Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, Yonghui Wu
Google Research
{adisid, ankurbpn, yuancao, orhanf, miachen, snehakudugunta, navari, yonghui}@google.com

Abstract

Over the last few years two promising research directions in low-resource neural machine translation (NMT) have emerged. The first focuses on utilizing high-resource languages to improve the quality of low-resource languages via multilingual NMT. The second direction employs monolingual data with self-supervision to pre-train translation models, followed by fine-tuning on small amounts of supervised data. In this work, we join these two lines of research and demonstrate the efficacy of monolingual data with self-supervision in multilingual NMT. We offer three major results: (i) Using monolingual data significantly boosts the translation quality of low-resource languages in multilingual models. (ii) Self-supervision improves zero-shot translation quality in multilingual models. (iii) Leveraging monolingual data with self-supervision provides a viable path towards adding new languages to multilingual models, getting up to 33 BLEU on WMT ro-en translation without any parallel data or back-translation.

1 Introduction

Recent work has demonstrated the efficacy of multilingual neural machine translation (multilingual NMT) on improving the translation quality of low-resource languages (Firat et al., 2016; Aharoni et al., 2019) as well as zero-shot translation (Ha et al., 2016; Johnson et al., 2017; Arivazhagan et al., 2019b). The success of multilingual NMT on low-resource languages relies heavily on transfer learning from high-resource languages for which copious amounts of parallel data is easily accessible. However, existing multilingual NMT approaches often do not effectively utilize the abundance of monolingual data, especially in low-resource languages. On the other end of the spectrum, self-supervised learning methods, consuming only monolingual data, have achieved great success on transfer learning (Devlin et al., 2019) and unsupervised NMT (Lample et al., 2018; Artetxe et al., 2018) without fully benefiting from the rich learning signals offered by the bilingual data of multiple languages.

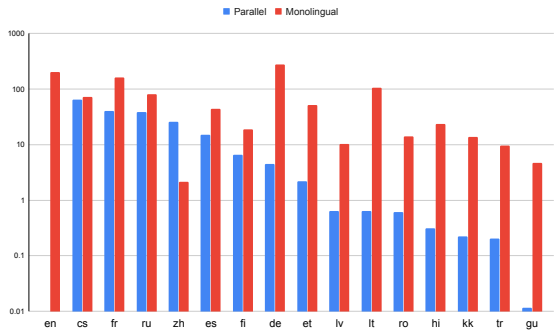


Figure 1: Number of parallel and monolingual training samples in millions for each language in WMT training corpora.

In this work, we propose to combine the beneficial effects of multilingual NMT with the self-supervision from monolingual data. Compared with multilingual models trained without any monolingual data, our approach shows consistent improvements in the translation quality of all languages, with greater than 10 BLEU points improvements on certain low-resource languages. We further demonstrate improvements in zero-shot translation, where our method has almost on-par quality with pivoting-based approaches, without using any alignment or adversarial losses. The most interesting aspect of this work, however, is that we introduce a path towards effectively adding new unseen languages to a multilingual NMT model, showing strong translation quality on several language pairs by leveraging only monolingual data with self-supervised learning, without the need for any parallel data for the new languages.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2827–2835
July 5 - 10, 2020.

xx	cs	fr	ru	zh	es	fi	de	et	lv	lt	ro	hi	kk	tr	gu
Any-to-English (xx→en)	31.3	37.2	36.0	21.7	32.7	27.3	31.7	23.1	15.0	21.3	30.1	8.5	11.5	15.9	1.0
English-to-Any (en→xx)	23.8	41.3	26.4	31.3	31.1	18.1	29.9	18.2	14.2	11.5	23.4	4.5	1.9	13.6	0.6

Table 1: Bilingual baselines. xx refers to language in the column header.

2 Method

We propose a co-training mechanism that combines supervised multilingual NMT with monolingual data and self-supervised learning. While several pre-training based approaches have been studied in the context of NMT (Dai and Le, 2015; Conneau and Lample, 2019; Song et al., 2019), we proceed with Masked Sequence-to-Sequence (MASS) (Song et al., 2019) given its success on unsupervised and low-resource NMT, and adapt it to the multilingual setting.

2.1 Adapting MASS for multilingual models

MASS adapts the masked de-noising objective (Devlin et al., 2019; Raffel et al., 2019) for sequence-to-sequence models, by masking the input to the encoder and training the decoder to generate the masked portion of the input. To utilize this objective function for unsupervised NMT, Song et al. (2019) enhance their model with additional improvements, including language embeddings, target language-specific attention context projections, shared target embeddings and softmax parameters and high variance uniform initialization for target attention projection matrices[1].

We use the same set of hyper-parameters for self-supervised training as described in (Song et al., 2019). However, while the success of MASS relies on the architectural *modifications* described above, we find that our multilingual NMT experiments are stable even in the absence of these techniques, thanks to the smoothing effect of multilingual joint training. We also forego the separate source and target language embeddings in favour of pre-pending the source sentences with a $< 2xx >$ token (Johnson et al., 2017).

We train our models simultaneously on supervised parallel data using the translation objective and on monolingual data using the MASS objective. To denote the target language in multilingual NMT models we prepend the source sentence with the $< 2xx >$ token denoting the target language.

[1]Verified from open-source Github implementation.

3 Experimental Setup

3.1 Datasets

We use the parallel and monolingual training data provided with the WMT corpus, for 15 languages to and from English. The amount of parallel data available ranges from more than 60 million sentence pairs as in En-Cs to roughly 10k sentence pairs as in En-Gu. We also collect additional monolingual data from WMT news-crawl, news-commentary, common-crawl, europarl-v9, news-discussions and wikidump datasets in all 16 languages including English.[2] The amount of monolingual data varies from 2 million sentences in Zh to 270 million in De. The distribution of our parallel and monolingual data is depicted in Figure 1.

3.2 Data Sampling

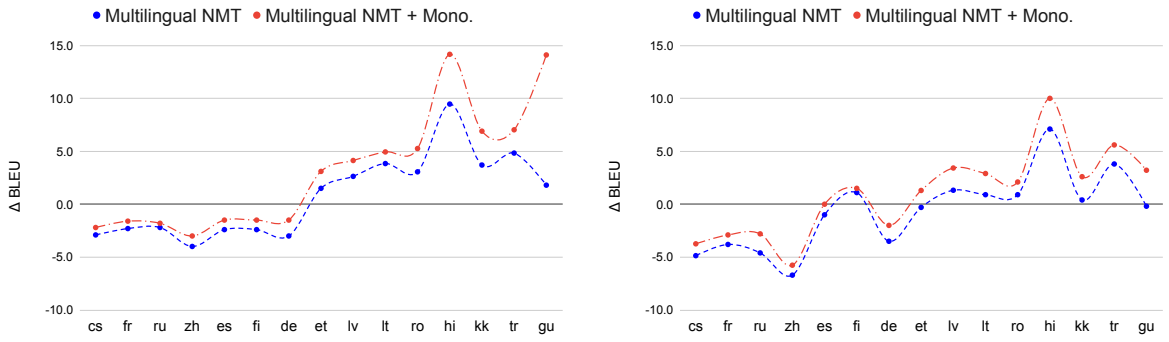
Given the data imbalance across languages in our datasets, we use a temperature-based data balancing strategy to over-sample low-resource languages in our multilingual models (Arivazhagan et al., 2019b). We use a temperature of $T = 5$ to balance our parallel training data. When applicable, we sample monolingual data uniformly across languages since this distribution is not as skewed. For experiments that use both monolingual and parallel data, we mix the two sources at an equal ratio (50% monolingual data with self-supervision and 50% parallel data).

3.3 Architecture and Optimization

All experiments are performed with the Transformer architecture (Vaswani et al., 2017) using the open-source Tensorflow-Lingvo implementation (Shen et al., 2019). Specifically, we use the Transformer Big model containing 375M parameters (6 layers, 16 heads, 8192 hidden dimension) (Chen et al., 2018) and a shared source-target SentencePiece model (SPM)[3] (Kudo and Richardson, 2018). We use a vocabulary size of 32k for the bilingual models and 64k for the multilingual mod-

[2]Followed the versions recommended by WMT'19 shared task, as in http://statmt.org/wmt19/translation-task.html

[3]https://github.com/google/sentencepiece



(a) Any-to-English (xx → en)

(b) English-to-Any (en → xx)

Figure 2: Translation quality of Multilingual NMT models relative to bilingual baselines with and without monolingual data. The left plot shows xx → en direction and right one shows en → xx direction. From left to right on x-axis, we go from high-resource to low-resource languages. The x-axis reflects the bilingual baselines.

els. Different SPMs are trained depending on the set of languages supported by the model.

4 Using Monolingual Data for Multilingual NMT

We evaluate the performance of the models using SacreBLEU (Post, 2018) on standard WMT validation and test sets (Papineni et al., 2002). The performance of our bilingual baselines for all 30 English-centric language pairs are reported in Table 1. We compare the performance of bilingual models, multilingual models trained with just supervised data for 30 language pairs (15 languages to and from English) and multilingual models trained with a combination of supervised and monolingual data in Figure 2.

High-Resource Translation Our results suggest that a single multilingual model is able to match the quality of individual bilingual models with a gap of less than 2 BLEU points for most high-resource languages, with the exception of Chinese (Zh). The slight quality regression is not surprising, given the large number of languages competing for capacity within the same model (Arivazhagan et al., 2019b). We find that adding additional monolingual data improves the multilingual model quality across the board, even for high-resource language pairs.

Low-Resource Translation From Figure 2, we observe that our supervised multilingual NMT model significantly improves the translation quality for most low and medium-resource languages compared with the bilingual baselines. Adding additional monolingual data leads to an additional improvement of 1-2 BLEU for most medium-resource languages. For the lowest-resource languages like Kazakh (kk), Turkish (tr) and Gujarati (gu), we can see that multilingual NMT alone is not sufficient to reach high translation quality. The addition of monolingual data has a large positive impact on very low resource languages, significantly improving quality over the supervised multilingual model. These improvements range from 3-5 BLEU in the en→xx direction to more than 5 BLEU for the xx→en translation.

Zero-Shot Translation We next evaluate the effect of training on additional monolingual data on zero-shot translation in multilingual models. Table 2 demonstrates the zero-shot performance of our multilingual model that is trained on 30 language pairs, and evaluated on French(fr)-German(de) and German(de)-Czech(cs), when trained with and without monolingual data. To compare with the existing work on zero-shot translation, we also evaluate the performance of multilingual models trained on just the relevant languages (en-fr-de for fr-de translation, en-cs-de for cs-de translation). We observe that the additional monolingual data significantly improves the quality of zero-shot translation, often resulting in 3-6 BLEU increase on all zero-shot directions compared to our multilingual baseline. We hypothesize that the additional monolingual data seen during the self-supervised training process helps better align representations across languages, akin to the smoothing effect in semi-supervised learning (Chapelle et al., 2010). We leave further exploration of this intriguing phenomenon to future work.

		fr_de	de_fr	cs_de	de_cs
4 lang.	w/ Parallel Data	27.7	35.3	—	—
	Translation via Pivot	21.9	29.2	20.4	19.0
	Arivazhagan et al. (2019a)	20.3	26.0	—	—
	Kim et al. (2019)	17.3	—	—	14.1
	Multilingual NMT	11.8	15.2	12.3	8.2
	Multilingual NMT + Mono.	18.5	27.2	16.9	12.6
30 lang.	Multilingual NMT	10.3	14.2	10.5	4.3
	Multilingual NMT + Mono.	16.6	22.3	14.8	7.9

Table 2: Zero-shot performance on non-English centric language pairs. We compare with pivot-based translation and two recent approaches from Arivazhagan et al. (2019a) and Kim et al. (2019). The translation quality between these language pairs when parallel data is available is also provided as a baseline. 4 lang. is a multilingual model trained on 4 language pairs (2 languages to and from English), while 30 lang. is our multilingual model trained on all English-centric language pairs.

	fr_en	en_fr	de_en	en_de	ro_en	en_ro	lt_en	en_lt	lv_en	en_lv	hi_en	en_hi
Multilingual NMT	34.9	37.5	28.7	26.4	33.2	24.3	25.1	12.4	17.6	15.5	18.0	11.6
Mono. Only	9.8	7.6	7.4	5.8	6.8	7.3	4.8	2.1	2.9	1.8	5.3	3.1
Multilingual NMT - xx	8.4	2.4	3.9	2.6	6.2	3.8	2.2	1.1	2.1	1.7	0.8	0.6
Multilingual NMT - xx + Mono.	30.7	9.8	24.2	8.9	33.0	9.3	21.3	6.7	18.8	6.1	14.6	5.4

Table 3: Translation quality of the new language added to Multilingual NMT using just monolingual data. Multilingual NMT here is a multilingual model with 30 language pairs, Mono. Only is a bilingual model used as a baseline trained with only monolingual data with self-supervised learning, Multilingual NMT-xx is a multilingual model trained on 28 language pairs (xx is the language not present in the model). Multilingual NMT-xx + Mono. is a multilingual model with 28 language pairs but only monolingual data for xx.

5 Adding New Languages to Multilingual NMT

Inspired by the effectiveness of monolingual data in boosting low-resource language translation quality, we continue with a stress-test in which we completely remove the available parallel data from our multilingual model, one language at a time, in order to observe the unsupervised machine translation quality for the missing language.

Results of this set of experiments are detailed in Table 3. We find that simply adding monolingual data for a new language to the training procedure of a multilingual model is sufficient to obtain strong translation quality for several languages, often attaining within a few BLEU points of the fully supervised multilingual baseline, without the need for iterative back-translation. We also notice significant quality improvements over models trained with just self-supervised learning using monolingual data for a variety of languages. On WMT ro-en, the performance of our model exceeds XLM (Conneau and Lample, 2019) by over 1.5 BLEU and matches bilingual MASS (Song et al., 2019), without utilizing any back-translation. This suggests that jump-starting the iterative back-translation process from multilingual models might be a promising avenue to supporting new languages.

6 Related Work

Our work builds on several recently proposed techniques for multilingual NMT and self-supervised representation learning. While massively multilingual models have obtained impressive quality improvements for low-resource languages as well as zero-shot scenarios (Aharoni et al., 2019; Arivazhagan et al., 2019a), it has not yet been shown how these massively multilingual models could be extended to unseen languages, beyond the pipelined approaches (Currey and Heafield, 2019; Lakew et al., 2019). On the other hand, self-supervised learning approaches have excelled at down-stream cross-lingual transfer (Devlin et al., 2019; Raffel et al., 2019; Conneau et al., 2019), but their success for unsupervised NMT (Conneau and Lample,

2019; Song et al., 2019) currently lacks robustness when languages are distant or monolingual data domains are mismatched (Neubig and Hu, 2018; Vulić et al., 2019). We observe that these two lines of research can be quite complementary and can compensate for each other's deficiencies.

7 Conclusion and Future Directions

We present a simple framework to combine multilingual NMT with self-supervised learning, in an effort to jointly exploit the learning signals from multilingual parallel data and monolingual data. We demonstrate that combining multilingual NMT with monolingual data and self-supervision (i) improves the translation quality for both low and high-resource languages in a multilingual setting, (ii) leads to on-par zero-shot capability compared with competitive bridging-based approaches and (iii) is an effective way to extend multilingual models to new unseen languages.

Future work should explore techniques like iterative back-translation (Hoang et al., 2018) for further improvement and scaling to larger model capacities and more languages (Arivazhagan et al., 2019b; Huang et al., 2019) to maximize transfer across languages and across data sources.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

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

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. 2010. *Semi-Supervised Learning*, 1st edition. The MIT Press.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Association for Computational Linguistics*.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*.

Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In *Workshop on Neural Generation and Translation*.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *International Workshop on Spoken Language Translation*.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Workshop on Neural Machine Translation and Generation*.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. 2019. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Advances in Neural Information Processing Systems*.

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

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-english languages. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Empirical Methods in Natural Language Processing: System Demonstrations*.

Surafel M Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. 2019. Adapting multilingual neural machine translation to unseen languages. In *International Workshop on Spoken Language Translation*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Conference on Machine Translation*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

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

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*.

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*.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*.

A Appendices

Language Pair	Data Sources			# Samples		
	Train	Dev	Test	Train	Dev	Test
cs→en	WMT'19	WMT'17	WMT'18	64336053	3005	2983
fr→en	WMT'15	WMT'13	WMT'14	40449146	3000	3003
ru→en	WMT'19	WMT'18	WMT'19	38492126	3000	2000
zh→en	WMT'19	WMT'18	WMT'19	25986436	3981	2000
es→en	WMT'13	WMT'13	WMT'13	15182374	3004	3000
fi→en	WMT'19	WMT'18	WMT'19	6587448	3000	1996
de→en	WMT'14	WMT'13	WMT'14	4508785	3000	3003
et→en	WMT'18	WMT'18	WMT'18	2175873	2000	2000
lv→en	WMT'17	WMT'17	WMT'17	637599	2003	2001
lt→en	WMT'19	WMT'19	WMT'19	635146	2000	1000
ro→en	WMT'16	WMT'16	WMT'16	610320	1999	1999
hi→en	WMT'14	WMT'14	WMT'14	313748	520	2507
kk→en	WMT'19	WMT'19	WMT'19	222424	2066	1000
tr→en	WMT'18	WMT'17	WMT'18	205756	3007	3000
gu→en	WMT'19	WMT'19	WMT'19	11670	1998	1016
en→cs	WMT'19	WMT'17	WMT'18	64336053	3005	2983
en→fr	WMT'15	WMT'13	WMT'14	40449146	3000	3003
en→ru	WMT'19	WMT'18	WMT'19	38492126	3000	2000
en→zh	WMT'19	WMT'18	WMT'19	25986436	3981	2000
en→es	WMT'13	WMT'13	WMT'13	15182374	3004	3000
en→fi	WMT'19	WMT'18	WMT'19	6587448	3000	1996
en→de	WMT'14	WMT'13	WMT'14	4508785	3000	3003
en→et	WMT'18	WMT'18	WMT'18	2175873	2000	2000
en→lv	WMT'17	WMT'17	WMT'17	637599	2003	2001
en→lt	WMT'19	WMT'19	WMT'19	635146	2000	1000
en→ro	WMT'16	WMT'16	WMT'16	610320	1999	1999
en→hi	WMT'14	WMT'14	WMT'14	313748	520	2507
en→kk	WMT'19	WMT'19	WMT'19	222424	2066	1000
en→tr	WMT'18	WMT'17	WMT'18	205756	3007	3000
en→gu	WMT'19	WMT'19	WMT'19	11670	1998	1016
fr→de	WMT'19	WMT'13	WMT'13	9824476	1512	1701
de→fr	WMT'19	WMT'13	WMT'13	9824476	1512	1701
cs→de	—-	WMT'13	WMT'13	—	1997	1997
de→cs	—-	WMT'13	WMT'13	—	1997	1997

Table 4: Data sources and number of samples for the parallel data in our corpus. Please note that we don't use parallel data in Fr-De for any of the experiments in the paper apart from training parallel data baseline in Table 2. We don't have any parallel data in Cs-De.

Language	Data Sources						# Samples		
	News Crawl	News Commentary	Common Crawl	Europarl	News Discussions	Wiki Dumps	Train	Dev	Test
en	✓						199900557	3000	3000
ro	✓						14067879	3000	3000
de	✓						275690481	3000	3000
fr	✓	✓		✓	✓		160933435	3000	3000
cs	✓						72157988	3000	3000
es	✓						43814290	3000	3000
et	✓		✓				51683012	3000	3000
fi	✓			✓			18847600	3000	3000
gu	✓		✓				4644638	3000	3000
hi	✓						23611899	3000	3000
kk	✓	✓	✓			✓	13825470	3000	3000
lt	✓		✓	✓		✓	106198239	3000	3000
lv	✓			✓			10205015	3000	3000
ru	✓						80148714	3000	3000
tr	✓						9655009	3000	3000
zh	✓	✓					2158309	3000	3000

Table 5: Data sources and number of samples for the monolingual data in our corpus.

Language Pair	Bilingual Baseline	Multilingual NMT	Multilingual NMT + Mono.	SOTA
cs→en	29.7	28.4	29.1	33.9
fr→en	35.5	34.9	35.6	39.5
ru→en	34.9	33.8	34.1	40.1
zh→en	21.7	17.7	18.7	39.3
es→en	30.1	28.9	29.6	31.4
fi→en	26.0	25.2	25.8	33.0
de→en	27.4	27.2	28.1	32.0
et→en	24.3	24.2	24.9	30.9
lv→en	15.0	17.6	18.8	36.3
lt→en	21.3	24.4	25.4	36.3
ro→en	30.1	33.0	34.1	38.5
hi→en	8.5	16.0	18.5	16.7
kk→en	4.7	11.2	17.6	30.5
tr→en	15.9	18.4	21.1	28.0
gu→en	2.0	3.0	15.1	24.9
en→cs	23.8	20.0	20.3	29.9
en→fr	38.1	36.2	36.6	43.8
en→ru	24.9	22.0	22.9	36.3
en→zh	31.3	5.0	5.9	36.3
en→es	32.8	29.7	30.0	30.4
en→fi	20.3	19.2	19.6	27.4
en→de	26.4	22.1	23.9	27.1
en→et	19.0	18.9	20.1	25.2
en→lv	14.2	14.9	16.5	21.1
en→lt	11.0	10.9	14.4	20.1
en→ro	23.7	23.6	24.8	33.3
en→hi	4.5	10.6	13.9	12.5
en→kk	0.2	1.1	4.3	11.1
en→tr	13.7	13.8	15.7	20.0
en→gu	0.6	0.4	4.0	28.2

Table 6: Absolute BLEU scores for results in Figure 2 in the paper.