Can Big Models Help Diverse Languages? Investigating Large Pretrained Multilingual Models for Machine Translation of Indian Languages

Telem Joyson Singh
IIT Guwahati
Assam, India
tjoyson@iitg.ac.in

Sanasam Ranbir Singh
IIT Guwahati
Assam, India
ranbir@iitg.ac.in

Priyankoo Sarmah
IIT Guwahati
Assam, India
priyankoo@iitg.ac.in

Abstract

Machine translation of Indian languages is challenging due to several factors, including linguistic diversity, limited parallel data, language divergence, and complex morphology. Recently, large pre-trained multilingual models have shown promise in improving translation quality. In this paper, we conduct a large-scale study on applying large pre-trained models for English-Indic machine translation through transfer learning across languages and domains. This study systematically evaluates the practical gains these models can provide and analyzes their capabilities for the translation of the Indian language by transfer learning. Specifically, we experiment with several models, including Meta's mBART, mBART-many-to-many, NLLB-200, M2M-100, and Google's MT5. These models are fine-tuned on small, high-quality English-Indic parallel data across languages and domains. Our findings show that adapting large pre-trained models to particular languages by fine-tuning improves translation quality across the Indic languages, even for languages unseen during pretraining. Domain adaptation through continued fine-tuning improves results. Our study provides insights into utilizing large pretrained models to address the distinct challenges of MT of Indian languages. [1]

1 Introduction

India is a linguistically diverse nation with 22 official scheduled languages and numerous dialects representing various language families like Indo-Aryan, Dravidian, and Tibeto-Burman. This linguistic diversity presents unique challenges for developing machine translation systems between Indian languages. Complex morphology, linguistic diversity, language divergence, and limited parallel data make translation of Indian languages difficult. Recent advances in neural machine translation (NMT) have improved translation quality for many language pairs. However, performance for low-resource Indian languages still lags behind European and East Asian languages. Recently, pretrained multilingual language models like mBART (Liu et al., 2020) and translation models like M2M-100 (Fan et al., 2021) have shown promising results by leveraging knowledge transferred from high-resource languages through large-scale pretraining. These models can translate between multiple languages with higher quality than previous approaches by leveraging their broad linguistic knowledge gained through pretraining. However, their capabilities have not been thoroughly explored for Indian languages on a large scale specifically.

This paper conducts a large-scale study on applying pretrained models to English-Indic machine translation across diverse languages and domains. Specifically, we investigate several models, including Meta's mBART (Liu et al., 2020), mBART50-many-to-many (Tang et al., 2020), NLLB-200 (team et al., 2022), M2M-100 (Fan et al., 2021), and Google's mT5 (Xue et al., 2021) for translation of Indian languages. The models are fine-tuned on high-quality human-translated parallel data between English and 16 Indic languages. The languages encompass those included during model pretraining along with unseen languages to cover high to low-resource scenarios. Through extensive experiments, we evaluate the improvements enabled by transfer learning and examine model capabilities for handling Indian language diversity. Our analysis provides practical insights into utilizing pretrained models for English-Indic translation.

Our contributions are as follows:

- We conduct a large-scale study on applying pretrained models like mBART, mBART-many-to-many, NLLB-200, M2M-100, and MT5 to English-Indic machine translation across 16 diverse Indic languages.

[1] Code is available at https://github.com/joysontelem/lpmm-indicmt

- Through extensive experiments fine-tuning models on small parallel data, we evaluate the improvements enabled by adapting pretrained models to a particular language or domain, and analyze model capabilities for handling Indian language diversity.

2 Related Work

Multilingual Neural Machine Translation (MNMT) traces its roots back to the early days of Neural Machine Translation (NMT), as evidenced by the works of Dong et al. (2015); Firat et al. (2016). The introduction of Google's end-to-end MNMT by Johnson et al. (2017) marked a significant milestone in the field, as it allowed for multilingual NMT within a single encoder-decoder model. Google's MNMT also demonstrated "zero-shot" translation, allowing for the translation of languages not explicitly provided during training. However, it had limitations, supporting only translation between languages seen during training.

Facebook AI expanded the scope of multilingual translation with Transformer-based NMT structures, such as mBART-m2m (Tang et al., 2020), M2M-100 (Fan et al., 2021), and NLLB (team et al., 2022). These models offered translation capabilities for 50, 100, and 200+ languages. Although they faced the challenge of translating entirely new languages not present in their pre-training data, they represented a significant development in the field of multilingual NMT. Additionally, transfer learning from pre-trained multilingual language models, such as BART (Liu et al., 2020) and MT5 (Xue et al., 2021), often outperforms traditional bilingual approaches.

Research on multilingual NMT for Indian languages has gained traction in recent years, driven by the linguistic diversity of the Indian subcontinent. Notable work includes that of Ramesh et al. (2022), who developed a multilingual NMT model for 11 Indian languages. AI4Bharat et al. (2023) further expanded to more Indian languages with their MultiMT model covering English and 22 Indian languages. Recent work by Dabre et al. (2022) investigates multilingual NMT for 11 Indian languages with a BART based pretrained model. Another works by Singh et al. (2021a) and Lee et al. (2022) also investigates MT5 and mBART pretrained language models for MT of three/four Indian languages. Unlike the previous works, which only use large pretrained denoising pretrained models like mBART and MT5, our work focuses on large pretrained translation models like mBART-m2m, M2M100, and NLLB on a diverse set of 16 Indian languages.

3 Focus Languages

We focus on 16 official languages of India with varying quantities of available data, including high-resource languages such as Hindi and Bengali and very low-resource languages such as Sanskrit, with ancient texts being its available corpus. Table 1 provides an overview of the focus languages, including the language families, location and number of speakers, and the source and original language for our corpus. The languages are from three language families: Indo-Aryan (e.g. Hindi), Dravidian (e.g. Tamil), and Tibeto-Burman (e.g. Manipuri). Most of the languages are from the Indo-Aryan family, India's largest language family. All languages, except for Sanskrit, are spoken by at least one million people.

Code	Language	Family	Script	#Speakers
asm	Assamese	Indo-Aryan	Bengali	15 million
ben	Bengali	Indo-Aryan	Bengali	205 million
brx	Bodo	Tibeto-Burman	Devanagari	1.5 million
doi	Dogri	Indo-Aryan	Devanagari	2.6 million
gom	Konkani	Indo-Aryan	Devanagari	2.2 million
hin	Hindi	Indo-Aryan	Devanagari	322 million
kan	Kannada	Dravidian	Kannada	40 million
kas	Kashmiri	Indo-Aryan	Arabic	5.5 million
mai	Maithili	Indo-Aryan	Devanagari	33 million
mal	Malayalam	Dravidian	Malayalam	38 million
mar	Marathi	Indo-Aryan	Devanagari	71 million
mni	Manipuri	Tibeto-Burman	Bengali	1.8 million
npi	Nepali	Indo-Aryan	Devanagari	16 million
san	Sanskrit	Indo-Aryan	Devanagari	24,821
tam	Tamil	Dravidian	Tamil	69 million
tel	Telugu	Dravidian	Telugu	74 million
urd	Urdu	Indo-Aryan	Arabic	60 million

Table 1: Overview of Focus Languages

4 Models and Methodology

4.1 Pretrained Multilingual Models

We conducted experiments using pre-trained multilingual models. Our selection of pre-trained models was based on their size, covering approximately 400 to 600 million parameters, as well as their ability to provide comparability across various models. Table 2 provides information on the size of the pre-trained models, the number of Indian languages they support, and the focus languages they cover.

Pretrained model	#parameters	#Focus Languages Covered
mT5	580M	ben, hin, kan, mal, mar, npi, tam, tel
mBART50	610M	ben, hin, mal, mar, npi, tam, tel
mBart50-m2m	610M	ben, hin, mal, mar, npi, tam, tel
M2M-100	418M	ben, hin, kan, mal, mar, npi, tam
NLLB-200	600M	asm, ben, hin, kan, kas, mai, mal, mar, mni, npi, san, tam, tel

Table 2: Overview of Pretrained Multilingual Models

mBART mBART is a multilingual Sequence-to-Sequence model that can be used for various natural language processing tasks, including translation. The model can be fine-tuned for specific applications, including translation and other NLP tasks, which provides flexibility for researchers and developers.

mT5 mT5 is a multilingual variant of the T5 model. It has been pre-trained on a large dataset that covers 101 different languages. Like T5, mT5-Base follows the text-to-text transformer architecture. This means it can handle various natural language processing tasks by framing them as text-to-text problems.

mBART50-many-to-many This model is a fine-tuned checkpoint of mBART (Liu et al., 2020). MBART50-many-to-many is fine-tuned for multilingual machine translation. It was introduced in Tang et al. (2020). The model can translate directly between any pair of 50 languages.

M2M-100 M2M-100 is a multilingual encoder-decoder model designed for Many-to-Many multilingual translation tasks. This model has been trained on an impressive 2,200 language directions. This extensive training data allows it to perform exceptionally well in diverse language pairs.

NLLB-200 NLLB-200 is a machine translation model, especially for low-resource languages. It's a powerful tool for multilingual translation tasks, supporting up to 200 different languages.

4.2 Fine-Tuning Strategies for Multilingual Models

Adaptation to a Particular Language We apply the transfer learning technique to adapt pretrained models to particular Indian languages. To achieve this, we fine-tune each of the pretrained models listed above on a small general domain human-translated English-Indic dataset. In cases where a language was not included in the pretrained models, we assigned it the language code of a related language that was included in the pretrained model. After considering the language family, syntactic and morphological similarities between languages, we have chosen the following pairs: ben for asm and mni, hin for brx, doi, gom, mai, and san, tam for kan, mal, and tel, and urd for kas.

Adaptation to a Particular Domain The MT model can be fine-tuned to fit a specific domain or style. A set of bilingual sentences representing the domain or style that the MT model should adapt to may be needed for fine-tuning. For a fixed domain MT problem, fine-tuning pre-trained models on a specific domain has grown to be a preferred strategy. We employ pretrained multilingual model for fine-tuning in both directions. We use the PMI dataset, an MT Indic dataset composed of political data sources, such as news commentary and speeches from the Indian Prime Minister.

5 Experimental Setup

In this section, we discuss the data and settings used in our experiments.

Language	Size	Language	Size
Assamese	44k	Maithili	24k
Bengali	48k	Malayalam	41k
Bodo	21k	Marathi	50k
Dogri	17k	Manipuri	20k
Konkani	17k	Nepali	45k
Hindi	40k	Sanskrit	27k
Kannada	32k	Tamil	21k
Kashmiri	21k	Telugu	29k

Table 3: Parallel Corpora used in Language Adaptation

Language	Size	Language	Size
Bengali	23k	Manipuri	5k
Hindi	50k	Tamil	32k
Marathi	25k		

Table 4: Parallel Corpora used in Domain Adaptation

We conducted experiments to fine-tune pre-trained models for 16 different languages. For 14 of these languages, we used the BPCC-wiki dataset (AI4Bharat et al., 2023) to focus on language-specific features. For the remaining two languages (Manipuri and Kashmiri), we used the

En ↦ X	asm	ben	brx	doi	gom	hin	kan	kas	mai	mal	mar	mni	npi	san	tam	tel
							0-SHOT									
mBART50-m2m	0.23	1.12	0.37	1.83	0.80	21.14	0.76	0.94	4.60	3.12	1.15	0.14	10.83	0.68	8.97	4.77
M2M-100	0.78	8.00	0.32	2.08	0.57	26.27	0.12	1.34	5.18	2.16	4.92	0.53	0.36	0.81	2.27	-
NLLB	8.39	17.05	0.40	2.47	0.74	31.36	17.68	5.18	13.13	12.87	14.49	5.32	16.61	1.07	16.18	20.27
							FINETUNED									
mBART50-m2m	10.28	15.85	7.97	21.66	9.82	29.53	14.58	5.23	11.13	14.02	13.78	6.55	18.26	2.08	14.75	17.36
M2M-100	10.93	14.81	8.18	15.62	9.51	28.24	10.74	5.58	10.92	10.79	11.51	6.88	16.35	1.87	11.55	-
NLLB	11.02	19.47	8.62	24.69	10.52	32.86	18.72	5.03	13.81	16.90	16.37	6.49	20.48	2.12	18.50	23.84
mBART	8.71	15.03	7.07	19.19	7.80	27.12	12.64	4.05	9.51	12.10	13.06	5.58	16.81	1.89	11.37	16.35
mT5	6.29	10.63	1.52	9.24	7.84	14.41	8.66	1.4	6.20	7.74	8.20	0.45	11.89	1.04	7.78	10.89

X ↦ En	asm	ben	brx	doi	gom	hin	kan	kas	mai	mal	mar	mni	npi	san	tam	tel
							0-SHOT									
mBART50-m2m	1.18	9.30	1.24	3.63	2.59	32.45	0.71	3.71	8.88	23.77	14.98	0.55	29.64	3.86	23.54	15.49
M2M-100	1.65	24.26	1.00	1.47	2.12	29.60	0.38	1.52	4.76	15.53	19.52	0.79	8.09	1.57	8.02	-
NLLB	28.27	33.98	1.54	11.20	7.15	38.96	32.15	28.55	39.30	34.39	33.97	23.23	38.25	18.82	31.19	37.96
							FINETUNED									
mBART50-m2m	20.13	24.58	18.40	26.38	18.86	33.42	21.63	21.28	28.84	27.36	27.83	16.76	32.97	14.71	24.55	27.97
M2M-100	20.90	24.76	19.36	26.30	18.46	30.47	19.68	21.25	29.04	22.77	25.16	18.41	28.15	14.80	17.48	-
NLLB	28.67	33.93	24.31	35.26	25.11	39.97	32.26	30.88	41.71	35.33	35.28	25.02	39.62	22.45	31.83	38.26
mBART	17.77	20.44	14.06	20.07	13.84	26.93	10.86	17.59	23.96	21.76	21.66	8.64	26.54	11.99	16.81	22.42
mT5	15.59	19.70	7.72	14.91	14.62	19.65	17.18	12.37	18.80	18.54	19.25	4.84	22.82	9.49	15.13	19.28

Table 5: **Adaptation to Particular language.** Translation performance (BLEU) of English-Indic languages in 0-shot and finetuned settings. Adaptation to particular languages by finetuning on small parallel data significantly improves over 0-shot transfer.

NLLB-seed (team et al., 2022) and ILCI (Choudhary and Jha, 2014) corpus. We used the FLORES (Goyal et al., 2022; team et al., 2022) dev set and devtest set for 13 out of 16 languages for development and testing. For the remaining three languages (Bodo, Konkani, and Dogri), we used a randomly selected 1k sentences as the dev set and the FLORES dev set translated by BPCC corpus as the test set.

To adapt the pretrained models to a specific domain, we used the PMI dataset (Haddow and Kirefu, 2020). It contains speeches and news from the Indian Prime Minister. For development and testing, we used the WAT 2021 development and test sets (Nakazawa et al., 2021). For languages not included in the WAT 2021 set, we randomly sampled sentences from the corpus. It is worth noting that the fine-tuning parallel data for all languages were less than 50k sentences, as shown in the Table 3 and 4.

We utilized the HuggingFace Transformers library (Wolf et al., 2020) for implementation and trained on NVIDIA Tesla V100 GPUs to take advantage of hardware-level parallelism. Our translations were evaluated using SacreBLEU (Post, 2018).

6 Result and Discussion

We discuss the results of transfer across languages and domain in this section.

6.1 Adaptation to Particular Language

Our experiments show that adapting pretrained models to a particular language improves translation performance across the Indic languages as shown in Table 5. The pretrained models are able to leverage knowledge transferred from high-resource languages during pretraining to boost translation quality for low-resource Indian languages. For example, mBART-m2m achieves an average improvement of over 7 BLEU points on the En ↦ X translation task. Similarly, for the X ↦ En translation direction, mBART-m2m outperforms baselines by over 10 BLEU points on average, with additional gains. The improvements are pronounced even for low-resource languages unseen during pretraining like Bodo, Dogri, and Manipuri. This shows the capability of models to generalize and adapt to new languages. For instance, NLLB boosts Bodo translations by over 8 BLEU points despite no Bodo data during pretraining. We also observe that translating into English outperforms translating from English in morphologically rich Indian languages

like Manipuri, Tamil, and Marathi (Singh et al., 2023). This can be attributed to BLEU's ignorance of subwords.

6.2 Adaptation to Particular Domain

Table 6 shows the results of evaluating the M2M-100 model on the PMIndia dataset, both with and without domain-specific fine-tuning.

En ↦ X	ben	hin	mar	mni	tam
M2M-100 0-shot	3.39	21.22	3.95	0.28	1.43
M2M-100 finetune	9.85	28.80	13.62	15.83	10.23

X ↦ En	ben	hin	mar	mni	tam
M2M-100 0-shot	13.78	26.05	13.40	1.89	3.43
M2M-100 finetune	22.38	36.78	25.39	29.65	24.00

Table 6: **Adaptation to Particular Domain.** Performance of M2M-100 on PMIndia before and after fine-tuning on in-domain data.

The results show that fine-tuning M2M-100 on in-domain pairs from PMIndia boosts performance by over 6 BLEU points in English-to-Indic translations and by over 9 BLEU points in Indic-to-English translations. Interestingly, translation to and from Manipuri scores higher than in other languages, except Hindi, despite having a smaller dataset. Translation data for this language may be more domain-specific than those for others (Singh et al., 2021b). The fine-tuned model demonstrates improved fluency and terminology usage for politically-related sentences. This indicates that adapting pretrained models to a particular domain is an effective technique to further specialize the models and achieve gains in domain-specific contexts beyond generic translation.

7 Conclusion and Future work

In conclusion, this paper shows that large pretrained models, such as mBART50-m2m, NLLB-200, and M2M-100, can significantly improve the quality of English-Indic translations through transfer learning, across diverse Indian languages. Our analysis shows adapting these models by fine-tuning on small amounts of high-quality parallel data for a particular language or domain substantially boosts translation quality. Improvements are seen even for low-resource languages unseen during model pretraining. This indicates pretrained models acquire linguistic knowledge that transfers across languages to handle complex morphology and diversity.

Future work includes expanding language coverage and exploring semi-supervised methods for continued improvements. As pretrained models grow in scale and linguistic breadth, their capabilities for Indic language translation will also improve. We hope our study provides insights to leverage these large pretrained models to address the unique challenges of Indian language diversity.

References

AI4Bharat, Jay Gala, Pranjal A. Chitale, AK Raghavan, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *ArXiv*, abs/2305.16307.

Narayan Choudhary and Girish Nath Jha. 2014. Creating multilingual parallel corpora in indian languages. In *Human Language Technology Challenges for Computer Science and Linguistics*, pages 527–537, Cham. Springer International Publishing.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1).

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

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Barry Haddow and Faheem Kirefu. 2020. Pmindia - a collection of parallel corpora of languages of india. *ArXiv*, abs/2001.09907.

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.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Salam Michael Singh, Loitongbam Sanayai Meetei, Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021a. On the transferability of massively multilingual pretrained models in the pretext of the indo-aryan and tibeto-burman languages. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 64–74, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Telem Joyson Singh, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2021b. English-manipuri machine translation: An empirical study of different supervised and unsupervised methods. *2021 International Conference on Asian Language Processing (IALP)*, pages 142–147.

Telem Joyson Singh, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2023. Subwords to word back composition for morphologically rich languages in neural machine translation. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

Nllb team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.