Machine Translation Advancements for Low-Resource Indian Languages in WMT23: CFILT-IITB's effort for bridging the Gap

Meet Doshi, Pranav Gaikwad, Sourabh Deoghare, Pushpak Bhattacharyya
Computation for Indian Language Technology Lab
Indian Institute of Technology, Bombay.
{meetdoshi, pranavgaikwad, sourabhdeoghare, pb}@cse.iitb.ac.in

Abstract

Machine translation for low-resource Indian languages has long been a challenge due to the scarcity of high-quality parallel corpora, demanding the development of effective translation models. The WMT23 Low-Resource Indic Language Translation task encourages us to utilize creative techniques to address this issue and enhance the performance of machine translation systems for these languages. We focused on the translation of two low-resource Indic languages: Assamese and Manipuri, enabling bidirectional translation between English and these languages. This paper presents CFILT-IITB's submission to WMT23, highlighting our exploration of transfer learning-based methodologies. Our experiments produced notable results of **47.54 BLEU** on MNI→EN, **18.15 BLEU** on EN→ASM and **35.24 BLEU** on ASM→EN, **26.36 BLEU** on EN→MNI test sets. These results not only demonstrate the effectiveness of transfer learning-based techniques but also contribute to advancing machine translation capabilities for low-resource Indian languages, addressing a critical need in bridging language barriers and facilitating cross-cultural communication.

1 Introduction

In the realm of machine translation, the WMT23 IndicMT shared task emerges as an arena where the boundaries of translation technology are stretched to their limits. Our efforts revolve around the translation of the 'En-X' pair in both directions, where 'En' signifies English and 'X' encompasses Assamese, a member of the Indo-Aryan language family, and Manipuri, a representative of the Tibeto-Burman family. As the task focused on English to and from low-resource Indian languages, we were provided with a small parallel corpus for each 'En-X' pair. Furthermore, participants had access to a substantial amount of monolingual data for Assamese and Manipuri, creating an ideal setting for trying out new and creative approaches.

In the realm of Machine Translation, the Neural Machine Translation paradigm has emerged as a dominant force, as evidenced by seminal works such as (Bahdanau et al., 2014) and the comprehensive survey by (Dabre et al., 2020). However, Neural Machine Translation models are notoriously data-hungry, leading to performance degradation when confronted with low-resource languages, as highlighted by (Dewangan et al., 2021). To tackle this challenge, we turn to the promising technique of transfer learning, a well-established approach in machine learning where knowledge gained from one task is leveraged to enhance performance in a related task. In our pursuit of improving translation capabilities for low-resource languages, we harness the multilingual IndicTrans2 model, as introduced by (AI4Bharat et al., 2023). Our methodology involves fine-tuning this model using the 'En-X' parallel data provided for the task. By adopting this approach, we aim to capitalize on the acquired knowledge during training to significantly bolster the performance of the model in the specific translation task at hand.

IndicTrans2 is rooted in the transformer-based encoder-decoder architecture pioneered by (Vaswani et al., 2017). It was trained on the extensive Bharat Parallel Corpus Collection (BPCC), a publicly accessible repository encompassing both pre-existing and freshly curated data for all 22 scheduled Indian languages, this model boasts a comprehensive understanding of the linguistic diversity within the Indian subcontinent. To enhance its linguistic prowess, IndicTrans2 has undergone auxiliary training utilizing the rich resource of back-translated monolingual data. The model was then trained on human-annotated data to achieve further improvements. We used this model and fine-tuned it on the training data provided by WMT23.

The fine-tuned IndicTrans2 achieves good scores; hence we are using it for our final submission. We hypothesize that its stellar performance

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 950–953
December 6–7, 2023.

can be attributed to the amalgamation of language knowledge acquired during its initial training, coupled with the domain-specific expertise gleaned from the fine-tuning process, facilitated by the training data made available through WMT23.

2 Data

We use the IndicTrans2 model and fine-tune it on the WMT23. The original IndicTrans2 was trained on the Bharat Parallel Corpus Collection (BPCC) corpus. They have used FLORES-200 as their validation set for Assamese and extended FLORES-200 (Team et al., 2022) for Manipuri. For auxiliary training which includes back-translated monolingual sentences, they have used IndicCorp v2 (Kakwani et al., 2020) and one side of NLLB data as monolingual corpus. They have used standard test sets like FLORES-200, but they have also created a new benchmark called the IN22 test set which is an n-way parallel corpus for all 22 Indian scheduled languages.

We have fine-tuned the model using the WMT23 parallel corpus. The 'English-Assamese' pair has 50K parallel sentences, and the 'English-Manipuri' pair has around 21.6K sentences. The validation set consisted of the WMT23 validation set. The size of the validation set for the 'English-Assamese' is 2K sentences; for the 'English-Manipuri' pair, it was 1k sentences. The test set for both pairs was the WMT23 test set.

3 System Overview

In the pursuit of enhancing machine translation for low-resource languages, various approaches have emerged, such as Phrase-Pair injection and Back-translation, aimed at enhancing performance. Our system, on the other hand, takes a distinct path and relies on the knowledge gained from the multilingual training of IndicTrans2 and applies it to different low-resource languages.

Phrase-Pair Injection (PTI) (Sen et al., 2021), (Dewangan et al., 2021) and (Banerjee et al., 2021) utilized a technique to combine both Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). The utilization of the phrase table during training is pivotal in Statistical Machine Translation (SMT) as it probabilistically maps phrases from the source to the target language. By incorporating these phrase mappings from the table into the existing parallel corpora, the training dataset for the Neural Machine Translation (NMT) model is significantly enriched. Consequently, this enrichment empowers the NMT model to excel in its translation performance.

Back-translation Back-translation (Sennrich et al., 2016; Conneau et al., 2020) is a technique that is used to improve the performance of low-resource translation systems using monolingual data. In this technique, a reverse model is employed to generate a parallel corpus from a monolingual corpus. This is a clever way to use the monolingual corpus to improve the translation performance of the NMT models. We do not include Back-translated sentences for training since we could not see any significant performance improvement.

Transfer Learning Transfer learning is a machine learning technique where a model trained on one task is adapted for a second related task. Instead of starting the training of a new model from scratch, transfer learning leverages the knowledge learned from the first task to improve learning on the second task. We have used IndicTrans2 (AI4Bharat et al., 2023), a powerful model that performs well for English-to-Indic and Indic-to-English translation for 22 scheduled Indian languages. This knowledge can be used to translate other Indian languages to and from English. Our approach entailed the fine-tuning of this model, leveraging the parallel corpus provided by the WMT23 for the IndicMT task. This fine-tuning process equipped the model with the expertise required to proficiently translate Assamese and Manipuri to and from English, ultimately yielding the most outstanding results. We do not inject phrase pairs since for such a low resource setting, it is difficult to see performance improvements even with phrase pair injections due to the inability of NMT models to capture the low resource language.

4 Experiments

4.1 Settings

All the experiments are conducted using two NVIDIA A100 GPUs each having 80GB of memory. Our models apply Adam (Kingma and Ba, 2015) as optimizer to update the parameters with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We employ a warm-up learning rate of 10^{-7} for 2000 update steps and a learning rate of $3 * 10^{-5}$. For normalization, we use a dropout value of 0.2 and normalize the proba-

Models	ASM→EN		EN→ASM		MNI→EN		EN→MNI	
	BLEU	ChrF2	BLEU	ChrF2	BLEU	ChrF2	BLEU	ChrF2
Baseline-1 (val)	2.32	-	1.64	-	3.12	-	2.67	-
IndicTrans2 (val)	25.60	47.20	14.70	41.40	33.40	58.50	11.90	43.50
FT IndicTrans2 (val)	34.60	52.40	**24.00**	46.00	47.00	67.30	**34.10**	62.20
FT IndicTrans2 (test)	**35.24**	**57.73**	18.15	**50.16**	**47.54**	**70.41**	26.36	**63.48**

Table 1: Comparison of results of Fined-tuned IndicTrans2 (AI4Bharat et al., 2023) on the test and val set. We compare val and test set results because we see that the EN-Indic model has overfitted for both languages and therefore we see a decrease in BLEU for EN-Indic models. We recommend readers to decrease the number of updates for better scores when the source is English.

bilities using smoothed label cross-entropy. We use GeLU activations (Hendrycks and Gimpel, 2016) for better learning. We train separate models for each language pair to avoid data imbalance and learn better low-resource representations.

We use the scareBLEU library[1] to calculate our BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) scores with a word order of 2. We choose the checkpoint with the highest validation BLEU score.

4.2 Results

Table 1 shows that the highest translation quality achieved is via the use of large monolingual and parallel corpora. Since IndicTrans2 is trained in many Indian languages, it enhances the translation quality via the power of multilingualism. With only some minor tuning of the model over the training and validation set, IndicTrans2 achieves remarkable performance on Indic-En translations. Our baseline-1 system is a WMT-14 En-De fairseq model trained that utilizes only the parallel data and shows substandard BLEU scores over all the language pairs. With our experiments, we see that with even the monolingual corpora and back translation, the translation models only see minor improvements. We realized the power of multilingualism and switched to pre-trained models which have been trained on a substantial amount of data like IndicTrans2 (AI4Bharat et al., 2023) and NLLB (Team et al., 2022). We analyze their vocabulary and merge it with a new vocabulary learned over the monolingual corpora provided in the task. Even for languages that are not seen by the model like Mizo and Khasi in the *latin* script, the IndicTrans2 model with its pre-trained English vocabulary gives a BLEU score of an average of 7.2 on the val set over these language pairs. We see that when we fine-tune the pre-trained model, we see large gains over both the *val* and the *test* set. Finally, after many experiments, we submit a fine-tuned version of a very powerful multilingual model for the shared task.

[1] https://github.com/mjpost/sacrebleu/blob/master/sacrebleu/metrics/bleu.py

5 Conclusion

In this paper, we present how CFILT-IITB utilized the power of multilingual models for the WMT23 IndicMT Low-Resource Machine Translation of Indian Languages Shared Task. Since, the data for low-resource languages is scarce, utilizing pre-trained multilingual translation models is very crucial. But to have reasonable to good performance over these models, it is helpful to have a model that is trained on similar languages. For this task, Indian languages like *Assamese* and *Manipuri* share similar structure and vocabulary with many Indian languages like *Bengali* which can be considered a high resource language for India. Training models over similar language does boost performance although to cover a wide variety of low-resource languages, one must face the curse of multilingualism. Our most proficient system attains an average BLEU score of 41.39 for Indic-English translation and 22.25 for English-Indic language pairs, specifically Assamese and Manipuri.

Limitations

Limitations of such powerful multilingual models are data extraction, enormous computing, and good data filtration techniques. Overcoming these obstacles is an open research problem.

References

AI4Bharat, Jay Gala, Pranjal A. Chitale, Raghavan AK, 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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Aakash Banerjee, Aditya Jain, Shivam Mhaskar, Sourabh Dattatray Deoghare, Aman Sehgal, and Pushpak Bhattacharyya. 2021. Neural machine translation in low-resource setting: a case study in english-marathi pair. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 35–47.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Shubham Dewangan, Shreya Alva, Nitish Joshi, and Pushpak Bhattacharyya. 2021. Experience of neural machine translation between indian languages. *Machine Translation*, 35(1):71–99.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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 of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 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.

Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. 2021. Neural machine translation of low-resource languages using smt phrase pair injection. *Natural Language Engineering*, 27(3):271–292.

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

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.