Strategies for Adapting Multilingual Pre-training for Domain-Specific Machine Translation

Neha Verma nverma7@jhu.edu
Kenton Murray kenton@jhu.edu
Kevin Duh kevinduh@cs.jhu.edu
Johns Hopkins University, Baltimore, USA

Abstract

Pretrained multilingual sequence-to-sequence models have been successful in improving translation performance for mid- and lower-resourced languages. However, it is unclear if these models are helpful in the domain adaptation setting, and if so, how to best adapt them to both the domain *and* translation language pair. Therefore, in this work, we propose two major fine-tuning strategies: our *language-first* approach first learns the translation language pair via general bitext, followed by the domain via in-domain bitext, and our *domain-first* approach first learns the domain via multilingual in-domain bitext, followed by the language pair via language pair-specific in-domain bitext. We test our approach on 3 domains at different levels of data availability, and 5 language pairs. We find that models using an mBART initialization generally outperform those using a random Transformer initialization. This holds for languages even outside of mBART's pretraining set, and can result in improvements of over +10 BLEU. Additionally, we find that via our domain-first approach, fine-tuning across multilingual in-domain corpora can lead to stark improvements in domain adaptation without sourcing additional out-of-domain bitext. In larger domain availability settings, our domain-first approach can be competitive with our language-first approach, even when using over 50X less bitext.

1 Introduction

Recent pretrained multilingual sequence-to-sequence (seq2seq) models have provided a basis to easily create neural machine translation (MT) systems via the pretrain and fine-tune paradigm ubiquitous throughout NLP (Liu et al., 2020; Xue et al., 2021). Due to the fact that fine-tuning these models generally requires less data than is needed for from-scratch translation models, pretrained models are great candidates for MT domain adaptation tasks, where domain-specific bitext is generally less available as compared to general bitext. However, these models have seldom been studied in domain-specific settings.

For MT domain adaptation, pretrained multilingual seq2seq models must be adapted to both 1) the language pair and 2) the domain of interest. Previous work has introduced several methods for adapting general translation models to domains, including training first on general bitext to bolster the total amount of bitext available, followed by training on smaller domain-specific bitext (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016). However, in the case of multilingual sequence models, the initial pretraining objective differs significantly from the task of machine translation, which suggests that alternative adaptation approaches are necessary. Additionally, approaches involving additional general bitext may not even benefit these models as they were already initially trained on large amounts of general pretraining data.

Therefore, it is currently unclear 1) if pretrained multilingual seq2seq models have useful properties for MT domain adaptation and 2) how to best adapt them to both the translation language pair *and* domain. As a result, in this work, we aim to systematically compare fine-tuning approaches for applying mBART to domain adaptation (Liu et al., 2020). We choose to focus on mBART as it has previously shown the most promising results in the MT setting among comparable models (Liu et al., 2021; Lee et al., 2022).

By framing language pair and domain as decoupled entities to learn during the fine-tuning process, we can compare two major approaches to the adaptation process. The first fine-tunes mBART on general bitext, followed by in-domain bitext. The second uses multilingual fine-tuning on in-domain bitext across several language pairs followed by bilingual fine-tuning on in-domain bitext in the language pair of interest (Tang et al., 2020). In other words, the first approach adapts to the *language pair* first, and the second approach adapts to the *domain* first, before both eventually fine-tune on the small amount of in-domain, language pair-specific bitext. We emphasize the importance of a multi-staged approach as we find that they are consistently better than naively fine-tuning mBART only on domain-specific bitext, especially when this data is limited.

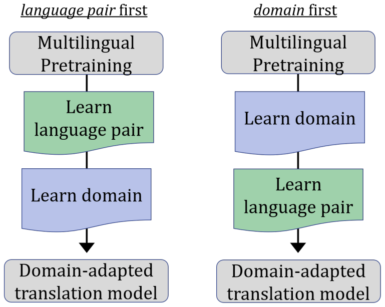

Figure 1: Summaries of our two major approaches. Our language pair first approach first fine-tunes the multilingual pretrained model on language-pair specific translation, and then on the domain. Our domain first approach first fine-tunes the multilingual pretrained model on the domain of interest, followed by the specific language pair.

We test our approach on 5 language pairs and 3 domains: TED Talks, Microblogs, and COVID-19 related information. We note that the amount of available in-domain bitext varies greatly across these domains. Because we want our method to be broadly applicable to new domains where data may be very limited and/or expensive to procure, we test our approaches on a small, fixed amount of domain data, as well as on the entirety of the domain data available. We find that in comparable approaches, those with an mBART initialization outperform those with a vanilla Transformer initialization in a majority of our language pairs and domains, and across our two domain availability settings. This holds even in the cases of higher-resourced language pairs originally unhelped by mBART's multilingual pretraining, and in language pairs outside of mBART's pretraining set. For our out-of-mBART Persian-English language pair, simply using an mBART initialization leads to +4.8 to 12.8 BLEU points across our domains. In addition, we find that our domain-first approach provides an efficient alternative to using additional general bitext by leveraging available multilingual in-domain corpora via multilingual fine-tuning. We show that in our whole domain availability setting, which is still several times smaller than the data needed for our language-first approach, our domain-first approach consistently shows improvements over baselines, and is sometimes competitive with our more data-heavy language-first approach.

This paper makes the following contributions:

- We explore various approaches for fine-tuning multilingual sequence models for specialized domains in machine translation. We demonstrate that our multi-step fine-tuning approaches can out-perform single-step and non-pretrained baselines—even for language pairs that normally do not see benefits from using multilingual models.
- We demonstrate the importance of in-domain data, showing that fine-tuning on this with multiple languages outperforms methods only using in-domain data in the target setting.
- We are able to get substantial BLEU point improvements on languages that are *not even included* during pretraining.

2 Background and Related Work

2.1 mBART

mBART is a pretrained multilingual sequence-to-sequence denoising autoencoder based on the Transformer architecture (Liu et al., 2020). Using the self-supervised BART objectives of masked language modeling and sentence permutations (Lewis et al., 2020), mBART is trained to recover noised Common Crawl texts across 25 languages. When fine-tuned on bitext for sentence-level machine translation, mBART's pretraining leads to performance gains across multiple low- and medium-resource language pairs. The shared, multilingual parameter space in the single encoder-decoder model are thought to be particularly helpful in lower-resourced language pairs. In the original release of mBART, fine-tuning on a single language pair, or bilingual fine-tuning, was the proposed method of adapting mBART to the translation task.

However, in follow up work, Tang et al. (2020) propose multilingual fine-tuning, where mBART is fine-tuned on bitext across multiple language pairs at the same time, creating a model capable of multilingual machine translation. Multilingual fine-tuning was shown to result in improvements over bilingual fine-tuning for translation, especially in the many-to-one setting where multiple languages are translated into the same target language.

2.2 Domain Adaptation

Generally, MT systems drop in performance when applied in a domain different from the training data, in a scenario known as domain mismatch (Koehn and Knowles, 2017). Additionally, while large amounts of general bitext may be available for a language pair, it is generally harder to find large amounts of data that fit a specific domain.

Continued training, or fine tuning, is a common training procedure-related approach to MT domain adaptation where a model first trains to convergence on general bitext, and then continues to train on domain-specific bitext (Luong and Manning, 2015). In this work, we expand upon the original multi-step fine-tuning ideas from continued training for domain adaptation.

Later work has focused on more complex ways to select and order data for domain adaptation. Xu et al. (2021) propose gradual fine-tuning for iteratively training a model on data that slowly approaches the distribution of the in-domain data. Xie et al. (2021) also use gradual fine-tuning to select data to adapt a multilingual MT model to in-domain data. Similar to gradual fine-tuning with respect to purposefully ordering samples for domain adaptation, curriculum learning based approaches have been proposed to sort and order samples based on their similarity to the domain of interest (Zhang et al., 2019). Dynamic data selection techniques have also been proposed to alter available training data between epochs in order to present more relevant data in later stages of training (van der Wees et al., 2017). While these methods enforce a stricter curriculum at a sample level, we draw inspiration from these methods by adhering to a coarse ordering of least-domain-relevant to most-domain-relevant (Saunders, 2021).

Recent work has introduced domain adaptation techniques for multilingual MT systems. One such work proposes methods for multilingual and multi-domain adaptation via domain-specific and language-specific adapter modules (Cooper Stickland et al., 2021). Dabre et al. (2019) exploit multi-parallel domain corpora in one-to-many multilingual MT setup to boost low-resource domain translation. Another closely related work, which specifically looks at the use of mBART for poetry translation, introduces multilingual fine-tuning for domain adaptation using mBART50 (Chakrabarty et al., 2021; Tang et al., 2020). In this domain, multilingual fine-tuning on available domain data was shown to outperform multilingual fine-tuning on general bitext, as well as bilingual fine-tuning on domain data, hinting at the use of multilingual in-domain data as an important tool in multilingual MT domain adaptation. In a comprehensive overview of the capabilities of mBART, Lee et al. (2022) find that mBART fine-tuned on smaller amounts of in-domain bitext can outperform a Transformer translation model trained on larger amounts of in-domain bitext, suggesting that mBART's pretraining may be valuable for domain adaptation in lower-resourced domains. In this work, we expand upon and formalize these initial results suggesting that mBART may be useful for domain adaptation, and provide a comparison of various techniques for pretrained model-specific domain adaptation.

3 Approach

Because mBART is a multilingual denoising autoencoder, it is trained only to reconstruct text in the source language given. We detail our two major approaches to domain adaptation using mBART, focusing on translation language pair and domain. Our approaches propose two different ways to learn these competencies. We summarize our approaches in Figures 2 and 3.

3.1 Language Pair First

In our language pair first approach, we first focus on adapting mBART to the specific language pair, and then to the domain of interest. We note that in all of our experiments, we have several source languages, and one target language. For our i^{th} source language S_i and target language T, we label general bitext as $B_{gen}(S_i, T)$, and in-domain bitext as $B_{in}(S_i, T)$. In the first stage, we fine-tune our original model, M_0 on $B_{gen}(S_i, T)$ to achieve a general-domain bilingual translation model, denoted as $M_{gen}(S_i, T)$. In the second stage, we fine-tune $M_{gen}(S_i, T)$ on $B_{in}(S_i, T)$, obtaining our final domain adapted bilingual model, $M_{in}(S_i, T)$, as desired. This approach is very similar to the conventional continued training approach for domain adaptation where a MT model is trained on out-of-domain bitext, and then subsequently fine tuned on the smaller in-domain bitext. The key difference between this approach and a general continued training approach for domain adaptation is the initialization of model parameters that mBART's pretraining provides.

Symbol	Reference
S_i	ith source language
T	target language
$B_{gen}(S_i, T)$	general bitext from S_i to T
$B_{in}(S_i, T)$	in domain bitext from S_i to T
M_0	original model, before fine-tuning
$M_{gen}(S_i, T)$	general domain translation model from S_i to T
$M_{in}(S_i, T)$	in-domain translation model from S_i to T

3.2 Domain First

In our domain-first approach, we first focus on adapting mBART to the domain of interest, and then adapt to the relevant language pair. Because at first we adapt only to the domain, and not yet language pair, we propose to perform the first stage via multilingual fine-tuning on

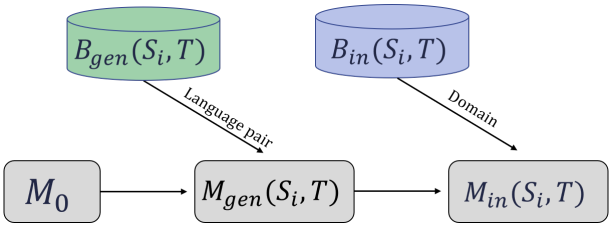

Figure 2: Our language pair first approach. We first fine-tune our original model, M_0 on general bitext, $B_{gen}(S_i, T)$, to create a language-pair adapted model, $M_{gen}(S_i, T)$. We then fine-tune our new interim model on in-domain bitext, $B_{in}(S_i, T)$, to achieve a both language pair- and domain-adapted translation model: $M_{in}(S_i, T)$.

available domain data (Tang et al., 2020). In particular, we focus on many-to-one fine-tuning. In this case, we denote a multilingual in-domain dataset as the union of all available bilingual in-domain datasets: $\bigcup_i B_{in}(S_i, T)$.

In the first stage, we multilingually fine-tune our original model, M_0 on $\bigcup_i B_{in}(S_i, T)$ to achieve a domain-specific and multilingual translation model, denoted as $M_{in}(\bigcup_i S_i, T)$. In our second step, we reintroduce $B_{in}(S_i, T)$ that matches our language pair of interest, and bilingually fine-tune $M_{in}(\bigcup_i S_i, T)$ on $B_{in}(S_i, T)$ to achieve our final domain-adapted bilingual model, $M_{in}(S_i, T)$. Because this approach only uses in-domain data and does not introduce external data, its training can use noticeably less data for domains with limited data, as compared to our language-first approach.

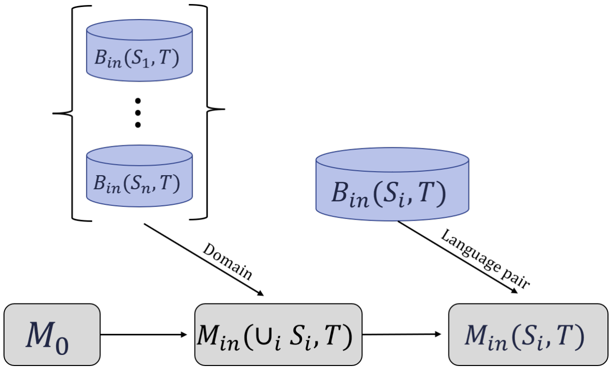

Figure 3: Our domain first approach. We first fine-tune our original model, M_0 on multilingual in-domain bitext, $\bigcup_i B_{in}(S_i, T)$, to create a domain adapted model, $M_{in}(\bigcup_i S_i, T)$. We then fine-tune our new interim model on translation pair-specific in-domain bitext, $B_{in}(S_i, T)$, to achieve a both domain- and language pair-adapted translation model: $M_{in}(S_i, T)$. We note that this approach can use far less data than our language-first approach.

3.3 Limiting the Amount of Domain Data

In comparing our approaches in adapting mBART for domain-specific MT, we also compare two scenarios in which 1) all available domain data is included, the size of which can very greatly by domain, and 2) domain data is heavily limited ($\leq$ 1000 lines). In our first scenario, where all domain data is used, we wish to provide comparisons of our domain adaptation techniques at original levels of domain availability. By keeping all data, we can make recommendations for domains that may be more available than in our limited setting. Our second scenario aims to compare our methods across each domains via a fixed amount of data, as some of our domain data is very limited ($\leq$ 1000 lines). This is the case of the Translation Initiative for COVID-19 challenge, where domain-specific translation is needed to quickly translate emergency content related to the COVID-19 pandemic (Anastasopoulos et al., 2020). Additionally, being able to create data-efficient methods helps reduce cost of dataset creation. For example, current translation services are priced around 0.06 to 0.12 US Dollars per word[1], and Germann (2001) note services costing up to 0.30 US Dollars per word. Assuming an average of 20 words per sentence, it can cost anywhere from approximately \$1,200 to \$6,000 to create a small 1000 line dataset. By including these two levels of domain availability, we hope to show the efficiency of our methods, as well as their generalizability to additional domains.

4 Experiments

4.1 Data

We translate 5 languages into English for our experiments: Arabic (ar), Persian (fa), Portuguese (pt), Russian (ru), and Chinese (zh). Arabic, Russian, and Chinese appear in mBART25's pretraining set, and Portuguese and Persian do not.

For each of our language pairs, we piece together general bitext from OPUS sources (Tiedemann, 2012). The general bitext make up part of our language-first adaptation approach. In particular, depending on availability, we sample bitext from Global Voices (GV), QCRI Educational Domain (QED), the United Nations Parallel Corpus (UN), Open Subtitles (OS), and Europarl (EP) (Nguyen and Daumé III, 2019; Abdelali et al., 2014; Ziemski et al., 2016; Lison and Tiedemann, 2016; Koehn, 2005). We first collect 1.5 million lines from these combined sources. We then remove sentences with more than 50% punctuation, deduplicate our data, remove all evaluation data from training data, and apply length ratio cleaning (Fan et al., 2021). We shuffle all lines and sample 1 million sentence pairs for a general bitext training set, and 2000 for a development set. The full composition of our general bitext is detailed in Table 1.

For domain adaptation, we choose 3 different domains with varying levels of data availability. We use translations of TED talks (Duh, 2018), Microblogs (McNamee and Duh, 2022), and documents from the Translation Initiative for COVID-19 (TICO-19) (Anastasopoulos et al., 2020). We note that originally, the TICO-19 dataset contains only 971 sentences in a development set, and 2100 in a test set. In our work, we split the original test set to create a new development and test set with 1050 lines each, and reallocate the original 971-line development set into our training set.[2] For each domain, we detail the amount available training data in Table 1. TED dev/test splits are 1958/1982 lines, and Microblog dev/test splits are 3000/3000 lines for ar-en and ru-en, and 2000/2000 lines for fa-en, pt-en, and zh-en.

For our multilingual fine-tuning experiments for our domain-first approach, we include additional languages that are available in the domain, included in mBART's pretraining set, and do not overlap with our language proxies for our out-of-mBART languages. For TED, we add Czech, German, French, Japanese, Korean, Romanian, and Vietnamese (12 languages total). For Microblogs, we add French and Korean (7 languages total). For TICO-19, we add French, Burmese, and Nepali (8 languages total).

To measure the amount of domain shift between our general bitext and our domain-specific bitext, we train a 5-gram language model with KenLM on our general bitext target-side training data, and evaluate the perplexity (including OOVs) on the target-side training data for each of our domains. We provide perplexity measures on our domain-specific bitext after applying

[1] https://gengo.com/pricing-languages/

[2] We create our own data split because the original TICO-19 data does not have a training set we could use for fine-tuning. Our TICO-19 results should not be directly compared with those from other papers.

	TED	Micro.	TICO	Gen.	GV	QED	UN	OS	EP
	# lines	# lines	# lines	# lines	%	%	%	%	%
ar-en	175377	18634	971	1M	3.5	33.4	31.6	31.5	0
fa-en	116525	2647	971	1M	0.7	1.1	0	98.2	0
pt-en	153357	2085	971	1M	5.8	28.8	0	32.7	32.7
ru-en	181465	36734	971	1M	11.4	37.6	25.5	25.5	0
zh-en	170341	1580	971	1M	8.5	0.9	45.3	45.3	0
add'l.	988691	35710	2991	-	-	-	-	-	-
total	1785756	97390	7976	-	-	-	-	-	-

Table 1: Sizes in # of lines for each of the domain and general corpora used in our work. We also provide the number of lines added with domain data from additional language pairs. We additionally provide a breakdown of our general bitext across 5 OPUS sources. For our limited domain experiments, we use 1K sentences per domain and language pair.

byte-pair encoding (Sennrich et al., 2016), and we measure vocabulary coverage on our data that is tokenized with the Moses tokenizer, but not byte-pair encoded (Koehn et al., 2007). We report vocabulary coverage and perplexity values in Table 2.

	TED	Micro.	TICO
vocab coverage	99.9%	95.7%	97.2%
perplexity	2.65	740.53	366.67

Table 2: Vocabulary coverage and perplexity for each of our domains. We train 5-gram language models on our general domain target data, and evaluate vocabulary coverage and perplexity on our domain target-side training data. We see that our Microblogs corpus has the largest domain shift while TED has the smallest, according to our perplexity measure.

4.2 Models

For all of our experiments using mBART, we use `mbart.cc25` which has 12 encoder and decoder layers, and covers 25 languages. We note that to begin decoding, mBART requires a language identification token. For our out-of-mBART languages, we choose a related language from the 25 mBART pretraining languages as a language identification token; we use ES as a proxy for PT, and HI for FA (Madaan et al., 2020; Cahyawijaya et al., 2021).

We train our language pair-first approach on 1M lines of general data for up to 10 epochs or 150,000 updates, whichever is first. We then fine-tune the model for up to an additional 10 epochs on domain data for the language pair, for both limited and whole domain availability. For our domain-first approach, we train on multilingual domain data for up to 200,000 updates or 60 epochs (whichever is first) in the whole domain approach, and up to 60 epochs in the limited domain approach. We then fine-tune these models for up to another 60 epochs.

We include two baseline models in our experiments. Baseline model 1 uses a Transformer with no pretraining, and trains on general bitext followed by domain bitext, much like our language first approach. This model uses the transformer_iwslt_de_en architecture as implemented by fairseq (Ott et al., 2019; Vaswani et al., 2017). This model has an embedding dimension of 512, feed-forward dimension of 1024, 4 attention heads, and 6 encoder/decoder layers each. For each language pair, we learn 16k subword operations per language on the general domain bitext, and use the subword vocabulary on our all of our Baseline 1 experiments (Sennrich et al., 2016). Baseline model 2 naively fine-tunes mBART only on in-domain bitext.

We train Baseline 1 first on our general bitext for up to 40 epochs, and then on our domain bitext for up to 10 additional epochs, keeping the best model. We train Baseline 2 for up to 40 epochs in the limited domain setting, and up to 100 epochs in the whole domain setting.

We evaluate all of our models with BLEU, as implemented by SacreBLEU[3] (Post, 2018).

5 Results

	Name	**Baseline 1**	**Language-First**	**Domain-First**	**Baseline 2**
	Initialization	Random	mBART	mBART	mBART
	Step 1	$B_{gen}(S_i, T)$	$B_{gen}(S_i, T)$	$\bigcup_i B_{in}(S_i, T)$	None
	Step 2	$B_{in}(S_i, T)$	$B_{in}(S_i, T)$	$B_{in}(S_i, T)$	$B_{in}(S_i, T)$
TED	ar-en	37.0	**37.4**	36.0	36.4
	fa-en	25.4	**30.9**	30.2	29.8
	pt-en	46.9	**48.2**	47.7	47.6
	ru-en	**30.3**	29.7	29.1	29.7
	zh-en	19.0	**22.0**	21.5	21.1
	avg	31.7	**33.6**	32.9	32.9
Microblogs	ar-en	40.0	**42.6**	40.8	40.3
	fa-en	17.6	**27.6**	20.9	10.5
	pt-en	40.0	**44.6**	39.7	33.9
	ru-en	43.2	**47.4**	46.8	45.8
	zh-en	19.7	**25.9**	25.4	22.1
	avg	32.1	**37.6**	34.7	30.5

Table 3: BLEU scores for the whole domain data experiments. In this resource setting, we have around 100K total lines in the Microblog domain, and 2M total lines in the TED domain. We find that our language pair-first approach is consistently our best system. We also note that both of our approaches outperform out baselines in a majority of language pair/domain combinations. Besides out-of-mBART languages in our Microblogs domain, our domain-first approach performs competitively, despite using 3X less data in TED, and 50X less data in the Microblogs domain.

5.1 mBART initialization improves domain adaptation

Results for our whole domain setting are summarized in Table 3, and limited domain results appear in Table 4. We recall that both Baseline 1 and our language pair-first setting are fine-tuned on 1M lines of out-of-domain data, followed by in-domain data. In this setting, their main difference is the initialization of parameters via either mBART or a random initialization. We see that in a majority of domain/language pair settings, our language-first approach is our best performing system, and in Table 5, we can see that this simple initialization can lead to drastic improvements of several BLEU points.

mBART initialization improves ru-en in TICO-19 and Microblogs, and improves zh-en in all domains. Both Chinese-English and Russian-English translation were reported to not have benefited from mBART's initialization in original fine-tuning experiments from Liu et al. (2020). We also note that this initialization improves out-of-mBART language pairs, and explain this further in Section 5.4.

5.2 The importance of in-domain data

In our limited domain setting, although our domain-first approach is not consistently competitive with our language-first approach, we do note a large BLEU difference between fine-tuning

[3] Signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.5.1

	Name	**Baseline 1**	**Language-First**	**Domain-First**	**Baseline 2**
	Initialization	Random	mBART	mBART	mBART
	Step 1	$B_{gen}(S_i, T)$	$B_{gen}(S_i, T)$	$\bigcup_i B_{in}(S_i, T)$	None
	Step 2	$B_{in}(S_i, T)$	$B_{in}(S_i, T)$	$B_{in}(S_i, T)$	$B_{in}(S_i, T)$
TED	ar-en	**36.3**	33.5	25.0	17.5
	fa-en	19.9	**24.7**	9.3	2.2
	pt-en	**44.4**	44.3	32.5	22.9
	ru-en	**29.3**	27.6	22.9	17.4
	zh-en	14.6	**17.4**	14.4	9.4
	avg	28.9	**29.5**	20.8	13.9
Microblogs	ar-en	32.8	**37.4**	31.4	26.9
	fa-en	16.0	**26.3**	12.8	8.2
	pt-en	38.8	**43.6**	32.8	22.7
	ru-en	37.0	**42.6**	37.4	30.1
	zh-en	19.5	**25.0**	23.2	20.1
	avg	28.8	**35.0**	27.5	21.6
TICO-19	ar-en	29.7	**32.5**	21.4	21.2
	fa-en	10.8	**23.6**	14.5	10.6
	pt-en	42.3	**45.6**	30.7	29.3
	ru-en	26.9	**30.3**	20.5	20.7
	zh-en	18.1	**23.2**	14.1	13.9
	avg	25.6	**31.0**	20.2	19.1

Table 4: BLEU scores for the limited domain data experiments. In this setting, we limit our bilingual in-domain data to <1k sentence pairs. In the limited domain setting, we find that our language pair-first approach consistently outperforms our baselines and our domain-first approach, with the exception of a few language pairs in TED. Additionally, although our domain-first approach does not perform competitively in this resource setting, we see benefits of multilingual in-domain learning by noting its improvements over Baseline 2.

	Limited Domain			Whole Domain	
	TED	**Microblogs**	**TICO-19**	**TED**	**Microblogs**
ar-en	-2.8	4.6	2.8	0.4	2.6
fa-en	4.8	10.3	12.8	5.5	10.0
pt-en	-0.1	5.1	3.3	1.3	4.6
ru-en	-0.9	3.8	0.7	-0.6	4.2
zh-en	2.8	5.5	5.1	3.0	6.2
avg	0.8	5.9	4.9	1.9	5.5

Table 5: ΔBLEU between initializing domain adaptation fine-tuning with mBART vs domain adaptation fine-tuning with a random Transformer initialization. Overall, mBART's initialization improves domain adaptation over a random Transformer initialization. This holds for the fa-en and pt-en language pairs, which are outside of mBART's pretraining set, sometimes leading to improvements of over 10 BLEU points.

on multilingual domain data (domain-first) and fine-tuning on bilingual in-domain data only (Baseline 2). We report these differences in Table 6. By using multilingual in-domain data, we can see up to 10 BLEU point improvements over using in-domain data only in the target setting. We note that we see a reduced efficacy in TICO-19, which may be in part due to its

	Limited Domain			Whole Domain	
	TED	**Microblogs**	**TICO-19**	**TED**	**Microblogs**
ar-en	7.5	4.5	0.2	-0.4	0.5
fa-en	7.1	4.6	3.9	0.4	10.4
pt-en	9.6	10.1	1.4	0.1	5.8
ru-en	5.5	7.3	-0.2	-0.6	1.0
zh-en	5.0	3.1	0.2	0.4	3.3
avg	6.9	5.9	1.1	0.0	4.2

Table 6: ΔBLEU between our domain-first approach using multilingual fine-tuning and our Baseline 2 system, which uses bilingual fine-tuning on our in-domain bitext. The addition of in-domain bitext outside of our source-target pair can be very useful for domain adaptation. The use of a pretrained multilingual model allows us to utilize additional in-domain corpora for improved in-domain performance via a shared parameter space.

multi-parallel nature. This multi-parallel nature also allows any improvement in this domain to be explained purely through multilingual parameter sharing, rather than other factors like increased diversity of tokens appearing in the target setting. In our TED and Microblog domains, using multilingual corpora can lead to much better unigram vocabulary coverage of the target language. For example, only 47% of the Microblog fa-en dev set unigrams are accounted for in the corresponding in-domain training set. However, 74% of these unigrams are accounted for across the in-domain multilingual training sets, providing a possible explanation for these large improvements in our domains besides TICO-19.

In the whole domain setting, we still see strong improvements over Baseline 2 with Microblogs, but modest improvement in the TED setting. However, utilizing multilingual fine-tuning results in ≤ 1 BLEU point of difference here, but is much more efficient by sharing parameters within one model. In Table 2, we see that the TED domain and our general bitext are far more similar than the Microblog or TICO-19 domain and our general bitext. The extent of domain shift may also explain why in the whole domain setting, we see drastic improvement in the Microblogs domain using multilingual fine-tuning, but modest improvement in TED.

5.3 Language-First vs Domain-First

In the limited domain setting, our language-first approach is consistently better than our domain-first approach, and our language-first approach outperforms our baselines in a majority of settings. We believe that in this setting, our limited domain data is insufficient to properly harness the in-domain transfer across languages that we seek to gain from our domain-first approach. Therefore, we recommend the use of additional general bitext in a low resource domain setting.

In the whole domain setting, we see a similar trend, however, the difference between the two proposed approaches is less pronounced. In a majority of language pair/domain combinations, both of our proposed approaches outperform our baselines. In both of our domains in our whole domain setting, and for languages in mBART's pretraining set, the difference between our proposed approaches is within 2 BLEU points. For languages outside of mBART's pretraining set, this difference is a bit more pronounced in the Microblogs domain. This may be due to their small corpus size, where both Microblogs corpora are <3K parallel sentences for both fa-en and pt-en. However, in the TED domain, even fa-en and pt-en have similar performance across the two approaches.

While these two approaches may be comparable in terms of performance, their data and parameter efficiencies are very different. In the language pair-first setting, we use 5M total lines of bitext to create 5 different general-domain fine-tuned models. Fine-tuning these models on in-domain bitext adds additional data overhead. In the domain-first setting, we use under 2M total lines of in-domain bitext across 12 language pairs for TED, and under 100K total lines of domain bitext across 7 language pairs for our Microblogs domain. This approach is also more parameter efficient due to the shared representations across languages for one domain. This in particular holds true for adapting to new languages, as we can reuse our fine-tuned multilingual domain model, rather than bilingually fine-tune mBART on a new language pair as in our language pair-first approach.

5.4 Out-of-mBART languages

As seen in Tables 5 and 6, both of our out-of-mBART language pairs benefit from multilingual training, whether it be at the pretraining stage, or at the fine-tuning stage. In Table 5, we see clear evidence of mBART's utility for both fa-en and pt-en leading to several BLEU point improvements, even in the whole domain setting, where more in-domain bitext in these language pairs is available. We also see clear benefits of multilingual fine-tuning in these language pairs, resulting in consistent improvements in Table 6. Therefore, for languages outside of mBART25 (with a related language within mBART25), we believe that both of our proposed methods could lead to effective domain adaptation.

5.5 Examples

We choose examples from our whole domain Persian-English TED translations to examine differences in outputs generated from our approaches.

Baseline 1:	*No agricultural products will take the reformist of England.*
Baseline 2:	*Without the genetically engineered crops, hunger will take over the U.K.*
Domain-first:	*Without genetically engineered crops, Britain will be hungry.*
Reference:	*Britain will starve without genetically modified crops.*

Baseline 1:	*How are we going to apply human resources?*
Baseline 2:	*How about the resources? How do we feed not billions of people?*
Language-first:	*How about the resources? How do we want to feed nine billion people?*
Reference:	*What about resources? How are we going to feed nine billion people?*

In our first example, we see that in our domain-first approach, the addition of multilingual in-domain bitext likely improves the in-domain style of the translation. While both generated outputs are similar in their "gisting", the style of the in domain-first most closely matches the overall style of TED Talks. In our second example, we see a clear improvement of translation quality at the lexical level as a result of additional bitext in the first fine-tuning step.

6 Conclusion

In this paper, we demonstrate that multilingual pretraining can be very effective in the domain adaptation setting, and we propose two methods of adaptation that are more useful than a naive adaptation approach. We also find that between our methods, our language-first approach where models are first customized to a specific bilingual setting, is consistently our best system, especially in limited domain scenarios. However, we also find that when we first customize our models to a domain, as in our domain-first approach, we achieve considerable translation quality at a fraction of the data needed in our language-first approach. Interestingly, we are also able to show that multilingual pretraining and fine-tuning continue to be effective domain adaptation techniques even when the pretrained model has not seen the language pair before.

References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The amara corpus: Building parallel language resources for the educational domain. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federmann, C., Genzel, D., Guzmán, F., Hu, J., Hughes, M., Koehn, P., Lazar, R., Lewis, W., Neubig, G., Niu, M., Öktem, A., Paquin, E., Tang, G., and Tur, S. (2020). TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M., Purwarianti, A., and Fung, P. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chakrabarty, T., Saakyan, A., and Muresan, S. (2021). Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cooper Stickland, A., Berard, A., and Nikoulina, V. (2021). Multilingual domain adaptation for NMT: Decoupling language and domain information with adapters. In *Proceedings of the Sixth Conference on Machine Translation*, pages 578–598, Online. Association for Computational Linguistics.

Dabre, R., Fujita, A., and Chu, C. (2019). Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416, Hong Kong, China. Association for Computational Linguistics.

Duh, K. (2018). The multitarget TED talks task. `http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/`.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Auli, M., and Joulin, A. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Freitag, M. and Al-Onaizan, Y. (2016). Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Germann, U. (2001). Building a statistical machine translation system from scratch: How much bang for the buck can we expect? In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Lee, E.-S. A., Thillainathan, S., Nayak, S., Ranathunga, S., Adelani, D. I., Su, R., and McCarthy, A. D. (2022). Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? *arXiv preprint arXiv:2203.08850*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Liu, Z., Winata, G. I., and Fung, P. (2021). Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. Association for Computational Linguistics.

Luong, M.-T. and Manning, C. (2015). Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Madaan, L., Sharma, S., and Singla, P. (2020). Transfer learning for related languages: Submissions to the WMT20 similar language translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 402–408, Online. Association for Computational Linguistics.

McNamee, P. and Duh, K. (2022). The multilingual microblog translation corpus: Improving and evaluating translation of user-generated text. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Nguyen, K. and Daumé III, H. (2019). Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Saunders, D. (2021). Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *arXiv preprint arXiv:2104.06951*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

van der Wees, M., Bisazza, A., and Monz, C. (2017). Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xie, W., Hu, B., Yang, H., Yu, D., and Ju, Q. (2021). TenTrans large-scale multilingual machine translation system for WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 439–445, Online. Association for Computational Linguistics.

Xu, H., Ebner, S., Yarmohammadi, M., White, A. S., Van Durme, B., and Murray, K. (2021). Gradual fine-tuning for low-resource domain adaptation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 214–221.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhang, X., Shapiro, P., Kumar, G., McNamee, P., Carpuat, M., and Duh, K. (2019). Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1903–1915, Minneapolis, Minnesota. Association for Computational Linguistics.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).