Iterative Dual Domain Adaptation for Neural Machine Translation

Jiali Zeng[1], **Yang Liu**[2], **Jinsong Su**[1]*, **Yubin Ge**[3], **Yaojie Lu**[4], **Yongjing Yin**[1], **Jiebo Luo**[5]
[1]Xiamen University, Xiamen, China [2]Tsinghua University, Beijing, China
[3]University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[4]Institute of Software, Chinese Academy of Sciences, Beijing, China
[5]Department of Computer Science, University of Rochester, Rochester NY 14627, USA
lemon@stu.xmu.edu.cn liuyang2011@tsinghua.edu.cn
jssu@xmu.edu.cn

Abstract

Previous studies on the domain adaptation for neural machine translation (NMT) mainly focus on the one-pass transferring out-of-domain translation knowledge to in-domain NMT model. In this paper, we argue that such a strategy fails to fully extract the domain-shared translation knowledge, and repeatedly utilizing corpora of different domains can lead to better distillation of domain-shared translation knowledge. To this end, we propose an iterative dual domain adaptation framework for NMT. Specifically, we first pre-train in-domain and out-of-domain NMT models using their own training corpora respectively, and then iteratively perform bidirectional translation knowledge transfer (from in-domain to out-of-domain and then vice versa) based on knowledge distillation until the in-domain NMT model convergences. Furthermore, we extend the proposed framework to the scenario of multiple out-of-domain training corpora, where the above-mentioned transfer is performed sequentially between the in-domain and each out-of-domain NMT models in the ascending order of their domain similarities. Empirical results on Chinese-English and English-German translation tasks demonstrate the effectiveness of our framework.

1 Introduction

Currently, neural machine translation (NMT) has become dominant in the community of machine translation due to its excellent performance (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017). With the development of NMT, prevailing NMT models become more and more complex with large numbers of parameters, which often require abundant corpora for effective training. However, for translation tasks in most domains, domain-specific parallel sentences are often scarce. If we only use domain-specific data to train the NMT model for such a domain, the performance of resulting model is usually unsatisfying. Therefore, NMT for low-resource domains becomes a challenge in its research and applications.

To deal with this issue, many researchers have conducted studies on the domain adaptation for NMT, which can be classified into two general categories. One is to transfer the rich-resource domain (out-of-domain) translation knowledge to benefit the low-resource (in-domain) NMT model. The other is to use the mixed-domain training corpus to construct a unified NMT model for all domains. Here, we mainly focus on the first type of research, of which typical methods include fine-tuning (Luong and Manning, 2015; Zoph et al., 2016; Servan et al., 2016), mixed fine-tuning (Chu et al., 2017), cost weighting (Chen et al., 2017), data selection (Wang et al., 2017a,b; Zhang et al., 2019a) and so on. The underlying assumption of these approaches is that in-domain and out-of-domain NMT models share the same parameter space or prior distributions, and the useful out-of-domain translation knowledge can be completely transferred to in-domain NMT model in a one-pass manner. However, it is difficult to achieve this goal due to domain differences. Particularly, when the domain difference is significant, such conventional brute-force transfer may be unsuccessful, facing the similar issue as the domain adaptation for other tasks (Pan and Yang, 2010).

In this paper, to tackle the above problem, we argue that corpora of different domains should be repeatedly utilized to fully distill domain-shared translation knowledge. To this end, we propose a novel **I**terative **D**ual **D**omain **A**daptation (IDDA) framework for NMT. Under this framework, we first train in-domain and out-of-domain NMT models using their own training corpora re-

*Corresponding author.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 845–855,
Hong Kong, China, November 3–7, 2019.

spectively, and then iteratively perform bidirectional translation knowledge transfer (from in-domain to out-of-domain and then vice versa). In this way, both in-domain and out-of-domain NMT models are expected to constantly reinforce each other, which is likely to achieve better NMT domain adaptation. Particularly, we employ a *knowledge distillation* (Hinton et al., 2015; Kim and Rush, 2016) based approach to transfer translation knowledge. During this process, the target-domain NMT model is first initialized with the source-domain NMT model, and then trained to fit its own training data and match the output of its previous best model simultaneously. By doing so, the previously transferred translation knowledge can be effectively retained for better NMT domain adaptation. Finally, we further extend the proposed framework to the scenario of multiple out-of-domain training corpora, where the above-mentioned bidirectional knowledge transfer is performed sequentially between the in-domain and each out-of-domain NMT models in the ascending order of their domain similarities.

The contributions of this work are summarized as follows:

- We propose an iterative dual domain adaptation framework for NMT, which is applicable to many conventional domain transfer approaches, such as fine-tune, mixed fine-tune. Compared with previous approaches, our framework is able to better exploit domain-shared translation knowledge for NMT domain adaptation.

- We extend our framework to the setting of multiple out-of-domain training corpora, which is rarely studied in machine translation. Moreover, we explicitly differentiate the contributions of different out-of-domain training corpora based on the domain-level similarity with in-domain training corpus.

- We provide empirical evaluations of the proposed framework on Chinese-English, German-English datasets for NMT domain adaptation. Experimental results demonstrate the effectiveness of our framework. Moreover, we deeply analyze impacts of various factors on our framework[1].

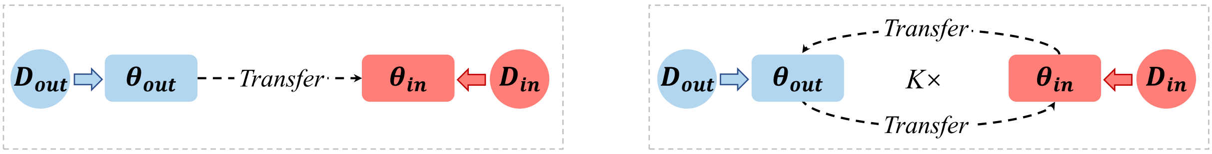
[1]We release code and results at https://github.com/DeepLearnXMU/IDDA.

2 Related Work

Our work is obviously related to the research on transferring the out-of-domain translation knowledge into the in-domain NMT model. In this aspect, fine-tuning (Luong and Manning, 2015; Zoph et al., 2016; Servan et al., 2016) is the most popular approach, where the NMT model is first trained using the out-of-domain training corpus, and then fine-tuned on the in-domain training corpus. To avoid overfitting, Chu et al. (2017) blended in-domain with out-of-domain corpora to fine-tune the pre-trained model, and Freitag and Al-Onaizan (2016) combined the fine-tuned model with the baseline via ensemble method. Meanwhile, applying data weighting into NMT domain adaptation has attracted much attention. Wang et al. (2017a) and Wang et al. (2017b) proposed several sentence and domain weighting methods with a dynamic weight learning strategy. Zhang et al. (2019a) ranked unlabeled domain training samples based on their similarity to in-domain data, and then adopts a probabilistic curriculum learning strategy during training. Chen et al. (2017) applied the sentence-level cost weighting to refine the training of NMT model. Recently, Vilar (2018) introduced a weight to each hidden unit of out-of-domain model. Chu and Wang (2018) gave a comprehensive survey of the dominant domain adaptation techniques for NMT. Gu et al. (2019) not only maintained a private encoder and a private decoder for each domain, but also introduced a common encoder and a common decoder shared by all domains.

Significantly different from the above methods, along with the studies of dual learning for NMT (He et al., 2016; Wang et al., 2018; Zhang et al., 2019b), we iteratively perform bidirectional translation knowledge transfer between in-domain and out-of-domain training corpora. To the best of our knowledge, our work is the first attempt to explore such a dual learning based framework for NMT domain adaptation. Furthermore, we extend our framework to the scenario of multiple out-of-domain corpora. Particularly, we introduce knowledge distillation into the domain adaptation for NMT and experimental results demonstrate its effectiveness, echoing its successful applications on many tasks, such as speech recognition (Hinton et al., 2015) and natural language processing (Kim and Rush, 2016; Tan et al., 2019).

Besides, our work is also related to the studies



(a) Non-iterative Domain Adaptation (b) Iterative Dual Domain Adaptation

Figure 1: Traditional approach vs IDDA framework for one-to-one NMT domain adaptation. D_{out}: out-of-domain training corpus, D_{in}: in-domain training corpus, θ_{out}: out-of-domain NMT model, θ_{in}: in-domain NMT model, $\boldsymbol{K}$ denotes the iteration number.

Algorithm 1 Iterative Dual Domain Adaptation for NMT

Input: Training corpora $\{D_{in}, D_{out}\}$, development sets $\{D^v_{in}, D^v_{out}\}$, and the maximal iteration number K.
Output: In-domain NMT model θ^*_{in}.
$\theta^{(0)}_{in} \leftarrow$ TrainModel(D_{in}), $\quad \theta^{(0)}_{out} \leftarrow$ TrainModel(D_{out})
$\theta^*_{in} \leftarrow \theta^{(0)}_{in}, \quad \theta^*_{out} \leftarrow \theta^{(0)}_{out}$
for $k = 1, 2, ..., K$ **do**
  $\theta^{(k)}_{out} \leftarrow$ TransferModel($\theta^{(k-1)}_{in}$, D_{out}, θ^*_{out})
  if EvalModel($D^v_{out}, \theta^{(k)}_{out}$) $>$ EvalModel($D^v_{out}, \theta^*_{out}$)
    $\theta^*_{out} \leftarrow \theta^{(k)}_{out}$
  end if
  $\theta^{(k)}_{in} \leftarrow$ TransferModel($\theta^{(k)}_{out}$, D_{in}, θ^*_{in})
  if EvalModel($D^v_{in}, \theta^{(k)}_{in}$) $>$ EvalModel(D^v_{in}, θ^*_{in})
    $\theta^*_{in} \leftarrow \theta^{(k)}_{in}$
  end if
end for

of multi-domain NMT, which focus on building a unified NMT model trained on the mixed-domain training corpus for translation tasks in all domains (Kobus et al., 2016; Tars and Fishel, 2018; Farajian et al., 2017; Pryzant et al., 2017; Sajjad et al., 2017; Zeng et al., 2018; Bapna and Firat, 2019). Although our framework is also able to refine out-of-domain NMT model, it is still significantly different from multi-domain NMT, since only the performance of in-domain NMT model is considered.

Finally, note that similar to our work, Tan et al. (2019) introduced knowledge distillation into multilingual NMT. However, our work is still different from (Tan et al., 2019) in the following aspects: (1) Tan et al. (2019) mainly focused on constructing a unified NMT model for multi-lingual translation task, while we aim at how to effectively transfer out-of-domain translation knowledge to in-domain NMT model; (2) Our translation knowledge transfer is bidirectional, while the procedure of knowledge distillation in (Tan et al., 2019) is unidirectional; (3) When using knowledge distillation under our framework, we iteratively update teacher models for better domain adaptation. In contrast, all language-specific teacher NMT models in (Tan et al., 2019) remain fixed.

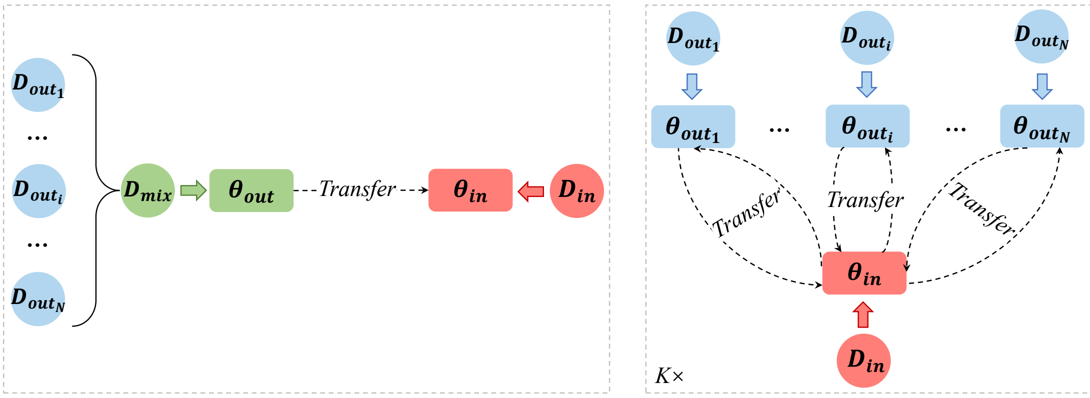
3 Iterative Dual Domain Adaptation Framework

In this section, we first detailedly describe our proposed framework for conventional one-to-one NMT domain adaptation, and then extend this framework to the scenario of multiple out-of-domain corpora (many-to-one).

3.1 One-to-one Domain Adaptation

As shown in Figure 1(a), previous studies mainly focus on the one-pass translation knowledge transfer from one out-of-domain NMT model to the in-domain NMT model. Unlike these studies, we propose to conduct iterative dual domain adaptation for NMT, of which framework is illustrated in Figure 1(b).

To better describe our framework, we summarize the training procedure of our framework



(c) Non-iterative Domain Adaptation (d) Iterative Dual Domain Adaptation

Figure 2: Traditional approach vs IDDA framework for many-to-one NMT domain adaptation. D_{mix}: a mixed out-of-domain training corpus.

in Algorithm 1. Specifically, we first individually train the initial in-domain and out-of-domain NMT models, respectively denoted by $\theta_{in}^{(0)}$ and $\theta_{out}^{(0)}$, via minimizing the negative likelihood of their own training corpora D_{in} and D_{out} (**Line 3**):

$$\mathcal{L}_{in}^{(0)} = \sum_{(\mathbf{x},\mathbf{y})\in D_{in}} -logP(\mathbf{y}|\mathbf{x};\theta_{in}^{(0)}), \quad (1)$$

$$\mathcal{L}_{out}^{(0)} = \sum_{(\mathbf{x},\mathbf{y})\in D_{out}} -logP(\mathbf{y}|\mathbf{x};\theta_{out}^{(0)}). \quad (2)$$

Then, we iteratively perform bidirectional translation knowledge transfer to update both in-domain and out-of-domain NMT models, until the maximal iteration number K is reached (**Lines 5-14**). More specifically, at the k-th iteration, we first transfer the translation knowledge of the previous in-domain NMT model $\theta_{in}^{(k-1)}$ to the out-of-domain NMT model $\theta_{out}^{(k)}$ trained on D_{out} (**Line 6**), and then reversely transfer the translation knowledge encoded by $\theta_{out}^{(k)}$ to the in-domain NMT model $\theta_{in}^{(k)}$ trained on D_{in} (**Line 10**). During this process, we evaluate the new models $\theta_{in}^{(k)}$ and $\theta_{out}^{(k)}$ on their corresponding development sets, and then record the best model parameters as θ_{in}^{*} and θ_{out}^{*} (**Lines 7-9, 11-13**).

Obviously, during the above procedure, one of important steps is how to transfer the translation knowledge from one domain-specific NMT model to the other one. However, if we directly employ conventional domain transfer approaches, such as fine-tuning, as the iterative dual domain adaptation proceeds, the previously learned translation knowledge tends to be ignored. To deal with this issue, we introduce knowledge distillation (Kim and Rush, 2016) to conduct the translation knowledge transfer. Specifically, during the transfer process from $\theta_{out}^{(k)}$ to $\theta_{in}^{(k)}$, we first initialize $\theta_{in}^{(k)}$ with parameters of $\theta_{out}^{(k)}$, and then train $\theta_{in}^{(k)}$ not only to match the references of D_{in}, but also to be consistent with probability outputs of the previous best in-domain NMT model θ_{in}^{*}, which is considered as the teacher model. To this end, we define the loss function as

$$\mathcal{L}_{in}^{(k)} = \sum_{(\mathbf{x},\mathbf{y})\in D_{in}} [-(1-\lambda)\cdot logP(\mathbf{y}|\mathbf{x};\theta_{in}^{(k)})+ \lambda\cdot \mathrm{KL}(P(\mathbf{y}|\mathbf{x};\theta_{in}^{(k)})||P(\mathbf{y}|\mathbf{x};\theta_{in}^{*}))], \quad (3)$$

where λ is the coefficient used to trade off these two loss terms, and it can be tuned on the development set. Notably, when λ=0, only the term of likelihood function affects the model training, and thus our transfer approach degenerate into fine-tuning at each iteration.

In this way, we enable in-domain NMT model $\theta_{in}^{(k)}$ to not only retain the previously learned effective translation knowledge, but also fully absorb the useful translation knowledge from out-of-domain NMT model $\theta_{out}^{(k)}$. Similarly, we employ the above method to transfer translation knowledge from $\theta_{in}^{(k-1)}$ to $\theta_{out}^{(k)}$ using out-of-domain corpus D_{out} and the previous best out-of-domain model θ_{out}^{*}. Due to the space limitation, we omit the specific description of this procedure.

3.2 Many-to-one Domain Adaptation

Usually, in practical applications, there exist multiple available out-of-domain training corpora simultaneously. As shown in Figure 2(a), previous studies usually mix them into one out-of-domain corpus, which is applicable for the conventional one-to-one NMT domain adaptation. However, various out-of-domain corpora are semantically related to in-domain corpus to different degrees, and thus intuitively, it is difficult to adequately play their roles without distinguishing them.

To address this issue, we extend the proposed framework to many-to-one NMT domain adaptation. Our extended framework is illustrated in Figure 2(b). Given an in-domain corpus and N out-of-domain corpora, we first measure the semantic distance between each out-of-domain corpus and the in-domain corpus using the proxy A-distance $\hat{d}_A$=$2(1-2\epsilon)$ (Ganin et al., 2015; Pryzant et al., 2017), where the ϵ is the generalization error of a linear bag-of-words SVM classifier trained to discriminate between the two domains. Then, we determine the transfer order of these out-of-domain NMT models as $\{\theta_{out_1}, \theta_{out_2}, ...\theta_{out_N}\}$, according to distances of their own training corpora to the in-domain corpus in a decreasing order. The reason behind this step is the translation knowledge of previously transferred out-of-domain NMT models will be partially forgotten during the continuous transfer. By setting transfer order according to their $\hat{d}_A$ values in a decreasing order, we enable the in-domain NMT model to fully preserve the translation knowledge transferred from the most relevant out-of-domain NMT model. Finally, we sequentially perform bidirectional knowledge transfer between the in-domain and each out-of-domain models, where this process will be repeated for K iterations.

4 Experiments

To verify the effectiveness of our framework, we first conducted one-to-one domain adaptation experiments on Chinese-English translation, where we further investigated impacts of various factors on our framework. Then, we carried out two-to-one domain adaptation experiments on English-German translation, so as to demonstrate the generality of our framework on different language pairs and multiple out-of-domain corpora.

4.1 Setup

Datasets. In the Chinese-English translation task, our in-domain training corpus is from IWSLT2015 dataset consisting of 210K *TED Talk* sentence pairs, and the out-of-domain training corpus contains 1.12M LDC sentence pairs related to *News* domain. For these two domains, we chose IWSLT dev2010 and NIST 2002 dataset as development sets. Finally, we used IWSLT tst2010, tst2011 and tst2012 as in-domain test sets. Particularly, in order to verify whether our framework can enable NMT models of two domains to benefit each other, we also tested the performance of out-domain NMT model on NIST 2003, 2004, 2005, 2006 datasets.

For the English-German translation task, our training corpora totally include one in-domain dataset: 200K *TED Talk* sentence pairs provided by IWSLT2015, and two out-of-domain datasets: 500K sentence pairs (*News* topic) extracted from WMT2014 corpus, and 500K sentence pairs (*Medical topic*) that are sampled from OPUS EMEA corpus[2]. As for development sets, we chose IWSLT tst2012, WMT tst2012 and 1K sampled sentence pairs of OPUS EMEA corpus, respectively. In addition, IWSLT tst2013, tst2014 were used as in-domain test sets, WMT news-test2014 (News topic) and 1K sampled sentence pairs of OPUS EMEA corpus were used as two out-of-domain test sets.

We first employed *Stanford Segmenter*[3] to conduct word segmentation on Chinese sentences and *MOSES script*[4] to tokenize English and German sentences. Then, we limited the length of sentences to 50 words in the training stage. Besides, we employed *Byte Pair Encoding* (Sennrich et al., 2016) to split words into subwords and set the vocabulary size for both Chinese-English and English-German as 32,000. We evaluated the translation quality with BLEU scores (Papineni et al., 2002) as calculated by `multi-bleu.perl` script.

Settings. We chose Transformer (Vaswani et al., 2017) as our NMT model, which exhibits excellent performance due to its flexibility in parallel computation and long-range dependency modeling. We followed Vaswani et al. (2017) to set the configurations. The dimensionality of all input and output layers is 512, and that of FFN layer is 2048. We employed 8 parallel attention heads in both encoder and decoder. Parameter optimization was performed using stochastic gradient descent, where *Adam* (Kingma and Ba, 2015) was used to automatically adjust the learning rate of

[2]http://opus.nlpl.eu/
[3]https://nlp.stanford.edu/
[4]http://www.statmt.org/moses/

each parameter. We batched sentence pairs by approximated length, and limited input and output tokens per batch to 25000 tokens. As for decoding we employed beam search algorithm and set the beam size as 4. Besides, we set the distillation coefficient λ as 0.4.

Contrast Models. We compared our framework with the following models, namely:

- **Single** A reimplemented Transformer only trained on a single domain-specific (in/out) training corpus.
- **Mix** A reimplemented Transformer trained on the mix of in-domain and out-of-domain training corpora.
- **Fine-tuning (FT)** (Luong and Manning, 2015). It first trains the NMT model on out-of-domain training corpus and then fine-tunes it using in-domain training corpus.
- **Mixed Fine-tuning (MFT)** (Chu et al., 2017). It also first trains the NMT model on out-of-domain training corpus, and then fine-tunes it using both out-of-domain and over-sampling in-domain training corpora.
- **Knowledge Distillation (KD)** (Kim and Rush, 2016). Using this method, we first train a out-of-domain and an in-domain NMT models using their own training corpus, respectively. Then, we use the in-domain training corpus to fine-tune the out-of-domain NMT model, supervised by the in-domain NMT model.

Besides, we reported the performance of some recently proposed multi-domain NMT models.

- **Domain Control (DC)** (Kobus et al., 2016). It is also based on the mix-domain NMT model. However, it adds an additional domain tag to each source sentence, incorporating domain information into source annotations.
- **Discriminative Mixing (DM)** (Pryzant et al., 2017). It jointly trains NMT with domain classification via multitask learning. Please note that it performs the best among three approaches proposed by Pryzant et al., (2017).
- **Word-level Domain Context Discrimination (WDCD)** (Zeng et al., 2018). It discriminates the source-side word-level domain specific and domain-shared contexts for multi-domain NMT by jointly modeling NMT and domain classifications.

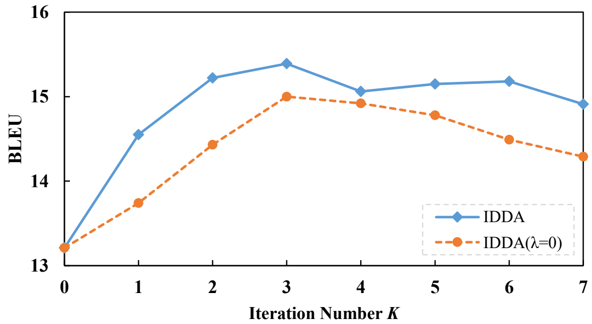


Figure 3: Effect of iteration number (K) on the Chinese-English in-domain development set.

4.2 Results on Chinese-English Translation

4.2.1 Effect of Iteration Number K

The iteration number K is a crucial hyper-parameter that directly determines the amount of the transferred translation knowledge under our framework. Therefore, we first inspected its impacts on the development sets. To this end, we varied K from 0 to 7 with an increment of 1 in each step, where our framework degrades to *Single* when K=0.

Figure 3 provides the experimental results using different Ks. We can observe that both *IDDA(λ=0)* and *IDDA* achieve the best performance at the 3-th iteration, respectively. Therefore, we directly used K=3 in all subsequent experiments.

4.2.2 Overall Performance

Table 1 shows the overall experimental results. On all test sets, our framework significantly outperforms other contrast models. Furthermore, we reach the following conclusions:

First, on the in-domain test sets, both *IDDA(λ=0)* and *IDDA* surpass *Single*, *Mix*, *FT*, *MFT* and *KD*, most of which are commonly used in the domain adaptation for NMT. This confirms the difficulty in completely one-pass transferring the useful out-of-domain translation knowledge to the in-domain NMT model. Moreover, the in-domain NMT model benefits from multiple-pass knowledge transfers under our framework.

Second, compared with *DC*, *DM* and *WDCD* that are proposed for multi-domain NMT, both *IDDA(λ=0)* and *IDDA* still exhibit better performance on the in-domain test sets. The underlying

Model	TED Talk (**In-domain**)					News (**Out-of-domain**)				
	Tst10	Tst11	Tst12	Tst13	AVE.	Nist03	Nist04	Nist05	Nist06	AVE.
Cross-domain Transfer Methods										
Single	15.82	20.80	17.77	18.33	18.18	45.38	45.93	42.80	42.70	44.20
Mix	16.46	20.85	19.13	19.87	19.08	44.87	45.71	42.24	42.02	43.71
FT	16.77	21.16	19.31	20.53	19.44	—	—	—	—	—
MFT	17.19	22.02	20.09	21.05	20.08	—	—	—	—	—
KD	17.62	21.88	19.97	20.43	19.98	—	—	—	—	—
Multi-domain NMT Methods										
DC	17.23	22.10	19.68	20.58	19.90	46.03	46.62	44.39	43.82	45.21
DM	16.45	21.35	18.77	20.27	19.21	45.12	45.83	42.77	42.59	44.08
WDCD	17.32	22.23	20.02	21.10	20.17	46.33	46.36	44.62	43.80	45.27
IDDA Framework										
IDDA(λ=0)	18.00	22.71	20.36	**21.82**	20.72	45.91	45.84	43.61	42.17	44.46
IDDA	**18.36**	**23.14**	**20.78**	21.79	**21.02**†	**47.17**	**47.44**	**45.38**	**44.04**	**46.01**†

Table 1: Experimental results on the Chinese-English translation task. † indicates statistically significantly better than (ρ <0.01) the result of *WDCD*.

reason is that these multi-domain models discriminate domain-specific and domain-shared information in encoder, however, their shared decoder are inadequate to effectively preserve domain-related text style and idioms. In contrast, our framework is adept at preserving these information since we construct an individual NMT model for each domain.

Third, *IDDA* achieves better performance than *IDDA(λ=0)*, demonstrating the importance of retaining previously learned translation knowledge. Surprisingly, *IDDA* significantly outperforms *IDDA(λ=0)* on out-of-domain data sets. We conjecture that during the process of knowledge distillation, by assigning non-zero probabilities to multiple words, the output distribution of teacher model is more smooth, leading to smaller variance in gradients (Hinton et al., 2015). Consequently, the out-of-domain NMT model becomes more robust by iteratively absorbing the translation knowledge from the best out-of-domain model.

Finally, note that even on the out-of-domain test sets, *IDDA* still has better performance than all listed contrast models in the subsequent experimental analyses. This result demonstrates the advantage of dual domain adaptation under our framework.

According to the reported performance of our framework shown in Table 1, we only considered *IDDA* in all subsequent experiments. Besides, we only chose *MFT*, *KD*, and *WDCD* as typical contrast models. This is because *KD* is the basic domain adaption approach of our framework, *MFT* and *WDCD* are the best domain adaptation method and multi-domain NMT model for comparison, respectively.

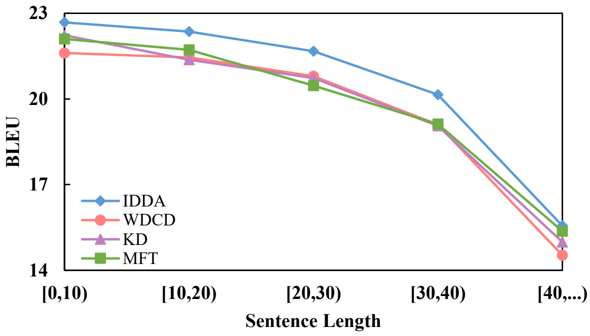


Figure 4: BLEU scores on different IWSLT test sets divided according to source sentence lengths.

4.2.3 Results on Source Sentences with Different Lengths

Following previous work (Bahdanau et al., 2015), we divided IWSLT test sets into different groups based on the lengths of source sentences and then investigated the performance of various models.

Figure 4 illustrates the results. We observe that our framework also achieves the best performance in all groups, although the performances of all models degrade with the increase of the length of source sentences.

4.2.4 Effect of Out-of-domain Corpus Size

In this group of experiments, we investigated the impacts of out-of-domain corpus size on our proposed framework. Specifically, we inspected the results of our framework using different sizes of out-of-domain corpora: 50K, 200K and 1.12M, respectively

Figure 5 shows the comparison results on the average BLEU scores of all IWSLT test sets. No matter how large out-of-domain data is used, *IDDA* always achieves better performance than

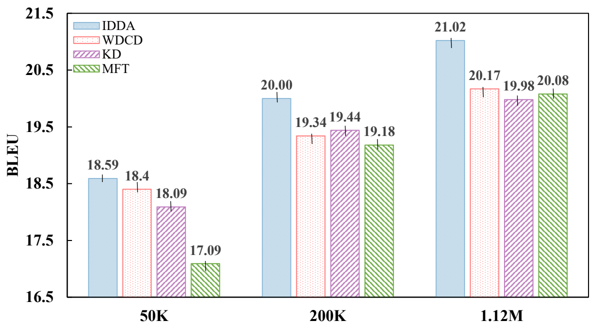


Figure 5: Experimental results with different sizes of out-of-domain corpora.

Model	AVE.
IDDA-unidir	20.43
IDDA-fixTea	20.60
IDDA	21.02

Table 2: Experimental results of comparing IDDA with its two variants.

other contrast models, demonstrating the effectiveness and generality of our framework. Specially, *IDDA* with 200K out-of-domain corpus is comparable to *KD* with 1.12M corpus. From this result, we confirm again that our framework is able to better exploit the complementary information between domains than *KD*.

4.2.5 Effects of Dual Domain Adaptation and Updating Teacher Models

Two highlights of our framework consist of the usage of bidirectional translation knowledge transfer and continuous updating teacher models θ^*_{out} and θ^*_{in} (See Line 6, 10 of Algorithm 1). To inspect their effects on our framework, we compared our framework with its two variants: (1) *IDDA-unidir*, where we only iteratively transfer out-of-domain translation knowledge to the in-domain NMT model; (2) *IDDA-fixTea*, where teacher models are fixed as the initial out-of-domain and in-domain NMT models, respectively.

The results are displayed in Table 2. We can see that our framework exhibits better performance than its two variants, which demonstrates that dual domain adaptation enables NMT models of two domains to benefit from each other, and updating teacher models is more helpful to retain useful translation knowledge.

4.2.6 Case Study

Table 3 displays the 1-best translations of a sampled test sentence generated by *MFT*, *KD*, *WDCD*, and *IDDA* at different iterations. Inspecting this

Model	Translation
Src	这(zhè) 是(shì) 第(dì) 一(yī) 种(zhǒng) 直立(zhílì) 行走(xíngzǒu) 的(de) 灵长类(língzhǎnglèi) 动物(dòngwù)
Ref	that was the first upright primate
MFT	this is the first *animal* to walk upright
KD	this is the first *growing* primate
WDCD	this is the first primate *walking around*
IDDA-1	this is the first *upright - walking primate*
IDDA-2	this is the first *upright - walking primate*
IDDA-3	this is the first *primates walking upright*
IDDA-4	this is the first *upright primate*
IDDA-5	this is the first *upright primate*
IDDA-6	this is the first *upright primate*
IDDA-7	this is the first *upright primate*

Table 3: Translation examples of different NMT models. **Src**: source sentence and **Ref**: target reference. *IDDA-k* represents the in-domain NMT model at the k-th iteration using our framework.

example provides the insight into the advantage of our proposed framework to some extent. Specifically, we observe that *MFT*, *KD*, *WDCD* are unable to correctly understand the meaning of "zhílì xíngzǒu de língzhǎnglèi dòngwù" and thus generate incorrect or incomplete translations, while *IDDA* successfully corrects these errors by gradually absorbing transferred translation knowledge.

4.3 Results on English-German Translation

4.3.1 Overall Performance

We first calculated the distance between the in-domain and each out-of-domain corpora: $\hat{d}_A$(*Ted Talk*, *News*) = 0.92 and $\hat{d}_A$(*Ted Talk*, *Medical*) = 1.92. Obviously, the *News* domain is more relevant to *TED Talk* domain than *Medical* domain, and thus we determined the final transfer order as $\{\theta_{\text{out}_{\text{medical}}}, \theta_{\text{out}_{\text{news}}}\}$ for this task. Then, as implemented in the previous Chinese-English experiments, we determined the optimal K=2 on the development set.

Table 4 shows experimental results. Similar to the previously reported experiment results, our framework still obtains the best performance among all models, which verifies the effectiveness of our framework on many-to-one domain adaptation for NMT.

As described above, we have two careful designs for many-to-one NMT domain adaptation: (1) We distinguish different out-of-domain corpora, and then iteratively perform bidirectional translation knowledge transfer between in-domain and each out-of-domain NMT models. (2) We determine the transfer order according to the seman-

Model	In-domain TED Talk			Out-of-domain1 News	Out-of-domain2 Medical
	IWSLT2013	IWSLT2014	AVE.	WMT14	EMEA
Cross-domain Transfer Methods					
Single	29.76	25.99	27.88	20.54	51.11
Mix	31.45	27.03	29.24	21.17	50.60
FT	30.54	27.02	28.78	—	—
MFT	31.86	27.49	29.67	—	—
KD	31.33	27.96	29.64	—	—
Multi-domain NMT Methods					
DC	31.13	28.02	29.57	21.61	52.25
DM	31.57	27.60	29.58	21.75	52.60
WDCD	31.87	27.82	29.84	21.86	52.84
IDDA Framework					
IDDA(λ=0)	32.11	28.10	30.11	22.01	52.07
IDDA	**32.93**	**28.88**	**30.91**†	**22.17***	**53.39**†

Table 4: Experimental results of the English-German translation task. * indicates statistically significantly better than (ρ <0.05) the result of *WDCD*.

Model	Transfer Order	AVE.
IDDA-mix	——	30.17
IDDA	$\{\theta_{\text{out}_{\text{news}}}, \theta_{\text{out}_{\text{medical}}}\}$	30.51
IDDA	$\{\theta_{\text{out}_{\text{medical}}}, \theta_{\text{out}_{\text{news}}}\}$	30.91

Table 5: Experimental results of IDDA using different configurations.

tic distance between each out-of-domain and in-domain training corpora. Here, we carried out two groups of experiments to investigate their impacts on our framework. In the first group of experiments, we first combined all out-of-domain training corpora into a mixed corpus, and then applied our framework to establish the in-domain NMT model. In the second group of experiments, we employed our framework in different transfer orders to perform domain adaptation.

Table 5 shows the final experimental results, which are in line with our expectations and verify the validity of our designs.

5 Conclusion

In this paper, we have proposed an iterative dual domain adaptation framework for NMT, which continuously fully exploits the mutual complementarity between in-domain and out-domain corpora for translation knowledge transfer. Experimental results and in-depth analyses on translation tasks of two language pairs strongly demonstrate the effectiveness of our framework.

In the future, we plan to extend our framework to multi-domain NMT. Besides, how to leverage monolingual sentences of different domains to refine our proposed framework. Finally, we will apply our framework into other translation models (Bahdanau et al., 2015; Su et al., 2018; Song et al., 2019), so as to verify the generality of our framework.

Acknowledgments

The authors were supported by National Natural Science Foundation of China (No. 61672440), Beijing Advanced Innovation Center for Language Resources, NSF Award (No. 1704337), the Fundamental Research Funds for the Central Universities (Grant No. ZK1024), and Scientific Research Project of National Language Committee of China (Grant No. YB135-49). We also thank the reviewers for their insightful comments

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR 2015*.

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. In *Proc. of NAACL 2019*.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proc. of WMT 2018*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation.

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

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proc. of WMT 2017*.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR abs/1612.06897*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2015. Domain-adversarial training of neural networks. *Machine Learning Research*, 17.

Shuhao Gu, Yang Feng, and Qun Liu. 2019. Improving domain adaptation translation with domain invariant and specific information. In *Proc. of NAACL 2019*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Proc. of NIPS 2016*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR abs/1503.02531*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proc. of EMNLP 2016*.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR 2015*.

Catherine Kobus, Josep Crego, and Jean Senellart. 2016. Domain control for neural machine translation. *CoRR abs/1612.06140*.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proc. of IWSLT 2015*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL 2002*.

Reid Pryzant, Denny Britz, and Q Le. 2017. Effective domain mixing for neural machine translation. In *Proc. of WMT 2017*.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. 2017. Neural machine translation training in a multi-domain scenario. *CoRR abs/1708.08712*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL 2016*.

Christophe Servan, Josep Crego, and Jean Senellart. 2016. Domain specialization: a post-training domain adaptation for neural machine translation. *CoRR abs/1612.06141*.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *TACL 2019*, 7.

Jinsong Su, Jiali Zeng, Deyi Xiong, Yang Liu, Mingxuan Wang, and Jun Xie. 2018. A hierarchy-to-sequence attentional neural machine translation model. *IEEE/ACM TALSP 2018*, 26(3).

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *Proc. of ICLR 2019*.

Sander Tars and Mark Fishel. 2018. Multi-domain neural machine translation. *CoRR abs/1805.02282*.

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

David Vilar. 2018. Learning hidden unit contribution for adapting neural machine translation models. In *Proc. of NAACL 2018*, pages 500–505.

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

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017b. Instance weighting for neural machine translation domain adaptation. In *Proc. of EMNLP 2017*.

Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. 2018. Dual transfer learning for neural machine translation with marginal distribution regularization. In *Proc. of AAAI 2018*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR abs/1609.08144*.

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 *Proc. of EMNLP 2018*.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019a. Curriculum learning for domain adaptation in neural machine translation. In *Proc. of NAACL 2019*.

Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2019b. Regularizing neural machine translation by target-bidirectional agreement.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *Proc. of EMNLP 2016*.