IRLab-Amsterdam at TREC 2021 Conversational Assistant Track

Antonios Minas Krasakis
a.m.krasakis@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Evangelos Kanoulas
e.kanoulas@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

This paper describes our participation (IRLab-Amsterdam) in TREC CAsT 2021. Our approach adapts a pre-trained token-level dense retriever (ColBERT) to perform zero-shot conversational search. Specifically, our query encoder reads the entire conversation history to contextualize the embeddings of the last user utterance/query, while the token-level matching function uses the contextualized embeddings to retrieve directly from the collection. The advantages of our method are two-fold: (a) it does not need any conversational data for training (ie. query resolutions, or conversational relevance judgements) and (b) it avoids complex pipeline systems based on rewriting that can affect performance (response latency) and robustness.

KEYWORDS

information retrieval, conversational search

1 INTRODUCTION

Data scarcity is one of the most important characteristics of conversational search, since most conversational queries are low-tailed (ie. they appear once) [11]. To deal with this problem, most methods first solve the surrogate task of conversational query resolution/rewriting by using human annotated query rewrites, which allows them to simplify conversational search to ad-hoc search [5, 6, 8, 10]. Despite their effectiveness in the current offline evaluation paradigm, those approaches (a) assume the presence of question reformulation data from a similar domain, which are not always available or easy to collect and (b) further complicate the retrieval pipeline by introducing higher response latency as well as robustness issues.

To overcome these issues, we adapt a pre-trained ad-hoc token-level dense retriever (ColBERT) to the conversational search setting, that uses no additional data specific to conversational search (ie. question rewriting or conversational relevance judgements). We achieve this in two steps: Firstly, our query encoder reads the entire conversation history and contextualizes the token-level embeddings of the last user utterance. Following that, our matching function uses the contextualized embeddings of the last turn's tokens to do dense retrieval directly from the corpus. Therefore, our method is zero-shot when it comes to conversational data, as it only relies on supervision from ad-hoc query relevance judgments, which are available at a large scale and much easier to collect.

Additionally, our method is much simpler and efficient in contrast to rewriting-based approaches, which have many different components that are often trained, tuned and evaluated in isolation. This increases the effort required for deployment and maintenance, but crucially calls into question the robustness and user satisfaction in the end of the pipeline.

Another serious shortcoming of pipeline systems is high response latency to a new query. Each component needs to run sequentially, as it takes input from the previous step. We should also note here, that in production systems the problem becomes even worse, as more components are usually added to those pipelines, such as speech-to-text or other post-processing modules.

2 METHODOLOGY

In this section, we describe our zero-shot dense retriever for Conversational Passage Retrieval.

2.1 Task & Notations

Let q_t be the user query to the system at the t-th turn, and p_t the corresponding canonical passage response provided by the competition organizers. We formulate our passage retrieval task as follows: Given the last user utterance q_t and the previous context of the conversation at turn t: $ctx_t = (q_0, p_0, ..., q_{t-1}, p_{t-1})$, we want return a ranking of K documents $R_{q_t} = (p_t^1, p_t^2, ..., p_t^K)$ from a collection C, that are most likely to satisfy the users' information need.

2.2 Token-level Dense Retrieval

In this section we briefly describe ColBERT[3], the dense retriever we adapted to our conversational task. In contrast to other dense retrievers that construct global query and document representations (eg. DPR[2] or ANCE[9]), ColBERT maintains embeddings of all query and document tokens and therefore performs matching on the token-level.

In practice, instead of relying on aggregated representations (ie. $[CLS]$ token), each token passes through multiple attention layers in a typical transformer architecture and is contextualized with respect to its surroundings [1, 7]. Then, those token embeddings are used to perform the matching. Specifically, each query token is matched with the most similar document token, using a $maxSimilarity$ operation. The score of a query-document pair is an aggregation over all query terms of the most similar term in this document:

$$S_{q,d} := \sum_{i \in [|E_q|]} \max_{j \in [|E_d|]} E_{q_i} \cdot E_{d_j}^T \tag{1}$$

Overall, this allows ColBERT to perform a more fine-grained matching on the term level, while computing soft term matches on contextualized token embeddings.

2.3 Conversational token-level Dense Retrieval

In our approach, we extend this idea of contextualizing embeddings of terms using their neighbors, to the task of conversational search. We argue that, when dealing with conversations, it is important for each turn to be contextualized with respect to the previous context,

Run name	Run type	NDCG@3	R@1000	MRR	MAP@500
Astypalaia256	canonical	0.24	0.46	0.52	0.14
histonly	raw	0.20	0.24	0.43	0.09
median	canonical	0.38	–	–	0.24
median	raw	0.33	–	–	0.18

Table 1: TREC CAsT '21 experimental results

as most conversational queries have continuity and even contain anaphoras to previous turns [6, 8, 10].

Therefore, the query encoder f_{QE} reads the previous conversational context ctx_t along with the last utterance q_t to produce the contextualized token embeddings of turn t:

$$E_{q_i^t} := f_{QE}(ctx_t \circ [SEP] \circ q_t) \tag{2}$$

Since the token embeddings of the last utterance are now contextualized with information from the previous history, we use ColBERT's token-level matching function (equation 1) to compute query-document relevance scores:

$$S_{q,d} := \sum_{i \in [|q^k|]} \max_{j \in [|E_d|]} E_{q_i^k} \cdot E_{d_j}^T \tag{3}$$

2.3.1 Zero-shot conversational search. When contextualizing user utterances q_i with respect to the history ctx_i, supervision from conversational tasks is not necessarily needed. Transformers pretrained with masked-language modelling and ad-hoc retrieval objectives have already been trained to change the token embeddings according to their surroundings.

Therefore, this method requires supervision only from the ad-hoc search task, in contrast to most other approaches that use human annotated query resolutions (CANARD etc.), or relevance judgements of conversational queries (ie. previous TREC tracks). This is important because conversational data are much harder to collect, in contrast to ad-hoc queries and judgements, which are typically available at a much larger scale. This is due to the fact that conversational queries are low-tailed (they become more rare and specific) as the user goes deeper into a conversation [11]. This makes it even harder, if ever possible to anonymize conversational queries, which can often be even more personal compared to ad-hoc queries.

In our experiments, we use the weights of a ColBERT retriever pre-trained on the MSMarco passage ranking dataset [4].

3 EXPERIMENTS

In this section, we describe the submitted runs and discuss our experimental results.

3.1 Runs

The primary difference between our two runs is the history context ctx_i that was used to contextualize the last turn embeddings.

- *historyonly*: uses only previous turns and ignore the canonical responses. Therefore the historical context at turn t becomes $ctx_t = q_1 \circ SEP \circ q_2 \circ ... \circ q_{t-1}$.
- *Astypalaia256* also includes the canonical passage response of the previous turn (p_{t-1}) in the historical context, which now becomes $ctx_i = q_1 \circ SEP \circ q_2 \circ ... \circ q_{t-1} \circ SEP \circ p_{t-1}$

In both cases, the maximum input length to the query encoder is set to 256, due to hardware limitations ($24GB$ GPU memory). When the total input exceeds this number, we ignore previous user utterances, until the length reduces enough. Specifically, we start deleting utterances from second turn onward. We do this to make sure we keep in the context some of the most important parts of the "conversation": (a) the last user utterance q_k(ie. needs to be answered), (b) the last canonical response p_{k-1} and (c) the first turn q_1, that often contains the overarching topic of the entire conversation.

3.2 Experimental Results

In this section, we discuss the official evaluation results of our submitted runs. Those can be found in Table 1. We also note that our results have been negatively affected by a bug that was discovered after the submission deadline, and therefore are tentative. Due to this reason, we are unable to provide additional baselines and oracles that further investigate the effectiveness of our method.

As we can see from the results in Table 1, our zero-shot retrievers perform lower than the average performance of the median run. The performance gap is roughly 40% for both the canonical and raw type submissions. Nonetheless, it is important to highlight that, in contrast to most other methods to be found in the literature our method: (a) does not take advantage of any additional training data and (b) is a first-stage ranker that does not use any cross-attentions between query and documents.

After comparing our two runs, it also becomes evident that canonical passages are an important part of the conversation and increasing the input length to our query encoder does not have such a detrimental effect.

4 CONCLUSIONS

In this paper, we describe our submissions in TREC CAsT 2021. We propose a zero-shot dense retriever, that uses supervision only from the ad-hoc ranking tasks and does not need any conversational-search related data. We show that our methods' performance as a zero-shot first-stage ranker is adequate, given that it significantly simplifies the previous complex conversational retrieval pipelines used in previous literature.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[3] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 39–48.

[4] 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. In *CoCo@ NIPS*.

[5] Ronak Pradeep, Xueguang Ma, Xinyu Zhang, Hang Cui, Ruizhou Xu, Rodrigo Nogueira, and Jimmy Lin. [n. d.]. H2oloo at TREC 2020: When all you got is a hammer... Deep Learning, Health Misinformation, and Precision Medicine. *Corpus* 5, d3 ([n. d.]), d2.

[6] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 355–363.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[8] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 921–930.

[9] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[10] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 1933–1936.

[11] Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-Shot Conversational Dense Retrieval. *arXiv preprint arXiv:2105.04166* (2021).