Retrieval-Augmented Domain Adaptation of Language Models

Benfeng Xu*, **Chunxu Zhao***, **Wenbin Jiang**†, **PengFei Zhu**
Songtai Dai, Chao Pang, Zhuo Sun, Shuohuan Wang, Yu Sun
Baidu Inc.

Abstract

Language models pretrained on general domain corpora usually exhibit considerable degradation when generalizing to downstream tasks of specialized domains. Existing approaches try to construct PLMs for each specific domains either from scratch or through further pretraining, which not only costs substantial resources, but also fails to cover all target domains at various granularity. In this work, we propose RADA, a novel Retrieval-Augmented framework for Domain Adaptation. We first construct a textual corpora that covers the downstream task at flexible domain granularity and resource availability. We employ it as a pluggable datastore to retrieve informative background knowledge, and integrate them into the standard language model framework to augment representations. We then propose a two-level selection scheme to integrate the most relevant information while alleviating irrelevant noises. Specifically, we introduce a differentiable sampling module as well as an attention mechanism to achieve both passage-level and word-level selection. Such a retrieval-augmented framework enables domain adaptation of language models with flexible domain coverage and fine-grained domain knowledge integration. We conduct comprehensive experiments across biomedical, science and legal domains to demonstrate the effectiveness of the overall framework, and its advantage over existing solutions.

1 Introduction

Language models pretrained on large scale of unsupervised corpora are capable of producing powerful representations as well as providing satisfactory performance when finetuned for general domain downstream tasks (Devlin et al., 2019; Brown et al., 2020). However, as such representations are learned from general domain distributions, their generalization performance are deteriorate on specialized domains where distribution are much more different, such as biomedicine, science, legal, etc.

Many works thus trying to construct domain-specific PLMs either from scratch or initialized from the original PLM, such as BioBERT (Lee et al., 2020), LegalBERT (Chalkidis et al., 2020), SciBERT (Beltagy et al., 2019), etc. Through the pretraining procedure, domain-specific knowledge are memorized and internalized into the parameter of the PLMs, thus providing a better initialization point for learning downstream tasks. However, such attempts not only consume much more costs, as it requires considerable computation to pretrain a language model, but also can not cover adaptive needs at task- or even instance-level granularity, as domain-specific PLMs are preliminarily trained with a fixed scale domain corpora but applied for all domain tasks. In fact, ***domain*** can be defined at various granularity w.r.t. various applications (Chronopoulou et al., 2022). For example, in the biomedical domain, texts might distribute across academic publication, medical record or medical-situated dialog, and exhibit completely different style and background knowledge.

In this paper, we novelly propose a **R**etrieval-**A**ugmented framework for **D**omain **A**daptation (**RADA**) to address the above challenges. Instead of internalizing a fixed corpora into PLM parameters via further pretraining, we construct and retrieve a datastore to dynamically enhance the learning of domain-specific downstream tasks. The flexibility of such a retrieval-augmented framework is in two-fold: 1) On one hand, we no longer need to maintain multiple PLMs respectively for each application domains, all domains simply share the same checkpoint for initialization but with domain-specialized datastore. 2) On the other hand, the datastore only consists of most relevant background passages tailored for each specific task. Such task-level granularity provides best trade-off between

*Equal contribution.
†Correspondence to: jiangwenbin@baidu.com

Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), pages 54–64

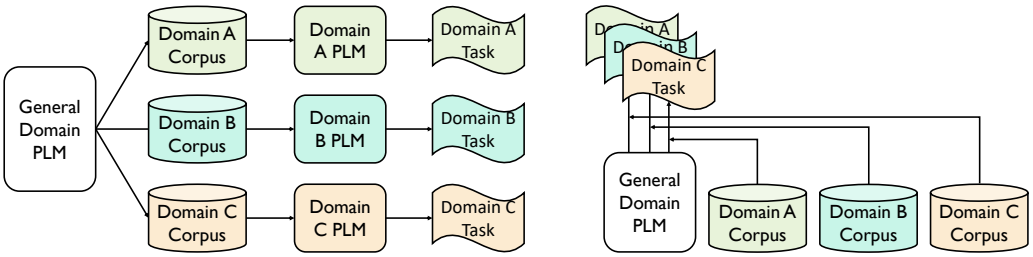

Figure 1: Illustration of the motivated retrieval-augmented framework for language model domain adaptation.

scale and coverage, we either do not need to collect a very large corpora in order to effectively cover all sub-domains, nor concerned with its coverage for this specific task. The illustration of the motivated framework compared to existing further pretraining solution is given in Figure 1.

The core component of such a retrieval-augmented framework remains how to effectively integrate the retrieved passages. One line of related works is knowledge-enhanced PLMs (Zhang et al., 2019; Peters et al., 2019; Liu et al., 2020), they mostly resort to attention mechanism to directly incorporate the matched knowledge triples. Differently, we consider textual domain corpus instead of structured knowledge in this paper as the latter only has very limited coverage, especially for domain specialized scenarios. As a consequence, we aim to integrate retrieved passages that are less knowledge-intensive and might be mixed with noisy information that are irrelevant to solve the task. We accordingly propose a two-level selection scheme: passage-level selection and word-level selection. Among the multiple retrieved passages given by a coarse-grained but efficient retriever, we first propose a differentiable sampling module that enables the model to dynamically select the most helpful background passage. We then adopt an attention mechanism to weigh all words inside the selected passage to achieve more fine-grained integration. The retrieval-augmentation interface works as a pluggable module that needs no further modification of the main encoding pipeline.

We conduct experiments on 7 popular tasks covering three different downstream domains and a wide range of task variety. The results and ablations demonstrate the effectiveness and advantage of the proposed framework. The contribution of this paper is three-fold:

- **Conceptual Contribution** We novelly propose a retrieval-augmented framework for domain-adaptation of PLMs. The framework is designed as a pluggable module, which not only saves the costs to construct and maintain specialized PLMs for each respective domain, but also provides domain specialty at flexible task-level granularity with no concern of the trade-off between domain scale and coverage.
- **Technical Contribution** We propose a two-level selection scheme to achieve fine-grained integration of retrieved passages while alleviate noise distraction. Specifically, we introduce a differentiable sampling module for passage-level selection and an attention mechanism to further assign word-level importance.
- **Experimental Contribution** We achieve consistent improvements across different domains including Science, Biomedicine and Legal. When combined with further pretraining, the improvements of the proposed framework are much more significant, surpassing competitive baseline of vanilla further pretraining.

2 Related Work

Domain Adaptation (Ramponi and Plank, 2020) is an important and widely investigated problem in the area of NLP. In the current pretrain-finetune paradigm, it becomes even more practical as most of NLP systems start with a general-purpose PLM but applied into various downstream scenarios (Guo and Yu, 2022). Many early works have found that continued pretraining of PLM on domain specific corpus could bring benefits (Alsentzer et al., 2019; Lee et al., 2020). Gururangan et al. (2020) take a step further to systematically investigate the effect of further pretraining on multiple domains, which they refer to as domain-adaptive

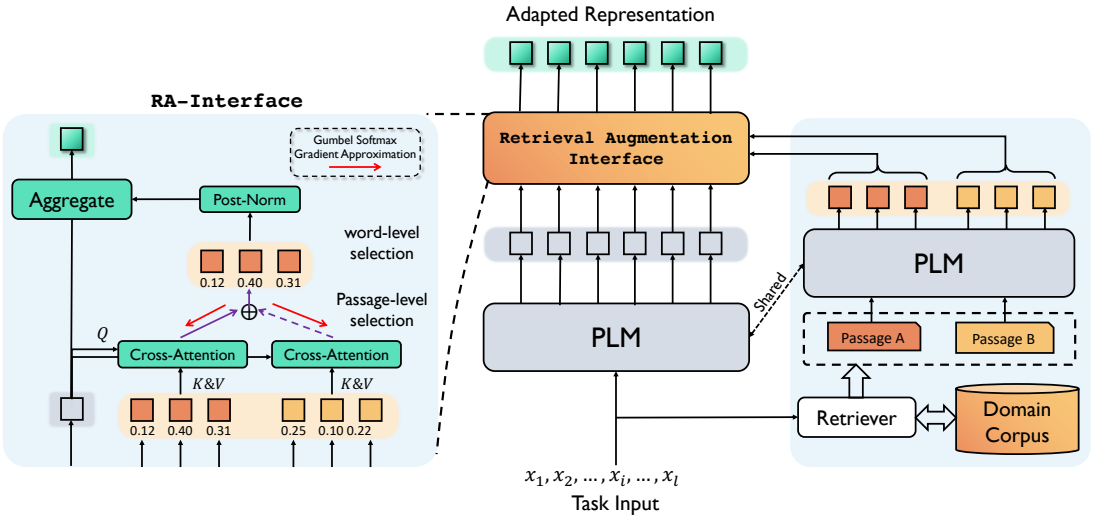

Figure 2: The overall retrieval-augmented framework of RADA. The **RA-Interface** serves as a pluggable module to augment task-agnostic input representations with domain-specialized background passages.

pretraining (DAPT). Besides, they also propose task-adaptive pretraining (TAPT) to perform further pretraining also on task data, which further brings significant improvements.

There are mainly two concerns for these further pretraining based methods. One is the cost to construct and maintain considerable number of PLMs for each specific domain, which can be quite a lot spreading over various application scenarios. The other is the complication to formally define the concept of domain because domains exhibit hierarchical taxonomy in practical scenarios (Reid et al., 2022). The scope of a domain should neither be too broad, which results in loss of domain specialty, nor be too narrow, which results in loss of generalization ability. Many recent works (Chronopoulou et al., 2022; Gururangan et al., 2022; Li et al., 2022) thus propose to incorporate multiple domains into one shared model, and introduce mixture-of-experts module to allow different sub-domains also share with each other. By contrast, RADA takes a very different retrieval-based road to resolve the above concerns.

Several other works also try to advance domain adaptation from various perspectives. For example, Yu et al. (2021) studied DAPT for the specific task of abstractive summarization. Xu et al. (2021) propose gradual finetuning for domain adaptation. Bai et al. (2021) investigates the trade-off between DAPT and data annotation under limited budget for domain adaptation.

Another very related line of work is retrieval-augmented language models. kNN-LM (Khandelwal et al., 2020) retrieves for similar language modeling probability distributions and interpolate them with the current step of token generation. Guu et al. (2020) and Lee et al. (2019) propose to learn differentiable retriever to better select helpful passages. Lewis et al. (2020) propose a non-parametric retriever and combine it with a generator model. Borgeaud et al. (2022) use attention mechanism to integrate retrieved passages and also explore the datastore in a very large scale. Izacard et al. (2022) investigate the usefulness of retrieval-augmentation in the context of few-shot prompting with LLMs. Although being technically related, most of these works focus on general-purpose language understanding, especially knowledge-intensive tasks. While in this paper, we focus on the challenge of LM domain adaptation, and also novelly propose a two-level selection scheme to further augment the framework.

3 Methodology

RADA is a general framework for task-agnostic domain adaptation, and can be readily integrated into the standard pipeline of PLM deployment. In the following section, we first brief the construction of domain specialized datastore in Section 3.1, and the framework formulation in Section 3.2. We then introduce the core component, a pluggable interface for retrieval-augmented domain adaptation in Section 3.3, and finally brief the training usage in Section 3.4. The overall framework is illustrated in Figure 2.

3.1 Datastore Construction

We consider textual domain corpus as a necessary recipe to adapt a general-domain language model. From their related sources, we collect domain specialized passages $\{\boldsymbol{p}_i\}^{\mathcal{D}}$, and refer to them as the datastore. In practical implementation, we truncate the length of each passage to 256 tokens for simplicity. We then build an Elasticsearch service[†], for any task-specific training or test input sequence x, we can retrieve for its K most relevant background passages $\{\boldsymbol{p}_k\}_{k=1}^{K}$. Although Elasticsearch only serves coarse-grained retrieving purpose, we can employ it as an efficient first-stage retriever.

For each specific downstream task, the datastore is constructed at **task-level** granularity. And the retrieving of most relevant passages further achieves **instance-level** granularity. This is in contrast with many existing works where datastore is constructed and utilized at **domain-level** granularity, once a target domain is determined, a fixed-scale datastore is collected and injected into PLM as an integral through further pretraining. As a result, RADA brings better flexibility and save the difficulty to deliberate the trade off between domain specialty and domain coverage.

3.2 Retrieval-Augmented PLM

In a standard PLM-based framework, we first take the task input sequence $\boldsymbol{x} = \{x_i\}$ and use the pretrained encoder to produce contextualized representations:

$$\boldsymbol{H}^{\boldsymbol{x}} = \text{encode}(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{H}^{\boldsymbol{x}} \in \mathbb{R}^{l*d}$ is the sequence level representation for l words and d dimensions. We accordingly refer to $\boldsymbol{h}_i^{\boldsymbol{x}}$ as the contextualized representation at each specific position i. The representations are then fed into a task-tailored head module, such as classification, regression or sequence tagging, etc.

In RADA, accompanied with each task input, we similarly construct the representations for retrieved passages $\{\boldsymbol{p}_k\}$:

$$\boldsymbol{H}^{\boldsymbol{p}_k} = \text{encode}(\boldsymbol{p}_k), \quad 1 \leq k \leq K \tag{2}$$

and we similarly get $\boldsymbol{H}^{\boldsymbol{p}_k}$ and $\boldsymbol{h}_j^{\boldsymbol{p}_k}$. The purpose of the proposed RADA framework can be formulated as a universal interface that augments and updates task representations $\boldsymbol{h}_i^{\boldsymbol{x}}$ using $\boldsymbol{H}^{\boldsymbol{p}_k}$, where $\boldsymbol{H}^{\boldsymbol{p}_k}$ serves as informative, relevant and domain-specialized knowledge source:

$$\widetilde{\boldsymbol{h}}_i = \texttt{RA-Interface}(\boldsymbol{h}_i^{\boldsymbol{x}}, \{\boldsymbol{H}^{\boldsymbol{p}_k}\}_{k=1}^{K}) \tag{3}$$

$\widetilde{\boldsymbol{h}}_i$ is then unchangeably used in the rest of the task pipeline. The overall framework remains task-agnostic and can be readily integrated into both pretraining and finetuning stage. The interface can also be easily extended to a more broad scenario with different backbone networks other than Transformer.

3.3 Retrieval Augmentation Interface

We elaborate the **`RA-Interface`** in this section. Different from many existing methods that resort to integrate structured knowledge (Zhang et al., 2019; Peters et al., 2019), the retrieved textual domain corpus are much less knowledge-intensive and thus might contain irrelevant noise distractions. However, Elasticsearch can only satisfy coarse-grained retrieving purpose, leaving it an important problem to effectively select and integrate actually helpful domain knowledge.

To address such challenge, we propose a two-level selection scheme which performs discrete sampling at passage-level and attentively aggregate at word-level. Such selection strategy strengthens the proposed interface with fine-grained selection capability to integrate helpful knowledge while filter out its noisy parts. More specifically, we introduce a gumbel sampling mechanism to select the most helpful passage, along with an attention mechanism to assign distinguishable weights for each word insides the selected passage.

3.3.1 Cross Attention

In order to cover both sequence-level and token-level downstream tasks, we would need to augment the representation $\boldsymbol{H}_{\boldsymbol{x}}$ at each specific position. For the i-th token representation $\boldsymbol{h}_{x_i}$, we calculate its relevancy with respect to all other words from the retrieved passages using a canonical cross attention mechanism:

$$\begin{aligned} \boldsymbol{q}_i^{\boldsymbol{x}} &= \boldsymbol{W}^Q \boldsymbol{h}_i^{\boldsymbol{x}} \\ \boldsymbol{k}_i^{\boldsymbol{p}_k} &= \boldsymbol{W}^K \boldsymbol{h}_j^{\boldsymbol{p}_k} \\ \boldsymbol{v}_i^{\boldsymbol{p}_k} &= \boldsymbol{W}^V \boldsymbol{h}_j^{\boldsymbol{p}_k} \end{aligned} \tag{4}$$

$$\alpha_{ij}^k = \frac{\boldsymbol{q}_i^{\boldsymbol{x}} \boldsymbol{k}_j^{\boldsymbol{p}_k}}{\sqrt{d}} \tag{5}$$

[†] https://github.com/elastic/elasticsearch

where α_{ij}^k denotes the attentive weight of j-th token from passage p_k for current position $\boldsymbol{h}_i^x$. Intuitively, this attentive weight can naturally measure each retrieved word its importance and relevance. To represent the usefulness of passages and their tokens at integral sequence level, we simply summing over all task input positions:

$$\alpha_j^k = \sum_i \alpha_{ij}^k \tag{6}$$

3.3.2 Differentiable Sampling

To alleviate the noise distraction in the retrieved passages, we first sample the most relevant passage, then integrate all its words according to respective attentive weights. Intuitively, the sampling operation should be an argmax operation across all candidate passages:

$$\alpha^k = \sum_j \alpha_j^k \tag{7}$$

$$\hat{k} = \underset{k \in \{1,2,...,K\}}{\text{argmax}} \ \alpha^k \tag{8}$$

Note that here we sum α_j^k up to α^k only temporarily for sampling most relevant passage, and will re-use α_j^k later in Equation 12 to aggregate representations. The passage level selection operation in Equation 8, although intuitive, is not differentiable, and are not allowed in an end-to-end trainable neural network. To enable such a learning-to-select ability inside the proposed framework, we novelly introduce the Gumbel-Softmax estimator (Jang et al., 2017) to replace Equation 8:

$$\beta^k = \frac{\exp((\alpha^k + g)/\tau)}{\sum_k \exp((\alpha^k + g)/\tau)} \tag{9}$$

where g is the noise sampled from Gumbel distribution:

$$g \sim \text{Gumbel}(0, 1) \tag{10}$$

This is practically implemented as:

$$\begin{aligned} g &= -\log(-\log(\boldsymbol{u})) \\ where \quad \boldsymbol{u} &\sim \text{Uniform}(0, 1) \end{aligned} \tag{11}$$

The resulting $\boldsymbol{\beta}$ distributed over $k \in \{1, 2, ..., K\}$ is an approximate **categorical** distribution when the temperature τ approaches 0. We follow the Straight-Through procedure (Jang et al., 2017) to discretize $\boldsymbol{\beta}$ into $\hat{k}$ in the forward pass but use gumbel approximation in the backward pass.

3.3.3 Aggregation

For the selected passage $\hat{k}$, we apply a post-normalization on the attentive weights:

$$\tilde{\alpha}_j^{\hat{k}} = \frac{\alpha_j^{\hat{k}}}{\sum_j \alpha_j^{\hat{k}}}, \quad \sum_j \tilde{\alpha}_j^{\hat{k}} = 1 \tag{12}$$

We then integrate the representations of passage $\hat{k}$ accordingly:

$$\hat{\boldsymbol{h}} = \sum_j \tilde{\alpha}_j^{\hat{k}} \boldsymbol{v}_j^{p_{\hat{k}}} \tag{13}$$

Note that in practical implementation, we use multi-head attentions and concatenate all outputs together. We then wrap $\hat{\boldsymbol{h}}$ within a residual connection and LayerNorm at each position i to produce the final output representation:

$$\widetilde{\boldsymbol{h}}_i = \text{LayerNorm}(\boldsymbol{W}\hat{\boldsymbol{h}} + \boldsymbol{h}_i^x) \tag{14}$$

where $\boldsymbol{W}$ is another feedforward layer that merges representations from multiple attention heads.

3.4 Retrieval-Augmented Training

The proposed **`RA-Interface`** serves as a **plug-gable** module that interacts with the representations in-place. It is also end-to-end trainable and does not modify the final training objective. In practical usages, we can either directly apply such a retrieval-augmented framework in downstream finetuning, or first adapt it with a domain-specific further pre-training stage. We investigate both settings in the experiments.

4 Experiments

4.1 Setup

Dataset We investigate the proposed framework on a variety of domain-specialized downstream tasks. Specifically, biomedical domain tasks including QIC (Intent Classification), QQR (Query-Query Relevance), CMeEE (Named Entity Recognition), CMeIE (Information Extraction), scientific domain tasks including SciERC (Relation Classification) (Luan et al., 2018), ACL-ARC (Citation Intent Classification) (Jurgens et al., 2018) and legal domain task CAIL2019-SCM (Similar Case Matching) (Xiao et al., 2019). And these 4 biomedical tasks are selected from the well established benchmark CBLUE (Zhang et al., 2022). Detailed statistics are listed in Table 4.

Domain	Biomedical				Science		Legal	Avg.
Method	QIC	QQR	CMeEE	CMeIE	SCI	ARC	SCM	
FT	$81.86_{0.07}$	$81.97_{0.81}$	$65.45_{0.07}$	$61.31_{0.20}$	$81.39_{1.03}$	$68.33_{2.05}$	$68.97_{0.21}$	72.75
RADA	$82.43_{0.21}$	$82.18_{0.74}$	$65.55_{0.18}$	$60.07_{0.49}$	$79.76_{0.48}$	$68.87_{2.86}$	$69.70_{0.46}$	72.87
DAPT	$81.74_{0.16}$	$82.06_{0.77}$	$65.57_{0.12}$	$61.42_{0.12}$	$81.36_{0.46}$	$73.14_{4.23}$	$69.10_{0.35}$	73.50
RADA†	$\mathbf{82.45}_{0.10}$	$\mathbf{82.75}_{0.67}$	$\mathbf{66.00}_{0.20}$	$\mathbf{61.53}_{0.24}$	$\mathbf{82.07}_{0.86}$	$\mathbf{75.42}_{1.48}$	$\mathbf{70.25}_{0.32}$	**74.34**

Table 1: Main Results. FT refers to standard finetuning. † denotes the RADA equipped with a pretraining stage.

Domain	Biomedical				Science		Legal	Avg.
Method	QIC	QQR	CMeEE	CMeIE	SCI	ARC	SCM	
FT	$81.07_{0.41}$	$80.52_{0.28}$	$63.24_{0.06}$	$53.97_{0.23}$	$70.77_{0.66}$	$56.46_{1.91}$	$61.20_{0.42}$	66.75
DAPT	$80.81_{0.35}$	$80.62_{0.43}$	$63.56_{0.13}$	$54.29_{0.13}$	$\mathbf{74.89}_{0.63}$	$62.15_{0.86}$	$61.31_{0.34}$	68.23
RADA	$\mathbf{81.16}_{0.19}$	$\mathbf{81.06}_{0.33}$	$\mathbf{63.62}_{0.22}$	$\mathbf{54.48}_{0.14}$	$73.96_{0.56}$	$\mathbf{62.59}_{2.80}$	$\mathbf{61.70}_{0.11}$	**68.36**

Table 2: Results under low resource scenario.

Domain	LANG	Size	Num. of Passages
Computer	English	6.2GB	4,602,628
Biomedicine	Chinese	0.76GB	960,595
Legal	Chinese	3.7GB	2,794,605

Table 3: Statistics of domain datastore.

Dataset	SCI	ARC	QIC	QQR	CMeEE	CMeIE	SCM
Domain	Science		Biomedical				Legal
Train	3219	1688	6238	13500	13500	12905	5102
Dev	455	144	693	1500	1500	1434	1500
Test	974	139	1955	1600	5000	3585	1500

Table 4: Statistics of domain tasks. For the biomedical tasks, as we do not have access to the test label, we use the released dev set for test, and split 10% of the train set for development.

Datastore We collect domain specialized corpus at scale as datastore. For scientific corpus, we use the S2ORC corpus constructed from semantic scholar (Lo et al., 2020), for biomedical and legal corpus, we use in-house data regarding medical reviews or legal documents crawled online. Detailed statistics are listed in Table 3.

Training For retrieval-augmented pretraining, we split 10% from the entire domain corpus as pretraining pretext task data while the rest 90% as pretraining datastore. The pretraining steps are set to 10k for both DAPT and RADA. For retrieval-augmented finetuning, all corpus are considered as datastore, but we only retrieval and keep the most relevant passages for each task instances. In ElasticSearch, we set the number of retrieved passages for each input data as $K = 10$. For each task, we search batch size through $\{16, 32\}$, learning rate through $\{1e-5, 2e-5, 5e-5\}$†, and set epoch to 10†. We also run with 3 different random seeds, and accordingly report the average and standard deviation. We use BERT (Devlin et al., 2019) as PLM for English tasks, and RoBERTa (Cui et al., 2020; Liu et al., 2019) for Chinese ones.

4.2 Main Results

Table 1 gives the main results for RADA. We can observe two conclusions: 1) RADA outperforms standard finetuning on 4 out of 7 selected tasks; 2) When combined with further pretraining, the proposed framework is further improved, significantly surpassing standard finetuning and DAPT. Specifically the absolute benefits over FT baseline are **+1.59**.

We further investigate a low-resource setting by subsampling 30% of the training set. This is a very practical scenario as we often need to deal with domain tasks at a relatively low annotation cost. Results in Table 2 also demonstrate the effectiveness of the proposed framework in such settings. On 6 out of 7 tasks, RADA achieves the best performance.

4.3 Ablations

We investigate the impact of various components in this section.

†For CMeIE, we search for {1e-5, 2e-5}.
†For CMeIE and CMeEE, we set to 35 and 15 respectively.

Domain	Biomedical				Science		Legal	Avg.
Method	QIC	QQR	CMeEE	CMeIE	SCI	ARC	SCM	
FT	$81.86_{0.07}$	$81.97_{0.81}$	$65.45_{0.07}$	$61.31_{0.20}$	$81.39_{1.03}$	$68.33_{2.05}$	$68.97_{0.21}$	72.75
RADA w/ Top 1	$82.31_{0.25}$	$82.56_{0.43}$	$\mathbf{66.08}_{0.26}$	$61.48_{0.05}$	$81.78_{0.25}$	$74.15_{0.96}$	$69.14_{0.26}$	73.93
RADA w/ DS	$\mathbf{82.45}_{0.10}$	$\mathbf{82.75}_{0.67}$	$66.00_{0.20}$	$\mathbf{61.53}_{0.24}$	$\mathbf{82.07}_{0.86}$	$\mathbf{75.42}_{1.48}$	$\mathbf{70.25}_{0.32}$	74.34

Table 5: Effects of Differentiable Sampling (DS). Top 1 means we simply select the passage according to their summed attention score.

Method	ARC	QIC	QQR
FT	$68.33_{2.05}$	$81.86_{0.07}$	$81.97_{0.81}$
TAPT	$\mathbf{72.43}_{2.36}$	$\underline{82.14}_{0.04}$	$\mathbf{82.71}_{0.57}$
RADA w/ Trainset	$\underline{70.65}_{0.38}$	$\mathbf{82.21}_{0.21}$	$\underline{82.52}_{0.46}$
RADA (Full)	$75.42_{1.48}$	$82.45_{0.10}$	$82.75_{0.67}$

Table 6: Results of using training data as datastore corpus. Full means extra domain corpus are used.

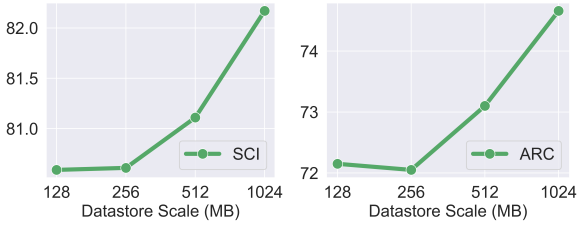

Figure 3: Ablation on scale of corpus.

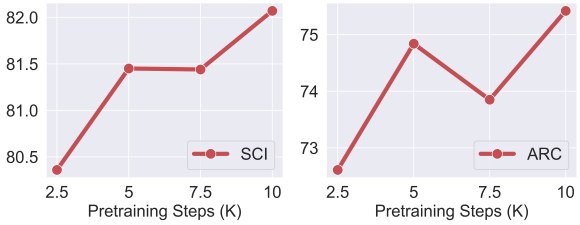

Figure 4: Ablation on pretraining steps.

4.3.1 Scale of Datastore Corpus

One essential component of the proposed framework is the datastore. We first look into the effects of its scale. At pretraining stage, we fix the pretraining steps to 5,000, and accordingly set the scale of datastore to 128, 256, 512 and 1024. The results are illustrated in Figure 3. We observe clear trends of increasing performance w.r.t. increased datastore scale. As a consequence, the proposed RADA framework can always benefit from a larger datastore.

4.3.2 Training Data as Datastore

We further investigate the feasibility of directly using downstream task data as datastore corpus. As previous study has demonstrated, training data itself can also provide useful background information to an extent (Wang et al., 2022). Similarly, Gururangan et al. (2020) have also used task training data to perform further pretraining, which they referred to as TAPT. In Table 6, we provide the results when using training data in the proposed retrieval-augmented framework, and also reproduce TAPT for more comprehensive comparison. The results show that both TAPT and RADA with Trainset can provide considerable benefits, but are outperformed by RADA (full). As for these two methods, they perform comparably on the investigated three datasets. We also find that training data is more useful in sentence-level tasks (as included in the table), but less helpful in other tasks such as classification, sequence tagging, etc.

4.3.3 Pretraining Steps

We have equipped RADA with a further pretraining stage with domain specialized data, and proved it to be very effective in Table 1. In Figure 4, we further ablate the effects of pretraining steps. We save and evaluate the intermediate checkpoint at respectively 2.5k, 5k and 7.5k pretraining steps. The results exhibit clear trends that with more pretraining steps, RADA continually brings better adaptation performance.

4.3.4 Differentiable Sampling

One key design of RADA is the two-level selection scheme. Specifically, the differentiable sampling module based on Gumbel Sampling trick enables the model to dynamically learn which passage to integrate. In Table 5 we investigate this choice of design. We provide an alternative baseline, where instead of learning to sample, we simply sum the cross attention weights over all positions as passage-level score, and select the maximum one to integrate into the retrieval-augmented interface. We refer to this as Top 1 selection strategy. The results show that both methods can bring improve-

• Input Text: This formalism is both elementary and powerful enough to strongly simulate many grammar formalisms, such as rewriting systems, dependency grammars , TAG , HPSG and LFG . • Relation Label: HYPONYM-OF			Sampling Score β^k
Retrieved Passages (ES Top 10)	1	… Although we have presented supertagging in the context of LTAG, it is applicable to other lexicalized grammar formalisms such as CCG (Steedrnan 1997), HPSG (Pollard and Sag 1987) , and LFG (Kaplan and Bresnan 1983) …	0.0992
	2	Scrambling in German poses a problem for most grammar formalisms. Neither Tree Adjoining Grammar (TAG, Joshi et al., 1975) nor even linear context-free rewriting systems (LCFRS, Weir, 1988) are powerful enough to deal with scrambling and the free word order in German (see Becker et al., 1992) …	0.1737
	3	… We would also expect that dependency grammars Mel'cuk and Pertsov 1987; Hudson 1984) and parsers (McDonald, Crammer, and Pereira 2005) could be trained and tested with little extra work on the dependencies in CCGbank. Finally, we believe that existing methods for translating the Penn Treebank from scratch into other grammar formalisms will benefit from …	0.1046
	…	…	…
	10	… The Broad-coverage Semantic Dependency Parsing shared task and corpora (Oepen et al., 2014 (Oepen et al., , 2015 include corpora annotated with the PDT-TL, and dependencies extracted from the HPSG grammars Enju (Miyao, 2006) and the LinGO English Reference Grammar (ERG; Flickinger, 2002) . Like the PDT-TL, projects based on CCG, HPSG, and other expressive grammars such as LTAG (Joshi and Vijay-Shanker, 1999) and LFG …	0.1018

Figure 5: Case study. Illustrated are top 10 retrieved passages using elasticsearch, and their sampling score produce inside the **`RA-Interface`**. Scores are re-normalized for better illustration.

	w/o RA	w/ RA
Efficiency	0.0083 sec/instance	0.0201 sec/instance
Times	2.4×	1.0×

Table 7: Inference efficiency. Measured using a single RTX TITAN GPU.

ments, but the proposed differentiable sampling is much more effective.

4.3.5 Efficiency Analysis

One potential concern for retrieval-augmented methods is their efficiency. We therefore investigate this factor in Table 7. We consider the inference efficiency at deployment time. For each task input, we retrieve 10 background passages using ElasticSearch, then compute and incorporate them into the **`RA-Interface`**. Note that at deployment time, it is practical to encode and cache all passages from the datastore in advance, so we only account for the time consumption starting from the interface. The results show that, the retrieval-augmentation framework only brings acceptable time increase. And the overall inference speed is maintained at around 0.02 seconds per instance on a production-level device.

4.3.6 Qualitative Analysis

We provide qualitative analysis in Table 5. The example is sampled from the SCI task, the target is to extract the relation between subject ***rewriting systems*** and object ***grammar formalisms***, i.e., ***HYPONYM-OF***. We can clearly see that elasticsearch can only provide coarse-grained, shallow-semantic retrieving capability based on keywords, such as *grammar, formalisms, HPSG, etc.* However, the proposed differentiable sampling module can more effectively identify the most helpful passage by reasoning over deep representations produced by shared encoders. From passage 2 with the highest sampling score, we can indeed reason, understand and accordingly infer the actual relationship between the target entity pair.

5 Conclusion

In this paper we propose a retrieval-augmented framework to novelly address the challenge of language model domain adaptation. We use domain specialized corpus as datastore and retrieve from it for informative and helpful domain knowledge. The key module of the framework is a retrieval-augmentation interface, where we design a two-level selection scheme to integrate the most relevant passage and its words while alleviating the noise. The overall framework enables flexible domain coverage and fine-grained domain knowledge integration. On a variety of downstream domains and tasks, we conduct comprehensive experiments and comparisons to demonstrate the effectiveness of the motivated framework and its components. In the future, we hope to further extend the framework to more scaled large language models and also more challenging few-shot prompting setting.

Limitation

We summarize two limitations which also serve as promising directions to be explored in our future work. RADA framework only considers textual domain corpus as the datastore, although this has greatly improve the coverage of domain knowledge as texts are always relatively easy to collect. However, it is widely investigated that structured knowledge such as knowledge base can also serve similar purpose. And such resources are generally in higher quality and are easier to match. Therefore, it would be benefiting to further integrate such resources at certain scenario where KB is available.

The other limitation regards to the scale of the RADA implementation. As large language models have becoming increasingly powerful, they have demonstrated quite impressive capability in memorizing and recalling a wide range of background knowledge existed in the massive corpora they have been pretrained on. This trends of development naturally raises question for the proposed framework: will it still be beneficial when scale up to LLMs? and on what kinds of scenario does it brings best improvements? These are very important questions to answer, and we can certainly expect them to be explored in future works.

Ethical Statement

We evaluate the proposed method on established and publicly available datasets. There is also no human evaluation involved. This paper is not concerned with the above ethical risks. When the proposed framework is deployed into domain specific production, the domain adapted language models might express ethical-related outputs, but just as any other language models do (Weidinger et al., 2021), and should be treated with according techniques to eliminate ethical risks such as bias, stereotypes.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Fan Bai, Alan Ritter, and Wei Xu. 2021. Pre-train or annotate? domain adaptation with a constrained budget. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5002–5015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2898–2904, Online. Association for Computational Linguistics.

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. 2022. Efficient hierarchical domain adaptation for pretrained language models. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1336–1351, Seattle, United States. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In Proceedings of the 2020 Conference on Empirical

Methods in Natural Language Processing: Findings, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xu Guo and Han Yu. 2022. On the domain adaptation and generalization of pretrained language models: A survey. arXiv preprint arXiv:2211.03154.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A. Smith, and Luke Zettlemoyer. 2022. DEMix layers: Disentangling domains for modular language modeling. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5557–5576, Seattle, United States. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 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, pages 8342–8360, Online. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning, ICML'20. JMLR.org.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. arXiv preprint arXiv:2208.03299.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In International Conference on Learning Representations.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. Transactions of the Association for Computational Linguistics, 6:391–406.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In International Conference on Learning Representations.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

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, Florence, Italy. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Advances in Neural Information Processing Systems, volume 33, pages 9459–9474. Curran Associates, Inc.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A Smith, and Luke Zettlemoyer. 2022. Branch-train-merge: Embarrassingly parallel training of expert language models. arXiv preprint arXiv:2208.03306.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. Proceedings of the AAAI Conference on Artificial Intelligence, 34(03):2901–2908.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4969–4983, Online. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Machel Reid, Victor Zhong, Suchin Gururangan, and Luke Zettlemoyer. 2022. M2d2: A massively multi-domain language modeling dataset. arXiv preprint arXiv:2210.07370.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3170–3179, Dublin, Ireland. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. arXiv preprint arXiv:1911.08962.

Haoran Xu, Seth Ebner, Mahsa Yarmohammadi, Aaron Steven White, Benjamin Van Durme, and Kenton Murray. 2021. Gradual fine-tuning for low-resource domain adaptation. In Proceedings of the Second Workshop on Domain Adaptation for NLP, pages 214–221, Kyiv, Ukraine. Association for Computational Linguistics.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5892–5904, Online. Association for Computational Linguistics.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.