Improving Conversational Passage Re-ranking with View Ensemble

Jia-Huei Ju*, Sheng-Chieh Lin†, Ming-Feng Tsai‡, and Chuan-Ju Wang*

*Research Center for Information Technology Innovation, Academia Sinica
†David R. Cheriton School of Computer Science, University of Waterloo
‡Department of Computer Science, National Chengchi Univeristy

Conversational Search (ConvSearch)

Information needs in ConvSearch have an unique **multi-turn structure**:

Turns	User's utterance
u_1	What is throat cancer?
u_2	Is it treatable?
u_3	Tell me about lung cancer.
u_4	What are its symptoms?
...	...
u_7	What is the first sign of it?

In this work, a conversational query at turn i is denoted as

$$\underbrace{q_i}_{\text{Conv. Info. need}} = \{ \underbrace{u_i}_{\text{Info. need}} ; \underbrace{u_1, u_2, u_3, ..., u_{i-1}}_{\substack{\text{context} \\ \text{e.g., historical utterances}}} \}.$$

Cascaded Architecture

Cascaded architecture for ad-hoc search

The retrieval-and-rerank pipeline:

$$R = \mathcal{F}_{\mathrm{RR}}\Big(q'; p \in \underbrace{\mathcal{F}_{\mathrm{RT}}(q', p \in \mathcal{D})}_{\text{Retrieving candidate passages}}\Big).$$

where R is a (re-)ranked list of relevant passages p for a given query q'.

To fit this effective architecture to ConvSearch, **conversational query reformulation** (CQR) has been recognized as an important module.

CQR reformulates the conversational query into a **de-contextualized** ad-hoc query via T5-rewriting, HQExp, etc.

$$q_i' = \mathcal{F}_{CQR}(u_i; u_1, u_2, ...u_{i-1})$$

Cascaded Architecture for ConvSearch

Recently, conversational dense retrieval (ConvDR) has shown the great success of integrating CQR into bi-encoder models.

Cascaded architecture for ConvSearch

$$R = \mathcal{F}_{\text{ConvRerank}}\Big(q; p \in \underbrace{\mathcal{F}_{\text{ConvDR}}(q, p \in \mathcal{D})}_{\text{Retrieving candidate passages}}\Big).$$

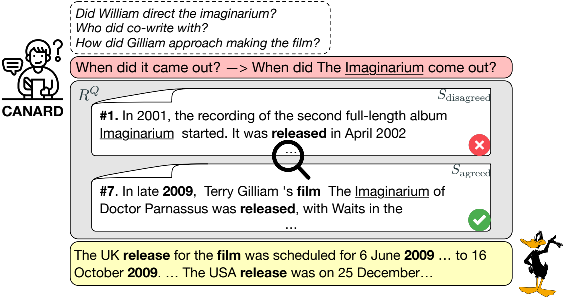
q is a raw conversational query **without** any reformulation.

Follow ConvDR's success, we want to build a **conversational passage re-ranker (ConvRerank)** for improving the top-ranking effectiveness.
$\Longrightarrow$ similar to ConvDR, perform re-ranking **without** reformulation.

Pseudo-labeling with View Ensemble

To collect the higher-quality training pairs for ConvRerank, we develop a pseudo-labeling approach with **view ensemble**.

The intuition is based on the empirical observation, for example



We hypothesize that **ground-truth answer** can provide more faithful signals of relevance.
(e.g., $\#1 \implies$ false positive vs. $\#7 \implies$ true positive)

Pseudo-labeling with View Ensemble

First, we use $\mathrm{BM25}$ and $\mathrm{monoT5}$ [4] with different query views,[1]

$$R^Q = \mathrm{monoT5}\Big(q^*; p \in \mathrm{BM25}\big(q^*; p \in \mathcal{D}\big)\Big),$$
$$R^A = \mathrm{monoT5}\Big(q^*; p \in \mathrm{BM25}\big(q^* \| a; p \in \mathcal{D}\big)\Big)$$

Second, we ensemble two ranked lists by simply pushing the **agreed** passages to the top; and down the **disagreed** passages to the bottom like

$$R^{\mathrm{EM}(R^Q|R^A)} = S_{\mathrm{agreed}} \,\|\, S_{\mathrm{disagreed}}.$$

Last, we use this re-ordered list to construct the pseudo training pairs; and fine-tune ConvRerank on this data from monoT5's checkpoint.

[1] Follow CQE paper, we use rewritten query q^* from CANARD dataset, and a refers to the ground-truth answer in QuAC dataset.

Experiments

Experiments: Full Ranking Results

Our baseline approach is setting #(e), which used the T5-rewriting model in advance of re-ranking

#	Retrieval ($\rightarrow$ Re-ranking)	Latency (ms/q)	CAsT'19 Eval nDCG@3 / 100	CAsT'20 Eval nDCG@3 / 100
	Upper-bound system w/ manual query			
	TCT-ColBERT [3] $\rightarrow$ monoT5	-	0.583 / 0.545	0.556 / 0.546
(a)	ConvDR $\rightarrow$ BERT (RRF) [7]	1900	0.541 / -	0.392 / -
(b)	CRDR [5]	1690	0.553 / -	0.381 / -
(c)	CTS+MVR† [1]	14630	**0.565** / -	- / -
(d)	CQE	-	0.492 / 0.447	0.319 / 0.350
(e)	CQE $\rightarrow$ T5-rewrite+monoT5	1910	0.549^{d} / 0.484^{d}	0.418^{d} / 0.395^{d}
(f)	CQE $\rightarrow$ ConvRerank	1675	$\mathbf{0.563}^{d}$ / $\mathbf{0.487}^{d}$	$\mathbf{0.432}^{d}$ / $\mathbf{0.456}^{de}$

$\Longrightarrow$ better top-ranking effectiveness(nDCG$\uparrow$); more efficient(latency $\downarrow$).

Experiments: Effect Analysis

Table 1: Fine-tune ConvRerank on different training data using different ranked list.

Ranked list	CAsT'19 Eval nDCG@3 / 100	CAsT'20 Eval nDCG@3 / 100
$R^{\mathrm{EM}(R^Q\|R^A)}$	$\mathbf{0.563}^{bcd}$ / $\mathbf{0.487}^{bcd}$	$\mathbf{0.432}^{bcd}$ / $\mathbf{0.456}^{bcd}$
R^Q	0.517 / 0.467	0.396 / 0.382
R^A	0.495 / 0.464	0.392 / 0.382
$R^{\mathrm{EM}(R^A\|R^Q)}$	0.519^{c} / 0.474^{bc}	0.403 / 0.389^{bc}

Table 2: Different first-stage retrieved passage candidates.

	Retrieval (→ Re-ranking)	CAsT'19 Eval nDCG@3 / 100	CAsT'20 Eval nDCG@3 / 100
Sparse	HQE [6]	0.261 / 0.308	0.164 / 0.204
	HQE → T5-rewrite + monoT5‡	0.553 / **0.519**	0.379 / 0.377
	HQE → ConvRerank‡	**0.558** / 0.511	**0.389** / **0.384**
Dense	CQE [2]	0.492 / 0.447	0.319 / 0.350
	CQE → T5-rewrite + monoT5	0.549 / 0.484	0.418 / 0.395
	CQE → ConvRerank	**0.563** / **0.487**	**0.432** / **0.456**
Hybrid	CQE-HYB [2]	0.498 / 0.494	0.330 / 0.368
	CQE-HYB → T5-rewrite + monoT5	0.556 / 0.531	**0.428** / **0.411**
	CQE-HYB → ConvRerank	**0.584** / **0.534**	0.424 / 0.410

Conclusion

Conclusion

Our conversational passage re-ranking (ConvRerank)

- uses the pseudo-labeling with the proposed view ensemble trick
- has better effectiveness and decent efficiency

Some future works include,

- Consolidating ConvDR and ConRerank (e.g., w/ co-training).
- Adopting candidate pruning (e.g., dynamically top-k candidates).
- Corpus-only data augmentation with high-quality training pairs.

References

References i

[1] V. Kumar and J. Callan. Making information seeking easier: An improved pipeline for conversational search. In *Proc. of EMNLP (Findings)*, pages 3971–3980, 2020. doi: 10.18653/v1/2020.findings-emnlp.354.

[2] S.-C. Lin, J.-H. Yang, and J. Lin. Contextualized query embeddings for conversational search. In *Proc. of EMNLP*, pages 1004–1015, 2021. doi: 10.18653/v1/2021.emnlp-main.77.

[3] S.-C. Lin, J.-H. Yang, and J. Lin. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proc. of ACL RepL4NLP workshop*, pages 163–173, Aug. 2021. doi: 10.18653/v1/2021.repl4nlp-1.17.

[4] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin. Document ranking with a pretrained sequence-to-sequence model. In *Proc. of EMNLP (Findings)*, pages 708–718, 2020. doi: 10.18653/v1/2020.findings-emnlp.63.

[5] H. Qian and Z. Dou. Explicit query rewriting for conversational dense retrieval. In *Pro. of EMNLP*, pages 4725–4737, 2022. URL https://aclanthology.org/2022.emnlp-main.311.

[6] J.-H. Yang, S.-C. Lin, C.-J. Wang, J. Lin, and M.-F. Tsai. Query and answer expansion from conversation history. In *Proc. of TREC*, 2019. URL https://trec.nist.gov/pubs/trec28/papers/CFDA_CLIP.C.pdf.

[7] S. Yu, Z. Liu, C. Xiong, T. Feng, and Z. Liu. Few-shot conversational dense retrieval. In *Proc. of SIGIR*, page 829–838, 2021. doi: 10.1145/3404835.3462856.

Thank You!

Are there any questions you'd like to ask?

Jia-Huei Ju	jhjoo@citi.sinica.edu.tw
Sheng-Chieh Lin	j587@uwaterloo.ca
Ming-Feng Tsai	mftsaics.nccu.edu.tw
Chuan-Ju Wang	cjwang@citi.sinica.edu.tw