Understanding and Mitigating Uncertainty in Zero-Shot Translation

Anonymous ACL submission

Abstract

Zero-shot translation is a promising direction for building a comprehensive multilingual neural machine translation (MNMT) system. However, its quality is still not satisfactory due to off-target issues. In this paper, we aim to understand and alleviate the off-target issues from the perspective of uncertainty in zero-shot translation. By carefully examining the translation output and model confidence, we identify two uncertainties that are responsible for the off-target issues, namely, extrinsic data uncertainty and intrinsic model uncertainty. Based on the observations, we propose two lightweight and complementary approaches to denoise the training data for model training and explicitly penalize the off-target translations during model training. Extensive experiments on both balanced and imbalanced datasets show that our approaches significantly improve the performance of zero-shot translation over strong MNMT baselines. Qualitative analyses provide insights into where our approaches reduce off-target translations.

1 Introduction

Multilingual neural machine translation (MNMT) aims to translate between any two languages with a unified model (Johnson et al., 2017; Aharoni et al., 2019; Wang et al., 2022). It is appealing due to the model efficiency, easy deployment, and knowledge transfer between languages. Previous studies (Johnson et al., 2017; Gu et al., 2019) suggest that knowledge transfer in MNMT significantly improves the performance of low-resource translation, and is potential for zero-shot translation between language pairs unseen in the training process. Since it is costly and even unrealistic to build parallel data for all language pairs, improving the quality of zero-shot translation is a promising direction for developing a comprehensive and well-performing MNMT system.

However, zero-shot translation suffers from serious **off-target issues** (Ha et al., 2016; Gu et al., 2019; Zhang et al., 2020), where the MNMT model tends to translate into other languages rather than the expected target language. As a result, the quality of zero-shot translation is far from satisfactory for practical application. A number of recent efforts have explored ways to improve zero-shot translation by mitigating the off-target issues. One thread of work focuses on modifying the model architecture (Zhang et al., 2020; Liu et al., 2021; Wu et al., 2021) or introducing auxiliary tasks (Al-Shedivat and Parikh, 2019; Yang et al., 2021; Wang et al., 2021b) to enhance the flexible translation relations in MNMT. Another thread of work aims to generate synthetic data for zero-shot translation pairs in either off-line (Gu et al., 2019) or on-line (Zhang et al., 2020) modes. However, these approaches require additional efforts for model modification and computational costs.

In this work, we target better understanding and mitigating the off-target issues in zero-shot translation. We first empirically connect the widely-cited off-target issues in zero-shot translation to the uncertain prediction of MNMT models, which assign high confidence to the off-target translations for zero-shot language pairs (§ 3). We then identify two language uncertainties that are responsible for the uncertain prediction on target languages:

- **extrinsic data uncertainty** (§ 4): we show that for 5.8% of the training examples in the commonly-used multilingual data OPUS (Zhang et al., 2020), the target sentences are in the source language. Previous studies have shown that such data noises can significantly affect the model uncertainty for bilingual NMT. Our study empirically reconfirms these findings for zero-shot translation, which is more sensitive to the data noises without supervision from parallel data.

- **intrinsic model uncertainty** (§ 5): we show that

MNMT models tend to spread too much probability mass over the vocabulary of off-target languages in zero-shot translation, resulting in an overall over-estimation of hypotheses in off-target languages. In contrast, the trend does not hold for supervised translations.

Starting from the above observations, we propose two lightweight and complementary approaches to mitigate the data and model uncertainties. For data uncertainty, we remove the off-target sentence pairs from training data to make sure the MNMT models can learn more correct mappings between languages during training. For model uncertainty, we propose unlikelihood training to explicitly penalize the off-target translations in training, which can perform better when the counteractive effect of data uncertainty is removed. Experimental results across different MNMT scenarios show that our approaches significantly improve zero-shot translation performance over strong MNMT baselines. Extensive analyses demonstrate that our approaches successfully reduce the ratios of off-target translations from more than 20% to as low as 1.1%.

Contributions The main contributions of our work are listed as follows:

- We identify two uncertainties, namely extrinsic data uncertainty and intrinsic model uncertainty, which are responsible for the off-target issues in zero-shot translation.
- We propose two effective approaches to mitigate the off-target issues, which introduce no or only marginal additional computational cost.

2 Preliminary

2.1 Definition of Off-Target Issue

Off-Target Issue is a type of translation error that commonly occurs in zero-shot translation (Ha et al., 2016; Zhang et al., 2020). It describes the phenomenon that MNMT models ignore the given target language information and translate the source sentence into wrong languages. Assume that $\mathcal{L}$ denotes the set of languages involved in the MNMT model, and $T \in \mathcal{L}$ is the target language, the off-target ratio (OTR) is calculated as:

$$\mathbf{OTR} = \frac{\sum_{i=1}^{N} \mathbb{1}_{l(\tilde{y}_i) \neq T}}{N}, \quad (1)$$

where N is the number of test samples, and $l(\tilde{y}_i)$ denotes the detected language of the translation $\tilde{y}_i$. We adopt OTR as one of the metrics to evaluate the performance of zero-shot translation in this work.

2.2 Experimental Setup

Training Data We mainly conduct experiments on two types of datasets:

- ***OPUS-100 Data*** is an **imbalanced** multilingual dataset, where some language pairs have more training instances than the others. (Zhang et al., 2020) propose OPUS-100 that consists of 55M English-centric sentence pairs covering 100 languages. We also choose five language pairs from OPUS-100, including English-German (En-De), -Chinese (En-Zh), -Japanese (En-Ja), -French (En-Fr), and -Romanian (En-Ro) to construct ***balanced OPUS-6 Data***. We follow (Zhang et al., 2020) to apply BPE (Sennrich et al., 2016) to learn a joint vocabulary size of 64K learned from the whole OPUS-100 dataset.
- ***WMT-6 Data*** is a large-scale **imbalanced** dataset. Specifically, we collect the language pairs same as OPUS-6 from the WMT competition tasks, including WMT20 En-De (45.2M), WMT20 En-Zh (19.0M), WMT20 En-Ja (11.5M), WMT14 En-Fr (35.5M), and WMT16 En-Ro (0.6M). We learn a joint BPE (Sennrich et al., 2016) model with 32K merge operations.

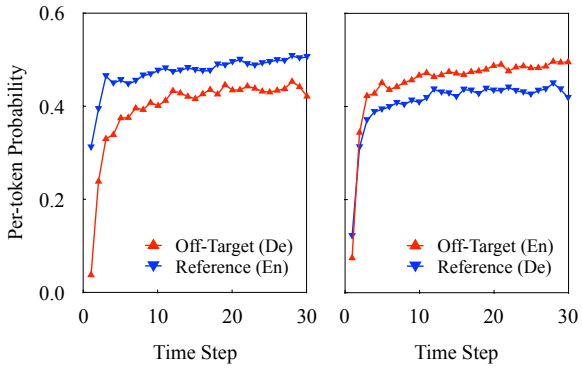
Multi-Source Test Set To eliminate the content bias across languages (Wang et al., 2021a), we evaluate the performance of multilingual translation models on the multi-source TED58 test set (Qi et al., 2018; Tran et al., 2020), where each sentence is translated into multiple languages. We select the above six languages and filter the original test set to ensure that each sentence has the translations in all the six languages. Finally, we obtain 3804 sentences in six languages, i.e., 22824 sentences in total. We use the filtered testset to evaluate the performance on both supervised and zero-shot translation. We report the results of both BLEU scores (Papineni et al., 2002) and off-target ratios (OTR) for both supervised and zero-shot translation. For example, the supervised translation and the zero-shot translation performance on OPUS-6 dataset are the average of 10 supervised directions (i.e., En-X and X-En) and 20 zero-shot directions (i.e., X_i-X_j), respectively. We employ the `langid`

Training Data	Supervised BLEU↑	Supervised OTR↓	Zero-Shot BLEU↑	Zero-Shot OTR↓
S-ENC-T-DEC Models				
OPUS-6	27.1	1.9	12.3	20.6
WMT-6	28.0	1.8	10.6	37.8
T-ENC Models				
OPUS-6	27.2	1.9	10.2	32.1
WMT-6	28.8	1.7	13.3	22.5

Table 1: BLEU scores and off-target ratios (OTR) of MNMT models on supervised and zero-shot test sets.

Target Lang.	BLEU↑ Sup.	BLEU↑ Zero	BLEU↑ △	OTR↓ Sup.	OTR↓ Zero	OTR↓ △
Ja	19.0	15.7	-3.3	0.4	2.1	+1.7
Zh	23.1	11.4	-11.7	0.4	32.6	+32.2
De	29.4	6.1	-23.3	2.8	49.6	+46.8
Fr	37.1	6.4	-30.7	2.7	61.8	+59.1
Ro	26.8	11.4	-15.4	3.7	14.3	+10.6

Table 2: Results on supervised ("Sup.") and zero-shot ("Zero") test sets for T-ENC model trained on OPUS-6.



(a) Supervised Fr-En (b) Zero-Shot Fr-De

Figure 1: Per-token probabilities of (a) supervised Fr-En and (b) zero-shot Fr-De translations. Higher probabilities are expected for the on-target references ("Reference"), and lower probabilities are expected for the off-target distractor translations ("Off-Target").

library[1], the most widely used language identification tool with 93.7% accuracy on 7 dataset across 97 languages, to detect the language of sentences and calculate the off-target ratio for zero-shot translation directions. We also adopt two widely used evaluation metrics, COMET (Rei et al., 2020) and chrF (Popovic, 2015) to validate our method.

Model All NMT models in this paper follow the Transformer-big settings, with 6 layers, 1024 hidden size and 16 heads. To distinguish languages, we add language tokens to the training samples by two strategies implemented in Fairseq, i.e., S-ENC-T-DEC and T-ENC. The S-ENC-T-DEC strategy adds source language tokens at encoder and target language tokens at decoder, while T-Enc only adds target language tokens at encoder. We regard T-ENC as a strong baseline which has been demonstrated better for zero-shot translation (Wu et al., 2021). For multilingual translation models, we train a Transformer-big model with 1840K tokens per batch for 50K updates. We conduct the experiments on 16 NVIDIA V100 GPUs and select the final model by the lowest loss on the validation set.

Our MNMT models consistently outperform their bilingual counterparts, demonstrating that our models are well trained so that the findings and improvement in this work are convincing. More details can be found in Appendix A.1.

3 Analyzing Uncertainty

In this section, we present poor zero-shot performance of our well-trained MNMT models due to off-target issues. Then we link the off-target issues to the uncertain prediction on target languages.

[1] https://github.com/saffsd/langid.py

Poor Zero-Shot Performance and Off-Target Issues. Table 1 lists the translation results. Compared with the supervised translation, the MNMT models produce lower-quality zero-shot translations (e.g., 15+ BLEU scores lower) due to much higher ratios of off-target translations (e.g., 32.1 vs. 1.9 on OPUS-6 with T-ENC). To further validate our claim, we list the detailed results in Table 2. As seen, the gap in BLEU score between supervised and zero-shot translations is highly correlated to that of OTR, showing the high correlation between translation performance and off-target issues.

Uncertain Prediction Causes Off-Target Issues. To investigate how MNMT models generate off-target translations, we follow (Ott et al., 2018) to analyze the model confidence on the target language. Specifically, we compute the average probability at each time step across a set of sentence pairs. In addition to the ground-truth reference sentence, we also consider a "distractor" translation in the off-target language for each source sentence.

Training Data	Language Paris (En-)					
	Zh	Ja	De	Fr	Ro	Ave.
OPUS-6	1.3	1.1	8.5	9.0	9.2	5.8
WMT-6	0.1	0.6	2.5	2.3	2.1	1.5

Table 3: Ratios(%) of off-target noises in the datasets.

Figure 1 plots the model confidence for both references ("Reference") and distractors ("Off-Target") on supervised Fr-En and zero-shot Fr-De tasks. We find that 94.7% of the off-target translations in the zero-shot Fr-De task are in English. Therefore, we only present the English off-target translation for simplicity. Different from the supervised translation, the zero-shot translation shows a surprisingly higher confidence on the off-target distractors. Accordingly, the model tends to generate more off-target translations (i.e., 74.9% vs. 1.6%).

In the following sections, we will connect the uncertain prediction problem to the language uncertainty in both data (§ 4) and model (§ 5). Based on these findings, we provide simple and effective solutions to mitigate the data and model uncertainty.

4 Extrinsic Data Uncertainty

Problem: Off-Target Noises in Multilingual Training Data The uncertainty in multilingual training data is an important reason for the uncertain prediction in zero-shot translation. As a data-driven approach, MNMT models learn the mappings between languages from the parallel data, which we assume that both the source and target sentences are in correct languages. However, we find that a quite portion of training data contains off-target translations, which are **mainly in the source language**. Table 3 lists the statistics, where we observe a high off-target ratio in both OPUS-6 (i.e., 5.8%) and WMT-6 (i.e., 1.5%). Previous study on bilingual MT (Ott et al., 2018) suggests that 1% to 2% of such data noises can make the NMT model highly uncertain and tend to produce translations in source language. We believe that similar uncertainty issues will also occur in MNMT models, especially for zero-shot translation where no supervision signal (from parallel data) exists.

Solution: Data Denoising We propose data denoising to make sure that MNMT learns a more correct language mapping from the training data. Specifically, we adopt the `langid` tool (Lui and Baldwin, 2012) to identify the off-target sentence pairs in the parallel training data of each language pair and remove them to build a clean dataset. The clean dataset is then used for training the MNMT models. Without the distraction from the off-target sentence pairs, the MNMT model is expected to be more confident on the target languages. As a result, we can reduce the off-target ratio and improve the performance of zero-shot translation.

Training Data	Supervised		Zero-Shot	
	BLEU↑	OTR↓	BLEU↑	OTR↓
OPUS-6 Data				
Raw Data	27.2	1.9	10.2	32.1
+ Denoise	27.1	1.5	14.0	10.0
WMT-6 Data				
Raw Data	28.8	1.7	13.3	22.5
+ Denoise	28.8	1.6	15.3	10.4

Table 4: Results of data denoising for T-ENC model.

Results Table 4 lists the results of removing off-target noises for both OPUS-6 and WMT-6 datasets. The data denoising method significantly improves the zero-shot translation performance by greatly reducing the off-target issues. However, there are still around 10% off-target translations unsolved, which we attribute to the intrinsic model uncertainty due to the nature of multilingual learning (§ 5).

5 Intrinsic Model Uncertainty

5.1 Problem: Over-Estimation on Off-Target Vocabulary

The uncertainty inside the MNMT model is another reason for the uncertain prediction in zero-shot translation. To enhance the knowledge transfer between languages, researchers seek to train a single model with parameters shared by different languages, including the vocabulary and the corresponding embedding matrix. However, the shared vocabulary also introduces uncertainty to the decoder output. Theoretically, the MNMT model is allowed to predict any token in the vocabulary, preserving the possibility of decoding into a wrong language. Such a language uncertainty can be avoided with the supervision of parallel data, which is unavailable for zero-shot translation.

Empirically, we compute the prediction distribution over the whole vocabulary for each token in the reference sentences. Then, we calculate how

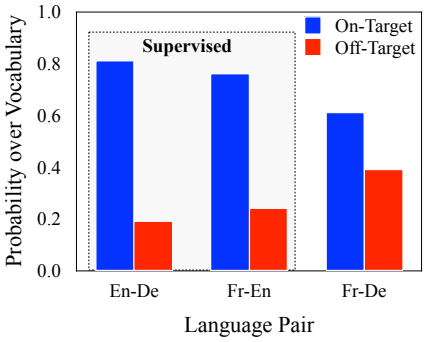


Figure 2: Probability over vocabulary of supervised (En-De, Fr-En) and zero-shot Fr-De translations. MNMT model over-estimates the off-target vocabulary (red column) for zero-shot translation.

much of the probability mass is assigned to the target language ("On-Target") based on its individual vocabulary, and how much to the others ("Off-Target"). Figure 2 plots the results on the zero-shot Fr-De translation. For reference, we also plot the related supervised En-De and Fr-En translations. Obviously, for supervised translation, the vast majority of the probability mass is assigned to the target language. However, for zero-shot translation, more probability mass (i.e., around 39%) is assigned to off-target languages, thus leading to serious off-target issues.

5.2 Solutions

Based on the above findings, we propose two methods to reduce the over-estimation on off-target vocabulary, which differ in whether to use the off-target vocabulary in training.

Vocabulary Masking One straightforward solution to model uncertainty is to constrain the probability distributions only on the vocabulary of target language by masking the output logits on off-target vocabulary. Specifically, we extract a language-specific vocabulary V_l for each language $l \in \mathcal{L}$ from the full vocabulary V ($V_l \subset V$). We first build a BPE vocabulary shared by all languages, which is the same one used for the vanilla MNMT model. We then construct the language-specific vocabulary by counting the BPE tokens in the segmented training data of the corresponding language. Note that different language-specific vocabularies can have shared tokens. For example, the English-specific vocabulary shares 33% tokens with German-specific vocabulary on the OPUS-6 data (see Table 11 in Appendix for more details).

This method can be applied in both training and inference. When predicting target tokens, we mask out the tokens that do not appear in the vocabulary V_T of target language T. Formally, the output probability of the token y is calculated as:

$$P_\theta(y|\mathbf{h}_t) = \begin{cases} \frac{\exp(\mathbf{h}_t^\top \mathbf{w}_y)}{\sum_{y' \in V_T} \exp(\mathbf{h}_t^\top \mathbf{w}_{y'})}, & y \in V_T \\ 0, & \text{otherwise} \end{cases}$$

where $\mathbf{h}_t$ is the hidden state at time step t, and $\mathbf{w}_y$ is the word embedding of the token y.

Unlikelihood Training While the *vocabulary masking* method can successfully reduce the probabilities of translations over the wrong languages, the performance may be limited by two factors: (1) The language-specific vocabularies need to be carefully partitioned for different languages, especially those similar ones (e.g., English and German). (2) The isolation of vocabularies may hinder knowledge transfer across languages. To avoid these limitations, we incorporate the *unlikelihood training objective* (Welleck et al., 2019) for MNMT, which forces the model to assign lower probability to unlikely generations.

Formally, the original likelihood training loss on a translation sentence pair is expressed as:

$$\mathcal{L}_{\text{Likelihood}} = -\sum_{t=1}^{T} \log P_\theta(y|\mathbf{x}, \mathbf{y}^{l_c}_{<t}),$$

where l_c denotes the correct language tag for the target sentence $\mathbf{y}$. This training loss encourages the model to generate *on-target translation*.

We design an additional unlikelihood loss to penalize the *off-target translation*. To simulate the off-target translation, for each sentence pair we change the target language tag to another wrong language l_w and form the negative candidate. Then the unlikelihood training loss is defined as:

$$\mathcal{L}_{\text{Unlikelihood}} = -\sum_{t=1}^{T} \log(1 - P_\theta(y|\mathbf{x}, \mathbf{y}^{l_w}_{<t})).$$

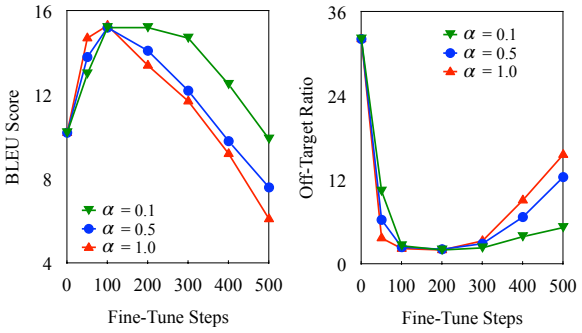
The final loss is the combination of the above two:

$$\mathcal{L} = \mathcal{L}_{\text{Likelihood}} + \alpha \mathcal{L}_{\text{Unlikelihood}}.$$

In this way, we provide supervision for zero-shot directions by penalizing the off-target translations (i.e., mismatch between target language tag and target sentence). We follow Welleck et al. (2019) to fine-tune pretrained MNMT model with the combined loss for K steps.

Mask in		Supervised		Zero-Shot	
Train	Infer.	BLEU↑	OTR↓	BLEU↑	OTR↓
×	×	27.2	1.8	10.2	32.1
×	✓	27.2	1.8	**13.1**	**12.7**
✓	✓	27.2	1.8	12.5	18.6

Table 5: Impact of vocabulary masking used in inference or both training and inference on OPUS-6 data for T-ENC model.



(a) BLEU↑ (b) Off-Target Ratio↓

Figure 3: Impact of interpolation weight α and fine-tune step K on zero-shot translations.

Model	Supervised		Zero-Shot	
	BLEU↑	OTR↓	BLEU↑	OTR↓
OPUS-6 Data				
Vanilla	27.2	1.9	10.2	32.1
+ Vocab Mask	27.2	1.8	13.1	12.7
+ Unlike Train	27.2	1.5	15.2	2.2
WMT-6 Data				
Vanilla	28.8	1.7	13.3	22.5
+ Vocab Mask	28.8	1.6	14.9	10.7
+ Unlike Train	28.8	1.6	16.3	5.6

Table 6: Results of mitigating model uncertainty for T-ENC model on raw data without denoising.

5.3 Ablation Study

Ablation Study on Vocabulary Masking The proposed vocabulary masking method can be used in both training and inference. Table 5 lists the results of different masking strategies. Applying vocabulary masking to the vanilla MNMT model during inference significantly improves zero-shot translation performance by remedying off-target issues, which demonstrates the effectiveness of vocabulary masking. However, further including vocabulary masking into the training process makes the improvement of zero-shot translation less significant. One possible reason is that isolating the vocabularies between languages during training may hinder cross-lingual knowledge transfer.

Ablation Study on Unlikelihood Training Figure 3 shows the impact of the interpolation weight α and fine-tune steps K on unlikelihood training. The zero-shot performance goes up with the increase of fine-tune steps K until $K = 100$ for all interpolation weights, and declines when fine-tuning for more steps. One possible reason is that the negative examples are semantically equivalent sentence pairs, while the target language tag is replaced with a wrong tag beyond the target language. The mismatch between target language tag and target sentence is a simple pattern, which can be easily learned by the model with as few as 100 steps. Fine-tuning with unlikelihood loss of higher interpolation weights or for more steps potentially harms the cross-lingual transfer ability among semantically equivalent sentences. In the following experiments, we use $\alpha = 0.1$ and $K = 100$ as default for its robust performance.

Comparison Results Table 6 lists the results of vocabulary masking and unlikelihood training. Clearly, unlikelihood training consistently outperforms vocabulary masking on zero-shot translation in all cases. We attribute to the superiority of unlikelihood training on directly penalizing simulated off-target translation during model training. In the following experiments, we use unlikelihood training to mitigate model uncertainty as default.

6 Main Results

6.1 Translation Performance

Table 7 lists the results of zero-shot translations on different benchmarks for different MNMT architectures. Clearly, using data denoising alone significantly improves the zero-shot translation performance in all cases, and unlikelihood training can even performs better by reducing more off-target issues. Combing them together achieves the best performance, demonstrating the complementarity between data uncertainty and model uncertainty. We also demonstrate the effectiveness of our proposed method using different metrics like COMET and ChrF, as shown in Table 8.

Larger-Scale Imbalanced Datasets In addition to the small-scale balanced OPUS-6 data, we also validate our approaches on the larger-scale imbal-

Model	OPUS-6 Data		WMT-6 Data		OPUS-100 Data	
	BLEU↑	OTR↓	BLEU↑	OTR↓	BLEU↑	OTR↓
S-ENC-T-DEC *MNMT Models*						
Vanilla	12.3	20.6	10.6	37.8	1.2	92.7
+ Data Denoising	14.1	7.0	11.1	23.9	4.8	45.0
+ Unlikelihood Training	15.3	1.7	16.4	4.0	12.5	2.3
+ Data Denoising	**15.6**	**1.1**	**17.2**	**2.4**	**12.6**	**1.6**
T-ENC *MNMT Models*						
Vanilla	10.2	32.1	13.3	22.5	7.5	38.4
+ Data Denoising	14.0	10.0	15.3	10.4	8.8	21.4
+ Unlikelihood Training	15.0	2.6	16.3	5.9	12.6	5.7
+ Data Denoising	**15.2**	**2.2**	**16.8**	**4.2**	**13.1**	**3.5**

Table 7: BLEU scores and off-target ratios (OTR) on the TED58 **zero-shot** (i.e., 20 non-English-centric pairs among Zh, Ja, De, Fr, and Ro) test sets. Our approaches consistently improve zero-shot translation performance without sacrificing the quality of supervised translation (shown in Table 12 in Appendix).

Model	Supervised		Zero-Shot	
	COMET↑	ChrF↑	COMET↑	ChrF↑
S-ENC-T-DEC *MNMT Models*				
Vanilla	0.168	51.0	-0.297	26.0
+ Data Denoise	0.169	51.4	-0.193	29.1
+ Vocab Mask	0.169	51.4	-0.125	30.0
+ Unlike Train	0.169	51.4	-0.098	30.8
T-ENC *MNMT Models*				
Vanilla	0.316	46.4	-0.336	22.9
+ Data Denoise	0.319	46.4	-0.187	28.8
+ Vocab Mask	0.319	46.4	-0.142	29.3
+ Unlike Train	0.320	46.4	-0.104	30.8

Table 8: Results of MNMT models trained on the OPUS-6 dataset measured by other evaluation metrics.

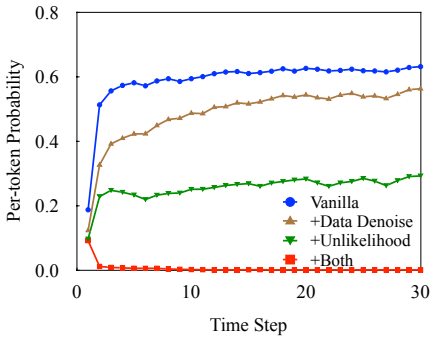


Figure 4: Per-token probabilities of off-target test sentences for zero-shot Fr-De translations for T-ENC model with our methods on OPUS-6 data.

anced datasets (i.e. 111.8M WMT-6 data of 6 languages and 55.0M OPUS-100 of 100 languages). Generally, the off-target issues are more severe in imbalanced scenarios. For example, the zero-shot translation almost crashes on imbalanced OPUS-100 data with 92.7% of off-target translation. Our approaches performs surprisingly well by reducing the off-target issues to as low as 1.6% to 2.4%, which are close to that on the small-scale balanced data (i.e. 1.1% on OPUS-6). These results demonstrate the scalability of our approaches to massively multilingual translation tasks.

Different Tagging Strategies There are considerable differences between T-ENC and S-ENC-T-DEC models, which differ in how to attach the language tags. T-ENC performs significantly better on imbalanced datasets (especially on OPUS-100), while performs worse on balanced OPUS-6 data than its S-ENC-T-DEC counterpart. Our approaches can consistently improve zero-shot performance on top of T-ENC in all cases, demonstrating the universality of the proposed approaches.

With the help of our approaches, S-ENC-T-DEC produces better overall performance than T-ENC. One possible reason is that S-ENC-T-DEC is better at modeling language mapping by explicitly identifying the source and target languages. Meanwhile, the side-effect of over-fitting on supervised mapping can be almost solved by our approaches.

6.2 Prediction Uncertainty

Figure 4 shows the prediction probabilities on the off-target translation that are produced by the vanilla T-ENC model on zero-shot translation. Clearly, data denoising and unlikelihood training consistently reduce the model confidence on the

Model	Supervised		Zero-Shot	
	BLEU↑	OTR↓	BLEU↑	OTR↓
Raw Data				
Vanilla	27.2	1.8	10.2	32.1
+ RemoveRes.	26.7	1.8	12.8	21.7
+ AE Loss	26.7	2.0	12.2	21.0
+ Ours	27.1	1.5	15.0	2.6
Clean Data (*with data denoising*)				
Vanilla	27.1	1.5	14.0	10.0
+ RemoveRes.	27.1	1.5	14.2	5.6
+ AE Loss	26.3	1.5	14.5	5.1
+ Ours	27.2	1.5	15.2	2.2

Table 9: Comparison with previous work on improving zero-shot translation. "Clean Data" denotes filtering off-target noises using our data denoising method.

off-target translation, which reconfirms our claim that extrinsic data uncertainty and intrinsic model uncertainty are responsible for the uncertain prediction on target languages. Specifically, we find that data denoising reduces the confidence on the first few tokens of off-target translations noticeably, while unlikelihood training consistently reduces all the tokens. Combining them together ("+Both") can surprisingly reduce the per-token probability of off-target translation to zero. The reason is that likelihood training on these off-target noises could encourage the model to generate off-target translation, which partially counteracts the effect of unlikelihood training that prevents the model from generating off-target translation. Therefore, denoising such off-target noises can further improve the performance of unlikelihood training.

6.3 Comparison with Previous Work

We compare our methods with two recent works on improving zero-shot translation: (1) RemoveRes. (Liu et al., 2021) that removes residual connections in an encoder layer to disentangle the positional information; (2) AE Loss (Wang et al., 2021b) that introduces a denoising autoencoding loss to implicitly maximize the probability distributions for zero-shot directions. We reimplemented these methods on top of the T-ENC model as done in the original papers. Table 9 lists the results on OPUS-6 data, which shows that our methods can consistently outperform their methods. The improvement is much larger on the noisy raw data, which we attribute to the advantage of our approach in directly penalizing off-target translations.

7 Related Work

Improving Zero-Shot Translation A number of recent efforts have explored ways to improve zero-shot translation by mitigating the off-target issues. One thread of work focuses on modifying the model architecture (Zhang et al., 2020; Liu et al., 2021; Wu et al., 2021). Another thread of work aims to generate synthetic data for zero-shot translation pairs in either off-line (Gu et al., 2019) or on-line (Zhang et al., 2020) modes. Our work is complementary to them: we remove the off-target noises in the original data rather than leveraging additional data. Besides, researchers also try to introduce auxiliary tasks with additional training losses (Al-Shedivat and Parikh, 2019; Yang et al., 2021; Wang et al., 2021b) to help the model training. We propose a novel and light-weight method to directly reduce the off-target translation via unlikelihood training.

Uncertainty in NMT Closely related to our work, (Ott et al., 2018) analyzed the uncertainty in bilingual machine translation, and attributed it to one specific type of data noise – copies of source sentences. In contrast, we analyze the uncertainty in multilingual machine translation, which is a more complicated scenario. Besides data uncertainty, we also reveal the intrinsic model uncertainty on the output distributions due to the shared vocabulary across multiple languages. In addition, the proposed methods for reducing model uncertainty by either masking out off-target vocabularies or penalizing off-target training examples are carefully designed for the multilingual scenario.

8 Conclusion

We present a comprehensive study of the off-target issues in zero-shot translation. We empirically show that the off-target noises in training examples and the shared vocabulary across languages bias MNMT models to over-estimate the translation hypotheses in off-target languages. In response to this problem, we propose several lightweight and complementary approaches to mitigate the uncertainty issues, which can significantly improve zero-shot translation performance with no or only marginal additional computational costs.

Future work will include investigating the uncertainty of large MNMT models trained on more complicated datasets (Fan et al., 2021; Schwenk et al., 2021) and also validating our approach.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *NAACL*.

Maruan Al-Shedivat and Ankur P. Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *NAACL*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and V. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. *ACL*.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. In *ACL/IJCNLP*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *ACL*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*, pages 3956–3965. PMLR.

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

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *WMT*.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *NAACL*.

Ricardo Rei, Craig Alan Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *EMNLP*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

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

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

Qian Wang, Jiajun Zhang, and Chengqing Zong. 2022. Synchronous inference for multilingual neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1827–1839.

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

Weizhi Wang, Zhirui Zhang, Yichao Du, Boxing Chen, Jun Xie, and Weihua Luo. 2021b. Rethinking zero-shot neural machine translation: From a perspective of latent variables. In *EMNLP (Findings)*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. In *ICLR*.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. In *ACL Findings*.

Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. 2021. Improving multilingual translation by representation and gradient regularization. In *EMNLP*.

Biao Zhang, P. Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *ACL*.

A Appendix

A.1 MNMT Models Are Well Trained

Model	English⇒X		X⇒English	
	OPUS	TED	OPUS	TED
Bilingual Model	32.0	21.6	31.0	22.9
Multilingual NMT Models				
S-ENC-T-DEC	34.8	26.9	33.5	27.2
T-ENC	34.8	27.1	33.5	27.4

Table 10: BLEU scores of bilingual and multilingual TRANSFORMER-BIG models trained on **OPUS-6** data for supervised translation. We report results on both the test sets provided by the OPUS data ("OPUS") and the multi-source TED test set used in this work ("TED").

Table 10 lists the supervised translation performance of our multilingual NMT models on both the OPUS test sets and the multi-source TED test set. For comparison, we also include the bilingual model for each language pair as baselines. For bilingual models, we train a Transformer-big model with 460K tokens per batch for 30K updates. Clearly, our MNMT models consistently and significantly outperform their bilingual counterparts, demonstrating that our models are well trained so that the findings and improvement in this work are convincing.

A.2 Statistics of Language-Specific Vocabulary

	En	De	Fr	Ja	Ro	Zh
En	**17.2K**					
De	5.7K	**16.2K**				
Fr	6.0K	5.5K	**14.9K**			
Ja	0.9K	0.8K	0.7K	**9.0K**		
Ro	3.8K	4.2K	5.0K	0.3K	**12.1K**	
Zh	2.0K	1.2K	1.2K	2.4K	1.2K	**15.4K**

Table 11: Statistics of language-specific vocabulary used in vocabulary masking on OPUS-6 data.

For the vocabulary masking approaches, we extract a language-specific vocabulary from the full vocabulary, and different language-specific vocabularies can have shared tokens. Table 11 lists the vocabulary statistics on OPUS-6 data. For example, the size of English vocabulary is 17.2K, which shares 5.7K tokens with the German vocabulary.

A.3 Supervised Translation Performance of Main Results (Table 7)

Model	OPUS-6 Data		WMT-6 Data		OPUS-100 Data	
	BLEU↑	OTR↓	BLEU↑	OTR↓	BLEU↑	OTR↓
S-ENC-T-DEC MNMT Models						
Vanilla	27.1	1.9	28.0	1.8	26.9	1.8
+ Data Denoise	27.2	1.5	28.0	1.7	27.0	1.5
+ Unlike Train	27.1	1.8	28.0	1.7	27.0	1.5
+ Data Denoise	27.2	1.5	28.0	1.6	27.0	1.5
T-ENC MNMT Models						
Vanilla	27.2	1.8	28.8	1.7	26.9	1.8
+ Data Denoise	27.1	1.5	28.8	1.6	26.9	1.6
+ Unlike Train	27.1	1.5	28.8	1.6	26.9	1.7
+ Data Denoise	27.2	1.5	28.8	1.6	26.9	1.7

Table 12: BLEU scores and off-target ratios (OTR) of multilingual translation models on the TED58 **supervised** (i.e., 10 English-centric language pairs) test sets that cover 6 languages.