IITD-DBAI: Multi-Stage Retrieval with Pseudo-Relevance Feedback and Query Reformulation

Shivani Choudhary
shivani@sire.iitd.ac.in
Indian Institute of Technology, Delhi

1 Introduction

Conversational systems have acquired the center stage in NLP research. Compared to the conventional information retrieval task where we have to extract the passage or document from a vast collection of documents, the Conversational system requires extracting related information to respond to a series of questions. The turns in the conversation may follow the previous question. Complexity in this task arises due to the way we form the queries, which often have a reference to previous information using pronouns, co-reference. The presence of pronouns and unresolved co-references induces ambiguity in the query. Resolving the contextual dependency is one of the most challenging tasks in the Conversational system.

The Conversational assistance track (CAsT) has started in 2019. The long-term vision of the CAsT is "to support natural conversations between a person and a search engine to satisfy information needs and support complex information tasks". The task in both years of CAsT remained the same, to retrieve relevant passages depending upon the context evolution from the subsequent queries. raw_utterance CAsT-2019 had two tacks viz: automatic and manual track. Under Automatic track, response extraction was based on the raw utterances while the manual track has queries rewritten by humans. Co-reference and pronouns were resolved in the manually rewritten queries based on the historical context. Manually rewritten queries contain all of the information required to represent the single turn of the underlying information need. In CAsT-2019, additional information like description and title for the session was also provided (Dalton et al., 2020). Three corpora, namely MSMARCO, TREC CAR (Wikipedia) paragraph Corpus V2.0[1] were used. Initially, it had also considered WaPo, but it was later dropped. CAsT-2020 had some changes,

[1]http://trec-car.cs.unh.edu/datareleases/

Turn	Queries
1	I just had a breast biopsy for cancer. What are the most common types?
2	Once it breaks out, how likely is it to spread?
3	How deadly is it?
4	What? No, I want to know about the deadliness of lobular carcinoma in situ.
5	Wow, that's better than I thought. What are common treatments?
6	...

Table 1: A sample query from CAsT-2021 Automatic Evaluation Topics

title and description was **removed from the query** info, and document id as a canonical response to the query is added in the task. CAsT-2020 had three tracks - automatic, automatic-canonical, and manual. Under automatic, only raw utterance is to be used, while automatic-canonical can use the provided canonical response. The manual track was similar to last year based on the manually rewritten queries (Dalton et al., 2021). The dataset was MSMARCO and CAR (Dietz et al., 2018). CAsT-2021 has three tracks similar to CAsT-2020, but with a modification that we need to extract a specific passage out of the extracted document. The idea of specific passage to be extracted from the document centered around the idea that responses should be crisp and could be used by automatic voice assistants like Alexa, Google, etc. CAsT-2021 used MSMARCO (2019/20 dump), WAPO-2020 and KILT (Petroni et al., 2021).

2 Problem Description

In CAsT track, A conversation session S has a series of utterances $\{u_1, u_2, u_3, u_4 \ldots\}$ called as turns. Our task is to predict a set of top-K passages from the collection for each turn.

Tracks: There are three tracks of submission

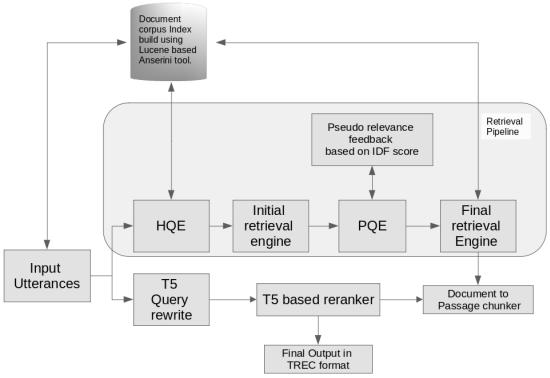

Figure 1: Multi-stage retrieval pipeline

for CAsT-2021 - Automatic, Automatic-canonical, and Manual. Manual track has rewritten queries as input.

Dataset: CAsT-2021 uses two dataset corpus for the task - MSMARCO, KILT (Petroni et al., 2021) and WAPO.

We have attempted the Automatic track. Each turn (Queries) in this session is raw in nature. The onus to resolve references and pronouns lies on the model itself.

3 Motivation and Method

We have gone through the CAsT-2020 submissions, and most of the models were multistage models. In the initial stage, a set of documents were retrieved from the collection, followed by re-ranking using neural model. Most of the models used query rewriting using generative model based on Transformer architecture (Vaswani et al., 2017). For query rewriting, T5, BART, GPT-2 was used, and the re-ranker engine was based on BERT, ALBERT, and T5.

Our retrieval framework has three components: Query rewrite, document retrieval with pseudo-relevance feedback, and neural engine-based re-ranker. The Query rewrite framework is based on chatty-goose [2] and re-ranker is based on pygaggle [3]. Core retrieval engine is based on the HQE (Yang et al., 2019) and PQE (Al-Thani et al.)

3.1 Historical Query expansion (HQE)

This step is based on the submission from CAsT-2019 (Yang et al., 2019). It is a three-stage algorithm. In the first step, it extracts the keyword for the session and query level followed by measurement of ambiguity in query in second step, and in the last stage, query expansion with the session keyword and query level keyword is done.

A keyword is deemed important if it strongly relates with the documents in the index. HQE uses BM25 score between a keyword and its highest-ranked document in the index as its measure of importance. HQE computes this measure for each token in every utterance when presented with a topic. The most informative ones are then selected for query expansion. For a given topic, keywords can be locally important (i.e., it strongly relates to the topic being discussed in the current utterance) or globally important (i.e., it relates to the overall conversation theme). HQE thus creates two sets of expansion keywords, session and query, during the extraction phase. Whether the keyword is considered a session keyword or a query keyword is decided by two separate cutoffs, Q_s, and Q_t. If the importance score of a keyword is greater than either of the cutoff scores, it is added to corresponding expansion sets. Note that session keywords are always used for expansion, while query keywords are only used when current utterance is identified as ambiguous. An utterance is deemed ambiguous when BM25 score between it and its highest ranked document in index is low (i.e. by itself utterance is not important). If the ambiguity score is less than a certain threshold θ, then the query expansion will take place with query keywords.

The cutoff values for Q_s, Q_t, and θ determine the performance of HQE. We follow a greedy approach to find the optimal values of these cutoffs. Note that a query is always expanded with current session keywords and occasionally with preceding query keywords. Thus, we start by tuning Q_s session cutoff. We set Q_t and θ to arbitrarily high values allowing query expansion only via session keywords. Then we do a line search for Q_s over the training set. Second, we set Q_t to some fixed value ($> Q_s$) and in similar fashion tune θ. Finally, we tune Q_s.

3.2 Passage Query Expansion (PQE)

Pseudo-relevance feedback enriches the query by incorporating features from top-k relevant documents. PQE uses a pseudo-relevance feedback mechanism as follows:

- The expanded query is from HQE is used to fetch an initial set of top-k documents. These

[2] https://github.com/castorini/chatty-goose
[3] https://github.com/castorini/pygaggle

Run Name	$K1$	b	NDCG@3	NDCG@5	P@5	AP@500
IITD-RAW_U_T5_1	0.9	0.4	0.3712	0.3631	0.5025	0.1759
IITD-RAW_U_T5_2	1.2	0.75	0.3801	0.3731	0.5203	0.1874

Table 2: Evaluation score of Submitted manual runs

documents combined together form the corpus of responses for the query.

- Individual tokens in the corpus are scored using TF-IDF. An IDF vector is pre-computed on complete MSMARCO documents.
- The topmost unique tokens are then considered for query expansion.

Note that PQE can be computationally prohibitive since it involves complete document retrieval. In order to avoid this, (Al-Thani et al.) proposes to use a simple rule to decide whether a query is to be expanded using PQE or not. (Al-Thani et al.) only perform PQE expansion when the query has at least one pronoun.

This process has two HyperParameters, top-k documents, and top-k tokens from the document based on the TF-IDF score. The TF-IDF score helps the system to select the important terms. While selecting the term, we have also placed a criterion that it should have a DF between 0.001 and 0.2. In the PQE step, we noticed that some digits also appeared in top-k terms. Later, we filter out those instances.

3.3 T5 query rewriter and reranker

In automatic track, query are presented in the bare format. In order to preserve the context of the query, we have used query rewriter. It uses T5 based model [4]. The model is trained on CANARD dataset (Elgohary et al., 2019) that contains a set of rewritten queries based upon the history. To produce a n^{th} rewitten utterance, we have supplied history of turns $\{u_1, u_2, u_3, u_4 \ldots, u_{n-1}\}$ concatenated with u_n. Reformulated queries were used for the final stage ranking of passages.

On the other hand T5 based ranking model (Nguyen et al., 2016) is trained on the MSMARCO, Robust04, Core17 and Core18. This model takes query Q and a list of passages $\{p_1, p_2, p_3, ..p_n\}$. It returns submitted passages according to the descending order of their relevance.

[4]https://huggingface.co/castorini/t5-base-canard

3.4 Document Indexing

All three datasets were pre-processed and converted into jsonl format. We have use Pyserini[5] to generate an index for faster retrieval of documents. Index was generated with an option to keep a copy of raw documents. We choose this option because the raw content was required at HQE and PQE stage.

4 Evaluation Matrix

CAsT-2021 used following matrix to present the results of the participants NDCG@3, NDCG@5, NDCG@500 and AP@500.

5 Results and Discussion

Index for the cleaned document was generated using Pyserini[6] with the default setting. Default setting did not restrict us to keeping $K1$ and b fixed. It can be changed during the retrieval of the documents. First, we performed our extraction and tuning on the 2019 training set. The best performance for HQE was obtained with Q_s = 4, Q_t = 4, θ = 10 and PQE with top-k = 5, top-k token = 3.

We have participated in automatic track and submitted two runs in the CAsT-2021. Two different set of parameters for BM25, $K1 = 0.9$, $b = 0.4$ and $K1 = 1.2$, $b = 0.75$ was selected for the retrieval of the document. The MAP is very low in both the results, and the main issue lies with the recall. Our retrieval engine extracted the top 100 documents from the corpus. Extracted documents were later chunked and re-ranked. In this process, lower retrieval numbers left the ranker with fewer relevant chunks, resulting in a lower than expected performance.

IITD-RAW_U_T5_2 has produced a mean NDCG@3 performance better than the median model. Model's performance on evaluation query has a noticeable variation in scores compared with median scores. Results for queries like – 115 and 119 were better than the median; however, results on the evaluation queries like – 111 and 117 were worse than the median benchmark for NDCG@3.

[5]https://github.com/castorini/pyserini
[6]https://github.com/castorini/pyserini

Analysis of query expansion term for the first stage retrieval suggests that query expansion term has carried the intent of the conversation to higher depths. While the poor performing queries have expansion terms with less relevant keywords, generic keywords. Query expansion terms for query 119 have terms like *swelling, shaking, infection, ear* that carried the conversation context to higher depth led to better performance.

We have also analyzed the hqe and pqe keywords for query expansion. We can take evaluation number 106; the first few queries are listed in Table-1. For turn 3, query expansion terms were *biopsy, breast, deadly, cancer, comment and cell*. This turn has only one term, "deadly", that can explain the intent of the query. But, it is too generic in nature which has led to poor recall. On the flip side, turn five is expanded with terms *carcinoma, wow, deadly, situ, lobular, deadliness, biopsy, breast* which has specific terms related to cancer like "lobular, situ, carcinoma" has a good recall. Here the word "specific" means relevance to the topic.

Our submission to CAsT-2021 aimed to preserve the key terms and the context in all subsequent turns and use classical Information retrieval methods. It was aimed to pull as relevant documents as possible from the corpus. It appears that it can retain some of the keywords in subsequent turns. But, it fails when the key term itself is generic. The performance of this model can be improved further by including the context word vectors in the HQE and PQE stages. Context vector-based selection may help to keep top-k terms that are relevant to the conversation theme.

References

Haya Al-Thani, Bernard J Jansen, and Tamer Elsayed. HBKU at TREC 2020: Conversational Multi-Stage Retrieval with Pseudo-Relevance Feedback; HBKU at TREC 2020: Conversational Multi-Stage Retrieval with Pseudo-Relevance Feedback.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2019: The Conversational Assistance Track Overview. Technical report.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021. CAsT 2020: The Conversational Assistance Track Overview. Technical report.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2018. Trec complex answer retrieval overview. In *TREC*.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Empirical Methods in Natural Language Processing*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Jheng-Hong Yang, Sheng-Chieh Lin, Chuan-Ju Wang, Jimmy Lin, and Ming-Feng Tsai. 2019. Query and answer expansion from conversation history. In *TREC*.