Unsupervised Domain Adaptation for Question Generation with Domain Data Selection and Self-training

Zhu, Peide; Hauff, Claudia

Publication date
2022

Document Version
Final published version

Published in
Findings of the Association for Computational Linguistics: NAACL 2022

Citation (APA)
Zhu, P., & Hauff, C. (2022). Unsupervised Domain Adaptation for Question Generation with Domain Data Selection and Self-training. In *Findings of the Association for Computational Linguistics: NAACL 2022: NAACL 2022 - Findings* (pp. 2388-2401). (Findings of the Association for Computational Linguistics: NAACL 2022 - Findings). Association for Computational Linguistics (ACL).

Unsupervised Domain Adaptation for Question Generation with Domain Data Selection and Self-training

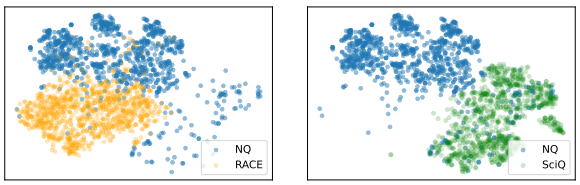
Peide Zhu and **Claudia Hauff**
Delft University of Technology
{p.zhu-1,c.hauff}@tudelft.nl

Abstract

Question generation (QG) approaches based on large neural models require (i) large-scale and (ii) high-quality training data. These two requirements pose difficulties for specific application domains where training data is expensive and difficult to obtain. The trained QG models' effectiveness can degrade significantly when they are applied on a different domain due to domain shift. In this paper, we explore an *unsupervised domain adaptation* approach to combat the lack of training data and domain shift issue with domain data selection and self-training. We first present a novel answer-aware strategy for domain data selection to select data with the most similarity to a new domain. The selected data are then used as *pseudo* in-domain data to retrain the QG model. We then present generation confidence-guided self-training with two generation confidence modeling methods: (i) generated questions' perplexity and (ii) the fluency score. We test our approaches on three large public datasets with different domain similarities, using a transformer-based pre-trained QG model. The results show that our proposed approaches outperform the baselines, and show the viability of unsupervised domain adaptation with answer-aware data selection and self-training on the QG task. The code is available at https://github.com/zpeide/transfer_qg.

1 Introduction

Natural language Question Generation (QG) aims to generate questions from given passages of text. It has been applied to a wide range of applications, such as question answering (Sultan et al., 2020; Fabbri et al., 2020), conversational systems (Gu et al., 2021), and education (Ma and Ma, 2019; Kurdi et al., 2020). Recently, pre-trained language models (LM) have advanced the state-of-the-art across a variety of natural language processing tasks (Devlin et al., 2018). Consequently, by modeling QG as a sequence-to-sequence task and fine-tuning on task-specific data, pre-trained LMs have substantially advanced the state-of-art performance on QG (Dong et al., 2019; Bao et al., 2020).



(a) (b)

Figure 1: 2D visualization of average-pool BERT hidden states of data from different domains using t-SNE. **(a)** Datasets NQ and RACE. **(b)** NQ and SciQ.

However, with billions of parameters, the performance of these deep neural models heavily relies on the quantity and quality of available training data. As the manual process of creating high-quality questions is expensive in terms of time and money, compared with abundant unlabeled data, the available data sources containing well-formed questions are insufficient, especially in the educational domain, where a lot of expertise is required to create questions geared towards human learning. To mitigate the lack of labeled training data, one solution is to pre-train models for QG on a data-abundant labeled domain (source domain) and transfer the learned knowledge to the unlabeled target domain, which is known as *unsupervised domain adaptation* (UDA) (Tan et al., 2018). It is a common challenge in machine learning research to learn knowledge in one domain and apply it in other domains with good generalization performance. One obstacle is the *domain shift* (Gretton et al., 2006) between the source domain and the target domain, as illustrated in Figure 1, which violates the assumption that the training set and the test set are independent and identically distributed (i.i.d.). This in turn limits the model's generalization and portability. To understand the effect of differences among domains on the performance of downstream QG tasks, fol-

Findings of the Association for Computational Linguistics: NAACL 2022, pages 2388 - 2401

lowing previous research (Hu et al., 2019; Aharoni and Goldberg, 2020), we perform a preliminary cross-domain study. We first train the QG model on all domains separately and evaluate them across different domain test sets. As shown in Table 1, the model achieves the best performance on the test set from the same domain and degrades dramatically on test sets of other domains, which poses a great challenge to the transferring task. We argue that based on these numbers further research into domain adaptation methods for QG is needed. There

Dataset	NQ	RACE	SciQ
NQ	**29.64**	13.76	14.32
RACE	16.59	**23.91**	12.37
SciQ	17.36	13.02	**29.47**

Table 1: Impact of domain shift on QG. Each row represents the METEOR score of the UniLM (Dong et al., 2019) model trained on one dataset (the row: NQ, SciQ and RACE) and tested on the test sets (the column).

is a growing interest in applying unsupervised domain adaptation to tackle the domain shift issue in natural language processing tasks, such as question answering (QA) (Rennie et al., 2020; Cao et al., 2020), or neural machine translation (NMT) (Van Der Wees et al., 2017; Rauf et al., 2019; Hu et al., 2019). However, UDA is under-examined in the context of question generation. Unlike the QA task which can be modeled as a multi-label classification problem, QG is a sequence generation problem, where it is hard to model the confidence or quality of generations (Niehues and Pham, 2019). Therefore, UDA methods for QA like pseudo-label generation and filtering cannot be directly extended to the QG area. Moreover, data augmentation UDA methods for the NMT task, such as domain mixing (Britz et al., 2017), back-translation (Sennrich et al., 2015), or target sentences copying (Currey et al., 2017) are not directly applicable to QG.

In this paper, we propose a two-stage unsupervised domain adaptation approach for QG to make use of the labeled source domain data, and abundant unlabeled data. In the first stage, we focus on unsupervised domain data selection. Although the definition of "domain" in QG is ambiguous, including the distribution of vocabulary, stylistic preferences, answer types etc, we first confirm that the learned BERT-based context paragraph representation can be used for robust domain data clustering as shown in Figure 1, and use Gaussian Mixture Models (GMMs) on the learned representations to find clusters, using methods proposed by Aharoni and Goldberg (2020). We perform domain data selection based on the distance of data example to cluster centers. To mitigate the gap of answer-type distributions, we further propose an answer-type aware data selection method (AADS) for pseudo in-domain data selection. The selected pseudo in-domain data are used to re-train the fine-tuned data to mitigate the domain shift.

In the second stage, we focus on self-training on the unlabeled target domain with the QG model trained in the first stage. The self-training approach is substantially hindered by noisy and low-quality generated pseudo labels. We first propose a normalization method to avoid re-enforcing poorly generated questions. We also explore using sentence perplexity and fluency scores to model the confidence of sequence generation. We filter pseudo labels with low sequence confidence during self-training to prevent the model from being degraded by wrong or low-quality predictions.

We conduct experiments across three domains, including the Natural Question dataset as the source domain, RACE as one target domain of education, and SciQ as the target domain of science. Our results show our proposed approach is effective even when the target domain is substantially different from the source domain and outperforms several baselines including Latent Dirichlet Allocation (LDA) (Druck et al., 2008), BERT discriminator based data selection (Ma et al., 2019), and unsupervised Gaussian mixture model(GMM) clustering on pre-trained language model features (Aharoni and Goldberg, 2020).

2 Background

In this section, we first present a short review for UDA and question generation, then we briefly discuss how our work is different from recent related research.

2.1 Unsupervised Domain Adaptation

The assumption that the training set and the test set are independent and identically distributed (i.i.d.) is a default assumption in many machine learning algorithms. When the underlying distributions do not match, the algorithms face the *domain shift* problem (Gretton et al., 2008; Ramponi and Plank, 2020), i.e. the *source* domain and the *target* domain data are not sampled from the same distribution. This issue happens in real-world scenarios, where

labeled training data are scarce while unlabeled data may be abundant since annotations are time-consuming and costly to acquire. It then translates into performance degradation. Unsupervised domain adaptation provides an elegant and scalable solution for mitigating this issue by learning only from unlabeled target data. In this paper, we focus on the data-centric methods: data selection and pseudo-labeling (Ramponi and Plank, 2020).

Data Selection for Domain Adaptation Not all samples in the source domain are equally important for adaptation. Data selection (Axelrod et al., 2011) aims to select the data that are most related to the target domain. It is attracting more attention, thanks to the abundance of data, and the large pre-trained models (Gururangan et al., 2020). It has been studied for several NLP tasks (Aharoni and Goldberg, 2020; Ma et al., 2019; Guo et al., 2020). Recently, Aharoni and Goldberg (2020) showed that sentence representation learned by pre-trained language models such as BERT (Devlin et al., 2018) and Roberta (Liu et al., 2019b) are capable of clustering textual data to domains in an unsupervised way with high precision. In our work, we follow this research and perform domain clustering and selection with BERT.

Self-Training Self-training is a bootstrapping method that has been used for domain adaptation in multiple NLP tasks (McClosky et al., 2006; Chattopadhyay et al., 2012; Bhatt et al., 2015; Sachan and Xing, 2018). The main idea of self-training (Lee et al., 2013) is to predict labels for unlabeled samples with a trained classifier as their 'pseudo' ground-truth, and use the synthetic data for further training.

2.2 Question Generation

Natural Question Generation (QG) aims to generate questions from given passages. Various neural models have been proposed for QG by formulating it as a sequence-to-sequence (Seq2Seq) learning problem (Du et al., 2017; Dong et al., 2019; Bao et al., 2020). QG has been applied to a range of application areas, such as conversational QA (Wang et al., 2018; Gu et al., 2021) and education (Kurdi et al., 2020). Although these approaches have made great strides in improving QG effectiveness, they are trained and tested with data from the same dataset. When there is domain shift between training and test data, the model performance deteriorates considerably. Liao and Koh (2020) explore this using supervised and semi-supervised domain adaptation but ignore the unsupervised setting.

The most related recent work to ours is by Kulshreshtha et al. (2021), who propose a new training protocol for UDA QG. However, it requires unlabeled questions in the target domain, which is not always available, and we focus on investigating a more effective self-training method. We compare this work in Appendix A.2.

In our work, we close the gap between source and target domain distributions by performing answer-type aware domain data selection.

3 Formalization

We now formulate the problem and present our notation. The data in the source domain with ground-truth questions are denoted as $\mathcal{D}_s = \{(C^s, Q^s)\}$, while unlabeled data in the target domain is $\mathcal{D}_t = \{(C^t\}$; here, C is denoting the context (the passages, and answer spans used for generating questions). The question generation task is then to generate a sequence $\hat{Q}$ that maximizes the conditional probability of the prediction $P(Q|C, \theta)$:

$$\begin{aligned}\hat{Q} &= \arg\max_{Q} P(Q|C, \theta) \\ &= \arg\min_{Q} \sum_{t=1}^{T} -\log P(Q_t|C, \theta, Q_{<t})\end{aligned} \tag{1}$$

where θ represents the parameters of the QG model, which is initially learned from training data in the source domain. In our work, we aim to learn to adapt the θ from a source domain $\mathcal{D}_S$ to the target domain $\mathcal{D}_T$ and achieve optimal performance.

4 Domains

4.1 Source Domain

We use the open-domain question answering corpus Natural Questions (NQ) (Kwiatkowski et al., 2019) as our source domain. It consists of aggregated questions issued to the Google search engine, and answers annotated by crowd-workers from the most related Wikipedia pages. It consists of a large amount of unique passages, and covers a range of topics, which makes it a good source domain for transferring. As there are many examples in NQ with tables as context, to use this dataset for QG, we select a subset which contains 89,453 samples in the training set and 3,726 samples in the test set, from the original NQ dataset.

4.2 Target Domains

Education The first target domain we choose is education, for which we use the RACE (Lai et al., 2017) dataset. RACE is a large dataset consisting of questions, answers and associated passages in English exams for middle-school and high-school Chinese students. Questions in RACE are designed by instructors (i.e. domain experts) for evaluating students' reading comprehension ability. There are three types of questions: cloze, general and specific. Following the practice of EQG-RACE (Jia et al., 2020), we only keep the specific questions in our work. For unsupervised QG, we use 18.6K data examples in the training set. The original dev and test sets are used for evaluation.

Science Our second target domain is science where we make use of the SciQ (Welbl et al., 2017) dataset. SciQ consists of 13.7K crowdsourced multiple-choice science exam questions, including 11.7K questions in the training set, and 1K for dev and test set each. Each SciQ question has an associated passage, the right answer, and the distractors. The SciQ passages are chosen from science study textbooks of different topics including biology, chemistry, earth science and physics. For unsupervised QG, we utilize the support passages in the training set without questions as unlabeled data; we use the original dev and test sets for QG evaluation.

Table 2 lists the basic statistics of our three datasets. On those datasets, we can make a thorough evaluation of the QG model's transfer performance and the effectiveness of proposed approach.

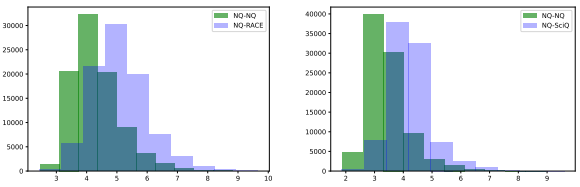
Features	NQ	SciQ	RACE
Question	Search Logs	Crowsourced	Experts
Context	Wikipedia	Textbook	Examinations
Train set	89,453	11,679	18,614
Test set	3,726	1,000	1036
#W/doc	106.27	78.05	318
#Sent./doc	4.43	4.84	17
#W/Sent.	26.81	16.13	17.96
#W/Q	10.20	14.31	10.8

Table 2: Overview of the source domain dataset NQ, and the selected datasets for target domains SciQ and RACE.

5 Domain Data Selection

Not all data are required or even useful for domain adaptation. Irrelevant data samples can add noise, and affect the learned model's performance and robustness towards cross-domain application considerably (Liu et al., 2019a). A solution to reduce the impact of irrelevant data is domain data selection, i.e. to retrieve the most appropriate data from the source domain data given the target domain data. Most proposed domain data selection approaches consider ranking training examples from $\mathcal{D}_S$ according to a domain similarity measure and select the top-n examples that are closest to $\mathcal{D}_T$.

We encode the context passage at the paragraph level with BERT, and perform average pooling of the last layer hidden state of each token to create its vector representation. To show that this is a robust representation for mapping sentences to domains in an unsupervised, data-driven approach, we first visualize them with t-SNE, as shown in Figure 1. We can observe the encoding vector representation with BERT indeed can cluster data examples to domains. Following the practice of Aharoni and Goldberg (2020), we then perform unsupervised clustering by fitting Gaussian Mixture Models (GMMs) to the vector context representations with k predefined clusters. We assign each cluster the domain class by measuring its purity (proportion of examples belonging to each domain). We use the Euclidean distance (Lee, 2001) of each example to cluster center as the measure of domain distance. Figure 2 shows the distribution of NQ dataset examples' distance to NQ's, RACE's and SciQ's domain center respectively. We sort source data examples based on their distance to the target domain center and select data examples with most domain similarity as the pseudo-in-domain data.



(a) (b)

Figure 2: Distribution of the distance between each data example to domain cluster center. **(a)** NQ and RACE. **(b)** NQ and SciQ.

Table 3 shows the unsupervised domain clustering results. We compare the proposed methods with Latent Dirichlet Allocation-based (LDA) clustering (Druck et al., 2008). We also compare different ways of creating paragraph vector representations, including using BERT [CLS] token

encoding ($_{CLS}$), average pooling of all BERT layer hidden states ($_{All}$), and average pooling of the last hidden states ($_{Last}$). Besides GMM clustering methods, we also compare the GMM method with K-Means (KM). To accelerate the clustering, we perform PCA over the paragraph representation first. Our results show the GMM method with pooling average of the last BERT hidden states to outperform the other methods.

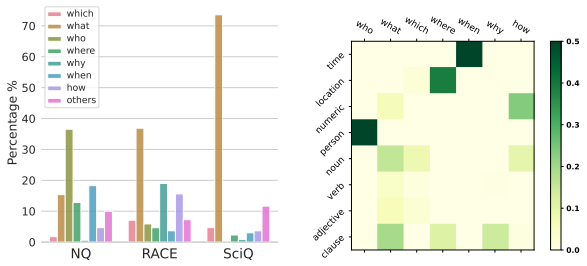
Method	RACE Acc	RACE F1	RACE Rc	SciQ Acc	SciQ F1	SciQ Rc
LDA	0.79	0.76	0.72	0.69	0.61	0.55
KM$_{CLS}$	0.37	0.35	0.98	0.33	0.25	0.97
KM$_{All}$	0.94	0.85	**0.99**	0.88	0.63	0.89
KM$_{Last}$	0.97	0.91	0.97	0.91	0.72	0.99
GMM$_{CLS}$	0.42	0.36	0.97	0.37	0.26	0.94
GMM$_{All}$	0.96	0.90	0.95	0.88	0.64	0.89
GMM$_{Last}$	**0.98**	**0.95**	0.96	**0.91**	**0.72**	**0.99**

Table 3: Unsupervised Domain Clustering Results.

5.1 Answer-Type Aware Data Selection

For different application domains, as shown in Figure 3a, the question type distributions vary a lot. For example, in NQ, the 'who' questions account for over 35% of all questions but in SciQ, 73.6% of questions have the 'what' type. Traditional data selection methods are based only on the similarity of context passages, which may suffer from unbalanced target label sampling. As there are no questions available in the target domain, it is a challenge to perform data selection according to the distribution of target question types. We first investigate the correlation between the answer types and question types. The question types are identified by the interrogative 'w'-word, such as 'who', 'what', etc. We identify the answer types such as 'time', 'location', etc. using the spacy[1] NER and POS tagger. The correlation matrix (expressed in Pearson correlation coefficient) is shown in Figure 3b. We find question types and answer types are strongly correlated to each other. For example, the correlation coefficient between 'time' and 'when' is 0.67, between 'person' and 'who' it is 0.63. Thus, we propose a heuristic answer-type aware data selection strategy for domain data selection from the source domain with a similar answer type distribution, in order to mitigate the label divergence. Specifically, we first group the data by answer types, and then conduct data selection on each group.

[1] https://spacy.io/



(a) (b)

Figure 3: **(a)** Question type distributions. **(b)** Correlation between answer type and question types.

6 Self-Training

When training the QG model with pseudo-labels, it is natural to put more emphasis on the labels that the model is more confident about. An intuitive solution is to weigh each pseudo-token according to its estimated probability in order to avoid reenforcing poor predictions. Thus, we propose the following normalized training objective for self-training:

$$\hat{Q} = \arg\min_{Q} \sum_{t=1}^{T} -\log \alpha_t P(Q'_t | C, \theta, Q'_{<t}) \quad (2)$$

where Q' is the pseudo-label, and α_t is the predicted probability of the t-th word Q'_t, and T is the length of the pseudo-label.

We apply the QG model to generate questions on unlabeled target-domain data, which are then used as 'pseudo' gold labels for further training. The self-training approach is substantially hindered by noisy, low-quality labels. How to deal with noisy pseudo labels is crucial to the final UDA effectiveness. Classical pseudo label generation methods (Mihalcea, 2004; Abney, 2007; Cui and Bollegala, 2019) filter generated labels by their 'confidence' which is the predicted probability of the label in those classification tasks. How to represent confidence of sequence generation in pseudo-labeling is insufficiently explored. Traditionally, confidence estimation has been defined as a task of assessing the quality of the whole sequence of words in the target sentence. Therefore, we propose a question quality guided pseudo labeling method to address this problem, with two confidence metrics: (i) the sentence perplexity, and (ii) the BERT-based fluency score.

Sentence Perplexity The first metric is the perplexity of the generated questions. The generation

with higher confidence should have lower perplexity. Here, perplexity (PPL) is defined as follows:

$$\text{PPL}(Q) = 2^{-\frac{1}{T}\log\prod_1^T P(Q_t|Q_{<t})} \quad (3)$$

BERT-based Fluency Score For our second metric, we use *fluency* as the question quality metric, which indicates whether the generation follows grammar rules and correct logic. The perplexity of a sentence under a well-trained language model usually serves as a good indicator of its fluency (Yang et al., 2018). We use a fine-tuned BERT language model as evaluator. The fluency metric $R_{fluency}$ for question Q is calculated as follows:

$$R_{fluency}(Q) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log \text{BERT}(Q_t|Q_{<t})\right). \quad (4)$$

During the unsupervised self-training, after each epoch, we perform beam search with the trained model, and the generated questions are ranked according to their fluency score. Only questions with confidence metrics better than the threshold ϕ and PPL are selected as pseudo-labels. If one data sample got selected in the last epoch, but its generated question's confidence metric in the current epoch is not higher than before, it is removed. In this way, only questions of high quality that improve over time are chosen for training.

Algorithm 1: Self-Training

Input : Target domain data: $\mathcal{D}_t = \{C^t\}$. QG model $\mathcal{M}_{QG}$ with parameters θ

repeat
  for $C^t \in \mathcal{D}_t$ **do**
    $[Q_t^{'}, \alpha_t]_1^T = \mathcal{M}_{QG}(C)$
    if *Use Fluency Score* **then**
      $f =$ $\exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log BERT(Q_t^{'}|Q_{<t}^{'})\right)$
    else if *Use PPL* **then**
      $f = 2^{-\frac{1}{T}\log\prod_1^T P(Q_t|Q_{<t})}$
    if $f > \phi$ **then**
      $\mathcal{L} =$ $\mathcal{L} + \sum_{t=1}^{T} -\log \alpha_t P(Q_t^{'}|C, \theta, Q_{<t}^{'})$
    end
  end
  $\theta \leftarrow \text{Adam}(\nabla_\theta \mathcal{L})$.
until *Convergence or Reach Maximum Epochs*;

7 Experiments

In this section, we describe the model and the training regime in more detail.

7.1 Experimental Settings

QG Model We use the state-of-art pre-trained transformer-based sequence-to-sequence natural language understanding and generating model **UniLM** (Dong et al., 2019) for question generation. Specifically, we choose the uncased pre-trained `unilm1.2-base-uncased` model for fine-tuning. It has 12 transformer layers and is jointly pre-trained on large amounts of text, optimized for bidirectional, unidirectional, and sequence-to-sequence language model objectives. We use the `s2s-ft` package[2] for fine-tuning. To fine-tune our model, the input context passage, the answer, and the generated question are combined together into a sequence: "*`[CLS]` context passage `[EOS]` answer span `[EOS]` question `[EOS]`*". Both the input passage and answer are regarded as the first text segment, while the generated question is the second segment in the unified LM.

Training Details The model is trained on a server consisting of 4 GeForce GTX 1080 gpus with a batch size of 32, a mask probability of 0.8, and the label smoothing rate of 0.1. The max_source_seq_length is set 464, the max_target_seq_length is 48. We first fine-tune UniLM with the `NQ` dataset for 10 epochs. We use the Adam optimizer with $\epsilon = 1e-8$, learning rate is $1e-4$ with 500 warmup steps.

Unsupervised Domain Data Clustering We use 4,500 examples randomly selected from `NQ`, `SciQ` and `RACE` for unsupervised data clustering. We set the number of clusters as 2, since we intend to investigate the separability between the source domain and the target domain.

Evaluation Metrics We compare the model performance along three automatic evaluation metrics: **BLEU** (Papineni et al., 2002), which is computed with the geometric average of the modified n-gram precision and the brevity penalty; **Meteor** (Denkowski and Lavie, 2014), which compares the generation with the gold question in terms of exact, stem, synonym, and paraphrase matches; and **Rouge-L** (Lin, 2004), which measures the shared longest common sub-sequence. We calculate these metrics with the package released by Du et al. (2017). We also conduct a human evaluation.As a sanity check and to evaluate the QG

[2] https://github.com/microsoft/unilm/tree/master/s2s-ft

	Method	RACE B-1	RACE B-4	RACE MT	RACE RG	SciQ B-1	SciQ B-4	SciQ MT	SciQ RG
	None	21.99	4.11	13.68	21.31	25.94	8.67	15.53	26.59
DDS	random	21.91	4.02	13.74	21.26	26.15	8.97	15.56	26.62
	LDA	21.97	4.29	13.72	21.47	26.57	8.88	15.67	27.07
	BERT-DDS	22.06	3.99	13.61	21.30	26.43	9.08	15.70	26.70
	KMeans	22.21	4.45	13.75	21.65	26.45	9.23⋆	15.72	27.15⋆
	GMM	22.38	4.58	14.05	21.70	26.51	9.08	15.79⋆	27.05
	AA-KMeans	22.28	4.40	13.92	21.71	26.26	8.85	15.66	26.82
	AA-GMM	22.79⋆	4.79⋆	14.23⋆	22.15⋆	26.61⋆	9.09	15.73	26.90
ST	w/o-Norm	23.34	4.82	14.45	22.89	27.89	10.37	16.51	28.26
	w/o-Filter	23.83	5.13	14.65	23.06	28.29	10.85	16.95	28.86
	Fluency	24.20	5.11	14.74	23.66	28.22	10.76	16.92	28.92
	PPL	24.38♡	5.22♡	14.85♡	23.43	28.30	11.04	17.12	29.03
	Fluency&PPL	24.32	5.14	14.73	23.52♡	28.30♡	11.04♡	17.12♡	29.03♡
DDS +ST	w/o-Filter	23.43	4.93	14.43	22.78	28.21	11.00	16.90	28.93
	Fluency	24.20	4.85	14.67	23.13	28.82	11.05	16.86	28.94
	PPL	24.43	5.40♣	15.08	23.49	29.12	11.04	16.92	29.38
	Fluency&PPL	24.71	5.20	14.96	23.78	29.40♣	11.23	17.13	29.52♣
	AA-Fluency	24.14	5.17	14.79	23.07	28.10	10.82	16.69	28.54
	AA-PPL	24.50	5.14	15.09♣	23.60	28.84	11.65	17.22♣	29.30
	AA-Fluency&PPL	24.71♣	5.16	14.87	23.80♣	28.68	11.70♣	17.17	29.36

Table 4: Results of unsupervised domain adaptation for QG with answer-type aware (**AA-**) domain data selection(**DDS**) and self-training(ST) on RACE and SciQ test set. We compare three baseline methods: LDA (Druck et al., 2008), BERT-DDS (Ma et al., 2019) , GMM (Aharoni and Goldberg, 2020). ⋆ denotes the best results for DDS, ♡ denotes best results for ST, and ♣ denotes best results for DDS+ST.

Dataset	B-1	B-4	MT	RG
NQ	60.05	30.31	29.64	59.26
SciQ	46.99	33.22	29.47	42.73
RACE	37.86	17.90	23.91	37.56

Table 5: In-domain test results of the QG model (fine-tuned and tested on the same dataset).

model's ability to generate questions based on these datasets, we first conduct in-domain tests on these three datasets separately, i.e. we fine-tune and test the model on the training/test set from the same dataset. As shown in Table 5, we achieve results comparable with state-of-art for the NQ, RACE and SciQ datasets.

7.2 Experiments on Data Selection

In this experiment, we compare the proposed answer-type aware data selection with several baselines. We train the QG model with the selected data and evaluate the data selection method by comparing its performance. The first baseline is random data selection (**random**). With this baseline, we randomly sample 1,000 samples from NQ. The second baseline is LDA-based clustering (Druck et al., 2008). We use the gensim (Řehůřek and Sojka, 2010) LDA implementation for this baseline. The third method (**BERT-DDS**) is proposed by Ma et al. (2019), where a BERT-based domain discriminator is used for data selection. The discriminator is first trained with randomly sampled data from the datasets. The baseline model achieved 99.85% for RACE and 92.35% accuracy for the SciQ dataset. The last baseline method we compare is adopted from the unsupervised domain clustering method (**GMM**) proposed by Aharoni and Goldberg (2020), as described in Section 5. We use the BERT-base model implementation of huggingface transformers (Wolf et al., 2020) to get the context passage encoding. In addition to GMM, we also compare the K-Means method (Sculley, 2010). The results are presented in Table 4.

Impact of Domain Data Selection Re-training with randomly selected data does not improve our model's generalization performance. All other data selection methods outperform random data selection, except BERT-DDS. One reason is that BERT-DDS training needs sampling data from different domains, its performance relies on the sampled data, and also label examples that are similar to the target domain as source domain. *Data selection with unsupervised domain clustering with BERT context encoding outperforms other methods,*

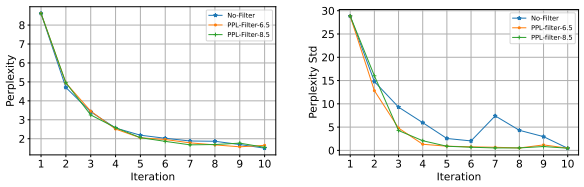
which confirms its effectiveness.

On the RACE dataset, answer-type aware data selection with K-Means (**AA-KMeans**) and GMM (**AA-GMM**) outperform the same selection method without answer-type awareness. We note this result does not always hold for the SciQ dataset. One possible reason is due to the extremely unbalanced answer type distribution in SciQ: we have to select examples with generally low domain similarities wrt. the source domain to create the same answer-type distributions.

7.3 Experiments on Self-Training

We conduct self-training with the target-domain unlabeled data on the QG model fine-tuned on the NQ dataset. We first verify the effectiveness of the proposed normalized training objective. As the results show in Table 4, self-training with normalization (**w/o-Filter**) outperforms self-training without any confidence filtering and normalization (**w/o-Norm**), which indicates its effectiveness.

Impact of Generation Confidence Guided Self-training We explore two generation confidence metrics for self-training, the sentence perplexity, and the question fluency score. To train the BERT LM for generating fluency scores for question quality evaluation, we combine all questions from NQ and the Quora Question Pairs dataset[3], creating a dataset consisting of 834,834 questions. The final model achieves a perplexity of 9.27 on the evaluation set. As the results in the **ST** part of Table 4 show, both proposed generation confidence metrics improve the performance considerably up to 6%. This can be explained by the removal of low-quality and noisy data, which hinders model training. As



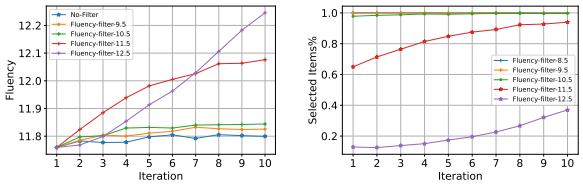
(a) (b)

Figure 4: Change of **(a)** average perplexity, and **(b)** standard deviation of generations along iterations.

Figure 4 shows, with perplexity filtering—although the changing curves of mean perplexity of the generated pseudo-labels in each iteration are similar—the standard deviation drops faster and more steady.

[3] https://www.kaggle.com/c/quora-question-pairs

As Figure 5 shows, the average fluency score improves along iterations even without fluency filtering, but with fluency filtering, the average fluency score improves more steadily and increases towards the threshold value ϕ. The proportion of questions with higher fluency score than ϕ increase along iterations. As reflected in Figure 5b and Table 6, if the threshold value is too low, fewer noisy pseudo examples can be filtered out. If the threshold is too high, there would be less supervision for the QG model. Both of these settings would lead to performance degradation.



(a) (b)

Figure 5: Change of **(a)** average fluency score, and **(b)** the percentage of generated questions whose fluency score is higher than ϕ along iterations.

	ϕ	B-1	B-4	MT	RG
RACE	8.5	23.83	**5.12**	14.65	23.06
	9.5	24.20	5.11	**14.74**	**23.66**
	10.5	**24.23**	5.06	14.68	23.29
	11.5	23.93	5.10	14.46	23.09
	12.5	23.78	4.55	14.41	23.05

Table 6: Influence of the fluency threshold (ϕ).

Impact of Joining Domain Data Selection and Self-Training We also conduct domain adaptation by joining domain data selection and self-training (**DDS+ST**). As shown in Table 4, joining DDS and self-training without filtering does not show performance improvement on both datasets, which implies with DDS, pseudo-labels during self-training may be noisier. With the proposed filtering with fluency score or question perplexity, the joint method outperforms DDS and self-training. On the RACE dataset, the answer-type aware joint methods generally achieves the best performance across all evaluation metrics.

7.4 Human Evaluation

In addition to the automatic evaluation results shown in Table 4, we also report on our human evaluation in Table 7. We randomly sampled 50 generated questions from the RACE and SciQ test

Method	RACE			SciQ		
	Syntax	Relevance	Answerability	Syntax	Relevance	Answerability
w/o-UDA	2.60 (0.66)	2.00 (0.78)	0.43 (0.49)	2.83 (0.40)	2.40 (0.64)	0.57 (0.50)
ST	2.78 (0.51)	2.12 (0.73)	0.46 (0.50)	**2.94 (0.26)**	2.49 (0.64)	**0.67 (0.47)**
DDS+ST	**2.81 (0.47)**	**2.12 (0.75)**	**0.51 (0.50)**	2.92 (0.27)	**2.53 (0.63)**	0.67 (0.47)

Table 7: Human evaluation (mean and standard deviation) on RACE and SciQ datasets. Syntax and Relevance evaluation adopt a 3-point scale. Higher is better; Answerability is boolean type (0-1).

set respectively and asked 3 domain experts (both male and female, ages ranging from 25 to 35) to rate the generated questions by the QG model without UDA (w/o-UDA), with self-training (ST), and self-training and domain data selection(DDS+ST). The experts are also presented with the context paragraphs, the answers, as shown in Figure 6 of the appendix. The generated questions are shown in Table 10 of the appendix. We rate questions along three dimensions: (i) syntax, i.e. the syntax correctness, in a 3-point scale, 1 for major syntax issues, 2 meaning minor issue and 3 is correct; (ii) relevance, i.e. whether the question is relevant to the context and the answer, also in a 3-point scale, 1 irrelevance, 2 for partial relevance and 3 meaning fully relevant; (iii) answerability, a boolean type value, indicating whether the question can be answered given the context and answer. As the results show, all QG with UDA methods outperform the QG model without domain adaptation. On the RACE dataset, the proposed unsupervised domain adaptation for QG with data selection and self-training (DDS+ST) achieves the best performance along with all metrics; although the performance of UDA with self-training only outperforms DDS+ST slightly in terms of syntax and answerability, DDS+ST outperforms self-training.

8 Conclusion

We proposed an unsupervised domain adaptation approach for question generation. Our approach includes an answer-type aware unsupervised domain data selection method and a sequence generation confidence guided self-training algorithm. We conduct experiments on three domains. We use the Natural Questions dataset as labeled source domain, RACE as target education domain and SciQ as target science domain. Our results suggest our approach is effective for this application settings. We find that it significantly improves domain adaptation performance of our QG model. In future work, we plan to expand our work to more domains and additional QG model types.

Acknowledgements

The first author has been supported by the China Scholarships Council (CSC). The second author has been supported by NWO project SearchX (639.022.722) and NWO Aspasia (015.013.027). We also would like to thank the anonymous ACL Rolling Review reviewers who kindly provided us with plenty of helpful comments for improving this paper and the experts who helped us with the human evaluation.

References

Steven Abney. 2007. *Semisupervised learning for computational linguistics*. CRC Press.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. *arXiv preprint arXiv:2004.02105*.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 355–362.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR.

Himanshu Sharad Bhatt, Deepali Semwal, and Shourya Roy. 2015. An iterative similarity based adaptation technique for cross-domain text classification. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 52–61.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126.

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.

Rita Chattopadhyay, Qian Sun, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. 2012. Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–26.

Xia Cui and Danushka Bollegala. 2019. Self-adaptation for unsupervised domain adaptation. *Proceedings-Natural Language Processing in a Deep Learning World*.

Anna Currey, Antonio Valerio Miceli-Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

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*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13063–13075.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.

Alexander R Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. *arXiv preprint arXiv:2004.11892*.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19:513–520.

Arthur Gretton, Karsten Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander J Smola. 2008. A kernel method for the two-sample problem. *arXiv preprint arXiv:0805.2368*.

Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. Chaincqg: Flow-aware conversational question generation. *arXiv preprint arXiv:2102.02864*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7830–7838.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. *arXiv preprint arXiv:1906.00376*.

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. 2020. Eqg-race: Examination-type question generation. *arXiv preprint arXiv:2012.06106*.

Devang Kulshreshtha, Robert Belfer, Iulian Vlad Serban, and Siva Reddy. 2021. Back-training excels self-training at unsupervised domain adaptation of question generation and passage retrieval. *arXiv preprint arXiv:2104.08801*.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

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.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *International Workshop on Artificial Intelligence and Statistics*, pages 176–183. PMLR.

Yin-Hsiang Liao and Jia-Ling Koh. 2020. Question generation through transfer learning. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 3–17. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Miaofeng Liu, Yan Song, Hongbin Zou, and Tong Zhang. 2019a. Reinforced training data selection for domain adaptation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1957–1968.

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

Lin Ma and Yuchun Ma. 2019. Automatic question generation based on mooc video subtitles and knowledge graph. In *Proceedings of the 2019 7th International Conference on Information and Education Technology*, pages 49–53.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with bert-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159.

Rada Mihalcea. 2004. Co-training and self-training for word sense disambiguation. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 33–40.

Jan Niehues and Ngoc-Quan Pham. 2019. Modeling confidence in sequence-to-sequence models. *arXiv preprint arXiv:1910.01859*.

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

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. *arXiv preprint arXiv:2006.00632*.

Sadaf Abdul Rauf, Kiran Kiani, Ammara Zafar, Raheel Nawaz, et al. 2019. Exploring transfer learning and domain data selection for the biomedical translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 156–163.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Steven Rennie, Etienne Marcheret, Neil Mallinar, David Nahamoo, and Vaibhava Goel. 2020. Unsupervised adaptation of question answering systems via generative self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1148–1157.

Mrinmaya Sachan and Eric Xing. 2018. Self-training for jointly learning to ask and answer questions. 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 629–640.

David Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for qa. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer.

Marlies Van Der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. *arXiv preprint arXiv:1708.00712*.

Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. *arXiv preprint arXiv:1805.04843*.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. *arXiv preprint arXiv:1805.11749*.

A Appendix

A.1 Examples Selected Data

Table 9 illustrates several data examples of selected from NQ dataset that are most similar to *education* domain, i.e. the RACE dataset, and to *science* domain, i.e. SciQ dataset using the GMM$_{last}$ BERT based domain data selection method. The RACE dataset is a large dataset of English exams for middle-school and high-school Chinese students. Its vocabulary is middle-school and high-school level. Many passages in it are story-style. As **NQ→RACE** data examples show, the selected data from NQ are close to SciQ in terms of both the vocabulary and text style. Meanwhile, SciQ passages are chosen from science study textbooks of different topics including biology, chemistry, earth science and physics. The examples of selected data (**NQ→RACE**) can be categorized into the biology domain, which includes a lot of the biology terms, elucidating biological processes. These examples show the effectiveness of the data selection method.

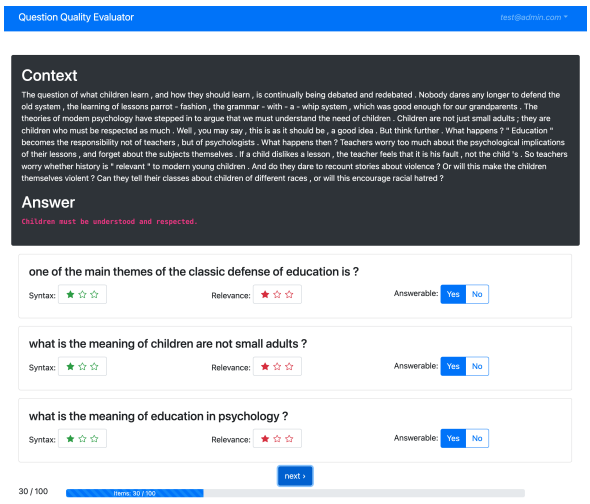


Figure 6: The interface for human annotation.

A.2 Experiments on MLQuestions

We conduct unsupervised domain adaptation experiments on MLQuestions (Kulshreshtha et al., 2021). We first conduct unsupervised domain data selection with GMM_{last} method and presents the confusion matrix in Figure 7 and select 1,000 data examples from NQ that are most close to MLQuestions clustering center. We set number of clusters as 2 because we want to directly investigate the unsupervised separability between NQ and MLQuestions. We use the provided development set and the test set of MLQuestions. Then we perform domain adaptation for QG, and show results in Table 5. Compared with the self-training method explored in (Kulshreshtha et al., 2021), the proposed method in this paper achieves more performance increase, e.g. **DDS+ST** method achieved 5.42 and 4.7 net gain in Rouge-L and BLEU-1 score respectively, compared with 0.58 and 0.80 net gain with self-training in (Kulshreshtha et al., 2021). We focus on the self-training method in this paper, so we consider conducting open-domain retrieval-based methods like **Back-Training** in future research.

Dataset	B-1	B-4	MT	RG
w/o-UDA	30.06	7.96	18.62	31.60
DDS	29.89	8.27	18.63	31.64
ST	32.58	9.41	19.41	34.20
DDS+ST	34.76	10.57	20.41	37.02
Net Gain	**4.7**↑	**2.61**↑	**1.79**↑	**5.42**↑

Table 8: Unsupervised domain adaptation results on MLQuestions dataset.

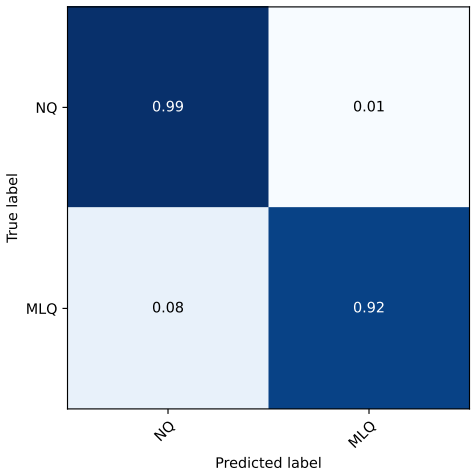


Figure 7: Confusion matrix for unsupervised domain data clustering results on MLQuestions and NQ datasets. We use 3,000 data examples from NQ and MLQuestions each.

NQ → RACE	**NQ → SciQ**
To expand the number of women smokers Hill decided to hire Edward Bernays, who today is known as the father of public relations, to help him recruit women smokers. Bernays decided to attempt to eliminate the social taboo against women smoking in public. ... The targeting of women in tobacco advertising led to higher rates of smoking among women. In 1923 women only purchased 5% of cigarettes sold, in 1929 that percentage increased to 12%, in 1935 to 18.1%, peaking in 1965 at 33.3%, and remaining at this level until 1977.	The lysosomes also act as the waste disposal system of the cell by digesting unwanted materials in the cytoplasm, both from outside the cell and obsolete components inside the cell. Material from outside the cell is taken - up through endocytosis, while material from the inside of the cell is digested through autophagy. Their sizes can be very different. They were discovered and named by Belgian biologist Christian de Duve, who eventually received the Nobel Prize in Physiology or Medicine in 1974.
A man named Bailey intends to take his family from Georgia to Florida for a summer vacation , but his mother , (referred to as "the grandmother" in the story) wants him to drive to East Tennessee , where the grandmother has friends ("connections"). She argues that his children, John Wesley and June Star, have never been to East Tennessee, and she shows him a news article in the Atlanta Journal Constitution ... He and the grandmother agree that things were much better in the past and that the world at present is degenerate; she concurs with Sammy's remark that "a good man is hard to find."	Decomposition is the process by which organic substances are broken down into simpler matter. The process is a part of nutrient cycle and is essential for recycling the finite matter that occupies physical space in the biosphere. Bodies of living organisms begin to decompose shortly after death. Animals, such as worms, also help decompose the organic materials. Organisms that do this are known as decomposers. Although no two organisms decompose in the same way, they all undergo the same sequential stages of decomposition. The science which studies decomposition is generally referred to as taphonomy from the Greek word taphos, meaning tomb.
The next day, just before Lincoln and Sara board a boat to escape to the Dominican Republic, Sucre gives Sara the $100,000 they stole from the General, apologizing for not being able to wire the money to them the night before as planned. Mahone gives Sara the paper Michael asked him to deliver, ..., but don't ever, say. He then says what he wants to say is that he loves them both, very much. He tells them to make sure his child is told every day how much he is loved and how lucky he is to be free. The video, and the entire series	An elater is a cell (or structure attached to a cell) that is hygroscopic, and therefore will change shape in response to changes in moisture in the environment. Elaters come in a variety of forms, but are always associated with plant spores. In many plants that do not have seeds, they function in dispersing the spores to a new location. Mosses do not have elaters, but peristome which also change shape with changes in humidity or moisture to allow for a gradual release of spores

Table 9: Examples of selected data from NQ dataset that are most similar to RACE dataset (NQ→RACE) and SciQ dataset (NQ→RACE).

	RACE	SciQ
Context	Jenny was a pretty five-year-old girl. One day when she and her mother were checking out at the grocery store , Jenny saw a plastic pearl necklace priced at $2.50. Her mother bought the necklace for her on condition that she had to do some homework to pay it off. Jenny agreed. She worked very hard every day, and soon Jenny paid off the necklace. Jenny loved it so much that she wore it everywhere except when she was in the shower. Her mother had told her it would turn her neck green! Jenny had a very loving daddy. When Jenny went to bed, he would read Jenny her favorite story. One night when he finished the story, he said, "Jenny, could you give me your necklace?" "Oh! Daddy, not my necklace!" Jenny said." But you can have Rosy, my favorite doll. Remember her?	Gamma rays are produced when radioactive elements decay. Radioactive elements are elements with unstable nuclei. To become more stable, the nuclei undergo radioactive decay. In this process, the nuclei give off energy and may also emit charged particles of matter. Types of radioactive decay include alpha, beta, and gamma decay. In alpha and beta decay, both particles and energy are emitted. In gamma decay, only energy, in the form of gamma rays, is emitted.
Answer	**She had to help her mother do some housework.**	**radioactive**
w/o-UDA	what's the meaning of the name jenny?	where do gamma rays come from when they decay?
ST	what is the name of jenny's necklace?	what type of element is the source of gamma rays?
DDS+ST	**how did jenny get her necklace in the movie?**	**what type of elements give off gamma rays?**
Context	Lawmakers in the United States have expanded an investigation into the use of location-tracking systems on mobile devices. The action follows recent reports about the storing of information on the Apple iPhone. Some people consider location tracking to be a threat to personal privacy and security. Allan Friedman, the research director, says, "All wireless companies do some location tracking as part of their networks. This information is usually stored by the companies, not the devices, and there are laws to protect it. Law enforcement agencies, ... There's the idea that because it's on my phone and on my computer, rogue applications that I pay for or that I'm tricked into downloading may be able to access this data and somehow misuse it." Apple says it is "not tracking the location of your iPhone". It is simply keeping a database of Wi-Fi hotspots and cell phone towers near the user's location.	Not all wetlands are alike, as you can see below (Figure below). Wetlands vary in how wet they are and how much of the year they are soaked. Wetlands also vary in the kinds of plants that live in them. This depends mostly on the climate where the wetland is found. Types of wetlands include marshes, swamps, and bogs.
Answer	**Because it is thought to threaten users' privacy.**	**wetland**
w/o-UDA	why is there a tracking system on my phone?	what do you call a place that is covered with water?
ST	why is there a location tracking system on apple?	what do you call marshes that are wet?
DDS+ST	**why do we not use location tracking on iphone?**	**what are marshes and bogs called?**

Table 10: Examples of generated questions with different methods.