Zero-shot Query Contextualization for Conversational Search

Krasakis, A.M.; Yates, A.; Kanoulas, E.

DOI
10.48550/arXiv.2204.10613
10.1145/3477495.3531769

Publication date
2022

Document Version
Final published version

Published in
SIGIR '22

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 426, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
Zero-shot Query Contextualization for Conversational Search

Antonios Minas Krasakis  
am.krasakis@uva.nl  
University of Amsterdam  
The Netherlands

Andrew Yates  
a.c.yates@uva.nl  
University of Amsterdam  
The Netherlands

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

ABSTRACT
Current conversational passage retrieval systems cast conversational search into ad-hoc search by using an intermediate query resolution step that places the user’s question in context of the conversation. While the proposed methods have proven effective, they still assume the availability of large-scale question resolution and conversational search datasets. To waive the dependency on the availability of such data, we adapt a pre-trained token-level dense retriever on ad-hoc search data to perform conversational search with ad-hoc fine-tuning. The proposed method allows to contextualize the user question within the conversation history, but restrict the matching only between question and potential answer. Our experiments demonstrate the effectiveness of the proposed approach. We also perform an analysis that provides insights of how contextualization works in the latent space, in essence introducing a bias towards salient terms from the conversation.

CCS CONCEPTS
• Information systems → Information retrieval; Users and interactive retrieval; Retrieval models and ranking; Information retrieval query processing.

KEYWORDS
information retrieval, conversational search, neural ranking

1 INTRODUCTION
The introduction of commercial voice assistants along with advances in natural language understanding have enabled users to interact with retrieval systems in richer and more natural ways through conversations. While those interactions can ultimately lead to increased user satisfaction, they are inherently complex to handle as they require an understanding of the entire dialogue semantics by the retrieval system. Hence, the retrieval of relevant passages within the context of a conversation has risen as a promising research direction [2, 7, 15, 19]. Understanding the semantics of language has been empowered by the availability of large-scale datasets in a variety of tasks [3, 20, 29], which are lacking when it comes to conversational retrieval. Constructing a large and diverse conversational retrieval dataset can be quite challenging. Conversational queries are tail queries. As conversations evolve, multi-turn queries are likely to be unique and therefore cannot be aggregated for anonymisation, making it unlikely that publicly available resources can be built from real user interactions. Therefore, datasets need to be built using human experts in controlled environments. However, this approach leads to small-scale datasets [6–8, 22] and requires explicit conversation development instructions which bias the nature of the constructed dataset and hurt the generalizability of models to new types of conversations [1].

On the other hand there is a plethora of data resources for ad-hoc retrieval, e.g. Craswell et al. [4]. Therefore, most conversational retrieval approaches so far introduce a query rewriting step, which essentially decomposes the conversational search problem into a query resolution problem and an ad-hoc retrieval problem. Regarding query resolution, the majority of methods perform an explicit query re-write attempting to place the user’s question in the context of the conversation, by either expanding queries with terms from recent history [27], or rewriting the full question using a sequence-to-sequence model [12, 16, 18, 25, 30]. Yu et al. [31] learns to better encode the user’s question in a latent space so that the learnt embeddings are close to human rewritten questions, while Lin et al. [17] uses human rewritten questions to generate large-scale pseudo-relevance labels and bring the user’s question embeddings closer to the pseudo-relevant passage embeddings. In all cases, supervision is necessary and it is performed against CANARD [11], which consists of 40K synthetic resolutions of conversational questions. The only approaches that do not use supervision simply add the user’s question by extracting general informative terms from the conversation history [2, 18].

In this work we pose the following key research question: to what extent can we transfer knowledge from ad-hoc retrieval to the domain of conversational retrieval, where data scarcity is and will likely remain an imminent problem? To answer this question we adapt ColBERT [14], the state-of-the-art BERT-based token-level dense retriever pre-trained on ad-hoc search data. We propose Zero-shot Conversational Contextualization (ZeCo²), a variant of ColBERT which on one hand contextualizes all embeddings within the conversation, but on the other hand matches only the contextualized terms of the last user’s question with potential answers (Figure 1). As such our approach is zero-shot in the conversational domain, that is, it does not use any conversational search data, neither rewritten queries nor relevance judgements, to retrieve relevant passages. It
is also different from the aforementioned unsupervised keyword extraction works, since it focuses on contextualizing embeddings rather than adding terms to queries. In this work we aim to answer the following research questions:

RQ1 Can zero-shot contextualization of conversational questions improve dense passage retrieval?
RQ2 How does zero-shot contextualization change the last turn’s question embeddings?
RQ3 How is zero-shot contextualization affected by the anaphora phenomena found in conversations?

To the best of our knowledge this is one of the few efforts for zero-shot conversational search. Our approach remains agnostic and unbiased to small conversational datasets and it can prove particularly useful in privacy-sensitive settings (eg. medical domain), where annotating rewrites of conversational questions is not an option. Further, our method is orthogonal to and can be applied in combination with existing query resolution techniques.

We open-source our code for reproducibility and future research purposes.

2 METHODOLOGY
In this section, we describe our zero-shot dense retriever for conversational retrieval. Our approach consists of two main components:

- An encoder that produces token embeddings of a document or query and a matching component that compares query and document token embeddings to produce a relevance score.

2.1 Task & Notation
Let \( q_t \) be the user utterance/query to the system at the \( t \)-th turn, and \( p_t \) the corresponding canonical passage response provided by the dataset. 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 produce a ranking of \( k \) passages \( R_{q_t} = (p_1^t, p_2^t, ..., p_k^t) \) 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 [14], a dense retrieval model that serves as our query and document encoder \( f_{Enc} \). In contrast to other dense retrievers that construct global query and document representations (eg. DPR [13] or ANCE [28]), ColBERT generates embeddings of all input tokens. This allows us to perform matching at the token-level. To generate token embeddings, ColBERT passes each token through multiple attention layers in a transformer encoder architecture, which contextualizes each token with respect to its surroundings [9, 26]. We use \( E_q \) to denote the embeddings produced for a query \( q \) and \( E_d \) to denote the embeddings produced for a document \( d \). To compute a query-document score, ColBERT performs a soft-match between the embeddings of a query token \( w_q \) and a document token \( w_d \) by taking their inner product. Specifically, each query token is matched with the most similar document token and the summation is computed:

\[
\text{Score}(q, d) := \max_{w_q \in q, w_d \in d} E_{w_q} \cdot E^T_{w_d}
\]

2.3 Conversational Token-level Dense Retrieval
Our approach extends the idea of token contextualization to multi-turn conversations. When dealing with conversations, it is crucial for each turn to be contextualized with respect to the conversation, because conversational queries have continuity and often contain phenomena of ellipsis or anaphoras to previous turns [24, 27, 30].

In practice, ColBERT serves as our query and document encoder \( f_{Enc} \). We encode documents in the usual way. However, to encode a query at turn \( t \), we concatenate the conversational context \( ctx_t \) with the last query utterance \( q_t \) before generating contextualized query token embeddings \( E^*_q \):

\[
E^*_q := f_{Enc}(ctx_t \circ [SEP] \circ q_t)
\]

While \( E^*_q \) constitutes token embeddings of the entire conversation (i.e., the input to \( f_{Enc} \)), our goal is to perform ranking based on only tokens from the last utterance \( q_t \). To do so, we (1) replace \( E_{w_q} \) with \( E^*_{w_q} \) in the token-level matching function (Eq 1) and (2) compute the score as \( \text{Score}(q_t, d) \), so that only query tokens from the last turn contribute to it.

\[
\text{Score}(q_t, d) := \max_{w_q \in q_t, w_d \in d} E^*_{w_q} \cdot E^T_{w_d}
\]

Note that, this approach of contextualizing \( q_t \) with respect to the conversation history \( ctx_t \) avoids the need for resolution supervision from conversational tasks. Instead, it relies on the pre-training of three different tasks: (a) Masked Language Modelling, (b) next sentence prediction tasks (pre-training of BERT [9]) and the (c) ad-hoc ranking task (pre-training of ColBERT [14]).

3 EXPERIMENTAL SETUP
In this section we outline our experimental setup.

3.1 Datasets and Evaluation
We test our approach on the TREC CAsT ’19, ’20 and ’21 [6–8] datasets. Each dataset consists of about 25 conversations, with an average of 10 turns per conversation. CAsT ’20 and ’21 include canonical passage responses to previous questions, that the user can refer to or give feedback. The corpus consists of the MSMarco
Passages and Documents, Wikipedia and Washington Post news articles [10, 20, 21]. TREC CAsT ’19 and ’20 includes relevance judgements at passage level, whereas CAsT ’21 at the document level. For CAsT ’21, we split the documents into passages and score each document based on its highest scored passage (MaxP [5]).

To quantify retrieval performance we use two metrics: NDCG@3 and Recall, R@100. The former quantifies effectiveness at the top ranks, which is important for a user, while the latter expresses the ability of a first-stage ranker to retrieve relevant passages that can be later re-ranked with a more effective second-stage ranker.

### 3.2 Methods & Baselines

**ColBERT.** ColBERT was trained to contextualize tokens through (a) the self-supervision of BERT’s masked language model and sentence prediction [9] and (b) the training to optimize ad-hoc ranking [14]. In our experiments, we use the weights of a ColBERT retriever pre-trained on the MSMarco passage ranking dataset [20]. We use ColBERT v1, while our method remains applicable to v2 [23], where the main novelty is optimizations to reduce the index size. To avoid any spill-over effects we perform matching only on the query tokens; we deactivate matching on the last-turn (last-turn history, we compare the performance of the non-contextualized utterance (last-turn) with our contextualized approach (ZeCo\(^2\)) in the model.

**Baselines.** To assess the effect of our contextualization method ZeCo\(^2\), we compare with the following baselines:

- **ColBERT-based:**
  - last-turn: embeddings are not contextualized in the conversation (i.e., \(ctx_i = \emptyset\) in Eq. (2)).
  - all-history: embeddings are contextualized in conversation, and the matching function includes terms across the entire history (i.e., \(Score(ctx_i \circ q_i, d)\)).
  - human: oracle, using human rewrites

- **ConDrD-based [31]:**
  - zero-shot: no supervision from conversational tasks/data
  - few-shot: trained with Knowledge-Distillation on query rewrites
  - human: oracle, using human rewrites

### 4 RESULTS AND ANALYSIS

To answer RQ1, which asks whether the last user utterance (question) can be effectively contextualized with respect to the conversation history, we compare the performance of the non-contextualized utterance (last-turn) with our contextualized approach (ZeCo\(^2\)) in Table 1. It is clear that contextualization helps in all cases, especially in terms of Recall with relative improvements of 37% - 73%. We further observe that our approach significantly outperforms the all-history baseline, which uses the entire conversation as the query, in the first two datasets and yields comparable performance on CAsT’21. We hypothesize that the baseline’s improved performance on CAsT’21 is due to its document-level annotations, with one document satisfying multiple turns of the conversation. We also observe that all-history performs worse than last-turn regarding NDCG@3 but better regarding R@100. Furthermore, ZeCo\(^2\) outperforms the zero-shot ConDrD in most cases, especially with respect to Recall. To assess the effect of conversational contextualization (ZeCo\(^2\)), we measure the token embedding changes in the user’s query and report the terms with the largest average change in Table 3. We define token embedding change as the cosine distance between a term before and after contextualization: \(\Delta tok = 1 - \cos(tok_{ZeCo^2}, tok_{last-turn})\). We observe that terms indicating anaphora (‘they’, ‘it’, etc.), punctuation symbols and special tokens are the ones most influenced. This is expected, since users often use anaphoras referring to previous

<table>
<thead>
<tr>
<th>base-retriever</th>
<th>variant</th>
<th>zero-shot</th>
<th>CAsT’19 NDCG@3</th>
<th>R@100</th>
<th>CAsT’20 NDCG@3</th>
<th>R@100</th>
<th>CAsT’21 NDCG@3</th>
<th>R@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>ColBERT</td>
<td>last-turn (^a)</td>
<td>✔</td>
<td>0.214</td>
<td>0.157</td>
<td>0.155</td>
<td>0.124</td>
<td>0.140</td>
<td>0.154</td>
</tr>
<tr>
<td></td>
<td>all-history (^b)</td>
<td>✔</td>
<td>0.190</td>
<td>0.165</td>
<td>0.150</td>
<td>0.166</td>
<td>0.237</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>ZeCo(^2) (ours)</td>
<td>✔</td>
<td>0.238 (^b)</td>
<td>0.216 (^a,b,c)</td>
<td>0.176 (^b)</td>
<td>0.200 (^a,b,c)</td>
<td>0.234 (^a)</td>
<td>0.267 (^a)</td>
</tr>
<tr>
<td></td>
<td>human</td>
<td></td>
<td>0.430</td>
<td>0.363</td>
<td>0.443</td>
<td>0.408</td>
<td>0.431</td>
<td>0.403</td>
</tr>
<tr>
<td>ConDrD [31]</td>
<td>zero-shot (^e)</td>
<td>✔</td>
<td>0.247</td>
<td>0.183</td>
<td>0.150</td>
<td>0.150</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>few-shot</td>
<td></td>
<td>0.466</td>
<td>0.362</td>
<td>0.340</td>
<td>0.345</td>
<td>0.361</td>
<td>0.376</td>
</tr>
<tr>
<td></td>
<td>human</td>
<td></td>
<td>0.461</td>
<td>0.389</td>
<td>0.422</td>
<td>0.454</td>
<td>0.548</td>
<td>0.451</td>
</tr>
</tbody>
</table>

Table 1: Effectiveness of zero-shot embedding contextualization on TREC-CAsT datasets. Bold font indicates the best zero-shot performing model. Superscripts indicate statistically significant improvements (paired t-test, \(p = value < 0.05\)) of ZeCo\(^2\) over zero-shot models: last-turn \(^a\), all-history \(^b\) and ConDrD zero-shot \(^e\).

<table>
<thead>
<tr>
<th>token</th>
<th>frequency</th>
<th>avg((\Delta tok))</th>
</tr>
</thead>
<tbody>
<tr>
<td>they</td>
<td>60</td>
<td>0.501</td>
</tr>
<tr>
<td>it</td>
<td>196</td>
<td>0.480</td>
</tr>
<tr>
<td>[SEP]</td>
<td>934</td>
<td>0.464</td>
</tr>
<tr>
<td>?</td>
<td>873</td>
<td>0.458</td>
</tr>
<tr>
<td>that</td>
<td>52</td>
<td>0.440</td>
</tr>
<tr>
<td>.</td>
<td>192</td>
<td>0.424</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 3: Average change of frequent token embeddings after zero-shot contextualization (all CAsT datasets).

Next we consider the effect of contextualization of the user’s query by looking into how this changes the last turn’s query embeddings so that we answer RQ2.

**What are the most influenced terms?** To assess the effect of conversational contextualization (ZeCo\(^2\)), we measure the token embedding changes in the user’s query and report the terms with the largest average change in Table 3. We define token embedding change as the cosine distance between a term before and after contextualization: \(\Delta tok = 1 - \cos(tok_{ZeCo^2}, tok_{last-turn})\). We observe that terms indicating anaphora (‘they’, ‘it’, etc.), punctuation symbols and special tokens are the ones most influenced. This is expected, since users often use anaphoras referring to previous
What is the role of positivity in it? sociolgy 0.44 +0.67 tell me about the origins of popular music? 0.46 ... the origins of popular music? 1
What technological developments enabled it? popular music 0.54 +0.42 tell me about the bronze age collapse. 0.64 tell me about the bronze age collapse. 1
What is the evidence for it? bronze age collapse 0.36 0.00
why did ben franklin want it to be the national symbol? turkeys 0.22 +0.29 where are turkeys from? 0.55 where are turkeys from? 0.85
what is it about? neverending story film 0.39 +0.12 the neverending story film. 0.58 the neverending story film. 0.88

Table 2: Examples of best term matches of anaphora terms in conversation history (before & after query contextualization).

To examine whether anaphora terms coming closer to resolutions is simply a by-product of being encoded together, we measure similarity between anaphoras (\(\vec{A}\)) and (a) resolution terms (\(\vec{R}\)) vs (b) random terms from the same conversation (\(\vec{R}^{\text{Rnd}}\)) in Table 4. For consistency, we also encode random terms independently. When anaphoras are contextualized within only the last-turn, they are more similar to random terms than to resolutions on average.

Figure 2: Correlation between \(\Delta \text{Recall}\) and similarity change of anaphoras towards resolutions (CAsT’19)

\[
\Delta \text{sim}(\vec{A} \rightarrow \vec{R}) = \text{sim}(\vec{A}_{\text{ZeCo}^2}, \vec{R}) - \text{sim}(\vec{A}_{\text{last-turn}}, \vec{R})
\]

where anaphoras are contextualized within the last turn (\(\vec{A}_{\text{last-turn}}\)) or the entire conversation (\(\vec{A}_{\text{ZeCo}^2}\)). We encode resolutions (\(\vec{R}\)) independently to ensure they retain their original representations. On queries with multi-token anaphoras or resolutions, we pick the highest match. In Figure 2 we observe the scatter plot of this \(\Delta \text{sim}\) against \(\Delta \text{Recall}\). In most cases, contextualization improves Recall (\(\Delta \text{Recall} > 0\)) and brings anaphoras closer to resolutions (\(\Delta \text{sim} > 0\)). Further, Recall correlates with this change in similarity towards resolutions (Pearson’s \(R = 0.31, p-value = 0.005\)).
However, our method (ZeCo$^3$) brings anaphoras closer to their resolutions, while pushing them away from other (random) words from the same conversation. This confirms that resolutions have a high impact on contextualizing anaphoras, in contrast to other random conversation words. The mechanism behind this effect requires further investigation. It could be that simply the lower frequency of resolution terms has an effect here, but it is also possible that pre-trained transformers have certain co-reference resolution capabilities (eg. by relating ‘it’ to a noun). Regardless, it is evident that our method induces a bias towards salient terms from the conversation, leading to improved ranking performance.

5 CONCLUSIONS

In this paper, we explore the possibility of performing conversational search in a zero-shot setting, by contextualizing the last user query with respect to the conversation history. We show that this method is highly effective for first-stage ranking, yielding consistent and significant improvements in R@100. Further, it is suitable for privacy-sensitive settings, and can be combined with existing query rewriting techniques. In addition, we shed light into how zero-shot contextualization changes the last turn embeddings and show that biasing them towards the previous conversation can help retrieval, since it brings them closer to the conversation topic and salient terms. For future work we aim to explore zero-shot re-ranking and extend this work to few-shot training.

ACKNOWLEDGMENTS

This research was supported by the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), and the H2020-EU.3.4. - Big Data / Digital Humanities (314-99-301), and the H2020-EU.3.4. - Big Data / Digital Humanities (314-99-301). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES