Exploring the Benefits and Limitations of Multilinguality for Non-autoregressive Machine Translation

Sweta Agrawal[1] **and Julia Kreutzer**[2] **and Colin Cherry**[2]
[1]Department of Computer Science, University of Maryland
[2]Google Research
sweagraw@umd.edu, {jkreutzer, colincherry}@google.com

Abstract

Non-autoregressive (NAR) machine translation has recently received significant developments, and now achieves comparable quality with autoregressive (AR) models on some benchmarks, while providing an efficient alternative to AR inference. However, while AR translation is often used to implement multilingual models that benefit from transfer between languages and from improved serving efficiency, multilingual NAR models remain relatively unexplored. Taking Connectionist Temporal Classification (CTC) as an example NAR model and IMPUTER as a semi-NAR model, we present a comprehensive empirical study of multilingual NAR. We test its capabilities with respect to positive transfer between related languages and negative transfer under capacity constraints. As NAR models require distilled training sets, we carefully study the impact of bilingual versus multilingual teachers. Finally, we fit a scaling law for multilingual NAR to determine capacity bottlenecks, which quantifies its performance relative to the AR model as the model scale increases.

1 Introduction

Non-autoregressive (NAR) models generate output tokens in parallel instead of sequentially, reducing potentially expensive inference dependencies. They rely on sequence-level knowledge distillation to reach the quality of autoregressive (AR) models (Gu et al., 2018). As the notion of NAR has expanded to include semi-NAR models that generate their outputs in multiple steps, with each step generating several tokens non-autoregressively (Lee et al., 2018; Ghazvininejad et al., 2019), we have begun to see NAR matching the quality of AR (Saharia et al., 2020). Prior works have benchmarked NAR models for machine translation (MT) on a number of language pairs, but with very few exceptions, the NAR models under test have been bilingual as opposed to multilingual.

Multilingual MT models (Dong et al., 2015; Firat et al., 2017; Johnson et al., 2017), translating between multiple languages, have two major advantages. First, they offer better parameter efficiency than bilingual models via multi-tasking. Second, they are able to transfer knowledge from high-resource languages to low-resource ones. Therefore they have become an attractive solution for expanding the language coverage of AR MT (Aharoni et al., 2019; Fan et al., 2021; Siddhant et al., 2022). The capability of multilingual modeling is a major feature of the AR regime, and it is one that we should seek to maintain in NAR models.

However, it is unclear to what extent the benefits of multilingual AR models transfer to NAR modeling (Caruana, 1997; Arivazhagan et al., 2019). Do related languages help each other as easily (*positive transfer*)? Do unrelated languages interfere with one another more (*negative transfer*)? Furthermore, NAR modeling raises a new issue of multilingual distillation. To retain the training-time efficiency of multilingual modeling, it is crucial that NAR works well with multilingual teachers; otherwise, the prospect of training many bilingual teachers would greatly increase the effective training cost. It may actually be the case that multilingual teachers are better suited than bilingual ones, as the effective capacity reduction may result in less complex (Zhou et al., 2019) and less multi-modal outputs (Gu et al., 2018).

We present an empirical study of multilingual NAR modeling. Taking CTC (Libovický and Helcl, 2018) as our canonical NAR method, and IMPUTER (Saharia et al., 2020) as our canonical semi-NAR model, we study how they respond to multilinguality through a series of "stress-tests", first in a six-language scenario designed to emphasize negative transfer (§4), and then in two-language scenarios designed to emphasize positive transfer under data resource constraints (§5). Lastly, we fit a scaling law for our six-language sce-

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 177–187
Abu Dhabi, December 7–8, 2022.

nario to measure the potential of increasing model sizes (§6). The main findings can be summarized as follows:

1. Multilingual NAR models work equally well whether datasets are distilled from bilingual or multilingual teachers.

2. Multilingual NAR models do benefit from positive transfer in scenarios that encourage it; however, in comparison to AR models, they suffer more from negative transfer and benefit less from positive transfer.

3. The scaling law demonstrates that this trend continues as model size increases.

Our extensive analysis on outputs from the NAR models suggest that they still struggle to generate "valid" tokens with desired output length. Furthermore, our results indicate that scaling up the NAR models is not going to close the gap to multilingual AR, but our analysis points to promising directions for future work throughout the paper.

2 Non-Autoregressive Multilingual NMT

Let, $D^l = (x, y) \in X \times Y$ denote the bilingual corpus of a language pair, l. Given an input sequence x of length T', an AR model (Bahdanau et al., 2015; Vaswani et al., 2017) predicts the target y with length T sequentially based on the conditional distribution $p(y_t \mid y_{<t}, x_{1:T'}; \theta)$. NAR models assume conditional independence in the output token space; that is, they model $p(y_t \mid x_{1:T'}; \phi)$. Due to this conditional independence assumption, training NAR models directly on the true target distribution leads to degraded performance (Gu et al., 2018). Hence, NAR models are typically trained with sequence-level knowledge distillation (Kim and Rush, 2016) to reduce the modeling difficulty.

2.1 Non-Autoregressive NMT with CTC

In this work, we focus on NAR modelling via CTC (Graves et al., 2006) due to its superior performance on NAR generation and the flexibility of variable length prediction (Libovický and Helcl, 2018; Saharia et al., 2020; Gu and Kong, 2021).

CTC models an alignment a that provides a mapping between a sequence of predicted and target tokens. Alignments can be constructed by inserting special *blank tokens* ("_") and token repetitions into the target sequence. The alignment is monotonic with respect to the target sequence and is always the same length as the source sequence x. However, in MT, the target sequence y can be longer than the source sequence x. This is handled via upsampling the source sequence x, to s times its original length. An alignment is valid only if when collapsed, i.e., merging repeated tokens and removing blank tokens, it results in the original target sequence. The CTC loss marginalizes over all possible valid alignments $\Gamma(y)$ compatible with the target y and is defined as:

$$p(y \mid x) = \sum_{a \in \Gamma(y)} \prod_{1 \le t' \le T'} p(a_{t'} \mid x_{1:T'}; \phi).$$

Note that each alignment token $a_{t'}$ is modeled independently. This conditional independence allows CTC to predict the single most likely alignment non-autoregressively at inference time, which can then be efficiently collapsed to an output sequence. This same independence assumption enables efficient minimization of the CTC loss via dynamic programming (Graves et al., 2006). While CTC enforces monotonicity between the target and the predictions, it does not require any cross- or self-attention layers inside the model to be monotonic. Hence, CTC should still be able to model language pairs with different word orders between the source and the target sequence. Following Saharia et al. (2020), we train encoder-only CTC models, using a stack of self-attention layers to map the source sequence directly to the alignments.

2.2 Iterative Decoding with Imputer

IMPUTER (Saharia et al., 2020) extends NAR CTC modeling by iterative refinement (Lee et al., 2018). At each inference step, it conditions on a previous partially generated alignment to emit a new alignment. While IMPUTER, like CTC, generates all tokens at each inference step, only a subset of these tokens is selected to generate a partial alignment, similar to iterative masking approaches (Ghazvininejad et al., 2019). This is achieved during training via marginalizing over partial alignments as follows:

$$p(y \mid x) = \sum_{a \in \Gamma(a)} p(a \mid a_{\text{Mask}}, x; \phi),$$

where a_{Mask} is a partially masked input-alignment. At training time, the a_{Mask} alignment is generated using a CTC model trained on the same dataset, and its masked positions are selected randomly. This training procedure enables IMPUTER to iteratively refine a partial alignment over multiple

	Tgt Word Order	Size	Script Difference	White Space	Avg. Src Length	Avg. Tgt Length
En-Kk	SOV	150K	✓	✓	26.7	20.0
En-De	SVO/SOV	4.6M	✗	✓	25.7	24.3
En-Pl	SVO	5M	✗	✓	16.2	14.6
En-Hi	SOV	8.6M	✓	✓	18.3	19.8
En-Ja	SOV	17.9M	✓	✗	21.4	25.9
En-Ru	Free	33.5M	✓	✓	23.2	21.5
En-Fr	SVO	38.1M	✗	✓	29.2	32.8

Table 1: Details on training data used. Target word orders are the ones that are dominating within the language according to (Dryer and Haspelmath, 2013), but there may be sentence-specific variations. English follows predominantly SVO (Subject-Verb-Object) order. Size is measured as the number of parallel sentences in the training data. Source (Src) and Target (Tgt) length are averaged across sentences after word-based tokenization.

decoding steps at inference time — consuming its own alignments as input to the next iteration. With $k > 1$ decoding steps, the IMPUTER becomes *semi*-autoregressive, requiring k times more inference passes than pure CTC models.

IMPUTER differs from Conditional Masked Language Modeling (CMLM) (Ghazvininejad et al., 2019) in that it uses the CTC loss instead of the standard cross-entropy loss, removing the need for explicit output length prediction. Also, IMPUTER is an encoder-only model that makes one prediction per source token, just like CTC. The cross-attention component from encoder-decoder is replaced by a simple sum between the embeddings of the source sequence and the input alignment (a_{Mask}) before the first self-attention layer.[1]

2.3 Multilingual Modeling

Multilingual AR and NAR models are trained on datasets from multiple language pairs, $\{D^l\}_{l=1}^{L}$. We prepend each source sequence with the desired target language tag (<2tgt>) and generate a shared vocabulary across all languages (Johnson et al., 2017). The models encode this tag as any other token, and uses it to guide the generation of the output sequence in the desired target language.

2.4 Efficiency

Inference We refrain from wallclock inference time measurements since these are dependent on implementation, low-level optimization and machines (Dehghani et al., 2021). We instead compare generation speed in terms of the number of tokens that get generated per iteration N_{gen} (Kreutzer et al., 2020), which is < 1 for AR models,[2] T for fully non-autoregressive models like CTC and $\frac{T}{k}$ for iterative semi-autoregressive models like IMPUTER. *While the potential for faster inference motivates our interest in* NAR*, our core contribution is a comparison of multilingual modeling capabilities; therefore, we do not measure inference speed experimentally.*

Training At training time, NAR models are less efficient than AR models because their quality depends on distillation (Gu and Kong, 2021). Extra cost is incurred to train a teacher model (usually AR) and to use it to decode the training set.

Multilinguality Multilingual models multi-task over language pairs, so that a single multilingual model can replace several bilingual models. Thanks to transfer across languages, model size needs to be increased less than m-fold for modeling m language pairs.

Considering all of the above factors, an ideal model needs only a few iterations (decoder passes or steps), requires no teacher or a cheap teacher, and covers several languages, while incurring the smallest drop in quality compared to less efficient models. CTC is desirable as it uses only one pass, while IMPUTER gives up some efficiency to improve quality. Both require a teacher, but we can try to reduce the cost by training fewer teachers.

3 Experimental Setup

Data We perform our main experiments on six language pairs, translating from English into WMT-14 German (De) (Bojar et al., 2014), WMT-15 French (Fr) (Bojar et al., 2015), WMT-19 Russian (Ru) (Barrault et al., 2019), WMT-20 Japanese (Ja), WMT-20 Polish (Pl) (Barrault et al., 2020) and Samanantar Hindi (Hi) (Ramesh et al., 2021). The lower-resourced WMT-19 English-Kazakh (kk) (Barrault et al., 2019) is used for an additional transfer experiment in Section 5. The properties

[1] We experimented with an encoder-decoder variant of IMPUTER but it did not change the overall output quality in multilingual scenarios or otherwise.

[2] 1 for greedy search, < 1 to account for scoring and expansion of multiple hypotheses in beam search.

of the datasets are listed in Table 1. Target word order and writing script notably differ across these languages, so we focus on translating *into* these languages as this is a more challenging direction. A shared sub-word vocabulary of 32k is trained with SentencePiece (Kudo and Richardson, 2018), with the number of sub-words allocated for each language being proportional to its data size.

Evaluation Metrics Translation quality is evaluated with BLEU (Papineni et al., 2002) as calculated by Sacrebleu (Post, 2018) with default tokenization ("13a") except for EN-JA, where we use character-level tokenization. [3]

Architecture We train the IMPUTER model using the same setup as described in Saharia et al. (2020): We follow their base model with d_{model} = 512, d_{hidden} = 2048, n_{heads} = 8, n_{layers} = 12, and $p_{dropout}$ = 0.1. AR models follow Transformer-base (Vaswani et al., 2017) and have similar parameter counts. We train both models using Adam with learning rate of 0.0001. We train CTC models with a batch size of 2048 and 8192 sentences for 300K steps for the bilingual and multilingual models respectively. We train the IMPUTER using CTC loss using a Bernoulli masking policy for next 300K steps with a batch size of 1024 and 2048 sentences for the bilingual and multilingual models respectively. We upsample the source sequence by a factor of 2 for all our experiments.[4] We pick the best checkpoint based on validation BLEU for bilingual models, and the last checkpoint for multilingual models, following Arivazhagan et al. (2019).

Distillation We apply sequence-level knowledge distillation (Kim and Rush, 2016) from AR teacher models as widely used in NAR generation (Gu et al., 2018). Specifically, when training the NAR models, we replace the reference sequences during training with translation outputs from Transformer-Big AR teacher model with a beam width of four. We also report the quality of the AR teacher models, both bilingual and multilingual. The configurations for training the big AR teacher models also follow Vaswani et al. (2017).

[3] SacreBLEU (short) signatures: nrefs:1|case:mixed|eff:no|tok:13a,char|smooth:exp|version:2.0.0

[4] We do not vary the upsampling ratio due to small difference in the performance of the resulting NAR models (see Table 6, Gu and Kong (2021)).

4 Negative Transfer Scenario

Our main experiment compares bilingual, multilingual, AR and NAR models for the six high-resource languages from Table 1. These languages are typologically diverse, and they each have enough data so that we do not expect them to benefit substantially from joint modeling. We use this challenging scenario to test the impact of multilingual teachers, and to measure each paradigm's ability to model several unrelated languages. Results are shown in Table 2.

4.1 Multilingual Teacher Comparison

Inspecting the AR teacher models (rows 1 and 2 of Table 2) confirms the negative transfer that we aimed to design: multilingual teachers have substantially reduced BLEU compared to bilingual teachers. How much is this drop in quality affecting NAR students? First of all, we see that bilingual CTC models trained from the multilingual teacher (5) do not reflect the entirety of this drop when compared to training with the bilingual teacher (4): An average teacher gap of −1.8 BLEU is causing −1.1 drop for the corresponding students.[5] The comparison becomes more interesting as we shift to multilingual students: multilingual CTC (8, 9) does not suffer at all from having a multilingual teacher (average BLEU gap of −0.1), and multilingual IMPUTER (10, 11) likewise suffers very little (−0.3). These three results taken together suggest that *datasets distilled from multilingual models are likely simpler, but easier to model non-autoregressively* by the multilingual NAR models, which makes up for the teacher's lower BLEU. Our analysis in Section 4.3 supports this hypothesis.

We hope that highly multilingual models, trained with similar target language pairs to enhance positive transfer (Tan et al., 2019), are even better suited to serve as teachers for multilingual NAR models, which we leave to future work.

4.2 Multilingual Student Comparison

Returning to the "Bilingual Models" section of Table 2 with `AR-big` teachers, we can see that we have reproduced the results of Saharia et al. (2020): Bilingual CTC (4) performs well for a fully NAR method, but does not reach AR quality (3). IMPUTER (6) ably closes the gap with AR, surpassing

[5] As the quality of CTC generated alignments from the multi-AR-big is worse than the alignments generated from the CTC with the AR-big teacher, we do not train IMPUTER on CTC-generated alignments from the multi-AR-big models.

	MODEL	TEACHER	N_{gen}	EN-FR	EN-DE	EN-PL	EN-RU	EN-HI	EN-JA	AVG.
	Teachers									
(1)	AR-big		< 1	38.8	29.0	21.4	27.2	34.6	35.4	31.1
(2)	multi-AR-big			38.5	27.0	21.6	25.3	32.6	33.6	29.3
	Bilingual Models									
(3)	AR-base		< 1	38.2	27.6	21.2	26.2	33.8	34.8	30.3
(4)	CTC	AR-big	T	35.7	25.2	18.0	21.4	31.6	31.6	27.3
(5)		multi-AR-big		35.1	24.0	17.7	20.8	30.8	28.9	26.2
(6)	IMPUTER	AR-big	$\frac{T}{8}$	38.5	27.2	21.2	25.6	32.0	32.0	29.4
	Multilingual Models									
(7)	multi-AR-base		< 1	35.2	24.8	19.7	23.2	30.8	31.2	27.5
(8)	CTC	AR-big	T	31.6	20.5	13.0	17.7	28.2	28.1	23.2
(9)		multi-AR-big		31.2	20.5	13.7	18.0	27.8	27.5	23.1
(10)	IMPUTER	AR-big	$\frac{T}{8}$	34.4	22.8	14.9	21.3	29.9	29.6	25.5
(11)		multi-AR-big		34.1	21.2	16.4	21.7	29.9	27.9	25.2

Table 2: Test BLEU scores for multilingual and bilingual AR and NAR models and their teachers.

or coming within 0.4 BLEU of the AR-base models on 3/6 language pairs, with the largest gap in performance for the distant EN-JA. Does this story hold as we move to multilingual NAR students?

To understand each model's multilingual capabilities, we can compare its bilingual performance to its multilingual performance. Comparing bilingual AR-base (3) to its multilingual counterparts (7) gives us a baseline average drop of −2.8 BLEU, confirming that this is indeed a difficult multilingual scenario that leads to negative transfer. Comparing bilingual CTC (4) to multilingual CTC (8) with AR-big teachers, we see an average drop of −4.1. This larger drop indicates that CTC *suffers more from negative interference than its* AR *counterpart*. We hypothesize that CTC models need more capacity than AR models to achieve similar multilingual performance, motivating our scaling law experiments in Section 6.

Performing the same bilingual-to-multilingual comparison for IMPUTER (6 vs. 10) shows a similar −3.9 average drop due to negative transfer. So although IMPUTER is indeed better than CTC (2 BLEU), it does not seem to be better suited for multilingual modeling in this difficult scenario.

4.3 How do the bilingual and the multilingual distilled datasets differ?

Table 3 summarizes different statistics for the original (R) and distilled datasets from both multilingual (M) and bilingual (B) AR teacher models. We report the number of types and average sequence length (in tokens) for the target side of the dataset. We compute the complexity of the dataset based on probabilities from a statistical word aligner (Zhou et al., 2019). The FRS (Talbot et al., 2011) score represents the average fuzzy reordering score over all the sentence pairs for the respective language pair as measured in Xu et al. (2021), with higher values suggesting that the target is more monotonic with the source sequence. We also report BLEU for the distilled datasets relative to the original training references.

The datasets distilled from the bilingual AR models (B) are shorter, less complex, have reduced lexical diversity (in number of types) and are more monotonic compared to the original corpora (R), which corroborates findings from prior work (Zhou et al., 2019; Xu et al., 2021). One exception is EN-JA, where the distilled translations are slightly less monotonic than the original references. Moving to multilingual teachers (M), the resulting datasets have further reduced types, are shorter and less complex than those distilled from bilingual teachers. In particular, their monotonicity increased (FRS) for the more distant language pairs, EN-JA and EN-HI. As shown in Xu et al. (2021) and Voita et al. (2021), reduced lexical diversity and reordering complexity can help bilingual NAR models to learn better alignments between source and target, improving the translation quality of the outputs. More work is needed to better understand

PROPERTY	R	B	M
	EN-FR		
# TYPES	522K	430K	396K
AVG. LENGTH	32.8	31.2	29.2
COMPLEXITY	1.529	1.167	0.944
FRS	0.463	0.541	0.536
BLEU (Train)	-	40.8	37.8
	EN-DE		
# TYPES	812K	616K	573K
AVG. LENGTH	24.3	23.4	22.2
COMPLEXITY	1.243	0.819	0.709
FRS	0.490	0.606	0.605
BLEU (Train)	-	35.0	26.4
	EN-PL		
# TYPES	636K	516K	503K
AVG. LENGTH	14.6	13.4	12.7
COMPLEXITY	1.435	0.942	0.591
FRS	0.590	0.678	0.695
BLEU (Train)	-	26.3	22.0
	EN-RU		
# TYPES	636K	516K	503K
AVG. LENGTH	21.5	20.5	19.5
COMPLEXITY	1.083	0.882	0.819
FRS	0.640	0.719	0.716
BLEU (Train)	-	43.2	40.0
	EN-HI		
# TYPES	346K	200K	185K
AVG. LENGTH	19.8	18.8	17.8
COMPLEXITY	1.438	1.256	1.138
FRS	0.347	0.363	0.366
BLEU (Train)	-	34.6	28.0
	EN-JA		
# TYPES	547K	440K	402K
AVG. LENGTH	25.9	23.5	22.2
COMPLEXITY	1.541	1.369	1.338
FRS	0.344	0.337	0.340
BLEU (Train)	-	35.9	30.6

Table 3: Comparison of datasets (1M samples) distilled from bilingual (B) or multilingual (M) AR models

the sweet-spot between the quality and complexity trade-off of the multilingual and bilingual distilled datasets for multilingual NAR modeling.

4.4 Which translation errors are made?

In this section, we analyze quantitatively how the output quality of NAR models differs across language pairs when trained in isolation (bilingual) or with other language pairs (multilingual).

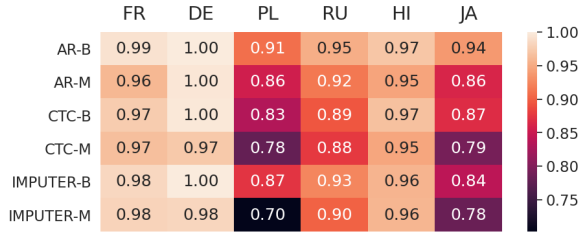

Figure 1: Brevity penalty scores for bilingual (-B) and multilingual (-M) models, the closer to 1 the better.

Effect of Length Figure 1 shows the brevity penalty (BP) scores (Papineni et al., 2002) for all languages. EN-PL and EN-JA have lowest BP scores across the board, meaning that their translations are shorter than the references. Manual inspection reveals that this could be attributed to the subject pronouns being dropped in both of these target languages. Multilingual modeling results in shorter outputs relative to bilingual models for both AR and NAR models and most language pairs. While IMPUTER models tend to have fewer issues with output length compared to CTC models, they still lag behind AR models, suggesting that the length might need to be controlled explicitly for these language pairs (Gu and Kong, 2021).

Invalid Words CTC frequently generates *invalid* words, i.e. tokens that are not present in the target side of the bitext but are being composed from multiple sub-words. These sub-words represent alternative translations that the model fails to distinguish. In the Hindi example below, the invalid (or made-up) word in the sentence is marked in red. The correct word should be जहरीले as the dependent vowel "ी" can only be used once.

Hindi: इससे ग्रामीण महिलाओं को जहरीले धुएं से मुक्ति मिली हैं।

English: This has relieved the rural women from the poisonous smoke.

Figure 2 reports the percentage of sequences that include at least one invalid word in the test set. CTC generates many invalid words compared to both AR and IMPUTER, with multilingual modeling leading to an average increase in invalid words by 37%. The shared vocabulary of the multilingual model results in shorter sub-words, hence longer sequences, and the conditionally independent generation leads to more clashing adjacent sub-words.[6]

[6]One might hope to alleviate this by increasing vocabulary size, but preliminary experiments showed that an increased vocabulary was less efficient in improving quality than increasing overall model size, which is explored in Section 6.

IMPUTER's iterative decoding alleviates this for some languages. Increasing the number of iterations could help, but would also erode the efficiency arguments that make NAR models attractive. As pointed out by Xiao et al. (2022), better modeling of target token dependencies is crucial to closing the gap in translation quality to AR models.

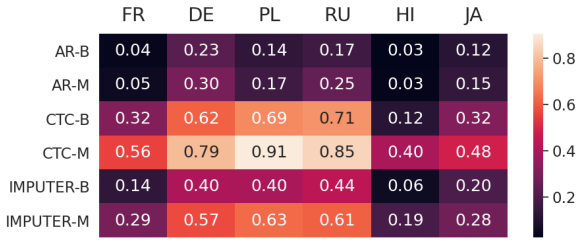

Figure 2: % of outputs with invalid words for bilingual (-B) and multilingual (-M) models, the lower the better.

5 Positive Transfer Scenario

In this section we present two experimental setups designed to emphasize positive transfer, where languages are related and training data is limited.

English→{German, French} To isolate the effect of transfer via multilingual modelling, we relax the capacity bottleneck and competition for parameters: We combine the two most related languages (DE, FR) (Kudugunta et al., 2019, Figure 2) and give them smaller, balanced training sets (1M sentences). We compare bilingual and multilingual AR and NAR models trained on this reduced data.

Table 4 shows that NAR *models benefit from training with multiple language pairs in this relaxed scenario* — all models exhibit positive transfer (in green). IMPUTER achieves higher positive transfer than CTC for both languages, but lags behind the AR multilingual model in EN-FR. However, for EN-FR the bilingual IMPUTER is already ahead of the bilingual AR model by 0.4 BLEU.

MODEL	EN-DE	EN-FR
Bilingual Models		
AR	22.8	27.7
CTC	21.5	26.5
IMPUTER	22.8	28.1
Multilingual Models		
AR	**24.3** +1.5	**29.0** +1.3
CTC	22.1 +0.6	26.9 +0.4
IMPUTER	23.7 +1.3	28.5 +0.4

Table 4: Results on subsampled (1M) training data.

English→{Russian, Kazakh} Does this positive transfer survive data imbalance? We test the performance of the multilingual NAR model on the low-resource task of translating English into Kazakh, for which the size of clean training data is insufficient to train a bilingual AR model from scratch. We instead distill translations from the publicly available multilingual AR model, PRISM (Thompson and Post, 2020). We then pair it with the higher-resource but related language Russian to encourage positive transfer to Kazakh. Given the huge difference in data sizes for Russian and Kazakh (see Table 1), we sample training data from the two languages based on the data size scaled by a temperature value τ, $p_l^{1/\tau}$ (Arivazhagan et al., 2019), where, $p_l = \frac{D_l}{\sum_k D_k}$. We experiment with multiple temperature values (1, 3, 5, 10, 20) and pick the best value ($\tau = 5; p_{\text{RU}}^{1/\tau} = 0.75, p_{\text{KK}}^{1/\tau} = 0.25$) based on the performance on the validation set.

MODEL	TEACHER	EN-KK	EN-RU
PRISM	-	**8.9**	**27.0**
Bilingual Models			
AR	PRISM	4.4	-
CTC		1.2	-
Multilingual Models			
AR	PRISM	7.1 +2.7	26.0
CTC		2.8 +1.6	20.4

Table 5: Results on English → Kazakh, Russian.

As can be seen in Table 5, both AR and CTC show positive transfer when translating into Kazakh when trained in combination with Russian. The multilingual CTC model is able to improve over the bilingual CTC model, but the overall quality of the outputs is very low compared to the teacher model (BLEU: -5.3). This experiment showcases that *current* NAR *models do not perform well on very low-resource language pairs* and might need further data augmentation (Song et al., 2022) or transfer from other similar languages.[7]

6 Impact of Model Scale

We hypothesized in Section 4 that CTC might require more capacity than AR models. If we increase the parameters for NAR models sufficiently, could we reach AR quality? Scaling laws can characterize the relationship between MT quality, the cross-entropy loss and the number of parameters

[7]We do not train IMPUTER for KK as the quality of the distilled dataset and alignments from CTC is very low.

used for training the model (Ghorbani et al., 2021; Gordon et al., 2021).

We derive the relationship between BLEU and the number of parameters for our AR and CTC models directly from the scaling laws proposed by Gordon et al. (2021) and Ghorbani et al. (2021) as follows:

$$L(N) \approx L_0 + \alpha_n (1/N)^{\alpha_k} \text{ (Ghorbani et al., 2021)}$$

$$\text{BLEU}(L) \approx C e^{-kL} \text{ (Gordon et al., 2021)}$$

$$\text{BLEU}(N) \approx a e^{-b(1/N)^c} \text{ (this work)}$$

where L is the test loss, $\{\alpha_n, \alpha_k, L_0, C, k\}$ are fitted parameters from previous power laws, and $\{a, b, c\}$ are the collapsed fitted parameters of our power law. Ghorbani et al. (2021)'s L_0 corresponds to the irreducible loss of the data (here: a).

Setup We train seven models with varying capacity for AR and CTC models. The number of layers and model sizes are varied as: (6, 128), (6, 256), (12, 256), *(12, 512)*,[8] (24, 512), (12, 1024), (24, 1024). The feed-forward size is 4× the model size. AR models have equal numbers of encoder and decoder layers. The number of attention heads is given by $(8/(512/\text{Model Size}))$. For a fair comparison, we train both AR and CTC models on distilled outputs from `AR-big` in Table 2. The evaluation is conducted in the challenging six-language negative-transfer scenario from Section 4, where capacity bottlenecks are likely to be most pronounced. We report BLEU averaged across six languages.

Results Figure 3 shows the fitted parameters using the scaling law, which can almost perfectly describe the relationship between the number of parameters and the development BLEU (R^2: 0.99). We can see that CTC*, even with many more parameters, do not come even close to the performance of* AR *models and plateaus early* at a BLEU of 26.7, while AR models plateau at 30.8. By projecting the curves out to 1 billion parameters, we show that increasing the capacity of NAR is insufficient to reach the quality of AR models.

7 Related Work

Multiple approaches with varying architectures (Gu et al., 2018, 2019; Chan et al., 2020; Xu and Carpuat, 2021), custom loss functions (Ghazvininejad et al., 2020; Du et al., 2021) and training strategies (Ghazvininejad et al., 2019; Qian et al., 2021)

[8] Size for experiments in Section 4.

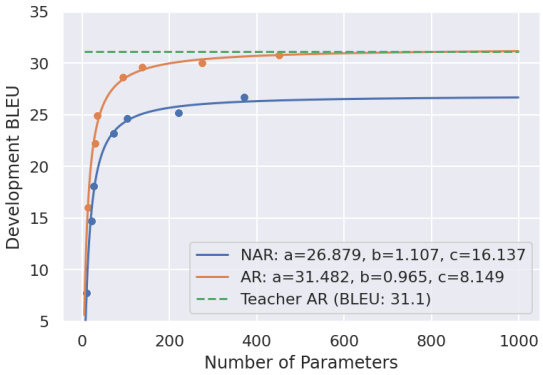

Figure 3: BLEU versus number of parameters and fitted power-law curves (R^2 AR: 0.99, R^2 CTC: 0.99).

have been used to enable parallel generation of output tokens for MT with sequence-level knowledge distillation as one of the key ingredient in the training of NAR models. Both supervised, and unsupervised (Sun et al., 2020) MT have benefitted from training with multiple languages, especially those that have tiny (Siddhant et al., 2020) to no training data (Zhang et al., 2020). However, multilingual modeling has not yet received any attention in the NAR literature, which we explore in this work. One limitation of our study is that we choose one representative system for NAR and semi-NAR modeling over the full breadth of NAR options.

8 Conclusion

Multilingual translation is a valuable feature of AR models, therefore, we have tested NAR models for that same capability. We focus on challenging scenarios to discover potential weaknesses and to identify areas for future work. In a relaxed setting with little interference between languages and balanced data, multilingual NAR models nicely exhibit positive transfer, practically closing the gap to AR models with a few decoding iterations. However, we do not see the same positive transfer in a true low-resource scenario. Experiments in a six-language scenario reveal that multilingual NAR models suffer proportionally more from negative interference than AR models. Our derived scaling laws show that scaling up CTC model parameters is not a sufficient remedy. Our analysis identified two issues that hurt translation quality and worsen with multilinguality, namely output length control and the generation of invalid words. We have also shown beneficial properties of using multilingual teachers for distillation. We hope that this work will serve as a call for increased focus on multilingual modeling in NAR research.

Acknowledgments

We thank George Foster, Aditya Gupta, and the anonymous reviewers for their helpful and constructive comments.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of NAACL-HLT*, pages 3874–3884.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 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.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

William Chan, Chitwan Saharia, Geoffrey Hinton, Mohammad Norouzi, and Navdeep Jaitly. 2020. Imputer: Sequence modelling via imputation and dynamic programming. In *International Conference on Machine Learning*, pages 1403–1413. PMLR.

Mostafa Dehghani, Anurag Arnab, Lucas Beyer, Ashish Vaswani, and Yi Tay. 2021. The efficiency misnomer. *arXiv preprint arXiv:2110.12894*.

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.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Cunxiao Du, Zhaopeng Tu, and Jing Jiang. 2021. Order-agnostic cross entropy for non-autoregressive machine translation. *arXiv preprint arXiv:2106.05093*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Computer Speech & Language*, 45:236–252.

Marjan Ghazvininejad, Vladimir Karpukhin, Luke Zettlemoyer, and Omer Levy. 2020. Aligned cross entropy for non-autoregressive machine translation. *CoRR*, abs/2004.01655.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. 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 6112–6121.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. 2021. Scaling laws for neural machine translation. *arXiv preprint arXiv:2109.07740*.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing, pages 5915–5922.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Jiatao Gu and Xiang Kong. 2021. Fully non-autoregressive neural machine translation: Tricks of the trade. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 120–133, Online. Association for Computational Linguistics.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11179–11189. Curran Associates, Inc.

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.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Julia Kreutzer, George Foster, and Colin Cherry. 2020. Inference strategies for machine translation with conditional masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5774–5782, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP (Demonstration)*.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Brussels, Belgium. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2018. End-to-end non-autoregressive neural machine translation with connectionist temporal classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3016–3021, 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.

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, Brussels, Belgium. Association for Computational Linguistics.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2021. Glancing transformer for non-autoregressive 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 1993–2003, Online. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *arXiv preprint arXiv:2104.05596*.

Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. 2020. Non-autoregressive machine translation with latent alignments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1098–1108.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia.

2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *CoRR*, abs/2201.03110.

Zhenqiao Song, Hao Zhou, Lihua Qian, Jingjing Xu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2022. switch-GLAT: Multilingual parallel machine translation via code-switch decoder. In *International Conference on Learning Representations*.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3525–3535, Online. Association for Computational Linguistics.

David Talbot, Hideto Kazawa, Hiroshi Ichikawa, Jason Katz-Brown, Masakazu Seno, and Franz Josef Och. 2011. A lightweight evaluation framework for machine translation reordering. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 12–21.

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.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. 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*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Language modeling, lexical translation, reordering: The training process of NMT through the lens of classical SMT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8478–8491, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-yan Liu. 2022. A survey on non-autoregressive generation for neural machine translation and beyond. *arXiv preprint arXiv:2204.09269*.

Weijia Xu and Marine Carpuat. 2021. Editor: An edit-based transformer with repositioning for neural machine translation with soft lexical constraints. *Transactions of the Association for Computational Linguistics*, 9:311–328.

Weijia Xu, Shuming Ma, Dongdong Zhang, and Marine Carpuat. 2021. How does distilled data complexity impact the quality and confidence of non-autoregressive machine translation? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4392–4400, Online. 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.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2019. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*.