DomainInv: Domain Invariant Fine Tuning and Adversarial Label Correction For Unsupervised QA Domain Adaptation

Anant Khandelwal
Applied Scientist, Microsoft India
anantk@microsoft.com

Abstract

Existing Question Answering (QA) systems are limited in their ability to answer questions from unseen domains or any out-of-domain distributions, making them less reliable for deployment in real scenarios. Importantly, all existing QA domain adaptation methods are either based on generating synthetic data or pseudo-labeling the target domain data. Domain adaptation methods relying on synthetic data and pseudo-labeling suffer from either the need for extensive computational resources or an additional overhead of carefully selecting the confidence threshold to distinguish noisy examples from the training dataset. In this paper, we propose unsupervised domain adaptation for an unlabeled target domain by transferring the target representation close to the source domain without using supervision from the target domain. To achieve this, we introduce the idea of domain-invariant fine-tuning along with adversarial label correction (DomainInv) to identify target instances that are distant from the source domain. This involves learning the domain invariant feature encoder to minimize the distance between such target instances and source instances class-wisely. This eliminates the possibility of learning features of the target domain that are still close to the source support but are ambiguous. The evaluation of our QA domain adaptation method, namely DomainInv, on multiple target QA datasets reveals a performance improvement over the strongest baseline.

1 Introduction

Over the past few years, machine learning models have been widely deployed in production. However, making them work satisfactorily in production requires a substantial amount of high-quality annotated data, which is expensive and time-consuming. Therefore, it is of utmost importance to build generalizable models that can perform well on unseen datasets. However, due to the mechanism of *domain shift* or *bias* in the training dataset (Ben-David et al., 2010, 2006), it is challenging to directly transfer knowledge from the model trained on the source domain to the unlabeled target domain. In this paper, we studied this phenomenon specifically for the case of extractive Question Answering (QA) systems.

Extractive QA systems perform the task of identifying the most relevant answer for a given question within a context or paragraph. The answer is represented as a sub-span of the context, with start and end positions predicted by the QA model. The training data for QA essentially consists of triplets specifying the question, answer, and context. The input to the model is a question and context presented as running text separated by a separator. The model is trained to predict the most relevant start and end positions in the context (Seo et al., 2016; Chen et al., 2017; Devlin et al., 2019; Kratzwald et al., 2019).

These QA systems also face performance degradation at test time, as questions and contexts can vary widely in complexity. The same question may be phrased in the simplest or most complex ways, and the answer may involve reasoning or follow a complex extraction pattern that is challenging to generalize with a limited annotated training dataset. Recent works (Fisch et al., 2019a; Miller et al., 2020; Zeng et al.) have explored this issue, proposing solutions such as using labeled target domain data or incorporating feedback during training (Daumé III, 2007; Kratzwald et al., 2020; Kamath et al., 2020). Others (Yue et al., 2022c, 2021) have employed synthetic or pseudo-labeled data to train these systems and enhance their generalization to out-of-domain distributions. However, it is important to note that pseudo-labeled data is prone to noise, and obtaining accurately labeled data requires considerable human labeling effort.

In this paper, we focus on unsupervised domain adaptation (UDA), which does not require labeled target domain data. There are numerous

Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024), pages 13–25
August 15, 2024 ©2024 Association for Computational Linguistics

works towards achieving domain-invariant representations, categorized into 1) optimizing the discrepancy between domain representations (Yue et al., 2022c, 2021), and 2) adversarial learning (Lee et al., 2019b; Cao et al., 2020). However, in general there exists different distance metrics for minimizing domain discrepancy, for example (Gretton et al., 2006) leverage maximum mean discrepancy (MMD) as the distance measure between source and target domain distributions. Similar to MMD, CMD (central moment discrepancy) (Zellinger et al., 2017), Wasserstein distance (WD) (Shen et al., 2018), sliced Wasserstein distance (SWD) (Kolouri et al., 2019), multi-kernel MMD (Long et al., 2015), joint MMD (Long et al., 2017) are other alternative measures.

Inspired by generative adversarial networks (GAN) (Goodfellow et al., 2014), adaptation methods based on adversarial learning have also shown promising results (Ganin et al., 2017; Xie et al., 2018; Pei et al., 2018; Saito et al., 2018; Lee et al., 2019a). Adversarial learning methods propose the idea of using the domain discriminator to distinguish whether the incoming sample is from the source or target domain, while the feature generator tries to fool the discriminator by generating domain-invariant features. During the process of creating domain-invariant representations, the generator positions the target representation near the source domain decision boundaries. However, these representations are misaligned with respect to the source classes, leading to a degradation in performance (Lee et al., 2019a).

Some works rely on high-confidence pseudo-labels (Yue et al., 2022c; Deng et al., 2019) for the target domain. However, this method of generating synthetic data for the target domain imposes an additional computational overhead. Moreover, target pseudo-labeling can have adverse effects on adaptation if it generates too many incorrect labels above the confidence threshold. Some works propose minimizing the distance between tokens from the target instances and those of the source support contrastively (Yue et al., 2022c). However, in practical scenarios, the target domain can be completely asymmetrical, necessitating the alignment of the pre-trained source model with the target domain before optimizing for domain-invariant representations. In this paper, we propose an adaptation framework called **DomainInv** (illustrated in Figure 1), which can perform domain adaptation without training an answer classifier with noisy pseudo-labeled data. This eliminates the need to filter out that noise before training on the target domain, as opposed to the existing SOTA method in (Yue et al., 2022c). Our approach involves learning domain-invariant features through domain-invariant fine-tuning along with adversarial label correction. This is done to identify target instances that are far apart from the source domain and optimize them to lie near the source support, class wisely. **Main Contributions of this paper are as follows**:

- We propose the unsupervised domain adaptation framework called DomainInv for extractive QA. The framework can address the *domain shift* phenomenon without the need for explicit training of an answer classifier with pseudo-labeled data. The noise in pseudo-labeled data, which is challenging to filter out, deteriorates the performance of the answer classifier and, consequently, hinders its ability to generalize well to the target domain.

- We propose the idea of 1) Domain Invariant Fine Tuning and 2) Adversarial Label Correction together, aiming to minimize the distance between the source and target domain representations class-wise (start and end) in an iterative manner.

- We evaluated our framework on multiple QA datasets as target domains without accessing their answers during training. DomainInv outperforms the strongest baseline for QA domain adaptation, which adapts the model by explicitly training on pseudo-labeled target domain.

2 Related Work

In the past few years, there has been an increasing interest in learning generalized representations through various learning paradigms, namely, unsupervised, multi-tasking, and transfer learning (Peters et al., 2018; McCann et al., 2018; Chronopoulou et al., 2019; Phang et al., 2018; Wang et al., 2018; Xu et al., 2019). Specifically, recent studies have explored the generalization capability of reading comprehension systems (Golub et al., 2017; Fisch et al., 2019b; Talmor and Berant, 2019; Yue et al., 2021, 2022c,b). Our interest in this paper lies solely in unsupervised approaches for domain adaptation, where target domain data is unlabelled. The approaches used for unsupervised

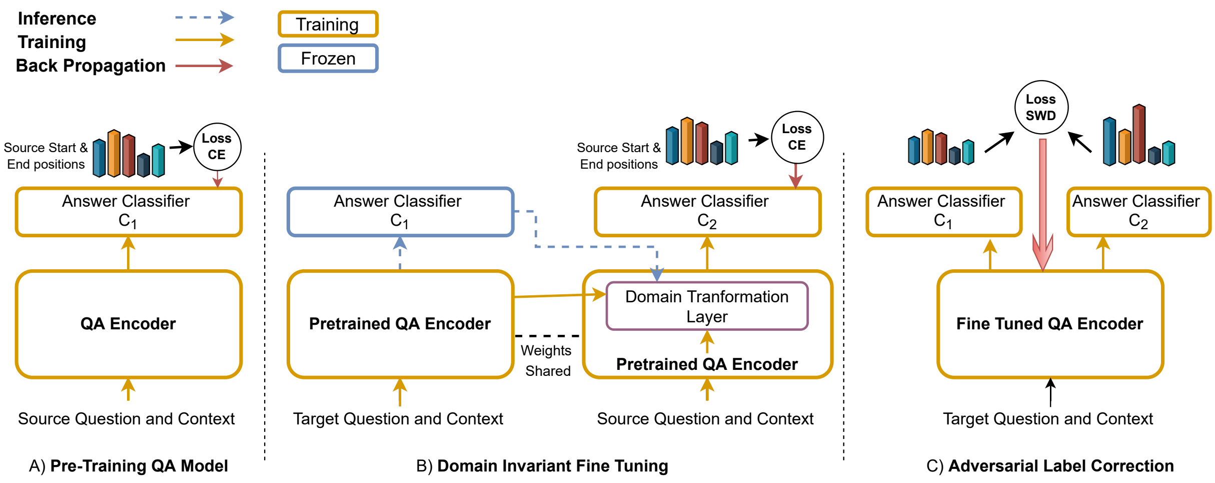

Figure 1: DomainInv: A Robust Framework for QA Domain Adaptation. It proposed to utilize domain invariant fine-tuning followed by adversarial label correction to overcome the limitations associated with domain invariant fine-tuning, demonstrating the noise free domain adaptation.

domain adaptation are broadly categorized into the following main themes: 1) Contrastive Learning, 2) Self-Supervision, and 3) Adversarial Learning.

Contrastive Learning: Contrastive learning methods (He et al., 2020; Caron et al., 2020; Chen et al., 2020; Yue et al., 2022c, 2021) aim to learn a feature encoder that generates similar features for the same input (obtained from different augmentations) and different features for any other input and its augmentations. Specifically for QA, (Sun et al., 2018; Du et al., 2017) have generated synthetic QA samples through Question Generation (QG). Leveraging these samples improves performance in out-of-domain distribution (Yue et al., 2021; Golub et al., 2017; Tang et al., 2017; Lee et al., 2020; Tang et al., 2018; Shakeri et al., 2020; Yue et al., 2022a; Zeng et al., 2022). Additionally, contrastive learning has been applied to minimize the discrepancy between the source and target domains using Maximum Mean Discrepancy (MMD) (Gretton et al., 2006). They learned to minimize the distance for averaged token features (Yue et al., 2022c) among answer and non-answer tokens in source and target domains and maximize the distance between them.

Self-Supervision: There are many works in computer vision that have explored the use of self-supervision for unsupervised domain adaptation, all aligned with the common objective of minimizing the discrepancy (distance) between domains (Kang et al., 2019; Wang et al., 2021; Thota and Leontidis, 2021). Although the objective is similar to that of contrastive learning, models learned through contrastive learning have been shown to perform better (Shen et al., 2022). Apart from the MMD (Gretton et al., 2006) criterion used in (Yue et al., 2022c), other metrics like central moment discrepancy (CMD) used in (Zellinger et al., 2017) directly match order-wise differences of central moments. Wasserstein distance, employed to measure the distance between two probability distributions, has been explored in (Shen et al., 2018; Kolouri et al., 2019). The method in (Yu et al., 2020) learns sentence representations for text matching between asymmetrical domains. In our approach, we consider the use of sliced Wasserstein distance (Kolouri et al., 2019). Instead of minimizing the distance between representations for domains, this distance is applied to minimize the distribution learned for the start and end tokens in QA domain adaptation.

Adversarial Learning: The objective of adversarial learning is also based on the idea of minimizing domain discrepancy. The main concept of domain adversarial learning is to learn domain-invariant representations through an adversarial loss between the feature generator and discriminator, similar to GANs (Goodfellow et al., 2014). Some works that use domain adversarial learning include (Ganin et al., 2017; Tzeng et al., 2017; Bousmalis et al., 2017; Yang et al., 2020; Long et al., 2018; Pei et al., 2018). Additionally, there are methods (Yue et al., 2022c) that explored the use of target data along with pseudo-labels to train the target classifier. In contrast to this, we have explored the use of adversarial loss to identify and correct mistakes in labels during target-aware source fine-tuning, aiming to learn domain-invariant representations for the target domain.

3 Setup

Problem setup for unsupervised domain adaptation(UDA) consider the labeled source domain $\mathcal{D}_s$ and unlabelled target domain $\mathcal{D}_t$. The goal is to maximize the performance on target domain by only training with labeled source domain data and unlabelled target domain data as in (Cao et al., 2020; Shakeri et al., 2020; Yue et al., 2021, 2022b,c).

Data: Specifically, for the case of QA domain adaptation we describe the labeled source domain $\mathcal{D}_s$ data as samples consisting of triplets, $\{c_s^{(i)}, q_s^{(i)}, a_s^{(i)}\} \in \mathbf{X}_s$, consisting of context $c_s^{(i)}$, question $q_s^{(i)}$ and answer $a_s^{(i)}$, where each triplet is obtained from the training data $\mathbf{X}_s$. Similarly, the unlabelled target domain $\mathcal{D}_t$ data consists of samples with pair $\{c_t^{(i)}, q_t^{(i)}\} \in \mathbf{X}_t$ consisting of only context $c_t^{(i)}$ and question $q_t^{(i)}$, obtained from unlabelled training data $\mathbf{X}_t$. Here, in our case of QA domain adaptation the answer is the start and end position in the context since we are working with extractive QA systems.

Model: We approach the problem of QA domain adaptation as training the model function f which predicts an answer $a_t^{(i)}$ given the context $c_t^{(i)}$ and question $q_t^{(i)}$ from $\mathbf{X}_t$, denoted as $a_t^{(i)} = f(c_t^{(i)}, q_t^{(i)})$. This requires to optimize the function f for maximum performance on target domain $\mathcal{D}_t$, given $\mathcal{D}_s$. Mathematically, this is denoted as:

$$\underset{f}{min}\, \mathcal{L}(f, \mathbf{X}_t; \mathbf{X}_s) \tag{1}$$

where $\mathcal{L}$ is the loss function. We adopt the two fold training scheme to maximize performance on target domain namely, Domain Invariant Fine Tuning and Adversarial Label Correction which will be discussed in the following sections.

4 DomainInv Framework

4.1 Overview

The proposed DomainInv framework consists of two main components: 1) **Domain Invariant Fine Tuning** and 2) **Adversarial Label Correction** for domain adaptation, as shown in Figure 1. We start with a pre-trained QA model f, fine-tuned on the source domain $\mathcal{D}_s$ as in (Cao et al., 2020), with an additional batch norm layer. The answer classifier $\mathcal{C}_1$ predicts the start and end indices in the context. During domain invariant fine-tuning, we incorporate the use of the target domain $\mathcal{D}_t$ to augment the style of pseudo-answer and non-answer tokens to the source domain. This results in another answer classifier $\mathcal{C}_2$, which possesses target domain style information while still being trained on the source domain. With the answer classifier $\mathcal{C}_2$, there are instances in the target domain $\mathcal{D}_t$ for which the answer differs from the one obtained using $\mathcal{C}_1$. We identify these instances as those which are far apart from the source domain, and the QA model is least confident about them. During adversarial correction, we minimize the distribution between these two classifiers and update the BERT encoder to generate features for the target domain closer to the source domain. This ensures that the classifier $\mathcal{C}_2$ predicts the answer as if it were operating on the source domain, aligning the features for the target domain with those of the source domain.

4.2 Domain Invariant Fine Tuning

In this section, we will explain in detail the process we have followed for domain invariant fine-tuning. Let the trained QA model on source domain is denoted as f, it is a BERT model with $\mathbf{L}$ layers of transformers (Vaswani et al., 2017). Specifically, let $\mathcal{C}_1$ be an answer classifier, and θ_g be the encoder parameters for this source-domain QA model.

During domain invariant fine-tuning (shown in Figure 2), we propose to feed the style information of the target domain $\mathcal{D}_t$ to the source domain QA model at each layer $l \in \mathbf{L}$, as illustrated in Figure 2. We keep the weights shared between the two encoders to allow the target domain information to be updated in the BERT encoder with the supervision of the source domain. Let $\phi(x, x')$ be the learnable domain shift vector between the source instance x and the target domain instance x', and $\mathcal{M}(x, \phi(x, x'))$ be a learnable domain transformation layer, which is introduced at the top of each transformer layer $l \in \mathbf{L}$ of model f. Cumulatively, it transforms the parameters of the source domain classifier $\mathcal{C}_1$ to the target-aware classifier $\mathcal{C}_2$ and updates encoder parameters θ_g with the style of the target domain.

Domain Transformation Layer: The domain transformation layer $\mathcal{M}$ is expected to fuse the domain shift vector with the hidden states (which were fine-tuned for the source domain) at each layer of the transformer. The domain shift vector ϕ should solely capture the information that is different from the source domain. This categorizes the vector containing any extra information in the target domain compared to the source domain, ir-

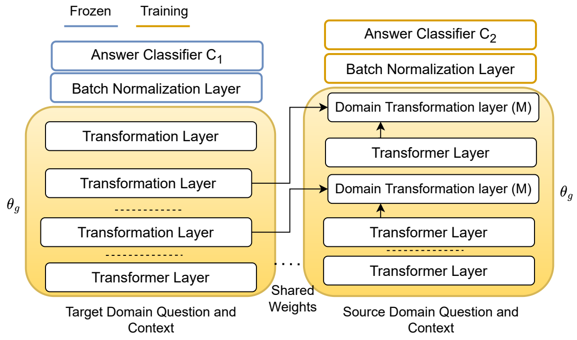

Figure 2: Domain Invariant Fine Tuning with Domain Transformation Layer

respective of its position in the context. This can be achieved by taking the difference between the average pooled vector of hidden states at each layer obtained for the source and target domains. Let $\mathbf{H}_s^{(l)}$ and $\mathbf{H}_t^{(l)}$ be the hidden states obtained at layer l for the source domain and target domain, respectively. Then, the domain shift vector at layer l between two instances is given as:

$$\phi^{(l)}(H_t^{(l)}, H_s^{(l)}) = avg(WH_t^{(l)}) - avg(WH_s^{(l)}) \tag{2}$$

where, $W \in \mathbf{R}^{k \times d}$ are the linear transform parameters shared across layers. Then, the domain transformation layer $\mathcal{M}$ is given as:

$$\begin{aligned} \mathcal{M}(H_s^{(l)}, \phi^{(l)}(H_t^{(l)}, H_s^{(l)})) = H_s^{(l)} + \\ W^T \phi^{(l)}(H_t^{(l)}, H_s^{(l)}) \end{aligned} \tag{3}$$

The expression $W^T \phi^{(l)}(H_t^{(l)}, H_s^{(l)}) \in \mathbf{R}^d$ is added to all hidden states at layer l corresponding to the source domain. This design of the domain shift vector follows the identity property, i.e., $\phi(x, x) = 0$. This allows us to plug in the domain transformation layer only at the time of training, while at the time of inference for the target domain, it is, $\phi^{(l)}(H_t^{(l)}, H_t^{(l)})) = 0$. However, it is required to carefully choose the value of k, which is a hyperparameter, because it assumes the domain shift information lies in the k-dimensional subspace where the difference between two domains can be minimized to make them appear similar.

QA Domain Transformation: Specifically, for the case of QA domain adaptation, we can't apply the domain shift across all the tokens uniformly, as there are underlying differences between the context, question, and answer tokens. Hence, it would be wise to calculate ϕ_A, ϕ_C, and ϕ_Q for Answer, Context, and Questions, respectively. Since we only have context and questions in the target domain, the answer is the pseudo-answer obtained from the classifier $\mathcal{C}_1$, where the weights of $\mathcal{C}_1$ are frozen, and the BERT encoder parameters θ_g are shared during fine-tuning. These will be updated jointly along with classifier $\mathcal{C}_2$ (initialized with the fine-tuned classifier $\mathcal{C}_1$), as shown in Figure 2. At the end of domain transformation fine-tuning, we obtain another target-aware classifier $\mathcal{C}_2$ and updated encoder parameters θ_g which are the fine-tuned parameters on the target-aware source domain with supervised cross-entropy loss $\mathcal{L}_{ce}$. Note that here pseudo labels on target domain is not directly involved in training answer classifier $\mathcal{C}_2$.

$$\underset{f}{min}\ \mathcal{L}_{ce}(f, \mathbf{X}_s || \mathbf{X}_t) \tag{4}$$

where f consists of parameters θ_g and parameters in classifier $\mathcal{C}_2$ and $\mathcal{M}$. During training, we randomly select samples from the target domain, ensuring the batch size matches that of the source domain, and an additional constraint to include parallel instances with the same question types (denoted as $\mathbf{X}_s || \mathbf{X}_t$). We employ the dependency parser, semantic role labeling, and named entity recognition (NER) to detect the question types, as done in (Keklik, 2018).

4.3 Adversarial Label Correction

We have introduced adversarial label correction based on the fact that during domain invariant fine-tuning, we relied on the pseudo-labels obtained from classifier $\mathcal{C}_1$. However, these labels are noisy and prone to error accumulation, potentially leading to erroneous alignment of the source and target domain. The domain shift information in the k-dimensional subspace may be captured incorrectly, resulting in similar performance degradation as observed in the approaches (Yue et al., 2022c, 2021) where pseudo-labels are used to train the answer classifier. However, these methods (Yue et al., 2022c, 2021) overlook the fact that domain adaptation can result in target features lie far apart from the source domain for similar semantics that are ambiguous and hence error-prone. To mitigate this, we propose the idea of reducing the inconsistency between domains by generating target features near the support of source classes. In our case, this refers to the starting and ending indices for the answer classifier for different question types.

In Equation 4, fine-tuning occurs with the target-aware source domain, resulting in the answer classifier $\mathcal{C}_2$. This learns the distribution of start and end classes given the style of the target domain $\mathcal{D}_t$, but these can be error-prone, mainly due to two reasons: 1) The target pseudo labels obtained from $\mathcal{C}_1$ are erroneous, and 2) The target domain information makes the classifier $\mathcal{C}_2$ difficult to learn the correct distribution of start and end classes on the source domain. Collectively, this happens when the target instances are far apart from the source domain and require explicit optimization for such cases. Hence, we make use of the adversarial loss to first identify the samples of the target domain that are far apart from the support of the source domain. We then update the parameters of answer classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$ with θ_g (obtained after domain invariant fine-tuning) frozen to maximize the discrepancy due to such instances. Specifically, we update $\mathcal{C}_1$ first keeping $\mathcal{C}_2$ fixed and then $\mathcal{C}_2$ with updated $\mathcal{C}_1$. Subsequently, we minimize the parameters of θ_g to generate target domain features near the source domain for start and end classes. Mathematically, this has been written as the minmax game of learning domain invariant representations:

$$\min_{\theta_g} \max_{\mathcal{C}_1,\mathcal{C}_2} \mathcal{L}_{lc}(X_t) \tag{5}$$

where $\mathcal{L}_{lc}$ is the label correction loss optimized for the target domain for maximum performance. The discrepancy maximizing term has been mathematically formulated as the Sliced Wasserstein Distance(SWD) between the representation as in (Kolouri et al., 2019) learnt for start and end classes by classifier $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively. Let $\mathcal{G}$ denote the BERT encoder with parameters θ_g, the loss is written as:

$$\begin{aligned} \min_{\mathcal{C}_1,\mathcal{C}_2} \mathcal{L}_{ce}(f, \mathbf{X}_s || \mathbf{X}_t) - \\ \sum_{k\in\{s,e\}} \mathcal{L}_{swd}(\mathcal{C}_1^k(G(X_t)), \mathcal{C}_2^k(G(X_t))) \end{aligned} \tag{6}$$

where $\mathcal{C}_1^k(.)$, $\mathcal{C}_2^k(.)$ for all $k \in \{s, e\}$ denotes the probability distribution obtained from classifier $\mathcal{C}_1$ and $\mathcal{C}_2$ for starting and ending indices s and e. This loss updates both classifiers on target aware source domain only for those instances where parallel target domain instances are inconsistent. Since both $\mathcal{C}_1$ and $\mathcal{C}_2$ trained to predict the start and end indices for the source domain, one with source domain features only and another on source augmented with target domain features. Hence, the classifier $\mathcal{C}_2$ can predict the different start and end indices for those instances where $\mathcal{C}_1$ is incorrect for target domain. Hence we update the classifier $\mathcal{C}_1$ first and then update $\mathcal{C}_2$ using updated $\mathcal{C}_1$. Adjusted answer classifiers results in updated domain discrepancy for the non-confident target predictions and hence we need to update the parameters θ_g so as to generate the target domain features near to source support class wisely. This has been achieved by minimizing the loss function:

$$\min_{\mathcal{G}} \sum_{k\in\{s,e\}} \mathcal{L}_{swd}(\mathcal{C}_1^k(G(X_t)), \mathcal{C}_2^k(G(X_t))) \tag{7}$$

Finally, at the end we use the Encoder $\mathcal{G}$ and answer classifier $\mathcal{C}_1$ as the domain adapted QA model. End-to-end training of DomainInv Framework is shown in Algorithm 1 (in Appendix B).

5 Experiments

Datasets: We consider the source domain $\mathcal{D}_s$ as **SQuAD v1.1** (Rajpurkar et al., 2016) following (Yue et al., 2022c, 2021; Shakeri et al., 2020; Cao et al., 2020; Lee et al., 2020). **SQuAD v1.1** is a well known annotated QA dataset where paragraphs (context) are from Wikipedia articles. Target Domains $\mathcal{D}_t$ are considered from MRQA Split 1 (Fisch et al., 2019b), namely, **NaturalQuestions** (Kwiatkowski et al., 2019), **HotpotQA** (Yang et al., 2018), **SearchQA** (Dunn et al., 2017), **TriviaQA** (Joshi et al., 2017), **NewsQA** (Trischler et al., 2016). The dataset details we have considered for target domain is given in Appendix A.

Baselines: We trained our DomainInv Framework on top of the fine-tuned QA model on source domain, adopted from BERT with an additional batch normalization layer, as in (Cao et al., 2020). This fine-tuned BERT model, trained on the source domain, acts as the naive baseline. However, to further assess the robustness of our framework in QA domain adaptation, we adopted the following state-of-the-art (SOTA) baselines: 1) **QADA** (Yue et al., 2022c): QA Domain Adaptation (QADA) leverages hidden space augmentation for enriching the training dataset and used attention-based contrastive learning for domain adaptation. 2) **CAQA** (Yue et al., 2021): Contrastive Domain Adaption for Question Answering (CAQA) combines question generation and contrastive domain adaptation to learn domain-invariant features, so that it can capture both domains and thus transfer knowledge to the target distribution 3) **DAT** (Tzeng et al., 2017; Lee et al., 2019b): Domain Adversarial Training (DAT) follows the known adversarial training and uses the [CLS] token in BERT as a discriminator to learn the generalized features from both source and target domains after training with labeled source domain 4) **CAQA*** (Yue et al., 2021, 2022c): Instead

Model	**HotpotQA**	**NaturalQ.**	**NewsQA**	**SearchQA**	**TriviaQA**
	EM / F1	EM / F1	EM / F1	EM / F1	EM / F1
(1) **Zero Shot Target Performance**					
BERT	43.34/60.42	39.06/53.7	39.17/56.14	16.19/25.03	49.70/59.09
(2) **QA Domain Adaptation Target Performance**					
DAT (Lee et al., 2019b)	44.25/61.10	44.94/58.91	38.73/54.24	22.31/31.64	49.94/59.82
CASe (Cao et al., 2020)	47.16/63.88	46.53/60.19	43.43/59.67	26.07/35.16	54.74/63.61
CAQA (Yue et al., 2021)	46.37/61.57	48.55/62.60	40.55/55.90	36.05/42.94	55.17/63.23
CAQA* (Yue et al., 2021, 2022c)	48.52/64.76	47.37/60.52	44.26/60.83	32.05/41.07	54.30/62.98
QADA (Yue et al., 2022c)	50.80/65.75	52.13/65.00	45.64/61.84	40.47/48.76	56.92/65.86
DomainInv(Ours)	**52.92/66.71**	**54.97/68.80**	**45.96/61.88**	**40.92/49.88**	**57.78/66.64**
(3) **Supervised Training Target Performance**					
BERT (10K Samples)	49.57/66.65	54.81/67.98	45.92/61.85	60.21/66.96	53.87/60.42
BERT (All Samples)	57.96/74.76	67.08/79.02	52.14/67.46	71.54/77.77	64.51/70.27

Table 1: QA Adaptation Performance on Target Domains

of question generation in CAQA, this baseline uses the same process of generating pseudo labels and self-supervised adaptation as in QADA. 5) **CASe** (Cao et al., 2020): Conditional Adversarial Self-Training (CASe) is an unsupervised domain adaptation method that iteratively performs self-training on high-confidence pseudo-labels and incorporates conditional adversarial learning.

Training, Evaluation and Implementation: Following (Cao et al., 2020; Yue et al., 2022c), we trained the naive baseline of the BERT model with an additional batch norm layer after the encoder (in PyTorch by Hugging Face, using the base-uncased pretrained model with 12 layers and 768-dim hidden state). Specifically, we used a learning rate of $3 \cdot 10^{-5}$ and trained for 2 epochs with a batch size of 12, optimized using the AdamW optimizer with 10% linear warm-up on the source domain $\mathcal{D}_s$. Following (Lee et al., 2020; Shakeri et al., 2020; Yue et al., 2021), we evaluated exact matches (EM) and F1 score on the dev sets. The rest of the baselines are implemented according to the methods described in their corresponding papers.

For the DomainInv Framework, we ran the domain invariant fine-tuning followed by adversarial label correction and repeated this for 10 epochs with the AdamW optimizer, learning rate of 10^{-5}, with 10% linear warm-up. During fine-tuning, we generated the labels for the target domain using the classifier $\mathcal{C}_1$, which is frozen during fine-tuning, and sampled parallel samples of the target domain with question types of source domain samples in a given batch size of 12. During fine-tuning, there is only one hyperparameter named k for the domain transformation layer, which has been searched for the best value in [64, 128, 256, 512, 768]. Eventually, the best value of 256 works for us in almost all cases and is the one with the maximum performance on the source domain $\mathcal{D}_s$ during fine-tuning. After domain invariant fine-tuning, the obtained classifiers $\mathcal{C}_1$, $\mathcal{C}_2$, and encoder $\mathcal{G}$ are trained with adversarial label correction. We stopped the training in between if there is no decrease in the loss described in equation 7 for the continuous 3 epochs in a row.

5.1 Experimental Results

Table 1 presents the results for QA domain adaptation performance on various target domains, as described in Section 5. We grouped our results and analysis into three main categories, namely: 1) **Zero short Target Performance**: This reports the results on the target domain with the BERT fine-tuned model without any domain adaptation on the target domain, serving as a *lower bound* for domain adaptation approaches. 2) **QA Domain Adaptation Target Performance**: This reports the results due to various domain adaptation methods, including DomainInv.. 3) **Supervised Training Target Performance**: This reports the results following the supervised training of BERT on the target domain using randomly selected 10K samples, along with all source domain samples, to establish the *upper bound* performance for QA domain adaptation approaches. QA domain adaptation performance (shown in Table 1) using the DomainInv

Model	HotpotQA EM / F1	NaturalQ. EM / F1	NewsQA EM / F1	SearchQA EM / F1	TriviaQA EM / F1
DomainInv(Ours)	**52.92/66.71**	**54.97/68.80**	**45.96/61.88**	**40.92/49.88**	**57.78/66.64**
w/o Adversarial Label Correction	51.60/64.07	53.91/65.21	45.88/61.86	39.81/46.98	56.98/65.32

Table 2: Ablation Study: QA Adaptation Performance on Target Domains by different components of DomainInv

Framework outperforms all the domain adaptation baselines across all target domains and is well beyond the naive baseline. In fact, almost all the domain adaptation baselines outperform the naive baseline by a significant margin on all target domains. However, BERT performs poorly (compared to domain adaptation baselines) on some target domains, namely, Natural Questions and SearchQA, due to two main reasons: 1) BERT does not understand the style of Natural Questions; even if the Wikipedia article is the same, the real user questions style is different from the one asked in SQuAD v1.1. 2) BERT does not understand the long form of contexts, which is usual in SearchQA, and it learns to focus on the nearby tokens similar to those in SQuAD v1.1.

However, the actual or more effective answer is also present in the long context. Compared to the worst QA adaptation baseline, DomainInv outperforms BERT on these two domains on average by 2.64% in EM and the other domains by 1.2% in EM. This is the main reason we adopted the domain style-based transformation layer and the corresponding fine-tuning, which can make BERT understand different contexts and questions. The DomainInv Framework outperforms all baselines; on average, it outperforms the best baseline by 2.59% and 2.17% in EM and F1, respectively. Moreover, our framework outperforms supervised training with 10K target data, additionally on Natural Questions and NewsQA, as compared to QADA (best baseline), which outperforms the supervised baseline of 10K target data only on HotpotQA and TriviaQA. We reported all the results after averaging the inference results from 10 rounds.

5.2 Ablation Studies

In Table 1, we compared the DomainInv framework against the strongest baseline named QADA. However, this comparison does not detail the importance of each component of DomainInv, namely, domain invariant fine-tuning and adversarial label correction. The absence of these components can cause a maximum drop of 5.17% compared to the best baseline QADA, highlighting the advantage of using DomainInv over other domain adaptation approaches. However, the contribution of each component towards performance gain is still unknown. Hence, we studied the performance (mentioned in Table 2) of DomainInv after removing the adversarial label correction component only, since removing the domain-invariant fine tuning as well results in the source-domain trained BERT model, for which the zero-shot performance is already mentioned in Table 1. The performance drop in Table 2 clearly depicts the advantage of adversarial label correction. For target domains, namely HotpotQA, Natural Questions, and NewsQA, the performance of DomainInv w/o adversarial label correction in terms of EM is still higher than that of QADA. However, in SearchQA, it goes below the performance of QADA. This indicates the important insight into the functioning of adversarial label correction. For long contexts like in SearchQA, where matching the answer exactly requires significant correction, and in Natural Questions, where the question style has changed but the context is still the same (i.e., Wikipedia articles), requiring only minor correction in labels. This proves the effectiveness of the label correction component in the DomainInv framework.

6 Conclusion

In this paper, we proposed a novel QA domain adaptation framework called DomainInv. It is an unsupervised algorithm that does not require the use of labeled target domain data, nor does it depend on synthetic data or pseudo-labeled target domain. DomainInv comprises two key components: 1) Domain Invariant Fine Tuning, which fine-tunes the QA model using the target style on the source domain, and 2) Adversarial Label Correction, which identifies target distributions that are far apart from the source domain and optimizes the feature generator to bring them closer to the source support class wisely. Evaluation of DomainInv showed that it outperforms all baselines, achieving superior performance and establishing a new benchmark.

Limitations

In this section, we highlights certain limitations of DomainInv that were not covered in the paper. In the domain invariant fine-tuning, we introduced a new layer called the domain adaptation layer, which computes the difference between the average pooled representations of the source and target domains. However, this design assumes equal importance for all tokens in both domains at each layer, overlooking the influence of the self-attention mechanism on token distribution. To rectify this, future work should explore incorporating attention-weighted representations before calculating the difference. Additionally, in the adversarial label correction, we proposed adjusting the feature encoder solely based on the target domain, neglecting the potential benefits of jointly aligning both source and target domains. Further research could explore these aspects for improvement.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79:151–175.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 95–104, Los Alamitos, CA, USA. IEEE Computer Society.

Yu Cao, Meng Fang, Baosheng Yu, and Joey Tianyi Zhou. 2020. Unsupervised domain adaptation on reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7480–7487.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. 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 2089–2095, Minneapolis, Minnesota. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Zhijie Deng, Yucen Luo, and Jun Zhu. 2019. Cluster alignment with a teacher for unsupervised domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9943–9952.

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.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *CoRR*, abs/1704.05179.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019a. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019b. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *EMNLP 2019 MRQA Workshop*, page 1.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2017. *Domain-Adversarial Training of Neural Networks*, pages 189–209. Springer International Publishing, Cham.

David Golub, Po-Sen Huang, Xiaodong He, and Li Deng. 2017. Two-stage synthesis networks for transfer learning in machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 835–844, Copenhagen, Denmark. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4888–4897, Los Alamitos, CA, USA. IEEE Computer Society.

Onur Keklik. 2018. *Automatic question generation using natural language processing techniques*. Ph.D. thesis, Izmir Institute of Technology (Turkey).

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. 2019. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32.

Bernhard Kratzwald, Anna Eigenmann, and Stefan Feuerriegel. 2019. RankQA: Neural question answering with answer re-ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6076–6085, Florence, Italy. Association for Computational Linguistics.

Bernhard Kratzwald, Stefan Feuerriegel, and Huan Sun. 2020. Learning a Cost-Effective Annotation Policy for Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3051–3062, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. 2019a. Sliced wasserstein discrepancy for unsupervised domain adaptation.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online. Association for Computational Linguistics.

Seanie Lee, Donggyu Kim, and Jangwon Park. 2019b. Domain-agnostic question-answering with adversarial training. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 196–202, Hong Kong, China. Association for Computational Linguistics.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 97–105. JMLR.org.

Mingsheng Long, ZHANGJIE CAO, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. 2017. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2208–2217. JMLR.org.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3723–3732, Los Alamitos, CA, USA. IEEE Computer Society.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. Association for Computational Linguistics.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z. Haochen, Tengyu Ma, and Percy Liang. 2022. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19847–19878. PMLR.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3930–3939, Brussels, Belgium. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2019. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921.

Duyu Tang, Nan Duan, Tao Qin, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *CoRR*, abs/1706.02027.

Duyu Tang, Nan Duan, Zhao Yan, Zhirui Zhang, Yibo Sun, Shujie Liu, Yuanhua Lv, and Ming Zhou. 2018. Learning to collaborate for question answering and asking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1564–1574, New Orleans, Louisiana. Association for Computational Linguistics.

Mamatha Thota and Georgios Leontidis. 2021. Contrastive domain adaptation. *CoRR*, abs/2103.15566.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset.

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. 2017. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, Los Alamitos, CA, USA. IEEE Computer Society.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. 2021. Cross-domain contrastive learning for unsupervised domain adaptation. *CoRR*, abs/2106.05528.

Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. 2018. Learning semantic representations for unsupervised domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5423–5432. PMLR.

Yichong Xu, Xiaodong Liu, Yelong Shen, Jingjing Liu, and Jianfeng Gao. 2019. Multi-task learning with sample re-weighting for machine reading comprehension. 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 2644–2655, Minneapolis, Minnesota. Association for Computational Linguistics.

Guanglei Yang, Haifeng Xia, Mingli Ding, and Zhengming Ding. 2020. Bi-directional generation for unsupervised domain adaptation. *CoRR*, abs/2002.04869.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Weijie Yu, Chen Xu, Jun Xu, Liang Pang, Xiaopeng Gao, Xiaozhao Wang, and Ji-Rong Wen. 2020. Wasserstein distance regularized sequence representation for text matching in asymmetrical domains. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2985–2994, Online. Association for Computational Linguistics.

Xiang Yue, Ziyu Yao, and Huan Sun. 2022a. Synthetic question value estimation for domain adaptation of question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1340–1351, Dublin, Ireland. Association for Computational Linguistics.

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. 2021. Contrastive domain adaptation for question answering using limited text corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9575–9593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022b. Domain adaptation for question answering via question classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1776–1790, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Bernhard Kratzwald, Stefan Feuerriegel, and Dong Wang. 2022c. Qa domain adaptation using hidden space augmentation and self-supervised contrastive adaptation. *arXiv preprint arXiv:2210.10861*.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (CMD) for domain-invariant representation learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Huimin Zeng, Zhenrui Yue, Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. 2022. Unsupervised domain adaptation for covid-19 information service with contrastive adversarial domain mixup.

Huimin Zeng, Zhenrui Yue, Yang Zhang, Ziyi Kou, Lanyu Shang, and Dong Wang. On attacking out-domain uncertainty estimation in deep neural networks.

A Benchmark Datasets

The dataset details we have considered for target domain is given as follows:

- **NaturalQuestions**: A real world QA dataset with questions that are actual user questions, and contexts as Wikipedia articles, which may or may not contain the answers (Kwiatkowski et al., 2019)

- **HotpotQA**: A reasoning based QA dataset with multi hop questions and supporting facts (Yang et al., 2018)

- **SearchQA**: QA dataset where context built by crawling through Google Search. However, this is based on existing QA pairs for which the context is extended. More details in (Dunn et al., 2017)

- **TriviaQA**: A reasoning based QA dataset containing evidences for questions asked (Joshi et al., 2017)

- **NewsQA**: QA dataset with news as contexts and questions with answers not from simple matching and entailment. (Trischler et al., 2016)

B Algorithm

We presented the end-to-end DomainInv algorithm as follows:

Algorithm 1 `DomainInv` Training for UDA

```
Require: Labeled Source $\{\mathcal{X}_s, \mathcal{Y}_s\}$; unlabelled Target $\{\mathcal{X}_t\}$,
  hyperparameter $k$, fine tuned QA model with encoder $\mathcal{G}$
  and Classifier $\mathcal{C}_1$ and classifier $\mathcal{C}_2$ initialized with $\mathcal{C}_1$.
  Step 1: Update $\mathcal{G}$, $\mathcal{C}_2$ on Source Domain (with target style augmentation) using Domain Invariant Fine-Tuning as in Equation 4
  while $\mathcal{G}, \mathcal{C}_1, \mathcal{C}_2$ still converging do
      Step 2: Update $\mathcal{C}_1, \mathcal{C}_2$ on target aware source set to maximize the sliced Wasserstein distance (SWD) on target instances as in Equation 6
      Step 3: Update $\mathcal{G}$ to minimize the SWD as calculated earlier according to Equation 7
  end while
```