IndicXNLI: Evaluating Multilingual NLI for Indian Languages

Divyanshu Aggarwal[1*] , Vivek Gupta[2] , Anoop Kunchukuttan[3,4]
[1]Delhi Technological University; [2]University of Utah; [3]AI4Bharat;
[4]Microsoft India

1. 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.

XNLI → IndicXNLI
en → hi, bn, as, gu, pa, kn, or, ta, te, mr, ml

2. 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.

3. Challenges and Premise

Premise

- Can we create a high quality NLI dataset for 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 resources 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?

4. 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 of **IndicXNLI** using **automatic scoring** techniques like **BertScore** and low cost **human evaluation** using **diverse sampling**.
- We asses various training strategies on various state of the art and **indic specific** and **multi-lingual** language models over **IndicXNLI**.

5. Why Indic Trans?

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.

6. 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.

Method

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.

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

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

7. 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.

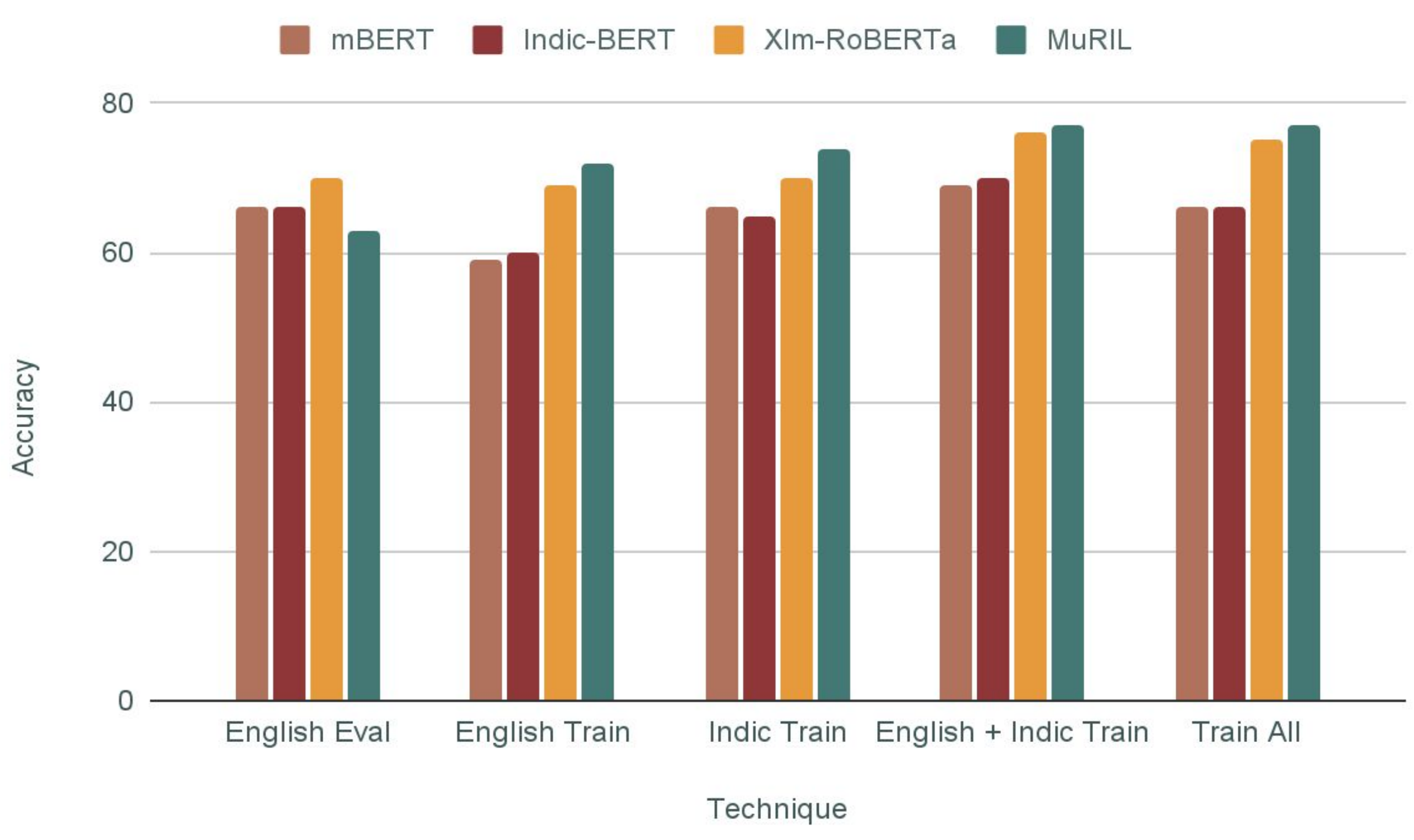
Score	hi	te	pa	bn	as	gu	ta	ml	kn	mr	or
Human Score 1	88	88	91	87	87	89	89	87	89	86	88
Human Score 2	81	84	93	83	84	89	87	87	87	87	90
Pearson Correlation	73	73	89	79	78	79	76	85	83	83	75
Spearman Correlation	82	87	94	90	88	85	88	93	86	89	85

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

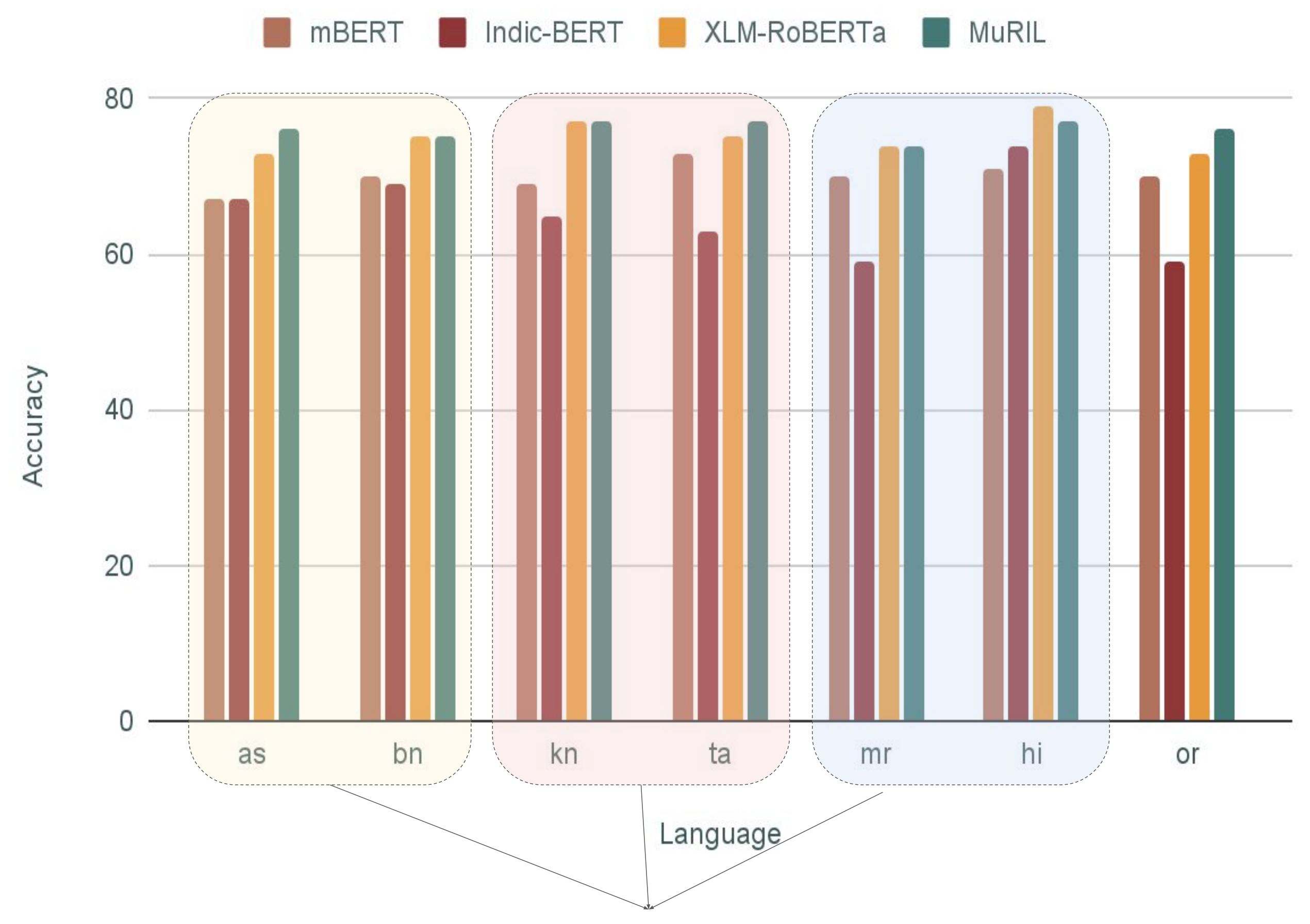
There is reasonably high pearson and spearman correlation between the 2 annotators, attesting to the quality of IndicXNLI.

8. Model Wise Analysis

MuRIL → best model
English+Indic Train performs best.
XLM-RoBERTa >> IndicBERT. Despite IndicBERT's indic specific training.
Size and Languages used in pre training >> Indic specific pre training.

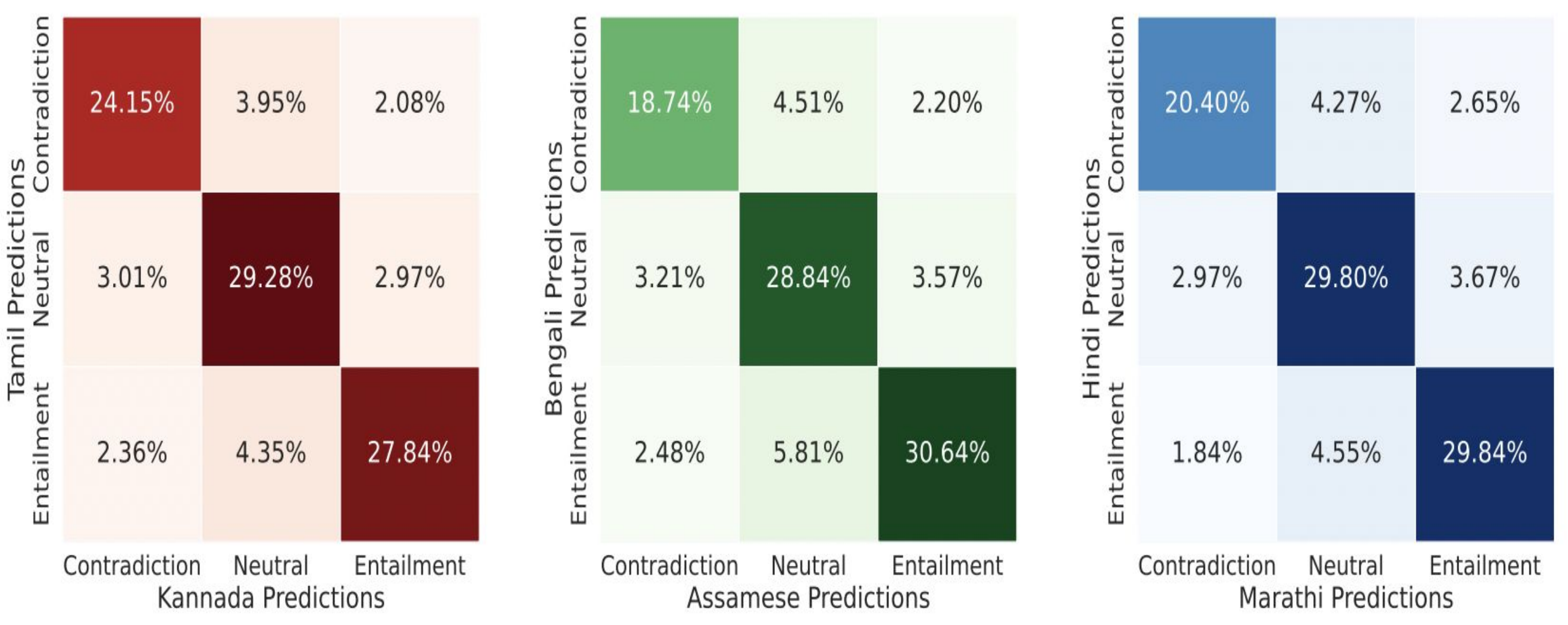


Similar Language perform similarly.
Model Size » Indic Specific Pretraining.
High Resource language > Mid resource language > Low

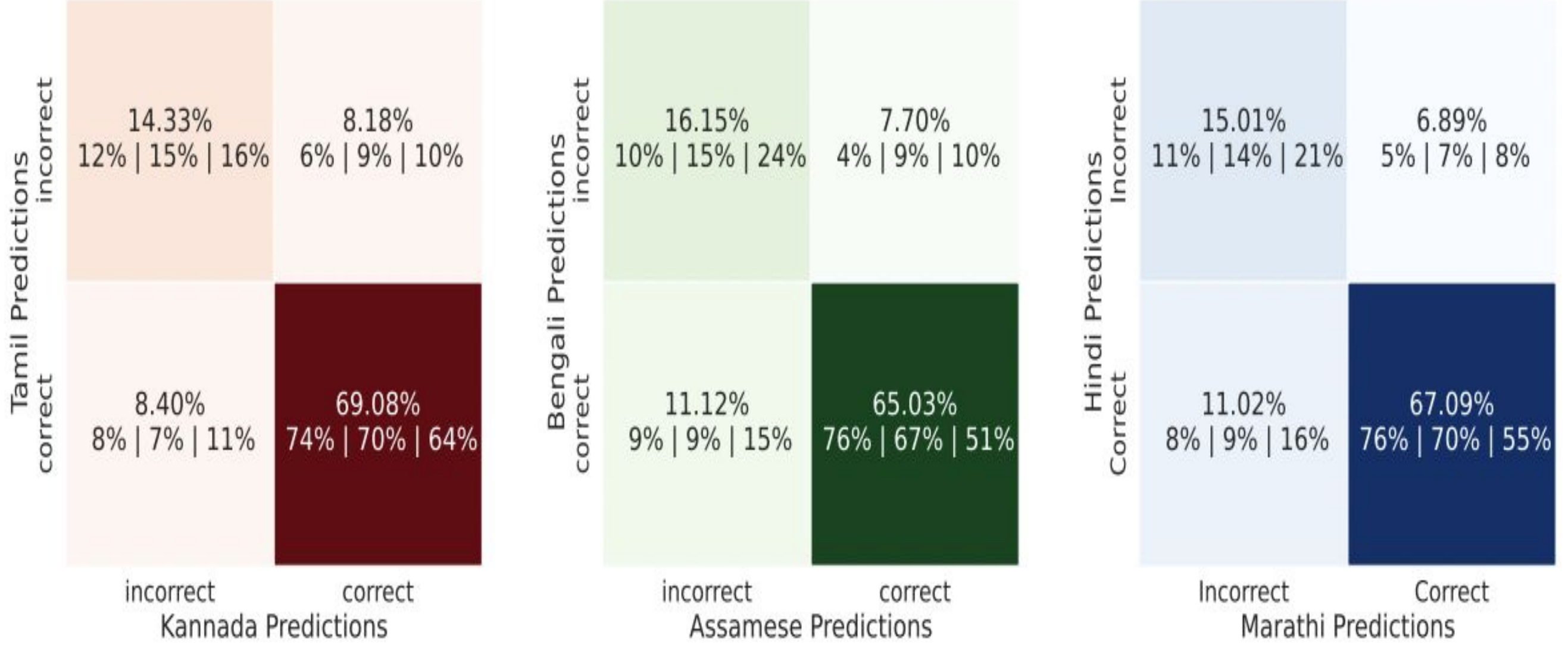


Pairs of languages with similar script

9. Error Analysis



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



(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.

Website:

https://indicxnli.github.io/