Natural Language Understanding & Generation

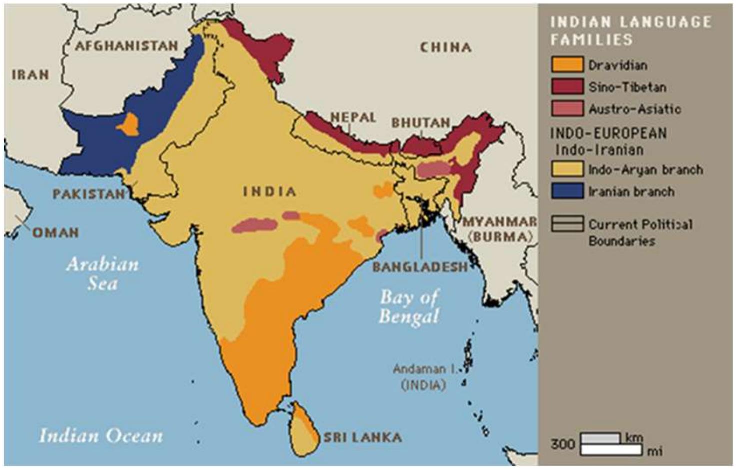
https://ai4bharat.org/language-understanding

https://ai4bharat.org/language-generation

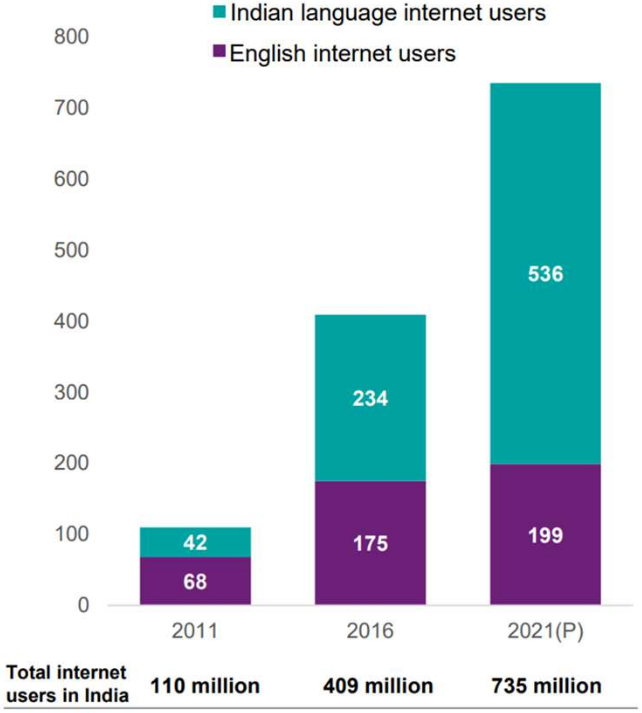
Workshop on 28th July 2022, IIT Madras

https://github.com/AI4Bharat/workshop-nlp-nlu-2022

Usage and Diversity of Indian Languages



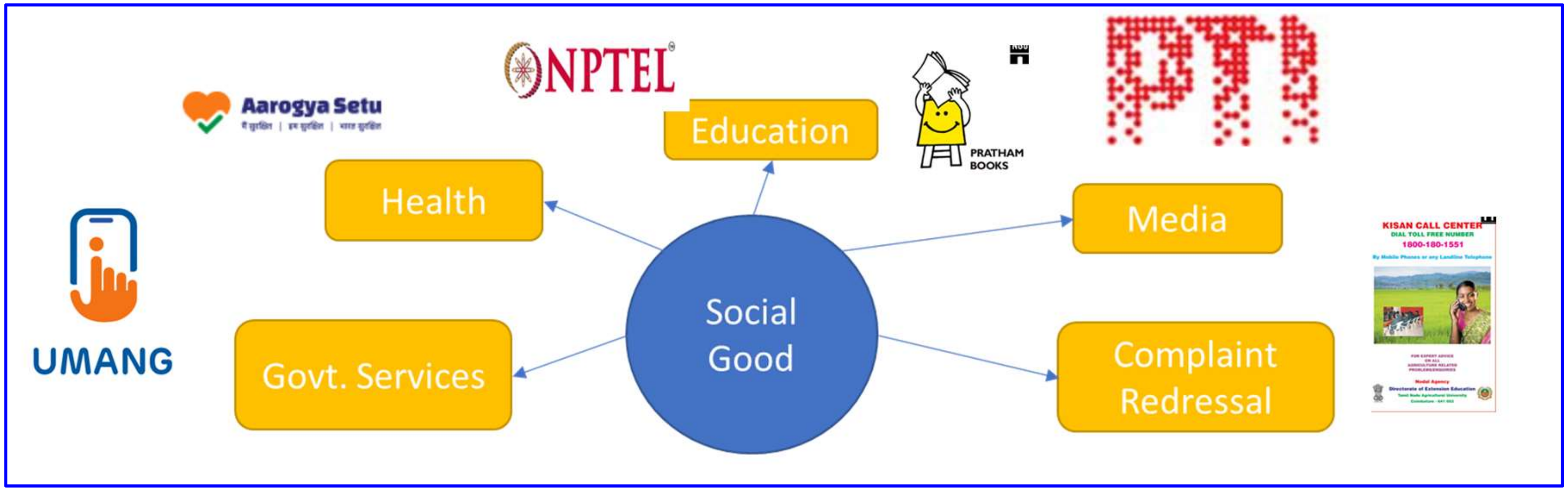
- *4 major language families*
- *22 scheduled languages*
- *125 million English speakers*
- *8 languages in the world's top 20 languages*
- *30 languages with more than 1 million speakers*



Source: KPMG in India's analysis, April 2017

(P): Projections

Internet User Base in India (in million)



Social Good

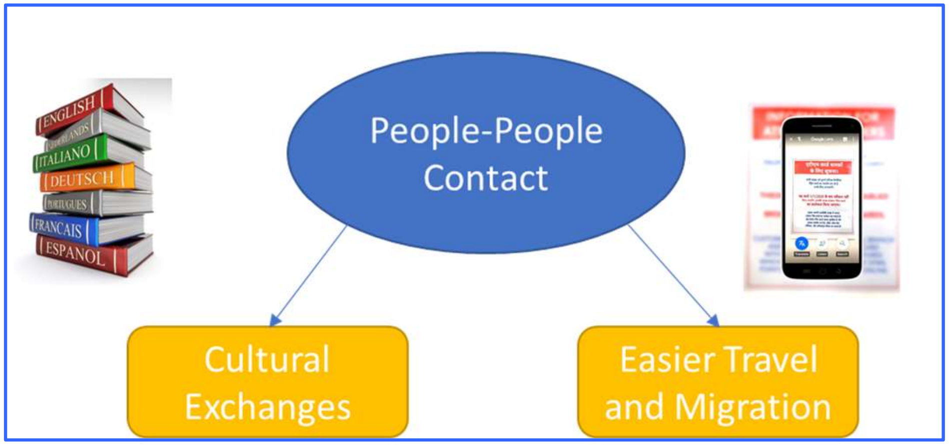
Health

Education

Media

Govt. Services

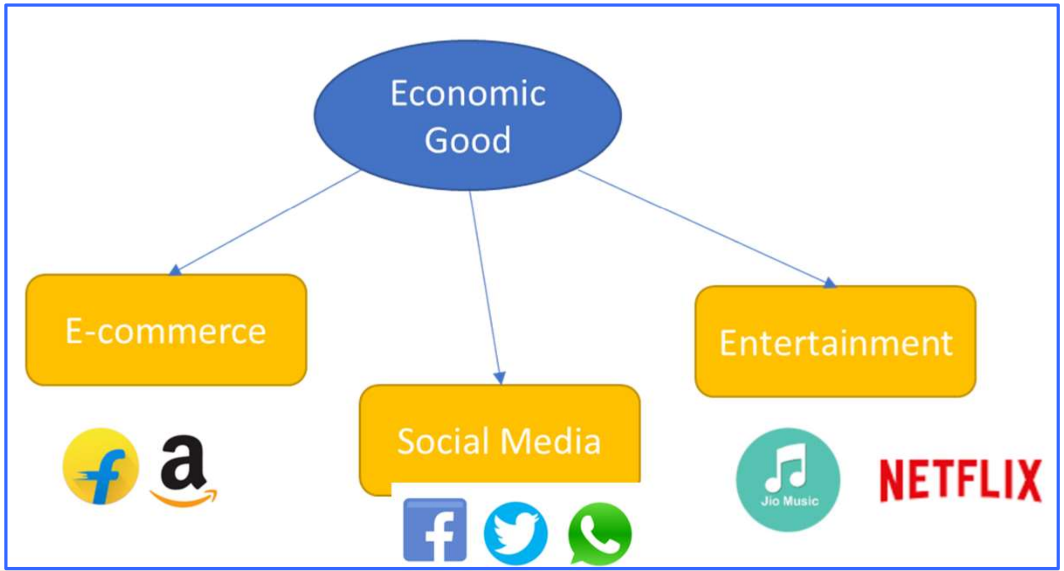
Complaint Redressal



People-People Contact

Cultural Exchanges

Easier Travel and Migration



Economic Good

E-commerce

Social Media

Entertainment

We are faced with a huge data skew

Raw Text Corpora	*Wikipedia articles*	*English*	***150k***
		Hindi	***6m***
Parallel Corpora	*Sentence pairs*	*En-fr (OPUS)*	***500m***
		En-hi	***1.5m***
NER Corpora	*Tokens*	*en (CoNLL 2003)*	***200k***
		hi	***40k***
QA	*Question-Answer Pairs*	*en (SQuAD 1.1)*	***100k***
		hi	***4.6k***

Journey of NLP systems so far

Hand-crafted rules

Atomic representations

Linguistic knowledge

Rule-based systems

Statistical ML Systems

Hand-crafted features

Atomic representations

Data Annotation

Deep Learning Systems

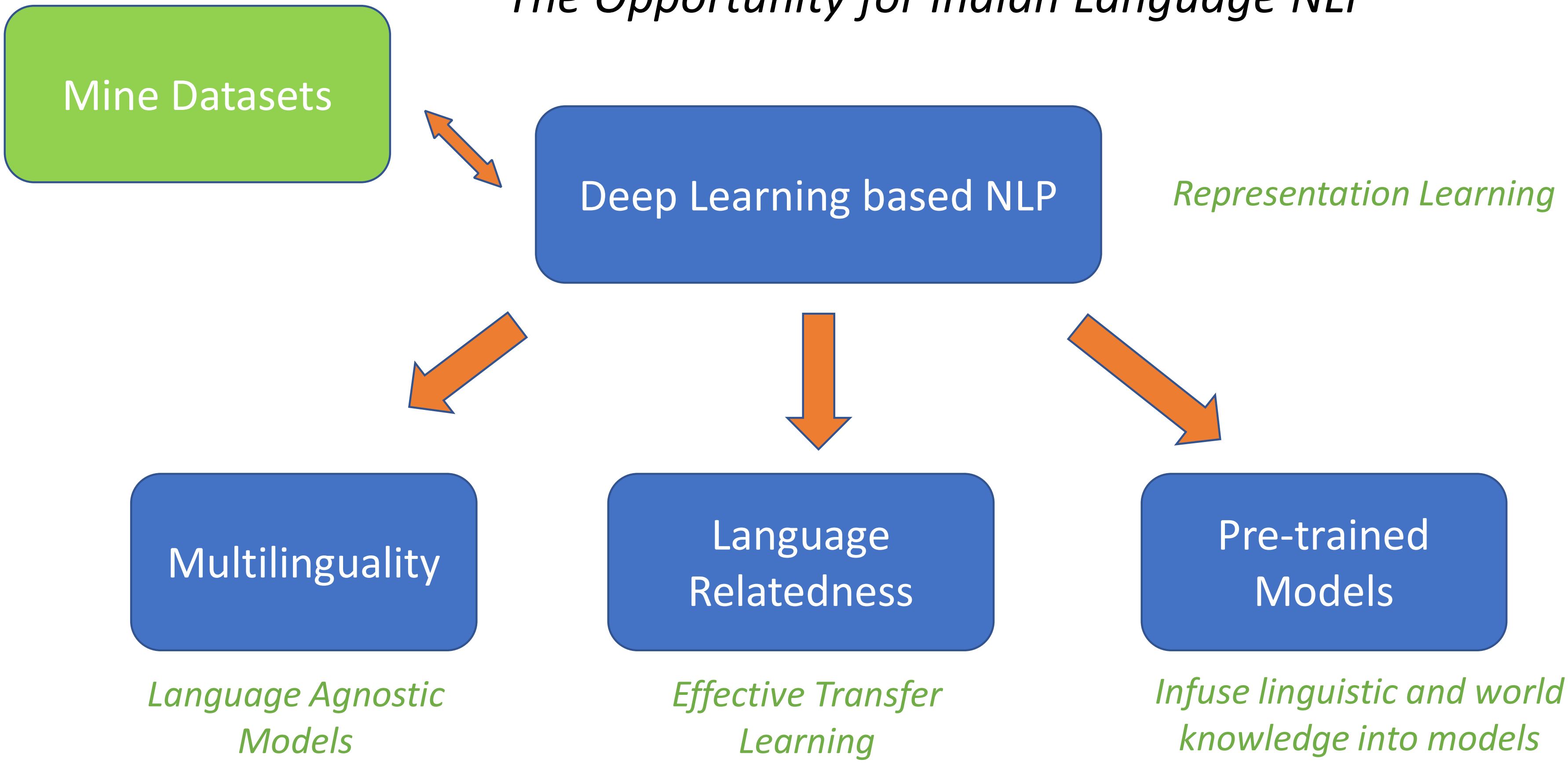
Automatically learnt features

Distributed representations

Data Annotation

Our Technical Direction

The Opportunity for Indian Language NLP



Mine Datasets

Deep Learning based NLP

Representation Learning

Multilinguality

Language Agnostic Models

Language Relatedness

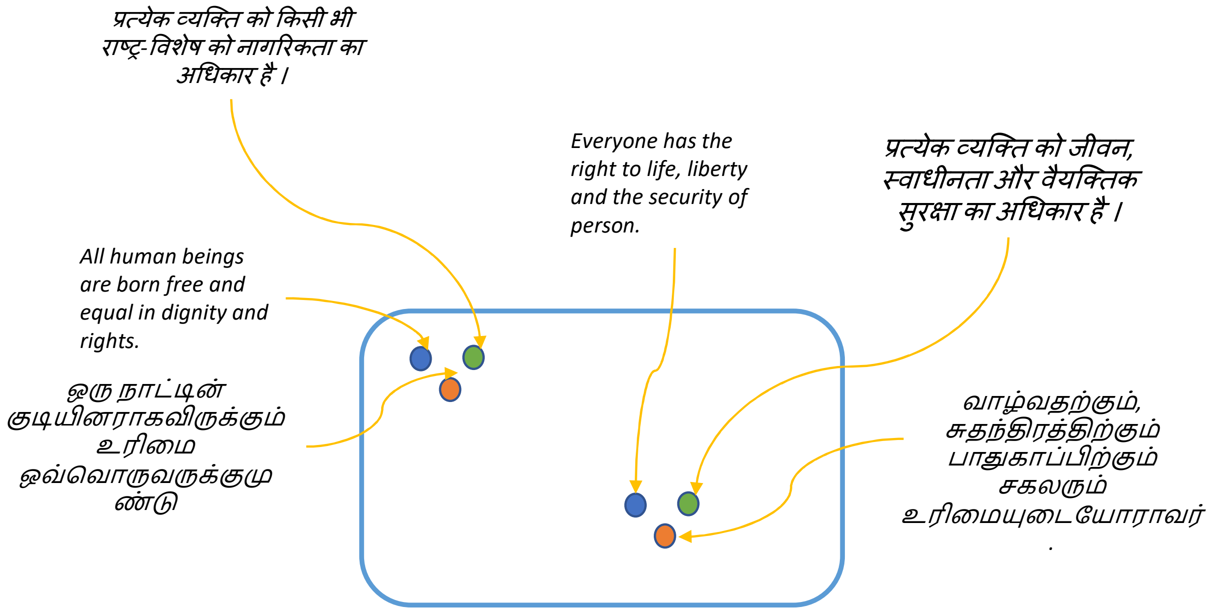
Effective Transfer Learning

Pre-trained Models

Infuse linguistic and world knowledge into models

What do multilingual models do?

Represent semantically similar language artifacts in the same vector space



प्रत्येक व्यक्ति को किसी भी राष्ट्र-विशेष को नागरिकता का अधिकार है ।

All human beings are born free and equal in dignity and rights.

ஒரு நாட்டின் குடியினராகவிருக்கும் உரிமை ஒவ்வொருவருக்குமுண்டு

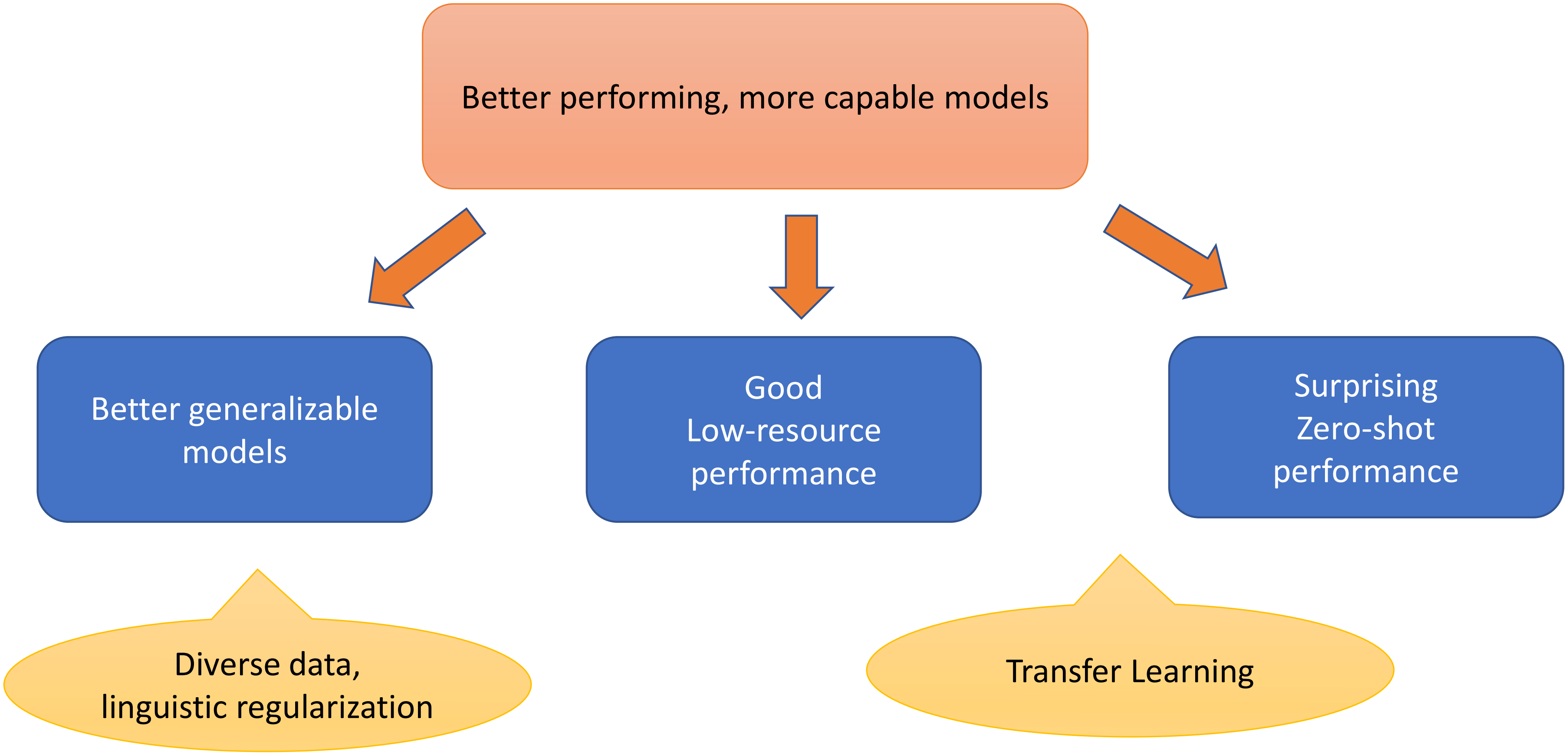
Everyone has the right to life, liberty and the security of person.

प्रत्येक व्यक्ति को जीवन, स्वाधीनता और वैयक्तिक सुरक्षा का अधिकार है ।

வாழ்வதற்கும், சுதந்திரத்திற்கும் பாதுகாப்பிற்கும் சகலரும் உரிமையுடையோராவர் .

Sentences

How does multilinguality help?



Better performing, more capable models

Better generalizable models

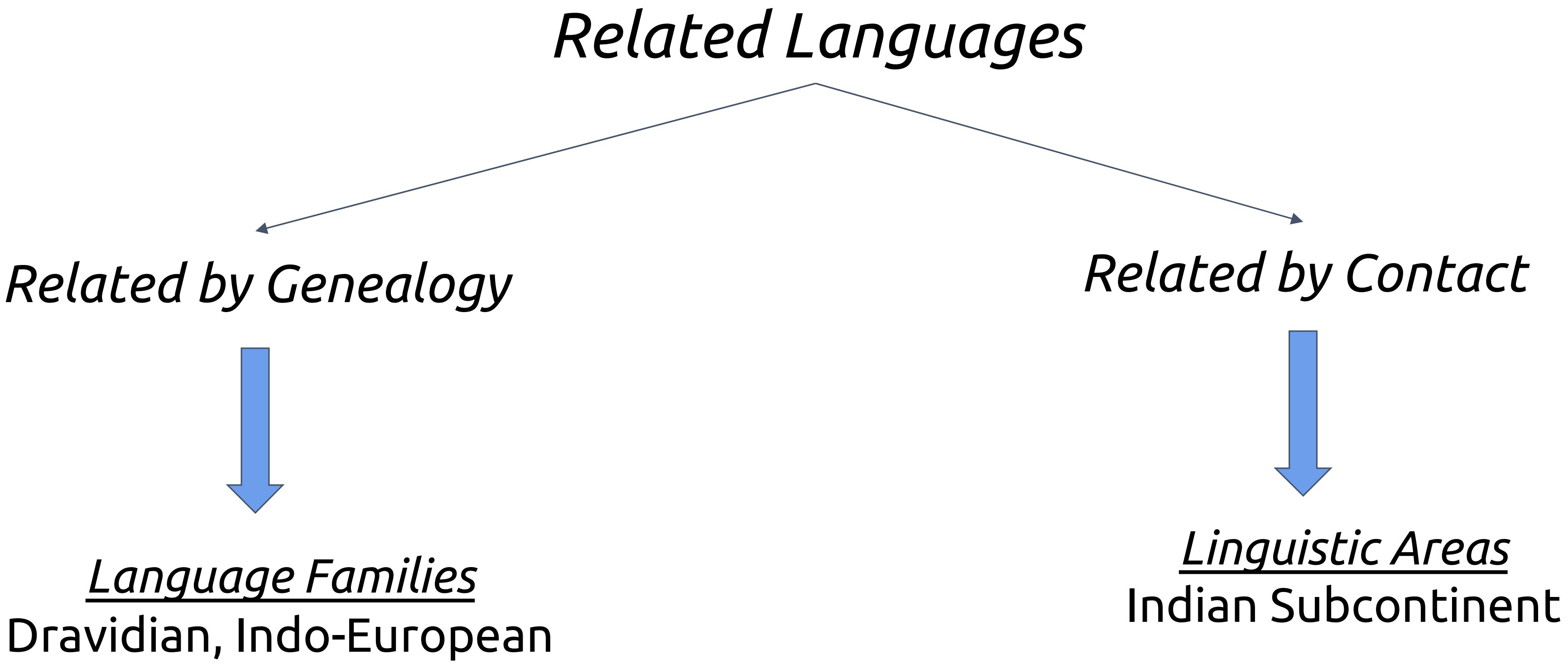
Good Low-resource performance

Surprising Zero-shot performance

Diverse data, linguistic regularization

Transfer Learning

Why are Indian languages related?



Related Languages

Related by Genealogy

Related by Contact

<u>Language Families</u>

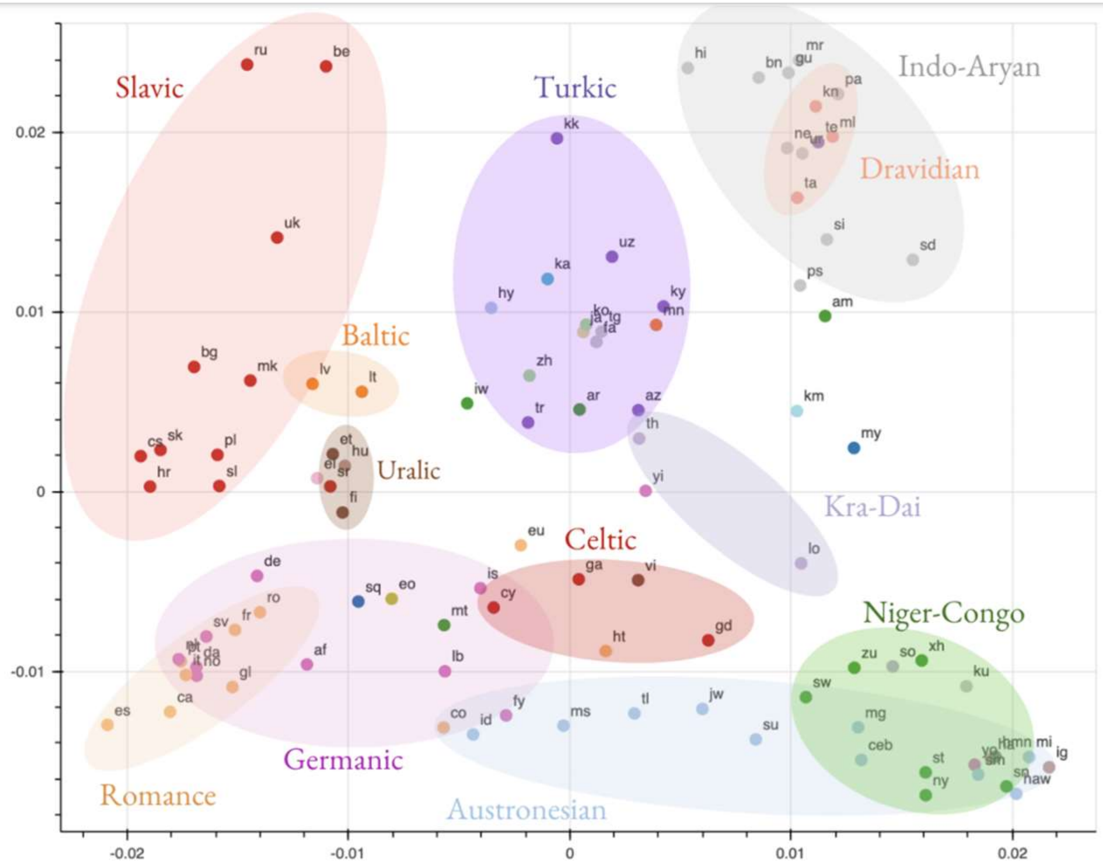
Dravidian, Indo-European

<u>Linguistic Areas</u>

Indian Subcontinent

Lexical, Syntactic & Orthographic similarities

How does language relatedness help?



(Kudungta et al, 2019) *Encoder Representations cluster by language family*

Transfer Learning works best for related languages
(+ use similarity priors)

Building multilingual systems systems specific to language families

How do pre-trained models help?

Supervised data not sufficient

How do we understand linguistics similarities?
synonymy, parts-of-speech, word categories, analogies

How do we know if the sentence is grammatically correct?

How do we know if the sentence makes sense?

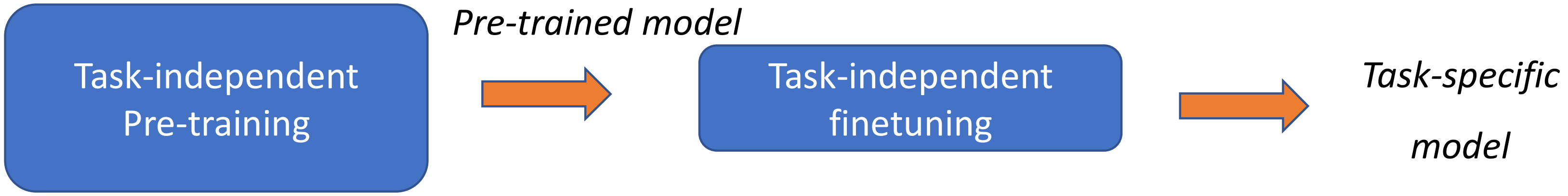
These capabilities are important for generalization

FAIRSEQ

BART

Task-independent models that know about language

Pre-train once, reuse for multiple downstream tasks



Task-independent Pre-training

Pre-trained model

Task-independent finetuning

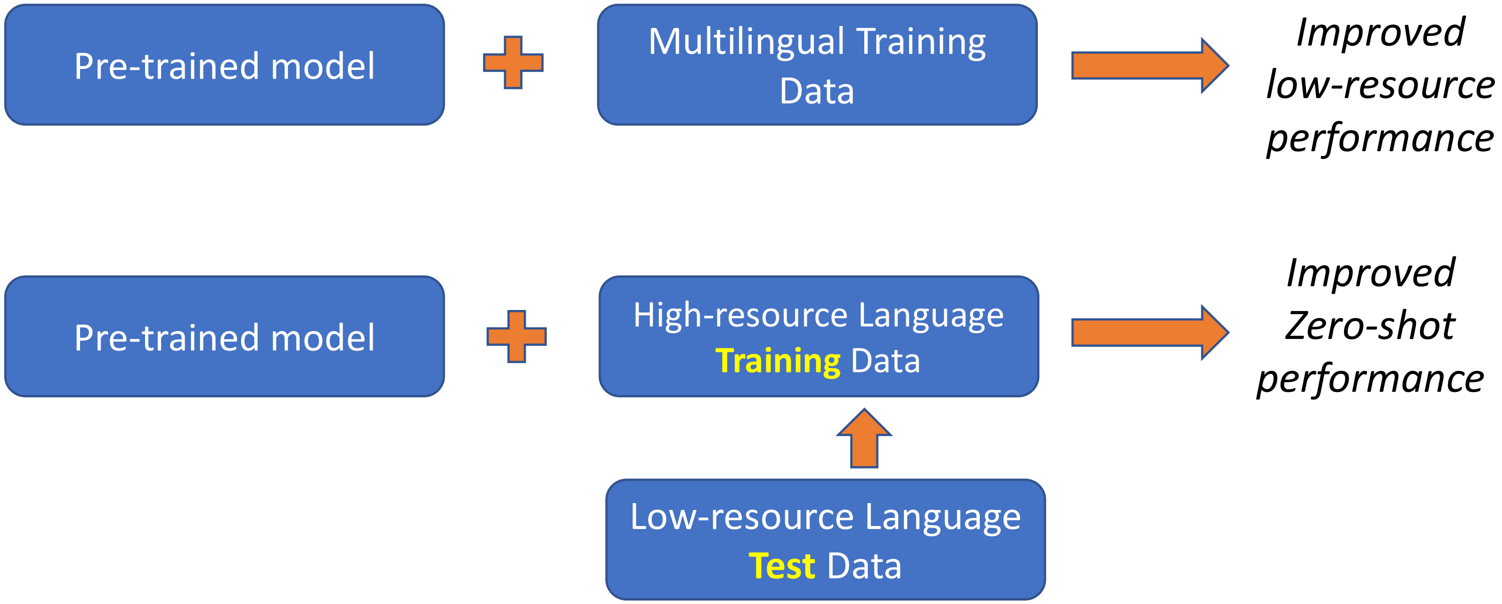
Task-specific model

Only task-specific training: less data & less computation

Multi-linguality and Pre-training are complementary

Language-family specific pre-trained model

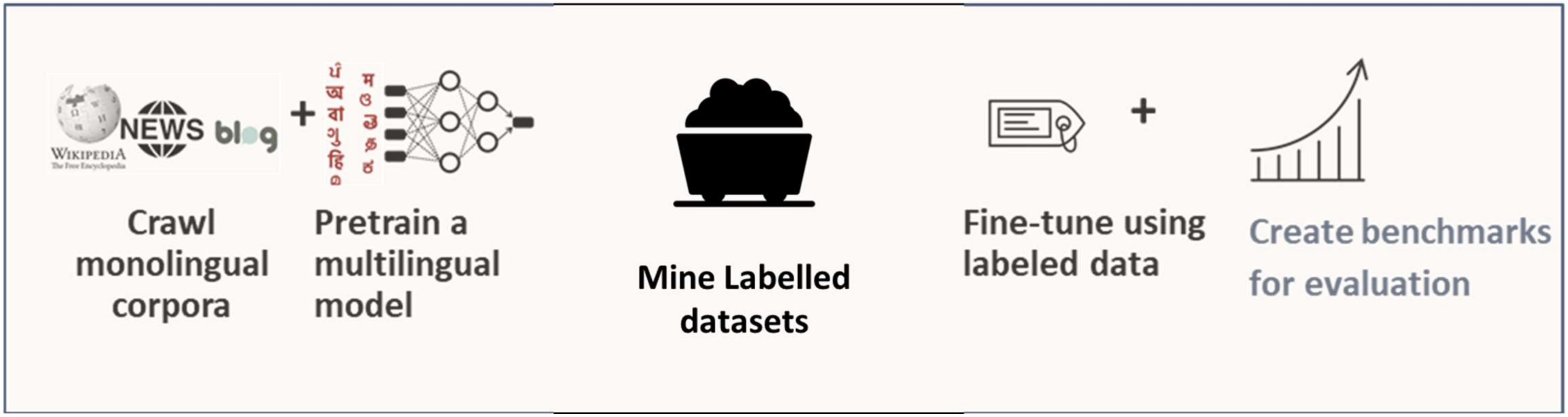
- *Compact pre-trained models*
- *Utilize language relatedness*
- *Better data representation*



Pre-trained model + Multilingual Training Data → *Improved low-resource performance*

Pre-trained model + High-resource Language **Training** Data → *Improved Zero-shot performance*

Low-resource Language **Test** Data



Crawl monolingual corpora

Pretrain a multilingual model

Mine Labelled datasets

Fine-tune using labeled data

Create benchmarks for evaluation

IndicCorp

IndicBERT

IndicBART

IndicGLUE

Naampadam

Indic NLG Benchmark

IndicNER

Agenda for today’s workshop

- Natural Language Understanding
- Natural Language Generation
- Named Entity Recognition

- *Overview*
- *Hands-on/Demo*
    - *Using AI4Bharat models*
    - *Finetuning models with datasets*
    - *Training from scratch with datasets*

Releasing

IndicBERT v2 alpha

IndicNER
Naampadam dataset

https://github.com/AI4Bharat/workshop-nlp-nlu-2022

NATURAL LANGUAGE UNDERSTANDING FOR INDIAN LANGUAGES

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, Pratyush Kumar, Sumanth Doddapaneni, Gowtham Ramesh, Rahul Aralikatte, Shreya Goyal GU0

Slide 15

GU0 affiliations?

Guest User, 2022-07-27T12:05:53.498

What is Natural Language Understanding?

Mask Filling

Paraphrase Detection

Sentiment Classification

Named Entity Recognition

Question Answering

Sentence Retrieval

News genre classification

… … … … …
…

A good language comprehension model is the backbone to perform these tasks

What is missing for Indian Languages?

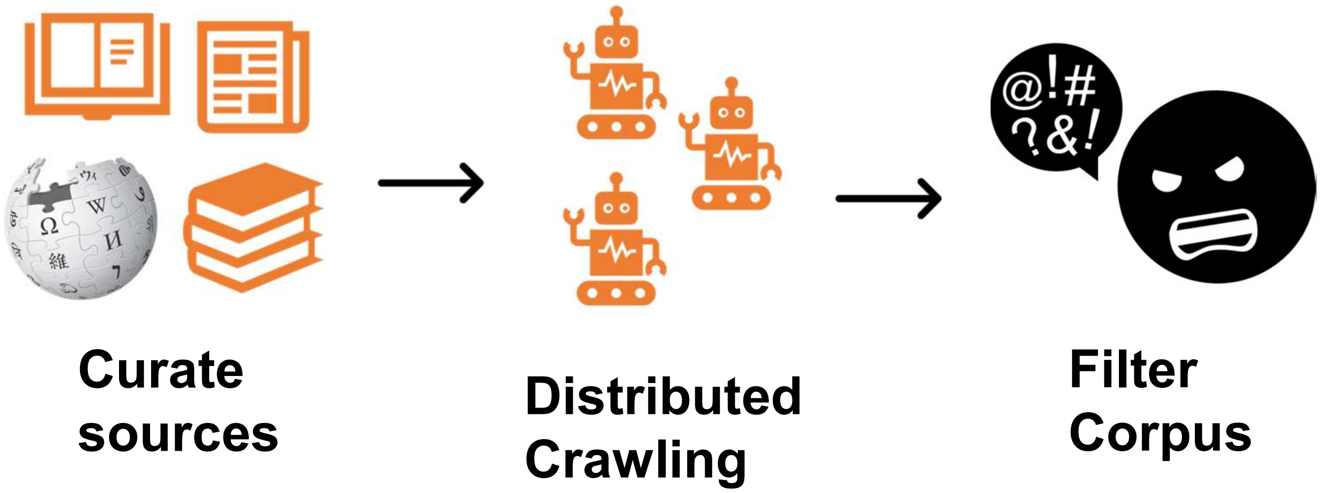
Large scale Monolingual Corpora → **IndicCorp** **450M** Sents.

Evaluation benchmarks → **IndicGLUE** **11** Tasks

Coming soon for IN-22

Multilingual Language Model for IN-22 → **IndicBERT** **18M** Parameter Model

Monolingual Corpora Creation



Curate sources

Distributed Crawling

Filter Corpus

https://github.com/AI4Bharat/webcorpus

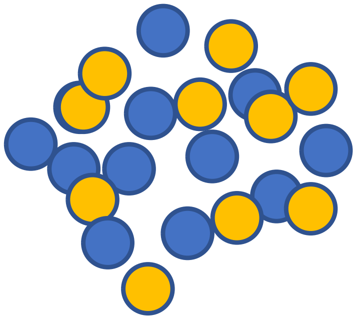
Multilingual Word Embeddings

மரத்தாலான (wooden)

மரத்தால் (tree) + ஆன (making)

Complex tense, verb embedded into a single word

→

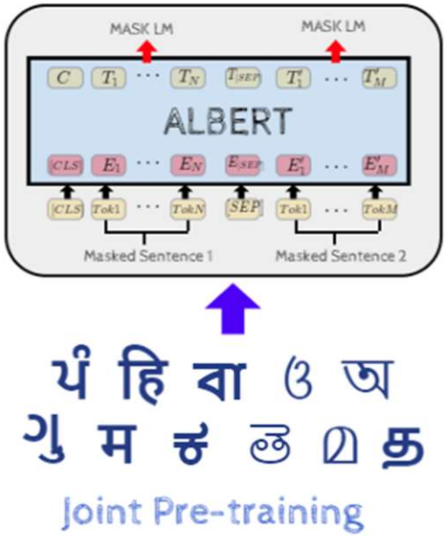


Indic FastText

IndicGLUE

Task Type	Task	N	Languages
Classification	News Article Classification	10	bn, gu, hi, kn, ml, mr, or , ta, te
	Sentiment Analysis	2	hi, te
	Discourse Mode Classification	1	hi
Diagnostics	WNLI	3	gu, hi, mr
	COPA	3	gu, hi, mr
Semantic Similarity	Headline Prediction	11	as, bn, gu, hi, kn ml, mr, or, pa, ta, te
	WIkipedia Section Titles	11	as, bn, gu, hi, kn ml, mr, or, pa, ta, te
	Close Style QA	11	as, bn, gu, hi, kn ml, mr, or, pa, ta, te
	Paraphrase Detection	4	hi, ml, pa, ta
	Named Entity Recognition	11	as, bn, gu, hi, kn ml, mr, or, pa, ta, te
Cross-lingual	Cross-lingual sentence retrieval	8	bn, gu, hi, ml, mr, or, ta, te

IndicBERT

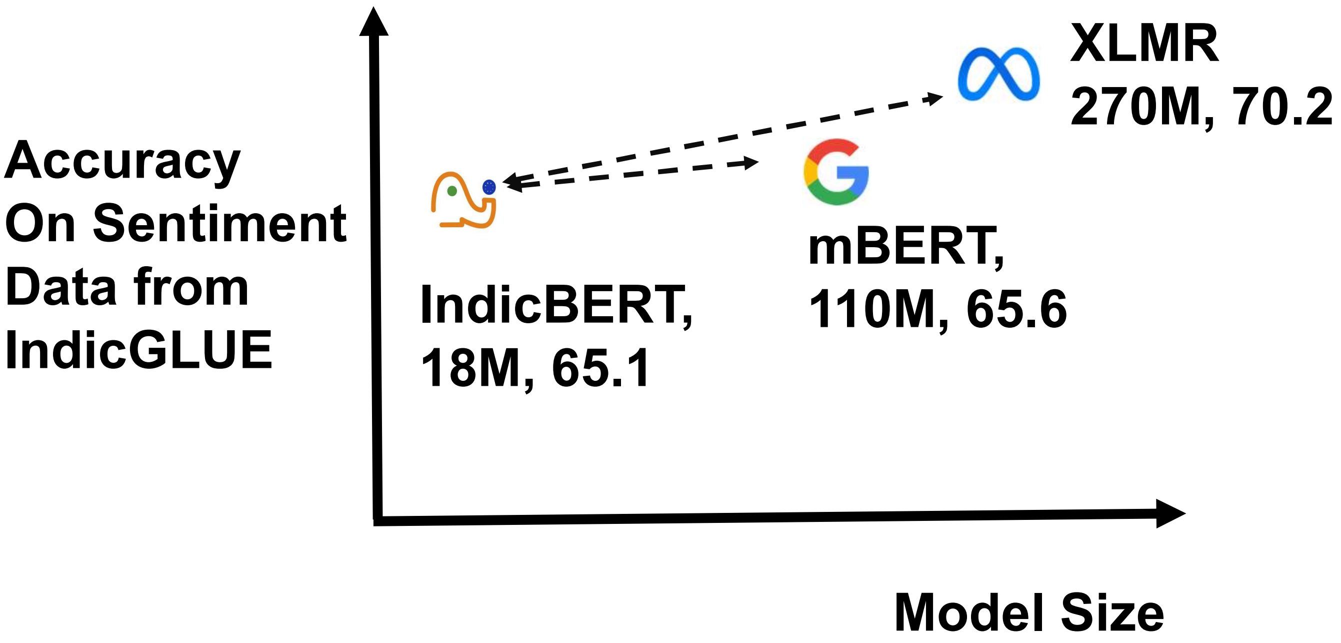


- Pre-trained Indic LM for NLU applications
- Large Indian language content (8B tokens)
  - 11 Indian languages
  - + Indian English content
- Multilingual Model
- Compact Model (~20m params)
- Competitive/better than mBERT/XLM-R
- Simplify fine-tune for your application
- 10k downloads per month on HuggingFace

https://indicnlp.ai4bharat.org/indic-berthttps://huggingface.co/ai4bharat/indic-bert

Results

IndicBERT is 6X smaller and yet very accurate



XLMR 270M, 70.2

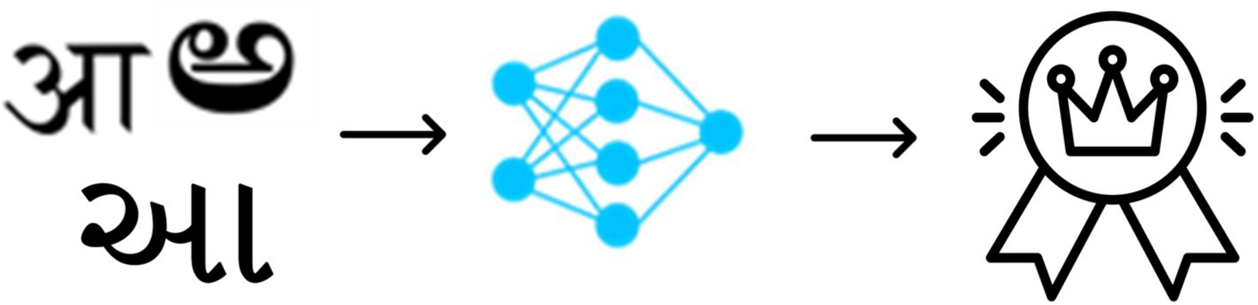
mBERT, 110M, 65.6

IndicBERT, 18M, 65.1

Accuracy On Sentiment Data from IndicGLUE

Model Size

No Training Data



IN-22 Input → **Model trained on English** → **Good Performance**

Work towards good zero-shot performance

Our Plan Ahead

- Support for IN-22 languages
- Evaluation benchmarks for multiple tasks
- Improve zero-shot performance
- Efficient pre-training and finetuning

Summary

- *IndicCorp:* Largest publicly available monolingual corpora for English and 11 Indian languages
- *IndicBERT:* Compact multilingual model trained on IndicCorp
- *IndicGLUE:* Natural Language Understanding benchmark with 11 tasks for Indian languages
- *IndicFT:* Multilingual word embeddings trained on IndicCorp
- We show that our multilingual IndicBERT is 6x smaller and still very accurate

NATURAL LANGUAGE GENERATION FOR INDIAN LANGUAGES

Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, Pratyush Kumar

What is Natural Language Generation?

Machine Translation

Automatic Summarization

Table-to-Text Generation

Dialog Generation

Paraphrase Generation

… … … … … …

What is missing for Indic languages?

- Pretraining Data and Model
- NLG Training Data
- Models across IN-22

Our Approach

1. Leverage IndicCorp with data in 11 langs to train IndicBART

2. Exploit lang. similarity by script unification

3. Devise methods to auto-create NLG training data

1. Train IndicBART on IndicCorp

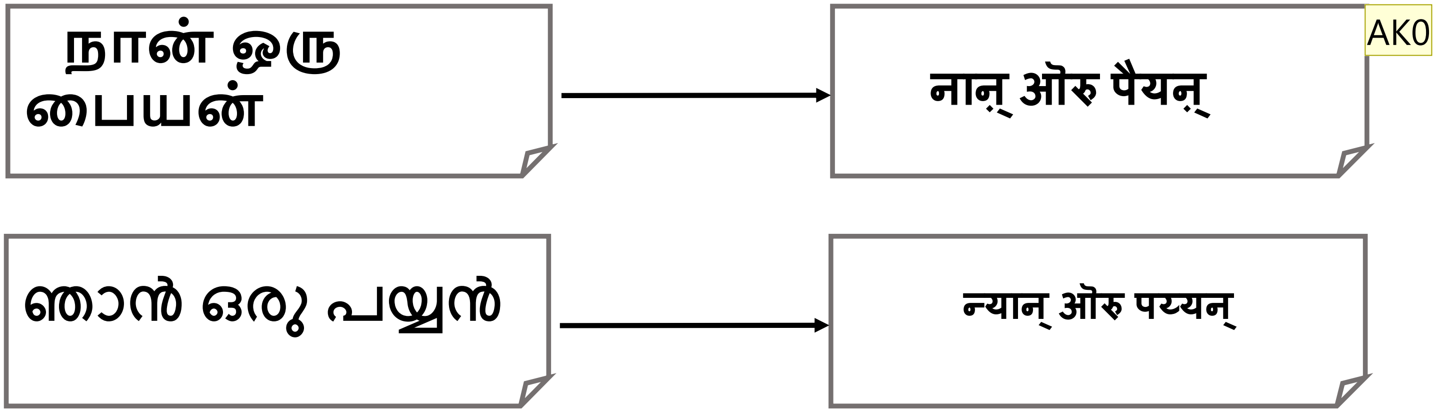
450M input sentences of training data

Compact models with 244M params

Covers 11 Indian languages

2. Script Unification

- Many languages need large vocabulary
- Script unification by converting to Devanagari
  - Increased vocabulary sharing
  - Compact vocabularies for compact models


நான் ஒரு பையன்

नाऩ् ऒरु पैयऩ्

AK0

ഞാൻ ഒരു പയ്യൻ

न्यान् ऒरु पय्यन्

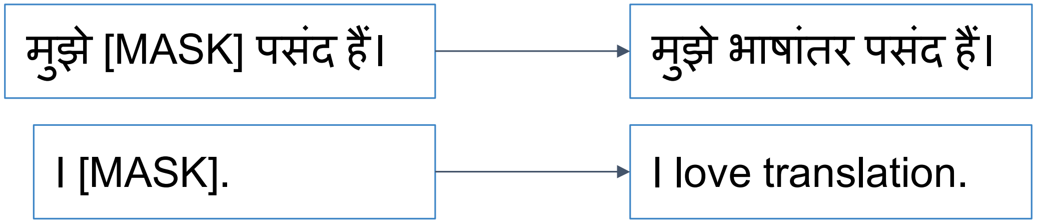
Slide 31

AK0 See the latest version of the poster - has a Tamil + Malayalam example

Anoop Kunchukuttan, 2022-07-27T06:52:01.899

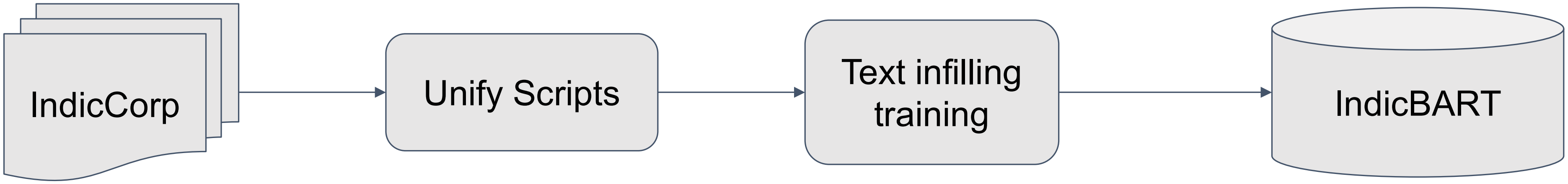
IndicBART Training

- Train models to do: (text infilling)



मुझे [MASK] पसंद हैं। → मुझे भाषांतर पसंद हैं।

I [MASK]. → I love translation.

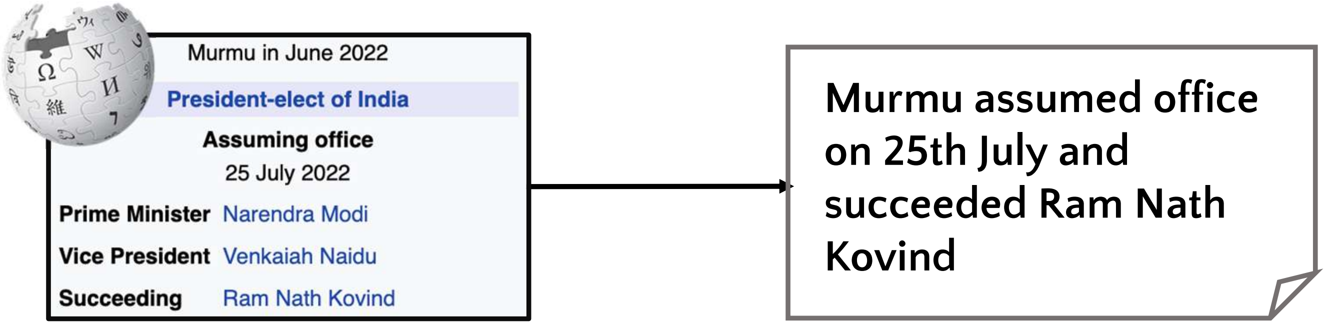


IndicCorp → Unify Scripts → Text infilling training → IndicBART

- IndicBART learns to infer a variation of input.
  - Learns generic NLG →Reduces need for task data (fine-tuning)
  - Variations: IndicALBART (compact)

3. Methods for creating training data

BIOGRAPHY GENERATION



Murmu in June 2022

President-elect of India

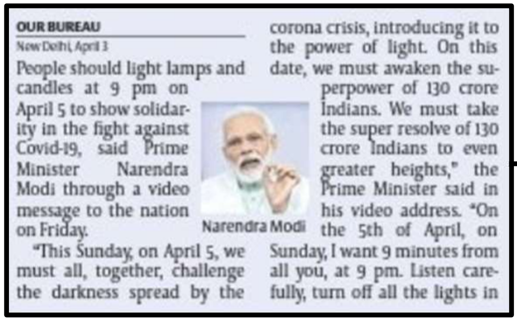
Assuming office
25 July 2022

Prime Minister	Narendra Modi
Vice President	Venkaiah Naidu
Succeeding	Ram Nath Kovind

Murmu assumed office on 25th July and succeeded Ram Nath Kovind

3. Methods for creating training data

HEADLINE GENERATION



OUR BUREAU

New Delhi, April 3

People should light lamps and candles at 9 pm on April 5 to show solidarity in the fight against Covid-19, said Prime Minister Narendra Modi through a video message to the nation on Friday.

"This Sunday, on April 5, we must all, together, challenge the darkness spread by the corona crisis, introducing it to the power of light. On this date, we must awaken the superpower of 130 crore Indians. We must take the super resolve of 130 crore Indians to even greater heights," the Prime Minister said in his video address. "On the 5th of April, on Sunday, I want 9 minutes from all you, at 9 pm. Listen carefully, turn off all the lights in

Narendra Modi

Dispel the virus darkness with light: PM to people

SENTENCE SUMMARISATION

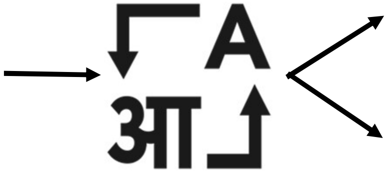
India 's financial markets are closed on Monday for a public holiday.

India markets closed for holiday

3. Methods for creating training data

PARAPHRASE GENERATION

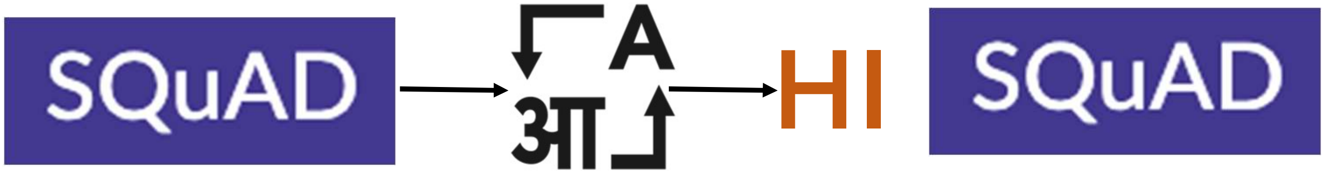
दिल्ली विश्वविद्यालय, भारत में उच्च शिक्षा के लिए एक प्रतिष्ठित संस्थान है।



The University of Delhi is a prestigious institution for higher education in India.

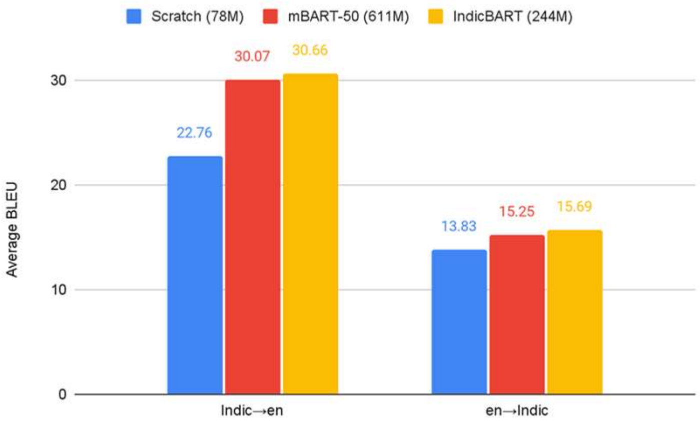
Delhi University is one of the famous universities of the country.

QUESTION GENERATION

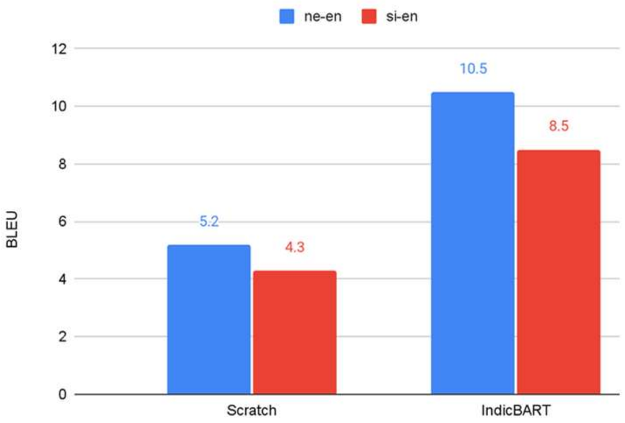


INDICTRANS

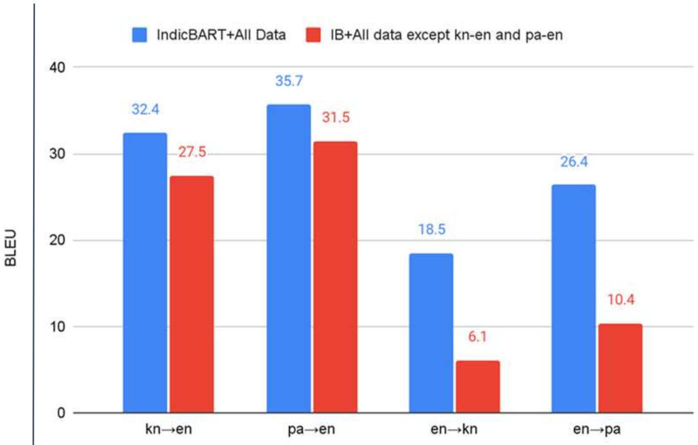
Machine Translation Results (BLEU)



- **Large impact of pre-training**
  - *Indic→En: 22.76→30.66*
  - *En→Indic: 13.83→15.69*
- **Indic→En gains more than En→Indic**



- **IndicBART helps Nepali and Sinhala translation**
- **Both were unseen by IndicBART**



- **IndicBART helps unseen language translation**
- **Punjabi and Kannada data not used**
  - Can still translate

Other NLG Task Results (Rouge and iBLEU)

Task	Scratch	mT5	IB
Biography Generation	47.8	54.6	53.7
Headline Generation	37.1	45.5	43.7
Sentence Summarization	48.9	55.2	54.5
Paraphrase Generation	8.7	5.1	10.6
Question Generation	20.0	25.2	26.0
Average	32.5	37.1	37.7

- IndicBART pre-training significantly improves quality
- Competitive or better than other generic pre-trained models (mT5)

Our Plan Ahead

- Support 22 Indian languages
- Train on diverse data
- More datasets particularly for open-ended generation tasks
- Generative language model like GPT
- Efficient pre-training & fine-tuning

Summary

- We contribute the first large-scale datasets, benchmarks, and models for Indic NLG.
- *IndicBART*: Compact Language model for 11 Indian languages
- *IndicNLG Benchmark*: Generation task datasets for 11 languages and 5 tasks
- We show that our models are 3x smaller yet competitive with large LMs

IndicNER

Named Entity Recognition Dataset and Models for Indic Languages

Harshit Kedia, Arnav Anil Maske, Anoop Kunchukuttan, Rudra Murthy, Mitesh M. Khapra, Pratyush Kumar

(Model) https://ai4bharat.org/indic-ner

(Dataset) https://ai4bharat.org/naamapadam

TL;DR

- Naamapadam Dataset
  - Large-Scale NER dataset for 11 Indic languages
    - As, Bn, Gu, Hi, Kn, Ml, Mr, Or, Pa, Ta, Te
    - Automated Creation via entity projection
  - Human annotated test-set for 8 Indic languages
    - Bn, Hi, Kn, Ml, Mr (large)
    - Ta, Te, Gu (small)
- Multilingual IndicNER model
  - 11 Indic languages (As, Bn, Gu, Hi, Kn, Ml, Mr, Or, Pa, Ta, Te)
  - Compact 159.05 M parameters
- Publicly available models and code

Named Entity Recognition

The task of identifying and extracting named entities in a given piece of text

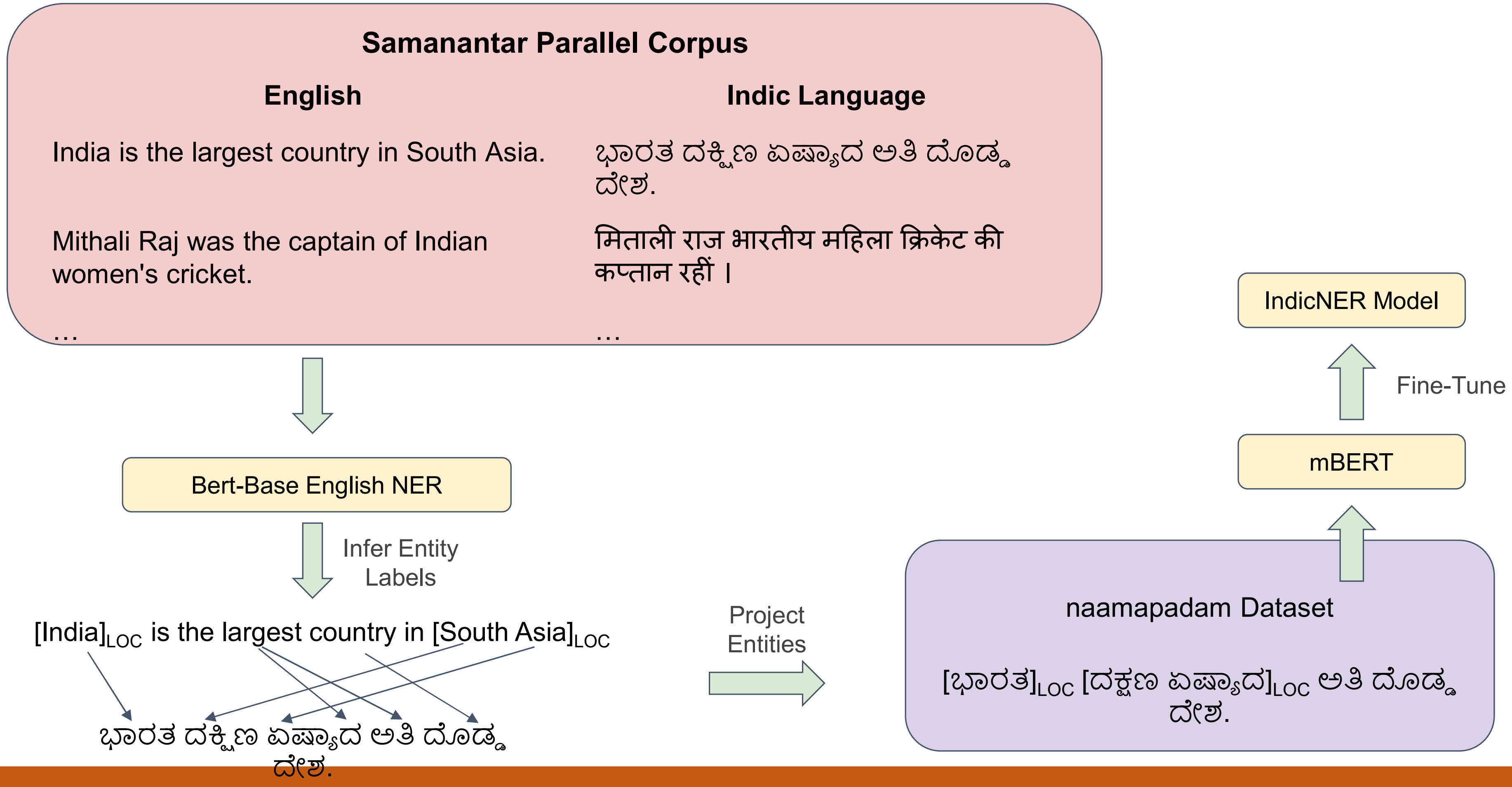
For example,

[Nilekani Center]$_{\text{LOCATION}}$ at [AI4Bharat]$_{\text{ORGANIZATION}}$ will be launched on [28th July]$_{\text{DATE}}$ at [IIT Madras]$_{\text{ORGANIZATION}}$

Challenges in Indic languages:

- Lack of capitalization feature
- Ambiguity between Proper nouns and common nouns
- Morphological variations
- Small labelled data

Naamapadam Dataset and IndicNER Model



Samanantar Parallel Corpus

English	Indic Language
India is the largest country in South Asia.	ಭಾರತ ದಕ್ಷಿಣ ಏಷ್ಯಾದ ಅತಿ ದೊಡ್ಡ ದೇಶ.
Mithali Raj was the captain of Indian women's cricket.	मिताली राज भारतीय महिला क्रिकेट की कप्तान रहीं ।
…	…

Bert-Base English NER

Infer Entity Labels

[India]$_{LOC}$ is the largest country in [South Asia]$_{LOC}$

ಭಾರತ ದಕ್ಷಿಣ ಏಷ್ಯಾದ ಅತಿ ದೊಡ್ಡ ದೇಶ.

Project Entities

naamapadam Dataset

[ಭಾರತ]$_{LOC}$ [ದಕ್ಷಣ ಏಷ್ಯಾದ]$_{LOC}$ ಅತಿ ದೊಡ್ಡ ದೇಶ.

mBERT

Fine-Tune

IndicNER Model

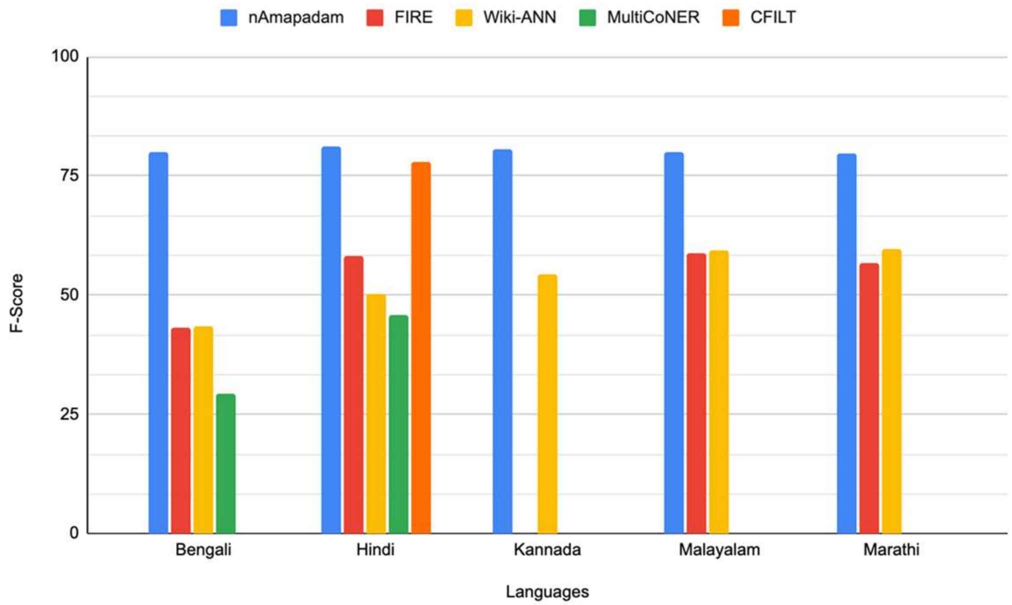
Naamapadam Dataset Statistics

	Train			Sentence Count		
Language	**Person**	**Location**	**Organization**	**Train**	**Dev**	**Test**
Bengali	214K	115K	144K	964K	4.8K	607
Hindi	197K	117K	143K	1335K	13.5K	437
Kannada	88K	42K	62K	471K	2.4K	1019
Malayalam	137K	61K	78K	716K	3.6K	974
Marathi	82K	39K	53K	455K	2.3K	1080
Gujarati	84K	42K	72K	473K	2.4K	50
Tamil	95K	68K	87K	553K	2.8K	49
Telugu	91K	49K	71K	535K	2.7K	53
Assamese	0.4K	1.1K	1.4K	10.2K	52	51
Odiya	43K	20K	31K	196K	1K	1K
Punjabi	99K	46K	88K	464K	2.3K	2.3K

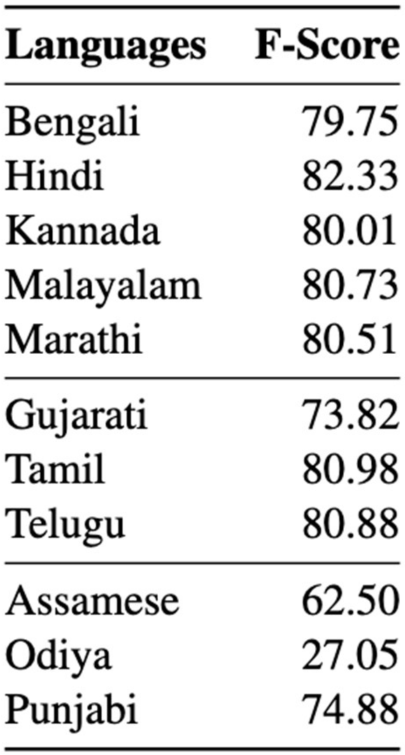
9 out of 11 of the languages have >400K sentences and >100K named entities.

Our projection based approach achieves >70 F-Score for many languages when evaluated against human annotations

Results



mBERT model fine-tuned on train split of existing available datasets and tested on our naamapadam test set. mBERT model fine-tuned on naamapadam train split achieves the best F-Score compared to mBERT model fine-tuned on existing datasets



Languages	F-Score
Bengali	79.75
Hindi	82.33
Kannada	80.01
Malayalam	80.73
Marathi	80.51
Gujarati	73.82
Tamil	80.98
Telugu	80.88
Assamese	62.50
Odiya	27.05
Punjabi	74.88

IndicNER multilingual model F-Score on naamapadam test set. Our multilingual model achieves >80 F-Score on many languages

Future Work

- Cover all 22 languages listed in the Indian constitution
- Wide coverage NER evaluation sets & high-quality seed training sets

Thank you!