INDICXNLI: Evaluating Multilingual Inference for Indian Languages

Divyanshu Aggarwal[1*], **Vivek Gupta**[2*†], **Anoop Kunchukuttan**[3, 4]
[1]Delhi Technological University; [2]University of Utah; [3]Microsoft IDC; [4]AI4Bharat
divyanshuggrwl@gmail.com ; vgupta@cs.utah.edu ; ankunchu@microsoft.com

Abstract

While Indic NLP has made rapid advances recently in terms of the availability of corpora and pre-trained models, benchmark datasets on standard NLU tasks are limited. To this end, we introduce INDICXNLI, an NLI dataset for 11 Indic languages. It has been created by high-quality machine translation of the original English XNLI dataset and our analysis attests to the quality of INDICXNLI. By fine-tuning different pre-trained LMs on this INDICXNLI, we analyze various cross-lingual transfer techniques with respect to the impact of the choice of language models, languages, multi-linguality, mix-language input, etc. These experiments provide us with useful insights into the behaviour of pre-trained models for a diverse set of languages.

1 Introduction

Natural Language Inference (NLI) is a well-studied NLP task (Dagan et al., 2013) that assesses if a premise entails, negates, or is neutral towards the hypothesis statement. The task is well suited for evaluating semantic representations of state-of-the-art transformers (Vaswani et al., 2017) models such as BERT (Devlin et al., 2019; Radford and Narasimhan, 2018). Two large scale datasets, such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018), has recently been developed to enhanced the relevance of the NLI task.

With the availability of multi-lingual pre-trained language models such as mBERT (Devlin et al., 2019), and XLM-RoBERTa (Conneau et al., 2020a) promising cross-lingual transfer and universal models, multi-lingual NLP has recently gained a lot of attention. However, most languages have a scarcity of datasets resources. Some multi-lingual datasets have attempted to fill this gap, including XNLI (Conneau et al., 2018) for NLI, XQUAD (Dumitrescu et al., 2021), MLQA (Lewis et al., 2020) for question answering, and PAWS-X for paraphrase identification (Yang et al., 2019). In many practical circumstances, training sets for non-English languages are unavailable, hence cross-lingual zero-shot evaluation benchmarks such as XTREME (Hu et al., 2020a), XTREME-R (Ruder et al., 2021), and XGLUE (Liang et al., 2020) have been suggested to use these datasets.

However, NLI datasets are not available for major Indic languages. The only exceptions are the test/validation sets in the XNLI (hi and ur), TaxiXNLI (hi) (K et al., 2021) and MIDAS-NLI (Uppal et al., 2020) datasets. Furthermore, because MIDAS-NLI is based on sentiment data recasting, hypotheses are not linguistically diverse and span limited reasoning. In this work, we address this gap by introducing INDICXNLI, an NLI dataset for *Indic* languages. INDICXNLI consists of English XNLI data translated into eleven *Indic* languages. We use INDICXNLI to evaluate *Indic*-specific models (trained only on *Indic* and English languages) such as IndicBERT (Kakwani et al., 2020) and MuRIL (Khanuja et al., 2021), as well as generic (train on non-*Indic* languages) such as mBERT and XLM-RoBERTa. Furthermore, we experimented with several training strategies for each multi-lingual model. Our experimental results answers multiple important questions regarding effective training for *Indic* NLI. Our contributions are as follows:

- We introduce INDICXNLI, an NLI benchmark dataset for eleven prominent Indo-Aryan *indic* languages from the Indo-European and Dravidian language families.
- We investigate several strategies to train multi-lingual models for NLI tasks on INDICXNLI. We also explore models cross-lingual NLI transfer ability across *Indic* languages and Intra-Bilingual NLI ability of pretrained multi-lingual language models.

*Equal Contribution †Corresponding Author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10994 - 11006

The INDICXNLI dataset, along with scripts, is available at https://indicxnli.github.io/.

2 The INDICXNLI dataset

We created INDICXNLI, a NLI data set for *Indic* languages. INDICXNLI is similar to existing XNLI dataset in shape/form, but focusses on *Indic* language family. INDICXNLI include NLI data for eleven major *Indic* languages that includes Assamese ('as'), Gujarat ('gu'), Kannada ('kn'), Malayalam ('ml'), Marathi ('mr'), Odia ('or'), Punjabi ('pa'), Tamil ('ta'), Telugu ('te'), Hindi ('hi'), and Bengali ('bn'). Next we describe the INDICXNLI construction and its validation in details.

INDICXNLI **Construction.** To create INDICXNLI, we follow the approach of the XNLI dataset and translate the English XNLI dataset (premises and hypothesis) to eleven *Indic*-languages. We use the IndicTrans (Ramesh et al., 2022), a state-of-the-art, publicly available translation model for Indic languages, for translating from English to *Indic* languages. The train (392,702), validation (2,490), and evaluation sets (5,010) of English XNLI were translated from English into each of the eleven *Indic* languages. IndicTrans is a large Transformer-based sequence to sequence model. It is trained on Samanantar dataset (Ramesh et al., 2022), which is the largest parallel multi-lingual corpus over eleven *Indic* languages. IndicTrans outperforms other open-source models based on mBART (Liu et al., 2020) and mT5 (Xue et al., 2021) for *Indic* language translations and is competitive with paid translation models such as Google-Translate or Microsoft-Translate on several benchmarks (Ramesh et al., 2022). Our choice of IndicTrans was motivated by *cost, language coverage and speed*, refer Appendix §A.

INDICXNLI **Validation.** While translation may lose the semantic link between the sentences, recent study by K et al. (2021) disproved this. K et al. (2021) qualitative analysis illustrate that when a high-quality machine translation system is utilized, classification labels and reasoning categories are only minimally altered for translated NLI datasets. We also demonstrate the high quality of IndicTrans translation for INDICXNLI in two ways (a.) manual human validation and, (b.) automatic metric BERTScore (Zhang* et al., 2020). Our validation approach guarantee correctness for the INDICXNLI labels. Next, we'll discuss on how to evaluate IndicTrans translations.

Score	hi	te	pa	bn	as	gu	ta	ml	kn	mr	or
HS1	88	88	91	87	87	89	89	87	89	86	88
HS2	81	84	93	83	84	89	87	87	87	87	90
PC	73	73	89	79	78	79	76	85	83	83	75
SC	82	87	94	90	88	85	88	93	86	89	85

Table 1: Human Validation Score ($\times 10^{-2}$): **HS1**, **HS2** represents human1, human2 annotation score respectively. **PC** and **SC** represents Pearson and Spearman correlation respectively.

HUMAN VALIDATION: We followed SemEval-2016 Task-I (Agirre et al., 2016) guidelines. We hired 2 annotators per languages and calculated the pearson (Kirch, 2008) and spearman (spe, 2008) correlation over annotations scores of sentences.

DIVERSE SAMPLING: Since human validation is time-consuming and expensive. We sampled 100 diverse sentences of the test set for validation. We apply the Determinantal Point Process (Kulesza, 2012) (DPP) over sentence representations for diverse sampling. DPP maximizes coverage volume using a minimal sampled set, thus guaranteeing diversity during sampling. We first used sentence transformers to convert data to BERT embeddings, and then use k-DPP (Kulesza and Taskar, 2011) with k = 100 to sample 100 examples. Using DPP for diverse sampling is a cost-effective method of evaluating translation quality. For scoring guidelines refer to Appendix §B.

HIRING EXPERTS: We recruited, 2 speakers for each of the 11 *indic* languages as annotators. These professional annotators are multilingual (English, *Indic*) and fluent in both mother-tongue *indic* and English language. The remuneration paid was 6.6 cents per sentence [1] for each *indic* language.

EVALUATION: Table 1 shows the final human evaluation scores. In general, we see that average human scores is more than 0.85 for all languages. The Pearson and Spearman Correlation values are more than 0.7 and 0.8 for all languages respectively. High human ratings and high correlation between the annotations support high quality IndicTrans translation, hence validating INDICXNLI quality.

AUTOMATIC VALIDATION: Given the absence of *Indic* language XNLI reference data, we use BERTScore similarity between the original English and English translated INDICXNLI for automatic evaluation. Here too, we use the IndicTrans model for translating INDICXNLI into English. This approach estimates the upper bound on error for the

[1] above minimum wage in India.

BS	hi	te	pa	bn	as	gu	ta	ml	kn	mr	or
$\textbf{ET}^{GT}$	94	93	92	94	NA	94	94	94	94	94	94
$\textbf{ET}^{IT}$	98	94	94	98	93	94	94	94	94	93	93
$\textbf{ML}^{GT}$	90	88	86	89	NA	89	86	85	88	87	82
$\textbf{ML}^{IT}$	96	87	88	96	85	96	87	87	87	86	86

Table 2: **BS** represent BERTScore (F1-Score $\times 10^{-2}$) for **EngTrans** (ET) and **Multilingual** (ML) strategies. Superscript ^{GT} and ^{IT} represent Google Translate and IndicTrans models respectively.

English to *Indic* translation (i.e. INDICXNLI quality), as it approximates the combined error of both English to *Indic* translation (INDICXNLI creation), and *Indic* to English translation (evaluation) (Rapp, 2009; Miyabe and Yoshino, 2015; Edunov et al., 2020; Behr, 2017). We utilize BERTScore for assessment since it correlates better with human judgment at the sentence level than BLEU (Zhang* et al., 2020; Papineni et al., 2002).

We evaluate two translation models, Google Translate and IndicTrans on the testsets of INDICXNLI dataset. We incorporate Google Translate to demonstrate IndicTrans's competitiveness in comparison to commercial translation approaches. In Table 2, we used two evaluation strategies for our evaluation (a.) *EngTrans*: which take the INDICXNLI sentence and translated it back to English using BERT model. (b.) *Multilingual*: directly compare the English sentences with multilingual INDICXNLI sentences using mBERT model.

On *Indic* languages, we notice that IndicTrans is comparable to, and sometimes outperforms, Google Translate. Additionally, when results are compared in a Multilingual setting, we observe a marginal decrement in scores. This can be because mBERT does not produce as precise multilingual embedding as BERT does for English. Additionally, we see a similar pattern in the distribution of scores across languages for both assessment strategies on both models. We also computed the BERTScore (using mBERT) between the Hindi test set of XNLI and INDICXNLI was found to be 0.87, supporting the high quality of INDICXNLI.

3 Experiments

EXPERMIMENTAL SETUP: Our experiments compare the performance of several multi-lingual models, including one particularly developed for *Indic* languages. We consider 2 broad categories, (a) ***Indic* Specific** which includes IndicBERT and MuRIL due to their indic specific pretraining, and (b) **Generic** which includes mBERT and XLM-Roberta due to their pretraining in more than 100 languages. We fine-tuned pre-trained multi-lingual models to develop NLI classifiers. The classifiers takes two sentence as input, i.e. the premise and the hypothesis and predicts the inference label. See Appendix §C and §D for models and hyper-parameters details respectively.

Training-Evaluation Strategies. To train the NLI classifier, we investigate several strategies. While the pre-trained multi-lingual models remain constant, the training and evaluation datasets vary. **1. *Indic* Train:** The models are trained and evaluated on INDICXNLI. The training set is translated from the XNLI English, thus a *translate-train* scenario. **2. English Train:** The models are trained on original English XNLI data and evaluated on INDICXNLI data. This is a *zero-shot evaluation* training scenario. **3. English Eval:** The model are trained on original English XNLI data, but evaluated on English translation of INDICXNLI data. This is the *translate-test* scenario. **4. English + *Indic* Train:** This approach combines approaches *(1)* and *(2)*. The model is first pre-finetuned (Lee et al., 2021; Aghajanyan et al., 2021) on English XNLI data and then finetuned on ***Indic* language** of INDICXNLI data. **5. Train All:** This approach begins by fine-tuning the pre-trained model on English XNLI data, followed by training on *all eleven* Indic *languages* of INDICXNLI sequentially. **6. Cross Lingual Transfer:** Additionally, we assess the models' capacity to transfer between languages. Where the model is trained on a single Indian language and then assessed on all other Indian languages as well as the training language. **7. Intra-Bilingual Inference:** Lastly, we also asses the model's capability to perform natural language inference with premise in English and hypothesis in Indic language.

RESULTS AND ANALYSIS. We summarizes our findings from Table §3 results across 4 categories:

ACROSS MODELS: In all experiments, MuRIL performs the best across all *indic* languages except in English Eval setup. This can be attributed to (a.) The large model size (b.) indic-specific pre-training data, (c.) A Mixture of Masked Language Modeling (MLM), Translation Language Modeling (TLM), and (d.) use of transliterated data in pre-training. XLM-RoBERTa beats MuRIL in rare scenarios, notably in which the model solely deals with English data (e.g. English Eval). XLM-RoBERTa outperforms MuRIL in such cases because it is better at assessing English than MuRIL,

Model	Indic Train											ModAvg	English Train											ModAvg
	as	gu	kn	ml	mr	or	pa	ta	te	bn	hi		as	gu	kn	ml	mr	or	pa	ta	te	bn	hi	
XLM-R	70	73	75	70	75	32	71	76	76	76	78	70	65	66	69	69	67	67	61	71	69	69	73	69
iBERT	67	69	68	60	68	69	73	37	62	70	68	65	57	63	53	42	59	57	66	41	56	48	63	60
mBERT	71	62	69	71	71	35	70	70	69	67	74	66	51	57	57	57	54	34	59	61	59	57	67	59
MuRIL	70	78	75	76	70	76	72	74	78	75	71	**74**	68	32	75	34	68	67	70	74	71	74	76	**72**
LangAvg	68	69	70	69	70	49	71	65	70	70	**72**	68	58	55	64	52	61	52	63	61	63	62	**68**	63
Model	**English Eval**											**ModAvg**	**English+*Indic* Train**											**ModAvg**
	as	gu	kn	ml	mr	or	pa	ta	te	bn	hi		as	gu	kn	ml	mr	or	pa	ta	te	bn	hi	
XLM-R	66	72	70	68	66	65	72	69	72	71	75	**70**	73	75	77	75	74	73	75	75	73	75	79	76
iBERT	63	66	68	61	65	65	66	63	63	72	72	66	67	72	65	62	59	59	74	63	66	69	74	70
mBERT	62	64	67	65	61	60	66	67	66	75	72	66	67	70	69	70	70	39	71	73	70	70	71	69
MuRIL	65	33	71	67	67	67	71	31	71	72	77	63	76	77	77	79	74	76	77	77	74	75	77	**77**
LangAvg	64	60	68	65	63	64	69	60	68	73	**74**	66	69	73	70	72	68	56	73	72	70	72	**75**	**72**
Model	**Train All**											**ModAvg**	**Cross Lingual Transfer**											**ModAvg**
	as	gu	kn	ml	mr	or	pa	ta	te	bn	hi		as	gu	kn	ml	mr	or	pa	ta	te	bn	hi	
XLM-R	73	77	74	76	72	73	77	77	76	77	77	75	66	70	33	34	70	35	68	70	70	71	72	60
iBERT	63	74	59	51	69	66	75	60	67	70	74	66	59	60	59	54	60	60	60	56	59	58	60	59
mBERT	63	69	69	71	70	33	71	69	70	74	72	66	57	59	60	59	58	33	59	60	59	60	60	57
MuRIL	73	76	74	76	74	78	81	78	76	80	78	**77**	75	73	75	76	71	33	75	76	73	75	73	**70**
LangAvg	68	73	69	68	71	58	75	71	71	**75**	74	70	63	64	57	56	63	39	64	64	64	**65**	**65**	60

Table 3: Here, LangAvg represents the language wise average score across models, while ModAvg average score represents the model average score across languages. Values in **Blue**, **Red** and **Green** represents the model average best score, language-wise average best score, and values where both model-wise and language-wise best score coincide. For Indic Cross Lingual Transfer, each row represent the average evaluation score of all *Indic* language when trained on the column language. For more detailed cross-lingual transfer results refer to Appendix §E. iBERT stand for *indic*BERT and XLM-R stand for XLM-RoBERTa.

which is designed mostly for indic language. Additionally, we discover that, compared to XLM-RoBERTa, MuRIL indic-specific training further enhances the model's performance. Despite indic-specific pretraining, IndicBERT performs worse than mBERT. This can be attributed to the smaller size of the IndicBERT model, i.e. only 33M compared to 167M mBERT (c.f. Table §5 in appendix).

ACROSS LANGUAGE: We see a strong positive correlation between language performance with their resource availability. Hindi and Bengali outperform, whereas Odia mostly underperform on majority of benchmarks. Low-resource languages such as Marathi, Assamese, and Kannada surprising also perform well. This can be attributed to the similarity of Marathi with Hindi script, Assamese with Bengali script, and Kannada with Tamil and Telugu scripts. This is discussed in detail in appendix 4. Odia, a low resource language, lacks script sharing language partners and hence performs poorly. Overall, English + *Indic* Train method outperforms, with MuRIL performing best.

ACROSS STRATEGIES: Our experiments show that models benefit from language-specific fine tuning. English + *Indic* train and Train All have the best results with minimal deviation across languages for XLM-R and MuRIL. Additionally, *Train All* follows a high-to-low resource hierarchy to mitigate the impact of catastrophic forgetting (Goodfellow et al., 2015). Due to the followed language order English + *Indic* train outperform Train All setting marginally for high resource languages. Overall, English + *Indic* Train strategy performs the best and MuRIL performs the best in that strategy. This can be attributed to the *indic* specific pre-training process of MuRIL which include both translation and transliteration. Furthermore, MuRIL has the second largest size after XLM-R.

CROSS-LINGUAL TRANSFER: Models favour high resource languages such as *Hindi* and *Bengali* training for cross-lingual transfer. These language are pre-trained on large mono-lingual corpora which enhanced performance (Conneau et al., 2020a). This setting can be thought equivalent of *Hindi* and *Bengali* substitution for English training. Additionally, when evaluated for all *indic* languages, model trains on non-*Hindi* and non-*Bengali* perform substantially better for *Hindi* and *Bengali*. Table 3 present results summary as average evaluation score across all *indic* language(rows) when train on the several *indic* languages (columns). [2]

INTRA-BILINGUAL INFERENCE: We also evaluate models on mixed input inference task EN-INDICXNLI, which consists of English *premises* paired with corresponding *indic* hypothesis. We

[2] For model-wise cross-lingual results c.f. Appendix §E.

Model	English+*Indic* Train											ModAvg	Train All											ModAvg
	as	gu	kn	ml	mr	or	pa	ta	te	bn	hi		as	gu	kn	ml	mr	or	pa	ta	te	bn	hi	
XLM-R	74	72	75	74	77	72	70	72	72	79	76	**74**	57	59	58	62	61	53	57	59	61	63	63	**59**
iBERT	70	68	63	65	69	68	**71**	64	64	69	69	67	49	53	46	37	52	51	59	39	51	**57**	50	50
mBERT	51	56	59	50	62	31	**63**	57	60	61	**63**	56	39	39	43	38	**43**	33	40	42	41	40	42	40
MuRIL	71	70	**73**	69	71	39	71	71	69	72	69	67	51	52	58	56	53	**55**	**58**	**65**	55	62	54	56
LangAvg	65	65	66	64	68	53	62	65	67	**71**	70	65	47	49	51	48	51	45	52	50	51	**54**	51	50

Table 4: EN-INDICXNLI model performance (refer §3) with English + *Indic* train and Train All setting. Here, ModAvg, LangAvg, and Color Code mean same as in table 3.

train model on mixed input using **English + *Indic* Train** and **Train All** strategies. Table 4 shows performance of **English + *Indic* Train** and **Train All** models on EN-INDICXNLI. Compared to uni-language inference task, mixed-language input task perform poorly. Furthermore, contrary to earlier observations, generic model such as XLM-R outperforms the *Indic* specific models. However, IndicBERT and MuRIL both perform substantially better than mBERT. Furthermore, English data augmentation enhance the **English + *Indic* Train** setting performance. This can be because, the model "meta-learns" the task successfully with English data training (premise language), and further prioritises the model's language-specific abilities with the follow-up indic data training. [3]

4 Error Analysis

In this section, we investigate the correlation between the language similarity and the model performance. We see that the model performs similarly on similar languages. We evaluate our results on MuRIL on the English+*Indic* finetuning Strategy.

In Figure 1 (Appendix), we observe that the overall Correct and Incorrect predictions, Bengali vs Assamese pair has the total of 81% overlap,Tamil vs Kannada has 83% overlap, Hindi vs Maratha has 82% overlap. All the language pairs have the largest overlap for entailment label for correct labels and largest overlap in contradiction label for incorrect overlaps. In Figure 2 (Appendix), interestingly Bengali vs Assamese pair and Hindi vs Marathi has the highest percentage of overlap in predictions where the most overlap is in entailment and minimum overlap is in contradiction. While for Tamil vs Kannada pair has the highest overlap for neutral and minimum for contradiction.

We have also done error analysis of model performance on original Hindi test data already present in XNLI and data obtained through translations from IndicTrans in Figure 3 (Appendix). We observe a total of 82% overlap in error consistency, and we observe that the greatest number of correct overlaps is for the entailment label, whereas the greatest number of incorrect predictions is for the contradiction label. We see the maximum overlap in neutral prediction and the least overlap in contradiction prediction in terms of consistency. This demonstrates that the model performs identically on both the original Hindi data and the machine-translated Hindi data, bolstering the legitimacy of our dataset.

[3] Further Analysis in appendix §F

5 Related Work

Recently many *Indic*-specific resources are developed such as IndicNLPSuite (Kakwani et al., 2020), which include (a.) word embeddings: IndicFT, (b.) transformer models: IndicBERT, (c.) monolingual corpora: IndicCorp, (d.) and, evaluation benchmark: IndicGLUE . Furthermore, *Indic*-specific pre-processing libraries such as iNLTK (Arora, 2020) and Indic-nlp-library (Kunchukuttan, 2020), other Indic monolingual corpora: Common Crawl Oscar Corpus (Wenzek et al., 2020; Ortiz Suárez et al., 2020), multilingual parallel corpora: PMIndia (Haddow and Kirefu, 2020) and Samantar (Ramesh et al., 2022), transformer model MuRIL (Khanuja et al., 2021) and language specific IndicTransformers (Jain et al., 2020) exists.

6 Conclusion

With INDICXNLI we extend the XNLI dataset for *Indic* languages family. We benchmark INDICXNLI with several multi-lingual models using various train-test strategies. We also study the use of English XNLI as pre-finetuning dataset. Furthermore, we also evaluate models on mixed-language inference input and cross-lingual transfer ability. We aim to integrate INDICXNLI and benchmark models in IndicGLUE (Kakwani et al., 2020). We also intend to enhance INDICXNLI with advanced translation techniques. Another direction is accessing model performance on INDIC-INDICXNLI task, where both premises and hypothesis are in two distinct *Indic* languages.

7 Limitations

One of our work's key limitations is that the dataset IndicXNLI was created by machine translation of the original English XNLI dataset. Although IndicXNLI is not human translated, it has been carefully evaluated for translation accuracy by a number of natural bilingual Indic speakers (2 for each language). Furthermore, as shown in our research (Table 2), employing automatic assessment measures such as round trip English-English evaluation via back translation and direct Indic-English sentence comparison is effective. In the past, such a metric has been shown to be highly beneficial for comparing without-reference machine translation (Bapna et al., 2022; Huang, 1990; Moon et al., 2020a,b). Furthermore, as did with the Hindi dataset in Appendix E, we might use correlation in the prediction score between human and machine translated sets for evaluating translation quality.

Second, adapting an existing dataset risks transferring biases and shortcomings from the original XNLI dataset into ours. However, it has been established that XNLI is a typical benchmark for evaluating multilingual and cross-lingual sentence representation, and it has been used to evaluate several multilingual models (Conneau et al., 2020b; Hu et al., 2020b). Morphological analysis of related languages, as well as insights into their performance behavior, may be useful. The authors, however, are not experts in that area, and such an assessment would have been outside the scope of the current work. This study might be expanded to include language groups other than Indian languages such as Indo-European. Third, because of limited resources, the current study did not include large versions of well-known models such as XLM-RoBERTa-Large and MuRIL-Large. However, for IndicBert, mBERT, XLM-RoBERTa, and MuRIL, we assessed model performance in relation to model size (#parameters) in Table 5.

Acknowledgement

We thank members of the Utah NLP group for their valuable insights and suggestions at various stages of the project; and reviewers their helpful comments. We would also like to thank Suhani Aggarwal, Shibani Krishnatraya and Ayush Dhall for participating in the dataset verification activity and helping us find fluent speakers in many different indic languages. Additionally, we appreciate the inputs provided by Vivek Srikumar and Ellen Riloff. Vivek Gupta acknowledges support from Bloomberg's Data Science Ph.D. Fellowship.

References

2008. *Spearman Rank Correlation Coefficient*, pages 502–505. Springer New York, New York, NY.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Gaurav Arora. 2020. iNLTK: Natural language toolkit for indic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 66–71, Online. Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Dorothée Behr. 2017. Assessing the use of back translation: the shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology*, 20(6):573–584.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. Language invariant properties in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 84–92, Dublin, Ireland. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. 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.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, and Viorica Patraucean. 2021. Liro: Benchmark and leaderboard for romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2015. An empirical investigation of catastrophic forgetting in gradient-based neural networks.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020a. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020b. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Xiuming Huang. 1990. A machine translation system for the target language inexpert. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Kushal Jain, Adwait Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-transformers: An analysis of transformer language models for indian languages.

Karthikeyan K, Aalok Sathe, Somak Aditya, and Monojit Choudhury. 2021. Analyzing the effects of reasoning types on cross-lingual transfer performance. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 86–95, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha

Talukdar. 2021. Muril: Multilingual representations for indian languages.

Wilhelm Kirch, editor. 2008. *Pearson's Correlation Coefficient*, pages 1090–1091. Springer Netherlands, Dordrecht.

Alex Kulesza. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286.

Alex Kulesza and Ben Taskar. 2011. K-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1193–1200, Madison, WI, USA. Omnipress.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Hung-yi Lee, Ngoc Thang Vu, and Shang-Wen Li. 2021. Meta learning and its applications to natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 15–20, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Mai Miyabe and Takashi Yoshino. 2015. Evaluation of the validity of back-translation as a method of assessing the accuracy of machine translation. In *2015 International Conference on Culture and Computing (Culture Computing)*, pages 145–150.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020a. Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020b. Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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.

Reinhard Rapp. 2009. The back-translation score: Automatic mt evaluation at the sentence level without

reference translations. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, page 133–136, USA. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shagun Uppal, Vivek Gupta, Avinash Swaminathan, Haimin Zhang, Debanjan Mahata, Rakesh Gosangi, Rajiv Ratn Shah, and Amanda Stent. 2020. Two-step classification using recasted data for low resource settings. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 706–719, Suzhou, China. Association for Computational Linguistics.

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.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Further Discussions

Why *Indic* languages? Indic languages are spoken by more than a billion people in the Indian subcontinent. With the introduction of IndicNLPSuite (Kakwani et al., 2020) by AI4Bharat[4] there has been has an increased interest and effort towards the research for *Indic* languages model. Recently, IndicBERT, MuRIL (Khanuja et al., 2021) based on BERT (Devlin et al., 2019) were introduced for the *Indic* languages. Furthermore, generation model IndicTrans (Ramesh et al., 2022) and IndicBART (Dabre et al., 2022) based on seq2seq architecture was also published recently. These model use the *Indic* enrich monolingual corpora: Common Crawl, Oscar and IndicCorp and parallel corpora: Samantar and PMIndia (Haddow and Kirefu, 2020) on *Indic* languages for training. Despite significant progress through large transformer-based *Indic* language models in addition to existing multilingual models e.g. mBERT (Devlin et al., 2019) , XLM-RoBERTa (Conneau et al., 2020a), and mBART (seq2seq) (Liu et al., 2020) there is currently a paucity of benchmark data-sets for evaluating these huge language models in the *Indic* language research field. Such benchmark dataset is necessary for studying the linguistic features of Indic languages and how well they are perceived by different multilingual models. Recently, IndicGLUE (Kakwani et al., 2020) was introduced to handle this scarcity. However, the scope of this benchmark, is confined to only few tasks and datasets.

Why INDICXNLI task? This research provides an excellent chance to investigate the efficacy of various Multilingual models on *Indic* languages that are rarely evaluated or explored before. Some of these *Indic* languages such as *'Assamese'* and *'Odia'* serve as unseen (zero-shot) evaluation for models such as mBERT (Pires et al., 2019), i.e. not pre-trained on *'Assamese'*. While other models, such as XLM-RoBERTa, IndicBERT and MuRIL covers all our languages but in widely varying proportions in their training data. Our work investigate the correlation effect of cross-lingual training

[4] https://ai4bharat.org

for English on these rare *Indic* languages, which are not explore by prior studies. Furthermore, we also investigate the cross-lingual transfer effect across *Indic* languages, also not explored before. We explore the impact of Multilingual training, english-data augmentation, unified Indic model performance, cross-lingual transfer of closely related *Indic* family and English-*Indic* NLI through our work. All the above mention topics are not explore for *Indic* language before. We aim to integrate INDICXNLI and benchmark models in IndicGLUE (Kakwani et al., 2020). Such a benchmark dataset is required for investigating the linguistic properties of Indian languages and how accurately they are interpreted by various multilingual models. Another direction is accessing model performance on INDIC-INDICXNLI task, where both premises and hypothesis are in two distinct *Indic* languages.

Why IndicTrans for Translation? We use the IndicTrans as a translation model for converting English XNLI to INDICXNLI because of the following reasons: (a.) **Open-Source:** IndicTrans is open-source to public for non-commercial usage without additional fees, while Google-Translate and Microsoft-Translate require a paid subscription. (b.) **Light Weight:** IndicTrans is the fastest and the lightest amongst mBART and mT5 on single GPU machines. Google-Translate and Microsoft-Translate are also relatively slower due to repeated network-intensive API calls. (c.) ***indic* Coverage:** Seq2Seq models like mBART and mT5 are not designed for all languages in the indic family. mBART supports seven (excludes kn,or,pa,as) while mT5 supports nine languages (excludes or,as) out of eleven indic languages. Google-Translate supports ten out of eleven *indic* languages (excludes Assamese). Microsoft Translate supports all the eleven *indic* languages. In future, we plan to enhance INDICXNLI with better translation.

B Human Validation Scoring

We provide English and *indic* language INDICXNLI (IndicTrans translated) sentence to the recruited native speaker of that *indic* language for validation. Before the annotation work, each expert was given a full explanation of the guidelines that needed to be followed. The validation instructions (mturk template and detailed examples) are taken from the Semeval-2016 Task-I. The native speaker access the sentence pairs assign an integer score between **0** and **5**, as follows: **0**: The two sentences are completely dissimilar. **1**: The two sentences are not equivalent, but are on the same topic. **2**: The two sentences are not equivalent, but share some details. **3**: The two sentences are roughly equivalent, but some important information differs/missing. **4**: The two sentences are mostly equivalent, but some unimportant details differ. **5**: The two sentences are exactly equivalent, as they mean the same. The score depicts the goodness of translated sentence in terms of semantics, i.e. same meaning as original English sentence[5]. Scores are then normalized to a probability range (between 0 and 1). The final validation score for each language is determined as the average of all 100 instances' scores.

Additionally, we also computed the BERTScore between the English and the Hindi test split of the XNLI[6], using multi-lingual strategy which came out to be 70 ($\times 10^{-}2$). We presume that the lower score is attributable to the fact that human-translated dataset encapsulates a large number of linguistic nuances, resulting in a change in the structure and tonality of the sentences, which is frequently overlooked by machine translation systems, as highlighted by Bianchi et al. (2022).

C Details: Multi-lingual Models

***Indic* Specific:** These models are specially pre-trained using Mask Language Modeling (MLM) or Translation Language Model (TLM) (CONNEAU and Lample, 2019) on monolingual / bilingual *Indic* language corpora. These include models such as MuRIL and IndicBERT trained on 17 and 11 *Indic* languages (+English) respectively. MuRIL is pre-trained using Common-Crawl Oscar Corpus (Ortiz Su'arez et al., 2019), PMIndia (Haddow and Kirefu, 2020) on the following languages: *en, hi, bn, gu, te, ta, or, ml, pa, kn, mni, as, ur*. IndicBERT is pre-trained using *Indic*-Corp(Kakwani et al., 2020) on the following languages: *en, hi, bn, ta, ml, te, Mr, kn, gu, pa, or, as*. Moreover, MuRIL is also pre-trained with TLM objective (with MLM objective) on machine translated data and machine transliterated data.

***Generic*:** These are massive multi-lingual models pre-trained on large number of languages with MLM. These include multi-lingual BERT i.e. mBERT (cased/uncased) and multi-lingual RoBERTa i.e. XLM-RoBERTa which are trained on more than 100 languages. XLM-RoBERTa also

[5] For NLI task, same syntax, i.e. grammar (e.g. Tense) lesser important than same Semantic, i.e. meaning preservation.

[6] XNLI hindi test splits was human translated.

includes pre-training on all eleven *Indic* languages. XLM-RoBERTa is pre-trained using the common crawl monolingual data. mBERT (cased/uncased) includes pre-training on nine of eleven *Indic* languages (Assamese and Odia excluded) and uses multi-lingual Wikipedia data for pre-training.

D Details: Hyper Parameters Settings

All the models were trained on google collaboratory [7] on TPU-v2 with 8 cores. The code was built in the PyTorch-lightning framework. We used accuracy as mentioned in the original XNLI paper (Conneau et al., 2018) as our metric of choice. The training was run with an early stopping callback with the patience of 3, validation interval of 0.5 epochs and AdamW as optimizer(Loshchilov and Hutter, 2019). In Table 5 the hyperparamaters are abbreviated as mentioned below: (a.) **PO:** Pre-training Objective. Where MLM stands for masked Language Modelling, TLM stands for Translation Language Modelling and TrLM stands for Transliteration Language Modelling, (b.) **CU:** Corpus Used, (c.) **LR:** Learning Rate, (d.) **BS:** Batch Size, (e.) **WD:** Weight Decay, (f.) **MSL:** Maximum Sequence Length, (g.) **MS:** Model Size described as number of parameters in millions, (h.) **WS**: Warm-up Step.

E Indic Cross-lingual Transfer

Table 6 (extension §3) are the cross-lingual transfer results of XLM-R, IndicBERT, mBERT and MuRIL respectively. The rows of the table consist of the languages on which the model is trained, while the columns represent the evaluation languages. E.g., in table 6 the first row represents that the model is trained on *"as"* and then tested on all the languages in the column. The values in the row are the accuracy scores of the model when trained on the language in its leftmost column and tested on the language in its top-most row column.

XLM-R. the model perform best for the *"bn"* language. The model gives the best performance average across all other languages if trained on *"bn"*. A model trained in other languages, on average, also performs best for *"bn"* language. XLM-R also struggles to correlate with *"kn"*, *"or"*, and *"ml"*, thus performs poorly on average if trained for them. At the same time, all models have poor cross-lingual ability transferability for the *"as"* language.

IndicBERT. the overall score is comparable to XLM-R despite it's smaller size. On average, across languages, the cross-lingual transfer ability for models trained on varying *indic* languages were consistently similar (b/w 0.5-0.6). However, the evaluation performance for cross-lingual models evaluated on *"ml"* were poor for all *indic* trained models. For model trained on some languages, *"kn"*, *"ml"* and *"pa"*, the best performance was across diagonal, i.e. indicating the model performs best on the trained language. This trend was, however, was not shown in other *indic* languages, indicating remarkable cross-lingual transfer ability of the IndicBERT model.

mBERT. the model performs worse for **"or"** on average for both when evaluated and train on. However, all models performs very consistently for other *indic* languages. Model trained on *kn*, *pa*, *ta*, *hi*, and *bn* perform best on average across languages. Here too, the best cross-lingual transfer ability was shown for *bn* language. mBERT also have best performance across diagonal for some languages e.g. *"as"*, *"gu"*, *"ml"*, *"pa"* and *"te"*.

MuRIL. shows the best overall cross-lingual transfer ability amongst all the models. MuRIL only fails to generalize well when trained for *"or"* language. However, model train on other *indic* language when evaluated on *"or"* performs well. Model trained on *"ta"* and *"ml"* performs best across all languages. The best cross-lingual transfer ability was shown for *"bn"* and *"hi"*. Overall, MuRIL has better cross-lingual transfer ability across all languages compared to other models. It also shows less performance bias for languages such as *"bn"* and *"hi"*, as compared to XLM-R.

F Intra-Bilingual Inference

We observed a performance loss except for XLM-RoBERTa when the model is evaluated on EN-INDICXNLI inference task. The inference models struggle to correlate and reason together on two different languages (English, *Indic*) sentences. Contrary to earlier observation, a generic model such as XLM-RoBERTa outperforms the *Indic* specific models. However, IndicBERT and MuRIL perform better than mBERT. *Bengali* perform best for both the training strategies. We also observe the benefit of English data augmentation **English + *Indic* Train** model, rather than all language augmentation **Train All** model.

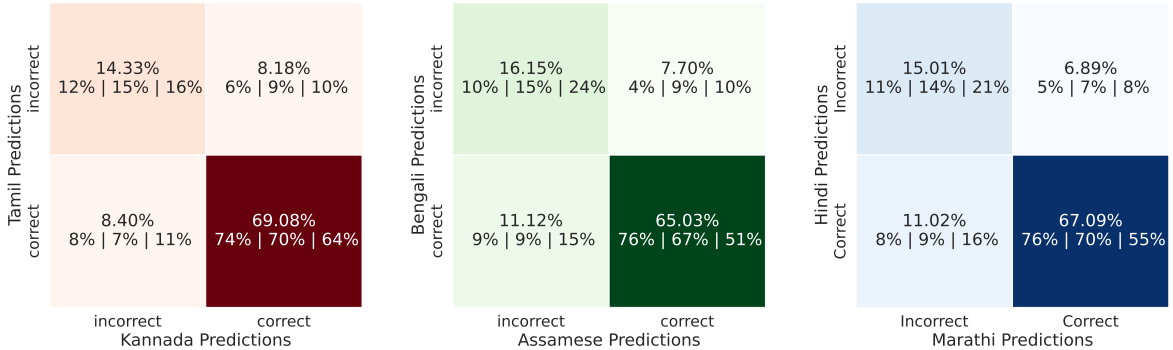
[7] https://colab.research.google.com/

Model	PO	CU	LR	BS	WD	MSL	MS	WS
XLM-R	MLM (Dynamic)	Wikipedia Corpus	2e-5	64	0.01	128	278M	1500
iBERT	MLM	IndicCorp	2e-5	128	0.01	128	33.7M	1500
MuRIL	MLM, TLM and TrLM	OSCAR and PM India	2e-5	64	0.01	128	237M	1500
mBERT	MLM	Wikipedia Corpus	2e-5	128	0.01	128	177M	1500

Table 5: Model Hyper-Parameters

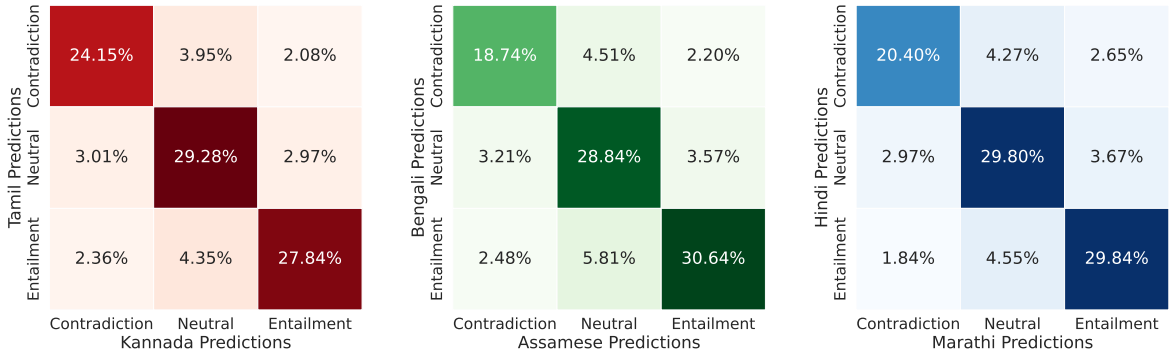
TrLang	XLM-RoBERTa											TrAvg	IndicBERT											TrAvg
	as	gu	kn	ml	mr	or	pa	ta	te	bn	hi		as	gu	kn	ml	mr	or	pa	ta	te	bn	hi	
as	64	67	66	67	63	63	**68**	**68**	64	66	65	66	65	63	54	46	61	60	66	48	57	**67**	60	58
gu	65	72	69	69	68	71	70	71	65	**74**	**74**	70	61	67	54	41	65	64	**70**	46	62	**70**	62	**60**
kn	33	31	**35**	**35**	31	34	32	31	32	33	32	33	58	64	**68**	48	59	59	65	46	59	63	63	59
ml	**35**	33	33	34	31	34	34	31	33	34	34	33	55	52	54	**60**	53	53	52	52	57	52	52	54
mr	66	74	70	72	72	68	70	69	65	**75**	73	71	62	65	54	48	61	61	67	52	60	**68**	63	**60**
or	35	33	32	36	35	34	34	**36**	34	**36**	**36**	35	61	66	57	49	61	66	65	48	60	**68**	64	**60**
pa	65	69	70	67	67	67	70	66	67	**73**	66	68	61	67	55	47	60	62	**74**	41	60	70	62	**60**
ta	64	67	69	72	71	68	71	70	70	**73**	70	70	55	**60**	53	49	56	54	58	59	55	58	55	56
te	61	70	71	70	70	71	68	68	**75**	**75**	72	71	61	63	53	45	59	63	**70**	46	63	68	58	59
bn	67	72	73	73	72	**74**	**74**	70	70	73	71	**72**	62	66	55	48	62	62	66	47	60	**68**	**68**	**60**
hi	66	70	69	72	69	68	71	71	71	**76**	73	71	58	63	53	49	61	61	66	43	57	**71**	61	59
TestAvg	56	60	60	61	59	59	60	59	59	**63**	61	60	60	63	55	48	60	60	65	48	59	**66**	61	59
TrLang	**mBERT**											**TrAvg**	**MuRIL**											**TrAvg**
	as	gu	kn	ml	mr	or	pa	ta	te	bn	hi		as	gu	kn	ml	mr	or	pa	ta	te	bn	hi	
as	**69**	59	61	53	57	36	61	57	52	59	64	56	73	**78**	75	74	74	73	75	75	75	76	77	75
gu	48	**70**	64	55	60	32	64	64	60	67	65	60	72	75	75	74	73	72	70	72	71	**76**	75	73
kn	49	62	68	64	60	35	65	64	59	**69**	62	**61**	72	75	76	76	73	73	74	75	76	**77**	**77**	75
ml	51	60	60	**71**	60	30	61	65	62	66	62	60	75	75	73	77	72	78	76	**79**	75	77	76	**76**
mr	45	61	63	56	69	35	64	56	57	**69**	66	60	69	70	72	71	**73**	68	76	70	69	73	74	72
or	34	33	29	32	36	35	34	35	33	33	34	33	33	36	35	30	32	**35**	30	30	33	32	36	33
pa	47	65	59	59	62	35	**70**	63	61	68	64	**61**	73	75	**76**	74	74	76	79	71	74	75	75	75
ta	48	64	**67**	63	60	32	65	66	63	69	62	**61**	74	76	76	77	75	72	74	77	76	**80**	78	**76**
te	51	59	63	63	60	32	61	64	**67**	66	62	60	70	72	74	71	73	70	**77**	74	**77**	**77**	75	74
bn	51	64	65	62	62	32	65	60	62	69	**67**	**61**	68	**76**	73	73	71	72	73	74	74	74	**76**	74
hi	50	66	65	61	62	30	65	63	61	**71**	63	**61**	73	76	73	75	74	73	76	74	74	75	**76**	75
Test Avg	49	60	60	58	59	33	61	60	58	**64**	61	58	68	71	71	70	69	69	71	70	70	**72**	**72**	71

Table 6: Indic Cross-lingual transfer



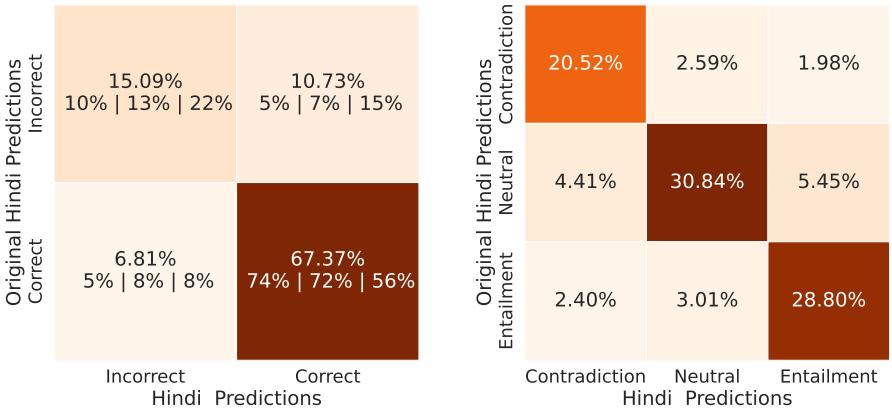
(a) Tamil vs Kannada (b) Bengali vs Assamese (c) Hindi vs Marathi

Figure 1: Consistency Matrix: Predictions of MuRIL for (a) Tamil vs Kannada (b) Bengali vs Assamese, , (c) Hindi vs Marathi. The percentage on top in each block represents the average across all three labels with each label percentage given below it in the order of Entailment, Neutral and Contradiction. (cf. Appendix § 4)



(a) Tamil vs Kannada (b) Bengali vs Assamese (c) Hindi vs Marathi

Figure 2: Confusion Matrix: for MuRIL (a) Tamil vs Kannada, (b) Bengali vs Assamese, (c) Hindi vs Marathi.



(a) Confusion Matrix (b) Consistency Matrix

Figure 3: Consistency Matrix and Confusion Matrix for Predictions of MuRIL on Original Hindi data in XNLI and Machine Translated Data generated from IndicTrans.