Qeios

Peer Review

Review of: "MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems"

Jian-Yun Nie[1]

1. Université de Montréal, Canada

The paper proposes a new benchmark mtRAG for multi-turn conversational RAG. It contains 4 datasets of different domains. The conversations are generated by human annotators in interaction with a RAG system. The annotators can modify the answers generated by the RAG system. In addition, another dataset is generated automatically by an LLM.

Experiments are carried out to test the ability of different retrievers to find relevant document given a query, which can be the original turn, its rewrite, and be provided with part or all the conversation history. To generate an answer, LLMs are provided at input the question, the reference (relevant) passages, as well as the retrieved passages. The experimental results show that LLMs still cannot deal with conversational RAG very well.

The dataset is interesting, and can complement the existing ones on some aspect, namely the manual repairs of the answers. The dataset also covers different types of questions and domains.

There are also several weaknesses.

1. While the dataset is new, it is unclear how it provides new possibilities for experiments, compared to the existing ones. There are multiple datasets for conversational IR, such as TopiOCQA and QReCC, which are datasets for conversational IR. The mtRAG dataset offers basically the same type of data. What mtRAG offers can also be done with the existing datasets (at least to a large part). By the way, it would be good to cite the datasets TopiOCQA and QReCC, which are widely used in conversational IR.

doi.org/10.32388/IG5ABM

2. The experimental results are mixed. In the experiments on retrieval, it is shown that adding conversation history does not help. The best strategy is to use the last turn as query (i.e. without previous turns). This result may be strongly related to the conversations generated in the dataset. It is mentioned that there are only a few turns that refer to previous turns - on average 1.3 turns per conversation. This is few and far less than that in the existing datasets (e.g. QReCC). In such a situation, it is obvious that for most cases, the past conversation history will add noise to the query. This raises the questions: To what extent, the dataset is truly conversational, i.e. there may be dependence among different turns in a conversation? How is it different from a traditional (non-conversational) dataset for IR?
3. The experimental results on answer generation are also mixed. It is observed that the generation based on the reference passages (the one selected or repaired by annotators) produces the best results, followed by reference+retrieved, and finally, retrieved (i.e. full RAG). It is also observed that adding a judgment on whether the question is answerable (IDK) does not always help the generation (Table 13) - the effect of providing IDK information produces mixed results. This set of results on answer generation do not show a clear trend. It is unclear if detecting IDK is useful for answer generation. The results with different LLMs also show different trends.

   Also of note, the observation that reference > reference+RAG > RAG may be problematic, because the three methods are not applied to the same set of questions. Therefore, the results should not be directly compared.
4. The additional synthetic dataset mtRAG-s is larger than the manually curated dataset. However, there is a question of quality control. It is difficult to be convinced that the synthetic dataset is reliable for experiments, despite the fact that some sort of correlation is observed between the manual and synthetic datasets.
5. The observation that the answers for the first turn are better than for later turns may be related to the intrinsic difficulty of the questions. It is possible that the annotators start with a question that is easier to answer than for later turns. This observation does not necessarily imply that “answering questions in a multi-turn setting is more challenging”.
6. The agreement among annotators is calculated. However, the way that the agreement is computed is quite particular. Why don't you use more common agreement measures such as Pearson coefficient? The notation of {F: 89.6, A: 92.1, N: 95.6, C: 90.3, WR: 87.4} is not explained.

Overall, the paper describes a new dataset that can potentially be used for experiments in conversational RAG. However, the experimental results with the retrievers and LLMs do not demonstrate that the dataset can be used to help design better approaches. The results are so mixed that it is difficult to see that a specifically designed approach can perform better. The results also raise the question about the nature of conversations in the dataset, i.e. whether later turns depend on the previous turns. It seems that the dependency is weak, which may reduce the value of the dataset.

Declarations

Potential competing interests: No potential competing interests to declare.