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):
Krasakis, A. M., Yates, A., & Kanoulas, E. (2022). Zero-shot Query Contextualization for Conversational Search. In SIGIR '22: proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval : July 11-15, 2022, Madrid, Spain (pp. 1880–1884). The Association for Computing Machinery. https://doi.org/10.48550/arXiv.2204.10613, https://doi.org/10.1145/3477495.3531769

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  
a.m.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 conversa-
tional search into ad-hoc search by using an intermediate query
resolution step that places the user’s question in context of the con-
versation. 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 no additional fine-tuning. The proposed method allows to con-
textualize 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 ap-
proach. 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
interacting. 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 ad-
ances 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 promis-
ing 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. Regard-
ing 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 em-
beddings 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 expand
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 conver-
sation, 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 (e.g., 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 \( t \): \( \text{ctx}_t = (q_0, p_0, ..., q_{t-1}, p_{t-1}) \), 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 \( f_{\text{Enc}} \). In contrast to other dense retrievers that construct global query and document representations (e.g., 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:

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

Figure 1: Zero-shot Conversational Dense Retriever (figure adapted from [14] with permission)

### 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 \( f_{\text{Enc}} \). We encode documents in the usual way. However, to encode a query at turn \( t \), we concatenate the contextual conversation context \( \text{ctx}_t \) with the last query utterance \( q_t \) before generating contextualized query token embeddings \( E^*_q \):

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

While \( E^*_q \) constitutes token embeddings of the entire conversation (i.e., the input to \( f_{\text{Enc}} \)), 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.

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

Note that, this approach of contextualizing \( q_t \) with respect to the conversation history \( \text{ctx}_t \) 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
Table 1 shows the 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, \( p = \text{value} < 0.05 \)) of ZeCo\(^2\) over zero-shot models: last-turn \(^ a\), all-history \(^ b\) and ConvDR zero-shot \(^ c\).

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 ConvDR in most cases, especially with respect to Recall. Last, while the supervised versions of ConvDR clearly outperform ZeCo\(^2\) in NDCG@3, ZeCo\(^2\) remains competitive in terms of Recall.

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: \( \Delta tok = 1 - \cos(tok_{ZeCo^2} \cdot tok_{last-turn}) \). 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 first sign of it?

What is the role of positivism in it?

What technological developments enabled it?

What is the evidence for it?

why did ben franklin want it to be the national symbol?

what is it about?

tell me about the story film

tell me about the neverending story film

tell me about the bronze age collapse.

what is taught in sociology?

... the origins of popular music?

... the origins of popular music?

we observe that in certain cases, zero-shot contextualization resolves anaphoras successfully, bringing anaphora embeddings very close to the referred term ("sociology", "popular music", "throat cancer").

Lastly, we see cases where embeddings come closer to punctuation symbols, indicating that those might preserve some sort of global query representation, or a multi-token concept (e.g., "the neverending story film").

To answer RQ3, we quantitatively explore whether contextualization brings anaphora terms closer to their corresponding resolutions and how this affects ranking. Bringing anaphora terms closer to resolutions is crucial for conversational search. We identify those terms automatically using the human rewrites and define the effect of contextualization in bringing anaphora embeddings ($\vec{A}$) closer to resolution embeddings ($\vec{R}$) as:

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

where anaphoras are contextualized within the last turn ($\vec{A}_{\text{last-turn}}$) or the entire conversation ($\vec{A}_{\text{ZeCo}}$). 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 $R = 0.31, p-value = 0.005$).

To examine whether anaphora terms coming closer to resolutions is simply a by-product of being encoded together, we measure similarities between anaphoras ($\vec{A}$) and (a) resolution terms ($\vec{R}$) vs (b) random terms from the same conversation ($\vec{R}_{\text{random}}$) 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.
However, our method ($Z\text{CeO}_3^5$) 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 Innovative 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. - Societal Challenges - Smart, Green And Integrated Transport (814961). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Vaibhav Adlakha, Shelzaa Dhiulialwa, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2021. TopiOQA: Open-domain Conversational Question Answering with Topic Switching. arXiv preprint arXiv:2110.06768 (2021).
[2] Oleg Borisov, Mohammad Aliannejadi, and Fabio Crestani. 2021. Keyword Extraction for Improved Document Retrieval in Conversational Search. arXiv preprint arXiv:2109.05799 (2021).
[3] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuaC: Question answering in context. arXiv preprint arXiv:1808.07036 (2018).
[4] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 20 million clicked query-document pairs for analyzing search. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2893–2898.
[5] Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In Proceedings of the 42nd ACM SIGIR Conference on Research and Development in Information Retrieval. 985–988.
[6] Jeffrey Dalton, Chenshen Xiong, and Jamie Callan. 2020. CAsT 2020: The Conversational Assistance Track Overview. Proceedings of TREC (2020).
[7] Jeffrey Dalton, Chenshen Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The conversational assistance track overview. arXiv preprint arXiv:2003.13624 (2020).
[8] Jeffrey Dalton, Chenshen Xiong, and Jamie Callan. 2022. CAsT 2021: The Conversational Assistance Track Overview. Proceedings of TREC (2022).
[9] 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).
[10] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC Complex Answer Retrieval Overview. In TREC.
[11] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? Learning to rewrite questions-in-context. Can You Unpack That? Learning to Rewrite Questions-in-Context (2019).
[12] Rafael Ferreira, Mariana Leite, David Semedo, and Joao Magalhaes. 2021. Open- Domain Conversational Search Assistant with Transformers. In European Conference on Information Retrieval. Springer. 130–145.
[13] Vladimir Karpukhin, Barlas Oguz, 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:2004.04986 (2020).
[14] 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.
[15] Antoninos Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides, and Evangelos Kanoulas. 2020. Analysing the effect of clarifying questions on document ranking in conversational search. In Proceedings of the 2020 ACM SIGIR International Conference on Theory of Information Retrieval. 129–132.
[16] Sheng-Chieh Lin, Jieh-Hong Yang, and Jimmy Lin. (n. d.). TREC 2020 Notebook: CAsT Track. (In d.)
[17] Sheng-Chieh Lin, Jieh-Hong Yang, and Jimmy Lin. 2021. Contextualized Query Embeddings for Conversational Search. arXiv preprint arXiv:2104.08707 (2021).
[18] Sheng-Chieh Lin, Jieh-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuang-Ju Wang, and Jimmy Lin. 2021. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. ACM Transactions on Information Systems (TOIS) 39, 4 (2021), 1–29.
[19] Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-Aware Self-Supervised Learning for Knowledge-Grounded Conversations. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 522–532.
[20] 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 CoCeg NIPS.
[21] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yardani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. KILT: a benchmark for knowledge intensive language tasks. arXiv preprint arXiv:2009.02252 (2020).
[22] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten De Rijke. 2021. Conversations with Search Engines: SERP-Based Conversational Response Generation. ACM Trans. Inf. Syst. 39, 4, Article 47 (aug 2021), 29 pages. https://doi.org/10.1145/3433276.
[23] Keshav Santhanam, Omar Khattab, Jun Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. arXiv preprint arXiv:2102.01489 (2021).
[24] Switlana 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.
[25] Switlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. A comparison of question rewriting methods for conversational passage retrieval. arXiv preprint arXiv:2105.04166 (2021).
[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems. 5998–6008.
[27] 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.
[28] Lee Xiong, Cheryn Xiong, Ye Li, Kew-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:2007.08088 (2020).
[29] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained Transformers for Text Ranking: BERT and Beyond. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 1154–1156.
[30] 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.
[31] Shu Yi, Zhenghao Liu, Chenshen Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-Shot Conversational Dense Retrieval. arXiv preprint arXiv:2105.04166 (2021).