Empowering Low-Resource Language Translation: Methodologies for Bhojpuri-Hindi and Marathi-Hindi ASR and MT

Harpreet Singh Anand and **Amulya Ratna Dash** and **Yashvardhan Sharma**
Dept. of Computer Science and Information Systems
Birla Institute of Technology and Science, Pilani, India

Abstract

This paper presents the methodologies implemented for the Automatic Speech Recognition and Machine Translation for the language pairs Bhojpuri-Hindi and Marathi-Hindi for the Dialectal and Low-Resource shared task proposed by The International Conference on Spoken Language Translation (IWSLT) for 2024. The implemented method uses the transcriptions generated through a fine-tuned Whisper models(for Marathi-Hindi) and vakyansh-wav2vec model (for Bhojpuri-Hindi) and generates the translations using fine-tuned NLLB(No Language Left Behind) Models for both the tasks. The selection of more accurate translation is done through sentence-embeddings generated using the MuRIL(Multilingual Representations for Indian Languages)(Khanuja et al., 2021) model for the Marathi-Hindi task.

1 Introduction

India boasts a vast linguistic variety, with more than 100 official languages and numerous dialects spoken all throughout the nation. Natural Language Generation (NLG) tasks such as automated speech recognition (ASR) and machine translation (MT), are greatly hampered by the tremendous variety. For millions of Indians who do not speak English or other commonly spoken languages, ASR and MT can be crucial in bridging the language gap and granting them access to information and services in a multi-linguistic country like India. However, the creation of ASR and MT systems is a challenging endeavour due to the inherent features of Indian languages, such as rich morphology, the occurrence of code-switching, and borrowing from other languages.

The 'Dialectal and Low-Resource Track'proposed by IWSLT 2024 requires the participants devise creative approaches to leverage the disparate resources available for 8 dialectal and low-resource languages. The participants are required to submit under two conditions - namely constrained and unconstrained. The constrained condition should contain systems that are trained only on the datasets provided by the organizers while the unconstrained condition can contain systems trained with any resource including pre-trained and multilingual models. Our team participated in the unconstrained condition for the language pairs - Marathi to Hindi and Bhojpuri to Hindi. This paper will discuss the implementation details of our ASR and MT systems for the above-mentioned language pairs.

2 Related Work

Automatic Speech Recognition(ASR) and Machine Translation(MT) in low-resource languages have been the subject of extensive research in recent years. Several approaches have been proposed to address the challenges associated with low-resource languages in ASR and MT. For instance, multilingual training has been identified as an effective approach for compensating for the limited amount of data in low-resourced ASR (Madikeri et al., 2020). Additionally, transfer learning methods have been used to develop end-to-end ASR systems for low-resource languages, demonstrating their influence in addressing the challenges of low data levels (Mamyrbayev et al., 2022). Furthermore, the use of self-supervised speech recognition models has been hindered by the requirement for considerable labeled training data, which poses a challenge for their application to low-resource languages (Hameed et al., 2022).

In the context of MT, the scarcity of parallel data for low-resource languages has been identified as a significant challenge (Gao et al., 2020). Neural Machine Translation (NMT) systems, which require large amounts of training data, face difficulties in creating high-quality systems for low-resource languages (Neubig and Hu, 2018). However, research efforts have been directed towards improv-

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 277–282
August 15-16, 2024 ©2024 Association for Computational Linguistics

ing low-resource NMT, with studies exploring techniques such as teacher-free knowledge distillation to enhance performance in low-resource languages (Zhang et al., 2020). The encoder-decoder framework for NMT has also been found to be less effective for low-resource languages, highlighting the need for specialized approaches to address the challenges of low-resource machine translation (Zoph et al., 2016).

The use of transfer learning has shown effectiveness in addressing the challenges of low-resource NMT, particularly in scenarios where parallel data is limited (Ji et al., 2019). Additionally, the development of multilingual NMT systems has contributed to improving the quality of translation, especially for low-resource language pairs, enabling zero-shot translation and allowing the translation of language pairs never seen in training (Escolano et al., 2021).

Automatic Speech Recognition (ASR) and Machine Translation (MT) for Indian languages have gained significant attention in recent years. The development of ASR systems for Indian languages has been a focus of research, with studies addressing low-resource challenges (Sailor et al., 2018), multilingual and code-switching ASR systems (Diwan, 2021), and the impact of multilingual representations on ASR and keyword search (Cui et al., 2015). Research has also been conducted on ASR for specific Indian languages such as Hindi, Marathi, Bengali, and Oriya (Dash et al., 2018). Furthermore, the potential of ASR to aid individuals with speech disabilities, such as dysarthria, has been explored (Shahamiri and Salim, 2014). In the realm of MT, efforts have been made to improve the quality of translations for Indian languages through techniques such as transliteration and part-of-speech tagging ((Durrani et al., 2014; Ameta et al., 2013). Moreover, the development of MT systems for Indian languages and their approaches have been a subject of interest (Saini and Sahula, 2015; Godase and Govilkar, 2015). Research has also delved into rule-based machine translation and inflection rules for specific Indian languages like Marathi (Kharate and Patil, 2021). The significance of ASR and MT for Indian languages is underscored by the need to break language barriers and facilitate inter-lingual communication (Godase and Govilkar, 2015). Furthermore, the development of ASR and MT systems for Indian languages is crucial for addressing the diverse linguistic landscape of the country and enabling access to information and services for non-English speakers.

In conclusion, the research on ASR and MT for Indian languages has made substantial progress, addressing challenges related to low-resource settings, multilingualism, and specific language requirements. These advancements are pivotal in enabling effective communication, accessibility, and inclusivity for Indian language speakers.

3 Datasets

We utilized the datasets provided by the track organizers as indicated below:

3.1 OpenSLR

OpenSLR(Open Speech and Language Resources) is a site devoted to hosting speech and language resources, such as training corpora for speech recognition, and software related to speech recognition. This data set[1](He et al., 2020) contains transcribed high-quality audio of Marathi sentences recorded by volunteers. The data set consists of .wav files, and a TSV file (line-index.tsv). The file line-index.tsv contains an anonymized FileID and the transcription of audio in the file. Following are some details about the dataset:

Identifier: SLR64

Summary: Dataset which contains recordings of native speakers of Marathi

Category: Speech

License: Attribution-ShareAlike 4.0 International

3.2 Common Voice

We used the Common Voice 11.0 dataset(Ardila et al., 2020) (Marathi) for the fine-tuning of Whisper. Common Voice is an open-source, multi-language dataset of voices that anyone can use to train speech-enabled applications. The dataset consists of a unique MP3 and corresponding text file. The dataset is available on the HuggingFace Datasets Hub[2] and can be directly imported from there.

3.3 Samanantar

We have used the Indic2Indic part of the Samanantar(Ramesh et al., 2022) dataset which is the largest publicly available parallel corpora collection for Indic languages.

[1] https://www.openslr.org/64/

[2] https://huggingface.co/datasets/mozilla-foundation/common_voice_11_0

4 Methodology

4.1 Marathi - Hindi

For the Marathi-Hindi track(unconstrained condition), we have utilized a cascaded approach consisting of two fine-tuned Whisper models for ASR and a fine-tuned NLLB(NLLB Team et al., 2022) model for MT.

4.1.1 ASR

In our submission, we have fine-tuned the Whisper-small(Radford et al., 2022) pre-trained checkpoint(244M parameters and Multilingual)[3] to obtain two fine-tuned models for ASR in Marathi. The first model was obtained after fine-tuning on the Common Voice 11.0 dataset while the second model was generated after fine-tuning on the OpenSLR dataset(SLR 64). We will call these models as "whisper-ft-cv" and "whisper-ft-slr" respectively. The following hyper-parameters were used during training of both the models:
Learning Rate: 1e-05
Train Batch Size: 16
Eval Batch Size: 8
Seed: 42
Optimizer: Adam with betas=(0.9,0.999)
LR scheduler type: linear
LR scheduler warmup steps: 500
Training Steps: 4000
Mixed Precision Training: Native AMP

We trained both the ASR models for 4000 steps since during experimentation, we found that there was not significant reduction in WER after 4000 steps. Table 1 below shows the results that we obtained on the evaluation dataset(in terms of WER score) after fine-tuning Whisper on the common voice dataset.

Step	Epoch	Training Loss	WER
1000	4.07	0.0658	46.3542
2000	8.13	0.004	44.7295
3000	12.2	0.0004	43.5046
4000	16.26	0.0002	43.3628

Table 1: Training results for whisper-ft-cv

Table 2 below shows the results that we obtained on the evaluation dataset(in terms of WER score) after fine-tuning Whisper on the OpenSLR dataset.

Step	Epoch	Training Loss	WER
1000	12.66	0.0018	16.6181
2000	25.32	0.0005	14.6303
3000	37.97	0.0002	14.4977
4000	50.63	0.0001	14.33917

Table 2: Training results for whisper-ft-slr

4.1.2 MT

For the Machine Translation of the transcriptions generated by the ASR model, we are using a fine-tuned NLLB model (600M-distilled)[4] trained on the Samanantar (Indic2Indic) dataset in the Marathi-Hindi direction.The fine-tuned model is then used to translate transcriptions obtained through both whisper-ft-cv and whisper-ft-slr. Following were the training arguments that were used to fine-tune both the NLLB models:
Learning Rate : 2e-5
Batch Size : 16
Weight Decay : 0.01
Epochs : 5
The model was fine-tuned for 5 epochs only since during experimentation we found out that the training loss and BLEU scores plateaued after 5 epochs.

4.1.3 Choice of Translation

The sentence embeddings of both the transcriptions and their respective translations are generated using MuRIL(Khanuja et al., 2021). The cosine-similarity of these translations with their respective transcriptions are then compared and the pair with higher value of cosine similarity is chosen as the more accurate transcription and translation.

4.2 Bhojpuri - Hindi

For the Bhojpuri-Hindi track(unconstrained condition), we have utilized cascaded approach consisting of a pre-trained wav2vec model(for ASR) and a fine-tuned NLLB model(for MT).

4.2.1 ASR

In our submission, we have used vakyansh-wav2vec2-bhojpuri-bhom-60[5](Chadha et al., 2022; Gupta et al., 2021) model for generating the transcriptions in Bhojpuri. It is a pre-trained wav2vec model available on HuggingFace.

[3] https://huggingface.co/openai/whisper-small

[4] https://huggingface.co/facebook/nllb-200-distilled-600M

[5] https://huggingface.co/Harveenchadha/vakyansh-wav2vec2-bhojpuri-bhom-60

4.2.2 MT

We translate the transcriptions generated in the previous step using NLLB (1.3B)[6] model with Bhojpuri as the source language and Hindi as the target language.

5 Results

The results of ASR have been calculated using WER and CER scores while those of MT are calculated using BLEU(Papineni et al., 2002) and chrF2(Popović, 2015; Zoph et al., 2016) metrics respectively.
Word Error Rate (WER) and Character Error Rate (CER) indicate the amount of text that was misread by the model. WER recognizes three different types of mistakes: substitutions, deletions, and insertions. It is possible to see mispredicted terms from word-level mistakes which can illustrate frequent word-level errors made by a model.Its formal definition is percentage of word-level mistakes in candidate text.Another statistic that measures correctness of a candidate text with regards to substitutions, deletions and insertions is Character Error Rate. Word-level errors focus on mispronounced words or wrong phonemes while character level mistakes help point out such mispronunciations.The number of character level mistake present in a candidate text is called CER. BLEU score measures the quality of predicted text, referred to as the candidate, compared to a set of references. BLEU score is a precision based measure and it ranges from 0 to 1. The closer the value is to 1, the better the prediction.

5.1 Marathi-Hindi

Table 3 below shows the results of our ASR System in which we are using fine-tuned models of Whisper-small. Here, "contrastive1" refers to the model fine-tuned on the Common Voice dataset whereas "contrastive2" refers to the model fine-tuned on the OpenSLR dataset.We chose "contrastive2" as our primary submission for ASR.

Submission	WER	CER
contrastive1	62.9	17.5
contrastive2	69.3	21.2
primary	69.3	21.2

Table 3: Results of our ASR System for Marathi-Hindi

Table 4 shows the results of our Speech Translation(ST) system (i.e ASR+MT). Here, "contrastive1" refers to ASR using "contrastive1" ASR system and MT using the fine-tuned NLLB model whereas "contrastive2" refers to ASR using "contrastive2" ASR system and MT using the fine-tuned NLLB model. Our "primary" submission consists of the translations which have higher cosine similarity to their respective transcriptions (from their respective ASR models).

Submission	BLEU	chrF2
contrastive1	25	50.1
contrastive2	19	44.8
primary	21.3	48.1

Table 4: Results for our ST System for Marathi-Hindi

5.2 Bhojpuri - Hindi

Table 5 below shows the results for our primary ST system for Bhojpuri-Hindi which consists of a vakyansh-wav2vec model for ASR and NLLB-1.3B distilled model for MT.

Submission	BLEU	chrF2
primary	12.9	41.1

Table 5: Results for our ST System for Bhojpuri-Hindi

6 Conclusion and Future Work

In this paper, we have presented our Speech Translation Systems for the dialectal and low-resource track of IWSLT 2024 employing a cascaded approach using fine-tuned models for both ASR and MT. Our submission trailed by a chrF2 score of 20 in comparison to the best submission in Marathi-Hindi task(unconstrained). In Bhojpuri-Hindi task(unconstrained) our submission trailed the best submission by a chrF2 score of 8.4 .
Our future work will comprise of using data-augmentation techniques and fine-tuning multiple pre-trained multilingual models and exploring more speech translation models for Low-Resource Indian Languages.

References

J. Ameta, N. Joshi, and I. Mathur. 2013. Improving the quality of gujarati-hindi machine translation through part-of-speech tagging and stemmer assisted transliteration. *International Journal on Natural Language Computing*, 2:49–54.

[6] https://huggingface.co/facebook/nllb-200-distilled-1.3B

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Harveen Singh Chadha, Anirudh Gupta, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2022. Vakyansh: Asr toolkit for low resource indic languages. *Preprint*, arXiv:2203.16512.

J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nußbaum-Thom, M. Picheny, Z. Tüske, P. Golik, R. Schlüter, H. Ney, M. Gales, K. Knill, A. Ragni, H. Wang, and P. Woodland. 2015. Multilingual representations for low resource speech recognition and keyword search.

D. Dash, M. Kim, K. Teplansky, and J. Wang. 2018. Automatic speech recognition with articulatory information and a unified dictionary for hindi, marathi, bengali and oriya.

A. Diwan. 2021. Multilingual and code-switching asr challenges for low resource indian languages.

N. Durrani, H. Sajjad, H. Hoang, and P. Koehn. 2014. Integrating an unsupervised transliteration model into statistical machine translation. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Pa*.

C. Escolano, M. Costa-jussà, J. Fonollosa, and C. Segura. 2021. Enabling zero-shot multilingual spoken language translation with language-specific encoders and decoders.

L. Gao, X. Wang, and G. Neubig. 2020. Improving target-side lexical transfer in multilingual neural machine translation.

A. Godase and S. Govilkar. 2015. Machine translation development for indian languages and its approaches. *International Journal on Natural Language Computing*, 4:55–74.

Anirudh Gupta, Harveen Singh Chadha, Priyanshi Shah, Neeraj Chimmwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2021. Clsril-23: Cross lingual speech representations for indic languages. *Preprint*, arXiv:2107.07402.

A. Hameed, I. Qazi, and A. Raza. 2022. Towards representative subset selection for self-supervised speech recognition.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6494–6503, Marseille, France. European Language Resources Association (ELRA).

B. Ji, Z. Zhang, X. Duan, M. Zhang, B. Chen, and W. Luo. 2019. Cross-lingual pre-training based transfer for zero-shot neural machine translation.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

N. G. Kharate and V. Patil. 2021. Inflection rules for marathi to english in rule based machine translation. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 10:780.

S. Madikeri, B. K. Khonglah, S. Tong, P. Motlíček, H. Bourlard, and D. Povey. 2020. Lattice-free maximum mutual information training of multilingual speech recognition systems. *Interspeech 2020*.

. Mamyrbayev, K. Alimhan, . , A. Bekarystankyzy, and B. Zhumazhanov. 2022. Identifying the influence of transfer learning method in developing an end-to-end automatic speech recognition system with a low data level. *Eastern-European Journal of Enterprise Technologies*, 1:84–92.

G. Neubig and J. Hu. 2018. Rapid adaptation of neural machine translation to new languages.

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.

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.

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022.

Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

H. Sailor, M. Krishna, D. Chhabra, A. Patil, M. Kamble, and H. Patil. 2018. Da-iict/iiitv system for low resource speech recognition challenge 2018.

S. Saini and V. Sahula. 2015. A survey of machine translation techniques and systems for indian languages. *2015 IEEE International Conference on Computational Intelligence Amp; Communication Technology*.

S. R. Shahamiri and S. S. Salim. 2014. A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artificial neural networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22:1053–1063.

X. Zhang, X. Li, Y. Yang, and R. Dong. 2020. Improving low-resource neural machine translation with teacher-free knowledge distillation. *Ieee Access*, 8:206638–206645.

B. Zoph, D. Yüret, J. May, and K. Knight. 2016. Transfer learning for low-resource neural machine translation.