RAC: Retrieval-augmented Conversation Dataset for Question Answering in Conversational Settings

Bonggeun Choi[1] Jeongjae Park[1] Yoonsung Kim[2] Jae-Hyun Park[2] and **Youngjoong Ko[3]***

[1]Department of Artificial Intelligence Sungkyunkwan University, Republic of Korea

[2]NCSOFT

[3]Department of Computer Science and Engineering Sungkyunkwan University, Republic of Korea

Introduction

- Conversational search and retrieval-augmented generation (RAG) have obtained substantial attention for their capacity to address two key challenges: query rewriting within conversational histories for the better retrieval and generating responses by employing retrieved knowledge.
- However, both fields are often independently studied so that comprehensive study on entire system remains underexplored due to the lack of datsaets covering overall processes.
- We introduce a novel **retrieval-augmented conversation (RAC) dataset** comprising conversations containing human-like responses with referenced knowledge as well as passage collections for retrieval purposes.
- We also provide **a baseline system** that consists of query rewriting, retrieval, reranking, and response generation with experimental results.

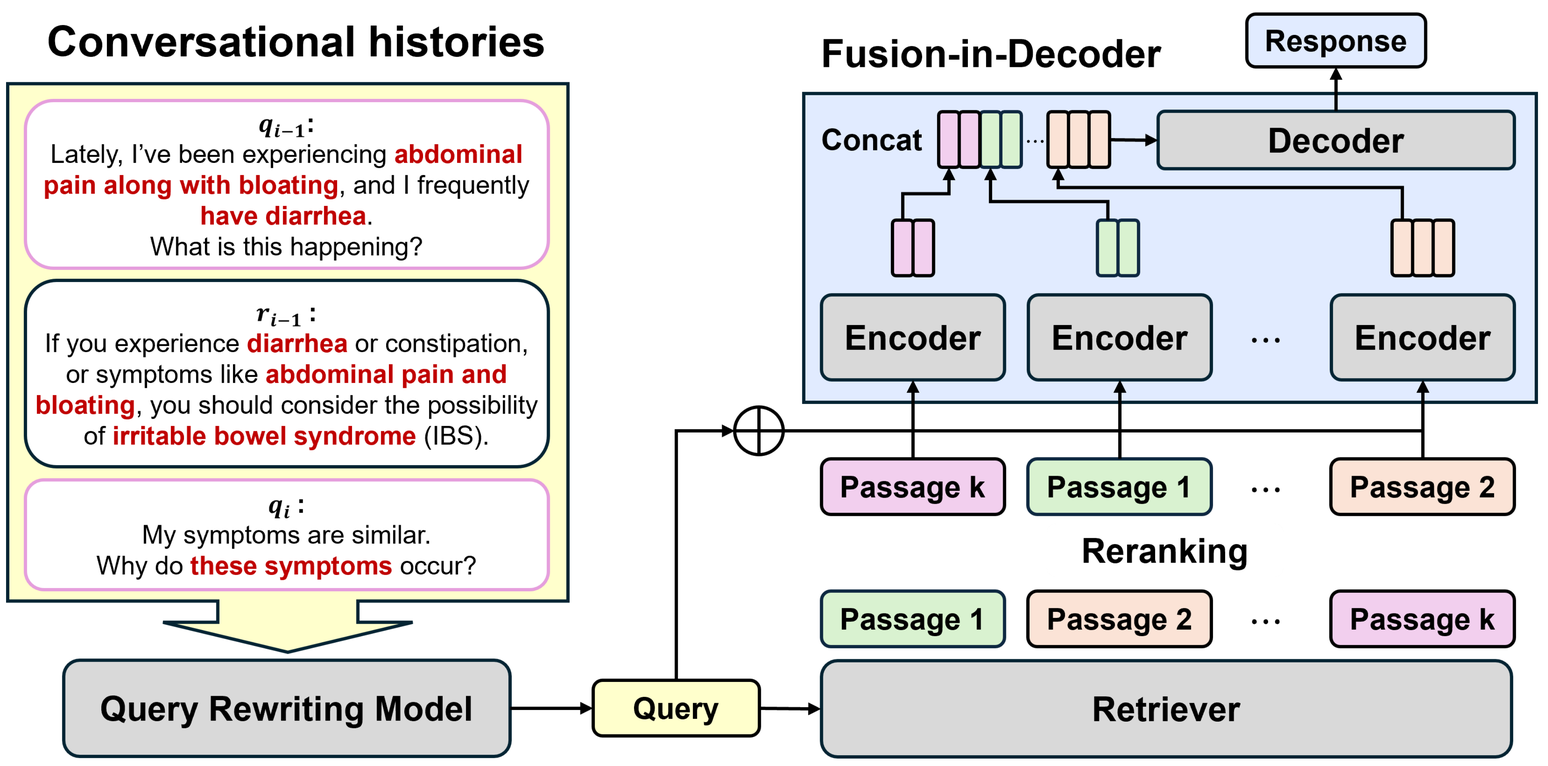
Data Construction

- Based on *Knowledge-retrieval conversation* dataset from AI-hub, a prominent Korean data platform, we construct a new dataset by supplementing the limitations of the basic datset.
- **[Passage collections]** For the retrieval purpose, we crawled whole Korean Wikipedia pages and about 1M publicly opened news data over 20 years. The crawled data were then chunked into passages of fixed length, resulting in 1,345,209 passages.
- **[Human-written query]** As colloquial questions often do not suit for retrieval purposes, proper queries are needed to deal with the query rewriting aspect of RAC. Consequently, 10,266 queries were written by human annotators.
- **[Relevant passage annotation]** To provide several relevant passages like real-world scenarios, we retrieved passages with the human-written queries and annotated relevance to top-5 retrieved passages.

Baseline System

- Query rewriting model is trained by typical teacher-forcing method using a encoder-decoder model. We utilize passages for inputs as well as questions to make the model generate tokens more related to relevant passages.
- The query rewriting model is also used for the reranking phase. By inputting retrieved passages from the first-stage retrieval, the model outputs probabilities of generating a query used for the retrieval. Then, the passages are reranked by the probabilities.
- BM25 and DPR models are adopted for the retrieval, and both Fusion-in-Decoder (FiD) and GPT-4o-mini are used for response generation.

<The overview of retrieval-augmented conversation system>



Experiments: Retrieval

Retriever	Stage	Metrics				
		MRR	Recall@5	MAP@5	NDCG@5	Hit@5
DPR	First-stage	0.272	0.213	0.143	0.192	0.382
	+Reranking	0.439	0.393	0.293	0.359	0.575
BM25	First-stage	0.332	0.272	0.192	0.249	0.460
	+Reranking	**0.453**	**0.414**	**0.310**	**0.377**	**0.595**
	Human-written	0.512	0.436	0.322	0.404	0.681

Experiments: Ablation Study

Retriever	Stage	Metrics			
		MRR	Recall@5	MAP@5	NDCG@5
BM25	Query rewriting	0.332	0.272	0.192	0.249
	- Passage learning	0.310	0.265	0.186	0.239

Experiments: Response Generation

Retriever	Generator	Stage	Metrics		
			ROUGE-L	BLEU	METEOR
Dense	FiD	First-stage	0.076	0.054	0.221
		+Reranking	0.101	**0.066**	0.244
BM25	FiD	First-stage	0.083	0.059	0.228
		+Reranking	0.102	0.065	0.241
	LLM	First-stage	0.134	0.056	0.309
		+Reranking	**0.154**	0.062	**0.324**

Human Evaluation

# Relevant	Rel.	Partial rel.	Partial irrel.	Irrel.
0	16.2%	30.7%	27.6%	25.5%
1	22.2%	26.7%	36.7%	14.4%
2	30.4%	22.4%	35.1%	12.1%
3	34.6%	22.6%	32.1%	10.7%
4	19.3%	29.8%	45.6%	5.3%
5	33.3%	8.3%	41.7%	16.7%
Total	22.3%	27.2%	32.7%	17.8%

Conclusion

- In this work, we **defined retrieval-augmented conversation (RAC)** task, **presented dataset** satisfying the all requirements of RAC, and finally **built the baseline system**.
- We employed both BM25 and DPR models to provide baseline performance of the retrieval. In addition, we used FiD and LLM for the response generation, comparing the results from the both models.
- Our empirical experiments and analyses discover the challenges of RAC and enlighten the future direction of developing the entire RAC system.