Language-Informed Beam Search Decoding for Multilingual Machine Translation

Anonymous ACL submission

Abstract

Beam search decoding is the de-facto method for decoding auto-regressive Neural Machine Translation (NMT) models, including multilingual NMT where the target language is specified as an input. However, decoding multilingual NMT models commonly produces "off-target" translations – yielding translation outputs not in the intended language. In this paper, we first conduct an error analysis of off-target translations for a strong multilingual NMT model and identify how these decodings are produced during beam search. We then propose Language-informed Beam Search (LiBS), a general decoding algorithm incorporating an off-the-shelf Language Identification (LiD) model into beam search decoding to reduce off-target translations. LiBS is an inference-time procedure that is NMT-model agnostic and does not require any additional parallel data. Results show that our proposed LiBS algorithm on average improves +1.1 BLEU and +0.9 BLEU on WMT and OPUS datasets, and reduces off-target rates from 22.9% to 7.7% and 65.8% to 25.3% respectively.[1]

1 Motivation

With Neural Machine Translation (NMT) (Bahdanau et al., 2014; Vaswani et al., 2017) becoming the state-of-the-art approach in the bilingual Machine Translation literature, Multilingual Neural Machine Translation (MNMT) has attracted much attention (Johnson et al., 2017). MNMT has two main advantages: a) it enables one model to translate between multiple language pairs and thus reduces the model and deployment complexity from $O(N^2)$ to $O(1)$, and b) it enables transfer learning between high-resource and low-resource languages. One attractive feature of such transfer learning is *zero-shot* translation, where the multilingual model is able to translate between language pairs unseen during training. For example, after training from French to English and English to German MT data, the model could directly translate French to German.

Despite the theoretical benefits, recent studies have found an overwhelming amount of *off-target translation* especially for the zero-shot directions (Zhang et al., 2020; Yang et al., 2021), where the translation is not in the intended language. Existing methods all aim to mitigate off-targets during training. Gu et al. (2019); Zhang et al. (2020) apply Back Translation (BT) to generate synthetic training data for the zero-shot pairs. Yang et al. (2021) introduces a language prediction loss and regularizes the training gradients with a held-out *oracle* set. Yet, none of the previous work has investigated the off-target issue at decoding time, i.e. how *off*-target translations emerge and come to outscore *on*-target translations during beam search decoding.

In this work, we first examine when and how off-target translation emerges during beam search decoding, and then propose Language-informed Beam Search (LiBS), a general algorithm to reduce off-target generation during beam search decoding by incorporating an off-the-shelf Language Identification (LiD) model. Our experiment results on two large-scale popular MNMT datasets (i.e. WMT and OPUS) demonstrate the effectiveness of LiBS in both reducing off-target rates and improving general translation performance. On average LiBS reduces off-target rates from 22.9% to 7.7% and 65.8% to 25.3% on WMT and OPUS respectively, which translates to +1.1 BLEU and +0.9 BLEU overall quality improvement. Moreover, LiBS can be added post-hoc to reduce off-target translation of any existing multilingual model without requiring any additional data or training.

2 Experiment Setup

In this section, we illustrate the data and model setup we used, and the experimental results of our

[1] Code to be released after publication.

Language-informed Beam Search algorithm.

2.1 Dataset

Following (Wang et al., 2020; Yang et al., 2021), we conduct experiments on two widely used large-scale MNMT datasets WMT[2] and OPUS-100[3], where the WMT dataset is concatenated from previous year WMT training data including English and 10 other languages. Since the WMT competition does not come with zero-shot evaluation data, we use the human labeled multi-way aligned test set from (Yang et al., 2021), based on the WMT-19 test set.

2.2 Model Training and Evaluation

For both WMT and OPUS-100, we tokenize the dataset with the SentencePiece model (Kudo and Richardson, 2018) to form a shared vocabulary of 64k tokens. We adopt the Transformer-Big setting (Vaswani et al., 2017) in our experiments on the open-sourced Fairseq codebase[4] (Ott et al., 2019). The model is optimized using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 5×10^{-4}, 4000 warm-up steps, and a total of 50k training steps. The multilingual model is trained on 8 V100 GPUs with a batch size of 8192 tokens and gradient accumulation of 8 steps, which essentially simulates the training on 64 V100 GPUs. To evaluate the baseline model, we employ beam search decoding with a beam size of 5 and a length penalty of 1.0. The BLEU score is then measured by the de-tokenized case-sensitive SacreBLEU[5] (Post, 2018).

To evaluate the off-target rates, we borrow the off-the-shelf LiD model[6] from FastText (Joulin et al., 2016) to detect the language for system translations. Similar to (Yang et al., 2021), we observe an overwhelming off-target rate (averaging 22.9%) across zero-shot pairs on our strong baseline model.

[2]Referred to as "WMT-10" in (Wang et al., 2020; Yang et al., 2021), we denoted it as WMT to disambiguate against the WMT 2010 campaign.

[3]We use the deduplicated version from (Yang et al., 2021).

[4]https://github.com/facebookresearch/fairseq

[5]BLEU+case.mixed+lang.src-tgt+numrefs.1+smooth.exp+tok.13a+version.1.4.14

[6]https://dl.fbaipublicfiles.com/fasttext/supervised-models/lid.176.bin

Direction	Beam size	BLEU	Off-Target Rate
De→Fr	5	17.3	23.1%
	10	16.1	31.7%
	20	14.3	41.4%
Cs→De	5	15.4	12.3%
	10	15.0	17.4%
	20	14.2	22.0%

Table 1: Multilingual beam search curse on WMT De→Fr and Cs→De, where larger beam widths consistently lead to more off-target translations.

3 Analyzing Off-Target Occurrence During Beam Search

To understand the off-target occurrence during beam search, we analyze the off-target error types on different language pairs, and conduct experiments with varying beam sizes.

3.1 Multilingual Beam Search Curse

The beam search curse phenomenon (Yang et al., 2018) is widely observed in bilingual NMT models. Given a larger beam size, the beam search process would explore a larger search space and choose from a larger candidate pool. Yet empirically, translation performance usually drops significantly with increasing beam sizes. In our study, we also found this phenomenon prevailing in the multilingual system and highly related to the off-target translation error.

As an example, we demonstrate the beam search curse on WMT De→Fr and Cs→De translation, since both are between high-resource languages and with decent translation performance (between 15 to 20 BLEU).

Table 1 illustrates the results on WMT De→Fr and Cs→De. We could clearly observe that the off-target rate grows sub-linearly with the beam size, and as a result the BLEU score drops significantly with increasing beam sizes. It then raises the curious question of why the off-target rate increases drastically with larger beam sizes, and whether the performance drop (i.e. BLEU decrease) is mainly due to the off-target errors.

3.2 Off-Target Error Analysis

As part of a detailed analysis, we study the off-target error type between six zero-shot pairs (i.e. 12 translation directions) from the WMT dataset. We categorize the off-target errors into three types:

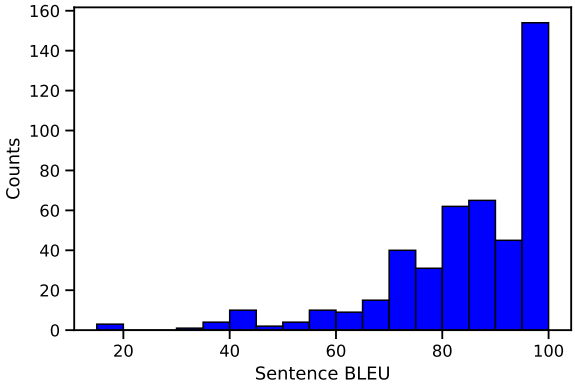

Figure 1: The sentence BLEU distribution between source and system translation from WMT Fr→De "→Source" errors, with an average BLEU of 85.3.

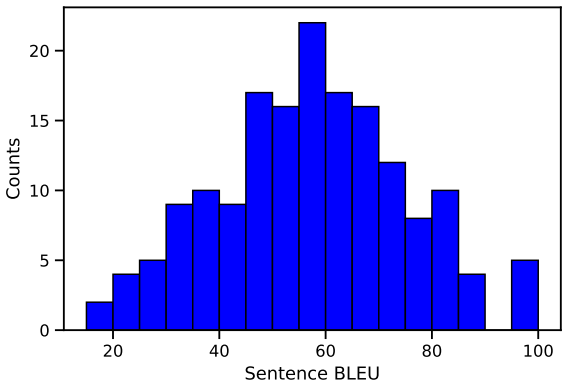

Figure 2: The sentence BLEU distribution between WMT Fr→De "→English" errors and Fr→En translation with the same source. The average BLEU is 55.9.

translating into English[7], translating into source language, and others.

The detailed off-target error analysis of WMT zero-shot direction is shown in Table 2. We find that even though the off-target error is overwhelming across languages, it could easily be categorized into mostly two types: translating into English and "translating" into source. The "Others" error type only comprises a negligible 1.1% of cases, given the FastText LiD model has an error margin of 0.81% (Yang et al., 2021).

"→Source" errors We hypothesize that this error is related to the previously studied "source copying" behavior (Ott et al., 2018) on the bilingual NMT model. We then sample three cases from this error type (shown in Table 3). The case study confirms that the "→Source" error type is the same as source copying behavior on bilingual models for these cases. To quantify the degree of source copying, we run Sentence BLEU evaluation[8] between source and system translation on WMT Fr→De "→Source" errors. The sentence BLEU distribution is shown in Figure 1 with an average sentence BLEU of 85.3. It clearly demonstrates that the "→Source" error strongly displays a source copying behavior and is somehow promoted by larger beam sizes.

"→English" errors Since none of our evaluated direction includes English as the target language, translating into English is never promoted and always trigger an off-target error. We similarly sampled three "→English" error cases from the WMT Fr→De test set. We also compare them against the real Fr→En translations with the same model and French input. This case study (in Table 4) hints that the "→English" generations from WMT Fr→De are generally *similar* but slightly *worse* "English" translations compared to Fr→En. To demonstrate the similarity, we plot the sentence BLEU distribution for all 172 "→English" errors between Fr→De and Fr→En translations in Figure 2. It demonstrates a strong similarity between Fr→De and Fr→En translations, with an average sentence BLEU of 55.9. Since the evaluation data of the WMT corpus is multi-way aligned, we can evaluate the →English translation quality for both Fr→De and Fr→En against the English human references (in Table 5). Results confirm our observation that the "→English" errors are generally poorer English translations.

3.3 Beam Search Process Analysis

To understand how "→English" and "→Source" errors emerge during beam search and why both errors dramatically increase with larger beam sizes, we investigate the step-by-step beam search process with case studies. Table 10 and 11 illustrates one representative decoding example from the WMT Fr→De test set with $b = 5, 20$ and French source "Nous avons maintenant une excellente relation. »". For $b = 20$, we only print the top-5 beams due to the space limit. From this example, we have a few observations:

- English candidates are live in the early steps

[7]English is never the correct target language in our 12 studied translation directions.

[8]We use the `sentence_bleu` function from (Post, 2018) with `smooth_method='floor'`: https://github.com/mjpost/sacrebleu/blob/master/sacrebleu/compat.py

Directions	$b = 5$				$b = 20$			
	Total	→English	→Source	Others	Total	→English	→Source	Others
De→Fr	23.1%	11.8%	11.1%	0.2%	41.4%	18.5%	22.8%	0.1%
Fr→De	39.9%	10.5%	29.4%	0.0%	62.7%	17.2%	45.5%	0.0%
Cs→De	12.3%	8.5%	3.6%	0.2%	22.0%	17.3%	4.5%	0.2%
De→Cs	19.0%	2.5%	15.8%	0.7%	27.6%	5.9%	21.3%	0.4%
De→Ro	1.6%	0.8%	0.5%	0.3%	1.9%	1.1%	0.5%	0.3%
Ro→De	7.3%	5.9%	0.7%	0.7%	16.3%	14.8%	0.7%	0.8%
Fr→Et	22.5%	8.1%	12.4%	2.0%	30.6%	13.6%	15.6%	1.4%
Et→Fr	26.1%	16.5%	6.3%	3.3%	36.6%	26.2%	6.7%	3.7%
Ro→Et	10.8%	6.3%	1.5%	3.0%	14.8%	10.4%	1.6%	2.8%
Et→Ro	2.0%	0.5%	0.2%	1.3%	1.9%	0.6%	0.3%	1.0%
Tr→Gu	73.7%	73.3%	0.2%	0.2%	78.7%	78.1%	0.2%	0.4%
Gu→Tr	36.4%	35.6%	0.1%	0.7%	41.4%	39.6%	0.0%	1.8%
Average	22.9%	15.0%	6.8%	1.1%	31.3%	20.3%	10.0%	1.1%

Table 2: Off-Target error analysis on 12 WMT zero-shot directions, where most are either →English or →Source.

Source (Fr)	**System Output (→De)**
Sa décision a laissé tout le monde sans voix.	Sa décision a laissé tout le monde sans voix.
Abandonnez Chequers et commencez à écouter. »	Abandonnez Chequers et commencez à écouter »
« C'est une très bonne chose », dit Jaynes.	C'est une très bonne chose, sagt Jaynes.

Table 3: Case studies for "→Source" errors. We sample three source-translation pairs from the WMT Fr→De test set (with translation LiD-ed as French). Token differences are colored in red.

(1-3) of $b = 5$ but tend to be dropped in later time steps. Meanwhile for $b = 20$, both English and French candidates are kept alive throughout the decoding process: even though they fall out of the top-5 beams at the 4th step, the off-target candidates quickly catch up and are ranked highest by the 7th step.

- Closely observing the winning English candidate of $b = 20$, we notice it suffers a heavy penalty in the first step (log prob is -3.58), yet all following steps experience small penalties.

- The final English translation by $b = 20$ is indeed a "better" candidate with greater probabilities (i.e. model score) compared to the final German translation by $b = 5$, therefore, if this off-target candidate is retained throughout the process it will naturally win out against all valid on-target translations.

From the above observations, we can try to answer our previous research questions.

RQ1. How do "→English" and "→Source" errors emerge during decoding?

We first observe that both "→English" and "→Source" candidates are easily accessible in the early steps of decoding. Meanwhile, the models place a low probability on decoding the first source or English token, but relatively high transition probabilities for the remaining source or English tokens can result in off-target sentences scoring more highly than on-target.

RQ2. Why do both errors dramatically increase with larger beam sizes?

With a larger beam size budget, it is more likely to retain off-target candidates in the earlier steps, when they receive heavy early step penalties. Yet since off-target candidates experience fewer penalties in the later steps, they tend to win out over on-target candidates in the long run. We found it

Source (Fr)	System Output (→De)	System Output (→En)
Comme la campagne était très avancée, elle avait pris du retard dans la collecte de fonds, et a donc juré qu'elle ne participerait pas à moins de recueillir 2 millions de dollars.	Since the campaign was very advanced, it had fallen behind in the collection of funds, and therefore swore that it would not participate to less than raise 2 million dollars.	Since the campaign was very advanced, it had lagged behind in raising funds and therefore swore that it would not participate unless it raised $2 million.
Woods a perdu ses quatre matchs en France et détient maintenant un record de 13-21-3 en carrière en Ryder Cup.	Woods has lost his four matches in France and now holds a record of 13-21-3 in career in Ryder Cup.	Woods lost his four matches in France and now holds a record of 13-21-3 in the Ryder Cup.
Le couple réfute être raciste, et assimile les poursuites à une « extorsion ».	The couple refutes being racist, and assimilates prosecutions to a "repression".	The couple refutes being racist and treats prosecutions as "extortion".

Table 4: Case studies for "→English" errors. We sample three source-translation pairs from the WMT Fr→De test set (with translation LiD-ed as English). As a comparison, we also show the output when the model is asked to translate into English. Token differences are colored in red.

Direction	BLEU	chrF2	TER*
Fr→De	26.92	0.572	0.62
Fr→En	34.91	0.611	0.53

Table 5: English translation quality for all WMT Fr→De "→English" errors. Both translation directions are evaluated against English human references. *TER score is lower the better.

to be the general case that the off-target continuations receive a higher probability (less penalty) than the on-target ones, even though the first off-target token receives a heavy penalty by the model. We hypothesize that it is due to the *recency bias* and poor calibration, yet it remains an interesting research question for future work.

Possible Solutions Off-target candidate gaining greater model score demonstrates that the model is poorly calibrated, especially for the later steps of autoregressive decoding. Methods with additional training data (Gu et al., 2019; Zhang et al., 2020) or regularizations (Yang et al., 2021) could alleviate this issue during training with a well-calibrated model. In this work, with the knowledge of how off-target cases emerge during decoding, we attempt to fix this issue solely at the decoding time by improving on beam search algorithm even with a proven poorly calibrated model.

4 Language-informed Beam Search (LiBS)

The standard beam search process originating from the bilingual NMT model is target-language-agnostic and is found to produce an overwhelming number of off-target translations (Zhang et al., 2020; Yang et al., 2021). Yet the target language (i.e. the desired language for generation) is always known during decoding, thus it is straightforward to enforce the desired language to reduce the off-target rates without any additional training or data. We thus propose Language-informed Beam Search (LiBS), a general decoding algorithm to inform the beam search process of the desired language during decoding.

To inform the beam search process of the desired language, we borrow an off-the-shelf Language Identification (LiD) model to score the running beam search candidates with their probabilities in the correct language. Since the candidates are normally ranked by the NMT model probabilities, we linearly combine the two log probabilities to ideally find the best candidate in the correct language.

The detailed algorithm is illustrated in Algorithm 1. For each step, we first pre-select top-w candidates from each beam. Then we sort all $b \cdot w$ active candidates by the linearly combined NMT and LiD log probabilities, where we tune the linear coefficient α on the dev set. Same as the Fairseq (Ott et al., 2019) implementation, we only

Algorithm 1: Langugage-informed Beam Search

Input : MNMT model θ, LiD model γ, source sentence $\mathbf{x}$, target language T, beam size b, pre-select window size w

Output : Finished candidate set $\mathbf{C} \leftarrow \emptyset$

▷ *Initialize each beam i with BOS symbols and zero score*

1 $\mathbf{B}_i \leftarrow \{\langle 0.0,\ \texttt{<s>}\rangle\}$

2 **repeat**

▷ *Pre-select top-w candidates from each beam i*

3 $\mathbf{W}_i \leftarrow \mathbf{top}_w\{\langle s \cdot p_\theta(y \mid \mathbf{x}, \mathbf{y}),\ \mathbf{y} \circ y\rangle \mid \langle s,\ \mathbf{y}\rangle \in \mathbf{B}_i,\ y \in \mathcal{V}\}$

▷ *Sort all candidates by the linearly combined NMT and LiD log probabilities*

4 $\mathbf{W} \leftarrow \mathbf{Sort}\{\langle \log s + \alpha \log p_\gamma(T,\ \mathbf{y}),\ s,\ \mathbf{y}\rangle \mid \langle s,\ \mathbf{y}\rangle \in \bigcup_{i=1}^{b} \mathbf{W}_i\}$

▷ *Store all finished ones from top-b candidates into $\mathbf{C}$*

5 $\mathbf{C} \leftarrow \{\langle s,\ \mathbf{y}\rangle \mid \langle s', s,\ \mathbf{y}\rangle \in \mathbf{top}_b\{\mathbf{W}\},\ \mathbf{y}_{|\mathbf{y}|} = \texttt{</s>}\}$

▷ *Store the top-b unfinished candidates into $\mathbf{B}$*

6 $\mathbf{B} \leftarrow \mathbf{top}_b\{\langle s,\ \mathbf{y}\rangle \mid \langle s', s,\ \mathbf{y}\rangle \in \mathbf{W},\ \mathbf{y}_{|\mathbf{y}|} \neq \texttt{</s>}\}$

7 **until** $\mathbf{C}$ *has b finished candidates (i.e. $|\mathbf{C}| = b$)*

▷ *Rerank finished candidates by the linearly combined NMT and LiD log probabilities*

8 $\mathbf{C} \leftarrow \mathbf{Sort}\{\langle \log s + \alpha \log p_\gamma(T,\ \mathbf{y}),\ \mathbf{y}\rangle \mid \langle s,\ \mathbf{y}\rangle \in \mathbf{C}\}$

9 **return** C

store the finished ones within the top-b candidates, meanwhile save the top-b active candidates into the beam for the next step[9]. The decoding process stops when we have found b finished candidates, and at the end of the generation, we again rerank all finished candidates with the linearly combined log probabilities.

Design Choice and Speed Concern We only pre-select $b \cdot w$ candidates for the LiD scoring, instead of considering all possible continuations, simply because we could not afford to run LiD model on all $b \cdot |\mathcal{V}|$ candidates.

Even though in our experiments we only pre-select the top-2 continuations from each beam (i.e. $w = 2$), the major slow down of the LiBS algorithm is still the un-BPE operation and LiD scoring on line 4 of Algorithm 1.

To speed up the LiBS algorithm, we use the FastText LiD model since it is both fast and accurate (in our case on translation prefixes). With its help, LiBS is only 7.5 times slower than the Fairseq beam search decoding on a single CPU and 3.5 times slower with parallelized LiD scoring on 20 CPUs.

[9] We only store the NMT model score instead of the linearly combined one to avoid overcounting LiD scores.

Model	De→Fr		Cs→De	
	BLEU	Off-Tgt	BLEU	Off-Tgt
Baseline	17.3	23.1%	15.4	12.3%
+LiBS, $\alpha = 0.7$	20.6	2.0%	16.1	1.6%
+LiBS, $\alpha = 0.8$	20.7	1.4%	16.2	1.6%
+LiBS, $\alpha = 0.9$	20.7	1.1%	16.2	1.6%
+LiBS, $\alpha = 1.0$	20.7	0.9%	16.2	1.4%
+LiBS, $\alpha = 1.1$	20.7	0.9%	16.2	1.4%
+LiBS, $\alpha = 1.2$	20.7	0.8%	16.2	1.3%

Table 6: Tuning the linear coefficient α on WMT De→Fr and Cs→De.

5 Experiment Results

To fully verify the performance of the LiBS algorithm, we compare LiBS against the baseline beam search decoding on both WMT and OPUS-100 datasets.

5.1 WMT Results

Tuning the Linear Coefficient α We tune the linear coefficient α on the dev set. As shown in Table 6, any α value from 0.8 to 1.2 performs similarly well. Because the linear coefficient α controls the weight of the LiD model score, as α increases, the off-target rate monotonically drops. We use $\alpha = 0.9$ for all the experiments on the WMT dataset.

Model		De→Fr			Cs→De		
		$b = 5$	$b = 10$	$b = 20$	$b = 5$	$b = 10$	$b = 20$
Baseline	BLEU	17.3	16.1	14.3	15.4	15.0	14.2
	Off-Tgt	23.1%	31.7%	41.4%	12.3%	17.4%	22.0%
+LiBS	BLEU	20.7	20.9	20.7	16.2	16.4	16.2
	Off-Tgt	1.1%	1.2%	1.1%	1.6%	1.7%	1.5%

Table 7: LiBS breaks the beam search curse on WMT De→Fr and Cs→De.

Multilingual Beam Search Curse As illustrated before, the beam search curse exists in Multilingual NMT models predominantly due to the increasing off-target errors with larger beam sizes. As shown in Table 7, LiBS successfully breaks the beam search curse by preventing off-target translations.

Zero-Shot Performance Table 8 illustrates the full results of LiBS on the WMT dataset. On average across all zero-shot directions, LiBS improves +1 BLEU score while reducing the off-target rates from 22.91% to 7.71%. We notice that for many directions the off-target rate is barely around the error margin of the FastText LiD model, which is 0.81% reported from (Yang et al., 2021). It hints that those translation directions do not suffer from off-target errors anymore, and the reported errors are largely due to the LiD model error. Meanwhile, the MNMT model still suffers from a large number of off-target errors, especially on Gu→Tr and Tr→Gu translations, which we hypothesize is due to the extremely low resources for both languages (WMT contains 180K and 80K parallel data for Tr-En and Gu-En respectively.).

5.2 OPUS-100 Results

To verify the effectiveness of our LiBS algorithm, we further compare it against the baseline beam search decoding on the large-scale OPUS-100 dataset, which includes a total of 100 languages.

Different from the WMT experiment, we tune and set $\alpha = 1.8$ for all directions. This is due to the challenging nature of the OPUS-100 dataset that it performs very poorly on the zero-shot directions with a massive amount of off-target translations. A higher α value for LiBS could effectively reduce the off-target rates and improve the translation performance. For example, Figure 3 plots the performance curve on the OPUS-100 Fr→De test set with increasing α values. It clearly shows a larger α value would consistently decrease the off-target rates and improve the overall performance (i.e. higher BLEU score)[10].

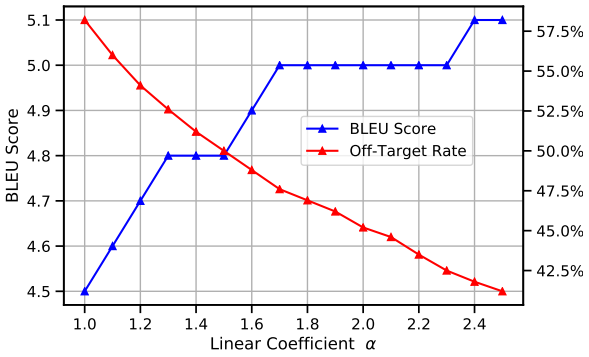

Figure 3: Translation performance (BLEU and off-target rates) with different α values on OPUS-100 Fr→De test set.

Zero-shot translation performance of LiBS on the OPUS-100 dataset is shown in Table 9. Across all directions, LiBS consistently improves an average of +0.9 BLEU and reduces the off-target rates from 65.79% to 25.34%.

Both WMT and OPUS-100 results clearly show our LiBS algorithm notably improves the zero-shot translation performance by significantly reducing the off-target translations.

6 Related Work

Off-Target Translation Off-target translation is a commonly observed failure mode in multilingual NMT models (Arivazhagan et al., 2019), and Rios et al. (2020) has linked it to the predominance of English in the training data of multilingual models. Gu et al. (2019); Zhang et al. (2019); Yang et al. (2021) all observe it under different data settings and propose to mitigate it using additional monolingual data or held-out oracle set. Similarly to our

[10]The flat BLEU curve is due to the one decimal digit precision of sacreBLEU evaluation.

Zero-Shot		Fr-De ←	Fr-De →	De-Cs ←	De-Cs →	Ro-De ←	Ro-De →	Et-Fr ←	Et-Fr →	Et-Ro ←	Et-Ro →	Gu-Tr ←	Gu-Tr →	Average
Baseline	BLEU	17.3	11.7	15.4	13.9	17.2	16.1	10.6	13.5	11.9	14.1	0.9	2.0	12.05
	Off-Tgt	23.1%	39.9%	12.3%	19.0%	1.6%	7.3%	22.5%	26.1%	10.8%	2.0%	73.7%	36.6%	22.91%
+LiBS	BLEU	20.7	15.7	16.2	15.3	17.1	16.5	11.8	14.6	12.4	13.9	1.2	2.3	13.14
	Off-Tgt	1.1%	6.5%	1.6%	3.8%	0.6%	0.6%	8.3%	4.6%	2.9%	0.3%	47.2%	15.0%	7.71%

Table 8: BLEU score and Off-Target rate of zero-shot translations on WMT dataset.

Zero-Shot		De-Fr ←	De-Fr →	Ru-Fr ←	Ru-Fr →	Nl-De ←	Nl-De →	Zh-Ru ←	Zh-Ru →	Zh-Ar ←	Zh-Ar →	Nl-Ar ←	Nl-Ar →	Average
Baseline	BLEU	3.3	3.0	5.4	4.0	5.9	5.2	5.7	11.8	3	11.6	1.2	3.2	5.28
	Off-Tgt	95.2%	93.7%	68.9%	91.2%	88.4%	89.7%	37.0%	20.2%	89.9%	8.0%	93.0%	14.3%	65.79%
+LiBS	BLEU	5.0	3.8	9.6	4.4	7.4	5.9	5.9	12.2	3.5	11.1	2.5	2.8	6.18
	Off-Tgt	46.9%	49.5%	22.5%	41.6%	37.9%	40.0%	5.8%	1.1%	28.3%	0.7%	28.4%	1.4%	25.34%

Table 9: BLEU score and Off-Target rate of zero-shot translations on OPUS-100 dataset.

work, Sennrich et al. (2023) proposes to mitigate off-target errors with constrastive decoding, yet their approach usually hurts the translation quality, on average -1.1 BLEU on high resource languages. Our work is the first to study off-target errors during decoding time, specifically how the off-target translations outscore on-target ones over time.

NMT Decoding Since beam search becomes the de-facto method for decoding NMT models (Bahdanau et al., 2014), studies has observed various flaws with it. Koehn and Knowles (2017) observes beam search curse, where the translation quality usually degrades with increasing beam sizes. Yang et al. (2018); Stahlberg and Byrne (2019) observe length bias, where the model heavily prefers shorter candidates. To address those issues, extensive work has proposed sampling-based decoding algorithms, where the most popular one is Minimum Bayes Risk (MBR) decoding (Eikema and Aziz, 2020). Yet, MBR decoding suffers severely from the quadratic complexity thus a slow inference speed. Another line of research adopts external Language models to NMT beam search. Yet this external LM usually interferes with NMT's *internal* LM (iLM). However, with iLM neutralization, this approach still lags behind leveraging the additional monolingual data through back-translation (Herold et al., 2023). Most similarly to our work, He et al. (2017); Ren et al. (2017) propose to incorporate a trained Value network during beam search decoding to improve the image-captioning task. Our work instead attempts to mitigate off-target translation errors with a small off-the-shelf LiD model, while keeping the inference overhead to the linear scale (3.5x slow down).

7 Conclusions

Our work conducts a comprehensive off-target error analysis with strong multilingual NMT models, to answer the question of how off-target translation wins over time during decoding. We additionally propose an empirical Language-informed Beam Search algorithm to mitigate off-target errors during decoding time and with linear-scale overhead.

8 Limitations

In this study, we utilize the widely adopted FastText LiD model, and the performance of LiBS may vary with the use of alternative LiD models. As our method is a modified beam search algorithm, it is not directly applicable to recent Language Language Models (Brown et al., 2020), which often do sampling during inference. Yet, we believe it will be particular interesting to adopt similar approach for LLM inference, as study shows LLMs are prune to hallucination (Zhang et al., 2023).

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bryan Eikema and Wilker Aziz. 2020. Is map decoding all you need? the inadequacy of the mode in neural machine translation. *arXiv preprint arXiv:2005.10283*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. *arXiv preprint arXiv:1906.01181*.

Di He, Hanqing Lu, Yingce Xia, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2017. Decoding with value networks for neural machine translation. In *Advances in Neural Information Processing Systems*, pages 178–187.

Christian Herold, Yingbo Gao, Mohammad Zeineldeen, and Hermann Ney. 2023. Improving language model integration for neural machine translation. *arXiv preprint arXiv:2306.05077*.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

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

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 *NAACL-HLT*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 290–298.

Annette Rios, Mathias Müller, and Rico Sennrich. 2020. Subword segmentation and a single bridge language affect zero-shot neural machine translation. *arXiv preprint arXiv:2011.01703*.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2023. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. *arXiv preprint arXiv:2309.07098*.

Felix Stahlberg and Bill Byrne. 2019. On nmt search errors and model errors: Cat got your tongue? *arXiv preprint arXiv:1908.10090*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Yiren Wang, ChengXiang Zhai, and Hany Hassan Awadalla. 2020. Multi-task learning for multilingual neural machine translation. *arXiv preprint arXiv:2010.02523*.

Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. 2021. Improving multilingual translation by representation and gradient regularization. *arXiv preprint arXiv:2109.04778*.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-) scoring methods and stopping criteria for neural machine translation. *arXiv preprint arXiv:1808.09582*.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019. Improving deep transformer with depth-scaled initialization and merged attention. *arXiv preprint arXiv:1908.11365*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Additional Tables

	$b = 5$			$b = 20$		
Step	Beam	LiD	LogProb	Beam	LiD	LogProb
1	_Wir	De	-1.10	_Wir	De	-1.10
	_"	En	-2.93	_"	En	-2.93
	_Jetzt	De	-3.02	_Jetzt	De	-3.02
	_We	En	-3.58	_We	En	-3.58
	_„	Ro	-3.61	_„	Ro	-3.61
2	_Wir _haben	De	-1.44	_Wir _haben	De	-1.44
	_Jetzt _haben	De	-3.48	_Jetzt _haben	De	-3.48
	_We _now	En	-4.10	_We _now	En	-4.10
	_Wir _verfügen	De	-4.29	_Nous _avons	Fr	-4.13
	_We _have	En	-4.83	_Wir _verfügen	De	-4.29
3	_Wir _haben _jetzt	De	-2.22	_Wir _haben _jetzt	De	-2.22
	_Wir _haben _nun	De	-2.82	_Wir _haben _nun	De	-2.82
	_Jetzt _haben _wir	De	-3.60	_Jetzt _haben _wir	De	-3.60
	_We _now _have	En	-4.20	_We _now _have	En	-4.20
	_Wir _haben _eine	De	-4.83	_Nous _avons _maintenant	Fr	-4.36
4	_Wir _haben _jetzt _eine	De	-2.65	_Wir _haben _jetzt _eine	De	-2.65
	_Wir _haben _nun _eine	De	-3.24	_Wir _haben _nun _eine	De	-3.24
	_Wir _haben _jetzt _ein	De	-3.93	_Wir _haben _jetzt _ein	De	-3.93
	_Jetzt _haben _wir _eine	De	-4.12	_Jetzt _haben _wir _eine	De	-4.12
	_Wir _haben _nun _ein	De	-4.58	_Wir _haben _nun _ein	De	-4.58
5	_Wir _haben _jetzt _eine _ausgezeichnete	De	-3.66	_Wir _haben _jetzt _eine _ausgezeichnete	De	-3.66
	_Wir _haben _nun _eine _ausgezeichnete	De	-4.23	_Wir _haben _nun _eine _ausgezeichnete	De	-4.23
	_Wir _haben _jetzt _eine _hervorragende	De	-4.25	_Wir _haben _jetzt _eine _hervorragende	De	-4.25
	_Wir _haben _nun _eine _hervorragende	De	-4.81	_We _now _have _an _excellent	En	-4.80
	_Wir _haben _jetzt _eine _exzellente	De	-4.95	_Wir _haben _nun _eine _hervorragende	De	-4.81

Table 10: Beam Search case study for $b = 5$ and $b = 20$ on one example from WMT Fr→De test set. English candidates ("→English" errors) are colored in red, while French candidates ("→Source" errors) are colored in blue.

	$b = 5$			$b = 20$		
Step	Beam	LiD	LogProb	Beam	LiD	LogProb
6	_Wir _haben _jetzt _eine _ausgezeichnete _Beziehung	De	-3.82	_Wir _haben _jetzt _eine _ausgezeichnete _Beziehung	De	-3.82
	_Wir _haben _nun _eine _ausgezeichnete _Beziehung	De	-4.39	_Wir _haben _nun _eine _ausgezeichnete _Beziehung	De	-4.39
	_Wir _haben _jetzt _eine _hervorragende _Beziehung	De	-4.42	_Wir _haben _jetzt _eine _hervorragende _Beziehung	De	-4.42
	_Wir _haben _nun _eine _hervorragende _Beziehung	De	-4.98	_We _now _have _an _excellent _relationship	En	-4.89
	_Wir _haben _jetzt _eine _exzellente _Beziehung	De	-5.12	_Wir _haben _nun _eine _hervorragende _Beziehung	De	-4.98
7	**_Wir _haben _jetzt _eine _ausgezeichnete _Beziehung .“**	De	-5.73	_Nous _avons _maintenant _une _excellent e _relation	Fr	-5.53
	_Wir _haben _jetzt _eine _ausgezeichnete _Beziehung .	De	-5.89	_Wir _haben _jetzt _ein _ausgezeichnete s _Verhältnis	De	-5.54
	_Wir _haben _jetzt _eine _ausgezeichnete _Beziehung ."	De	-5.95	**_We _now _have _an _excellent _relationship ."**	En	-5.70
	_Wir _haben _jetzt _eine _hervorragende _Beziehung .“	De	-6.28	_Wir _haben _jetzt _eine _ausgezeichnete _Beziehung .“	De	-5.73
	_Wir _haben _nun _eine _ausgezeichnete _Beziehung .“	De	-6.31	_Wir _haben _jetzt _ein _hervorragende s _Verhältnis	De	-5.82

Table 11: Beam Search case study for $b = 5$ and $b = 20$ on one example from WMT Fr→De test set. English candidates (“→English” errors) are colored in red, while French candidates (“→Source” errors) are colored in blue. Final translations (at step 7) are in bold, where $b = 5$ generates a German translation, and $b = 20$ generates an off-target English translation at the 7th step.