SRIB-NMT's Submission to the Indic MT Shared Task in WMT 2024

Pranamya Patil , **Raghavendra HR** , **Aditya Raghuwanshi** and **Kushal Verma**
Samsung Research India Bangalore, Bangalore, India
{pran.patil,raghav.hr,aditya.r1,kushal.verma}@samsung.com

Abstract

In the context of the Indic Low Resource Machine Translation (MT) challenge at WMT-24 ((Pakray et al., 2024)), we participated in four language pairs: English-Assamese (en-as), English-Mizo (en-mz), English-Khasi (en-kh), and English-Manipuri (en-mn). To address these tasks, we employed a transformer-based sequence-to-sequence architecture (Vaswani et al., 2017). In the PRIMARY system, which did not utilize external data, we first pretrained language models (low resource languages) using available monolingual data before finetuning them on small parallel datasets for translation. For the CONTRASTIVE submission approach, we utilized pretrained translation models like Indic Trans2 (Gala et al., 2023) and applied LoRA Fine-tuning (Hu et al., 2021) to adapt them to smaller, low-resource languages, aiming to leverage cross-lingual language transfer capabilities (CONNEAU and Lample, 2019). These approaches resulted in significant improvements in SacreBLEU scores(Post, 2018) for low-resource languages.

1 Introduction

With increasing digital connectivity there is huge demand for good translations systems for people to access wide array of digital information in their native local languages. This gives people flexibility and ease of access. For any machine learning task, the quality and quantity of data is of paramount importance. There are multiple languages which have either negligible or zero digital footprint. On top of that presence of good quality parallel/ bitext data is even more rare event.

Due to increased demand and intangible benefits from translation systems there has been lot of research in the field of machine translation. One such area is the low resource machine translation system. In the above statement "resource" refers to data resource. We are working on handling translations for languages which have very less data available (both monolingual and parallel).

Some major works for multilingual Machine translation (approx 200*200 languages) is NLLB (Costa-jussà et al., 2022). Here the authors use Mixture of Experts (MoE) to train single large multilingual model capable of handling approx 200+ languages as source(src) and target(tgt) languages. One of the objectives behind NLLB is handling low resource languages.

The details of WMT23 Indic MT findings can be found here (Pal et al., 2023). For our PRIMARY approach (no additional data apart from what was shared), we pretrained language model (using monolingual data). We explored 2 pretraining objectives namely Causal Language Modeling (Radford et al., 2019) and denoising (Lewis et al., 2020). Using these Pretrained Language Models as initial model weights we finetuned for tranlstaion task using available parallel corpus.

For the CONTRASTIVE submission approach, we utilized pretrained translation models like Indic Trans2 (Gala et al., 2023) and applied LoRA Fine-tuning (Hu et al., 2021) to adapt them to smaller, low-resource languages, aiming to leverage cross-lingual language transfer capabilities (CONNEAU and Lample, 2019).

2 Related Work

Our submissions use the concepts like transfer learning, denoising pretraining (Lewis et al., 2020), Causal Language Modeling (Radford et al., 2019). We use denoising pretraining and causal language modeling as pretraining tasks. Using transfer learning we use the pretrained model for new task like translation. We use pretrained indic translation model like Indic Trans2 (Gala et al., 2023) for contrastive submission and adapted them to low resource languages using LoRA Fine-tuning (Hu et al., 2021). We aim to take advantage of cross-lingual language transfer capabilities (CONNEAU and Lample, 2019) and hence used IndicTrans2

Proceedings of the Ninth Conference on Machine Translation, pages 747–750
November 15-16, 2024 ©2024 Association for Computational Linguistics

Monolingual Data	# lines
Assamese	2,624,715
Manipuri	2,144,897
Mizo	1,909,823
Khasi	182,737
Parallel Train Data	**# lines**
English <-> Assamese	50,000
English <-> Manipuri	21,687
English <-> Mizo	50,000
English <-> Khasi	24,000

Table 1: Dataset Sizes shared as part of Indic MT Task

machine translation model.

3 System Description

We have submitted in 2 categories i) PRIMARY ii) CONTRASTIVE

3.1 PRIMARY System

For PRIMARY System we trained the model from scratch using only the shared data as part of workshop. We trained separate model for each direction.

3.1.1 Tokenizer

First we need to train tokenizer for each english <-> language pair (ie:- one of Assamese, Manipuri, Mizo, Khasi). We use sentence piece[1] tokenizer library to train joint dictionary (combined vocab) for each of english <-> (Assamese, Manipuri, Mizo, Khasi) language pairs. We use mono data(each from as,mn,kh,mz languages) + almost equal amount of english data to train sentence piece tokenizer (subset choosen from AI4Bharat Samanantaar dataset (Ramesh et al., 2022)).

3.1.2 Pretraining

We Pretrain the model using monolingual data. We explored 2 subtasks for the same

1. Causal Language Modeling (Radford et al., 2019) - Here we train decoder only model for causal language modeling ie:- predicting the next word using context words. The pretrained decoder from this task is utilized to initialize the decoder component of our seq2seq architecture employed for the translation task.

2. Denoising Task (Lewis et al., 2020) - We integrate a seq2seq transformer model that takes a noisy version of the input (such as perturbed mono data, added tokens, or shuffled data) and expects the output from the decoder to be the original, unperturbed input. By utilizing this denoising objective task, we aim for the model to understand language patterns and structures. The pretrained model from this task can be used for translation task.

Both of the above pretraining tasks are explored independently and we used each tasks pretrained checkpoints for finetuning separately.

3.1.3 Finetuning

Using the pretrained checkpoint we finetune the models for translation task (with small amount of parallel data). Pretraining helps the model to understand the language nuances and leads to faster converging of models for translation tasks.

3.2 CONTRASTIVE System

For CONTRASTIVE Submission (where external data etc is allowed). We use translation model of other languages eg:- IndicTrans2 (Gala et al., 2023). As the 4 low resource languages ie:- Mizo, Manipuri, Khasi and Assamese are near to Indic Languages supported by (Gala et al., 2023) we believe the the model will benefit from shared parameters, vocabs and hence map Cross Lingual language references (CONNEAU and Lample, 2019). We use the same tokenizer as used by IndicTrans2 Model.

LoRA (Hu et al., 2021) adaptation is a lightweight and resource-friendly technique for customizing pretrained models. It involves adding small adapter weights (to certain layers) alongside the existing model weights. During training, the original model weights remain unchanged while only the adapter weights are updated. At test time, the adapter weights and the original model weights are combined to generate predictions. This approach allows for efficient customization without requiring extensive modifications to the original model. Since only a small number of parameters are updated during training, the overall training time is reduced. Additionally, this approach helps mitigate the issue of catastrophic forgetting to some degree.

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

Parameter	Value
encoder_layers	4
decoder_layers	4
attention_heads	8
embedding_dimension	512
ffn_embedding_dimension	4096

Table 2: PRIMARY Submission Model Architecture details

Parameter	Value
lora_rank	32
lora_alpha	32
lora_dropout	0.1
device_batch_size	16
device_grad_accumulation_steps	2
max_steps	100,000
eval_steps	5,000
patience	10

Table 3: CONTRASTIVE Submission Model details

4 Experiments

4.1 Implementation

4.1.1 PRIMARY Submission

For Primary submission we use fairseq[2] framework for both pretraining and finetuning stage. The model architecture details can be found in Table 2. We experimented with lesser #encoder, #decoder layers as compared to standard (6 encoder and 6 decoder layers) to reduce model complexity and hence training time.

4.1.2 CONTRASTIVE Submission

For CONTRASTIVE submission we use IndicTrans2(Gala et al., 2023) and use huggingface peft library for LoRA finetuning. The model details can be found in Table 3

5 Results

The results as shared by conference committee are attached below. PRIMARY submission results Table 4 and CONTRASTIVE submission results in Table 5.

We use SacreBLEU (Post, 2018) for validation evaluation. The validation scores (development set) reported during training for Primary system are attached in Table 6

Direction	BLEU	TER	RIBES	METEOR	ChrF
en -> as	1.32	101.83	7.1	7.44	22.15
en -> mn	0	101.83	1.91	3.07	18.89
en -> mz	0	102.98	3.61	6.20	16.46
en -> kh	0.54	103.72	8.21	9.69	17.78

Table 4: PRIMARY Submission Scores on test suite

Direction	BLEU	TER	RIBES	METEOR	ChrF
as -> en	29.59	34.92	35.05	74.09	64.88
mn -> en	18.89	53.05	29.17	59.43	57.1
mz -> en	11.27	64.94	20.26	47.84	44.82
kh -> en	4.2	80.89	12.05	32.83	31.8

Table 5: CONTRASTIVE Submission Scores on test suite

5.1 Learnings

Following are the learnings from our Experiments

1. Transfer Learning benefits translation task. We saw it in PRIMARY submissions in which language models are pretrained on denoising/ causal language modeling(CLM) task and then transferred for translation task. Especially its evident from initial bleu score and loss. We saw denoising task led to faster converging of models (lower initial loss) relative to CLM task objective.

2. Languages that share a common linguistic ancestor or follow similar word order patterns (such as SVO or SOV) can benefit from using the same vocabulary and sharing parameters during initialization. This allows for more efficient training and better performance across related languages.

3. Using Translation for related language benefits from cross lingual language reference.(Bleu scores of CONTRASTIVE submission)

4. LoRA finetuning is effective for adapting a translation model to new low resource language (lesser training time and resources).

Direction	SacreBLEU
en -> as	8
en -> mn	16.9
en -> mz	22.3
en -> kh	11.1

Table 6: PRIMARY Submission Scores on development suite shared along with training data

[2] https://github.com/facebookresearch/fairseq

5.2 Conclusion

The adpatation of another language translation model to similar but low resource language is benefitted by sharing params, vocabs etc across languages (due to cross lingual language learning). LoRA finetuning leds to quicker converging for low resource languages (18-19 hours on A100 GPU with 40GB of RAM).

We have described our submission to WMT2024 Indic Translation Task, leveraging various concepts like Denoising task(Lewis et al., 2020), Cross Lingual Transfer Learning(CONNEAU and Lample, 2019), IndicTrans2 Model(Gala et al., 2023), LoRA adaptation(Hu et al., 2021) etc.

Limitations

1. Exploring impact of Iterative Backtranslation(Hoang et al., 2018) benefits using intermediate models in PRIMARY setting.

2. Exploring more pretraining task objectives for PRIMARY System.

3. Exploring multi task learning impact for PRIMARY systems.

4. Exploring the difference in scores, training resources for full precision finetuning vs LoRA finetuning.

References

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Mohana Krishna, Arnab Kumar Maji, Sandeep Kumar Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of wmt 2024 shared task on low-resource indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

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.