UNIVERSITY OF AMSTERDAM

Performance Prediction for Conversational Search Using Perplexities of Query Rewrites

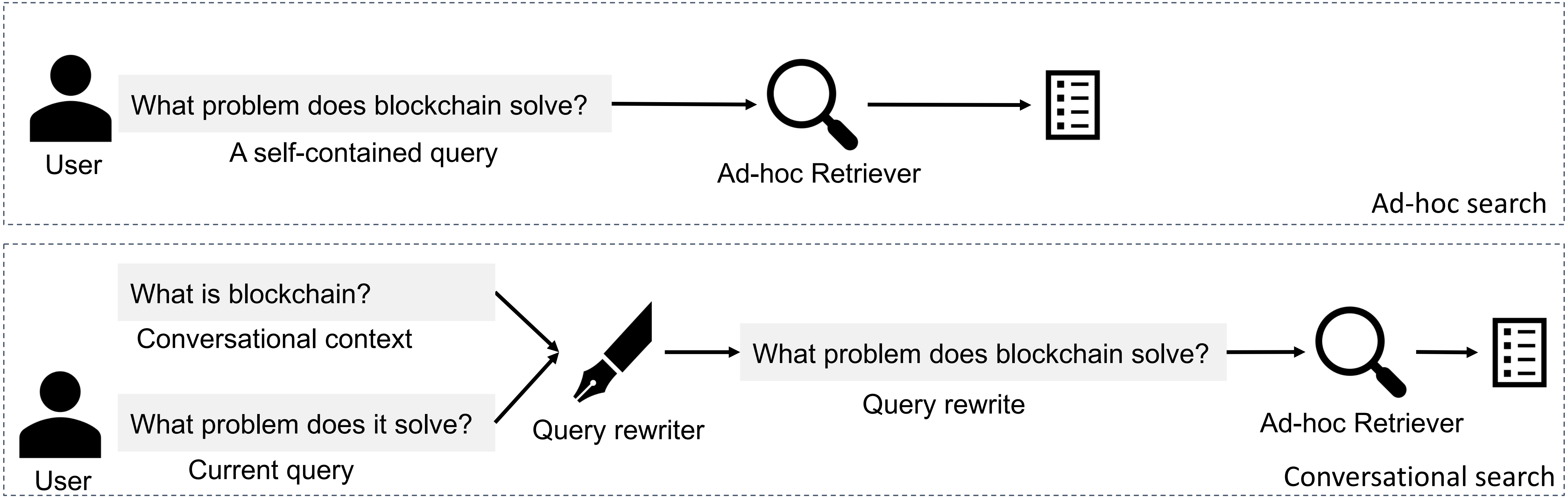
Chuan Meng, Mohammad Aliannejadi and Maarten de Rijke

Outline

- ❑ **Background**
- ❑ Motivation
- ❑ Methodology
- ❑ Experiments
- ❑ Conclusion and Future Work

Background—Conversational Search

- A big difference between ad-hoc and conversational search [1,2]:
  - self-contained query vs. context-dependent query
- A popular pipeline for conversational search:
  - query rewriting + ad-hoc retrieval



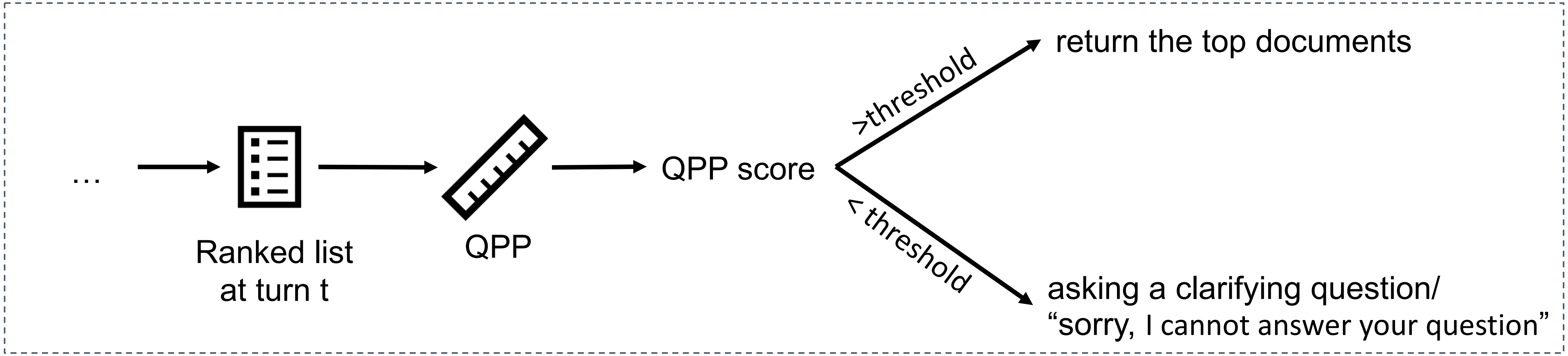
[1] Mao et al. Learning Denoised and Interpretable Session Representation for Conversational Search. In WWW 2023.
[2] Qian et al. Explicit Query Rewriting for Conversational Dense Retrieval. In EMNLP, 2022.

Outline

- [x] Background
- [ ] **Motivation**
- [ ] Methodology
- [ ] Experiments
- [ ] Conclusion and Future Work

Motivation

- Why do we need QPP for conversational search?
  - help a conversational search system take appropriate action at the current turn [1,2]



[1] Arabzadeh et al. Unsupervised Question Clarity Prediction Through Retrieved Item Coherency. In CIKM 2022.
[2] Roitman et al. A Study of Query Performance Prediction for Answer Quality Determination. In ICTIR 2019.

Motivation

- Lower query rewriting quality tends to result in lower retrieval quality
- Query rewriting quality provides evidence for QPP

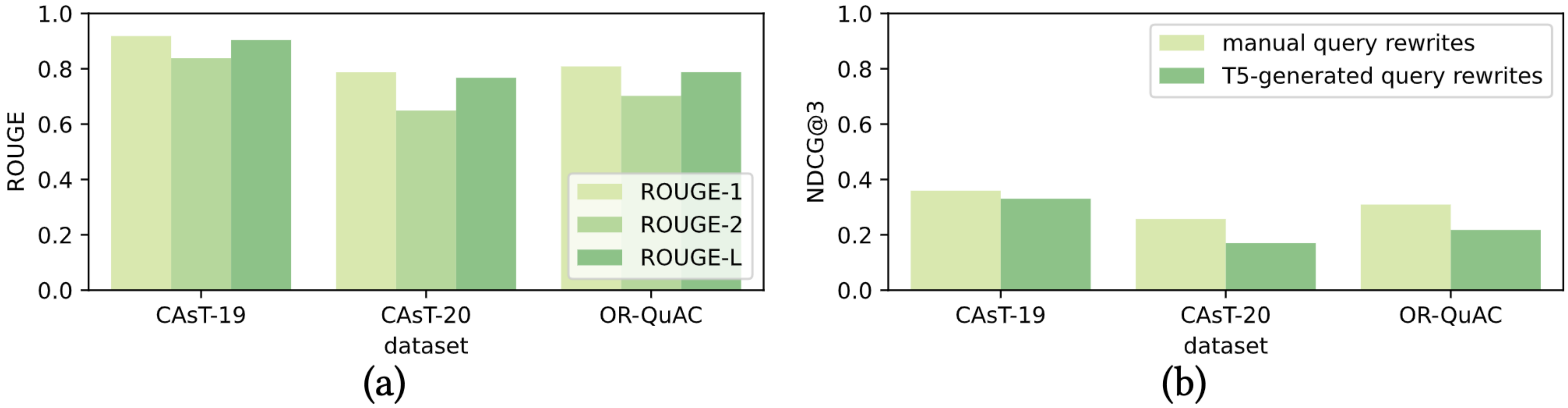


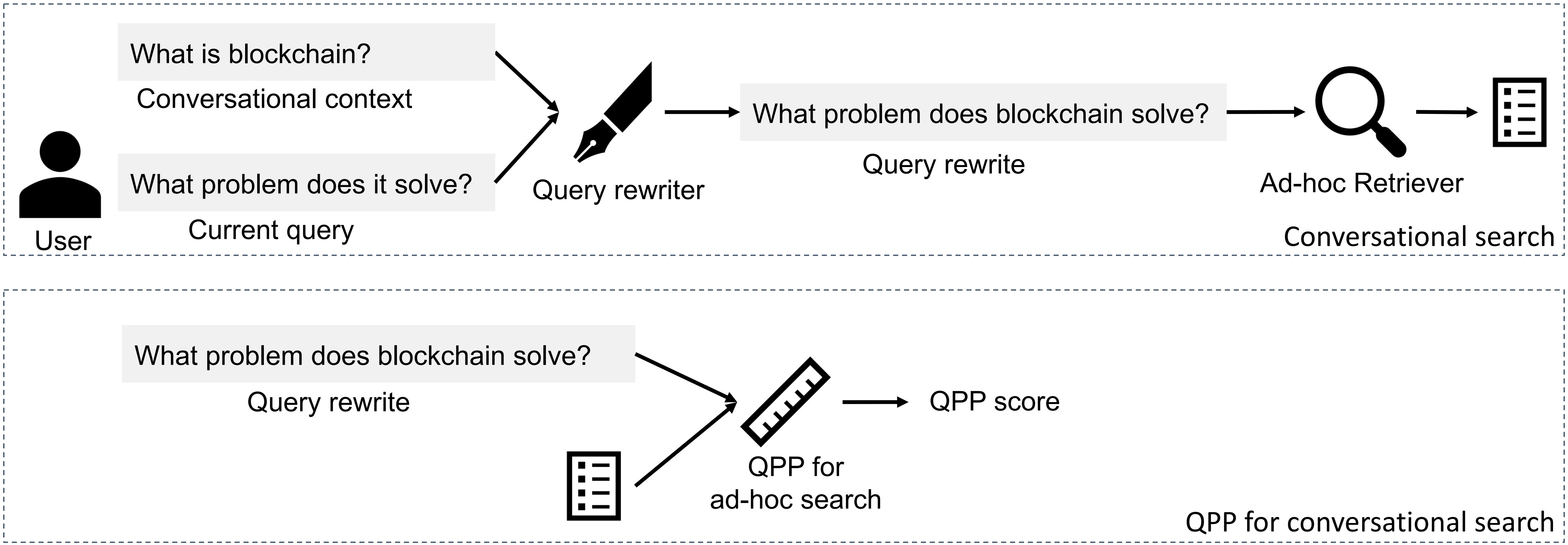
Figure 1: The similarity between manual and T5-generated query rewrites in terms of ROUGE (a) and the retrieval quality of BM25 for manual/T5-generated query rewrites in terms of NDCG@3 (b).

Outline

- ☑ Background
- ☑ Motivation
- ❑ **Methodology**
- ❑ Experiments
- ❑ Conclusion and Future Work

Methodology

- Feed query rewrites to QPP methods designed for ad-hoc search



Methodology

- How?
  - evaluate the query rewriting quality
    - perplexity
  - inject the quality into the QPP
    - linear interpolation
  - $final\ QPP\ score = \alpha \cdot \frac{1}{perplexity} + (1-\alpha) \cdot QPP\ score$

Outline

- [x] Background
- [x] Motivation
- [x] Methodology
- [ ] **Experiments**
- [ ] Conclusion and Future Work

Experiments

- Experimental settings:
  - baselines: QS, SCS, avgICTF, IDF, PMI, SCQ, VAR
  - retriever: T5 query rewriter [1] + BM25
  - target metric: nDCG@3
  - perplexity measurer: GPT-2 XL (1.5B parameters) [2]

[1] https://huggingface.co/castorini/t5-base-canard
[2] https://huggingface.co/gpt2-xl

Experiments

- Observations:
  - lower quality tends to lead to worse QPP effectiveness
  - PPL-QPP improves QPP effectiveness on CAsT-19 and, in particular, CAsT-20

Methods	CAsT-19			CAsT-20		
	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*
SCS	0.191	0.134	0.191	0.173	0.102	0.140
avgICTF	0.266	0.180	0.257	0.142	0.107	0.144
IDF (avg, avg, sum)	0.271	0.187	0.267	0.149	0.114	0.152
PMI (max, avg, max)	0.320	0.208	0.293	0.136	0.113	0.155
SCQ (avg, avg, max)	0.174	0.127	0.178	0.224	0.167	0.226
VAR (sum, avg, sum)	0.321	0.221	0.310	0.210	0.162	0.221
PPL-QPP	**0.324**	**0.225**	**0.315**	**0.231**	**0.191**	**0.256**

Outline

- ☑ Background
- ☑ Motivation
- ☑ Methodology
- ☑ Experiments
- ☐ **Conclusion and Future Work**

Conclusion and Future Work

- Contributions
  - propose PPL-QPP that incorporates query rewriting quality into QPP methods.
  - PPL-QPP improves QPP effectiveness if the query rewriting quality is limited.
- Future work
  - incorporate query rewriting quality into post-retrieval QPP methods
  - the choice of evaluator for measuring the quality of query rewrites

Code:
https://github.com/ChuanMeng/QPP4CS

Conclusion and Future Work

- Our reproducibility paper has been accepted at SIGIR 2023

Query Performance Prediction:
From Ad-hoc to Conversational Search

A Reproducibility Study

Chuan Meng
University of Amsterdam
Amsterdam, The Netherlands
c.meng@uva.nl

Negar Arabzadeh
University of Waterloo
Waterloo, Canada
narabzad@uwaterloo.ca

Mohammad Aliannejadi
University of Amsterdam
Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.deRijke@uva.nl

ABSTRACT

Query performance prediction (QPP) is a core task in information retrieval. The QPP task is to predict the retrieval quality of a search system for a query without relevance judgments. Research has shown the effectiveness and usefulness of QPP for ad-hoc search, where a retrieval system can adapt its ranking strategies according to the predicted difficulty of a query. Recent years have witnessed considerable progress in conversational search (CS). Effective QPP could help a CS system to decide an appropriate action to be taken at the next turn. Despite its potential, QPP for CS has been little studied. We address this research gap by reproducing and studying the effectiveness of multiple existing QPP methods in the context of CS. While the task of passage retrieval remains the same in the two settings, a user query in CS depends on the conversational history,

ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan. ACM, New York, NY, USA, 12 pages.

1 INTRODUCTION

Query performance prediction (QPP) is an essential task in information retrieval (IR). It is about estimating the retrieval quality of a search system for a given query without relevance judgments [13, 15, 21, 25, 52, 55]. QPP has been long studied in the IR community [9]. Numerous benefits of QPP have been identified, including selecting the most effective ranking algorithm for a query [25, 26, 52] based on the difficulty of the input query.

In conversational search (CS) there has been significant progresson multiple subtasks [54], including passage retrieval [12, 51],

Thanks!

Q & A