Attentive fine-tuning of Transformers for Translation of low-resourced languages @LoResMT 2021

Karthik Puranik[1] **Adeep Hande**[1] **Ruba Priyadharshini**[2] **Thenmozi Durairaj**[3]
Anbukkarasi Sampath[4] **Kingston Pal Thamburaj**[5] **Bharathi Raja Chakravarthi**[6]
[1]Department of Computer Science, Indian Institute of Information Technology Tiruchirappalli
[2]ULTRA Arts and Science College, Madurai, Tamil Nadu, India
[3]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India
[4]Kongu Engineering College, Erode, Tamil Nadu, India
[5]Sultan Idris Education University, Tanjong Malim, Perak, Malaysia
[6]Insight SFI Research Centre for Data Analytics, National University of Ireland Galway
{{karthikp, adeeph}18c@iiitt.ac.in, rubapriyadharshini.a@gmail.com, theni_d@ssn.edu.in,
anbu.1318@gmail.com, fkingston@gmail.com, bharathi.raja@insight-centre.org}

Abstract

This paper reports the Machine Translation (MT) systems submitted by the IIITT team for the English→Marathi and English⇔Irish language pairs LoResMT 2021 shared task. The task focuses on getting exceptional translations for rather low-resourced languages like Irish and Marathi. We fine-tune IndicTrans, a pretrained multilingual NMT model for English→Marathi, using external parallel corpus as input for additional training. We have used a pretrained Helsinki-NLP Opus MT English⇔Irish model for the latter language pair. Our approaches yield relatively promising results on the BLEU metrics. Under the team name IIITT, our systems ranked 1, 1, and 2 in English→Marathi, Irish→English, and English→Irish respectively. The codes for our systems are published[1].

1 Introduction

Today, a large number of text and written materials are present in English. However, with roughly around 6,500 languages in the world[2] (Chakravarthi, 2020; Hande et al., 2021a; Sarveswaran et al., 2021), every native monoglot should not be deprived of this knowledge and information. The manual translation is a tedious job involving much time and human resources, giving rise to Machine Translation (MT). Machine Translation involves the automated translation of text from one language to another by using various algorithms and resources to produce quality translation predictions (Pathak and Pakray, 2018; Krishnamurthy, 2015, 2019). Neural Machine Translation (NMT) brought about a great improvement in the field of MT by overcoming flaws of rule-based and statistical machine translation (SMT) (Revanuru et al., 2017; Achchuthan and Sarveswaran, 2015; Parameswari et al., 2012; Thenmozhi et al., 2018; Kumar et al., 2020b). NMT incorporates the training of neural networks on parallel corpora to predict the likeliness of a sequence of words. sequence-to-sequence neural models (seq2seq)

[1]https://github.com/karthikpuranik11/LoResMT
[2]https://blog.busuu.com/most-spoken-languages-in-the-world/

(Sutskever et al., 2014; Kalchbrenner and Blunsom, 2013) are the widely adopted as the standard approach by both industrial and research communities (Jadhav, 2020a; Bojar et al., 2016; Cheng et al., 2016).

Even though NMT performs exceptionally well for all the languages, it requires a tremendous amount of parallel corpus to produce meaningful and successful translations (Kumar et al., 2020a). With little research on low resourced languages, finding a quality parallel corpus to train the models can be arduous. The two low-resourced languages worked on in this paper are Marathi (mr) and Irish (ga). With about 120 million Marathi speakers in Maharashtra and other states of India, Marathi is recognized as one of the 22 scheduled languages of India[3]. The structural dissimilarity which occurs while translating from English (Subject-Verb-Object) to Marathi (Subject-Object-Verb) or vice versa adds up to issues faced while translation (Garje, 2014). The Irish language was recognized as the first official language of Ireland and also by the EU(Dowling et al., 2020a; Scannell, 2007). Belonging to the Goidelic language family and the Celtic family(Scannell, 2007; Lynn et al., 2015), Irish is also claimed as one of the low resourced languages due to its limited resources by the META-NET report (Dhonnchadha et al., 2012; Scannell, 2006).

Our paper represents the work conducted for the LoResMT @ MT Summit 2021[4] shared task to build MT systems for the low-resourced Marathi and Irish languages on COVID-19 related parallel corpus. We implement Transformer-based (Vaswani et al., 2017) NMT models to procure BLEU scores(Papineni et al., 2002) of 24.2, 25.8, and 34.6 in English→Marathi, Irish→English, and English→Irish respectively.

2 Related works

Neural Machine Translation has been exhaustively studied over the years (Kalchbrenner and Blunsom, 2013), with several intuitive approaches involving collective learning to align and translate (Bahdanau et al., 2016), and a language-independent attention bridge for multilingual translational systems (Vázquez et al., 2019). There have been several approaches to NMT, with zero-shot translational systems, between language pairs that have not seen the parallel training data during training (Johnson et al., 2017). The introduction of artificial tokens has reduced the architectural changes in the decoder (Ha et al., 2016). There have been some explorations towards neural machine translation in low resource languages, with the development of a multi-source translational system that targets the English string for any source language (Zoph and Knight, 2016).

There has been subsequent research undertaken by researchers for machine translation in low-resource Indian languages. Chakravarthi et al. surveyed orthographic information in machine translation, examining the orthography's influence on machine translation and extended it to under-resourced Dravidian languages (Chakravarthi et al., 2019a). Another approach of leveraging the information contained in rule-based machine translation systems to improve machine translation of low-resourced languages was employed (Torregrosa et al., 2019). Several approaches involving the improvement of WordNet for low-resourced languages have been explored (Chakravarthi et al., 2018, 2019b). Chakravarthi et al. constructed MMDravi, a multilingual multimodal machine translation dataset for low-resourced Dravidian languages, extending it from the Flickr30K dataset, and generating translations for the captions using phonetic transcriptions (U Hegde et al., 2021).

There have been relatively fewer approaches experimented with and benchmarked when it comes to translating from Marathi to English and vice versa. (Aharoni et al., 2019) tried to build multilingual NMT systems, comprising 103 distinct languages and 204 translational directions

[3] https://en.wikipedia.org/wiki/Marathi_language

[4] https://sites.google.com/view/loresmt/

simultaneously. (Jadhav, 2020b; Puranik et al., 2021) developed a machine translation system for Marathi to English using transformers on a parallel corpus. Other works on improving machine translation include (Adi Narayana Reddy et al., 2021) proposing an English-Marathi NMT using local attention. The same can be stated to Irish, as it is a poorly resourced language, as the quality of the MT outputs have struggled to achieve the same level as well-supported languages (Dowling et al., 2016; Rehm and Uszkoreit, 2012). In recent years, several researchers tried to overcome the resource barrier by creating artificial parallel data through back-translation (Poncelas et al., 2018), exploiting out-of-domain data (Imankulova et al., 2019), and leveraging other better-resourced languages as a pivot (Dowling et al., 2020b; Wu and Wang, 2007).

3 Dataset

We use the dataset provided by the organizers of LoResMT @ MT Summit 2021. The datasets can be found here[5]. It is a parallel corpus for English and the low resourced language, i.e., Irish and Marathi, mostly containing text related to COVID-19 (Ojha et al., 2021).

Language pair	English⇔Irish	English⇔Marathi
Train	8,112	20,933
Dev	502	500
Test	1,000	1,000
Total	9,614	22,433

Table 1: Number of sentences distribution

We have used bible-uedin[6] (Christodoulopoulos and Steedman, 2015) an external dataset for Marathi. It is a multilingual parallel corpus dataset containing translations of the Bible in 102 languages(Christodoulopoulos and Steedman, 2014) and shows the possibility of using the Bible for research and machine translation. English-Marathi corpus contains 60,495 sentences. CVIT PIB[7] (Philip et al., 2020) has also been used for the purpose of this research. It contains 1,14,220 parallel corpora for English-Marathi.

4 Methodology

4.1 IndicTrans

Fairseq PyTorch[8] (Ott et al., 2019) is an open-source machine learning library supported as a sequence modeling toolkit. Custom models can be trained for various tasks, including summarization, language, translation, and other generation tasks. Training on fairseq enables competent batching, mixed-precision training, multi-GPU and multi-machine training. IndicTrans (Ramesh et al., 2021), a Transformer-4x multilingual NMT model by AI4Bharat, is trained on the Samanantar dataset. The architecture of our approach is displayed in Fig.1. Samanantar[9] is the most extensive collection of parallel corpora for Indic languages available for public use. It includes 46.9 million sentence pairs between English and 11 Indian languages. IndicTrans is claimed to successfully outperform the existing best performing models on a wide variety of benchmarks. Even commercial translation systems and existing publicly available systems were surpassed for the majority of the languages. IndicTrans is based on fairseq, and it was

[5] https://github.com/loresmt/loresmt-2021
[6] https://opus.nlpl.eu/JRC-Acquis.php
[7] http://preon.iiit.ac.in/~jerin/bhasha/
[8] https://github.com/pytorch/fairseq
[9] https://indicnlp.ai4bharat.org/samanantar/

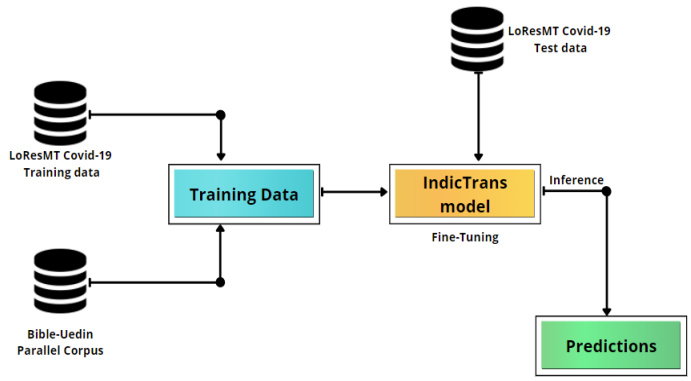

Figure 1: Our approach for the English→Marathi language pair.

fine-tuned on the Marathi training dataset provided by the organizers and the external datasets. The model was fine-tuned with the cross-entropy criterion to compute the loss function, Adam optimizer (Zhang, 2018), dropout of 0.2, fp16 (Micikevicius et al., 2017), maximum tokens of 256 for better learning, and a learning rate if 3e-5 in GPU. The process was conducted for a maximum of 3 epochs.

4.2 Helsinki-NLP Opus-MT

OPUS-MT (Tiedemann and Thottingal, 2020) supports both bilingual and multilingual models. It is a project that focuses on the development of free resources and tools for machine translation. The current status is a repository of over 1,000 pretrained neural MT models. We fine-tune a transformer-align model that was fine-tuned for the Tatoeba-Challenge[10]. ***Helsinki-NLP/opus-mt-en-ga*** model from the HuggingFace Transformers (Wolf et al., 2020) for English→ Irish and ***Helsinki-NLP/opus-mt-ga-en*** for Irish→ English were used.

Language pair	Method	BLEU
English→Marathi	IndicTrans baseline	14.0
English→Marathi	IndicTrans TRA	17.8
English→Marathi	IndicTrans CVIT-PIB	23.4
English→Marathi	IndicTrans bible-uedin	27.7
English→Irish	Opus MT	30.4
English→Irish	M2M100	25.6
Irish→English	Opus MT	37.2
Irish→English	M2M100	30.4

Table 2: BLEU scores obtained for the various models for the development set

5 Results and Analysis

For Marathi, it is distinctly visible that our system model, i.e., IndicTrans fine-tuned on the training data provided by the organizers or TRA and the bible-uedin dataset, gave the best BLEU

[10] https://github.com/Helsinki-NLP/Tatoeba-Challenge

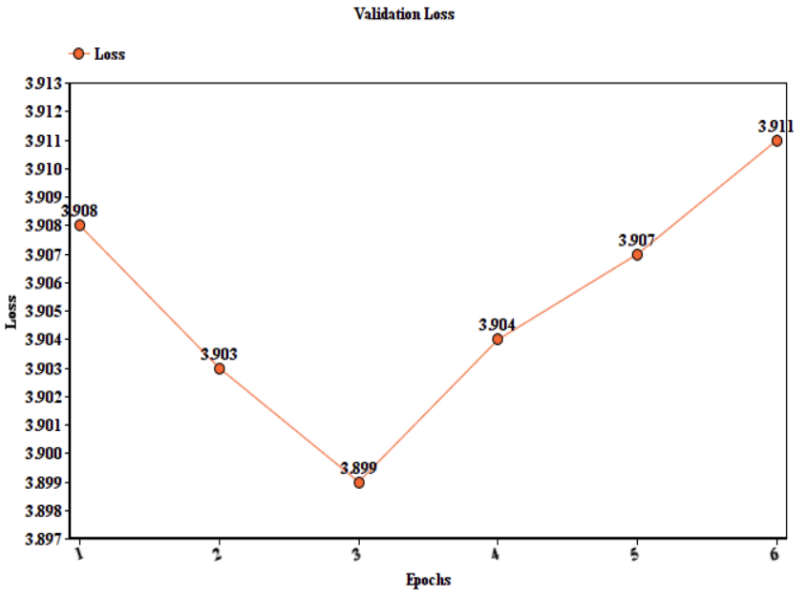

Figure 2: The graph depicting the increase in val lossafter the third epoch

scores. It was surprising how the model fine-tuned on a parallel corpus of 60,495 sentences of bible-uedin surpassed the model fine-tuned on 1,14,220 sentences from the CVIT PIB dataset. The possible explanation is the higher correlation between the sentences of the bible-uedin dataset with the test dataset than the CVIT PIB dataset. Another reason could be the presence of excessive noise in the CVIT PIB dataset. The other reason for this could be noise and a lower quality of translations in the CVIT PIB dataset compared to bible-uedin.

To infer the same, 1000 random pairs of sentences were picked from the datasets, and the average LaBSE or language-agnostic BERT Sentence Embedding (Feng et al., 2020) scores were found out. LaBSE gives a score between 0 and 1, depending on the quality of the translation. It was seen that the average was 0.768 for the bible-uedin dataset, while it was 0.58 for the CVIT PIB dataset. This might have been one of the reasons for the better BLEU scores. The model also showed constant overfitting after the second and third epoch, as the BLEU scores reduced considerably as they reached the 6th epoch. The BLEU scores decreased by a difference of 6.The validation loss starts increasing after the third epoch, thus, showing the overfitting occurring in training. So, the model was fine-tuned for three epochs while maintaining a low learning rate of around 3e-5 to get a BLEU score of 24.2.

Language pair	BLEU	CHRF	TER	Rank
English→Marathi	24.2	0.59	0.597	1
Irish→English	34.6	0.61	0.711	1
English→Irish	25.8	0.53	0.629	2

Table 3: Result and ranks obtained for the test dataset(Popović, 2015; Snover et al., 2006)

Training a model to predict for a low-resourced language was highly challenging due to the absence of prominent pretrained models (Kalyan et al., 2021; Yasaswini et al., 2021; Hande

et al., 2021b). However, as an experiment, two models from HuggingFace Transformers[11], M2M100 (Fan et al., 2020) and Opus-MT from Helsinki NLP (Tiedemann, 2020) were compared. For the dev data, Opus MT produced a BLEU score of 30.4 while M2M100 gave 25.62 for translations from English to Irish and 37.2 and 30.37 respectively for Irish to English translations. Probably, the individual models pretrained on numerous datasets gave Opus MT an edge over M2M100. This led us to submit the Opus MT model for the LoResMT Shared task 2021. The model gave exceptional BLEU scores of 25.8 for English to Irish, which ranked second in the shared task, while 34.6 for Irish to English stood first.

6 Conclusion

It is arduous and unyielding to get accurate translations for low-resourced languages due to limited datasets and pretrained models. However, our paper puts forward a few methods to better the already existing accuracies. Ranked 1, 1, and 2 in English→Marathi, Irish→English, and English→Irish respectively in the LoResMT 2021 shared task, IndicTrans fine-tuned on the bible-uedin, and the dataset provided by the organizers manages to surpass the other models due to its high correlation with the test set and minimal noise for the Marathi language. The Irish language task was dominated by the Opus MT model by Helsinki-NLP, outperforming other Transformer models, M2M100.

References

Achchuthan, Y. and Sarveswaran, K. (2015). Language localisation of Tamil using statistical machine translation. In *2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 125–129. IEEE.

Adi Narayana Reddy, K., Shyam Chandra Prasad, G., Rajashekar Reddy, A., Naveen Kumar, L., and Kannaiah (2021). English-marathi neural machine translation using local attention. In Garg, D., Wong, K., Sarangapani, J., and Gupta, S. K., editors, *Advanced Computing*, pages 280–287, Singapore. Springer Singapore.

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Chakravarthi, B. R. (2020). *Leveraging orthographic information to improve machine translation of under-resourced languages*. PhD thesis, NUI Galway.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2018). Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86.

[11] https://huggingface.co/transformers/

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019a). Comparison of different orthographies for machine translation of under-resourced dravidian languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019b). Wordnet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7.

Chakravarthi, B. R., Priyadharshini, R., Stearns, B., Jayapal, A. K., Sridevy, S., Arcan, M., Zarrouk, M., and McCrae, J. P. (2019c). Multilingual multimodal machine translation for dravidian languages utilizing phonetic transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63.

Chakravarthi, B. R., Rani, P., Arcan, M., and McCrae, J. P. (2021). A survey of orthographic information in machine translation. *SN Computer Science*, 2(4):1–19.

Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., and Liu, Y. (2016). Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.

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

Christodoulopoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49:375 – 395.

Dhonnchadha, E., Judge, J., Chasaide, A., Dhubhda, R., and Scannell, K. (2012). The irish language in the digital age: An ghaeilge sa ré dhigiteach.

Dowling, M., Castilho, S., Moorkens, J., Lynn, T., and Way, A. (2020a). A human evaluation of english-irish statistical and neural machine translation.

Dowling, M., Castilho, S., Moorkens, J., Lynn, T., and Way, A. (2020b). A human evaluation of English-Irish statistical and neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 431–440, Lisboa, Portugal. European Association for Machine Translation.

Dowling, M., Judge, J., Lynn, T., and Graham, Y. (2016). English to irish machine translation with automatic post-editing.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2020). Beyond english-centric multilingual machine translation.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding.

Garje, G. (2014). Marathi to english machine translation for simple sentences. *International Journal of Science and Research (IJSR)*, 3.

Ha, T.-L., Niehues, J., and Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. *ArXiv*, abs/1611.04798.

Hande, A., Puranik, K., Priyadharshini, R., and Chakravarthi, B. R. (2021a). Domain identification of scientific articles using transfer learning and ensembles. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25*, pages 88–97. Springer International Publishing.

Hande, A., Puranik, K., Priyadharshini, R., Thavareesan, S., and Chakravarthi, B. R. (2021b). Evaluating pretrained transformer-based models for covid-19 fake news detection. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 766–772.

Imankulova, A., Dabre, R., Fujita, A., and Imamura, K. (2019). Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 128–139, Dublin, Ireland. European Association for Machine Translation.

Jadhav, S. (2020a). Marathi to english neural machine translation with near perfect corpus and transformers. *ArXiv*, abs/2002.11643.

Jadhav, S. A. (2020b). Marathi to english neural machine translation with near perfect corpus and transformers.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Kalyan, P., Reddy, D., Hande, A., Priyadharshini, R., Sakuntharaj, R., and Chakravarthi, B. R. (2021). IIITT at CASE 2021 task 1: Leveraging pretrained language models for multilingual protest detection. In *Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 98–104, Online. Association for Computational Linguistics.

Krishnamurthy, P. (2015). Development of Telugu-Tamil transfer-based machine translation system: With special reference to divergence index. In *Proceedings of the 1st Deep Machine Translation Workshop*, pages 48–54, Praha, Czechia. ÚFAL MFF UK.

Krishnamurthy, P. (2019). Development of Telugu-Tamil transfer-based machine translation system: An improvization using divergence index. *Journal of Intelligent Systems*, 28(3):493–504.

Kumar, A., Mundotiya, R. K., and Singh, A. K. (2020a). Unsupervised approach for zero-shot experiments: Bhojpuri–Hindi and Magahi–Hindi@LoResMT 2020. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 43–46, Suzhou, China. Association for Computational Linguistics.

Kumar, B. S., Thenmozhi, D., and Kayalvizhi, S. (2020b). Tamil paraphrase detection using encoder-decoder neural networks. In *International Conference on Computational Intelligence in Data Science*, pages 30–42. Springer.

Lynn, T., Scannell, K., and Maguire, E. (2015). Minority language twitter: Part-of-speech tagging and analysis of irish tweets.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al. (2017). Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Ojha, A. K., Liu, C.-H., Kann, K., Ortega, J., Satam, S., and Fransen, T. (2021). Findings of the LoResMT 2021 Shared Task on COVID and Sign Language for Low-Resource Languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation.

Parameswari, K., Sreenivasulu, N., Uma Maheshwar Rao, G., and Christopher, M. (2012). Development of Telugu-Tamil bidirectional machine translation system: A special focus on case divergence. In *proceedings of 11th International Tamil Internet conference*, pages 180–191.

Pathak, A. and Pakray, D. P. (2018). Neural machine translation for indian languages. *Journal of Intelligent Systems*, 28.

Philip, J., Siripragada, S., Namboodiri, V. P., and Jawahar, C. V. (2020). Revisiting low resource status of indian languages in machine translation. *8th ACM IKDD CODS and 26th COMAD*.

Poncelas, A., Shterionov, D., Way, A., Wenniger, G. M. D. B., and Passban, P. (2018). Investigating backtranslation in neural machine translation. *ArXiv*, abs/1804.06189.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Puranik, K., Hande, A., Priyadharshini, R., Thavareesan, S., and Chakravarthi, B. R. (2021). IIITT@LT-EDI-EACL2021-hope speech detection: There is always hope in transformers. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 98–106, Kyiv. Association for Computational Linguistics.

Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., J, M., Kakwani, D., Kumar, N., Pradeep, A., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., and Khapra, M. S. (2021). Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Rehm, G. and Uszkoreit, H. (2012). The irish language in the digital age.

Revanuru, K., Turlapaty, K., and Rao, S. (2017). Neural machine translation of indian languages. In *Proceedings of the 10th Annual ACM India Compute Conference*, Compute '17, page 11–20, New York, NY, USA. Association for Computing Machinery.

Sarveswaran, K., Dias, G., and Butt, M. (2021). Thamizhimorph: A morphological parser for the Tamil language. *Machine Translation*, 35(1):37–70.

Scannell, K. P. (2006). Machine translation for closely related language pairs. In *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, pages 103–109. Citeseer.

Scannell, K. P. (2007). The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks.

Thenmozhi, D., Kumar, B. S., and Aravindan, C. (2018). Deep learning approach to English-Tamil and Hindi-Tamil verb phrase translations. In *FIRE (Working Notes)*, pages 323–331.

Tiedemann, J. (2020). The tatoeba translation challenge – realistic data sets for low resource and multilingual mt.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Torregrosa, D., Pasricha, N., Masoud, M., Chakravarthi, B. R., Alonso, J., Casas, N., and Arcan, M. (2019). Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 125–133, Dublin, Ireland. European Association for Machine Translation.

U Hegde, S., Hande, A., Priyadharshini, R., Thavareesan, S., and Chakravarthi, B. R. (2021). UVCE-IIITT@DravidianLangTech-EACL2021: Tamil troll meme classification: You need to pay more attention. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 180–186, Kyiv. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Vázquez, R., Raganato, A., Tiedemann, J., and Creutz, M. (2019). Multilingual NMT with a language-independent attention bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (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.

Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic. Association for Computational Linguistics.

Yasaswini, K., Puranik, K., Hande, A., Priyadharshini, R., Thavareesan, S., and Chakravarthi, B. R. (2021). IIITT@DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194, Kyiv. Association for Computational Linguistics.

Zhang, Z. (2018). Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pages 1–2. IEEE.

Zoph, B. and Knight, K. (2016). Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.