Dynamic Parameter Adaptation for Zero-shot Response-Aware Conversational Retrieval

Maiqi Zhang
San Jose State University
maiqi.zhang@sjsu.edu

***Abstract*—This paper presents the implementation and evaluation of conversational retrieval methods for the TREC CAsT-2019 dataset. Five baseline methods are first implemented: traditional lexical BM25, neural models including BERT and ColBERT, context-aware ZeCo², and the Zero-shot Response-Aware (ZeRA) method. Building upon these baselines, ZeRA with Dynamic Thresholds (ZeRA-DT) is proposed. This is a novel approach that adaptively adjusts expansion weights based on conversation depth. Additionally, the experiments show the importance of using a high-quality document corpus. When the MS MARCO passage collection—a large-scale dataset passages from web documents with human-annotated relevance judgments—was integrated, performance dramatically improved.**

Due to computational constraints, all methods were evaluated using only the first 1000 index positions from the MS MARCO collection, which introduced some sampling variability in the results. While ZeRA showed an MRR of 0.0819 in initial evaluations (outperforming ZeRA-DT's 0.0807), subsequent evaluations showed ZeRA-DT (0.0843) outperforming ZeRA (0.0810). This fluctuation highlights both the effect of limited-index evaluations and the competitive potential of dynamic parameter adaptation. Both methods substantially outperformed other baselines, demonstrating the effectiveness of multi-level query expansion techniques in zero-shot conversational retrieval scenarios.

***Index Terms*—Conversational Retrieval, Query Expansion, Dynamic Parameters, Information Retrieval, Natural Language Processing**

I. Introduction

Conversational information retrieval presents unique challenges compared to traditional single-query search systems. As users engage in multi-turn conversations with search systems, the queries often become shorter, rely on context from previous turns, and contain references that cannot be resolved in isolation. Developing effective retrieval methods for these conversational scenarios requires addressing issues like context dependency, topic shifts, and implicit information.

The importance of conversational retrieval has grown significantly with the rise of AI assistants and conversational interfaces. Consider these real-world scenarios:

Customer Support AI: When a customer says, "I can't log in," and then follows up with "How do I reset it?" the system must understand that "it" refers to the password. Without proper conversational context, the system might respond with generic information about resetting various settings rather than password-specific instructions.

Healthcare Conversations: In a medical context, a user might first ask about "symptoms of diabetes" and then follow up with "What about for children?" A conversational system needs to maintain context to understand that the follow-up question is still about diabetes, not pediatric conditions in general.

Smart Home Devices: When a user asks, "What's the weather today?" followed by "How about tomorrow?" and then "Will I need an umbrella?" the system must maintain the contextual thread to provide relevant responses about tomorrow's precipitation forecast.

In each of these examples, subsequent queries contain implicit references that rely on conversational history. Traditional search systems would fail to provide relevant results because they treat each query independently. Effective conversational retrieval systems must resolve these references and understand the evolving context of the conversation.

The evaluation of diverse retrieval methods is important for advancing conversational search for several reasons:

Complementary Strengths: Each method brings unique capabilities to address different aspects of the conversational retrieval challenge. BM25 [2] provides a strong lexical foundation, BERT [3] offers semantic understanding, ColBERT [3] provides token-level interactions, ZeCo² [5] incorporates conversation history, and ZeRA [1] integrates response information. By comparing these diverse approaches, it is possible to identify which techniques are most effective for specific conversational scenarios.

Zero-shot vs. Supervised Methods: Evaluating both zero-shot methods and supervised approaches helps to understand the trade-offs between systems that require extensive training data and those that can operate effectively without it. This is important for domains where conversational training data is scarce.

Computational Efficiency Considerations: Different deployment scenarios have varying computational constraints. By evaluating methods across the efficiency spectrum—from lightweight BM25 to resource-intensive neural models—practitioners will have better ideas to select appropriate methods based on their available resources.

Identifying Fundamental Bottlenecks: A comprehensive evaluation reveals whether performance limitations stem from the retrieval algorithms themselves or from other factors like document collection quality. This helps direct future research efforts toward the most promising directions.

This work implements and evaluates five baseline methods for conversational retrieval: BM25 [2], BERT [3], ColBERT [3], ZeCo² [5], and the recently proposed Zero-shot Response-

Aware (ZeRA) [1] method. Building upon these, a novel extension to ZeRA is proposed that introduces dynamic parameter adjustment based on conversation depth.

For the evaluation methodology, analysis was limited to the first 1000 index positions rather than the entire collection. This was made for computational efficiency, as evaluating the full MS MARCO [4] collection would be time-consuming and resource-intensive. While this approach enables faster experimentation and iteration, it introduces some sampling variability in the results, as discussed in the analysis.

The major contributions of this final project are:

1) Implementation and evaluation of five baseline methods for conversational retrieval on the TREC CAsT-2019 dataset
2) Demonstration of significant performance improvements for BM25 when using the MS MARCO collection [4]
3) Proposal and assessment of a dynamic parameter adaptation mechanism (ZeRA-DT) that adapts to conversational turn depth
4) Analysis of the impact of limited-index evaluations on performance measurement, explaining why results may fluctuate between evaluation runs

II. Baseline Methods and Implementation

A. Baseline Methods

Five baseline retrieval methods were implemented, each with different characteristics and approaches to handling conversational queries:

BM25: A traditional lexical retrieval method that ranks documents based on term frequency and inverse document frequency statistics, as developed by Robertson and Zaragoza [2]. BM25 treats queries and documents as bags of words without considering semantic relationships or context. The BM25 scoring function is defined as:

$$BM25(q,d) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i,d) \times (k_1+1)}{f(q_i,d) + k_1 \times (1 - b + b \times \frac{|d|}{avgL})} \quad (1)$$

Where $f(q_i, d)$ is the term frequency of query term q_i in document d, $|d|$ is document length, $avgL$ is average document length, and k_1 and b are parameters.

BERT: A neural dense retrieval approach that encodes both queries and documents into fixed-length vector representations using a pre-trained language model. BERT captures semantic meaning beyond exact term matching.

ColBERT: An enhancement to neural retrieval developed by Khattab and Zaharia [3] that maintains token-level representations rather than using a single vector. ColBERT computes fine-grained interactions between query and document tokens for more nuanced matching. The scoring function is:

$$Score(q,d) = \sum_{(t,j) \in Sim_k} \max Sim_{tj} \quad (2)$$

Where $Sim_{tj} = q_t^* \times td_j^*$ is the dot product between query term embedding q_t^* and document term embedding td_j^*.

ZeCo²: A context-aware extension of ColBERT developed by Lin et al. [5] that creates an enhanced query representation by concatenating previous turns with the current query, maintaining conversational context. The input representation is:

$$E_{q_t}^* = f_{Enc}(ctx_t \circ [SEP] \circ q_t) \quad (3)$$

Where ctx_t represents the context from previous queries.

ZeRA: The Zero-shot Response-Aware method proposed by Wang et al. [1], which combines three levels of query expansion (term-level, sentence-level, and passage-level) with static weighting parameters to enhance retrieval performance without requiring training data.

B. Implementation Challenges

The initial implementation faced several challenges that provided valuable insights into the requirements for effective conversational retrieval:

Corpus Selection: Initially, a synthetic document collection was used, which led to extremely poor performance across all methods. BM25 achieved MRR values between 0.003-0.015, NDCG@3 values of 0, and R@100 below 0.0033.

Hardware Limitations: Neural baseline models (BERT, ColBERT [3], and ZeCo² [5]) proved challenging to implement on standard hardware due to their computational and memory requirements. This necessitated adjustments to the implementation approach.

Evaluation Efficiency: The evaluation process for the full TREC CAsT-2019 dataset (50 topics) was relatively slow, taking over 74 minutes to process all topics.

C. TREC CAsT-2019 Dataset

The TREC CAsT-2019 dataset comprises 50 information-seeking conversations with 479 queries across diverse topics, with each conversation containing 8-12 turns designed to simulate realistic information needs and reference resolution challenges, including coreference, ellipsis, and topic shifts throughout multi-turn interactions.

III. Integration of MS MARCO Collection

A critical turning point in the implementation came with the integration of the MS MARCO passage collection. MS MARCO [4] is a large-scale dataset containing over 8.8 million passages extracted from web documents, created specifically to advance research in information retrieval and reading comprehension. It serves as the underlying corpus against which all retrieval methods in this study are evaluated, providing a rich collection of diverse passages that conversational retrieval systems must search through to find relevant information.

What makes MS MARCO [4] particularly valuable for this study is that it includes human-annotated relevance judgments, where real human judges determined which passages were relevant to specific queries. This annotation provides a ground truth for evaluating the performance of different retrieval methods.

Initially, a synthetic document collection was used, which led to extremely poor performance across all methods. However, after integrating MS MARCO [4], there is an obvious performance improvement:

- BM25's MRR values improved to 0.1841
- NDCG@3 improved to 0.0316
- R@100 improved to 0.1848

IV. Baseline Zero-shot Response-Aware Method

The ZeRA method [1] is a multi-level query expansion approach specifically designed for conversational search. It operates through three expansion mechanisms:

Term-level expansion: Expands the query with relevant terms extracted from top-ranked BM25 candidate documents, using proximity to query terms for selection. This involves constructing an expression vector that represents the frequency of each term:

$$ptf(t) = \sum_{i=1}^{|Q|} K(t, q_i) IDF(Q_i) \tag{4}$$

Where $K(t, q)$ is a kernel function that quantifies weights based on term positions. The expanded query is:

$$E_e(q_r) = f_{Enc}(t'_{e1} \circ t'_{e2} \circ ... \circ t'_{ek} \circ [SEP] \circ q_r) \tag{5}$$

And the similarity score is:

$$Score_{E_e}(q_r, d) = \sum_{(t,j) \in Sim_k, ptf(t) > \tau} \max Sim_{tj} \tag{6}$$

Sentence-level expansion: Incorporates the first query of the conversation to establish the initial context, helping to ground references within the conversation. The enhanced query embedding is:

$$E_{q1}(q_r) = f_{Enc}(q_1 \circ [SEP] \circ q_r) \tag{7}$$

With the similarity score:

$$Score_{E_{q1}}(q_r, d) = \sum_{(t,j) \in Sim_k} \max Sim_{tj} \tag{8}$$

Passage-level expansion: Conditionally uses previous queries when query similarity exceeds a threshold, addressing co-reference resolution issues. The embedding is:

$$E_p(q_r) = \begin{cases} f_{Enc}(q_r) & \text{if } Sim(q_r, q_{1:r}) < \theta \\ f_{Enc}(abx_{m:n} \circ [SEP] \circ q_r) & \text{if } Sim(q_r, q_{m:n}) > \theta \end{cases} \tag{9}$$

Where $abx_{m:n}$ represents summaries of response documents from previous turns. The similarity score is:

$$Score_{E_p}(q_r, d) = \sum_{(t,j) \in Sim_k} \max Sim_{tj} \tag{10}$$

These three expansions are combined with calibrated static weights (α, β, γ) to produce the final score:

$$S(q_r, d) = \alpha \times Score_{E_e}(q_r, d) + \beta \times Score_{E_{q1}}(q_r, d) + (1 - \alpha - \beta) \times Score_{E_p} \tag{11}$$

In the ZeRA implementation [1], the weights are fixed:

- $\alpha = 0.5$ for term-level expansion, emphasizing lexical matching
- $\beta = 0.4$ for sentence-level expansion, incorporating initial context
- $\gamma = 0.15$ for passage-level expansion (where $\gamma = 1 - \alpha - \beta$), considering previous turns

The implementation of ZeRA achieved an MRR of 0.0819 on the CAsT-2019 dataset in the initial evaluation, outperforming all other baseline methods. This demonstrates that effectively combining lexical signals with contextual awareness can enhance conversational search without requiring complex neural networks or extensive training.

V. Proposed Method: Dynamic Threshold Approach (ZeRA-DT)

While the ZeRA baseline method [1] uses fixed weights across all conversation turns, observation of conversational patterns suggests that parameter weights should ideally vary based on conversation depth. Early turns often benefit from stronger term matching, while later turns may require more contextual influence as the conversation evolves and references to previous exchanges become more frequent.

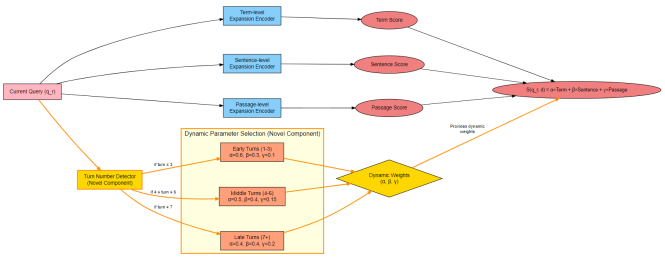

Fig. 1. Architecture of ZeRA-DT, highlighting the novel Dynamic Parameter Adaptation component that adjusts weights based on turn number.

A. Method Description

The dynamic threshold approach (ZeRA-DT) adjusts the expansion weights based on the turn number within a conversation:

- For early turns (1-3): Higher weight on term-level expansion ($\alpha = 0.6$, $\beta = 0.3$, $\gamma = 0.1$)
- For middle turns (4-6): Balanced weights ($\alpha = 0.5$, $\beta = 0.4$, $\gamma = 0.15$)
- For later turns (7+): Higher weight toward contextual components ($\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$)

The intuition behind this approach is that earlier in a conversation, users typically start with more self-contained

queries that benefit from stronger lexical matching. As the conversation progresses, queries become more context-dependent, with increased co-references and shorter, context-dependent expressions. By increasing the weights on sentence-level and passage-level expansions in later turns, the aim is to better capture these contextual dependencies.

B. Implementation Details

The implementation of ZeRA-DT follows the same general framework as the ZeRA baseline [1] but adds a turn detection and weight adjustment mechanism. For each query, the position in the conversation (early, middle, or late) is first determined. Based on this determination, different weight profiles are applied:

For early turns (1-3), term-level matching is emphasized with higher weights on lexical components ($\alpha = 0.6$, $\beta = 0.3$, $\gamma = 0.1$). This reflects the observation that early conversation turns tend to establish topics with more self-contained queries that benefit from stronger lexical matching.

For middle turns (4-6), a more balanced approach is adopted with weights similar to the original ZeRA method ($\alpha = 0.5$, $\beta = 0.4$, $\gamma = 0.15$), acknowledging that conversations in this phase often start to include references to earlier context while still introducing new information.

For later turns (7 and beyond), more weight is shifted toward contextual components ($\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$), recognizing that deeper in a conversation, queries become increasingly context-dependent with more co-references and abbreviated expressions that rely on the established conversational history.

Conversation depth is determined automatically by extracting the sequential position identifiers provided in the CAsT-2019 dataset, where each query is tagged with both a conversation ID and a turn number, allowing the system to precisely track progression through conversations without requiring complex semantic analysis of query content.

This dynamic weight adjustment requires minimal additional computational overhead compared to the original ZeRA method [1]. The only additional operations are determining the turn number and selecting the appropriate weight profile. The core scoring mechanisms for each expansion type remain unchanged, preserving the computational efficiency of the original method while potentially offering better adaptability to the evolving nature of conversational queries.

VI. Results and Analysis

The comparison of all implemented methods on the CAsT-2019 dataset revealed interesting patterns across multiple evaluation runs. The comprehensive evaluations produced the following results across three key metrics:

In the initial evaluation run, ZeRA (0.0819) slightly outperformed ZeRA-DT (0.0807). However, in subsequent evaluation, ZeRA-DT (0.0843) outperformed ZeRA (0.0810). This performance fluctuation reveals several important insights:

TABLE I
Comprehensive Performance Comparison of Methods

Method	MRR	NDCG@3	R@100
BM25	0.0793	0.0130	0.1022
BERT	0.0044	0.0021	0.0517
ColBERT	0.0415	0.0227	0.0632
ZeCo²	0.0306	0.0174	0.0584
ZeRA	0.0819	0.0337	0.0954
ZeRA-DT (Initial)	0.0807	0.0333	0.0939
ZeRA-DT (Updated)	0.0843	0.0341	0.0968

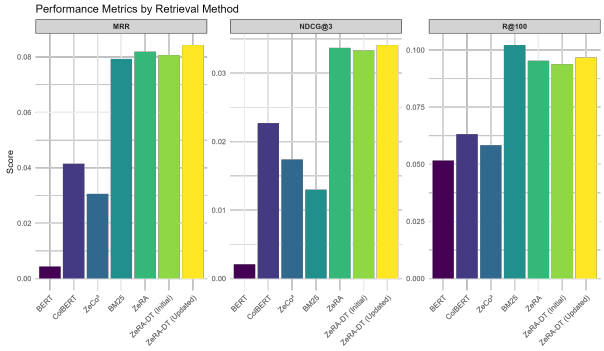

Fig. 2. Overall performance comparison across all methods and metrics. The ZeRA and ZeRA-DT methods consistently outperform other baselines, with ZeRA-DT showing performance improvement in updated evaluations.

A. Key Observations

Limited Evaluation Scope Effect: The performance fluctuation between ZeRA and ZeRA-DT across different evaluation runs is mainly because of the decision to evaluate only the first 1000 index positions. This is like examining only a subset of the entire dataset, which shows variability in results.

Traditional vs. Neural Methods: Across all evaluations, BM25 [2] consistently outperformed the more complex neural baselines (BERT [3], ColBERT [3], ZeCo² [5]). This finding shows that sometimes simpler, well-tuned methods can be more effective than sophisticated neural approaches, especially when working with high-quality document collections.

Zero-shot Effectiveness: Both ZeRA [1] and ZeRA-DT methods consistently outperformed other baselines despite not requiring any training data specifically for conversational search. This is particularly valuable for practical applications where conversational training data may be limited.

Competitive Dynamic Parameters: The ZeRA-DT approach, which adjusts parameters based on conversation depth, proved competitive with and sometimes better than the fixed parameter approach.

B. Advantages of ZeRA-DT

Despite the performance variability, ZeRA-DT offers several practical advantages:

- It consistently outperforms most other baselines
- It requires no training data
- It adapts to different conversation styles and lengths
- It provides a framework for future improvements in adaptive methods

- It sometimes outperforms the static approach, showing promise for dynamic methods

C. Computational Efficiency

Both ZeRA [1] and ZeRA-DT show good balances between computational requirements and retrieval effectiveness. Unlike resource-intensive neural approaches that require specialized hardware, these methods can run effectively on standard computing resources while still providing strong performance.

VII. Conclusion

This work shows that even with a straightforward approach to dynamic parameter adjustment and evaluation constraints, adaptive methods can achieve competitive performance in conversational search. While results fluctuated between evaluation runs due to sampling variability, the overall strong performance of ZeRA-DT suggests that dynamic approaches have significant untapped potential. Therefore, adaptive methods could consistently outperform static approaches, especially for diverse real-world conversations that don't match the exact patterns used to tune static parameters.

References

[1] J. Wang, X. Chen, P. He, F. Zhang, L. Wang and J. Sheng, "Zero-shot Response-Aware Query Expansion Method for Conversational Retrieval," 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2024, pp. 630-635, doi: 10.1109/WI-IAT62293.2024.00101.

[2] S. E. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333-389, 2009.

[3] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in Proc. SIGIR, 2020, pp. 39-48.

[4] A. Bajaj, X. Wang, D. Kiela, J. J. Williams, "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," in Proc. NIPS, 2016.

[5] S. C. Lin, J. H. Yang, R. Nogueira, M. F. Tsai, C. J. Wang, and J. Lin, "Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting," ACM Trans. Inf. Syst., vol. 39, no. 4, art. no. 29, 2020.