Understanding and Improving Sequence-to-Sequence Pretraining for Neural Machine Translation

Anonymous ACL submission

Abstract

In this paper, we present a substantial step in better understanding the SOTA sequence-to-sequence (Seq2Seq) pretraining for neural machine translation (NMT). We focus on studying the impact of the jointly pretrained decoder, which is the main difference between Seq2Seq pretraining and previous encoder-based pretraining approaches for NMT. By carefully designing experiments on three language pairs, we find that Seq2Seq pretraining is a double-edged sword: On one hand, it helps NMT models to produce more diverse translations and reduce adequacy-related translation errors. On the other hand, the discrepancies between Seq2Seq pretraining and NMT finetuning limit the translation quality (i.e., domain discrepancy) and induce the over-estimation issue (i.e., objective discrepancy). Based on these observations, we further propose simple and effective strategies, named in-domain pretraining and input adaptation to remedy the domain and objective discrepancies, respectively. Experimental results on several language pairs show that our approach can consistently improve both translation performance and model robustness upon Seq2Seq pretraining.

1 Introduction

There has been a wealth of research over the past several years on self-supervised pre-training for natural language processing tasks (Devlin et al., 2019; Liu et al., 2019; Conneau et al., 2020), which aims at transferring the knowledge of large-scale unlabeled data to downstream tasks with labeled data. Despite its success in other understanding and generation tasks, self-supervised pretraining is not a common practice in machine translation (MT). One possible reason is the architecture discrepancy between pretraining model (e.g., Transformer **encoder**) and NMT models (e.g., Transformer **encoder-decoder**).

To remedy the architecture gap, several researchers propose sequence-to-sequence (Seq2Seq) pretraining models for machine translation, e.g., MASS (Song et al., 2019) and BART (Zhu et al., 2019; Lewis et al., 2020). Recently, Liu et al. (2020) extend BART by training on large-scale multilingual language data (i.e., mBART), leading to significant improvement on translation performance across various language pairs. While previous pretraining approaches for NMT generally focus only on Transformer encoder (Lample and Conneau, 2019), mBART pretrains a complete autoregressive Seq2Seq model by recovering the input sentences that are noised by masking phrases. One research question naturally arises: *how much does the jointly pretrained decoder matter?*

In this work, we present a substantial step in better understanding the SOTA Seq2Seq pretraining model. We take a fine-grained look at the impact of the jointly pretrained decoder by carefully designing experiments, which are conducted on several WMT and IWSLT benchmarks across language pairs and data scales using the released mBART-25 model (Liu et al., 2020). By carefully examining the translation outputs, we find that (§ 2.2):

- Jointly pretraining decoder produces more diverse translations with different word orders, which calls for multiple references to accurately evaluate its effectiveness on large-scale data.
- Jointly pretraining decoder consistently reduces adequacy-related translation errors over pretraining encoder only.

Although jointly pretraining decoder consistently improves translation performance, we also identify several side effects due to the discrepancies between pretraining and finetuning (§2.3):

- **domain discrepancy**: Seq2Seq pretraining model is generally trained on general domain data while the downstream translation models are trained on specific domains (e.g., news). The domain discrepancy requires more efforts for the

finetuned model to adapt the knowledge in pretrained models to the target in-domain.

- **objective discrepancy**: NMT training learns to translate a sentence from one language to another, while Seq2Seq pretraining learns to reconstruct the input sentence. The objective discrepancy induces the over-estimation issue and tends to generate more hallucinations with noisy input. The over-estimation problem along with more copying translations induced by Seq2Seq pretraining (Liu et al., 2021) make it suffer from more serious beam search degradation problem.

To remedy the above discrepancies, we propose simple and effective strategies, named in-domain pretraining and input adaptation in finetuning (§3). In in-domain pretraining, we propose to reduce the domain shift by continuing the pretraining of mBART on in-domain monolingual data, which is more similar in data distribution with the downstream translation tasks. For input adaptation, we add noises to the source sentence of bilingual data, and combine the noisy data with the clean bilingual data for finetuning. We expect the perturbed inputs to better transfer the knowledge from pretrained model to the finetuned model. Experimental results on the benchmark datasets show that in-domain pretraining improves the translation performance significantly and input adaptation enhances the robustness of NMT models. Combining the two approaches gives us the final solution to a well-performing NMT system. Extensive analyses show that our approach can narrow the domain discrepancy, particularly improving the translation of low-frequency words. Besides, our approach can alleviate the over-estimation issue and mitigate the beam search degradation problem of NMT models.

2 Understanding Seq2Seq Pretraining

In this section, we conduct experiments and analyses to gain a better understanding of current Seq2Seq pretraining for NMT. We first present the translation performance of the pretrained components (§2.2), and then show the discrepancy between pretraining and finetuning (§2.3).

2.1 Experimental Setup

Data. We conduct experiments on several benchmarks across language pairs, including high-resource WMT19 English-German (W19 En-De, 36.8M instances), and low-resource WMT16 English-Romanian (W16 En-Ro, 610K instances) and IWSLT17 English-French (I17 En-Fr, 250K instances). To eliminate the effect of different languages, we also sample a subset from WMT19 En-De (i.e., W19 En-De (S), 610K instances) to construct a low-resource setting for ablation studies.

For the proposed *in-domain pretraining*, we collect the NewsCrawl monolingual data as the in-domain data for WMT tasks (i.e., 200M English, 200M German, and 60M Romanian), and the TED monolingual data for IWSLT tasks (i.e., 1M English and 0.9M French). Since the monolingual data from TED is rare, we expand it with pseudo in-domain data, OpenSubtitle (Tiedemann, 2016), which also provides spoken languages as TED. Specifically, we use the latest 200M English subtitles and all the available French subtitles (i.e., 100M). We follow Liu et al. (2020) to use their released sentence-piece model (Kudo and Richardson, 2018) with 250K subwords to tokenize both bilingual and monolingual data. We evaluate the translation performance using the SacreBLEU (Post, 2018).

Models. As for the pretrained models, we adopt the officially released mBART25 model (Liu et al., 2020)[1], which is trained on the large-scale CommonCrawl (CC) monolingual data in 25 languages. As a result, the vocabulary is very large in mBART25, including 250K words. mBART uses a larger Transformer model which extends both the encoder and decoder of Transformer-Big to 12 layers. We use the parameters of either encoder or encoder-decoder from the pretrained mBART25 for finetuning. Then, in the following section, we use pretrained encoder, and pretrained encoder-decoder for short. We follow the officially recommended finetuning setting with dropout of 0.3, label smoothing of 0.2, and warm-up of 2500 steps. We finetune on the high-resource task for 100K steps and the low-resource tasks for 40K steps, respectively.

We also list the results of vanilla Transformer without pretraining as baseline. The vocabulary is built on the bilingual data, hence is much smaller (e.g., En-De 44K) than mBART25. Specifically, for high-resource tasks we train 6L-6L Transformer-Big with 460K tokens per batch for 30K steps, and for low-resource tasks we train 6L-6L Transformer-Base with 16K tokens per batch for 50K steps.

[1] https://github.com/pytorch/fairseq/tree/main/examples/mbart

Pretraining			W19 En-De		W19 En-De (S)		W16 En-Ro		I17 En-Fr	
Model	*Enc*	*Dec*	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐
no pretrain			39.6	41.0	29.7	30.1	34.5	34.3	37.3	38.0
	×	×	39.4	40.1	26.7	27.1	30.0	29.6	35.3	35.1
mBART	✓	×	40.8	41.1	31.7	33.5	35.0	35.6	38.4	38.4
	✓	✓	40.8	41.4	35.3	35.7	37.1	37.4	39.2	40.2

Table 1: BLEU scores on MT benchmarks. "Enc:×, Dec:×" represents that we use only the pre-trained embeddings for fair comparisons, and we highlight performance improvement over this setting in red color.

2.2 Impact of Jointly Pretrained Decoder

The main difference of Seq2Seq pretraining models (e.g., mBART) from previous pretraining models (e.g., BERT and XLM-R) lies in whether to train the decoder together. In this section, we investigate the impact of the jointly pretrained decoder in terms of BLEU scores, and provide some insights on where the jointly pretrained decoder improves performance.

Translation Performance. Table 1 lists the BLEU scores of pretraining different components of NMT models, where we also include the results of NMT models trained on the datasets from scratch ("no pretrain"). For fair comparisons, we use the same vocabulary size for all variants of pretraining NMT components. We use the pretrained word embedding for the model variant with randomly initialized encoder-decoder ("Enc:×, Dec:×"), which makes it possible to train 12L-12L NMT models on the small-scale datasets. Accordingly, the results of ("Enc:×, Dec:×") is worse than the "no pretrain" model due to the larger vocabulary (e.g., 250K vs. 44K) that makes the model training more difficult.

Pretraining encoder only ("Enc:✓, Dec:×") significantly improves translation performance, which is consistent with the findings in previous studies (Zhu et al., 2019; Weng et al., 2020). We also conduct experiments with the pretrained encoder XLM-R (Conneau et al., 2020), which achieves comparable performance as the mBART encoder (see Appendix A.1). For fair comparisons, we only use the mBART encoder in the following sections. Encouragingly, jointly pretraining decoder can further improve translation performance, although the improvement is not significant on the large-scale WMT19 En-De data. These results seem to provide empirical support for the common cognition – pretraining is less effective on large-scale data.

Src	Sie bezichtigt die Erwachsenen Kinderhandel zu betreiben.
Ref	She accuses the adults of child trafficking.
	Large-Scale Data
no pre.	It accuses (the) adults of children trafficking.
(×, ×)	It accuses (the) adults of children trafficking.
(✓, ×)	She accuses the adults of children trafficking.
(✓, ✓)	She accuses the adults of trafficking in children.
	Small-Scale Data
no pre.	It accuses the adults to trade children.
(×, ×)	It requires adult trafficking on children.
(✓, ×)	It accuses (the) adults of children trafficking.
(✓, ✓)	She accuses the adults of trafficking in children.

Table 2: Translation examples on WMT19 De⇒En test set. The translation errors are highlighted in red and changes of word order are highlighted in blue.

However, we have some interesting findings of the generated outputs, which may draw different conclusions. To eliminate the effect of language and data bias, we use the full set and sampled subset of WMT19 De⇒En data as representative large-scale and small-scale data scenarios.

Table 2 shows some translation examples. Firstly, jointly pretraining decoder can produce good translations that are different in the word order from the ground-truth reference (e.g., "trafficking in children" vs. "child trafficking"), thus are assigned low BLEU scores. This may explain why jointly pretraining decoder only marginally improves performance on large-scale data. Secondly, jointly pretraining decoder can reduce translation errors, especially on small-scale data (e.g., correct the mistaken translation of "*It*" to "*She*"). We empirically validate the above two findings in the following experiments.

Impact on Translation Diversity. We follow Du et al. (2021) to better evaluate the translation quality for different word orders using multiple references. We use the test set released by Ott et al.

Pretrain	Single		Multiple	
	BLEU	△	BLEU	△
Large-Scale Data				
no pretrain	39.5	-	77.1	-
(×, ×)	38.6	-0.9	75.7	-1.4
(✓, ×)	39.5	+0.0	77.8	+0.7
(✓, ✓)	39.9	+0.4	79.1⇑	**+2.0**
Small-Scale Data				
no pretrain	27.0	-	53.1	-
(×, ×)	27.0	+0.0	52.3	-0.8
(✓, ×)	32.3	+5.3	63.4	+10.3
(✓, ✓)	35.3⇑	+8.3	69.1⇑	**+16.0**

Table 3: BLEU scores on En⇒De testset with single and multiple references. "⇑" denotes significantly better (with $p < 0.01$) than No mBART pretraining.

Pretrain		Large			Small		
Enc	*Dec*	*Ut*	*Mt*	*Ot*	*Ut*	*Mt*	*Ot*
×	×	4	9	0	25	45	0
✓	×	3	3	0	5	21	5
✓	✓	2	0	0	3	15	0

Table 4: Human evaluation of mBART pretrained NMT models in terms of under-translation (Ut), mis-translation (Mt), and over-translation (Ot) errors.

(2018), which consists of 10 human translations for 500 sentences taken from the WMT14 En⇒De test set. As shown in Table 3, the pretrained decoder achieves more significant improvement in all cases when measured by multiple references. These results provide empirical support for our claim that jointly pretraining decoder produces more diverse translations with different word orders, which can be better measured by multiple references. These results may renew our cognition of pretraining, that is, *they are also effective on large-scale data when evaluated more accurately*.

Impact on Adequacy. We conduct a human evaluation to provide a more intuitive understanding of how jointly pre-training decoder improves translation quality. Specifically, we ask two annotators to annotate under-translation, mis-translation and over-translation on 100 sentences randomly sampled from WMT19 De⇒En test set. As listed in Table 4, inheriting the pretrained decoder reduces more translation errors on small data than on large data, which is consistent with the results of BLEU score in Table 1. Interestingly, inheriting only the pretrained encoder introduces more over-translation errors on small data, which can be solved by combining the pretrained decoder. One possible reason is that inheriting only the pretrained encoder excessively enlarges the impact of source context.[2] This problem does not happen on large data, since the large amount of in-domain data can balance the relation between encoder and decoder to accomplish the translation task well.

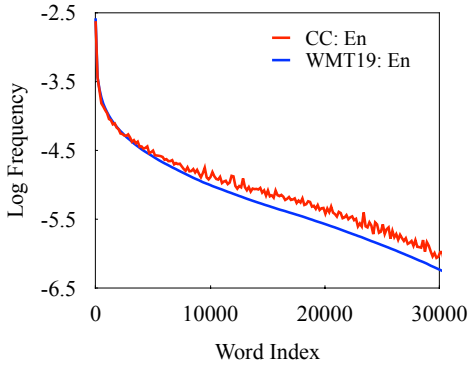


Figure 1: Word distributions of English corpora from general domain (i.e., CC data) and in-domain (i.e., WMT19 En-De news domain), respectively. The word frequency is normalized and reported in log-scale.

2.3 Pretraining-and-Finetuning Discrepancy

Although Seq2Seq pretraining consistently improves translation performance across data scales, we find several side effects of Seq2Seq pretraining due to the discrepancy between pretraining and finetuning. In this section, we present two important discrepancies: *domain discrepancy* and *objective discrepancy*. Unless otherwise stated, we report results on WMT19 En-De test set using small data.

2.3.1 Domain Discrepancy

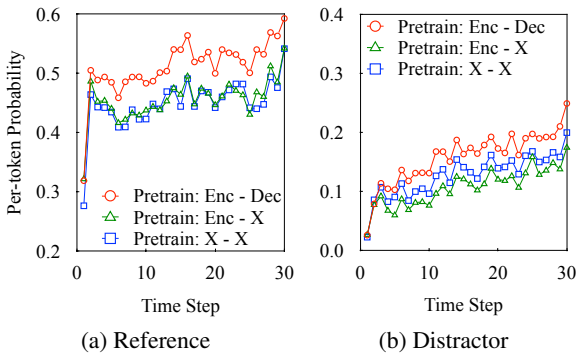
Seq2Seq pretraining model is generally trained on **general domain** data while the downstream translation models are trained on **specific domains** (e.g., news). Such a domain discrepancy requires more efforts for the finetuned models to adapt the knowledge in pretrained models to the target in-domain. We empirically show the domain discrepancy in terms of lexical distribution and domain classifier.

Lexical Distribution in Training Data. Inspired by lexicon distribution analysis (Ding et al., 2021), we first plot the word distributions of English corpora from general domain (i.e., CC data) and in-domain (i.e., WMT19 En-De news domain)

[2] Tu et al. (2017a) showed that more impact of source context leads to over-translation errors.

Set	En⇒De	De⇒En
Source	77.5	73.7
Target	71.0	75.4

Table 5: Ratio of sentences in WMT19 En-De test sets that are classified as WMT news domain.



(a) Reference (b) Distractor

Figure 2: Per-token generation probability on the test set of WMT19 En⇒De (S). Higher probabilities are expected for the groundtruth references (a), and lower probabilities are expected for the distractors (b).

to study their difference at the lexicon level. The words are ranked according to their frequencies in the WMT19 En-De training data. As shown in Figure 1, we observe a clear difference between WMT news data and CC data in the long tail region, which is supposed to carry more domain-specific information. Accordingly, there will be a domain shift from pretraining to finetuning.

Domain Classifier for Test Data. We further demonstrate that the test data also follows a consistent domain as the training data. To distinguish general domain and in-domain, we build a domain classifier based on the WMT19 En-De training data and the CC data. We select a subset from the WMT training data with some trusted data (Wang et al., 2018; Jiao et al., 2020), which includes 22404 sample from WMT newstest2010-2017 (see Appendix A.2 for details). Specifically, we select 1.0M samples from the WMT training data and the CC data, respectively, to train the domain classifier. The newstest2018 is combined with an equally sized subset of CC data for validation. We adopt the domain classifier to classify each sample in the test sets of WMT19 En-De. As shown in Table 5, most of the sentences (e.g., 70% - 80%) are recognized as WMT news domain, which demonstrates the domain consistency between the training data and test data in the downstream tasks.

2.3.2 Objective Discrepancy

The learning objective discrepancy between Seq2Seq pretraining and NMT training is that NMT learns to translate a sentence from one language to another, while Seq2Seq pretraining learns to reconstruct the input sentence (Liu et al., 2021). In this section, we study the side effects of the objective discrepancy by evaluating the predicting behaviors that are highly affected by the learning objective.

Model Uncertainty. We follow Ott et al. (2018) to analyze the model's uncertainty by computing the average probability at each time step across a set of sentence pairs. To evaluate the capability of LM modeling on the target language, we also follow Wang and Sennrich (2020) to consider a set of "distractor" translations, which are random sentences from the CC data that match the corresponding reference translation in length. Figure 2 plots model uncertainties for both references (Y) and distractors ($\hat{Y}$). We find that jointly pretraining decoder significantly improves model certainty after the first few time steps (Figure 2a). As for the distractors, pretraining encoder only results in certainties even lower than training from scratch (Figure 2b), which suggests that the corresponding NMT model is more dominated by the source context. It reconfirms the finding in our human evaluation (Table 4). In contrast, jointly pretraining decoder leads to a significant improvement of certainties, suggesting that the pretrained decoder tends to induce the **over-estimation issue** of NMT models. A possible reason is that Seq2Seq pretraining does not establish the connection between languages, such that its strong capability of LM modeling still recognizes the distractor as a valid target sentence even though it is mismatched with the source sentence in semantics.

Hallucination under Perturbation. One translation problem associated with over-estimation is hallucination (Wang and Sennrich, 2020), where NMT models generate fluent translation but is unrelated to the input. In this section, we follow Lee et al. (2018) to evaluate the model's tendency of generating hallucination under noisy input, to which NMT models are highly sensitive (Belinkov and Bisk, 2018). Specifically, we employ two different perturbation strategies: (1) First position insertion (FPI) that inserts a single additional input token into the source sequence, which can

Pretrain		FPI (%)		RSM (%)	
Enc	*Dec*	$\triangle_{\text{BLEU}}$	HUP	$\triangle_{\text{BLEU}}$	HUP
×	×	-1.3	0.5	-8.8	2.4
✓	×	-0.3	0.5	-8.3	0.5
✓	✓	**-3.2**	**7.8**	**-17.8**	**15.5**

Table 6: BLEU change of model performance under perturbed inputs over the standard inputs, and hallucinations under perturbation (HUP) score.

Pretrain		BLEU		Copy (%)	
Enc	*Dec*	5	100	5	100
×	×	26.7	26.6	12.9	13.9
✓	×	31.7	31.6	12.7	12.9
✓	✓	35.3	33.5	**13.2**	**19.4**

Table 7: Beam search degradation and ratio of copying tokens in translation outputs.

completely divorce the translation from the input sentence (Lee et al., 2018). (2) Random span masking (RSM) that simulates the noisy input in the Seq2Seq pretraining of mBART (Liu et al., 2020). We follow Lee et al. (2018) to count a translation as hallucination under perturbation (HUP) when: (1) BLEU between reference sentence and translation of unperturbed sentence is bigger than 5 and (2) BLEU between the translation of perturbed sentence and the translation of unperturbed sentence is lower than 3. We calculate the percentage of hallucination as the HUP score. Table 6 lists the BLEU change and HUP score for the perturbed inputs. As expected, jointly pretraining decoder is less robust to perturbed inputs (more decline of BLEU scores), and produces more hallucinations than the other two model variants.

Beam Search Problem. One commonly-cited weakness of NMT model is the beam search problem, where the model performance declines as beam size increases (Tu et al., 2017b). Previous studies demonstrate that over-estimation is an important reason for the beam search problem (Ott et al., 2018; Cohen and Beck, 2019). We revisit this problem for NMT models with Seq2Seq pretraining, as shown in Table 7. We also list the ratio of copying tokens in translation outputs (i.e., directly copy source words to target side without translation) for different beam sizes, which has been shown as a side effect of Seq2Seq pretraining models (Liu et al., 2021). As seen, jointly pretraining decoder suffers from more serious beam search degradation problem, which reconfirms the connection between beam search problem and over-estimation. In addition, larger beam size introduces more copying tokens than the other model variants (i.e., 19.4 vs. 13.9, 12.9), which also links copying behaviors associated with Seq2Seq pretraining to the beam search problem.

3 Improving Seq2Seq Pretraining

3.1 Approach

To bridge the above gaps between Seq2Seq pretraining and finetuning, we introduce *in-domain pretraining* and *input adaptation* to improve the translation quality and model robustness.

In-Domain Pretraining. To bridge the domain gap, we propose to continue the training of mBART (Liu et al., 2020) on the in-domain monolingual data. Specifically, we first remove spans of text and replace them with a mask token. We mask 35% of the words in each sentence by random sampling a span length according to a Poisson distribution ($\lambda = 3.5$). We also permute the order of sentences within each instance. The training objective is to reconstruct the original sentence at the target side. We expect the in-domain pretraining to reduce the domain shift by re-pretraining on the in-domain data, which is more similar in data distribution with the downstream translation tasks.

Input Adaptation in Finetuning. To bridge the objective gap and improve the robustness of models, we propose to add noises (e.g., mask, delete, permute) to the source sentences during finetuning, and keep target sentences as original ones. Empirically, we add noises to 10% of the words in each source sentence, and combine the noisy data with the clean data by the ratio of 1:9, which are used to finetune the pretraining model. We expect the introduction of perturbed inputs in finetuning can help to better transfer the knowledge from pretrained model to the finetuned model, thus alleviate over-estimation and improve the model robustness.

3.2 Experimental Results

Main Results on Translation Performance and Robustness. The main results are listed in Table 8. We report the results of input adaptation, in-domain pretraining, and the combination of these two approaches, respectively. For input adaptation, it achieves comparable translation quality as the

Approach	W19 En⇒De		W19 En⇒De (S)		W16 En⇒Ro		I17 En⇒Fr	
	BLEU	HuP	BLEU	HuP	BLEU	HuP	BLEU	HuP
Baseline	39.4	2.6	26.7	2.4	30.0	1.1	35.3	1.6
General	40.8	3.3	35.3	15.5	37.1	6.5	39.2	7.8
+ Input Adapt	40.8	2.7	35.6	5.7	37.2	2.4	39.4	1.5
+ In-Domain	**42.2**	9.2	**36.4**	10.4	**38.0**	8.2	39.9	5.5
+ Input Adapt	41.3	4.1	36.1	3.6	37.8	2.9	**40.1**	3.0
Approach	**W19 De⇒En**		**W19 De⇒En (S)**		**W16 Ro⇒En**		**I17 Fr⇒En**	
	BLEU	HuP	BLEU	HuP	BLEU	HuP	BLEU	HuP
Baseline	40.1	2.8	27.1	1.3	29.6	1.3	35.1	1.7
General	**41.4**	7.7	35.7	4.9	37.4	6.0	40.2	4.7
+ Input Adapt	41.2	2.6	35.9	2.8	37.1	3.5	40.7	2.5
+ In-Domain	41.3	8.2	**36.9**	7.4	**38.1**	7.7	**41.1**	4.2
+ Input Adapt	**41.4**	3.1	36.8	2.9	37.9	3.9	41.0	1.7

Table 8: BLEU and HuP scores of our approaches for downstream translation tasks.

Approach	W19 En-De		W19 En-De (S)	
	BLEU	△	BLEU	△
Baseline	75.7	-	52.3	-
General	79.1	+3.4	69.1	+16.8
+ Input Adapt	79.2	+3.5	71.7	+19.4
+ In-Domain	**80.1**	**+4.4**	73.7	+21.4
+ Input Adapt	79.8	+4.1	**75.6**	**+23.3**

Table 9: BLEU scores with multiple references.

Approach	W19 En-De (S)		W16 En-Ro	
	⇒	⇐	⇒	⇐
Baseline	26.7	27.1	30.0	29.6
In-Domain	35.2	35.7	36.1	36.3

Table 10: BLEU scores of in-domain pretraining only.

general domain pretrained model and significantly reduces the ratio of HuP, indicating the enhancement of model robustness. In-domain pretraining generally improves the translation quality but does not make the model more robust. On the contrary, it may increase the ratio of HuP in some cases (e.g., En⇒Ro 5.6 vs. 8.2). Conducting input adaptation right after in-domain pretraining will combine the advantages of these two approaches, and improve both the translation quality and model robustness. The effectiveness of our approaches, especially input adaptation, is more significant when evaluated with multiple references, as shown in Table 9.

In-Domain Only. Given the promising performance of in-domain pretraining, we investigate whether pretraining on in-domain data only can also obtain significant improvement. We report the results in Table 10. We can observe that pretraining solely on the in-domain data can improve the translation performance noticeably over the models without pretraining. However, the improvement is less competitive than the pretrained mBART25 (e.g., En⇒Ro: 36.1 v.s. 37.1 in Table 8), which may result from the much larger scale of multilingual data used in general pretraining.

3.3 Analysis

We provide some insights into how our approach improves model performance over general pretraining. We report results on WMT19 En⇒De test set using small-scale data.

Narrowing Domain Gap. Since the difference of lexical distribution between general domain and in-domain data mainly lies in the long tail region (see Figure 1), we study how our approach performs on low-frequency words. Specifically, we calculate the word accuracy of the translation outputs for WMT19 En-De (S) by the `compare-mt`[3] tool. We follow Wang et al. (2021) to divide words into three categories based on their frequency in the bilingual data, including High: the most 3,000 frequent words; Medium: the most 3,001-12,000 frequent words; Low: the other words. Table 11 lists

[3] https://github.com/neulab/compare-mt

Approach	Frequency		
	Low	*Med*	*High*
Baseline	36.8	45.3	57.5
General	44.5	54.3	64.2
+ In-Domain	46.2	54.3	64.9

Table 11: F-measures of word prediction for different frequencies that are calculated in the bilingual data.

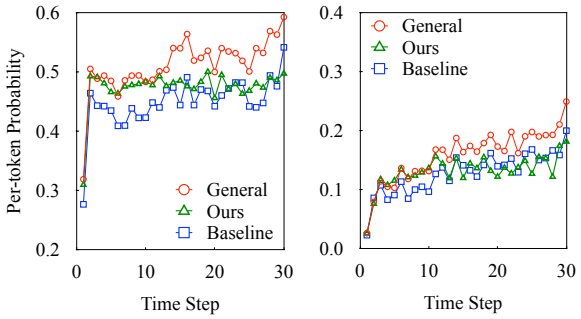
the results. The improvements on low-frequency words are the major reason for the performance gains of in-domain pretraining, where it outperforms general pretraining on the translation accuracy of low/medium/high- frequency words by 1.7, 0.0, and 0.7 BLEU scores, respectively. These findings confirm our hypothesis that in-domain pretraining can narrow the domain gap with in-domain data, which is more similar in the lexical distribution as the test sets.

Alleviating Over-Estimation. Figure 3 shows the impact of our approach on model uncertainty. Clearly, our approach successfully alleviates the over-estimation issue of general pretraining in both the groundtruth and distractor scenarios.

Mitigating Beam Search Degradation. We recap the beam search degradation problem with the application of our approaches in Table 12. The input adaptation approach can noticeably reduce the performance decline when using a larger beam size (e.g., from -1.8 to -0.9), partially due to a reduction of copying tokens in generated translations (e.g., from 19.4% to 15.3%). Although in-domain pretraining does not alleviate the beam search degradation problem, it can be combined with input adaptation to build a well-performing NMT system.

4 Related Work

Pretraining for NMT. Previous pretraining approaches for NMT generally focus on how to effectively integrate pretrained BERT (Devlin et al., 2019) or GPT (Radford et al., 2019) to NMT models. For example, Yang et al. (2020) propose a concerted training framework, and Weng et al. (2020) propose a dynamic fusion mechanism and a distillation paradigm to acquire knowledge from BERT and GPT. In this work, we aim to provide a better understanding of how Seq2Seq pretraining model works for NMT, and propose a simple and effective approach to improve model performance based on these observations.



(a) Reference (b) Distractor

Figure 3: Per-token generation probability on WMT19 En⇒De (S) test set when adopting our approaches.

Approach	BLEU		Copy (%)	
	5	100	5	100
General	35.3	33.5	**13.2**	19.4
+ Input Adapt	35.6	34.7	12.5↓	15.3↓
+ In-Domain	36.4	33.9	12.9	**19.8**
+ Input Adapt	36.1	35.0	12.6↓	15.6↓

Table 12: Beam search degradation and "copy" translations when adopting our approaches.

Intermediate Pretraining. Our in-domain pretraining approach is related to recent successes on intermediate pretraining and intermediate task selection in NLU tasks. For example, Ye et al. (2021) investigate the influence of masking policies in intermediate pretraining. Poth et al. (2021) explore to select tasks for intermediate pretraining. Closely related to our work, Gururangan et al. (2020) propose to continue the pretraining of RoBERTa (Liu et al., 2019) on task-specific data. Inspired by these findings, we employ in-domain pretraining to narrow the domain gap between general Seq2Seq pretraining and NMT training. We also show the necessity of target-side monolingual data on in-domain pretraining (see Appendix A.3), which has not been studied in previous works of in-domain pretraining.

5 Conclusion

In this paper we provide a better understanding of Seq2Seq pretraining for NMT by showing both the benefits and side effects. We propose simple and effective approaches to remedy the side effects by bridging the gaps between Seq2Seq pretraining and NMT finetuning, which further improves translation performance and model robustness. Future directions include validating our findings on more Seq2Seq pretraining models and language pairs.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *ICLR*.

Eldan Cohen and Christopher Beck. 2019. Empirical analysis of beam search performance degradation in neural sequence models. In *ICML*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, E. Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Understanding and improving lexical choice in non-autoregressive translation. In *Proc. of ICLR*.

Cunxiao Du, Zhaopeng Tu, and Jing Jiang. 2021. Order-agnostic cross entropy for non-autoregressive machine translation. In *ICML*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proc. of ACL*.

Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. 2020. Data rejuvenation: Exploiting inactive training examples for neural machine translation. In *EMNLP*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. of EMNLP*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. In *NeurIPS-IRASL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021. On the copying behaviors of pre-training for neural machine translation. In *Proc. of ACL*.

Yinhan Liu, Jiatao Gu, Naman Goyal, X. Li, Sergey Edunov, Marjan Ghazvininejad, M. Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *TACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proc. of ICML*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? efficient intermediate task selection. *arXiv*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proc. of ICML*.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *LREC*.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017a. Context gates for neural machine translation. *TACL*.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017b. Neural machine translation with reconstruction. In *AAAI*.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *ACL*.

Shuo Wang, Zhaopeng Tu, Zhixing Tan, Shuming Shi, Maosong Sun, and Yang Liu. 2021. On the language coverage bias for neural machine translation. In *Proc of ACL*.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *WMT*.

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

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards making the most of bert in neural machine translation. In *AAAI*.

Qinyuan Ye, Belinda Z Li, Sinong Wang, Benjamin Bolte, Hao Ma, Wen-tau Yih, Xiang Ren, and Madian Khabsa. 2021. On the influence of masking policies in intermediate pre-training. *arXiv*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2019. Incorporating bert into neural machine translation. In *ICLR*.

A Appendix

A.1 Comparison of XLM-R and mBART

Pre-Train		En-De		En-De (S)		En-Ro	
Enc	*Dec*	⇒	⇐	⇒	⇐	⇒	⇐
mBART model							
✓	×	40.8	41.0	31.7	33.5	35.0	35.6
✓	✓	40.8	**41.4**	**35.3**	**35.7**	**37.1**	**37.4**
XLM-R model							
✓	×	**41.6**	**41.4**	27.7	30.1	32.8	30.0
✓	✓	41.1	40.4	31.4	32.3	34.4	33.4

Table 13: Comparison between 12L-12L mBART and XLM-R in terms of BLEU scores on MT benchmarks.

Throughout our paper, we mainly rely on mBART to investigate the pretrained encoder only setting. Here, we report our results on the same benchmark datasets with the popular pretrained encoder, XLM-R (Conneau et al., 2020). We also try to initialize the decoder of NMT models with XLM-R. The results are listed in Table 13. We find that XLM-R achieves comparable translation performance as mBART on the large-scale WMT19 En-De data but under-performs mBART on small-scale data significantly. The strong results of mBART ensure the reliability of our findings.

A.2 Domain Classifier

To distinguish general domain and in-domain, we build a domain classifier based on the WMT19 En-De training data and the CC data. We select a subset from the full training data of WMT with some trusted data (Wang et al., 2018; Jiao et al., 2020) , i.e., WMT newstest2010-2017 consisting of 22404 samples, to reduce the impact of possible noises in the training data. Specifically, we first train a language model on the full WMT training data as the noisy model and then finetune it on the trusted data to obtain the denoised model. For a sentence $\mathbf{x}$, the difference of confidence between the two models, i.e., $\log P_{noisy}(\mathbf{x}) - \log P_{denoised}(\mathbf{x})$, represents the noise score. We select 1.0M samples with the lowest noise score from the WMT training data and randomly select 1.0M samples from the CC data to train the domain classifier. The newstest2018 combined with an equally sized subset of CC data is used as the validation data to select the best classifier.

A.3 Involved Languages

Lang	BLEU	Frequency		
		Low	*Med*	*High*
None	40.8	50.9	58.0	67.0
En	41.3	51.1	58.1	67.5
En,De	42.2	52.2	59.2	67.7

Table 14: Effect of languages involved in in-domain pretraining, evaluated on WMT19 En-De dataset.

We investigate whether the languages involved in the in-domain pretraining process affect the final performance of our approach. In Table 14, we present the results of in-domain pretraining with only one language involved, i.e., English. While the translation quality can also be improved slightly, the improvements of accuracy on medium- and low-frequency words are very limited. It indicates that in-domain pretraining on the target language (i.e., German here) is critical for medium- and low-frequency words.