Simple Domain Adaptation for Sparse Retrievers

Mathias Vast, Yuxuan Zong, Benjamin Piwowarski, Laure Soulier

Simple Domain Adaptation for Sparse Retrievers

Mathias Vast[1,2][0009−0007−4612−717X], Yuxuan Zong[2][0009−0002−0376−1369], Benjamin Piwowarski[2][0000−0001−6792−3262], and Laure Soulier[2][0000−0001−9827−7400]

[1] Sinequa
[2] Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
{firstname.lastname}@isir.upmc.fr

Abstract. In Information Retrieval, and more generally in Natural Language Processing, adapting models to specific domains is conducted through fine-tuning. Despite the successes achieved by this method and its versatility, the need for human-curated and labeled data makes it impractical to transfer to new tasks, domains, and/or languages when training data doesn't exist. Using the model without training (zero-shot) is another option that however suffers an effectiveness cost, especially in the case of first-stage retrievers. Numerous research directions have emerged to tackle these issues, most of them in the context of adapting to a task or a language. However, the literature is scarcer for domain (or topic) adaptation. In this paper, we address this issue of cross-topic discrepancy for a sparse first-stage retriever by transposing a method initially designed for language adaptation. By leveraging pre-training on the target data to learn domain-specific knowledge, this technique alleviates the need for annotated data and expands the scope of domain adaptation. Despite their relatively good generalization ability, we show that even sparse retrievers can benefit from our simple domain adaptation method.

Keywords: Pretrained Language Models · Cross-Topic Adaptation · Zero-Shot.

1 Introduction

Nowadays, most of the models that achieve state-of-the-art results in Natural Language Processing (NLP) rely on the "*pre-train then fine-tune*" pipeline [4,22]. In this setup, a model is first trained with self-supervised objectives, such as Masked Language Modeling, on a large unlabeled corpus, before being fine-tuned on a labeled dataset for a specific task. In Information Retrieval (IR), this method has delivered huge improvements over the previous "pre-BERT" models and has given birth to a large variety of frameworks including, but not limited to, dense retrievers [13], learned sparse retrievers [5] and cross-encoders [22]. For a given task, when enough good-quality labeled data is available, fine-tuning an already pre-trained model is the best option. Thanks to datasets such as MS MARCO [21], this condition might be met for the majority of the downstream tasks in the generic domain for the English language, but in other languages and/or specific domains, these resources may not exist.

This poses a problem since a fine-tuned model evaluated outside its training domain might perform poorly [29]. When the volume of training data is insufficient for fine-tuning, it can even degrade the model's performance [32]. Besides, the increasing size of these models coupled with the difficulty of collecting a satisfying amount of high-quality labeled data has led to a growing interest in zero-shot or few-shot methods.

It is possible to tackle this issue by leveraging unlabeled data and pursuing pre-training over task or domain-related data [7,14] before fine-tuning. Providing a better initialization point alleviates the dependency on the amount of in-domain labeled data available for fine-tuning. Differently, Artetxe et al. [2] leverage pre-training as a means to adapt a model *subpart* for a specific language, different from the source language used for pre-training and fine-tuning the base model. In this paper, we consider a setup of cross-domain adaptation where the model to be transferred can be trained on a source domain D_{source}, where an annotated corpus containing enough documents for both pre-training and fine-tuning is available, whereas the target domain D_{target} contains enough data for pre-training but no annotation. We extend the work of [2] in two ways to deal with ad-hoc IR. First, inspired by [3], we study which subparts of the model should be pre-trained or fine-tuned – we do not only consider the embeddings to be domain-dependent. Second, we propose a new pre-training procedure on both the source and target domains, and not only on the target one. This allows a model to learn task-specific parameters (on the source domain) when (and only when) its domain-specific ones are well set. As domain shift impacts more strongly first-stage retrievers, we study a first-stage sparse retriever, namely SPLADE [5]. Our approach allows one to share the fine-tuning resources across multiple domains with less drop in effectiveness due to domain discrepancies.

2   Related works

Transfer learning and adaptation to a specific context are long-standing research topics. With the recent arrival of large language models, research shifted to transferring these models at the lowest computational and annotation cost possible [1]. To get rid of human annotation, works inspired by Doc2Query [23] propose to generate queries for which a document in the target domain would be relevant [16,30]. Inversely, generative models can also be used to produce pseudo-relevant documents [6] before fine-tuning a model on it. [9] even proposed a method that generates a complete collection of both documents and queries from a simple description of the target domain. For each of these methods, one can apply the traditional "*pretrain then fine-tune*" framework. Besides the fact that this type of approach can be computationally costly, studying, as in our work, pre-training techniques, is complementary. Indeed, as highlighted in [15], better pre-training, either by using different self-supervised objectives [12] or pursuing it longer [7,14], often provides better results compared to fine-tuning for first-stage methods.

Other works have tried to ease the transfer of (large) pre-trained models. We can separate them into two categories, even though they share the same
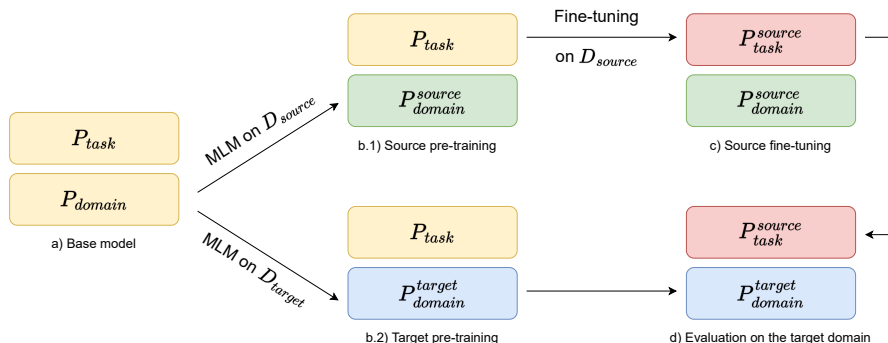
Fig. 1. Illustration of the cross-domain adaptation process

underlying principle of distinguishing different sets of parameters during training. The first family, referred to as Parameter-Efficient Fine-Tuning (or PEFT) [10,11,17,19], achieves this objective by modifying the original model with an *adapter* whose parameters are fine-tuned. Its success relies on the quality of the underlying pre-trained model, whose generality is worth being conserved while tuning only a small portion of parameters on the target task. In IR, these methods have started to be explored recently and, so far, have proven to be successful [18,24,28]. Orthogonal to these works, some papers do not use freezing to preserve generality or to spare computations, but to adapt specific parts of the model for a given language or task. In particular, [31] applies PEFT methods to Dense Retrievers to disentangle domain and relevance learning. For each domain, a Transformer backbone is trained to learn its linguistic features (MLM). Separately, a distinct module is trained once and then shared, to predict relevance based on these extracted features. Artetxe et al. [2] adapt a similar reasoning but apply the disentangling within the Transformer backbone's layers, thus introducing a zero inference overhead. In this work, we extend their approach to domain adaptation in the context of IR. Note that using PEFT following [31] is an alternative that we will explore in the future.

3   Cross-Domain Adaptation of a Neural Network model

We describe here how we adapted the pipeline introduced by Artetxe et al. [2] to cross-domain adaptation. This approach can be extended to any model in any setting by establishing a distinction between the model's parameters dedicated to learning the domain/language model, denoted P_{domain}, and the ones dedicated to learning the task denoted P_{task}, which are supposed to be distinct, i.e. $P_{domain} \cap P_{task} = \emptyset$. Given the IR setting, we define P_{domain} to be the embeddings together with the first k layers of the transformer architecture – whereas in the language adaptation setting, [2] only considered P_{domain} to be the embeddings.

Figure 1 describes the training process made of two distinct pre-trainings; source (b.1) and target (b.2); and one fine-tuning on the source (c), before evaluating (d) on the target domain:

Pre-training for D_{source} (resp. D_{target}): As in [2], we continue to pre-train BERT over D_{source} (resp. D_{target}) while keeping the parameters' subset P_{task} frozen. This step adapts the parameters P_{domain} to the specificity of the domain D_{source} (resp. D_{target}). We denote P_{domain}^{source} (resp. P_{domain}^{target}) the parameters after the MLM pre-training. Note that contrary to [2], we keep the base model's vocabulary unchanged throughout the pipeline.

Fine-tuning on D_{source}: We leverage the previously pre-trained subset P_{domain}^{source} but keep it frozen while only fine-tuning the P_{task} parameters. The objective of this step is to specialize the subset of parameters P_{task} to the IR task. We denote as P_{task}^{source} the fine-tuned parameters.

Contrary to [2], we add an additional pre-training step over the source corpus (b.1 in Figure 1). We suppose that thanks to this step, the model can better learn the downstream task as it is not fine-tuned on the same dataset (MS MARCO) that it was originally pre-trained on (Wikipedia).

Finally, for inference on the D_{target}'s dataset, we combine the two subsets P_{domain}^{target} and P_{task}^{source}. The underlying intuition is that the subset P_{task}^{source} is task-specific while P_{domain}^{target} is domain-specific. The model should be able to generalize better when used forIR in the target domain.

4   Experiments

We use $BERT_{base}$ pre-trained on the English version of Wikipedia[1] as our base model and share its base vocabulary through every domain. We focus on a sparse first-stage ranker, SPLADE [5], as first-stage rankers particularly suffer from domain shifts and SPLADE is known to be at the state-of-the-art among them.

4.1   Datasets, baselines, and ablations

Two datasets in IR might differ in 1) domain/vocabulary, 2) document types, and/or 3) topic types [25]. For a neural model, any of these distinctions can impact the generalization power. The source dataset we consider across all our experiments is MS MARCO [21] as it is the defacto dataset for fine-tuning pre-trained models in IR and its domain can be regarded as "generic Web search". We assemble a collection of target datasets from a subset of the BEIR benchmark (*) [29], the LoTTE's collection (**) [27], as well as other sources. The decision about using or not some datasets from the BEIR benchmark is based on multiple criteria, including the existence of distinct train and test sets and the task affiliated with the dataset as we are only interested in pure IR in this case. We processed them with the *ir-datasets* library [20]. Table 1 gives details about the characteristics of each dataset and shows how we covered various domains, ranging from bio-medical to literature.

To verify the benefits of our adaptation method, we experimented with multiple variants, differing on the P_{domain} / P_{task} split by the number of layers $k \in \{0, 1, 2, 3, 4, 6, 8, 10\}$ after the embedding belonging to P_{domain}. We also included two baselines and two ablations that we describe in what follows:

---

[1] The model is made available by Google on the HuggingFace' Hub: bert-base-uncased

Table 1. Datasets considered in the study. Statistics on the average topic and document length for each dataset are computed using BERT tokens.

Dataset	Domain	Dataset Statistics		
		Avg top./doc. len	# Top	# Doc
MSM-Passage	Generic	7.5 / 74.9	808K	8.8M
TREC-COVID (*)	Bio-Medical	16.0 / 243.5	50	171.3K
NFCorpus (*)	Bio-Medical	5.0 / 338.1	325	5.4K
BioASQ (*)	Bio-Medical	13.1 / 320.3	500	14.9M
FiQA-2018 (*)	Finance	13.6 / 175.1	648	57.6K
TREC-NEWS (*)	News	18.2 / 652.5	50	595K
Robust04 (*)	News	18.7 / 638.5	250	528.1K
ANTIQUE [8]	Web	11.9 / 52.3	200	403.7K
LoTTe-Wri. (**)	Literature	8.7 / 165.2	1071	200K
LoTTe-Tec. (**)	Technology	9.7 / 216.6	596	638.5K
LoTTe-Rec. (**)	Entertainment	9.3 / 181.6	924	167K

– **BM25** [26], is a very strong baseline in domain adaptation [29] and is also a sparse first-stage retriever;
– **Zero-Shot Learning** evaluation of the SPLADE model fine-tuned on MS MARCO to quantify the benefits we can obtain with our adaptation method;
– Ablation of the source pre-training, referred to as **w/o source**, as this corresponds to the original approach described in Artetxe et al. (with $k = 0$);
– Ablation of the source *and* target pre-trainings, referred to as **w/o pre-training** (again with $k = 0$). Note that this is a zero-shot model. As we are starting from a model pre-trained on Wikipedia, this ablation evaluates whether our proposition has an impact.

All the project code, based on the library *experimaestro-ir* [33], including the experimental details, is freely accessible[3].

4.2 Results

Table 2 contains the results of these methods from which we can draw the following conclusions: (i) Compared to the Zero-Shot Learning baseline, our adaptation approach allows us to gain on average between 0.7 and 1.4 points in nDCG@10, and more than 1 point for some datasets; (ii) Comparing "w/o source" to our approach, it seems that the additional pre-training step over the source domain helps the model to generalize better, with an average improvement between 0.1 and 0.8 points (for nDCG@10). This is sensible since the model learns to adapt the task-specific parameters from domain-specific parameters; (iii) Similarly, no pre-training at all can hurt the performance up to 1.2 points in nDCG@10 ("w/o pre-training"), showing the benefit of our pre-training procedure. However, in a

---

[3] https://git.isir.upmc.fr/mat_vast/cross_domain_adaptation

Table 2. Performance in nDCG@10 of our approach versus BM25 and Zero-shot (SPLADE). Bold values are strictly superior to the Zero-shot baseline. † and ‡: Improves upon BM25 baseline and Zero-Shot baseline respectively with statistical significance ($p \leq 0.05$) under the two-tailed Student's t-test.

Method ($\rightarrow$)	Baselines		Proposed pipeline ($k = \ldots$)				Ablations	
Dataset ($\downarrow$)	BM25	0-shot	0 layer	1 layer	2 layers	4 layers	w/o pre-training	w/o source
TREC-COVID	58.1	71.3	68.8†	70.5†	**72.1†**	70.9†	67.9†	67.7†
NFCorpus	24.5	24.9	**25.6**	**25.4**	**25.1**	24.1	**25.1**	**25.6**
BioASQ	**52.3**	43.1	45.4‡	44.2	45.2‡	46.2‡	44.8‡	45.3‡
FiQA-2018	23.6	27.3	27.2†	**27.5†**	**29.5†‡**	**29.0†‡**	25.2†	**28.3†**
TREC-NEWS	31.4	32.1	**33.3**	**32.8**	**33.0**	**36.5‡**	31.9	**33.1**
Robust04	40.8	39.6	**42.9‡**	**42.9‡**	**43.5‡**	**42.6‡**	**41.2**	**42.1‡**
ANTIQUE	**45.4**	43.3	42.9	**44.0**	**43.7**	**44.1**	**45.0‡**	**45.0‡**
LoTTE-Rec.	39.3	46.2	**47.6†‡**	**47.7†‡**	**47.8†‡**	**46.9†**	**46.8†**	**46.7†**
LoTTE-Wri.	41.3	52.7	**53.0†**	**53.8†‡**	**53.4†**	**54.1†‡**	52.0†	**52.9†**
LoTTE-Tec.	25.3	36.8	**38.2†**	**37.6†**	**37.4†**	**36.8†**	36.8†	36.8†
Avg.	38.2	41.7	**42.5**	**42.6**	**43.1**	**43.1**	**41.9**	**42.4**

Table 3. Evolution of the performance in nDCG@10 and of the sparsity of SPLADE when the number of layers dedicated to the learning of the target domain increases. Best result is highlighted in bold.

Method ($\rightarrow$)	0-shot	k=0	k=1	k=2	k=3	k=4	k=6	k=8	k=10
Average Performance (nDCG@10)	41.7	42.5	42.6	43.1	42.6	**43.1**	43.1	43.0	42.6
Sparsity ($\times 10^3$)	9.3	7.3	9.2	11.5	8.4	10.3	10	9.8	9.3

zero-shot setting, we observe that fine-tuning only the Transformer's layers performs ("w/o pre-training") on par with fine-tuning the whole model ("0-shot"). Both setups are evaluated without using the target dataset, and can thus be considered as Zero-Shot Learning. It indicates that we could save some computations during fine-tuning without hurting performance. (iv) Pre-training over additional layers can provide additional gains of up to 0.5 points in nDCG@10 (columns "1 layer", "2 layers" and "4 layers" compared to the column "0 layer"). Results however seemed to plateau beyond $k = 2$ as illustrated by the minor differences between the columns $k = 4$, with the best result in average, and $k = 2$. Table 3 summarizes the evolution of the average performance along with the number of additional layers reserved for learning the target domain. These results seem to comfort recent findings on the role of the first layers' attention head in a Transformer model [3]. It also highlights that even though SPLADE-base models highly rely on the Masked Language Modeling task, fine-tuning nevertheless remains an important part of the process, and increasing the share of the parameters dedicated to the former can infringe the model's performance on the IR task.

Table 3 also gives additional insights into the sparsity achieved by the document encoder of our SPLADE variants. Query encoder sparsity isn't specified

as it says consistent along the variants and the baseline. We note that the best models usually have a higher sparsity level, which tends to indicate that the associated model is able to build a more appropriate representation of each target domain. However, more work is needed to understand why some other variants happen to have lower sparsity values than the baseline while performing better.

Discussion and limitations. This work is a preliminary study of how pre-training can be used to ease domain adaptation in IR. We led the same series of experiments with a second-stage ranker, namely monoBERT [22]. Interestingly, the results showed that our approach didn't provide any improvements compared to the Zero-Shot Learning approach in this case. Previous work [15] already mentioned the fact that second-stage rankers benefit more from larger collections compared to pre-training on specific datasets. Our conclusion is that our approach is too naive to provide any improvement at all, given the generalization power of monoBERT, and suggests that a better understanding of which parameters are domain or task-specific, as well as their interplay, is necessary. Further work is also needed to explore the impact of pre-training time given corpus length on the final results. We leave both of these directions for future works as well as the understanding of why second-stage rankers cannot draw any benefits from specific pre-training. Finally, we did not compare with generative methods (i.e., generating a query matching a document in the target domain) such as [16] – these methods are costly, and we also need to investigate whether the improvements are complementary with our (more cost-efficient) approach.

5   Conclusion

We presented a simple approach to domain adaptation for IR, based on a language adaptation approach [2]. Results from our experiments highlight the potential of our method with learned sparse retrievers and could be particularly helpful when transferring fine-tuned models in contexts where high-quality annotated data isn't available or impossible to collect because it is either too expensive or complex. Another benefit of our approach is its re-usability. Once the expensive fine-tuning has been performed, the subset P_{task} can be re-used with different subsets P_{domain} pre-trained over different domains, which eliminates the need to perform a fine-tuning over all the model for every new domain. In the future, we would like to include dense retrievers as well as more costly methods in our study. In addition, we also want to extend it to more sophisticated pre-training approaches.

6   Ackowledgements

References

1. Althammer, S., Zuccon, G., Hofstätter, S., Verberne, S., Hanbury, A.: Annotating data for fine-tuning a neural ranker? current active learning strategies are not better than random selection. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. p. 139–149. SIGIR-AP '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3624918.3625333, https://doi.org/10.1145/3624918.3625333

2. Artetxe, M., Ruder, S., Yogatama, D.: On the cross-lingual transferability of monolingual representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. p. 4623–4637 (2020). https://doi.org/10.18653/v1/2020.acl-main.421, http://arxiv.org/abs/1910.11856, arXiv:1910.11856 [cs]

3. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does bert look at? an analysis of bert's attention. Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (2019). https://doi.org/10.18653/v1/w19-4828, http://dx.doi.org/10.18653/v1/W19-4828

4. Devlin, J., Chang, M.W., Lee, K., Toutanov, K.: Bert: Pre-training of deep bidirectional transformers for language understandin. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistic (2018). https://doi.org/10.48550/arXiv.1810.04805

5. Formal, T., Lassance, C., Piwowarski, B., Clinchant, S.: Splade v2: Sparse lexical and expansion model for information retrieval (2021). https://doi.org/10.48550/ARXIV.2109.10086, https://arxiv.org/abs/2109.10086

6. Gao, L., Ma, X., Lin, J., Callan, J.: Precise zero-shot dense retrieval without relevance labels (2022)

7. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8342–8360. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.740, https://aclanthology.org/2020.acl-main.740

8. Hashemi, H., Aliannejadi, M., Zamani, H., Croft, W.B.: Antique: A non-factoid question answering benchmark. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 166–173. Springer International Publishing, Cham (2020)

9. Hashemi, H., Zhuang, Y., Kothur, S.S.R., Prasad, S., Meij, E., Croft, W.B.: Dense retrieval adaptation using target domain description. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval. p. 95–104. ICTIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3578337.3605127, https://doi.org/10.1145/3578337.3605127

10. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-Efficient Transfer Learning for NLP (Jun 2019), http://arxiv.org/abs/1902.00751

11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models (Oct 2021), http://arxiv.org/abs/2106.09685, arXiv:2106.09685 [cs]

12. Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E.: Unsupervised Dense Information Retrieval with Contrastive Learning. Transactions on Machine Learning Research (May 2022), https://openreview.net/forum?id=jKN1pXi7b0

13. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.550, https://aclanthology.org/2020.emnlp-main.550

14. Krishna, K., Garg, S., Bigham, J., Lipton, Z.: Downstream datasets make surprisingly good pretraining corpora. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12207–12222. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.682, https://aclanthology.org/2023.acl-long.682

15. Lassance, C., Dejean, H., Clinchant, S.: An Experimental Study on Pretraining Transformers from Scratch for IR. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval, vol. 13980, pp. 504–520. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-28244-7_32, https://link.springer.com/10.1007/978-3-031-28244-7_32, series Title: Lecture Notes in Computer Science

16. Li, M., Gaussier, E.: Domain adaptation for dense retrieval through self-supervision by pseudo-relevance labeling (2022)

17. Li, X.L., Liang, P.: Prefix-Tuning: Optimizing Continuous Prompts for Generation (Jan 2021). https://doi.org/10.48550/arXiv.2101.00190, http://arxiv.org/abs/2101.00190, arXiv:2101.00190 [cs]

18. Litschko, R., Vulić, I., Glavaš, G.: Parameter-efficient neural reranking for cross-lingual and multilingual retrieval. In: Calzolari, N., Huang, C.R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.S., Ryu, P.M., Chen, H.H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T.K., Santus, E., Bond, F., Na, S.H. (eds.) Proceedings of the 29th International Conference on Computational Linguistics. pp. 1071–1082. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022), https://aclanthology.org/2022.coling-1.90

19. Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks (Mar 2022). https://doi.org/10.48550/arXiv.2110.07602, http://arxiv.org/abs/2110.07602, arXiv:2110.07602 [cs]

20. MacAvaney, S., Yates, A., Feldman, S., Downey, D., Cohan, A., Goharian, N.: Simplified data wrangling with ir_datasets. In: SIGIR (2021)

21. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: Ms marco: A human generated machine reading comprehension dataset. CoRR **abs/1611.09268** (2016), http://dblp.uni-trier.de/db/journals/corr/corr1611.html#NguyenRSGTMD16

22. Nogueira, R., Cho, K.: Passage re-ranking with bert (2019), http://arxiv.org/abs/1901.04085

23. Nogueira, R., Yang, W., Lin, J.J., Cho, K.: Document expansion by query prediction. ArXiv **abs/1904.08375** (2019), https://api.semanticscholar.org/CorpusID:119314259

24. Pal, V., Lassance, C., Déjean, H., Clinchant, S.: Parameter-efficient sparse retrievers and rerankers using adapters. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval. pp. 16–31. Springer Nature Switzerland, Cham (2023)
25. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010). https://doi.org/10.1109/TKDE.2009.191
26. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at trec-3. In: Text Retrieval Conference (1994), https://api.semanticscholar.org/CorpusID:3946054
27. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., Zaharia, M.: ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3715–3734. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.272, https://aclanthology.org/2022.naacl-main.272
28. Tam, W.L., Liu, X., Ji, K., Xue, L., Zhang, X., Dong, Y., Liu, J., Hu, M., Tang, J.: Parameter-Efficient Prompt Tuning Makes Generalized and Calibrated Neural Text Retrievers (Jul 2022). https://doi.org/10.48550/arXiv.2207.07087, http://arxiv.org/abs/2207.07087
29. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models (Oct 2021). https://doi.org/10.48550/arXiv.2104.08663, http://arxiv.org/abs/2104.08663
30. Wang, K., Thakur, N., Reimers, N., Gurevych, I.: Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2022). https://doi.org/10.18653/v1/2022.naacl-main.168, http://dx.doi.org/10.18653/v1/2022.naacl-main.168
31. Zhan, J., Ai, Q., Liu, Y., Mao, J., Xie, X., Zhang, M., Ma, S.: Disentangled Modeling of Domain and Relevance for Adaptable Dense Retrieval (Aug 2022), http://arxiv.org/abs/2208.05753, arXiv:2208.05753 [cs]
32. Zhang, X., Yates, A., Lin, J.: A little bit is worse than none: Ranking with limited training data. In: Moosavi, N.S., Fan, A., Shwartz, V., Glavaš, G., Joty, S., Wang, A., Wolf, T. (eds.) Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing. pp. 107–112. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.sustainlp-1.14, https://aclanthology.org/2020.sustainlp-1.14
33. Zong, Y., Piwowarski, B.: Xpmir: A modular library for learning to rank and neural ir experiments. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 3185–3189. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539618.3591818, https://doi.org/10.1145/3539618.3591818