IndicXNLI: Evaluating Multilingual NLI for Indic Languages

Divyanshu Aggarwal[1], Vivek Gupta[2],
Anoop Kunchukkutan[3,4]

[1]Delhi Technological University; [2]University of Utah; [3]AI4Bharat; [4]Microsoft India

Natural Language Inference Task

Premise	Hypothesis	Label
They told me that, uh, that I would be called in a guy at the end for me to meet.	I was never told anything about meeting anyone.	Contradiction
They told me that, uh, that I would be called in a guy at the end for me to meet.	We had a great talk.	Entailment
They told me that, uh, that I would be called in a guy at the end for me to meet.	The guy showed up a bit late.	Neutral

Natural Language Inference Task

Premise	Hypothesis	Label
They told me that, uh, that I would be called in a guy at the end for me to meet.	I was never told anything about meeting anyone.	Contradiction
They told me that, uh, that I would be called in a guy at the end for me to meet.	We had a great talk.	Entailment
They told me that, uh, that I would be called in a guy at the end for me to meet.	The guy showed up a bit late.	Neutral

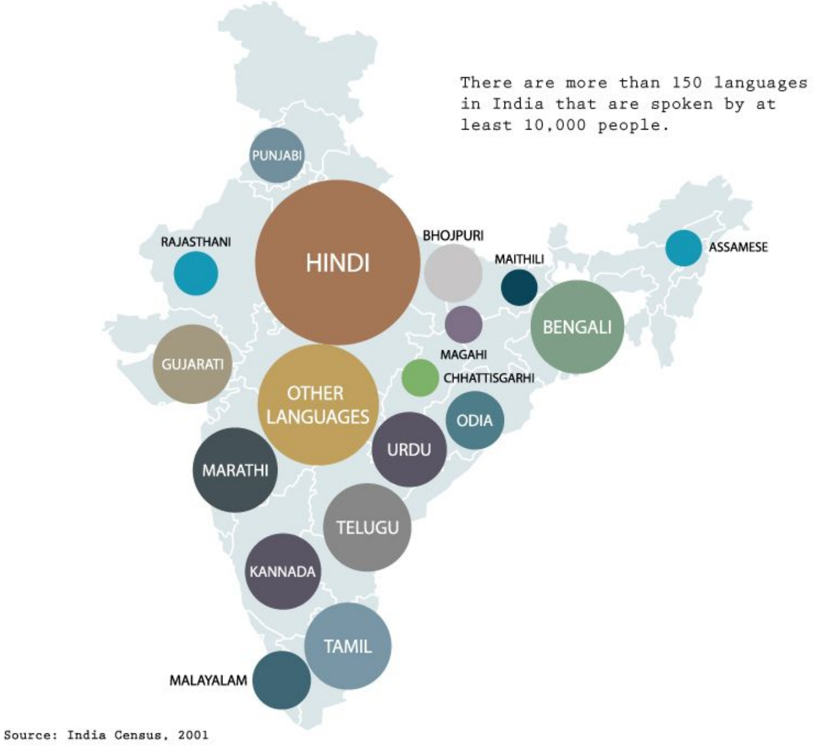
➔ IndicXNLI is an NLI dataset but for Indic Languages.

Motivation

- Indian Languages are a diverse yet closely related set of languages which are spoken by more than billion people in the world in the south asian region.

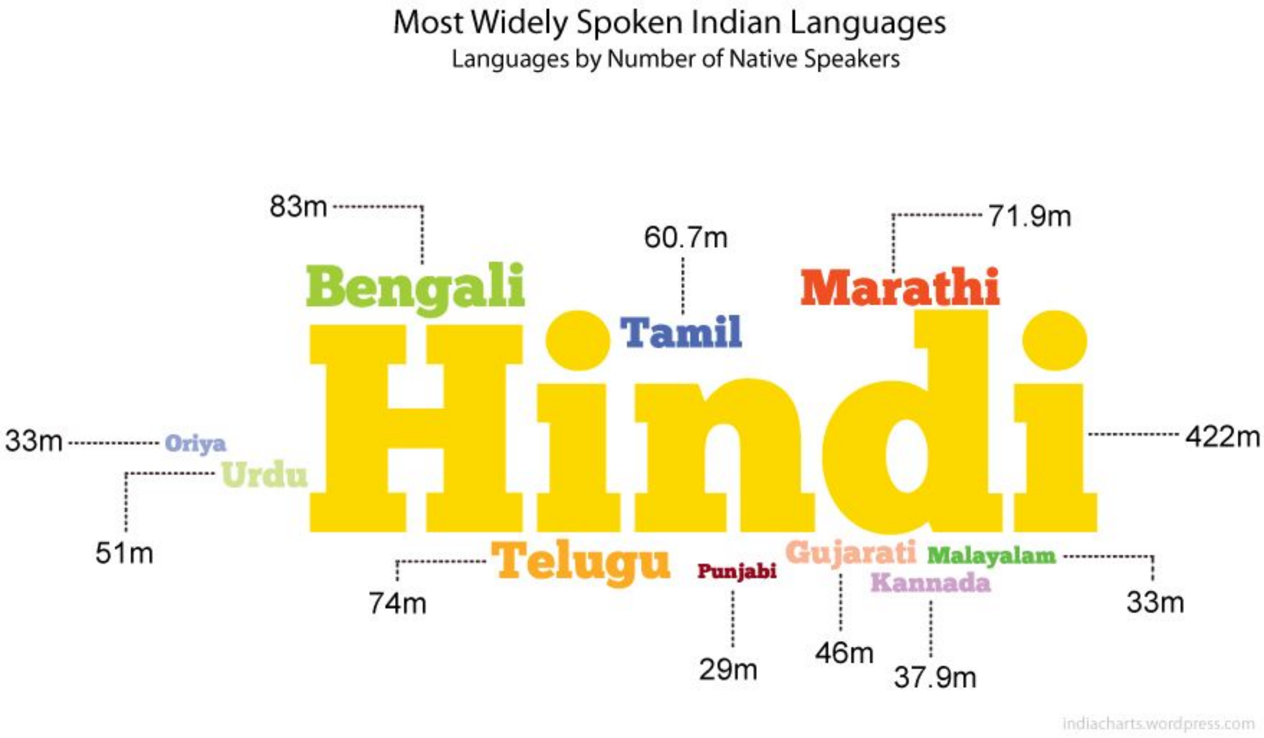
- They are also one of the largest set of internet users in the world who can leverage the current advancements in NLP in their native languages.

- There has been significant advancements in indic specific resources (e.g. IndicCorp) and transformers models (IndicBERT, IndicBART etc), we still lack good quality benchmarks due to lack of expert annotators in these languages.



Reference: Wikipedia

Speakers of Indian Languages



Reference: India charts

Premise & Challenges

Premise

- Can we create a high quality NLI dataset with minimal human supervision?
- Can we leverage current translation resources and generate a high quality NLI dataset for Indic Languages?
- How well can current pre-trained multilingual language models reason on IndicXNLI?

Premise & Challenges

Premise

- Can we create a high quality NLI dataset with minimal human supervision?
- Can we leverage current translation resources and generate a high quality NLI dataset for Indic Languages?
- How well can current pre-trained multilingual language models reason on IndicXNLI?

Challenges

- Lack of benchmarking techniques for machine translation without reference text.
- Lack of fluent Indic and English bilingual speakers.
- How to verify meaning preservation in translated sentences to preserve inference labels?

Our Contributions

- We created **IndicXNLI** which is a high quality **NLI dataset** created by translating the english **XNLI dataset** to indic languages Using **IndicTrans**.

- We verified the quality using **automatic scoring** techniques using **BertScore** and low cost **human evaluation** using **diverse sampling**.

- We asses various training strategies on various state of the art **indic specific** and **multi-lingual** language models over **IndicXNLI**.

Why IndicTrans for Machine Translation?

Open Source

It is open source with an MIT License making it free for access for research and non-commercial use.

Why IndicTrans for Machine Translation?

Open Source

It is open source with an MIT License making it free for access for research and non-commercial use.

Light Weight

Despite being a 4x transformer model it is still lighter than mBART and mT5 with full indic coverage.

Why IndicTrans for Machine Translation?

Open Source

It is open source with an MIT License making it free for access for research and non-commercial use.

Light Weight

Despite being a 4x transformer model it is still lighter than mBART and mT5 with full indic coverage.

Indic Coverage

IndicTrans covers all 11 major Indic languages which are only covered by azure translate other than IndicTrans.

Azure translate is not free for research.

Automatic Evaluation

English Translated (Round Trip)

- Capture similarity between Back translated english sentence and original english sentence.

- We used BertScore to compare back translated and original english sentence.

- We compared google translate and IndicTrans where IndicTrans performed better

Automatic Evaluation

English Translated (Round Trip)

- Capture similarity between Back translated english sentence and original english sentence.
- We used BertScore to compare back translated and original english sentence.
- We compared google translate and IndicTrans where IndicTrans performed better.

Multilingual (Single Trip)

- Capture similarity between forward translated indic sentence and original english sentence.
- We used BertScore with mBERT as base model to compare forward translated Indic sentence and original english sentence.
- We compared google translate and IndicTrans where IndicTrans performed better.

Automatic Evaluation Scores

BertScore	hi	te	pa	bn	as	gu	ta	ml	kn	mr	or
English translated (Google Translate)	94	93	92	94	NA	94	94	94	94	94	94
English translated (IndicTrans)	98	94	94	98	93	94	94	94	94	93	93
Multi Lingual (Google translate)	90	88	86	89	NA	89	86	85	88	87	82
Multi Lingual (IndicTrans)	96	87	88	96	85	96	87	87	87	86	86

Table 1: Automatic Evaluation Scores Using BertScore ($X10^{-2}$)

We observed that IndicTrans fairs better than google translate in our automatic evaluation setup.

Human Evaluation

Problem

It is both time consuming and expensive to get all 10,000 samples evaluated.

Furthermore, it require expert fluent speakers in all 11 Indic languages and English.

Human Evaluation

Problem

It is both time consuming and expensive to get all 10,000 samples evaluated.

Furthermore, it require expert fluent speakers in all 11 Indic languages and English.

Solution

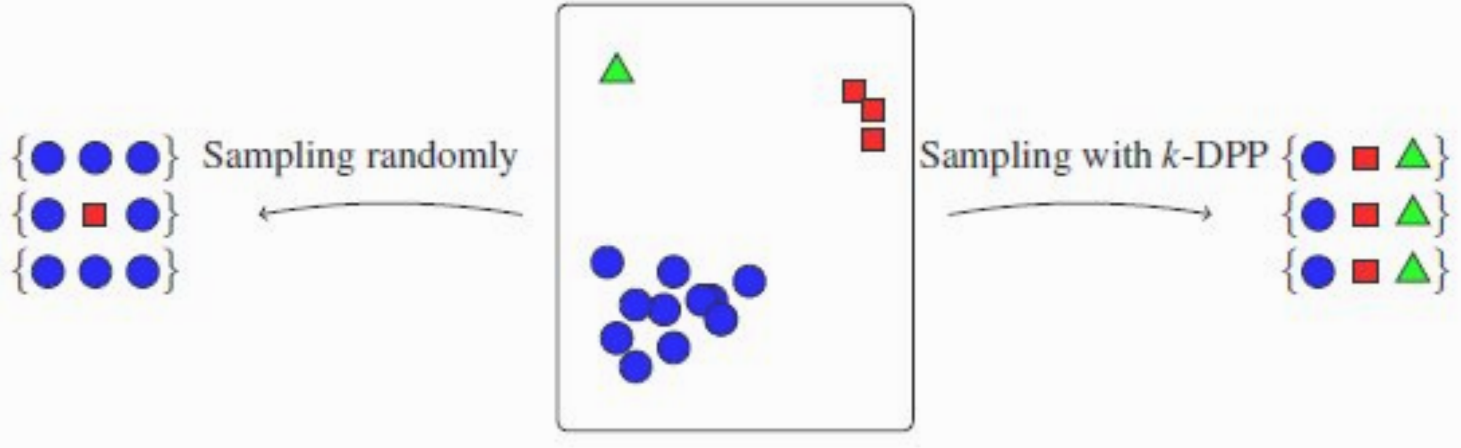
Sample a relatively small diverse set (~100 samples) of examples with maximum coverage in the test set.

Human Evaluation

Problem	Solution	Method
It is both time consuming and expensive to get all 10,000 samples evaluated. Furthermore, it require expert fluent speakers in all 11 Indic languages and English.	Sample a relatively small diverse set (~100 samples) of examples with maximum coverage in the test set.	Sampled 50 sentences from the bert embeddings of the test set using dppy library[1] i.e. **DPP** Added the premise of hypothesis and hypothesis of premise obtained from DPP Sampling. Increasing our sample count to 100.

[1] https://github.com/guilgautier/DPPy

Diverse Sampling: What and Why?



K-DPP Process (Reference: Disney Research Studios)

Human Score Labelling

- 22 evaluators (2 for each language),
- fluent in both english and Indic mother tongue
- use Semeval-2016 Task-I guidelines[2].
- 5 Indian Rupees per sentence.

[2] https://web.eecs.umich.edu/~mihalcea/papers/agirre.semeval16.pdf

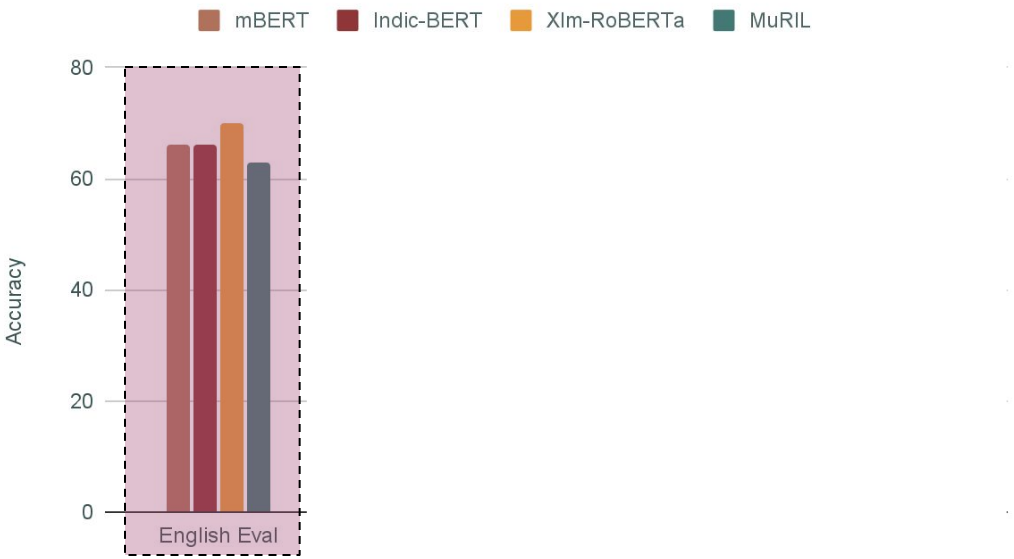
Human Evaluation Scores

Score	hi	te	pa	bn	as	gu	ta	ml	kn	mr	or
Human Score 1	88	88	91	87	87	89	89	87	89	86	88
Human Score 2	81	84	93	83	84	89	87	87	87	87	90
Pearson Correlation	73	73	89	79	78	79	76	85	83	83	75
Spearman Correlation	82	87	94	90	88	85	88	93	86	89	85

Table 2: Human Validation Score ($X10^{-2}$)

There is reasonably high pearson and spearman correlation between the 2 annotators, attesting to the quality of IndicXNLI.

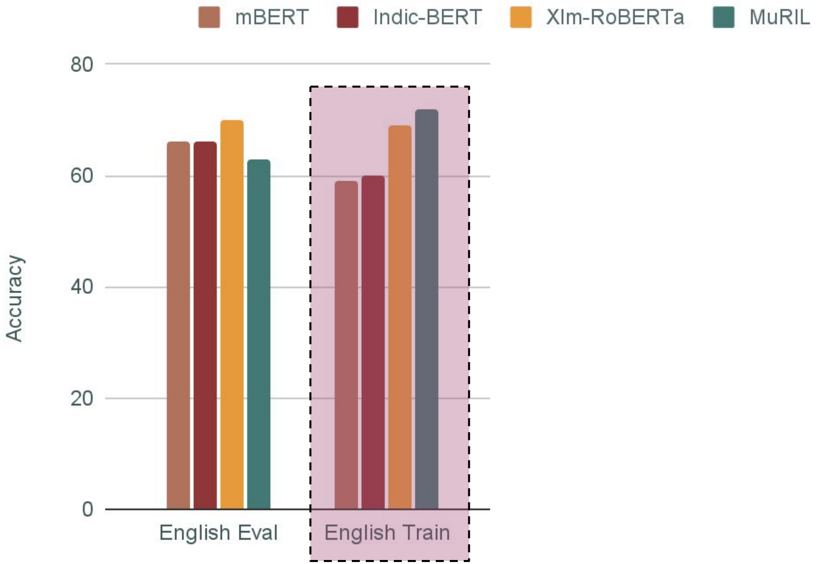
Results and Analysis



English Eval

- The model are trained on original English XNLI train data.
- The model is evaluated on English translation of INDICXNLI test data.
- This Translate-Test Scenario

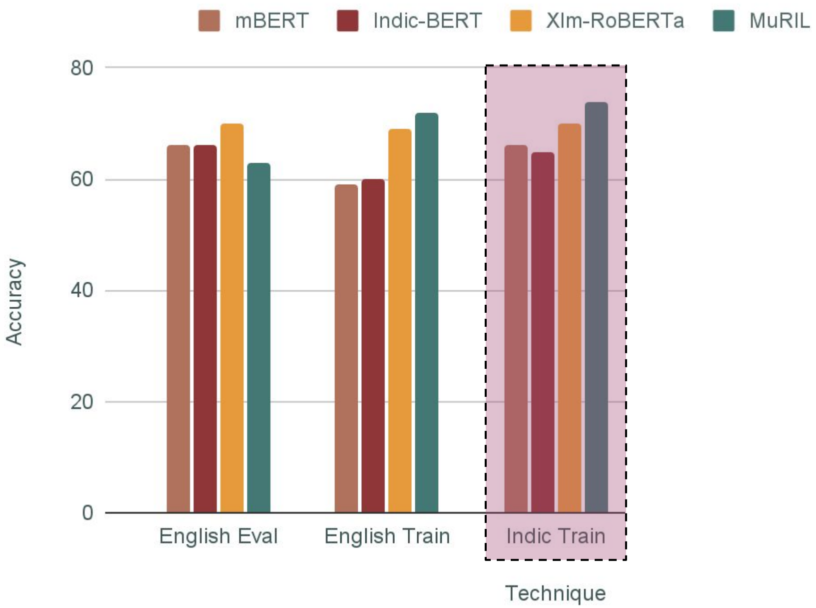
Results and Analysis



English Train

- The model are trained on original English XNLI train set data.
- The model is evaluated on INDICXNLI test set data.
- This is a zero-shot evaluation training scenario.

Results and Analysis

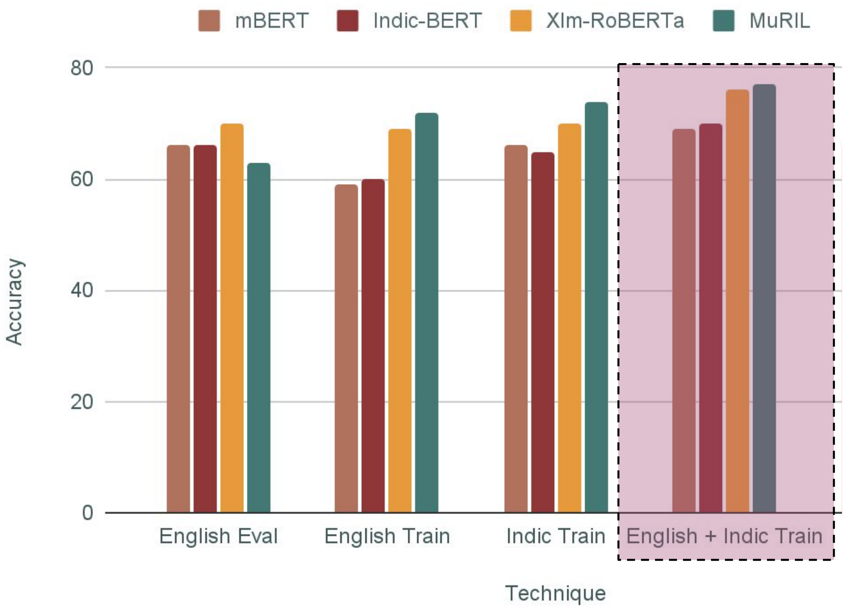


Indic Train[4]

- The model are trained on IndicXNLI train set data.
- The model is evaluated on INDICXNLI test set data.
- This is Translate-Train Scenario

4 We also tested models trained on this technique on all other indic languages. You can find it Indic Cross lingual Transfer Section.

Results and Analysis



English + Indic Train

- The model is first finetuned on English data of XNLI and then finetuned on Indic data of IndicXNLI.
- The model is evaluated on INDICXNLI test set data.

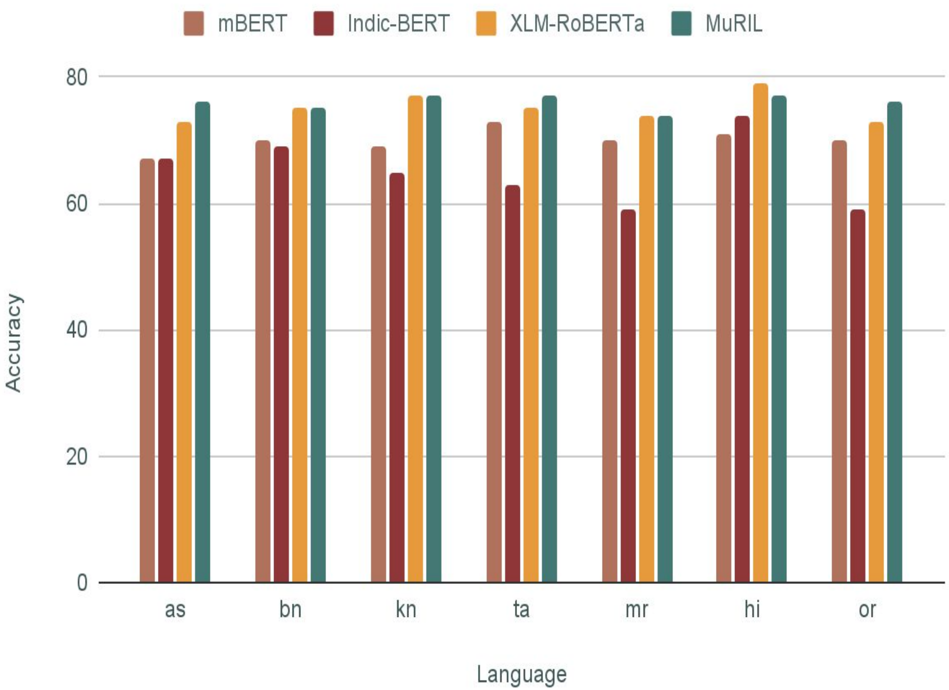
Results and Analysis

Train All[3]

- This approach begins by finetuning the model on English XNLI data, followed by training on all eleven Indic languages of INDICXNLI sequentially.
- The model is evaluated on INDICXNLI test set data.
- This is Train-all scenario.

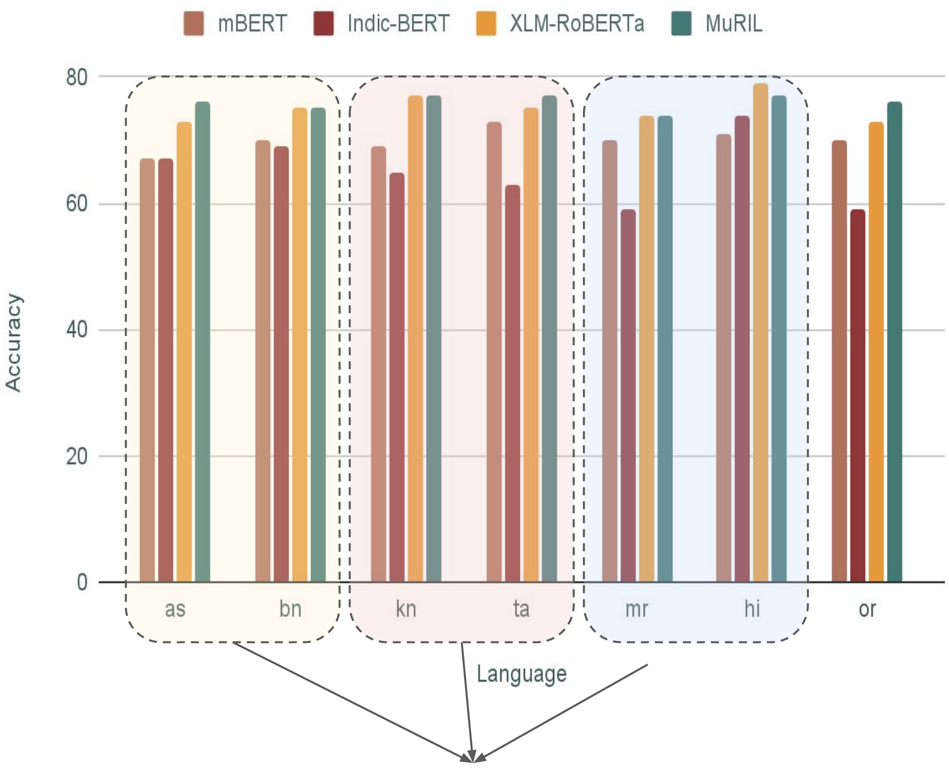
3 We also fine-tuned models on intra-bilingual NLI task which contains mixed language input. More on it is given in the paper.

Language Wise Comparison



- MuRIL > other models
- High Resource language > Mid resource language >> Low Resource Language
- Low resource languages with similar script to high resource languages perform well.
- XLM-RoBERTa >> IndicBERT. Despite IndicBERT's indic specific training.
    - Size and Languages used in pre training >> Indic specific pre training.

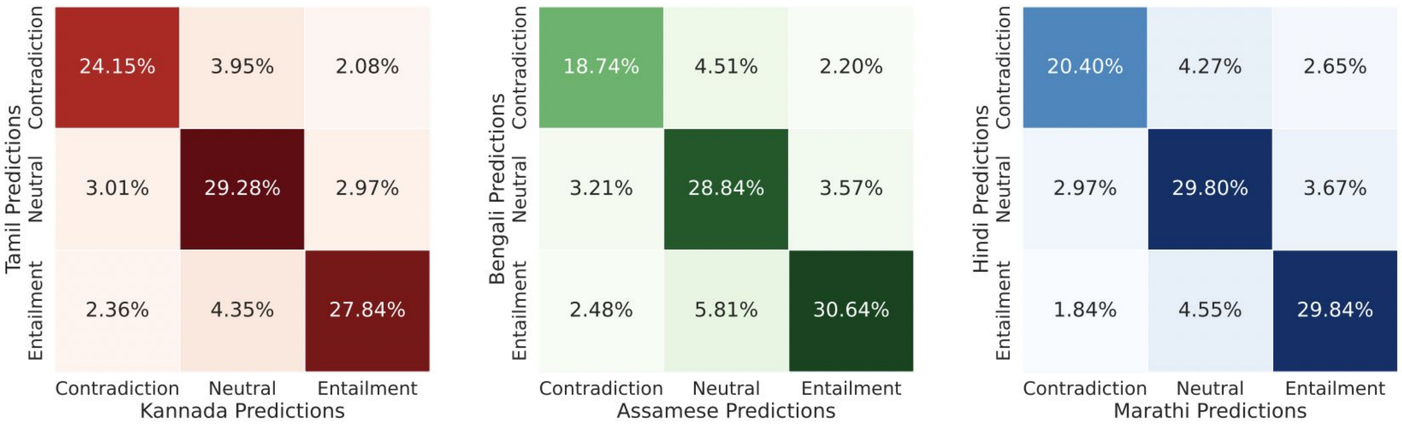
Language Wise Comparison



Pairs of languages with similar script

- MuRIL > other models
- High Resource language > Mid resource language >> Low Resource Language
- Low resource languages with similar script to high resource languages perform well.
- XLM-RoBERTa >> IndicBERT. Despite IndicBERT's indic specific training.
  - Size and Languages used in pre training >> Indic specific pre training.

Error Analysis



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

- Similar languages predicts similarly regardless of resource variability.
- Low resource languages which are similar to High Resource Languages (in terms of Script) performs as good as high resource languages.
- Similar languages usually agree better upon entailment / contradiction difference as compared to neutral / contradiction and neutral / entailment difference.

Key Takeaways

- With **IndicXNLI** we extend the **XNLI dataset** for **Indic languages family.**

- We Evaluate the quality of our dataset with various **automatic** and **human evaluation** techniques which are **less expensive and time consuming**.

- 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**.
    - **Future Work:** Accessing model performance on **INDIC-INDIC XNLI task**, where both premises and hypothesis are in two distinct Indic languages.