

# Meta-Learning for Few-Shot NMT Adaptation

**Amr Sharaf**

University of Maryland  
amr@cs.umd.edu

**Hany Hassan**

Microsoft  
hanyh@microsoft.com

**Hal Daumé III**

Microsoft Research &  
University of Maryland  
me@hal3.name

## Abstract

We present META-MT, a meta-learning approach to adapt Neural Machine Translation (NMT) systems in a few-shot setting. META-MT provides a new approach to make NMT models easily adaptable to many target domains with the minimal amount of in-domain data. We frame the adaptation of NMT systems as a meta-learning problem, where we learn to adapt to new unseen domains based on simulated offline meta-training domain adaptation tasks. We evaluate the proposed meta-learning strategy on ten domains with general large scale NMT systems. We show that META-MT significantly outperforms classical domain adaptation when very few in-domain examples are available. Our experiments shows that META-MT can outperform classical fine-tuning by up to 2.5 BLEU points after seeing only 4,000 translated words (300 parallel sentences).

## 1 Introduction

Neural Machine Translation (NMT) systems (Bahdanau et al., 2016; Sutskever et al., 2014) are usually trained on large general-domain parallel corpora to achieve state-of-the-art results (Barrault et al., 2019). Unfortunately, these generic corpora are often qualitatively different from the target domain of the translation system. Moreover, NMT models trained on one domain tend to perform poorly when translating sentences in a significantly different domain (Koehn and Knowles, 2017; Chu and Wang, 2018). A widely used approach for adapting NMT is *domain adaptation by fine-tuning* (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Sennrich et al., 2016), where a model is first trained on general-domain data and then adapted by continuing the training on a smaller amount of in-domain data. This approach often leads to empirical improvements in the targeted

domain; however, it falls short when the amount of in-domain training data is insufficient, leading to model over-fitting and catastrophic forgetting, where adapting to a new domain leads to degradation on the general-domain (Thompson et al., 2019). Ideally, we would like to have a model that is easily adaptable to many target domains with minimal amount of in-domain data.

We present a meta-learning approach that *learns* to adapt neural machine translation systems to new domains given only a small amount of training data in that domain. To achieve this, we simulate many domain adaptation tasks, on which we use a *meta-learning* strategy to learn how to adapt. Specifically, based on these simulations, our proposed approach, META-MT (Meta-learning for Machine Translation), learns model parameters that should generalize to future (real) adaptation tasks (§3.1).

At training time (§3.2), META-MT simulates many small-data domain adaptation tasks from a large pool of data. Using these tasks, META-MT simulates what would happen after fine-tuning the model parameters to each such task. It then uses this information to compute parameter updates that will lead to efficient adaptation during deployment. We optimize this using the Model Agnostic Meta-Learning algorithm (MAML) (Finn et al., 2017).

The contribution of this paper is as follows: first, we propose a new approach that enables NMT systems to effectively adapt to a new domain using few-shots learning. Second, we show what models and conditions enable meta-learning to be useful for NMT adaptation. Finally, We evaluate META-MT on ten different domains, showing the efficacy of our approach. To the best of our knowledge, this is the first work on adapting large scale NMT systems in a few-shot learning setup<sup>1</sup>.

<sup>1</sup>**Code Release:** We make the code publicly available online: <https://www.dropbox.com/s/jguxb75utg1dmx1/meta-mt.zip?dl=0>

## 2 Related Work

Our goal for few-shot NMT adaptation is to adapt a pre-trained NMT model (e.g. trained on general domain data) to new domains (e.g. medical domain) with a small amount of training examples. [Chu et al. \(2018\)](#) surveyed several recent approaches that address the shortcomings of traditional fine-tuning when applied to domain adaptation. Our work distinguishes itself from prior work by learning to fine-tune with tiny amounts of training examples.

Most recently, [Bapna et al. \(2019\)](#) proposed a simple approach for adaptation in NMT. The approach consists of injecting task specific adapter layers into a pre-trained model. These adapters enable the model to adapt to new tasks as it introduces a bottleneck in the architecture that makes it easier to adapt. Our approach uses a similar model architecture, however, instead of injecting a new adapter for each task separately, META-MT uses a single adapter layer, and meta-learns a better initialization for this layer that can easily be fine-tuned to new domains with very few training examples.

Similar to our goal, [Michel and Neubig \(2018\)](#) proposed a space efficient approach to adaptation that learns domain specific biases to the output vocabulary. This enables large-scale personalization for NMT models when small amounts of data are available for a lot of different domains. However, this approach assumes that these domains are static and known at training time, while META-MT can dynamically generalize to totally new domains, previously unseen at meta-training time.

Several approaches have been proposed for lightweight adaptation of NMT systems. [Vilar \(2018\)](#) introduced domain specific gates to control the contribution of hidden units feeding into the next layer. However, [Bapna et al. \(2019\)](#) showed that this introduced a limited amount of per-domain capacity; in addition, the learned gates are not guaranteed to be easily adaptable to unseen domains. [Khayrallah et al. \(2017\)](#) proposed a lattice search algorithm for NMT adaptation, however, this algorithm assumes access to lattices generated from a phrase based machine translation system.

Our meta-learning strategy mirrors that of [Gu et al. \(2018\)](#) in the low resource translation setting, as well as [Wu et al. \(2019\)](#) for cross-lingual named entity recognition with minimal resources, [Mi et al. \(2019\)](#) for low-resource natural language generation in task-oriented dialogue systems, and [Dou et al. \(2019\)](#) for low-resource natural language un-

derstanding tasks. To the best of our knowledge, this is the first work using meta-learning for few-shot NMT adaptation.

## 3 Approach: Meta-Learning for Few-Shot NMT Adaptation

Neural Machine Translation systems are not robust to domain shifts ([Chu and Wang, 2018](#)). It is a highly desirable characteristic of the system to be adaptive to any domain shift using weak supervision without degrading the performance on the general domain. This dynamic adaptation task can be viewed naturally as a learning-to-learn (meta-learning) problem: how can we train a global model that is capable of using its previous experience in adaptation to learn to adapt faster to unseen domains? A particularly simple and effective strategy for adaptation is fine-tuning: the global model is adapted by training on in-domain data. One would hope to improve on such a strategy by decreasing the amount of required in-domain data. META-MT takes into account information from previous adaptation tasks, and aims at learning how to update the global model parameters, so that the resulting learned parameters after meta-learning can be adapted faster and better to previously unseen domains via a weakly supervised fine-tuning approach on a tiny amount of data.

Our goal in this paper is to learn how to adapt a neural machine translation system from experience. The training procedure for META-MT uses offline simulated adaptation problems to learn model parameters  $\theta$  which can adapt faster to previously unseen domains. In this section, we describe META-MT, first by describing how it operates at test time when applied to a new domain adaptation task (§3.1), and then by describing how to train it using offline simulated adaptation tasks (§3.2).

### 3.1 Test Time Behavior of META-MT

At test time, META-MT adapts a pre-trained NMT model to a new given domain. The adaptation is done using a small in-domain data that we call the *support set* and then tested on the new domain using a *query set*. More formally, the model parametrized by  $\theta$  takes as input a new adaptation task  $T$ . This is illustrated in [Figure 1](#): the adaptation task  $T$  consists of a standard domain adaptation problem:  $T$  includes a support set  $T_{\text{support}}$  used for training the fine-tuned model, and a query set  $T_{\text{query}}$  used for evaluation. We’re particularly

	Support Set	Query Set
Meta-training	<b>Domain: Books</b> En: Chapter I De: Erstes Kapitel	En: Conclusion De: Schluß
	<b>Domain: TED Talks</b> En: A garden in my apartment De: Ein Garten in meiner Wohnung	En: Following the mercury trail De: Der Spur des Quecksilbers folgen
Meta-testing	<b>Domain: Medical</b> En: Swirl gently De: Schwenken Sie behutsam	En: Remove the filter needle De: Die Filtrnadel entfernen
	<b>Domain: News</b> En: European Inflation De: Europäischer Inflation	En: Red Tide Update De: Neues zu den Algenblüten.

Figure 1: Example meta-learning set-up for few-shot NMT adaptation. The top represents the meta-training set  $\mathcal{D}_{\text{meta-train}}$ , where inside each box is a separate dataset  $T$  that consists of the support set  $T_{\text{support}}$  (left side of dashed line) and the query set  $T_{\text{query}}$  (right side of dashed line). In this illustration, we are considering the books and TED talks domains for meta-training. The meta-test set  $\mathcal{D}_{\text{meta-test}}$  is defined in the same way, but with a different set of domains not present in any of the datasets in  $\mathcal{D}_{\text{meta-train}}$ : Medical and News.

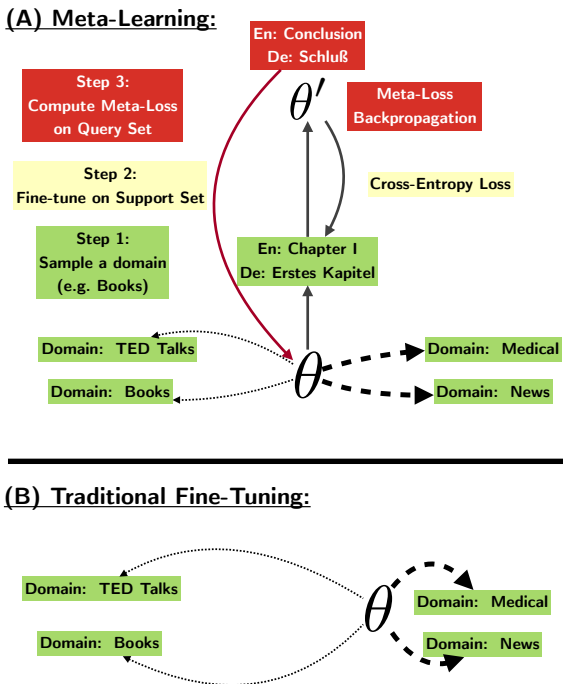


Figure 2: **[Top-A]** a training step of META-MT. **[Bottom-B]** Differences between meta-learning and Traditional fine-tuning. Wide lines represent high resource domains (Medical, News), while thin lines represent low-resource domains (TED, Books). Traditional fine-tuning may favor high-resource domains over low-resource ones while meta-learning aims at learning a good initialization that can be adapted to any domain with minimal training samples. <sup>2</sup>

interested in the distribution of tasks  $P(\mathcal{T})$  where the support and query sets are very small. In our experiments, we restrict the size of these sets to only few hundred parallel training sentences. We consider support sets of sizes: 4k to 64k source words (i.e.  $\sim 200$  to 3200 sentences). At test time, the meta-learned model  $\theta$  interacts with the world as follows (Figure 2):

- Step 1:** The world draws an adaptation task  $T$  from a distribution  $P$ ,  $T \sim P(\mathcal{T})$ ;
- Step 2:** The model adapts from  $\theta$  to  $\theta'$  by fine-tuning on the task's support set  $T_{\text{support}}$ ;
- Step 3:** The fine-tuned model  $\theta'$  is evaluated on the query set  $T_{\text{query}}$ .

Intuitively, meta-training should optimize for a representation  $\theta$  that can quickly adapt to new tasks, rather than a single individual task.

### 3.2 Training META-MT via Meta-learning

The meta-learning challenge is: how do we learn a good representation  $\theta$ ? We initialize  $\theta$  by training an NMT model on global-domain data. In addition, we assume access to meta-training tasks on which we can train  $\theta$ ; these tasks must include support/query pairs, where we can simulate a domain adaptation setting by fine-tuning on the support set and then evaluating on the query. This is a weak assumption: in practice, we use purely simulated data as this meta-training data. We construct this data as follows: given a parallel corpus for the desired language pair, we randomly sample training example to form a few-shot adaptation task. We build tasks of 4k, 8k, 16k, 32k, and 64k training words. Under this formulation, it's natural to think of  $\theta$ 's learning process as a process to learn a good parameter initialization for fast adaptation, for which a class of learning algorithms to consider are Model-agnostic Meta-Learning (MAML) and it's first order approximations like First-order MAML (FoMAML) (Finn et al., 2017) and Reptile (Nichol et al., 2018).

Informally, at training time, META-MT will treat one of these simulated domains  $T$  as if it were a domain adaptation dataset. At each time step, it will update the current model representation from  $\theta$  to  $\theta'$  by fine-tuning on  $T_{\text{support}}$  and then ask: what is the meta-learning loss estimate given  $\theta$ ,  $\theta'$ , and  $T_{\text{query}}$ ? The model representation  $\theta$  is then updated to minimize this meta-learning loss. More formally, in meta-learning, we assume access to a

<sup>2</sup>colorblind friendly palette was selected from Neuwirth and Brewer (2014).

---

**Algorithm 1** META-MT (trained model  $f_\theta$ , meta-training dataset  $\mathcal{D}_{\text{meta-train}}$ , learning rates  $\alpha, \beta$ )

---

```

1: while not done do
2:   Sample a batch of domain adaptation tasks
    $T \sim \mathcal{D}_{\text{meta-train}}$ 
3:   for all  $T_i \in T$  do
4:     Evaluate  $\nabla_\theta L_{T_i}(f_\theta)$  on the support set
      $T_{i,\text{support}}$ 
5:     Compute adapted parameters with gradient
     descent:  $\theta'_i = \theta - \alpha \nabla_\theta L_{T_i}(f_\theta)$ 
6:   end for
7:   Update  $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_i \in T} L_{T_i}(f_{\theta'_i})$  on the
   query set  $T_{i,\text{query}} \forall T_i \in T$ 
8: end while

```

---

distribution  $P$  over different tasks  $\mathcal{T}$ . From this, we can sample a meta-training dataset  $\mathcal{D}_{\text{meta-train}}$ . The meta-learning problem is then to estimate  $\theta$  to minimize the meta-learning loss on  $\mathcal{D}_{\text{meta-train}}$ .

The meta-learning algorithm we use is MAML by Finn et al. (2017), and is instantiated for the meta-learning to adapt NMT systems in Alg 1. MAML considers a model represented by a parametrized function  $f_\theta$  with parameters  $\theta$ . When adapting to a new task  $T$ , the model’s parameters  $\theta$  become  $\theta'$ . The updated vector  $\theta'$  is computed using one or more gradient descent updates on the task  $T$ . For example, when using one gradient update:

$$\theta' = \theta - \alpha \nabla_\theta L_T(f_\theta) \quad (1)$$

where  $\alpha$  is the learning rate and  $L$  is the task loss function. The model parameters are trained by optimizing for the performance of  $f_{\theta'}$  with respect to  $\theta$  across tasks sampled from  $P(\mathcal{T})$ . More concretely, the meta-learning objective is:

$$\min_{\theta} \sum_{T \sim P(\mathcal{T})} L_T(f_{\theta'}),$$

$$L_T(f_{\theta'}) = L_T(f_{\theta - \alpha \nabla_\theta L_T(f_\theta)}) \quad (2)$$

Following the MAML template, META-MT operates in an iterative fashion, starting with a trained NMT model  $f_\theta$  and improving it through optimizing the meta-learning loss from Eq 2 on the meta-training dataset  $\mathcal{D}_{\text{meta-train}}$ . Over learning rounds, META-MT selects a random batch of training tasks from the meta-training dataset and simulates the test-time behavior on these tasks (Line 2). The core functionality is to observe how the current

model representation  $\theta$  is adapted for each task in the batch, and to use this information to improve  $\theta$  by optimizing the meta-learning loss (Line 7). META-MT achieves this by simulating a domain adaptation setting by fine-tuning on the task specific support set (Line 4). This yields, for each task  $T_i$ , a new adapted set of parameters  $\theta'_i$  (Line 5). These parameters are evaluated on the query sets for each task  $T_{i,\text{query}}$ , and a meta-gradient w.r.t the original model representation  $\theta$  is used to improve  $\theta$  (Line 7).

Our pre-trained baseline NMT model  $f_\theta$  is a sequence to sequence model that parametrizes the conditional probability of the source and target sequences as an encoder-decoder architecture using self-attention Transformer models (Vaswani et al., 2017)).

## 4 Experimental Setup and Results

We seek to answer the following questions experimentally:

1. How does META-MT compare empirically to alternative adaptation strategies? (§4.4)
2. What is the impact of the support and the query sizes used for meta-learning? (§4.5)
3. What is the effect of the NMT model architecture on performance? (§4.6)

In our experiments, we train META-MT only on simulated data, where we simulate a few-shot domain adaptation setting as described in §3.2. This is possible because META-MT learns model parameters  $\theta$  that can generalize to future adaptation tasks by optimizing the meta-objective function in Eq 2.

We train and evaluate META-MT on a collection of ten different datasets. All of these datasets are collected from the Open Parallel Corpus (OPUS) (Tiedemann, 2012), and are publicly available online. The datasets cover a variety of diverse domains that should enable us to evaluate our proposed approach. The datasets we consider are:

1. Bible: a parallel corpus created from translations of the Bible (Christodouloupoulos and Steedman, 2015).
2. European Central Bank: website and documentations from the European Central Bank.
3. KDE: a corpus of KDE4 localization files.
4. Quran: a collection of Quran translations compiled by the Tanzil project.
5. WMT news test sets: a parallel corpus of

News Test Sets provided by WMT.

6. Books: a collection of copyright free books.
7. European Medicines Agency (EMA): a parallel corpus made out of PDF documents from the European Medicines Agency.
8. Global Voices: parallel news stories from the Global Voices web site.
9. Medical (ufal-Med): the UFAL medical domain dataset from [Yepes et al. \(2017\)](#).
10. TED talks: talk subtitles from [Duh \(2018\)](#).

We simulate the few-shot NMT adaptation scenarios by randomly sub-sampling these datasets with different sizes. We sample different data sets with sizes ranging from 4k to 64k training words (i.e.  $\sim 200$  to 3200 sentences). This data is the only data used for any given domain across all adaptation setups. It is worth noting that different datasets have a wide range of sentence lengths. We opted to sample using number of words instead of number of sentences to avoid introducing any advantages for domains with longer sentences.

#### 4.1 Domain Adaptation Approaches

Our experiments aim to determine how META-MT compares to standard domain adaptation strategies. In particular, we compare to:

- (A) **No fine-tuning:** The non-adaptive baseline. Here, the pre-trained model is evaluated on the meta-test and meta-validation datasets (see [Figure 1](#)) without any kind of adaptation.
- (B) **Fine-tuning on a single task:** The domain adaptation by fine-tuning baseline. For a single adaptation task  $T$ , this approach performs domain adaptation by fine-tuning only on the support set  $T_{\text{support}}$ . For instance, if  $|T_{\text{support}}| = K$  words, we fine tune the pre-trained model  $f_{\theta}$  only on  $K$  training words to show how classical fine-tuning behaves in few-shot settings.
- (C) **Fine-tuning on meta-train:** Similar to (B), however, this approach fine-tunes on much more data. This approach fine-tunes on all the support sets used for meta-training:  $\{T_{\text{support}}, \forall T \in \mathcal{D}_{\text{meta-train}}\}$ . The goal of this baseline is to ensure that META-MT doesn't get an additional advantage by training on more data during the meta-training phase. For instance, if we are using  $N$  adaptation tasks each with a support set of size  $K$ , this will be using  $N * K$  words for classical fine-tuning. This establishes a fair baseline to evaluate how

classical fine-tuning would perform using the same data albeit in a different configuration.

- (D) **META-MT:** Our proposed approach from [Alg 1](#). In this setup, we use  $N$  adaptation tasks  $T$  in  $\mathcal{D}_{\text{meta-train}}$ , each with a support set of size  $K$  words to perform Meta-Learning. Second order meta-gradients are ignored to decrease the computational complexity.

#### 4.2 Model Architecture and Implementation Details

We use the Transformer Model ([Vaswani et al., 2017](#)) implemented in fairseq ([Ott et al., 2019](#)). In this work, we use a transformer model with a modified architecture that can facilitate better adaptation. We use “Adapter Modules” ([Houlsby et al., 2019](#); [Bapna et al., 2019](#)) which introduce an extra layer after each transformer block that can enable more efficient tuning of the models. Following [Bapna et al. \(2019\)](#), we augment the Transformer model with feed-forward adapters: simple single hidden-layer feed-forward networks, with a nonlinear activation function between the two projection layers. These adapter modules are introduced after the Layer Norm and before the residual connection layers. It is composed of a down projection layer, followed by a ReLU, followed by an up projection layer. This bottle-necked module with fewer parameters is very attractive for domain adaptation as we will discuss in [§4.6](#). These modules are introduced after every layer in both the encoder and the decoder. All experiments are based on the “base” transformer model with six blocks in the encoder and decoder networks. Each encoder block contains a self-attention layer, followed by two fully connected feed-forward layers with a ReLU non-linearity between them. Each decoder block contains self-attention, followed by encoder-decoder attention, followed by two fully connected feed-forward layers with a ReLU non-linearity between them.

We use word representations of size 512, feed-forward layers with inner dimensions 2,048, multi-head attention with 8 attention heads, and adapter modules with 32 hidden units. We apply dropout ([Srivastava et al., 2014](#)) with probability 0.1. The model is optimized with Adam ([Kingma and Ba, 2014](#)) using  $\beta_1 = 0.9$ ,  $\beta_2 = 0.98$ , and a learning rate  $\alpha = 7e - 4$ . We use the same learning rate schedule as [Vaswani et al. \(2017\)](#) where the

learning rate increases linearly for 4,000 steps to  $7e - 4$ , after which it is decayed proportionally to the inverse square root of the number of steps. For meta-learning, we used a meta-batch size of 1. We optimized the meta-learning loss function using Adam with a learning rate of  $1e - 5$  and default parameters for  $\beta_1, \beta_2$ .

All data is pre-processed with joint sentence-pieces (Kudo and Richardson, 2018) of size 40k. In all cases, the baseline machine translation system is a neural English to German (En-De) transformer model (Vaswani et al., 2017), initially trained on 5.2M sentences filtered from the standard parallel data (Europarl-v9, CommonCrawl, NewsCommentary-v14, wikititles-v1 and Rapid-2019) from the WMT-19 shared task (Barrault et al., 2019). We use WMT14 and WMT19 newtests as validation and test sets respectively. The baseline system scores 37.99 BLEU on the full WMT19 newtest which compares favorably with strong single system baselines at WMT19 shared task (Ng et al., 2019; Junczys-Dowmunt, 2019).

For meta-learning, we use the MAML algorithm as described in Alg 1. To minimize memory consumption, we ignored the second order gradient terms from Eq 2. We implement the First-Order MAML approximation (FoMAML) as described in Finn et al. (2017). We also experimented with the first-order meta-learning algorithm Reptile (Nichol et al., 2018). We found that since Reptile doesn’t directly account for the performance on the task query set, along with the large model capacity of the Transformer architecture, it can easily over-fit to the support set, thus achieving almost perfect performance on the support, while the performance on the query degrades significantly. Even after performing early stopping on the query set, Reptile didn’t account correctly for learning rate scheduling, and finding suitable learning rates for optimizing the meta-learner and the task adaptation was difficult. In our experiments, we found it essential to match the behavior of the dropout layers when computing the meta-objective function in Eq 2 with the test-time behavior described in §3.1. In particular, the model has to run in “*evaluation mode*” when computing the loss on the task query set to match the test-time behavior during evaluation.

### 4.3 Evaluation Tasks and Metrics

Our experimental setup operates as follows: using a collection of simulated machine translation

adaptation tasks, we train an NMT model  $f_\theta$  using META-MT (Alg 1). This model learns to adapt faster to new domains, by fine-tuning on a tiny support set. Once  $f_\theta$  is learned and fixed, we follow the test-time behavior described in §3.1. We evaluate META-MT on the collection of ten different domains described in §4. We simulate domain adaptation problems by sub-sampling tasks with 4k English tokens for the support set, and 32k tokens for the query set. We study the effect of varying the size of the query and the support sets in §4.5. We use  $N = 160$  tasks for the meta-training dataset  $\mathcal{D}_{\text{meta-train}}$ , where we sample 16 tasks from each of the ten different domains. We use a meta-validation  $\mathcal{D}_{\text{meta-test}}$  and meta-test  $\mathcal{D}_{\text{meta-test}}$  sets of size 10, where we sample a single task from each domain. We report the mean and standard-deviation over three different meta-test sets. For evaluation, we use BLEU (Papineni et al., 2002). We measure case-sensitive de-tokenized BLEU with SacreBLEU (Post, 2018). All results use beam search with a beam of size five.

### 4.4 Experimental Results

Here, we describe our experimental results comparing the several algorithms from §4.1. The overall results are shown in Table 1 and Figure 3. Table 1 shows the BLEU scores on the meta-test dataset for all the different approaches across the ten domains. From these results we draw the following conclusions:

1. The pre-trained En-De NMT model performs well on general domains. For instance, BLEU for WMT-News<sup>3</sup>, GlobalVoices, and ECB is at least 26 points. However, performance degrades on closed domains like Books, Quran, and Bible. [Column A].
2. Domain adaptation by fine-tuning on a single task doesn’t improve the BLEU score. This is expected, since we’re only fine-tuning on 4k tokens (i.e.  $\sim 200 - 300$  sentences) [A vs B].
3. Significant leverage is gained by increasing the amount of fine-tuning data. Fine-tuning on all the available data used for meta-learning improves the BLEU score significantly across all domains. [B vs C]. To put this into perspective, this setup is tuned on all data aggregated from all tasks:  $160 * 4k$  words which is approximately 40K sentences.

<sup>3</sup>This is subset of the full test set to match the sizes of query sets from other domains

Domain	A. No fine-tuning	B. Fine-tuning on task	C. Fine-tuning on meta-train	D. META-MT
Books	11.338 ± 0.25	11.34 ± 0.24	<u>12.49 ± 0.15</u>	<b>12.92 ± 0.94</b>
Tanzil	11.25 ± 0.04	11.33 ± 0.04	<u>13.62 ± 0.05</u>	<b>15.16 ± 0.94</b>
Bible	12.93 ± 0.93	12.95 ± 0.94	<u>17.19 ± 0.54</u>	<b>24.70 ± 0.61</b>
KDE4	20.53 ± 0.34	20.54 ± 0.32	26.61 ± 0.16	<b>27.26 ± 0.36</b>
Med	19.30 ± 0.24	19.53 ± 0.28	28.31 ± 0.04	<b>29.59 ± 0.05</b>
GlobalVoices	25.10 ± 0.11	25.17 ± 0.23	25.83 ± 0.25	<b>26.03 ± 0.13</b>
WMT-News	26.93 ± 0.36	26.92 ± 0.48	<b>27.26 ± 0.55</b>	27.23 ± 0.12
TED	27.69 ± 0.05	27.85 ± 0.06	28.78 ± 0.03	<b>29.37 ± 0.03</b>
EMEA	27.81 ± 0.01	27.79 ± 0.05	29.77 ± 0.59	<b>32.38 ± 0.01</b>
ECB	29.18 ± 0.03	29.21 ± 0.04	31.18 ± 0.01	<b>33.23 ± 0.40</b>

Table 1: BLEU scores on meta-test split for different approaches evaluated across ten domains. Best results are highlighted in bold, results with-in two standard-deviations of the best value are underlined.

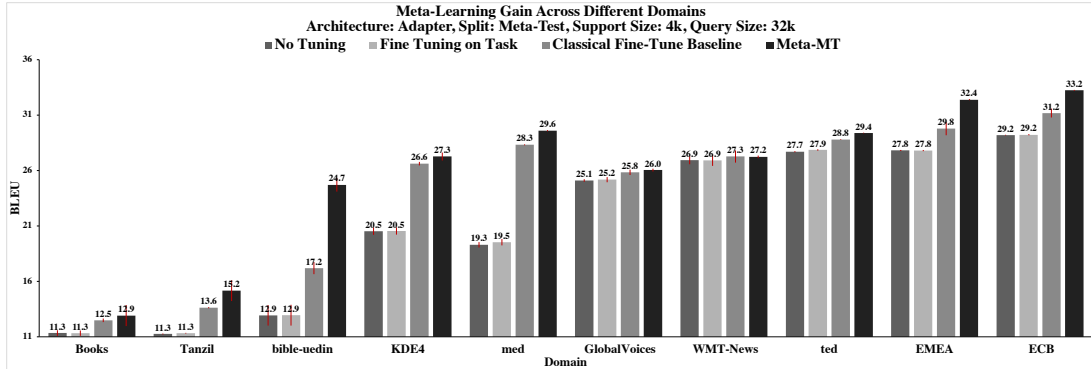


Figure 3: BLEU scores on meta-test split for different approaches evaluated across ten domains.

- META-MT outperforms alternative domain adaptation approaches on all domains with negligible degradation on the baseline domain. META-MT is better than the non-adaptive baseline [A vs D], and succeeds in learning to adapt faster given the same amount of fine-tuning data [B vs D, C vs D]. Both **Fine-tuning on meta-train** [C] and **META-MT** [D] have access to exactly the same amount of training data, and both use the same model architecture. The difference however is in the learning algorithm. META-MT uses MAML (Alg 1) to optimize the meta-objective function in Eq 2. This ensures that the learned model initialization can easily be fine-tuned to new domains with very few examples.

#### 4.5 Impact of Adaptation Task Size

To evaluate the effectiveness of META-MT when adapting with small in-domain corpora, we further compare the performance of META-MT with classical fine-tuning on varying amounts of training data per adaptation task. In Figure 4 we plot the overall adaptation performance on the ten domains when using different data sizes for the support set. In this experiment, the only parameter that

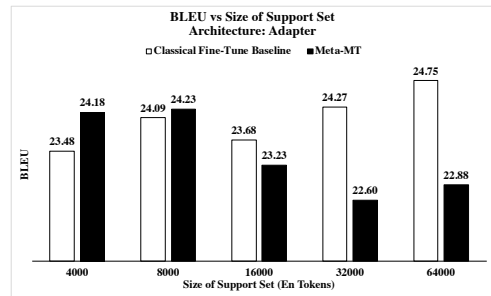


Figure 4: META-MT and fine-tuning adaptation performance on the meta-test set  $\mathcal{D}_{\text{meta-test}}$  vs different support set sizes per adaptation task.

varies is the size of the task support set  $T_{\text{support}}$ . We fix the size of the query set per task to  $16k$  tokens, and we vary the size of the support set from  $4k$  to  $64k$ . To ensure that the total amount of meta-training data  $\mathcal{D}_{\text{meta-train}}$  is the same, we use  $N = 160$  tasks for meta-training when the support size  $T_{\text{support}}$  is  $4k$ ,  $N = 80$  tasks when the support size is  $8k$ ,  $N = 40$  tasks for support size of  $16k$ ,  $N = 20$  tasks when the support size is  $32k$ , and finally  $N = 10$  meta-training tasks when the support size is  $64k$ . This controlled setup ensures that no setting has any advantage by getting access to additional amounts of training data. We notice

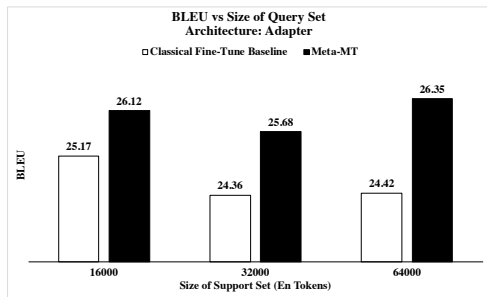


Figure 5: META-MT and fine-tuning adaptation performance on the meta-test set  $\mathcal{D}_{\text{meta-test}}$  vs different query set sizes per adaptation task.

that for reasonably small size of the support set ( $4k$  and  $8k$ ), META-MT outperforms the classical fine-tuning baseline. However, when the data size increase ( $16k$  to  $64k$ ), META-MT is outperformed by the fine-tuning baseline. This happens because for a larger support size, e.g.  $64k$ , we only have access to 10 meta-training tasks in  $\mathcal{D}_{\text{meta-train}}$ , this is not enough to generalize to new unseen adaptation tasks, and META-MT over-fits to the training tasks from  $\mathcal{D}_{\text{meta-train}}$ , however, the performance degrades and doesn't generalize to  $\mathcal{D}_{\text{meta-test}}$ .

To understand more directly the impact of the query set on META-MT's performance, in Figure 5 we show META-MT and fine-tuning adaptation performance on the meta-test set  $\mathcal{D}_{\text{meta-test}}$  on varying sizes for the query set. We fix the support size to  $4k$  and vary the query set size from  $16k$  to  $64k$ . We observe that the edge of improvement of META-MT over fine-tuning adaptation increases as we increase the size of the query set. For instance, when we use a query set of size  $64k$ , META-MT outperforms fine-tuning by 1.93 BLEU points, while the improvement is only 0.95 points when the query set is  $16k$ .

#### 4.6 Impact of Model Architecture

In our experiments, we used the Adapter Transformer architecture (Bapna et al., 2019). This architecture fixes the parameters of the pre-trained Transformer model, and only adapts the feed-forward adapter module. Our model included  $\sim 66M$  parameters, out of which we adapt only  $405K$  (only 0.6%). We found this adaptation strategy to be more robust to meta-learning. To better understand this, Figure 6 shows the BLEU scores for the two different model architectures. We find that while the meta-learned Transformer architecture (Right) slightly outperforms the Adapter model (Left), it

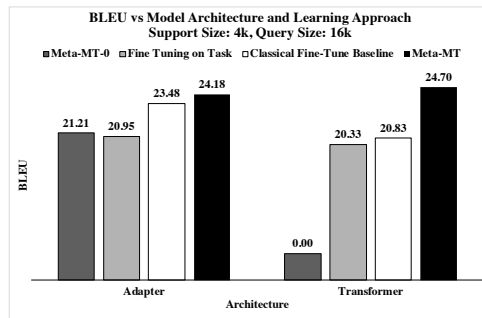


Figure 6: BLEU scores reported for two different model architectures: Adapter Transformer (Bapna et al., 2019) (Left), and the Transformer base architecture (Vaswani et al., 2012) (Right).

suffers from catastrophic forgetting: **META-MT-0** shows the zero-shot BLEU score before fine-tuning the task on the support set. For the Transformer model, the score drops to zero and then quickly improves once the parameters are tuned on the support set. This is undesirable, since it hurts the performance of the pre-trained model, even on the general domain data. We notice that the Adapter model doesn't suffer from this problem.

## 5 Conclusion

We presented META-MT, a meta-learning approach for few shot NMT adaptation. We formulated few shot NMT adaptation as a meta-learning problem, and presented a strategy that learns better parameters for NMT systems that can be easily adapted to new domains. We validated the superiority of META-MT to alternative domain adaptation approaches. META-MT outperforms alternative strategies in most domains using only a small fraction of fine-tuning data.

## Acknowledgements

The authors would like to thank members of the Microsoft Machine Translation Team as well as members of the Computational Linguistics and Information Processing (CLIP) lab for reviewing earlier versions of this work. Part of this work was conducted when the first author was on a summer internship with Microsoft Research. This material is based upon work supported by the National Science Foundation under Grant No. 1618193. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



## References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. [Neural machine translation by jointly learning to align and translate](#). *arXiv preprint arXiv:1409.0473*.
- Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*.
- Loïc Barrault, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.
- Christos Christodoulopoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.
- Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2018. A comprehensive empirical comparison of domain adaptation methods for neural machine translation. *Journal of Information Processing*, 26:529–538.
- Chenhui Chu and Rui Wang. 2018. [A survey of domain adaptation for neural machine translation](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. [Investigating meta-learning algorithms for low-resource natural language understanding tasks](#). 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 1192–1197, Hong Kong, China. Association for Computational Linguistics.
- Kevin Duh. 2018. The multitarget ted talks task. <http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/>.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. [Model-agnostic meta-learning for fast adaptation of deep networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia. PMLR.
- Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *ArXiv*, abs/1612.06897.
- Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. [Meta-learning for low-resource neural machine translation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzekski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for NLP](#). *CoRR*, abs/1902.00751.
- Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *WMT*.
- Huda Khayrallah, Gaurav Kumar, Kevin Duh, Matt Post, and Philipp Koehn. 2017. [Neural lattice search for domain adaptation in machine translation](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 20–25, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- 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.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*.
- Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. 2019. [Meta-learning for low-resource natural language generation in task-oriented dialogue systems](#). In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI’19*, pages 3151–3157. AAAI Press.
- Paul Michel and Graham Neubig. 2018. [Extreme adaptation for personalized neural machine translation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.
- Erich Neuwirth and R Color Brewer. 2014. Colorbrewer palettes. *R package version*, pages 1–1.
- Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. [Facebook FAIR’s WMT19 news translation task submission](#).

- In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.
- Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.
- 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*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.
- Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of 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 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey. European Language Resources Association (ELRA).
- Ashish Vaswani, Liang Huang, and David Chiang. 2012. Smaller alignment models for better translations: Unsupervised word alignment with the l0-norm. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–319, Jeju Island, Korea. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- David Vilar. 2018. Learning hidden unit contribution for adapting neural machine translation models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 500–505, New Orleans, Louisiana. Association for Computational Linguistics.
- Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F Karlsson, Biqing Huang, and Chin-Yew Lin. 2019. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. *arXiv preprint arXiv:1911.06161*.
- Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondrej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, et al. 2017. Findings of the wmt 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247.

# Supplementary Material For: Meta-Learning for Few-Shot NMT Adaptation

## A Background

### A.1 Neural Machine Translation

Neural Machine Translation (NMT) is a sequence to sequence model that parametrizes the conditional probability of the source and target sequences as a neural network following encoder-decoder architecture (Bahdanau et al., 2016; Sutskever et al., 2014). Initially, the encode-decoder architecture was represented by recurrent networks. Currently, this has been replaced by self-attention models aka Transformer models (Vaswani et al., 2017)). Currently, Transformer models achieves state-of-the-art performance in NMT as well as many other language modeling tasks. While transformers models are performing quite well on large scale NMT tasks, the models have huge number of parameters and require large amount of training data which is really prohibitive for adaptation tasks especially in few-shot setup like ours.

### A.2 Few Shots Domain Adaptation

Traditional domain adaptation for NMT models assumes the availability of relatively large amount of in domain data. For instances most of the related work utilizing traditional fine-tuning experiment with hundred-thousand sentences in-domain. This setup is quite prohibitive, since practically the domain can be defined by few examples. In this work we focus on few-shot adaptation scenario where we can adapt to a new domain not seen during training time using just couple of hundreds of in-domain sentences. This introduces a new challenge where the models have to be quickly responsive to adaptation as well as robust to domain shift. Since we focus on the setting in which very few in-domain data is available, this renders many traditional domain adaptation approaches inappropriate.

### A.3 Meta-Learning

Meta-learning or Learn-to-Learn is widely used for few-shot learning in many applications where a model trained for a particular task can learn another task with a few examples. A number of approaches are used in Meta-learning, namely: Model-agnostic Meta-Learning (MAML) and its first order approximations like First-order MAML (FoMAML) (Finn et al., 2017) and Reptile (Nichol et al., 2018). In

Domain	# sentences	# En Tokens
bible-uedin	62195	1550431
ECB	113174	3061513
KDE4	224035	1746216
Tanzil	537128	9489824
WMT-News	912212	5462820
Books	51467	1054718
EMEA	1108752	12322425
GlobalVoices	66650	1239921
ufal-Med	140600	5527010
TED	51368	1060765

Table 2: Dataset statistics for different domains.

this work, we focus on using MAML to enable few-shots adaptation of NMT transformer models.

## B Statistics of in-domain data sets

Table 2 lists the sizes of various in-domain datasets from which we sample our in-domain data to simulate the few-shot adaptation setup.