TREC CAsT 2022 - CFDA & CLIP Lab

The Conversational Encoded Multi-stage Pipeline & Question Generation

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

Presenter: (Dylan) Jia-Huei Ju

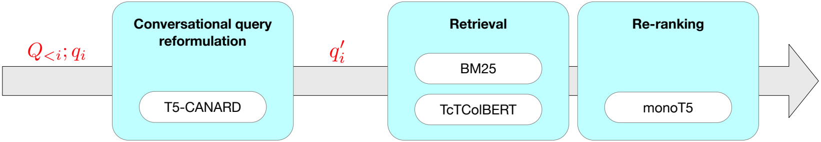
* Research Center for Information Technology Innovation, Academia Sinica,
† David R. Cheriton School of Computer Science University of Waterloo,
‡ Department of Computer Science, National Chengchi University

Outline (Main task/MI-subtask)

Preliminary

The multi-stage pipeline

The multi-stage pipeline for conversational search[1]:



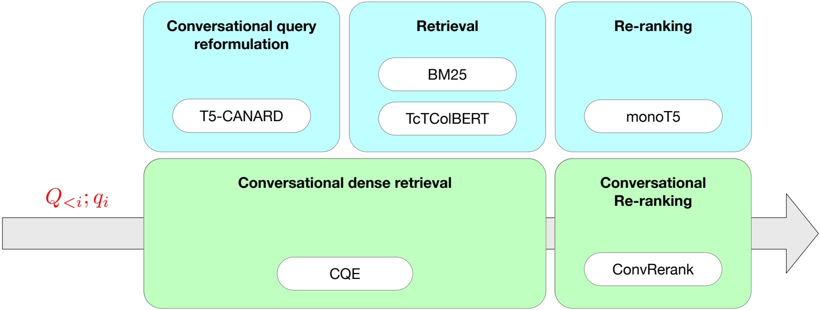
With **conversational query reformulation (CQR)**, we can thereby regard the conversational search as a standard passage retrieval task.

[1] We treated the multi-stage pipeline with CQR as our baseline.

Our Pipeline

The conversationally encoded representation

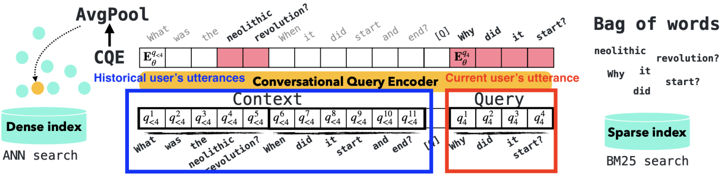
Without CQR module, we encoded the multi-turn queries $(Q_{<i} = \{q_1, q_2, ...q_{i-1}\}; q_i)$ on embedding space.



To achieve, we integrated our pipeline with conversational dense retriever and re-ranker (e.g. CQE and ConvRerank).

ConvDR: Contextualized query embeddings (CQE)

The CQE [3] approach is basically representing $Q_{<i}$ and q_i as a dense vector (using the fine-tuned conversational query encoder).



Besides CQE, we adopted CQE-hybrid[1] for top-1000 candidate passages:

- Dense: CQE (using ANN search)
- Sparse: CQE's query expansion (using BM25 search)

[1] The important tokens with greater $L2$-norm of token embeddings (see detail in the CQE paper)

ConvRerank: monoT5 with conversational query

We then predict their relevance scores using point-wise re-rankers.

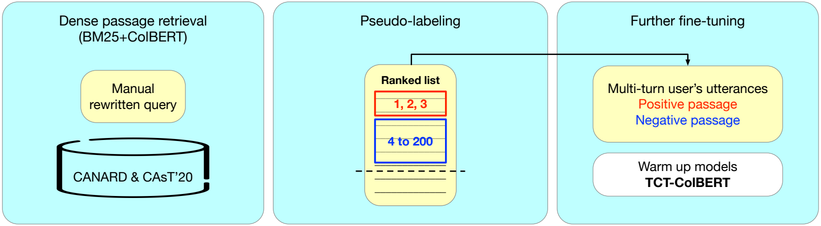
Specifically, we followed monoT5 [4] and further transform the model into a **conversational** passage re-ranker (ConvRerank) with $Q_{<i}$ and q_i.

ConvRerank's T5 text-to-text formulation
Processed input
Query: q_i Context: q_1 ||| q_2 ||| ... q_{i-1} Document: d Relevant:
Target (for training)
true/false
Processed output (for inferencing)
$P(\text{"true"})$ (from logit normalization techniques)

Fine-tuning (weakly-supervised)

Pseudo labeling of CQE [3]

We use the rewritten and multi-turn query from CANARD [2] and CAsT'20 passage collections.

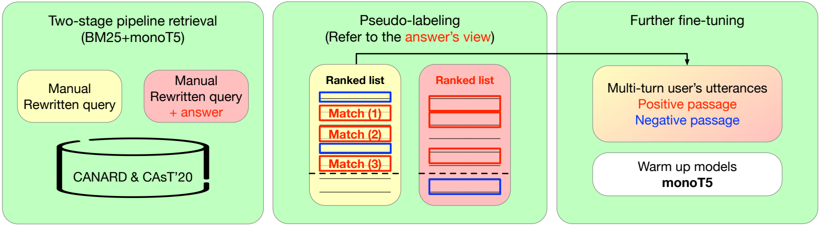


Finally, we acquired the training pairs (i.e., multi-turn query, passage):

$$(\{q_1, q_2, ..., q_{i-1}\}; q_i), p_i^+, p_i^-$$

Higher quality pseudo labeling for ConvRerank

Again, we use the rewritten and multi-turn query from CANARD [2] and CAsT'20 collections.



Finally, we acquired the training pairs (i.e., multi-turn query, passage):

$$(\{q_1, q_2, ..., q_{i-1}\}; q_i)\, ,\, p_i^+ / p_i^-\, ,$$

Results On CAsT'20

Full ranking performance

Pipeline	Rewriting	nDCG			
		3	5	500	Overall
BM25	✓	0.1464	0.1432	0.2582	0.2824
+ monoT5	✓	0.3701	0.3613	0.4067	0.4089
TctColBERT	✓	0.3381	0.3271	0.4349	0.4520
+ monoT5	✓	0.3819	0.3786	0.4801	0.4888
CQE	✗	0.3416	0.3288	0.4335	0.4532
+ monoT5	✓	0.3987	0.3876	0.4838	0.4946
+ ConvRerank	✗	0.4026	0.3973	0.4818	0.4977
CQE-hybrid	✗	0.3676	0.3506	0.4752	0.4954
+ monoT5	✓	0.3939	0.3857	0.5051	0.5196
+ ConvRerank	✗	0.4087	0.3993	0.5097	0.5273

Table 1: The settings in boldface indicate the first-stage retrieval. ✓: the queries used are rewritten by CQR module.

Ablation experiments (monoT5 wo/ rewrite)

What if we predict the relevance scores for conversational query without fine-tune a new re-ranker (ConvRerank)?

Pipeline	Query used	nDCG		
		3	5	500
BM25+monoT5	$\mathcal{F}_{CQR}(Q_{<i}; q_i)$	0.3343	0.3192	0.3913
BM25+monoT5	$(Q_{<i}; q_i)$	0.3563	0.3449	0.3926
BM25+ConvRerank	$(Q_{<i}; q_i)$	0.3777	0.3616	0.3954

For the better effectiveness, we need to fine-tune a re-ranking model for conversational query.

Ablation experiments

Why we adopted **another** pseudo-labeling ?

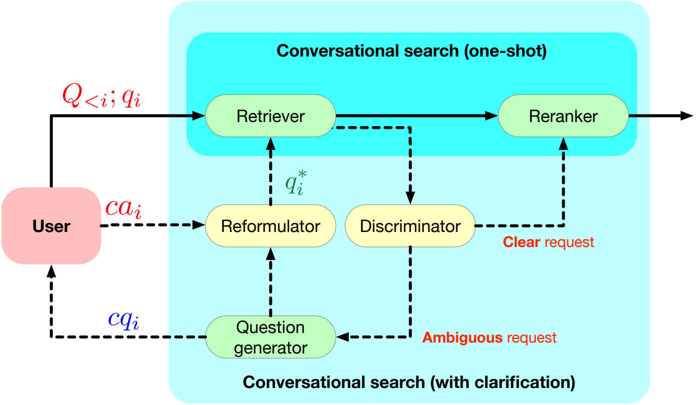
Pipeline	Pseudo-labeling	nDCG		
		3	5	500
BM25+monoT5	-	0.3343	0.3192	0.3913
BM25+ConvRerank	CQE's pseudo labels.	0.3639	0.3473	0.3859
BM25+ConvRerank	**Pseudo labels w/ answer**	0.3777	0.3616	0.3954

For better quality of positive and negative training pairs, we adopted aforementioned pseudo-labeling with answer's view.

Mixed-initiative interactions

The workflow of conversational search system

We were planing to include the MI information into our pipeline (the dotted lines)



However, so far we have only (roughly) fine-tune the question generator.

Clarification question generation (CQG)

For the mixed-initiative sub-task, we fine-tune a CQG model:

- Generative model: T5
- Dataset: ClariQ [1]
  - Initial question: q_i
  - Context: historical clarification cycle (if any), including system asked question $cq^{(j)}$ and user's feedback $ca^{(j)}$.
  - Keywords: augmented 10 words from top-30 relevant passages.
  - Clarification question: $cq_i^{(j+1)}$

CQG: T5 text-to-text formulation
Input source   Context: q_i ||| $cq_i^{(j)}$ ||| $cq_i^{(j)}$ ||| ... Keywords: $kw_1, kw_2, \ldots$ Clarifying:
Output target $cq_i^{(j)}$

Conclusion

Conclusion

For the main task,

- Open-retrieval question answering (ORConvQA [5])
  - Re-ranker with summarization
- More effective fine-tuning framework
  - Knowledge distillation (bi-encoder $\leftrightarrow$ cross-encoder)

For MI-subtask,

- Integrating modules (discriminator, generator, ...etc)

References i

[1] M. Aliannejadi, J. Kiseleva, A. Chuklin, J. Dalton, and M. Burtsev. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.367. URL https://aclanthology.org/2021.emnlp-main.367.

[2] A. Elgohary, D. Peskov, and J. Boyd-Graber. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1605. URL https://aclanthology.org/D19-1605.

[3] S.-C. Lin, J.-H. Yang, and J. Lin. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.77. URL https://aclanthology.org/2021.emnlp-main.77.

[4] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.63. URL https://aclanthology.org/2020.findings-emnlp.63.

[5] C. Qu, L. Yang, C. Chen, M. Qiu, W. B. Croft, and M. Iyyer. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 539–548, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401110. URL https://doi.org/10.1145/3397271.3401110.

Thank You!

Are there any questions you'd like to ask?

Jia-Huei Ju	jhjoo@citi.sinica.edu.tw
Ming-Feng Tsai	mftsai@nccu.edu.tw
Chuan-Ju Wang	cjwang@citi.sinica.edu.tw

Follow-up modules

As our future works, we will start with different directions.

1. **Discriminator (When to clarify)**
   - Query performance predictor.
   - Relevance scores.
2. **Question generator (What to ask)** .
   - Generating questions that can help first-stage retrieval.
3. **Conversation reformulator** .
   - Fine-tune the ConvDR (e.g. CQE) with additional clarification turns (i.e., system asked questions and the feedbacks).
   - Dialogue summarization.