Can Domains Be Transferred Across Languages in Multi-Domain Multilingual Neural Machine Translation?

Thuy-Trang Vu[◇*] and **Shahram Khadivi**[†]
Xuanli He[◇] and **Dinh Phung**[◇] and **Gholamreza Haffari**[◇]
[◇]Department of Data Science and AI, Monash University, Australia
[†] eBay Inc.
{trang.vu1,xuanli.he1,first.last}@monash.edu
skhadivi@ebay.com

Abstract

Previous works mostly focus on either multilingual or multi-domain aspects of neural machine translation (NMT). This paper investigates whether the domain information can be transferred across languages on the composition of multi-domain and multilingual NMT, particularly for the incomplete data condition where in-domain bitext is missing for some language pairs. Our results in the curated leave-one-domain-out experiments show that multi-domain multilingual (MDML) NMT can boost zero-shot translation performance up to +10 gains on BLEU, as well as aid the generalisation of multi-domain NMT to the missing domain. We also explore strategies for effective integration of multilingual and multi-domain NMT, including language and domain tag combination and auxiliary task training. We find that learning domain-aware representations and adding target-language tags to the encoder leads to effective MDML-NMT.

1 Introduction

Multilingual NMT (MNMT), which enables a single model to support translation across multiple directions, has attracted a lot of interest both in the research community and industry. The gap between MNMT and bilingual counterparts has been reduced significantly, and even for some settings, it has been shown to surpass bilingual NMT (Tran et al., 2021). MNMT enables knowledge sharing among languages, and reduces model training, deployment, and maintenance costs. On the other hand, multi-domain NMT aims to build robust NMT models, providing high-quality translation on diverse domains. While multilingual and multi-domain NMT are highly appealing in practice, they are often studied separately.

To accommodate the domain aspect, previous MNMT works focus on learning a domain-specific

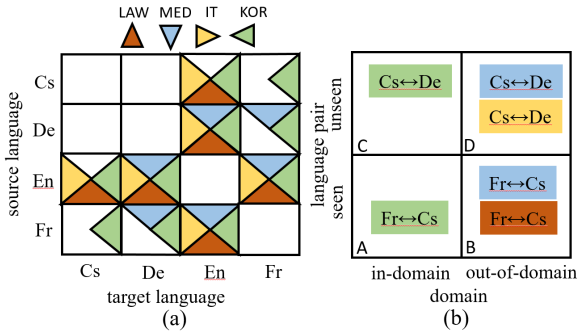

Figure 1: An example of the multi-domain multilingual incomplete data condition (best seen in colours). (a) The colour indicates the availability of bitext in the corresponding domain for each language. (b) Domain and language-pair matrix for the data condition in (a).

MNMT by finetuning a general NMT model on the domain of interest (Tran et al., 2021; Bérard et al., 2020). Recently, Cooper Stickland et al. (2021) propose to unify multilingual and multi-domain NMT into a holistic system by stacking language-specific and domain-specific adapters with a two-phase training process. Thanks to the plug-and-play ability of adapters, their system can handle translation across multiple languages and support multiple domains. However, as each domain adapter is learned independently, their adapter-based model lacks the ability of effective knowledge sharing among domains.

In this paper, we take a step further toward unifying multilingual and multi-domain NMT into a single setting and model, *i.e.,* multi-domain multilingual NMT (MDML-NMT), and enable effective knowledge sharing across both domains and languages. Unlike the *complete* data assumption in the multi-domain single language-pair setting where training data is available in all domains, we assume the existence of bitext in all domains for only a subset of language-pairs, as illustrated in Figure 1(a). In fact, it is highly improbable to obtain in-domain bitext for all domains and all language pairs in

[*]Work done while doing internship at eBay Inc.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 381 - 396

many real-life settings. Depending on the availability of parallel data, we categorise a translation task from a source to a target language into four categories based on the following dimensions:

- *in-domain/out-of-domain*, wrt to the domain of interest, and
- *seen/unseen*, wrt to the translation direction during training.

Please note the domain and language-pair matrix in Figure 1(b). In this figure, parallel data available in the training set specifies the group A, the *in-domain seen* tasks. Given this training dataset, most MNMT research focuses on cross-lingual transfer to *in-domain unseen* translation tasks (A→C), while the studies on multi-domain NMT and domain adaptation seek to generalise to *out-of-domain seen* translation tasks (A→B). Integrating domain and language aspects in the incomplete data condition gives rise to an interesting and more challenging setting that transfers to *out-of-domain unseen* translation tasks (A→D). We hypothesise that the out-of-domain "seen and unseen" translation tasks (A→B+D) can benefit from the in-domain translation tasks if there exists the domain transfer across languages in MDML-NMT.

Specifically, we ask the following research questions: (1) Do out-of-domain translation tasks benefit from the out-of-domain and in-domain bitext in other seen translation pairs? and (2) What is effective method to handle the composition of domains and languages? Furthermore, beyond the cross-lingual transfer (A→C) and the out-of-domain generalisation (A→B), we also consider the challenging setting where the translation direction of interest may not have any bitext in any domain, i.e. the zero-shot setting (A→D).

In general, we can vary the degree of domain transfer based on the number of domains in which parallel data for a translation task is available. Combining with the number of language pairs of interest, there are large numbers of incomplete data conditions, even for our toy examples in Figure 1. In this study, we assume the highest degree of domain transfer and carefully design controlled experiments where one domain is left out for some language pairs (Table 1). We then examine the potential of MDML-NMT on this incomplete data condition. We also explore training strategies for effective integration of multi-domain and multilingual NMT, mainly on (i) how to combine the language and domain tags, and (ii) using auxiliary task training to learn effective representations. Our contributions are as follows:

	LAW	IT	KORAN	MED	SUB
En-Fr	✔	✔	✔	✔	✔
En-De	✔	✔	✔	✔	✔
De-Fr	✔	✔	✔	✔	✔
En-Cs	✗	✔	✔	✔	✔
En-Pl	✗	✔	✔	✔	✔

Table 1: Illustration of leave-one-out LAW experiment setting. ✗, ✔describes whether there is bitext in the corresponding domain for the given language pairs.

- We investigate effective strategies to jointly learn multi-domain and multilingual NMT models under the incomplete data condition.
- Our empirical results show that MDML-NMT model can improve translation quality in the zero-shot directions by mitigating the **off-target translation** issue that an MNMT model translates the input sentence to a wrong target language. Additionally, MDML-NMT exhibits domain transfer ability by achieving up to +4 BLEU improvement over the multi-domain NMT on the translation direction where in-domain training data is absent. Thanks to the effective cross-domain and cross-lingual knowledge sharing, MDML-NMT outperforms the adapter-based method (Cooper Stickland et al., 2021) by a large margin in the language-domain zero-shot setting.
- Our study sheds light on effective MDML-NMT training. Our experimental results reveal that: (i) for the domain, it is important to make the encoder domain-aware by either providing the domain tags or training with the auxiliary task; and (ii) for the language, the best practice is to prepend the target language tag to the encoder.

2 Multi-domain Multilingual NMT

In this section, we first provide the necessary background on multilingual NMT (MNMT) and multi-domain NMT individually. We then describe effective modelling approaches for the integration of multi-domain and multilingual NMT (MDML-NMT).

2.1 Multilingual NMT

Given a set of languages L, the primary goal of MNMT is to learn a single NMT model that can handle all translation directions of interest in this set of languages (Dabre et al., 2020). According to the parameter sharing strategy, MNMT can be categorised into: 1) partial parameter sharing (Dong et al., 2015; Firat et al., 2016; Zhang et al., 2021), and 2) full parameter sharing (Ha et al., 2016; Johnson et al., 2017). The latter has been widely adopted because of its simplicity, lightweight, and its zero-shot capability. Thus, we adopt the full parameter sharing strategy in our work.

In the fully parameter-shared MNMT, all parameters of encoders, decoders and attentions are shared across tasks. Special language tags are introduced to indicate the target languages. One can prepend the target language tags to either the source or target sentences. The model is then trained jointly to minimise the negative log-likelihood across all training instances:

$$\mathcal{L}_{\text{ML}}(\boldsymbol{\theta}) := -\sum_{(s,t)\in T}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{C}_{s,t}} \log P(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) \quad (1)$$

where $\boldsymbol{\theta}$ is model parameters, $\mathcal{C}_{s,t}$ denotes a bilingual corpus for the source language s and the target language t, $(\boldsymbol{x},\boldsymbol{y})$ is a pair of parallel sentences in the source and target language, and T denotes the translation tasks for which we have bitext available. Among all possible language pairs $(s,t) \in L \times L$, we often only have access to bilingual data for a subset of them. We denote these pairs as *seen* (observed) translation tasks, and the rest as *unseen* tasks corresponding to the zero-shot setting.

2.2 Multi-domain NMT

Multi-domain NMT aims to handle translation tasks across multiple domains for a given language pair. Similar to MNMT, tagging the training corpus is the most popular approach, where a tag indicates the domain of a sentence pair. We also minimise the negative log-likelihood across all domains to train the model:

$$\mathcal{L}_{\text{MD}}(\boldsymbol{\theta}) := -\sum_{d\in D}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{C}^d_{s,t}} \log P(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) \quad (2)$$

where D is the set of domains, and $\mathcal{C}^d_{s,t}$ denotes the parallel bitext in the source language s, target language t, and the domain d.

Apart from tagging, some auxiliary tasks have also been incorporated into the training process. A common practice is the use of domain discrimination, which aims to force the encoder to capture *domain-aware* characteristics (Britz et al., 2017). For this purpose, a domain discriminator is added to the NMT model at training time. The input to the discriminator is the encoder output, and its output predicts the probability of the domain of the source sentence. The discriminator is jointly trained with the NMT model, and is discarded at inference time.

Let $\mathbf{h} = \text{enc}(\boldsymbol{x})$ be the representation of sentence $\boldsymbol{x}$ computed by the mean-pooling over the hidden states of the top layer of the encoder. The training objective for the domain-aware encoder is as follows:

$$\mathcal{L}_{\text{disc}}(\boldsymbol{\theta},\boldsymbol{\psi}) := -\sum_{d\in D}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{C}^d_{s,t}} \log \Pr(d|\mathbf{h};\boldsymbol{\psi}) \quad (3)$$

$$\mathcal{L}_{\text{MD-aware}}(\boldsymbol{\theta},\boldsymbol{\psi}) := \mathcal{L}_{\text{MD}}(\boldsymbol{\theta}) + \lambda\mathcal{L}_{\text{disc}}(\boldsymbol{\theta},\boldsymbol{\psi}) \quad (4)$$

where $\boldsymbol{\psi}$ is the parameter of the domain discriminator classifier, and λ controls the contribution of the domain discriminator into the training objective of the multi-domain NMT model.

Alternatively, one can design an adversarial training objective in order to learn domain-agnostic representations by the encoder. This is achieved by inserting a gradient reversal layer (Ganin and Lempitsky, 2015) between the encoder and the domain discriminator. The gradient reversal layer behaves as an identity layer in the forward pass but reverses the gradient sign during back-propagation. It has the opposite effect on the encoder, forcing it to learn domain-agnostic representations. This encourages the domain specific characteristic to be learned mainly by the decoder of the NMT model.

2.3 Composition of Domains and Languages

In this paper, we explore strategies for composing multi-domain and multilingual NMT. We consider the incomplete multi-domain multilingual data condition where in-domain data may be only available in a subset of language pairs. For example, Table 1 shows one of the data conditions explored in our experiments in Section 3. Given the five language pairs and five domains, we assume that the domain data in some language pairs are missing. Our goal is to investigate effective techniques to train a high-quality MDML-NMT model covering all combinations of domains and language pairs.

Given a specific domain, we define *in-domain languages* as those having data available in the domain as part of some bilingual corpora; the rest

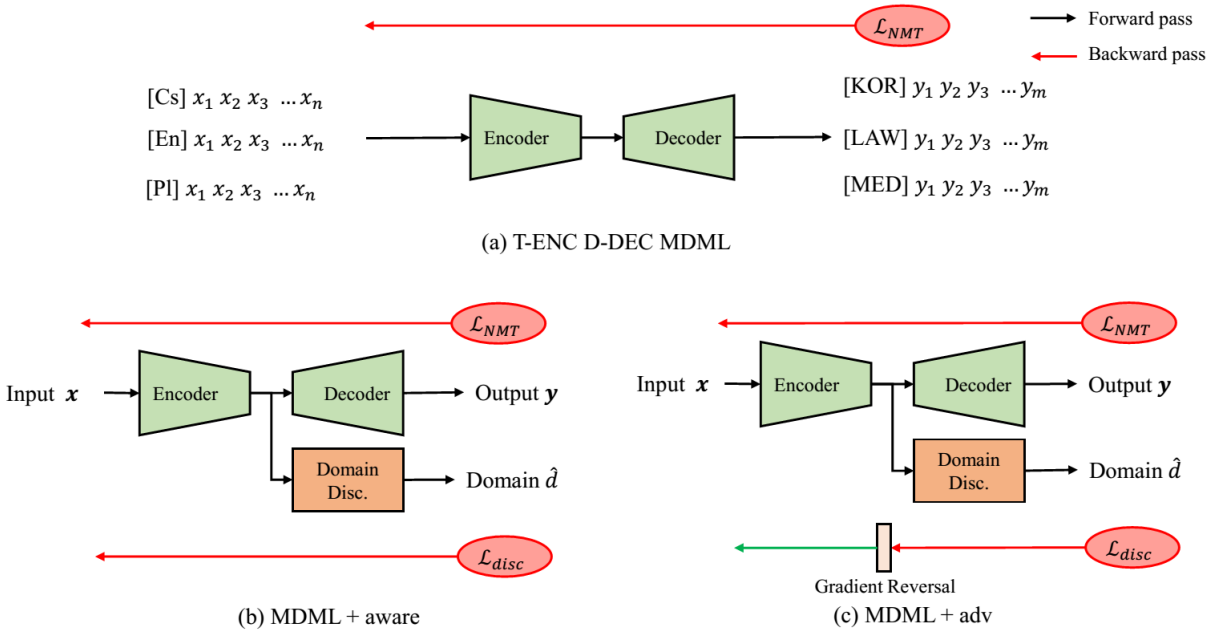

Figure 2: Illustration of domain and languages composition strategies: (a) prepending domain (D) and target language (T) tag to encoder (ENC) or decoder (DEC). This example shows a T-ENC D-DEC model where the target language tag and domain tag are added to encoder and decoder respectively; (b) combining the tagging method with the domain aware auxiliary task (MDML + aware) to learn domain-aware representation; and (c) combining the tagging method with the domain adversarial auxiliary task (MDML + adv) to learn domain-agnostic representation.

Trans. direction	Eval. domain	MDML task type
En→De	LAW	seen in→in
En→Cs	LAW	seen in→out
Pl→En	LAW	seen out→in
De→Cs	LAW	unseen (zero-shot) in→out
Cs→De	LAW	unseen (zero-shot) out→in
Pl→Cs	LAW	unseen (zero-shot) out→out

Table 2: Examples of MDML task types in the leave-one-domain-out LAW training scenario of Table 1. Please refer to Table 1 for the in/out and seen/unseen settings.

of the languages are referred to as *out-of-domain languages*. We consider all combinations of in-domain/out-of-domain source/target languages for both seen and unseen translation directions (see examples in Table 2) in Section 3.

We investigate different combinations of the tagging strategy and auxiliary task training to effectively train MDML-NMT models, as shown in Figure 2.

Language and Domain Tags. We explore different ways of injecting the target language tags and domain tags into the translation process. Following the standard convention, we explore inserting the target language tag at the beginning of either the source sentence or the translation. Furthermore, the domain tag can also be added to either the source or the target side.

Auxiliary Task Training. We investigate the effect of encoder-based auxiliary tasks on MDML-NMT. As described in Section 2.2, we consider two types of auxiliary objectives to train encoder which are domain-aware or domain-agnostic. The former aims to amplify the domain-related features, while the latter focuses on the domain invariant representation in the encoder.

3 Experiments

In this section, we evaluate the MDML-NMT approaches and seek to answer the following research questions (**RQs**):

- **RQ1**: *Do out-of-domain translation tasks benefit from the out-of-domain and in-domain bitext in other translation pairs?*

  We explore the benefits of having a single MDML model trained on all available training data from multiple languages and domains over the multi-domain bilingual (MDBL) and the single domain multilingual (SDML) models learned on a subset of training data from a

single language pair or domain. We carefully design controlled experiments to build incomplete data conditions and study the translation quality of the unified MDML-NMT model on both seen and unseen (zero-shot) translation directions. We hypothesise that the translation involving the out-of-domain languages can be beneficial from the in-domain languages thanks to the knowledge sharing across domain and languages.

- **RQ2**: *What is effective method to handle composition of domains and languages?*

  We investigate strategies for effective integration of existing multi-domain and multilingual NMT methods, including the use of language and domain tags and auxiliary task training.

3.1 Setup

We describe the experimental setup in this section, and then present our results.

Dataset. We conduct experiments with translation directions among five languages English (En), Czech (Cs), German (De), French (Fr) and Polish (Pl). Following the recipe in Koehn and Knowles (2017), we create five domains: Law (LAW) , IT (IT), Koran (KOR), Medical (MED), and Subtitles (SUB) from OPUS (Tiedemann, 2012). These corpora are deduplicated and randomly selected, from each corpus 2K sentences extracted as the development and test sets in all possible translation pairs. The statistics of the training dataset are reported in Appendix A.

Seen vs Unseen Language Pairs. We categorise the evaluated languages into two groups, high-resource languages including En, De, and Fr, for which bilingual data among these languages is easy to obtain. We also consider low-resource languages, including Cs and Pl, for which only English-centric data is available, resulting in two language pairs. As a result, there are five *seen* language pairs, consisting of ten seen translation directions.[1] There are also five *unseen* language pairs, resulting in ten unseen translation directions; they are the ones for which we do not have any bitext in the dataset.[2]

Leave-one-domain-out (LODO). We curate the incomplete MDML data condition by removing the data of one domain for the translations tasks involving low-resource languages. An example of the leave-one-domain-out data condition is shown in Table 1. In total, there are five LODO conditions, each of which corresponding to removing the bitext of one domain for both En-Cs and En-Pl (*i.e.,* our low-resource language pairs). For each of these LODO conditions, we have five seen language-pairs and five unseen language-pairs, hence a total of 20 translation tasks in both directions.

In the multi-domain NMT literature, this setting is related to domain generalisation which evaluates the NMT model on out-of-domain data in a zero-shot manner. By carefully removing only a specific domain, we would like to examine whether extra data (*i.e.,* the in-domain and out-of-domain data for high-resource languages, and out-of-domain data for low-resource languages) can boost the generalisation of MDML-NMT to the domain of interest.

Models. We use Transformer (Vaswani et al., 2017) as the NMT model architecture and Fairseq implementation (Ott et al., 2019). For all MDML-NMT models, we initialise them with mBART_large (Liu et al., 2020). We describe the model training details in Appendix B.

As described in Section 2.3, our approaches to MDML problem include combining language and domain tags, and adding domain auxiliary task to the standard multilingual NMT objective. In the first approach, the target language tags can be inserted to the source sentence (T-ENC) or the target sentence (T-DEC). The domain tags can also be handled in similar manners denoted as D-ENC and D-DEC respectively. On combining these tags, the language tag always appears first in the sentence. In addition to the domain and language tag combination, we also explore whether learning domain-aware or domain-agnostic representation in the encoder with auxiliary task can aid MDML-NMT performance. Figure 2 summarises the MDML-NMT approaches evaluated in this paper.

We also report the results of the adapter-based domain-specific MNMT, proposed by Cooper Stickland et al. (2021). Language adapters (Bapna and Firat, 2019) are firstly injected to each layer of a pre-trained MNMT model and then trained while freezing the backbone. Then, domain adapters are stacked on top of the language adapters and trained without backpropagating to the MNMT backbone and the language adapters. Since we do not consider any additional parallel

[1]This set consists of En-Fr, En-De, De-Fr, En-Cs, En-Pl.
[2]This set consists of De-Cs, De-Pl, Fr-Cs, Fr-Pl, Cs-Pl.

	D-ENC	D-DEC
MDBL	12.43	10.21
+adv	12.91	9.90
+aware	13.13	10.13

	D-ENC		D-DEC	
	T-ENC	T-DEC	T-ENC	T-DEC
MDML	14.48	13.21	14.11	8.16
+adv	14.91	14.30	14.72	8.44
+aware	15.00	14.59	**15.35**	7.99

Table 3: Average BLEU score of En→Cs translation across all leave-out domains for multi-domain multilingual (MDML) models and multi-domain bilingual (MDBL) models. The best score on overall and within each tagging group are marked in **bold** and underline respectively.

		seen -both	unseen -SDML	unseen -both
T-ENC	SDML	**41.40**	6.80	7.73
	MDML	37.25	**21.72**	**9.27**
T-DEC	SDML	**41.03**	7.79	8.16
	MDML	35.44	**21.43**	**14.73**

Table 4: Average BLEU scores of single-domain multilingual (SDML) and multi-domain multilingual (MDML) on the leave-out domains for three groups: (i) *seen-both* - the three seen high-resource language pairs (En-De, En-Fr, De-Fr); (ii) *unseen-SDML* - the two low-resource language pairs which are seen by MDML but unseen to SDML (En-Cs, En-Pl); and (iii) *unseen-both* - the other five unseen language pairs.

data apart from the multi-domain dataset, we train the MNMT backbone as well as the language and domain adapters using this multi-domain multilingual dataset (instead of Paracrawl) for fair comparison.

Evaluation. We report the detokenised BLEU scores calculated by SacreBLEU (Post, 2018) (Post, 2018) and the micro-average of BLEU score in a group as the measure of overall performance.[3]

3.2 Results and Discussions

Can multilinguality help the multi-domain learning? (MDBL vs. MDML) We first examine the potential of MDML over the counterpart multi-domain NMT model. Table 3 shows the BLEU scores of MDBL and MDML for En→Cs translation on various LODO settings. A breakdown of BLEU scores on leave-out domains is shown in Table 11 in the Appendix C. The MDBL models are trained on all En→Cs bilingual data except of the domain of interest. Within the same tagging method, augmenting the NMT training with the domain auxiliary objectives (*i.e.,* domain-aware and domain-agnostic encoders) enhances the translation performance. The MDML models consistently surpass the corresponding MDBL settings, with an exceptional case, where both domain and language tags are applied to the decoder (*i.e.,* T-DEC D-DEC). This observation suggests there is knowledge sharing from in-domain languages to out-of-domain languages.

	MDML	+adv	+aware
T-ENC	25.17	28.91	**30.14**
T-ENC D-ENC	24.36	28.90	29.23
T-ENC D-DEC	22.10	29.43	29.94
T-DEC	24.82	29.14	29.52
T-DEC D-ENC	24.95	28.56	29.01
T-DEC D-DEC	19.19	17.68	14.37
Adapter-based		23.26	

Table 5: Average BLEU score of MDML-NMT models across all five leave-one-out scenarios. The best score overall and within each tagging group are marked in **bold** and underline respectively.

Can multi-domain data help multilingual NMT? (SDML vs. MDML) SDML models are domain-specific multilingual NMT models trained on the multilingual dataset in a given domain. As in-domain parallel data is absent for several language pairs, the MDML models are exposed to more seen translation tasks than SDML models thanks to the availability of out-of-domain data. Hence, for a given domain, we divide the evaluation translation tasks into three groups: seen-both, unseen-SDML and unseen-both. The seen-both and unseen-both groups consist of translation directions which are observed and unobserved respectively by both models in training. The unseen-SDML group corresponds to those unseen by SDML, but seen by MDML models. We report the average performance of the MDML and SDML model on the

[3] nrefs:1|case:mixed|eff:no|tok:none|smooth:exp|version:2.0.0

		seen (10)			unseen (zero-shot) (10)			AVG
		in→in (6)	in→out (2)	out→in (2)	in→out (4)	out→in (4)	out→out (2)	
Adapter-based		34.32	11.76	**33.34**	7.38	6.86	6.84	16.75
T-ENC	MDML	37.25	**14.63**	29.05	7.93	10.35	9.79	18.17
	+adv	36.81	13.88	28.33	**10.91**	22.05	11.38	20.56
	+aware	37.50	14.31	29.09	10.61	24.50	11.94	21.33
T-ENC D-ENC	MDML	32.32	11.52	24.23	7.22	17.25	7.85	16.73
	+adv	37.24	13.66	31.17	10.20	24.21	11.67	**21.36**
	+aware	**37.57**	13.15	31.15	8.65	25.20	11.24	21.16
T-ENC D-DEC	MDML	31.94	10.55	22.14	5.88	8.63	5.83	14.16
	+adv	36.70	12.85	25.38	10.61	22.57	9.52	19.61
	+aware	37.47	12.08	25.59	10.08	22.41	9.01	19.44
T-DEC	MDML	31.44	11.25	23.62	7.63	20.39	8.59	17.15
	+adv	36.92	13.94	28.83	8.93	24.48	12.14	20.87
	+aware	37.20	14.00	28.62	10.30	23.95	**12.18**	21.04
T-DEC D-ENC	MDML	31.80	10.40	22.13	5.97	18.81	7.47	16.10
	+adv	36.35	13.22	27.96	8.46	24.35	10.21	20.09
	+aware	37.00	13.32	29.34	9.57	**25.89**	11.43	21.09
T-DEC D-DEC	MDML	30.17	4.77	24.72	3.65	14.08	4.43	13.64
	+adv	25.18	6.04	25.94	5.88	14.72	6.37	14.02
	+aware	20.61	5.50	23.27	5.72	7.64	6.40	11.52

Table 6: Average BLEU score on leave-out domain for different translation tasks. We categorise 20 translation direction into *seen* where the training data for the translation direction is available, otherwise *unseen*. *in* and *out* show whether the corresponding domain is observed during training or not (see Table 2 for a concrete example). The number in parentheses shows how many translation directions are in the corresponding category. The best score of each column overall and within each tagging group are marked in **bold** and underline respectively.

leave-out domains in Table 4. The detailed results on each leave-out domain can be found in Table 12 in the Appendix C. As expected, SDML works well on the seen directions (seen-both) but behaves badly on the zero-shot settings (unseen-SDML and unseen-both). We speculate it is due to the negative inference among domains. On the other hand, MDML outperforms SDML in unseen-SDML by a large margin thanks to the out-of-domain parallel data. Additionally, leveraging multi-domain data also helps to improve multilingual NMT on unseen-both tasks up to +6 BLEU score on average.

What is an effective method to MDML? We have previously shown the benefits of MDML over multi-domain and multilingual NMT models. The remaining question is how to integrate the multi-domain and multilingual approaches effectively. We report the average BLEU scores of different MDML methods across all five LODO scenarios and 20 translation tasks in Table 5. Similar to the previous observation on En→Cs translation, models with domain discriminator outperform the vanilla MNMT model in all tagging methods. More specifically, the domain-aware MNMT models (+aware) are the winning method in most scenarios. These results emphasise the importance of having domain-aware representation in the encoder. Furthermore, it shows MDML is more effective than the adapter-based approach.

As illustrated in Table 2, translation tasks in MDML setting can be categorised into seen and unseen (zero-shot) tasks involving the in-domain or out-of-domain languages. Table 6 reports the performance of MDML-NMT models in the leave-out domains on different task categories, e.g. LAW in the example in Table 1. The results for other domains, i.e. excluding the leave-out domains, can be found in Appendix C. Consistent with previous findings, the domain discriminative mixing methods outperform the other models. While the best multilingual NMT model (MDML T-ENC) performs comparably with other MDML-NMT models on seen translation tasks, the main benefit of MDML-NMT models comes from unseen translation tasks. As expected, for both seen and unseen tasks, the quality of translation when translating into in-domain languages is consistently higher than into out-of-domain languages. Stacking the

		seen					unseen (zeroshot)			
		En	De	Fr	Cs	Pl	De	Fr	Cs	Pl
T-ENC	MDML	94.72	95.99	95.54	92.10	94.50	48.66	49.38	32.73	40.78
	+adv	94.81	96.01	95.33	91.62	95.06	75.56	85.93	59.18	66.57
	+aware	94.85	96.09	95.56	91.60	94.69	80.92	90.86	64.77	74.67
T-ENC D-ENC	MDML	92.55	95.54	95.06	91.21	94.12	73.99	72.85	44.69	58.83
	+adv	94.65	96.11	95.42	90.51	93.60	80.30	81.16	59.13	67.17
	+aware	94.67	96.18	95.44	90.35	92.69	81.21	84.05	61.10	66.92
T-ENC D-DEC	MDML	94.33	95.22	95.10	91.26	94.98	43.53	46.94	36.10	44.04
	+adv	94.86	95.81	95.44	91.55	94.64	87.23	90.67	69.01	74.42
	+aware	95.01	96.01	95.49	91.34	94.46	82.00	91.38	68.75	75.69
T-DEC	MDML	94.03	95.32	95.04	90.70	93.83	90.44	92.74	60.44	70.93
	+adv	94.68	96.05	95.44	91.72	94.84	86.03	88.64	52.45	64.01
	+aware	94.72	96.21	95.50	92.22	95.13	77.20	87.61	58.17	70.86
T-DEC D-ENC	MDML	92.72	95.51	95.06	89.72	92.23	85.75	89.82	56.74	68.56
	+adv	93.82	96.14	95.53	91.41	94.27	84.70	87.99	51.59	63.66
	+aware	94.19	96.12	95.54	91.54	93.80	79.93	87.00	60.22	72.77
T-DEC D-DEC	MDML	93.44	90.30	93.44	74.49	83.06	64.33	58.96	21.42	25.74
	+adv	80.29	17.71	94.36	49.29	16.07	3.37	47.72	1.04	0.62
	+aware	69.89	14.25	85.62	52.34	10.10	2.89	9.52	2.07	0.28
		0	25	50	75	100				

Table 7: On-target translation ratio of MDML-NMT models on the seen and unseen translation tasks.

language and domain adapters works particularly well in seen translation direction to in-domain target languages. Aligned with previous findings, the adapter-based method struggles to translate to out-domain target languages due to the unobserved combination of language and domain adapters during training (Cooper Stickland et al., 2021).

4 Analysis

4.1 Domain-specific token generation

In this section, we will look at how well MDML models are in generating domain-specific tokens. We concatenate all training data in a given domain in each language, remove stopwords, and extract the top 1000 domain-specific tokens with TF-IDF. The stopwords for each language are obtained from stopwords-iso[4]. Table 8 reports the F1 score of MDML models in generating leave-out domain-specific tokens. As expected, translation to in-domain languages (in→in, out→in) has a higher F1 score than translation to out-of-domain languages (in→out, out→out). Compared to MDML, both MDML-aware and MDML-adv models are able to generate more domain-specific tokens.

4.2 On-target translation ratio

One challenge of multilingual NMT (MNMT) is the off-target translation in zero-shot direction. Off-target translation is an issue that the MNMT model translates the input sentence to the wrong language, causing low BLEU scores. In this section, we assess the ability to alleviate the off-target issue in MDML models. Table 7 reports the on-target translation ratio of MDML models on seen and unseen translation for different target languages. We detect the language of translated targets using langdetect[5] tool and calculate the on-target translation ratio as the percentage of translated sentences having the target language detected correctly. As expected,

		in→in	in→out	out→in	out→out
T-ENC	MDML	63.22	21.45	35.58	16.42
	+adv	62.71	26.73	44.48	22.06
	+aware	63.45	25.85	47.53	23.96
T-ENC D-ENC	MDML	58.93	20.58	35.17	16.10
	+adv	63.14	24.24	46.75	23.55
	+aware	63.48	20.80	47.82	23.17
T-ENC D-DEC	MDML	58.82	20.32	30.34	13.37
	+adv	62.83	27.68	47.02	26.21
	+aware	63.59	27.64	47.21	25.73
T-DEC	MDML	58.35	21.70	43.69	19.98
	+adv	62.83	23.72	47.55	25.56
	+aware	63.08	25.90	47.31	25.49
T-DEC D-ENC	MDML	58.94	18.34	38.85	17.56
	+adv	62.37	21.86	45.67	20.31
	+aware	62.98	24.23	47.87	24.17
T-DEC D-DEC	MDML	56.74	12.52	40.01	10.98
	+adv	46.71	9.18	34.73	8.34
	+aware	39.20	8.93	26.32	8.06
	0	25	50	75	100

Table 8: In-domain token generation F1 score.

[4] https://github.com/stopwords-iso/stopwords-iso

[5] https://github.com/Mimino666/langdetect

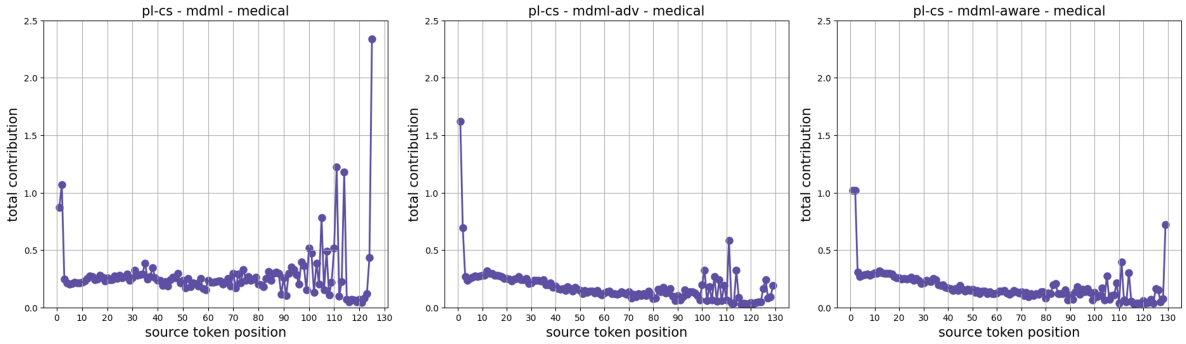

Figure 3: Source token contribution on Pl→Cs MDML with T-ENC D-ENC. The target language and domain tag are the first two tokens.

		En	De	Fr	Cs	Pl
LO	En		91.60	93.27	33.30	43.99
	De	93.26		90.79	9.88	10.82
	Fr	90.59	83.86		2.53	4.90
	Cs	91.51	68.39	64.78		7.87
	Pl	92.18	66.58	64.96	18.83	
others	En		95.38	94.79	84.79	92.83
	De	94.58		92.78	37.27	46.95
	Fr	91.70	86.50		22.04	28.20
	Cs	94.37	63.79	58.49		15.48
	Pl	94.66	63.29	56.46	13.21	
		0	25	50	75	100

Table 9: On target ratio of T-DEC D-DEC MDML on the leave-out (LO) and other domains. Rows and columns correspond to the source and target languages.

the seen translation tasks have more than 90% sentences in the correct target language, except T-DEC D-DEC models. On the other hand, the unseen tasks suffer from a low ratio, especially for Cs and Pl. We also observe significant improvement from MDML-aware and MDML-adv over the MDML models on unseen translation tasks to Cs and Pl.

Generally, T-DEC D-DEC model always underperforms other models and have a much lower on-target ratio on unseen tasks. Table 9 further confirms this phenomenon on the leave-out domains. While heavily suffering from the off-target issue in the leave-out domains, it has comparable ratios to other methods in other domains on seen tasks En-Pl and En-Cs. One possible explanation is that the combination of the target language and domain tags has never been observed during training for the unseen tasks with out-of-domain languages.

4.3 Language and domain tag contribution

To understand the role of the target and language tags to the generated prediction, we estimate the total contribution of source tokens at each position to the whole target sentence using Layerwise Relevance Propagation (Voita et al., 2021). We filter out the pairs having too short or too long target sentences and compute the contribution to target sentences of length between 10 and 100.

Results of T-ENC D-ENC MDML models on Pl→Cs translation in the leave-out medical domain are shown in Figure 3. The language and domain tag are the first two source tokens in respective order. It can be seen that all models have a similar trend in which the contribution of source tokens decreases toward later positions and suddenly increases at a few last positions. Additionally, the target language tags play an important role in the final prediction of all MDML models. Interestingly, while still having a fairly high contribution compared to other tokens, the domain tag seems less important for the domain adversarial models. It can be explained that the encoder learns to produce domain agnostic representation; hence less depends on the domain tags.

5 Related works

Multilingual NMT. As a remarkable branch of NMT, multilingual NMT (MNMT) has been appealing for its capability of supporting translations among different language pairs. Dong et al. (2015) opened the door to the MNMT by conducting a one-to-many translation. Firat et al. (2016) effectively extend this approach to a many-to-many setting. Since these approaches consider each translation as an independent system, they suffer from two major drawbacks. First, as the parameter size is proportional to the language size, it is not parameter-efficient when scaling to tens or hundreds of languages. In addition, the separate architectures cannot fully benefit from cross-lingual knowledge transfer. Johnson et al. (2017); Ha et al.

(2016) devise a universal MNMT system to alleviate these issues by prepending a target language tag to the inputs and training a shared SEQ2SEQ model on the concatenation of all bitext. However, owing to the negative interference, high-resource languages suffer from translation inferiority, compared to the corresponding bilingual NMT models. As a remedy, Zhang et al. (2021); Kudugunta et al. (2021) leverage a mixture-of-experts design to separate language-specific features from the generic features by incorporating language-specific components into the universal MNMT model. Besides, Bapna and Firat (2019); Zhu et al. (2021) propose to fine-tune a lightweight adapter as a means of compensation for the quality loss caused by the adverse effect.

Multidomain NMT. While both involving training on dataset coming from multiple domains, NMT domain adaption is different from multi-domain NMT. The former aims to transfer the knowledge of out-of-domain data into the in-domain data (Luong and Manning, 2015; Zoph et al., 2016; Freitag and Al-Onaizan, 2016), while the latter focuses on building a system, performing well on multiple domains (Pham et al., 2021). Since lexical and topic variations have been observed in different domains, it is challenging to handle the mixed-domain data with a generic NMT model (Farajian et al., 2017). To operate translation in multiple domains, recent research focuses on exploiting domain-shared and domain-specific knowledge by introducing a domain tag to the source sentence (Kobus et al., 2017), using auxiliary objectives such as domain discrimination loss (Britz et al., 2017; Gu et al., 2019), domain knowledge distillation (Currey et al., 2020), and modifying the architecture to capture this information explicitly (Zeng et al., 2018). Rather than using a heavy domain-specific encoder-decoder architecture, Wang et al. (2020) introduce lightweight domain transformation layers between the shared encoder and decoder.

Multilingual & multi-domain NMT. Previous works have mainly considered multilingual and multi-domain NMT models as two disjoint systems. Until recently, Cooper Stickland et al. (2021) propose to unify these two settings into a holistic system, but focus more on the domain adaptation angle. They investigate the combination of language and domain adapters by superimposing domain adapters on language adapters. They noticed that domain adapters and back-translation could boost the translation quality on the out-of-domain languages. In contrast, our work creatively stitches multilingual and multi-domain NMT together and explores the capability of a cross-lingual domain transfer within a unified model without adaption.

6 Conclusion

We study the problem of MDML-NMT for which a single NMT can support multiple translation directions and domains. We investigate whether the tagging and auxiliary task learning method can be combined for MDML-NMT. Our empirical results reveal a positive transfer from in-domain to out-of-domain languages, especially in the zero-shot scenario. This study provides insights into the synergy of the domain and language aspects of training an MDML-NMT model. The main findings include: (i) it is crucial to make the encoder domain-aware; and (ii) it is best to prepend the target language tag to the encoder in MDML. These findings lay the groundwork for future research in this direction.

Acknowledgments

This research is supported by an eBay Research Award and the ARC Future Fellowship FT190100039. This work is partly sponsored by 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 Minghao Wu, Michelle Zhao and the anonymous reviewers for their helpful comments to improve the manuscript.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Alexandre Bérard, Zae Myung Kim, Vassilina Nikoulina, Eunjeong Lucy Park, and Matthias Gallé. 2020. A multilingual neural machine translation model for biomedical data. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

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

Asa Cooper Stickland, Alexandre Berard, and Vassilina Nikoulina. 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.

Anna Currey, Prashant Mathur, and Georgiana Dinu. 2020. Distilling multiple domains for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4500–4511, Online. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, Nicola Bertoldi, and Marcello Federico. 2017. Neural vs. phrase-based machine translation in a multi-domain scenario. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 280–284, Valencia, Spain. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Shuhao Gu, Yang Feng, and Qun Liu. 2019. Improving domain adaptation translation with domain invariant and specific information. 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 3081–3091, Minneapolis, Minnesota. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Philipp Koehn and Rebecca Knowles. 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.

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. 2021. Beyond distillation: Task-level mixture-of-experts for efficient inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3577–3599.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Minh-Thang Luong and Christopher Manning. 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.

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 NAACL-HLT 2019: Demonstrations*.

MinhQuang Pham, Josep Maria Crego, and François Yvon. 2021. Revisiting multi-domain machine translation. *Transactions of the Association for Computational Linguistics*, 9:17–35.

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

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

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook ai's wmt21 news translation task submission. In *Proc. of WMT*.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Yong Wang, Longyue Wang, Shuming Shi, Victor OK Li, and Zhaopeng Tu. 2020. Go from the general to the particular: Multi-domain translation with domain transformation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9233–9241.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, Brussels, Belgium. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Serial or parallel? plug-able adapter for multilingual machine translation. *arXiv preprint arXiv:2104.08154*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.

Domain	Cs-En	De-En	Fr-En	Pl-En	De-Fr
LAW	1.3M	467K	596K	1.0M	1.3M
IT	73K	158K	230K	97K	146K
MED	686K	705K	705K	666K	707K
KOR	117K	17.8K	28K	30K	10K
SUB	595K	494K	492K	491K	590K

Table 10: Number of training sentences in the evaluation datasets. Each dataset contains 2K dev and test sentences.

A Data statistics

Table 10 shows the statistics of dataset used in the experiments.

B Training Details

For all MDML-NMT models, we initialise them with mBART_large (Liu et al., 2020) and train with mixed-precision training up to 200K update steps (around 13 epochs) using a batch size of 8192 tokens and early stopping on 8 V100 GPUs. The multi-domain NMT (MDBL) is trained in a similar manner, except with the total update steps of 60K which is equivalent to around 30 epochs. We apply Adam with an inverse square root schedule, a linear warmup of 5000 steps and a learning rate of 3e-5. We set dropout and label smoothing with a rate of 0.3 and 0.2. We use temperature-based sampling with $T = 5$ to balance training size between domains and languages (Arivazhagan et al., 2019).

For the NMT model with auxiliary task, the domain discriminator is a 2-layer feed-forward network with hidden size of 1024. We set the mixing hyperparameters λ in Equation 4 to 1, *i.e.,* the domain discriminative loss and NMT loss contributes equally to the training signal.

Followed (Cooper Stickland et al., 2021), we use adapter bottle-neck of 1024 for the adapter-based models. The monolingual language adapters are trained all together on the multi-domain dataset while the NMT backbone are frozen. In contrast, we train domain adapters separately for each domain and build homogeneous batches containing sentences from the same language direction and domain. We also apply domain-adapter dropout (DADrop) where the domain adapters are skipped 20% of time.

C Additional Results

MDBL vs. MDML. Table 11 shows the BLEU scores of different models for En→Cs translation on various LODO settings. Each domain column reports the results corresponding to the LODO setting in which the bitext of that domain is removed.

SDML vs. MDML. We report the performance of the MDML and SDML model on each leave-out domains in Table 12.

MDML Result. The average BLEU scores on each domain across all five LODO scenarios and 20 translation tasks are reported in Table 13. Table 14 reports the performance of MDML-NMT models on other domains (excluding the leave-out domains) on different task categories.

		LAW	IT	KOR	MED	SUB	AVG
D-ENC	MDBL	10.52	20.37	7.63	19.46	4.16	12.43
	+adv	10.62	19.43	8.16	20.79	5.57	12.91
	+aware	10.37	21.70	8.25	20.07	5.26	13.13
D-DEC	MDBL	9.21	12.39	6.64	19.56	3.27	10.21
	+adv	9.78	11.01	6.76	18.03	3.94	9.90
	+aware	9.51	12.25	7.00	18.46	3.41	10.13
T-ENC D-ENC	MDML	11.98	22.64	8.08	21.05	8.63	14.48
	+adv	12.11	**23.43**	**9.26**	21.59	8.16	14.91
	+aware	11.82	23.07	9.08	21.54	9.51	15.00
T-DEC D-ENC	MDML	10.57	18.32	6.87	20.04	10.25	13.21
	+adv	11.36	22.69	7.13	20.88	9.44	14.30
	+aware	11.25	21.94	8.73	20.69	10.34	14.59
T-DEC D-DEC	MDML	5.21	17.56	4.53	9.12	4.36	8.16
	+adv	2.25	17.47	4.89	12.41	5.18	8.44
	+aware	3.37	18.85	4.23	9.35	4.14	7.99
T-ENC D-DEC	MDML	9.39	21.29	8.28	22.06	9.54	14.11
	+adv	10.84	22.92	8.45	22.27	9.10	14.72
	+aware	**12.29**	22.70	8.39	**22.62**	**10.75**	**15.35**

Table 11: BLEU score of En→Cs translation on leave-out domains for multi-domain multilingual (MDML) models and multi-domain bilingual (MDBL) models. +adv and +aware denote MDML models trained with domain-agnostic or domain-aware auxiliary tasks, respectively. The best score on each domain overall and within each tagging group are marked in **bold** and underline respectively.

		LAW	IT	KOR	MED	SUB	AVG	
T-ENC	(I)	SDML	**49.21**	**41.63**	**32.33**	**51.84**	**32.01**	**41.40**
		MDML	45.87	35.76	29.01	47.30	28.30	37.25
	(II)	SDML	1.98	13.29	3.03	12.57	3.11	6.80
		MDML	**23.40**	**27.27**	**13.29**	**31.19**	**13.44**	**21.72**
	(III)	SDML	2.68	14.89	4.26	11.70	5.10	7.73
		MDML	**5.07**	**15.32**	**6.26**	**12.87**	**6.85**	**9.27**
T-DEC	(I)	SDML	**48.42**	**41.36**	**29.50**	**54.00**	**31.88**	**41.03**
		MDML	44.48	30.43	28.53	45.75	27.99	35.44
	(II)	SDML	2.07	14.01	3.95	14.54	4.36	7.79
		MDML	**21.73**	**28.84**	**13.77**	**29.18**	**13.65**	**21.43**
	(III)	SDML	2.66	15.37	4.38	12.94	5.44	8.16
		MDML	**14.57**	**15.57**	**14.39**	**20.52**	**8.61**	**14.73**

Table 12: Average BLEU scores of single-domain multilingual (SDML) and multi-domain multilingual (MDML) on the leave-out domains for three groups: (I) the three seen high-resource language pairs (En-De, En-Fr, De-Fr); (II) the two low-resource language pairs which are seen by MDML but unseen to SDML (En-Cs, En-Pl); and (III) the other five unseen language pairs.

	model	LAW	IT	KOR	MED	SUB	AVG
Adapter-based		23.02	29.37	19.52	28.87	15.51	23.26
T-ENC	MDML	23.09	27.86	22.83	34.19	17.88	25.17
	+adv	28.74	31.14	25.68	40.08	18.89	28.91
	+aware	31.56	**32.00**	26.63	**40.88**	19.62	**30.14**
T-ENC D-ENC	MDML	21.14	26.92	21.84	34.61	17.31	24.36
	+adv	26.09	31.91	26.09	40.72	19.67	28.90
	+aware	27.10	31.85	26.40	40.77	**20.03**	29.23
T-ENC D-DEC	MDML	20.04	24.71	19.57	30.77	15.43	22.10
	+adv	30.14	31.26	26.03	41.04	18.68	29.43
	+aware	**31.61**	31.49	26.56	40.84	19.20	29.94
T-DEC	MDML	25.92	26.81	20.49	34.34	16.56	24.82
	+adv	29.64	30.91	26.22	40.31	18.62	29.14
	+aware	29.45	31.54	**26.81**	40.82	19.01	29.52
T-DEC D-ENC	MDML	24.89	27.23	20.97	34.83	16.85	24.95
	+adv	27.66	31.02	25.54	39.55	19.02	28.56
	+aware	28.15	31.31	25.77	40.27	19.56	29.01
T-DEC D-DEC	MDML	21.24	19.24	16.06	27.42	12.01	19.19
	+adv	20.58	20.60	11.75	24.09	11.40	17.68
	+aware	13.91	17.75	9.49	20.97	9.74	14.37

Table 13: Average BLEU score of MDML-NMT models on each domain across all five leave-one-out scenarios and 20 (seen and unseen) translation tasks. The best score on each domain overall and within each tagging group are marked in **bold** and underline respectively.

		seen (10)						unseen (zero-shot) (10)					
		LAW	IT	KOR	MED	SUB	AVG	LAW	IT	KOR	MED	SUB	AVG
Adapter-based		43.38	36.15	26.92	50.06	27.27	36.76	10.97	23.23	27.69	35.05	12.16	21.82
T-ENC	MDML	43.07	36.66	26.32	49.37	26.28	36.34	4.16	21.07	23.45	22.67	11.12	16.49
	+adv	42.83	35.73	26.60	49.04	25.46	35.93	17.32	28.73	28.58	35.38	14.17	24.84
	+aware	43.36	36.54	27.08	49.84	25.97	36.56	22.88	29.64	29.91	36.62	15.38	26.88
T-ENC D-ENC	MDML	36.02	36.50	22.29	43.00	22.48	32.06	7.01	26.46	22.79	28.34	12.28	19.38
	+adv	43.08	36.41	26.76	49.46	25.99	36.34	10.63	29.77	28.31	36.49	15.02	24.05
	+aware	43.41	36.71	27.02	49.74	26.15	36.61	12.95	29.20	29.25	36.64	15.37	24.68
T-ENC D-DEC	MDML	36.06	36.32	22.24	42.74	21.86	31.85	3.87	21.95	19.36	21.10	9.92	15.24
	+adv	42.80	35.67	26.66	49.10	25.23	35.89	20.29	29.92	29.45	37.91	14.79	26.47
	+aware	43.53	36.50	27.39	49.94	25.84	36.64	23.20	29.72	29.82	36.98	15.32	27.01
T-DEC	MDML	35.30	35.45	21.60	42.39	21.68	31.28	17.02	27.49	20.37	28.55	12.00	21.08
	+adv	42.76	35.95	26.90	49.36	25.56	36.10	18.73	28.04	29.25	36.22	13.33	25.11
	+aware	43.22	36.28	27.33	49.70	25.72	36.45	17.84	29.02	30.25	36.75	14.23	25.62
T-DEC D-ENC	MDML	35.94	35.74	21.64	42.31	22.11	31.55	17.00	28.23	21.68	29.42	11.60	21.59
	+adv	42.08	35.45	26.85	48.55	25.41	35.67	15.87	28.49	29.12	34.95	13.71	24.43
	+aware	43.12	35.62	25.98	48.69	25.90	35.86	16.02	28.99	28.40	36.31	14.62	24.87
T-DEC D-DEC	MDML	33.99	32.09	20.85	39.35	18.75	29.01	9.10	13.17	11.46	16.37	4.82	10.98
	+adv	35.67	27.11	19.13	35.53	17.76	27.04	6.92	15.26	4.74	14.39	5.75	9.41
	+aware	26.89	23.03	17.07	31.83	16.26	23.02	3.14	13.03	2.58	10.93	3.92	6.72

Table 14: Average BLEU score on other domains, i.e. excluding the leave-out domains, for different translation tasks. We categorise 20 translation direction into *seen* where the translation direction in which training data are available, otherwise *unseen*. The number in parentheses shows how many translation directions in the corresponding category.