Relevance-assisted Generation for Robust Zero-shot Retrieval

Jihyuk Kim, Minsoo Kim, Joonsuk Park, Seung-won Hwang

Yonsei University, jihyukkim@yonsei.ac.kr
Seoul National University, {minsoo9574, seungwonh}@snu.ac.kr
{University of Richmond, NAVER AI Lab, NAVER Cloud}, park@joonsuk.org

Abstract

Zero-shot retrieval tasks such as the BEIR benchmark reveal out-of-domain generalization as a key weakness of high-performance dense retrievers. As a solution, domain adaptation for dense retrievers has been actively studied. A notable approach is synthesizing domain-specific data, by generating pseudo queries (PQ), for fine-tuning with domain-specific relevance between PQ and documents. Our contribution is showing that key biases can cause sampled PQ to be irrelevant, negatively contributing to generalization. We propose to preempt their generation, by dividing the generation into simpler subtasks, of generating relevance explanations and guiding the generation to avoid negative generalization. Experiment results show that our proposed approach is more robust to domain shifts, validated on challenging BEIR zero-shot retrieval tasks.

1 Introduction

Despite strong in-domain performance, dense retrievers have shown poor generalization to out-of-domain (OOD) zero-shot tasks where no training queries are available (Thakur et al., 2021). To enable training, pseudo-query generation (PQG) (Ma et al., 2021; Liang et al., 2020) has shown promising results, by generating in-domain pseudo queries $\tilde{Q}$ from a target corpus $D$.

However, we show $\tilde{Q}$ are often irrelevant to the documents for which they were generated, and generating a single document vector from the fine-tuned document encoder using $\tilde{Q}$ is often insufficient. Figure 1(a) illustrates the two limitations of the standard PQG approach, and Figure 1(b) our solutions, discussed as follows.

To tackle the two limitations, we propose Relevance-assisted Multi-query Domain Adaptation, or RaMDA. First, for relevance-guided generation, we first generate relevance explanations $Z_{d}$ (e.g., keywords explaining the relevance of the given document to queries to be generate). Second, guided by $Z_{d}$, we generate multiple queries, that form a more relevant and comprehensive $\tilde{Q}$. To address the second issue, we augment the single vector from $d$ with varying numbers of vectors from $\tilde{Q}$, denoted by $v_{d}$ and $v_{\tilde{q}_{i}}$, respectively. This enables the document to be matched to diverse relevant queries at test time.

We conduct experiments on BEIR benchmarks,
which include diverse out-of-domain retrieval tasks. The results show that, compared to the baseline PQG, our proposed RaMDA increases nDCG@10 and Recall@100, by 2.4 pt and 4.6 pt on average, respectively. Further analyses show that our generated queries approximate gold queries better, and capture diverse queries.

Our contributions are threefold:

1. We analyze existing PQG and identify their term frequency bias and diversity bias (§3.1).
2. Inspired, we disentangle generation into relevance explanation, and relevance-guided generation, for relevance-guided multi-query generation (§3.2).
3. Our relevance-guided generation is robust to distribution shifts (§4.2.1), complements the document, and thus improves document representation (§4.2.2).

2 Related Work

To address retrieval across diverse domains, dense retrievers have been trained using open-domain large-scale corpus, in supervised manner (Karpukhin et al., 2020; Xiong et al., 2021; Hofstätter et al., 2021) when relevance annotations are available (e.g., MS MARCO (Nguyen et al., 2016)), or using self-supervised learning in cases where such annotations are absent (Lee et al., 2019; Izacard et al., 2022). However, dense retrievers have shown poor performance when tested on specialized out-of-domain datasets, due to distribution shifts (Thakur et al., 2021; Yu et al., 2022).

Towards improved generalization, we discuss two approaches that tackle the challenge of distribution shifts: 1) improving training and 2) robustifying inference.

2.1 Improving Training for Better Adaptation

For improved training, existing work can be categorized into those pursuing domain invariant and domain-tailored learning.

The former aims to reduce the representation gap between source and target domains, by training a domain classifier, distinguishing source from target, based on which the encoder adversarially learns features that are domain independent (Xin et al., 2022). Recently, CONTinuous CONtrastive pretraining (COCO) (Yu et al., 2022) of a language model on target corpus, followed by implicit Distributionally Robust optimization (iDRO), achieved state-of-the-arts in this direction. However, as universal features from COCO-DR may not be effective for some target corpus, we adopt COCO-DR, but with domain-specific adaptation, by combining it with domain-tailored learning.

In contrast, domain-tailored learning aims to produce a domain-specific encoder, by fine-tuning the encoder to better fit each target domain. To enable fine-tuning, relevant query-document pairs should be constructed to build a training dataset, by devising pseudo-queries for each document in the corpus. To this end, pseudo-queries have been generated by either heuristic rules or a trained generator. For the former, TSDAE (Wang et al., 2021a) randomly injects noise into the document, while for the latter, GenQ (Ma et al., 2021) or GPL (Wang et al., 2021b) leverage a pseudo-query generator trained using MS MARCO, resulting in better adaptation performance compared to the former. While employing a trained generator, our distinction is ensuring the relevance of pseudo-queries.

2.2 Robustifying Inference by Increasing Model Capacity

In another dimension, domain shifts can be tackled by increasing the model capacity, through enriching query-document interactions or ensembling multiple retrievers, discussed as follows.

Beyond the similarity between a pair of single vectors having limited capacity (Luan et al., 2021), matching between query-document can be extended to term-level interaction. Cross-encoder (Guo et al., 2016) can capture full interactions between query and document terms, though not scalable to our target tasks as documents are not indexable. ColBERT (Khattab and Zaharia, 2020), with late interaction, is an indexable alternative with comparable performance, which we adopt as a baseline. Ours shares the same motivation of enriching interaction but distinguishes itself by making the interaction more concise via $\hat{Q}$, showing better performance with little index overhead.

While the term-level interaction enriches relevance signals via multiple terms, such signals can be captured from multiple retrievers by ensembling (Gao et al., 2021). With this view, ours can be viewed as introducing another retriever, to gain the benefit of such signals from two complementary text representations, $\hat{Q}$ and $D$. While showing comparable performance to the standard ensemble, when combined ours further enhances state-of-the-
art performance.

3 Method

Given a document \( d \) in a target corpus \( D \), PQG aims to generate pseudo-queries \( \tilde{Q}_d = \{ q_i \}_{i=1} \), as alternatives to gold queries \( Q_d \). Following previous work, we employ T5 (Raffel et al., 2020) as the backbone generator.

3.1 Motivation: Distribution Shift on PQG

Desirably, a robust PQG method should model \( p(\tilde{Q}_d|D) \), such that the sampled \( \tilde{Q}_d \) should closely approximate \( Q_d \). However, as we will show, PQG often fails to generalize to OOD settings. We hypothesize that this failure is driven by two biases in the source domain. First, term frequency bias: PQG can be biased to generate terms that occur frequently in the source domain, and thus fail to generate rarely observed terms. Second, diversity bias: The source domain may have a short passage, where the topic of queries naturally coincides. When target domains have a long document covering a diverse set of topics, PQG trained from the source domain would generate a homogeneous set of queries covering only a single main topic.

We conduct a preliminary analysis of existing PQG approaches with respect to such biases. We first quantify the two biases and categorize OOD datasets in terms of the biases. Similar to Wang et al. (2022), the term frequency bias is measured by \( \frac{1}{\sum_{t \in q} \log(1 + DF_t)} \), where \( q \) denotes a query (or a pseudo-query) in the target domain and \( DF_t \) denotes document frequency of \( t \), i.e., how many documents in the source domain contain \( t \) in their relevant queries. For diversity bias, we measure the maximum cosine distance between pairs of embeddings of any two relevant queries (or pseudo-queries) for the same document. Figure 2(a) visualizes datasets regarding the two bias metrics. We can observe that, while some OOD datasets share similar distributions to MS MARCO, others deviate significantly from it in terms of bias characteristics, namely Climate-FEVER, TREC-COVID, SCIDOCS, and NFCorpus. With the goal of de-biasing PQG, we adopt these four datasets, which demonstrate clear distribution shifts, denoted as “BEIR-BiasShift”, in our experiments.

For an efficient preliminary analysis, we focus on the small corpus datasets among the four, which are NFCorpus and TREC-COVID, denoted as “BEIR-BiasShift-Small”. Figure 2(b) compares terms in gold and synthesized queries in the target domain in terms of term frequency bias, denoted as \( Q \) (x-axis) and \( \tilde{Q} \) (y-axis) in the figure, respectively. Desirably, the two distributions should be identical (as in the dotted diagonal line \( y = x \)). The figure shows that gold query terms \( Q \) are rarely observed in the source distribution, but \( \tilde{Q} \) from the baseline PQG model (shown in red skewing lower than the optimal line) fails to generate the rare terms. In terms of diversity bias, Figure 2(c) compares the semantic diversity of \( Q \) and \( \tilde{Q} \), where \( \tilde{Q} \) should be as diverse as \( Q \). Results show that the baseline PQG suffers from the bias, showing significantly lower diversity compared to that of gold queries.

Our hypothesis is that biased queries, as observed above, negatively affect the generalization and should be pruned off, to allow the retriever to learn from an unbiased set of \( \tilde{Q} \).

3.2 Relevance-assisted Multi-query Generation

To this end, our distinction is to decompose the generation of \( \tilde{Q}_d \) into relevance explanation, and relevance-guided generation. First, we generate an explanation of the relevance between \( d \) and the query to be generated, as the set of terms \( Z_d \) which are shared by the relevant \( d-Q_d \) pairs. Next, we leverage \( Z_d \) to guide the generator to sample improved \( \tilde{Q}_d \) that includes relevant terms for \( d \), thereby enhancing generalization.

Alternatively, one may over-generate-then-filter (i.e. post filtering), which we denote as GenQ + RTF in Figure 2. RTF refers to round-trip filtering (Dai et al., 2022), approximating the relevance of generated \( \tilde{q}_i \) if a dense retriever ranks \( d \) at top-1 using \( \tilde{q}_i \) as a query. However, this is not only expensive, requiring repetitive decoding and ranking, but also aggravates the biases by filtering out rarely observed query terms and diverse query terms, as shown in blue lines in Figure 2 (b) and (c).

In contrast, we propose to filter preemptively, by decoding \( q_i \) guided by \( Z_d \). Among many relevance explanations surveyed in Anand et al. (2022), we employ SPLADE (Formal et al., 2021), generating \( Z_d \) terms with weights \( \lambda_{Z_d} \) on terms, based on strong empirical results. Given \( Z_d \), our pseudo-query generator decodes \( q_i \in \tilde{Q} \) guided by \( Z_d \) via \( \arg \max_{q'} p(q'|Z_d, d) \). For \( p(q'|Z_d, d) \), we add

\[ \text{We used COCO-DR for encoding queries, which is one of the state-of-the-art dense retrievers.} \]

\[ \text{We used COCO-DR for the dense retriever.} \]
λZ_d to output logits of the decoder, followed by softmax normalization, such that terms with high scores given by SPLADE will be more likely to be generated as a pseudo-query. As a result, ours better alleviates the distribution shifts as shown in green lines in Figure 2 (b) and (c).

Given ̂Q, standard PQG approaches fine-tune the document encoder for each domain adaptively, to enable it to represent the dense vector v_d of d, yet often limited to a single vector representation. In contrast, as we observed in §3.1, relevant queries for d in target domains are often diverse, where the capacity of the fixed-size v_d becomes the bottleneck. Our distinction is increasing the representation capacity, by appending varying numbers of vectors from ̂Q_d to v_d. Specifically, we first partition tokens in d into S segments \( \{ s^d_l \}_{l=1}^S \), where each segment \( s^d_l \) has a fixed number of tokens and has a sub-topic of \( d^3 \). We then generate pseudo-queries \( \{ q^d_l \}_{l=1}^S \) for each segment, such that pseudo-queries from the whole segments can cover diverse topics in d. Finally, to augment v_d, we encode ̂q_d into the dense vector v̂_d and append it to v_d. In our experiments, to maximize the coverage, we sample 50 pseudo-queries per segment and then do mean pooling of embeddings of those, to have the single vector v̂_d.

Given v_d and \( \{ v^d_l \}_{l=1}^S \), the relevance to the test-time q is measured by \( v^T_q v_d + \max_{l \in [1,S]} v^T_q v^d_l \), where \( v_q \) denotes the dense vector of q and \( v^T \) denotes inner product. We employ the max-pooling on varying numbers of pseudo-query embeddings, to capture the most relevant sub-topic to q (Khattab and Zaharia, 2020).

4 Experiments

4.1 Experimental Setup

Dataset and Evaluation Metrics To evaluate the effectiveness of our method, we conduct experiments on BEIR, a benchmark designed to evaluate the zero-shot generalization of retrieval systems across an array of different information retrieval tasks on various domains. Among BEIR datasets, we adopt BEIR-BiasShift datasets showing the largest distribution shifts from MS MARCO in terms of the two biases (Figure 2(a)), which are NFCorpus, SCIDOCS, TREC-COVID, and Climate-Fever. For evaluation metrics, following Thakur et al. (2021), we adopt nDCG@10 and Recall@100, which measure the overall quality of predicted top-10 ranking and the completeness of top-100 documents on relevant documents, respectively.

Baselines We compare RaMDA to both domain-invariant retrievers and domain-adaptive retrievers. For the former, we compare COCO-DR and SPLADE, as the state-of-the-art models among dense retrievers and sparse retrievers, respectively. While employed to guide PQG in RaMDA, SPLADE can serve as a retriever, by assessing the relevance of d via the sum of \( \lambda Z_d \). In addition, we also compare Contriever, which is the state-of-the-art model among retrievers trained using self-supervised learning.

As domain-adaptive retrievers, we compare
Retriever | NDCG@10 on BEIR-BiasShift datasets
---|---
| NFCorpus | SCIDOCS | TREC-COVID | Climate-Fever | Average
---|---|---|---|---|---
domain invariant retriever
Contriever | 32.8 | 16.5 | 59.6 | **23.7** | 33.2
SPLADEv2 | 33.4 | 15.8 | 71.0 | 23.5 | 35.9
COCO-DR | 35.5 | 16.0 | 78.9 | 21.1 | 37.9
domain adaptive retriever
GenQ | 31.9 | 14.3 | 61.9 | 17.5 | 31.4
GPL | 34.5 | 16.9 | 70.0 | 23.5 | 36.2
(Ours) RaMDA † | **38.6** | **17.8** | **81.4** | **23.5** | **40.3**
GenQ † | 34.5 | 13.3 | 72.3 | 20.5 | 35.2
GenQ + RTF † | 34.2 | 13.2 | 72.0 | 19.5 | 34.7

Table 1: NDCG@10 on BEIR-BiasShift datasets. † denotes retrievers that employ $\tilde{Q}_d$-augmented document representations (i.e., $\{v_{\tilde{Q}_d}\}_{i=1}^S$ in addition to $v_d$) with different generators. The best and the second-best results are denoted in **bold**-faced and *underlined*, respectively.

GenQ and GPL. GenQ utilizes a pseudo-query generator, initially trained on MS MARCO and subsequently adapted for each target domain to produce domain-specific pseudo-queries. GPL is a similar query generation approach, but additionally utilizes an expensive cross-encoder to label the generated pseudo-queries, for better adaptation.

Implementation Details For the pseudo-query generator, we fine-tune T5 (Base) using MS MARCO for 50k steps with 1k warm-up steps, by employing AdamW (Loshchilov and Hutter, 2019) optimizer with learning rate 1e-5 and batch size 32.

4.2 Results
We first validate the effectiveness of RaMDA in retrieval performance on BEIR, by comparing RaMDA with domain-adaptive retrievers as well as the state-of-the-art domain-invariant retriever. Following previous work, we adopt nDCG@10 as the evaluation metric. Results are shown in Table 1.

4.2.1 Analysis on pseudo-query generation
In this section, we study how PQG affects the adaptation to out-of-domain tasks.

Poor PQG does not help domain adaptation. Both existing domain adaptive retrievers (GenQ and GPL) exhibit lower average performance than the domain-invariant retriever, COCO-DR. This is because $\tilde{Q}$ is often different from gold queries in target domains, as observed in Figure 2.

RaMDA’s preemptive filtering helps, while post-filtering is detrimental. We compare RaMDA with the post-filtering approach, denoted as “GenQ + RTF†”6. While RTF produces similar, or even worse queries than blind generation, in contrast, our preemptive RTF consistently outperforms both “GenQ” and “GenQ + RTF”, as well as domain invariant baselines such as COCO-DR and SPLADE.

Biases on PQG negatively affect retrieval performance. We conduct ablation studies where we remove half of the pseudo-queries that account for each of the biases, to compare with RaMDA (with all pseudo-queries) in bias-amplified settings. Regarding the frequency bias, pseudo-queries that have the most rarely observed terms in MS MARCO are removed. Regarding the diversity bias, we repeatedly remove a pair of pseudo-queries whose distance is the farthest among the remaining pseudo-queries, until only half of the pseudo-queries remain. To demonstrate the significance of alleviating the two biases, we also compare performance of randomly removing pseudo-queries. The results are reported in Table 2.

Removing randomly sampled pseudo-queries shows the least degradation, indicating that alleviating two biases has a significant contribution to performance. The contribution of the two varies depending on the dataset characteristics. As shown in Figure 2(a), between the two datasets, NFCorpus and TREC-COVID, NFCorpus exhibits a more pronounced distribution shift in terms of diversity bias, with diversity scores of 0.83 and 0.42 for NFCorpus and TREC-COVID, respectively. Conversely, TREC-COVID demonstrates a more significant shift in term frequency bias, with rarity scores of 0.25 and 0.43 for NFCorpus and TREC-

---

For fair comparisons, we employ the same $\tilde{Q}_d$-augmented document representation for all methods.
Table 2: Ablation study on BEIR-BiasShift-Small datasets, targeting the biases introduced in §3.1, by removing pseudo-queries, from the full set of pseudo-queries, that contribute to alleviating (a) frequency bias and (b) diversity bias. Red numbers denote the largest performance drop from RaMDA.

<table>
<thead>
<tr>
<th>(Ours) RaMDA</th>
<th>NFCorpus</th>
<th>TREC-COVID</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) abl. rarely observed Q</td>
<td>36.5</td>
<td>76.5</td>
</tr>
<tr>
<td>(b) abl. diverse Q</td>
<td>35.6</td>
<td>77.2</td>
</tr>
<tr>
<td>abl. random Q</td>
<td>37.7</td>
<td>80.3</td>
</tr>
</tbody>
</table>

Figure 3: Comparing PQG models on BEIR-BiasShift-Small datasets, regarding efficiency-effectiveness. X-axis (efficiency) denotes the number of decoded generation samples for pseudo-queries. Y-axis (effectiveness) denotes the retrieval performance.

Figure 4: Recall and precision of tokens from different fields, on gold query terms.

Efficiency-Effectiveness trade-offs. Figure 3 compares efficiency-effectiveness trade-offs between preemptive and post-filtering, where efficiency is measured by the number of pseudo-query samples (x-axis), and effectiveness (y-axis) by the retrieval performance.

Ours shows high effectiveness consistently over all sample numbers, and tends to show performance improvements as more pseudo-queries are sampled. In contrast, when using GenQ, sampling more pseudo-queries rather decreases the retrieval performance on TREC-COVID, indicating the biased PQG negatively contributes to the generalization. GenQ + RTF shows a similar trend, indicating that RTF fails to filter such harmful pseudo-queries.

Compared to the domain-invariant retriever COCO-DR, ours requires only 10 or 15 pseudo-query samples to outperform COCO-DR, showing better efficiency. In contrast, both GenQ and GenQ + RTF consistently show worse performance than COCO-DR, indicating poor PQG makes the domain adaptation ineffective.

4.2.2 Analysis on document representation

We now examine whether $\hat{Q}$ can enhance document representations by complementing $D$, in Figure 4.

$\hat{Q}$ complements $D$. Since documents are much longer than pseudo-queries, $D$ shows better recall than $\hat{Q}$ alone. However, even terms from relevant queries often do not appear in documents. $\hat{Q}$ adds missing terms to complement $D$, further increasing recall on gold query terms when combined with $D$. On the other hand, regarding precision, $\hat{Q}$ is consistently better than $D$. This indicates that $\hat{Q}$ can take the role of a summary of $D$, to rectify the noisy vocabulary of $D$.

We further show that $\hat{Q}$ alleviates a well-known problem of dense vector representation, called token amnesia (Ram et al., 2023), where the single dense vector of a document often fails to capture its salient terms, due to occlusion by noisy terms.

Specifically, to see whether gold query terms in $d$ can be retained by the dense vector of $d$ (or $\hat{Q}_d$), we project $v_d$ (or $v_{\hat{Q}_d}$) into interpretable BERT
Table 3 compares ours with baselines with higher model capacity – enriching query-document interactions or ensembling multiple retrievers. Beyond single vector representation, ColBERT (Khattab and Zaharia, 2020) enables the document representation to have adaptive capacity by indexing all terms in \( d \), with a memory overhead on index size. On the other hand, ensemble methods increase the capacity by introducing another dense retriever. We compare with an ensemble of COCO-DR and GTR (Ni et al., 2022).

Surprisingly, though increasing the capacity, ColBERT underperforms COCO-DR on all datasets except SCIDOCS, and the ensemble often shows comparable or worse performance to individual retrievers. This is because the capacity increase in both methods is constrained by the quality of \( d \), which often produces noisy lexicons and semantics, as observed in Figures 4 and 5. While sharing the same goal, we leverage \( \tilde{Q} \) to complement \( d \). As a result, with only minimal overhead on the index compared to dense retrievers, our method outperforms all compared baselines on all tested datasets.

5 Conclusion
We investigated PQG for overcoming domain shifts in zero-shot retrieval, motivated by the observation that generated PQs often negatively affect such a goal. We show term frequency and diversity bias as a cause, and propose a novel PQG method that preempts negative PQG. We validate with extensive experiments on the BEIR benchmark, that through relevance-guidance and multi-query generation, our proposed model effectively addresses the challenges of domain shifts in zero-shot retrieval.

Acknowledgements
This work has been financially supported by SNU-NAVER Hyperscale AI Center, and Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (NO.2022-0-00077, AI Technology Development for Commonsense Extraction, Reasoning, and Inference from Heterogeneous Data).

---

For further details, refer to Ram et al. (2023).


