Domain Adaptation for Dense Retrieval and Conversational Dense Retrieval through Self-Supervision by Meticulous Pseudo-Relevance Labeling

Minghan Li[1,*], **Eric Gaussier**[2]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]Univ. Grenoble Alpes, CNRS, LIG, Grenoble, France
mhli@suda.edu.cn, eric.gaussier@imag.fr

Abstract

Recent studies have demonstrated that the ability of dense retrieval models to generalize to target domains with different distributions is limited, which contrasts with the results obtained with interaction-based models. Prior attempts to mitigate this challenge involved leveraging adversarial learning and query generation approaches, but both approaches nevertheless resulted in limited improvements. In this paper, we propose to combine the query-generation approach with a self-supervision approach in which pseudo-relevance labels are automatically generated on the target domain. To accomplish this, a T5-3B model is utilized for pseudo-positive labeling, and meticulous hard negatives are chosen. We also apply this strategy on conversational dense retrieval model for conversational search. A similar pseudo-labeling approach is used, but with the addition of a query-rewriting module to rewrite conversational queries for subsequent labeling. This proposed approach enables a model's domain adaptation with real queries and documents from the target dataset. Experiments on standard dense retrieval and conversational dense retrieval models both demonstrate improvements on baseline models when they are fine-tuned on the pseudo-relevance labeled data. Source code is available at https://github.com/lmh0921/DoDress.

Keywords: dense retrieval, domain adaptation, conversational search

1. Introduction

Neural information retrieval (IR) has significantly improved IR systems through deep neural networks. It can be classified into two categories: interaction-based and representation-based (dense retrieval) approaches (Guo et al., 2020). While interaction-based models generally outperform dense retrieval models, the latter are preferred if one needs to deploy a model at large scale due to their speed advantage. However, recent studies, such as BEIR (Thakur et al., 2021), have shown that dense retrieval (DR) models trained on a source domain generalize less well than traditional models as BM25 and interaction-based models on out-of-distribution (OOD) data sets. Training on target datasets with gold labels requires expensive annotations, posing limitations in real-world scenarios. Thus, addressing OOD scenarios for dense retrieval is crucial.

Domain adaptation aims to enable a model trained on a source domain to perform well on a target domain without using human labels (Wang and Deng, 2018; Wang et al., 2022a). Several domain adaptation techniques have been proposed for dense retrieval. One approach is through data generation, as demonstrated by QGen (Ma et al., 2021), which generates queries for the target domain using a query generator. However, the synthetic queries may not resemble real target queries. Another approach is domain adversarial learning (Wang et al., 2022a), exemplified by MoDIR (Xin et al., 2022), which adversarially trains a dense retrieval encoder to learn domain-invariant representations. However, such a learning objective may result in poor embedding spaces and unstable performance (Wang et al., 2022b).

In this paper, we address **do**main generalization for **d**ense **re**trieval through **s**elf-**s**upervision by pseudo-relevance labeling (in short, DoDress). We aim to build pseudo-relevance labels on the target domain using interaction-based models solely trained on the source domain, such as T5-3B (Nogueira et al., 2020), acting as re-rankers. This method eliminates the need for human annotations and allows the model to use genuine queries and documents from the target domain. Additionally, we investigate different negative sampling strategies (Zhou et al., 2022) to further enhance the final dense retrieval model on the target domain.

Conversational search has become a prominent research area within information retrieval, involving natural conversations for information retrieval purposes (Zamani et al., 2023; Culpepper et al., 2018). Conversations exhibit contextualization, conciseness, and reliance on prior knowledge, presenting challenges for search systems in accurately understanding information needs. Figure 1 shows an example of conversational search's query format: given a query, for example the third one, the

* Most of the work was done while the author was at Univ. Grenoble Alpes.

LREC-COLING 2024, pages 5247–5259
20-25 May, 2024.

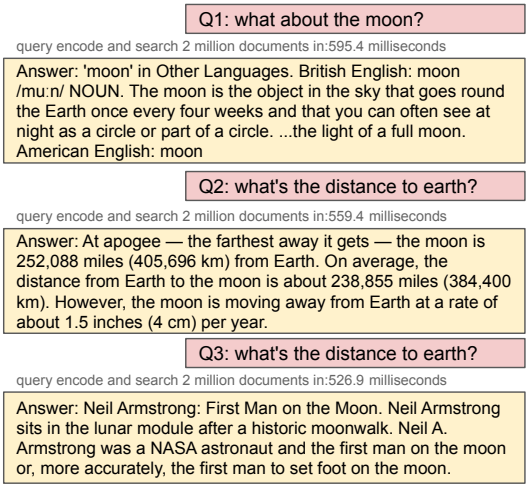

Figure 1: An example of conversational search (with our model deployed, top 1 as answer).

system needs to understand the omission or user intention, by taking account into previous queries. Conversational dense retrieval (CDR) models have been developed to address this problem. However, CDR models require a large amount of training data and annotating relevance labels for target conversational search datasets is expensive.

To overcome these challenges, researchers have proposed approaches that leverage source domain data to mitigate data scarcity, such as (Yu et al., 2021; Lin et al., 2021). However, these approaches do not utilize queries and documents from the target domain, leaving space for improvement when adapting to target domain data. In this paper, we propose a method that combines a query rewriting module with the pseudo-relevance approach for CDR models to alleviate the data and label scarcity issue of the target domain. This enables CDR models that are trained on a source domain like CANARD (Elgohary et al., 2019) to better adaptation on the target domain using pseudo-relevance labels.

Our contributions are threefold: First, we propose a pseudo-relevance labeling approach for a target IR dataset, which can be used to fine-tune a model to adapt better on the target domain. Second, we adopt the effective interaction-based model T5-3B trained on the source domain to generate pseudo-positive labels for the target domain. Besides, we explore different negative sampling strategies to enhance the final DR model. To the best of our knowledge, this is the first attempt to combine and investigate the two strategies. Third, we further apply the pseudo-relevance labeling approach to CDR models for conversational search by incorporating a query rewriting module, which is naturally a further step to apply the proposed pseudo-relevance labeling strategy. This pseudo-relevance data complements CDR models and enables domain adaptation on the target dataset. Experimental results demonstrate the effectiveness of our approach, showing that fine-tuning DR models on the target pseudo-labeled data improves their performance, particularly benefiting the state-of-the-art approach GPL. Furthermore, further training CDR models on the generated training data from the target dataset leads to improved effectiveness (Yu et al., 2021; Lin et al., 2021).

2. Related Work

2.1. Pseudo-Quesries and Pseudo-Labeling for DR or IR

QGen (Ma et al., 2021) proposes a generation approach for zero-shot learning in dense passage retrieval, using synthetic query generation. Similarly, Liang et al. (2020) suggests using synthetic queries for unsupervised domain adaptation in dense passage retrieval. These papers highlight the effectiveness of query generation, which is also utilized in the GPL model (Wang et al., 2022b), leveraging a pre-trained T5 encoder-decoder (Raffel et al., 2020). (Izacard et al., 2021) build positive pairs from a single document, with inverse cloze task, independent cropping and contrastive learning for unsupervised dense information retrieval. Recently, Dai et al. (2022) prompt large language models (LLM) to create queries and train task-specific retrievers. However, they focus on the few-shot setting where a few annotated examples are required and do not focus on domain adaptation. Sun et al. (2021) generate discriminative queries based on contrastive documents. Their approach also focus on the few-shot setting where a small volume of target data is required.

Dehghani et al. (2017) suggest training a neural ranking model with weak supervision, wherein labels are automatically acquired without human annotators. This is achieved by employing the output of an unsupervised ranking model, such as BM25, as a signal for weak supervision. Mokrii et al. (2021) evaluate the transfer ability of BERT-based neural ranking models and use BM25 to generate pseudo-relevance labels. However, these two approaches don't focus on DR models and using only BM25 for pseudo-relevance labels may not be sufficient. Hashemi et al. (2023) assume that IR models have access to a brief textual description that explains the target domain, and produce a synthetic document collection, query set. They then generate a synthetic document collection, query set, and pseudo-relevance labels based on this textual domain description. Qu et al. (2021) propose RocketQA which uses pseudo-labels for data augmentation, but this approach needs human label

and does not focus on the domain adaptation of DR models.

2.2. Conversational Dense Retrieval

Conversational search presents unique challenges (see Section 1). Two commonly proposed approaches for addressing these challenges are query rewriting and conversational dense retrieval (CDR). The query rewriting approach involves a module that rewrites conversational queries into a standard format for better handling by existing information retrieval systems (Mele et al., 2020; Ren et al., 2018; Vakulenko et al., 2021). The second approach is CDR with a query encoder to understand the conversational queries directly. Mao et al. (2022) propose ConvTrans, that transforms web search sessions into conversational search sessions to address data scarcity of CDR. Yu et al. (2021) introduce ConvDR, a teacher-student framework that improves the few-shot ability of CDR by learning from a well-trained ad hoc dense retriever. CQE (Lin et al., 2021) uses annotated queries of the conversational query reformulation dataset CANARD (Elgohary et al., 2019) for the target datasets to train CDR. While these approaches still face domain gaps in the training data.

Another approach is CoSPLADE (Hai Le et al., 2023). The authors train a first-stage ranker based on SPLADE (Formal et al., 2022) model. They leverages the gold queries in CANARD dataset to learn to generate SPLADE representations using previous queries and answers, which is a strategy similar to CQE. CoSPLADE performs reranking using T5Mono and finetunes T5Mono with gold queries (human reformulated query) in the target domain. This differs from our approach which does not require gold queries.

3. Background

Dense retrieval DR seeks to encode both queries and documents into a low-dimensional space with an encoder g, typically a BERT-like model Karpukhin et al. (2020); Xin et al. (2022). The retrieval status value (RSV) of a query and a document is then calculated:

$$RSV(q,d)_{DR} = g(q) \cdot g(d) \\ (or\ RSV(q,d)_{DR} = cos(g(q), g(d)))\,,$$

where $g(q)$ (resp. $g(d)$) denotes the encoding of the query (resp. document). This enables a fast retrieval through a nearest neighbour search strategy (Xiong et al., 2020).

BM25 BM25 (Robertson and Zaragoza, 2009) is a widely used standard IR algorithm based on term matching, without requiring to be trained. The RSV of a document with respect to a query is given by:

$$RSV(q,d)_{BM25} = \\ \sum_{w \in q \cap d} IDF(w) \cdot \frac{tf_w}{k_1 \cdot (1 - b + b \cdot \frac{l_d}{l_{avg}}) + tf_w},$$

where $IDF(w)$ is the inverse document frequency, l_d is the length of document d, l_{avg} the average length of the documents in the data set, and k_1 and b two hyper-parameters.

T53B By establishing a uniform framework that transforms all text-based language problems into a text-to-text format, T5 (Raffel et al., 2020) explores the landscape of transfer learning for NLP and achieves state-of-the-art results on many benchmarks. Nogueira et al. (2020) proposed to use T5 (Raffel et al., 2020) as an interaction-based model for IR by learning:

Query: [q] Document: [d] Relevant: true or false

where $[q]$ and $[d]$ are replaced with the query and document texts. During training, the T5 model learns to generate the word "true" when the document is relevant to the query, and the word "false" when it is not. The relevance score for inference is then determined by the likelihood of producing "true" (Nogueira et al., 2020). The RSV is determined by:

$$RSV(q,d)_{T5} = \frac{e^{Z_{true}}}{e^{Z_{true}} + e^{Z_{false}}},$$

where Z_{true} and Z_{false} are the logits of output tokens.

Conversational Dense Retrieval The conversational dense retrieval (CDR) is similar to dense retrieval (DR), except the query format. A CDR architecture with pairwise learning is shown in Figure 4, where the query encoder accepts the concatenation of conversational queries, to understand the user intention.

4. Pseudo-Relevance Labeling for Dense Retrieval

To enhance the domain generalization ability of DR models, we employ the pseudo-relevance labeling strategy. This involves finding pseudo-positive and pseudo-negative documents for a target dataset's queries, which are used to fine-tune the DR models. Our approach includes BM25 hard negative sampling (Figure 2) and our best approach using SimANS hard negative sampling (Figure 3). We will discuss them further below.

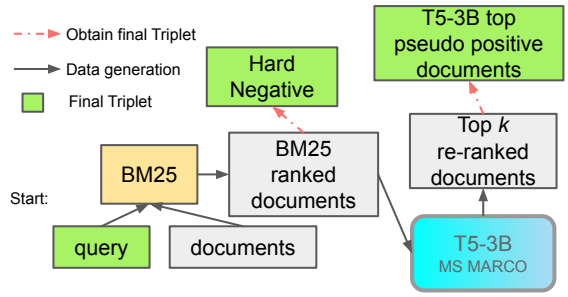

Figure 2: The overall pipeline with BM25 hard negative sampling for pseudo-relevance labeling.

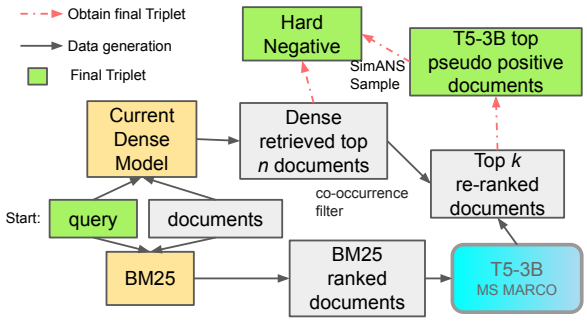

Figure 3: The overall pipeline of generating self-supervised data with meticulous pseudo-relevance labeling using SimANS hard negative sampling.

4.1. Pseudo-Positive Sampling

We propose here to consider, for each query, the top k documents obtained with the combination BM25&T53B, in which T5-3B serves as a re-ranker, as relevant (or positive). k is a hyper-parameter which can be set according to different information, as, *e.g.*, the number of available queries and documents. T5-3B, which has been shown to be a good zero-shot IR model in (Nogueira et al., 2020), is fine-tuned on MS MARCO collection.

4.2. Pseudo-Negative Sampling

Furthermore, for each relevant query-document pair, we sample m documents and consider them as non-relevant (or negative). Different negative mining strategies can be used, as described below. For each query, $k \times m$ query-document triplets (query, positive document, negative document) can be formed. The green blocks in Figure 2 and Figure 3 represent the elements constituting these triplets, the training data for domain adaptation.

Global and BM25 Hard Negative Sampling A simple negative mining strategy is global random negative sampling which consists in sampling, from all non-positive documents in the dataset, m documents which are considered as negative.

A key challenge in DR is to construct proper negative instances for learning its representations (Karpukhin et al., 2020). Previous global random negative instances might be too simple for the DR models. So, for each query, we further propose to use the BM25 top ranking documents, again excluding the positive documents, as hard negative instances for training the DR models. The architecture is shown in Figure 2.

Meticulous Hard Negative Sampling Recently, SimANS (Zhou et al., 2022) shows existing negative sampling strategies (Karpukhin et al., 2020; Xiong et al., 2020) suffer from the uninformative or false negative problem, and the authors show that the negatives ranked around the positives are generally more informative and less likely to be false negatives. They propose SimANS approach and this leads to a sampling probability distribution of the form (Zhou et al., 2022):

$$p_i \propto \exp\left(-a(s(q,d_i) - s(q,\tilde{d}^+) - b)^2\right), \forall\, d_i \in \widetilde{\mathcal{D}}^-, \quad (1)$$

where a controls the density of the distribution, b controls the peak of the distribution, $\tilde{d}^+ \in \mathcal{D}^+$ is a randomly sampled positive, and $\widetilde{\mathcal{D}}^-$ is the top-k ranked negatives.

In this paper, we use this SimANS approach with the positive documents obtained in Section 4.1. The architecture is shown in Figure 3: we select hard negatives that are around the positive instances in the top ranking of current DR models (i.e., D-BERT and GPL respectively), thus more ambiguous and informative negatives can be sampled. The green blocks in Figure 3 correspond to the queries and the associated positive and hard negative documents.

4.3. Improving GPL: Combining Pseudo-Relevance Labels and Pseudo-Queries

To enhance both the QGen and GPL approaches which rely on pseudo-queries, we suggest further training such models like GPL using the proposed pseudo-relevance triplets. We believe one can gain from this additional training on the target collection as pseudo-queries and pseudo-relevance labels rely on different sources of information and are complementary to each other. In our experiments, we demonstrate that this combination significantly improves the pseudo-query generation approach.

5. Pseudo-Relevance Labeling for Conversational Dense Retrieval

The architecture for training the CDR model with pairwise loss is shown in Figure 4, which is similar to the DR model except with a different query format.

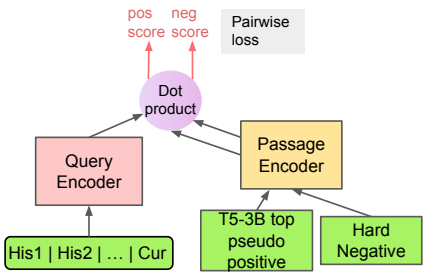

Figure 4: CDR architecture with training.

CDR models face data scarcity issues and have potential for improvement through domain adaptation. Naturally, the pseudo-relevance labeling approach can be used to help the CDR models by modifying the previous section slightly.

To train the CDR model, we concatenate the history and current queries, aiming to teach the query encoder to generate a de-contextualized query representation. Additionally, we need annotations for positive and negative documents. Our solution is pseudo-relevance labeling with conversational queries.

To achieve this, we train a T5-Large sequence-to-sequence model on CANARD (Elgohary et al., 2019), which is a dataset for learning to rewrite conversational queries. The overall architecture for the proposed approach in this section is shown in Figure 5 and the detailed procedures are outlined below.

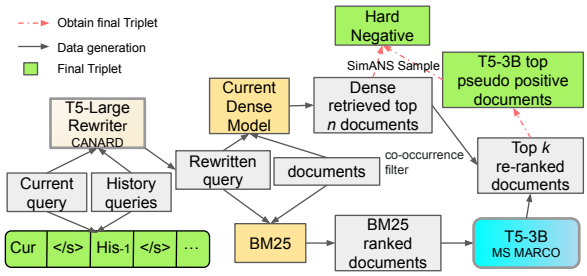

Figure 5: Overall pipeline of generating pseudo-data for conversational dense retrieval.

5.1. T5-Large Query Rewriter Module

This module is shown in the left part of Figure 5. We utilize the T5 model's special token "</s>" to concatenate the current query and history queries. The current query is placed at the beginning, followed by the recent history queries (with farther history queries located towards the end), following a similar approach as described in (Mao et al., 2022):

$$< /s > cur < /s > his_{-1} < /s > his_{-2} ...$$

The T5-Large model is trained using CANARD (Elgohary et al., 2019) with ground-truth human rewritten queries serving as labels, to generate rewritten queries that comprehensively capture the user intentions. The T5-Large query rewritten model can effectively rewrite conversational queries for target datasets into the desired de-contextualized queries. Then with them, we can sample pseudo-positive and pseudo-negative documents for them, as illustrated in Figure 5.

5.2. Pseudo-Data Format of Conversational Dense Retrieval Model

The sampling step is similar to the approach shown in Section 4: with T5-3B and SimANS for the generation of pseudo-labels given the rewritten queries.

Consequently, for training a CDR model as shown in Figure 4, each line of the training triplet file can be presented in the following format:

ConcatenateQ \t Texts of a Pseudo-Positive Doc \t Texts of a Pseudo-Negative Doc

where $ConcatenateQ$ follows the format in (Lin et al., 2021):

$$hisQ_1|hisQ_2|...|curQ$$

The triplet training file can serve as training data for finetuning a CDR model from a source domain. If the generated data is large enough, it even may be used to train a query encoder from scratch.

6. Experiments

We conducted experiments on both DR and CDR models for domain adaptation based on the approaches described in the previous sections. In the remainder of this section, we first describe the setup of DR experiments and analyze their results; we then detail the CDR experiments.

6.1. Domain Adaptation for Dense Retrieval Experimental Setup

Datasets The MS MARCO passage ranking data set (Nguyen et al., 2016) is used as the source domain data. We want to experiment in an extreme scenario where no test queries can be seen during training even without human labels. This is to say, we need to generate the pseudo training data with the training queries which is not in the test set. To do so, we experiment on 3 target domain data sets from the BEIR benchmark (Thakur et al., 2021). They are FiQA, finance question answering (Maia et al., 2018) which contains 6000 training queries, BioASQ biomedical question answering (Tsatsaronis et al., 2015) (following (Wang et al., 2022b), irrelevant documents are randomly eliminated, leaving 1M documents) which contains 3243

training queries from original collection[1], and Robust04, news documents (Voorhees, 2005) which contains 250 queries. Different topics and tasks are covered by these chosen data sets. For Robust04, we select the first 100 queries as training and development set, and the last 150 queries are used as test set.

Experimental Setting The DR model, called D-BERT, is based on the DistilBERT (Sanh et al., 2019) with 6 layers. D-BERT is initially trained on the source domain. Two groups of experiments are conducted: one with D-BERT alone and the other with the GPL model (Wang et al., 2022b) as start points. Both models are trained using the RankNet pairwise loss (Burges, 2010; Li and Gaussier, 2022) on the generated triplets.

Table 1 provides details on the number of queries, the selected value of k (top documents as relevant), and the number m of non-relevant documents per relevant document for each dataset. A development set is created to select hyperparameters, consisting of 10 relevant documents and 90 randomly selected non-relevant documents per query. The best model is saved based on the NDCG@10 score on the development set.

A maximum sequence length of 350 with mean pooling and dot-product similarity is used. A batch size of 8 and a learning rate of 2e-6 with Adam optimizer are employed for 10,000 training steps. Cosine learning rate decay (Loshchilov and Hutter, 2017b) is utilized. For SimANS, the hyperparameters a and b are set to 0.5 and 0, respectively. Experiments are done on a server with a RTX 6000 GPU, Intel Xeon E5-2623 v4 @ 2.60GHz CPU, and 148 GB memory (also for Section 6.2).

Table 1: The top k selected as positive and m as negative for each data set. The number in parentheses is used for generating training data, remaining for dev set.

data set	#queries (exclude test)	#docs	k	m
FiQA	6000 (5960)	57K	1	10
BioASQ	3243 (3193)	1M	2	15
Robust04	100 (90)	528K	15	67
CAsT-19	269 (219)	2M	5	100

Baselines We compare our proposed approaches with various existing methods in the field, including:

- Zero-shot models: BM25 based on Anserini (Yang et al., 2018) and D-BERT trained solely on the source collection.
- Pre-training based models: we use three state-of-the-art models, namely SimCSE (Gao et al., 2021), ICT (Lee et al., 2019), and TSDAE (Wang et al., 2021).
- Domain adaptation approaches: MoDIR (Xin et al., 2022), UDALM (Karouzos et al., 2021), QGen (Ma et al., 2021), and GPL (Wang et al., 2022b). The combination of GPL with TSDAE is currently considered as the best approach.

Furthermore, we include interaction-based models BM25+CE and BM25+T53B as strong baselines, which re-rank the top 100 BM25 ranked list using *ms-marco-MiniLM-L-6-v2* and T5-3B cross encoders[2], Cross encoders are viewed as upper bound baselines compared with dense retrieval models because they are known as interaction-based models for re-ranking and perform better. In this paper, this T5-3B model is also used for pseudo-positive labeling. These models are known for their good performance in out-of-domain settings (Thakur et al., 2021).

Results and Analysis Table 2 displays the results obtained with different models and approaches. The results reported for BM25+CE, UDALM, MoDIR, SimCSE, ICT, TDSAE, QGen and TSDAE+GPL are from (Wang et al., 2022b). Since we test the Robust04 on the last 150 queries, for BM25+CE, GPL and TSDAE+GPL, we load the trained checkpoints of D-BERT from (Wang et al., 2022b)[3], and evaluate them on the last 150 queries. The notation "DoDress-T53B (D-BERT)" corresponds to the D-BERT dense retrieval model pre-trained on MS MARCO and fine-tuned on the target data using the pseudo-relevance labels generated with BM25+T53B. The notation (GPL) means the same for GPL, which is first trained on the target pseudo queries it generates and associated documents prior to be trained on the target triplets.

We address three main research questions, denoted as **RQ**.

RQ1 Do BM25+T53B top positives help domain generalization for dense retrieval models?

The results in Table 2 shows that DoDress-T53B (D-BERT) and DoDress-T53B (GPL) outperform D-BERT and GPL, respectively, on the FiQA and Robust04 datasets using different negative sampling strategies. Notably, DoDress-T53B (D-BERT) with the SimANS negative mining strategy achieves an 11.5% improvement over D-BERT, while DoDress-T53B (GPL) shows an 8.6% improvement over GPL.

[1] http://participants-area.bioasq.org/Tasks/8b/trainingDataset/

[2] https://huggingface.co/castorini/monot5-3b-msmarco which is trained on MS MARCO for 100K steps.

[3] https://huggingface.co/GPL

On the BioASQ dataset, the approach with global random negative sampling fails, but the proposed approach with the other two negative sampling strategies improves D-BERT and GPL, respectively. The reasons of this success can be explained by the fact that the proposed pseudo-relevance labeling approach enables the DR models to see and be trained with real queries and documents of the target dataset. This labeling approach is further improved when combined with query generation approach of GPL. These results demonstrate that the proposed pseudo-relevance labeling approach helps dense retrieval models generalize to new domains. As the reader may have noticed, the choice of the negative sampling strategy is crucial for its effectiveness.

RQ2 What is the best negative sampling strategy?

From Table 2, we observe an overall ascending trend in performance with the three different negative sampling strategies. The global random negative method shows improvements on FiQA and Robust04 but fails on the BioASQ dataset. This may be due to uninformative negatives that are too easy for the dense retrieval models on the target domain. In contrast, the BM25 hard negative and SimANS negative sampling strategies outperform the global random negative strategy, improving D-BERT and GPL on all three datasets. This highlights the importance of sampling hard negatives in the proposed pseudo-relevance labeling data generation approach. Among the three strategies, SimANS hard negative sampling consistently performs the best on all datasets, surpassing the global random negative and BM25 hard negative strategies.

RQ3 What is the best overall approach?

As one can note from Table 2, the DoDress-T53B models outperform all other models on all collections but BM25 on BioASQ and of course the models consisting in re-ranking BM25 results with cross-encoders, which constitute an upper bound and are too costly to be used in practice. In addition, if BM25 is a strong competitor for domain adaptation, as reported in (Thakur et al., 2021), its performance vary significantly from one collection to the other (very good on BioASQ, very poor on FiQA). Lastly, it is interesting to note that DoDress-T53B (GPL) outperforms the previous state-of-the-art model (TSDAE+GPL) by a large margin on BioASQ and Robust04, showing the effectiveness of the proposed approach.

6.2. Conversational Dense Retrieval Experimental Setup

Dataset Used We utilize the TREC CAsT 2019 (CAsT-19) dataset (Dalton et al., 2020). CAsT-19 comprises 30 training topics and 20 test topics, with each topic representing a conversational search session consisting of queries from multiple turns. The dataset contains a total of 269 training queries. Notably, in this paper, we investigate an extreme scenario in which we do not have access to human rewritten queries and relevance labels for CAsT-19, resulting in an almost zero-shot scenario.

For efficient experiments, we furthermore follow the experimental protocol defined in (Wang et al., 2022b): we randomly remove irrelevant passages from the whole 38M TREC CAsT-19 corpus to obtain a smaller corpus consisting of 2M passages.

Table 2: Domain adaptation result of FiQA, BioASQ and Robust04 (during training only use train queries).

Method	FiQA	BioASQ	Robust04	Avg.
Zero-Shot Models				
D-BERT	26.7	53.6	39.1	39.8
BM25 (Anserini)	23.6	73.0	44.4	47.0
Re-Ranking with Cross-Encoders (Upper Bound)				
BM25 + CE	33.1	72.8	45.8	50.6
BM25 + T53B	39.2	76.1	51.8	55.7
Previous Domain Adaptation Methods				
UDALM	23.3	33.1	-	-
MoDIR (ANCE)	29.6	47.9	-	-
Pre-Training based: Target → D-BERT				
SimCSE	26.7	53.2	-	-
ICT	27.0	55.3	-	-
TSDAE	29.3	55.5	-	-
Generation-based (Previous SOTA)				
QGen	28.7	56.5	-	-
GPL	32.8	62.8	41.9	45.8
TSDAE + GPL	34.4	61.6	40.7	45.6
Proposed: T53B, Global Random Neg				
DoDress-T53B (D-BERT)	27.3	52.9	40.5	40.2
DoDress-T53B (GPL)	33.0	62.0	43.2	46.1
Proposed: T53B, BM25 Hard Neg				
DoDress-T53B (D-BERT)	30.4	58.6	41.6	43.5
DoDress-T53B (GPL)	34.2	64.7	43.3	47.4
Proposed: T53B, SimANS Hard Neg				
DoDress-T53B (D-BERT)	**31.0**	**60.6**	**43.6**	**45.1**
DoDress-T53B (GPL)	**34.9**	**65.3**	**45.5**	**48.6**

T5 Rewriter The conversational rewriter used is the T5-Large version[4]. We train the model on CANARD dataset (Elgohary et al., 2019) which contains 31526 training instances repeatedly for 80K instances, and evaluate the T5-Large model on development set for every 20000 instances and save the best model. The learning rate is 5e-5 and batch size is 4 using AdamW optimizer (Loshchilov and Hutter, 2017a). The BLEU (Papineni et al., 2002) results of the final T5-Large model on CANARD are presented in Table 3. We can see the T5-Large model can obtain near human accuracy for rewriting conversational queries on CANARD dataset. Besides, we have computed its BLEU score on TREC CAsT-19, obtaining 64.35 , which indicates a favorable outcome.

Baselines

[4] https://huggingface.co/t5-large

Table 3: BLEU scores of different approaches for rewriting conversational queries on CANARD dataset. The first four methods are baseline approaches used in (Elgohary et al., 2019).

Method	Dev	Test
Copy	33.84	36.25
Pronoun Sub	47.72	47.44
Seq2Seq	51.37	49.67
Human Rewrites	59.92	
T5-Large	59.5	57.9

- Zero-Shot baselines: The BERT-dot-v5[5] model trained on MS MARCO is used as zero-shot baselines. We experiment with four methods: (1) using only the current query, (2) concatenating the history and current queries (it should be noted that the query encoder is not specifically trained to handle this format), (3) using T5-Large rewritten queries for retrieval (may not be efficient in real-world scenarios), and (4) the upper bound method with human rewritten queries from the dataset. In addition, we also include BM25 using Anserini (Yang et al., 2018) with official rewritten queries and T5-Large rewritten queries as baselines.
- Cross-Encoder: We adopt T5-3B model which is trained on MS MARCO (same version as before) to rerank the BM25 list from T5 Rewritten queries.
- Related work: ConvDR (Yu et al., 2021) and CQE (Lin et al., 2021) are compared, which are CDR models aiming to address the data scarcity issue. They are trained from the BERT-dot-v5 checkpoint. We train ConvDR for 40k steps with the batch size of 8. We observed that the performance remained similar across different training intervals, namely 10k, 40k, and 80k steps. Following (Lin et al., 2021), we train CQE for 120k steps with a batch size of 8, which is comparable to their original paper's training process of 20k steps with a batch size of 96. The learning rate used in our training process is set to 2e-6.

CDR Training We further conduct experiments based on the learned checkpoints of baseline models ConvDR and CQE to deal with the few-shot learning scenario. We fine-tune them with our generated pseudo-labels. Following (Yu et al., 2021; Lin et al., 2021), the passage encoder is fixed and the query encoder is trained. All history turns are used since they are short. The parameters for generating pseudo-relevance labels are the same as before (Table 1). Our fine-tuning strategy is similar to the previous DR experiments, using the RankNet pairwise loss with a batch size of 8 and learning rate of 2e-6. We train on our (few shot) pseudo-target data for 2000 steps and evaluate on the development set every 500 steps, saving the best models. Finally, we report the results of the trained models using NDCG@3, which is consistent with previous research (Yu et al., 2021; Lin et al., 2021).

Experiment Result Experiment results are presented in Table 4. Besides, we also show the deployed demo of our best trained CDR model (58.0) dealing with real world conversational queries in Table 1 (deployed on a RTX 6000 GPU). We address here two main research questions.

Table 4: Domain adaptation result of CAsT-19.

model	nDCG@3 (%)
Zero-Shot Models	
BERT-dot-v5(current)	33.4
BERT-dot-v5(concatenation)	27.2
BERT-dot-v5(T5Rewrite)	53.2
BERT-dot-v5(Human) (Upper Bound)	58.9
BM25(Human)	37.0
BM25(T5Rewrite)	31.2
Re-Ranking with Cross-Encoders	
T5-3B rerank T5Rewrite	56.7
Related Work	
ConvDR (BERT-dot-v5)	55.4
CQE (BERT-dot-v5)	53.7
Proposed Approach	
T53B, SimANS Neg, based on ConvDR	
DoDress-T53B (BERT-dot-v5)	**58.0**
T53B, SimANS Neg, based on CQE	
DoDress-T53B (BERT-dot-v5)	**57.6**

RQ1 In the context of conversational search, how do IR models exclusively trained on MS MARCO perform as zero-shot models?

The best baseline among zero-shot models is BERT-dot-v5(Human), which constitutes an upper bound as it uses human rewritten queries on the target dataset. This model largely outperforms BM25(Human). In comparison, using the current query or the concatenation of queries with the BERT-dot-v5 model yields lower results than the human rewritten queries, respectively 33.4 and 27.2 compared to 58.9. This means that, for this dataset, although DR models can successfully retrieve documents when given real human rewritten queries and outperform BM25 approach, they do not perform well when they are not specifically fine-tuned to understand the conversational queries or when they are not given good rewritten queries. When using T5 rewritten queries, the performance becomes closer to the one obtained with human rewritten queries. In addition, using T5-3B with the T5 rewritten queries as a reranking model yields very good results, close to the upper bound.

[5] https://huggingface.co/sentence-transformers/msmarco-bert-base-dot-v5

RQ2 Does the proposed in-domain pseudo-relevance data generation approach effectively enhance the performance of CDR models?

Firstly, let's discuss related works on CDR. ConvDR and CQE achieve scores of 55.4 and 53.7, respectively, outperforming standard zero-shot DR model baselines except the upper bound using human rewritten queries. These models benefit from the CANARD dataset, which enables effective representations for conversational queries. However, as the queries in CANARD differ from the ones in the target dataset, training with in-domain queries becomes crucial. The proposed models based on ConvDR and CQE achieve scores of 58.0 and 57.6, respectively, representing improvements of 4.7% and 7.3% over ConvDR and CQE. This is due to the fact that the proposed pseudo-relevance labeling approach enables the CDR models to see real queries and respective documents on the target domain, resulting in better adaptation. Overall, our proposed approach achieves comparable performance to the one obtained by the BERT-dot-v5 model with real human rewritten queries, without requiring human annotations on the target dataset.

Explainability of the Models Let's explore the explainability of the models and their efficacy on unlabeled datasets. A significant factor contributing to their success is the adoption of the pseudo-relevance labeling approach. The T53B, a large cross-encoder model, demonstrates good performance across both familiar and unfamiliar domains. Acting as expert annotators for previously unseen domain data, it also serves as a teacher, distilling its knowledge to the student DR models through knowledge distillation like ways. Through mining hard negatives, the models undergo more effective training, thereby enhancing their capacity to generate representations. Additionally, through the utilization of the T5 large rewriter module, the conversational search dataset is transformed to a standard IR dataset and same strategy can be used and can also be explained.

7. Conclusion

This paper first studied whether one can benefit from existing re-ranking based IR models, pre-trained on MS MARCO, to generate pseudo-relevance labels for an unannotated, target collection. These labels, along with sampled positives and negatives, are used to fine-tune dense retrieval models on the target collection. The experiments revealed that carefully generating pseudo-labels improves the generalization results of DR models and that additional improvements can be obtained with the query generation approach of GPL. We also investigated several negative sampling strategies, based on BM25 and SimANS, and confirmed the importance of identifying useful hard negative documents. The proposed pseudo-relevance labeling approach has also been applied to CDR models for conversational search. In particular, we incorporated a query rewritten module that utilizes T5-Large to deal with conversational queries and relied on pseudo-relevance labels generated using T5-3B and SimANS on the rewritten queries. The experiments revealed that this approach yields state-of-the-art CDR models for domain adaptation. Overall, by making use of real queries and documents of the target domain, the simple labeling approach we have followed, combined with query generation or query rewriting, has proved to be very effective for adapting or further improving a DR or CDR model to new domains.

8. Acknowledgements

This work has been partly supported by the French project ANR-19-P3IA-0003 MIAI@Grenoble Alpes, the Chinese Scholarship Council (CSC) grant No.201906960018 and Soochow University grant No.NH11801824. The first author also thanks the IDEX for funding ML mobility grant.

9. Bibliographical References

Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. 2021. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100.

Christopher J. C. Burges. 2010. From RankNet to LambdaRank to LambdaMART: An overview. Technical report, Microsoft Research.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238.

J. Shane Culpepper, Fernando Diaz, and Mark D Smucker. 2018. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, volume 52, pages 34–90. ACM New York, NY, USA.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu,

Keith Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Trec cast 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624*.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 65–74.

Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. 2020. Learning to learn with variational information bottleneck for domain generalization. In *European Conference on Computer Vision*, pages 200–216. Springer.

Antonio D'Innocente and Barbara Caputo. 2018. Domain generalization with domain-specific aggregation modules. In *German Conference on Pattern Recognition*, pages 187–198. Springer.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2353–2359.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. 2020. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067.

Nam Hai Le, Thomas Gerald, Thibault Formal, Jian-Yun Nie, Benjamin Piwowarski, and Laure Soulier. 2023. Cosplade: Contextualizing splade for conversational information retrieval. In *European Conference on Information Retrieval*, pages 537–552. Springer.

Helia Hashemi, Yong Zhuang, Sachith Sri Ram Kothur, Srivas Prasad, Edgar Meij, and W Bruce Croft. 2023. Dense retrieval adaptation using target domain description. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 95–104.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Seogkyu Jeon, Kibeom Hong, Pilhyeon Lee, Jewook Lee, and Hyeran Byun. 2021. Feature stylization and domain-aware contrastive learning for domain generalization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 22–31.

Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potamianos. 2021. Udalm: Unsupervised domain adaptation through language modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2579–2590.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage

retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550.

Minghan Li and Eric Gaussier. 2022. Bert-based dense intra-ranking and contextualized late interaction via multi-task learning for long document retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2347–2352.

Davis Liang, Peng Xu, Siamak Shakeri, Cicero Nogueira dos Santos, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Embedding-based zero-shot retrieval through query generation. *arXiv preprint arXiv:2009.10270*.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015.

Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. 2021. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34:6155–6170.

Ilya Loshchilov and Frank Hutter. 2017a. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ilya Loshchilov and Frank Hutter. 2017b. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 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*, pages 1075–1088.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. 2018. Best sources forward: domain generalization through source-specific nets. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 1353–1357. IEEE.

Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng, and Zhao Cao. 2022. Convtrans: Transforming web search sessions for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2935–2946.

Ida Mele, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, and Ophir Frieder. 2020. Topic propagation in conversational search. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 2057–2060.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2018. Mixed precision training. In *International Conference on Learning Representations*.

Iurii Mokrii, Leonid Boytsov, and Pavel Braslavski. 2021. A systematic evaluation of transfer learning and pseudo-labeling with bert-based ranking models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2081–2085.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. 2017. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. 2021. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. 2019. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7249–7255. IEEE.

Fengchun Qiao, Long Zhao, and Xi Peng. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. 2019. Multi-component image translation for deep domain generalization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 579–588. IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. 2018. Conversational query understanding using sequence to sequence modeling. In *Proceedings of the 2018 World Wide Web Conference*, pages 1715–1724.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. 2018. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*.

Si Sun, Yingzhuo Qian, Zhenghao Liu, Chenyan Xiong, Kaitao Zhang, Jie Bao, Zhiyuan Liu, and Paul Bennett. 2021. Few-shot text ranking with meta adapted synthetic weak supervision. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5030–5043.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 355–363.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. 2018. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31.

Ellen Voorhees. 2005. Overview of the trec 2004 robust retrieval track. Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022a. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022b. 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*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

M. Wang and W. Deng. 2018. Deep visual domain adaptation: a survey. *Neurocomputing*.

Ji Xin, Chenyan Xiong, Ashwin Srinivasan, Ankita Sharma, Damien Jose, and Paul Bennett. 2022. Zero-shot dense retrieval with momentum adversarial domain invariant representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4008–4020.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using lucene. *Journal of Data and Information Quality (JDIQ)*, 10(4):1–20.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval*, pages 829–838.

Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational information seeking.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-rong Wen, and Nan Duan. 2022. SimANS: Simple ambiguous negatives sampling for dense text retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 548–559, Abu Dhabi, UAE. Association for Computational Linguistics.