UNIVERSITY OF AMSTERDAM

Query Performance Prediction for Conversational Search

Chuan Meng

Information Retrieval Lab (IRLab)

University of Amsterdam

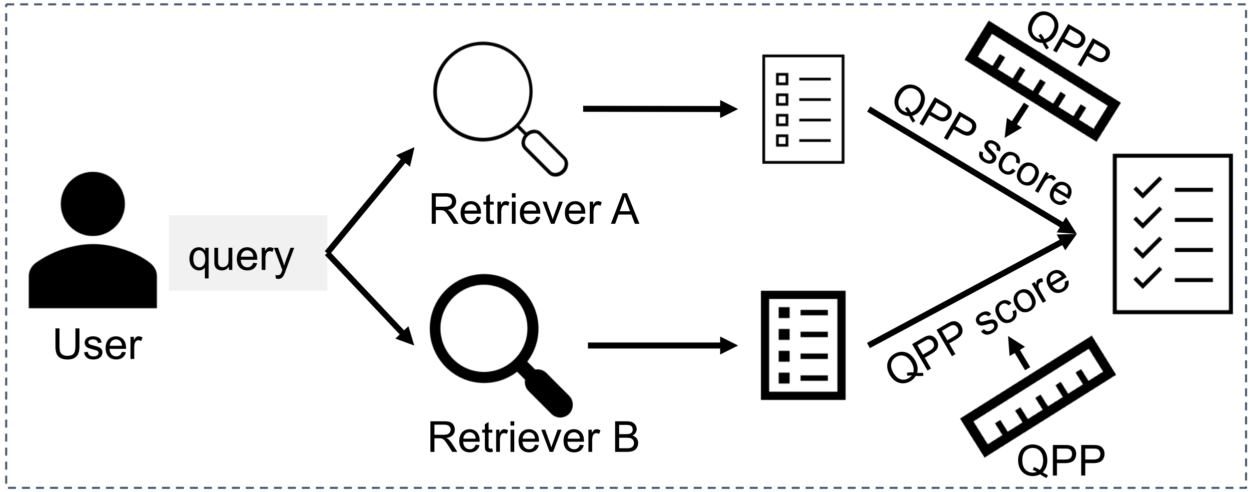
22nd May 2023

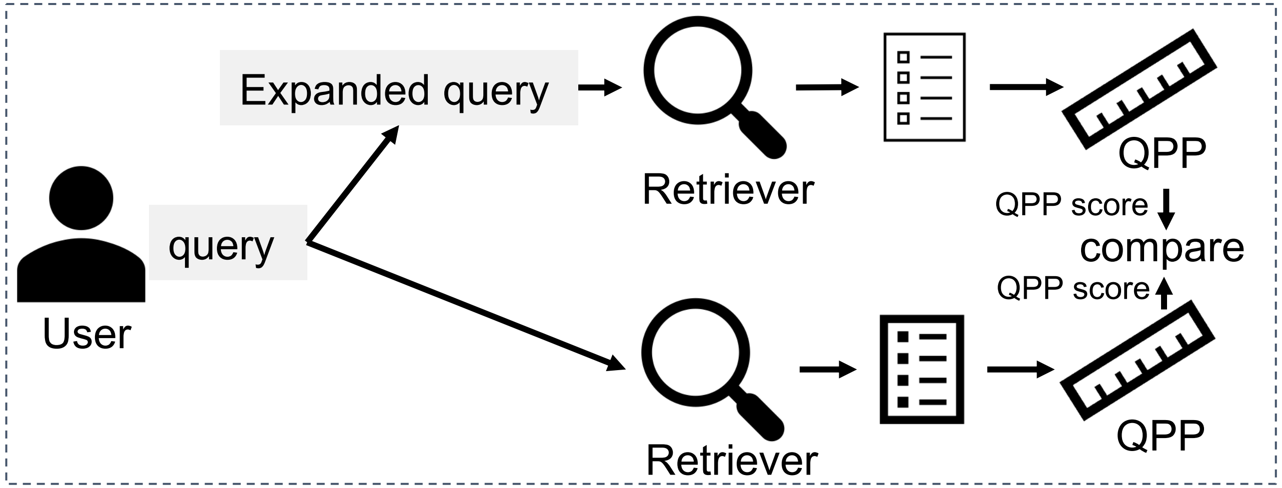
Outline

- **Background**
  - Query Performance Prediction (QPP)
  - Conversational Search (CS)
- Motivation
- Study 1: Reproducing existing QPP methods in CS (SIGIR 2023)
- Study 2: Improve QPP for CS using query rewriting quality (QPP++2023)
- Conclusion

Background—Query Performance Prediction

- Query performance prediction (QPP)
  - Estimates the retrieval quality of a search system for a given query without relevance judgments [1,2,3].
  - Widely studied in the fields of ad-hoc search [1,2] and retrieval-based non-factoid question answering [3]
- QPP is beneficial for many reasons:
  - Ranking fusion [4], selective query expansion [5], etc.





[1] Datta et al. A 'Pointwise-Query, Listwise-Document based Query Performance Prediction Approach. In SIGR 2022.
[2] Negar et al. BERT-QPP: Contextualized Pre-trained Transformers for Query Performance Prediction. In CIKM, 2021.
[3] Hashemi et al. Performance Prediction for Non-Factoid Question Answering. In ICTIR 2019.
[4] Mackenzie et al. Query-Performance Prediction: Setting the Expectations Straight. In SIGIR 2014.
[5] Amati et al. Query Difficulty, Robustness, and Selective Application of Query Expansion. In ECIR 2014.

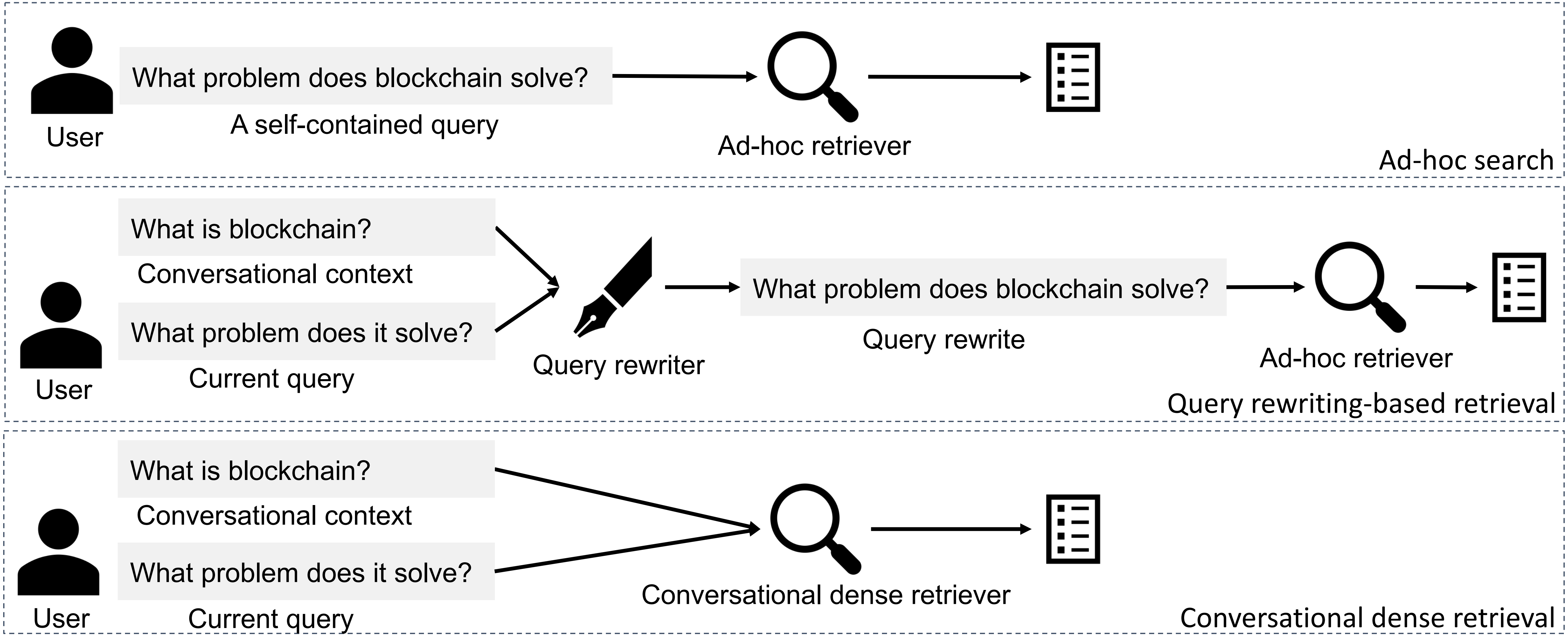
Background—Query Performance Prediction

- There are two types of QPP methods
  - pre-retrieval QPP methods [1,2]
    - $f(query) \rightarrow QPP\ \text{score}$
  - post-retrieval QPP methods [3,4,5]
    - $f(query, a\ ranked\ list) \rightarrow QPP\ score$

[1] Arabzadeh et al. Neural embedding-based specificity metrics for pre-retrieval query performance prediction. In IPM 2020.
[2] Roy et al. Estimating Gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. In IPM 2019.
[3] Chen et al. Groupwise Query Performance Prediction with BERT. In ECIR 2022.
[4] Datta et al. A 'Pointwise-Query, Listwise-Document based Query Performance Prediction Approach. In SIGR 2022.
[5] Arabzade et al. BERT-QPP: Contextualized Pre-trained Transformers for Query Performance Prediction. In CIKM, 2021.

Background—Conversational Search (CS)

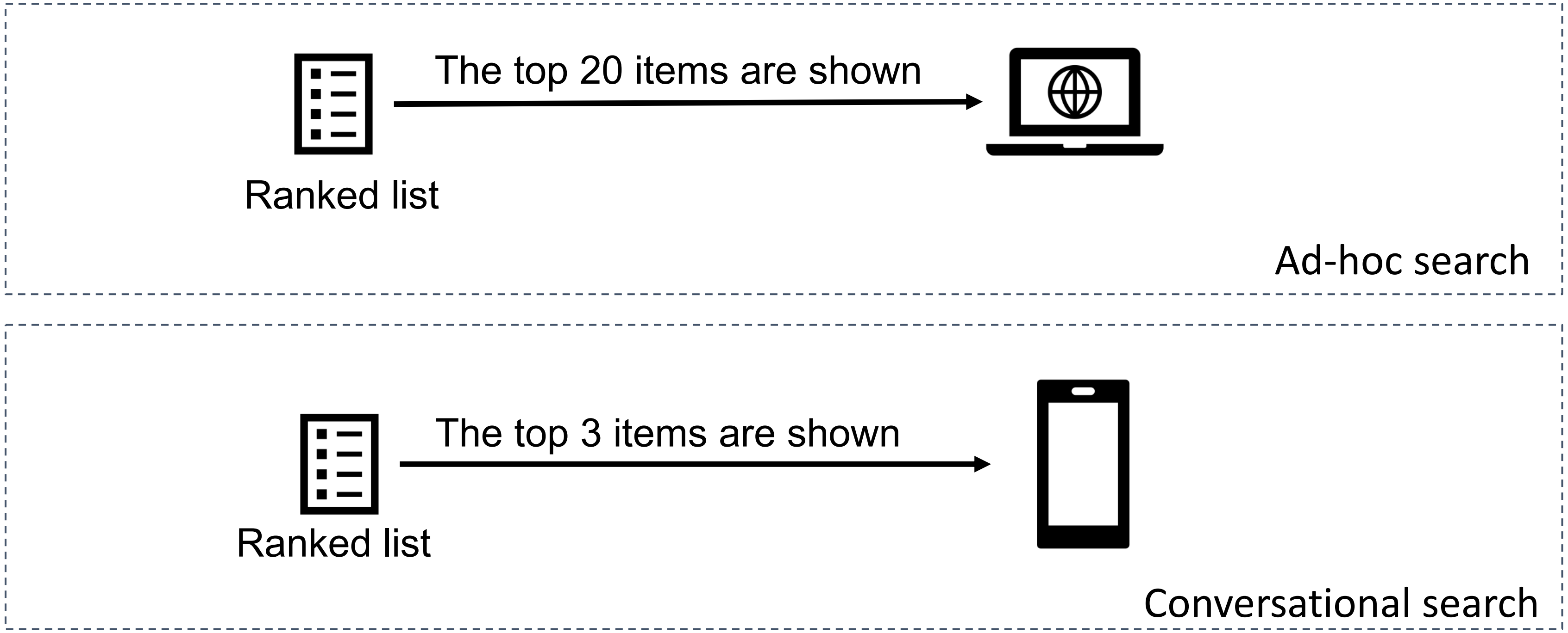
- Queries are different between ad-hoc and CS [1,2]:
  - Self-contained query vs. context-dependent query



[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.

Background—Conversational Search

- Preferred ranking depth is different between ad-hoc and CS [1]:
  - large cut-off (nDCG@20) vs. small cut-off (nDCG@3)



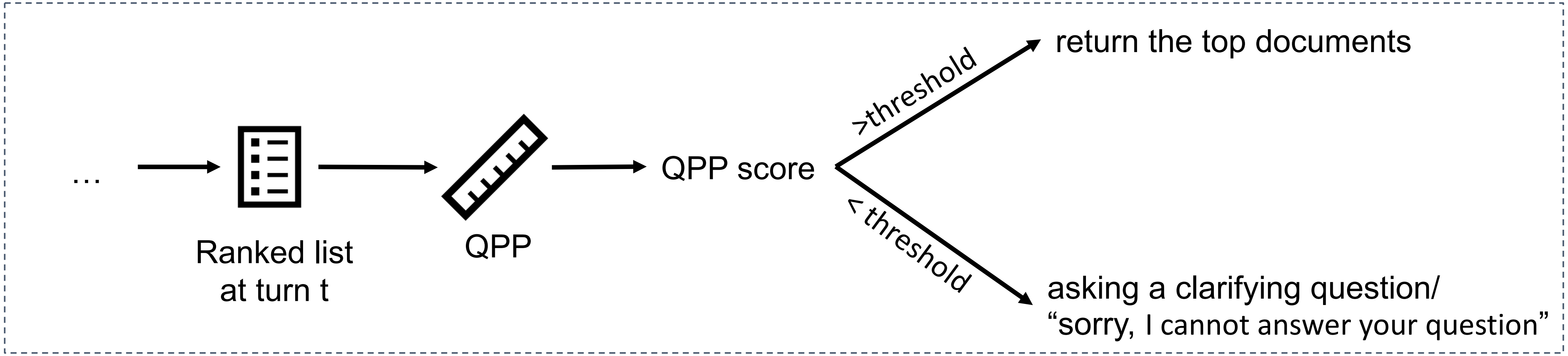
[1] Dalton et al. Cast-19: A Dataset for Conversational Information Seeking. In SIGIR 2020.

Outline

- ☑ Background
  - ☑ Query Performance Prediction (QPP)
  - ☑ Conversational Search (CS)
- ☐ **Motivation**
- ☐ Study 1: Reproducing existing QPP methods in CS (SIGIR 2023)
- ☐ Study 2: Improve QPP for CS using query rewriting quality (QPP++2023)
- ☐ Conclusion

Motivation

- Why do we need QPP for CS? QPP can benefit CS in terms of
  - Action prediction [1,2]
  - Query expansion determination [3]
  - Query rewrite selection [4]
  - Clarifying question selection in CS [5]
  - Conversation contextualization [6]



[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.
[3] Lin et al. Multi-Stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting. In TOIS 2021.
[4] Al-Thani,et al. Improving Conversational Search with Query Reformulation Using Selective Contextual History. DIM 2022.
[5] Aliannejadi et al. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In SIGIR 2019.
[6] Dipasree et al. Effective Query Formulation in Conversation Contextualization : A Query Specificity-based Approach. In ICTIR 2021.

Motivation

- We already know that simply applying QPP to CS benefits CS [1-6]
- However, we still do not know:

- How well various existing ad-hoc QPP methods perform in CS
  - **Motivate a comprehensive reproducibility study**
- A QPP method specifically designed for CS
  - **Motivate a new QPP method for CS**

[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.
[3] Lin et al. Multi-Stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting. In TOIS 2021.
[4] Al-Thani,et al. Improving Conversational Search with Query Reformulation Using Selective Contextual History. DIM 2022.
[5] Aliannejadi et al. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In SIGIR 2019.
[6] Dipasree et al. Effective Query Formulation in Conversation Contextualization : A Query Specificity-based Approach. In ICTIR 2021.

Outline

- ☑ Background
  - ☑ Query Performance Prediction (QPP)
  - ☑ Conversational Search (CS)
- ☑ Motivation
- ❑ **Study 1: Reproducing existing QPP methods in CS (SIGIR 2023)**
- ❑ Study 2: Improve QPP for CS using query rewriting quality (QPP++2023)
- ❑ Conclusion

Query Performance Prediction: From Ad-hoc to Conversational Search

Chuan Meng, Negar Arabzadeh, Mohammad Aliannejadi and Maarten de Rijke

Got accepted at SIGIR 2023

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Examine whether the three findings on QPP for ad-hoc search still hold in CS
  1. Supervised QPP methods outperform unsupervised QPP methods [1-6]
  2. List-wise supervised QPP methods outperform point-wise ones [1,2]
  3. Retrieval score-based unsupervised QPP methods perform badly in estimating the retrieval quality of neural-based retrievers [5,7]

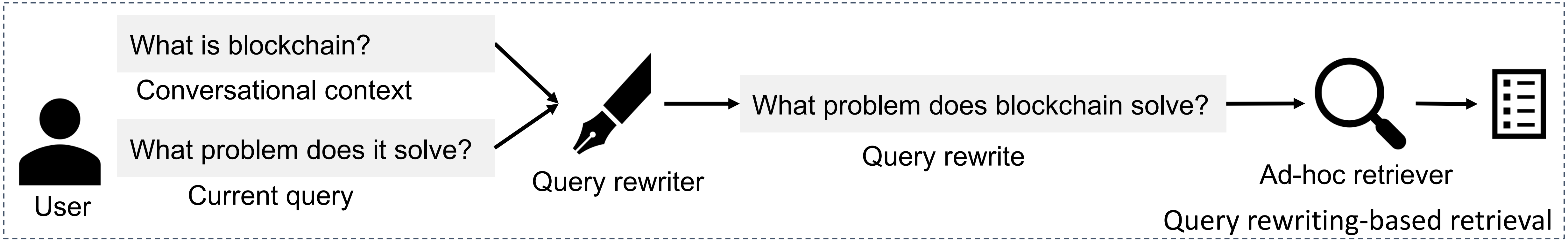
[1] Datta et al. A ‘Pointwise-Query, Listwise-Document based Query Performance Prediction Approach. In SIGR 2022.
[2] Chen et al. Groupwise Query Performance Prediction with BERT. In ECIR 2022.
[3] Datta et al. Deep-QPP: A Pairwise Interaction-based Deep Learning Model for Supervised Query Performance Prediction. In WSDM 2022.
[4] Arabzadeh et al. BERT-QPP: Contextualized Pre-trained Transformers for Query Performance Prediction. In CIKM 2021.
[5] Hashemi et al. Performance Prediction for Non-Factoid Question Answering. In ICTIR 2019.
[6] Zamani et al. Neural Query Performance Prediction Using Weak Supervision from Multiple Signals. In SIGIR 2018.
[7] Datta et al. A Relative Information Gain-based Query Performance Prediction Framework with Generated Query Variants. In TOIS 2022.

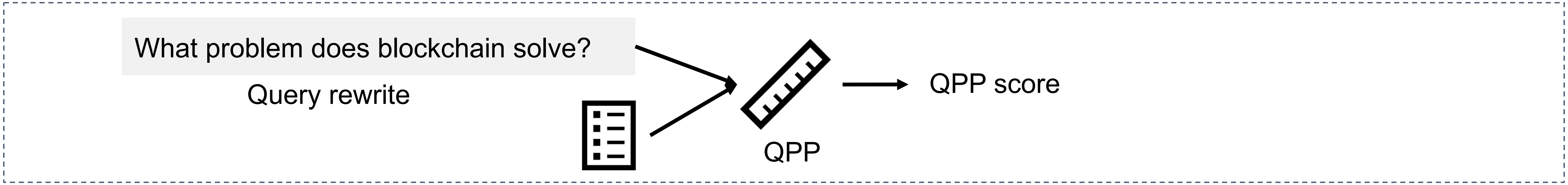
Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Research Questions:
  1. Does the performance of QPP methods for ad-hoc search generalize to CS when estimating the retrieval quality of different query rewriting-based retrieval methods?
  2. Does the performance of QPP methods for ad-hoc search generalize to CS when estimating the retrieval quality of a conversational dense retrieval method?
  3. What is the performance difference between QPP methods when predicting the retrieval quality for top-ranked items vs. for longer-ranked lists?

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Experimental design for RQ1:
  - Estimate the retrieval quality of
    - T5-based query rewriter + BM25 [1]
    - QuReTeC-based query rewriter+BM25 [2]
    - Human query rewriter + BM25
  - QPP methods and BM25 always share the same query rewrites.

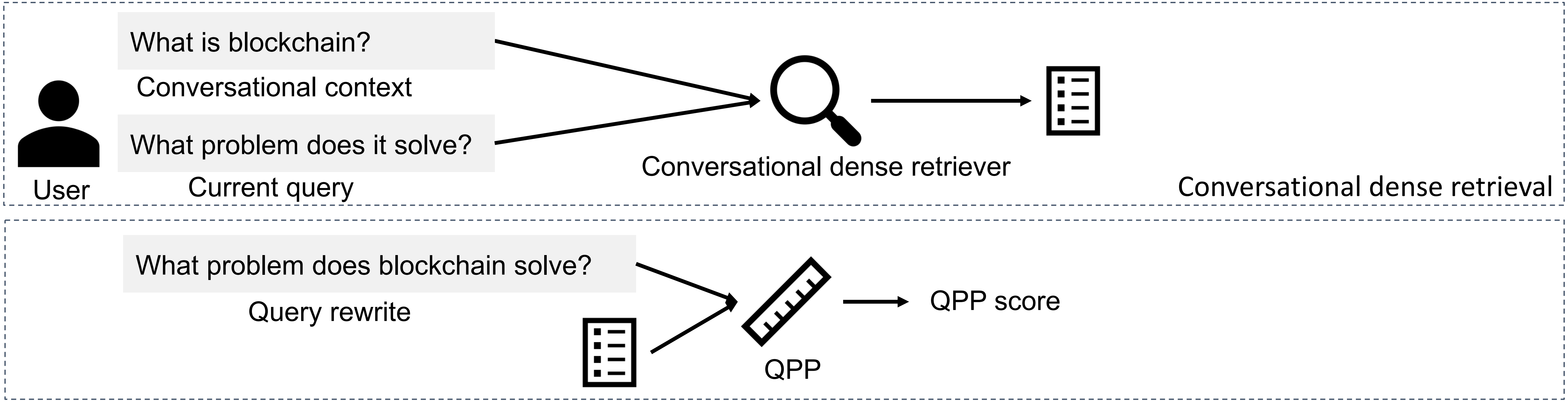




[1] Lin et al. Multi-Stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting. In TOIS 2021.
[2] Voskarides et al. Query Resolution for Conversational Search with Limited Supervision. In SIGIR 2020.

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Experimental design for RQ2:
  - Estimate the retrieval quality of a conversational dense retriever, ConvDR [1]
  - Study the effect of feeding three different query rewrites into QPP methods
    - T5-based query rewrites [2]
    - QuReTeC-based query rewrites [3]
    - Human-rewritten query rewrites



[1] Yu et al. Few-Shot Conversational Dense Retrieval. In SIGIR 2021.
[2] Lin et al. Multi-Stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting. In TOIS 2021.
[3] Voskarides et al. Query Resolution for Conversational Search with Limited Supervision. In SIGIR 2020.

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Experimental design for RQ3:
  - Estimate the retrieval quality in terms of ranking metrics with different cut-offs
    - nDCG@3 [1]
    - nDCG@100
    - Recall@100, for first-stage CS rankers

[1] Dalton et al. Cast-19: A Dataset for Conversational Information Seeking. In SIGIR 2020.

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Experimental settings:
  - QPP methods
    - Unsupervised:
      - Clarity [1]
      - WIG [2]: magnitude of retrieval scores
      - NQC [3], σ_{max} [4], $n(\sigma_{x\%})$ [5]: standard deviation of retrieval scores
      - SMV [6]: consider magnitude and standard deviation
    - Supervised:
      - NQA-QPP [7], BERT-QPP [8]: point-wise methods
      - qppBERT-PL [9]: a listwise-document method

[1] Cronen-Townsend et al, Predicting Query Performance. In SIGIR 2002
[2] Zhou et al. Query Performance Prediction in Web Search Environments. In SIGIR 2007.
[3] Shtok et al. Predicting Query Performance by Query-Drift Estimation. In TOIS 2012.
[4] Pérez-Iglesias et al. Standard Deviation as a Query Hardness Estimator. In SPIRE 2010.
[5] Cummins et al. Improved Query Performance Prediction Using Standard Deviation. SIGIR 2010.
[6] Tao et al. Query Performance Prediction by Considering Score Magnitude and Variance Together. In CIKM 2014.
[7] Hashemi et al. Performance Prediction for Non-Factoid Question Answering. In ICTIR 2019.
[8] Arabzadeh et al. BERT-QPP: Contextualized Pre-trained Transformers for Query Performance Prediction. In CIKM 2021.
[9] Datta et al. A ‘Pointwise-Query, Listwise-Document based Query Performance Prediction Approach. In SIGR 2022..

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Experimental settings:
  - Datasets:
    - CAsT-19 [1]
    - CAsT-20 [2] with harder information needs and query rewriting
    - OR-QuAC [3]

	CAsT-19	CAsT-20	OR-QuAC		
	test	test	train	valid	test
#conversations	50	25	4,383	490	771
#conversations (judged)	20	25	–	–	–
#questions	479	216	31,526	3,430	5,571
#questions (judged)	173	208	–	–	–
#documents	38M		11M		

	CAsT-19	CAsT-20	OR-QuAC
T5-based query rewriter + BM25	0.330	0.170	0.218
QuReTeC-based query rewriter + BM25	0.338	0.172	0.249
Human query rewriter + BM25	0.360	0.257	0.309
ConvDR	0.471	0.343	0.614

[1] Dalton et al. Cast-19: A Dataset for Conversational Information Seeking. In SIGIR 2020.
[2] Dalton et al. CAsT 2020: The Conversational Assistance Track Overview. In Text Retrieval Conference 2020.
[3] Qu et al. Open-retrieval Conversational Question Answering. In SIGIR 2020.

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Experimental settings:
  - Evaluation metrics
    - Pearson's ρ, Kendall's τ , and Spearman's ρ correlation coefficients

[1] Cronen-Townsend et al, Predicting Query Performance. In SIGIR 2002
[2] Zhou et al. Query Performance Prediction in Web Search Environments. In SIGIR 2007.
[3] Shtok et al. Predicting Query Performance by Query-Drift Estimation. In TOIS 2012.
[4] Pérez-Iglesias et al. Standard Deviation as a Query Hardness Estimator. In SPIRE 2010.
[5] Cummins et al. Improved Query Performance Prediction Using Standard Deviation. SIGIR 2010.
[6] Tao et al. Query Performance Prediction by Considering Score Magnitude and Variance Together. In CIKM 2014.
[7] Hashemi et al. Performance Prediction for Non-Factoid Question Answering. In ICTIR 2019.
[8] Arabzadeh et al. BERT-QPP: Contextualized Pre-trained Transformers for Query Performance Prediction. In CIKM 2021.
[9] Datta et al. A 'Pointwise-Query, Listwise-Document based Query Performance Prediction Approach. In SIGR 2022..

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Results for RQ1:
  - Feeding T5/QuReTeC query rewrites into QPP methods is effective
  - Supervised methods perform best when large-scale training data is available
  - NQA-QPP and BERTQPP outperform qppBERT-PL

Datasets	QPP methods	T5+BM25			QuReTeC+BM25			Human+BM25		
		P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ
OR-QuAC	Clarity	0.090	0.085	0.110	0.110	0.103	0.133	0.076	0.069	0.091
	WIG	0.247	0.235	0.304	0.290	0.270	0.350	0.257	0.241	0.316
	NQC	0.251	0.274	0.355	0.290	0.311	0.404	0.276	0.291	0.381
	σ_{max}	0.317	0.279	0.359	0.367	0.316	0.406	0.412	0.367	0.474
	n($\sigma_{x\%}$)	0.181	0.172	0.223	0.229	0.209	0.270	0.245	0.193	0.252
	SMV	0.204	0.239	0.310	0.239	0.273	0.355	0.194	0.232	0.304
	NQA-QPP	**0.781**	**0.566**	**0.695**	**0.792**	**0.591**	**0.725**	**0.809**	**0.621**	**0.767**
	BERTQPP	0.678	0.434	0.546	0.692	0.476	0.598	0.725	0.527	0.666
	qppBERT-PL	0.594	0.507	0.576	0.617	0.526	0.597	0.618	0.525	0.600

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Results for RQ1:
  - Feeding T5/QuReTeC query rewrites into QPP methods is effective
  - Supervised methods are comparable/inferior to unsupervised ones
  - qppBERT-PL has a slight advantage in a few-shot setting

Datasets	QPP methods	T5+BM25 P-ρ	T5+BM25 K-τ	T5+BM25 S-ρ	QuReTeC+BM25 P-ρ	QuReTeC+BM25 K-τ	QuReTeC+BM25 S-ρ	Human+BM25 P-ρ	Human+BM25 K-τ	Human+BM25 S-ρ
CAsT-19	Clarity	0.321	0.234	0.330	0.327	0.211	0.304	0.359	0.231	0.335
	WIG	0.436	0.232	0.452	0.354	0.250	0.356	0.409	0.293	0.414
	NQC	0.348	0.246	0.354	0.286	0.190	0.275	0.334	0.234	0.335
	σ_{max}	0.442	0.354	0.501	0.351	0.251	0.357	0.410	**0.312**	**0.441**
	n($\sigma_{x\%}$)	0.430	0.332	0.466	0.348	0.259	0.364	0.407	0.307	0.430
	SMV	0.344	0.250	0.360	0.289	0.188	0.273	0.326	0.230	0.333
	NQA-QPP	0.188	*0.047*	*0.072*	*-0.016*	*0.010*	*0.014*	0.152	*0.069*	*0.099*
	BERTQPP	0.440	0.307	0.424	0.352	0.272	0.395	0.270	0.188	0.271
	qppBERT-PL	0.414	0.296	0.421	0.392	0.298	0.406	0.292	0.196	0.280
CAsT-20	Clarity	0.258	0.191	0.259	*0.099*	*0.061*	*0.085*	*0.127*	*0.089*	*0.121*
	WIG	0.248	**0.251**	**0.339**	0.245	0.163	0.222	0.307	0.222	0.317
	NQC	0.150	0.235	0.316	0.198	0.189	0.259	0.286	**0.266**	**0.370**
	σ_{max}	0.179	0.221	0.304	0.207	0.168	0.230	0.241	0.199	0.283
	n($\sigma_{x\%}$)	0.178	0.225	0.304	0.182	0.133	0.181	0.213	0.167	0.237
	SMV	0.139	0.219	0.298	0.189	0.163	0.227	0.264	0.260	0.363
	NQA-QPP	*0.001*	*0.067*	*0.093*	*-0.064*	*-0.082*	*-0.111*	*0.086*	*-0.011*	*-0.012*
	BERTQPP	*0.042*	*-0.009*	*-0.007*	0.172	0.145	0.196	0.194	0.110	0.159
	qppBERT-PL	*0.131*	0.125	0.159	0.175	0.150	0.185	*0.043*	*0.015*	*0.021*

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Results for RQ1:
  - Supervised methods perform better after warming up
  - They still do not have a distinct advantage on CAsT-20

Datasets	QPP methods	T5+BM25			QuReTeC+BM25			Human+BM25		
		P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ
CAsT-19	Clarity	0.321	0.234	0.330	0.327	0.211	0.304	0.359	0.231	0.335
	WIG	0.436	0.232	0.452	0.354	0.250	0.356	0.409	0.293	0.414
	NQC	0.348	0.246	0.354	0.286	0.190	0.275	0.334	0.234	0.335
	σ_{max}	0.442	0.354	0.501	0.351	0.251	0.357	0.410	**0.312**	**0.441**
	n($\sigma_{x\%}$)	0.430	0.332	0.466	0.348	0.259	0.364	0.407	0.307	0.430
	SMV	0.344	0.250	0.360	0.289	0.188	0.273	0.326	0.230	0.333
	NQA-QPP (warm-up)	**0.538**	**0.357**	**0.510**	**0.420**	**0.301**	**0.428**	0.331	0.230	0.336
	BERTQPP (warm-up)	0.526	**0.357**	0.503	0.369	0.264	0.384	**0.418**	0.282	0.411
	qppBERT-PL (warm-up)	0.317	0.218	0.313	0.330	0.232	0.326	0.297	0.190	0.277
CAsT-20	Clarity	0.258	0.191	0.259	*0.099*	*0.061*	*0.085*	*0.127*	*0.089*	*0.121*
	WIG	0.248	**0.251**	**0.339**	0.245	0.163	0.222	0.307	0.222	0.317
	NQC	0.150	0.235	0.316	0.198	0.189	0.259	0.286	**0.266**	**0.370**
	σ_{max}	0.179	0.221	0.304	0.207	0.168	0.230	0.241	0.199	0.283
	n($\sigma_{x\%}$)	0.178	0.225	0.304	0.182	0.133	0.181	0.213	0.167	0.237
	SMV	0.139	0.219	0.298	0.189	0.163	0.227	0.264	0.260	0.363
	NQA-QPP (warm-up)	**0.274**	0.170	0.227	0.190	0.149	0.201	0.231	0.155	0.222
	BERTQPP (warm-up)	0.207	0.171	0.236	**0.403**	**0.301**	**0.409**	**0.336**	0.227	0.318
	qppBERT-PL (warm-up)	0.228	0.213	0.275	0.317	0.268	0.335	*0.094*	*0.095*	*0.130*

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Results for RQ2:
  - Supervised methods perform best when large-scale training data is available
  - NQA-QPP and BERTQPP outperform qppBERT-PL

		T5			QuReTeC			Human		
Datasets	QPP methods	P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ
OR-QuAC	Clarity	-0.050	-0.029	-0.038	-0.050	-0.029	-0.038	-0.050	-0.029	-0.038
	WIG	0.137	0.107	0.145	0.116	0.088	0.120	0.140	0.111	0.149
	NQC	0.227	0.163	0.221	0.227	0.163	0.221	0.227	0.163	0.221
	σ_{max}	0.442	0.339	0.443	0.442	0.339	0.443	0.442	0.339	0.443
	n($\sigma_{x\%}$)	-0.032	*-0.003*	*-0.004*	-0.073	-0.035	-0.045	*-0.022*	*0.008*	*0.011*
	SMV	0.098	0.076	0.106	0.098	0.076	0.106	0.098	0.076	0.106
	NQA-QPP	**0.615**	**0.479**	**0.615**	**0.639**	**0.499**	**0.638**	**0.600**	**0.470**	**0.601**
	BERTQPP	<u>0.481</u>	<u>0.417</u>	<u>0.541</u>	<u>0.505</u>	<u>0.435</u>	<u>0.563</u>	<u>0.481</u>	<u>0.408</u>	<u>0.529</u>
	qppBERT-PL	0.391	0.250	0.287	0.424	0.294	0.335	0.437	0.306	0.349

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Results for RQ2:
  - Retrieval score-based methods NQC/WIG perform best in most cases
  - Supervised methods tend to perform better when fed with human-rewritten queries
  - qppBERT-PL has a slight advantage in a few-shot setting

		T5			QuReTeC			Human		
Datasets	QPP methods	P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ
CAsT-19	Clarity	0.257	0.176	0.257	0.257	0.176	0.257	0.257	0.176	0.257
	WIG	<u>0.387</u>	0.274	0.395	<u>0.388</u>	0.266	0.381	<u>0.412</u>	<u>0.285</u>	<u>0.408</u>
	NQC	**0.431**	**0.307**	**0.438**	**0.431**	**0.307**	**0.438**	**0.431**	**0.307**	**0.438**
	σ_{max}	0.378	0.267	0.381	0.378	0.267	0.381	0.378	0.267	0.381
	n($\sigma_{x\%}$)	0.187	0.175	0.262	0.181	0.170	0.256	0.216	0.196	0.288
	SMV	0.386	<u>0.285</u>	<u>0.405</u>	0.386	<u>0.285</u>	<u>0.405</u>	0.386	<u>0.285</u>	0.405
	NQA-QPP	*0.121*	*0.075*	*0.115*	*0.118*	*0.073*	*0.109*	0.150	0.109	0.153
	BERTQPP	0.167	0.107	0.169	0.220	0.145	0.217	0.298	0.193	0.296
	qppBERT-PL	0.344	0.225	0.324	0.316	0.197	0.284	0.276	0.178	0.255
CAsT-20	Clarity	*0.126*	*0.088*	*0.127*	*0.126*	*0.088*	*0.127*	*0.126*	*0.088*	*0.127*
	WIG	**0.377**	**0.277**	**0.386**	**0.377**	**0.263**	**0.373**	<u>0.384</u>	0.264	0.368
	NQC	<u>0.339</u>	<u>0.261</u>	<u>0.360</u>	<u>0.339</u>	<u>0.261</u>	<u>0.360</u>	0.339	0.261	0.360
	σ_{max}	0.282	0.219	0.310	0.282	0.219	0.310	0.282	0.219	0.310
	n($\sigma_{x\%}$)	0.199	0.168	0.236	0.197	0.156	0.224	0.201	0.156	0.220
	SMV	0.275	0.216	0.299	0.275	0.216	0.299	0.275	0.216	0.299
	NQA-QPP	*-0.037*	*-0.037*	*-0.058*	*-0.081*	*-0.063*	*-0.092*	*0.059*	*0.023*	*0.032*
	BERTQPP	0.223	0.157	0.226	0.216	0.146	0.212	**0.404**	**0.281**	**0.395**
	qppBERT-PL	0.185	0.144	0.191	*0.029*	*0.023*	*0.031*	0.251	0.171	0.232

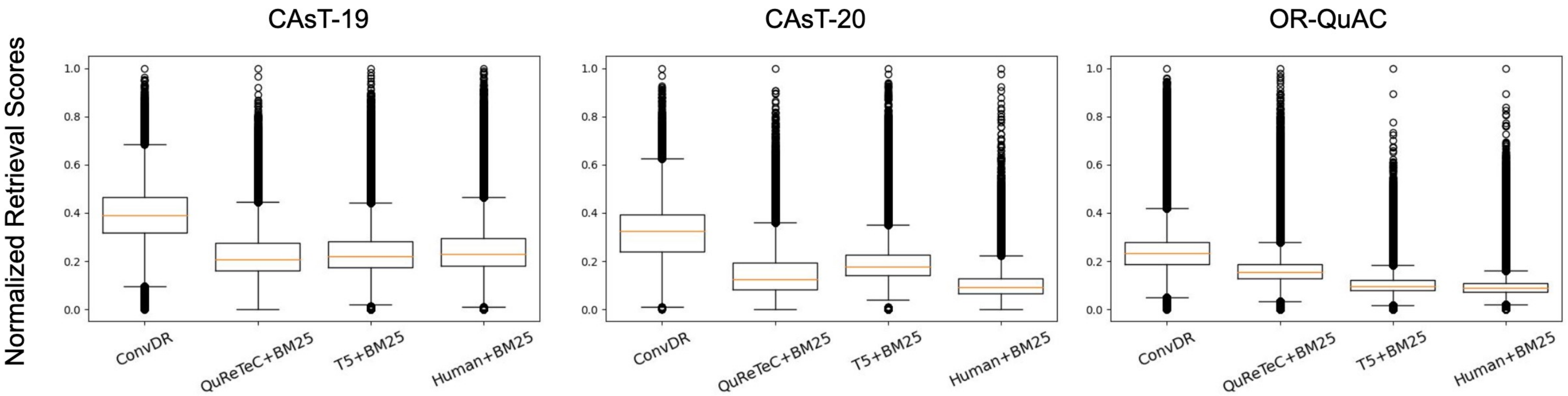
Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Results for RQ2:
  - Retrieval score-based methods NQC/WIG still perform best in most cases
  - Supervised methods tend to perform better when fed with human-rewritten queries

Datasets	QPP methods	T5			QuReTeC			Human		
		P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ
CAsT-19	Clarity	0.257	0.176	0.257	0.257	0.176	0.257	0.257	0.176	0.257
	WIG	0.387	0.274	0.395	0.388	0.266	0.381	0.412	0.285	0.408
	NQC	**0.431**	**0.307**	**0.438**	**0.431**	**0.307**	**0.438**	**0.431**	**0.307**	**0.438**
	σ_{max}	0.378	0.267	0.381	0.378	0.267	0.381	0.378	0.267	0.381
	n($\sigma_{x\%}$)	0.187	0.175	0.262	0.181	0.170	0.256	0.216	0.196	0.288
	SMV	0.386	0.285	0.405	0.386	0.285	0.405	0.386	0.285	0.405
	NQA-QPP (warm-up)	0.187	0.128	0.186	0.161	0.107	0.157	0.287	0.191	0.282
	BERTQPP (warm-up)	0.282	0.187	0.277	0.234	0.157	0.233	0.371	0.251	0.361
	qppBERT-PL (warm-up)	0.212	0.151	0.213	0.167	0.117	0.170	0.172	0.115	0.154
CAsT-20	Clarity	*0.126*	*0.088*	*0.127*	*0.126*	*0.088*	*0.127*	*0.126*	*0.088*	*0.127*
	WIG	**0.377**	**0.277**	**0.386**	**0.377**	**0.263**	**0.373**	0.384	0.264	0.368
	NQC	0.339	0.261	0.360	0.339	0.261	0.360	0.339	0.261	0.360
	σ_{max}	0.282	0.219	0.310	0.282	0.219	0.310	0.282	0.219	0.310
	n($\sigma_{x\%}$)	0.199	0.168	0.236	0.197	0.156	0.224	0.201	0.156	0.220
	SMV	0.275	0.216	0.299	0.275	0.216	0.299	0.275	0.216	0.299
	NQA-QPP (warm-up)	0.315	0.218	0.313	0.240	0.178	0.245	0.374	0.267	0.375
	BERTQPP (warm-up)	0.253	0.183	0.257	0.320	0.236	0.338	0.349	0.244	0.346
	qppBERT-PL (warm-up)	0.218	0.164	0.227	0.140	0.115	0.157	0.348	0.268	0.376

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Results for RQ2:
  - *[1] found that the retrieval scores from neural-based retrievers, such as ColBERT, are restricted within a shorter range compared to lexical-based retrievers, limiting the performance of score-based unsupervised QPP methods.*
  - The retrieval score distribution of ConvDR displays a higher variance
  - Score-based methods tend to be less impacted by the query understanding challenge



[1] Datta et al. A Relative Information Gain-based Query Performance Prediction Framework with Generated Query Variants. In TOIS 2022.

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Results for RQ3:
  - Supervised methods perform best when large-scale training data is available
  - qppBERT-PL performs best when assessing ConvDR in terms of Recall@100

		T5 + BM25						ConvDR (QPP fed with T5 query rewrites)					
		nDCG@3		nDCG@100		Recall@100		nDCG@3		nDCG@100		Recall@100	
	QPP methods	P-ρ	K-τ	P-ρ	K-τ	P-ρ	K-τ	P-ρ	K-τ	P-ρ	K-τ	P-ρ	K-τ
OR-QuAC	Clarity	0.090	0.085	0.197	0.196	0.362	0.312	-0.050	-0.029	-0.029	-0.015	0.053	0.057
	WIG	0.247	0.235	0.376	0.370	0.482	0.450	0.137	0.107	0.195	0.130	0.298	0.261
	NQC	0.251	0.274	0.356	0.409	0.414	0.461	0.227	0.163	0.302	0.194	0.402	0.333
	σ_{max}	0.317	0.279	0.418	0.393	0.438	0.437	0.442	0.339	0.490	0.359	0.434	0.370
	n($\sigma_{x\%}$)	0.181	0.172	0.295	0.302	0.415	0.401	-0.032	*-0.003*	*-0.001*	*0.010*	0.102	0.106
	SMV	0.204	0.239	0.311	0.383	0.396	0.456	0.098	0.076	0.170	0.109	0.313	0.277
	NQA-QPP	**0.781**	**0.566**	**0.783**	**0.602**	**0.603**	**0.486**	**0.615**	**0.479**	**0.644**	**0.475**	0.446	0.323
	BERTQPP	0.678	0.434	0.767	0.551	0.589	0.484	0.481	0.417	0.595	0.453	0.447	0.313
	qppBERT-PL	0.594	0.507	0.655	0.552	0.451	0.440	0.391	0.250	0.449	0.277	**0.455**	**0.383**

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Results for RQ3:
  - Unsupervised methods perform better with deeper ranked lists

		T5 + BM25						ConvDR (QPP fed with T5 query rewrites)					
		nDCG@3		nDCG@100		Recall@100		nDCG@3		nDCG@100		Recall@100	
	QPP methods	P-ρ	K-τ	P-ρ	K-τ	P-ρ	K-τ	P-ρ	K-τ	P-ρ	K-τ	P-ρ	K-τ
CAsT-19	Clarity	0.321	0.234	0.326	0.257	0.214	0.187	0.257	0.176	0.342	0.227	0.335	0.216
	WIG	0.436	0.232	**0.608**	0.429	**0.579**	0.426	0.387	0.274	0.542	0.398	0.451	0.347
	NQC	0.348	0.246	0.548	0.397	0.545	0.444	**0.431**	**0.307**	**0.647**	**0.481**	0.557	0.421
	σ_{max}	0.442	0.354	0.574	**0.433**	0.494	0.399	0.378	0.267	0.637	0.456	**0.591**	**0.441**
	n($\sigma_{x\%}$)	0.430	0.332	0.569	0.406	0.505	0.365	0.187	0.175	0.358	0.292	0.362	0.288
	SMV	0.344	0.250	0.548	0.417	0.541	**0.466**	0.386	0.285	0.619	0.471	0.556	0.423
	NQA-QPP (warm-up)	**0.538**	**0.357**	0.542	0.392	0.537	0.377	0.187	0.128	0.401	0.275	0.364	0.263
	BERTQPP (warm-up)	0.526	**0.357**	0.532	0.391	0.463	0.325	0.282	0.187	0.378	0.249	0.261	0.194
	qppBERT-PL (warm-up)	0.317	0.218	0.412	0.279	0.363	0.263	0.212	0.151	0.354	0.233	0.345	0.249
CAsT-20	Clarity	0.258	0.191	0.452	0.343	0.467	0.332	*0.126*	*0.088*	0.270	0.195	0.264	0.178
	WIG	0.248	**0.251**	**0.494**	**0.453**	**0.478**	**0.438**	**0.377**	**0.277**	**0.549**	0.389	**0.465**	0.320
	NQC	0.150	0.235	0.363	0.399	0.320	0.380	0.339	0.261	0.544	**0.404**	0.463	**0.357**
	σ_{max}	0.179	0.221	0.339	0.372	0.339	0.382	0.282	0.219	0.496	0.364	0.440	0.328
	n($\sigma_{x\%}$)	0.178	0.225	0.413	0.422	0.420	0.410	0.199	0.168	0.409	0.309	0.397	0.285
	SMV	0.139	0.219	0.362	0.400	0.333	0.387	0.275	0.216	0.503	0.380	0.454	0.352
	NQA-QPP (warm-up)	**0.274**	0.170	0.471	0.362	0.466	0.370	0.315	0.218	0.310	0.237	0.324	0.223
	BERTQPP (warm-up)	0.207	0.171	0.404	0.301	0.364	0.246	0.253	0.183	0.349	0.242	0.221	0.133
	qppBERT-PL (warm-up)	0.228	0.213	0.367	0.305	0.312	0.287	0.218	0.164	0.378	0.272	0.313	0.229

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Takeaway
  - *Previous finding 1: Supervised QPP methods outperform unsupervised ones [1-6]*
  - We found
    - Supervised ones distinctly outperform unsupervised ones only when a large amount of training data is available
    - Compared to supervised ones, Unsupervised ones show strong performance
      - In a few-shot setting
      - When predicting the retrieval quality for deeper-ranked lists

[1] Datta et al. A ‘Pointwise-Query, Listwise-Document based Query Performance Prediction Approach. In SIGR 2022.
[2] Chen et al. Groupwise Query Performance Prediction with BERT. In ECIR 2022.
[3] Datta et al. Deep-QPP: A Pairwise Interaction-based Deep Learning Model for Supervised Query Performance Prediction. In WSDM 2022.
[4] Arabzadeh et al. BERT-QPP: Contextualized Pre-trained Transformers for Query Performance Prediction. In CIKM 2021.
[5] Hashemi et al. Performance Prediction for Non-Factoid Question Answering. In ICTIR 2019.
[6] Zamani et al. Neural Query Performance Prediction Using Weak Supervision from Multiple Signals. In SIGIR 2018.

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Takeaway
  - *Previous finding 2: List-wise supervised QPP methods outperform point-wise ones [1,2]*
  - We found
    - Point-wise ones outperform list-wise ones in most cases
    - List-wise ones
      - Are more data-efficient
      - Show a slight advantage for deeper-ranked lists

[1] Datta et al. A 'Pointwise-Query, Listwise-Document based Query Performance Prediction Approach. In SIGR 2022.
[2] Chen et al. Groupwise Query Performance Prediction with BERT. In ECIR 2022.

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Takeaway
  - *Previous finding 3: Retrieval score-based unsupervised QPP methods perform badly in estimating the retrieval quality of neural-based retrievers [1,2]*
  - We found
    - Retrieval score-based methods show great effectiveness in assessing ConvDR, either for top ranks or deeper-ranked lists
    - The effectiveness of score-based methods relies on the retrieval score distribution of a specific retriever

[1] Hashemi et al. Performance Prediction for Non-Factoid Question Answering. In ICTIR 2019.
[2] Datta et al. A Relative Information Gain-based Query Performance Prediction Framework with Generated Query Variants. In TOIS 2022.

Reproducing Existing QPP Methods in CS (SIGIR 2023)

- Other takeaways
  - Feeding query rewrites into QPP methods to estimate the retrieval quality of CS methods shows great performance
  - Improve query understanding for supervised QPP methods
    - Improve query rewriting quality
    - Develop a mechanism of conversational context understanding for QPP
  - Design supervised QPP methods using few-shot learning techniques

Outline

- [x] Background
  - [x] Query Performance Prediction (QPP)
  - [x] Conversational Search (CS)
- [x] Motivation
- [x] Study 1: Reproducing existing QPP methods in CS (SIGIR 2023)
- [ ] **Study 2: Improve QPP for CS using query rewriting quality (QPP++2023)**
- [ ] Conclusion

Performance Prediction for Conversational Search Using Perplexities of Query Rewrites

Chuan Meng, Mohammad Aliannejadi and Maarten de Rijke

Got accepted at QPP++ 2023:

Query Performance Prediction and Its Evaluation in New Tasks Workshop co-located with The 45th European Conference on Information Retrieval (ECIR)

Improve QPP for CS using Query Rewriting Quality

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

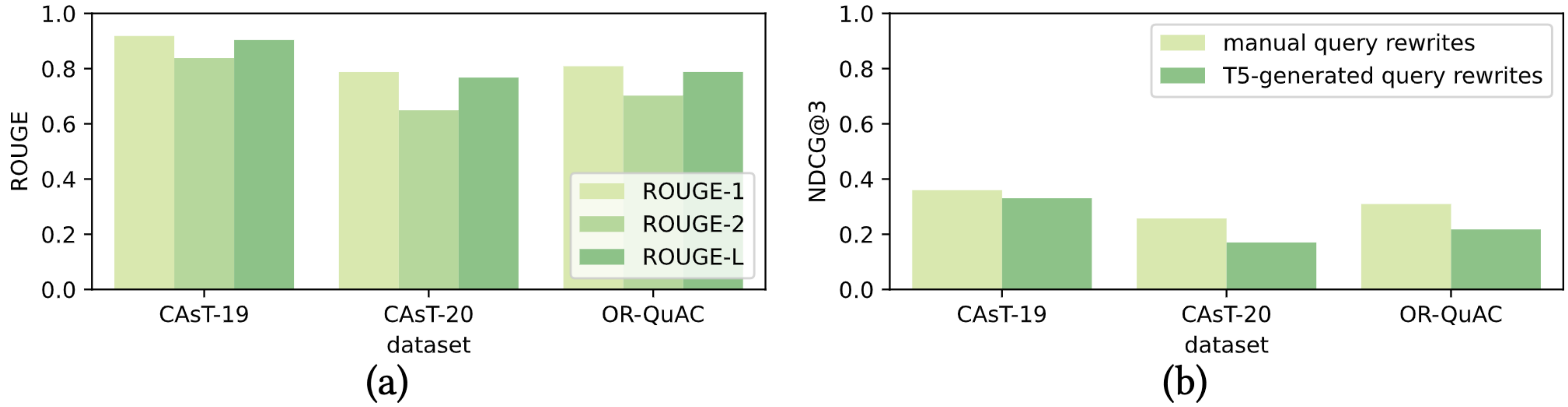


Figure 1: The similarity between manual and T5-generated query rewrites in terms of ROUGE (a) and the retrieval quality of BM25 for manual/T5-generated query rewrites in terms of NDCG@3 (b).

Improve QPP for CS using Query Rewriting Quality

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

Improve QPP for CS using Query Rewriting Quality

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

[1] Lin et al. Multi-Stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting. In TOIS 2021.
[2] https://huggingface.co/gpt2-xl

Improve QPP for CS using Query Rewriting Quality

- 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**

Improve QPP for CS using Query Rewriting Quality

- Takeaways:
  - Propose PPL-QPP that incorporates query rewriting quality into QPP methods.
  - PPL-QPP improves QPP effectiveness when 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

Outline

- [x] Background
  - [x] Query Performance Prediction (QPP)
  - [x] Conversational Search (CS)
- [x] Motivation
- [x] Study 1: Reproducing existing QPP methods in CS (SIGIR 2023)
- [x] Study 2: Improve QPP for CS using query rewriting quality (QPP++2023)
- [ ] **Conclusion**

Conclusion and Future Work

- Contributions
  - A comprehensive reproducibility study into existing ad-hoc QPP methods in CS
  - A new QPP method for CS using query rewriting quality
  - The data and code are open-sourced, https://github.com/ChuanMeng/QPP4CS

Query Performance Prediction for Conversational Search (QPP4CS)

VISITORS 213

This is the repository for the papers:

- Query Performance Prediction: From Ad-hoc to Conversational Search (SIGIR 2023)
- Performance Prediction for Conversational Search Using Perplexities of Query Rewrites (QPP++ 2023)

This repository allows the replication of all results reported in the papers. In particular, it is organized as follows:

Thanks!

Q & A

Appendix

$$Clarity(q, D^k_{q;M}, D) = \sum_{w \in V} P(w|D^k_{q;M}) \log \frac{P(w|D^k_{q;M})}{P(w|D)},$$

$$WIG(q, D^k_{q;M}, D) = \frac{1}{k} \sum_{d \in D^k_{q;M}} \frac{1}{\sqrt{|q|}} (Score(q; d) - Score(q; D)),$$

$$NQC(q, D^k_{q;M}, D) = \frac{1}{Score(q; D)} \sqrt{\frac{1}{k} \sum_{d \in D^k_{q;M}} (Score(q; d) - \mu)^2},$$

$$SMV(q, D^k_{q;M}, D) = \frac{\frac{1}{k} \sum_{d \in D^k_{q;M}} (Score(q; d) |\ln \frac{Score(q;d)}{\mu}|)}{Score(q; D)},$$

Appendix

Table 11
Frequency-based pre-retrieval QPP baselines.

Baseline	Formula	Description
SCQ	$SCQ(t) = (1 + \log(TF(t, D))). IDF(t)$	D denotes the collection. $TF(t, D)$ denotes the term frequency of term t in collection D.
IDF	$IDF(t) = \log(\frac{N}{N_t})$	N denotes the number of documents in the collection. N_t is the number of documents containing query term t.
SCS	$SCS(q) = \log(\frac{1}{\lvert q \rvert}) + avgICTF(q)$	$avgICTF(q) = \frac{1}{\lvert q \rvert} \sum_{t \in q} \log(\frac{\lvert D \rvert}{TF(t.D)})$
PMI	$PMI(t_i, t_j) = \log \frac{P_r(t_i, t_j \mid D)}{P_r(t_i \mid D) P_r(t_j \mid D)}$	$P_r(t_i, t_j \lvert D)$ denotes the probability of two terms co-occurring in the collection.
VAR	$VAR(w(t, d))$	$VAR(w(t, d))$ is the variance of term weights over documents $d \in D$ containing query term t, where : $w(t, d) = \frac{\log(1 + TF(t, d)). IDF(t)}{\lvert d \rvert}$