Iterative Domain-Repaired Back-Translation

Hao-Ran Wei, Zhirui Zhang, Boxing Chen and **Weihua Luo**
Machine Intelligence Technology Lab
Alibaba Group
Hangzhou, China
{funan.whr, zhirui.zzr, boxing.cbx, weihua.luowh}@alibaba-inc.com

Abstract

In this paper, we focus on the domain-specific translation with low resources, where in-domain parallel corpora are scarce or nonexistent. One common and effective strategy for this case is exploiting in-domain monolingual data with the back-translation method. However, the synthetic parallel data is very noisy because they are generated by imperfect out-of-domain systems, resulting in the poor performance of domain adaptation. To address this issue, we propose a novel iterative domain-repaired back-translation framework, which introduces the Domain-Repair (DR) model to refine translations in synthetic bilingual data. To this end, we construct corresponding data for the DR model training by round-trip translating the monolingual sentences, and then design the unified training framework to optimize paired DR and NMT models jointly. Experiments on adapting NMT models between specific domains and from the general domain to specific domains demonstrate the effectiveness of our proposed approach, achieving 15.79 and 4.47 BLEU improvements on average over unadapted models and back-translation.[1]

1 Introduction

Neural Machine Translation (NMT) has achieved impressive performance when large amounts of parallel sentences are available (Wu et al., 2016; Vaswani et al., 2017; Hassan et al., 2018). However, some previous works have shown that NMT models perform poorly in specific domains, especially when they are trained on the corpora from very distinct domains (Koehn and Knowles, 2017; Chu and Wang, 2018). The fine-tuning method (Luong and Manning, 2015) is a popular way to mitigate the effect of domain drift. However, it is not realistic to collect large amounts of high-quality parallel data in every domain we are interested in. Since monolingual in-domain data are usually abundant and easy to obtain, it is essential to explore the unsupervised domain adaptation scenario that utilizes large amounts of out-of-domain bilingual data and in-domain monolingual data.

One straightforward and effective solution for unsupervised domain adaptation is to build in-domain synthetic parallel data, including copying monolingual target sentences to the source side (Currey et al., 2017) or back-translation of in-domain monolingual target sentences (Sennrich et al., 2016; Dou et al., 2019). Although the back-translation approach has proven the superior effectiveness in exploiting monolingual data, directly applying this method in this scenario brings low-quality in-domain synthetic data. Table 1 gives two incorrect translation sentences generated by back-translation method. The main reason for this situation is that the synthetic parallel data is built by imperfect out-of-domain NMT systems, which leads to inappropriate word expressions or wrong translations. Fine-tuning on such synthetic data is very likely to hurt the performance of domain adaptation.

In this paper, we extend back-translation by a Domain-Repair (DR) model to explicitly remedy this issue. Specifically, the DR model is designed to re-generate in-domain source sentences given the synthetic data. In this way, the pseudo parallel data's source side can be re-written with the in-domain style, and some wrong translations are fixed. To optimize the DR model, we use the round-trip translation of monolingual source sentences to construct the corresponding training data.

Since source monolingual data is involved, it is natural to extend the back-translation method to bidirectional setting (Zhang et al., 2018), which

[1] Our code is released in https://github.com/whr94621/Iterative-Domain-Repaired-Back-Translation

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 5884–5893,
November 16–20, 2020.

SRC:	eine Gewichtszunahme wurde nach *Markteinführung* bei Patienten berichtet , denen ABILIFY verschrieben wurde .
REF:	weight gain has been reported **post-marketing** among patients prescribed ABILIFY .
MT:	a weight gain has been reported **after market introduction** in patients who have been prescribed ABILIFY .
SRC:	es werden möglicherweise nicht alle Packungsgrößen *in den Verkehr gebracht* .
REF:	not all pack sizes may be **marketed**.
MT:	it may not all pack sizes may be **added to the pack** .

Table 1: Two incorrect medical translations caused by the law-domain NMT model in German-English multi-domain datasets (Tiedemann, 2012), in which "*Markteinführung*" and "*in den Verkehr gebracht*" are translated to "after market introduction" and "added to the pack" respectively.

jointly optimizes source-to-target and target-to-source NMT models. Based on this setting, we propose the iterative domain-repaired back-translation (iter-DRBT) framework to fully exploit both source and target in-domain monolingual data. The whole framework starts with pre-trained out-of-domain bidirectional NMT models, and then these models are adopted to perform round-trip translation on monolingual data to obtain initial bidirectional DR models. Next, as illustrated in Figure 1, we design a unified training algorithm consisting of translation repair and round-trip translation procedures to jointly update DR and NMT models. More particularly, in the translation repair stage, the back-translated synthetic data can be well rewritten as in-domain sentences by the well-trained DR models to further improve NMT models. Then enhanced NMT models run the round-trip translation on monolingual data to build domain-mapping data, which helps DR models better identify mistakes made by the latest NMT models. This training process is iteratively carried out to make full use of the advantage of DR models to improve NMT models.

We evaluate our proposed method on German-English multi-domain datasets (Tiedemann, 2012). Experimental results on adapting NMT models between specific domains and from the general domain to specific domains show that our proposed method obtains 15.79 and 4.47 BLEU improvements on average over unadapted models and back-translation, respectively. Further analysis demonstrates the ability of DR models to repair the synthetic parallel data.

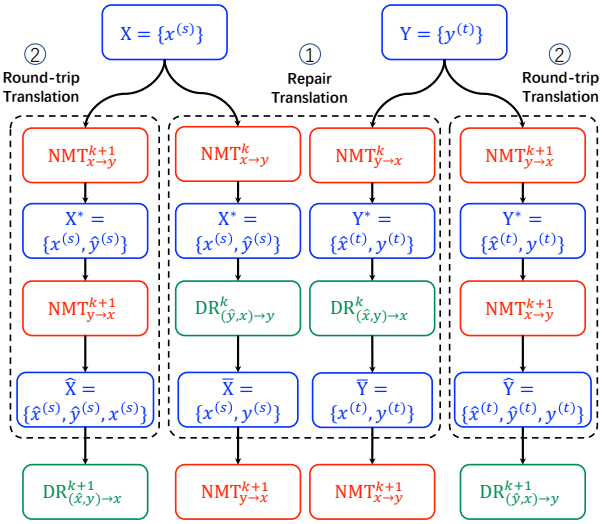

Figure 1: The training process of the iterative domain-repaired back-translation (iter-DRBT) framework at epoch k, where x and y represent the source and target sentences respectively, $\widehat{x}$ and $\widehat{y}$ denote the translation generated by NMT models. The whole framework consists of translation repair and round-trip translation procedures, which are used to generate corresponding training data for NMT and DR models respectively.

2 Related Work

Since in-domain parallel corpora are usually hard to obtain, many studies attempt to improve the performance of NMT models without any in-domain parallel sentences. One research line is to extract pseudo in-domain data from large amounts of out-domain parallel data. Biçici and Yuret (2011) use an in-domain held-out set to obtain parallel sentences from out-domain parallel sentences by computing n-gram overlaps. Instead, Moore and Lewis (2010), Axelrod et al. (2011) and Duh et al. (2013) use LMs score to select data similar to in-domain text. Recently, Chen et al. (2017) train a domain classifier to weight the out-domain training samples. There are also work on adaptation via retrieving sentences or n-grams in the training data similar to the test set (Farajian et al., 2017; Bapna and Firat, 2019). However, these methods cannot always guarantee to find domain-specific samples from out-domain data.

Another research direction is to exploit plenty of in-domain monolingual data, e.g., integrating a language model during decoding (Çaglar Gülçehre et al., 2015), copy method (Currey et al., 2017), back-translation (Sennrich et al., 2016) or obtaining domain-aware feature embedding via an auxiliary language modeling (Dou et al., 2019). Among

these approaches, back-translation is a widely used and effective method in exploiting monolingual data. Our proposed method is also based on back-translation and makes the most of it by improving the data quality with the DR model.

The methods of exploiting monolingual data in NMT can be naturally applied in unsupervised domain adaptation. Some studies are working on exploiting source-side monolingual data by self-training (Zhang and Zong, 2016; Chinea-Ríos et al., 2017) or pre-training (Yang et al., 2019; Weng et al., 2020; Ji et al., 2020), and leveraging both source and target monolingual data simultaneously by semi-supervised learning (Cheng et al., 2016), dual learning (He et al., 2016) and joint training (Zhang et al., 2018; Hoang et al., 2018). Our method utilizes both source and target data as well, with different that we use monolingual data to train bidirectional DR models, and then these models are used to fix pseudo data.

As back-translation is widely considered more effective than the self-training method, several works find that performance of back-translation degrades due to the less rich translation or domain mismatch at the source side of the synthetic data (Edunov et al., 2018; Caswell et al., 2019). Edunov et al. (2018) attempt to use sampling instead of maximum a-posterior when decoding with the reverse direction model. Imamura et al. (2018) add noises to the results of beam search. Caswell et al. (2019) propose to add a tag token at the source side of the synthetic data. Unlike their methods, our method leverages the DR model to re-generate the source side of the synthetic data, which can also increase translation diversity and mitigate the effect of different domains.

3 Iterative Domain-Repaired Back-Translation

In this section, we first illustrate the overview of iter-DRBT framework, then describe the architecture of DR model and the joint training strategy.

3.1 Overview

Suppose that we have non-parallel in-domain monolingual sentences $X = \{x^{(s)}\}$ and $Y = \{y^{(t)}\}$ in two languages respectively, as well as two pre-trained out-of-domain translation models $\text{NMT}^0_{x\to y}$ and $\text{NMT}^0_{y\to x}$, where x and y denote the source and target sentences respectively. The purpose of unsupervised domain adaptation is to train in-domain models $\text{NMT}_{x\to y}$ and $\text{NMT}_{y\to x}$.

In this work, we incorporate a Domain-Repair (DR) model in the iterative back-translation process to fully exploit in-domain monolingual data, in which the DR model is used to refine translation sentences given the synthetic bilingual sentences. The whole framework consists of translation repair and round-trip translation procedures, which are used to generate corresponding training data for NMT and DR models, respectively. For convenience, we take source-to-target translation ($x \to y$) as an example to explain the usage of our proposed method.

Translation Repair Stage. The basic process of back-translation method is to first translate $y^{(t)}$ into $\widehat{x}^{(t)}$ with $\text{NMT}^0_{y\to x}$, and then fine-tune $\text{NMT}^0_{x\to y}$ on the synthetic parallel data $Y^* = \{\widehat{x}^{(t)}, y^{(t)}\}$. As the model $\text{NMT}^0_{y\to x}$ is not trained on truly in-domain bilingual data, there exists domain mismatch between $\widehat{x}^{(t)}$ and the genuine in-domain sentences x. Given the synthetic parallel data $Y^* = \{\widehat{x}^{(t)}, y^{(t)}\}$, we apply the corresponding DR model ($\text{DR}_{(\widehat{x},y)\to x}$) to repair errors in translated sentences, e.g. wrong translations of in-domain phrases or domain-inconsistent expressions, and then obtain the new synthetic parallel data $\overline{Y} = \{x^{(t)}, y^{(t)}\}$ to train $\text{NMT}_{x\to y}$ initialized with $\text{NMT}^0_{x\to y}$.

Round-Trip Translation Stage. In order to optimize $\text{DR}_{(\widehat{x},y)\to x}$, we use the round-trip translation of monolingual source sentences $X = \{x^{(s)}\}$ to construct the corresponding training data $\widehat{X} = \{\widehat{x}^{(s)}, \widehat{y}^{(s)}, x^{(s)}\}$, where $\widehat{y}^{(s)}$ and $\widehat{x}^{(s)}$ are generated by $\text{NMT}^0_{x\to y}$ and $\text{NMT}^0_{y\to x}$ respectively ($x^{(s)} \to \widehat{y}^{(s)} \to \widehat{x}^{(s)}$). In this way, $\text{DR}_{(\widehat{x},y)\to x}$ learns to identify mistakes made by $\text{NMT}^0_{y\to x}$ and corresponding mapping rules, which helps to better fix the errors in synthetic parallel data.

Similarly, these two stages are also applied in the reverse translation direction to train target-to-source NMT model ($\text{NMT}_{y\to x}$) and corresponding DR model ($\text{DR}_{(\widehat{y},x)\to y}$). As illustrated in Figure 1, it is natural to extend such a training process to a joint training framework, which alternately carries out the translation repair and round-trip translation procedures to make full use of the advantage of DR models to improve NMT models.

3.2 Domain-Repair Model

Since the DR model takes the synthetic bilingual sentences as input to produce the in-domain sen-

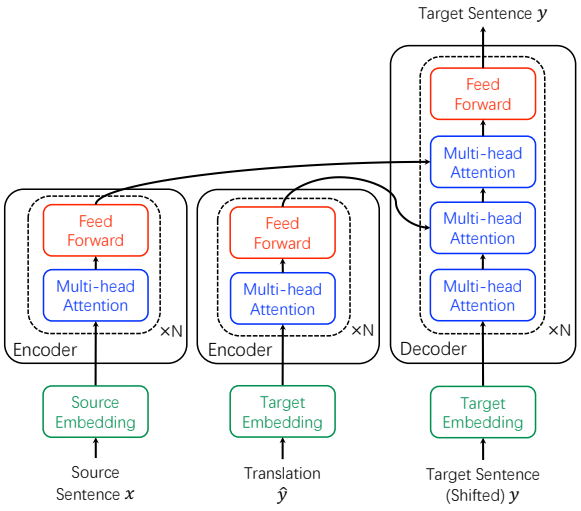

Figure 2: The dual-source transformer architecture of the Domain-Repair model ($\text{DR}_{(\widehat{y},x)\to y}$). For simplicity, we omit some architecture details such as layer normalization and residual connection.

tences, we parameterize the DR model as a dual-source sequence-to-sequence model. As illustrated in Figure 2, the dual-source transformer model naturally extends the original architecture from Vaswani et al. (2017) by adding another encoder for translated sentences and stacking an additional multi-head attention component above the multi-head self-attention component. As usual for the transformer architecture, each block is followed by a skip connection from the previous input and layer normalization. For simplicity, we omit these architecture details in Figure 2.

Our proposed framework involves two DR models ($\text{DR}_{(\widehat{x},y)\to x}$ and $\text{DR}_{(\widehat{y},x)\to y}$), both of which are optimized by maximizing the conditional log likelihood on the training corpus $\widehat{X} = \{\widehat{x}^{(s)}, \widehat{y}^{(s)}, x^{(s)}\}$ and $\widehat{Y} = \{\widehat{x}^{(t)}, \widehat{y}^{(t)}, y^{(t)}\}$ built by round-trip translation respectively:

$$\mathcal{L}_1(\theta_1) = \sum_{s=1}^{|\widehat{X}|} \log P(x^{(s)}|\widehat{y}^{(s)}, \widehat{x}^{(s)}; \theta_1) \quad (1)$$

$$\mathcal{L}_2(\theta_2) = \sum_{t=1}^{|\widehat{Y}|} \log P(y^{(t)}|\widehat{x}^{(t)}, \widehat{y}^{(t)}; \theta_2) \quad (2)$$

where θ_1 and θ_2 denote the model parameters of $\text{DR}_{(\widehat{x},y)\to x}$ and $\text{DR}_{(\widehat{y},x)\to y}$ respectively.

3.3 Joint Training Strategy

We design the iterative training framework to jointly optimize DR and NMT models, as illustrated in Algorithm 1. The whole training framework starts with pre-trained out-of-domain bidirectional NMT models ($\text{NMT}^0_{x\to y}$ and $\text{NMT}^0_{y\to x}$) and in-domain monolingual data ($X = \{x^{(s)}\}$ and $Y = \{y^{(t)}\}$). To train initial DR models, we use $\text{NMT}^0_{x\to y}$ and $\text{NMT}^0_{y\to x}$ to run round-trip translation on X and Y to construct dataset $\widehat{X} = \{\widehat{x}^{(s)}, \widehat{y}^{(s)}, x^{(s)}\}$ and $\widehat{Y} = \{\widehat{x}^{(t)}, \widehat{y}^{(t)}, y^{(t)}\}$;

Based on initial NMT and DR models, a joint training process is iteratively carried out to further optimize these models. This process consists of translation repair and round-trip translation stages. In the translation repair stage, we first adopt NMT models to translate monolingual data, based on which the DR models are used to further re-write the translated sentences as in-domain sentences. In this way, we can obtain better in-domain synthetic data to further improve NMT models. Next, in the round-trip translation stage, we perform round-trip translation on monolingual data with enhanced NMT models to re-build training data for DR models. The DR models trained on such datasets can better identify mistakes made by latest NMT models ($\text{NMT}^{k+1}_{x\to y}$ and $\text{NMT}^{k+1}_{y\to x}$) and learn correspond-

Algorithm 1: Joint Training Algorithm for NMT and DR Models

Input: pre-trained out-of-domain models $\text{NMT}^0_{x\to y}$ and $\text{NMT}^0_{y\to x}$, in-domain monolingual sentences $X = \{x^{(s)}\}$ and $Y = \{y^{(t)}\}$, maximum iteration number T
Use $\text{NMT}^0_{x\to y}$ and $\text{NMT}^0_{y\to x}$ to perform round-trip translation on X and Y to construct dataset $\widehat{X} = \{\widehat{x}^{(s)}, \widehat{y}^{(s)}, x^{(s)}\}$ and $\widehat{Y} = \{\widehat{x}^{(t)}, \widehat{y}^{(t)}, y^{(t)}\}$;
Train $\text{DR}^0_{(\widehat{x},y)\to x}$ and $\text{DR}^0_{(\widehat{y},x)\to y}$ with $\widehat{X}$ and $\widehat{Y}$;
$k = 0$;
for $k \leq T$ **do**
  Translation Repair Stage:
  Use $\text{NMT}^k_{x\to y}$ and $\text{NMT}^k_{y\to x}$ to build synthetic data $X^* = \{x^{(s)}, \widehat{y}^{(s)}\}$ and $Y^* = \{\widehat{x}^{(t)}, y^{(t)}\}$ for X and Y respectively;
  Use $\text{DR}^k_{(\widehat{y},x)\to y}$ and $\text{DR}^k_{(\widehat{x},y)\to x}$ to repair X^* and Y^* to construct in-domain synthetic data $\overline{X} = \{x^{(s)}, y^{(s)}\}$ and $\overline{Y} = \{x^{(t)}, y^{(t)}\}$;
  Update NMT Models:
  $\text{NMT}^{k+1}_{x\to y} \leftarrow$ Fine-tune $\text{NMT}^k_{x\to y}$ with $\overline{Y}$;
  $\text{NMT}^{k+1}_{y\to x} \leftarrow$ Fine-tune $\text{NMT}^k_{y\to x}$ with $\overline{X}$;
  Round-Trip Translation Stage:
  Use $\text{NMT}^{k+1}_{x\to y}$ and $\text{NMT}^{k+1}_{y\to x}$ to perform round-trip translation on X and Y to construct corresponding dataset $\widehat{X} = \{\widehat{x}^{(s)}, \widehat{y}^{(s)}, x^{(s)}\}$ and $\widehat{Y} = \{\widehat{x}^{(t)}, \widehat{y}^{(t)}, y^{(t)}\}$;
  Update DR Models:
  $\text{DR}^{k+1}_{(\widehat{x},y)\to x} \leftarrow$ Fine-tune $\text{DR}^k_{(\widehat{x},y)\to x}$ with $\widehat{X}$;
  $\text{DR}^{k+1}_{(\widehat{y},x)\to y} \leftarrow$ Fine-tune $\text{DR}^k_{(\widehat{y},x)\to y}$ with $\widehat{Y}$;
  $k = k + 1$

Domains	LAW	MEDICAL
#Bi.	377,114	328,132
#Mono. (de)	187,550	171,906
#Mono. (en)	189,564	156,226
#Dev	4,233	1,141
#Test	4,063	1,272

Table 2: Statistics on bilingual, monolingual, development and test data of medical and law domains.

ing mapping rules, which helps to better fix the synthetic parallel data in the next iteration. Note that we fine-tune the NMT and DR models in each iteration to speed up the whole training process.

4 Experiments

4.1 Setup

Datasets. To evaluate the performance of our proposed method, we adopt a multi-domain dataset released by Koehn and Knowles (2017), which is further built as an unaligned monolingual corpus in Hu et al. (2019). However, there are two issues in the train/dev/test splits used in Hu et al. (2019). First, Ma et al. (2019) and Dou et al. (2020) find that some same sentence pairs exist between the training and test data. Second, Hu et al. (2019) randomly shuffle the bi-text data and split it into halves, which may bring more overlap than in natural monolingual data, i.e., bilingual sentences from a document are probably selected into monolingual data (e.g., one sentence on the source split and its translation on the target split).

To address the impact of the above two issues, we re-collect in-domain monolingual data and test sets in the following steps:

- Download the XML files from OPUS[2], extract parallel corpus from each documents and record the document boundaries.
- Randomly take some documents as dev/test sets and use the rest as training data.
- Divide the training set into two parts, where the number of sentences in the two parts is similar. Then the source and target sentences of the first and second halves are chosen as monolingual data, respectively.
- De-duplicate all overlap sentences within train/dev/test sets.

We choose medical (EMEA) and law (JRC-Acquis) domains for our experiments. All the data statistics are illustrated in Table 2.

[2]http://opus.nlpl.eu/

Experimental Details. We implement all NMT models with *Transformer_base* (Vaswani et al., 2017). More specifically, the number of layers in the encoder and decoder is set to 6, with 8 attention heads in each layer. Each layer in both encoder and decoder has the same dimension of input and output $d_{\text{model}} = 512$, dimension of feed-forward layer's inner-layer $d_{\text{hidden}} = 2048$. Besides, DR models follow the same setting as the NMT model.

The Adam (Kingma and Ba, 2014) algorithm is used to update DR and NMT models. For training initial NMT and DR models, following the setting of Hu et al. (2019), we set the dropout as 0.1 and the label smoothing coefficient as 0.2. Besides, we adopt the setting of *Fairseq* (Ott et al., 2019) on IWSLT'14 German to English to fine-tune NMT and DR models. During training, we schedule the learning rate with the inverse square root decay scheme, in which the warm-up step is set as 4000, and the maximum learning rate is set as 1e-3 and 5e-4 for pre-training and fine-tuning, respectively.

For the joint training strategy, we set the maximum iteration number T in Algorithm 1 as 2 for balancing speed and performance. In practice, we train our framework on 2 Tesla P100 GPUs for all tasks, and it takes 2 days to finish the whole training.

Methods. We compare our approach with several baseline methods in our experiment:

- **Base:** Directly use out-of-domain NMT models to evaluate on in-domain test sets.
- **Copy:** Copy the target in-domain monolingual data to the source side as parallel data.
- **BT:** Back-translation method, which fine-tunes the out-domain model on synthetic training data generated by a target-to-source out-domain NMT model.
- **DALI-BT:** Using word translation instead of back-translation to generate synthetic parallel data. Such data can be mixed with common back-translation for domain adaptation (Hu et al., 2019).
- **iter-BT:** Iterative back-translation, which alternatively generates synthetic data and optimizes NMT models at both side (Hoang et al., 2018). We adopt the same iteration number as iter-DRBT.
- **DRBT:** The simplified version of our proposed method, in which we only use the DR model to repair synthetic data once.

All experimental results are evaluated by *Sacre-*

Methods	MED2LAW		LAW2MED		Ave.	WMT2LAW		WMT2MED		Ave.
	DE2EN	EN2DE	DE2EN	EN2DE		DE2EN	EN2DE	DE2EN	EN2DE	
Base	19.81	19.91	27.27	25.46	23.11	42.17	36.46	37.01	34.94	37.65
Copy	20.34	20.51	29.59	27.95	24.60	42.52	36.71	37.43	37.39	38.51
BT	35.84	32.47	42.84	38.13	37.32	49.07	42.50	49.72	43.04	46.08
DALI-BT	36.38	33.40	44.76	39.20	38.44	49.58	42.85	50.23	43.23	46.47
DRBT	39.64	35.42	45.81	41.17	40.51	50.41	45.24	50.69	45.13	47.87
iter-BT	40.72	33.29	45.66	40.51	40.05	49.97	44.73	51.15	45.70	47.89
iter-DRBT	**43.42**	**37.94**	**48.69**	**44.60**	**43.66**	**51.15**	**46.14**	**51.37**	**46.04**	**48.68**

Table 3: BLEU scores(%) under different settings. The left four columns are results of adapting between two distinct domains, while the right four domains are results of adapting from the general domain (WMT) to specific domains.

BLEU (Post, 2018) in terms of case-sensitive tokenized BLEU (Papineni et al., 2002).

4.2 Main Results

Adapting between Specific Domains. We verify our approach by adapting NMT models from one distinct domain to another. As illustrated in the left four columns of Table 3, the unadapted models perform poorly on the out-of-domain test sets. Besides, the Copy and BT can improve the performance on target domains, in which the back-translation method achieving more improvements consistently. We reproduce Hu et al. (2019)'s work, and their method combined with back-translation (DALI-BT) gains better performance. Our proposed method (DRBT) significantly outperforms all previous methods on all four translation tasks, achieving up to average 17.40 and 2.08 BLEU improvements compared to Base and DALI-BT, respectively. It demonstrates that the DR model effectively repairs the errors occurred by out-of-domain models, improving the performance of unsupervised domain adaptation.

As the back-translation method suffers from low-quality synthetic data, iter-BT is used to improve the quality of synthetic data and achieves 2.73 BLEU improvements on average, but it still has 0.46 BLEU behind DRBT. This result indicates that the DR model shows a better ability to repair the imperfections of synthetic data. The joint training of DR and NMT models (iter-DRBT) can further obtain 3.15 BLEU improvements compared to DRBT. It also proves that the joint training process helps DR models to better identify mistakes made by the latest NMT models and fix the synthetic parallel data in the following iteration.

#Para.	BT	DRBT	iter-DRBT	Sup.
1K	46.03	48.98	51.30	61.56
5K	49.30	53.59	54.93	61.74
10K	51.32	54.30	56.04	62.07
50K	57.99	59.29	60.03	62.81

Table 4: BLEU scores(%) of DRBT and iter-DRBT under semi-supervised scenario with varied size of in-domain parallel data. We also report supervised results with all the in-domain parallel (Sup.) as upper bound.

Adapting from General to Specific Domains. We further evaluate our method when adapting a model trained on large amounts of general domain data. We use out-of-domain models trained on the WMT14 German-English dataset and adapt them to the Medical and Law domains, respectively. All results are shown in the right half of Table 3.

These results show a similar pattern as previous experiments, except that the gap between our method and BT/iter-BT is reduced. We attribute this reduction to the improvements of general models on in-domain translation. Even so, the iter-DRBT yields the best performance on all test sets, with 11.03 and 0.79 BLEU improvements on average compared to Base and iter-BT, respectively.

Semi-supervised Adaptation. Our method can be easily applied in semi-supervised domain adaptation, with a limited number of in-domain parallel data available. The implementation in this setting is to mix the in-domain parallel data with the generated synthetic data for NMT models training. In addition to the round-translation on monolingual data, we conduct back-translation on parallel data to construct corresponding training data for DR models training.

We conduct experiments on adapting German-to-English NMT models from the Law domain to

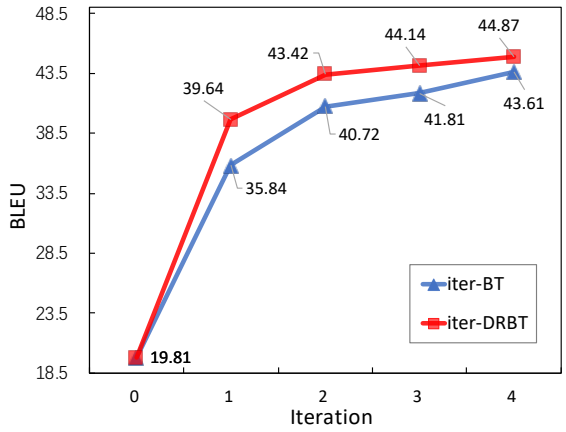

Figure 3: BLEU scores(%) at different iterations of joint training. The model at '0'-th iteration is the unadapted model.

the Medical domain. To assess performance under different scales of in-domain parallel data, we fix the number of monolingual in-domain sentences and vary the number of in-domain parallel sentences in 1K, 5K, 10K, and 50K. We also report the results of fine-tuning on full in-domain parallel data, including additional in-domain parallel data and monolingual data paired with its original translations, to indicate the upper bound of semi-supervised training. All the results are listed in Table 4. We observe the consistent improvement of our proposed method. It is worth noting that given 50K in-domain parallel data, the gap between using repaired synthetic data and using the actual parallel data is rapidly reduced from 12.58 to 3.52 BLEU, and further decreased to only 2.78 by joint-training with one more iteration, demonstrating the effectiveness of our method in the semi-supervised scenario.

4.3 Effect of Joint Training

We further investigate the effect of joint training with more iterations. Specifically, we conduct experiments on adapting from the Medical domain to the LAW domain from German to English, in which iterative back-translation is used for comparison.

We plot the BLEU curve of these two methods over the number of iterations. From Figure 3, we can observe that our proposed method (iter-DRBT) consistently outperforms iterative back-translation (iter-BT) under the same number of iterations. As the number of iterations increases, BLEU improvement achieved by iter-DRBT and iter-BT gradually decreases, but the gap remains.

	w/o DR	w/ DR	Δ
LAW2MED	24.84/26.54	36.10/41.06	11.2/14.5
MED2LAW	18.45/18.46	29.80/34.53	11.3/16.0
WMT2MED	32.62/35.59	41.50/46.57	8.8/10.8
WMT2LAW	34.61/39.87	39.48/46.96	4.8/7.0

Table 5: BLEU scores(%) (German/English) on development sets before and after applying DR models.

4.4 Analysis of Domain Repair Models

In this section, we mainly discuss how DR models repair the source side of synthetic data to improve its quality. Compared to the original back-translation data, we find that the change comes from three main points: an improvement in the overall quality of the source side, an improvement in the accuracy of the in-domain lexical translation, and a closer in-domain style of the source side.

Improvement of Translation Quality. We first assess the change in translation quality at the source side of back-translation data. We report the BLEU changes on all the development sets before and after using the DR model. All the results are listed in Table 5. We can see that the source side of the back translation data generated by the out domain model is inferior at the initial stage. The DR model significantly improves its quality, which improves the effectiveness of back-translation.

Improvement of Lexical Translation. We then assess the change in lexical translation at the source side of synthetic data before and after domain repair. Based on the frequency of words that appear in the out-of-domain training data, we allocate target side words of development sets into three buckets (< 1, $[1, 20)$ and ≥ 20, which represent zero-shot words, few-shot words, and frequent words, respectively), and compute the word translation f-scores within each bucket. We use *compare-mt* (Neubig et al., 2019) to do all the analysis and plot the results in Figure 4. We can see that the synthetic data repaired by DR models show better word translation in all the buckets. It is worth noting that the improvement of word translation f-scores on zero/few-shot (< 20) words dramatically exceeds that on frequent words, which shows that DR models are especially good at repairing in-domain lexical mistranslations.

Improvement of Domain Consistent Style. We further evaluate how can DR models remedy the domain mismatch issue at the source side of back-

SRC:	Arzneimittel , deren Plasmaspiegel bei gemeinsamer Anwendung mit Telzir erhöht sein können
REF:	Medicinal products whose plasma levels may be increased when **co-administered** with Telzir
w/o DR:	Medicinal products whose plasma ponds may be increased if they are **commonly used** by telzir
w/ DR:	Medicinal products whose plasma aspiegel may be increased when **co-administered** with Telzir
SRC:	Johanniskraut (Hypericum perforatum) Die Serumspiegel von Amprenavir und Ritonavir können durch die gleichzeitige Anwendung von pflanzlichen Zubereitungen mit Johanniskraut (Hypericum perforatum) erniedrigt werden .
REF:	**St John's wort** (Hypericum perforatum) Serum levels of amprenavir and ritonavir can be reduced by **concomitant use of the herbal** preparation St John's wort (Hypericum perforatum) .
w/o DR:	**Johanniskraut** (Hypericum perforatum) The serum levels of Amprenavir and Ritonavir can be reduced by **the simultaneous use of plant** preparations with currant (hypericum perforatum) .
w/ DR:	**St. John's wort** (Hypericum perforatum) Serum levels of amprenavir and ritonavir can be stratified by **concomitant use of herbal** preparations containing St John's wort (Hypericum perforatum) .

Table 6: Cases of sentences that are repaired by DR Model. Inappropriate translations are marked with blue wave lines while corresponding corrections are marked with red underlines.

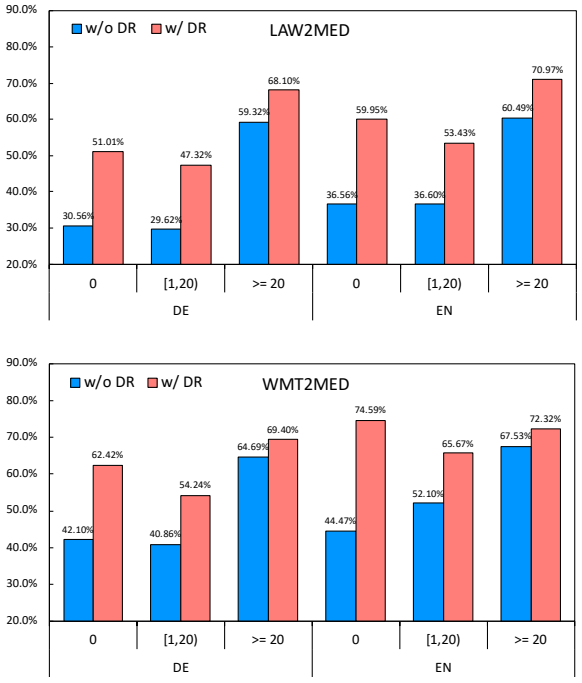

Figure 4: F-measures of the word translation on medical development set bucketed by the frequency of words occurring in the out-Of-domain training data.

translated data, including domain inconsistent word selection and language style. We evaluate them by observing the perplexity change measured by in-domain and out-of-domain language models before and after being repaired, in which all the language models are trained with *KenLM* (Heafield, 2011). The out-of-domain language models are trained on out-of-domain training data, while in-domain language models are trained on the original translations of in-domain monolingual data. We list all the perplexity scores in Table 7. On both MED2LAW and WMT2LAW, we observe a consistent bias of perplexity scores towards in-domain language models, which demonstrates that DR models correct the expression of the source side of synthetic data to be more domain consistent.

	Out-of-domain LM	In-domain LM
	MED2LAW	
w/o DR	15.04/11.16	10.93/9.13
w/ DR	**21.17/18.03** ↑↑	**7.27/6.57** ↓↓
	WMT2LAW	
w/o DR	12.29/9.23	8.30/6.54
w/ DR	**13.60/9.96** ↑↑	**7.31/5.69** ↓↓

Table 7: Perplexity of synthetic data's source side scores by both in/out domain language models before and after domain repair.

Case Study. We provide some examples to display how DR models improve the synthetic data. As shown in Table 6, the DR model can reduce some mistranslation, such as correcting the translation of "Johanniskraut" into "St John's wort", as well as generating more domain-related expressions, like "co-administered" and "concomitant use of herbal preparations". This shows the ability of domain repair models to improve the quality and domain consistency of synthetic data generated by imperfect out-of-domain NMT models.

5 Conclusion

In this paper, we argue that back-translation, the predominant unsupervised domain adaptation method in neural machine translation, suffers from the domain shift, restricting the performance of unsupervised domain adaptation. We propose to remedy this mismatch by leveraging a domain repair model that corrects the errors in back-translation sentences. Then the iterative domain-repaired back-

translation framework is designed to make full use of the advantage of the domain repair model. Experiments on adapting translation models between specific domains and from general domain to specific domains demonstrate the effectiveness of our method, achieving significant improvements over strong back-translation baselines.

In the future, we would like to extend our method to enhance the back-translation method in multi-domain settings.

Acknowledgments

We would like to thank the anonymous reviewers for the helpful comments. This work is supported by National Key R&D Program of China (2018YFB1403202).

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *EMNLP*.

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

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *EMNLP*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63.

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

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. *ArXiv*, abs/1606.04596.

Mara Chinea-Ríos, Álvaro Peris, and Francisco Casacuberta. 2017. Adapting neural machine translation with parallel synthetic data. In *Proceedings of the Second Conference on Machine Translation*, pages 138–147.

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

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

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. *ArXiv*, abs/2004.03672.

Zi-Yi Dou, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. 2019. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *ACL*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

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

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *ArXiv*, abs/1503.03535.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mengnan Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *ArXiv*, abs/1803.05567.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime G. Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *ACL*.

Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita. 2018. Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 55–63.

Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. 2020. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In *AAAI*.

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 *NMT@ACL*.

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

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with BERT-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*.

Robert C. Moore and William D. Lewis. 2010. Intelligent selection of language model training data. In *ACL*.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL) Demo Track*.

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 *NAACL HLT 2019*, page 48.

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

Matt Post. 2018. A call for clarity in reporting bleu scores. In *WMT*.

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.

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

Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, and Weihua Luo. 2020. Acquiring knowledge from pre-trained model to neural machine translation. In *AAAI*.

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, Gregory S. 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.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Yong Yu, Weinan Zhang, and Lei Li. 2019. Towards making the most of bert in neural machine translation. *arXiv preprint arXiv:1908.05672*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.