Unsupervised Domain Adaptation for Neural Machine Translation with Iterative Back Translation

Di Jin[1] , **Zhijing Jin**[2] , **Joey Tianyi Zhou**[3*] and **Peter Szolovits**[1]
[1]Computer Science & Artificial Intelligence Laboratory, MIT
[2]Amazon Web Services
[3]A*STAR, Singapore
jindi15@mit.edu, zhijing.jin@connect.hku.hk, zhouty@ihpc.a-star.edu.sgu, psz@mit.edu

Abstract

State-of-the-art neural machine translation (NMT) systems are data-hungry and perform poorly on domains with little supervised data. As data collection is expensive and infeasible in many cases, unsupervised domain adaptation methods are needed. We apply an **I**terative **B**ack **T**ranslation (**IBT**) training scheme on in-domain monolingual data, which repeatedly uses a Transformer-based NMT model to create in-domain pseudo-parallel sentence pairs in one translation direction on the fly and then use them to train the model in the other direction. Evaluated on three domains of German-to-English translation task with no supervised data, this simple technique alone (without any out-of-domain parallel data) can already surpass all previous domain adaptation methods—up to $+9.48$ BLEU over the strongest previous method, and up to $+27.77$ BLEU over the unadapted baseline. Moreover, given available supervised out-of-domain data on German-to-English and Romanian-to-English language pairs, we can further enhance the performance and obtain up to $+19.31$ BLEU improvement over the strongest baseline, and $+47.69$ BLEU increment against the unadapted model.

1 Introduction

Current neural machine translation (NMT) systems are data-hungry. While some domains with large amounts of parallel translation data (i.e., high-resource domains) see the benefit of NMT, other domains with small parallel corpora (i.e., low-resource domains, such as law, medicine, and technology) still suffer from poor model performance. One solution is to collect the same kind of large parallel data in such domains, which are both expensive and time-consuming, and in some cases even impractical.

A one-cure-for-all solution is to develop effective methods to generalize the NMT models to new domains, namely domain adaptation. For example, we can translate English medical documents to German, by training an NMT model on parallel translations from another high-resource domain such as parliament hearing transcripts. The majority of previous works in domain adaptation for NMT focus on the supervised setting, where a small amount of parallel in-domain data is still available [Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Chu and Wang, 2018]. However, collection of translation pairs in some domains requires skilled human writers, and not every party can afford or wait for the supervised data. Therefore, our work focuses on unsupervised domain adaptation, where no in-domain parallel data is available. Existing approaches can be categorized into two lines: *data-centric* and *model-centric* methods. Data-centric methods rely on a pseudo-parallel corpus constructed either by copying the in-domain target corpus to the source side, termed as *Copy* [Currey *et al.*, 2017], or by pairing the in-domain target-side sentences with their translated counterparts by a well-trained NMT model, termed as *back-translation* [Sennrich *et al.*, 2015]. Model-centric methods mainly focus on multitask learning of the translation task on the out-of-domain parallel data and the language modeling task on the target-side monolingual in-domain data [Domhan and Hieber, 2017; Dou *et al.*, 2019].

Despite the various previous approaches, the rich value of in-domain monolingual data is still under-explored; only language modeling and one-time back translation have been attempted on them. We highlight an **I**terative **B**ack **T**ranslation (**IBT**) scheme to make better use of in-domain text in our approach, inspired by the great success of unsupervised NMT and text style transfer in the past two years [Lample *et al.*, 2018; Artetxe *et al.*, 2017; Jin *et al.*, 2019]. This scheme iteratively generates translations from the source language to the target language and train the translation model to map the generated data back to the original source sentences, and then we perform the same generation and training process but in the reversed direction. This process is repeated in the training until convergence. Without any parallel or out-of-domain data, we found this method alone already surpasses all previous state-of-the-art domain adaptation methods which rely on data from other domains, proving that the in-domain corpus can be more directly beneficial to the task than the out-of-domain paired data. Notably, our proposed method is model-centric, and thus is independent to the data-centric methods, combined with which we can obtain further significant improvements. Moreover, we find that pre-training both the encoder and decoder of the NMT model on large-scale unsuper-

*Corresponding Author

vised corpora through masked language modeling [Lample and Conneau, 2019] can benefit both our method and other popular unsupervised domain adaptation methods.

We evaluate our approach by a Transformer-based NMT system [Vaswani *et al.*, 2017] under two different data settings and on two language pairs. For the first setting where the model adapts between two specific domains, our vanilla model achieved up to +9.48 BLEU score improvement over the strongest out of four baseline models, and +27.77 BLEU over the unadapted baseline. For the second setting, provided the availability of general domain data (WMT), our model combined with the data augmentation through back-translation achieves up to +19.31 BLEU improvement over the best previous method and +47.69 BLEU improvement over the unadapted model.

2 Methods

In this section, we first illustrate the architecture of IBT, and then formulate the overall training strategy.

2.1 Model Architecture

To fit in our proposed framework, we adopt the Transformer [Vaswani *et al.*, 2017] with the encoder-decoder structure for the sequence-to-sequence translation model, as shown in Figure 1. Following the practice in [Lample and Conneau, 2019], we add the language embeddings to the standard token and position embeddings via the element-wise summation operation. This language embedding can inform both encoder and decoder which language it is processing. For instance, when translating from German to English, we set the language embedding of the encoder to German (through a look-up table) while setting that of the decoder to English. For the reversed direction of translation (i.e., from English to German), we just need to reverse the language embedding settings for the encoder and decoder without changing the model architecture. In this way, the same model architecture can be used to translate any language pair at the same time.

Three key properties of our model are introduced in the following paragraphs:

Shared Sub-Word Vocabulary In our experiments, we process all languages with the same shared vocabulary created through Byte Pair Encoding (BPE) [Sennrich *et al.*, 2016]. This not only enables us to translate between any language pair with the same model, but improves the alignment of embedding spaces across languages that share either the same alphabet or anchor tokens such as digits. The BPE splits are learned on the concatenation of sentences sampled randomly from the monolingual corpora.

Shared Latent Representations All encoder parameters (including the embedding matrices since we perform joint tokenization) are shared across the source and target languages so that the encoder can map the input of any source language into a shared latent representation space, which is then translated to the target language by the decoder. Furthermore, we share the decoder parameters across the two languages to reduce parameter size with the language embeddings as the language identifier. We also share the encoder and decoder between the translation and language modeling tasks, which ensures that the benefits of language modeling, implemented via the denoising auto-encoder objective, nicely transfer to translation from noisy sources and eventually help the NMT model to translate more fluently.

Initialization Both the encoder and decoder are initialized by pre-trained parameters, which are obtained by pre-training a masked language model on large-scale monolingual corpora of both languages in the language pair [Lample and Conneau, 2019]. Such initialization can not only accelerate the model convergence but also improve the adaptation performance, which will be discussed in Section 5.1.

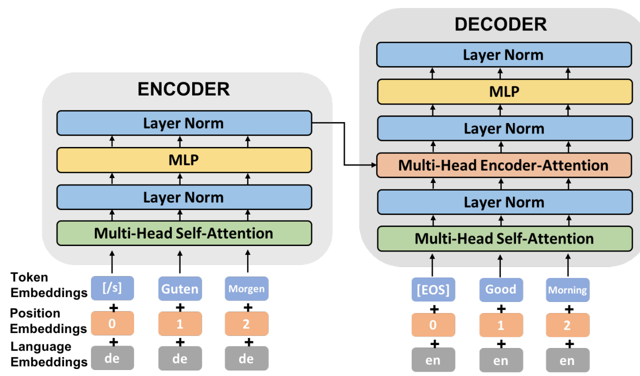


Figure 1: Transformer-based model architecture.

2.2 Training Objectives

In the unsupervised domain adaptation setting, we assume access to an out-of-domain *parallel* training corpus $(X_{\mathrm{out}}, Y_{\mathrm{out}})$ and in-domain *monolingual* data X_{in} and Y_{in}. We have the following three training objectives:

Language Modeling In NMT, language modeling is usually implemented via denoising autoencoders, by minimizing

$$\begin{aligned}\mathcal{L}_{\mathrm{lm}} =& \mathbb{E}_{\boldsymbol{x}\in X_{\mathrm{in}}}[-\log P_{\mathrm{s}\to\mathrm{s}}(\boldsymbol{x}|C(\boldsymbol{x});\boldsymbol{\theta})]+\\ & \mathbb{E}_{\boldsymbol{y}\in Y_{\mathrm{in}}}[-\log P_{\mathrm{t}\to\mathrm{t}}(\boldsymbol{y}|C(\boldsymbol{y});\boldsymbol{\theta})], \end{aligned} \quad (1)$$

where C is a word corruption function with some words randomly dropped, blanked, and swapped; $P_{\mathrm{s}\to\mathrm{s}}$ and $P_{\mathrm{t}\to\mathrm{t}}$ are the composition of encoder and decoder both operating on the source and target sides, respectively; θ denotes the model parameters including both the encoder and decoder.

Iterative Back-Translation We have two NMT models, $\mathrm{Model}_{\mathrm{s2t}}(\cdot)$ which translates from the source language to the target language, and $\mathrm{Model}_{\mathrm{t2s}}(\cdot)$ vice versa (they are implemented by one model architecture). In each iteration, we translate on the fly from each source language sentence $\boldsymbol{x} \in X_{\mathrm{in}}$ to a target language sentence $\mathrm{Model}_{\mathrm{s2t}}(\boldsymbol{x})$. Similarly, we translate from every target sentence $\boldsymbol{y} \in Y_{\mathrm{in}}$ to its counterpart in the source language $\mathrm{Model}_{\mathrm{t2s}}(\boldsymbol{y})$. Then the pairs of $(\boldsymbol{x}, \mathrm{Model}_{\mathrm{s2t}}(\boldsymbol{x}))$ and $(\mathrm{Model}_{\mathrm{t2s}}(\boldsymbol{y}), \boldsymbol{y})$ can be used as synthetic parallel data to train the two NMT models by minimizing the following loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{back}} =& \mathbb{E}_{\boldsymbol{x}\in X_{\mathrm{in}}}[-\log P_{\mathrm{t}\to\mathrm{s}}(\mathrm{Model}_{\mathrm{s2t}}(\boldsymbol{x})|\boldsymbol{x};\boldsymbol{\theta})]+\\ & \mathbb{E}_{\boldsymbol{y}\in Y_{\mathrm{in}}}[-\log P_{\mathrm{s}\to\mathrm{t}}(\boldsymbol{y}|\mathrm{Model}_{\mathrm{t2s}}(\boldsymbol{y});\boldsymbol{\theta})]. \end{aligned} \quad (2)$$

Algorithm 1 Training Strategy of IBT

Input: In-domain monolingual data X_{in} and Y_{in}, parallel data $(X_{\text{para}}, Y_{\text{para}})$, and model parameters $\boldsymbol{\theta}$

while $\boldsymbol{\theta}$ has not converged **do**
  Sample $\boldsymbol{x}$ from X_{in} and $\boldsymbol{y}$ from Y_{in}
  Create pairs $(C(\boldsymbol{x}), \boldsymbol{x})$ and $(C(\boldsymbol{y}), \boldsymbol{y})$ via word corruption
  Update $\boldsymbol{\theta}$ by minimizing Eq. (1)
  Sample $\boldsymbol{x}$ from X_{in} and $\boldsymbol{y}$ from Y_{in}
  Create $(\boldsymbol{x}, \text{Model}_{\text{s2t}}(\boldsymbol{x}))$ and $(\text{Model}_{\text{t2s}}(\boldsymbol{y}), \boldsymbol{y})$ via back-translation
  Update $\boldsymbol{\theta}$ by minimizing Eq. (2)
  Sample $(\boldsymbol{x}, \boldsymbol{y})$ from $(X_{\text{para}}, Y_{\text{para}})$
  Update $\boldsymbol{\theta}$ by minimizing Eq. (3)
end while

This pseudo-supervised training paradigm is termed as iterative back-translation. To be noted, when minimizing this objective function, we do not back propagate through the models that are used to generate translations.

Supervised Machine Translation When given parallel data, denoted as $(X_{\text{para}}, Y_{\text{para}})$, we can also minimize the supervised translation loss:

$$\mathcal{L}_{\text{sup}} = \mathbb{E}_{\boldsymbol{x} \in X_{\text{para}}, \boldsymbol{y} \in Y_{\text{para}}}[-\log P_{\text{s}\to\text{t}}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta})]. \qquad (3)$$

The parallel data can be the out-of-domain supervised data $(X_{\text{out}}, Y_{\text{out}})$ or the back-translated synthetic pairs by an NMT model trained on the out-of-domain data.

2.3 Training Strategy

As shown in Algorithm 1, in each iteration, we randomly draw a batch of data to minimize the aforementioned three loss equations 1, 2 and 3, respectively. The iterations will continue until the validation set BLEU score does not increase for a certain number of epochs.

3 Experiments

3.1 Data Setup

We validate our model under two different data settings. For the first setting, we train on the law (LAW), medical (MED), and Information Technology (IT) datasets of the German-English OPUS corpus[1] [Tiedemann, 2012] and test the domain adaptation ability of our model on every pair of domains. The second setting is to adapt models trained on the general-domain WMT datasets to TED [Duh, 2018], LAW, and MED datasets. Two language pairs are used for this setting, German-English (de-en) and Romanian-English (ro-en). The general-domain WMT datasets for de-en and ro-en come from WMT-14[2] and WMT-16[3] tasks, respectively. Data statistics for the train set are shown in Table 1. The size of validation and test sets for WMT-14 are both 3K, and all the other domains are 2K.

Note that to build an unaligned monolingual corpus for each domain, we randomly shuffle the original parallel corpus and split the corpus into two halves with the same number of parallel sentences. We use the target and source sentences of the first and second halves respectively. Hence, there are no parallel sentences in the monolingual corpus.

[1]https://github.com/khayrallah/domain-adaptation-data
[2]https://www.statmt.org/wmt14/translation-task.html
[3]https://www.statmt.org/wmt16/translation-task.html

Lang.	Corpus	Words	Sentences	W/S
De-En	MED	14,533,613	1,104,752	13.2
	LAW	18,461,140	715,372	25.8
	IT	3,212,130	337,817	9.5
	TED	3,110,970	151,627	20.5
	WMT-14	126,735,962	4,468,840	28.4
Ro-En	MED	13,142,512	990,220	13.3
	LAW	10,631,517	450,715	23.6
	TED	3,328,621	161,291	20.6
	WMT-16	10,796,138	399,375	27.0

Table 1: Statistics of the corpora used for training (target side).

3.2 Baselines

We compare our models with four strongest baselines described below.

COPY The target-side sentences of in-domain data are copied to the source side and then combined with out-of-domain data as train data to train an NMT model [Currey *et al.*, 2017].

BACK A target-to-source NMT model is first trained on the out-of-domain data and then used to generate pseudo in-domain parallel data for model training by translating the target in-domain monolingual sentences [Sennrich *et al.*, 2015].

DALI Lexicon induction is first performed to extract an in-domain lexicon, and then a pseudo-parallel in-domain corpus is constructed by performing word-to-word back-translation of monolingual in-domain target sentences, which is used for fine-tuning a pre-trained out-of-domain NMT model [Hu *et al.*, 2019].

DAFE It performs multi-task learning on a translation model on out-of-domain parallel data and a language model on in-domain target-side monolingual data, while inserting domain and task embedding learners into the transformer-based model [Dou *et al.*, 2019].

To be noted, COPY, BACK, and DALI are data-centric methods, while DAFE and our approach are model-centric.

3.3 Settings

The architecture of the encoder and decoder in the transformer model follows the common practice, using 6 layers, 8 heads, and a dimension of 1024. For the word corruption function, word dropping and blanking adopts an uniform distribution with probability of 0.1 and word shuffling is implemented with a window of 3 tokens. The Adam optimizer uses a learning rate of 0.0001.

4 Experiments

4.1 Main Results

Adapting between Specific Domains Our main results are shown in Table 2, with the left six columns showing the set-

Methods	DE to EN									RO to EN		
	LAW		MED		IT		WMT-14			WMT-16		
	MED	IT	LAW	IT	LAW	MED	TED	LAW	MED	TED	LAW	MED
(1) UNADAPTED	18.76	6.62	7.92	5.94	6.19	10.90	23.36	23.77	24.42	23.59	33.26	18.39
(2) COPY	23.57	10.58	11.44	12.83	9.39	18.19	24.32	25.25	27.67	29.29	38.23	27.37
(3) BACK	33.94	22.21	23.74	23.56	22.43	31.00	31.02	31.27	35.69	36.98	49.28	43.70
(4) DALI	11.32	8.75	26.98	19.49	11.65	10.99	–	–	–	–	–	–
(5) DAFE	26.96	15.41	14.28	13.03	11.67	21.30	**34.89**	31.46	38.79	37.05†	49.63†	46.77†
(6) IBT	38.67	31.69	27.89	31.69	27.89	38.67	30.88	27.89	38.67	34.48	49.45	61.55
(7) IBT+OUTD	**41.22**	34.33	29.54	32.47	30.20	39.77	33.23	32.81	41.40	38.68	53.49	60.98
(8) IBT+BACK	40.40	**35.41**	**30.27**	**35.76**	**30.49**	**40.28**	34.15	**33.35**	**42.08**	**38.90**	**54.39**	**66.08**
(9) MT (Sup.)	48.95	59.38	37.72	59.38	37.72	48.95	38.97	37.72	48.95	42.22	61.69	80.32

Table 2: Translation accuracy (BLEU) under different settings. The second and third columns are source and target domains respectively. "DE", "EN", and "RO" denote German, English, and Romanian, respectively. DALI and DAFE results are the best results from the original papers, except that numbers marked by † are from our re-implementation. Settings (7) IBT+OUTD and (8) IBT+BACK uses the out-of-domain data and back-translated data to minimize the supervised machine translation loss, respectively. (9) "MT (Sup.)" results are obtained by training an NMT model on the supervised target domain data.

tings where models are adapted between specific domains. From this table, we see the unadapted baseline model, (1) UNADAPTED, performs very poorly, verifying the previous statement that current NMT models cannot generalize well to test data from a new domain. In contrast, the copy method, (2) COPY, and back-translation method, (3) BACK, can significantly improve the adaptation performance, and BACK showing much superior performance. To be noted, BACK even outperforms the other two baselines: DALI and DAFE, by a large margin in the majority of cases, although it was proposed earlier and is much simpler.

Our method IBT shown in row (6) of Table 2 achieves higher performance than all baselines, with absolute gains of $+0.91$ to $+9.48$ BLEU scores over the strongest baseline, and $+19.91$ to $+27.77$ BLEU scores over the unadapted baseline. Notably, our method IBT only needs the in-domain monolingual data but can still substantially outperform those baseline models that rely on the parallel out-of-domain data, indicating that previous methods have not exhausted the potential contained within the in-domain data.

Adapting from a General to a Specific Domain In the second data setting, the right 6 columns of Table 2 show the results by adapting an NMT model trained on the general domain corpus (WMT) to specific domains (TED, LAW, and MED) for two language pairs: from German to English and from Romanian to English. In this setting, both COPY and BACK achieve better performance compared to the previous setting, indicating that the out-of-domain data is in general helpful in the case of little in-domain parallel data. Our method IBT surpasses UNADAPTED by a large margin but it does not outperform the baselines BACK and DAFE in some cases. These two settings together form a comprehensive comparison between the two types of methods: one focuses on exploring the in-domain monolingual data internally (ours) and the other aims to extract information from out-of-domain parallel data and transfer it to in-domain data (previous methods). In cases where out-of-domain is of low resource or distant to in-domain data, our method can perform much better, while in the other cases, our method underperforms by a small margin.

Combining IBT with Out-of-Domain or Back-Translated Data Our method is model-centric and can be combined with any data-centric approaches, such as adding out-of-domain data or back-translated data. In row (7) of Table 2, during the training of IBT, we also insert the supervised translation training task on the out-of-domain data, which can bring in around $+1$ to $+4$ BLEU improvements consistently (except for the MED target domain in the ro-en language pair). Furthermore, instead of directly using the out-of-domain data, we can implement the supervised translation training on the back-translated pairs of in-domain data synthesized by the NMT model trained on the out-of-domain data, which achieves even better performance, as shown in the row (8) of Table 2. Although the data used for training in row (7) and (8) are all the same, IBT+Back still outperforms, indicating that training the model on in-domain data is always a better option than that on out-of-domain data. Overall, our best setting can harvest up to $+19.31$ BLEU improvement over the best baseline model and $+47.69$ BLEU improvement over the unadapted model. We have also tried combining the out-of-domain parallel data with the back-translated data together for supervised translation training but found it performs worse than current settings.

4.2 Ablation Study

We implemented an ablation study to check the contributions of the key components in our model, as summarized in Table 3. We first adapted the IBT+BACK model from the LAW domain to MED and IT, of which the validation set BLEU scores are reported in row (1) of Table 3. Comparing row (1) and row (2) in this table, we can see that if the model is not initialized with pre-trained parameters, the performance would suffer a substantial degradation by 10 to 20 BLEU scores. Moreover, both the language modeling (LM) as in Eq. (1) and iterative back-translation (IBT) as in Eq. (2) are integral to the model, especially IBT, as shown in row (3) and (4). Most interestingly, in addition to removing the iterative back-translation part, if we further remove the language modeling on the source-side sentences, as shown in row (6), the performance of adaptation is much better than that by removing only the iterative back-translation part. This indicates that when there is no back-translation training, language modeling on the source-side has a negative effect, which actually agrees with the intuition that in this case the decoder just needs to learn how to output fluent language in the target side.

Model Configuration	LAW	
	MED	IT
(1) IBT+BACK	**42.13**	**47.64**
(2) – Pre-training	31.80	27.71
(3) – IBT	33.75	25.37
(4) – LM	40.97	40.82
(5) – Source-side LM	40.04	42.66
(6) – BT – Source-side LM	37.29	35.06

Table 3: Ablation study by removing some key components from the IBT+BACK model when adapting from LAW domain to MED and IT domains. Validation set BLEU scores are reported. "– Pre-training" represents the model without pre-trained parameters. "– LM" and "– IBT"mean we remove the language modeling and iterative back-translation training for both source and target sides, respectively. "– Source-side LM" means we only remove the language modeling training for the source side.

5 Discussion

5.1 Pre-Training is Always Helpful

Through the experiments, we find that initializing the translation model with pre-trained parameters can benefit both our method and the baselines, as shown in Table 4. We compared three baselines (UNADAPTED, COPY, BACK) under two settings: with and without pre-training, when we adapted from the LAW domain to MED and IT. For all three adaptation methods, pre-training consistently leads to large improvements. This shows that good initialization of models is crucial to the unsupervised domain adaptation problem to circumvent the lack of supervised data.

Target	UNADAPTED		COPY		BACK	
	w/	w/o	w/	w/o	w/	w/o
MED	**17.44**	9.69	**24.38**	17.42	**36.61**	26.75
IT	**5.35**	3.87	**8.73**	5.07	**28.32**	9.21

Table 4: The comparison of three baselines: UNADAPTED, COPY, and BACK, between cases with and without pre-training when adapting them from the domain of LAW to MED and IT. Validation set BLEU scores are reported. Results show that pre-training can also benefit these baselines.

5.2 Do More In-Domain Monolingual Data Help?

One advantage of our method is that it keeps gaining higher performance if the monolingual data get larger. To verify this statement, we collected additional monolingual data and analyzed the performance difference before and after adding these extra data in Table 5. We considered two sources of extra monolingual data: out-of-domain and in-domain data. Specifically, we studied the adaptation from the WMT data to TED for both de-en and ro-en language pairs. The out-of-domain data source can be the WMT data itself, whereas for in-domain data, we collected an additional dataset of TED talks. After scraping all the TED talk web-pages[4] until the beginning of January, 2020, we extracted the transcripts in three languages, English, German and Romanian, and kept the unique TED talk identifier of the transcript.[5] Note that for any language pair, we made sure that the transcript of a TED talk only appeared once in either the source or the target side so as to avoid any parallel sentences. When combining the extra monolingual data with the original in-domain data, we always up-sample the in-domain data via replication so that it can have the same size as the extra data.

By comparing row (1) and (2) in Table 5, we see that the combination of out-of-domain and original in-domain monolingual data can be better than in-domain data alone for our method. And row (3) reveals that the addition of in-domain monolingual data can yield even larger improvement, which agrees with the intuition that in-domain data should always be more helpful to adaptation. We also added the in-domain data to our best setting IBT+BACK, and achieved even higher BLEU scores that are very close to supervised translation performance, as shown in row (6) and (7) in Table 5. This set of experiments have shown the great potential of our method: we can always seek to collecting more out-of-domain or in-domain monolingual data that are easily accessible to keep improving the unsupervised adaptation performance.

Model	Extra Data	WMT14	WMT16
		TED (DE-EN)	TED (RO-EN)
(1) IBT	None	30.88	34.48
(2) IBT	WMT	32.45	37.03
(3) IBT	TED	**33.34**	**39.01**
(4) IBT+BACK	None	34.15	38.90
(5) IBT+BACK	WMT	34.74	38.79
(6) IBT+BACK	TED	**36.45**	**40.92**
(7) MT (Sup.)	None	38.97	42.22

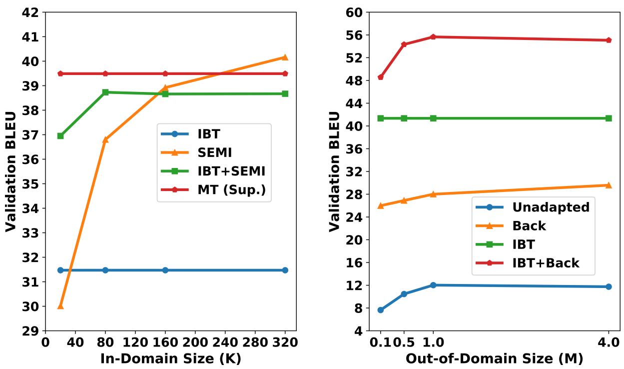
Table 5: We added more monolingual data to the original in-domain data and found they can further improve the adaptation performance (test set BLEU scores are reported). "WMT" extra data represent adding the out-of-domain data as monolingual data to the in-domain data, while "TED" means that we extracted extra monolingual data from the TED official website and added them to the in-domain data.

5.3 How Does Our Model Perform on Semi-supervised Adaptation?

Although this work targets at unsupervised domain adaptation, we are also curious to see the effect of our method on semi-supervised domain adaptation, with a limited number

[4]https://www.ted.com/

[5]Our newly collected TED dataset is available at https://github.com/zhijing-jin/ted_talk_downloader.



(a) IT-LAW

(b) WMT14-IT

Figure 2: Effect of training on increasing number of (a) in-domain and (b) out-of-domain parallel data.

of in-domain parallel sentence pairs. We first tested a standard semi-supervised setting, SEMI, by first pre-training an NMT model on the parallel corpus of the IT domain, and then fine-tuning it on the data that concatenate the out-of-domain IT data with sub-sampled in-domain LAW data. We sub-sampled 20K, 80K, 160K, and 320K real in-domain sentences. Note that during concatenation, if there are fewer sub-sampled in-domain data than the out-of-domain data, we repeat the in-domain data until it reaches the same size as the out-of-domain data. For comparison, we also use this combined data to train the supervised loss $\mathcal{L}_{\text{sup}}$ in our model (denoted as IBT+SEMI). We also plotted the unaffected performances of IBT and the supervised MT for clearer comparison in Figure 2a. As we can see, the performance of SEMI soars with the increasing data size of in-domain parallel data, while IBT+SEMI increases only sightly and to a limited extent. This indicates the limitation of our method in the semi-supervised domain adaptation task. On the other hand, in low-resource cases, where there are fewer than 160K in-domain data, our model IBT+SEMI is the best choice, as it prevails the normal semi-supervised training by a large margin.

5.4 How do Different Sizes of Out-of-Domain Data Affect?

In this section, we want to examine this question: *how does fine-tuning on different unadapted NMT models affect the adaptation performance?* To this end, we sub-sampled different sizes of out-of-domain parallel data, trained an NMT model (UNADAPTED) on them, and fine-tuned it in three ways: BACK, IBT, and IBT+BACK. Specifically, we adapted from the general domain (WMT14) to the IT domain for the German-English language pair, used all the in-domain monolingual data, and sub-sampled 0.1, 0.5, 1 and 4 million out-of-domain parallel data. As shown in Figure 2b, IBT does not use any out-of-domain data so it stays unchanged, while all the other settings demonstrate improved performance with the increasing number of out-of-domain data, and the performances slightly fluctuate when the number of out-of-domain data exceeds 1 million. Notably, IBT+BACK consistently outperforms all others by a large margin, and its performance also increases at a higher rate than others, indicating that our method makes better use of the out-of-domain supervised data.

6 Related Work

Most previous domain adaptation works for NMT focus on the supervised setting, where there is large out-of-domain parallel data and small amount of in-domain parallel data. Sequential fine-tuning [Luong and Manning, 2015; Freitag and Al-Onaizan, 2016] methods first train an NMT model on out-of-domain data and subsequently fine-tune it on the in-domain data. Britz *et al.* proposes to jointly train the translation task and the domain discrimination task to mitigate the domain shift. Kobus *et al.* uses the domain tokens and domain embeddings to force the NMT model to take into account the domain information. Joty *et al.* assigns higher weight to those out-of-domain data that more assemble the in-domain ones so as to remove unwanted noise. Our proposed method focuses on solving the adaptation problem with no in-domain parallel sentences, which is a more strict setting.

Unsupervised domain adaptation for NMT can be divided into two threads: data-centric and model-centric. Data-centric methods mainly propose to select or generate domain-related pseudo-parallel data for model training. In terms of data selection, Duh *et al.* uses language models to rank the out-of-domain data and select the top ranked parallel sentences as synthetic data. More representative methods are back translation-based [Sennrich *et al.*, 2015] and copy-based [Currey *et al.*, 2017], which are simple yet have been widely demonstrated to be effective. On the other hand, model-based methods have not been fully explored. This line of methods change model architectures to leverage monolingual corpus by introducing an extra learning objective, such as auto-encoding objective [Cheng *et al.*, 2016] and language modeling objective [Ramachandran *et al.*, 2017; Dou *et al.*, 2019] on the target-side sentences. In contrast, we propose utilizing online iterative back-translation to make better use of the in-domain data via an unsupervised manner.

7 Conclusion

In this paper, we identify that the iterative back-translation training scheme can bring large improvement to unsupervised domain adaptation for NMT. On three low-resource domains, this vanilla approach demonstrates improvements of up to +9.48 BLEU scores over the strongest out of four previous models, and up to +27.77 BLEU over the unadapted baseline. By further combining with popular data augmentation methods and utilizing supervised data from general domains on two language pairs, our model shows prevailing improvement of up to +19.31 BLEU higher than the strongest baseline model, and up to +47.69 over the unadapted model.

References

[Artetxe *et al.*, 2017] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

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

[Cheng *et al.*, 2016] Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Chu and Wang, 2018] Chenhui Chu and Rui Wang. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*, 2018.

[Currey *et al.*, 2017] Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. Copied monolingual data improves low-resource neural machine translation. In *WMT*, 2017.

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

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

[Duh *et al.*, 2013] Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[Duh, 2018] Kevin Duh. The multitarget ted talks task. http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/, 2018.

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

[Hu *et al.*, 2019] Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy, July 2019. Association for Computational Linguistics.

[Jin *et al.*, 2019] Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. Imat: Unsupervised text attribute transfer via iterative matching and 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3095–3107, 2019.

[Joty *et al.*, 2015] Shafiq Joty, Hassan Sajjad, Nadir Durrani, Kamla Al-Mannai, Ahmed Abdelali, and Stephan Vogel. How to avoid unwanted pregnancies: Domain adaptation using neural network models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1259–1270, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[Kobus *et al.*, 2016] Catherine Kobus, Josep Maria Crego, and Jean Senellart. Domain control for neural machine translation. In *RANLP*, 2016.

[Lample and Conneau, 2019] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *ArXiv*, abs/1901.07291, 2019.

[Lample *et al.*, 2018] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *EMNLP*, 2018.

[Luong and Manning, 2015] Minh-Thang Luong and Christopher D Manning. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79, 2015.

[Ramachandran *et al.*, 2017] Prajit Ramachandran, Peter Liu, and Quoc Le. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[Sennrich *et al.*, 2015] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

[Sennrich *et al.*, 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Tiedemann, 2012] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.