Combining Multiple Supervision for Robust Zero-shot Dense Retrieval

Yan Fang[1], Qingyao Ai[1]*, Jingtao Zhan[1], Yiqun Liu[1], Xiaolong Wu[2], Zhao Cao[2]

[1] Quan Cheng Laboratory & Department of Computer Science and Technology, Tsinghua University & Zhongguancun Laboratory & Beijing, China
[2]Huawei Poisson Lab
fangy21@mails.tsinghua.edu.cn, aiqy@tsinghua.edu.cn

Abstract

Recently, dense retrieval (DR) models, which represent queries and documents with fixed-width vectors and retrieve relevant ones via nearest neighbor search, have drawn increasing attention from the IR community. However, previous studies have shown that the effectiveness of DR critically relies on sufficient training signals, which leads to severe performance degradation when applied in out-of-domain scenarios, where large-scale training data are usually unavailable. To solve this problem, existing studies adopt a data-augmentation-plus-joint-training paradigm to construct weak/pseudo supervisions on the target domain and combine them with the large-scale human annotated data on the source domain to train the DR models. However, they don't explicitly distinguish the data and the supervision signals in the training process and simply assume that the DR models are mighty enough to capture and memorize different domain knowledge and relevance matching patterns without guidance, which, as shown in this paper, is not true. Based on this observation, we propose a Robust Multi-Supervision Combining strategy (RMSC) that decouples the domain and supervision signals by explicitly telling the DR models how the domain data and supervision signals are combined in the training data with specially designed soft tokens. With the extra soft tokens to store the domain-specific and supervision-specific knowledge, RMSC allows the DR models to conduct retrieval based on human-like relevance matching patterns and target-specific language distribution on the target domain without human annotations. Extensive experiments on zero-shot DR benchmarks show that RMSC significantly improves the ranking performance on the target domain compared to strong DR baselines and domain adaptation methods, while being stable during training and can be combined with query generation or second-stage pre-training.

1 Introduction

Dense Retrieval (Guo et al. 2021; Lin, Nogueira, and Yates 2021) has become increasingly popular in recent years and has achieved the state-of-the-art ranking performance. It effectively leverages the pre-trained language models (Devlin et al. 2018) to create dense representations for text, and relevant documents can be efficiently retrieved by conducting nearest neighbor search with the encoded query vector. Experimental results show that dense retrieval substantially outperforms the traditional lexical retrieval methods like BM25 (Robertson and Walker 1994).

Studies have shown that dense retrieval methods are often data hungry and require large-scale annotated data in order to achieve reliable performance (Gulrajani and Lopez-Paz 2021; Gururangan et al. 2020). In practice, however, obtaining such large-scale training data are usually prohibitive due to the high cost of data annotations. Without new annotated data, existing dense retrieval methods can hardly adapt to a target domain that is different from their original training corpus, which usually results in poor performance on the target dataset. This is often refer to as the out-of-domain (OOD) problem of dense retrieval. Therefore, how to construct effective dense retrieval models on a specific target domain without using annotated data, i.e., zero-shot dense retrieval, has become an important question to the IR community (Thakur et al. 2021).

To tackle this problem, previous studies on zero-shot DR and domain adaptation mostly adopt different training algorithms with data augmentation techniques to improve the performance of dense retrieval models in OOD scenarios. Specifically, they first extract weak or pseudo supervision data with data augmentation methods on the target domain, and then train dense retrieval models with both the source domain (which provides the original large-scale annotated data for the dense retrieval model) and the target domain together. A variety of data augmentation methods and training algorithms have been explored, including unsupervised Sequence Contrastive Learning (SCL) (Yu et al. 2022) and Inverse Cloze Task (ICT) (Lee, Chang, and Toutanova 2019; Izacard et al. 2021), which are mainly used for second-stage pre-training, and query generation methods (Wang et al. 2022), which are combined with continuous fine-tuning or knowledge distillation. Despite their differences in algorithm design, most existing zero-shot DR methods assume that, with proper training algorithms, dense retrieval models are mighty enough to model and memorize different language distributions and relevance matching patterns. In other words, there should be no need to explicitly distinguish the data and supervision signals we feed into the dense retrieval model.

However, we consider that such assumptions usually lead to two limitations in OOD scenarios. First, as existing stud-

*corresponding author

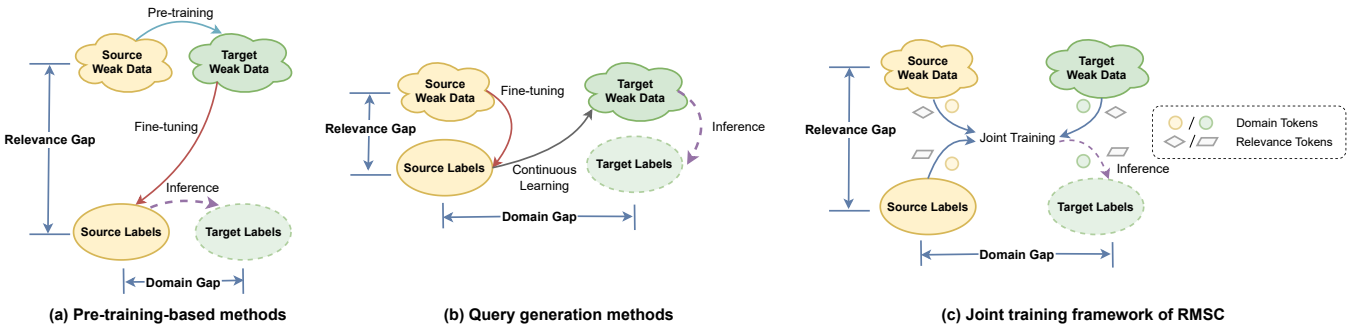

Figure 1: An illustration of previous methods and the proposed RMSC. Sub-figure (a) illustrates the pre-training and fine-tuning process of previous pre-training-based methods. Sub-figure (b) demonstrates the source fine-tuning and target continuous fine-tuning process of previous query generation methods. Sub-figure (c) presents the joint training process of RMSC.

ies do not explicitly tell DR models where the training data came from, DR models tend to model the data from different domains using shared parameters. When the DR models are not powerful enough, which is mostly true in practice, this means that the DR models are likely to focus on the common characteristics between the source and target domain data while ignoring the domain-specific knowledge in the training data. It essentially limits the model's ability to fit the target data, which is problematic when we care about the retrieval performance on the target domain but not the source domain. Second, in OOD scenarios, we have imbalanced prior knowledge for relevance matching on the source and target domain. Human annotated relevance data is easily accessed on the source domain while only pseudo/weak supervision signals are available on the target domain. The goal of domain adaptation methods is to learn a DR model that can predict relevance like human (i.e., predict the human annotated relevance) on the target domain, but the human annotated relevance labels are only observed on the source domain. As shown in Figure 1, previous pre-training-based methods such as Contriever follow a training paradigm that couples the source and target domains, while distinguishing pseudo supervision signals and human annotated data by dividing the training process into pre-training and fine-tuning (Figure 1a). Without proper treatments, it's difficult to teach the DR model to capture the relevance patterns expressed in the human labeled data without overfitting the data patterns of the source domain. Different from pre-training-based methods, QGen and GPL ignore the differences between pseudo supervision data and human labeled data, and distinguish domains by first training on the source labels, and then continuous learning or knowledge distillation on the target pseudo supervision data (Figure 1b). Similarly, it's difficult to teach the DR model to capture the data patterns of the target domain without overfitting the relevance patterns of pseudo supervision data. In the inference process, the DR models are directly used on the target domain without knowing which data are feed and what types of relevance it should looking for. Therefore, as shown in this paper, dense retrieval models usually have unstable target domain performance during training and are highly sensitive to the settings of training steps and early stop criteria.

In order to tackle those problems and develop reliable training algorithms for zero-shot DR, we propose RMSC, a Robust Multi-Supervision Combining strategy for dense retrieval. Similar to existing studies, RMSC follows a data-augmentation-plus-joint-training paradigm to train DR models on both the target domain weak/pseudo supervision data and the source domain human labeled data. However, compared to existing methods that implicitly or explicitly prevent the DR models from distinguishing the source/target domain and the human/pseudo relevance supervision data, we design soft tokens to tell the DR models how the domain data and supervision signals are combined in the training data. As shown in Figure 1c, RMSC decouples domains and supervision signals by explicitly telling the DR models about the current input data combination with soft tokens. In the inference process, we can simply feed the tokens representing "the target domain" and "human annotated relevance" to guide the DR model to retrieve desired documents. With the soft tokens to store the domain-specific and relevance-specific knowledge, RMSC allows the joint training of data from different domains and relevance types, thus enabling DR models to focus more on general relevance matching rather than overfitting the data patterns of a specific domain.

To verify the effectiveness of the proposed RMSC, we conduct extensive experiments on publicly available zero-shot DR benchmarks and compare RMSC against a wide range of existing dense retrieval models and domain adaptation methods. Experimental results show that: 1) RMSC significantly improves the retreival performance on the target domain. 2) RMSC is stable during training and can substantially outperform other dense retrieval methods without sophisticated training strategies and negative mining techniques. 3) RMSC is generic as it can adopt and combine any data augmentation methods.

2 Related Works

In this section, we recap related work in dense retrieval and domain adaptation.

Dense Retrieval

Dense retrieval models encode the query and the document into dense vectors and use nearest neighbor search to retrieve documents. Earlier works focused on exploring training strategies for DR models, such as hard negative mining techniques (Xiong et al. 2020; Zhan et al. 2021a,b) and knowledge distillation from a strong cross-encoder (Qu et al. 2020; Hofstätter et al. 2021; Lin, Yang, and Lin 2021). Recent works investigate how to perform retrieval-oriented second-stage pre-training on large language models (Gao and Callan 2021a,b; Izacard et al. 2021; Liu and Shao 2022; Li et al. 2023; Dong et al. 2023).

Zero-Shot Dense Retrieval

Thakur et al. (Thakur et al. 2021) collected the BEIR benchmark, which consists of diverse retrieval tasks from different domains. They evaluated dense retrieval models and challenged the generalization ability of DR models. Many later works follow this Zero-shot DR setting, where the DR model is trained using a diverse and richly supervised retrieval dataset and then evaluated on the out-of-domain search corpus, which is usually accessible during training.

Domain Adaptation

Previous work on domain adaptation for IR can be roughly divided into two categories. The first category mainly focuses on query generation techniques. These methods generate additional auxiliary training data on the target domain with the help of a well-tuned generation model (Ma et al. 2021), or further employs a powerful cross-encoder to generate pseudo-labels for distillation (Wang et al. 2022). The second category is retrieval-oriented pre-training. These methods employ pre-training tasks specially designed to improve the retrieval performance of DR models. Condenser (Gao and Callan 2021a) and coCondenser (Gao and Callan 2021b) enhance the representation of [CLS] token and introduces the Sequence Contrastive Learning (SCL) task for pre-training. Contriever (Izacard et al. 2021) pre-train DR models on crawled large-scale web pages with Inverse Cloze Task (ICT) and Independent cropping task. COCO-DR (Yu et al. 2022) continues to use SCL on target domain for pre-training and leverages implicit DRO during fine-tuning to improve model robustness.

3 Problem Formulation

Dense Retrieval

Dense retrieval models encode the query and the document into dense vectors and use nearest neighbor search to retrieve documents. With the development of large-scale pre-trained language models such as BERT, advanced dense retrieval models in recent years have followed the Transformer's structure. More specifically, given a query q and a document d, the text encoder f represents them as dense vectors and use inner product to model relevance:

$$s(q, d) = \langle f(q; \theta), f(d; \theta) \rangle \tag{1}$$

Here, θ denotes the parameter of the text encoder.

Zero Shot DR

Different from in-domain Dense Retrieval, where a substantial amount of supervised signals are available, zero-shot DR designates a retrieval task in the absence of manually labeled relevance signals. In many practical scenarios where search systems need to be built, it is costly to acquire a large amount of manual annotations. For example, medical, legal, and other specialized fields require relevant professionals to complete the annotation, or in personalized scenarios, manual annotation may involve user privacy issues. Therefore, zero-shot DR challenges to construct an effective DR model when the target domain lacks manually labeled relevance signals.

In fact, not only the relevance annotation of query-document pairs is unavailable in the target domain, but also the query set is scarce in most cases. The only information that can be easily accessed in large quantities is the document collection of the target domain, namely the target corpus. Therefore, we focus on how to build a retrieval system with strong generalization performance where the target domain corpus is accessible and the supervised training signals only come from the source domain.

Instablility of Target Performance

We first fine-tune dense retrieval models and observe the out-of-domain performance during the training process. Previous work has shown that the mined hard negative samples significantly improve the retrieval performance of dense retrieval models. Common negative sampling strategies include using top irrelevant documents from unsupervised BM25 or mining from the previous episode of DR models(Xiong et al. 2020; Zhan et al. 2021a). However, for simplicity and to avoid the impact of different negative sampling strategies, the simplest random negative sampling method is applied in this paper.

Following standard DR training process, we select three representative DR models: Condenser, QGen, and COCO-DR. We use the source domain human labeled data (MS-MARCO) to fine-tune Condenser and COCO-DR, and use the target domain generated pseudo data to continuous fine-tune QGen. Figure 2 exhibits the performance of different training steps on five BEIR datasets from different domains. As for Condenser and COCO-DR, we can observe that during the training process, the retrieval performance on MS-MARCO (in-domain) keeps improving, while the performance on BEIR (out-of-domain) starts oscillating and even decreasing after the initial increase. We believe that the DR models can learn general relevance matching patterns from the supervised signals, which leads to the initial increase on out-of-domain performance. However, the DR models will inevitably overfit to the domain features and data patterns of the source domain during the succeeding process of training, which causes the degradation of the out-of-domain performance. As for QGen, we can see that the in-domain performance keeps decreasing, since the DR models are fine-tuned with target pseudo data. However, the out-of-domain performance also oscillates, and we attribute this to the DR models overfitting to the relevance patterns of pseudo super-

Figure 2: The performance of Condenser, COCO-DR and QGen over different training steps on 5 of BEIR datasets. The blue dotted line indicates the nDCG@10 on MSMARCO (in-domain) and the orange solid line indicates the performance on the corresponding BEIR dataset. The x-axis in the third row is not aligned with the first two rows because QGen uses the target domain generated pseudo data for continuous fine-tuning.

vision data, while the out-of-domain performance requires human annotated relevance.

4 Proposed Method

In order to tackle the problems above and develop reliable training algorithms for zero-shot DR, we propose RMSC, a Robust Multi-Supervision Combining strategy for dense retrieval. RMSC jointly trains weak supervision signals extracted from the target domain corpus and human labeled data from the source domain, and constructs context-aware vector representations with designed soft tokens. This section describes the components of RMSC in detail.

Weak Supervision Extraction

In this paper, we consider weak supervision to be a technique where training labels are obtained automatically without human annotators or any external resources (e.g., click data). Many works have used unsupervised ranking models such as BM25 to generate pseudo-labels for specific query sets and document sets as weak supervision signals (Dehghani et al. 2017). However, since these methods require a given query set, which is scarce under the zero-shot DR setting, we consider using existing document-based unsupervised methods, which have been widely explored in works on second-stage pre-training for IR tasks, or query generation models to construct training data when only corpus is available. Formally, given a document d_i, we unsupervisedly extract weak supervision signals as:

$$\hat{q}_i, \hat{d_{i+}} = h_q(d_i), h_d(d_i) \tag{2}$$

Here $\hat{q}$ and $\hat{d_+}$ denote the query and the positive document extracted. $h_q(\cdot)$ and $h_d(\cdot)$ denote the corresponding extraction method. Notice that we only extract positive documents and not negative ones, since we follow the general format of retrieval task annotations, where the supervision signals are presented in the form of query and positive document pairs.

Specifically, we explored the following three extraction methods:

Sequence Contrastive Learning (SCL) improves the alignment and the uniformity of the text sequences embedding space over the target search corpus in a query-agnostic manner, which has been proved to be beneficial to in-domain dense retrieval models. For each document, SCL splits it into text spans and randomly extracts two spans to form a training pair.

Inverse Cloze Task (ICT) selects a sentence in the document as a pseudo question, and its context is treated as pseudo evidence, or the positive document. Given a pseudo-question, ICT requires selecting the corresponding pseudo-evidence out of the candidates in a batch.

Query Generation (QGen) generates relevant queries for a certain document by employing a well-tuned generation model. The generated data are usually used directly to further fine-tune a DR model for domain adaptation. Since the generation model is not directly built based on the actual queries on the target domain, it can be considered as weak/pseudo supervision on the target domain.

Soft Tokens and Context-aware Representation

After weak supervision extraction of the target corpus, we can get the training set $\{\hat{q}_i^t, \hat{d}_{i+}^t\}_{i=1}^N$, together with the supervised signals of the source domain as $\{q_i^s, d_{i+}^s\}_{i=1}^M$. With the weak supervision training set, we can keep the target domain information continuously accessible to the DR model during the training process with the source labels, thus alleviating the model's potential overfitting to the source domain features while improving the generalization ability specially to the target domain.

However, we argue that it is sub-optimal to directly mix the two training sets at the data level in a multi-task training manner. This is attributed to the existence of two gaps in these two training sets. Firstly, these two sets of data are from the source domain and the target domain respectively, which differ in domain features such as term distribution and language pattern. Secondly, the relevant query-document pair of weak supervision is generated by the unsupervised method, reflecting semantic text similarity, while the source labels annotated by humans reflecting relevance under real information needs, thus they could be different in terms of relevance matching patterns. Direct joint training would require the DR model to process text inputs that follow a mixed distribution and to model two related but not identical training objectives simultaneously, which may conflict with each other and lead to ranking performance degradation on the target domain.

To address such challenge, RMSC resorts to the special token technique, which is widely used in the Transformer architecture language models. Two special tokens, [CLS] and [SEP], are first added to the beginning and end of a text sequence input $x = [x_1, x_2, ..., x_l]$, which is then passed through the Transformer backbone. The [CLS] representation from the last layer is considered to aggregate the information of the entire text sequence, which directly serves as the dense vector in the Transformer-based DR models:

$$v = f(x; \theta) = \text{Transformer}([\text{CLS}, x, \text{SEP}]; \theta) \quad (3)$$

Specifically, RMSC introduces two sets of special soft tokens, namely, domain tokens and relevance tokens. The domain tokens enable the DR model to explicitly distinguish text inputs from different domains, which substantially reduces the modeling difficulty and enables the DR models to focus on the common characteristics between the source and target domain data while keeping the domain-specific knowledge by leveraging the soft tokens. The relevance tokens allow the model to tackle the two different training objectives in a uniform manner, which can avoid potential conflicts between weak supervision and human labeled data.

Formally, RMSC employs k independent domain tokens, where k is set as a hyper-parameter, for source and target domain respectively:

$$\text{source domain}: \ [S_1], [S_2], ..., [S_k] \quad (4)$$
$$\text{target domain}: \ [T_1], [T_2], ..., [T_k] \quad (5)$$

Similarly, k independent relevance tokens are utilized to model weak supervision and human annotated labels respectively:

$$\text{weak supervision}: \ [W_1], [W_2], ..., [W_k] \quad (6)$$
$$\text{human labels}: \ [H_1], [H_2], ..., [H_k] \quad (7)$$

RMSC then combines the text inputs with the special tokens, and thus construct context-aware representations:

$$v_x = f([S_1]...[S_k], x, [H_1]...[H_k]; \theta) \quad (8)$$
$$v_{\hat{x}} = f([T_1]...[T_k], \hat{x}; [W_1]...[W_k]; \theta) \quad (9)$$

where x denotes text inputs from the source supervised training set, $x \in \{q_i^s, d_{i+}^s\}_{i=1}^M$ and $\hat{x}$ denotes text inputs from the target weak supervision training set, $x \in \{\hat{q}_i^t, \hat{d}_{i+}^t\}_{i=1}^{N_t}$, while θ is the parameter of the Transformer backbone.

Joint Training and Inference

In order to better mitigate the two gaps between the two parts of data mentioned above, RMSC employs exactly the same unsupervised method on the source domain to extract weak supervision and constructs corresponding context-aware representations as:

$$\{\hat{q_i^s}, d_{i+}^{\hat{s}}\}_{i=1}^{N_s} = h(D_s) \quad (10)$$
$$v_{\hat{x}} = f([S_1]...[S_k], \hat{x}, [W_1]...[W_k]; \theta) \quad (11)$$

where $D_s = \{d_i^s\}_{i=1}^{N_s}$ denotes the source corpus and N_s is the number of documents in the corpus, while $\hat{x}$ denotes text inputs from the source weak supervision training set. We thus obtain three sets of the training data, $\{q_i^s, d_{i+}^s\}_{i=1}^M$, $\{\hat{q_i^s}, d_{i+}^{\hat{s}}\}_{i=1}^{N_s}$, $\{\hat{q}_i^t, \hat{d}_{i+}^t\}_{i=1}^{N_t}$, which are source supervised signals and the weak supervision extracted from the two domains. Among these three sets of training data, the two weak supervision sets differ only in the domain features, since they are extracted by the same method. Meanwhile, the source supervised signals and the source weak supervision differ only in the relevance matching signals. By comparing the former, we expect the DR model to learn the shared domain features into the backbone parameters, while storing the information of the distinct parts into the embedding of the corresponding domain tokens. By comparing the latter, the DR model can integrate the general matching signals into the backbone parameters and utilize the embedding of the relevance tokens to represent the different relevance matching signals. We can control the size of the weaker supervision set such that it is comparable to the size of the source supervised signals. We then mix the three sets of training data and treat them as a unified dataset.

During the training stage, we adopt random negative sampling and in-batch negative techniques. We choose this strategy mainly to rule out the potential influence of different training strategies and hard negative mining techniques when comparing RMSC with other zero-shot DR algorithms. Please note that other training and negative sampling strategies are also applicable to RMSC and could potentially lead to better experiment performance, but this is not the focus of this paper. For each query-document pair q_i, d_i, we randomly select a different document d_j, $j \neq i$

from the corpus as the negative. Then we can optimize the following ranking loss:

$$\mathcal{L}(\theta) = \sum_{i=1}^{B} \frac{S_i^+}{S_i^+ + \sum_{\phi_i=\phi_j, \xi_i=\xi_j} S_{ij}^-} \tag{12}$$

$$S_i^+ = \exp\left(s(q_i, d_{i+}; \theta)\right) \tag{13}$$

$$S_{ij}^- = \exp\left(s(q_i, d_{j-}; \theta)\right) \tag{14}$$

where B denotes the training batch size and $s(q, d; \theta)$ is the context-aware score of query and document:

$$s(q, d; \theta) = \langle v_q, v_d \rangle \tag{15}$$

$$= \langle f(\phi, q, \xi; \theta), f(\phi, d, \xi; \theta) \rangle \tag{16}$$

Here, ϕ and ξ denote the domain tokens and relevance tokens corresponding to the given query-document pair, respectively. Notice that in the loss function, the in-batch negatives is only calculated for samples with the same domain and relevance matching patterns. Thus, RMSC can integrate weak supervision and source labels by leveraging joint training.

During the inference stage, we set the domain tokens to target domain tokens and the relevance tokens to human label tokens.

$$\tilde{s}(q, d; \theta) = \langle \tilde{v_q}, \tilde{v_d} \rangle \tag{17}$$

$$\tilde{v_q} = f([T_1]...[T_k], q, [H_1]...[H_k]; \theta) \tag{18}$$

$$\tilde{v_d} = f([T_1]...[T_k], d, [H_1]...[H_k]; \theta) \tag{19}$$

In this way, RMSC can effectively combine the matching signals learned from source labels with the domain features learned from target weak supervision, thus achieving a promising retrieval performance on the target corpus.

5 Experiment Setup

In this section, we present our experimental settings, including datasets, baselines, and implementation details.

Dataset and Metrics

Following recent zero-shot DR research, we use the MS MARCO Passage Ranking dataset (Nguyen et al. 2016) as the source domain, with a corpus of 8.8M passages from web pages and 0.5M training queries. Each training query is coupled with a manually labeled positive passage, which together constitute the source supervised signals.

As for the out-of-domain test sets, we select the BEIR dataset (Thakur et al. 2021) and the Lotte benchmark (Santhanam et al. 2021). BEIR is a heterogeneous evaluation benchmark for information retrieval and contains 18 datasets from diverse text retrieval tasks, from which we select publicly available datasets to conduct experiments. The Lotte benchmark consists of a collection of questions and answers sourced from the StackExchange platform, which are then divided into five distinct topics: writing, recreation, science, technology, and lifestyle. In this benchmark, the relevance of the answers is determined by their acceptance status or upvote count on the original platform.

To measure the retrieval performance of DR models, we use NDCG@10 as the evaluation metrics. We also report Recall@100, Recall@1000 to reflect the retrieval capacity over the entire search corpus.

Baselines

We consider various baselines, including standard sparse and dense retrieval models. For sparse models, we select BM25(Robertson and Walker 1994) as representative. For dense models, we select representative dense retrieval models as baselines. Details can be seen in the Appendix.

6 Experiments

Now we empirically evaluate the proposed RMSC to address the following three research questions:

- **RQ1**: Can RMSC substantially improve the retrieval performance over DR models in the target domain?
- **RQ2**: Is RMSC stable during training and is RMSC scale with other methods?

Main Results

This section compares our model with other baselines on Lotte dataset to answer RQ1.

We initialize RMSC with Condenser and COCO-DR, namely RMSC(CD) and RMSC(CO) respectively, and summarize the retrieval performance of different baseline models in Table 1. Note that the pre-trained COCO-DR model is fine-tuned with random negative sampling strategy for fair comparison. According to the results, RMSC outperforms all baseline models of BERT-base scale on the Lotte dataset, which confirms the efficacy of the proposed RMSC method.

Compared to DR baselines such as DPR, ANCE, TAS-B and Condenser, RMSC considerably improves the retrieval performance by a large margin, and on average boosts 14% and 20% over the best ANCE on NDCG@10 and Recall@1000, respectively. Notice that although ANCE and TAS-B employ sophisticated training strategies such as self-mined hard negatives and knowledge distillation, RMSC still outperforms them with a simple random negative sampling during training. Meanwhile, due to the relevance gap discussed above, weak supervision, to a certain extent, can be viewed as supervised signals with noise. This can also explain that RMSC improves more on the coarse-grained Recall metrics compared to NDCG@10, which requires the fine-grained top ranking ability.

As for coCondenser and Contriever, which both utilize unsupervised contrastive learning, their performance is dramatically enhanced compared to other DR models. Since Contriever is pre-trained on a large amount of crawled web pages with mixed domains, its generalization ability and performance is slightly superior compared to coCondenser. Nevertheless, RMSC improves 5.6% and 10% over Contriever on NDCG@10 and Recall@1000, which further illustrates the benefit of introducing target domain-specific weak supervision.

As for COCO-DR, similar to Contriever, it also uses unsupervised contrastive learning, but directly on the target corpus, thus its retrieval capability is marginally stronger. When initialized with COCO-DR, RMSC can further improve the retrieval performance and we attribute this to that the joint training methodology can effectively mitigate the overfitting problem. The improvement also indicates that RMSC can be

Datasets	Metrics	Sparse	Dense						Ours	
		BM25	DPR	ANCE	TAS-B	Cond	Contr.	COCO	RMSC(CD)	RMSC(CO)
Lotte-Wri	N@10	0.352	0.357	0.388	0.346	0.373	0.424	0.422	$\underline{0.466}^{\ddagger\natural}$	$\mathbf{0.474}^{\ddagger\natural}$
	R@100	0.541	0.547	0.573	0.564	0.593	0.639	0.656	$\underline{0.723}^{\ddagger\natural}$	$\mathbf{0.732}^{\ddagger\natural}$
	R@1000	0.681	0.705	0.704	0.712	0.741	0.777	0.806	$\underline{0.866}^{\ddagger\natural}$	$\mathbf{0.871}^{\ddagger\natural}$
Lotte-Life	N@10	0.305	0.385	0.423	0.406	0.384	0.459	0.427	$\underline{0.463}^{\ddagger}$	$\mathbf{0.470}^{\ddagger\flat}$
	R@100	0.548	0.648	0.680	0.686	0.668	0.746	0.718	$\underline{0.784}^{\ddagger\natural}$	$\mathbf{0.794}^{\ddagger\natural}$
	R@1000	0.754	0.848	0.855	0.872	0.861	0.907	0.895	$\underline{0.939}^{\ddagger\natural}$	$\mathbf{0.946}^{\ddagger\natural}$
Lotte-Rec	N@10	0.325	0.362	0.403	0.390	0.376	0.443	0.424	$\underline{0.462}^{\ddagger\natural}$	$\mathbf{0.475}^{\ddagger\natural}$
	R@100	0.570	0.598	0.629	0.651	0.629	0.709	0.689	$\underline{0.764}^{\ddagger\natural}$	$\mathbf{0.778}^{\ddagger\natural}$
	R@1000	0.736	0.770	0.775	0.811	0.801	0.857	0.849	$\underline{0.903}^{\ddagger\natural}$	$\mathbf{0.906}^{\ddagger\natural}$
Lotte-Sci	N@10	0.156	0.131	0.149	0.133	0.123	0.158	$\underline{0.159}$	0.157	$\mathbf{0.176}^{\ddagger\natural}$
	R@100	0.303	0.249	0.282	0.281	0.252	0.304	0.326	$\underline{0.346}^{\ddagger\natural}$	$\mathbf{0.379}^{\ddagger\natural}$
	R@1000	0.498	0.418	0.453	0.463	0.427	0.497	0.556	$\underline{0.578}^{\ddagger\natural}$	$\mathbf{0.632}^{\ddagger\natural}$
Lotte-Tech	N@10	0.151	0.155	0.188	0.167	0.157	0.199	0.205	$\underline{0.220}^{\dagger\flat}$	$\mathbf{0.234}^{\ddagger\natural}$
	R@100	0.335	0.318	0.367	0.360	0.345	0.415	0.445	$\underline{0.491}^{\ddagger\natural}$	$\mathbf{0.526}^{\ddagger\natural}$
	R@1000	0.539	0.524	0.576	0.595	0.567	0.660	0.708	$\underline{0.753}^{\ddagger\natural}$	$\mathbf{0.787}^{\ddagger\natural}$
Avg	N@10	0.258	0.278	0.310	0.288	0.282	0.336	0.327	$\underline{0.353}^{\ddagger\natural}$	$\mathbf{0.366}^{\ddagger\natural}$
	R@100	0.459	0.472	0.506	0.508	0.497	0.562	0.566	$\underline{0.621}^{\ddagger\natural}$	$\mathbf{0.642}^{\ddagger\natural}$
	R@1000	0.641	0.653	0.672	0.690	0.679	0.739	0.762	$\underline{0.807}^{\ddagger\natural}$	$\mathbf{0.828}^{\ddagger\natural}$

Table 1: Results of RMSC and different baseline models on the Lotte dataset. $\flat/\natural$ and $\dagger/\ddagger$ indicates statistically significant results over the strongest baselines Contriever and COCO-DR with $p < 0.05/0.01$, respectively. The best results are marked bold. The second-best results are underlined.

combined with continues contrastive pre-training introduced in COCO-DR.

We also conduct experiments on the BEIR benchmark. The results can be found in the Appendix.

Training Robustness

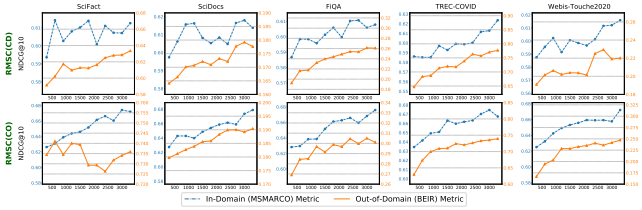

Figure 3: Performance of RMSC(CD) and RMSC(CO) over different training steps on 5 of BEIR datasets. The blue dotted line indicates the nDCG@10 on MSMARCO (in-domain) and the orange solid line indicates the performance on the corresponding BEIR dataset.

In order to study the training robustness of RMSC, we plot the performance of RMSC on MSMARCO and on BEIR datasets during the training process. As we can see in Figure 3, on most datasets, the retrieval performance on MSMARCO (in-domain) keeps improving, while the performance on BEIR (out-of-domain) maintains a steady upward trend without the instability and fluctuation during direct fine-tuning. We attribute this to that RMSC utilizes target weak data and joint training methodology, thus alleviating the overfitting to the source domain features during training, which further validates the stability and robustness of the proposed RMSC.

7 Conclusions and Future Work

In this paper, we propose RMSC, which combines multiple supervision signals for robust zero-shot dense retrieval. RMSC follows a data-augmentation-plus-joint-training paradigm to train DR models, while decouples the domain and supervision signals by explicitly telling the DR models how the domain data and supervision signals are combined in the training data with specially designed soft tokens. With the soft tokens as extra memory, RMSC enables the models to focus not only on common characteristics, but also on domain-specific and supervision-specific knowledge, instead of ignoring them. Meanwhile, the introduction of joint training with target weak supervision largely eliminates the problem of model overfitting during training. Finally, at the inference stage, RMSC can guide the DR models to retrieve desired documents on the target corpus through employing corresponding soft tokens.

Experiment results show that RMSC can outperform other dense retrieval methods and domain adaptation methods, without employing sophisticated training strategies. However, we believe that these strategies, including hard negative mining, knowledge distillation and second-stage pre-training, can also be combined with RMSC, which we leave for future research.

Datasets	Sparse	Dense							Ours	
	BM25	ANCE	TAS-B	GenQ	GPL	Cond.	Contr.	COCO	RMSC(CD)	RMSC(CO)
TREC-C	0.656	0.654	0.482	0.619	0.700	0.739	0.611	0.743	**0.812**‡	0.772
HotpotQA	0.603	0.456	0.584	0.534	0.582	0.537	**0.638**	0.585	0.455	0.583
FiQA	0.236	0.295	0.300	0.308	**0.344**	0.256	0.329	0.304	0.317†	0.325‡
Touche	**0.367**	0.240	0.162	0.182	0.255	0.175	0.209	0.198	0.245‡	0.252‡
NFCorpus	0.325	0.237	0.319	0.319	0.345	0.274	0.328	**0.357**	0.308	0.355
NQ	0.329	0.446	0.463	0.358	0.483	0.435	**0.498**	0.479	0.391	0.464
DBPedia	0.313	0.281	0.384	0.328	0.384	0.343	**0.413**	0.380	0.340	0.377
SciFact	0.575	0.507	0.640	0.644	0.674	0.583	0.678	0.734	0.649	**0.742**
SciDocs	0.158	0.122	0.148	0.143	0.169	0.141	0.163	0.161	0.186‡	**0.192**‡
Quora	0.789	0.852	0.835	0.830	0.836	0.851	**0.865**	0.864	0.839	0.859
Fever	0.753	0.669	0.700	0.669	**0.759**	0.683	0.758	0.725	0.715	0.748
C-Fever	0.213	0.198	0.228	0.175	0.235	0.221	0.237	0.199	**0.237**†	0.214†
ArguAna	0.414	0.415	0.429	0.493	0.557	0.345	0.446	0.456	0.515‡	**0.564**‡
Avg.	0.441	0.413	0.436	0.431	0.486	0.429	0.475	0.476	0.462	**0.497**†

Table 2: Comparison to DR baselines and domain adaptation methods on the BEIR dataset. We report NDCG@10 in the table. †/‡ indicates statistically significant results over the strongest baseline COCO-DR with $p < 0.05/0.01$. The best results are marked bold. The second-best results are underlined.

A Baselines

We consider various baselines, including standard sparse and dense retrieval models. For sparse models, we select BM25(Robertson and Walker 1994) as representative. For dense models, we consider three types of baselines.

Dense retrieval methods: The models of this type are standard dual-encoder structures. ANCE (Xiong et al. 2020) is trained using hard negatives retrieved from the previous epoch. TAS-B (Hofstätter et al. 2021) employs knowledge distillation to improve the retrieval performance. Although it is not fair to compare ANCE and TAS-B with our method since we don't employ sophisticated training strategies, we still report the results for reference.

Domain adaption methods: We also compare RMSC with a variety of domain adaption methods. Contriever (Izacard et al. 2021) pre-trains DR models on crawled large-scale web pages with Inverse Cloze Task (ICT) and Independent cropping task. Condenser (Gao and Callan 2021a) enforces the [CLS] token to be the information bottleneck to aggregate attention over the entire input texts during pre-training. Thereby, the DR model can better capture the semantic information using the representation of the [CLS] token. COCO-DR (Yu et al. 2022) continues to use Span Contrastive Learning (SCL) on target domain for pre-training to alleviate the distribution shift, while employing implicit DRO during fine-tuning to improve model robustness.

Query generation methods: QGen (Ma et al. 2021) uses a T5-based query generation model, which is fine-tuned on MS MARCO, to generate 5 queries for each document in the target corpus as additional training data. A well-tuned DR model is then fine-tuned on the generated data for domain adaptation. Following QGen, GPL (Wang et al. 2022) first generates queries for documents in the target corpus, and then leverages an additional cross-encoder to score the generated query-document pairs as pseudo labels for efficient denoising, which enables training with mined hard negatives.

B Implementation Details

We build our models based on PyTorch Framework and HuggingFace Library. Details can be found in the supplementary materials. When implementing RMSC, the number of special tokens k is set to 1, and each special token is randomly initialized. For weak supervision extraction, we use the NLTK library to divide a document into sentences. As for the training settings of RMSC, we adopt random negative sampling and select one negative document for each query. The ICT method is used for the Lotte dataset and SCL for the BEIR dataset. We use AdamW optimizer during training. The batch size is set to 512 and the learning rate is set to 1e-5. We use the same random negative sampling strategy to fine-tune Condenser, coCondenser and COCO-DR for fair comparison.

C Results on BEIR

We report the retrieval performance of different DR baselines and a variety of domain adaptation methods on BEIR datasets in Table 2. It should be noted that the performance of COCO-DR differs from that reported in the original paper. This is due to that we utilized random negative sampling during training instead of the iDRO and hard negative techniques used in the original COCO-DR paper, which we believe is more fair and appropriate for our research objectives.

Acknowledgments

This work is supported by Quan Cheng Laboratory (Grant No. QCLZD202301) and Huawei Poisson Lab.

References

Dehghani, M.; Zamani, H.; Severyn, A.; Kamps, J.; and Croft, W. B. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 65–74.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, Q.; Liu, Y.; Ai, Q.; Li, H.; Wang, S.; Liu, Y.; Yin, D.; and Ma, S. 2023. I3 Retriever: Incorporating Implicit Interaction in Pre-Trained Language Models for Passage Retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, 441–451. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701245.

Gao, L.; and Callan, J. 2021a. Condenser: a pre-training architecture for dense retrieval. *arXiv preprint arXiv:2104.08253*.

Gao, L.; and Callan, J. 2021b. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*.

Gulrajani, I.; and Lopez-Paz, D. 2021. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*.

Guo, J.; Cai, Y.; Fan, Y.; Sun, F.; Zhang, R.; and Cheng, X. 2021. Semantic models for the first-stage retrieval: A comprehensive review. *arXiv preprint arXiv:2103.04831*.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360. Online: Association for Computational Linguistics.

Hofstätter, S.; Lin, S.-C.; Yang, J.-H.; Lin, J.; and Hanbury, A. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 113–122.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Towards unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Lee, K.; Chang, M.-W.; and Toutanova, K. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6086–6096. Florence, Italy: Association for Computational Linguistics.

Li, H.; Ai, Q.; Chen, J.; Dong, Q.; Wu, Y.; Liu, Y.; Chen, C.; and Tian, Q. 2023. SAILER: Structure-Aware Pre-Trained Language Model for Legal Case Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, 1035–1044. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394086.

Lin, J.; Nogueira, R.; and Yates, A. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4): 1–325.

Lin, S.-C.; Yang, J.-H.; and Lin, J. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, 163–173. Online: Association for Computational Linguistics.

Liu, Z.; and Shao, Y. 2022. RetroMAE: Pre-training Retrieval-oriented Transformers via Masked Auto-Encoder. *arXiv preprint arXiv:2205.12035*.

Ma, J.; Korotkov, I.; Yang, Y.; Hall, K.; and McDonald, R. 2021. Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1075–1088. Online: Association for Computational Linguistics.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Qu, Y.; Ding, Y.; Liu, J.; Liu, K.; Ren, R.; Zhao, W. X.; Dong, D.; Wu, H.; and Wang, H. 2020. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

Robertson, S. E.; and Walker, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*, 232–241. Springer.

Santhanam, K.; Khattab, O.; Saad-Falcon, J.; Potts, C.; and Zaharia, M. 2021. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. arXiv:2112.01488.

Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; and Gurevych, I. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Wang, K.; Thakur, N.; Reimers, N.; and Gurevych, I. 2022. GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2345–2360. Seattle, United States: Association for Computational Linguistics.

Xiong, L.; Xiong, C.; Li, Y.; Tang, K.-F.; Liu, J.; Bennett, P.; Ahmed, J.; and Overwijk, A. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Yu, Y.; Xiong, C.; Sun, S.; Zhang, C.; and Overwijk, A. 2022. COCO-DR: Combating Distribution Shifts in Zero-Shot Dense Retrieval with Contrastive and Distributionally Robust Learning. *arXiv preprint arXiv:2210.15212*.

Zhan, J.; Mao, J.; Liu, Y.; Guo, J.; Zhang, M.; and Ma, S. 2021a. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1503–1512.

Zhan, J.; Mao, J.; Liu, Y.; Guo, J.; Zhang, M.; and Ma, S. 2021b. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1503–1512.