From Baseline to Top Performer: A Reproducibility Study of Approaches at the TREC 2021 Conversational Assistance Track

Weronika Łajewska and Krisztian Balog

University of Stavanger, Norway

ECIR'23, Dublin

Motivation and objectives

Why did we choose to reproduce TREC systems?

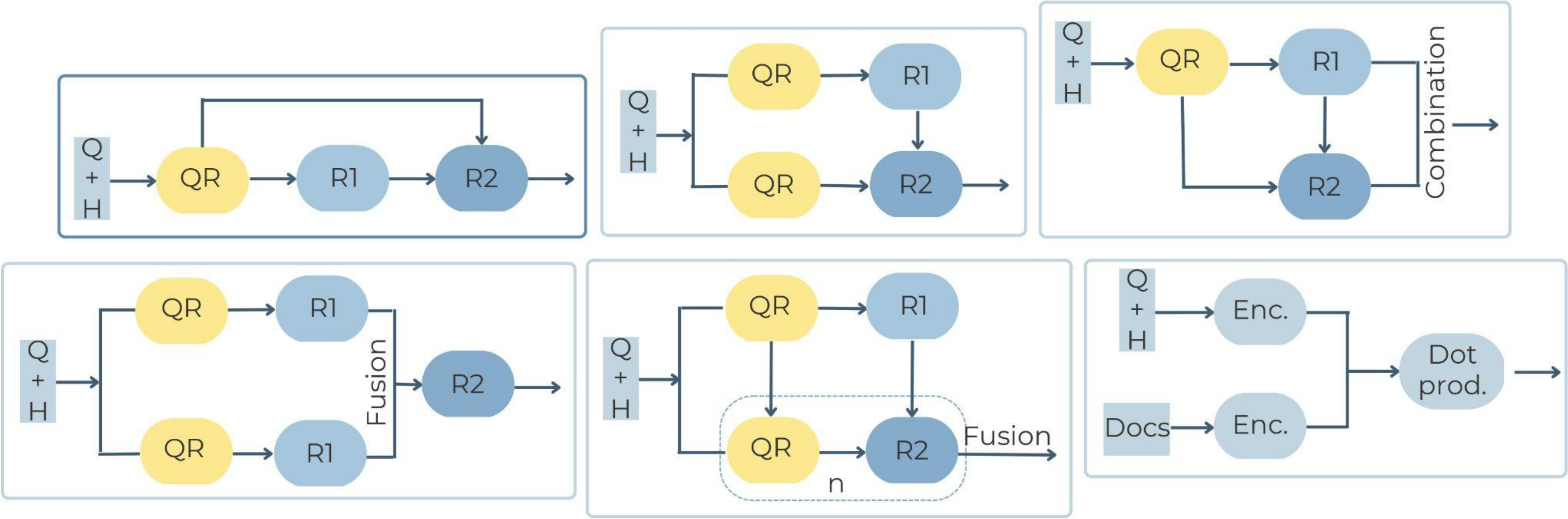
- TREC systems are reference points for effectiveness comparison
- TREC papers have less strict requirements than peer-reviewed publications

What systems did we reproduce?

- Organizers' baseline [1]
- The top performing participant submission at the 2021 edition [2]

[1] J. Dalton, C. Xiong, and J. Callan. TREC CAsT 2021: The Conversational Assistance track overview. In The Thirtieth Text REtrieval Conference Proceedings, TREC '21, 2021.
[2] X. Yan, C. L. Clarke, and N. Arabzadeh. WaterlooClarke at the TREC 2021 conversational assistant track. In The Thirtieth Text REtrieval Conference Proceedings, TREC '21, 2021.

Conversational search system architectures

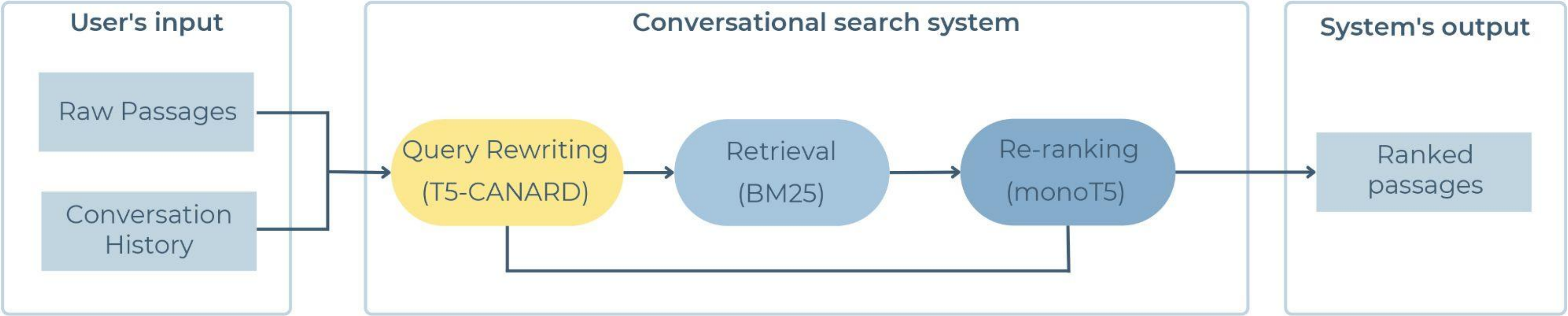


QR Query Rewriter

R1 First-pass retriever

R2 Re-ranker

Baseline system (OrganizersBaseline)



Baseline system

- Reproducibility attempted based on overview paper
- Aspects of the reproduced system modified in our implementation:
  - Context given as input to the query rewriter:

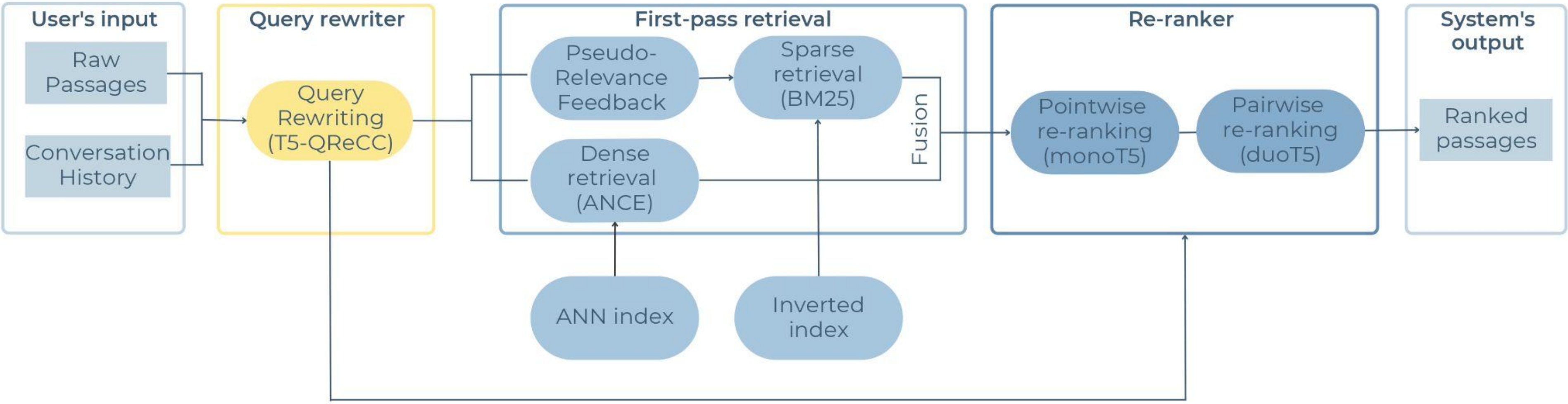
$$\hat{q}_i = Rewrite(q_1, q_2, ..., q_{i-3}, r_{i-3}, q_{i-2}, r_{i-2}, q_{i-1}, r_{i-1}, q_i)$$

$$\downarrow$$

$$\hat{q}_i = Rewrite(\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_{i-1}, trim(r_{i-1}), q_i)$$

  - Parameters in BM25 first-pass retrieval → parameters reported by the organizers: (k1=4.46, b=0.82), default parameters: (k1=1.2, b=0.75)

The state-of-the-art system (WaterlooClarke)



The state-of-the-art system

- Reproducibility attempted based on working notes paper plus communication with authors
- Aspects of the reproduced system modified in out implementation:
  - Question-answering system in the first-pass retrieval
  - Tuning of BM25 parameters
  - Implementation of PRF algorithm

Reproducibility experiments

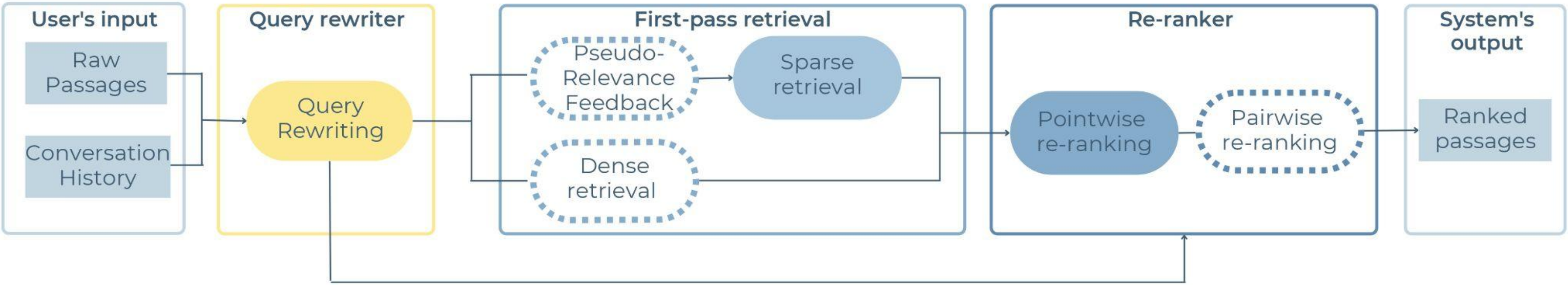
Possible reasons for discrepancies in the results:

- **BaselineOrganizers**
  -9% NDCG@3; +2% Recall@500
  - possibly different formulation of input sequences for query rewriting with regards to exceeding the length limits of the model
- **WaterlooClarke**
  -19% NDCG@3; -20% Recall@500
  - missing C4-based question-answering step performed in first-pass retrieval

Approach	R@500	NDCG@3
BaselineOrganizers@TREC'21	0.636	0.436
BaselineOrganizers	0.647	0.397
WaterlooClarke@TREC'21	0.869	0.514
WaterlooClarke reproduced by us	0.692	0.415

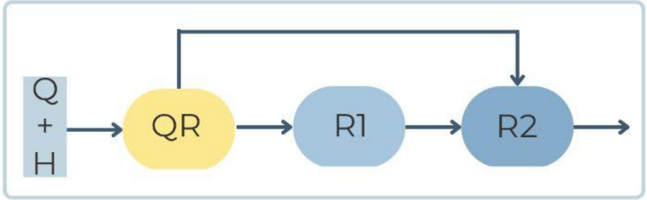
Additional experiments

Additional experiments (1)

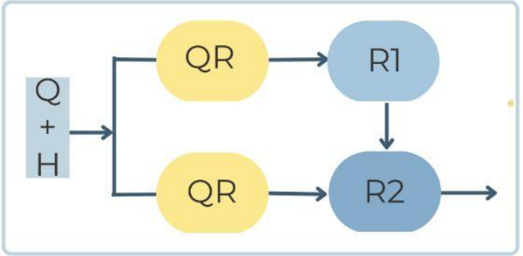


- **How specific components of the pipeline contribute to the overall performance?**
  - Adding PRF and combining sparse and dense retrieval methods for first-pass retrieval improves performance (+12%–29% in recall and +3%–12% in NDCG@3)
  - T5-CANARD used for query rewriting achieves better results than T5-QReCC (+3%–7% in recall, +1% in NDCG@3)

Additional experiments (2)



VS



- **Is impact of the query rewriting the same for both ranking steps?**
  - Using T5-CANARD for first-pass retrieval results in the higher recall
  - The overall best combination in terms of final ranking (NDCG@3) is when T5-QReCC is employed in first-pass retrieval and T5 CANARD is used in re-ranking (+6% in recall, +1% in NDCG@3 over WaterlooClarke system)

Conclusions from the reproducibility study

- Our reproducibility efforts have met with moderate success
- We have managed to come closer to reproducing the organizers' baseline than the participant's submission (-9% vs. -19% in NDCG@3 w.r.t. official results)
- Key missing information includes:
  - the names of specific algorithms and models used
  - descriptions of procedures of constructing inputs to neural models
  - methods of obtaining models' parameters

Practical suggestions for the community

- Sharing model parameters in some cases is not enough
- Details on collection preprocessing or collection statistics are needed
- Sharing intermediate results from the different pipeline components would be helpful

Thank you for your attention!

Questions?

Results and code: https://github.com/iai-group/ecir2023-reproducibility

Technical details of WaterlooClarke system

- Technical details obtained via email communication:
  - query rewriting model and its parameters
  - BM25 parameters
  - PRF parameters
  - fusion method used for sparse retrieval rankings

- Still missing information:
  - PRF algorithm
  - question-answering system employed
  - approach used for tuning the BM25 parameters
  - preprocessing employed for the inverted index
  - method used for combining sparse and dense rankings

Reproducibility results

Approach	R@500	NDCG@3
BaselineOrganizers@TREC’21	0.636	0.436
BaselineOrganizers-QR-BM25	0.563	0.346
BaselineOrganizers-BM25	0.589	0.397
BaselineOrganizers	0.647	0.397
WaterlooClarke@TREC’21	0.869	0.514
WaterlooClarke reproduced by us	0.692	0.415

Discrepancies in runfiles evaluation

Results reported in the overview paper:

Approach	R@500	NDCG@3
BaselineOrganizers@TREC'21	0.636	0.436
WaterlooClarke@TREC'21	0.869	0.514

```
{TREC_EVAL_PATH}/trec_eval trec_eval -q -c -m map -m P.1,3 -m ndcg_cut.1,3,5 -m
recip_rank -m all_trec -l2 -M500 data/qrels/{YEAR}.txt data/runs/{YEAR}/{RUNID}.trec
```

Results obtained by evaluating official runfiles:

Approach	R@500	NDCG@3
BaselineOrganizers@TREC'21 (runfile)	0.623	0.424
WaterlooClarke@TREC'21 (runfile)	0.861	0.495

Component-based analysis

Approach	TREC CAsT 2020		TREC CAsT 2021	
	R@500	NDCG@3	R@500	NDCG@3
T5 CANARD + BM25 + monoT5	0.528	0.379	0.647	0.397
T5 QReCC + BM25 + monoT5	0.510	0.362	0.602	0.393
T5 CANARD + ANCE/BM25 + mono/duoT5	0.678	0.405	0.726	0.407
T5 QReCC + ANCE/BM25 + mono/duoT5	0.645	0.406	0.680	0.416
T5 CANARD + ANCE/BM25/PRF + mono/duoT5	0.688	0.409	0.731	0.406
T5 QReCC + ANCE/BM25/PRF + mono/duoT5	0.661	0.405	0.692	0.415

Variants of a two-stage retrieval pipeline

R2	Recall	NDCG@3	Recall	NDCG@3
R1	T5 CANARD		T5 QReCC	
T5 CANARD	2020: 0.6878 2021: 0.7306	2020: 0.4086 2021: 0.4061	2020: 0.6878 2021: 0.7267	2020: 0.3923 2021: 0.4166
T5 QReCC	2020: 0.6608 2021: 0.6879	2020: 0.4086 2021: 0.4176	2020: 0.6608 2021: 0.6915	2020: 0.4052 2021: 0.4151