TREC 2020 Conversational Assistance : Query Reformulation

Group 5

Fatema Tuz Zohora
ft.zohora@stud.uis.no

Daniel Grønner
da.gronner@stud.uis.no

Dipanjan Banik
d.banik@stud.uis.no

ABSTRACT

TREC 2020 Conversational Assistance provides us with a series of con- conversational utterances. We have to retrieve the most relevant passages for each user's utterances from the MS MARCO and TREC CAR corpus collection. We are focusing on the reformulation of the raw utterance of the queries in each conversation turns to relieve them from various context dependencies and hence, more relevant retrieval of the passages. The challenges we handled included context dependence, co-reference to previous queries, ellipsis, conversational turn shift, etc. We have implemented the T5 model trained with the CANARD dataset for query reformulation, which was our primary focus. Elasticsearch was implemented to create an inverted index with a BM25 score for fast and relevant retrieval of the passages. We measured our system on the metrics of MAP@1000, Precision@1000 and NDCG@1000 scores. The result of our method outperformed the baseline.

KEYWORDS

query reformulation, T5 model, CANARD, elasticsearch, conversational information retrieval, context dependency, machine learning

1 INTRODUCTION

The goal of the Text Retrieval Conference, also known as TREC, is to promote research within the information retrieval community by offering the infrastructure required for conducting comprehensive evaluations of text retrieval methods [11]. TREC is divided into several tracks and specialized fields of study in which particular retrieval tasks are outlined. One of those specialized tracks is CAsT. This track aims to create a reusable benchmark for open-domain conversational search where answers are retrieved passages from a large text corpus and also pursue research on Conversational Information Seeking (CIS). CIS has recently gained popularity due to the rise in popularity of conversational assistant systems and recent developments in automatic speech recognition. Conversational assistant systems help users with various tasks, including checking the weather forecast, managing music streaming services, and conducting e-commerce transactions. These implementations are utilized in chatbots, intelligent home devices, wearable devices, and smartphones [9].

Given the text understanding challenges in the task, one significant development was applying generative query models and ranking models using pre-trained neural language models. The corpus consists of two standard TREC collections: MS MARCO Ranking passages and Wikipedia (TREC CAR). The task for CAsT 2020 is relatively similar to that of 2019; the corpora are updated in 2021 and 2022. Conversational Information Seeking is the sequential interaction between one or more users and an information system [6]. As most conversational interactions are conducted with natural language dialogue, most conversational assistants are good at simple, well-defined actions, but their information-seeking ability is limited.

On the other hand, CIS can perform in various forms like Conversational Search, Conversational Recommendation, Conversational Question Answering, etc. [14]. Among these fields, we are going to focus on Conversational Question Answering, where answers are retrieved ranked paragraphs.

In conversational question answering, there can be more than one turn for each topic in the raw utterances on which a system retrieves relevant answers. On each turn, the users might introduce a context dependence to the previous query. Additionally, the utterances with human interaction contain co-reference pronouns, subject replacement by the objects, conversational turn shifts, ellipsis, etc., easily detectable by humans. However, understanding these dependencies is difficult for a system and can result in poor performance. As a result, query reformulation is required to make these context-dependent queries to be self-contained for answering more relevantly and accurately.

In query reformulation, firstly, the importance and relation of each word in the current context of the query are evaluated, then the relation of each word to the contexts of previous turns is determined and evaluated to find the dependencies. As the volume of comparison and decontextualization increases in each turn, the challenges increases because of the subtlety and vague shift in context. In this project, we will focus on reformulating the raw utterances of the conversational queries in CAsT so that we get better results on information retrieval than using the raw utterances with its initial context dependencies.

2 PROBLEM STATEMENT

For the task in our hand, we have data sets with a conversation average of eight utterances and raw responses extracted from various sources like TREC CAR, MS MARCO Ranking Passage etc. The utterances could be linked based on the previous turns and reformulated to retrieve information most relevant to that turn. As a result, we can focus more on co-referencing these turns of utterances, or our primary focus can be a better retrieving mechanism. We would focus more on reformulation of the utterances so that we can retrieve more accurate information from our data sets.

Our overall task would be to reformulate the utterances, index the documents, retrieve the output of each utterance after reformulating, and re-ranking the documents based on the relevance to that query and repeat these steps for each topic.

2.1 Data Set

CAsT-20 corpus contains passages collected from TREC Complex Answer Retrieval(CAR) and Microsoft Machine Reading Comprehension(MARCO) datasets. The corpus contains 38,636,520 passages, 25 information needs(topics), with an average of 8.6 utterance length, and a total of 256 turns. The queries recorded contain ellipsis, implied context, mild topic shifts, and other context dependence. Finally, this corpus provides relevance assessment based on BM25 and BERT re-ranking for the evaluation.

Title: Environmental Cost of Food Production
Description: Discusses about environmental cost of food production, being vegetarian and their proteins sources

Turn	Conversational Utterance
1	What is the environmental cost of food production?
2	Oh that much water? How is that for meat?
3	How much less is used for vegetables?
4	What are the benefits of being vegetarian?
5	Are there any health concerns?
6	What are their sources of proteins?
7	Oh almonds? Can you show me recipes with it?
8	How do you make the flour?

Table 1: CAsT 2020 Topic 98.

The dataset presented at 1 shows how each turn evolves as the conversation progresses. These topics are related, and humans can easily detect the similarity, but not the machines. We can observe the shift into a subtopic in this case in turn 4. Machines can better understand with clear context. This can be achieved by reformulating the utterances properly. The best retrieval score of this system is if queried with the **manually rewritten utterances**. Our goal is to reach as close as possible to that score.

3 BASELINE METHOD

The baseline evaluation provided [7] for this system is based on the quality of the retrieved result. To make our system retrieve results on queries, we first index the documents to be retrieved and implemented the indexing through elasticsearch. We choose BM25 to score our baseline retrieval as it is the default scorer for this probabilistic retrieval model [8] like elasticsearch, and as our focus is not towards better re-ranking mechanism, we used which is most suitable for our retrieval model.

In the project description, the queries are presented in their raw utterances, automatically rewritten utterances, and manually rewritten utterances. As our baseline for query reformulating, we have selected the raw utterances for querying the indexed documents and to evaluate the quality of scores received on these queries, we are using queries with manual utterances to evaluate our reformulations. However, we only retrieved the top 1000 documents based on our queries. As a result, as the conversation progresses, the score of the retrieved documents also gets worse with the raw utterances presented as it is in each turn. This is why our group is focusing on reformulating the queries so that we can receive a much better score on retrieval with these queries, even when it progresses in each turn, and be consistent.

On the other hand, relevancy ranking is the process of sorting the document results so that those documents which are most likely to be relevant to the query are shown at the top. The higher the score, the more relevancy it has to our intended query. However, as our main target is to reformulate the queries for better retrieval at each turn of the conversation, we have not set our focus on re-ranking the documents on this baseline implementation.

To evaluate the baseline, we used the python library "trectools" [12]. The image below shows the average NDCG@1000 for all the topics at each turn.

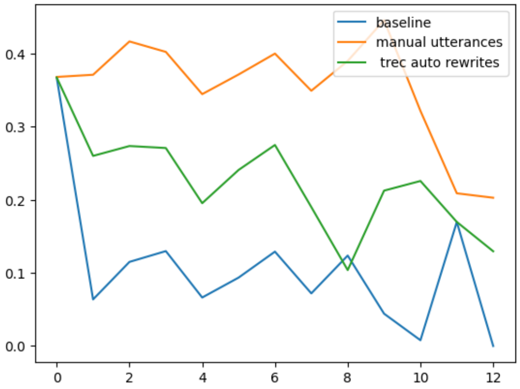


Figure 1: Average NDCG@1000 per topic

The table below shows a sample of the mean average precision(MAP)

Run	MAP@1000	Precision@1000	NDCG@1000
Baseline	0.042	0.009	0.123
Manually	0.149	0.020	0.365
Trec auto	0.085	0.014	0.241

4 ADVANCED METHOD

As the advanced method for reformulating the queries, we have chosen the T5-Model pre-trained with the Canard dataset. The implementation was inspired by the *Castorini/t5-base-canard* from the documentation of the *Hugging Face*, a community working on improving the Information Retrieval systems. As we observed T5 model is fine tuned to the CANARD dataset for query reformulation, this was our selected setup for evaluating the reformulation of the raw utterances of the CAst queries.

After reformulating the queries, we retrieved the top 1000 relevant documents using Elasticsearch. The results of this implementation are coherent with our expectations. The models we have selected for our implementation are briefly described below.

4.1 T5 Model

The T5 model or Text-to-Text Transfer Transformer model is a supervised neural model for handling different challenges of information retrieval with natural languages. As this is a transfer learning model, T5 is trained on data-rich tasks with C4 corpus

(Colossal Clean Crawled Corpus) [13] and can be fine-tuned for a definite Text-to-Text transfer task. As the authors proposed in [13], T5 is a unified model for performing various natural language processing on the input text. The output is also a text, as shown in figure 2.

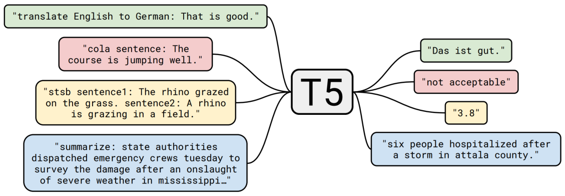
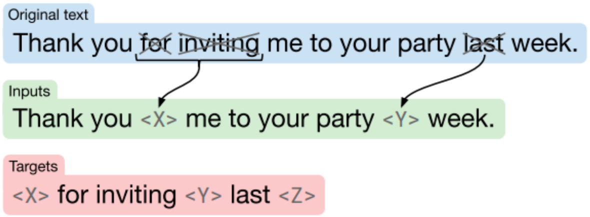

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 2: Text-to-Text Transfer Trasformer Model

In essence, T5 is an encoder-decoder model for performing specified tasks as depicted in figure 3. For this purpose, T5, like the BERT model, uses the Masked Language Model. However, BERT uses mask separately for each word, and T5 uses single mask keyword to replace multiple consecutive tokens [10].


Original text
Thank you ~~for inviting~~ me to your party ~~last~~ week.
Inputs
Thank you <X> me to your party <Y> week.
Targets
<X> for inviting <Y> last <Z>

T5's mask language modeling (Raffel et al., 2019)

Figure 3: T5 Encoding-Decoding

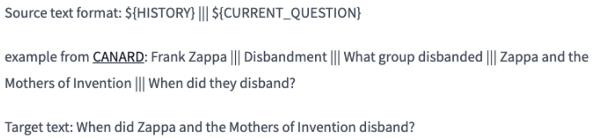
For our query reformulation purpose, we fine-tuned the model by training it with the castorini-t5-based-canard dataset. We performed a supervised training with T5Tokenizer and T5ForConditionalGeneration on this to generate the desired reformulated queries.

4.2 CANARD dataset

CANARD is a crowdsourced dataset for question-in-context rewriting, constructed based on QuAc [5], which worked up the answers from sections in a Wikipedia article. The dataset was built up with 40,527 questions with different context lengths. It consists of questions given in a dialouge utterances that precede the question. This dataset is suitable for evaluating query rewriting models that handle natural language phenomena like coreference and ellipsis resolution.

4.3 Rewrite - Castorini/T5 Model with CANARD dataset

Castorini-T5-based-CANARD model is specifically trained for conversational question rewriting [4]. The model is trained to resolve context dependent queries in the following way: As we can see from figure 4, this model is using historical information and the current information and passing them to the T5 transformer and uses CANARD dataset examples to resolve those co-dependence to generate a context independent query as the output.


Source text format: ${HISTORY} ||| ${CURRENT_QUESTION}

example from CANARD: Frank Zappa ||| Disbandment ||| What group disbanded ||| Zappa and the Mothers of Invention ||| When did they disband?

Target text: When did Zappa and the Mothers of Invention disband?

Figure 4: castorini-t5-based-canard training.

4.4 Retrieval - ElasticSearch(BM25)

Elasticsearch is a distributed, open, and free search and analytics engine for all data kinds, including textual, numerical, geographic, structured, and unstructured data [2]. Elasticsearch is built with Apache Lucene [2]. Elasticsearch is capable of searching any document. Mainly, Elasticsearch is utilized for indexing the data collections. An index can be viewed as an efficient collection of documents, with each document consisting of a set of fields, which are the key-value pairs contained in our dataset. Elasticsearch is so quick because it can utilize per-field data structures to build and return search results [1]. Elasticsearch indexes all data in every field by default, and each field has its own optimal data structure. The inverted index takes the passage and tokenizes each word according to the original document. The inverted index is the primary factor contributing to Elasticsearch's robustness and speed in its search capabilities [1].

In the retrieval phase, we search the Document corpus to get the top 1000 results using the BM25 information retrieval method. BM-25 is a ranking function that calculates a score to represent a document's relevance with respect to the query [3]. We used Elasticsearch for information retrieval since this software deploys the BM25 algorithm and is scalable for a large number of records.

4.5 Methodology

We have passages from CAsT 2020 dataset (MS-MARCO and TREC CAR). The CAsT 2020 also contains evaluation dialogues. The dialogues are provided with topics, and each topic has multiple turns. The file consists of turn number, raw_utterance, manual_rewritten_utterance, and automatic_rewritten_utterance [7]. We used the T5 model [10] to reformulate our queries and Elasticsearch to create inverted index on the MS-MARCO and TREC CAR corpus. Next, we used our trained T5 models to reformulate queries using raw utterances from the file in [7]. For passage retrieval, we utilized a standard pipeline that consists of initial retrieval followed by BM25 scoring and obtained the top 1000 passages returned by initial retrieval. Finally, We have used TREC TOOLS [12] to evaluate our results. Our methodology and workflow are described in figure 5.

5 RESULTS

To evaluate our method, we have looked at different metrics. We have focused on MAP@1000, precision@1000, and NDCG@1000. The main focus of this project was to reformulate the conversational queries, and as mentioned in the baseline section, we only used

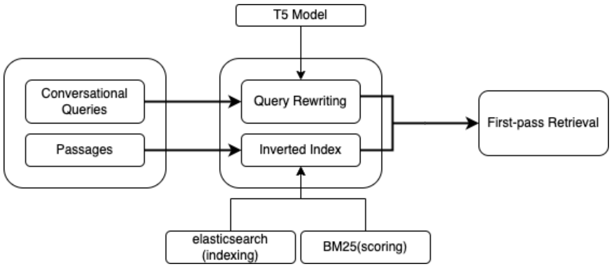


Figure 5: TREC query reformulation.

BM25 as a first retrieval. We used the python library "trectools" to calculate the different metrics [12].

The image below shows the mean NDCG@1000 for all the topics at each turn.

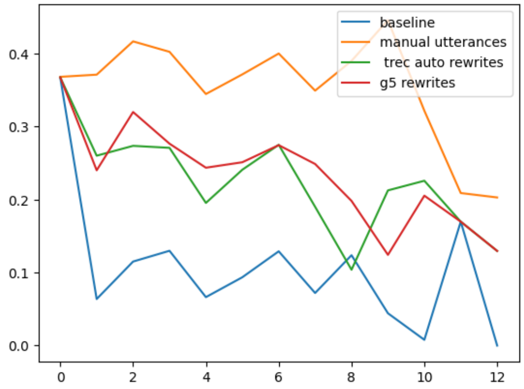


Figure 6: Average NDCG@1000 per topic

As the image shows, our methods outperform the baseline. This is because our approach keeps the context of the topic enclosed as the "conversation" evolves.

Run	MAP@1000	Precision@1000	NDCG@1000
Baseline	0.042	0.009	0.123
Manually	0.149	0.020	0.365
Trec auto	0.085	0.014	0.241
G5 approach	0.095	0.015	0.259

6 DISCUSSION

We attained a very close score to the automatically rewritten utterances.However, manually rewritten utterances is the best reformulation possible for these conversational queries. We have not done any fine-tuning of the parameters while training our advanced method. In the future works, we can adjust the parameter weights for training the T5 model as well as can introduce a model that recognizes named entities. We believe, we can attain even better score in the future with these adjustments.

7 CONCLUSION

This project was exquisite as a conversational assistant. However, The intial challenge was to understand how TREC 2020 dataset works. We later found out that it has it's own in-house reading method and using that method, reading the data was not as difficult as we thought. Additionally, it is a huge dataset. Therefore, parsing and indexing it was time consuming. We used elasticsearch for this purpose, as we believe it was the fastest way available to create the inverted index we needed. Lastly, we distributed among ourselves which model we want to implement as our advanced method - T5, GPT-2 or AllenNLP. We agreed upon using T5, as we understood this model better than the other two and this model is trained with CANARD dataset for query reformulation.

Finally, we observed from the evaluation of our reformulated context independent queries with respect to raw utterances, automatically rewritten utterances, and manually rewritten utterances that our system attains better accuracy for each turn than the original raw utterances. We attained an score of 0.259 on NDCG@1000 which outperforms the baseline socre of 0.123.

REFERENCES

[1] Elasticsearch . [n.d.]. Data in: documents and indices | Elasticsearch Guide [7.17] | Elastic. https://www.elastic.co/guide/en/elasticsearch/reference/7.17/documents-indices.html

[2] Elasticsearch . 2010. What is Elasticsearch | Elastic. https://www.elastic.co/what-is/elasticsearch

[3] Elasticsearch . 2018. Practical BM25 - Part 2: The BM25 Algorithm and its Variables. https://www.elastic.co/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables

[4] castorini. [n.d.]. castorini/t5-base-canard. https://huggingface.co/castorini/t5-base-canard?fbclid=IwAR2izTWj7uwGDyNE2M5WVCMGN9H1_7xedqCjv_5OUsr6P45ftBGm-1aAPCU

[5] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[6] Jeffrey Dalton, Sophie Fischer, Paul Owoicho, Filip Radlinski, Federico Rossetto, Johanne R Trippas, and Hamed Zamani. 2022. Conversational Information Seeking: Theory and Application. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3455–3458.

[7] daltonj. 2020. TREC Conversational Assistance Track (CAsT). https://github.com/daltonj/treccastweb/blob/master/2020/2020_manual_evaluation_topics_v1.0.json

[8] Venkat N. Gudivada, Dhana L. Rao, and Amogh R. Gudivada. 2018. Chapter 11 - Information Retrieval: Concepts, Models, and Systems. In *Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications*, Venkat N. Gudivada and C.R. Rao (Eds.). Handbook of Statistics, Vol. 38. Elsevier, 331–401. https://doi.org/10.1016/bs.host.2018.07.009

[9] Ida Mele, Cristina Muntean, Franco Nardini, Raffaele Perego, and Nicola Tonellotto. 2020. Topical Enrichment of Conversational Search Utterances Participation of the HPCLab-CNR Team in CAsT 2020. https://trec.nist.gov/pubs/trec29/papers/HPCLab-CNR.C.pdf

[10] Prakhar Mishra. 2020. *Understanding T5 Model : Text to Text Transfer Transformer Model.* https://towardsdatascience.com/understanding-t5-model-text-to-text-transfer-transformer-model-69ce4c165023

[11] NIST. 2000. Text REtrieval Conference (TREC) Overview. https://trec.nist.gov/overview.html

[12] Joao Palotti, Harrisen Scells, and Guido Zuccon. 2019. TrecTools: an open-source Python library for Information Retrieval practitioners involved in TREC-like campaigns *(SIGIR'19)*. ACM.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.

[14] Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational Information Seeking: An Introduction to Conversational Search, Recommendation, and Question Answering.