How far can we get with one GPU in 100 hours?
CoAStaL at MultiIndicMT Shared Task

Rahul Aralikatte **Héctor Ricardo Murrieta Bello** **Miryam de Lhoneux**
Daniel Hershcovich **Marcel Bollmann** **Anders Søgaard**
Department of Computer Science
University of Copenhagen
{rahul,ml,dh,marcel,soegaard}@di.ku.dk xhd160@alumni.ku.dk

Abstract

This work shows that competitive translation results can be obtained in a constrained setting by incorporating the latest advances in memory and compute optimization. We train and evaluate large multilingual translation models using a single GPU for a maximum of 100 hours and get within 4-5 BLEU points of the top submission on the WAT 2021 leaderboard. We also benchmark standard baselines on the PMI corpus and re-discover well-known shortcomings of current translation metrics.

1 Introduction

Machine Translation is one of the few tasks in NLP which has the luxury of data. Due to the efforts of the community over the last two decades (Koehn, 2005; Tiedemann, 2012, 2020), most major languages of the world have millions of translated sentence pairs with English. With the introduction of sequence to sequence models (Sutskever et al., 2014; Cho et al., 2014), transformers (Vaswani et al., 2017), and large pre-trained language models (Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019; Liu et al., 2019), the accuracy of machine translation models has almost risen to that of humans (Wu et al., 2016). Yet, the ability to train such models is limited by the availability of compute. Today's state-of-the-art models are trained by industry research labs, using large compute infrastructure which is usually unavailable or unaffordable to others. Such training is also shown to have large carbon footprints (Strubell et al., 2019).

In this work, we show that competitive translation performance can be achieved even with limited resources. We first train a statistical MT system that does not require GPUs, as a baseline. Next, we run inference on the best publicly available pre-trained models to benchmark their performance. Finally, we train graph2seq, seq2seq, and text2text models, which progressively perform better. All our experiments are constrained both in compute[1] and training time: we use one NVIDIA Titan RTX GPU for a maximum of 100 hours. Our main findings are: (i) pre-trained seq2seq and text2text models perform the best, especially when trained only on the PMI corpus, (ii) the benefits of pre-trained multilingual language models diminish for Indic language decoding due to their under-representation in pre-training data, and (iii) a small empirical evaluation on 2 languages shows that the prediction fluency and faithfulness start plateauing at 100 hours.

Source	Language(s)
CVIT-PIB (2020)	BN,GU,HI,ML,MR,OR,PA,TA,TE
JW (2019)	BN,GU,HI,KN,ML,MR,PA,TA,TE
TED (2012)	BN,GU,HI,KN,ML,MR,PA,TA,TE
PMIndia (2020)	BN,GU,HI,KN,ML,MR, OR,PA,TA,TE
Bible-uedin (2014)	GU,HI,KN,ML,MR,TE
OpenSubtitles (2016)	BN,HI,ML,TA,TE
WikiMatrix (2019)	HI,ML,MR,TA,TE
Wiki Titles (2021)	GU,TA
ALT (2016)	BN,HI
IITB 3.0 (2018)	HI
NLPC (2020)	TA
UFAL EnTam (2012)	TA
Uka Tarsadia (2019)	GU
MTEnglish2Odia (2018)	OR
OdiEnCorp 2.0 (2020)	OR

Table 1: Sources of MultiIndicMT data.

2 Data

The MultiIndicMT data is a combination of parallel corpora from different sources as shown in Tab. 1. It contains translations from 10 Indic languages to English. The Indic languages included are Bengali (BN), Gujarati (GU), Hindi (HI), Kannada (KN), Malayalam (ML), Marathi (MR), Oriya (OR), Punjabi (PA), Tamil (TA), and Telugu (TE). The training split contains about 11 million translation pairs

[1] which in turn constrains the number of model parameters

Proceedings of the 8th Workshop on Asian Translation, pages 205–211
Bangkok, Thailand (online), August 5-6, 2021. ©2021 Association for Computational Linguistics

Sentence	Translation	Fluency	Faithfulness	Comment
Big Business!	बड़ा व्यवसाय !	5	5	-
It affects all of us . It helps us — and it harms us .	हम सब पर इसका प्रभाव पड़ता है ।	5	3	Translation of second sentence missing
And there are things we can do about it . A giant , or " big , " corporation may have assets worth $ 1,500,000,000 .	यह हमारी सहायता करता है — और हमें हानि भी पहुँचाता है ।	5	0	Translation of the previous line's second sentence
Many have far more . That kind of money represents power .	और कुछ ऐसी बातें हैं जो हम उसके बारे में कर सकते हैं ।	5	0	Misalignment
Giant corporations have tussled with countries — and won . No wonder so many are suspicious of them !	एक विशाल , या " बड़े " निगम के पास शायद १,५०,००,००,००० डॉलर की सम्पत्ति हो ।	5	0	Misalignment

Figure 1: A sample from the JW corpus which shows misalignments between the translation pairs.

from these languages. The development and test splits contain 1000 and 2390 11-way parallel sentences taken from the PMIndia corpus (Haddow and Kirefu, 2020), respectively.

Analysis To understand the data better, a small analysis is performed by randomly sampling 100 sentences from each language the authors can read (HI and KN). Overall, the translations are of high quality, except in a few sources where the parallel sentences are automatically extracted. For example, we found that JW (Agić and Vulić, 2019) has alignment issues, where a part of the translation is moved to the next line, thereby starting a chain of misalignments, as shown in Fig. 1. We manually annotate 100 translations for fluency and faithfulness on a scale of 0-5 and get a score of 4.01 for fluency and 3.54 for faithfulness.

3 Models

We train four types of models: (i) a phrase-based statistical model, (ii) a graph-to-text model, (iii) a sequence-to-sequence model, and (iv) a text-to-text model. Brief descriptions of the models are given below.

3.1 Moses

We train a statistical phrase-based model with Moses (Koehn et al., 2007) using default settings, following the guidelines for training a baseline.[2] We prune words that occur less than three times in the corpus and use the same tokenizer as for the other models and de-tokenize predictions before evaluating. We train a separate model for each language pair and use the respective development set for tuning before translating the test set.

3.2 GRAPH-TO-TEXT model

We also train a graph2seq model with a GCN (Kipf and Welling, 2016) encoder and LSTM decoder. In addition to text, we input the source syntax trees obtained from a parser trained on Universal Dependencies (Nivre et al., 2016). We borrow hyperparameter settings from Bastings et al. (2017) and input a bag of source words to the encoder and expect subword units from the decoder. We train separate models for each language pair.

3.3 SEQ2SEQ model

For training multilingual models, we use pre-trained transformer-based language models to initialize the encoder and decoder of our seq2seq models. For English, we use standard uncased BERT-Base (Devlin et al., 2019) and for Indic languages, we use MuRIL (Khanuja et al., 2021). MuRIL's architecture is similar to BERT and is pre-trained on 17 Indic languages including all ten required for our translation task. It is pre-trained on publicly available corpora from Wikipedia and Common Crawl. It also uses automatically translated and transliterated data for pre-training. We add cross-attention between the encoder and decoder following Rothe et al. (2020).

The model has 375M trainable parameters. When the decoder is multilingual, we follow previous works and force a language identifier as the BOS token. We use a learning rate of 5×10^{-5} and a batch size of 12. We truncate sequences to a maximum length of 128 and use a cosine learning rate scheduler with a warmup of 10,000 steps. We denote our models as BERT2MURIL and MURIL2BERT when translating from and to En-

[2] http://www.statmt.org/moses/?n=Moses.Baseline

Model	Bn		Gu		Hi		Kn		Ml		Mr		Or		Pa		Ta		Te	
	chrF	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**
m2m100 (418M)	0.30	2.98	0.11	0.40	0.48	21.21	0.15	0.05	0.21	0.69	0.31	3.96	0.05	0.06	0.11	0.66	0.20	1.43	-	-
m2m100 (1.2B)	0.35	4.48	0.16	1.35	0.49	22.22	0.18	0.15	0.28	1.29	0.35	6.19	0.05	0.04	0.16	1.84	0.23	1.26	-	-
Moses (PMI)	0.40	4.90	0.46	12.4	0.48	15.7	0.44	8.00	0.39	2.60	0.41	7.50	0.42	8.40	0.44	14.1	0.42	5.20	0.35	3.40
Moses (all)	0.38	5.00	0.47	13.0	0.51	18.0	0.43	7.90	0.41	3.50	0.45	9.50	0.44	10.5	0.44	14.5	0.42	7.00	0.36	3.60
GCN (PMI)	0.40	5.20	0.48	14.3	0.50	17.1	0.44	9.10	0.36	2.10	0.41	7.30	0.40	8.90	0.46	16.7	0.48	8.20	0.35	4.90
mT5-large (PMI)	0.40	7.14	**0.52**	**20.8**	**0.55**	**26.5**	**0.52**	15.0	**0.46**	5.37	0.46	12.6	-	-	0.48	20.8	**0.49**	**10.1**	0.39	3.89
mT5-large (all)	0.36	5.52	0.46	16.0	0.54	**26.5**	0.45	9.18	0.42	3.83	0.41	9.40	-	-	0.43	16.9	0.47	8.46	0.35	3.79
bert2muril (PMI)	0.42	7.68	0.51	19.6	0.53	23.5	0.49	14.0	0.43	5.62	0.46	12.8	0.46	13.6	0.49	21.8	0.46	8.30	0.39	6.04
bert2muril (all)	0.37	5.09	0.50	18.9	0.53	23.3	0.45	11.0	0.38	3.95	0.46	12.3	0.48	14.8	0.48	19.4	0.43	7.03	0.36	4.68
+FT on PMI	**0.44**	**8.89**	**0.52**	20.2	**0.55**	25.5	**0.52**	**16.0**	**0.46**	**5.91**	**0.48**	**14.3**	**0.49**	**15.3**	**0.52**	**24.1**	**0.49**	9.83	**0.41**	**6.54**

Table 2: Character F1 and BLEU scores of English to Indic translations.

Model	Bn		Gu		Hi		Kn		Ml		Mr		Or		Pa		Ta		Te	
	chrF	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**	**chrF**	**bleu**
m2m100 (418M)	0.43	13.8	0.10	0.18	0.55	26.0	0.10	0.08	0.32	7.45	0.40	13.4	0.15	0.75	0.36	9.65	0.25	2.44	-	-
m2m100 (1.2B)	0.44	14.7	0.10	0.16	0.55	26.7	0.09	0.16	0.36	11.0	0.41	14.1	0.15	0.42	0.36	9.08	0.21	2.91	-	-
Moses (PMI)	0.37	6.00	0.46	10.6	0.49	12.6	0.42	8.30	0.39	5.90	0.41	7.50	0.41	7.60	0.44	10.0	0.38	6.20	0.41	7.00
Moses (all)	0.40	8.00	0.48	12.5	0.52	16.1	0.43	8.80	0.43	8.50	0.44	10.4	0.43	10.6	0.48	13.0	0.42	9.60	0.43	8.80
GCN (PMI)	0.40	8.30	0.49	12.8	0.54	14.8	0.44	10.9	0.48	11.5	0.48	13.5	0.46	7.20	0.45	12.6	0.43	14.3	0.45	14.7
mT5-large (PMI)	**0.51**	**24.2**	**0.60**	**34.5**	**0.62**	**36.3**	**0.57**	**30.9**	**0.55**	**28.4**	**0.54**	**27.5**	-	-	**0.61**	**35.7**	**0.53**	**26.6**	**0.57**	**30.4**
mT5-large (all)	0.49	21.5	0.59	31.9	**0.62**	35.2	0.55	27.9	0.53	25.7	0.52	25.3	-	-	0.59	33.4	0.51	24.4	0.54	26.5
muril2bert (PMI)	0.48	16.6	0.56	24.3	0.59	26.9	0.54	22.1	0.52	20.5	0.51	19.8	**0.51**	**20.1**	0.57	26.1	0.50	19.2	0.53	21.3
muril2bert (all)	0.37	11.0	0.41	13.8	0.46	17.1	0.41	13.3	0.40	13.0	0.39	12.0	0.38	11.8	0.42	14.6	0.39	12.1	0.41	13.1
+FT on PMI	0.47	16.6	0.55	24.0	0.58	26.5	0.53	21.7	0.52	20.5	0.51	19.6	0.50	19.7	0.57	25.5	0.50	19.1	0.52	21.2

Table 3: Character F1 and BLEU scores of Indic to English translations.

glish, respectively.[3]

3.4 TEXT2TEXT model

To push the extent to which a single GPU can be utilized, we also train the large multilingual-T5 (mT5-large; Xue et al., 2020) model on our translation task. This model is pre-trained on mC4, a multilingual version of the Common Crawl consisting of text from 101 languages. It contains 1.2B trainable parameters which do not fit on our 24GB GPU, even if trained with mixed-precision and a batch size of one. Therefore, we resort to optimizer state and gradient partitioning with ZeRO (Rajbhandari et al., 2020). ZeRO is a zero-redundancy optimizer that offloads some computations and memory to the host's CPU and provides a better GPU management system that uses smart allocation methods to reduce memory fragmentation. For more details, see Rasley et al. (2020). With these modifications, we train the model with a learning rate of 3×10^{-5}. All other hyper-parameters remain unchanged.

4 Results

We report results in both English to Indic, and Indic to English directions. We use character F1 and BLEU (Papineni et al., 2002), which are standard metrics to evaluate translations. We train two variants of all models: (i) only on the PMI corpus, and (ii) on the full training data. The English to Indic results are shown in Tab. 2 and the Indic to English results, in Tab. 3.[4]

m2m100 We first benchmark the performance of the Many-to-Many multilingual model (m2m100; Fan et al., 2020) which is trained on non-English centric translation. It can translate to and from all Indic languages in our task, except Telugu. As expected, with no finetuning, both the small (418M parameters) and large (1.2B parameters) models perform poorly, on all languages except Hindi. This is expected as the other languages are under-represented in the mC4 dataset.

Moses We see that simple phrase-based translation works relatively well. Though significantly worse than the best model, Moses produces results comparable to that of mT5-large (all) in both directions. Although this can be attributed to mT5-large being under-trained, it gives us an insight into

[3] This is the largest model we could train on our GPU without using optimization tricks.

[4] Note that we report local evaluation metrics which do not exactly match with the leaderboard numbers because of the differences in tokenization. We do this to avoid uploading multiple prediction files and overloading the evaluation server.

Language	En→*		*→En	
	Loc.	Off.	Loc.	Off.
Bengali	8.89	11.1	24.2	24.4
Gujarati	20.8	20.4	34.5	34.6
Hindi	26.5	31.7	36.3	36.5
Kannada	16.0	16.1	30.9	31.0
Malayalam	5.91	6.27	28.4	28.5
Marathi	14.3	14.5	27.5	27.7
Oriya	15.3	15.7	20.1	19.6
Punjabi	24.1	27.2	35.7	35.9
Tamil	10.1	10.0	26.6	26.7
Telugu	6.54	12.9	30.4	30.5

Table 4: Comparison of BLEU scores obtained during local and official evaluations.

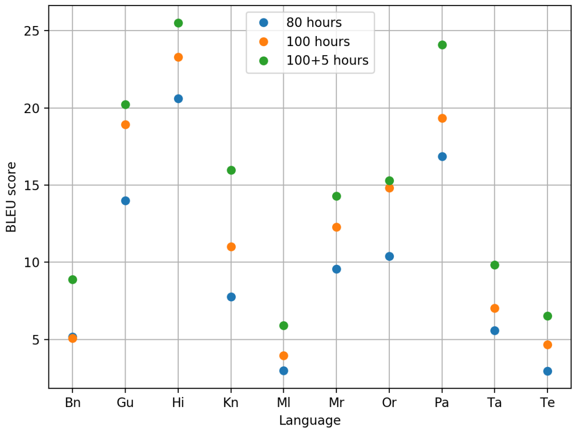

Figure 2: Increase in BLEU score across languages when trained on the full training data, at different intervals of time.

the ability of simpler models to learn quickly in constrained environments. We also note that just training on the PMI corpus gives a result that is almost on par with the results obtained by training on the entire training split. The model trained on PMI even surpasses the other model, on Kannada indicating a strong in-domain training bias.

GCN In this setup, we only train on the PMI corpus due to time constraints. We find that while it comfortably surpasses Moses, it also comes close to the much bigger models, especially when translating to Indic languages. It is to be noted that, this small gap in results can be mainly attributed to the lack of convergence of the bigger models, as discussed next.

mT5 mT5 can translate to and from all Indic languages required by our task, except Oriya. We note that the model trained only on the PMI corpus is always better than the model trained on the complete data. We postulate that 100 hours is not enough time for the model to converge on the full data. We also see that mT5's performance is far superior compared to all other models for Indic to English translation. This may be expected as the model is pre-trained to generate fluent English text. For English to Indic translation, mT5 performs on-par or slightly worse than bert2muril finetuned on PMI data, except for Hindi and Tamil, where it is better.

MuRIL and BERT Following the mT5 models, these models also perform better when trained only on the PMI corpus as it fails to converge on the larger data in the given time. As an additional step, we finetune these under-fit models on the PMI data for 5 hours and see a significant performance gain in the English to Indic direction (bert2muril). The model outperforms the much bigger mT5 on 7 languages with Gujarati, Hindi, and Tamil being the exceptions. However, finetuning does not seem to have a major effect in the other direction (muril2bert). As in the case of mT5, we believe that the BERT decoder's pre-training subsumes any gains from extra finetuning.

Official Evaluation Since Tab. 2 and 3 show BLEU scores obtained by evaluating the generated predictions locally, they do not exactly match the official scores on the leaderboard.[5] For a fair comparison, we present both local and official BLEU scores of our best submissions in Tab. 4. We see that the scores are similar when translating from Indic languages to English. But when translating from English, the official scores are often significantly higher. This is a result of our use of minimal tokenization (mteval-v13a) before computing BLEU, while the official evaluation uses the Indic-tokenizer (Kunchukuttan, 2020).

5 Discussion

As reported in §4, the text2text and seq2seq models perform better when trained only on the PMI corpus when compared to them being trained on the entire train split. Though it can be argued that they perform better since the test set also comes from the same domain,[6] we hypothesize that 100

[5] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

[6] The development and test sets are taken from the PMI corpus.

hours is not enough time for the models to converge when trained on the full training set. Fig 2 shows the BLEU scores of BERT2MuRIL model after 80 and 100 hours of training, respectively. We see that the model gets significantly better in the last 20 hours. A 5 hour finetuning with the PMI corpus, further increases its performance. This clearly shows that the model would become more accurate if it is trained for a longer period or with more compute.

To establish whether an increase in BLEU scores corresponds to an increase in the fluency and faithfulness of the translations, we manually annotate 50 Hindi and Kannada test predictions from the best model to find that the increase in both cases is marginal. In the 20 additional training hours, the fluency and faithfulness increased by 0.005 and 0.01 respectively which suggests that BLEU may not be the best metric to quantify the goodness of translation systems, as shown in works like Zhang et al. (2004); Callison-Burch et al. (2006).

6 Conclusion

In this work, we show that it is possible to get competitive translation results using a single GPU for a limited amount of time by carefully selecting and training large pre-trained encoder-decoder models. We also show that we can train models which have more than 10^9 trainable parameters using the latest advances in GPU resource optimization. Finally, through a small empirical study, we find that while longer training can increase BLEU scores, it may not increase their fluency and faithfulness.

Acknowledgements

We thank the reviewers for their valuable feedback. Rahul Aralikatte and Anders Søgaard were funded by a Google Focused Research Award. Miryam de Lhoneux was funded by the Swedish Research Council (grant 2020-00437). Marcel Bollmann was funded from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 845995.

References

2018. MTEnglish2Odia. https://odianlp.github.io/. Accessed: 2021-05-20.

2021. Wiki Titles. http://data.statmt.org/wikititles/v3/. Accessed: 2021-05-20.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967, Copenhagen, Denmark. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Christos Christodoulopoulos and Mark Steedman. 2014. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49:1–21.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2020. Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. *CoRR*, abs/2011.02821.

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

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Shantipriya Parida, Satya Ranjan Dash, Ondřej Bojar, Petr Motlicek, Priyanka Pattnaik, and Debasish Kumar Mallick. 2020. OdiEnCorp 2.0: Odia-English parallel corpus for machine translation. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 14–19, Marseille, France. European Language Resources Association (ELRA).

Jerin Philip, Shashank Siripragada, Vinay P. Namboodiri, and C. V. Jawahar. 2020. Revisiting low resource status of indian languages in machine translation. *CoRR*, abs/2008.04860.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press.

Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012. Morphological processing for english-tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, pages 113–122.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

Parth Shah and Vishvajit Bakrola. 2019. Neural machine translation system of indic languages - an attention based approach. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pages 1–5.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the Asian language treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1574–1578, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *LREC*.