Generalised Unsupervised Domain Adaptation of Neural Machine Translation with Cross-Lingual Data Selection

Thuy-Trang Vu **Xuanli He** **Dinh Phung** **Gholamreza Haffari**
Department of Data Science and AI
Faculty of Information Technology, Monash University, Australia
{trang.vuthithuy,xuanli.he1,first.last}@monash.edu

Abstract

This paper considers the unsupervised domain adaptation problem for neural machine translation (NMT), where we assume the access to only monolingual text in either the source or target language in the new domain. We propose a cross-lingual data selection method to extract in-domain sentences in the missing language side from a large generic monolingual corpus. Our proposed method trains an adaptive layer on top of multilingual BERT by contrastive learning to align the representation between the source and target language. This then enables the transferability of the domain classifier between the languages in a zero-shot manner. Once the in-domain data is detected by the classifier, the NMT model is then adapted to the new domain by jointly learning translation and domain discrimination tasks. We evaluate our cross-lingual data selection method on NMT across five diverse domains in three language pairs, as well as a real-world scenario of translation for COVID-19. The results show that our proposed method outperforms other selection baselines up to +1.5 BLEU score.

1 Introduction

Unsupervised domain adaptation (UDA) aims to generalise MT models trained on domains with typically large-scale bilingual parallel text to new domains without parallel data (Chu and Wang, 2018). Most prior works in UDA of NMT assume the availability of either non-parallel texts of both languages or only the *target*-language monolingual text in the new domain to adapt the NMT model. The adaptation is achieved by modifying the model architecture and joint training with other auxiliary tasks (Gulcehre et al., 2015; Domhan and Hieber, 2017; Dou et al., 2019), or constructing a parallel corpus for the new domain from a general-domain parallel text using data-selection methods (Silva et al., 2018; Hu et al., 2019). However, very little attention has been paid to the UDA problem with only the *source*-language monolingual text in the new domain. In practice, this setting is not very rare, e.g. building a translation system from English to Shona (a low-resource African language) in a specific domain such as healthcare and disaster. While it would be very time consuming to collect in-domain text in Shona, English corpora are more accessible.

In this paper, we consider the *generalised* problem of UDA in NMT where we assume the availability of monotext in only one language, either the source or target, in the new domain. We propose a generalised approach to the problem using cross-lingual data selection to extract sentences in the new domain for the missing language side from a large monolingual generic corpus. Our proposed data selection method trains an adaptive layer on top of multilingual BERT by contrastive learning (Chen et al., 2020), such that the representations of source and target language are aligned. The aligned representations enable the transferability of a domain classifier trained on one language side to the other language for in-domain data detection. Previous works have explored filtering data of the same language for MT (Moore and Lewis, 2010; Axelrod et al., 2011; Duh et al., 2013; Junczys-Dowmunt, 2018); however, utilising data in one language to detect in-domain data in the other language is under-explored.

With selected sentences in the new domain of the missing language side, the original adaptation problem is transformed to the usual setting of UDA problem, and can be approached by the existing UDA methods. In this paper, we extend the discriminative domain mixing method for supervised domain adaptation (Britz et al., 2017) which jointly learns domain discrimination and translation to the unsupervised setting. More specifically, the NMT model jointly learns to translate with the translation loss on pseudo bitext, and captures the characteris-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3335–3346
November 7–11, 2021.

tics of the new domain by the domain discrimination loss on data from the old and new domains.

Our contributions can be summarised as follows:

- We introduce a generalised UDA (GUDA) problem for NMT which unifies both the usual setting of having only target language monotext and the under-explored setting with only source language monotext in the new domain.
- We propose a cross-lingual data selection method to address GUDA by retrieving in-domain sentences of the missing language from a generic monolingual corpus.
- We augment the discriminative domain mixing method to UDA by constructing an in-domain pseudo bitext via forward-translation and back-translation.
- We empirically verify the effectiveness of our approach on translation tasks across five diverse domains in three language-pairs, as well as a real-world translation scenario for COVID-19. The experimental results show that our method achieves up to +1.5 BLEU improvement over other data selection baselines. The visualisation of the representations generated by the adaptive layer demonstrates that our method is not only able to align the representation of the source and target language, but it also preserves characteristics of the domains in each space[1].

2 Generalised Unsupervised Domain Adaptation

Domain adaptation is an important problem in NMT as it is very expensive to obtain training data that are both large and relevant to all possible domains. Supervised adaptation problem requires the existence of out-of-domain (OOD) bitext and in-domain bitext. Unsupervised domain adaptation problem assumes OOD and in-domain monotext, usually in the target language.

A domain is defined as a distribution $P(X, Y)$ where X ranges over sentences in the *source* language s, and Y is its translation in the *target* language t. We define the generalised unsupervised domain adaptation (GUDA) for NMT as the problem of adapting an NMT model trained on an old domain $P_{old}(X, Y)$ to a new domain $P_{new}(X, Y)$, where only *either* the source *or* target language text is available in the new domain. Since $P(X, Y) = P^s(X)P^{s,t}(Y|X)$, let us consider $P^s_{old}(X)$ which is the distribution over sequences on the source language s in the old domain. It is usually much richer (i.e., containing diverse categories such as news, politics, etc.) than $P^s_{new}(X)$ which is typically a much more specific category where we aim to adapt the NMT model. The conditional distribution $P^{s,t}_{old}(Y|X)$ specifies the encoder-decoder NMT network to be adapted to the new domain.

Given parallel bitext $\mathcal{D}_{\text{old}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ in the old domain, we consider two settings in GUDA:

- An initial monolingual text $\mathcal{X}_{\text{new}} = \{\boldsymbol{x}_j\}$ of the source language in the new domain and a *generic* monolingual text $\mathcal{D}_t$ of the target language.
- An initial monolingual text $\mathcal{Y}_{\text{new}} = \{\boldsymbol{y}_k\}$ of the target language in the new domain and a *generic* monolingual text $\mathcal{D}_s$ of the source language.

Crucially, in both cases we **do not** require any parallel text in the new domain, hence the term *unsupervised domain adaptation*. The goal is to adapt an NMT model, parametrised by $\boldsymbol{\theta}$, trained on the old domain bitext $\mathcal{D}_{\text{old}}$ to the new domain.

In the setting involving $\mathcal{Y}_{\text{new}}$, it can be used to create pseudo-parallel data via back-translation (Sennrich et al., 2016), or to adapt the decoder via multi-task learning (Gulcehre et al., 2015; Domhan and Hieber, 2017). This setting is the usual formulation in UDA for NMT (Chu and Wang, 2018). In contrast, the setting involving the source monotext $\mathcal{X}_{\text{new}}$ is not well explored in the literature.

Our approach for addressing GUDA is to create *in-domain* monotext for the language side, where the data in the new domain is missing. That is, if given $\mathcal{X}_{\text{new}}$, we build a *classifier* to select in-domain monotext $\mathcal{Y}_{\text{new}}$ in the target language from the generic monotext $\mathcal{D}_t$. We perform a similar procedure for the other case where only in-domain $\mathcal{Y}_{\text{new}}$ is present. We then adapt the NMT model based on the bitext from the old domain as well as the source and target language monotext in the new domain. The challenge, however, is how to train a classifier for data selection for the language-side with missing data. We address this problem in Section 3, then mention how to adapt the NMT model to the new domain in Section 4.

[1] Source code is available at https://github.com/trangvu/guda.

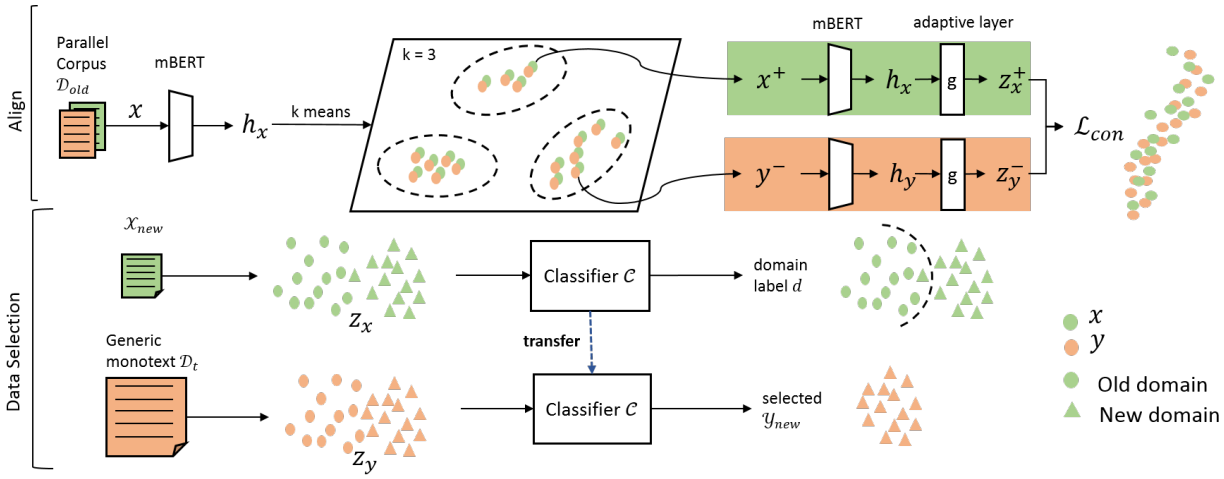

Figure 1: Our proposed cross-lingual data selection method for GUDA with source monotext $\mathcal{X}_{\text{new}}$.

3 Cross-lingual In-domain Data Selection

Aharoni and Goldberg (2020) have shown that the emergent domain clusters via BERT (Devlin et al., 2019) can be used to select in-domain *bitext* for NMT. Inspired from that observation, we leverage the sentence representations produced by the multilingual BERT (mBERT) for cross-lingual *monotext* selection. We first align the source and target language representation space while preserving the domain clustering characteristics in each space. Using the available monotext in one language, we train a binary classifier to detect old and new domains on the aligned semantic spaces. This classifier is then transferred to pick in-domain sentences in the other language (fig. 1).

Representation Alignment. We encode the representation of a sentence $\boldsymbol{x}$ by $\boldsymbol{h}(\text{mBERT}(\boldsymbol{x}))$, where $\boldsymbol{h}$ computes the mean-pooled top-layer hidden states obtained from mBERT. To align the representation space of the source and target language, we learn an *adaptive* layer $\boldsymbol{g_\phi}(.)$, a feed-forward network parametrised by $\boldsymbol{\phi}$, on top of the mBERT by contrastive learning (Chen et al., 2020). The intuition is that the representation of a translation pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ should be close to each other in the semantic space, while the representation of non-translation pairs should be far apart. Specifically, we aim to optimise a contrastive loss,

$$\mathcal{L}_{\text{con}}(\boldsymbol{z_x^+}, \boldsymbol{z_y^+}) = -\log \frac{\exp(\text{sim}(\boldsymbol{z_x^+}, \boldsymbol{z_y^+}))/\tau}{\sum \exp(\text{sim}(\boldsymbol{z_x^+}, \boldsymbol{z_y^-}))/\tau} \quad (1)$$

where $\boldsymbol{z_x} := \boldsymbol{g_\phi}(\boldsymbol{h}(\text{mBERT}(\boldsymbol{x})))$ and $\boldsymbol{z_y} := \boldsymbol{g_\phi}(\boldsymbol{h}(\text{mBERT}(\boldsymbol{y})))$ are the output of the adaptive layer for the source and target sentences; $(\boldsymbol{z_x^+}, \boldsymbol{z_y^+})$ and $(\boldsymbol{z_x^+}, \boldsymbol{z_y^-})$ denote the positive and negative example pairs, τ is a temperature parameter, and sim(.) is the cosine similarity following Aharoni and Goldberg (2020). While training $\boldsymbol{\phi}$ of the adaptive layer, other layers including embedding and transformer layers are frozen.

Given a batch of N training examples from the old domain $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^n \sim \mathcal{D}_{\text{old}}$, these translation pairs from the bitext are the positive examples. Instead of blindly treating those from non-translation pairs as negative examples, we create domain labels by clustering the mBERT representations of the bitext into k clusters. For a given $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ pair in the training batch, we consider the pairs from distinct clusters in the same batch as the negative examples. This helps the computational complexity by encoding and using all positive and negative examples in the same batch (Chen et al., 2020). We will show the benefit of this setting in § 6.1.

In-domain Data Selection. Using the adaptive layer's encoding, we learn a domain classifier for the language-side in which we are given the monotext in the new domain. Let us assume we are given source language monotext $\mathcal{X}_{\text{new}}$ in the new domain, and the bitext $\mathcal{D}_{\text{old}}$ in the old domain.[2] The domain classifier $\boldsymbol{c_\psi}(\boldsymbol{z})$ produces the probability of belonging to the new domain for an input vector $\boldsymbol{z}$. We train the parameter $\boldsymbol{\psi}$ for the domain classifier by

[2]The other case where we are given $\mathcal{Y}_{\text{new}}$ is similar, and is omitted due to the space constraints.

minimising the following loss (fig. 1),

$$\mathcal{L}_{\text{disc}}(\boldsymbol{\psi}) = -\sum_{\boldsymbol{x}\in\mathcal{X}_{\text{new}}} \log(\boldsymbol{c}_{\boldsymbol{\psi}}(\boldsymbol{g}(\boldsymbol{h}(\boldsymbol{x})))) - \sum_{\boldsymbol{x}\in\mathcal{D}^s_{\text{old}}} \log(1-\boldsymbol{c}_{\boldsymbol{\psi}}(\boldsymbol{g}(\boldsymbol{h}(\boldsymbol{x})))) \quad (2)$$

where $\mathcal{D}^s_{\text{old}}$ denotes source language side of the parallel bitext $\mathcal{D}_{\text{old}}$. Thanks to the aligned semantic spaces, we then *transfer* the trained domain classifier cross-lingually to the other language-side to select a subcorpus of in-domain monotext. We select the top-k probable sentences from the given generic corpus of the other language-side.

4 NMT Adaptation to the New Domain

Given the parallel data in the old domain $\mathcal{D}_{\text{old}}$ and monolingual data in the new domain for both the source language $\mathcal{X}_{\text{new}}$ and target language $\mathcal{Y}_{\text{new}}$, we adapt the NMT model by minimising the loss,

$$\mathcal{L} = \mathcal{L}^{s,t}_{\text{NMT}} + \mathcal{L}^{s}_{\text{disc}} + \mathcal{L}^{t}_{\text{disc}} \quad (3)$$

as illustrated in fig. 2 and explained below.

Bitext Loss. We create *pseudo*-bitext $\mathcal{D}_{\text{new}}$ by back-translating $\mathcal{Y}_{\text{new}}$ using a reverse-direction translation model trained on $\mathcal{D}_{\text{old}}$. The quality of the pseudo-bitext depends on the quality of the reverse-direction NMT model in the new domain. We further mix the pseudo-bitext $\mathcal{D}_{\text{new}}$ with the old-domain bitext $\mathcal{D}_{\text{old}}$ to form the bitext loss function

$$\mathcal{L}^{s,t}_{\text{NMT}}(\boldsymbol{\theta}) = -\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}_{\text{new}}} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) - \lambda_1 \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}_{\text{old}}} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) \quad (4)$$

where $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ is the translation probability according to the NMT model, and λ_1 controls effect of the old domain.

Source Monotext Loss. To take into account the clean text in source language of the new domain, we apply the discriminative domain mixing method (Britz et al., 2017) to force the encoder towards capturing new domain's characteristics. For this purpose, we build a classifier $\boldsymbol{c}_{\boldsymbol{\psi}_e}(\boldsymbol{z}_e)$, a feedforward network parametrised by $\boldsymbol{\psi}_e$, whose output is the new domain's probability. $\boldsymbol{z}_e = \boldsymbol{h}(\text{enc}_{\boldsymbol{\theta}}(\boldsymbol{x}))$ is the representation of the sentence $\boldsymbol{x}$, computed by the mean-pooled average of the top layer's states

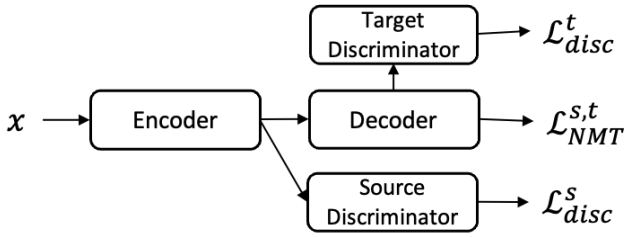

Figure 2: Discriminative domain mixing approach to UDA for NMT

of the NMT's encoder. The source monotext loss is then defined as,

$$\mathcal{L}^{s}_{\text{disc}}(\boldsymbol{\theta},\boldsymbol{\psi}_e) = -\sum_{\boldsymbol{x}\in\mathcal{X}_{\text{new}}} \log \boldsymbol{c}_{\boldsymbol{\psi}_e}(\boldsymbol{h}(\text{enc}_{\boldsymbol{\theta}}(\boldsymbol{x}))) - \lambda_2 \sum_{\boldsymbol{x}\in\mathcal{D}^s_{\text{old}}} \log(1-\boldsymbol{c}_{\boldsymbol{\psi}_e}(\boldsymbol{h}(\text{enc}_{\boldsymbol{\theta}}(\boldsymbol{x})))) \quad (5)$$

where λ_2 controls the effect of the old domain.

Target Monotext Loss. Similarly, the target monotext loss is defined as,

$$\mathcal{L}^{t}_{\text{disc}}(\boldsymbol{\theta},\boldsymbol{\psi}_d) = -\sum_{\boldsymbol{y}\in\mathcal{Y}_{\text{new}}} \log \boldsymbol{c}_{\boldsymbol{\psi}_d}(\boldsymbol{h}(\text{dec}_{\boldsymbol{\theta}}(\boldsymbol{y}))) - \lambda_3 \sum_{\boldsymbol{y}\in\mathcal{D}^t_{\text{old}}} \log(1-\boldsymbol{c}_{\boldsymbol{\psi}_d}(\boldsymbol{h}(\text{dec}_{\boldsymbol{\theta}}(\boldsymbol{y})))) \quad (6)$$

where $\text{dec}_{\boldsymbol{\theta}}$ is the NMT's decoder, $\boldsymbol{c}_{\boldsymbol{\psi}_d}$ is the domain classifier parametrised by $\boldsymbol{\psi}_d$ for the decoder, $\mathcal{D}^t_{\text{old}}$ is the target sentences in the old domain's bitext, and λ_3 controls the effect of the old domain.

5 Experiments

We evaluate our proposed approach for GUDA on the three language pairs covering five domains, and a real-world translation task, namely, TICO-19.

5.1 Setup

Datasets. Table 1 shows data statistics. The general domain datasets come from WMT2014 for English-French, WMT2020 for English-German, news parallel corpus from OPUS for Arabic-English[3]. We appraise our proposed methods on following specific domains: TED talk, Law, Medical, IT, Koran from OPUS (Tiedemann, 2012) following the recipe in Koehn and Knowles (2017). We sample 10M English sentences from Newcrawl 2007-2019 as the generic monolingual corpus. Data pre-processing is described in Appendix A.

[3]GlobalVoices, News-Commentary, UN, WikiMatrix, UNPC

Domain	Fr-En	De-En	Ar-En
NEWS	35.7M	40.5M	21.2M
LAW	625K	454K	-
MED	689K	231K	-
IT	362K	158K	246K
KORAN	128K	17.8K	183K
TED	190K	164K	199K

Table 1: Number of training sentences in the evaluation datasets. Each dataset contains 2K dev and test sentences.

Baselines. We evaluate the effectiveness of our GUDA framework over the zero-shot baseline (**base**) where the old-domain model is evaluated without any further training on the new domain. We also evaluate our method against a pseudo-translation baseline (**trans**) where the old-domain model is further trained on the pseudo-translation of monolingual data from the new domain. More specifically, the pseudo-translation training data contains sentences in the source language and its forward-translated sentences in English for to-English translation direction. Otherwise, it contains sentences in the target language and their back-translated sentences in English for from-English translation direction. We also train fully-supervised models (**sup.**) which further trains the old-domain models on in-domain parallel data and yields approximately the upperbound BLEU scores.

We compare our proposed in-domain data selection method against several baselines including,

- **random**: we randomly select English sentences from the generic monolingual pool and treat them as in-domain sentences.

- **cross entropy difference (CED)** (Moore and Lewis, 2010) which is a widely used data selection method in MT. The CED score of a given sentence x in the generic corpus is calculated as $\text{CED}(x) = H_S(x) - H_G(x)$, where $H_S(x)$ and $H_G(x)$ are the cross-entropy of the sentence x according to the specific domain and generic domain LMs respectively. The lower the CED score is, the more likely the sentence belongs to this specific domain. In our GUDA setting, to enable cross-lingual data selection, we train a multilingual neural LM on the bitext in the old domain then further finetune it on the available monotext in the new domain and use it to rank the generic corpus. We only run CED methods for En↔Fr and En↔De translation since we do not share vocabulary between Ar and En.

- **domain-finetune** (Aharoni and Goldberg, 2020) which trains a domain classifier on mBERT representations and selects the top-k in-domain sentences scored by the classifier. Despite of having similar selection mechanism to our method, the classifier in the domain-finetune technique operates on the pretrained representation space of mBERT without alignment between languages.

GUDA setup. We assume the availability of non-English language data and evaluate our method to select 500K English sentences from the generic monolingual pool. We use the multilingual DistillBERT(mDistillBERT) (Sanh et al., 2019) to encode the sentence representation. We sample and cluster 2M sentences from the old-domain bitext into k=5 clusters for negative example creation. To train the domain classifier, we extract the top 500K sentences from the old domain with low similarity scores between their representation and the mean representation of the monotext in the new domain.

The adaptive layer is a 2-layer feed-forward network with hidden size 128. We set the temperature parameter τ in the contrastive loss to 0.2. We train the adaptive layer using the Adam optimiser with learning rate 1e-5 , batch size of 64 sentences, up to 20 epochs with early stopping if there is no improvement for 5 epochs on the loss of the dev set in the old domain. The domain discriminator is also a 2-layer feed-forward network with the same hyperparameters as the adaptive layer. We use the Transformer (Vaswani et al., 2017) as NMT model and set the mixing hyperparameters $\lambda_1, \lambda_2, \lambda_3$ to 1, i.e. the old domain parallel data as well as source and target monotext contributes equally to the training signal for the NMT model. Detail of the model hyperparameters can be found in the Appendix A.

5.2 Main Results

Table 2 presents the result of translations to and from English, according to GUDA with source and target language monotext respectively. There is a significant gap between the fully supervised (sup.) and zero-shot (base) scores. It can be seen that GUDA is able to reduce this gap, especially when the in-domain data are selected intelligently.

	Fr-En					De-En					Ar-En		
	law	med	IT	Koran	TED	law	med	IT	Koran	TED	IT	Koran	TED
Translate to English													
base	42.64	37.81	28.79	7.68	34.27	39.37	37.97	35.66	14.08	36.55	15.32	1.91	21.84
trans	43.33	40.70	31.15	7.94	33.60	38.35	37.50	35.48	14.08	36.40	3.91	0.23	14.61
rand	46.33	42.08	35.48	11.15	35.48	44.32	39.81	37.53	18.52	36.92	16.52	6.43	20.35
CED	46.60	43.33	37.66	14.56	37.41	48.03	45.00	42.72	19.20	38.46	-	-	-
DF	**47.92**	44.06	38.79	16.34	37.48	49.87	45.37	42.52	**21.86**	39.22	18.98	7.74	22.39
our	47.88	**44.36**†	**38.89**	**17.25**†	**38.79**†	**51.01**†	**46.61**†	**42.73**†	21.45†	**39.34**	**20.22**†	**10.90**†	**22.71**†
sup.	49.81	53.82	63.68	19.53	41.56	61.02	53.38	43.42	20.98	40.19	41.61	17.44	36.71
Translate from English													
base	23.73	25.32	20.51	5.58	35.73	34.60	34.52	29.35	11.23	31.32	13.66	0.30	12.77
trans	33.71	28.82	35.28	12.91	35.11	35.13	38.80	31.50	12.35	32.58	12.65	0.89	12.83
rand	32.43	28.53	40.02	13.77	34.97	33.86	36.47	30.29	12.57	32.77	12.37	3.01	14.46
CED	33.19	29.14	40.82	14.04	35.86	34.81	40.62	31.02	12.52	32.83	-	-	-
DF	34.63	29.99	41.09	14.97	36.18	35.23	41.28	**31.93**	13.19	33.69	14.32	6.42	15.07
our	**35.67**†	**30.59**†	**41.48**†	**16.10**†	**37.79**†	**35.65**†	**42.67**†	31.81	**13.72**†	**33.86**†	**14.51**†	**7.99**†	**16.33**†
sup.	40.95	41.09	53.24	22.72	40.47	46.82	46.09	34.03	14.29	34.53	26.64	15.74	21.85

Table 2: BLEU score of GUDA under various selection strategies: random (*rand*), cross-entropy difference (*CED*), domain-finetune (*DF*), and our cross-lingual data selection. *base* and *sup.* are the scores of zero-shot and fully supervised on in-domain parallel data. *trans* is the NMT model trained on pseudo bitext where monolingual in-domain data is machine translated in the missing side. Highest scores of GUDA are marked in **bold**. † indicates that our method is statistically significant difference to the domain-finetune baseline (p-value ≤ 0.05).

	Fr-En	En-Fr	Ar-En	En-Ar
base	32.35	25.07	34.11	24.52
rand	30.59	24.61	32.30	24.20
CED	32.55	25.13	-	-
DF	33.25	26.24	34.56	25.36
Our	**34.17**†	**27.45**†	**35.24**†	**26.10**†

Table 3: Results on TICO-19 translation task. † indicates that our method is statistically significant difference to the domain-finetune baseline (p-value ≤ 0.05).

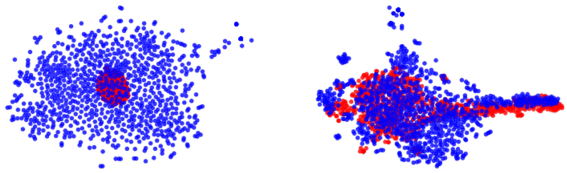

Figure 3: t-SNE visualisation of the Fr (red) and En (blue) of TICO-19 dev set, encoded by multilingual DistillBERT (left) and the adaptive layer (right).

Overall, our selection method consistently outperforms both the domain-finetune and CED strategy.

We further assess our approach on the translation initiative for COVID-19 task (TICO-19) for En-Fr and En-Ar (Anastasopoulos et al., 2020). The task contains a dev set and a test set of 971 and 2100 sentences. As an emerging domain, there is no training set. We collect additional 49K and 17K in-domain French and Arabic monotext[4]. As shown in Table 3, surprisingly, GUDA on random selection deteriorates the BLEU score. It is possible that pandemic related words have not appeared often before. Consistent with previous results, our method outperforms other methods up to +1.2 BLEU score.

To evaluate our alignment method, we visualise the representation of the TICO-19 dev set produced by mDistillBERT and the adaptive layer in Figure 3. It can be seen that the adapted French and English representations are better aligned in the semantic space than the mDistillBERT.

6 Ablation and Analysis

6.1 Ablation

Clustering-based negative sampling. The intuition of the clustering-based negative sampling is to preserve the domain clustering characteristics emerged in mBERT. We assess the importance of this clustering step and the effect of the number of

[4] https://github.com/neulab/covid19-datashare

k	De-En			En-De		
	law	med	TED	law	med	TED
1	50.76	46.32	38.61	34.98	41.45	31.92
2	50.60	46.44	38.43	35.14	41.71	33.01
3	50.86	46.62	39.20	**35.77**	42.34	33.51
5	51.01	46.61	39.34	35.65	**42.67**	**33.86**
7	**51.04**	46.62	**39.67**	35.02	42.05	33.06
10	50.81	**46.70**	39.39	35.21	42.24	35.35

Table 4: Cluster-based negative sampling ablation. k is the number of clusters.

	De-En			En-De		
	law	med	TED	law	med	TED
sup.	61.02	53.38	40.19	46.82	46.09	34.53
True Bitext						
BI	53.59	49.35	40.01	45.69	43.10	30.78
BI+S	54.69	**51.47**	40.33	46.98	45.59	**31.83**
BI+T	54.70	51.31	40.28	46.84	45.62	31.63
BI+S+T	**54.73**	51.38	**40.38**	**47.11**	**45.79**	31.67
Pseudo Bitext - Warm Start						
BI	48.57	45.62	38.61	35.04	39.98	31.99
BI+S	50.65	46.50	39.05	35.27	41.97	33.33
BI+T	50.22	46.27	38.88	35.16	40.68	33.16
BI+S+T	**51.01**	**46.61**	**39.34**	**35.65**	**42.67**	**33.86**
Pseudo Bitext - Cold Start						
BI	29.33	35.02	30.83	30.98	35.28	28.19
BI+S	35.39	37.28	33.02	32.03	**37.81**	30.35
BI+T	35.68	37.13	**33.49**	32.37	37.50	**30.80**
BI+S+T	**36.07**	**37.78**	33.47	**33.24**	37.73	30.47

Table 5: Domain discriminative mixing ablation

cluster k on the En↔De translation performance in law, med and TED domains. Table 4 reports the BLEU score of the NMT model in the new domain with $k = \{1, 2, 3, 5, 7, 10\}$ where $k = 1$ corresponds to perform negative sampling without pre-clustering mBERT representation space. Overall, the NMT model trained on the selected data with clustering-based negative sampling $k > 1$ outperforms the one without clustering $k = 1$. On the other hand, the effect of number clusters k varies, depending on the domains and languages. From the empirical results, we found that $k = \{5, 7\}$ works better than other values.

Discriminative domain mixing. We run ablation experiments to verify the contribution of each loss term in the discriminative domain mixing training objective presented in eq. (3). Particularly, we evaluate the NMT adapted to the new domain using (i) only the bitext loss (**BI**); (ii) the combination of the bitext loss and either the source monotext loss (**BI+S**) or the target monotext loss (**BI+T**); and (iii) the joint of all three loss terms (**BI+S+T**). Table 5 shows the results under both supervised domain adaptation where we have access to the true bitext, and UDA in which the model is trained on pseudo bitext generated by back-translation (warm-start). The size of the ground-truth bitext is shown in Table 1. The size of the pseudo-bitext is 500K which is approximately double the size of the ground-truth bitext of TED and med domains, and roughly the same for law domain. We also further evaluate the contribution of the discriminative domain loss when the NMT model is trained from scratch (cold-start).

Consistent with Britz et al. (2017), training NMT on mixed domain data (BI) degrades performance versus models fit to a single domain (sup.). Adding the discriminative domain loss can mitigate this negative effect in multi-domain NMT. We observe similar outcomes in both domain adaption with the true bitext and the pseudo bitext. Overall, we found that the source monotext loss plays a more critical role than the target monotext loss. Combining both monotext loss achieves the best BLEU score in most of domain adaptation scenarios.

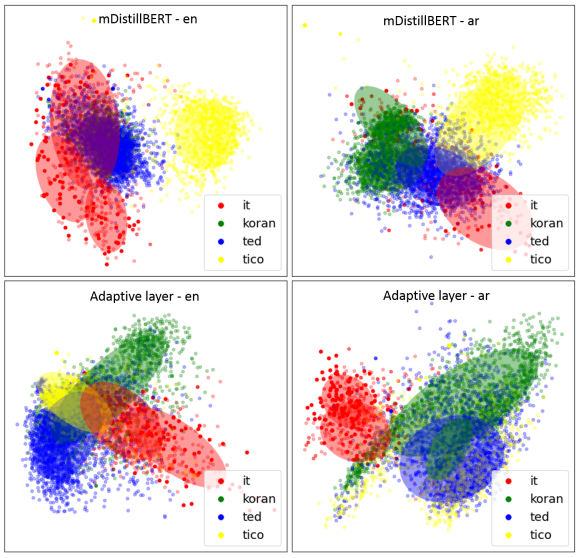

Figure 4: 2D visualisation of the unsupervised GMM-based clustering of En-Ar representations.

6.2 Analysis

Domain cluster visualisation. To demonstrate the ability of our approach in preserving the domain clustered characteristics of mBERT, we plot 2D visualisation of the mean-pooling BERT hidden

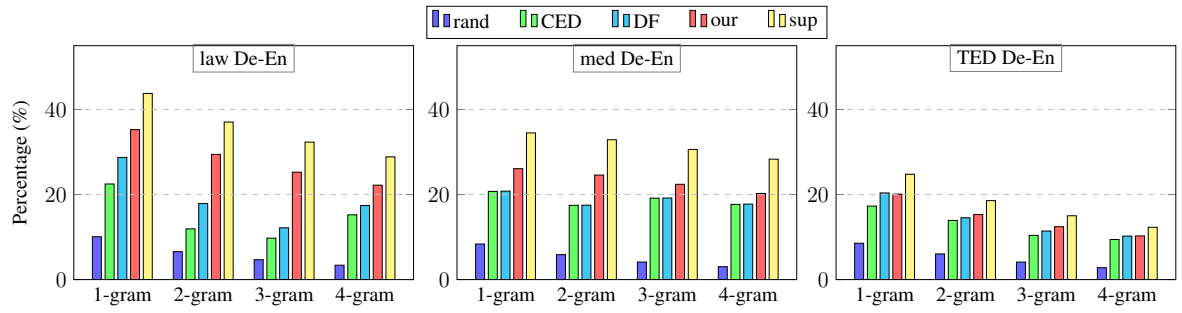

Figure 5: Percentage of newly introduced correct ngram.

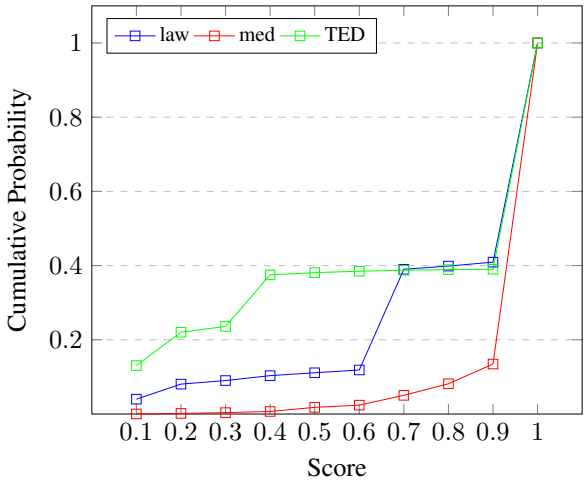

Figure 6: CDF of predicted score produced by the domain classifier. The smaller the score is, the higher probability the sentence belongs to the new domain.

state sentence representation and our constrastive-based sentence representations using PCA. Following Aharoni and Goldberg (2020), we combined the development set of all the new domain dataset and cluster the representations using a Gaussian Mixture Model (GMM) with k pre-defined clusters where k is number of domains.

Figure 4 visualises the obtained clusters in semantic space of mDistillBERT and the adaptive layer for each language in the translation pairs. The ellipses describe the mean and variance parameters learned for each cluster. In line with the finding in Aharoni and Goldberg (2020), the mDistillBERT representation of English sentences can be clustered by their domains with a small overlap region. In contrast, Arabic sentences are not well-clustered according to their domains where their domain clusters exhibit a high overlap rate. As can be seen, our contrastive-based representation alignment method is not only able to preserve the domain clusters in English sentences but also learn domain clustered representations of Arabic sentences in which the clusters are less overlapped.

Distribution of domain predictive score. Figure 6 plots the cumulative distribution for the domain predictive score over the generic English corpus. It can be seen that only a small portion of the generic corpus are predicted to belong to the new domains. As expected, the more specific-domains such as med and law have smaller number of anticipated sentences than the TED domain.

ngram analysis. A domain can be considerred as a distribution over ngram. The data selection methods mitigate the domain shift in NMT by introducing ngrams of the new domain to the training corpus. We estimate the new in-domain ngram contribution of each selection method by calculating the overlap of ngrams in the translation hypothesis and the translation reference. The new ngram contribution is calculated as

$$\frac{\sum_i \mathcal{G}(\tilde{y}^{\text{GUDA}}_{i,\text{new}}) \cap (\mathcal{G}(y^{\text{ref}}_{i,\text{new}}) \backslash \mathcal{G}(\tilde{y}^{\text{zero}}_{i,\text{new}}))}{\sum_i \mathcal{G}(y^{\text{ref}}_{i,\text{new}}) \backslash \mathcal{G}(\tilde{y}^{\text{zero}}_{i,\text{new}})} \quad (7)$$

where $\mathcal{G}(y^{\text{ref}}_{i,\text{new}}), \mathcal{G}(\tilde{y}^{\text{zero}}_{i,\text{new}}), \mathcal{G}(\tilde{y}^{\text{GUDA}}_{i,\text{new}})$ are the set of ngrams in the reference, the zero-shot and the GUDA translation hypothesis of the sentence i in the test set in the new domain, respectively.

Figure 5 presents the percentage of new ngram contribution, $1 \leq n \leq 4$, of each data selection methods as well as the fully supervised model for De-En translation in law, med, ted domains. As expected, the fully-supervised model has the highest correct in-domain ngram rate to the translation hypothesis. Our proposed selection method contributes a higher percentage of in-domain ngrams than other selection methods in all domains.

7 Related Works

Unsupervised Domain Adaptation. Previous works in UDA has been focused on aligning domain distribution by minimising the discrepancy

between representations of source and target domains (Shen et al., 2018; Wang et al., 2018); learning domain-invariant representation via adversarial learning (Ganin and Lempitsky, 2015; Shah et al., 2018; Moghimifar et al., 2020); bridging the domain gap by adaptive pretraining of contextualised word embeddings (Han and Eisenstein, 2019; Vu et al., 2020). In this paper, we adapt the NMT model from the old to new domain by learning domain-invariant representations of both encoder and decoder via domain discrimination loss.

Unsupervised Domain Adaptation of NMT. There are two main approaches in UDA for NMT, including model-centric and data-centric methods (Chu and Wang, 2018). In the model-centric approach, the model architecture is modified and jointly trained on MT tasks, and other auxiliary tasks such as language modelling (Gulcehre et al., 2015). On the other hand, the data-centric methods focus on constructing in-domain parallel corpus by data-selection from general corpus (Domhan and Hieber, 2017), and back-translation (Jin et al., 2020; Mahdieh et al., 2020). Most prior works in UDA of NMT often assume the availability of in-domain data in the target language. While there are few studies on the UDA problem with in-domain source-language data in statistical MT (Mansour and Ney, 2014; Cuong et al., 2016), this problem remains unexplored in NMT.

Data selection for NMT. To address the scarcity problem of MT parallel data in specific-domain, data selection methods utilise an initial in-domain training data to select relevant additional sentences from a generic parallel corpus. Previous research has used n-gram language model (Moore and Lewis, 2010; Axelrod et al., 2011; Duh et al., 2013), count-based methods (Way et al., 2018; Parcheta et al., 2018), similarity score of sentence embeddings (Wang et al., 2017; Junczys-Dowmunt, 2018; Dou et al., 2020) to rank the generic corpus. The ranking and selection process often operate in the same language, either source or target language, and take advantage of the parallel corpus to retrieve the paired translation (Farajian et al., 2017). When such generic parallel corpus is unavailable, cross-lingual data selection which uses data in one language to detect in-domain data in the other language is under-explored.

8 Conclusion

We have proposed a cross-lingual data selection method to the GUDA problem for NMT where only monolingual data from one language side is available in the new domain. We first learn an adaptive layer to align the BERT representation of the source and target languages. We then utilise a domain classifier trained on one language to select in-domain data for another. Experiments on translation tasks of several language pairs and domains show the effectiveness of our method over other baselines.

Acknowledgments

This material is based on research sponsored by the ARC Future Fellowship FT190100039; the Air Force Research Laboratory and DARPA under agreement number FA8750-19-2-0501. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The authors are grateful to the anonymous reviewers for their helpful comments. The computational resources of this work are supported by the Multimodal Australian ScienceS Imaging and Visualisation Environment (MASSIVE)[5].

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of ACL*.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation initiative for COvid-19. arXiv:2007.01788.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP*.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICLR*.

[5] www.massive.org.au

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of COLING*.

Hoang Cuong, Khalil Sima'an, and Ivan Titov. 2016. Adapting to all domains at once: Rewarding domain invariance in SMT. *TACL*, 4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HTL*.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of EMNLP*.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. *arXiv preprint arXiv:2004.03672*.

Zi-Yi Dou, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. 2019. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings. In *Proceedings of EMNLP-IJCNLP*.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of ACL*.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the EMNLP-IJCNLP*.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of ACL*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. A simple baseline to semi-supervised domain adaptation for machine translation. *arXiv e-prints*, pages arXiv–2001.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL: Demonstrations.*

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of WMT*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the ACL*.

Mahdis Mahdieh, Mia Xu Chen, Yuan Cao, and Orhan Firat. 2020. Rapid domain adaptation for machine translation with monolingual data. *arXiv preprint arXiv:2010.12652*.

Saab Mansour and Hermann Ney. 2014. Unsupervised adaptation for statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Farhad Moghimifar, Gholamreza Haffari, and Mahsa Baktashmotlagh. 2020. Domain adaptative causality encoder. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of ACL*.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of LREC*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the NAACL: Demonstrations*.

Zuzanna Parcheta, Germán Sanchis-Trilles, and Francisco Casacuberta. 2018. Data selection for nmt using infrequent n-gram recovery.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of ACL*.

Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial domain adaptation for duplicate question detection. In *Proceedings of the EMNLP*.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Catarina Cruz Silva, Chao-Hong Liu, Alberto Poncelas, and Andy Way. 2018. Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of LREC*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Thuy-Trang Vu, Dinh Phung, and Gholamreza Haffari. 2020. Effective unsupervised domain adaptation with adversarially trained language models. In *Proceedings of the EMNLP*.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of ACL*.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the NAACL*.

Andy Way, Alberto Poncelas, and Gideon Maillette de Buy Wenniger. 2018. Data selection with feature decay algorithms using an approximated target side. IWSLT.

A Training Procedure

Data preprocessing. We tokenise English, French, German sentences using Moses tokenizer (Koehn et al., 2007) and remove the sentences with more than 175 tokens. Arabic text are tokenised using CAMeL (Obeid et al., 2020). For Arabic, we first filter out the sentences containing more than 50% Latin characters, then remove those with more than 175 tokens.

Model hyperparameters. The adaptive layer is a 2-layer feed-forward net with hidden size 128. We set the temperature parameter τ in the contrastive loss to 0.2. We train the adaptive layer using the Adam optimiser with learning rate 1e-5, batch size of 64 sentences, up to 20 epochs with early stopping if there is no improvement for 5 epochs on the loss of the dev set in the old domain. The domain discriminator is also a 2-layer feed-forward net. We train it with the same hyperparameters as in the adaptive layer.

We use the Transformer as NMT model, which consists of 6 encoder and decoder layers, 4 self-attention heads, hidden size of 256, feed-forward hidden size of 1024, implemented in Fairseq framework (Ott et al., 2019). Number of parameters is 64.3M. We use the Adam optimiser with learning rate 5e-4 (Kingma and Ba, 2015) and an inverse square root schedule with warm-up 1000 steps. We apply dropout and label smoothing with a rate of 0.3 and 0.1 respectively. We learn the vocabulary of size 32000 using unigram language model (Kudo, 2018), implemented in SentencePiece[6]. For En-Fr, En-De, and En-Cs, the source and target embeddings are shared and tied with the last layer. We set the mixing hyperparameters $\lambda_1, \lambda_2, \lambda_3$ to 1, i.e. the old domain parallel data as well as source and target monotext contributes equally to the training signal for the NMT model. We train the NMT with the batch size of 32768 tokens and up to 30 epochs with early stopping if there is no improvement on dev set for 5 epochs.

Our model is trained on a V100 GPU, and took up to 4 days for the NMT trained in old domain, and 1 day for other experiments.

[6] https://github.com/google/sentencepiece