Performance Prediction for Conversational Search Using Perplexities of Query Rewrites

Chuan Meng[1], Mohammad Aliannejadi[1] and Maarten de Rijke[1]

[1]*University of Amsterdam, The Netherlands*

Abstract

We consider query performance prediction (QPP) task for conversational search (CS), i.e., to estimate the retrieval quality for queries in multi-turn conversations. We reuse QPP methods from ad-hoc search for CS by feeding them self-contained query rewrites generated by T5. Our experiments on three CS datasets show that (i) lower query rewriting quality may lead to worse QPP performance, and (ii) incorporating query rewriting quality (as measured by perplexity) improves the effectiveness of QPP methods for CS if the query rewriting quality is limited. Our implementation is publicly available at https://github.com/ChuanMeng/QPP4CS.

Keywords

Query performance prediction, conversational search, perplexity

1. Introduction

We consider the task of *query performance prediction* (QPP) [1, 2] for conversational search (CS) [3], i.e., estimating the retrieval quality for a query in a multi-turn conversation. Little research has been done into QPP for CS. A unique aspect of CS is that each conversational query may contain omissions or coreferences, making it hard for ad-hoc search systems or QPP methods to capture the underlying information need. A popular two-stage CS pipeline [3] can effectively solve this issue by (i) rewriting a conversational query into a self-contained query, and (ii) reusing ad-hoc search systems fed with the query rewrite.

Inspired by the two-stage pipeline, we model QPP for CS by feeding query rewrites to QPP methods designed for ad-hoc search. However, our experiments on CS datasets show that low-quality query rewrites reduce the effectiveness of QPP methods. Based on the fact that lower query rewriting quality tends to result in lower retrieval quality, we argue that query rewriting quality provides evidence for estimating retrieval quality. To incorporate query rewriting quality into QPP methods, we propose a *perplexity-based pre-retrieval QPP framework* (PPL-QPP) for CS. PPL-QPP first evaluates the quality of a query rewrite by its perplexity measured by a pre-trained language model, and then combines the perplexity with a state-of-the-art pre-retrieval QPP

QPP++ 2023: Query Performance Prediction and Its Evaluation in New Tasks, co-located with The 45th European Conference on Information Retrieval (ECIR) April 2, 2023, Dublin, Ireland

c.meng@uva.nl (C. Meng); m.aliannejadi@uva.nl (M. Aliannejadi); m.derijke@uva.nl (M. de Rijke)
https://chuanmeng.github.io/ (C. Meng); https://aliannejadi.com/ (M. Aliannejadi); https://staff.fnwi.uva.nl/m.derijke/ (M. de Rijke)
0000-0002-1434-7596 (C. Meng); 0000-0002-9447-4172 (M. Aliannejadi); 0000-0002-1086-0202 (M. de Rijke)

Table 1
Performance of QPP methods on three CS datasets, in terms of Pearson's ρ, Kendall's τ, and Spearman's ρ correlation coefficients. IDF, PMI, SCQ, and VAR are defined for a single query term; aggregation functions over terms are needed; we report the performance of each method using the optimal aggregation function on each dataset; the aggregation functions used by each method on CAsT-19, CAsT-20, and OR-QuAC are listed sequentially in the brackets. All values are statistically significant (t-test, $p < 0.05$) except the ones in *italics*. The best value in each column is marked in bold.

Methods	CAsT-19			CAsT-20			OR-QuAC		
	P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ
QS	*-0.054*	*-0.011*	*-0.017*	*0.125*	*0.086*	*0.118*	-0.040	-0.038	-0.049
SCS	0.191	0.134	0.191	0.173	0.102	0.140	0.116	0.109	0.141
avgICTF	0.266	0.180	0.257	0.142	0.107	0.144	0.206	0.178	0.229
IDF (avg, avg, sum)	0.271	0.187	0.267	0.149	0.114	0.152	0.259	0.212	0.273
PMI (max, avg, max)	0.320	0.208	0.293	0.136	0.113	0.155	0.180	0.176	0.227
SCQ (avg, avg, max)	0.174	0.127	0.178	0.224	0.167	0.226	0.212	0.159	0.204
VAR (sum, avg, sum)	0.321	0.221	0.310	0.210	0.162	0.221	**0.308**	**0.251**	**0.324**
PPL-QPP	**0.324**	**0.225**	**0.315**	**0.231**	**0.191**	**0.256**	**0.308**	**0.251**	**0.324**

method [2]. Experiments show that PPL-QPP improves the effectiveness of QPP methods in the context of CS in cases when the query rewriting quality is limited.

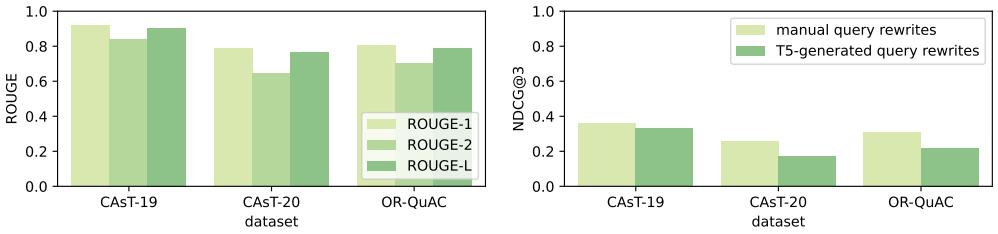
2. Experiments

Experimental setup. We use seven widely used pre-retrieval QPP methods [2] on three CS datasets: CAsT-19 [4], CAsT-20 [4], and OR-QuAC [5]. The retriever to be evaluated by the QPP methods is *T5-based query rewriter*[1]*+BM25*, a widely-used CS method [3]. The T5-generated query rewrites used by BM25 are fed into all QPP methods. We evaluate QPP methods by calculating the correlation between the NDCG@3 scores of the queries in the test set and the estimated retrieval quality. Note that NDCG@3 is the primary metric in CAsT [4, 6].

Performance of QPP methods for CS. Experimental results are presented in Table 1. Our leading observation is that the overall performance of QPP methods on CAst-19 and OR-QuAC is better than on CAsT-20. The difference in results seems to be due to the difference in query rewriting quality on the three datasets. We measure query rewriting quality using the similarity between manual and T5-generated query rewrites in terms of ROUGE, and the BM25 retrieval quality gap between using manual and T5-generated query rewrites. Fig. 1a shows that the ROUGE scores on CAsT-20 are lower than those on CAsT-19 and OR-QuAC; Fig. 1b shows that the gap is larger on CAsT-20 than the gap on CAsT-19. We conclude that the quality of T5-generated query rewrites is lower on CAsT-20 than on the other datasets and that lower query rewriting quality may lead to worse QPP effectiveness.

Incorporating query rewriting quality into QPP for CS. Based on our observation that lower query rewriting quality tends to result in lower retrieval quality, we argue that query rewriting quality can provide evidence for estimating retrieval quality. We propose PPL-QPP,

[1]https://huggingface.co/castorini/t5-base-canard



(a) (b)

Figure 1: The similarity between manual and T5-generated query rewrites in terms of ROUGE (a) and the retrieval quality of BM25 for manual/T5-generated query rewrites in terms of NDCG@3 (b).

which incorporates query rewriting quality into QPP methods. Since we cannot obtain manual query rewrites during estimation, we regard the perplexity of generated query rewrites as a measure of quality. PPL-QPP first uses GPT-2 XL[2] to measure the perplexity of a T5-generated query rewrite and combines the perplexity with a pre-retrieval QPP method through linear interpolation: $\alpha \cdot \frac{1}{PPL} + (1-\alpha) \cdot QPP$. Here, α is a trade-off parameter; the perplexity and QPP values are first normalized prior to fusion. For the QPP method to be combined, we use the state-of-the-art VAR (sum) on CAsT-19 and OR-QuAC, and SCQ (avg) on CAsT-20. The performance of PPL-QPP is presented in Table 1. The results show that PPL-QPP improves the effectiveness of QPP methods in the context of CS on CAsT-19 and, in particular, on CAsT-20, where the query rewriting quality is limited. Interestingly, and different from CAsT-19/20, PPL-QPP does not bring improvements on the OR-QuAC dataset; we plan to further investigate this in our future work.

3. Conclusion

In this paper, we have targeted QPP for CS. We have reused QPP methods for ad-hoc search in the context of CS by feeding them self-contained query rewrites generated by T5. Our experiments on three CS datasets show that (i) lower query rewriting quality may lead to worse QPP performance, and (ii) incorporating query rewriting quality into QPP methods improves their effectiveness in the context of CS when query rewriting quality is limited.

Acknowledgement. We want to thank our reviewers for their feedback. This research was partially supported by the China Scholarship Council (CSC).

[2]https://huggingface.co/gpt2-xl

References

[1] D. Ganguly, S. Datta, M. Mitra, D. Greene, An analysis of variations in the effectiveness of query performance prediction, in: ECIR, Springer, 2022, pp. 215–229.

[2] D. Carmel, E. Yom-Tov, Estimating the query difficulty for information retrieval, Morgan & Claypool Publishers, 2010.

[3] S.-C. Lin, J.-H. Yang, R. Nogueira, M.-F. Tsai, C.-J. Wang, J. Lin, Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting, TOIS 39 (2021) 1–29.

[4] J. Dalton, C. Xiong, J. Callan, CAsT 2020: The conversational assistance track overview, in: Text Retrieval Conference, 2020.

[5] C. Qu, L. Yang, C. Chen, M. Qiu, W. B. Croft, M. Iyyer, Open-retrieval conversational question answering, in: SIGIR, 2020, pp. 539–548.

[6] J. Dalton, C. Xiong, V. Kumar, J. Callan, CAsT-19: A dataset for conversational information seeking, in: SIGIR, 2020, pp. 1985–1988.