Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation

Biao Zhang[1] **Philip Williams**[1] **Ivan Titov**[1,2] **Rico Sennrich**[3,1]

[1]School of Informatics, University of Edinburgh
[2]ILLC, University of Amsterdam
[3]Department of Computational Linguistics, University of Zurich
B.Zhang@ed.ac.uk, {pwillia4,ititov}@inf.ed.ac.uk, sennrich@cl.uzh.ch

Abstract

Massively multilingual models for neural machine translation (NMT) are theoretically attractive, but often underperform bilingual models and deliver poor zero-shot translations. In this paper, we explore ways to improve them. We argue that multilingual NMT requires stronger modeling capacity to support language pairs with varying typological characteristics, and overcome this bottleneck via language-specific components and deepening NMT architectures. We identify the off-target translation issue (i.e. translating into a wrong target language) as the major source of the inferior zero-shot performance, and propose random online backtranslation to enforce the translation of unseen training language pairs. Experiments on OPUS-100 (a novel multilingual dataset with 100 languages) show that our approach substantially narrows the performance gap with bilingual models in both one-to-many and many-to-many settings, and improves zero-shot performance by ∼10 BLEU, approaching conventional pivot-based methods.[1]

1 Introduction

With the great success of neural machine translation (NMT) on bilingual datasets (Bahdanau et al., 2015; Vaswani et al., 2017; Barrault et al., 2019), there is renewed interest in multilingual translation where a single NMT model is optimized for the translation of multiple language pairs (Firat et al., 2016a; Johnson et al., 2017; Lu et al., 2018; Aharoni et al., 2019). Multilingual NMT eases model deployment and can encourage knowledge transfer among related language pairs (Lakew et al., 2018; Tan et al., 2019), improve low-resource translation (Ha et al., 2016; Arivazhagan et al., 2019b), and enable zero-shot translation (i.e. direct translation between a language pair never seen in training) (Firat et al., 2016b; Johnson et al., 2017; Al-Shedivat and Parikh, 2019; Gu et al., 2019).

Source	Jusqu'à ce qu'on trouve le moment clé, celui où tu as su que tu l'aimais.
Reference	Bis wir den unverkennbaren Moment gefunden haben, den Moment, wo du wusstest, du liebst ihn.
Zero-Shot	Jusqu'à ce qu'on trouve le moment clé, celui où tu as su que tu l'aimais.
Source	Les États membres ont été consultés et ont approuvé cette proposition.
Reference	Die Mitgliedstaaten wurden konsultiert und sprachen sich für diesen Vorschlag aus.
Zero-Shot	Les Member States have been consultedés and have approved this proposal.

Table 1: Illustration of the off-target translation issue with French→German zero-shot translations with a multilingual NMT model. Our baseline multilingual NMT model often translates into the wrong language for zero-shot language pairs, such as copying the source sentence or translating into English rather than German.

Despite these potential benefits, multilingual NMT tends to underperform its bilingual counterparts (Johnson et al., 2017; Arivazhagan et al., 2019b) and results in considerably worse translation performance when many languages are accommodated (Aharoni et al., 2019). Since multilingual NMT must distribute its modeling capacity between different translation directions, we ascribe this deteriorated performance to the deficient capacity of single NMT models and seek solutions that are capable of overcoming this capacity bottleneck. We propose language-aware layer normalization and linear transformation to relax the representation constraint in multilingual NMT models. The linear transformation is inserted in-between the encoder and the decoder so as to facilitate the induction of language-specific translation correspon-

[1]We release our code at https://github.com/bzhangGo/zero. We release the OPUS-100 dataset at https://github.com/EdinburghNLP/opus-100-corpus.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1628–1639
July 5 - 10, 2020.

dences. We also investigate deep NMT architectures (Wang et al., 2019a; Zhang et al., 2019) aiming at further reducing the performance gap with bilingual methods.

Another pitfall of massively multilingual NMT is its poor zero-shot performance, particularly compared to pivot-based models. Without access to parallel training data for zero-shot language pairs, multilingual models easily fall into the trap of *off-target translation* where a model ignores the given target information and translates into a wrong language as shown in Table 1. To avoid such a trap, we propose the random online backtranslation (ROBT) algorithm. ROBT finetunes a pretrained multilingual NMT model for unseen training language pairs with pseudo parallel batches generated by back-translating the target-side training data.[2] We perform backtranslation (Sennrich et al., 2016a) into randomly picked intermediate languages to ensure good coverage of ∼10,000 zero-shot directions. Although backtranslation has been successfully applied to zero-shot translation (Firat et al., 2016b; Gu et al., 2019; Lakew et al., 2019), whether it works in the massively multilingual set-up remained an open question and we investigate it in our work.

For experiments, we collect OPUS-100, a massively multilingual dataset sampled from OPUS (Tiedemann, 2012). OPUS-100 consists of 55M English-centric sentence pairs covering 100 languages. As far as we know, no similar dataset is publicly available.[3] We have released OPUS-100 to facilitate future research.[4] We adopt the Transformer model (Vaswani et al., 2017) and evaluate our approach under one-to-many and many-to-many translation settings. Our main findings are summarized as follows:

- Increasing the capacity of multilingual NMT yields large improvements and narrows the performance gap with bilingual models. Low-resource translation benefits more from the increased capacity.
- Language-specific modeling and deep NMT architectures can slightly improve zero-shot translation, but fail to alleviate the off-target translation issue.
- Finetuning multilingual NMT with ROBT substantially reduces the proportion of off-target translations (by ∼50%) and delivers an improvement of ∼10 BLEU in zero-shot settings, approaching the conventional pivot-based method. We show that finetuning with ROBT converges within a few thousand steps.

2 Related Work

Pioneering work on multilingual NMT began with multitask learning, which shared the encoder for one-to-many translation (Dong et al., 2015) or the attention mechanism for many-to-many translation (Firat et al., 2016a). These methods required a dedicated encoder or decoder for each language, limiting their scalability. By contrast, Lee et al. (2017) exploited character-level inputs and adopted a shared encoder for many-to-one translation. Ha et al. (2016) and Johnson et al. (2017) further successfully trained a single NMT model for multilingual translation with a target language symbol guiding the translation direction. This approach serves as our baseline. Still, this paradigm forces different languages into one joint representation space, neglecting their linguistic diversity. Several subsequent studies have explored different strategies to mitigate this representation bottleneck, ranging from reorganizing parameter sharing (Blackwood et al., 2018; Sachan and Neubig, 2018; Lu et al., 2018; Wang et al., 2019c; Vázquez et al., 2019), designing language-specific parameter generators (Platanios et al., 2018), decoupling multilingual word encodings (Wang et al., 2019b) to language clustering (Tan et al., 2019). Our language-specific modeling continues in this direction, but with a special focus on broadening normalization layers and encoder outputs.

Multilingual NMT allows us to perform zero-shot translation, although the quality is not guaranteed (Firat et al., 2016b; Johnson et al., 2017). We observe that multilingual NMT often translates into the wrong target language on zero-shot directions (Table 1), resonating with the 'missing ingredient problem' (Arivazhagan et al., 2019a) and the spurious correlation issue (Gu et al., 2019). Approaches to improve zero-shot performance fall into two categories: 1) developing novel cross-lingual regularizers, such as the alignment regularizer (Arivazhagan et al., 2019a) and the consistency regularizer (Al-

[2]Note that backtranslation actually converts the zero-shot problem into a zero-resource problem. We follow previous work and continue referring to *zero-shot* translation, even when using synthetic training data.

[3]Previous studies (Aharoni et al., 2019; Arivazhagan et al., 2019b) adopt in-house data which was not released.

[4]https://github.com/EdinburghNLP/opus-100-corpus

Shedivat and Parikh, 2019); and 2) generating artificial parallel data with backtranslation (Firat et al., 2016b; Gu et al., 2019; Lakew et al., 2019) or pivot-based translation (Currey and Heafield, 2019). The proposed ROBT algorithm belongs to the second category. In contrast to Gu et al. (2019) and Lakew et al. (2019), however, we perform online back-translation for each training step with randomly selected intermediate languages. ROBT avoids decoding the whole training set for each zero-shot language pair and can therefore scale to massively multilingual settings.

Our work belongs to a line of research on massively multilingual translation (Aharoni et al., 2019; Arivazhagan et al., 2019b). Aharoni et al. (2019) demonstrated the feasibility of massively multilingual NMT and reported encouraging results. We continue in this direction by developing approaches that improve both multilingual and zero-shot performance. Independently from our work, Arivazhagan et al. (2019b) also find that increasing model capacity with deep architectures (Wang et al., 2019a; Zhang et al., 2019) substantially improves multilingual performance. A concurrent related work is (Bapna and Firat, 2019), which introduces task-specific and lightweight adaptors for fast and scalable model adaptation. Compared to these adaptors, our language-aware layers are jointly trained with the whole NMT model from scratch without relying on any pretraining.

3 Multilingual NMT

We briefly review the multilingual approach (Ha et al., 2016; Johnson et al., 2017) and the Transformer model (Vaswani et al., 2017), which are used as our baseline. Johnson et al. (2017) rely on prepending tokens specifying the target language to each source sentence. In that way a single NMT model can be trained on the modified multilingual dataset and used to perform multilingual translation. Given a source sentence $\mathbf{x}=(x_1, x_2, \ldots, x_{|\mathbf{x}|})$, its target reference $\mathbf{y}=(y_1, y_2, \ldots, y_{|\mathbf{y}|})$ and the target language token t[5], multilingual NMT translates under the encoder-decoder framework (Bahdanau et al., 2015):

$$\mathbf{H} = \text{Encoder}([t, \mathbf{x}]), \quad (1)$$

$$\mathbf{S} = \text{Decoder}(\mathbf{y}, \mathbf{H}), \quad (2)$$

[5] t is in the form of "<2X>" where X is a language name, such as <2EN> meaning *translating into English*.

where $\mathbf{H} \in \mathbb{R}^{|\mathbf{x}| \times d}/\mathbf{S} \in \mathbb{R}^{|\mathbf{y}| \times d}$ denote the encoder/decoder output. d is the model dimension.

We employ the Transformer (Vaswani et al., 2017) as the backbone NMT model due to its superior multilingual performance (Lakew et al., 2018). The encoder is a stack of $L = 6$ identical layers, each containing a self-attention sublayer and a point-wise feedforward sublayer. The decoder follows a similar structure, except for an extra cross-attention sublayer used to condition the decoder on the source sentence. Each sublayer is equipped with a residual connection (He et al., 2015), followed by layer normalization (Ba et al., 2016, $\text{LN}(\cdot)$):

$$\bar{\mathbf{a}} = \text{LN}(\mathbf{a} \mid \mathbf{g}, \mathbf{b}) = \frac{\mathbf{a} - \mu}{\sigma} \odot \mathbf{g} + \mathbf{b}, \quad (3)$$

where $\odot$ denotes element-wise multiplication, μ and σ are the mean and standard deviation of the input vector $\mathbf{a} \in \mathbb{R}^d$, respectively. $\mathbf{g} \in \mathbb{R}^d$ and $\mathbf{b} \in \mathbb{R}^d$ are model parameters. They control the sharpness and location of the regularized layer output $\bar{\mathbf{a}}$. Layer normalization has proven effective in accelerating model convergence (Ba et al., 2016).

4 Approach

Despite its success, multilingual NMT still suffers from 1) *insufficient modeling capacity*, where including more languages results in reduction in translation quality (Aharoni et al., 2019); and 2) *off-target translation*, where models translate into a wrong target language on zero-shot directions (Arivazhagan et al., 2019a). These drawbacks become severe in massively multilingual settings and we explore approaches to alleviate them. We hypothesize that the vanilla Transformer has insufficient capacity and search for model-level strategies such as deepening Transformer and devising language-specific components. By contrast, we regard the lack of parallel data as the reason behind the off-target issue. We resort to data-level strategy by creating, in online fashion, artificial parallel training data for each zero-shot language pair in order to encourage its translation.

Deep Transformer One natural way to improve the capacity is to increase model depth. Deeper neural models are often capable of inducing more generalizable ('abstract') representations and capturing more complex dependencies and have shown encouraging performance on bilingual translation (Bapna et al., 2018; Zhang et al., 2019; Wang

et al., 2019a). We adopt the depth-scaled initialization method (Zhang et al., 2019) to train a deep Transformer for multilingual translation.

Language-aware Layer Normalization Regardless of linguistic differences, layer normalization in multilingual NMT simply constrains all languages into one joint Gaussian space, which makes learning more difficult. We propose to relax this restriction by conditioning the normalization on the given target language token t (LaLN for short) as follows:

$$\bar{\mathbf{a}} = \text{LN}(\mathbf{a} \mid \mathbf{g}_t, \mathbf{b}_t). \quad (4)$$

We apply this formula to all normalization layers, and leave the study of conditioning on source language information for the future.

Language-aware Linear Transformation Different language pairs have different translation correspondences or word alignments (Koehn, 2010). In addition to LaLN, we introduce a target-language-aware linear transformation (LaLT for short) between the encoder and the decoder to enhance the freedom of multilingual NMT in expressing flexible translation relationships. We adapt Eq. (2) as follows:

$$\mathbf{S} = \text{Decoder}(\mathbf{y}, \mathbf{H}\mathbf{W}_t), \quad (5)$$

where $\mathbf{W}_t \in \mathbb{R}^{d \times d}$ denotes model parameters. Note that adding one more target language in LaLT brings in only one weight matrix.[6] Compared to existing work (Firat et al., 2016b; Sachan and Neubig, 2018), LaLT reaches a better trade-off between expressivity and scalability.

Random Online Backtranslation Prior studies on backtranslation for zero-shot translation decode the whole training set for each zero-shot language pair (Gu et al., 2019; Lakew et al., 2019), and scalability to massively multilingual translation is questionable – in our setting, the number of zero-shot translation directions is 9702.

We address scalability by performing online backtranslation paired with randomly sampled intermediate languages. Algorithm 1 shows the detail of RoBT, where for each training instance $(\mathbf{x}_k, \mathbf{y}_k, t_k)$, we uniformly sample an intermediate language t'_k ($t_k \neq t'_k$), back-translate $\mathbf{y}_k$ into t'_k to obtain $\mathbf{x}'_k$, and train on the new instance $(\mathbf{x}'_k, \mathbf{y}_k, t_k)$. Although $\mathbf{x}'_k$ may be poor initially (translations are produced on-line by the model being trained), RoBT still benefits from the translation signal of $t'_k \rightarrow t_k$. To reduce the computational cost, we implement batch-based greedy decoding for line 7.

[6] We also attempted to factorize $\mathbf{W}_t$ into smaller matrices/vectors to reduce the number of parameters. Unfortunately, the final performance was rather disappointing.

Algorithm 1: Algorithm for Random Online Backtranslation

Input : Multilingual training data, D;
Pretrained multilingual model, M;
Maximum finetuning step, N;
Finetuning batch size, B;
Target language set, $\mathcal{T}$;
Output: Zero-shot enabled model, M

1 $i \leftarrow 0$
2 **while** $i \leq N \wedge$ *not converged* **do**
3 $\mathcal{B} \leftarrow$ sample batch from D
4 **for** $k \leftarrow 1$ ***to*** B **do**
5 $(\mathbf{x}_k, \mathbf{y}_k, t_k) \leftarrow \mathcal{B}_k$
6 $t'_k \sim \text{Uniform}(\mathcal{T})$ such that $t'_k \neq t_k$
7 $\mathbf{x}'_k \leftarrow M([t'_k, \mathbf{y}_k])$
 `// backtrans` $t_k \rightarrow t'_k$ `to produce training example for` $t'_k \rightarrow t_k$
8 $\mathcal{B} \leftarrow \mathcal{B} \cup (\mathbf{x}'_k, \mathbf{y}_k, t_k)$
9 Optimize M using $\mathcal{B}$
10 $i \leftarrow i + 1$
11 **return** M

5 OPUS-100

Recent work has scaled up multilingual NMT from a handful of languages to tens or hundreds, with many-to-many systems being capable of translation in thousands of directions. Following Aharoni et al. (2019), we created an English-centric dataset, meaning that all training pairs include English on either the source or target side. Translation for any language pair that does not include English is zero-shot or must be pivoted through English.

We created OPUS-100 by sampling data from the OPUS collection (Tiedemann, 2012). OPUS-100 is at a similar scale to Aharoni et al. (2019)'s, with 100 languages (including English) on both sides and up to 1M training pairs for each language pair. We selected the languages based on the volume of parallel data available in OPUS.

The OPUS collection is comprised of multiple corpora, ranging from movie subtitles to GNOME

ID	Model Architecture	L	#Param	BLEU_{94}	WR	BLEU_4
1	Transformer, Bilingual	6	106M	-	-	20.90
2	Transformer, Bilingual	12	150M	-	-	**22.75**
3	Transformer	6	106M	24.64	*ref*	18.95
4	3 + MATT	6	99M	23.81	20.2	17.95
5	4 + LALN	6	102M	24.22	28.7	18.50
6	4 + LALT	6	126M	27.11	72.3	20.28
7	4 + LALN + LALT	6	129M	27.18	75.5	20.08
8	4	12	137M	25.69	81.9	19.13
9	7	12	169M	28.04	91.5	19.93
10	7	24	249M	**29.60**	**92.6**	21.23

Table 2: Test BLEU for one-to-many translation on OPUS-100 (100 languages). "*Bilingual*": bilingual NMT, "L": model depth (for both encoder and decoder), "*#Param*": parameter number, "*WR*": win ratio (%) compared to *ref* (③), MATT: the merged attention (Zhang et al., 2019). LALN and LALT denote the proposed language-aware layer normalization and linear transformation, respectively. "$BLEU_{94}$/$BLEU_4$": average BLEU over all 94 translation directions in test set and En→De/Zh/Br/Te, respectively. Higher BLEU and WR indicate better result. Best scores are highlighted in **bold**.

documentation to the Bible. We did not curate the data or attempt to balance the representation of different domains, instead opting for the simplest approach of downloading all corpora for each language pair and concatenating them. We randomly sampled up to 1M sentence pairs per language pair for training, as well as 2000 for validation and 2000 for testing.[7] To ensure that there was no overlap (at the monolingual sentence level) between the training and validation/test data, we applied a filter during sampling to exclude sentences that had already been sampled. Note that this was done cross-lingually, so an English sentence in the Portuguese-English portion of the training data could not occur in the Hindi-English test set, for instance.

OPUS-100 contains approximately 55M sentence pairs. Of the 99 language pairs, 44 have 1M sentence pairs of training data, 73 have at least 100k, and 95 have at least 10k.

To evaluate zero-shot translation, we also sampled 2000 sentence pairs of test data for each of the 15 pairings of Arabic, Chinese, Dutch, French, German, and Russian. Filtering was used to exclude sentences already in OPUS-100.

6 Experiments

6.1 Setup

We perform one-to-many (English-X) and many-to-many (English-X ∪ X-English) translation on OPUS-100 ($|\mathcal{T}|$ is 100). We apply byte pair encoding (BPE) (Sennrich et al., 2016b; Kudo and Richardson, 2018) to handle multilingual words with a joint vocabulary size of 64k. We randomly shuffle the training set to mix instances of different language pairs. We adopt BLEU (Papineni et al., 2002) for translation evaluation with the toolkit SacreBLEU (Post, 2018)[8]. We employ the *langdetect* library[9] to detect the language of translations, and measure the translation-language accuracy for zero-shot cases. Rather than providing numbers for each language pair, we report average BLEU over all 94 language pairs with test sets (BLEU_{94}). We also show the win ratio (WR), counting the proportion where our approach outperforms its baseline.

Apart from multilingual NMT, our baselines also involve bilingual NMT and pivot-based translation (only for zero-shot comparison). We select four typologically different target languages (German/De, Chinese/Zh, Breton/Br, Telugu/Te) with varied training data size for comparison to bilingual models as applying bilingual NMT to each language pair is resource-consuming. We report average BLEU over these four languages as BLEU_4. We reuse the multilingual BPE vocabulary for bilingual NMT.

We train all NMT models with the Transformer base settings (512/2048, 8 heads) (Vaswani et al., 2017). We pair our approaches with the merged attention (MATT) (Zhang et al., 2019) to reduce training time. Other details about model settings are in the Appendix.

6.2 Results on One-to-Many Translation

Table 2 summarizes the results. The inferior performance of multilingual NMT (③) against its

[7] For efficiency, we only use 200 sentences per language pair for validation in our multilingual experiments.

[8] Signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.1

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

ID	Model Architecture	L	#Param	w/o ROBT			w/ ROBT		
				$BLEU_{94}$	WR	$BLEU_4$	$BLEU_{94}$	WR	$BLEU_4$
1	Transformer, Bilingual	6	110M	-	-	**20.28**	-	-	-
2	Transformer	6	110M	19.50	*ref*	15.35	18.75	4.3	14.73
3	2 + MATT	6	103M	18.49	5.3	14.90	17.85	6.4	14.38
4	3 + LALN + LALT	6	133M	21.39	78.7	18.13	20.81	69.1	17.45
5	3	12	141M	20.77	94.7	16.08	20.24	84.0	15.80
6	4	12	173M	22.86	97.9	19.25	22.39	97.9	18.23
7	4	24	254M	**23.96**	**100.0**	19.83	23.36	97.9	19.45

Table 3: English→X test BLEU for many-to-many translation on OPUS-100 (100 languages). "*WR*": win ratio (%) compared to *ref* (② w/o ROBT). ROBT denotes the proposed random online backtranslation method.

ID	Model Architecture	L	#Param	w/o ROBT			w/ ROBT		
				$BLEU_{94}$	WR	$BLEU_4$	$BLEU_{94}$	WR	$BLEU_4$
1	Transformer, Bilingual	6	110M	-	-	21.23	-	-	-
2	Transformer	6	110M	27.60	*ref*	23.35	27.02	14.9	22.50
3	2 + MATT	6	103M	26.90	2.1	22.78	26.28	4.3	21.53
4	3 + LALN + LALT	6	133M	27.50	37.2	23.05	27.22	23.4	23.30
5	3	12	141M	29.15	**98.9**	24.15	28.80	91.5	24.03
6	4	12	173M	29.49	97.9	24.53	29.54	96.8	25.43
7	4	24	254M	**31.36**	**98.9**	26.03	30.98	95.7	**26.78**

Table 4: X→English test BLEU for many-to-many translation on OPUS-100 (100 languages). "*WR*": win ratio (%) compared to *ref* (② w/o ROBT).

bilingual counterpart (①) reflects the capacity issue (-1.95 $BLEU_4$). Replacing the self-attention with MATT slightly deteriorates performance (-0.83 $BLEU_{94}$ ③→④); we still use MATT for more efficiently training deep models.

Our ablation study (④-⑦) shows that enriching the language awareness in multilingual NMT substantially alleviates this capacity problem. Relaxing the normalization constraints with LALN gains 0.41 $BLEU_{94}$ with 8.5% WR (④→⑤). Decoupling different translation relationships with LALT delivers an improvement of 3.30 $BLEU_{94}$ and 52.1% WR (④→⑥). Combining LALT and LALN demonstrates their complementarity (+3.37 $BLEU_{94}$ and +55.3% WR, ④→⑦), significantly outperforming the multilingual baseline (+2.54 $BLEU_{94}$, ③→⑦), albeit still behind the bilingual models (-0.82 $BLEU_4$, ①→⑦).

Deepening the Transformer also improves the modeling capacity (+1.88 $BLEU_{94}$, ④→⑧). Although deep Transformer performs worse than LALN+LALT under a similar number of model parameters in terms of BLEU (-1.49 $BLEU_{94}$, ⑦→⑧), it shows more consistent improvements across different language pairs (+6.4% WR). We obtain better performance when integrating all approaches (⑨). By increasing the model depth to 24 (⑩), Transformer with our approach yields a score of 29.60 $BLEU_{94}$ and 21.23 $BLEU_4$, beating the baseline (③) on 92.6% tasks and outperforming the base bilingual model (①) by 0.33 $BLEU_4$. Our approach significantly narrows the performance gap between multilingual NMT and bilingual NMT (20.90 $BLEU_4$ → 21.23 $BLEU_4$, ①→⑩), although similarly deepening bilingual models surpasses our approach by 1.52 $BLEU_4$ (⑩→②).

6.3 Results on Many-to-Many Translation

We train many-to-many NMT models on the concatenation of the one-to-many dataset (English→X) and its reversed version (X→English), and evaluate the zero-shot performance on X→X language pairs. Table 3 and Table 4 show the translation results for English→X and X→English, respectively.[10] We focus on the translation performance w/o ROBT in this subsection.

Compared to the one-to-many translation, the many-to-many translation must accommodate twice as many translation directions. We observe that many-to-many NMT models suffer more se-

[10]Note that the one-to-many training and test sets were not yet aggressively filtered for sentence overlap as described in Section 5, so results in Table 2 and Table 3 are not directly comparable.

ID	Model Architecture	L	#Param	English→X			X→English		
				High	Med	Low	High	Med	Low
1	Transformer	6	110M	20.69	20.82	15.18	26.99	28.60	27.49
2	1 + MATT	6	103M	19.70	19.77	14.17	26.32	27.81	26.84
3	2 + LALN + LALT	6	133M	21.07	22.88	19.99	27.03	28.60	26.97
4	2	12	141M	21.67	22.17	16.95	28.39	30.24	29.26
5	3	12	173M	22.48	24.38	21.58	28.66	30.73	29.50
6	3	24	254M	**23.69**	**25.61**	**22.24**	**30.29**	**32.58**	**31.90**

Table 5: Test BLEU for High/Medium/Low (*High/Med/Low*) resource language pairs in many-to-many setting on OPUS-100 (100 languages). We report average BLEU for each category.

ID	Model Architecture	L	#Param	w/o ROBT		w/ ROBT	
				BLEU_{zero}	ACC_{zero}	BLEU_{zero}	ACC_{zero}
1	Transformer, Pivot & Bilingual	6	110M	12.98	84.87	-	-
2	Transformer	6	110M	3.97	36.04	10.11	86.08
3	2 + MATT	6	103M	3.49	31.62	9.67	85.87
4	3 + LALN + LALT	6	133M	4.02	45.43	11.23	87.40
5	3	12	141M	4.71	39.40	11.87	87.44
6	4	12	173M	5.41	51.40	12.62	**87.99**
7	4	24	254M	5.24	47.91	14.08	87.68
8	7 + Pivot	24	254M	14.71	84.81	**14.78**	85.09

Table 6: Test BLEU and translation-language accuracy for zero-shot translation in many-to-many setting on OPUS-100 (100 languages). "*BLEU_{zero}/ACC_{zero}*": average BLEU/accuracy over all zero-shot translation directions in test set, "*Pivot*": the pivot-based translation that first translates one source sentence into English (X→English NMT), and then into the target language (English→X NMT). Lower accuracy indicates severe off-target translation. The average Pearson correlation coefficient between language accuracy and the corresponding BLEU is 0.93 (significant at $p < 0.01$).

rious capacity issues on English→X tasks (-4.93 BLEU_4, ①→② in Table 3 versus -1.95 BLEU_4 in Table 2), where the deep Transformer with LALN + LALT effectively reduces this gap to -0.45 BLEU_4 (①→⑦, Table 3), resonating with our findings from Table 2. By contrast, multilingual NMT benefits X→English tasks considerably from the multitask learning alone, outperforming bilingual NMT by 2.13 BLEU_4 (①→②, Table 4). Enhancing model capacity further enlarges this margin to +4.80 BLEU_4 (①→⑦, Table 4).

We find that the overall quality of English→X translation (19.50/23.96 BLEU_{94}, ②/⑦, Table 3) lags far behind that of its X→English counterpart (27.60/31.36 BLEU_{94}, ②/⑫, Table 4), regardless of the modeling capacity. We ascribe this to the highly skewed training data distribution, where half of the training set uses English as the target. This strengthens the ability of the decoder to translate into English, and also encourages knowledge transfer for X→English language pairs. LALN and LALT show the largest benefit for English→X (+2.9 BLEU_{94}, ③→④, Table 3), and only a small benefit for X→English (+0.6 BLEU_{94}, ③→④, Table 4). This makes sense considering that LALN and LALT are specific to the target language, so capacity is mainly increased for English→X. Deepening the Transformer yields benefits in both directions (+2.57 BLEU_{94} for English→X, +3.86 BLEU_{94} for X→English; ④→⑦, Tables 3 and 4).

6.4 Effect of Training Corpus Size

Our multilingual training data is distributed unevenly across different language pairs, which could affect the knowledge transfer delivered by language-aware modeling and deep Transformer in multilingual translation. We investigate this effect by grouping different language pairs in OPUS-100 into three categories according to their training data size: High ($\geq$ 0.9M, 45), Low ($<$ 0.1M, 18) and Medium (others, 31). Table 5 shows the results.

Language-aware modeling benefits low-resource language pairs the most on English→X translation (+5.82 BLEU, Low versus +1.37/+3.11 BLEU, High/Med, ②→③), but has marginal impact on X→English translation as analyzed in Section 6.3. By contrast, deep Transformers yield similar benefits across different data scales (+2.38 average BLEU, English→X and +2.31 average BLEU, X→English, ②→④). We obtain the best perfor-

mance by integrating both (①→⑥) with a clear positive transfer to low-resource language pairs.

6.5 Results on Zero-Shot Translation

Previous work shows that a well-trained multilingual model can do zero-shot X→Y translation directly (Firat et al., 2016b; Johnson et al., 2017). Our results in Table 6 reveal that the translation quality is rather poor (3.97 BLEU_{zero}, ② w/o ROBT) compared to the pivot-based bilingual baseline (12.98 BLEU_{zero}, ①) under the massively multilingual setting (Aharoni et al., 2019), although translations into different target languages show varied performance. The marginal gain by the deep Transformer with LALN + LALT (+1.44 BLEU_{zero}, ②→⑥, w/o ROBT) suggests that weak model capacity is not the major cause of this inferior performance.

In a manual analysis on the zero-shot NMT outputs, we found many instances of off-target translation (Table 1). We use translation-language accuracy to measure the proportion of translations that are in the correct target language. Results in Table 6 show that there is a huge accuracy gap between the multilingual and the pivot-based method (-48.83% ACC_{zero}, ①→②, w/o ROBT), from which we conclude that the off-target translation issue is one source of the poor zero-shot performance.

We apply ROBT to multilingual models by finetuning them for an extra 100k steps with the same batch size as for training. Table 6 shows that ROBT substantially improves ACC_{zero} by 35%∼50%, reaching 85%∼87% under different model settings. The multilingual Transformer with ROBT achieves a translation improvement of up to 10.11 BLEU_{zero} (② w/o ROBT→⑦ w/ ROBT), outperforming the bilingual baseline by 1.1 BLEU_{zero} (① w/o ROBT→⑦ w/ ROBT) and approaching the pivot-based multilingual baseline (-0.63 BLEU_{zero}, ⑧ w/o ROBT→⑦ w/ ROBT).[11] The strong Pearson correlation between the accuracy and BLEU (0.92 on average, significant at $p < 0.01$) suggests that the improvement on the off-target translation issue explains the increased translation performance to a large extent.

Results in Table 3 and 4 show that ROBT's success on zero-shot translation comes at the cost of sacrificing ∼0.50 BLEU_{94} and ∼4% WR on English→X and X→English translation. We also note that models with more capacity yield higher language accuracy (+7.78%/+13.81% ACC_{zero}, ③→⑤/③→④, w/o ROBT) and deliver better zero-shot performance before (+1.22/+0.53 BLEU_{zero}, ③→⑤/③→④, w/o ROBT) and after ROBT (+2.20/+1.56 BLEU_{zero}, ③→⑤/③→④, w/ ROBT). In other words, increasing the modeling capacity benefits zero-shot translation and improves robustness.

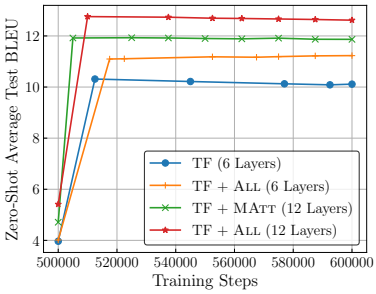

Figure 1: Zero-shot average test BLEU for multilingual NMT models finetuned by ROBT. ALL = MATT + LALN + LALT. Multilingual models with ROBT quickly converge on zero-shot directions.

Setting	BLEU_{zero}
6-to-6	11.98
100-to-100	11.23

Table 7: Zero-short translation quality for ROBT under different settings. "*100-to-100*": the setting used in the above experiments; we set $\mathcal{T}$ to all target languages. "*6-to-6*": $\mathcal{T}$ only includes the zero-shot languages in the test set. We employ 6-layer Transformer with LALN and LALT for experiments.

Convergence of ROBT. Unlike prior studies (Gu et al., 2019; Lakew et al., 2019), we resort to an online method for backtranslation. The curve in Figure 1 shows that ROBT is very effective, and takes only a few thousand steps to converge, suggesting that it is unnecessary to decode the whole training set for each zero-shot language pair. We leave it to future work to explore whether different back-translation strategies (other than greedy decoding) will deliver larger and continued benefits with ROBT.

Impact of $\mathcal{T}$ *on* ROBT. ROBT heavily relies on $\mathcal{T}$, the set of target languages considered, to distribute the modeling capacity on zero-shot directions. To study its impact, we provide a comparison by constraining $\mathcal{T}$ to 6 languages in the zero-shot test set. Results in Table 7 show that the biased ROBT outperforms the baseline by 0.75 BLEU_{zero}. By narrowing $\mathcal{T}$, more capacity is scheduled to the focused languages, which results in performance improvements. But the small scale of this improve-

[11] Note that ROBT improves all zero-shot directions due to its randomness in sampling the intermediate languages. We do not bias ROBT to the given zero-shot test set.

ment suggests that the number of zero-shot directions is not ROBT's biggest bottleneck.

7 Conclusion and Future Work

This paper explores approaches to improve massively multilingual NMT, especially on zero-shot translation. We show that multilingual NMT suffers from weak capacity, and propose to enhance it by deepening the Transformer and devising language-aware neural models. We find that multilingual NMT often generates off-target translations on zero-shot directions, and propose to correct it with a random online backtranslation algorithm. We empirically demonstrate the feasibility of backtranslation in massively multilingual settings to allow for massively zero-shot translation for the first time. We release OPUS-100, a multilingual dataset from OPUS including 100 languages with around 55M sentence pairs for future study. Our experiments on this dataset show that the proposed approaches substantially increase translation performance, narrowing the performance gap with bilingual NMT models and pivot-based methods.

In the future, we will develop lightweight alternatives to LALT to reduce the number of model parameters. We will also exploit novel strategies to break the upper bound of ROBT and obtain larger zero-shot improvements, such as generative modeling (Zhang et al., 2016; Su et al., 2018; García et al., 2020; Zheng et al., 2020).

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements 825460 (ELITR) and 825299 (GoURMET). This project has received support from Samsung Electronics Polska sp. z o.o. - Samsung R&D Institute Poland. Rico Sennrich acknowledges support of the Swiss National Science Foundation (MUTAMUR; no. 176727).

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.

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot 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 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ankur Bapna, Mia Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. 2018. Training deeper neural machine translation models with transparent attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3033, Brussels, Belgium. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational*

Linguistics, pages 3112–3122, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Xavier García, Pierre Forêt, Thibault Sellam, and Ankur P. Parikh. 2020. A multilingual view of unsupervised machine translation. *ArXiv*, abs/2002.02955.

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.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*, Seattle, USA.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

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. In *International Conference on Learning Representations*.

Philipp Koehn. 2010. *Statistical Machine Translation*, 1st edition. Cambridge University Press, New York, NY, USA.

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.

Surafel M. Lakew, Marcello Federico, Matteo Negri, and Marco Turchi. 2019. Multilingual Neural Machine Translation for Zero-Resource Languages. *arXiv e-prints*, page arXiv:1909.07342.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435, Brussels, Belgium. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. 2018. Variational recurrent neural machine translation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao QIN, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2019. Multilingual NMT with a language-independent attention bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019a. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019b. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations*.

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019c. A compact and language-sensitive multilingual translation method. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1213–1223, Florence, Italy. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019. Improving deep transformer with depth-scaled initialization and merged attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 898–909, Hong Kong, China. Association for Computational Linguistics.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. Association for Computational Linguistics.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xin-Yu Dai, and Jiajun Chen. 2020. Mirror-generative neural machine translation. In *International Conference on Learning Representations*.

A OPUS-100: The OPUS Multilingual Dataset

Table 8 lists the languages (other than English) and numbers of sentence pairs in the English-centric multilingual dataset.

B Model Settings

We optimize model parameters using Adam ($\beta_1 = 0.9, \beta_2 = 0.98$) (Kingma and Ba, 2015) with label smoothing of 0.1 and scheduled learning rate (warmup step 4k). We set the initial learning rate to 1.0 for bilingual models, but use 0.5 for multilingual models in order to stabilize training. We apply dropout to residual layers and attention weights, with a rate of 0.1/0.1 for 6-layer Transformer models and 0.3/0.2 for deeper ones. We group sentence

Table 8: Numbers of training, validation, and test sentence pairs in the English-centric multilingual dataset.

Language		Train	Valid	Test
af	Afrikaans	275512	2000	2000
am	Amharic	89027	2000	2000
an	Aragonese	6961	0	0
ar	Arabic	1000000	2000	2000
as	Assamese	138479	2000	2000
az	Azerbaijani	262089	2000	2000
be	Belarusian	67312	2000	2000
bg	Bulgarian	1000000	2000	2000
bn	Bengali	1000000	2000	2000
br	Breton	153447	2000	2000
bs	Bosnian	1000000	2000	2000
ca	Catalan	1000000	2000	2000
cs	Czech	1000000	2000	2000
cy	Welsh	289521	2000	2000
da	Danish	1000000	2000	2000
de	German	1000000	2000	2000
dz	Dzongkha	624	0	0
el	Greek	1000000	2000	2000
eo	Esperanto	337106	2000	2000
es	Spanish	1000000	2000	2000
et	Estonian	1000000	2000	2000
eu	Basque	1000000	2000	2000
fa	Persian	1000000	2000	2000
fi	Finnish	1000000	2000	2000
fr	French	1000000	2000	2000
fy	Western Frisian	54342	2000	2000
ga	Irish	289524	2000	2000
gd	Gaelic	16316	1605	1606
gl	Galician	515344	2000	2000
gu	Gujarati	318306	2000	2000
ha	Hausa	97983	2000	2000
he	Hebrew	1000000	2000	2000
hi	Hindi	534319	2000	2000
hr	Croatian	1000000	2000	2000
hu	Hungarian	1000000	2000	2000
hy	Armenian	7059	0	0
id	Indonesian	1000000	2000	2000
ig	Igbo	18415	1843	1843
is	Icelandic	1000000	2000	2000
it	Italian	1000000	2000	2000
ja	Japanese	1000000	2000	2000
ka	Georgian	377306	2000	2000
kk	Kazakh	79927	2000	2000
km	Central Khmer	111483	2000	2000
kn	Kannada	14537	917	918
ko	Korean	1000000	2000	2000
ku	Kurdish	144844	2000	2000
ky	Kyrgyz	27215	2000	2000
li	Limburgan	25535	2000	2000
lt	Lithuanian	1000000	2000	2000
lv	Latvian	1000000	2000	2000
mg	Malagasy	590771	2000	2000
mk	Macedonian	1000000	2000	2000
ml	Malayalam	822746	2000	2000
mn	Mongolian	4294	0	0
mr	Marathi	27007	2000	2000
ms	Malay	1000000	2000	2000
mt	Maltese	1000000	2000	2000
my	Burmese	24594	2000	2000
nb	Norwegian Bokmål	142906	2000	2000
ne	Nepali	406381	2000	2000
nl	Dutch	1000000	2000	2000
nn	Norwegian Nynorsk	486055	2000	2000
no	Norwegian	1000000	2000	2000
oc	Occitan	35791	2000	2000
or	Oriya	14273	1317	1318
pa	Panjabi	107296	2000	2000
pl	Polish	1000000	2000	2000
ps	Pashto	79127	2000	2000
pt	Portuguese	1000000	2000	2000
ro	Romanian	1000000	2000	2000
ru	Russian	1000000	2000	2000
rw	Kinyarwanda	173823	2000	2000
se	Northern Sami	35907	2000	2000
sh	Serbo-Croatian	267211	2000	2000
si	Sinhala	979109	2000	2000
sk	Slovak	1000000	2000	2000
sl	Slovenian	1000000	2000	2000
sq	Albanian	1000000	2000	2000
sr	Serbian	1000000	2000	2000
sv	Swedish	1000000	2000	2000
ta	Tamil	227014	2000	2000
te	Telugu	64352	2000	2000
tg	Tajik	193882	2000	2000
th	Thai	1000000	2000	2000
tk	Turkmen	13110	1852	1852
tr	Turkish	1000000	2000	2000
tt	Tatar	100843	2000	2000
ug	Uighur	72170	2000	2000
uk	Ukrainian	1000000	2000	2000
ur	Urdu	753913	2000	2000
uz	Uzbek	173157	2000	2000
vi	Vietnamese	1000000	2000	2000
wa	Walloon	104496	2000	2000
xh	Xhosa	439671	2000	2000
yi	Yiddish	15010	2000	2000
yo	Yoruba	10375	0	0
zh	Chinese	1000000	2000	2000
zu	Zulu	38616	2000	2000

pairs of roughly 50k target tokens into one training/finetuning batch, except for bilingual models where 25k target tokens are used. We train multilingual and bilingual models for 500k and 100k steps, respectively. We average the last 5 checkpoints for evaluation, and employ beam search for decoding with a beam size of 4 and length penalty of 0.6.