Source-Free Unsupervised Domain Adaptation for Question Answering via Prompt-Assisted Self-learning

Maxwell J. Yin and **Boyu Wang** and **Charles Ling***
Western University
jyin97@uwo.ca, bwang@csd.uwo.ca, charles.ling@uwo.ca

Abstract

This work addresses source-free domain adaptation (SFDA) for Question Answering (QA), wherein a model trained on a source domain is adapted to unlabeled target domains without additional source data. Existing SFDA methods only focus on the adaptation phase, overlooking the impact of source domain training on model generalizability. In this paper, we argue that source model training itself is also critical for improving the adaptation performance and stability. To this end, we investigate the role of prompt learning as an effective method to internalize domain-agnostic QA knowledge, which can be integrated into source training. After source training, an interactive self-learning strategy is proposed to further fine tune both model and prompt in the model adaptation phase. This leads to the Prompt-Assisted Self-Adaptive Learning (PASAL), an innovative SFDA approach for QA. Empirical evaluation on four benchmark datasets shows that PASAL surpasses existing methods in managing domain gaps and demonstrates greater stability across various target domains, validating the significance of source domain training for effective domain adaptation.

1 Introduction

Question-answering (QA) systems have significantly advanced with the advent of pretrained language models (PLMs). Despite this, research shows that PLMs often grapple with domain shifts, when training and evaluation datasets have different distributions (Yue et al., 2021). Additionally, the success of PLMs is heavily reliant on human-annotated data within the relevant domain for specialized tasks (Devlin et al., 2018; Liu et al., 2019). In many real-world applications, however, the expense of obtaining annotated data is substantial, making the use of unlabeled data more feasible. Consequently, there is a necessity to fine-tune models—originally trained on general datasets—to specific domains, despite the substantial domain gaps.

In response to these challenges, unsupervised domain adaptation (UDA) for QA aims to leverage knowledge from a well-labeled source domain to enhance performance in unlabeled target domains. While existing approaches (Wang et al., 2019; Cao et al., 2020; Vaswani et al., 2017; Nishida et al., 2019; Yue et al., 2021) show promise, they often assume ongoing access to source domain data during target domain adaptation. This reliance presents a challenge in contexts involving sensitive data. For example, clinical data is often subject to stringent privacy regulations due to patient confidentiality, thus only model trained on the source domain is available for public use after initial training (Laparra et al., 2021; Su et al., 2022). Our study contributes to this domain by applying source-free unsupervised domain adaptation (SFDA) to the realm of QA, eliminating the dependency on source domain data after the initial phase of model pretraining. This approach allows the adaptation process to adhere to strict privacy constraints while still leveraging the extensive knowledge gained from source domain pretraining, ensuring a balance between data privacy and the effectiveness of domain adaptation in sensitive contexts.

While SFDA has been explored in the field of computer vision (Liang et al., 2020; Li et al., 2020; Kundu et al., 2020; Yang et al., 2023b), these studies predominantly concentrate on classification tasks, employing methods like clustering which are not directly transferable to QA tasks. Furthermore, such research tends to focus solely on the target domain adaptation phase, overlooking the potential benefits of enhancing the source training phase. In contrast, we argue that improving the initial training phase can significantly boost the generalization capabilities of the model across different domains, thus facilitating more effective domain adaptation.

To fulfill our objectives, it is essential for the model to internalize fundamental QA knowledge

Findings of the Association for Computational Linguistics: NAACL 2024, pages 700–713

that is not bound by the specificities of individual domains, maximizing the use of source data. Our research indicates that prompt learning (Liu et al., 2023; Lester et al., 2021) is particularly effective in this regard. We observe that QA samples from different domains can vary significantly in vocabulary and sentence structure. Consequently, fine-tuning an entire model from one domain to another often necessitates substantial adjustments in the weights of the model. However, considering the nature of prompts, which essentially frame the task description, their scope is relatively fixed once the task is defined. Thus, the prompt for a QA task remains consistent across various training domains. For instance, a prompt like "please answer the question given the context" is applicable irrespective of the domain. This uniformity in prompts, despite divergent training domains, underscores their potential in streamlining domain adaptation and minimizing the need for extensive model retraining.

In our work, we have innovated Prompt-Assisted Self-Adaptive Learning (PASAL), a methodology that seamlessly incorporates prompt learning into the domain adaptation process of QA models. Initially, during the pretraining phase on the source domain, we employ a PLM and augment it with an additional prompt specifically designed to assimilate key, domain-agnostic QA concepts. Additionally, we train an auxiliary Question Generation (QG) model for creating questions from given contexts.

The adaptation to the target domain employs a distinctive self-learning strategy. Commencing with unlabeled data from the target domain, our QG model initially generates pseudo-questions. Subsequently, the QA model, leveraging the initially trained prompt, produces corresponding pseudo-answers. These synthesized question-answer pairs initiate our iterative self-learning cycle: We begin by fine-tuning the prompt, keeping the QA model constant, to closely align with the subtleties of the target domain. Following the optimization of the prompt, it then directs the focused fine-tuning of the QA model, which is conducted with the prompt remaining static. After this, the QG model also undergoes fine-tuning to refine its question-generation capabilities within the target domain. This cycle perpetuates in an alternating fashion—fine-tuning the prompt, then the QA model, and finally the QG model—in a consistent rhythm. Each phase utilizes the pseudo-labeled samples, progressively honing the prompt, QA model, and QG model to foster a more robust and effective domain adaptation.

The core contributions of our research are:

1. We pioneer the application of source-free unsupervised domain adaptation (SFDA) in the realm of QA, emphasizing data privacy and addressing the challenges of sensitive data usage.

2. We introduce the Prompt-Assisted Self-Adaptive Learning (PASAL) framework, an innovative integration of prompt learning with SFDA for QA. This framework leverages prompts to enhance the learning of domain-agnostic knowledge, effectively managing domain shifts.

3. We develop a comprehensive self-learning strategy for the iterative fine-tuning of prompts, QA, and QG models within the target domain. This strategy significantly enhances the adaptability of the model to new domains without the need for source domain data.

2 Related Work

2.1 Unsupervised Domain Adaptation for Question Answering

Historically, UDA for QA has employed adversarial training, multitask learning, and contrastive learning, as seen in the seminal work of Wang et al. (2019). These methods strive to align domain features and answer spans to facilitate the transfer of knowledge. Successive studies by Nishida et al. (2019) and Cao et al. (2020) have built upon this foundation, integrating multitask learning and self-training techniques. Nonetheless, a common limitation of these approaches is the necessity for simultaneous access to both source and target data.

In contrast, source-free unsupervised domain adaptation (SFDA), introduced by Liang et al. (2020), removes the dependency on source domain data during the adaptation phase, addressing data privacy and accessibility issues. Existing SFDA research, as evidenced by Li et al. (2020); Huang et al. (2021); Zeng et al. (2022); Yi et al. (2023); Wang et al. (2023), typically focuses on the adaptation of models post-training in the source domain, with a particular emphasis on classification tasks. Attempts to apply SFDA in NLP have been made, yet discussions on extending it to QA are sparse, as noted by Zhang et al. (2021) and Su et al. (2022).

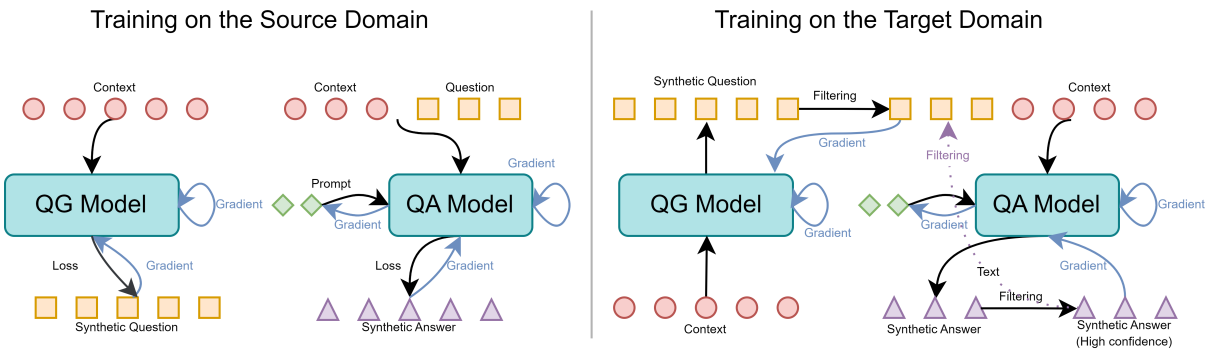

Figure 1: Schematic Overview of the PASAL Framework for Source-Free Unsupervised Domain Adaptation in QA. The left panel illustrates the initial training phase on the source domain. The right panel shows the self-adaptive learning phase on the target domain, highlighting the cyclical fine-tuning process. The QG Model generates pseudo-questions, which are filtered and used by the QA Model to obtain synthetic answers, with the prompt guiding the adaptation. This iterative approach ensures progressive enhancement and adaptability of the model to the target domain.

The distinctive challenges of QA, which resist the direct application of classification-based clustering, underscore the need for further exploration. Our study enhances this dialogue by not only refining the adaptability of the model in the target domain but also by innovating how the model is initially trained in the source domain to bolster the effectiveness of SFDA for QA.

2.2 Prompt Learning

The field of prompt learning has seen significant strides, offering new approaches for enhancing NLU tasks. "Prefix tuning", introduced by Li and Liang (2021), enriches model understanding by appending prefixes to input sequences. The "WARP" method by Hambardzumyan et al. (2021) and "P-tuning" by Liu et al. (2023) modify language model outputs and input sequences, respectively, demonstrating a robustness comparable to traditional fine-tuning while utilizing fewer task-specific parameters. Additionally, "soft prompts" by Qin and Eisner (2021) represent an adaptive strategy that can dynamically tailor prompt tokens for pre-trained models. Despite these breakthroughs, the integration of prompt learning into UDA for QA remains under-explored, signaling a significant opportunity for future research to enhance domain adaptation in QA.

3 Problem Definition

The problem of UDA for QA is defined as follows. Given a context $c = (c_1, c_2, \ldots, c_{L_1})$ with L_1 tokens, and a query $q = (q_1, q_2, \ldots, q_{L_2})$ with L_2 tokens, the system must identify an answer $a = (c_{a_s}, c_{a_s+1}, \ldots, c_{a_e})$ within the context c. Here, a_s and a_e represent the starting and ending indices of the answer within c.

In the scenario of SFDA for QA, the source domain D_S provides labeled data accessible only during the initial training phase of the model. Post this phase, the data from D_S becomes unavailable. Conversely, the target domain D_T offers unlabeled data without such restrictions. Our methodology involves using n labeled samples $\{c_i, q_i, a_i\}_{i=1}^{n}$ from D_S. Additionally, we employ n' unlabeled samples $\{c'_j\}_{j=1}^{n'}$ from D_T, adhering to the same QA task as in D_S. We postulate that D_S and D_T have different data distributions. The main goal is to adapt a pre-trained model from D_S to D_T, ensuring the model can effectively bridge the domain gap after access to D_S ceases. This adaptation is critical to enhancing the generalization ability of the model to the new domain D_T, thus addressing the domain shift challenge.

4 Method

4.1 Overview

The PASAL framework (Fig. 1), designed for effective SFDA in QA tasks, comprises three principal components: the QG model, denoted as f_{gen}, the QA model, denoted as f, and a specifically designed prompt, denoted as π. Each of these elements—f_{gen}, f, and π—undergoes initial training with source domain data. Within the target domain D_T, f_{gen} is leveraged to generate pseudo-questions $q' = f_{\text{gen}}(c')$, which, in

turn, elicits pseudo-answers $a' = f(\pi, c', q')$ from the QA model using the prompt. The generated pseudo-triplets (c', q', a') instigate a systematic self-learning cycle: The prompt is refined first, followed by the QA model, and subsequently the QG model, each step leveraging pseudo-labeled data to progressively enhance adaptability of the model to the target domain. This recursive pattern of alternation between fine-tuning the prompt and the models embodies the core of our self-learning strategy, leading to robust domain adaptation.

4.2 Question Generation Model

In our QG model, denoted as f_{gen}, we employ a Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020). For each context, the prefix token "`generate question:`" is prepended to signal the generation task to the model. This modified context serves as the input, and the model is trained to output the corresponding question q. The training objective for f_{gen} is defined by the cross-entropy loss function:

$$\mathcal{L}_{\text{qg}}(D) = \sum_{i=1}^{|D|} -\log p_{f_{\text{gen}}}(q^{(i)}|c^{(i)}), \qquad (1)$$

where $p_{f_{\text{gen}}}(q^{(i)}|c^{(i)})$ represents the conditional probability of generating the correct question $q^{(i)}$ given the context $c^{(i)}$, as predicted by the QG model f_{gen}.

4.3 Question Answering Model

While previous research on UDA for QA has largely utilized BERT-based, encoder-only models for their NLU capabilities (Yue et al., 2021; Laparra et al., 2021; Su et al., 2022), the current work employs the T5 architecture. The integration of T5, with its encoder and decoder components, is strategically chosen to optimize the use of prompt learning for improving domain adaptation in QA tasks.

We employ a soft prompt strategy as documented in recent research (Li and Liang, 2021; Qin and Eisner, 2021; Zhong et al., 2021; Liu et al., 2023). For a given question-context pair, the data is formatted as follows:

`question: xxx context: xxx`

This input is subsequently prefixed with a sequence of artificial tokens, resulting in the structure:

$\langle v_1, v_2, \ldots, v_k \rangle$ `question: xxx context: xxx`

Algorithm 1 PASAL Training Procedure

Require: Question Generation model f_{gen}, Question Answering model f, soft prompt π, number of iterations N

Pre-train f_{gen} on source domain $\mathcal{S}$ to generate pseudo-questions q
Pre-train π on $\mathcal{S}$
Pre-train f on $\mathcal{S}$ with the trained π
for $i = 1$ to N **do**
  for each context $c' \in \mathcal{D}_T$ **do**
    $q' \leftarrow f_{\text{gen}}(c')$
    Apply LM filtering to q'
    $a' \leftarrow f(\pi, c', q')$
    Apply LM filtering to a'
    Fine-tune π using (c', q', a'), keeping f fixed
    $a'' \leftarrow f(\pi, c', q')$
    Fine-tune f using (c', q', a'') with the refined π
    Update f_{gen} using (c', q')
  end for
end for

Here, each v_i, for $i \in \{1, 2, ..., k\}$, is a trainable vector $v_i \in \mathbb{R}^d$ and collectively, the sequence $\langle v_1, v_2, \ldots, v_k \rangle$ forms the soft prompt π. These vectors are initialized randomly and positioned in the lowest embedding layer of the PLMs. The dimensionality of these vectors is denoted by d, which aligns with the dimensionality of the hidden layers in the PLM. The hyperparameter k designates the number of tokens comprising the prompt π, and thus, also the length of the prompt.

The prompt π and the QA model, denoted as f, is also trained using the cross-entropy loss function:

$$\mathcal{L}_{\text{qa}}(D) = \sum_{i=1}^{|D|} -\log p_f(a^{(i)}|c^{(i)}, q^{(i)}, \pi) \qquad (2)$$

Here, $p_f(a^{(i)}|c^{(i)}, q^{(i)}, \pi)$ is the conditional probability that the QA model f, with the aid of the prompt π, assigns to generating the correct answer $a^{(i)}$ given the context $c^{(i)}$ and the question $q^{(i)}$.

4.4 Training Procedure

The training methodology for both the Question Generation model f_{gen} and the Question Answering model f encompasses two phases: initial training within the source domain and subsequent adaptation within the target domain. In the source domain,

we begin by training the prompt π while keeping the QA model f static, to capture domain-agnostic knowledge crucial for domain adaptation. Following this, we train f with the now-tuned prompt π remaining fixed. Additionally, f_{gen} is trained to generate pseudo questions for use in the target domain. In the target domain phase, given a specific context, f_{gen} is first employed to produce pseudo questions. These questions are then subjected to LM filtering, as per (Shakeri et al., 2020), to select those with high scores. The process continues with the combination of these contexts and pseudo questions, which are then fed into the QA model f to elicit pseudo answers. LM filtering is again utilized, this time to sieve out answers with low confidence. The resulting (c, q', a') pairs are initially used to fine-tune the prompt π, keeping f static. Subsequently, f, in conjunction with the newly refined π, is used to generate a fresh set of pseudo answers. Post-filtering, the remaining (c, q', a') pairs are directed towards fine-tuning f, while keeping the prompt π fixed. Following this, the identical (c, q) pairs are employed to fine-tune f_{gen}. These steps constitute a training loop, which is iterated multiple times to augment the adaptability and performance of the models in the target domain. The training procedure is delineated in Algorithm 1.

5 Experiment Setup

5.1 Datasets

In accordance with previous studies (Nishida et al., 2019; Yue et al., 2021, 2022), this research utilizes datasets from the MRQA (Fisch et al., 2019). The SQuAD dataset (Rajpurkar et al., 2016) is chosen as the source domain for our experiments. For target domain datasets, only unlabeled samples are accessible. In this paper, we employ HotpotQA (Yang et al., 2018), Natural Questions (NQ) (Kwiatkowski et al., 2019), NewsQA (Trischler et al., 2017), and BioASQ (Tsatsaronis et al., 2015), which are frequently used in the field.

5.2 Implementation Details

We utilize the T5 model developed by Google (Chung et al., 2022) and implement the PASAL framework using the Huggingface Transformers library (Wolf et al., 2020). The batch size is set to 8, with each training epoch spanning 8 iterations. For optimization, the AdamW optimizer (Loshchilov and Hutter, 2017) is employed. We set the learning rate to 1e-3 for the prompt and 5.6e-5 for the model. The self-training loop is executed 5 times. The LM filtering threshold is determined by model selection strategy (Nguyen et al., 2020; Yang et al., 2023a). The default prompt length is established at 100 tokens, and the maximum input sequence length is limited to 412 tokens, with a document stride of 128. Text pieces excluding the answers will be discarded in training. Other hyperparameters follow the default settings provided by the Transformers library.

5.3 Baselines

We compare PASAL with the following baselines.

- **Source Only** This baseline involves training the model on the source domain and evaluating it on target domains without employing any UDA techniques.
- **Pseudo Labeled** This approach fine-tunes the model, initially trained on the source domain, utilizing samples from the target domain that are augmented with pseudo questions generated by an off-the-shelf QG tool (Alberti et al., 2019).
- **AdaMRC** (Wang et al., 2019): This use the domain adversarial neural network (Ganin et al., 2016) to align the feature between the source and target domains.
- **UDARC** (Nishida et al., 2019): This research engages in multitask learning by performing the QA task in the source domain and the LM task in the target domain concurrently.
- **CAQA** (Yue et al., 2021): This research design a contrastive adaptation loss that enhances domain-invariant learning.

6 Results

6.1 Overall Results

Table 1 presents the primary experimental results, underscoring the consistent superiority of the PASAL method over baseline methods across all domains. The PASAL method exhibits an improvement in Exact Match (EM) by at least 6.18% and up to 11.07%, and in F1 score by a minimum of 7.12% and a maximum of 14.99%, when compared to the CAQA baseline. This underpins the robust UDA capabilities of PASAL. Furthermore, a variance

Methods	HotpotQA		NQ		NewsQA		BioASQ		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Source Only	47.84	65.69	46.34	60.13	41.13	57.19	47.88	59.62	45.80	60.66
Pseudo Labeled	49.32	66.71	47.52	60.88	41.20	57.26	50.43	62.87	47.12	61.93
UDARC	48.51	66.40	46.93	60.49	41.18	57.26	50.11	62.51	46.68	61.67
AdaMRC	50.13	67.47	48.40	61.11	41.35	57.31	51.85	63.25	47.93	62.29
CAQA	51.28	68.85	51.10	64.01	44.29	59.02	52.64	63.09	49.83	63.74
PASAL	**60.52**	**75.97**	**57.57**	**71.75**	**55.36**	**74.01**	**58.82**	**71.12**	**58.07**	**73.21**

Table 1: Main results on comparing question-answering performance while performing domain adaptation from SQuAD to MRQA datasets. EM denotes the exact match.

Methods	HotpotQA		NQ		NewsQA		BioASQ		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Full Model	60.52	75.97	57.57	71.75	55.36	74.01	58.82	71.12	58.07	73.21
- psf	58.55	74.38	55.61	69.93	53.93	72.76	57.07	69.37	56.29	71.61
- prompt	56.19	70.92	53.49	67.32	50.99	68.10	56.27	66.66	54.24	68.25
- msf	58.47	73.38	52.68	67.39	51.85	70.98	54.79	67.08	54.45	69.71
- LM filtering	58.52	73.20	52.72	67.50	52.05	70.97	55.05	67.41	54.71	69.53

Table 2: Ablation study results on question-answering performance for domain adaptation from SQuAD to MRQA datasets.

in performance across target domains is observed, with all methods achieving their best results on HotpotQA and their least effective performance on NewsQA. This disparity may be attributed to the varying degrees of domain alignment, with HotpotQA potentially being more akin to the source domain than NewsQA. This specific aspect will receive further examination in Section 6.3. Despite these domain disparities, the PASAL method exhibits more consistent outcomes across all domains. For example, the standard deviation in EM and F1 for CAQA is 3.25 and 3.49, respectively, in contrast to the PASAL method, which demonstrates a substantially lower standard deviation at 1.88 and 1.92. This reaffirms the robustness of PASAL in domain-invariant knowledge retention.

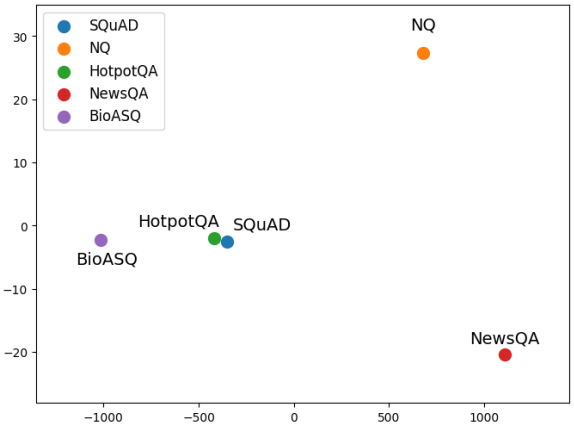

Figure 2: PCA visualization of embedding layer weights for domain specific models.

6.2 Ablation Study

To better understand our proposed framework, we conduct ablation studies to see the effectiveness of each component. The results are shown in Table 2. The notation "- psf" denotes the absence of self-learning of prompts in the target domain, as delineated in Line 10 of Alg. 1. The notation "- prompt" indicates the complete removal of the prompt module, "- msf" signifies the omission of self-learning within both the QA and QG models in the target domain, corresponding to Lines 11 and 12 of Alg. 1, and "- LM filtering" refers to the exclusion of the language model filtering process. Our findings reveal that prompt-tuning is pivotal to the success of PASAL; its removal leads to a notable degradation in performance. The self-learning mechanisms for both the prompt and language models are crucial, as their removal significantly impacts outcomes. The LM filtering process is also vital for maintaining high-quality pseudo samples by filtering out low-quality labels, thereby preventing a detrimental effect on performance.

Token length	SQuAD		HotpotQA		NQ		NewsQA		BioASQ	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
10 tokens	68.50	83.01	54.20	74.02	51.14	66.50	53.83	72.84	56.71	68.81
30 tokens	69.11	84.02	59.45	75.03	56.61	70.54	54.78	73.20	56.99	69.50
50 tokens	69.60	84.07	59.56	75.53	56.76	70.38	54.12	72.03	57.28	69.71
100 tokens	69.38	83.95	60.52	75.97	57.57	71.75	55.36	74.01	58.82	73.21

Table 3: Performance of PASAL with various prompt lengths across different domains.

6.3 Relationships between Domains

Previous studies (Cao et al., 2020; Yue et al., 2021) have engaged in qualitative analyses through manual inspection of question and context structures. However, it is argued that capturing the complex patterns and domain disparities is challenging when simply reviewing examples. In response, the current study leverages machine learning methodologies. The T5 model is trained on individual domains and employ principal component analysis for the visualization of encoder layer weights. These results are illustrated in Fig 2.

The visual evidence from the figure indicates a close similarity between HotpotQA and SQuAD, whereas Natural Questions and NewsQA exhibit substantial domain divergence from SQuAD. This observation corroborates the quantitative outcomes detailed in Table 1, where all methods achieve their best performance on HotpotQA, with NewsQA trailing with the lowest scores. This trend aligns with earlier studies (Cao et al., 2020; Yue et al., 2021) as well as our manual analysis, which highlights the unique textual styles and complex sentence structures of Natural Questions and NewsQA, distinct significantly from those in SQuAD. More analysis can be found in Appendix B.

6.4 Impact of Prompt Length

The adaptability of the PASAL system, when analyzed through the variation in prompt length, reveals a consistent trend: increasing the prompt length tends to correspond with enhanced EM and F1 scores, as indicated in Table 3. It is noteworthy that the performance gains are particularly substantial when the prompt length is expanded from 10 to 30 tokens. Beyond this point, the rate of improvement moderates. Furthermore, this trend is not consistent across different domains, suggesting that domain-specific characteristics significantly affect the efficacy of the prompts. A pronounced improvement is observed for the HotpotQA and NQ datasets with extended prompts, whereas the performance for SQuAD demonstrates a more subdued progression. This disparity suggests that the complexity of questions in HotpotQA and NQ benefits from extended prompts, which are perhaps necessary to encapsulate the requisite knowledge for defining the task.

6.5 Impact of Self-Training Loops

An experimental investigation was undertaken to evaluate the influence of the number of self-training loops on model performance, with the results presented in Figure 3. The analysis reveals a generally positive trend in performance enhancement with an increase in the number of loops, highlighting the efficacy of self-training in improving the accuracy and precision for PASAL. However, performance tends to plateau or even slightly fluctuate in the later stages, indicating potential instability. To mitigate the risk of overfitting, a decision was made to implement early stopping after the fifth loop.

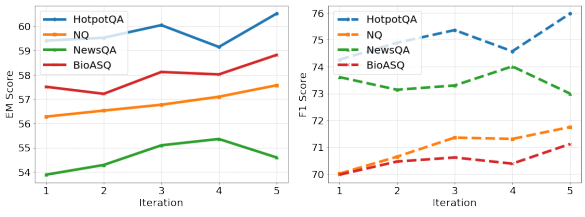

Figure 3: Comparative analysis of performance for PASAL across successive self-training loops.

6.6 Analysis of Prompt Embeddings

To better understand the influence of prompt learning on domain adaptation, we employed t-SNE visualization for the question-context pair embeddings, utilizing methodologies aligned with Wang et al. (2019) and Zhu and Hauff (2022). The analysis was conducted on the embeddings from the PLM that, along with the prompt, trained exclusively on the source domain data. The visualizations, displayed in Figures 4 and 5, reveal distinctive clustering patterns. Without the prompt, embeddings from different domains naturally clus-

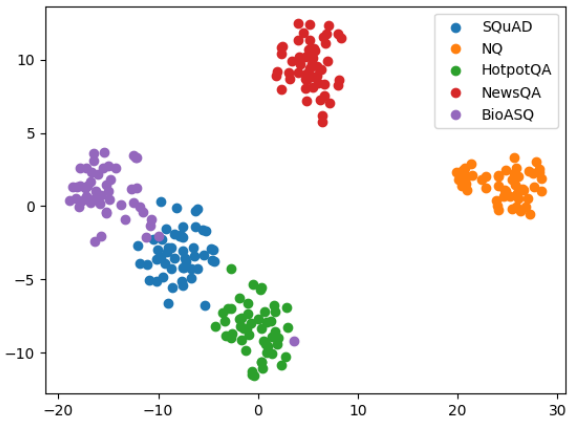

Figure 4: T-SNE visualization of question-context embeddings without prompt

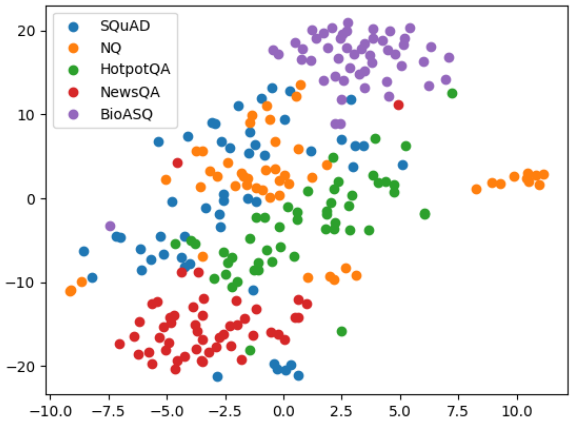

Figure 5: T-SNE visualization of question-context embeddings with prompt

ter into separate groups, as depicted in Figure 4, mirroring the domain relationships we previously noted at Table 1 and Fig. 2. In stark contrast, when question-context pairs are amalgamated with prompts, Figure 5 exhibits a convergence of these previously distinct clusters, clearly demonstrating that prompt learning significantly mitigates domain variances among samples from diverse domains.

HotpotQA	NQ	NewsQA	BioASQ
0.9994	0.9988	0.9945	0.9998

Table 4: Cosine similarity of prompt embeddings for different target domains compared with the source domain.

Additionally, we calculated the cosine similarity between the target domain-adapted prompt embeddings and those from the source domain, as shown in Table 4. Remarkably, these embeddings retain a high degree of similarity, as evidenced by the similarity scores, despite the application of a relatively high learning rate (1e-3). This finding substantiates our earlier assertion about the inherent stability of prompts across different domains. Furthermore, the similarity trends we observed corroborate our previous findings: prompts for HotpotQA and BioASQ are more similar to the source domain than those for NQ and NewsQA. This insight underscores the necessity of further fine-tuning prompts through our self-learning framework for enhanced domain adaptation.

6.7 Diverse Source Domain Analysis

In the pursuit of understanding the impact of source domain diversity on the performance of SFDA in QA systems, we conducted an exhaustive analysis across all domain datasets. As depicted in Table 5, the dataset from SQuAD was distinguished as a notably effective source domain, significantly enhancing performance across all domains under evaluation. This enhancement is likely due to the broad and varied collection of questions in SQuAD, which includes a wide range of topics and question styles. Such diversity provides a solid, versatile foundation for training the model, enabling effective domain adaptation and knowledge transfer.

6.8 Examples of Generated Questions

The percentage of generated questions starting with "what", "who", "when", "where" and "how" are 46.29%, 26.83%, 10.97%, 7.49% and 6.51%, respectively. We provide several examples of generated questions in Table 6. In the given examples, we observe a trend that the generated questions, while syntactically correct, sometimes miss the nuance of the GT questions. For instance, the GT question regarding the Del Mar Fair inquires about a specific historical name change, while the pseudo question focuses on the reinstatement of the San Diego County Fair name, which corresponds accurately to the answer provided. However, in the case of Amir Zaki, the GT question asks specifically about the club Zaki failed to return to, which implies a negative event, while the pseudo question merely asks about the current club, losing the context of the event in question. This suggests that while our question generator is adept at formulating syntactically coherent and contextually appropriate questions, enhancing its sensitivity to the nuances of situational context could further refine its output. Overall, the generator exhibits a commendable level of proficiency in synthesizing questions,

Datasets	SQuAD		HotpotQA		NQ		NewsQA		BioASQ	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
SQuAD	-	-	**60.52**	**75.97**	**57.57**	**71.75**	**55.36**	**74.01**	**58.82**	**71.12**
HotpotQA	69.89	80.80	-	-	52.92	67.20	49.00	68.24	51.51	68.72
NQ	68.23	79.03	58.01	73.14	-	-	52.18	71.87	56.73	71.02
NewsQA	65.76	77.54	50.02	69.54	52.45	68.23	-	-	49.13	67.85
BioASQ	**79.80**	**81.15**	59.01	73.90	54.34	68.02	48.71	67.56	-	-

Table 5: Cross-dataset performance evaluation of PASAL.

In 1954, the fair's name was changed to the Southern California Exposition and San Diego County Fair. In 1970, this was shortened to the Southern California Exposition. The fair was again renamed in 1984 to the Del Mar Fair, which lasted until 2002 when the name San Diego County Fair was reinstated. It is sometimes still referred to as the "Del Mar Fair" by locals. **Answer:** 2002 **GT Question:** When did the Del Mar Fair change its name? **Pseudo Question:** When was the name San Diego County Fair reinstated?
The surname Keith has several origins. In some cases, it is derived from Keith in East Lothian, Scotland. In other cases, the surname is originated from a nickname, derived from the Middle High German kīt a word meaning "sprout", "offspring". **Answer:** a nickname, derived from the Middle High German kīt **GT Question:** Where did the last name Keith come from? **Pseudo Question:** What is the surname Keith derived from?
LONDON, England (CNN) – After a week when he could not be traced, Egyptian striker Amir Zaki is back at his Premier League club side Wigan Athletic in northern England. ... Wigan and Egypt striker Amir Zaki has mended relations with his club manager. ... Zaki told Al-Hayat TV that the pair "ended up laughing" about his absence – when he failed to return from international duty and had a hamstring strain which no one knew the seriousness of. ... But, it wasn't all laughs a week ago. ... On Wigan's club Web site, Bruce had said of Zaki: "I just feel it's time that we went public on just what a nightmare he has been to deal with. ..." **Answer:** Wigan Athletic **GT Question:** Which club did Amir Zaki fail to return to? **Pseudo Question:** What Premier League club is Amir Zaki back at?

Table 6: Examples of generated questions compared with the ground-truth human-written questions.

but there remains a spectrum of improvement opportunities, particularly in the realm of semantic precision. This insight underscores the potential impact of synthetic data on the fine-tuning process of the answer module, where the fidelity of question-answer pairing is paramount.

7 Conclusion

In this paper, we propose the Prompt-Assisted Self-Adaptive Learning (PASAL) framework, a novel approach to SFDA for QA systems. By combining prompt learning with a self-learning strategy, PASAL enhances adaptability across various domains while upholding data privacy. The empirical results on various benchmark datasets demonstrate the superiority of PASAL over existing methods, particularly in its stability and performance across diverse target domains. The findings underscore the significance of incorporating prompt learning and self-learning strategies in the domain adaptation process, offering new avenues for future research in QA systems.

8 Limitations

While the PASAL framework marks a significant advancement in SFDA for QA, it is not without limitations. One notable constraint is its dependency on the quality of pseudo-questions generated during self-learning. If these questions are not sufficiently diverse or contextually relevant, the adaptation may not fully capture the nuances of the target domain. Furthermore, despite improve-

ments in domain adaptation, the performance of the model still varies across domains, indicating a need for further optimization in handling complex and highly divergent domains like NewsQA. Future research should focus on enhancing the question generation process and exploring more sophisticated methods for addressing diverse domain characteristics. Additionally, the computational demands of the iterative fine-tuning process necessitate consideration of efficiency improvements, especially for large-scale implementations.

Acknowledgement

The authors gratefully acknowledge the support received from the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Discovery Grants Program. Appreciation is also extended to the anonymous reviewers whose constructive feedback significantly contributed to the enhancement of this manuscript.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. 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.

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

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4921–4933, Online. Association for Computational Linguistics.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. 2021. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34:3635–3649.

Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. 2020. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Egoitz Laparra, Xin Su, Yiyun Zhao, Özlem Uzuner, Timothy Miller, and Steven Bethard. 2021. SemEval-2021 task 10: Source-free domain adaptation for semantic processing. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 348–356, Online. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. 2020. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. Gpt understands, too. *AI Open*.

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*.

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

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. 2020. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pages 7294–7305. PMLR.

Kosuke Nishida, Kyosuke Nishida, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Unsupervised domain adaptation of language models for reading comprehension. *arXiv preprint arXiv:1911.10768*.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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.

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.

Xin Su, Yiyun Zhao, and Steven Bethard. 2022. A comparison of strategies for source-free domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8352–8367, Dublin, Ireland. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

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.

Huazheng Wang, Zhe Gan, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Hongning Wang. 2019. Adversarial domain adaptation for machine reading comprehension. 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 2510–2520, Hong Kong, China. Association for Computational Linguistics.

Wei Wang, Zhun Zhong, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, and Nicu Sebe. 2023. Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24090–24099.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jianfei Yang, Hanjie Qian, Yuecong Xu, and Lihua Xie. 2023a. Can we evaluate domain adaptation models without target-domain labels? a metric for unsupervised evaluation of domain adaptation. *arXiv preprint arXiv:2305.18712*.

Shiqi Yang, Yaxing Wang, Luis Herranz, Shangling Jui, and Joost van de Weijer. 2023b. Casting a bait

for offline and online source-free domain adaptation. *Computer Vision and Image Understanding*, page 103747.

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.

Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A Ian McLeod, and Boyu Wang. 2023. When source-free domain adaptation meets learning with noisy labels. *arXiv preprint arXiv:2301.13381*.

Xiang Yue, Ziyu Yao, and Huan Sun. 2022. 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.

Qiuhao Zeng, Tianze Luo, and Boyu Wang. 2022. Domain-augmented domain adaptation. *arXiv preprint arXiv:2202.10000*.

Bo Zhang, Xiaoming Zhang, Yun Liu, Lei Cheng, and Zhoujun Li. 2021. Matching distributions between model and data: Cross-domain knowledge distillation for unsupervised domain adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5423–5433, Online. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Peide Zhu and Claudia Hauff. 2022. Unsupervised domain adaptation for question generation with DomainData selection and self-training. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2388–2401, Seattle, United States. Association for Computational Linguistics.

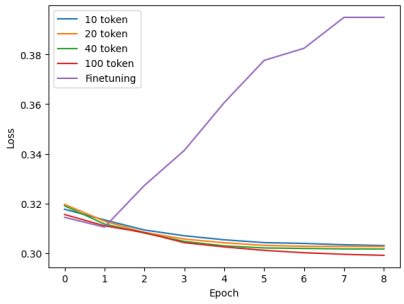

Figure 6: Evaluation losses for fine-tuning and prompt-tuning on SQuAD dataset.

A Comparison of Evaluation Loss

Figure 6 presents a comparison of evaluation losses for fine-tuning versus prompt-tuning on the SQuAD dataset. Throughout the training epochs, fine-tuning shows an initial reduction in loss, which then rises, indicating potential overfitting. Conversely, prompt-tuning demonstrates a steady, downward trend in loss, highlighting its consistent improvement and capacity for better generalization. This contrast reinforces the premise that prompt-tuning can offer superior generalization over fine-tuning in the context of PLM training. This observation aligns with our approach, emphasizing the critical role of domain-agnostic learning facilitated by prompt learning in adapting to new domains. It underscores the potential of prompt-tuning, not merely as a training technique but as a strategic tool to foster adaptability across domain shifts, reaffirming the core tenets of our PASAL framework.

B Examples Across Domains

This section offers a selection of examples from the MROQ datasets, and undertakes a qualitative analysis of the relationships and distinctive characteristics that define each domain.

B.1 SQuAD

Example 1

Question: To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?

Context: Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately

behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.

Answer: Saint Bernadette Soubirous

Example 2

Question: How many BS level degrees are offered in the College of Engineering at Notre Dame?

Context: The College of Engineering was established in 1920, however, early courses in civil and mechanical engineering were a part of the College of Science since the 1870s. Today the college, housed in the Fitzpatrick, Cushing, and Stinson-Remick Halls of Engineering, includes five departments of study – aerospace and mechanical engineering, chemical and biomolecular engineering, civil engineering and geological sciences, computer science and engineering, and electrical engineering – with eight B.S. degrees offered. Additionally, the college offers five-year dual degree programs with the Colleges of Arts and Letters and of Business awarding additional B.A. and Master of Business Administration (MBA) degrees, respectively.

Answer: eight

B.2 HotpotQA

Example 1

Question: Where did the form of music played by Die Rhöner Säuwäntzt originate?

Context: Die Rhöner Säuwäntzt are a Skiffle-Bluesband from Eichenzell-Lütter in Hessen, Germany. The line-up consists of Martin Caba, Christoph Günther and Christoph Leipold playing Skiffle-Blues with lyrics based on Rhön Mountains dialect and other Hessian dialects varieties. The expression "Säuwäntzt" means pork belly and refers also to untidy or unruly children and youth. Skiffle is a music genre with jazz, blues, folk and American folk influences, usually using a combination of manufactured and homemade or improvised instruments. Originating as a term in the United States in the first half of the 20th century, it became popular again in the UK in the 1950s.

Answer: United States

Example 2

Question: Who is the American internet entrepreneur who founded the company featured on 24 Hours on Craigslist?

Context: 24 Hours on Craigslist is a 2005 American feature-length documentary that captures the people and stories behind a single day's posts on the classified ad website Craigslist. The film, made with the approval of Craigslist's founder Craig Newmark, is woven from interviews with the site's users, all of whom opted in to be contacted by the production when they submitted their posts on August 4, 2003.

Answer: Craig Newmark

B.3 Natural Questions

Example 1

Question: Where did they hike in "Just Go With It"?

Context: The film was shot in Los Angeles and the Hawaiian islands of Maui and Kauai between March 2, 2010, and May 25, 2010. The film is deliberately vague about which Hawaiian island its latter portion depicts; thus, the characters hike across a rope bridge on Maui and arrive in the next scene at a spectacular waterfall on Kauai, rather than the ordinary irrigation dam and pond on Maui where the actual trail terminates.

Answer: Maui

Example 2

Question: Who did the motorcycle jump in "The Great Escape"?

Context: James Sherwin "Bud" Ekins (May 11, 1930 – October 6, 2007) was an American professional stuntman in the U.S. film industry. He is considered to be one of the film industry's most accomplished stuntmen with a body of work that includes classic films such as "The Great Escape" and "Bullitt". Ekins, acting as stunt double for Steve McQueen while filming "The Great Escape", was the rider who performed what is considered to be one of the most famous motorcycle stunts ever performed in a movie.

Answer: James Sherwin "Bud" Ekins

B.4 NewsQA

Example 1

Question: Where was Michael Strank born?

Context: WASHINGTON (CNN) – One of the Marines shown in a famous World War II photograph raising the U.S. flag on Iwo Jima was posthumously awarded a certificate of U.S. citizenship on Tuesday. Sgt. Michael Strank, who was born in Czechoslovakia and came to the United States when he was 3, derived U.S. citizenship when his

father was naturalized in 1935. However, U.S. Citizenship and Immigration Services recently discovered that Strank never was given citizenship papers.
Answer: Czechoslovakia
Example 2
Question: How many attacks have been done since July?
Context: BAGHDAD, Iraq (CNN) – Iraqi Security Forces captured 66 people believed to be connected to al Qaeda in Iraq terror cells, the U.S. military said Thursday. One of the suspects is believed to have conducted more than 12 attacks since July.
Answer: 12

B.5 BioASQ

Example 1
Question: What type of enzyme is peroxiredoxin 2 (PRDX2)?
Context: In melanoma, transition to the vertical growth phase is the critical step in conversion to a deadly malignant disease. The antioxidant enzyme peroxiredoxin-2 (Prx2) has a key role in this transition, inversely correlating with the metastatic capacity of human melanoma cells.
Answer: Antioxidant
Example 2
Question: What nerve is involved in carpal tunnel syndrome?
Context: This study aimed to determine the efficacy of median nerve epineurectomy in the surgical management of carpal tunnel syndrome (CTS). The median nerve is commonly implicated in CTS, showing flattening along with hypervascularization.
Answer: Median

B.6 Conlusion

Based on these examples, we can see that SQuAD and HotpotQA share a historical and fact-oriented focus, this is in align with our findings in Figure 2, which shows that HotpotQA is extremly close to the SQuAD dataset. Conversely, BioASQ is steeped in scientific and medical discourse, necessitating advanced technical understanding, which accounts for its notable distinction from SQuAD in the PCA space. Moreover, Natural Questions and NewsQA are characterized by their intricate structures and inferential demands, with Natural Questions covering pragmatic, real-life situations and NewsQA focusing on topical events and granular details. These complexities and the unique textual nuances contribute to their discernible departure from the SQuAD domain. Notably, the comparison between NewsQA and BioASQ underscores that contextual structure and complexity exert a greater impact on domain adaptation than the presence of specialized terminology.

C Justification for Selection of Baselines

The baseline methods were selected to represent the spectrum of UDA strategies pertinent to the QA context. UDARC is included for its fundamental approach using self-supervised learning within transformer models. AdaMRC, utilizing the Domain-Adversarial Neural Network (DANN) framework, offers a robust comparison due to its extensive validation and significant domain adaptation capabilities. CAQA, a recent advancement in UDA for QA, serves as a benchmark against current state-of-the-art methods. This selection provides a comprehensive overview of the application of UDA to QA, from foundational approaches to cutting-edge techniques.

D Details of Model Parameters, Computational Resources, and Infrastructure

The T5 model, used in our research, contains approximately 220 million parameters. These parameters are integral to the architecture of the model, encompassing the transformer encoder and decoder blocks. Each block in the model is composed of layers of self-attention mechanisms and fully connected neural network layers. Adhering to the standard T5 architecture, it features 12 layers each in both the encoder and decoder, a hidden size of 768, and 12 attention heads.

Our experiments were conducted on a server powered by an Intel(R) Xeon(R) Silver 4210R CPU at 2.40GHz with an x86_64 architecture, featuring 40 CPUs across 2 sockets, each with 10 cores and 2 threads per core. The system boasts a substantial memory capacity of 251 GB, with 185 GB available for use, and runs on a Linux kernel version 6.5.6-100.fc37.x86_64. The computational tasks, specifically model training and testing, were accelerated using an NVIDIA RTX A5000 GPU, chosen for its proficiency in handling demanding machine learning applications. The training process on the source domain was completed in approximately 30 hours. Additionally, the self-learning phase for a single target domain was conducted over a span of around 10 hours.