Understanding and Mitigating the Uncertainty in Zero-Shot Translation

Wenxuan Wang, Wenxiang Jiao, Shuo Wang, Zhaopeng Tu, and Michael R. Lyu, Life Fellow, IEEE

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 by unlikelihood training 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.

Index Terms—Neural machine translation, zero-shot translation, uncertainty.

I. INTRODUCTION

MULTILINGUAL neural machine translation (MNMT) aims to translate between any two languages with a unified model [1], [2], [3], [4]. It is appealing due to the model’s efficiency, easy deployment, and knowledge transfer between languages. Previous studies [1], [5], [6] suggest that knowledge transfer in MNMT significantly improves the performance of low-resource translation, and has the potential for zero-shot translation between language pairs unseen in the training process. For example, a widely-adopted setting is that the training data contains non-English to English (X-En) and English to non-English (En-X) sentence pairs [1], [7], [8], [9], [10]. The MNMT model trained on such data can conduct zero-shot translation between two non-English languages (X-X). 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 [5], [7], [11], where the MNMT model tends to translate into other languages rather than the expected target language. Table 1 shows off-target examples for zero-shot translation. 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 off-target issues. One thread of work focuses on modifying the model architecture [7], [12], [13] or introducing auxiliary training losses [9], [14], [15] to enhance the flexible translation relations in MNMT. Another thread of work aims to generate synthetic data for zero-shot translation pairs in either an offline [5] or online [7] manner. These approaches require additional efforts for model modification and computational costs.

In this work, we aim to better understand and mitigate 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 (Section III). In uncertainty theory, there are different sources of uncertainty for engineering systems. These uncertainties can come from model design, named intrinsic uncertainties, or from model calibrations using uncertain data, named extrinsic uncertainty [16]. Following this framework, we then identify two language uncertainties that are responsible for the uncertain prediction of target languages. Extrinsic uncertainty is the uncertainty caused outside the model while intrinsic uncertainty is the uncertainty caused by model design:

- **extrinsic data uncertainty** (Section IV): we show that for 5.8% of the training examples in the commonly used multilingual data OPUS [7], the target sentences are in the source

| Category | Examples |
|----------|----------|
| to-Source | src (Fr): Tout semble s’être bien passé. |
| to-Source | tgt (En): Alles scheint gut gelaufen zu sein. |
| to-Others | src (Fr): Metter-le sur quatre assiettes. |
| to-Others | tgt (En): Legen Sie ihn auf vier Teller. |
| to-Others | hyp (En): Put it on four plates. |

The MNMT model often copies the source sentence or translates into other languages (e.g. English rather than German).

TABLE I OFF-TARGET EXAMPLES FOR ZERO-SHOT FR-DE IN MULTILINGUAL TRANSLATION

Received 4 March 2023; revised 2 April 2023, and 29 July 2024; accepted 17 October 2024. Date of publication 31 October 2024; date of current version 21 November 2024. This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Grant CUHK14206921 of the General Research Fund. The associate editor coordinating the review of this article and approving it for publication was Dr. Yue Zhang. (Corresponding author: Zhaopeng Tu.)

Wenxuan Wang and Michael R. Lyu are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong 999077, China.

Wenxiang Jiao and Zhaopeng Tu are with the Tencent AI Lab, Shenzhen 518057, China (e-mail: tuzhaopeng@gmail.com).

Shuo Wang is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

Digital Object Identifier 10.1109/TASLP.2024.3485555
language. Previous studies have shown that such data noises [17], [18], [19] can significantly affect the model uncertainty for bilingual NMT. Our study empirically re-confirms these findings for zero-shot translation, which is more sensitive to the data noises without supervision from parallel data.

- **intrinsic model uncertainty (Section V):** 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 unlikelyhood 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.

The rest of the paper is organized as follows: We first introduce the background of multilingual translation and off-target translation, existing working, as well as the experimental setup in Section II; Then in Section III, we present poor zero-shot performance of well-trained MNMT models due to off-target issues; In Section IV and Section V, we identify two uncertainties that are responsible for the off-target issues, i.e., extrinsic data uncertainty and intrinsic model uncertainty, as well as the corresponding mitigating methods; Next, in Section VI, we empirically evaluate the effectiveness of the proposed methods; Finally, we review the previous work related to ours in Section VII.

## II. PRELIMINARY

### A. Multilingual Neural Machine Translation

Multilingual neural machine translation (MNMT) aims to translate between any two languages with a unified translation model. Early studies [20] on MNMT follow a multi-task learning scheme, such as learning one-to-many translations with a shared encoder and separate decoders or many-to-one translations with separate encoders and a shared encoder. This kind of method impedes the knowledge transfer between languages and is also very inefficient in model storage when the languages scale.

Later, [1] successfully realized MNMT with a single model, achieving competitive performance with individual bilingual NMT models. It also enables zero-shot translation, which translates between language pairs unseen in the training process. For example, the training data contains non-English to English (X-En) and English to non-English (En-X) sentence pairs. The MNMT model trained on such data can conduct zero-shot translation between two non-English languages (X-X). In this paper, we follow this unified architecture to study the zero-shot translation of MNMT models.

### B. Definition of Off-Target Issue

**Off-Target Issue** is a type of translation error that commonly occurs in zero-shot translation [7]. It describes the phenomenon that MNMT models ignore the given target language information and translate the source sentence into the wrong language. Assume that $L$ denotes the set of languages involved in the MNMT model, and $T \in L$ is the target language, the off-target ratio (OTR) is calculated as:

$$\text{OTR} = \frac{\sum_{i=1}^{N} \mathbb{I}_{\hat{y}_i \neq T}}{N},$$  

(1)

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

### C. Existing Works on Off-Target Issue

A number of recent efforts have explored ways to improve zero-shot translation by mitigating the off-target issue. One thread of work focuses on modifying the model architecture, such as adding a target language-aware linear transformation between the encoder and the decoder [7] and removing residual connections in an encoder layer [12]. Another thread of work introduces auxiliary tasks with additional training losses to help the model training, such as a likelihood training objective that encourages the model to produce equivalent translations of parallel sentences in auxiliary languages [15], an auxiliary target language prediction task to regularize decoder outputs to retain information about the target language [9], and an additional denoising autoencoder objective [10]. However, the effectiveness of these methods is limited. The off-target ratios are still high according to our experiments.

Besides, researchers also try to generate synthetic data for zero-shot translation pairs in either an offline [5] or online [7] manner. However, such methods require additional computational costs in generating data and model training. Also, adding data pairs in the zero-shot translation direction could hurt the performance of supervised translation, which is known as the curse of representation bottleneck in the multilingual translation field [7].

In this paper, we will identify two uncertainties that are responsible for the off-target issues, extrinsic data uncertainty, i.e., for a portion of the training examples, the target sentences are in the source languages, and intrinsic model uncertainty, i.e., MNMT models tend to spread too much probability mass over the vocabulary of off-target languages. Based on these,
we provide simple and effective solutions, data denoising and unlikelihood training, to mitigate the data and model uncertainty, respectively.

D. Experimental Setup

In this section, we introduce the experimental setup throughout this work, including the training data, evaluation data, and model configurations.

1) Training Data: We mainly conduct experiments on three datasets across different data distributions and corpus sizes:

- **OPUS-100 Data** is an unbalanced multilingual dataset, where some language pairs have more training instances than others. [7] propose OPUS-100 which consists of 55M English-centric sentence pairs covering 100 languages. We also choose five language pairs from OPUS-100, including English-German (En-De), English-Chinese (En-Zh), English-Japanese (En-Ja), English-French (En-Fr), and English-Romanian (En-Ro) to construct balanced OPUS-6 Data (5M in total with 1M each). We follow [7] to apply BPE [21] to learn a joint vocabulary size of 64K from the whole OPUS-100 dataset.

- **WMT-6 Data** is a large-scale unbalanced dataset. Specifically, we collect the language pairs same as OPUS-6 from the widely-used WMT competition dataset [22], [23], [24], [25], [26], 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 [21] model with 32 K merge operations.

- **Validation Set:** For OPUS data, we use our original validation data provided by OPUS-100 as the validation set. For WMT data, we use WMT 19 En-De test set, WMT 19 En-Zh test set, WMT 20 En-Ja dev set, WMT14 En-Fr dev set, and WMT16 En-Ro dev set as validation set.

2) Multi-Source Test Set: To eliminate the content bias across languages [27], we evaluate the performance of multilingual translation models on the multi-source TED58 test set [28], [29], 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 translations in all six languages. Finally, we obtain 3804 sentences in six languages, i.e., 22824 sentences in total. We use the filtered test set to evaluate the performance on both supervised and zero-shot translations. We report the results of both BLEU scores [30] 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., X₁-X₂), respectively. We employ the langid library,¹ the most widely used language identification tool with 93.7% accuracy on 7 datasets 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 [31] and chrF [32] to validate our method.

3) Model: We adopt the Fairseq² toolkit for experiments. All NMT models in this paper follow the Transformer-big settings, with 6 layers, 1024 hidden sizes, 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 the encoder and target language tokens at the decoder, while T-Enc only adds target language tokens at the encoder. 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 model by the lowest loss on the validation set.

III. ANALYZING UNCERTAINTY

In this section, we present the 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 of target languages.

A. MNMT Models are Well Trained

In Table II, we list the supervised translation performance of our multilingual NMT models, which are evaluated 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 on high-resource datasets, we train a Transformer-big model with 460 K tokens per batch for 30 K updates. As for low-resource datasets, we train a Transformer-base model with 16 K tokens per batch for 50 K 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.

B. Poor Zero-Shot Performance and Off-Target Issues

Table III 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., 32.1vs. 1.9 on OPUS-6 with T-Enc). To further validate our claim, we list the detailed results in Table IV. As seen, the gap in BLEU score between supervised and zero-shot translations is highly

---

¹https://github.com/saffsd/langid.py

²https://github.com/pytorch/fairseq

---

### Table II: BLEU Scores of Bilingual and Multilingual Transformer-Big Models Trained on OPUS-6 Data for Supervised Translation

| Model          | English→X | X→English |
|----------------|-----------|-----------|
|                | OPUS      | TED       | OPUS      | TED       |
| Bilingual Model| 32.0      | 21.6      | 31.0      | 22.9      |

**Multilingual NMT Models**

|                  | OPUS | TED |
|------------------|------|-----|
| S-Enc-T-Dec      | 34.8 | 26.9 |
| T-Enc            | 34.8 | 27.1 |

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 III

| Training Data | Supervised | Zero-Shot | Supervised | Zero-Shot |
|---------------|------------|-----------|------------|-----------|
| S-Enc-T-Enc 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 IV

| Target | BLEU↑ | OTR↓ | BLEU↑ | OTR↓ |
|--------|-------|------|-------|------|
|        | Sup.  | Zero | Sup.  | Zero |
| OPUS-6 | 19.0  | 15.7 | -3.3  | 2.1  |
| Zh     | 23.1  | 11.4 | -11.7 | 32.6 |
| De     | 29.4  | 6.1  | -23.3 | 49.6 |
| Fr     | 37.1  | 6.4  | -30.7 | 68.1 |
| Ro     | 26.8  | 11.4 | -15.4 | 14.3 |
| WMT-6  | 26.1  | 18.1 | -8.0  | 2.0  |
| Zh     | 26.4  | 17.1 | -9.3  | 6.7  |
| De     | 33.2  | 8.1  | -25.1 | 41.8 |
| Fr     | 38.9  | 12.3 | -26.6 | 39.0 |
| Ro     | 24.1  | 11.0 | -13.1 | 23.0 |

C. Uncertain Prediction Causes Off-Target Issues

To investigate how MNMT models generate off-target translations, we follow [17] to analyze the model confidence in 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. Fig. 1 plots the model confidence for both references (“Refer- ence”) 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. One thing that needs to be mentioned is when calculating the per-token probability of En/De, we do not need to consider which tokens belong to English, German, or both. For each French sentence, we have one English reference sentence and one German reference sentence. Given the same French sentences as the input, we calculate the average per-token probability of generating the English/German reference sentences.

From Fig. 1 we can find that, different from the supervised translation, the zero-shot translation shows a surprisingly higher confidence in the off-target distractors. Accordingly, the model tends to generate more off-target translations (i.e., 74.9% vs. 1.6%).

In uncertainty theory [16], the uncertainties for an engineering system can come from model design, named intrinsic uncertainties, or from model calibrations using uncertain data, named extrinsic uncertainty. In the following sections, we will connect the uncertain prediction problem to the language uncertainty from both data (Section IV) and model (Section V). Based on these findings, we provide simple and effective solutions to mitigate the data and model uncertainty.

IV. EXTRINSIC DATA UNCERTAINTY

A. Problem: Data Uncertainty

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, and we assume that both the source and target sentences are in the correct languages. However, we find that a portion of training data contains off-target translations, mainly in English. We guess it is because the dataset is collected by translated corpus mainly from English, such as the TED subscripts translated from English. For some sentences, the translators don’t know how or don’t think they need to be translated, so they leave the text in English, leading to the English-English pairs. Table V 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%). A previous study on bilingual MT [17] suggests that 1% to 2% of such data noises can make the NMT model highly uncertain and tend to produce translations in the 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.

B. Solution: Data Denoising

We utilize data denoising to make sure the MNMT model learns a more correct mapping between languages from the training data. Specifically, we adopt the langid tool to identify the off-target sentence pairs in the training data of each parallel data 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 in the target languages. As a result, we can reduce the off-target ratio and improve the performance of zero-shot translation.

Table VI 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 (Section V).

V. INTRINSIC MODEL UNCERTAINTY

A. 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 (e.g., a word in the source language can correspond to multiple translations in different forms and languages [33]), preserving the possibility of decoding into a wrong language. Such 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 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”). Fig. 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.

B. Solution

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

1) Vocabulary Masking: One straightforward solution to model uncertainty is to constrain the probability distributions only on the vocabulary of the target language by masking the output logits on the off-target vocabulary. [34] proposed a target vocabulary filtering method that filters the model’s target vocabulary and embeddings to only contain the token from target languages, aiming to accelerate the inference times. Similar to their method, we filter the tokens from other languages to mitigate the off-target issues. 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 VII for more details).

This method can be applied in both training and inference. When predicting target tokens, we mask 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:

$$P_\theta(y|h_t) = \begin{cases} \frac{\exp(h_t^\top w_y)}{\sum_{y' \in V_T} \exp(h_t^\top w_{y'})}, & y \in V_T \\ 0, & \text{otherwise} \end{cases}$$

where $h_t$ is the hidden state at time step $t$, and $w_y$ is the word embedding of the token $y$ in [35], [36].
2) Unlikelihood Training: While the vocabulary masking method can successfully reduce the probabilities of translations in 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 for 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 \[37\] for MNMT, which forces the model to assign lower probabilities to unlikely generations.

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

\[
L_{\text{Likelihood}} = - \sum_{t=1}^{T} \log P_{\theta}(y|\mathbf{x}, y^{l_{c}}),
\]

where \(l_{c}\) denotes the correct language tag for the target sentence \(y\). 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:

\[
L_{\text{Unlikelihood}} = - \sum_{t=1}^{T} \log (1 - P_{\theta}(y|\mathbf{x}, y^{l_{w_{<t}}})).
\]

The final loss is the combination of the above two:

\[
L = L_{\text{Likelihood}} + \alpha L_{\text{Unlikelihood}}.
\]

In this way, we provide supervision for zero-shot directions by penalizing the off-target translations (i.e., the mismatch between the target language tag and the target sentence). We follow Welleck et al. [37] to fine-tune the pre-trained MNMT model with the combined loss for \(K\) steps.

C. Ablation Study

1) Ablation Study on Vocabulary Masking:

a) Statistics of language-specific vocabularies: For the language masking approach, we need to extract a language-specific vocabulary for each language from the full vocabulary, and different language-specific vocabularies can have shared tokens. Table VII lists the vocabulary statistics on OPUS-6 data.

| Language | En | De | Fr | Ja | Ro | Zh |
|----------|----|----|----|----|----|----|
| EN       | 17.2K | 5.7K | 16.2K | 6.0K | 0.9K | 2.0K |
| DE       | 6.0K | 5.5K | 14.9K | 9.0K | 5.0K | 1.2K |
| FR       | 0.9K | 0.8K | 0.7K | 0.3K | 2.4K | 1.2K |
| JA       | 3.8K | 4.2K | 5.0K | 3.0K | 12.1K | 15.4K |

Each number denotes the number of the shared tokens between the vocabulary of the languages in rows and columns.

b) Language masking in training or inference: As aforementioned, the proposed vocabulary masking method can be used in both training and inference. Table VIII presents the results of different masking strategies. Applying vocabulary masking to the vanilla MNMT model during inference significantly improves zero-shot translation performance by reme-\(y\)ing 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.

2) Ablation Study on Unlikelihood Training: Fig. 3 shows the impact of the interpolation weight \(\alpha\) and fine-tune step \(K\) on zero-shot translations.

For example, the size of English vocabulary is 17.2 K, which shares 5.7 K tokens with the German vocabulary.

b) Language masking in training or inference: As aforementioned, the proposed vocabulary masking method can be used in both training and inference. Table VIII presents the results of different masking strategies. Applying vocabulary masking to the vanilla MNMT model during inference significantly improves zero-shot translation performance by reme-\(y\)ing 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.

For example, the size of English vocabulary is 17.2 K, which shares 5.7 K tokens with the German vocabulary.
TABLE IX
RESULTS OF MITIGATING MODEL UNCERTAINTY FOR T-ENC MODEL ON RAW DATA WITHOUT DENOISING

| 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  |

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.

3) Comparison Results: Table IX 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 the superiority of unlikelihood training to directly penalizing simulated off-target translation during model training. In the following experiments, we use unlikelihood training to mitigate model uncertainty as default.

VI. OVERALL EMPIRICAL ASSESSMENT

In this section, we investigate whether our approaches can alleviate the off-target issue in zero-shot translation.

A. Translation Performance

Table X lists the results of both supervised and zero-shot translations. Clearly, both data denoising, vocabulary masking, and unlikelihood training can significantly improve the zero-shot translation performance in all cases. Combining 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 XI.

1) Larger-Scale Imbalanced Datasets: In addition to the small-scale balanced OPUS-6 data, we also validate our approaches on the larger-scale imbalanced datasets (i.e. 111.8M WMT-6 data with 6 languages and 55.0M OPUS-100 with 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 perform 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.

2) 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. S-Enc-S-Dec is a not widely adopted tagging manner, which encodes both source and target language in the encoder, due to its worse zero-shot translation performance [13]. In Appendix Table XIII, we also show that our proposed methods can improve the zero-shot translation performance and mitigate the off-target issues on top of the S-Enc-S-Dec model.

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.

B. Prediction Uncertainty

In this section, we present a qualitative analysis to provide insights into where our approaches reduce off-target translations.

Fig. 4 shows the prediction probabilities of 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 off-target translation, which reconfirms our claim that extrinsic data uncertainty and intrinsic model uncertainty are responsible for the uncertain prediction of target languages. Specifically, we find that data denoising reduces the confidence on the first few tokens of off-target translations noticeably, while unlikehood 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 unlkedelihood training that prevents the model from generating off-target translation. Therefore, denoising such off-target noises can further improve the performance of unlikelihood training.
C. Comparison With Previous Work

We compare our methods with three recent works on improving zero-shot translation: (1) RemoveRes. [12] that removes residual connections in an encoder layer to disentangle the positional information; (2) AE Loss [10] that introduces a denoising autoencoding loss to implicitly maximize the probability distributions for zero-shot directions; (3) Contrastive loss [38] that maximize the cosine similarly between the encoder representation of sentences of different languages. We reimplement these methods and compare the translation performance. Table XII 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.

VII. RELATED WORK

A. Improving Zero-Shot Translation

A number of recent efforts have explored ways to improve zero-shot translation by mitigating the off-target issue. One thread of work focuses on modifying the model architecture. Zhang, et al. [7] added a target language-aware linear transformation between the encoder and the decoder to enhance the freedom of multilingual NMT in expressing flexible translation relationships. Liu, et al. [12] found that the off-target issue can be alleviated by removing residual connections in an encoder layer. And Wu [13] showed that adding language tags properly to the model can enhance the consistency of semantic representations and alleviate the off-target issue. Compared with these methods, our work is orthogonal to them and mitigates the off-target issue from an uncertainty perspective without the need to modify the architecture.

Another thread of work introduces auxiliary tasks with additional training losses to help the model training. For example, Al-Shedivat et al. [15] proposed an agreement-based likelihood training objective that encourages the model to produce equivalent translations of parallel sentences in auxiliary languages. Yang et al. [9] leveraged an auxiliary target language prediction...
task to regularize decoder outputs to retain information about the target language. And Wang et al. [10] introduced an additional denoising autoencoder objective into the traditional training objective to improve the translation accuracy on zero-shot directions. Our work proposes a novel and lightweight method to directly reduce the off-target translation via unlabeled training.

Besides, researchers also try to generate synthetic data for zero-shot translation pairs in either an off-line [5] or on-line [7] manner. However, such methods require additional computational costs in generating data and model training. Also, adding data pairs in the zero-shot translation direction could hurt the performance of supervised translation, which is known as the curse of representation bottleneck in the multilingual translation field [7]. Our work also tackles the off-target issue from a data perspective, but we remove the off-target noises in the original data rather than leveraging additional data, which is not a common practice in MNMT.

Recently, a concurrence work [11] also proposed a method to manipulate the distribution of the multilingual vocabulary, aiming to reduce the off-target issues in zero-shot translation. However, the motivation and method are different from this paper. [11] contributes the off-target issues to the closer lexical distance between two languages’ vocabulary and tries to enlarge the lexical distance by adding language-specific tokens to the shared tokens. However, according to our experiments, even between two languages that already have large lexical distances, e.g., German and Chinese, we still observe a high off-target ratio. Our paper contributes the off-target issues to the data uncertainty and model uncertainty, and then proposes several methods to mitigate the uncertainties. Our vocabulary masking method, which masks the tokens of other languages when generating a specific language, aims to reduce the probability assigned to the tokens from other languages. Hence, the motivation and specific operation of our paper are different from [11].

### B. Uncertainty in NMT

Closely related to our work, Ott et al. [17] 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, for understanding and mitigating the off-target issues. Besides data uncertainty, we also reveal the intrinsic model uncertainty on the output distributions due to the shared vocabulary across multiple languages.

Recently, Chen et al. [39] proposed Masked Label Smoothing, which masks the soft label probability of source-side words to zero. The operation is similar to our vocabulary masking method. However, they focus on supervised NMT for better calibration and mitigating copy behavior, while we focus on zero-shot translation to mitigate the off-target issues. In addition, our 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.

### VIII. 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.

In the future, we plan to explore the uncertainty of large MNMT models trained on more complicated datasets [40], [41], as well as that of large language models (LLMs), which also support the zero-shot translation [42].

### APPENDIX

#### TABLE XIII

| Model | Supervised | Zero-Shot |
|-------|------------|-----------|
|       | BLEU↑ | OTR↑ | BLEU↑ | OTR↑ |
| OPUS-6 Data |     |     |     |     |
| Vanilla                                  | 27.1 | 2.0 | 8.5  | 43.7 |
| + Data Denoise                           | 27.1 | 1.8 | 10.3 | 30.9 |
| + Vocab Mask                             | 27.0 | 1.9 | 7.7  | 18.8 |
| + Unlike Train                           | 27.1 | 1.9 | 14.9 | 3.8  |
| WMT-6 Data                               |     |     |     |     |
| Vanilla                                  | 28.6 | 1.7 | 7.8  | 44.2 |
| + Data Denoise                           | 28.6 | 1.7 | 9.4  | 35.4 |
| + Vocab Mask                             | 28.5 | 1.7 | 13.9 | 12.1 |
| + Unlike Train                           | 28.6 | 1.7 | 15.3 | 5.0  |

#### REFERENCES

[1] M. Johnson et al., “Google’s multilingual neural machine translation system: Enabling zero-shot translation,” Trans. Assoc. Comput. Linguistics, vol. 5, pp. 339–351, 2017.

[2] R. Aharoni, M. Johnson, and O. Firat, “Massively multilingual neural machine translation,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2019, pp. 3874–3884.

[3] Q. Wang, J. Zhang, and C. Zong, “Synchronous inference for multilingual neural machine translation,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 30, pp. 1827–1839, 2022.

[4] W. Jiao, Z. Tu, J. Li, W. Wang, J.-T. Huang, and S. Shi, “Tencent’s multilingual machine translation system for WMT22 large-scale African languages,” in Proc. 7th Conf. Workshop Mach. Transl., 2022, pp. 1049–1056.

[5] J. Gu, Y. Wang, K. Cho, and V. Li, “Improved zero-shot neural machine translation via ignoring spurious correlations,” in Proc. Assoc. Comput. Linguistics, 2019, pp. 1258–1268.

[6] Y. Hou, W. Jiao, M. Liu, Z. Tu, C. Allen, and M. Sachan, “Adapters for enhanced modeling of multilingual knowledge and text,” in Proc. Conf.: Findings Assoc. Comput. Linguistic, 2022, pp. 3902–3917.

[7] B. Zhang, P. Williams, I. Titov, and R. Sennrich, “Improving massively multilingual neural machine translation and zero-shot translation,” in Proc. Ann. Conf. Assoc. Comput. Linguistics, 2020, pp. 1628–1639.

[8] Y. Tang et al., “Multilingual translation with extensible multilingual pre-training and finetuning,” in Proc. Findings Assoc. Comput. Linguistics, 2021, pp. 3450–3466.

[9] Y. Yang, A. Eriguchi, A. Muzio, P. Tadepalli, S. Lee, and H. Hassan, “Improving multilingual translation by representation and gradient regularization,” in Proc. 2021 Conf. Empirical Methods Natural Lang. Process., 2021, pp. 7266–7279.
WANG et al.: UNDERSTANDING AND MITIGATING THE UNCERTAINTY IN ZERO-SHOT TRANSLATION 4903

[10] W. Wang, Z. Zhang, Y. Du, B. Chen, J. Xie, and W. Luo, “Rethinking zero-shot neural machine translation: From a perspective of latent variables,” in Proc. Conf. Empir. Methods Natural Lang. Process. (Findings), 2021, pp. 4321–4327.

[11] J. Chen, S. Ma, D. Zhang, F. Wei, and B. Chang, “On the off-target problem of zero-shot multilingual neural machine translation,” in Proc. Findings Assoc. Comput. Linguistics, Jul. 2023, pp. 9542–9558.

[12] D. Liu, J. Niehues, J. Cross, F. Guzmán, and X. Li, “Improving zero-shot disentanglement by dissentrating positional information,” in Proc. 39th Annua. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., 2021, pp. 1259–1273.

[13] L. Wu, S. Cheng, M. Wang, and L. Li, “Language tags matter for zero-shot neural machine translation,” in Proc. Findings Assoc. Comput. Linguistics, 2021, pp. 3001–3007.

[14] N. Arivazhagan et al., “Massively multilingual neural machine translation in the wild: Findings and challenges,” 2019, arXiv:1907.05019.

[15] M. Al-Shedivat and A. P. Parikh, “Consistency by agreement in zero-shot neural machine translation,” in Proc. 2019 Conf. North, 2019, pp. 1184–1197.

[16] G. Steck, F. C. Hurlbut, and D. A. Dornfeld, “Experimentation and uncertainty analysis for engineers,” 1989. [Online]. Available: https://api.semanticscholar.org/CorpusID:113632031

[17] M. Ott, M. Auli, D. Grangier, and M. Ranzato, “Analyzing uncertainty in neural machine translation,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 3956–3965.

[18] W. Jiao, X. Wang, S. He, I. King, M. Lyu, and Z. Tu, “Data rejuvenation: Exploiting inactive training examples for neural machine translation,” in Proc. 2020 Conf. Empirical Methods Natural Lang. Process., 2020, pp. 2255–2266.

[19] W. Jiao, X. Wang, S. He, Z. Tu, I. King, and M. R. Lyu, “Exploiting inactive examples for natural language generation with data rejuvenation,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 30, pp. 931–943, 2022.

[20] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, “Multi-task learning for multiple language translation,” in Proc. 53rd Assoc. Comput. Linguistics, 2015, pp. 1723–1732.

[21] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in Proc. Assoc. Comput. Linguistics, 2016, pp. 1715–1725.

[22] R. Sennrich and B. Haddow, “Edinburgh neural machine translation systems for WMT’16,” in Proc. Conf. Workshop Mach. Transl., 2016, pp. 371–376.

[23] K. Chen, R. Wang, M. Utiyama, and E. Sumita, “Integrating prior translation knowledge into neural machine translation,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 30, pp. 330–339, 2022.

[24] K. Chen et al., “Towards more diverse input representation for neural machine translation,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 1586–1597, 2020.

[25] X. Liu, D. F. Wong, L. S. Chao, and Y. Liu, “Latent attribute based hierarchical decoder for neural machine translation,” IEEE/ACM Trans. Audio,Speech, Lang. Process., vol. 27, no. 12, pp. 2103–2112, Dec. 2019.

[26] J. Yang et al., “GTrans: Grouping and fusing transformer layers for neural machine translation,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 31, pp. 1489–1498, 2022.

[27] S. Wang, Z. Tu, Z. Tan, S. Shi, M. Sun, and Y. Liu, “On the language coverage bias for neural machine translation,” in Proc. Findings Assoc. Comput. Linguistics, 2021, pp. 4778–4790.

[28] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig, “When and why are pre-trained word embeddings useful for neural machine translation?” in Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics, 2018, pp. 529–535.

[29] C. Tran, Y. Tang, X. Li, and J. Gu, “Cross-lingual retrieval for iterative self-supervised training,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 2207–2219.

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 311–318.

[31] R. Rei, C. A. Stewart, A. C. Farinha, and A. Lavie, “COMET: A neural framework for MT evaluation,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2020, pp. 2685–2702.

[32] M. Popovic, “IchrF: Character N-gram F-score for automatic MT evaluation,” in Proc. 10th Workshop Statist. Mach. Transl., 2015, pp. 392–395.

[33] W. Jiao, X. Wang, Z. Tu, S. Shi, M. Lyu, and I. King, “Self-training sampling with monolingual uncertainty for neural machine translation,” in Proc. 59th Annua. Meeting Assoc. Comput.Linguistics 11th Int. Joint Conf. Natural Lang. Process., 2021, pp. 2840–2850.

[34] A. Bérand, D. Lee, S. Clinchant, K. W. Jung, and V. Nikouliina, “Efficient inference for multilingual neural machine translation,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2021.

[35] A. Vaswani et al., “Attention is all you need,” in Proc. Neural Inf. Process. Syst., 2017, pp. 6000–6010. [Online]. Available: https://api.semanticscholar.org/CorpusID:13756489

[36] O. Press and L. Wolf, “Using the output embedding to improve language models,” in Proc. Eur. Conf. Field Comput. Linguistics, 2016, pp. 157–163.

[37] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, “Neural text generation with unlikelihood training,” in Proc. Int. Conf. Learn. Representations, 2019.

[38] X. Pan, M. Wang, L. Wu, and L. Li, “Contrastive learning for many-to-many multilingual neural machine translation,” in Proc. Assoc. Comput. Linguistics, 2021, pp. 244–258.

[39] L. Chen, R. Xu, and B. Chang, “Focus on the target’s vocabulary: Masked label smoothing for machine translation,” in Proc. Assoc. Comput. Linguistics, 2022, pp. 669–676.

[40] A. Fan et al., “Beyond english-centric multilingual machine translation,” J. Mach. Learn. Res., vol. 22, pp. 107:1–107:48, 2021.

[41] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, and F. Gazmán, “WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from wikipedia,” in Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics: Main Vol. Apr. 2021, pp. 1351–1361. [Online]. Available: https://aclanthology.org/2021.eacl-main.115

[42] W. Jiao, W. Wang, J.-T. Huang, X. Wang, and Z. Tu, “Is ChatGPT a good translator? A preliminary study,” 2023, arXiv:2301.08745.

Wenxiang Jiao received the bachelor’s and M.Phil. degrees from Nanjing University, Nanjing, China, in 2015 and 2017, respectively, and the Ph.D. degree from the Chinese University of Hong Kong, Hong Kong, advised by Prof. Michael R. Lyu. His research interests include reliability of natural language processing models and software, such as machine translation models and pre-training language models.

Wenxuan Wang received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2017. He is currently working toward the Ph.D. degree with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, advised by Prof. Michael R. Lyu. His research interests include conversational emotion recognition, neural machine translation, and multilingual pretraining, and has published papers in top-tier conferences and journals such as ACL, EMNLP, NAACL, AAAI, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He was the recipient of the 1st place at WMT 2022 Large-Scale Machine Translation Evaluation for African Languages (Constrained Track), and the Best Paper Award of Multilingual Representation Learning Workshop in EMNLP 2022.

Shuo Wang received the B.S. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2018. He is currently working toward the Ph.D. (fifth-year) degree. He is advised by Prof. Yang Liu. His research interests include trustworthy neural machine translation and multilingual natural language processing.
Zhaopeng Tu received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2013. He is currently a Principal Researcher with the Tencent AI Lab, Shenzhen, China. He was a Postdoctoral Researcher with the University of California at Davis, Davis, CA, USA, from 2013 to 2014. He was a Researcher with Huawei Noah’s Ark Lab, Hong Kong, from 2014 to 2017. He is currently working on neural machine translation and Seq2Seq learning for other NLP tasks, such as dialogue and question answering.

Michael R. Lyu (Life Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981, the M.S. degree in computer science from the University of California, Santa Barbara, Santa Barbara, CA, USA, in 1985, and the Ph.D. degree in computer science from the University of California, Los Angeles, Los Angeles, CA, in 1988. He is currently a Professor with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. His research interests include software engineering, dependable computing, distributed systems, cloud computing, mobile networking, Big Data, and machine learning. He was elected to IEEE Fellow in 2004, AAAS Fellow in 2007, and ACM Fellow in 2015 for his contributions to software reliability engineering and software fault tolerance.