WaterlooClarke at the Trec 2020 Conversational Assistant Track

Negar Arabzadeh , Charles L. A. Clarke

David R. Cheriton School of Computer Science, University of Waterloo

1 Introduction

This report describes the methodology and results of the runs submitted by the WaterlooClarke group to TREC Conversational Assistant Track (CAsT) 2020. Our runs this year were based solely on the raw utterances. We did not submit any runs using the manually rewritten utterances or canonical response. All in all, our team submitted the four following runs :

1. `WatACBase`
2. `WatACBaseRe`
3. `WatACGPT2Re`
4. `WatACReAll`

The overall approach is based on last year's approach [1]: 1) Refining the query, 2) retrieving the passages and 3) reranking the passages. We did not apply the reranking step for the `WatACBase` run. Compared to last year, we tried to improve our performance by: 1) expanding the pool of the retrieved documents by merging the retrieved documents from two query variations and 2) re-ranking the passages with Bert [2]. Based on preliminary experiments on the TREC CAsT 2019 data set, employing these two approaches showed statistically significant improvement in performance. In the following, we will explain the details of our methodology and discuss the results.

2 Methodology

We created a combined pool of retrieved documents using two distinct query variants; Base and GPT2, which we will explain below. We used the first query variant for `WatACBase`, `WatACBaseRe` and `WatACReAll`. We used the second query variant for `WatACGPT2Re` and `WatACReAll` .

2.1 Query Generation

Base Query Variant: This query variant was based on our query construction approach from last year[1] . Similar to last year, we first filtered terms appearing in our track-specific stopword lists. Then, we employed a simple, yet surprisingly well-performing, trick to maintain the topic of the conversation. We prepend the first utterance in each conversation to all of the utterances.

Table 1: An example of Base Query Refinement and GPT2 Query Refinement

#	Raw utterance	Base query refinement:	GPT2 Query refinement:	Manually Rewritten Queries
82-1	I would like to learn about GMO Food labeling.	I would like to learn about GMO Food labeling.	I would like to learn about GMO Food labeling.	I would like to learn about GMO Food labeling.
82-2	What are the pros and cons?	**I would like to learn about GMO Food labeling.** What are the pros and cons ? **GMO labeling.**	What are the pros and cons of GMO Food labeling?	What are the pros and cons of GMO food labeling?
82-3	And what about the cons?	**I would like to learn about GMO Food labeling.** And what about the cons ? **GMO labeling.**	What about the GMO Food labeling pros and cons?	What are the cons of GMO food labeling?
82-4	What are the EU rules?	**I would like to learn about GMO Food labeling.** What are the EU rules ? **GMO labeling.**	What are the EU rules on GMO food labeling?	What are the GMO food labeling rules in the EU?"
82-5	Tell me more about traceability tools.	**I would like to learn about GMO Food labeling.** traceability tools . **traceability tools.**	Tell me more about traceability tools.	Tell me about traceability tools for GMO foods in the EU.

In addition, this year we appended the sentence *"Tell me more about it"* to the end of each utterance. We then applied AllenNLP coreference resolution to each conversation to the point of each utterance.

GPT2 Query Variant: We used the automatically rewritten queries based on GPT-2 model trained on TREC CAst 2019 data as an alternative to the Base Query Variant. While these automatic rewritten utterances work amazingly well most of the time, but there are some cases where they still do not maintain the thread of conversation. We determined the Base query variant a suitable complement for this variant.

Table 1 illustrates a few examples of the two query generation methods applied to a set of raw queries. As shown, appending the first utterance to the rest of the utterances in the conversation is helpful for keeping track of the topic. For example in utterance $\#82-2$, the raw query "What are the pros and cons?" has been modified to "I would like to learn about GMO Food labeling. What are the pros and cons? GMO labeling." In other words, the Base Query Variant helped to find out what is the topic for which we are seeking pros and cons. While the GPT2-rewriter also addressed this problem, in some cases such as utterance $\#82-5$, the Base Query refinement provides more details and is closer to the manually rewritten version of the queries. One of the limitations of the Base Query Variant appears when the main topic changes in a conversation. In those cases, appending the initial utterance to other queries, will not help and may have harmful effects on the performance. We hope to tackle this problem for next year's track.

2.2 Passage Retrieval

n order to retrieve the document at the very first stage, we utilized BM25 ranking with pseudo relevance feedback with the same exact experimental setup as in last year's experiments [1]. The second column in Table 2 explains the query variations we utilized for each of the runs. For `WatACReAll` run, we retrieved documents for both variations of query and then merged the pool of documents retrieved by both together. On the TREC CAsT 2019 dataset, broadening the pool of the retrieved documents led to a significant improvement in the performance.

2.3 Passage Re-ranking

We applied passage re-ranking with Bert [2] on all the runs except `WatACBase`. In each of the re-ranked runs, we re-ranked the pool of the documents retrieved by different query variants by Pygaggle library[1]. Based on recent literature and our experiments on TREC CAst 2019 data, the

[1] `https://github.com/castorini/pygaggle`

Table 2: Comparing the results of our four submitted runs

Run	Query Refinement method	ndcg@5	map@5	$\Delta ndcg@5$
`WatACBase`	Base	0.1547	0.0300	
`WatACBaseRe`	Base	0.2913	0.0680	+0.1366
`WatACGPT2Re`	GPT-2	0.3161	0.0681	+0.1614
`WatACReAll`	Base + GPT2	0.3265	0.0720	+0.1718

re-ranking has a significant improvement on the performance.

3 Results

Table 2 compares our four different runs based on ndcg@5 and map@5. In addition, we investigated whether the improvement made in each of the runs is statistically significant. As it is demonstrated in Table 2, the proposed reranking method in [2], showed a statistically significant improvement on `WatACBaseRe` compared to `WatACBase`. More interestingly, map@5 did not have statistically significant improvement when using GPT2-writer compared to Base Query refinement (`WatACGPT2Re` Vs `WatACBaseRe`).

4 Conclusion

In TREC CAst 2020, not only we made re-ranking work, but also we ran experiments on two different query variations and compared their performance along with combining them. In future, we will consider generating different query variations to expand the pool of the retrieved documents. Moreover, detecting topic changes in the conversations will be helpful to modify our proposed Base Query Refinement methods. We look forward to participating in TREC CAst 2021.

References

[1] Clarke, C. L. Waterlooclarke at the trec 2019 conversational assistant track. In *TREC* (2019).

[2] Nogueira, R. & Cho, K. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085* (2019).