Massively Multilingual Neural Machine Translation

Roee Aharoni[*]
Bar Ilan University
Ramat-Gan
Israel
roee.aharoni@gmail.com

Melvin Johnson and **Orhan Firat**
Google AI
Mountain View
California
melvinp,orhanf@google.com

Abstract

Multilingual neural machine translation (NMT) enables training a single model that supports translation from multiple source languages into multiple target languages. In this paper, we push the limits of multilingual NMT in terms of the number of languages being used. We perform extensive experiments in training massively multilingual NMT models, translating up to 102 languages to and from English within a single model. We explore different setups for training such models and analyze the trade-offs between translation quality and various modeling decisions. We report results on the publicly available TED talks multilingual corpus where we show that massively multilingual many-to-many models are effective in low resource settings, outperforming the previous state-of-the-art while supporting up to 59 languages. Our experiments on a large-scale dataset with 102 languages to and from English and up to one million examples per direction also show promising results, surpassing strong bilingual baselines and encouraging future work on massively multilingual NMT.

1 Introduction

Neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014; Sutskever et al., 2014) is the current state-of-the-art approach for machine translation in both academia (Bojar et al., 2016, 2017, 2018) and industry (Wu et al., 2016; Hassan et al., 2018). Recent works (Dong et al., 2015; Firat et al., 2016a; Ha et al., 2016; Johnson et al., 2017) extended the approach to support multilingual translation, i.e. training a single model that is capable of translating between multiple language pairs.

Multilingual models are appealing for several reasons. First, they are more efficient in terms of the number of required models and model parameters, enabling simpler deployment. Another benefit is transfer learning; when low-resource language pairs are trained together with high-resource ones, the translation quality may improve (Zoph et al., 2016; Nguyen and Chiang, 2017). An extreme case of such transfer learning is zero-shot translation (Johnson et al., 2017), where multilingual models are able to translate between language pairs that were never seen during training.

While very promising, it is still unclear how far one can scale multilingual NMT in terms of the number of languages involved. Previous works on multilingual NMT typically trained models with up to 7 languages (Dong et al., 2015; Firat et al., 2016b; Ha et al., 2016; Johnson et al., 2017; Gu et al., 2018) and up to 20 trained directions (Cettolo et al., 2017) simultaneously. One recent exception is Neubig and Hu (2018) who trained many-to-one models from 58 languages into English. While utilizing significantly more languages than previous works, their experiments were restricted to many-to-one models in a low-resource setting with up to 214k examples per language-pair and were evaluated only on four translation directions.

In this work, we take a step towards practical "universal" NMT – training massively multilingual models which support up to 102 languages and with up to one million examples per language-pair simultaneously. Specifically, we focus on training "English-centric" many-to-many models, in which the training data is composed of many language pairs that contain English either on the source side or the target side. This is a realistic setting since English parallel data is widely available for many language pairs. We restrict our experiments to Transformer models (Vaswani et al., 2017) as they were shown to be very effective in recent benchmarks (Ott et al., 2018), also in

[*] Work carried out during an internship at Google AI.

Proceedings of NAACL-HLT 2019, pages 3874–3884
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

the context of multilingual models (Lakew et al., 2018; Sachan and Neubig, 2018).

We evaluate the performance of such massively multilingual models while varying factors like model capacity, the number of trained directions (tasks) and low-resource vs. high-resource settings. Our experiments on the publicly available TED talks dataset (Qi et al., 2018) show that massively multilingual many-to-many models with up to 58 languages to-and-from English are very effective in low resource settings, allowing to use high-capacity models while avoiding overfitting and achieving superior results to the current state-of-the-art on this dataset (Neubig and Hu, 2018; Wang et al., 2019) when translating into English.

We then turn to experiment with models trained on 103 languages in a high-resource setting. For this purpose we compile an English-centric in-house dataset, including 102 languages aligned to-and-from English with up to one million examples per language pair. We then train a single model on the resulting 204 translation directions and find that such models outperform strong bilingual baselines by more than 2 BLEU averaged across 10 diverse language pairs, both to-and-from English. Finally, we analyze the trade-offs between the number of involved languages and translation accuracy in such settings, showing that massively multilingual models generalize better to zero-shot scenarios. We hope these results will encourage future research on massively multilingual NMT.

2 Low-Resource Setting: 59 Languages

2.1 Experimental Setup

The main question we wish to answer in this work is how well a single NMT model can scale to support a very large number of language pairs. The answer is not trivial: on the one hand, training multiple language pairs together may result in transfer learning (Zoph et al., 2016; Nguyen and Chiang, 2017). This may improve performance as we increase the number of language pairs, since more information can be shared between the different translation tasks, allowing the model to learn which information to share. On the other hand, adding many language pairs may result in a bottleneck; the model has a limited capacity while it needs to handle this large number of translation tasks, and sharing all parameters between the different languages can be sub-optimal (Wang et al., 2018) especially if they are not from the same typological language family (Sachan and Neubig, 2018).

We begin tackling this question by experimenting with the TED Talks parallel corpus compiled by Qi et al. (2018)[1], which is unique in that it includes parallel data from 59 languages. For comparison, this is significantly "more multilingual" than the data available from all previous WMT news translation shared task evaluations throughout the years – the latest being Bojar et al. (2016, 2017, 2018), which included 14 languages so far.[2]

We focus on the setting where we train "English-centric" models, i.e. training on all language pairs that contain English in either the source or the target, resulting in 116 translation directions. This dataset is also highly imbalanced, with language pairs including between 3.3k to 214k sentence pairs for training. Table 9 in the supplementary material details the languages and training set sizes for this dataset. Since the dataset is already tokenized we did not apply additional preprocessing other than applying joint subword segmentation (Sennrich et al., 2016) with 32k symbols.

Regarding the languages we evaluate on, we begin with the same four languages as Neubig and Hu (2018) – Azerbeijani (Az), Belarusian (Be), Galician (Gl) and Slovak (Sk). These languages present an extreme low-resource case, with as few as 4.5k training examples for Belarusian-English. In order to better understand the effect of training set size in these settings, we evaluate on four additional languages that have more than 167k training examples each – Arabic (Ar), German (De), Hebrew (He) and Italian (It).

2.2 Model Details

Using the same data, we trained three massively multilingual models: a many-to-many model which we train using all 116 translation directions with 58 languages to-and-from English, a one-to-many model from English into 58 languages, and a many-to-one model from 58 languages into English. We follow the method of Ha et al. (2016); Johnson et al. (2017) and add a target-language

[1] github.com/neulab/word-embeddings-for-nmt

[2] Chinese, Czech, English, Estonian, Finnish, French, German, Hindi, Hungarian, Latvian, Romanian, Russian, Spanish, Turkish. According to http://www.statmt.org/wmtXX

prefix token to each source sentence to enable many-to-many translation. These different setups enable us to examine the effect of the number of translation tasks on the translation quality as measured in BLEU (Papineni et al., 2002). We also compare our massively multilingual models to bilingual baselines and to two recently published results on this dataset (Neubig and Hu (2018); Wang et al. (2019)).

Regarding the models, we focused on the Transformer in the "Base" configuration. We refer the reader to Vaswani et al. (2017) for more details on the model architecture. Specifically, we use 6 layers in both the encoder and the decoder, with model dimension set at 512, hidden dimension size of 2048 and 8 attention heads. We also applied dropout at a rate of 0.2 in the following components: on the sum of the input embeddings and the positional embeddings, on the output of each sub-layer before added to the previous layer input (residual connection), on the inner layer output after the ReLU activation in each feed-forward sub-layer, and to the attention weight in each attention sub-layer. This results in a model with approximately 93M trainable parameters. For all models we used the inverse square root learning rate schedule from Vaswani et al. (2017) with learning-rate set at 3 and 40k warmup steps. All models are implemented in Tensorflow-Lingvo (Shen et al., 2019).

In all cases we report test results for the checkpoint that performed best on the development set in terms of BLEU. For the multilingual models we create a development set that includes examples we uniformly sample from a concatenation of all the individual language pair development sets, resulting in 13k development examples per model. Another important detail regarding multilingual training is the batching scheme. In all of our multilingual models we use heterogeneous batching, where each batch contains examples which are uniformly sampled from a concatenation of all the language pairs the model is trained on. Specifically, we use batches of 64 examples for sequences shorter than 69 tokens and batches of 16 examples for longer sequences. We did not use oversampling as the dataset is relatively small.

	Az-En	Be-En	Gl-En	Sk-En	Avg.
# of examples	5.9k	4.5k	10k	61k	20.3k
Neubig & Hu 18					
baselines	2.7	2.8	16.2	24	11.42
many-to-one	11.7	18.3	29.1	28.3	21.85
Wang et al. 18	11.82	18.71	30.3	28.77	22.4
Ours					
many-to-one	11.24	18.28	28.63	26.78	21.23
many-to-many	**12.78**	**21.73**	**30.65**	**29.54**	**23.67**

Table 1: X→En test BLEU on the TED Talks corpus, for the language pairs from Neubig and Hu (2018)

	Ar-En	De-En	He-En	It-En	Avg.
# of examples	213k	167k	211k	203k	198.5k
baselines	27.84	30.5	**34.37**	33.64	31.59
many-to-one	25.93	28.87	30.19	32.42	29.35
many-to-many	**28.32**	**32.97**	33.18	**35.14**	**32.4**

Table 2: X→En test BLEU on the TED Talks corpus, for language pairs with more than 167k examples

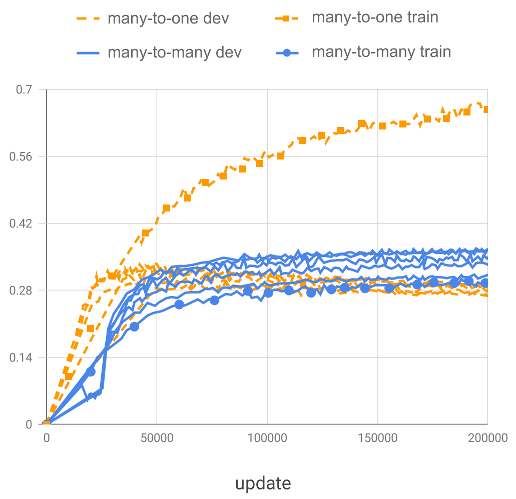

Figure 1: Development BLEU on {It,Ro,Nl,De,Ar}→En vs. training BLEU for the many-to-one and many-to-many models. Best viewed in color.

2.3 Results

We use tokenized BLEU in order to be comparable with Neubig and Hu (2018). Table 1 shows the results of our experiments when evaluating on the same language pairs as they did. The results under "Neubig & Hu 18" are their bilingual baselines and their best many-to-one models. Their many-to-one models use similar-language-regularization, i.e. fine-tuning a pre-trained many-to-one model with data from the language pair of interest together with data from a language pair that has a typologically-similar source language and more training data (i.e. Russian and Belarusian, Turkish and Azerbaijani). The results under "Ours" are our many-to-one and many-to-many models we trained identically in terms of model architecture and hyper-parameters.

We first note that our many-to-many model outperforms all other models when translating into English, with 1.82 BLEU improvement (when averaged across the four language pairs) over the best fine-tuned many-to-one models of Neubig and Hu (2018) and 2.44 BLEU improvement over our many-to-one model when averaged across the four low-resource language pairs (Table 1). This is surprising as it uses the same X→En data, model architecture and capacity as our many-to-one model, while handling a heavier burden since it also supports 58 *additional* translation tasks (*from* English *into* 58 languages). Our models also outperform the more complex models of Wang et al. (2019) which use ”Soft Decoupled Encoding” for the input tokens, while our models use a simple subword segmentation.

One possible explanation is that the many-to-one model overfits the English side of the corpus as it is multi-way-parallel: in such setting the English sentences are overlapping across the different language pairs, making it much easier for the model to memorize the training set instead of generalizing (when enough capacity is available). On the other hand, the many-to-many model is trained on additional target languages other than English, which can act as regularizers for the X→En tasks, reducing such overfitting.

To further illustrate this, Figure 1 tracks the BLEU scores on the individual development sets during training for Italian (It), Romanian (Ro), Dutch (Nl), German (De) and Arabic (Ar) into English (left), together with BLEU scores on a subset of the training set for each model. We can see that while the many-to-one model degrades in performance on the development set, the many-to-many model still improves. Note the large gap in the many-to-one model between the training set BLEU and the development set BLEU, which points on the generalization issue that is not present in the many-to-many setting. We also note that our many-to-one model is on average 0.75 BLEU behind the best many-to-one models in Neubig and Hu (2018). We attribute this to the fact that their models are fine-tuned using similar-language-regularization while our model is not.

We find an additional difference between the results on the resource-scarce languages (Table 1) and the higher-resource languages (Table 2). Specifically, the bilingual baselines outperform the many-to-one models only in the higher-resource setting. This makes sense as in the low-resource setting the baselines have very few training examples to outperform the many-to-one models, while in the higher resource setting they have access to more training data. This corroborates the results of Gu et al. (2018) that showed the sensitivity of such models to similar low resource conditions and the improvements gained from using many-to-one models (however with much fewer language pairs).

	En-Az	En-Be	En-Gl	En-Sk	Avg.
# of examples	5.9k	4.5k	10k	61k	20.3k
baselines	2.16	2.47	3.26	5.8	3.42
one-to-many	**5.06**	**10.72**	**26.59**	**24.52**	**16.72**
many-to-many	3.9	7.24	23.78	21.83	14.19

	En-Ar	En-De	En-He	En-It	Avg.
# of examples	213k	167k	211k	203k	198.5k
baselines	12.95	23.31	23.66	30.33	22.56
one-to-many	**16.67**	**30.54**	**27.62**	**35.89**	**27.68**
many-to-many	14.25	27.95	24.16	33.26	24.9

Table 3: En→X test BLEU on the TED Talks corpus

Table 3 shows the results of our massively multilingual models and bilingual baselines when evaluated out-of-English. In this case we see an opposite trend: the many-to-many model performs worse than the one-to-many model by 2.53 BLEU on average. While previous works (Wang et al., 2018; Sachan and Neubig, 2018) discuss the phenomena of quality degradation in English-to-many settings, this shows that increasing the number of *source* languages also causes additional degradation in a many-to-many model. This degradation may be due to the English-centric setting: since most of the translation directions the model is trained on are into English, this leaves less capacity for the other target languages (while still performing better than the bilingual baselines on all 8 language pairs). We also note that in this case the results are consistent among the higher and lower resource pairs – the one-to-many model is better than the many-to-many model, which outperforms the bilingual baselines in all cases. This is unlike the difference we saw in the X→ En experiments since here we do not have the multi-way-parallel overfitting issue.

2.4 Discussion

From the above experiments we learn that NMT models can scale to 59 languages in a low-resource, imbalanced, English-centric setting, with the following observations: (1) massively multilingual many-to-many models outperform many-to-one and bilingual models with similar capacity and identical training conditions when averaged over 8 language pairs into English. We attribute this improvement over the many-to-one models to the multiple target language pairs which may act as regularizers, especially in this low-resource multi-way-parallel setting that is prone to memorization. (2) many-to-many models are inferior in performance when going out-of-English in comparison to a one-to-many model. We attribute this to English being over-represented in the English-centric many-to-many setting, where it appears as a target language in 58 out of 116 trained directions, which may harm the performance on the rest of the target languages as the model capacity is limited.[3]

It is important to stress the fact that we compared the different models under *identical training conditions* and did not perform extensive hyper-parameter tuning for each setting separately. However, we believe that such tuning may improve performance even further, as the diversity in each training batch is very different between the different settings. For example, while the baseline model batches include only one language in the source and one language in the target, the many-to-many model includes 59 languages in each side with a strong bias towards English. These differences may require tailored hyper-parameter choices for each settings (i.e. different batch sizes, learning rate schedules, dropout rates etc.) which would be interesting to explore in future work.

In the following experiments we investigate whether these observations hold using (1) an even larger set of languages, and (2) a much larger, balanced training corpus that is not multi-way-parallel.

3 High-Resource Setting: 103 Languages

3.1 Experimental Setup

In this setting we scale the number of languages and examples per language pair further when training a single massively multilingual model. Since we are not aware of a publicly available resource for this purpose, we construct an in-house dataset. This dataset includes 102 language pairs which we "mirror" to-and-from English, with up to one million examples per language pair. This results in 103 languages in total, and 204 translation directions which we train simultaneously. More details about this dataset are available in Table 4, and Table 10 in the supplementary material details all the languages in the dataset.[4]

Similarly to our previous experiments, we compare the massively multilingual models to bilingual baselines trained on the same data. We tokenize the data using an in-house tokenizer and then apply joint subword segmentation to achieve an open-vocabulary. In this setting we used a vocabulary of 64k subwords rather than 32k. Since the dataset contains 24k unique characters, a 32k symbol vocabulary will consist of mostly characters, thereby increasing the average sequence length. Regarding the model, for these experiments we use a larger Transformer model with 6 layers in both the encoder and the decoder, model dimension set to 1024, hidden dimension size of 8192, and 16 attention heads. This results in a model with approximately 473.7M parameters.[5] Since the model and data are much larger in this case, we used a dropout rate of 0.1 for our multilingual models and tuned it to 0.3 for our baseline models as it improved the translation quality on the development set.

We evaluate our models on 10 languages from different typological families: *Semitic* – Arabic (Ar), Hebrew (He), *Romance* – Galician (Gl), Italian (It), Romanian (Ro), *Germanic* – German (De), Dutch (Nl), *Slavic* – Belarusian (Be), Slovak (Sk) and *Turkic* – Azerbaijani (Az) and Turkish (Tr). We evaluate both to-and-from English, where each language pair is trained on up to one million examples. As in the previous experiment, we report test results from the model that performed best in terms of BLEU on the development set.

# of language pairs	102
examples per pair	
min	63,879
max	1,000,000
average	940,087
std. deviation	188,194
total # of examples	95,888,938

Table 4: Training set details for the 103 langauges corpus, X→En data.

[3] This issue may be alleviated by over-sampling the non-English-target pairs, but we leave this for future work.

[4] The average number of examples per language pair is 940k, as for 13 out of the 102 pairs we had less than one million examples available.

[5] This is larger than the Transformer "Big" configuration, which includes approximately 213M trained parameters.

	Ar	Az	Be	De	He	It	Nl	Ro	Sk	Tr	Avg.
baselines	23.34	16.3	21.93	30.18	31.83	**36.47**	36.12	34.59	25.39	27.13	28.33
many-to-one	**26.04**	**23.68**	**25.36**	35.05	**33.61**	35.69	**36.28**	36.33	28.35	**29.75**	**31.01**
many-to-many	22.17	21.45	23.03	**37.06**	30.71	35.0	36.18	**36.57**	**29.87**	27.64	29.97

Table 5: X→En test BLEU on the 103-language corpus

	Ar	Az	Be	De	He	It	Nl	Ro	Sk	Tr	Avg.
baselines	10.57	8.07	15.3	23.24	19.47	31.42	28.68	27.92	11.08	15.54	19.13
one-to-many	**12.08**	**9.92**	**15.6**	**31.39**	**20.01**	**33**	**31.06**	**28.43**	**17.67**	**17.68**	**21.68**
many-to-many	10.57	9.84	14.3	28.48	17.91	30.39	29.67	26.23	18.15	15.58	20.11

Table 6: En→X test BLEU on the 103-language corpus

3.2 Results

Table 5 describes the results when translating into English. First, we can see that both multilingual models perform better than the baselines in terms of average BLEU. This shows that massively multilingual many-to-many models can work well in realistic settings with millions of training examples, 102 languages and 204 jointly trained directions to-and-from English. Looking more closely, we note several different behaviors in comparison to the low-resource experiments on the TED Talks corpus. First, the many-to-one model here performs better than the many-to-many model. This shows that the previous result was indeed due to the pathologies of the low-resource dataset; when the training data is large enough and not multi-way-parallel there is no overfitting in the many-to-one model, and it outperforms the many-to-many model in most cases while they are trained identically.

One particular outlier in this case is German-to-English, where the many-to-one model is 2 BLEU points below the many-to-many model. We examine the BLEU score of this language pair on its dedicated German-English development set during training in the many-to-one model and find that it highly fluctuates. We then measure the performance on the test set for this language pair by choosing the best checkpoint on the dedicated German-English development set (instead of on the mixed multilingual development set) and find it to be 38.07, which is actually *higher* in 1 BLEU than the best result of the many-to-many model. This shows that while training many languages together, there is no “silver bullet”: some languages may suffer from severe interference during training (i.e. a reduction of 3 BLEU in this case, from 38.07 to 35.05) while other languages continue to improve with more updates.

Table 6 describes the results when translating out-of-English. Again, both of the massively multilingual models perform better than the baselines when averaged across the 10 evaluated language pairs, while handling up to 102 languages to-and-from English and 204 translation tasks simultaneously. In this case the results are similar to those we observed on the TED talks corpus, where the one-to-many model performs better than the many-to-many model. Again, this advantage may be due to the one-to-many model handling a smaller number of tasks while not being biased towards English in the target side like the many-to-many model.

4 Analysis

The above results show that massively multilingual NMT is indeed possible in large scale settings and can improve performance over strong bilingual baselines. However, it was shown in a somewhat extreme case with more than 100 languages trained jointly, where we saw that in some cases the joint training may harm the performance for some language pairs (i.e. German-English above). In the following analysis we would like to better understand the trade-off between the number of languages involved and the translation accuracy while keeping the model capacity and training configuration fixed.

4.1 Multilinguality & Supervised Performance

We first study the effect of varying the number of languages on the translation accuracy in a supervised setting, where we focus on many-

	Ar-En	En-Ar	Fr-En	En-Fr	Ru-En	En-Ru	Uk-En	En-Uk	Avg.
5-to-5	**23.87**	**12.42**	**38.99**	**37.3**	29.07	**24.86**	**26.17**	16.48	**26.14**
25-to-25	23.43	11.77	38.87	36.79	**29.36**	23.24	25.81	**17.17**	25.8
50-to-50	23.7	11.65	37.81	35.83	29.22	21.95	26.02	15.32	25.18
75-to-75	22.23	10.69	37.97	34.35	28.55	20.7	25.89	14.59	24.37
103-to-103	21.16	10.25	35.91	34.42	27.25	19.9	24.53	13.89	23.41

Table 7: Supervised performance while varying the number of languages involved

	Ar-Fr	Fr-Ar	Ru-Uk	Uk-Ru	Avg.
5-to-5	1.66	4.49	3.7	3.02	3.21
25-to-25	1.83	**5.52**	**16.67**	4.31	7.08
50-to-50	**4.34**	4.72	15.14	**20.23**	**11.1**
75-to-75	1.85	4.26	11.2	15.88	8.3
103-to-103	2.87	3.05	12.3	18.49	9.17

Table 8: Zero-Shot performance while varying the number of languages involved

to-many models. We create four subsets of the in-house dataset by sub-sampling it to a different number of languages in each subset. In this way we create four additional English-centric datasets, containing 5, 25, 50 and 75 languages each to-and-from English. We make sure that each subset contains all the languages from the next smaller subsets – i.e. the 25 language subset contains the 5 language subset, the 50 language subset contains the 25 language subset and so on. We train a similar-capacity large Transformer model (with 473.7M parameters) on each of these subsets and measure the performance for each model on the 8 supervised language pairs from the smallest subset – {Arabic, French, Russian, Ukrainian}↔English. In this way we can analyze to what extent adding more languages improves or harms translation quality while keeping the model capacity fixed, testing the capacity vs. accuracy "saturation point".

Table 7 shows the results of this experiment, reporting the test results for the models that performed best on the multilingual development set. We can see that in most cases the best results are obtained using the 5-to-5 model, showing that there is indeed a trade off between the number of languages and translation accuracy when using a fixed model capacity and the same training setup. One may expect that the gaps between the different models should become smaller and even close with more updates, as the models with more languages see less examples per language in each batch, thus requiring more updates to improve in terms of BLEU. However, in our setting these gaps did not close even after the models converged, leaving 2.73 average BLEU difference between the 5-to-5 and the 103-to-103 model.

4.2 Multilinguality & Zero-Shot Performance

We then study the effect of the number of languages on zero-shot translation accuracy. Since we find zero-shot accuracy as an interesting measure for model generalization, we hypothesize that by adding more languages, the model is forced to create a more generalized representation to better utilize its capacity, which may improve zero-shot performance. We choose four language pairs for this purpose: Arabic↔French which are distant languages, and Ukrainian↔Russian which are similar. Table 8 shows the results of our models on these language pairs. For Arabic↔French the BLEU scores are very low in all cases, with the 50-to-50 and 25-to-25 models being slightly better than rest on Ar-Fr and Fr-Ar respectively. On Russian↔Ukrainian we see clear improvements when increasing the number of languages to more than five.

Figure 2 further illustrates this, showing the better generalization performance of the massively multilingual models under this zero-shot setting. While the zero-shot performance in this case is low and unstable for the 5-to-5 and 25-to-25 mod-

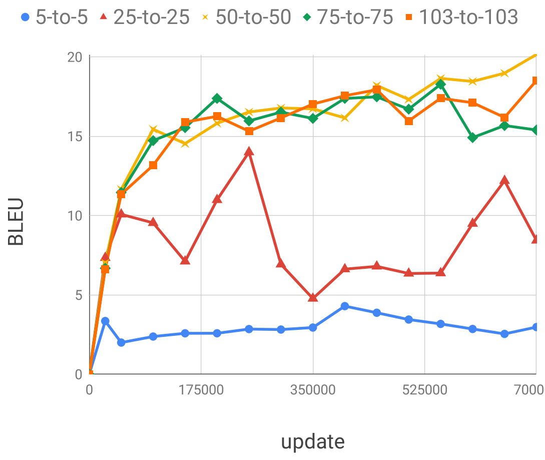

Figure 2: Zero-shot BLEU during training for Ukranian to Russian

els, it is much better for the 50-to-50, 75-to-75 and 103-to-103 models. Given these results we can say that the balance between capacity and generalization here favors the mid range 50-to-50 model, even when using models with more than 473M trained parameters. This may hint at the necessity of even larger models for such settings, which is a challenging avenue for future work. We also note that our 103 language corpus includes up to one million examples per language pair – while in real-world MT deployments, systems are trained on much more examples per pair. This again emphasizes the need for better techniques for training such massively multilingual models as we may already be hitting the capacity barrier in our setting.

5 Related Work

Dong et al. (2015) extended the NMT model of Bahdanau et al. (2014) to one-to-many translation (from English into 4 languages) by adding a dedicated decoder per target language, showing improvements over strong single-pair baselines. Firat et al. (2016a,b) proposed many-to-many models (with up to 6 languages) by using separate encoders and decoders per language while sharing the attention mechanism. They also introduced the notion of zero-resource translation, where they use synthetic training data generated through pivoting to train translation directions without available training data. Ha et al. (2016) and Johnson et al. (2017) proposed to use a shared encoder-decoder-attention model for many-to-many translation (with up to 7 languages in the latter). In order to determine the target language in such scenarios they proposed adding dedicated target-language symbols to the source. This method enabled zero-shot translation, showing the ability of the model to generalize to unseen pairs.

Recent works propose different methods for parameter sharing between language pairs in multilingual NMT. Blackwood et al. (2018) propose sharing all parameters but the attention mechanism and show improvements over sharing all parameters. Sachan and Neubig (2018) explore sharing various components in self-attentional (Transformer) models. Lu et al. (2018) add a shared "interlingua" layer while using separate encoders and decoders. Zaremoodi et al. (2018) utilize recurrent units with multiple blocks together with a trainable routing network. Platanios et al. (2018) propose to share the entire network, while using a contextual parameter generator that learns to generate the parameters of the system given the desired source and target languages. Gu et al. (2018) propose a "Universal Language Representation" layer together with a Mixture-of-Language-Experts component to improve a many-to-one model from 5 languages into English.

While the mentioned studies provide valuable contributions to improving multilingual models, they apply their models on only up to 7 languages (Johnson et al., 2017) and 20 trained directions (Cettolo et al., 2017) in a single model, whereas we focus on scaling NMT to much larger numbers of languages and trained directions. Regarding massively multilingual models, Neubig and Hu (2018) explored methods for rapid adaptation of NMT to new languages by training multilingual models on the 59-language TED Talks corpus and fine-tuning them using data from the new languages. While modeling significantly more languages than previous studies, they only train many-to-one models, which we show are inferior in comparison to our proposed massively multilingual many-to-many models when evaluated into English on this dataset.

Tiedemann (2018) trained an English-centric many-to-many model on translations of the bible including 927 languages. While this work pointed to an interesting phenomena in the latent space learned by the model where it clusters representations of typologically-similar languages together, it did not include any evaluation of the produced translations. Similarly, Malaviya et al. (2017) trained a many-to-English system including 1017 languages from bible translations, and used it to infer typological features for the different languages (without evaluating the translation quality). In another relevant work, Artetxe and Schwenk (2018) trained an NMT model on 93 languages and used the learned representations to perform cross-lingual transfer learning. Again, they did not report the performance of the translation model learned in that massively multilingual setting.

6 Conclusions and Future Work

We showed that NMT models can successfully scale to 102 languages to-and-from English with 204 trained directions and up to one million examples per direction. Such models improve the translation quality over similar single-pair baselines when evaluated to and from English by more than 2 BLEU when averaged over 10 diverse language pairs in each case. We show a similar result on the low-resource TED Talks corpus with 59 languages and 116 trained directions. We analyze the trade-offs between translation quality and the number of languages involved, pointing on capacity bottlenecks even with very large models and showing that massively multilingual models can generalize better to zero-shot settings.

We hope this work will encourage future research on massively multilingual NMT, enabling easier support for systems that can serve more people around the globe. There are many possible avenues for future work, including semi-supervised learning in such settings, exploring ways to reduce the performance degradation when increasing the number of languages, or using such models for multilingual transfer learning (McCann et al., 2017; Eriguchi et al., 2018; Artetxe and Schwenk, 2018). Understanding and improving zero-shot performance in such scenarios is also a promising direction for future work.

Acknowledgments

We would like to thank the Google Brain and Google Translate teams for their useful inputs and discussions. We would also like to thank the entire Lingvo development team for their foundational contributions to this project.

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

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

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*, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, et al. 2016. Findings of the 2016 conference on machine translation. In *ACL 2016 FIRST CONFERENCE ON MACHINE TRANSLATION (WMT16)*, pages 131–198.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*.

Mauro Cettolo, Federico Marcello, Bentivogli Luisa, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*.

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*.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

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*.

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*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Hany Hassan, Anthony Aue, Chang Chen, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Melvin Johnson, Mike Schuster, Quoc V Le, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA.

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*, Santa Fe, New Mexico, USA.

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*, Belgium, Brussels.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proc. IJCNLP*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

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

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*.

Ye Qi, Sachan Devendra, Felix Matthieu, Padmanabhan Sarguna, and Neubig Graham. 2018. When and why are pre-trained word embeddings useful for neural machine translation. In *HLT-NAACL*.

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*, Belgium, Brussels.

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

Jonathan Shen, Patrick Nguyen, Yonghui Wu, et al. 2019. Lingvo: a modular and scalable framework for sequence-to-sequence modeling. *arXiv preprint arXiv:1902.08295*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. *arXiv preprint arXiv:1802.00273*.

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.

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

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Yonghui Wu, Mike Schuster, Zhifeng Chen, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

A Supplementary Material

Language	Train set size
Arabic	214111
Hebrew	211819
Russian	208458
Korean	205640
Italian	204503
Japanese	204090
Chinese-Taiwan	202646
Chinese-China	199855
Spanish	196026
French	192304
Portuguese-Brazil	184755
Dutch	183767
Turkish	182470
Romanian	180484
Polish	176169
Bulgarian	174444
Vietnamese	171995
German	167888
Persian	150965
Hungarian	147219
Serbian	136898
Greek	134327
Croatian	122091
Ukrainian	108495
Czech	103093
Thai	98064
Indonesian	87406
Slovak	61470
Swedish	56647
Portuguese	51785
Danish	44940
Albanian	44525
Lithuanian	41919
Macedonian	25335
Finnish	24222
Burmese	21497
Armenian	21360
French-Canadian	19870
Slovenian	19831
Hindi	18798
Norwegian	15825
Kannada	13193
Estonian	10738
Kurdish	10371
Galician	10017
Marathi	9840
Mongolian	7607
Esperanto	6535
Tamil	6224
Urdu	5977
Azerbaijani	5946
Bosnian	5664
Chinese	5534
Malay	5220
Basque	5182
Bengali	4649
Belarusian	4509
Kazakh	3317

Table 9: Language pairs in the TED talks dataset (58 languages, paired with English) with the train-set size for each pair.

Languages	
Afrikaans	Laothian
Albanian	Latin
Amharic	Latvian
Arabic	Lithuanian
Armenian	Luxembourgish*
Azerbaijani	Macedonian
Basque	Malagasy
Belarusian	Malay
Bengali	Malayalam
Bosnian	Maltese
Bulgarian	Maori
Burmese	Marathi
Catalan	Mongolian
Cebuano	Nepali
Chichewa*	Norwegian
Chinese	Pashto
Corsican*	Persian
Croatian	Polish
Czech	Portuguese
Danish	Punjabi
Dutch	Romanian
Esperanto	Russian
Estonian	Samoan*
Finnish	Scots Gaelic*
French	Serbian
Frisian	Sesotho
Galician	Shona*
Georgian	Sindhi*
German	Sinhalese
Greek	Slovak
Gujarati	Slovenian
Haitian Creole	Somali
Hausa*	Spanish
Hawaiian*	Sundanese
Hebrew	Swahili
Hindi	Swedish
Hmong*	Tagalog
Hungarian	Tajik*
Icelandic	Tamil
Igbo	Telugu
Indonesian	Thai
Irish	Turkish
Italian	Ukrainian
Japanese	Urdu
Javanese	Uzbek
Kannada	Vietnamese
Kazakh	Welsh
Khmer	Xhosa
Korean	Yiddish
Kurdish	Yoruba*
Kyrgyz	Zulu

Table 10: Language pairs in the in-house dataset (102 languages, paired with English). For languages marked with * we had less than 1M examples, while for the rest we used exactly 1M.