Bridging the Data Gap between Training and Inference for Unsupervised Neural Machine Translation

Zhiwei He*  
Shanghai Jiao Tong University  
zwhe.cs@sjtu.edu.cn  

Xing Wang  
Tencent AI Lab  
brightxwang@tencent.com  

Rui Wang†  
Shanghai Jiao Tong University  
wangrui12@sjtu.edu.cn  

Shuming Shi  
Tencent AI Lab  
shumingshi@tencent.com  

Zhaopeng Tu  
Tencent AI Lab  
zptu@tencent.com  

Abstract

Back-translation is a critical component of Unsupervised Neural Machine Translation (UNMT), which generates pseudo parallel data from target monolingual data. A UNMT model is trained on the pseudo parallel data with translated source, and translates natural source sentences in inference. The source discrepancy between training and inference hinders the translation performance of UNMT models. By carefully designing experiments, we identify two representative characteristics of the data gap in source: (1) style gap (i.e., translated vs. natural text style) that leads to poor generalization capability; (2) content gap that induces the model to produce hallucination content biased towards the target language.

To narrow the data gap, we propose an online self-training approach, which simultaneously uses the pseudo parallel data \{\text{natural source, translated target}\} to mimic the inference scenario. Experimental results on several widely-used language pairs show that our approach outperforms two strong baselines (XLM and MASS) by remedying the style and content gaps.

1 Introduction

In recent years, there has been a growing interest in unsupervised neural machine translation (UNMT), which requires only monolingual corpora to accomplish the translation task (Lample et al., 2018a,b; Artetxe et al., 2018b; Yang et al., 2018; Ren et al., 2019). The key idea of UNMT is to use back-translation (BT) (Sennrich et al., 2016) to construct the pseudo parallel data for translation modeling. Typically, UNMT back-translates the natural target sentence into the synthetic source sentence (translated source) to form the training data. A BT loss is calculated on the pseudo parallel data \{\text{translated source, natural target}\} to update the parameters of UNMT models.

In Supervised Neural Machine Translation (SNMT), Edunov et al. (2020) found that BT suffers from the translationese problem (Zhang and Toral, 2019; Graham et al., 2020) in which BT improves BLEU score on the target-original test set with limited gains on the source-original test set. Unlike authentic parallel data available in the SNMT training data, the UNMT training data entirely comes from pseudo parallel data generated by the back-translation. Therefore in this work, we first revisit the problem in the UNMT setting and start our research from an observation (§2): with comparable translation performance on the full test set, the BT based UNMT models achieve better translation performance than the SNMT model on the target-original (i.e. translationese) test set, while achieves worse performance on the source-
In summary, the contributions of this work are detailed as follows:

- Our empirical study demonstrates that the back-translation based UNMT framework suffers from the translationese problem, causing the inaccurate evaluation of UNMT models on standard benchmarks.
- We empirically analyze the data gap between training and inference for UNMT and identify two critical factors: style gap and content gap.
- We propose a simple and effective approach for incorporating the self-training method into the UNMT framework to remedy the data gap between the training and inference.

2 Translationese Problem in UNMT

2.1 Background: UNMT

Notations. Let $X$ and $Y$ denote the language pair, and let $X = \{x_i\}_{i=1}^M$ and $Y = \{y_j\}_{j=1}^N$ represent the collection of monolingual sentences of the corresponding language, where $M, N$ are the size of the corresponding set. Generally, UNMT method that based on BT adopts dual structure to train a bidirectional translation model (Artetxe et al., 2018b, 2019; Lample et al., 2018a,b). For the sake of simplicity, we only consider translation direction $X \rightarrow Y$ unless otherwise stated.

Online BT. Current mainstream of UNMT methods turn the unsupervised task into the synthetic supervised task through BT, which is the most critical component in UNMT training. Given the translation task $X \rightarrow Y$ where target corpus $Y$ is available, for each batch, the target sentence $y \in Y$ is used to generate its synthetic source sentence by the backward model $MT_{Y \rightarrow X}$:

$$x^* = \arg \max_x P_{Y \rightarrow X}(x \mid y; \tilde{\theta}),$$  (1)

where $\tilde{\theta}$ is a fixed copy of the current parameters $\theta$ indicating that the gradient is not propagated through $\tilde{\theta}$. In this way, the synthetic parallel sentence pair $\{x^*, y\}$ is obtained and used to train the forward model $MT_{X \rightarrow Y}$ in a supervised manner by minimizing:

$$\mathcal{L}_B = \mathbb{E}_{y \sim Y}[- \log P_{X \rightarrow Y}(y \mid x^*; \theta)].$$  (2)

It is worth noting that the synthetic sentence pair generated by the BT is the only supervision signal of UNMT training.
**Objective function.** In addition to BT, denoising auto-encoding (DAE) is an additional loss term of UNMT training, which is denoted by $L_D$ and is not the main topic discussed in this work.

In all, the final objective function of UNMT is:

$$L = L_B + \lambda_D L_D,$$

where $\lambda_D$ is the hyper-parameter weighting DAE loss term. Generally, $\lambda_D$ starts from one and decreases as the training procedure continues\(^3\).

### 2.2 Translationese Problem

To verify whether the UNMT model suffers from the input gap between training and inference and thus is biased towards translated input while against natural input, we conduct comparative experiments between SNMT and UNMT models.

**Setup** We evaluate the UNMT and SNMT models on WMT14 En-Fr, WMT16 En-De and WMT16 En-Ro test sets, following Lample and Conneau (2019) and Song et al. (2019). We first train the UNMT models on the above language pairs with model parameters initialized by XLM and MASS models. Then, we train the corresponding SNMT models whose performance on the full test sets is controlled to be approximated to UNMT by undersampling training data. Finally, we evaluate the UNMT and SNMT models on the target-original and source-original test sets, whose inputs are translated and natural respectively. Unless otherwise stated, we follow previous work (Lample and Conneau, 2019; Song et al., 2019) to use case-sensitive BLEU score (Papineni et al., 2002) with the `multi-bleu.perl`\(^4\) script as the evaluation metric. Please refer to Appendix B for the results of SacreBLEU, and refer to Appendix A for the training details of SNMT and UNMT models.

**Results** We present the translation performance in terms of the BLEU score in Table 2 and our observations are:

- UNMT models perform close to the SNMT models on the full test sets with 0.3 BLEU difference at most on average (33.5/33.9 vs. 33.6).

- UNMT models outperform the SNMT models on target-original test sets (translated input) with average BLEU score improvements of 3.6 and 4.4 BLEU points (36.8/37.6 vs. 33.2).

- UNMT models underperform the SNMT models on source-original test sets (natural input) with an average performance degradation of 4.4 and 4.2 BLEU points (28.7/28.9 vs. 33.1).

The above observations are invariant concerning the pre-trained model and translation direction. In particular, the unsatisfactory performance of UNMT under natural input indicates that UNMT is overestimated on the previous benchmark. We attribute the phenomenon to the data gap between training and inference for UNMT: there is a mismatch between natural inputs of source-original test data and the back-translated inputs that UNMT employed for training. This work focuses on the experiments on the source-original test sets (i.e., the input of an NMT translation system is generally natural), which is closer to the practical scenario.\(^5\)

### 3 Data Gap between Training and Inference

In this section, we identify two representative data gaps between training and inference data for UNMT.

| Model | En-Fr | En-De | En-Ro | Avg. |
|-------|-------|-------|-------|------|
| **Full Test Set** | | | | |
| SNMT | 38.4 | 33.6 | 29.5 | 33.9 | 33.7 | 32.5 | 33.6 |
| XLM | 37.4 | 34.5 | 27.2 | 34.3 | 34.6 | 32.7 | 33.3 |
| MASS | 37.8 | 34.9 | 27.1 | 35.2 | 35.1 | 33.4 | 33.9 |
| **Target-Original Test Set / Translated Input** | | | | |
| SNMT | 37.4 | 32.4 | 25.6 | 37.1 | 37.2 | 28.2 | 33.2 |
| XLM | 39.1 | 36.5 | 26.6 | 42.2 | 42.1 | 34.4 | 36.8 |
| MASS | 39.2 | 37.6 | 27.0 | 42.9 | 43.1 | 35.6 | 37.6 |
| **Source-Original Test Set / Natural Input** | | | | |
| SNMT | 38.2 | 34.1 | 32.3 | 28.8 | 29.4 | 35.9 | 33.1 |
| XLM | 34.7 | 30.4 | 26.6 | 22.5 | 27.4 | 30.6 | 28.7 |
| MASS | 35.2 | 30.2 | 26.1 | 23.6 | 27.4 | 30.8 | 28.9 |

Table 2: Translation performance of SNMT and UNMT models on full / target-original / source-original test sets. SNMT denotes the supervised translation models trained on undersampled parallel data and their performance on full test data are controlled to be approximate to the UNMT counterparts.

\(^3\)Verified from open-source XLM Github implementation. 
\(^4\)https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

\(^5\)From WMT19, the WMT community proposes to use the source-original test with natural input sets to evaluate the translation performance.
3.1 Style Gap

To perform the quantitative analysis of the style gap, we adopt KenLM\(^6\) to train a 4-gram language model on the UNMT translated source sentences\(^7\) and use the language model to calculate the perplexity (PPL) of natural and translated input sentences in the test sets. The experimental results are shown in Table 3. The lower perplexity value (219 < 242) indicates that compared with the natural inputs, the UNMT translated training inputs have a more similar style with translated inputs in the test sets.

In order to further reveal the influence of the style gap on UNMT, we manually eliminated it and re-evaluated the models on the natural input portion of WMT16 De⇒En. Concretely, We first take the third-party Google Translator to translate the target English sentences of the test sets into the source German language to eliminate the style gap. And then we conduct translation experiments on the natural input portion and its Google translated portion to evaluate the impact of the style gap on the translation performance. We list the experimental results in Table 4. We can find that by converting from the natural inputs (natural De) to the translated inputs (translated De\(^*\)), the UNMT model achieves more improvement than the SNMT model (-2.8 > -6.3), demonstrating that the style gap inhibits the UNMT translation output quality.

3.2 Content Gap

In this section, we show the existence of the content gap by (1) showing the most high-frequency name entities, (2) calculating content similarity using term frequency-inverse document frequency (TF-IDF) for the training and inference data.

We use spaCy\(^8\) to recognize German named entities for the UNMT translated source sentences, natural inputs and translated inputs in test sets, and show the ten most frequent name entities in Table 5. From the table, we can observe that the UNMT translated source sentences have few named entities biased towards source language German (words in red color), while having more named entities biased towards target language English, e.g., USA, Obama. It indicates that the content of the UNMT translated source sentences is biased towards the target language English.

Meanwhile, the natural input portion of the inference data has more named entities biased towards source language German (words in red color), demonstrating that the content gap exists between the natural input portion of the inference data and the UNMT translated training data.

Next, we remove the stop words and use the term frequency-inverse document frequency (TF-IDF) approach to calculate the content similarity between the training and inference data. Similarity scores are presented in Table 6. We can observe that the UNMT translated source data has a more significant similarity score with translated inputs which are generated from the target English sentences. This result indicates that the content of UNMT translated source data is more biased towards the target language, which is consistent with the findings in Table 5.

As it is difficult to measure the name entities

---

\(^6\)https://github.com/kpu/kenlm

\(^7\)To alleviate the content bias problem, we generate the training data 50% from En⇒De translation and 50% from round trip translation De⇒En⇒De.

\(^8\)https://github.com/explosion/spaCy
Table 5: Ten most frequent entities in the source sentences (i.e., German) of back-translated training data (“BT Train Data”). For reference, we also list the most frequent entities in the natural and translated inference inputs. The BT training data has more entities biased towards the target language English (blue words) rather than the expected source language German (red words).

| Inference Input       | Natural | Translated |
|-----------------------|---------|------------|
| Natural               | 0.95    | 0.85       |
| Translated            | 0.84    | 0.93       |

Table 6: Content similarity between different kinds of training and inference data.

translation accuracy in terms of BLEU evaluation metric, we provide a translation example in Table 7 to show the effect of the content gap in the UNMT translations (more examples in Appendix C). We observe that the UNMT model outputs the hallucinated translation “U.S.”, which is biased towards target language English. We present a quantitative analysis to show the impact of the content gap on UNMT translation performance in Section 6.2.

5 Experiments

5.1 Setup

Data We follow the common practices to conduct experiments on several UNMT benchmarks: WMT14 En-Fr, WMT16 En-De, WMT16 En-Ro. The details of monolingual training data are delineated in Appendix A.2. We adopt En-Fr newsdev2014, En-De newsdev2016, En-Ro newsdev2016 as the validation (development) sets, and En-Fr newstest2014, En-De newstest2016, En-Ro newstest2016 as the test sets. In addition to the full test set, we split the test set into two parts: target-original and source-original, and evaluate the model’s performance on the three kinds of test sets. We use the released XLM BPE codes and vocabulary for all language pairs.

Model We evaluate the UNMT model fine-tuned on XLM\(^9\) and MASS\(^10\) pre-trained model (Lample and Conneau, 2019; Song et al., 2019). For XLM models, we adopt the pre-trained models released by Lample and Conneau (2019) for all language pairs. For MASS models, we adopt the pre-trained

\(^9\)https://github.com/facebookresearch/XLM
\(^10\)https://github.com/microsoft/MASS
Table 8: Translation performance on WMT14 En-Fr, WMT16 En-De, WMT16 En-Ro and their corresponding source-original (natural input) and target-original (translated input) subset. 

### 5.2 Main Result

Table 8 shows the translation performance of XLM and MASS baselines and our proposed models. We have the following observations:

- Our re-implemented baseline models achieve comparable or even better performance as reported in previous works. The reproduced XLM+UNMT model has an average improvement of 1.4 BLEU points compared to the original report in Lample and Conneau (2019) and MASS+UNMT model is only 0.1 BLEU lower on average than Song et al. (2019).

- Our approach with online self-training significantly improves overall translation performance (+0.8 BLEU on average). This demonstrates the universality of the proposed approach on both large-scale (En-Fr, En-De) and data imbalanced corpus (En-Ro).

- In the translated input scenario, our approach achieves comparable performance to baselines. It demonstrates that although the sample of self-training is source-original style, our approach does not sacrifice the performance on the target-original side.

- In the natural input scenario, we find that our proposed approach achieves more significant improvements, with +1.1 and +1.3 average BLEU on both baselines. The reason is that the source-original style sample introduced by self-training alleviates model bias between natural and translated input.

### 5.3 Comparison with Offline Self-training and CBD

We compare online self-training with the following two related methods, which also incorporate natural inputs in training:

- **Offline Self-training** model distilled from the forward and backward translated data generated by the trained UNMT model.

- **CBD** (Nguyen et al., 2021) model distilled from the data generated by two trained UNMT models through cross-translation, which embraces data diversity.
| Model | Approach | WMT19 | WMT20 | Avg. | ∆ | Training Cost |
|-------|----------|-------|-------|------|---|----------------|
|       |          | ⇒    | ⇐    | ⇒    | ⇐ |                |
| XLM   | UNMT     | 26.6 | 24.4 | 22.9 | 26.6 | 25.1 | –   | 1.0 |
|       | +Offline ST | 26.9 | 24.2 | 23.2 | 25.9 | 25.1 | +0.0 | ×1.8 |
|       | +CBD     | 28.3 | 25.6 | 24.2 | 26.9 | 26.3 | +1.2 | ×7.3 |
|       | +Online ST | 28.3 | 26.0 | 24.3 | 27.6 | 26.6 | +1.5 | ×1.2 |
| MASS  | UNMT     | 26.7 | 24.6 | 23.1 | 26.9 | 25.4 | +0.1 | ×1.8 |
|       | +Offline ST | 27.2 | 24.6 | 23.1 | 26.9 | 25.4 | +0.1 | ×1.8 |
|       | +CBD     | 28.3 | 25.6 | 24.0 | 27.0 | 26.2 | +0.9 | ×7.3 |
|       | +Online ST | 28.5 | 26.1 | 23.8 | 27.8 | 26.6 | +1.3 | ×1.1 |

Table 9: Comparison with offline self-training and CBD\(^1\). “↑”/“↑↑”: significant over the corresponding baseline model (p < 0.05/0.01), tested by bootstrap resampling (Koehn, 2004). The training cost is estimated by the time required for training one epoch where the cost of data generation is also considered.

Dataset

Previous studies have recommended restricting test sets to natural input sentences, a methodology adopted by the 2019-2020 edition of the WMT news translation shared task (Edunov et al., 2020). In order to further verify the effectiveness of the proposed approach, we also conduct the evaluation on WMT19 and WMT20 En-De test sets. Both test sets contain only natural input samples.

Results

Experimental results are presented in Table 9. We also show the training costs of these methods. We find that

- Unexpectedly, the offline self-training has no significant improvement over baseline UNMT. Sun et al. (2021) have demonstrated the effectiveness of offline self-training in UNMT under low-resource and data imbalanced scenarios. However, in our data-sufficient scenarios, offline self-training may suffer from the data diversity problem while online self-training can alleviate the problem through the dynamic model parameters during the training process. We leave the complete analysis to future work.

- CBD achieves a significant improvement compared to baseline UNMT, but the training cost is about six times that of online self-training.

- The proposed online self-training achieves the best translation performance in terms of BLEU score, which further demonstrates the superiority of the proposed method under natural input.

\(^1\)Our re-implemented CBD model can not achieve comparable performance with Nguyen et al. (2021), with 28.4 and 35.2 BLEU scores on WMT16 En-De and De-En test sets.

6 Analysis

6.1 Translationese Output

Since the self-training samples are translated sentences on the target side, there is concern that the improvement achieved by self-training only comes from making the model outputs better match the translated references, rather than enhancing the model’s ability on natural inputs. To dispel the concern, we conducted the following experiments: (1) evaluate the fluency of model outputs in terms of language model PPL and (2) evaluate the translation performance on Google Paraphrased WMT19 En⇒De test sets (Freitag et al., 2020).

Output fluency

We exploit the monolingual corpora of target languages to train the 4-gram language models. Table 10 shows the language models’ PPL on model outputs of test sets mentioned in §5.2. We find that online self-training has only a slight impact on the fluency of model outputs, with the average PPL of XLM and MASS models only increasing by +3 and +6, respectively. We attribute this phenomenon to the translated target of self-training samples, which is model generated and thus less fluent than natural sentences. However, since the target of BT data is natural and the BT loss term is the primary training objective, the output fluency does not decrease significantly.

Translation performance on paraphrased references

Freitag et al. (2020) collected additional human translations for newstest2019 with the ultimate aim of generating a natural-to-natural test set. We adopt the HQ(R) and HQ(all 4), which have higher human adequacy rating scores, to re-
Table 10: Automatic fluency analysis in terms of perplexity (PPL). Language models are trained on the natural monolingual data in the respective target language.

evaluate our proposed models.

We present the experimental results in Table 11. Our proposed method outperforms baselines on both kinds of test sets. Therefore, we demonstrate that our proposed method improves the UNMT model performance on natural input with limited translationese outputs.

| Model        | Supervised Model (Freitag et al., 2020) | HQ(R) | HQ(all 4) |
|--------------|----------------------------------------|-------|-----------|
| XLM+UNMT     |                                          | 24.5  | 19.6      |
| +Self-training|                                        | 25.9  | 20.7      |
| MASS+UNMT    |                                          | 24.3  | 19.6      |
| +Self-training|                                        | 26.0  | 20.8      |

Table 11: Translation performance on WMT19 En⇒De test sets with additional human translation references provided by Freitag et al. (2020). We report sacreBLEU for comparison with supervised model.

Table 12: Accuracy of NER translation on natural input portion of test sets.

| Model | Approach | NER Acc. |
|-------|----------|----------|
| XLM   | UNMT     | 0.46     |
|       | +Self-training | 0.53     |
| MASS  | UNMT     | 0.44     |
|       | +Self-training | 0.52     |

Table 12: Accuracy of NER translation on natural input portion of test sets.

6.2 Data Gap

Style Gap  From Table 8, our proposed approach achieves significant improvements on the natural input portion while not gaining on the translated input portion over the baselines. It indicates our approach has better generalization capability on the natural input portion of test sets than the baselines.

Content Gap  To verify that our proposed approach bridges the content gap between training and inference, we calculate the accuracy of NER translation by different models. Specifically, we adopt spaCy to recognize the name entities in reference and translation outputs and treat the name entities in reference as the ground truth to calculate the accuracy of NER translation. We show the results in Table 12. Our proposed method achieves a significant improvement in the translation accuracy of NER compared to the baseline. The result demonstrates that online self-training can help the model pay more attention to the input content rather than being affected by the content of the target language training corpus.

6.3 Target Quality

Next, we investigate the impact of target quality on ST. We use the SNMT model from §2.2 to generate ST data rather than the current model itself and keep the process of BT unchanged. As shown in Table 2, the SNMT models perform well on source-original test set and thus yield higher quality target in ST data. We denote this variant as “knowledge distillation (KD)” and report the performance on WMT19/20 E⇌De in Table 13. When target quality gets better, model performance improves significantly, as expected. Therefore, reducing the noise on the target side of the ST data may further improve the performance. Implementing in an unsupervised manner is left to future work.

| Approach | WMT19 | WMT20 |
|----------|-------|-------|
| XLM      |       |       |
| UNMT     | 26.6  | 24.4  |
| +ST      | 28.3  | 26.0  |
| +KD      | 33.8  | 31.0  |
| MASS     |       |       |
| UNMT     | 26.7  | 24.6  |
| +ST      | 28.5  | 26.1  |
| +KD      | 32.9  | 31.0  |

Table 13: Translation performance on WMT19/20 En⇌De. “KD” denotes the variant that exploits SNMT model to generate ST data with higher quality target.
7 Related Work

Unsupervised Neural Machine Translation

Before attempts to build NMT model using monolingual corpora only, unsupervised cross-lingual embedding mappings had been well studied by Zhang et al. (2017); Artetxe et al. (2017, 2018a); Conneau et al. (2018). These methods try to align the word embedding spaces of two languages without parallel data and thus can be exploited for unsupervised word-by-word translation. Initialized by the cross-lingual word embeddings, Artetxe et al. (2018b) and Lample et al. (2018a) concurrently proposed UNMT, which achieved remarkable performance for the first time using monolingual corpora only. Both of them rely on online back-translation and denoising auto-encoding. After that, Lample et al. (2018b) proposed joint BPE for related languages and combined the neural and phrase-based methods. Artetxe et al. (2019) warmed up the UNMT model by an improved statistical machine translation model. Lample and Conneau (2019) proposed cross-lingual language model pretraining, which obtained large improvements over previous works. Song et al. (2019) extended the pretraining framework to sequence-to-sequence. Tran et al. (2020) induced data diversification in UNMT via cross-model back-translated distillation.

Data Augmentation

Back-translation (Sennrich et al., 2016; Edunov et al., 2018; Marie et al., 2020) and self-training (Zhang and Zong, 2016; He et al., 2020; Jiao et al., 2021) have been well studied in the supervised NMT. In the unsupervised scenario, Tran et al. (2020) have shown that multilingual pre-trained language models can be used to retrieve the pseudo parallel data from the large monolingual data. Han et al. (2021) use generative pre-training language models, e.g., GPT-3, to perform zero-shot translations and use the translations as few-shot prompts to sample a larger synthetic translations dataset. The most related work to ours is that offline self-training technology used to enhance low-resource UNMT (Sun et al., 2021). In this paper, the proposed online self-training method for UNMT can be applied to both high-resource and low-resource scenarios without extra computation to generate the pseudo parallel data.

Translationese Problem

Translationese problem has been investigated in machine translation evaluation (Lembersky et al., 2012; Zhang and Toral, 2019; Edunov et al., 2020). These works aim to analyze the effect of translationese in bidirectional test sets. In this work, we revisit the translationese problem in UNMT and find it causes the inaccuracy evaluation of UNMT performance since the training data entirely comes from the translated pseudo-parallel data.

8 Conclusion

Pseudo parallel corpus generated by back-translation is the foundation of UNMT. However, it also causes the problem of translationese and results in inaccuracy evaluation on UNMT performance. We attribute the problem to the data gap between training and inference and identify two data gaps, i.e., style gap and content gap. We conduct the experiments to evaluate the impact of the data gap on translation performance and propose the online self-training method to alleviate the data gap problems. Our experimental results on multiple language pairs show that the proposed method achieves consistent and significant improvement over the strong baseline XLM and MASS models on the test sets with natural input.

Acknowledgements

Zhiwei He and Rui Wang are with MT-Lab, Department of Computer Science and Engineering, School of Electronic Information and Electrical Engineering, and also with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200204, China. Rui is supported by General Program of National Natural Science Foundation of China (6217020129), Shanghai Pujiang Program (21PJ1406800), and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). Zhiwei is supported by CCF-Tencent Open Fund (RAGR20210119).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In ACL.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In ACL.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In ACL.
Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In ICLR.

Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jéou. 2018. Word translation without parallel data. In ICLR.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In EMNLP.

Sergey Edunov, Myle Ott, Marc’Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In ACL.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. In EMNLP.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In EMNLP.

Jesse Michael Han, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, et al. 2021. Unsupervised neural machine translation with generative language models only. arXiv.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc’Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In ICLR.

Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. In ACL-IJCNLP.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In EMNLP.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In ICLR.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In EMNLP.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fanjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. In NeurIPS 2018 Workshop on Interpretability and Robustness for Audio, Speech, and Language.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves smt. In EACL, pages 255–265.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In ACL.

Xuan-Phi Nguyen, Shaﬁq Joty, Thanh-Tung Nguyen, Wu Kui, and Ai Ti Aw. 2021. Cross-model back-translated distillation for unsupervised machine translation. In ICML.

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 Proceedings of the Third Conference on Machine Translation: Research Papers.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. In AAAI.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In ACL.

Kaitào Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In ICML.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2021. Self-training for unsupervised neural machine translation in unbalanced training data scenarios. In NAACL.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In NeurIPS.

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 Findings of ACL.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In ACL.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In EMNLP.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In ACL.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In WMT.
A Training Details

A.1 Training Details of SNMT Model

Training Data We use WMT16 parallel data for En-De and En-Ro and WMT14 for En-Fr. We randomly undersample the full parallel corpus. The final sizes of En-De and En-Fr training corpus are 2M respectively, the size of En-Ro corpus is 400K.

Model We initialize the model parameter by XLM pre-trained model and adopt 2500 tokens/batch to train the SNMT model for 40 epochs. We select the best model by BLEU score on the validation set mentioned in §5.1. Note that in order to avoid introducing other factors, our SNMT models are bidirectional, which is consistent with the UNMT models.

A.2 Training Details of UNMT Model

Training data Table 14 lists the monolingual data used in this study to train the UNMT models. We filter the training corpus based on language and remove sentences containing URLs.

Model We adopt the pre-trained XLM models released by Lample and Conneau (2019) and MASS models released by Song et al. (2019) for all language pairs. In order to better reproduce the results for MASS on En-De, we use monolingual data to continue pre-training the MASS pre-trained model for 300 epochs and select the best model by perplexity (PPL) on the validation set. We adopt 2500 tokens/batch to train the UNMT model for 70 epochs and select the best model by BLEU score on the validation set.

Hyper-parameter The target of self-training samples is the translation of the model, which may be noisy in comparison with the reference. Therefore, we adopted the strategy of linearly increasing $\lambda_S$ and keeping it at a small value to avoid negatively affecting the online back-translation training. We denote the beginning and final value of $\lambda_S$ by $\lambda_S^0$ and $\lambda_S^1$, respectively. We tune the $\lambda_S^0$ within $\{0, 1e^{-3}, 1e^{-2}, 2e^{-2}\}$ and $\lambda_S^1$ within $\{5e^{-3}, 5e^{-2}, 1e^{-1}, 1.5e^{-1}\}$ based on the BLEU score on validation sets.

B Sacrebleu Results

To be consistent with previous works (Lample and Conneau, 2019; Song et al., 2019; Nguyen et al., 2021), we use multi-bleu.perl script in the main text to measure translation performance. However, Post (2018) has pointed out that multi-bleu.perl requires user-supplied preprocessing, which cannot be directly compared and provide a sacreBLEU tool to facilitate this. Although we adopted the same preprocessing steps for all models, we still report BLEU scores calculated with sacreBLEU in this section. Tables 15 to 19 show the sacreBLEU results of Tables 2, 4, 8, 9 and 13, respectively.

C Translation Examples

Table 20 presents several example translations that the UNMT model outputs the hallucinated translations, which are biased towards the target language.

---

12 All the data is available at http://www.statmt.org/wmt20/translation-task.html except for En-De which we will release in our github repo.

13 https://github.com/mjpost/sacrebleu

14 BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.5.1
| Model   | En-Fr | En-De | En-Ro | Avg. |
|---------|-------|-------|-------|------|
| **Full Test Set** |       |       |       |      |
| SNMT    | 37.3  | 33.4  | 29.7  | 33.8 |
| XLM     | 36.5  | 34.3  | 27.4  | 34.1 |
| MASS    | 36.6  | 34.7  | 27.3  | 35.1 |
| **Target-Original Test Set / Translated Input** |       |       |       |      |
| SNMT    | 36.1  | 32.2  | 25.7  | 36.9 |
| XLM     | 37.8  | 36.2  | 26.9  | 42.0 |
| MASS    | 37.9  | 37.3  | 27.3  | 42.7 |
| **Source-Original Test Set / Natural Input** |       |       |       |      |
| SNMT    | 34.2  | 30.1  | 26.3  | 23.6 |
| XLM     | 37.3  | 33.8  | 32.5  | 28.6 |
| MASS    | 34.2  | 30.1  | 26.3  | 23.6 |

Table 15: SacreBLEU results of Table 2.

| Testset   | Model | Approach | En-Fr | En-De | En-Ro | Avg. | $\Delta$ |
|-----------|-------|----------|-------|-------|-------|------|----------|
|           |       |          |       |       |       |      |          |
| **Our Implementation** |       |          |       |       |       |      |          |
| Full set  | XLM   | UNMT     | 36.3  | 34.3  | 27.4  | 34.1 | 34.8    |
|           |       | +Self-training | 36.7 | 34.9  | **28.3** | **34.6** | **36.3** | **33.7** | **34.1** | +0.9 |
|           | MASS  | UNMT     | 36.6  | 34.7  | 27.3  | 35.1 | 35.2    |
|           |       | +Self-training | 36.8 | **35.0** | **29.1** | **35.5** | **36.6** | **33.7** | **34.4** | +0.7 |
| Trg-Ori   | XLM   | UNMT     | 37.8  | 36.2  | 26.9  | 42.0 | 42.2    |
|           |       | +Self-training | **38.0** | **37.5** | **26.7** | **42.1** | **42.9** | **33.8** | **36.8** | +0.3 |
|           | MASS  | UNMT     | 37.9  | **37.3** | 27.3  | **42.7** | **43.2** | **35.2** | **37.3** |      |
|           |       | +Self-training | 37.7 | 37.0  | 27.9  | 42.5 | 43.0    |
| Src-Ori   | XLM   | UNMT     | 33.8  | **30.2** | 26.8  | 22.5 | 27.6    |
|           |       | +Self-training | 34.4 | 30.1  | **23.6** | **28.2** | **23.2** | **29.7** | **32.4** | +1.2 |
|           | MASS  | UNMT     | 34.2  | 30.1  | 26.3  | 23.6 | 27.5    |
|           |       | +Self-training | **34.9** | **30.7** | **28.9** | **24.9** | **30.3** | **31.5** | **30.2** | +1.5 |

Table 17: SacreBLEU results of Table 8.

| Model | WMT19 | WMT20 | Avg. | $\Delta$ |
|-------|-------|-------|------|----------|
| XLM   | +Approach |       |       |          |
| +UnMT | 25.8  | 24.1  | 21.8  | 26.3  | 24.5 | −    |
| +Offline ST | 26.0  | 23.9  | 22.0  | 25.8  | 24.4 | -0.1 |
| +CBD  | **27.4** | **25.2** | **23.0** | **26.7** | 25.6 | +1.1 |
| +Online ST | **27.4** | **25.8** | 22.8  | **27.1** | **25.8** | +1.3 |
| MASS  | +UnMT | 26.0  | 24.3  | 22.1  | 26.5  | 24.7 | −    |
| +Offline ST | 26.4  | 24.2  | 22.1  | 26.4  | 24.8 | +0.1 |
| +CBD  | 27.4  | 25.2  | **22.9** | 26.6  | 25.5 | +0.8 |
| +Online ST | **27.7** | **25.7** | **22.8** | **27.4** | **25.9** | +1.2 |

Table 18: SacreBLEU results of Table 9.

| Model | Natural De | Translated De* |
|-------|------------|---------------|
|       | BLEU | $\Delta$ | BLEU | $\Delta$ |
| SNMT  | 28.6 | −    | 44.9 | −    |
| UNMT  | **22.5** | −6.1 | **42.0** | −2.9 |

Table 16: SacreBLEU results of Table 4.

| Approach | WMT19 | WMT20 |
|----------|-------|-------|
| XLM      |       |       |
| UnMT     | 25.8  | 24.1  |
| +ST      | 27.4  | 25.8  |
| +KD      | **32.4** | **30.6** |
| MASS     |       |       |
| UnMT     | 26.0  | 24.3  |
| +ST      | 27.7  | 25.7  |
| +KD      | **31.8** | **30.5** |

Table 19: SacreBLEU results of Table 13.
| Source                                      | Reference                                      | UNMT                                           |
|---------------------------------------------|------------------------------------------------|------------------------------------------------|
| Mindestens ein Bayern-Fan wurde verletzt aus dem Stadion transportiert. | At least one Bayern fan was taken injured from the stadium. | At least one Scotland fan was transported injured from the stadium. |
| Übrigens : München liegt hier ausnahmsweise mal nicht an der Spitze. | Incidentally , for once Munich is not in the lead. | Remember , Edinburgh is not at the top of the list here for once. |
| Justin Bieber in der Hauptstadt : Auf Bieber-Expedition in Berlin | Justin Bieber in the capital city : on a Bieber expedition in Berlin | Justin Bieber in the capital : On Bieber-inspired expedition in NYC |
| Zum Vergleich : In diesem Jahr werden in Deutschland 260.000 Einheiten fertig. | In comparison , 260,000 units were completed in this year in Germany. | To date , 260,000 units are expected to be finished in the UK this year. |
| Deutschland schiebe ein Wohnungsdefizit vor sich her , das von Jahr zu Jahr größer wird. | Germany has a housing deficit which increases every year. | The U.S. was shooting ahead of a housing deficit that is expected to grow from year to year. |

Table 20: Example translations in WMT16 De⇒En. the UNMT model outputs the hallucinated translations which are biased towards the target language En.