Overview of MTIL Track at FIRE 2023: Machine Translation for Indian Languages

Surupendu Gangopadhyay[a], Ganesh Epili[a], Prasenjit Majumder[a], Baban Gain[b], Ramakrishna Appicharla[b], Asif Ekbal[b], Arafat Ahsan[c] and Dipti Sharma[c]

[a]*Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India*
[b]*Indian Institute of Technology Patna, Patna, India*
[c]*International Institute of Information Technology - Hyderabad, Hyderabad, India*

Abstract

The objective of the MTIL track in FIRE 2023 was to encourage the development of Indian Language to Indian Language (IL-IL) Neural Machine Translation models. The languages covered included Hindi, Gujarati, Kannada, Odia, Punjabi, Urdu, Telugu, Kashmiri, and Sindhi. The track comprised of two tasks: (i) a General Translation Task and (ii) a Domain specific Translation Task with Governance and Healthcare being the chosen domains. For the listed languages, we proposed 12 diverse language directions for the general domain translation task and 8 each for healthcare and governance domains. Participants were encouraged to submit models for one or more language pairs. Consequently, we witnessed the creation of 34 distinct models spanning various language pairs and domains. Model assessments were conducted using five evaluation metrics: BLEU, CHRF, CHRF++, TER, and COMET. The submitted model outputs were ultimately ranked using the CHRF score.

Keywords

Neural Machine Translation, Domain specific Machine Translation, Machine Translation for Low resource Indic languages

1. Introduction

Research on translation of low-resource languages opens up new challenges in the field of neural machine translation. Many Indian languages, especially, when the translation directions are IL-IL fall under a low resource scenario, hence the need for experimentation, and discovery of new techniques, that can help effectively translate between low resource language pairs. While some shared tasks previously have focused on Indic-English[1] low resource language translation settings, the Indic-Indic translation directions need further exploration. This shared task, titled as Machine Translation for Indian Languages (MTIL) aims to fill this gap by proposing a number of Indic-Hindi and Hindi-Indic translation directions making test data available for a number of these pairs. Furthermore, the shared task also proposes domain-specific translation with

Forum for Information Retrieval Evaluation, 15-18 December 2023, Panjim, India
surupendu.g@gmail.com (S. Gangopadhyay); ganeshepili1998@gmail.com (G. Epili); prasenjit.majumder@gmail.com (P. Majumder); gainbaban@gmail.com (B. Gain); ramakrishnaappicharla@gmail.com (R. Appicharla); asif.ekbal@gmail.com (A. Ekbal); arafat.ahsan@iiit.ac.in (A. Ahsan); dipti@iiit.ac.in (D. Sharma)


[1]https://www2.statmt.org/wmt23/indic-mt-task.html

Governance and Healthcare being the domains in focus.

1.1. Task 1: General Translation Task

Task 1 is meant to be a general domain translation task where participants are required to build a model to translate the language pairs shown in Table 1. They are provided pointers to existing training data that may be mined for Hindi-Indic direction pairs. The task is unconstrained and participants are free to adapt or leverage existing data and models to create the best models for each translation direction.

Table 1
Languages Pairs (Task 1)

Sr. No.	Language Pairs
1.	Hindi-Gujarati
2.	Gujarati-Hindi
3.	Hindi-Kannada
4.	Kannada-Hindi
5.	Hindi-Odia
6.	Odia-Hindi
7.	Hindi-Punjabi
8.	Punjabi-Hindi
9.	Hindi-Sindhi
10.	Urdu-Kashmiri
11.	Telugu-Hindi
12.	Hindi-Telugu

1.2. Task 2: Domain-specific Translation Task

- Task 2a (Governance): In this subtask, the participants have to build a model to translate sentences in the Governance domain.
- Task 2b (Healthcare): In this subtask, the participants have to build a model to translate sentences in the Healthcare domain.

The language pairs used in both the subtasks are shown in Table 2.

We evaluate the submissions using the following evaluation metrics: BLEU, CHRF, CHRF++, TER, and COMET. However, the final ranking is based on the CHRF scores. The evaluation metrics are discussed in Section 3.

2. Dataset

We encouraged participants to leverage existing publicly available parallel or monolingual data for this shared task. Specifically, we encouraged the use of the Bharat Parallel Corpus Collection BPCC[1] released by AI4Bharat. The Bharat Parallel Corpus Collection (BPCC) is currently the largest English-Indic parallel corpus encompassing data for all 22 scheduled Indian

Table 2
Languages Pairs (Task 2)

Sr. No.	Language Pairs
1.	Hindi-Gujarati
2.	Gujarati-Hindi
3.	Hindi-Kannada
4.	Kannada-Hindi
5.	Hindi-Odia
6.	Odia-Hindi
7.	Hindi-Punjabi
8.	Punjabi-Hindi

languages. This collection comprises of two sections, BPCC-Mined and BPCC-Human, and contains approximately 230 million pairs of bitext. The BPCC-Mined section incorporates about 228 million pairs, with nearly 126 million pairs freshly added as part of this initiative. This component plays a pivotal role in augmenting the available data for all 22 scheduled Indian languages. On the other hand, BPCC-Human consists of 2.2 million gold standard English-Indic pairs. Additionally, it includes 644K bitext pairs sourced from English Wikipedia sentences, forming the BPCC-H-Wiki subset, and 139K sentences covering everyday use cases, forming the BPCC-H-Daily subset. The statistics of the dataset is shown in Table 3.

Table 3
Statistics of BPCC

BPCC-Mined	Existing	Samanantar	19.4M
		NLLB	85M
	Newly Added	Samanantar++	121.6M
		Comparable	4.3M
BPCC-Human	Existing	NLLB	18.5K
		ICLI	1.3M
		Massive	115K
	Newly Added	Wiki	644K
		Daily	139K

The participants were free to utilize the entire collection or a subset of the corpus based on their needs.

Test Data To ensure accurate evaluation of model performance, we make available a manually translated test corpus for each language pair listed in either of the sub-tasks. The test set of Task 1 comprises of 2000 sentences, while the test set of Task 2 comprises 1000 sentences for each language pair. Test sets are blind and only the Source is released for translation submissions.

3. Evaluation Metrics

Evaluation is performed utilizing multiple metrics. Canonical string-based metrics like BLEU, CHRF and TER are used and a pre-trained metric (COMET) is also utilized. The choice of metrics

was influenced by two factors: that the languages under evaluation exhibited considerable morphological variation, thus the metric must not be biased against morphological complexity; that while recently popular pre-trained metrics have shown greater correlations with human judgements, they are yet to be proven to scale to lower resource languages, thus providing an opportunity to test them for certain low resource languages that made up this shared task. We use the SacreBLEU [2] library to evaluate the submissions. The evaluation metrics that we use in this shared task are described below:

1. BLEU: The BLEU [3] score evaluates the quality of a translation based on the overlap of n-grams between the hypothesis and reference and the length of the hypothesis w.r.t. the reference. It uses unigrams, bigrams, trigrams, and four-grams to measure the overlap of the n-grams. The BLEU score uses a brevity score to penalize the hypothesis that is shorter in length than the reference. Since we are measuring the BLEU score w.r.t. percentage, the value lies between 0-100, wherein a high BLEU indicates a better quality of translation.
2. CHRF: The CHRF [4] score evaluates the quality of a translation based on the overlap of character level n-gram between the hypothesis and reference. It calculates the F score using the character level n-gram precision and recall. The CHRF score is better for evaluating translations of morphologically rich languages. The formula of CHRF is shown in Equation 1 wherein $CHRP$ is the percentage of n-grams present in the hypothesis, which is present in the reference, and $CHRR$ is the percentage of n-grams present in the reference, which is present in the hypothesis. The scores have a higher correlation to human-level judgments as compared to BLEU. We used character level n-gram of 6 characters, and β was set to 1 to calculate the CHRF score.

$$CHRF(\beta) = (1 + \beta^2)\frac{CHRP.chrR}{\beta^2.CHRP + CHRR} \tag{1}$$

3. CHRF++: The CHRF++ [5] is a modification over the CHRF wherein it considers the overlap of character level and word level n-grams between the hypothesis and reference. It takes an average of the F score of character level n-grams and the F score of word level n-grams. We used the character level n-gram of 6 characters, word level n-gram of 1, and β set to 1.
4. TER: The TER [6] score measures the number of edits (insert, delete, substitution, and shift) required to match the hypothesis to the reference. The lower the TER score, the better the performance of the model.
5. COMET: COMET [7] is an embedding-based metric to measure the similarity between the hypothesis and reference. It uses an encoder to get the source (s), hypothesis (h), and reference (r) sentence embeddings. The COMET score uses Equation 2 to calculate a harmonic mean over the Euclidean distances between the reference and hypothesis and source and hypothesis. It uses Equation 3 to calculate the similarity between the hypothesis and reference. We use Unbabel/wmt22-comet-da as the COMET evaluation model.

$$f(s, h, r) = \frac{2.d(s, h).d(r, h)}{d(s, h) + d(r, h)} \tag{2}$$

$$f(s,h,r) = \frac{1}{1 + f(s,h,r)} \tag{3}$$

4. Results

We have received 34 submissions across various domains and language pairs. However, only three of the teams submitted a paper detailing the methodology they employed. The results are shown in Tables 4, 5, and 6.

In Task 1 and 2 for the language pairs Hindi-Odia and Odia-Hindi the team IIIT-BH-MT is the best performing team followed by BITSP. IIIT-BH-MT team used a custom Hindi-Odia and Odia-Hindi dataset which consists of 36100 parallel sentences. The authors finetuned the distilled version of NLLB 200 model which consists of 600M parameters for the translation task. The BITSP team use the BPCC dataset wherein they use English as the pivot language while translating from Odia to Hindi and vice versa. In Task 1 the authors use a combination of IndicTrans2 and NLLB models to generate the translations from the source language and MuRIL to generate sentence embeddings from the translations. The authors then select the best translation based on the cosine similarity. For Task 2 the authors create a domain specific dataset from BPCC by using BART-MNLI to assign the class labels. They finetune the NLLB model to perform task specific translation.

Table 4
Results of Task 1

Language Pair	Team Name	Score				
		BLEU	**CHRF**	**CHRF++**	**TER**	**COMET**
Hindi-Odia	BITSP	20.057	56.389	51.836	63.967	0.842
	SLPBV	11.858	46.741	42.540	83.247	0.752
	IIIT-BH-MT	30.505	**61.998**	59.048	51.192	0.842
Odia-Hindi	BITSP	29.374	55.572	53.309	56.188	0.804
	SLPBV	20.756	47.645	45.513	67.952	0.672
	IIIT-BH-MT	44.154	**66.395**	64.929	41.876	0.837
Gujarati-Hindi	SLPBV	22.989	**48.687**	46.746	62.295	0.702
Punjabi-Hindi	SLPBV	38.700	63.351	61.653	56.460	0.792
	CDACN-Punjabi	62.195	**77.456**	76.601	22.231	0.837
Hindi-Punjabi	SLPBV	33.966	60.406	58.528	56.667	0.818
	CDACN-Punjabi	50.939	**69.790**	68.184	38.088	0.845
Hindi-Sindhi	SLPBV	0.465	1.029	1.137	119.441	0.385

In case of Punjabi-Hindi and Hindi-Punjabi the CDACN-Punjabi team is the best performing team except in Task 2a where SLPBV team is the best performing team. The CDACN-Punjabi team use a custom dataset of Punjabi-Hindi and vice versa language pairs to train their model. The authors finetune the NLLB-200 model for the translation task.

In Hindi-Sindhi the scores for SLPBV team's submission was unusually low. However, on closer examination, we found that the submitted model outputs are using the Devanagri script

Table 5
Results of Task 2a (Governance)

Language Pair	Team Name	Score				
		BLEU	CHRF	CHRF++	TER	COMET
Hindi-Odia	BITSP	23.039	60.327	55.885	61.224	0.867
	SLPBV	13.317	50.251	45.852	82.538	0.802
	IIIT-BH-MT	32.607	**65.137**	61.974	49.936	0.865
Odia-Hindi	BITSP	20.031	42.329	40.916	65.476	0.822
	SLPBV	14.795	35.984	35.112	70.730	0.708
	IIIT-BH-MT	30.252	**49.201**	48.752	54.244	0.858
Gujarati-Hindi	SLPBV	23.673	**49.721**	47.654	66.682	0.682
Punjabi-Hindi	SLPBV	32.236	**59.285**	57.212	69.074	0.822
	CDACN-Punjabi	33.119	56.169	54.636	51.554	0.818
Hindi-Punjabi	SLPBV	39.633	58.536	57.768	51.914	0.795
	CDACN-Punjabi	56.894	**73.695**	72.859	25.757	0.817
Hindi-Sindhi	SLPBV	0.465	1.029	1.137	119.441	0.385

Table 6
Results of Task 2b (Healthcare)

Language Pair	Team Name	Score				
		BLEU	CHRF	CHRF++	TER	COMET
Hindi-Odia	BITSP	15.225	53.323	48.381	69.468	0.823
	SLPBV	7.812	41.393	37.275	90.843	0.713
	IIIT-BH-MT	23.373	**56.749**	53.537	58.438	0.806
Odia-Hindi	BITSP	31.931	55.342	53.620	53.791	0.739
	SLPBV	15.443	40.298	38.321	76.198	0.590
	IIIT-BH-MT	39.216	**60.786**	59.174	49.566	0.763
Gujarati-Hindi	SLPBV	24.300	**50.180**	48.113	60.720	0.719
Punjabi-Hindi	SLPBV	34.270	59.177	57.439	61.152	0.793
	CDACN-Punjabi	37.518	**60.854**	59.521	42.060	0.838
Hindi-Punjabi	SLPBV	42.328	66.430	64.824	48.029	0.590
	CDACN-Punjabi	65.055	**79.577**	78.815	20.454	0.852
Hindi-Sindhi	SLPBV	0.465	1.029	1.137	119.441	0.385

for Sindhi, whereas our ground truth has Sindhi in the Perso-Arabic script. In Gujarati-Hindi we had only one submission by the SLPBV team.

In Task 1 from Table 4 we observe that in Hindi-Odia language pair the COMET score of IIIT-BH-MT and BITSP teams are same. Similarly in Task 2a as well the COMET scores of the both the teams are same and in Task 2b the COMET score of BITSP is higher than IIIT-BH-MT by 0.017. However if we observe the other metrics then it shows a difference in performance of both the teams. However in other language pairs we do not observe this anomalous behaviour.

5. Concluding Discussions

Our shared task focuses on Indic-Indic language translation instead of Indic-English language translation. From the submission we observe that NLLB model is widely used among the participants for the translation task. Among the three teams we found one team which used BPCC as the training dataset and English as the pivot language. While the other two teams used custom dataset for the shared tasks. We observed some anomalous behaviour between the chrF and COMET scores in case of Hindi-Odia language pair wherein the first and second team had the same COMET score, and one case in the Governance domain sub-task, where the CHRF and COMET scores were discordant. We need to investigate this further by using human annotators to evaluate the translations. However in other language pairs we did not observe this behaviour.

Acknowledgments

The track organizers thank all the participants for their interest in this track. We also thank the FIRE 2023 organizers for their support in organizing the track. We also thank BHASHINI (https://bhashini.gov.in/en) for enabling the HIMANGY consortium to create the test dataset which helped us to conduct this track. We thank the Principal Investigator, Co-Principal Investigators and Host Institute (IIIT Hyderabad) for providing us with this opportunity of using the dataset in the track. We also thank Ministry of Electronics and Information Technology (MeitY) and Ministry of Human Resource Development, Government of India for providing this opportunity to develop the dataset and other resources.

References

[1] AI4Bharat, J. Gala, P. A. Chitale, R. AK, S. Doddapaneni, V. Gumma, A. Kumar, J. Nawale, A. Sujatha, R. Puduppully, V. Raghavan, P. Kumar, M. M. Khapra, R. Dabre, A. Kunchukuttan, Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages, arXiv preprint arXiv: 2305.16307 (2023).

[2] M. Post, A call for clarity in reporting BLEU scores, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 186–191. URL: https://www.aclweb.org/anthology/W18-6319.

[3] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[4] M. Popović, chrf: character n-gram f-score for automatic mt evaluation, in: Proceedings of the tenth workshop on statistical machine translation, 2015, pp. 392–395.

[5] M. Popović, chrf++: words helping character n-grams, in: Proceedings of the second conference on machine translation, 2017, pp. 612–618.

[6] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, 2006, pp. 223–231.

[7] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2685–2702. URL: https://aclanthology.org/2020.emnlp-main.213. doi:10.18653/v1/2020.emnlp-main.213.