2]. 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 pointed out by Hashemi et al. [7] about Question Answering (QA), we observe that CS involves *i)* highly

IIR2023: 13th Italian Information Retrieval Workshop, 8th - 9th June 2023, Pisa, Italy

*This is an extended abstract of [1].

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 Neural Information Retrieval (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, making the evaluation of CS systems intrinsically different from classical ad-hoc IR [8]. 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 [9] or to determine if the answer provided to the user contains the relevant information [10], 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.

2. Hyper-Volume Based Predictors

In this work, we focus on dense representations of the documents and the queries. In particular, we consider two well-known dense representation approaches STAR [11] and ConvDR [12]. 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 call the top-k retrieved documents $\mathcal{D}@k$.

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 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. In particular, we compute 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 ν_q^k of the hyper-cube constructed around the top-k documents for query q is computed as: $\nu_q^k = \prod_{h=1}^{d} l_h$. 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, 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.

The reference measure used most often in conversational search [13, 14] is normalize Discounted Cumulative Gain (nDCG) [15]. nDCG is based on the model of a user browsing the ranked list of retrieved documents and accruing utility proportional to the relevance of the document and inversely proportional to its position [16]. Inspired by it, we propose a second predictor, dubbed Discounted Matryoshka (DM), defined as follows:

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

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 $DM_j(q) \leq DM_{j+1}(q)$ – 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. 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.

3. Experimental Evaluation

Our experiments are based on 2019, 2020, and 2021 TREC Conversational Assistant Track (CAsT)[1] datasets[2]. In our experiments, we try to predict nDCG@3, the most commonly used measure in CS [13, 14]. 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. For ConvDR we used publicly available weights[3]. For all models employing it, the cutoff hyperparameter k has been selected from the set {3, 5, 10, 50, 100, 500}. QPP models have been fine-tuned using two-fold repeated sampling [17, 18, 19, 20] with 30 repetitions.

Table 1 reports our experimental findings. First of all, it is interesting to notice that, when using STAR vectors 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

[1]Conversational Assistant Track, https://www.treccast.ai/
[2]for space reasons, we report results only on CAsT 2019. Full results are available in [1]
[3]https://github.com/thunlp/ConvDR

Table 1
Prediction Performance on CAsT 2019. Results for other collections available in [1]

	Pearson			Kendall			RBO		
	CDR-o	CDR	STAR	CDR-o	CDR	STAR	CDR-o	CDR	STAR
CAsT 2019									
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†

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. 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, since they would not be available to a real CS agent. Considering 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 Rank-Biased Overlap (RBO) as correlation, where the performance is statistically not diverse from the best method (WIG). If we consider the most conversational and realistic scenario, 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 second, behind RV but statistically they are equivalent.

4. Conclusion and Future Work

In this study, we explore the potential of a geometric framework for performance prediction in the CS domain. We propose two geometric post-retrieval coherency predictors, which measure the proximity of retrieved documents to the query by encapsulating them within a hypercube. The predictors are applied to two conversational dense retrieval models, ConvDR and STAR, on three established conversational collections, using the evaluation procedures defined. The results demonstrate that our proposed methodology outperforms QPP baselines on CAsT 2019 and CAsT 2021. 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.

References

[1] G. Faggioli, N. Ferro, C. Muntean, R. Perego, N. Tonellotto, A Geometric Framework for Query Performance Prediction in Conversational Search, in: Proceedings of 46th international ACM SIGIR conference research development in information retrieval, 2023.

[2] D. Carmel, E. Yom-Tov, Estimating the Query Difficulty for Information Retrieval, Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers, 2010. doi:10.2200/S00235ED1V01Y201004ICR015.

[3] P. Thomas, F. Scholer, P. Bailey, A. Moffat, Tasks, queries, and rankers in pre-retrieval performance prediction, in: B. Koopman, G. Zuccon, M. J. Carman (Eds.), Proceedings of the 22nd Australasian Document Computing Symposium, ADCS 2017, Brisbane, QLD, Australia, December 7-8, 2017, ACM, 2017, pp. 11:1–11:4. doi:10.1145/3166072.3166079.

[4] H. Scells, L. Azzopardi, G. Zuccon, B. Koopman, Query variation performance prediction for systematic reviews, in: K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, E. Yilmaz (Eds.), The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, ACM, 2018, pp. 1089–1092. doi:10.1145/3209978.3210078.

[5] H. Roitman, Enhanced Performance Prediction of Fusion-based Retrieval, in: D. Song, T. Liu, L. Sun, P. Bruza, M. Melucci, F. Sebastiani, G. H. Yang (Eds.), Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2018, Tianjin, China, September 14-17, 2018, ACM, 2018, pp. 195–198. URL: https://doi.org/10.1145/3234944.3234950. doi:10.1145/3234944.3234950.

[6] N. Tonellotto, C. Macdonald, Using an inverted index synopsis for query latency and performance prediction, ACM Trans. Inf. Syst. 38 (2020). doi:10.1145/3389795.

[7] H. Hashemi, H. Zamani, W. B. Croft, Performance prediction for non-factoid question answering, in: Y. Fang, Y. Zhang, J. Allan, K. Balog, B. Carterette, J. Guo (Eds.), Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2019, Santa Clara, CA, USA, October 2-5, 2019, ACM, 2019, pp. 55–58. doi:10.1145/3341981.3344249.

[8] G. Faggioli, M. Ferrante, N. Ferro, R. Perego, N. Tonellotto, Hierarchical dependence-aware evaluation measures for conversational search, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 1935–1939. doi:10.1145/3404835.3463090.

[9] M. Aliannejadi, H. Zamani, F. Crestani, W. B. Croft, Asking Clarifying Questions in Open-Domain Information-Seeking Conversations, in: B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, F. Scholer (Eds.), Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, ACM, 2019, pp. 475–484. URL: https://doi.org/10.1145/3331184.3331265. doi:10.1145/3331184.3331265.

[10] H. Roitman, S. Erera, G. Feigenblat, A Study of Query Performance Prediction for Answer Quality Determination, in: Y. Fang, Y. Zhang, J. Allan, K. Balog, B. Carterette, J. Guo (Eds.), Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2019, Santa Clara, CA, USA, October 2-5, 2019, ACM, 2019, pp. 43–46.

URL: https://doi.org/10.1145/3341981.3344219. doi:10.1145/3341981.3344219.

[11] J. Zhan, J. Mao, Y. Liu, J. Guo, M. Zhang, S. Ma, Optimizing dense retrieval model training with hard negatives, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 1503–1512. URL: https://doi.org/10.1145/3404835.3462880. doi:10.1145/3404835.3462880.

[12] S. Yu, Z. Liu, C. Xiong, T. Feng, Z. Liu, Few-shot conversational dense retrieval, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 829–838. URL: https://doi.org/10.1145/3404835.3462856. doi:10.1145/3404835.3462856.

[13] J. Dalton, C. Xiong, J. Callan, TREC cast 2019: The conversational assistance track overview, CoRR abs/2003.13624 (2020) 1–10. URL: https://arxiv.org/abs/2003.13624. arXiv:2003.13624.

[14] J. Dalton, C. Xiong, J. Callan, Cast 2020: The conversational assistance track overview, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020, volume 1266 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2020, pp. 1–10. URL: https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.C.pdf.

[15] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: http://doi.acm.org/10.1145/582415.582418. doi:10.1145/582415.582418.

[16] B. Carterette, System effectiveness, user models, and user utility: a conceptual framework for investigation, in: W. Ma, J. Nie, R. Baeza-Yates, T. Chua, W. B. Croft (Eds.), Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011, ACM, 2011, pp. 903–912. URL: https://doi.org/10.1145/2009916.2010037. doi:10.1145/2009916.2010037.

[17] A. Shtok, O. Kurland, D. Carmel, F. Raiber, G. Markovits, Predicting query performance by query-drift estimation, ACM Trans. Inf. Syst. 30 (2012) 11:1–11:35. URL: https://doi.org/10.1145/2180868.2180873. doi:10.1145/2180868.2180873.

[18] S. Datta, D. Ganguly, M. Mitra, D. Greene, A Relative Information Gain-Based Query Performance Prediction Framework with Generated Query Variants, ACM Transactions on Information Systems 41 (2022) 1–31.

[19] O. Zendel, A. Shtok, F. Raiber, O. Kurland, J. S. Culpepper, Information needs, queries, and query performance prediction, in: B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, F. Scholer (Eds.), Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, ACM, 2019, pp. 395–404. URL: https://doi.org/10.1145/3331184.3331253. doi:10.1145/3331184.3331253.

[20] H. Zamani, W. B. Croft, J. S. Culpepper, Neural Query Performance Prediction using Weak Supervision from Multiple Signals, in: K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, E. Yilmaz (Eds.), The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, ACM, 2018, pp. 105–114. doi:10.1145/3209978.3210041.