On the Off-Target Problem of Zero-Shot Multilingual Neural Machine Translation

Anonymous ARR submission

Abstract

While multilingual neural machine translation has achieved great success, it suffers from the off-target issue, where the translation is in the wrong language. This problem is more pronounced on zero-shot translation tasks. In this work, we explore the major cause of the off-target problem and find that a closer lexical distance (i.e., KL-divergence) between two languages' vocabularies leads to a higher off-target rate. Motivated by the finding, we propose LAVS, a simple and effective algorithm to construct the multilingual vocabulary, that greatly alleviates the off-target problem of the translation model by increasing the KL-divergence between languages. We conduct experiments on a multilingual machine translation benchmark in 11 languages. Experiments show that the off-target rate for 81 translation tasks is reduced from 29% to 8%, while the overall BLEU score is improved by an average of 1.9 points.[1]

1 Introduction

Multilingual NMT makes it possible to do the translation among multiple languages using only one model, even for zero-shot directions (Johnson et al., 2017; Aharoni et al., 2019). It has been gaining increasing attention since it provides insights for multilinguality studies and greatly reduces the MT system's deployment cost. Despite its success, the off-target phenomenon is a harsh and widespread problem in the existing multilingual models. For the zero-shot translation directions, MT system translates the source sentence to a wrong language, which severely degrades the system's credibility. As shown in Figure 1, the off-target rate could be up to nearly 45% for high-resource languages and even up to 95% for low-resource languages.

Researchers have been noticing and working on solving the problem from different perspectives like data augmentation (Gu et al., 2019; Zhang et al., 2020) and regularization (Yang et al., 2021). While most of the existing work focus on addressing the problem by improving the data or the optimization, the importance of vocabulary, which reflects the token distribution among languages, is often neglected.

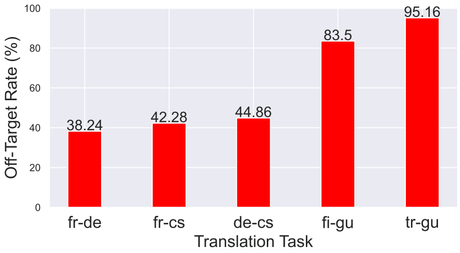


Figure 1: Off-target rate of some directions tested on Flores-101 dataset. The off-target problem is widespread with a maximum of 95.16% error translation language for low-resource language (tr, gu) pairs and 44.86% for high-resource language (fr,de,cs,fi) pairs.

In this work, we perform a comprehensive analysis of the off-target problem, finding that the off-target rate is positively related to the proximity of the language pair. We quantify the proximity within language pairs using KL-divergence between token distribution. It turns out that translation direction with lower KL divergence is related to a higher off-target rate and the correlation coefficient could be as high as -0.92.

A simple solution by separating the vocabulary of different languages can greatly increase the KL divergence between languages. Although it proves to improve the zero-shot translation performance, it also greatly increases the model size and costs the cross-lingual transferability.

To address these problems, we propose Language-Aware Vocabulary Sharing (LAVS), a novel algorithm to construct the multilingual vocabulary that increases the KL-divergence of token distributions among languages while preserving

[1] We will release the code for reproducibility.

		Source										
	OTR	cs	fr	de	fi	lv	et	ro	hi	tr	gu	AVG
Target	cs		43%	45%	21%	13%	11%	13%	12%	10%	33%	22%
	fr	20%		30%	22%	18%	21%	12%	10%	15%	12%	18%
	de	15%	38%		19%	13%	16%	14%	36%	28%	36%	23%
	fi	14%	32%	28%		12%	9%	13%	44%	19%	64%	26%
	lv	8%	34%	24%	7%		5%	10%	33%	19%	58%	22%
	et	16%	32%	15%	8%	15%		23%	47%	23%	74%	28%
	ro	2%	2%	3%	2%	0%	3%		10%	8%	50%	9%
	hi	15%	13%	6%	13%	20%	14%	16%		54%	78%	25%
	tr	2%	1%	0%	1%	0%	1%	18%	33%		70%	14%
	gu	77%	60%	53%	84%	80%	74%	80%	92%	95%		77%
	AVG	19%	28%	23%	19%	19%	17%	22%	35%	30%	53%	**29%**

Table 1: Zero-shot off-target rate of the baseline model. While the average OTR of supervised directions is about 0%, the average OTR of 81 zero-shot directions increases to 29%.

An Off-Target Case
Direction: FR -> DE
Input: <DE> Un sondage effectué auprès de 1 400 personnes avant les élections fédérales de 2010 a révélé que le nombre d'opposants à la transformation de l'Australie en république avait augmenté de 8 % depuis 2008.
Output: A survey of 1400 people prior to the 2010 federal elections revealed that the number of opponents of Australia's transformation into a republic had increased by 8 % since 2008.
Gold: Von den 1.400 Personen, die vor den Bundeswahlen 2010 befragt wurden, hat der Anteil derjenigen, die sich dagegen aussprechen, dass Australien zur Republik wird, seit 2008 um 8 Prozent zugenommen.

Figure 2: A real Off-Target case observed in our multilingual NMT system. In this case, the output is literally English while the real target is German.

the cross-lingual transferability. It is simple and can be applied to any existing multilingual translation model without introducing any extra data or parameters. Our empirical experiments prove that LAVS reduces the off-target rate from 29% to 8% and improves the BLEU score by 1.9 points on the average of 81 translation directions.

2 Delving into the Off-Target Problem

In this section, we start by briefly introducing our baseline multilingual NMT system and analyze the result of off-target phenomena. Then, we explore the causes of the off-target problem and reveal its relation to language vocabulary.

2.1 Multilingual NMT System

We adopt the Transformer-Big (Vaswani et al., 2017) model as the baseline model. For multilingual translation, we add a target language identifier <XX> at the beginning of input tokens to combine direction information. We train the model on an English-centric dataset WMT'10 (Callison-Burch et al., 2010). Zero-shot translation performance is evaluated on Flores-101 (Goyal et al., 2021) dataset. We use a public language detector[2] to identify the sentence-level language and compute the off-target rate (OTR) which denotes the ratio of translation that deviates to wrong languages. Full information about training can be found in Section 4.1.

2.2 Off-Target Statistics Safari

Off-Target Rate Differs in Directions We first train the multilingual NMT model in 10 EN-X directions and 10 inverse directions from WMT'10 simultaneously. Then we test the model on 81 X-Y zero-shot directions using semantic parallel sentences from the previous 10 languages provided by Flores-101. We compute the off-target rate of all directions and list the result in Table 1.

In addition to the individual score, we next split the languages into High (cs, fr, de, fi, >5M), Mid (lv, et, 1M-5M), and Low (ro, tr, hi, gu, <1M) resources according to data abundance degree. Then we compute the average OTR of High-to-High, High-to-Low, Low-to-High, and Low-to-Low directions and rank the result. The ranked result is: Low-to-Low (50.28%) > High-to-High (27.16%) > Low-to-High (23.18%) > High-to-Low (20.78%). Based on the observation, we can see that language with the lowest resource (gu) contributes to a large portion of off-target cases. This is reasonable since the model might not be familiar with the language identifier <GU> and the same situation goes for Low-to-Low translations.

The Hidden Reason for Off-Target However, it is surprising to see that translations between high-resource languages suffer from more severe off-target than those directions involving one low-resource language. There seem to be other factors influencing the off-target phenomena.

In other words, if data imbalance is not the key factor for off-targets between high-resource languages, what are the real reasons and possible solutions? To answer these questions, we need to delve deeper into the real off-target cases.

2.3 The Major Symptom of Off-Target

When the model encounters an off-target issue, a natural question is which language the model most possibly deviates to. We find that among different directions, a majority(77%) of the off-target cases are wrongly translated to English, which is the

[2] https://github.com/Mimino666/langdetect

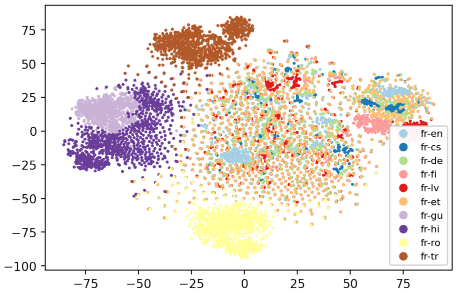


Figure 3: Encoder pooled output visualization using TSNE for French-to-Many translations. The input French sentences are the same for all directions. Note that there are only French sentences in the encoder side.

centric language in the dataset. It raises our interest that why most off-target cases deviate to English.

2.4 Failing in Encoding Discriminative Target Language Signal Leads to Off-Target

Considering the encoder-decoder structure of the model, we have one hypothesis for a possible reason for off-target: *The encoder fails to encode discriminative target language information to the hidden representations before passing to the decoder.*

To test the hypothesis, we start by analyzing the output of the transformer's encoder trained on the WMT'10 dataset.

1) We choose French as the source language and conduct a French-to-Many translation (including all languages in WMT'10) on Flores-101.

2) We collect all the pooled encoder output representations of the French-to-Many translation and project them to 2D space using TSNE. The visualization result is shown in Figure 3.

The visualization result justifies our hypothesis. We can tell from the distribution that only representations belonging to "fr-tr" and "fr-ro" directions have tight cluster structures with boundaries. *The representations from high/mid-resource language pairs are completely in chaos and they are also mixed with fr-en representations.* And those languages generally have a higher off-target rate in French-to-Many Translation according to Table 1.

The decoder cannot distinguish the target language signal from the encoder's output when it receives representations from the "chaos" area. Moreover, during the training process, the decoder generates English far more frequently than other languages and it allocates a higher prior for English.

The above two factors could cause that passing hidden representation similar to English one will possibly confuse the decoder to generate English no matter what the given target language is. It could explain the relatively high off-target rate in H-H directions and why most cases deviate to English.

Now we have a key clue for the off-target issue. The left question is ***what causes the degradation of target language signal in some directions*** and whether we can make the representations of different target languages more discriminative to eliminate the off-target cases.

2.5 Language Proximity Correlates with Zero-Shot Off-Target Rate

To explore how off-target occurs differently in different language pairs, we conduct experiments using a balanced subset of WMT'10 dataset where we hope to preclude the influence of data size. We randomly sampled 500k sentences from different directions to form a balanced training set and remove the directions(hi, tr and gu) that do not have enough sentences.

Language Proximity is an Important Characteristic of Translation Direction Languages themselves have different relations. For example, German and English are more close because they both belong to Germanic language and we hope to find the relation between inner-characteristic of a certain language pair and its off-target rate.

Token Distribution Similarity Reflects Language Proximity Our motivation is quite intuitive that if two languages are rather close, the probability distribution of different n-grams in the two languages' tokenized corpus should be nearly identical. Considering a large number of different n-grams in the corpus that burdens computing, we only consider 1-grams to compute the distribution. We call the result "Token Distribution."

We use Kullback–Leibler divergence from Token Distribution of Language B to Language A to reflect the degree of difficulty[3] if we hope to encode sentence from B using A.

$$D_{\mathrm{KL}}(A\|B) = \sum_{x\in\mathcal{V}} A(x)\log\left(\frac{A(x)}{B(x)}\right) \quad (1)$$

[3] In information theory, a simple interpretation of the KL divergence from B to A is the expected self-information increment from using A as a model when the actual distribution is B. We need more extra information if A is less similar to B. This amount of extra information is equivalent to our definition of degree of difficulty.

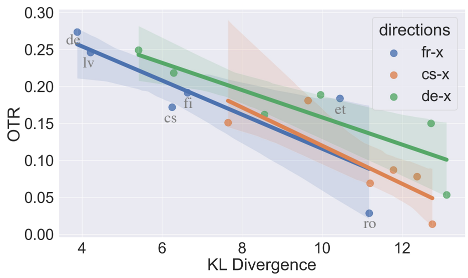


Figure 4: Scatter plot of off-target rate and KL-divergence for different language pairs. We draw the linear regression result with 95% confidence interval.

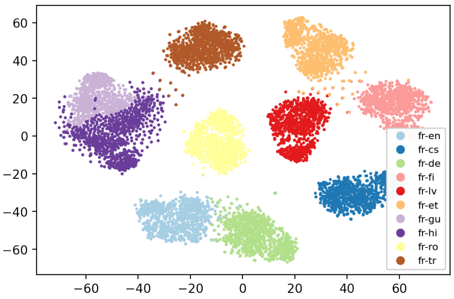


Figure 5: Encoder pooled output visualization using TSNE for French-to-Many translation using separate vocab. The result is comparable to Figure 3, which shows result with shared vocab.

Method	Size	OTR	BLEU
Vocab Sharing	308M	29%	10.2
Separate Vocab (Dec)	515M	**5%**	**12.4**
Separate Vocab (Enc,Dec)	722M	84%	2.1

Table 2: Average zero-shot result for models with different vocab. (Dec) means only the decoder uses the separate vocab. (Enc,Dec) means both the encoder and the decoder use the separate vocab.

where $\mathcal{V}$ denotes the shared vocabulary, $A(x)$ is the probability of token x in language A. To avoid zero probability during computing Token Distribution, we add 1 to the frequency of all tokens in the vocabulary as a smoothing factor.

Lower KL Divergence is related to Higher Off-Target Rate We compute the KL divergence between language pairs with the training data. After training on the balanced dataset, a zero-shot translation experiment is conducted on the Flores-101 dataset. We collect the result of French-to-Many, German-to-Many, and Czech-to-Many for analysis.

As shown in Figure 4, we can observe from the statistics that language proximity is highly related to the off-target rate. The Pearson correlation coefficients between the off-target rate and the KL-Divergence from target to source of the three x-to-many translations[4] are -0.75. -0.9. and -0.92. It indicates that language pair which has lower KL-Divergence from target to source has a higher chance to encounter off-target than those language pairs which has less similar languages.

It further implies that language proximity is one hidden reason other than data balance for off-target, which means we cannot avoid off-target solely with data balancing methods.

To better justify our finding, we involve a high-resource non-alphabet language Chinese to the training. We randomly extract 10M Chinese-English sentence pairs from WMT'19 dataset and add them to the WMT'10 training set. We train a new model on the combined dataset with the same configuration in section 2.1.

Zero-shot translation is also conducted on Flores-101. It turns out that directions involving Chinese have the lowest average off-target rate(9%) compared to other high-resource languages(fr: 33% cs: 29% de: 28% fi: 31%). This result further proves our findings that language proximity is an important factor influencing off-target since Chinese almost has no vocab overlap with other languages.

2.6 Separating Vocab of Different Languages is Effective yet Expensive

Based on the previous conclusion, we now have an idea that maybe we can ease the off-target problem by raising the KL divergence between languages. However, the token distribution is fixed when the tokenization process is done. In other words, the tokenization model and vocabulary directly influence the token distribution.

When building the vocabulary, current multilingual NMT studies tend to regard all languages as one and learn a unified sub-word-based tokenization model. We argue that this may lead to low divergence of token distribution since many sub-words are shared across languages.

There is an easy method to increase the KL divergence without changing the tokenization. We can separate the vocab of different languages as shown in Figure 9 from Appendix. Under such condition, no two languages share the same token.

As shown in Table 2, with separate decoder vocab the average off-target rate in 81 directions is

[4] To ensure robustness, we resample the datasets for 5 times and give the average results.

reduced from 29% to 5% and the BLEU score is raised from 10.2 to 12.4. We conduct the same probing experiment on encoder representation with the original WMT'10 dataset. As shown in Figure 5, representations for different target are divided. The "chaos" area does not exist anymore. We think a possible explanation for the drop down in OTR is that, the model is more sensitive to the language identifier during decoding when each output language has individual tokens.

We also train the model with separated encoder&decoder vocab and finds it suffers from worse zero-shot performance compared to baseline. We think that without any vocabulary sharing among languages, the model will learn a "spurious correlation" between input language and output language and ignore the target language identifier during the English-centric training process.

Though achieving great improvement in zero-shot translation's performance, there is a problem that cannot be ignored with the current method. When the number of languages arises, keeping isolating all vocabulary will be really parameter-consuming. In fact, in our experiment, the number of parameters increases from 308M to 515M.

3 Language-Aware Vocabulary Sharing

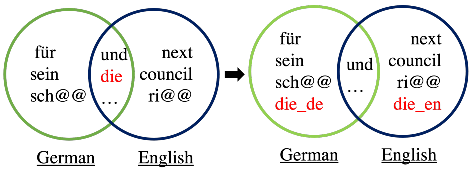


Figure 6: Illustration of LAVS. Tokens with higher shared frequency are split into language-specific ones.

We propose to deal with off-target in a parameter-efficient way. We start by introducing the methods, defining the optimization objective, and propose a greedy-selection algorithm to address the problem.

3.1 Adding Language-Specific Tokens

Based on previous observation, language pairs that have low vocabulary KL Divergence tend to encounter off-target during zero-shot translation. Thus our goal is to increase the vocabulary KL Divergence between languages. We can achieve it without changing the original tokenizer by splitting the shared tokens into language-specific ones.

As shown in Figure 6, instead of splitting all shared tokens, we can choose specfic tokens to split. After decoding, we could simply remove all language-specific tags to restore the literal output sentence. By adding language-specific tokens, the number of shared tokens between different languages decreases and makes the token distribution more different thus increasing the KL Divergence.

Algorithm 1 Language-Aware Vocabulary Sharing

Input: Shared vocabulary set V', language list L, language's token distributions P and the number of extra language-specific tokens N.
Output: V_{out} is the output vocabulary set.

```
MaxFreqs = PriorQueue(length=$N$)  ▷ queue that ranks the input elements E from high to low based on E[0].
for $i$ in $V'$ do
    for $m$ in $L$, $n$ in $L$ do
        if $m < n$ then
            freq = min($P_m^{V'}(i)$,$P_n^{V'}(i)$)
            MaxFreqs.add([freq,$m$,$n$,$i$])
$V_{out}$ = $V'$
for T in MaxFreqs do
    $m, n, i$ = T[1], T[2], T[3]
    $V_{out}$ = $V_{out} \cup (V'[i], L[m]) \cup (V'[i], L[n])$
return $V_{out}$
```

3.2 Optimization Goal

Given original vocab set V' and language list L, we aim at creating new vocab V to maximize the average KL divergence within each language pair under the new vocabulary with the restriction of adding N new language-specific tokens. Thus, our objective becomes:

$$V^* = \arg\max_V \frac{1}{|L|^2} \sum_{m \in L} \sum_{n \in L} D_{KL}(P_m^V || P_n^V)$$
$$s.t. \quad V' \subseteq V, \quad |V| - |V'| = N \tag{2}$$

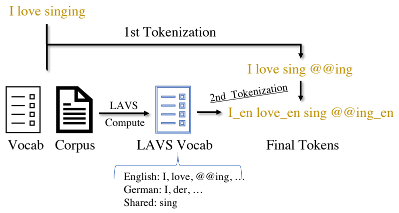
where P_m^V denotes the m-th language's token distribution on vocabulary V, add-one smoothing is applied to avoid zero probability. It is a combinatorial optimization problem. The searching space of V has an astronomical size of $C_{|V'| \cdot |L|}^N$.

3.3 Separating Tokens by Frequency

We start from only two languages J and Q and compute KL-divergence's change if we only split one shared token to two language-specific tokens.

$$\begin{aligned}\Delta D_{KL}^i &= -J(i)log\frac{J(i)}{Q(i)} - Q(i)log\frac{Q(i)}{J(i)} + \lambda \\ &= [J(i) - Q(i)]log\frac{Q(i)}{J(i)} + \lambda\end{aligned} \tag{3}$$

where we will have two i-th tokens for the different languages from the original vocabulary. λ is the



Input Sentence: I love singing.
Sentencepiece Tokenization: I love sing @@ing .
LAVS Tokenization: I_en love_en sing @@ing_en .
Training/Inference
LAVS Detokenization: I love sing @@ing .
Sentencepiece Detokenization : I love singing.

Figure 7: Illustration of tokenization and detokenization process with Language-Aware Vocabulary Sharing.

smoothing factor that can be seen as a constant. According to equation 3, splitting token that has more similar occurrence probability in the two languages will lead to higher increment in language's KL-Divergence. Also considering the fact that the tokens with high frequency influence the training process much more than the near-zero ones, we should first split the tokens that appear in *two or more* languages with similar *high frequency*.

3.4 Greedy Selection Algorithm that Maximizes Divergence Increment

Based on the previous discussion, we propose the Language-Aware Vocabulary Sharing algorithm as listed in Algorithm 1 to add language-specific tokens. First, we adopt a prior queue to keep the token candidates. Second, for each token in the shared vocabulary, we compute the shared token frequency in each language pair and add the (frequency, languageA, languageB, token) tuple to the queue. Last, since the queue ranks the elements by frequency, we create language-specific tokens for the top N tuples and return the new vocab.

Figure 7 illustrates the whole tokenization process with LAVS. In practice, given an original shared vocab with M tokens, we can always first learn a vocab with $M - N$ tokens and conduct LAVS to add N language-specific tokens to maintain the vocab size M unchanged.

4 Experiments

4.1 Datasets

Following Wang et al. (2020), we collect WMT'10 datasets for training. The devtest split of Flores-101 is used to conduct evaluation. Full information of datasets is in Appendix C.

4.2 Vocabulary Building

Vocab Sharing We adopt Sentencepiece (Kudo and Richardson, 2018) as the tokenization model. We randomly sample 10M examples from the training corpus with a temperature of 5(Arivazhagan et al., 2019) on different directions and learn a shared vocabulary of 64k tokens.

Separate Vocab Based on the sharing vocab of the baseline model, we separate the vocab of each language forming a 266k vocab.

LAVS We first learn a 54k vocabulary using the same method as the baseline model's and add 10k language-specific tokens using LAVS.

4.3 Training Details of MNMT

Architecture We use the Transformer-big model (Vaswani et al., 2017) implemented by fairseq (Ott et al., 2019) with $d_{model} = 1024$, $d_{hidden} = 4096$, $n_{heads} = 16$, $n_{layers} = 6$. We add a target language identifier <XX> at the beginning of input tokens to indicate the translation directions as suggested by Wu et al. (2021).

Optimization We train the models using Adam (Kingma and Ba, 2015), with a total batch size of 524,288 tokens for 100k steps in all experiments on 8 Tesla V100 GPUs. The sampling temperature, learning rate and warmup steps are set to 5, 3e-4 and 4000.

Evaluation We report detokenized BLEU using sacrebleu[5]. We also report the Off-Target rate with language detector[6] and conduct model-based evaluation using Bert-Score[7] (Zhang* et al., 2020).

4.4 Results

LAVS improves zero-shot translation by a large margin. Table 3 and 4 list the overall results on both zero-shot and supervised directions. According to Table 3, we can see that LAVS improves all the x-to-many and many-to-x directions with a maximum average improvement of -61.6% OTR, +3.7 BLEU and +0.036 Bert-Score compared to the baseline vocab. It gains an average of -21% OTR, +1.9 BLEU and +0.02 Bert-Score improvement on 81 zero-shot directions. Compared with the Separate Vocab (Dec) method which also leads to significant improvement in x-y directions, LAVS does not increase any model size.

[5] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0
[6] https://github.com/Mimino666/langdetect
[7] https://github.com/Tiiiger/bert_score

Method	Size	Zero-Shot Off-Target Rate					BLEU Score						
		x-y	H-H	L-L	H-L	L-H	x-y	H-H	L-L	H-L	L-H	en-x	x-en
Vocab Sharing	308M	29%	27%	50%	21%	23%	10.2	11.26	5.03	9.18	9.95	24.8	30.2
Separate Vocab (Dec)	515M	**5%**	4%	19%	**1%**	**1%**	12.4	14.69	6.54	**10.10**	**12.22**	24.6	**30.5**
LAVS (Enc, Dec)	308M	12%	**3%**	33%	13%	6%	**12.5**	**15.90**	6.26	9.91	12.14	24.8	30.3
LAVS (Dec)	308M	8%	13%	**14%**	3%	4%	12.1	13.33	**7.81**	9.80	12.01	**24.9**	30.3

Table 3: Overall performance comparison. x-y denotes all zero-shot directions. H and L denotes High/Low-resources. All evaluation are done with Flores-101 dataset. (Dec) suggests vocab only changes in decoder and (Enc, Dec) suggests changing in both encoder and decoder. LAVS outperforms baseline in zero-shot setting on both BLEU and OTR by a large margin while maintaining the en-x and x-en performance.

Metric	Method	cs-x	fr-x	de-x	fi-x	lv-x	et-x	ro-x	hi-x	tr-x	gu-x
OTR	Vocab Sharing	18.8%	28.3%	22.6%	19.5%	19.2%	17.1%	22.0%	35.2%	30.1%	52.8%
	LAVS(Dec)	**4.2%**	**14.4%**	**11.5%**	**6.2%**	**3.7%**	**4.7%**	**2.9%**	**9.7%**	**10.2%**	**6.1%**
	Δ ↓	-14.6%	-13.9%	-11.1%	-13.3%	-15.5%	-12.4%	-19.1%	-25.5%	-19.9%	-46.7%
BLEU	Vocab Sharing	10.9	10.5	11.3	9.0	9.4	10.0	11.7	6.9	7.3	4.7
	LAVS(Dec)	**12.0**	**12.0**	**12.2**	**9.6**	**10.9**	**11.0**	**14.0**	**9.3**	**9.1**	**8.4**
	Δ ↑	+1.1	+1.5	+0.9	+0.6	+1.5	+1.0	+2.3	+2.4	+1.8	+3.7
BERT Score	Vocab Sharing	0.781	0.808	0.787	0.766	0.783	0.774	0.791	0.771	0.643	0.677
	LAVS(Dec)	**0.799**	**0.829**	**0.806**	**0.786**	**0.790**	**0.798**	**0.796**	**0.777**	**0.660**	**0.713**
	Δ ↑	0.018	0.021	0.019	0.020	0.007	0.024	0.005	0.006	0.017	0.036

Metric	Method	x-cs	x-fr	x-de	x-fi	x-lv	x-et	x-ro	x-hi	x-tr	x-gu
OTR	Vocab Sharing	22.4%	17.8%	23.9%	26.0%	21.9%	28.1%	8.9%	25.4%	14.0%	77.0%
	LAVS(Dec)	**8.7%**	**5.9%**	**6.6%**	**9.2%**	**8.4%**	**7.8%**	**3.0%**	**1.7%**	**7.0%**	**15.4%**
	Δ ↓	-13.7%	-11.9%	-17.3%	-16.8%	-13.5%	-20.3%	-5.9%	-23.7%	-7.0%	-61.6%
BLEU	Vocab Sharing	11.0	17.9	13.2	8.3	12.2	9.9	14.0	8.3	8.8	3.3
	LAVS(Dec)	**12.5**	**20.1**	**15.7**	**9.4**	**13.3**	**11.7**	**14.2**	**9.9**	**9.0**	**6.7**
	Δ ↑	+1.5	+2.2	+2.5	+1.1	+1.1	+1.8	+0.2	+1.6	+0.2	+3.4
BERT Score	Vocab Sharing	0.772	0.776	0.781	0.749	0.757	0.759	0.771	0.743	0.750	0.723
	LAVS(Dec)	**0.791**	**0.799**	**0.796**	**0.770**	**0.777**	**0.774**	**0.797**	**0.756**	**0.768**	**0.726**
	Δ↑	0.019	0.023	0.015	0.021	0.020	0.015	0.026	0.013	0.018	0.003

Table 4: The zero-shot translation performance (Off-Target Rate, BLEU and BERT-Score) on average x-to-many and many-to-x directions using LAVS (Dec) compared to baseline.

LAVS in encoder benefits more to the high-resource languages. LAVS (Enc,Dec) also splits the vocabulary in the encoder. Compared with LAVS (Dec), this leads to larger improvement in H-H directions while smaller improvement in directions involving low-resource language according to Table 3. Vocabulary sharing in the encoder has more advantages for low-resource languages since those directions desperately need knowledge transfer from other directions, which would be blocked by adding language-specific tokens.

Constrained decoding further improves the performance of LAVS. Given the vocabulary of different languages, we propose another method to prevent off-target, which is through constrained decoding (CD). During decoding, the decoder only considers tokens that belong to the target vocab in softmax. The target vocab could be computed using the training corpus. CD is orthogonal to LAVS so they can be jointly applied. We implement CD for both original vocab sharing and LAVS.

Method	DE->CS		FR->DE	
	OTR	BLEU	OTR	BLEU
Vocab Sharing	45.1%	9.7	38.3%	12.7
w/ CD	30.9%	11.4	36.4%	12.8
LAVS (Dec)	18.9%	13.0	15.4%	17.2
w/ CD	**11.1%**	**14.2**	**11.3%**	**17.8**

Table 5: The results of constrained decoding (CD) combined with LAVS. Constrained decoding could further improve the performance of LAVS.

As shown in Table 5, it turns out that constrained decoding can further improve the zero-shot performance for both methods. It is worth noticing that, in some direction like FR->DE, the benefit of CD is rather small for the baseline model. We think the reason is that the original vocab sharing generates many shared tokens between languages, which will weaken the influence of the constraint. Thus, with more language-specific tokens, LAVS can work better with constrained decoding.

4.5 Discussion

How does LAVS calibrate the translation direction? During zero-shot translation, the language identifier token "<XX>" is the only element indi-

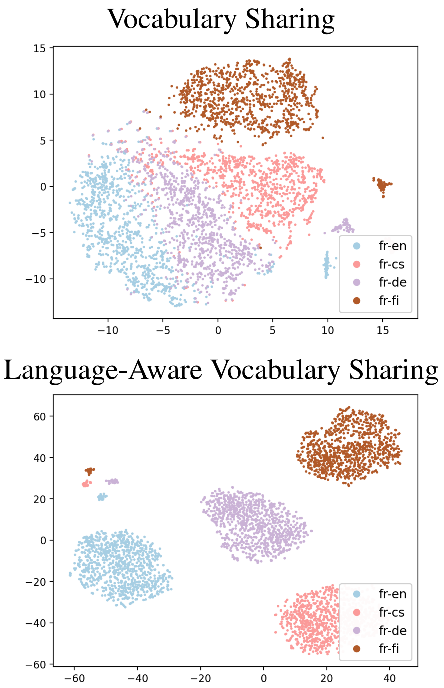


Figure 8: Encoder's hidden output for language identifier token <XX>, visualized using TSNE.

cating the correct direction. Similar to the visualization in Section 2.4, as shown in Figure 8, we visualize the <XX> tokens' hidden output(instead of the pooled result from all input tokens) during French-to-Many translation among high-resource languages and compare the results of the original Vocabulary Sharing and LAVS. It turns out that LAVS encodes more discriminative target language information into the <XX> token's hidden output, while the original Vocabulary Sharing fails on that.

In original Vocabulary Sharing the mapping between the target language identifier <XX> and output token is Many-to-One since different language could share output tokens. While for LAVS, the mapping becomes One-to-One for a part of tokens, impulsing the encoder to learn more discriminative representations for the target language identifier and make the model more sensitive to the target language identifier during zero-shot translation. We also give it a case study as shown in Appendix B.

How many Language-Specific tokens do we need? As shown in Table 6 from Appendix, we conduct an ablation study on how the number of language specific(LS) tokens influence the zero-shot performance. The result shows that the OTR keeps decreasing when the number of LS tokens increases. It suggests that more LS tokens can better relieve the off-target issue.

Shared Tokens(M)	LS Tokens(N)	OTR
64k	0	29.4%
54k	0	33.1%
54k	10k	8.2%
54k	20k	7.4%
54k	50k	5.9%

Table 6: Ablation Study on the number of Language-Specific tokens and the Off-Target Rate on Flores-101. We report the average OTR on 81 zero-shot directions.

5 Related Work

Off-Target Problem in Multilingual NMT Several methods are proposed to eliminate the off-target problem. Gu et al. (2019) introduced decoder pretraining to prevent the model from capturing spurious correlations. Zhang et al. (2020); Gu et al. (2019) resorted back-translation technique to generate data for non-English directions ,Wu et al. (2021) explored how language tag settings influence zero-shot translation and Yang et al. (2021) introduced extra optimization objective to address the problem. However, the cause for off-target still remains underexplored and the contribution of LAVS is orthogonal to previous studies.

Vocabulary of NMT In the early stages, several word-split methods like Byte-Pair Encoding (Sennrich et al., 2016), Wordpiece (Wu et al., 2016) and Sentencepiece (Kudo and Richardson, 2018), are proposed to handle rare words using a limited vocab size. In the background of multilingual NMT, most current studies and models (Devlin et al., 2018; Conneau et al., 2019; Liu et al., 2020; Ma et al., 2021) regard all languages as one and learn a shared vocabulary for different languages. Recently, Xu et al. (2021a) adopted optimal transport to find the vocabulary with most marginal utility. Chen et al. (2022) studied the relation between vocabulary sharing and label smoothing. To the best of our knowledge, we are the first to explore how vocabulary affects off-target in multilingual NMT.

6 Conclusion

In this paper, we delve into the hidden reason for the off-target problem in zero-shot multilingual NMT and propose Language-Aware Vocabulary Sharing (LAVS) which could significantly alleviate the off-target problem without extra parameters. Our experiments justify that LAVS creates a better multilingual vocab than the original Vocabulary Sharing method for multiple languages.

7 Limitation

LAVS is proposed to overcome the off-target problem among languages that share alphabets because those languages tend to have more sharing tokens after the sub-word tokenization process. As for language pair that does not have shared tokens, LAVS might not have a direct influence on the zero-shot translation though it can also increase the overall performance for those languages, which might need further exploration.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan, editors. 2010. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, Uppsala, Sweden.

Liang Chen, Runxin Xu, and Baobao Chang. 2022. Focus on the target's vocabulary: Masked label smoothing for machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 665–671, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *ArXiv*, abs/2106.13736.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034, Online. Association for Computational Linguistics.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3001–3007, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021a. Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7361–7373, Online. Association for Computational Linguistics.

Weijia Xu, Yuwei Yin, Shuming Ma, Dongdong Zhang, and Haoyang Huang. 2021b. Improving multilingual neural machine translation with auxiliary source languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3029–3041, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. 2021. Improving multilingual translation by representation and gradient regularization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7266–7279, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Method for Completely Separating Vocab

It is easy to turn a shared vocabulary into a separate vocabulary for different languages. As shown in Figure 9, we can split the shared token into language specific token if it appears in more than one language.

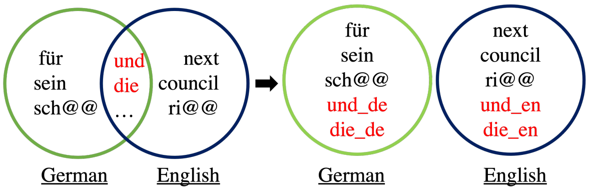


Figure 9: Illustration of completely separating vocabulary of different languages.

B Case Study

We compare different model's outputs as shown in Figure 10. The baseline output has off-target problem while LAVS output generates in the correct language. From the direct token output of LAVS, we can see that many of which are language-specific tokens. Models with LAVS could learn the relation between the target language signal and corresponding language-specific tokens, which further decreases the probability of off-target.

Direction: DE-> FR
Input: <FR> Apia wurde in den 50ern des 18. Jahrunderts gegründet und ist seit 1959 die offizielle Hauptstadt von Samoa.
Output(baseline): Apia was founded in the 50s of the 18th century and is the official capital of Samoa since 1959. (Off-Target to English)
Gold: Apia a été fondée dans les années 1850 et est la capitale officielle des Samoa depuis 1959.

Output(LAVS-token): Apia_fr a_fr été fondée dans les_fr 50 ans_fr du_fr 18e siècle et_fr est_fr depuis 1959 la_fr capitale officielle de_fr Samoa.
Output(LAVS-literal): Apia a été fondée dans les 50 ans du 18e siècle et est depuis 1959 la capitale officielle de Samoa.

Figure 10: Case study of DE->FR zero-shot translation. The baseline model off-target to English. Tokens in blue belong to language-specific tokens.

C Datasets

C.1 WMT'10

Following Wang et al. (2020); Yang et al. (2021); Xu et al. (2021b), we collect data from freely-accessible WMT contests to form a English-Centric WMT10 dataset.

Direction	Train	Test	Dev
Fr↔En	10.00M	newstest15	newstest13
Cs↔En	10.00M	newstest18	newstest16
De↔En	4.60M	newstest18	newstest16
Fi↔En	4.80M	newstest18	newstest16
Lv↔En	1.40M	newstest17	newsdev17
Et↔En	0.70M	newstest18	newsdev18
Ro↔En	0.50M	newstest16	newsdev16
Hi↔En	0.26M	newstest14	newsdev14
Tr↔En	0.18M	newstest18	newstest16
Gu↔En	0.08M	newstest19	newsdev19

Table 7: Description for WMT'10 Dataset.

C.2 Flores-101

Flores-101 (Goyal et al., 2021; Guzmán et al., 2019) is a Many-to-Many multilingual translation benchmark dataset for 101 languages. It provides parallel corpus for all languages, which makes it suitable to test the zero-shot translation performance of multilingual NMT model. We use the devtest split of the dataset, and only test on the languages that appear during supervised training.

Language	Code	Split	Size
French	Fr	devtest	1012
Czech	Cs	devtest	1012
German	De	devtest	1012
Finnish	Fi	devtest	1012
Latvian	Lv	devtest	1012
Estonian	Et	devtest	1012
Romanian	Ro	devtest	1012
Hindi	Hi	devtest	1012
Turkish	Tr	devtest	1012
Gujarati	Gu	devtest	1012

Table 8: Description for Flores-101 Dataset.