AI4Bharat

An IIT Madras Initiative

https://indicnlp.ai4bharat.org



Mitesh M. Khapra
Associate Professor
IIT Madras

Pratyush Kumar
Researcher, *Microsoft*
Adjunct Faculty, *IIT Madras*

Anoop Kunchukuttan
Senior Applied Researcher
Microsoft

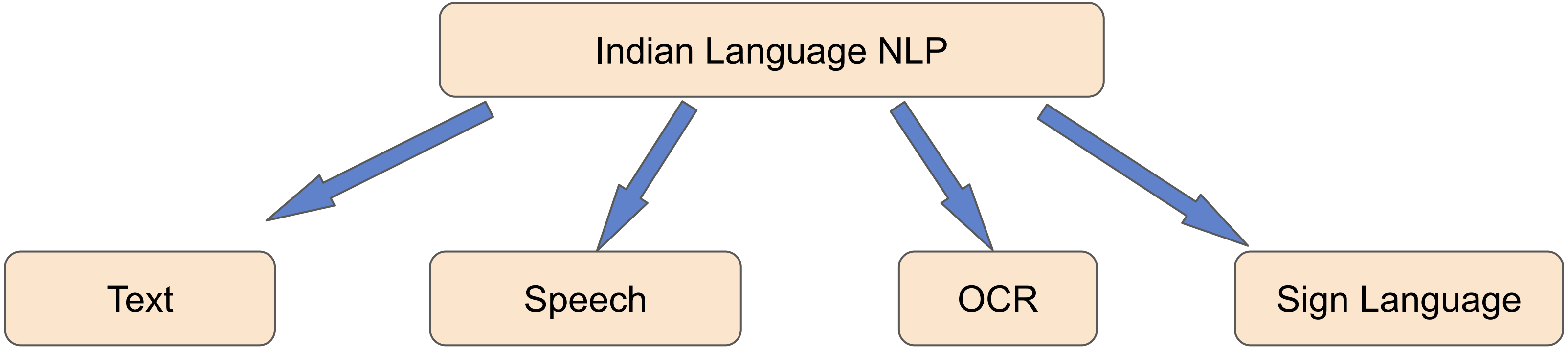
+ Many hard-working students and volunteers

Tutorial at ICON 2021, Dec 2021

AI4Bharat

Let us solve India's challenges with AI

AI4Bhārat is a non-profit, open-source community of engineers, domain experts, policy makers, and academicians collaborating to build AI solutions to solve India's problems, today.



Multimodal NLP

https://ai4bharat.org

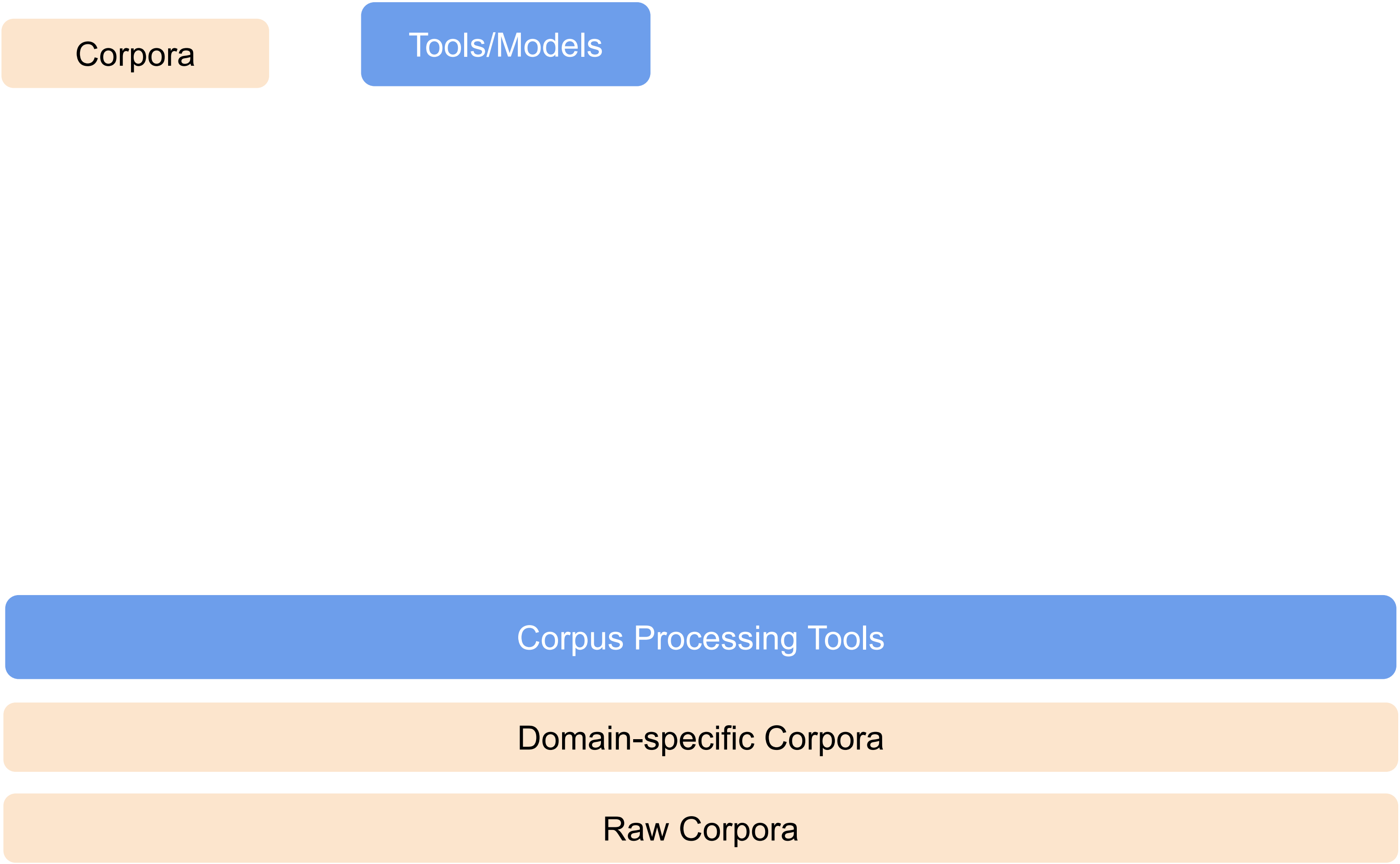
Mission Statement

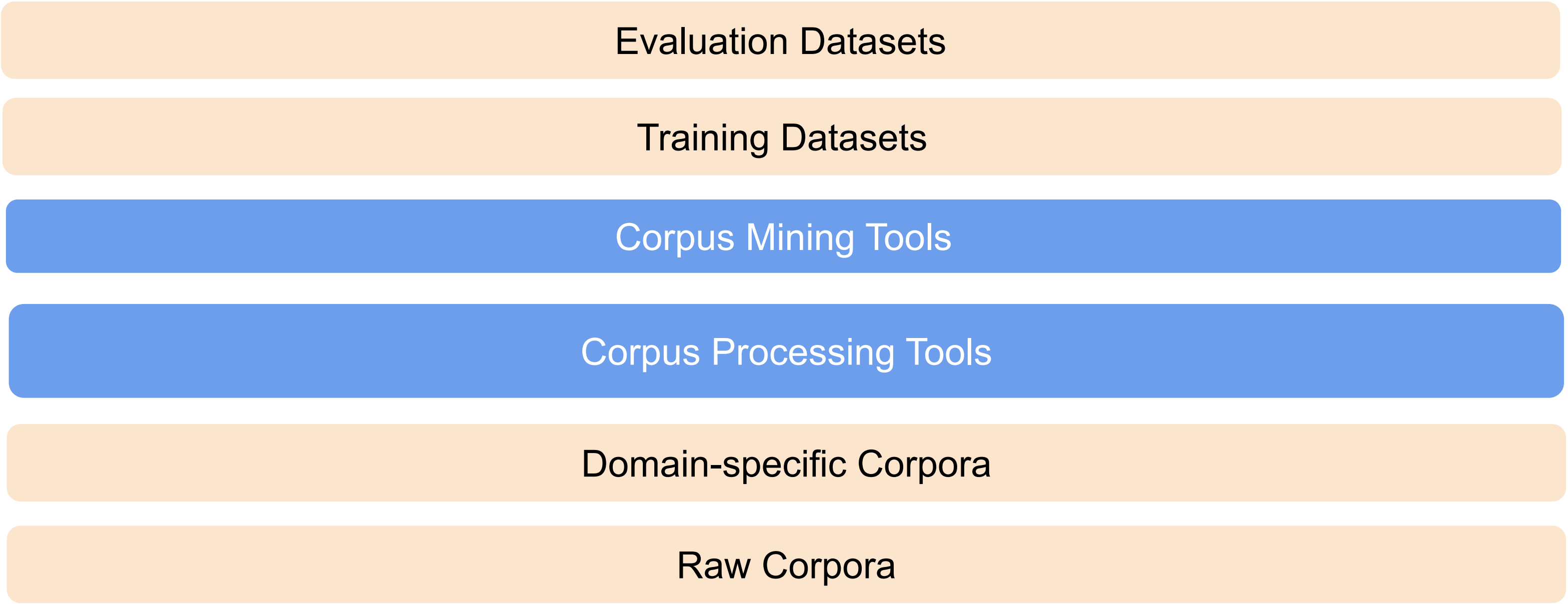
Bring parity with English
in AI tech for Indian languages
with open data and open source contributions

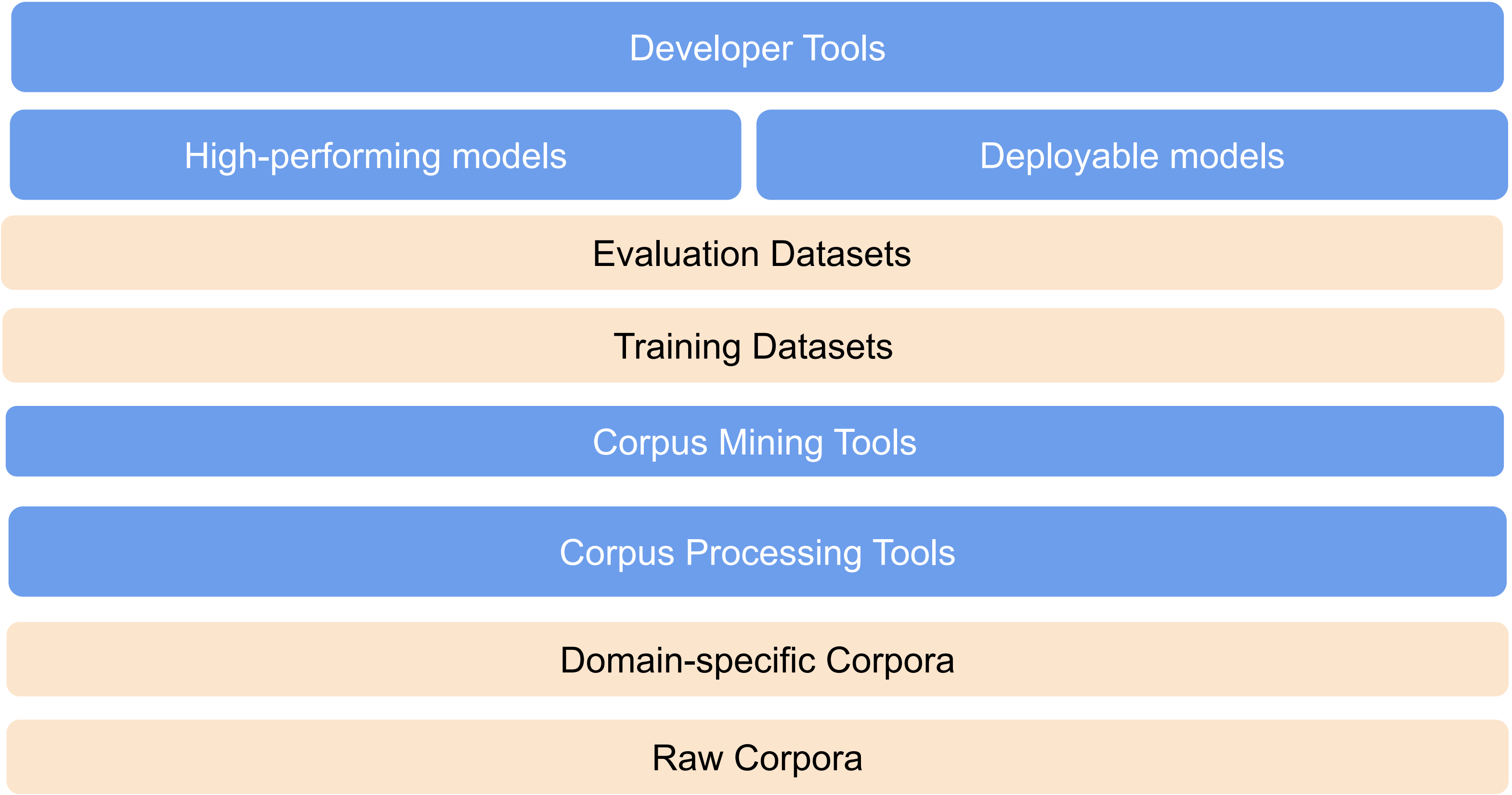
THE APACHE® SOFTWARE FOUNDATION

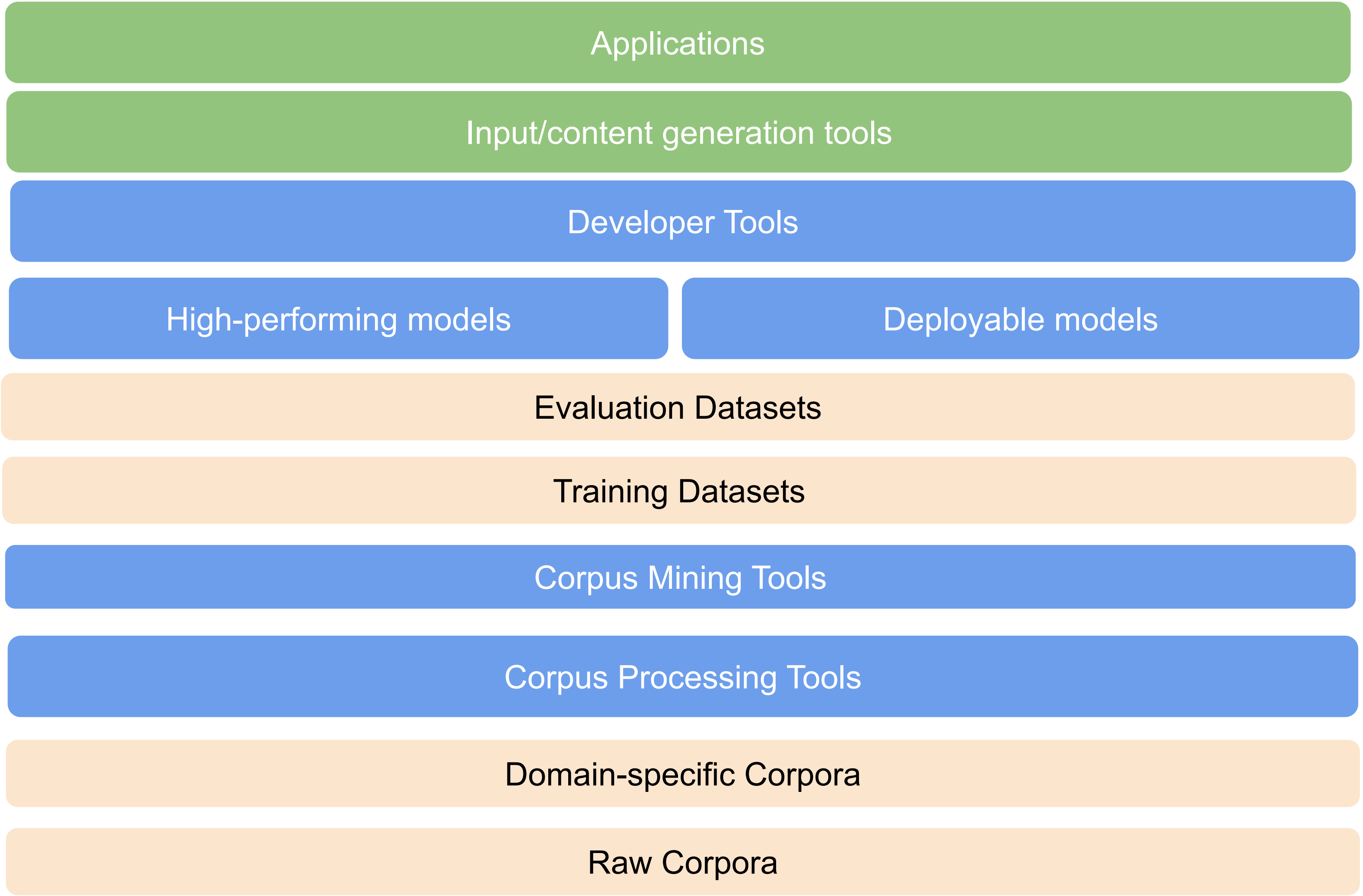
We want to be the Apache for Indian Languages AI stack

Build an ecosystem of datasets, models, partners and stakeholders to advance IndicNLP


Corpora
Tools/Models
Corpus Processing Tools
Domain-specific Corpora
Raw Corpora


Evaluation Datasets
Training Datasets
Corpus Mining Tools
Corpus Processing Tools
Domain-specific Corpora
Raw Corpora


Developer Tools
High-performing models
Deployable models
Evaluation Datasets
Training Datasets
Corpus Mining Tools
Corpus Processing Tools
Domain-specific Corpora
Raw Corpora


Applications
Input/content generation tools
Developer Tools
High-performing models
Deployable models
Evaluation Datasets
Training Datasets
Corpus Mining Tools
Corpus Processing Tools
Domain-specific Corpora
Raw Corpora

Goal

for 22 languages

Full NLP stack
Text Generators
Inference Engines
Text Analysers
Input Tools

Translation   Dialog   Summarisation   ....

QA   NLI   Paraphrase Detection   ....

Named Entity Recognition   Sentiment Analysis   Topic Classification   Content Filters   ....

Keyboards   Spell checkers   Standardise fonts

Founding Principles

Move Fast	solve problems today
Mine Data	collective intelligence, smart annotation
Think Big	22 languages, industry-scale models
Deliver Performance	Best-in-class models, compute-efficiency
Democratize Indic NLP	**open-source models and datasets**
Do Good Science	publish in top-tier conferences

What have we done so far?

What have we done so far?

Basic Infrastructure: Raw corpora & core language models

IndicCorp

Large Monolingual corpora for 11 Indian languages

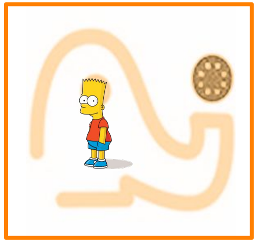
IndicGLUE

NLU benchmark datasets

IndicBERT

Compact pre-trained language models for NLU

IndicFT: word embeddings



IndicBART

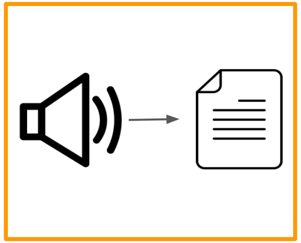
Compact pre-trained seq2seq models for NLG

What have we done so far?

Data and models for various end tasks

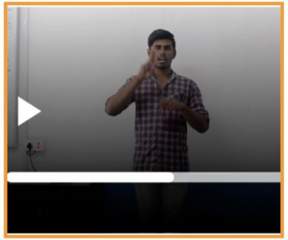
Samanantar

Parallel corpus, translation models between English & 11 Indic languages



IndicWav2Vec

Large-scale raw audio corpora & ASR models for 9 Indian languages



INCLUDE

Datasets and efficient models for isolated Indian Sign Language

Input Tools

Romanized keyboards for under-represented languages https://github.com/AI4Bharat/indicnlp_catalog

IndicNLP Catalog

Evolving, collaborative catalog of Indian language NLP resources

Please add resources you know of and send a pull request

- Major Indic Language NLP Repositories
- Libraries and Tools
- Evaluation Benchmarks
- Standards
- Text Corpora
  - Unicode Standard
  - Monolingual Corpus
  - Language Identification
  - Lexical Resources
  - NER Corpora
  - Parallel Translation Corpus
  - Parallel Transliteration Corpus
  - Text Classification
  - Textual Entailment/Natural Language Inference
  - Paraphrase
  - Sentiment, Sarcasm, Emotion Analysis
  - Question Answering
  - Dialog
  - Discourse
  - Information Extraction
  - POS Tagged corpus
  - Chunk Corpus
  - Dependency Parse Corpus
  - Co-reference Corpus
- Models
  - Word Embeddings
  - Sentence Embeddings
  - Multilingual Word Embeddings
  - Morphanalyzers
  - SMT Models
- Speech Corpora
- OCR Corpora
- Multimodal Corpora
- Language Specific Catalogs

Featured Resources

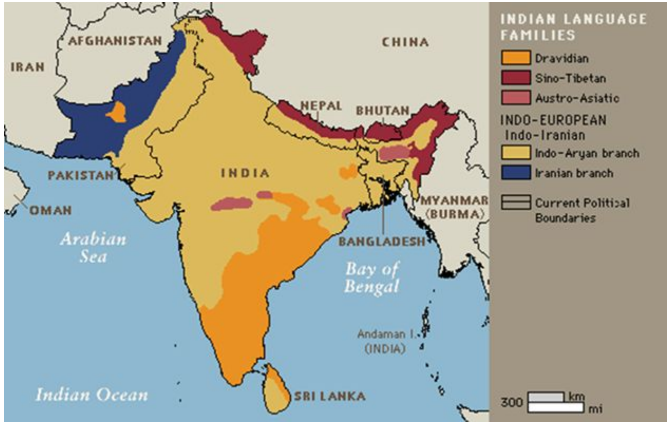
- AI4Bharat IndicNLPSuite: Text corpora, word embeddings, BERT for Indian languages and NLU resources for Indian languages.
- IIT Bombay English-Hindi Parallel Corpus: Largest en-hi parallel corpora in public domain (about 1.5 million semgents)
- CVIT-IIITH PIB Multilingual Corpus: Mined from Press Information Bureau for many Indian languages. Contains both English-IL and IL-IL corpora (IL=Indian language).
- CVIT-IIITH Mann ki Baat Corpus: Mined from Indian PM Narendra Modi's *Mann ki Baat* speeches.
- iNLTK: iNLTK aims to provide out of the box support for various NLP tasks that an application developer might need for Indic languages.
- Dakshina Dataset: The Dakshina dataset is a collection of text in both Latin and native scripts for 12 South Asian languages. Contains an aggregate of around 300k word pairs and 120k sentence pairs. Useful for transliteration.

Parallel Translation Corpus

- IIT Bombay English-Hindi Parallel Corpus: Largest en-hi parallel corpora in public domain (about 1.5 million semgents)
- CVIT-IIITH PIB Multilingual Corpus: Mined from Press Information Bureau for many Indian languages. Contains both English-IL and IL-IL corpora (IL=Indian language).
- CVIT-IIITH Mann ki Baat Corpus: Mined from Indian PM Narendra Modi's *Mann ki Baat* speeches.
- PMIndia: Parallel corpus for En-Indian languages mined from *Mann ki Baat* speeches of the PM of India (paper).
- Indian Language Corpora Initiative: Available on TDIL portal on request
- OPUS corpus
- WAT 2018 Parallel Corpus: There may significant overlap between WAT and OPUS.
- Charles University English-Hindi Parallel Corpus: This is included in the IITB parallel corpus.
- Charles University English-Tamil Parallel Corpus
- Charles University English-Odia Parallel Corpus v1.0
- Charles University English-Odia Parallel Corpus v2.0
- Charles University English-Urdu Religious Parallel Corpus
- IndoWordnet Parallel Corpus: Parallel corpora mined from IndoWordNet gloss and/or examples for Indian-Indian language corpora (6.3 million segments, 18 languages).
- MTurk Indian Parallel Corpus
- TED Parallel Corpus
- JW300 Corpus: Parallel corpus mined from jw.org. Religious text from Jehovah's Witness.
- ALT Parallel Corpus: 10k sentences for Bengali, Hindi in parallel with English and many East Asian languages.
- FLORES dataset: English-Sinhala and English-Nepali corpora
- Uka Tarsadia University Corpus: 65k English-Gujarati sentence pairs. Corpus is described in this paper
- NLPC-UoM English-Tamil Corpus: 9k sentences, 24k glossary terms

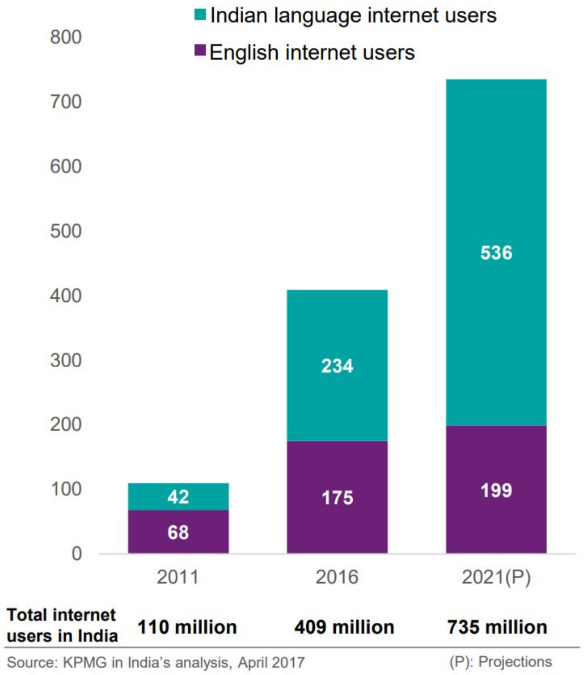
Where and Why do we need Indic NLP solutions today?

Usage and Diversity of Indian Languages



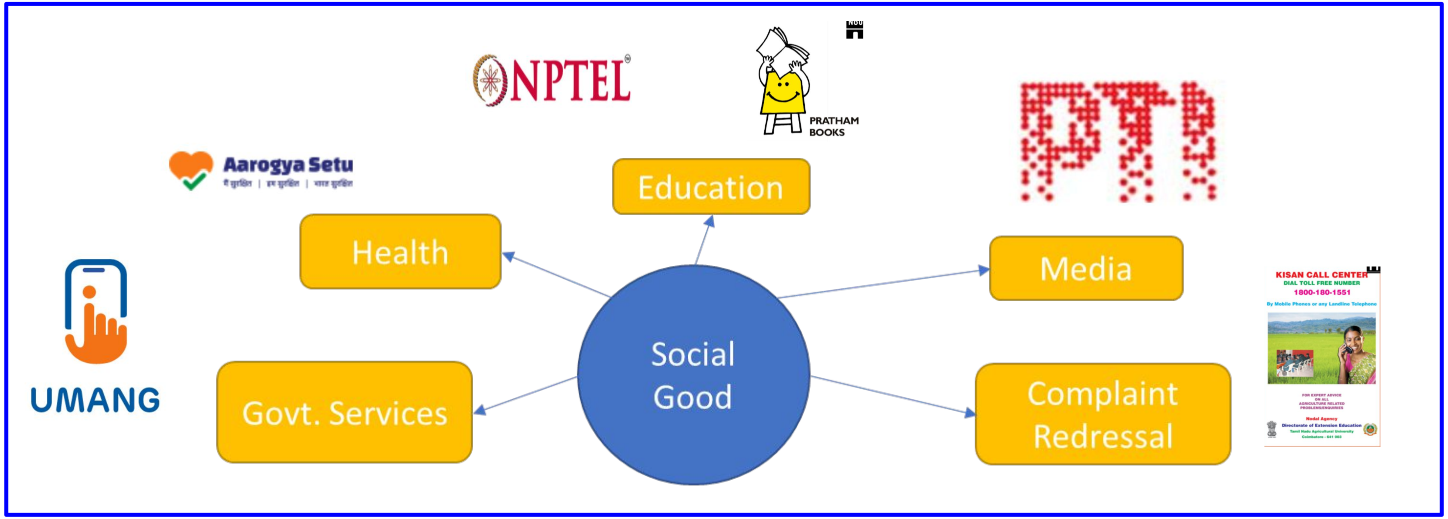
- *4 major language families*
- *22 scheduled languages*
- *125 million English speakers*
- *8 languages in the world's top 20 languages*
- *30 languages with more than 1 million speakers*

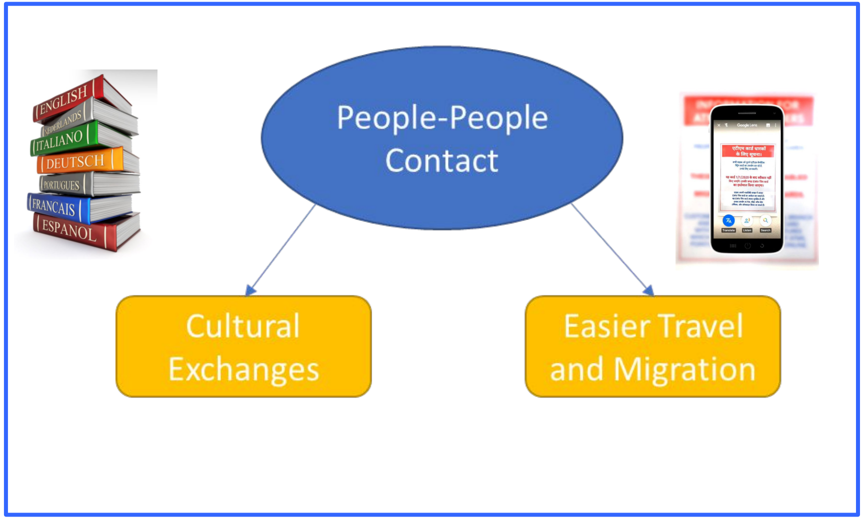
Sources: Wikipedia, Census of India 2011

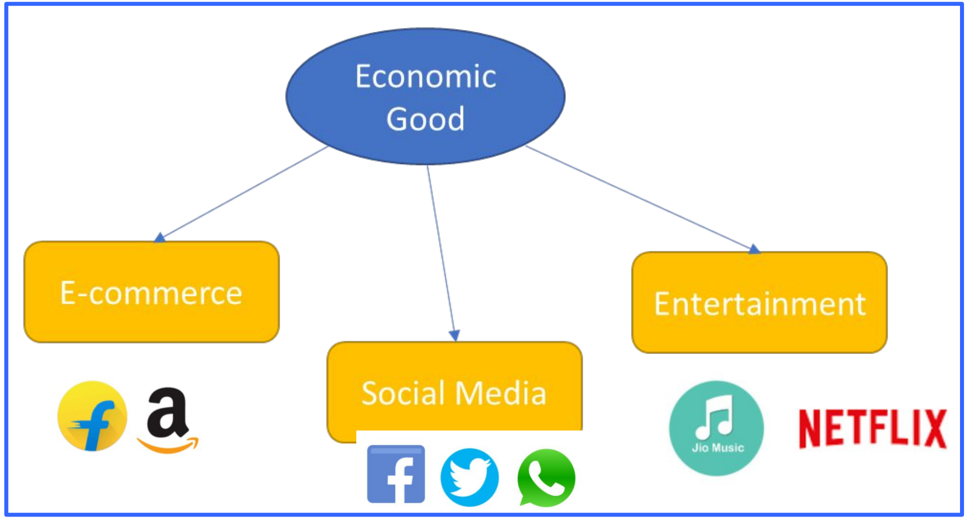


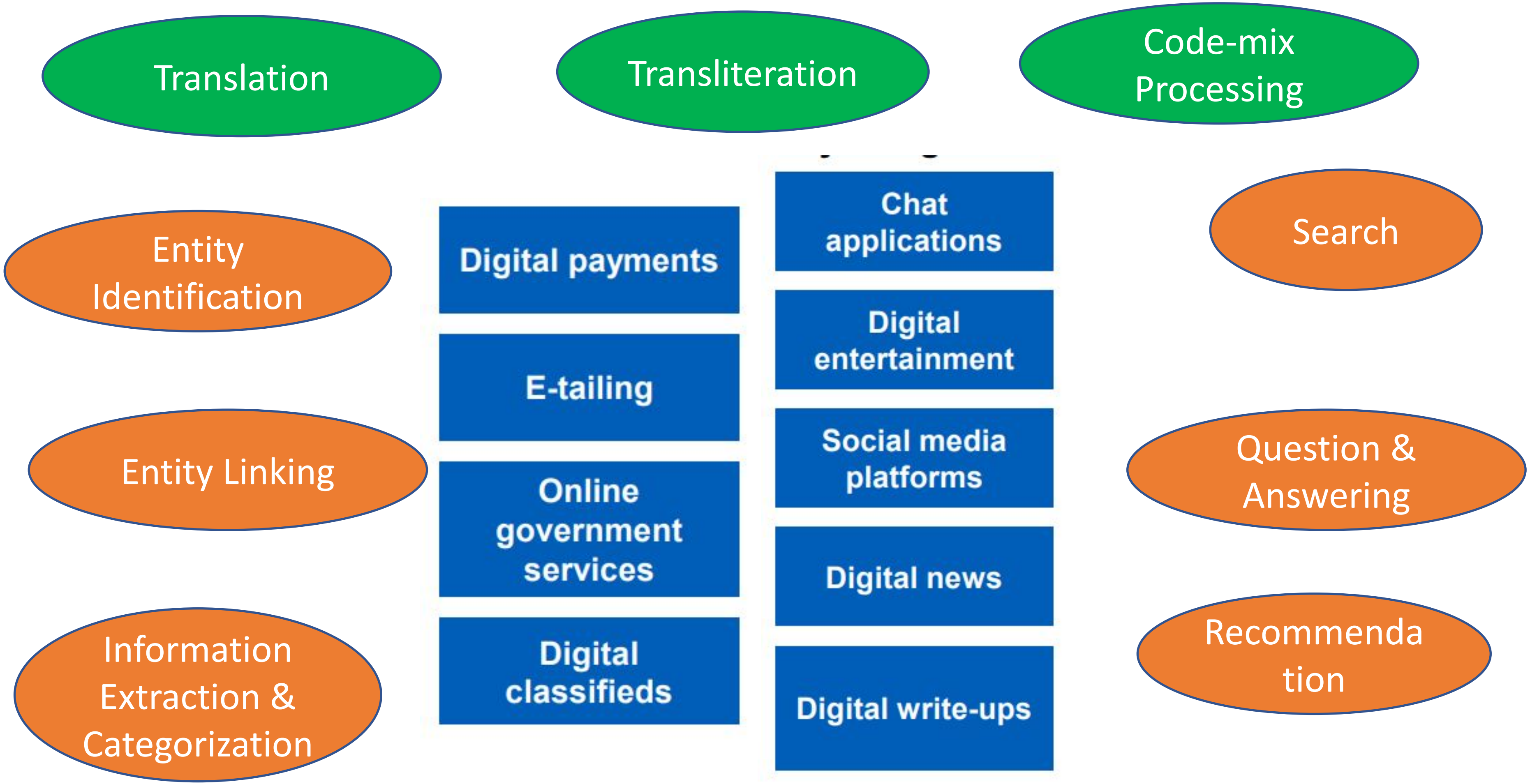
Internet User Base in India (in million)

Source: Indian Languages:
Defining India's Internet KPMG-Google Report 2017


NPTEL
PRATHAM BOOKS
Aarogya Setu
Education
Health
Media
UMANG
Social Good
Govt. Services
Complaint Redressal
KISAN CALL CENTER
DIAL TOLL FREE NUMBER
1800-180-1551
By Mobile Phones or any Landline Telephone

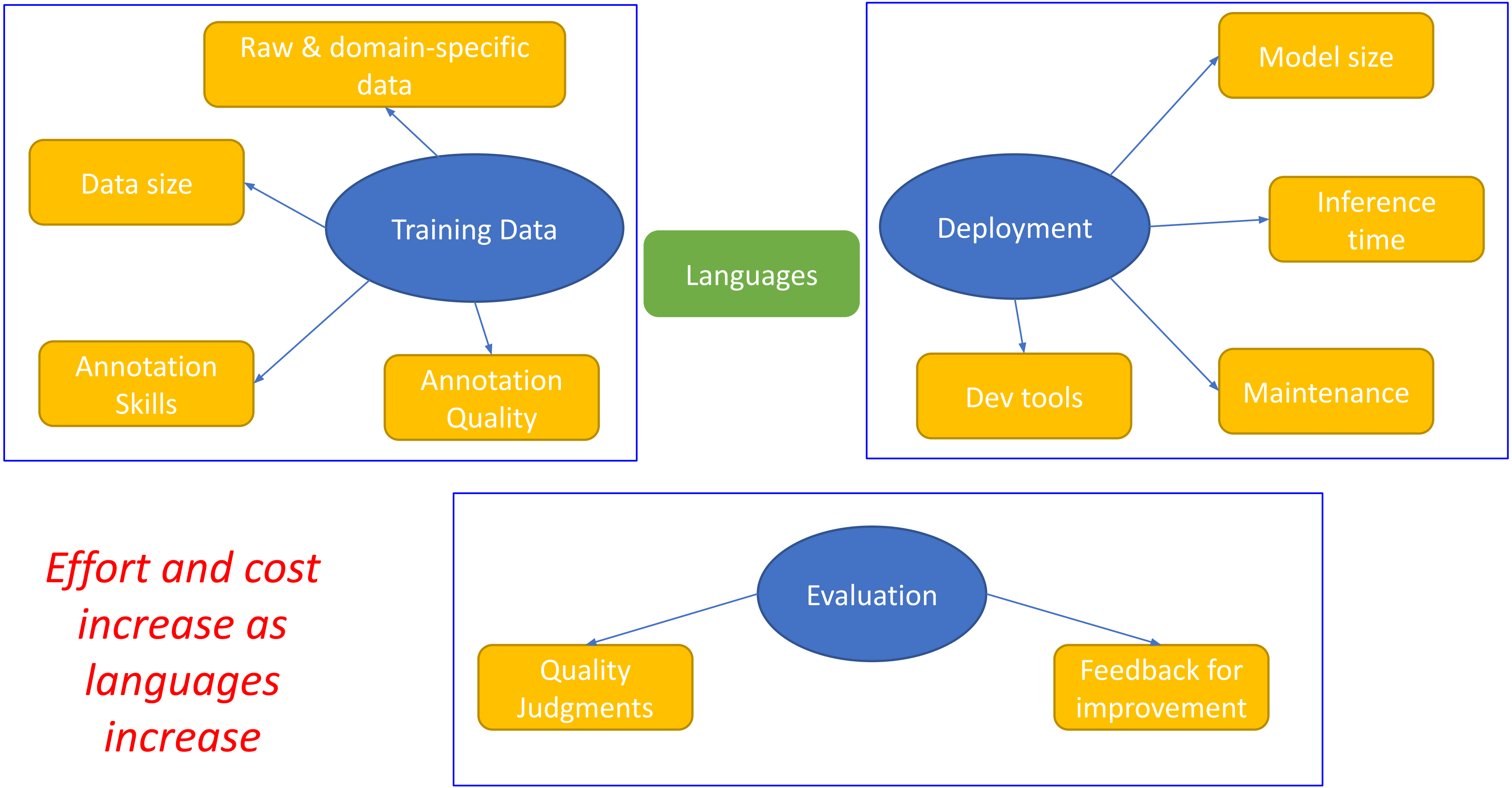

ENGLISH
ITALIANO
DEUTSCH
PORTUGUES
FRANCAIS
ESPANOL
People-People Contact
Cultural Exchanges
Easier Travel and Migration


Economic Good
E-commerce
Entertainment
Social Media
Jio Music
NETFLIX



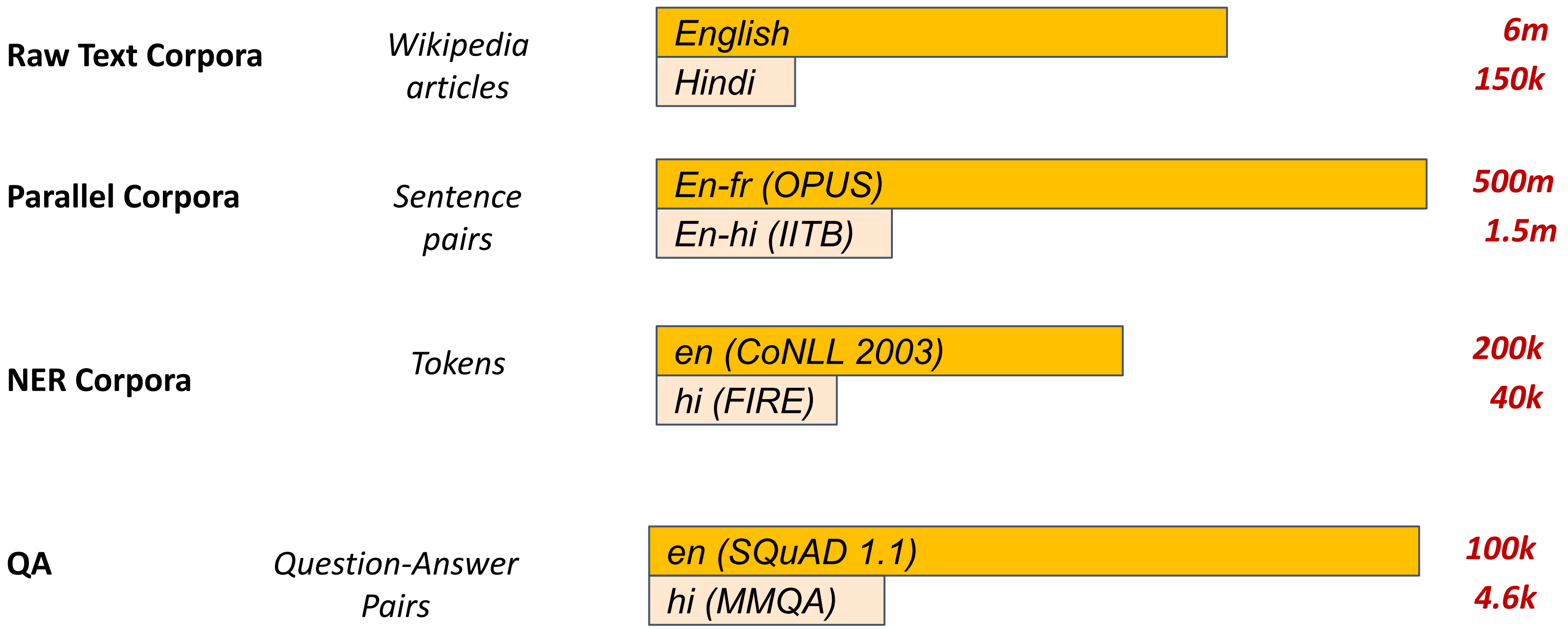
Applications requiring Indian language support

Scalability Challenges for NLP solutions



Effort and cost increase as languages increase

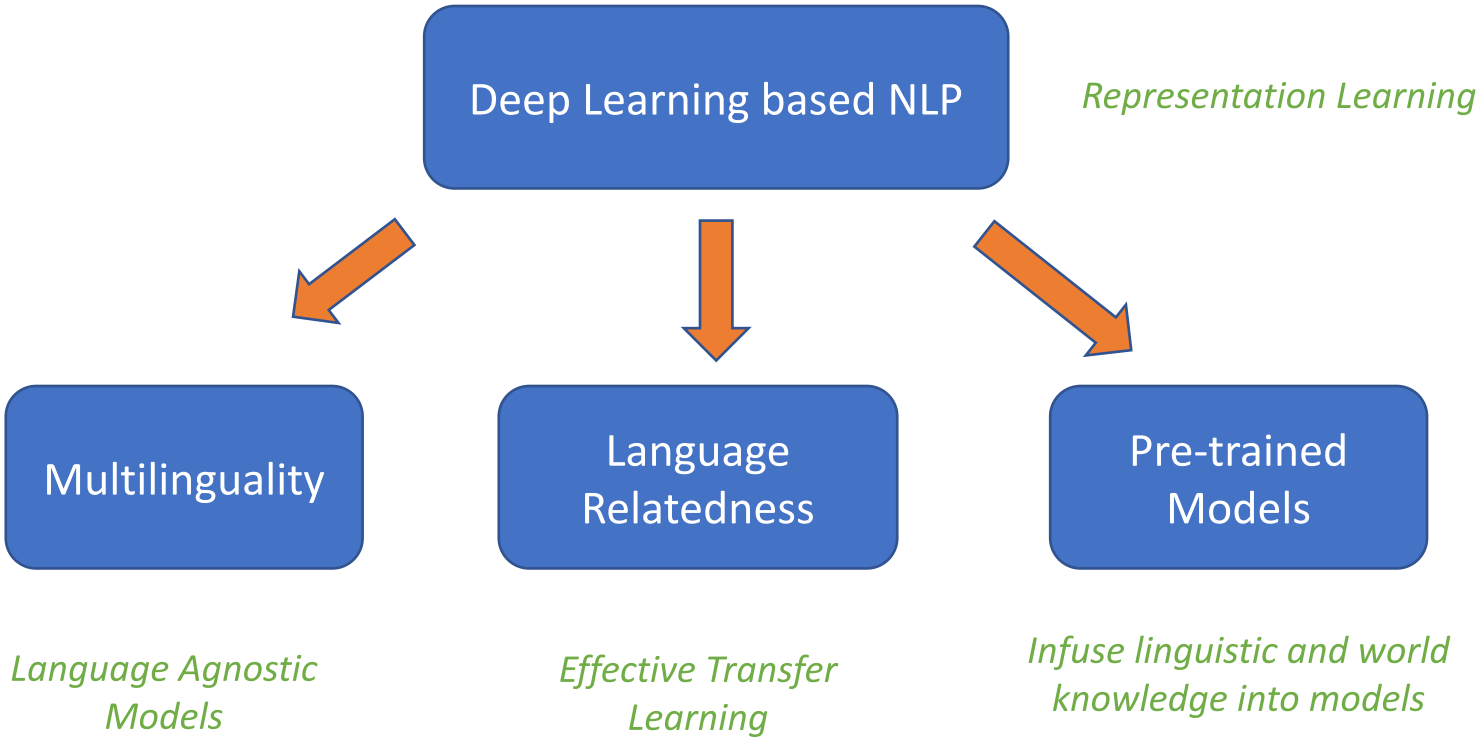
We are faced with a huge data skew



What is our approach?

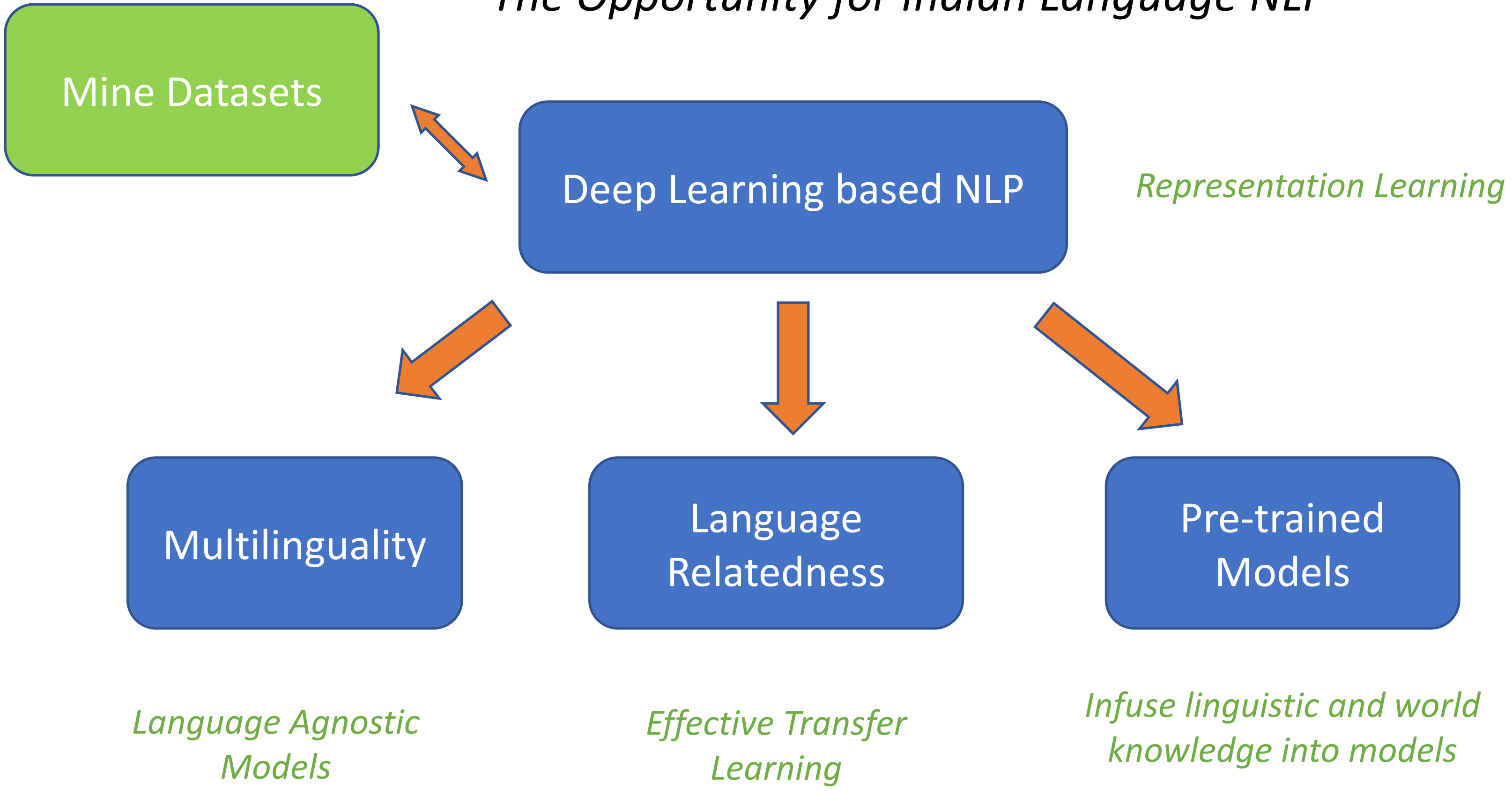
Our Technical Direction

The Opportunity for Indian Language NLP



Our Technical Direction

The Opportunity for Indian Language NLP



Representation Learning

Traditional feature engineering requires linguistic resources

Let us look at a simple NLP application – Sentiment Analysis

Kabir Singh Movie Review

Review by **Bollywood Hungama News Network**
20 June 2019 23:39 pm IST

4.0
Kabir Singh Movie Rating

Listen to this article in audio 00:00/00:00

One of the most loved love stories of Bollywood is DEVDAS. It has been remade several times and ten years ago, Anurag Kashyap gave a different touch to the tale through DEV D [2009]. All the interpretations have been liked as there's a charm in the story of a man who goes on a self-destructive path when he fails to get the girl he loves. Two years ago, Sandeep Reddy Vanga made a Telugu film named ARJUN REDDY, which had a kind of a deja vu of DEVDAS. Yet, it stood out due to the treatment, execution and performances. ARJUN REDDY became a cult success and now its Hindi remake KABIR SINGH is all set to hit theatres. So does KABIR SINGH turn out to be as good as or better than ARJUN REDDY? Or does it fail to stir the emotions of the viewers? Let's analyse.

?

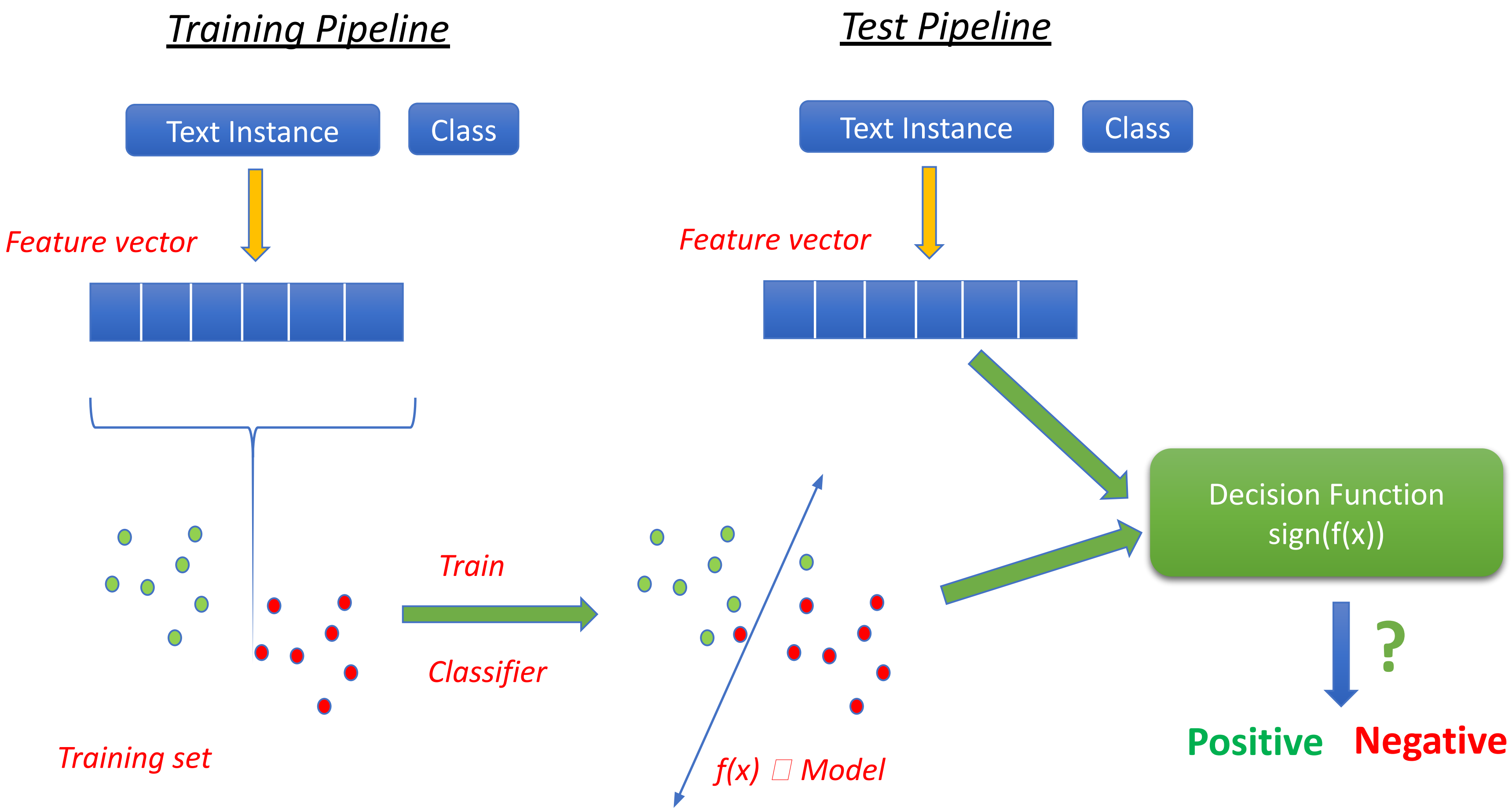
Positive

Negative

Neutral

An example of a text classification problem

A Machine Learning Pipeline for Text Classification



Simple Features

Bag-of-words (presence/absence)

Well-made	hit	script	lovely	boring	music
1	1	1	1	0	1

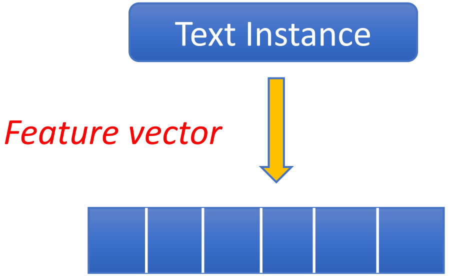
Large and sparse feature vector: size of vocabulary
Each feature is atomic similarity between features, synonyms not captured

More features

- Bigrams: e.g. *lovely_script*
- Presence in [positive/negative] sentiment word list
- Negation words
- Is the sentence sarcastic (output from saracasm classifier?)

- *These features have to be* ***hand-crafted manually*** *– repeat for domains and tasks*
- ***Need linguistic resources*** *like POS, lexicons, parsers for building features*
- *Can some of these features be discovered from the text in an unsupervised manner using raw corpora?*

Distributed Representations



Can we replace the
high-dimensional, resource-heavy document feature vector
with

- *low-dimensional vector*
- *learnt in an unsupervised manner*
- *subsumes many linguistic features*

Distributional Hypothesis

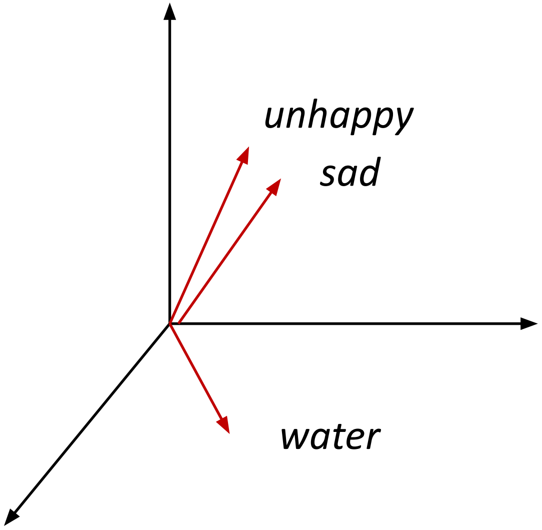
"A word is known by the company it keeps" *- Firth (1957)*

"Words that occur in similar contexts tend to have similar meanings"
- Turney and Pantel (2010)

He is unhappy about the failure of the project

The failure of the team to successfully finish the task made him sad

- The distribution of the context defines the word
- Can define notion of similarity based on contextual distributions

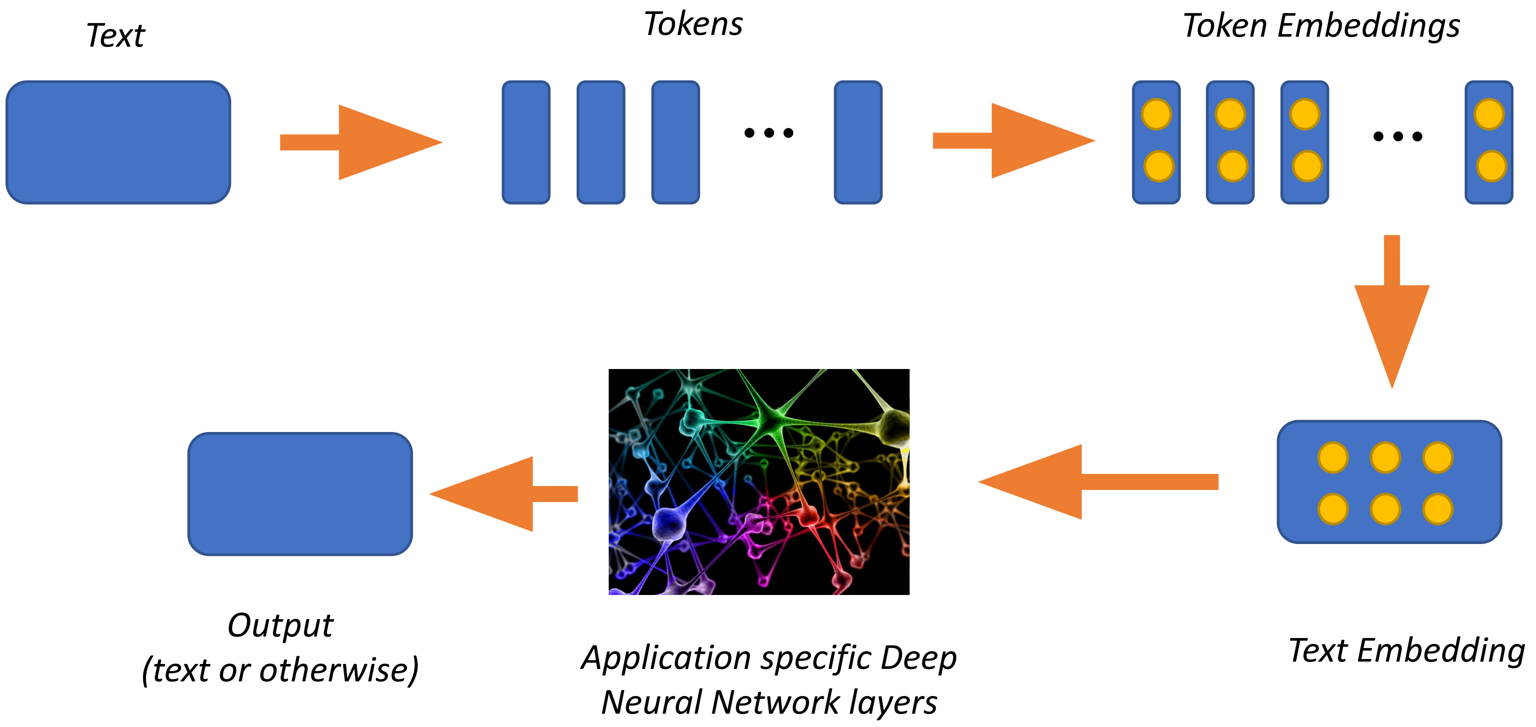


Similarity of words can be defined in terms of vector similarity: Cosine similarity, Euclidean distance, Mahalanobis distance

Similarity across languages

Contextual representation of words

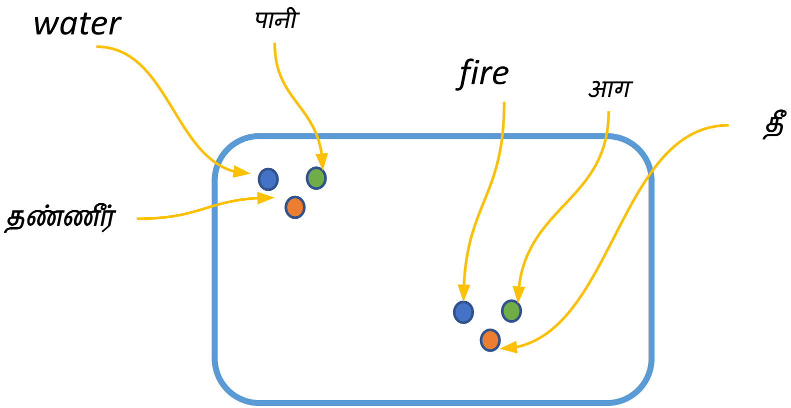
A Typical Deep Learning NLP Pipeline



Multilinguality

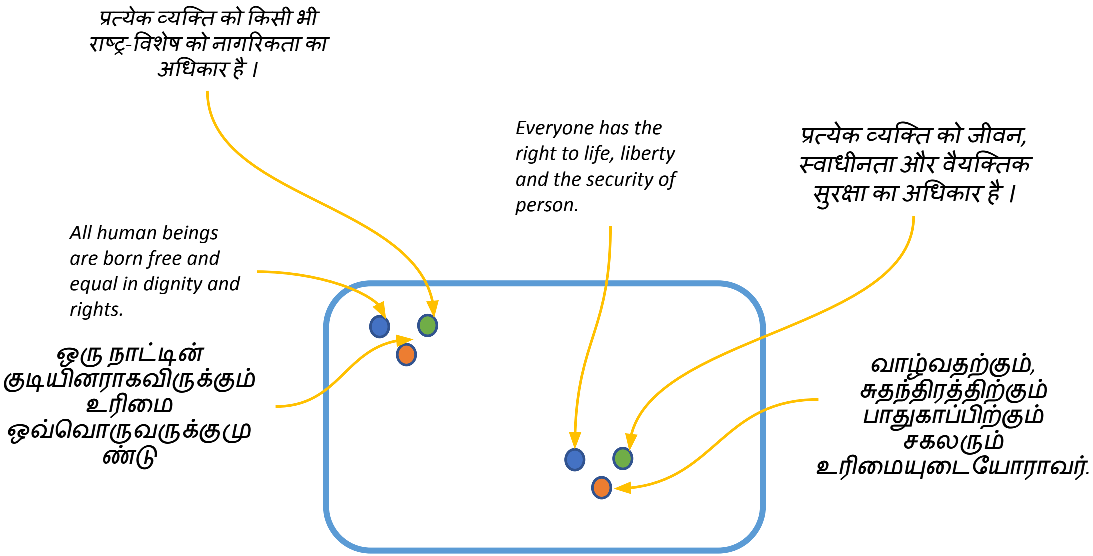
But we still need training data for each language?

Represent semantically similar language artifacts in the same vector space



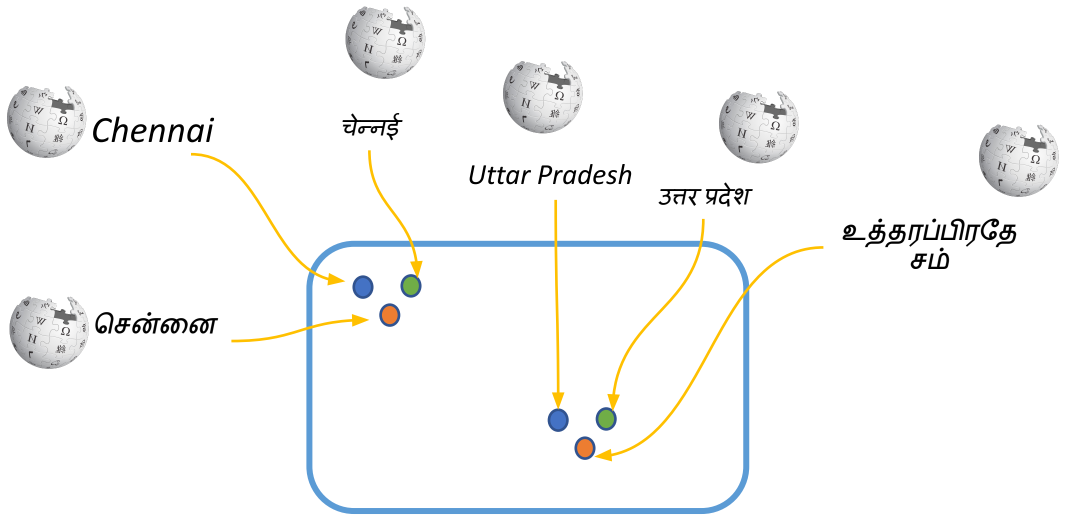
Words

Represent semantically similar language artifacts in the same vector space



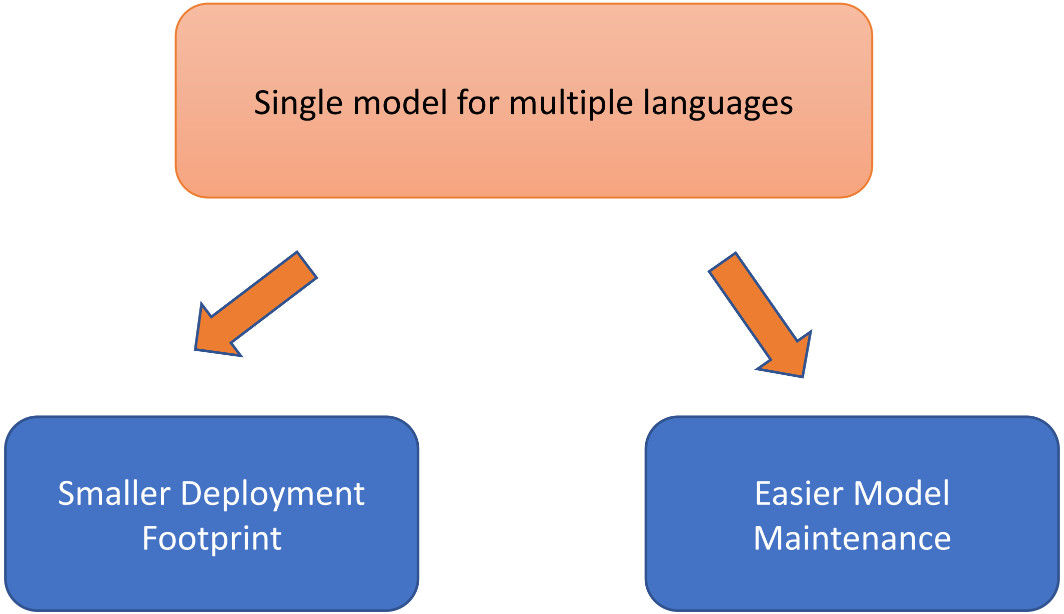
Sentences

Represent semantically similar language artifacts in the same vector space

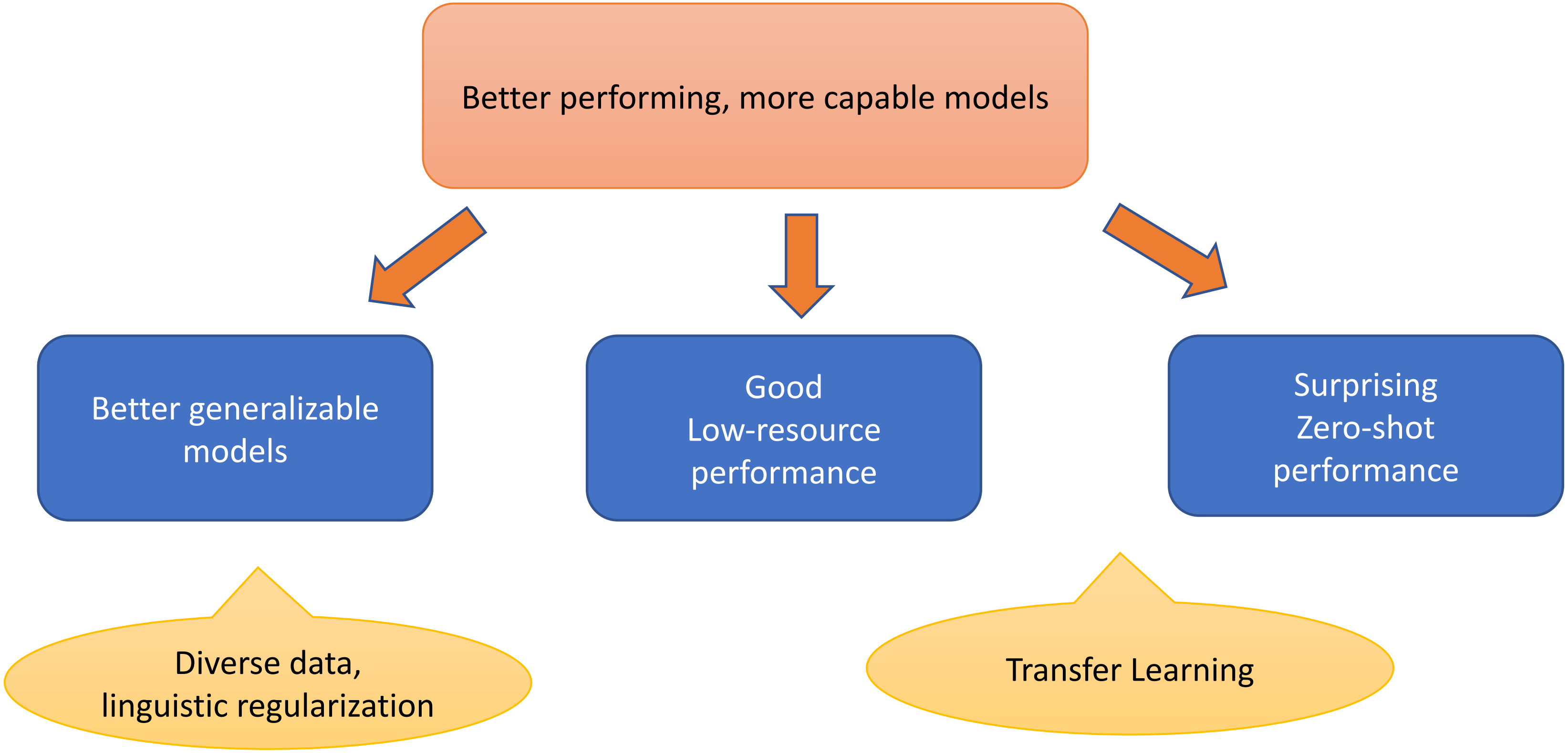


Documents

How does multilinguality help?

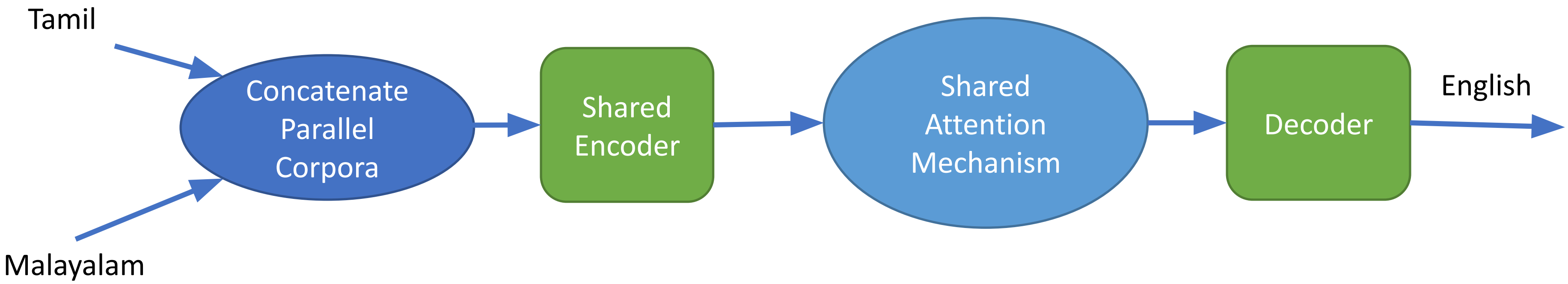


How does multilinguality help?



Multilingual Indian Language en Translation Models

(Zoph et al., 2016; Nguyen et al., 2017; Lee et al., 2017; Dabre et al., 2018)



We want Malayalam English translation but little parallel corpus is available
We have lot of Tamil English parallel corpus

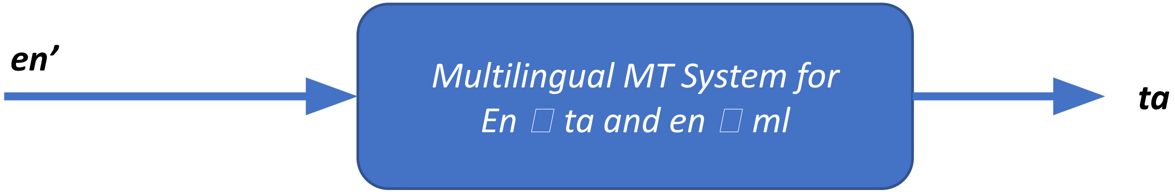
English □ Indian Languages

How do we support multiple target languages with a single decoder?

A simple trick!: Append input with special token indicating the target language

<u>Original Input</u>: *France and Croatia will play the final on Sunday*

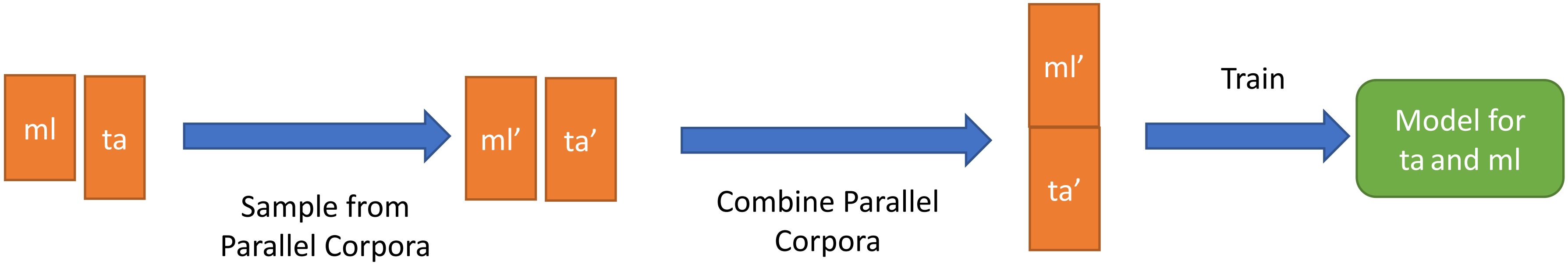
<u>Modified Input</u>: *France and Croatia will play the final on Sunday* *<ta>*



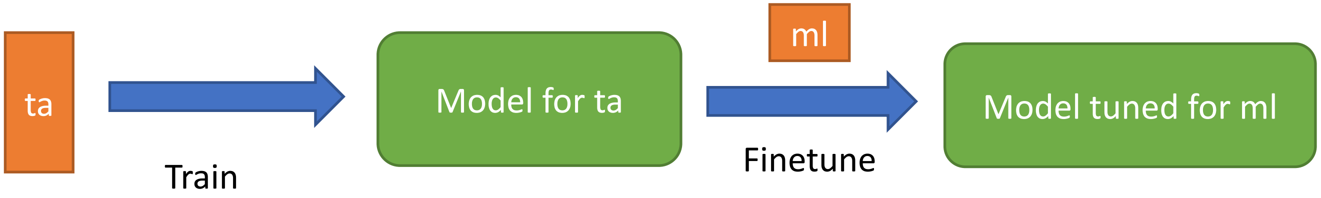
Still a challenging problem

Training Multilingual NMT systems

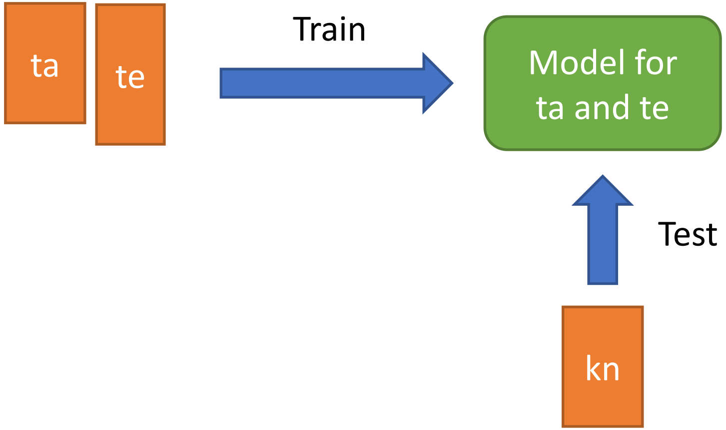
Joint Training



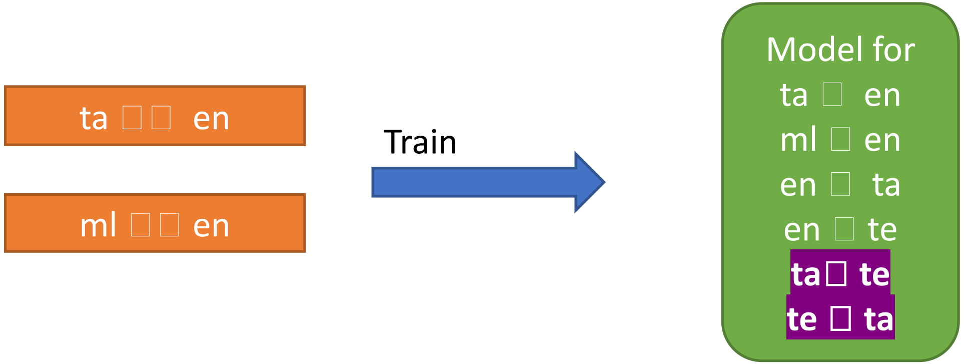
Transfer Learning



Zeroshot Translation into English



Zeroshot Translation between Indian languages



Language Relatedness

Why are Indian languages related?

Related Languages

Related by Genealogy

Related by Contact

__Language Families__

Dravidian, Indo-European, Turkic

(Jones, Rasmus, Verner, 18th & 19th centuries, Raymond ed. (2005))

__Linguistic Areas__

Indian Subcontinent, Standard Average European

(Trubetzkoy, 1923)

Related languages may not belong to the same language family!

Cognates & Borrowed words in Indian Languages

Indo-Aryan

English	Vedic Sanskrit	Hindi	Punjabi	Gujarati	Marathi	Odia	Bengali
bread	Rotika	chapātī, roṭī	roṭi	paũ, roṭlā	chapāti, poli, bhākarī	pauruṭi	(pau-)ruṭi
fish	Matsya	Machhlī	machhī	māchhli	māsa	mācha	machh
hunger	bubuksha, kshudhā	Bhūkh	pukh	bhukh	bhūkh	bhoka	khide

Dravidian

English	Tamil	Malayalam	Kannada	Telugu
fruit	pazham , kanni	pazha.n , phala.n	haNNu , phala	pa.nDu , phala.n
ten	pattu	patt,dasha.m,dashaka.m	hattu	padi

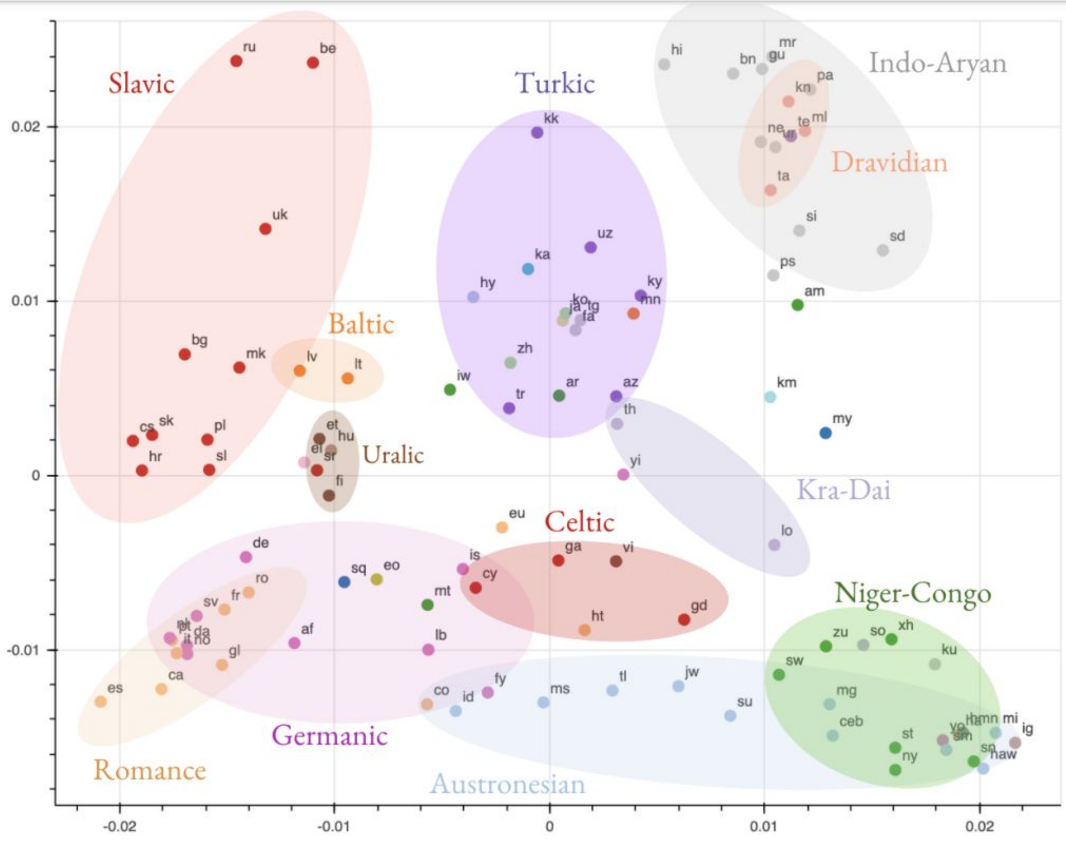
Indo-Aryan words in Dravidian languages

Sanskrit word	Language	Loanword	English
cakram	Tamil	cakkaram	wheel
matsyah	Telugu	matsyalu	fish
ashvah	Kannada	ashva	horse
jalam	Malayalam	jala.m	water

Other borrowings like echo words, retroflex sounds in other direction. (Subbarao, 2012)

Source: Wikipedia and IndoWordNet

Transfer Learning works best for related languages



Transformer models are powerful enough to learn multilingual representation □
but similarity priors (natural or induced) help

Motivation for:

- *Building multilingual systems systems specific to language families*
- *Transfer learning from a related parent*

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

Key Similarities between related languages

On the occasion of India's Independence day, a programme was organized in American city of Los Angeles

भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला *bhAratAcyA svAta.ntryadinAnimitta ameriketIla lOsa enjalsa shaharAta kAryakrama Ayojita karaNyAta AlA*	*Marathi*
भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला *bhAratA cyA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA*	*Marathi segmented*
भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया *bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsa shahara me.n kAryakrama Ayojita kiyA gayA*	*Hindi*

Lexical: share significant vocabulary (cognates & loanwords)

Morphological: correspondence between suffixes/post-positions

Syntactic: share the same basic word order

Orthographic Similarity

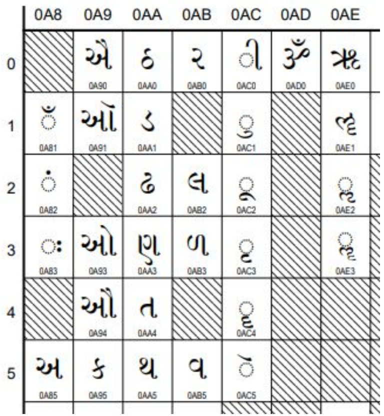
Brahmi-derived Indic scripts are orthographically similar

- Largely overlapping character set, but the visual rendering differs
- *highly overlapping phoneme sets*
- Highly consistent grapheme-to-phoneme mapping

Script Conversion

- Read any script in any script
- Unicode standard enables consistent script conversion with a single rule

$$\textit{unicode\_codepoint(char) - Unicode\_range\_start}(L_1) + \textit{Unicode\_range\_start}(L_2)$$



केरला kerala

কেরলা કેરલા

As a developer, you can read text in a script you understand

Only a single mapping needed for Romanization too

Indian Language Speech sound Label set *(Samudravijaya & Murthy, 2012)*

A simple and powerful property to utilize

relatedness between Indian languages

***Pre-requisite to Neural Transfer Learning*: Represent all data in a common script**

Multilingual Transliteration

(Kunchukuttan, et al, 2018;2021)

Pool training sets

Malayalam	കോഴിക്കോട്	kozhikode
Hindi	केरल	kerala
Kannada	ಬೆಂಗಳೂರು	bengaluru

Convert to a common script

Malayalam	कोऴिक्कोट्	kozhikode
Hindi	केरल	kerala
Kannada	बेंगळूरु	bengaluru

Train a joint transliteration model for multiple Indian languages to English & vice-versa

Example of Multi-task Learning

Similar tasks help each other

Zero-shot transliteration is possible

Perform Telugu English transliteration even if network has not seen that data

On the other hand, we cannot pool Hindi and Urdu data
Though they are pretty much the same language The scripts are very different

Traditionally organized as per sound phonetic principles

shows various symmetries

- *Useful for unsupervised transliteration*

Primary vowels

1

	Short				Long				Diphthongs			
	Initial		Diacritic		Initial		Diacritic		Initial		Diacritic	
Unrounded low central	अ	a	प	pa	आ	ā	पा	pā				
Unrounded high front	इ	i	पि	pi	ई	ī	पी	pī				
Rounded high back	उ	u	पु	pu	ऊ	ū	पू	pū				
Syllabic variants	ऋ	ṛ	पृ	pṛ	ॠ	ṝ	पॄ	pṝ				
	ऌ	ḷ	पॢ	pḷ	ॡ	ḹ	पॣ	pḹ				
Secondary vowels												
Unrounded front					ए	e	पे	pe	ऐ	ai	पै	pai
Rounded back					ओ	o	पो	po	औ	au	पौ	pau

3

Occlusives

2

	Voiceless plosives				Voiced plosives				Nasals	
	unaspirated		aspirated		unaspirated		aspirated			
Velar	क	ka	ख	kha	ग	ga	घ	gha	ङ	ṅa
Palatal	च	ca	छ	cha	ज	ja	झ	jha	ञ	ña
Retroflex	ट	ṭa	ठ	ṭha	ड	ḍa	ढ	ḍha	ण	ṇa
Dental	त	ta	थ	tha	द	da	ध	dha	न	na
Labial	प	pa	फ	pha	ब	ba	भ	bha	म	ma

4

5

Sonorants and fricatives

6

	Palatal		Retroflex		Dental		Labial	
Sonorants	य	ya	र	ra	ल	la	व	va
Sibilants	श	śa	ष	ṣa	स	sa		

Other letters

ह	ha	ळ	ḷa

Lexical Similarity

Lexical Similarity
*(Words having similar **form** and **meaning**)*

- *Cognates*

  a common etymological origin

roTI (hi)	*roTIA (pa)*	*bread*
bhai (hi)	*bhAU (mr)*	*brother*

- *Loan Words*

  borrowed without translation

matsya (sa)	*matsyalu (te)*	*fish*
pazha.m (ta)	*phala (hi)*	*fruit*

- *Named Entities*

  do not change across languages

mu.mbaI (hi)	mu.mbaI (pa)	mu.mbaI (pa)
keral (hi)	k.eraLA (ml)	keraL (mr)

- *Fixed Expressions/Idioms*

  MWE with non-compositional semantics

dAla me.n kuCha kAlA honA	(hi)	
		Something fishy
dALa mA kAIka kALu hovu	(gu)	

Enables sharing of data across languages

Why it matters

भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला

bhAratA cyA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA

भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया

bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsa shahara me.n kAryakrama Ayojita kiyA gayA

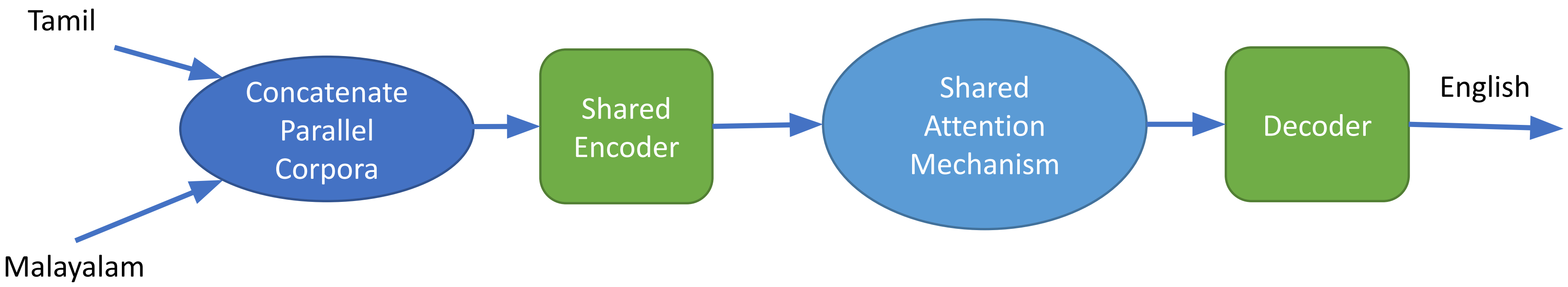
Lexical Overlap Representation overlap
Makes it easier for the model to learn

On the occasion of India's Independence day, a programme was organized in American city of Los Angeles

Multilingual Indian Language en Translation Models

(Zoph et al., 2016; Nguyen et al., 2017; Lee et al., 2017; Dabre et al., 2018; Ramesh et al 2021;)

We want Malayalam English translation but little parallel corpus is available
We have lot of Tamil English parallel corpus



- *Train models at the subword-level (BPE etc).*
- *Represent data in a common script*

Similar trends for NLU and language models: Khemchandani et al. (2021), Dhamecha et al. (2021)

Syntactic Similarity

भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला

bhAratA cyA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA

भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया

bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsa shahara me.n kAryakrama Ayojita kiyA gayA

Syntactic Divergence Makes it more difficult for the model to learn common representations

India ke Independence day ke **occasion** par america ke los angeles city me programme organize kiya gaya

On the **occasion** of India's Independence day, a programme was organized in American city of Los Angeles

Source reordering for SMT

(Kunchukuttan et al., 2014)

Change order of words in input sentence to match word order in the target language

Bahubali earned more than 1500 crore rupees at the boxoffice

↓

Bahubali the boxoffice at 1500 crore rupees earned

बाहुबली ने बॉक्सओफिस पर 1500 करोड रुपए कमाए

	Indo-Aryan				
	pan	**hin**	**guj**	**ben**	**mar**
Baseline	15.83	21.98	15.80	12.95	10.59
Generic	17.06	23.70	16.49	13.61	11.05
Hindi-tuned	**17.96**	**24.45**	**17.38**	**13.99**	**11.77**

A common set of rules can be written for all Indian languages

Rules from (Ramanathan et al. 2008, Patel et al. 2013) for Hindi.

https://github.com/anoopkunchukuttan/cfilt_preorder

Language Relatedness can be successfully utilized between languages where contact relation exists

Experiment	BLEU
Baseline	12.91
+ Hindi as helper language	**16.25**

Tamil to English NMT with transfer-learning using Hindi

Pre-trained Models

Representation Learning

Automatic Feature Extraction
Continuous Space Representation
Numerical Optimization at disposal

Multilingual learning

Transfer Learning
Better generalizability across languages

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

Pre-trained Models

Task-independent models that know about language

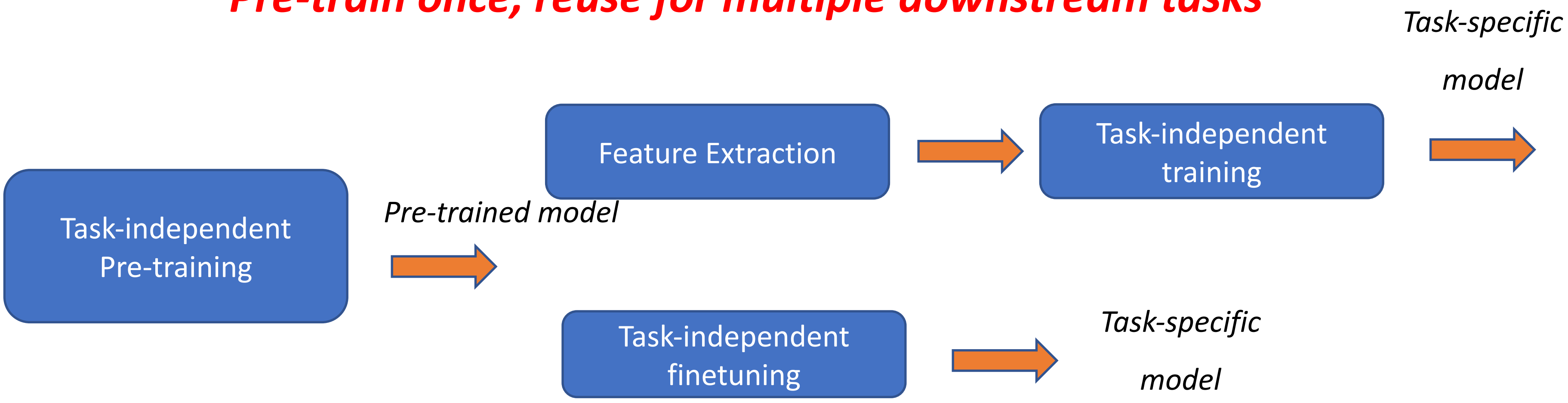
Word Embeddings	Encoder Language Model for NLU	Decoder Language Model for NLG	Encoder-decoder Language Model for NLU+NLG
fastText	Google BERT	OpenAI GPT-3	FAIRSEQ BART
	+ Multilinguality		
MUSE	mBERT		mBART

Trained on a large amount of raw text corpora with unsupervised objectives

Language models are

- *computationally intensive to train*
- *trained on a large amount of raw text corpora*
- *giant models*

Pre-train once, reuse for multiple downstream tasks



Only task-specific training: less data & less computation

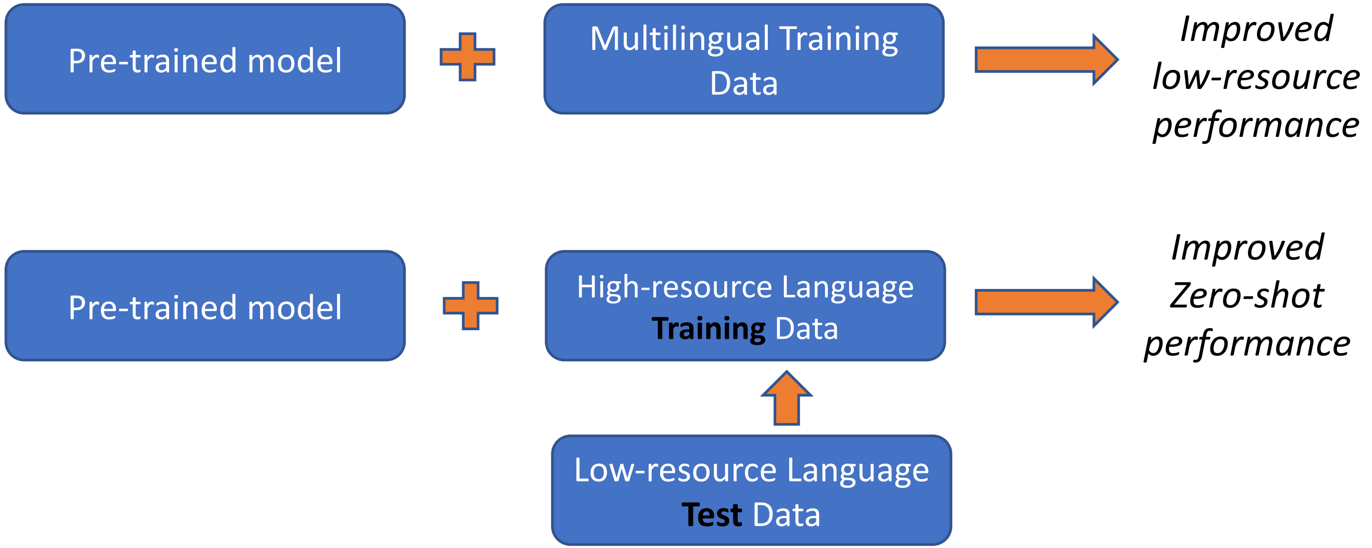
Language understanding for tasks like sentiment analysis, question answering, paraphrase detection

Language modeling & Language generation for tasks like summarization, ASR, question generation

Multi-linguality and Pre-training are complementary

Language-family specific pre-trained model

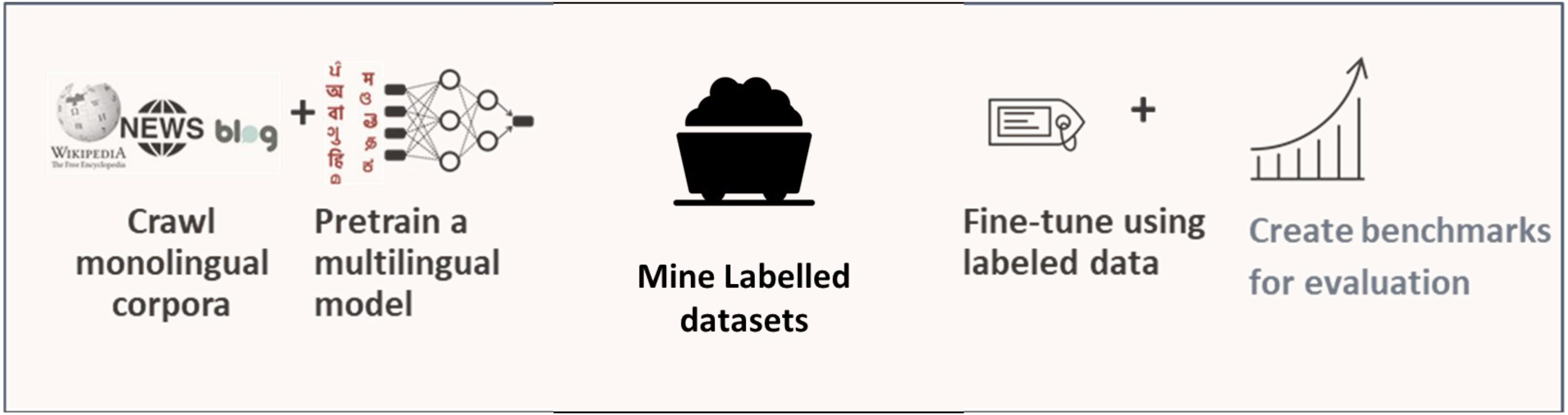
- *Compact pre-trained models*
- *Utilize language relatedness*
- *Better data representation*



Summary

- Deep Learning presents a unique opportunity to build NLP technologies at scale for Indian languages
- Utilizing language relatedness is important to this mission
- The orthographic similarity of Indian languages is a strong starting point for utilizing language relatedness.
- Contact as well as genetic relatedness are useful in the context of Indian languages.
- Multilingual pre-trained models trained on large corpora needed for transfer learning in NLU and NLG tasks.

Our Approach



Some Projects

IndicNLPSuite

Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages

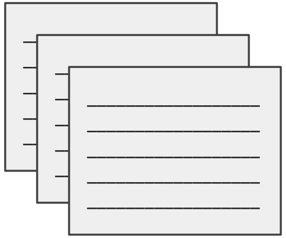
Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, Pratyush Kumar

AI4Bharat, IITM, Microsoft, RBCDSAI,

EMNLP Findings 2020

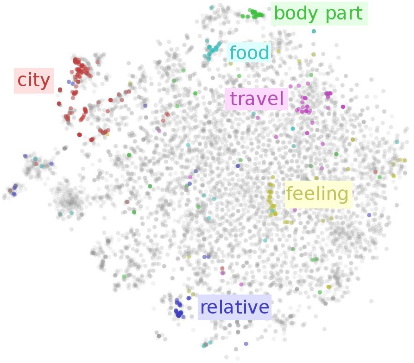
IndicNLPSuite

Monolingual Corpora



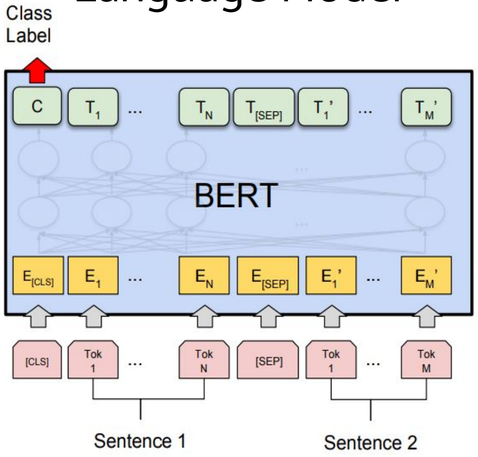
IndicCorp

Embeddings



IndicFT

Language Model



IndicBERT

NLU Benchmark

IndicGLUE

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, Pratyush Kumar. *IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages*. Findings of EMNLP. 2020.

IndicCorp

https://indicnlp.ai4bharat.org/corpora

11 Indic languages

(+Indian English)

8.8B tokens

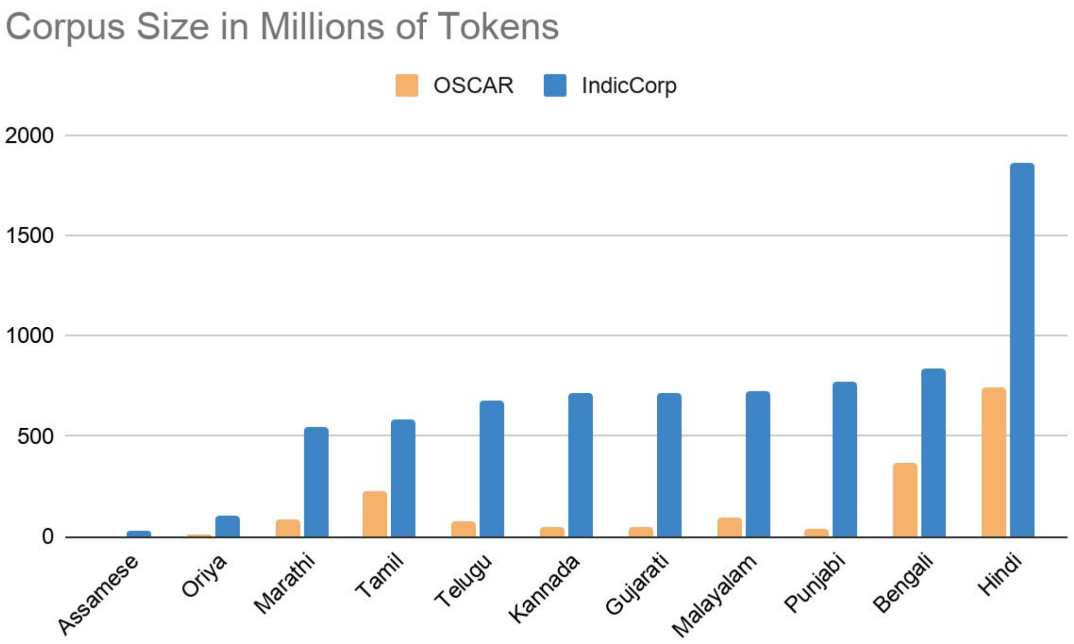
450M sentences

57M pages

General domain

1000+ Sources

~6 months of crawl



9x increase, **Largest Corpora**

Models

IndicBERT

IndicBART

n-gram LM

IndicWav2Vec

MT Models

IndicCorp is a central resource

Mined Datasets

Parallel Translation Corpus

Parallel Transliteration Corpus

NER Corpus

Text Classification

Language Generation

Benchmark Datasets

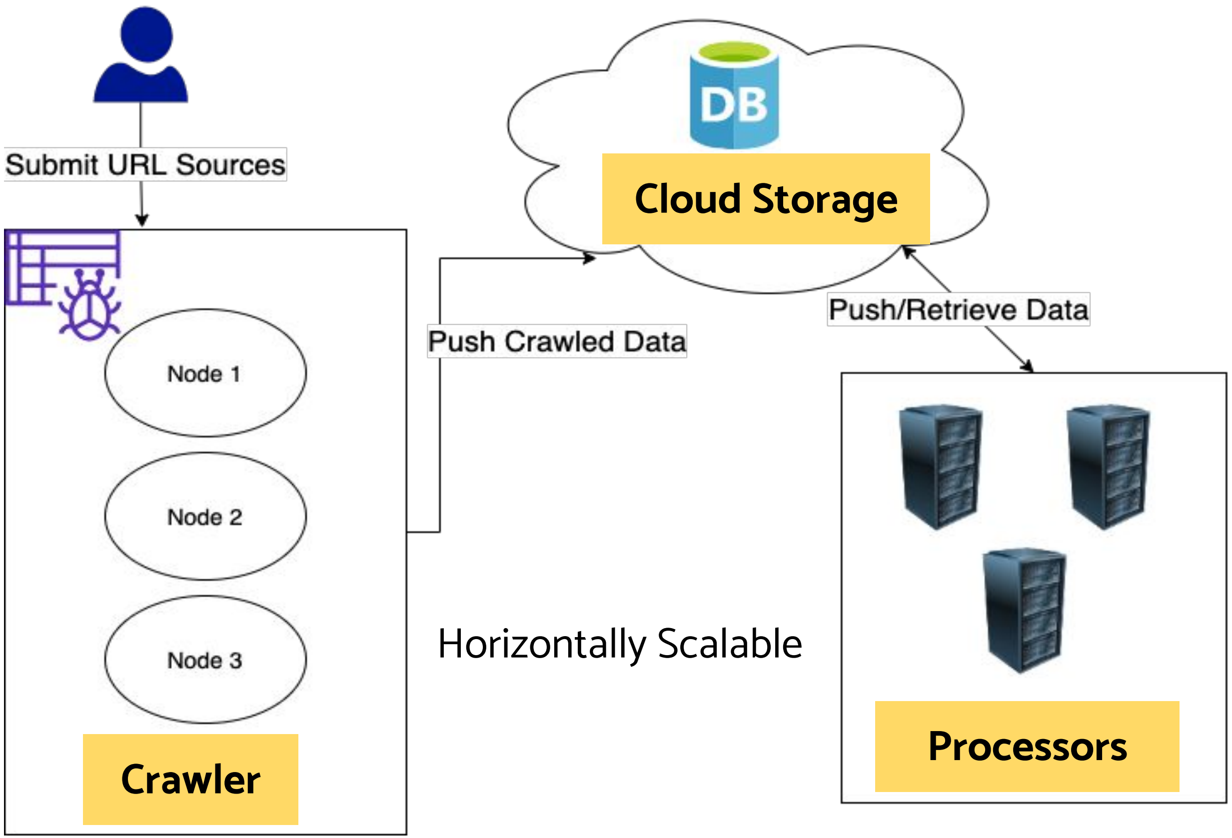
Webcorpus

(a scalable web crawler)

https://github.com/AI4Bharat/webcorpus

Distributed,
Multi-threaded

Dashboard



Processing HTML Pages to Get Sentences

https://github.com/AI4Bharat/webcorpus

HTML Page → (1.) Extracted Article → (2., 3.) Extracted Sentences

<html>
<head>...</head>
<body>...</body>
</html>

RCB batting coach Sanjay Bangar, ahead of their contest against a struggling Punjab Kings...

RCB appear to be a well-oiled machine in this season of IPL 2021

There are multiple players who on their day can win the match on their own

Boilerpipe
Custom Extractors

IndicNLP library
Filters (language, length,script)

IndicGLUE *(Indic General Language Understanding Evaluation Benchmark)*

Task Type	Task	N	Languages
Classification	News Article Classification	10	bn, gu, hi, kn, ml, mr, or, pa, ta, te
	Headline Classification	4	gu, ml, mr, ta
	Sentiment Analysis	2	hi, te
	Discourse Mode Classification	1	hi
Diagnostics	Winograd Natural Language Inference	3	gu, hi, mr
	Choice of Plausible Alternatives	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
	Cloze-style Question Answering	11	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te
	Paraphrase Detection	4	hi, ml, pa, ta
Sequence Labelling	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

IndicGLUE

New tasks

Difficult tasks

Span all languages

Task Type	Task	N	Languages
Classification	News Article Classification	10	bn, gu, hi, kn, ml, mr, or, pa, ta, te
	Headline Classification	4	gu, ml, mr, ta
	Sentiment Analysis	2	hi, te
	Discourse Mode Classification	1	hi
Diagnostics	Winograd Natural Language Inference	3	gu, hi, mr
	Choice of Plausible Alternatives	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
	Cloze-style Question Answering	11	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te
	Paraphrase Detection	4	hi, ml, pa, ta
Sequence Labelling	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

Creation of IndicGLUE

Lack of Evaluation Datasets !

Data Sources
News Crawls
Wikipedia
Public Datasets

Unsupervised Mining

Task 1

Task 2

Task 3

IndicGLUE Tasks

6 Tasks

4 Types

Semantic	
News Articles Headline Prediction	**News Crawls**
Wikipedia Section Title Prediction	**Wikipedia**
Article Genre Classification	**News Crawls**
Knowledge	
Cloze-style multiple-choice QA	**Wikipedia**
Syntax	
Named Entity Recognition	**Public Dataset**
Cross-lingual	
Cross-Lingual Sentence Retrieval	**Public Dataset**

Additional Tasks (Paraphrase Detection, Movie Reviews etc.)

IndicGLUE: News Article Headline Prediction

Created From: News Crawls

Task: Predict the correct headline

IPL 2021: Australian Cricketers, Support Staff Expected To Head To Maldives

-ve

With their country shut for all those flying from India, the now-suspended IPL's Australian contingent, comprising players, support staff and commentators, is expected to head to Maldives before taking a connecting flight for home. The IPL was "indefinitely suspended" on Tuesday after multiple cases of COVID-19 emerged from Kolkata Knight Riders, Delhi Capitals, SunRisers Hyderabad and Chennai Super Kings. There are 14 Australian players along with coaches and commentators who might now take a detour as the Australian government has imposed strict sanctions for people returning from India.

IPL 2021: Mayank Agarwal's 99* In Vain As Delhi Capitals Thrash Punjab Kings To Go Top Of The Table

+ve

Shikhar Dhawan's delightful 69 dwarfed Mayank Agarwal's unbeaten 99 as Delhi Capitals defeated Punjab Kings by seven wickets in the IPL, on Sunday to go atop the points table. Agarwal, leading the side in the absence of regular skipper K L Rahul, used the straight bat effectively in his lone hand to take Punjab Kings to 166 for six. Delhi Capitals hardly broke a sweat in the run chase, cantering to victory in 17.4 overs overs, their sixth win in eight matches.

Input

Careful Negative Sampling

SRH vs MI, IPL 2021: SunRisers Hyderabad Players To Watch Out For

-ve

Bottom-placed SunRisers Hyderabad take on a high-flying Mumbai Indians team at the Arun Jaitley Stadium in Delhi on Tuesday. SunRisers Hyderabad have had a torrid time in IPL 2021 so far, winning a solitary game after playing seven matches. They have just two

Sri Lanka All-Rounder Thisara Perera Bids Adieu To International Cricket

-ve

Sri Lankan all-rounder Thisara Perera, on Monday, announced his retirement from international cricket with immediate effect. In a letter to Sri Lanka Cricket (SLC), Perera said that he wanted to focus on his family, before adding that it was the right time for him

IndicGLUE: Cloze-style multiple-choice QA

Created From: Wikipedia **Task**: Predict the masked entity

Homi Bhabha was born in 1949 in Mumbai to a Parsi family. After receiving his early education at St. Mary's, he went on to graduate from Bombay University . He then moved to [MASK] for higher education . He received his MA and M.Phil degrees from Oxford University .

Candidate 1: Britain [correct answer]
Candidate 2: India
Candidate 3: Chicago
Candidate 4: Pakistan

IndicGLUE: Article Genre Classification

Created From: News Crawl

Task: Predict the genre of news article

IPL 2021: Mayank Agarwal's 99* In Vain As Delhi Capitals Thrash Punjab Kings To Go Top Of The Table

Shikhar Dhawan's delightful 69 dwarfed Mayank Agarwal's unbeaten 99 as Delhi Capitals defeated Punjab Kings by seven wickets in the IPL, on Sunday to go atop the points table. Agarwal, leading the side in the absence of regular skipper K L Rahul, used the straight bat effectively in his lone hand to take Punjab Kings to 166 for six. Delhi Capitals hardly broke a sweat in the run chase, cantering to victory in 17.4 overs overs, their sixth win in eight matches.

Category: Sports

=> Mined from URL

IndicFT https://indicnlp.ai4bharat.org/indicft

- *Pre-trained word embeddings trained with FastText.*
- ***300 dimension vectors, suitable for morphologically rich languages.***
- *Outperforms embeddings from the FastText project on word analogy, similarity and classification tasks.*

Lang	**FT-W**	**FT-WC**	***IndicFT***
Word Similarity (*Pearson Correlation*)			
pa	**0.467**	0.384	0.445
hi	0.575	0.551	**0.598**
gu	0.507	0.521	**0.600**
mr	0.497	**0.544**	0.509
te	0.559	0.543	**0.578**
ta	**0.439**	0.438	0.422
Average	0.507	0.497	**0.525**
Word Analogy (*% accuracy*)			
hi	19.76	**32.93**	29.65

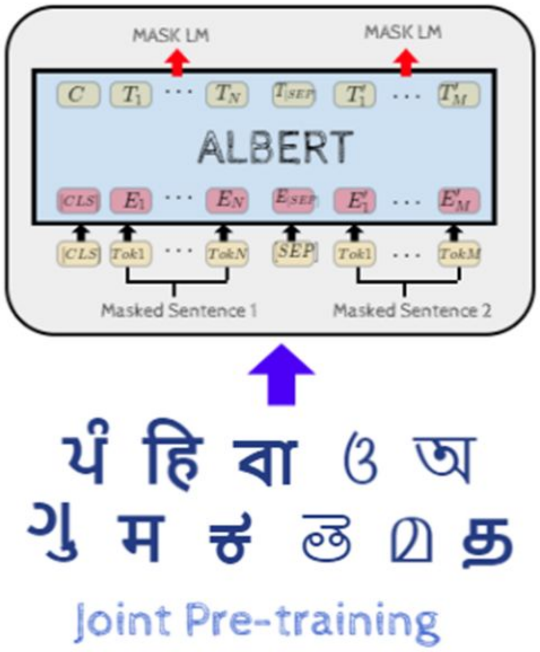
Lang	**Dataset**	**FT-W**	**FT-WC**	***IndicFT***
hi	BBC Articles	72.29	67.44	**77.02**
	IITP+ Movie	41.61	44.52	**45.81**
	IITP Product	58.32	57.17	**61.57**
bn	Soham Articles	62.79	64.78	**71.82**
gu	iNLTK Headlines	81.94	84.07	**90.74**
ml		86.35	83.65	**95.87**
mr		83.06	81.65	**91.40**
ta		90.88	89.09	**95.37**
te	ACTSA	46.03	42.51	**52.58**
	Average	69.25	68.32	**75.80**

FT-W: pre-trained FastText (Wikipedia). FT-WC: pre-trained FastText (Wikipedia+CommonCrawl)

IndicBERT

https://indicnlp.ai4bharat.org/indic-bert

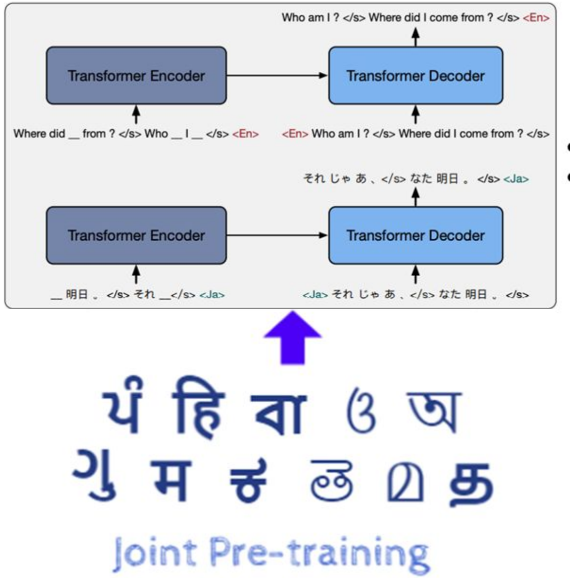
https://huggingface.co/ai4bharat/indic-bert



- 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

IndicBART

https://indicnlp.ai4bharat.org/indic-bart



- Pre-trained Indic S2S for NLG applications
- Large Indian language content (8B tokens)
  - 11 Indian languages
  - + Indian English content
- Multilingual Model
- Compact Model (~224m params)
- Single Script
- Competitive with mBART50 for MT and summarization
- Simply fine-tune for your application

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, Pratyush Kumar. *IndicBART: A Pre-trained Model for Natural Language Generation of Indic Languages.* Arxiv preprint 2109.02903. 2021.

Future Possibilities

Monolingual Data

- Language coverage
- Larger Monolingual Crawls
- Release more metadata
- Offensive Text Filtering

Pre-trained models

- Language coverage
- Train on larger data
- Incorporate parallel data
- Model distillation recipes

Benchmark datasets

- Diverse & challenging tasks
- Language coverage
- Zeroshot evaluation
- NLG datasets

Samanantar

The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages

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, Mitesh Shantadevi Khapra

AI4Bharat, EkStep, IITM, Microsoft, RBCDSAI, Tarento

TACL 2022

11 Languages + English

- Assamese, Bengali, Hindi, Gujarati, Marathi, Odia, Punjabi
- Kannada, Malayalam, Telugu, Hindi

	#lang-pair	#sent-pair (million)
English-Indic languages	11	49.7
Indic-Indic languages	55	83.4

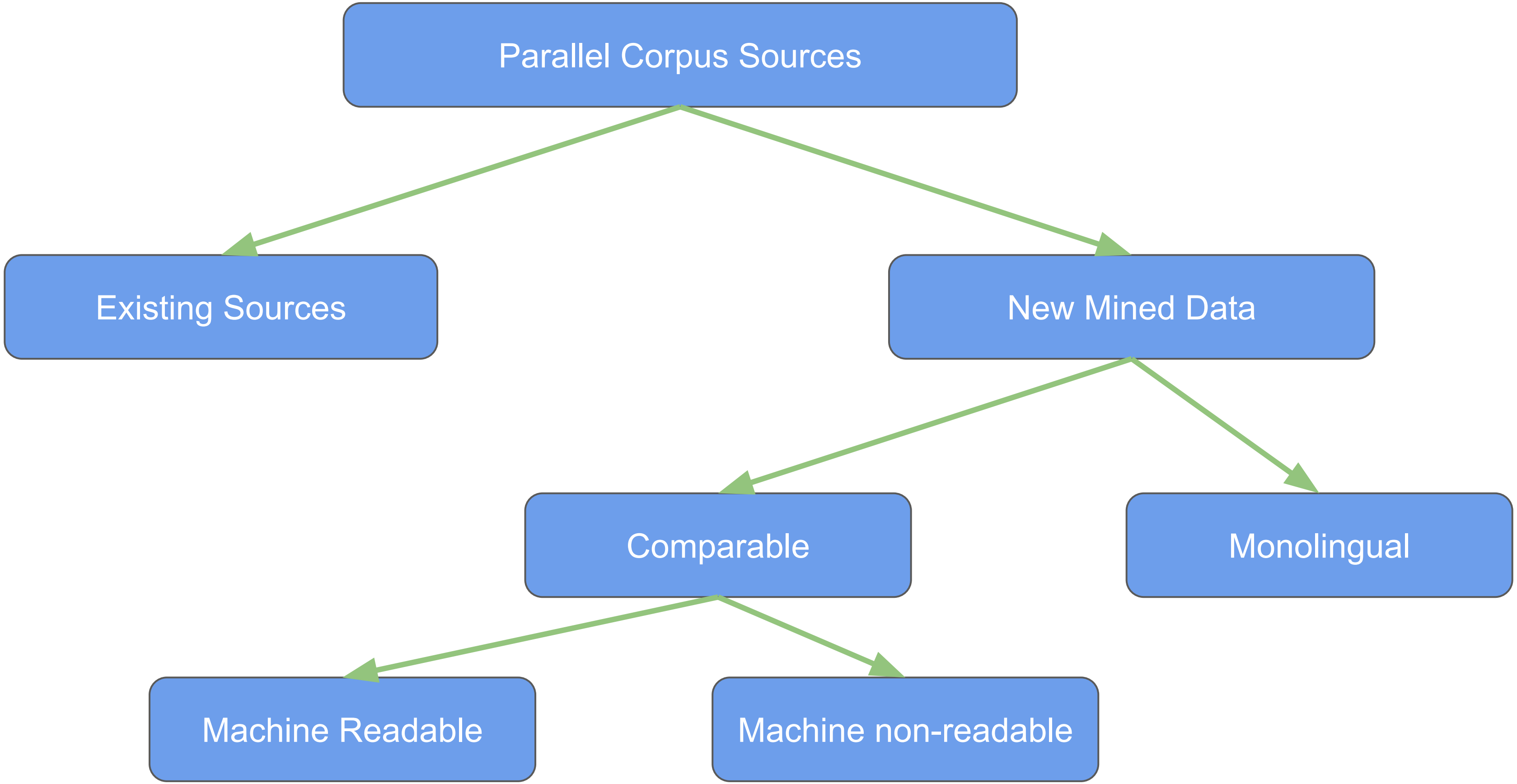
4x increase over existing corpora

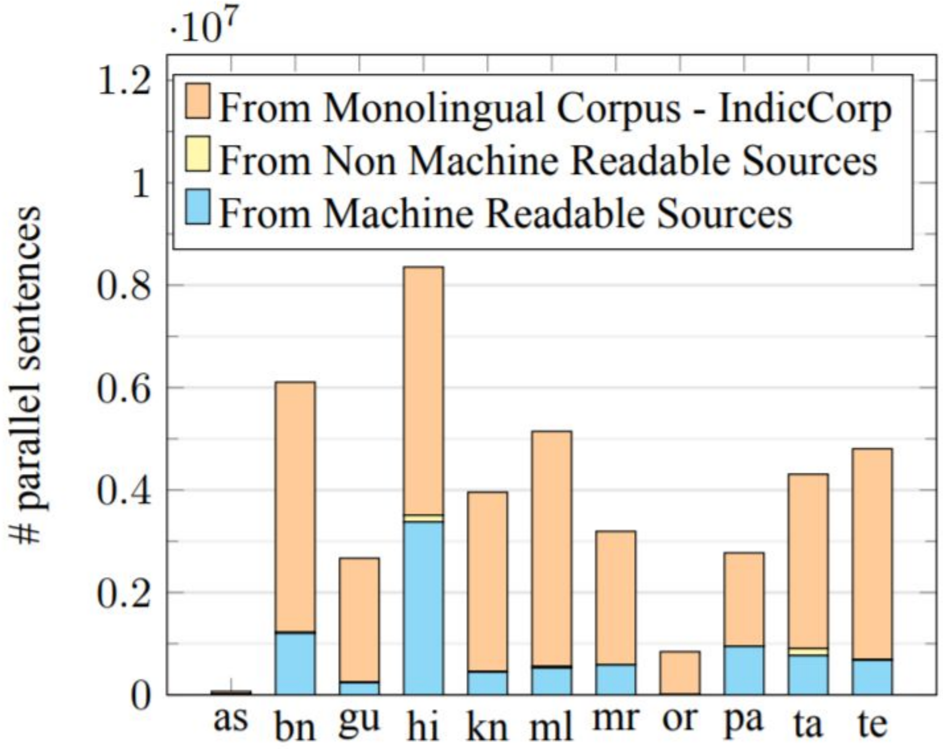
Sentence pair similarity scores available

Source	en-as	en-bn	en-gu	en-hi	en-kn	en-ml	en-mr	en-or	en-pa	en-ta	en-te	Total
Existing Sources	108	3,496	611	2,818	472	1,237	758	229	631	1,456	593	12,408
New Sources	34	5,109	2,457	7,308	3,622	4,687	2,869	769	2,349	3,809	4,353	37,366
Total	141	8,605	3,068	10,126	4,094	5,924	3,627	998	2,980	5,265	4,946	49,774
Increase Factor	1.3	2.5	5	3.6	8.7	4.8	4.8	4.4	4.7	3.6	8.3	4

#sentences (in millions)

https://indicnlp.ai4bharat.org/samanantar


Parallel Corpus Sources
Existing Sources
New Mined Data
Comparable
Monolingual
Machine Readable
Machine non-readable



Mining from monolingual corpora is the largest contributor to Samanantar

Compiling all the existing sources

https://github.com/AI4Bharat/indicnlp_catalog#ParallelTranslationCorpus

1. All the sources from **OPUS** were selected as of 21st March 2021 (13 Sources)
    a. ELRC_2922, GNOME, KDE, Ubuntu, Global Voices, JW300, Mozilla-I10n, Open subtitles, TED 2020, Tanzil,Tatoeba, Bible-eudin, Wiki-Matrix
2. All the recent releases of **shared tasks, papers, and open sources data** (14 Sources)
    a. ALT, BanglaNMT, CVIT-PIB, IITB, MTEnglish2Odia, NLPC, OdiEnCorp 2.0, PMIndia V1, SIPC, TICO19, UFAL, URST, WMT-2019-wiki24 , WMT-2019-govin
3. Total of 12.4M sentences from English to 11 Indian Languages

There's no public model available trained on all the existing parallel data

*cited at the end

Existing sources of parallel data

	en-as	en-bn	en-gu	en-hi	en-kn	en-ml	en-mr	en-or	en-pa	en-ta	en-te	Total
JW300	46	269	305	510	316	371	289	-	374	718	203	3400
banglanmt	-	2380	-	-	-	-	-	-	-	-	-	2380
iitb	-	-	-	1603	-	-	-	-	-	-	-	1603
cvit-pib	-	92	58	267	-	43	114	94	101	116	45	930
wikimatrix[6]	-	281	-	231	-	72	124	-	-	95	92	895
OpenSubtitles	-	372	-	81	-	357	-	-	-	28	23	862
Tanzil	-	185	-	185	-	185	-	-	-	92	-	647
KDE4	6	35	31	85	13	39	12	8	78	79	14	402
PMIndia V1	7	23	42	50	29	27	29	32	28	33	33	333
GNOME	29	40	38	30	24	23	26	21	33	31	37	332
bible-uedin	-	-	16	62	61	61	60	-	-	-	62	321
Ubuntu	21	28	27	25	22	22	26	20	29	25	24	269
ufal	-	-	-	-	-	-	-	-	-	167	-	167
sipc	-	21	-	38	-	30	-	-	-	35	43	166
GlobalVoices	-	138	-	2	-	-	-	326	1	-	-	142
TED2020	< 1	10	16	46	2	6	22	-	752	11	5	120
Mozilla-I10n	7	21	-	< 1	12	13	15	8	-	17	25	119
odiencorp 2.0	-	-	-	-	-	-	-	91	-	-	-	91
Tatoeba	< 1	5	< 1	11	< 1	< 1	53	< 1	< 1	< 1	< 1	71
urst	-	-	65	-	-	-	-	-	-	-	-	65
alt	-	20	-	20	-	-	-	-	-	-	-	40
mtenglish2odia	-	-	-	-	-	-	-	35	-	-	-	35
nlpc	-	-	-	-	-	-	-	-	-	31	-	31
wmt-2019-wiki	-	-	18	-	-	-	-	-	-	-	-	18
wmt2019-govin	-	-	11	-	-	-	-	-	-	-	-	11
tico19	-	< 1	< 1	< 1	< 1	< 1	< 1	-	< 1	< 1	< 1	6
ELRC_2922	-	< 1	-	< 1	-	< 1	-	-	-	< 1	< 1	1
Total	108	3496	611	2818	472	1237	758	229	631	1456	593	**12408**

Mining from Machine Readable Sources

1. Identified 12 websites which publish content in multiple Indian languages
   a. DriveSpark, OneIndia, NativePlanet, MyKhel, Newsonair, DW, TimesofIndia, IndianExpress, GoodReturns, CatchNews, DD National

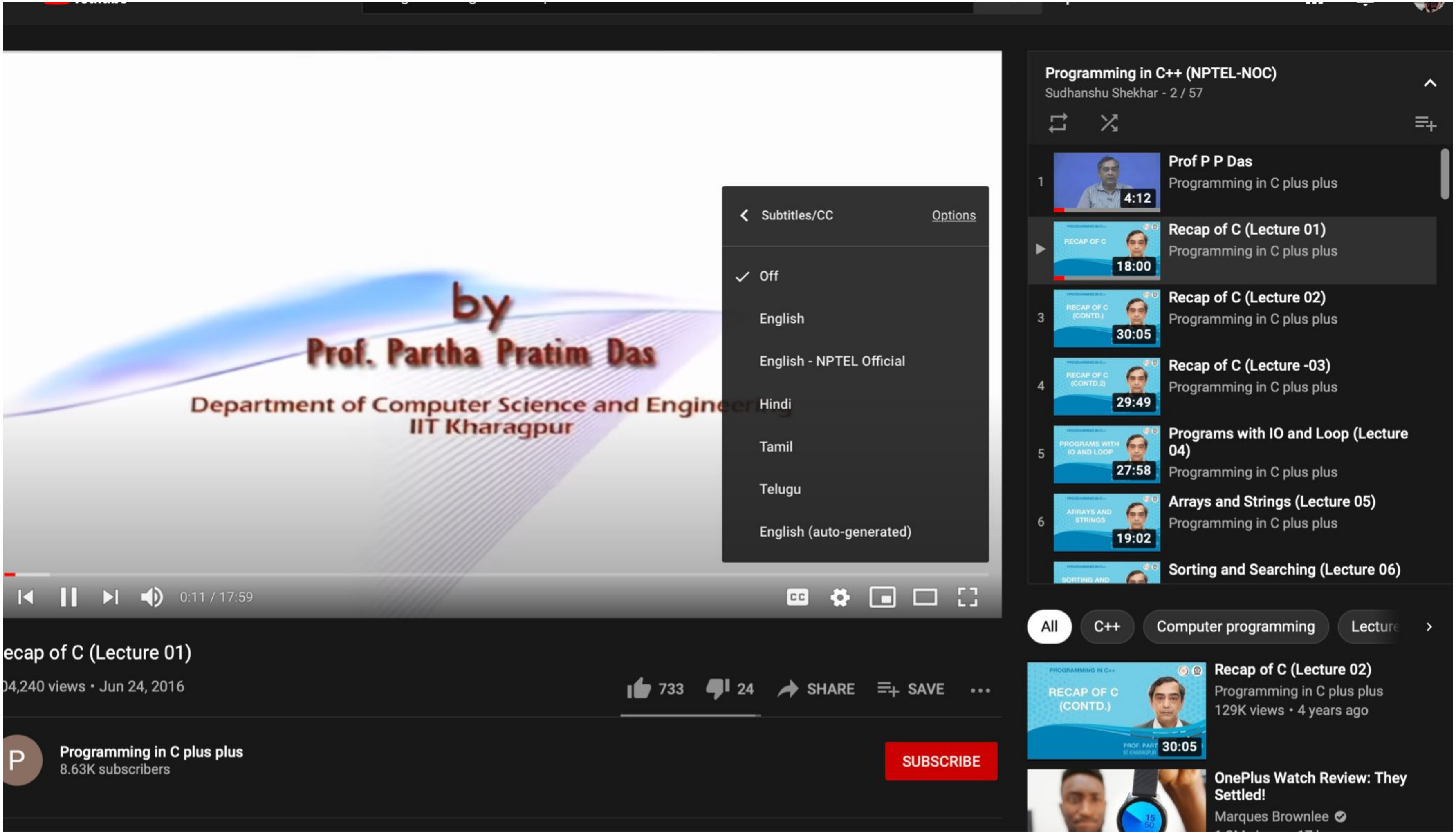
2. Identified 2 Educational sources
   a. NPTEL, Khan Academy

IPL 2021: RCB vs CSK: Highlights: Ravindra Jadeja show helps CSK maul RCB by 69 runs, climb at top

By Avinash Sharma Updated: Sunday, April 25, 2021, 19:44 [IST]

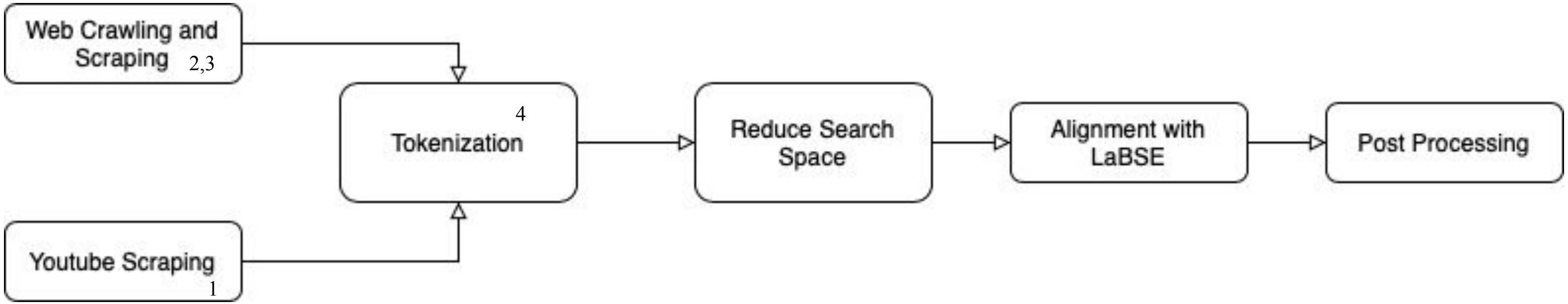
CSK vs RCB: బ్యాట్, బంతితో 'సర్' జడేజా ఆల్‌రౌండ్ షో.. బెంబేలెత్తిన బెంగళూరు! కోహ్లీసేనకు తొలి ఓటమి!

By Sampath Kumar Updated: Sunday, April 25, 2021, 19:53 [IST]



by
Prof. Partha Pratim Das
Department of Computer Science and Engineering
IIT Kharagpur
Subtitles/CC
Options
Off
English
English - NPTEL Official
Hindi
Tamil
Telugu
English (auto-generated)
0:11 / 17:59
ecap of C (Lecture 01)
04,240 views • Jun 24, 2016
733
24
SHARE
SAVE
Programming in C plus plus
8.63K subscribers
SUBSCRIBE
Programming in C++ (NPTEL-NOC)
Sudhanshu Shekhar - 2 / 57
1
4:12
Prof P P Das
Programming in C plus plus
18:00
Recap of C (Lecture 01)
Programming in C plus plus
3
30:05
Recap of C (Lecture 02)
Programming in C plus plus
4
29:49
Recap of C (Lecture -03)
Programming in C plus plus
5
27:58
Programs with IO and Loop (Lecture 04)
Programming in C plus plus
6
19:02
Arrays and Strings (Lecture 05)
Programming in C plus plus
Sorting and Searching (Lecture 06)
All
C++
Computer programming
Lecture
RECAP OF C (CONTD.)
30:05
Recap of C (Lecture 02)
Programming in C plus plus
129K views • 4 years ago
OnePlus Watch Review: They Settled!
Marques Brownlee

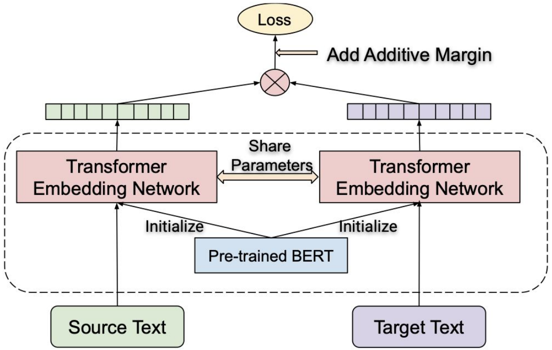
Pipeline from Extraction to Alignment



1. https://youtube-dl.org
2. https://www.crummy.com/software/BeautifulSoup
3. https://pypi.org/project/selenium
4. https://pypi.org/project/indic-nlp-library/

Alignment with LaBSE

1. Language agnostic BERT Sentence Embedding
2. LaBSE is a multilingual model trained on 17B monolingual sentences and 6B parallel sentences using the MLM (Masked Language Modelling), TLM (Translation Language Modelling) and margin-based task
3. Translation Ranking Task
4. LaBSE provides high-dimensional vector(768) for a given input sentence
5. We use cosine similarity as the similarity metric
6. Select the sentence with the maximum similarity



Feng, F., Yang, Y., Cer, D.M., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT Sentence Embedding. *ArXiv, abs/2007.01852*.

https://tfhub.dev/google/LaBSE/2

Post Processing

1. Select only sentences with LaBSE alignment score greater than 0.75
2. The LaBSE Alignment Score (LAS) threshold of 0.75 has been chosen based on manual verification of the sentences
3. More about the threshold in Analysis section
4. Dedup the sentence pairs
5. Drop sentences with less than 4 words
6. Drop sentences if language identified as anything other than intended language [polyglot]

https://github.com/aboSamoor/polyglot

Mining from Non-Machine Readable Sources

1. Documents published from parliament proceedings
2. Speeches from AP and TS Legislative Assemblies
3. Speeches from Bangladesh Parliament

1. https://cloud.google.com/vision/docs/ocr
2. https://pypi.org/project/indic-nlp-library/

- அடுத்த 7 வருடங்களில், உலகளவில் உயர் வருமானத்தைக் கொண்ட நாடுகளுக்கு நிகராக, தனிநபர் வருமானத்தில் 3 மடங்கு வளர்ச்சியினை அடைந்து, 2023 ஆம் ஆண்டிற்குள் இந்தியாவின் பொருளாதாரத்தில் வளமிக்க மாநிலங்களில் ஒன்றாக தமிழ்நாடு இருக்கும்.
- தமிழ்நாடு அனைவரையும் உள்ளடக்கிய வளர்ச்சி முறையை வெளிப்படுத்தும் – இலாபகரமான மற்றும் பயனுள்ள வேலைகளைக் தேடும் அனைவருக்கும், வாய்ப்புகளை வழங்கி, வறுமையில்லா மாநிலமாக தமிழ்நாடு திகழ்ந்து, பாதிக்கப்பட்டவர்கள், நலிவுற்ற பிரிவினர் மற்றும் ஆதரவற்றோர்களுக்கு பராமரிப்பு அளிக்கும்.
- சமுதாய மேம்பாட்டில் தமிழ்நாடு முன்னிலை மாநிலமாக விளங்கி, இந்தியாவில் உள்ள அனைத்து மாநிலங்களிடையே மனித மேம்பாட்டு குறியீட்டில் உயரிய இடத்தைப் பெறும்.
- தமிழ்நாடு, பல்வேறு துறைகளில் உலகத்தரம் வாய்ந்த நிறுவனங்கள் மற்றும் உயர் மனித திறமையின் மூலம் புதுமை மையமாகவும் அறிவாற்றலில் இந்தியாவின் தலைநகரமாகவும் விளங்கும்.
- தமிழ்நாடு, அதனுடைய சூழலியல் மற்றும் அதனுடைய பாரம்பரியத்தை என்றென்றும் பாதுகாக்கும்.

- **Tamil Nadu will be amongst India's most economically prosperous states by 2023, achieving a three-fold growth in per capita income (in real terms) over the next 7 years to be on par with the Upper Middle Income countries globally.**
- **Tamil Nadu will exhibit a highly inclusive growth pattern – it will largely be a poverty free state with opportunities for gainful and productive employment for all those who seek it, and will provide care for the disadvantaged, vulnerable and the destitute in the state.**
- **Tamil Nadu will be India's leading state in social development and will have the highest Human Development Index (HDI) amongst all Indian States.**
- **Tamil Nadu will be known as the innovation hub and knowledge capital of India, on the strength of world class institutions in various fields and the best human talent.**
- **Tamil Nadu will preserve and care for its ecology and heritage.**

- About 65% of the persons targeted for skill development, who would have studied upto secondary school, would be provided with training on basic

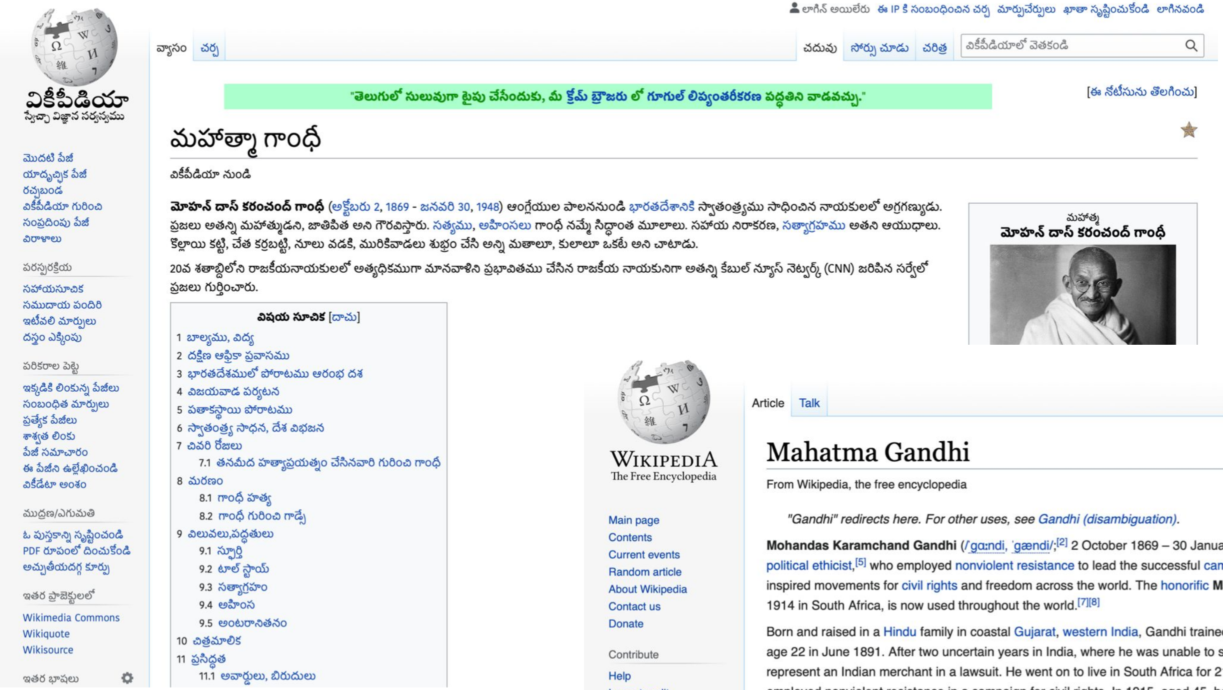
skills for a variety of livelihoods. About 33% would have already undergone formal education as part of vocational training programmes or in colleges, while the remaining top 2% would be top echelon professionals.

Going beyond comparable corpora

- Discovering parallel sources is non-trivial
- Not necessarily Regular URL patterns across websites
- Parallel content can exist across different domains
- Sometimes, it is difficult to say that the websites are parallel

Audacious goal: can we mine parallel data from just large monolingual corpora

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB. 2019. arXiv:1911.04944



మహాత్మా గాంధీ

వికీపీడియా నుండి

మోహన్ దాస్ కరంచంద్ గాంధీ (అక్టోబరు 2, 1869 - జనవరి 30, 1948) ఆంగ్లేయుల పాలననుండి భారతదేశానికి స్వాతంత్ర్యము సాధించిన నాయకులలో అగ్రగణ్యుడు. ప్రజలు అతన్ని మహాత్ముడని, జాతిపిత అని గౌరవిస్తారు. సత్యము, అహింసలు గాంధీ నమ్మే సిద్ధాంత మూలాలు. సహాయ నిరాకరణ, సత్యాగ్రహము అతని ఆయుధాలు. కొల్లాయి కట్టి, చేత కర్రబట్టి, నూలు వడకి, మురికివాడలు శుభ్రం చేసి అన్ని మతాలూ, కులాలూ ఒకటే అని చాటాడు.

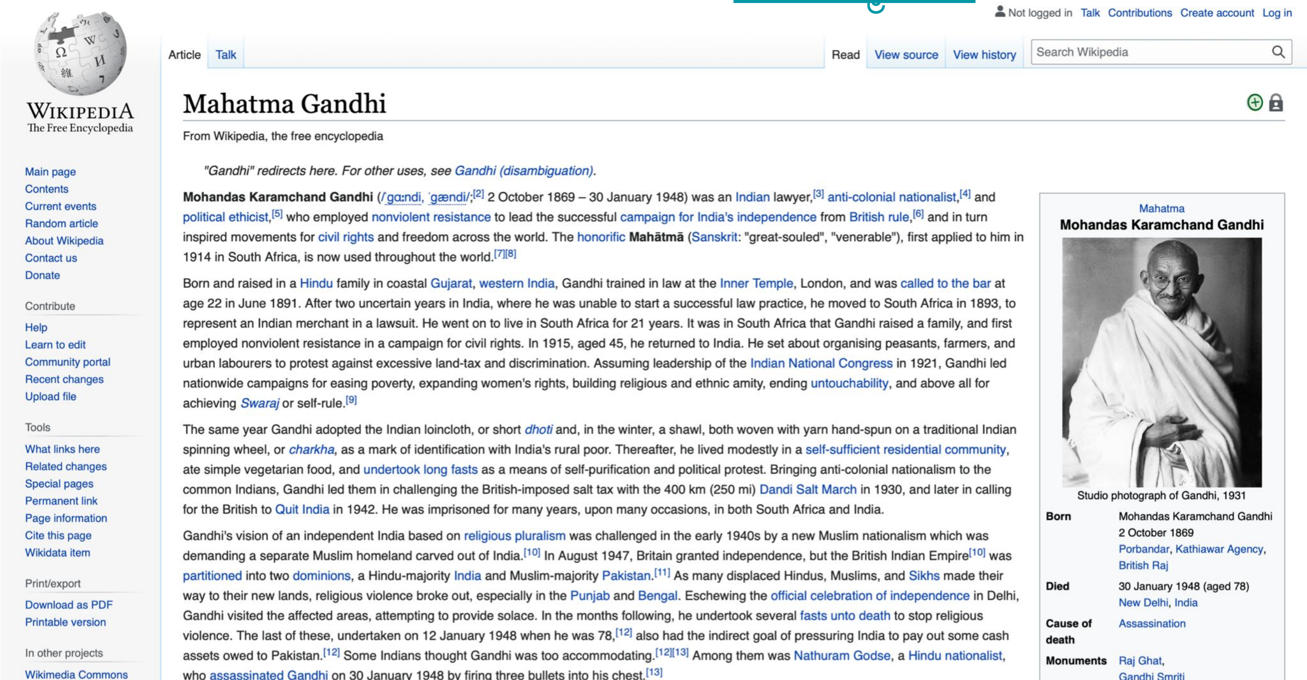
20వ శతాబ్దిలోని రాజకీయనాయకులలో అత్యధికముగా మానవాళిని ప్రభావితము చేసిన రాజకీయ నాయకునిగా అతన్ని కేబుల్ న్యూస్ నెట్వర్క్ (CNN) జరిపిన సర్వేలో ప్రజలు గుర్తించారు.

విషయ సూచిక [దాచు]

1 బాల్యము, విద్య
2 దక్షిణ ఆఫ్రికా ప్రవాసము
3 భారతదేశములో పోరాటము ఆరంభ దశ
4 విజయవాడ పర్యటన
5 పతాకస్థాయి పోరాటము
6 స్వాతంత్ర్య సాధన, దేశ విభజన
7 చివరి రోజులు
7.1 తనమీద హత్యాప్రయత్నం చేసినవారి గురించి గాంధీ
8 మరణం
8.1 గాంధీ హత్య
8.2 గాంధీ గురించి గాడ్సే
9 విలువలు,పద్ధతులు
9.1 స్ఫూర్తి
9.2 టాల్ స్టాయ్
9.3 సత్యాగ్రహం
9.4 అహింస
9.5 అంటరానితనం
10 చిత్రమాలిక
11 ప్రసిద్ధత
11.1 అవార్డులు, బిరుదులు

En -
https://en.wikipedia.org/wiki/Mahatma_Gandhi

Te -
https://te.wikipedia.org/wiki/మహాత్మా_గాంధీ



Mahatma Gandhi

From Wikipedia, the free encyclopedia

"Gandhi" redirects here. For other uses, see Gandhi (disambiguation).

Mohandas Karamchand Gandhi (/ˈɡɑːndi, ˈɡændi/;[2] 2 October 1869 – 30 January 1948) was an Indian lawyer,[3] anti-colonial nationalist,[4] and political ethicist,[5] who employed nonviolent resistance to lead the successful campaign for India's independence from British rule,[6] and in turn inspired movements for civil rights and freedom across the world. The honorific Mahātmā (Sanskrit: "great-souled", "venerable"), first applied to him in 1914 in South Africa, is now used throughout the world.[7][8]

Born and raised in a Hindu family in coastal Gujarat, western India, Gandhi trained in law at the Inner Temple, London, and was called to the bar at age 22 in June 1891. After two uncertain years in India, where he was unable to start a successful law practice, he moved to South Africa in 1893, to represent an Indian merchant in a lawsuit. He went on to live in South Africa for 21 years. It was in South Africa that Gandhi raised a family, and first employed nonviolent resistance in a campaign for civil rights. In 1915, aged 45, he returned to India. He set about organising peasants, farmers, and urban labourers to protest against excessive land-tax and discrimination. Assuming leadership of the Indian National Congress in 1921, Gandhi led nationwide campaigns for easing poverty, expanding women's rights, building religious and ethnic amity, ending untouchability, and above all for achieving *Swaraj* or self-rule.[9]

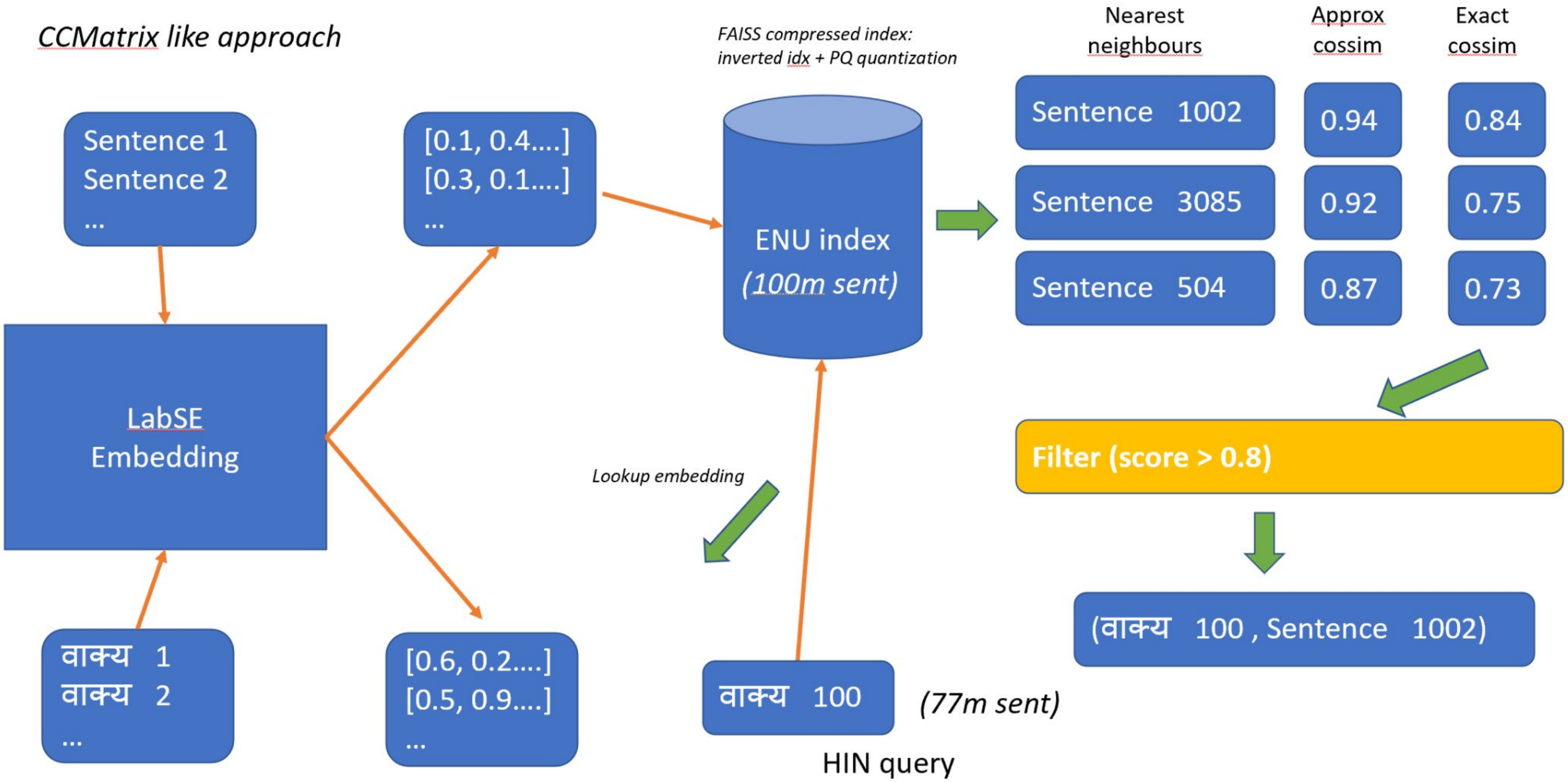
The same year Gandhi adopted the Indian loincloth, or short *dhoti* and, in the winter, a shawl, both woven with yarn hand-spun on a traditional Indian spinning wheel, or *charkha*, as a mark of identification with India's rural poor. Thereafter, he lived modestly in a self-sufficient residential community, ate simple vegetarian food, and undertook long fasts as a means of self-purification and political protest. Bringing anti-colonial nationalism to the common Indians, Gandhi led them in challenging the British-imposed salt tax with the 400 km (250 mi) Dandi Salt March in 1930, and later in calling for the British to Quit India in 1942. He was imprisoned for many years, upon many occasions, in both South Africa and India.

Gandhi's vision of an independent India based on religious pluralism was challenged in the early 1940s by a new Muslim nationalism which was demanding a separate Muslim homeland carved out of India.[10] In August 1947, Britain granted independence, but the British Indian Empire[10] was partitioned into two dominions, a Hindu-majority India and Muslim-majority Pakistan.[11] As many displaced Hindus, Muslims, and Sikhs made their way to their new lands, religious violence broke out, especially in the Punjab and Bengal. Eschewing the official celebration of independence in Delhi, Gandhi visited the affected areas, attempting to provide solace. In the months following, he undertook several fasts unto death to stop religious violence. The last of these, undertaken on 12 January 1948 when he was 78,[12] also had the indirect goal of pressuring India to pay out some cash assets owed to Pakistan.[12] Some Indians thought Gandhi was too accommodating.[12][13] Among them was Nathuram Godse, a Hindu nationalist, who assassinated Gandhi on 30 January 1948 by firing three bullets into his chest.[13]

Mahatma Mohandas Karamchand Gandhi	
Studio photograph of Gandhi, 1931	
Born	Mohandas Karamchand Gandhi 2 October 1869 Porbandar, Kathiawar Agency, British Raj
Died	30 January 1948 (aged 78) New Delhi, India
Cause of death	Assassination
Monuments	Raj Ghat, Gandhi Smriti

Mining from Monolingual Corpora

Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with GPUs. *ArXiv, abs/1702.08734*.

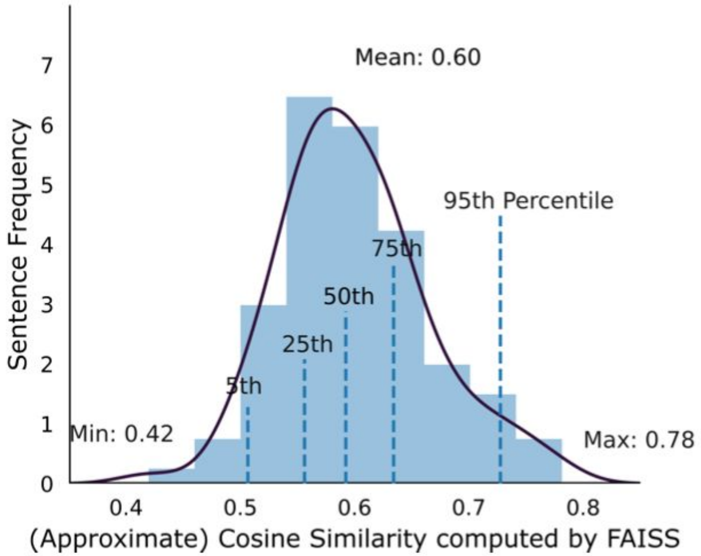


What helps scaling to large datasets

- Simple similarity metric (cosine similarity)
    - Distance from binary argument functions can’t scale (e.g. COMET score)
- Approximate nearest-neighbourhood search
- Compressed indexes to fit indices in GPU memory
- Distributing indices over multiple GPUs
- Searching over multiple indices *(to speed up searches)*

Recomputing the Cosine Similarity

1. Variance on cosine similarity computed on the low-dimension vectors
2. Recompute the cosine similarity on the high-dimensional vector for the top-1 FAISS match
3. Here we use a higher LAS of 0.8



Qualitative Analysis of the parallel corpus

10000 samples manually evaluated using 30+ annotators across 11 languages
Using SemEval-1 guidelines for cross-lingual semantic textual similarity
Available for **cross-lingual STS** studies (https://storage.googleapis.com/samanantar-public/human_annotations.tsv)

1. Sentence pairs included in *Samanantar* have high semantic textual similarity (STS)
   a. avg: 4.17, min: 3.83, max: 4.82 (out of 5)
2. Quality depends on resource size
   a. Highest: hi, bn
   b. Lowest : as, or

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.* SemEva.

Task Instructions

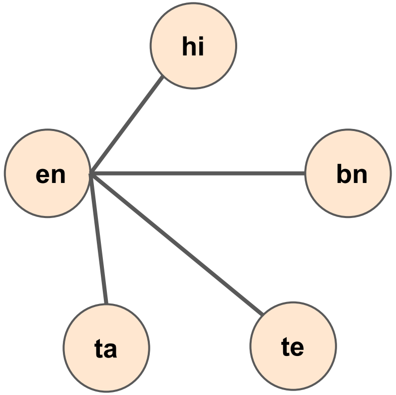
Instruction	Score	Sample sentence pairs
Sentences are completely dissimilar	0	He is a strokemaker. இவர் ஒரு செயின் ஸ்மோக்கர் (he is a chain smoker)
Sentences are dissimilar but topically related	1	Can we save our lakes from global warming? ठंडे पानी के कोरल जलवायु परिवर्तन से बच पायेंगे? (Will cold water corals survive climate change?)
Sentences are dissimilar but agree on some details	2	Going smoke-free புகையில்லா போகி (smoke free Boghi festival)
Sentences have differences in important details	3	The province is divided into ten districts. இந்த மாவட்டத்தை ஆறு மண்டலங்களாகப் பிரித்துள்ளனர். (The province is divided into 6 districts.)
Differences in details are not important	4	Maruti Suzuki To Add More CNG Models, Hybrids मारूति सुजुकी सीएनजी मॉडलों में करेगी इजाफा (Maruti Suzuki to increase CNG models)
Complete semantic similarity	5	They can’t come out from their houses. वे घर से निकल नहीं पाते. (They can’t get out of their homes)

SemEval-2016 Task 1 Cross-lingual STS annotation guidelines

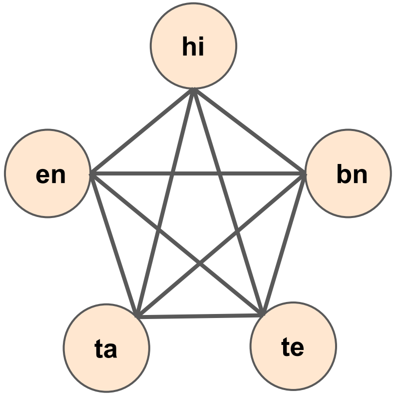
Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. *SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation.* SemEval.

Mining between Indic Languages

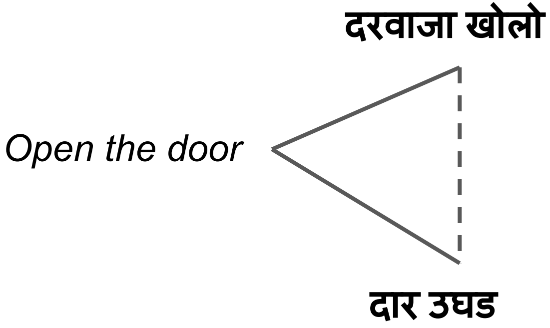
Mine Indic-Indic parallel corpora from English to Indic corpora



English-centric



Complete

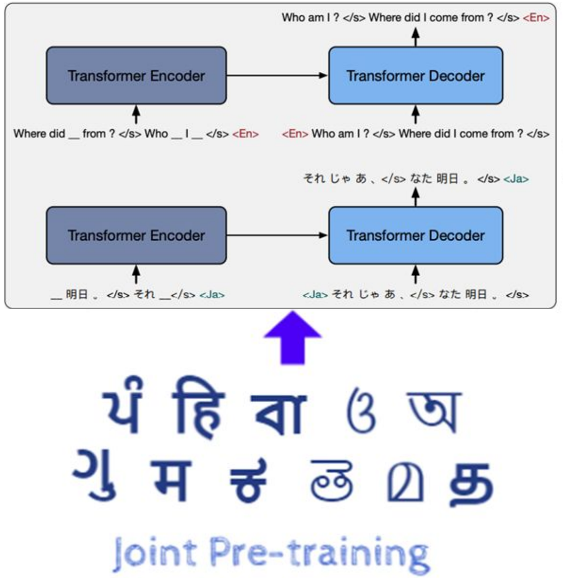


83.7 million sentence pairs for 55 language pairs

Markus Freitag and Orhan Firat. 2020. *Complete multilingual neural machine translation*.WMT.
Annette Rios, Mathias Müller, and Rico Sennrich. 2020. *Subword segmentation and a single bridge language affect zero-shot neural machine translation*. WMT

IndicTrans

https://indicnlp.ai4bharat.org/indic-trans



- Trained on Samanantar parallel corpus
- Multilingual Model (en→IL, IL→en, IL→IL)
- Single Script
- Model size: (~430m params)
- Best open-source model
- Deployed in the Supreme Court of India & Bangladesh
- Training/fine-tuning/inference scripts available

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh Khapra & others. *Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages*. TACL. 2022.

Key Results

- Compilation of existing resources was a fruitful exercise.
- IndicTrans trained on Samanantar outperforms all publicly available open source models.
- IndicTrans trained on Samanantar is competitive with commercial systems.
- Pre-training needs further investigation. formance gains are higher for low resource languages.

	x-en									en-x								
Model	GOOG	MSFT	CVIT	OPUS	mBART	TF	mT5	IT	Δ	GOOG	MSFT	CVIT	OPUS	mBART	TF	mT5	IT	Δ
WAT2021																		
bn	20.6	21.8	-	11.4	4.7	24.2	24.8	**29.6**	4.8	7.3	11.4	12.2	-	0.5	13.3	13.6	**15.3**	1.7
gu	32.9	34.5	-	-	6.0	33.1	34.6	**40.3**	5.7	16.1	22.4	22.4	-	0.7	21.9	24.8	**25.6**	0.8
hi	36.7	38.0	-	13.3	33.1	38.8	39.2	**43.9**	4.7	32.8	34.3	34.3	11.4	27.7	35.9	36.0	**38.6**	2.6
kn	24.6	23.4	-	-	-	23.5	27.8	**36.4**	8.6	12.9	16.1	-	-	-	12.1	17.3	**19.1**	1.8
ml	27.2	27.4	-	5.7	19.1	26.3	26.8	**34.6**	7.3	10.6	7.6	11.4	1.5	1.6	11.2	7.2	**14.7**	3.3
mr	26.1	27.7	-	0.4	11.7	26.7	27.6	**33.5**	5.9	12.6	15.7	16.5	0.1	1.1	16.3	17.7	**20.1**	2.4
or	23.7	27.4	-	-	-	23.7	-	**34.4**	7.0	10.4	14.6	16.3	-	-	14.8	-	**18.9**	2.6
pa	35.9	35.9	-	8.6	-	36.0	37.1	**43.2**	6.1	22	28.1	-	-	-	29.8	31.	**33.1**	2.1
ta	23.5	24.8	-	-	26.8	28.4	27.8	**33.2**	4.8	9.0	11.8	11.6	-	11.1	12.5	13.2	**13.5**	0.3
te	25.9	25.4	-	-	4.3	26.8	28.5	**36.2**	7.7	7.6	8.5	8.0	-	0.6	12.4	7.5	**14.1**	1.7
WAT2020																		
bn	17.0	17.2	18.1	9.0	6.2	16.3	16.4	**20.0**	1.9	6.6	8.3	8.5	-	0.9	8.7	9.3	**11.4**	2.1
gu	21.0	22.0	23.4	-	3.0	16.6	18.9	**24.1**	0.7	10.8	12.8	12.4	-	0.5	9.7	11.8	**15.3**	2.5
hi	22.6	21.3	23.0	8.6	19.0	21.7	21.5	**23.6**	0.6	16.1	15.6	16.0	6.7	13.4	17.4	17.3	**20.0**	2.6
ml	17.3	16.5	18.9	5.8	13.5	14.4	15.4	**20.4**	1.5	5.6	5.5	5.3	1.1	1.5	5.2	3.6	**7.2**	1.6
mr	18.1	18.6	19.5	0.5	9.2	15.3	16.8	**20.4**	0.9	8.7	10.1	9.6	0.2	1.0	9.8	10.9	**12.7**	1.8
ta	14.6	15.4	17.1	-	16.1	15.3	14.9	**18.3**	1.3	4.5	5.4	4.6	-	5.5	5.0	5.2	**6.2**	0.7
te	15.6	15.1	13.7	-	5.1	12.1	14.2	**18.5**	2.9	5.5	7.0	5.6	-	1.1	5.0	5.4	**7.6**	0.7
WMT																		
hi	31.3	30.1	24.6	13.1	25.7	25.3	26.0	**29.7**	-1.6	24.6	24.2	20.2	7.9	18.3	23.	23.8	**25.5**	0.9
gu	30.4	29.9	24.2	-	5.6	16.8	21.9	**25.1**	-5.4	15.2	17.5	12.6	-	0.5	9.0	12.3	**17.2**	-0.3
ta	27.5	27.4	17.1	-	20.7	16.6	17.5	**24.1**	-3.4	9.6	10.0	4.8	-	6.3	5.8	7.1	**9.9**	-0.1
UFAL																		
ta	25.1	25.5	19.9	-	24.7	26.3	25.6	**30.2**	3.9	7.7	10.1	7.2	-	9.2	11.3	**11.9**	10.9	-1.0
PMI																		
as	-	16.7	-	-	-	7.4	-	**29.9**	13.2	-	10.8	-	-	-	3.5	-	**11.6**	0.8

	x-en							en-x						
Model	GOOG	MSFT	CVIT	OPUS	mBART	IT†	IT	GOOG	MSFT	CVIT	OPUS	mBART	IT†	IT
as	-	<u>24.9</u>	-	-	-	17.1	**23.3**	-	<u>13.6</u>	-	-	-	**7.0**	6.9
bn	<u>34.6</u>	31.2	-	17.9	9.4	30.1	**32.2**	<u>28.1</u>	22.9	7.9	-	1.4	18.2	**20.3**
gu	<u>40.2</u>	35.4	-	-	4.8	30.6	**34.3**	25.6	<u>27.7</u>	14.1	-	0.7	19.4	**22.6**
hi	<u>44.2</u>	36.9	-	18.6	32.6	34.3	**37.9**	<u>38.7</u>	31.8	25.7	13.7	22.2	32.2	**34.5**
kn	<u>32.2</u>	30.5	-	-	-	19.5	**28.8**	<u>32.6</u>	22.0	-	-	-	9.9	**18.9**
ml	<u>34.6</u>	34.1	-	9.5	24.0	26.5	**31.7**	<u>27.4</u>	21.1	6.6	4.4	3.0	10.9	**16.3**
mr	<u>36.1</u>	32.7	-	0.6	14.8	27.1	**30.8**	<u>19.8</u>	18.3	8.5	0.1	1.2	12.7	**16.1**
or	<u>31.7</u>	31.0	-	-	-	26.1	**30.1**	<u>24.4</u>	20.9	7.9	-	-	11.0	**13.9**
pa	<u>39.0</u>	35.1	-	9.9	-	30.3	**35.8**	27.0	<u>28.5</u>	-	-	-	21.3	**26.9**
ta	<u>31.9</u>	29.8	-	-	22.3	24.2	**28.6**	<u>28.0</u>	20.0	7.9	-	8.7	10.2	**16.3**
te	<u>38.8</u>	37.3	-	-	15.5	29.0	**33.5**	<u>30.6</u>	30.5	8.2	-	4.5	17.7	**22.0**

Table 7: BLEU scores for En-X and X-En translation for FLORES devtest Benchmark. IT† is IndicTrans trained only on existing data. We bold the best public model and underline the overall best model.

Future Possibilities

Training Data

- Language Coverage
- Use larger monolingual corpora
- Mine longer sentences
- Filtering strategies
  - COMET, PRISM, etc.

Benchmark data

- Create benchmark testsets
  - Source-original
  - Multi-domain
- Create human judgment pool for studying evaluation metrics

Model

- Language Coverage
- Romanized/code-mixed input
- Compact/distilled models
- Better multilingual transfer

IndicWav2Vec

Towards Building ASR Systems for the Next Billion Users

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra

AI4Bharat, IITM, Microsoft, RBCDSAI,

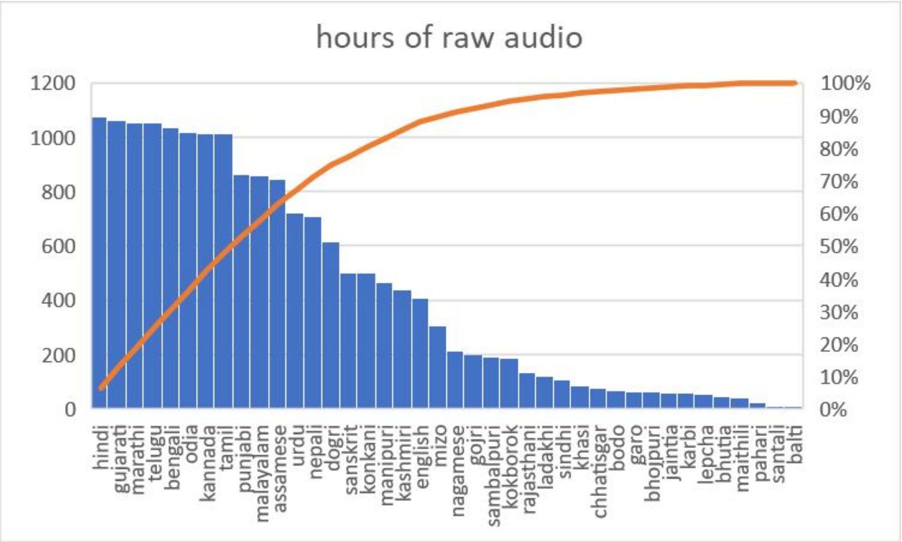
AAAI 2022

Raw Speech Corpora

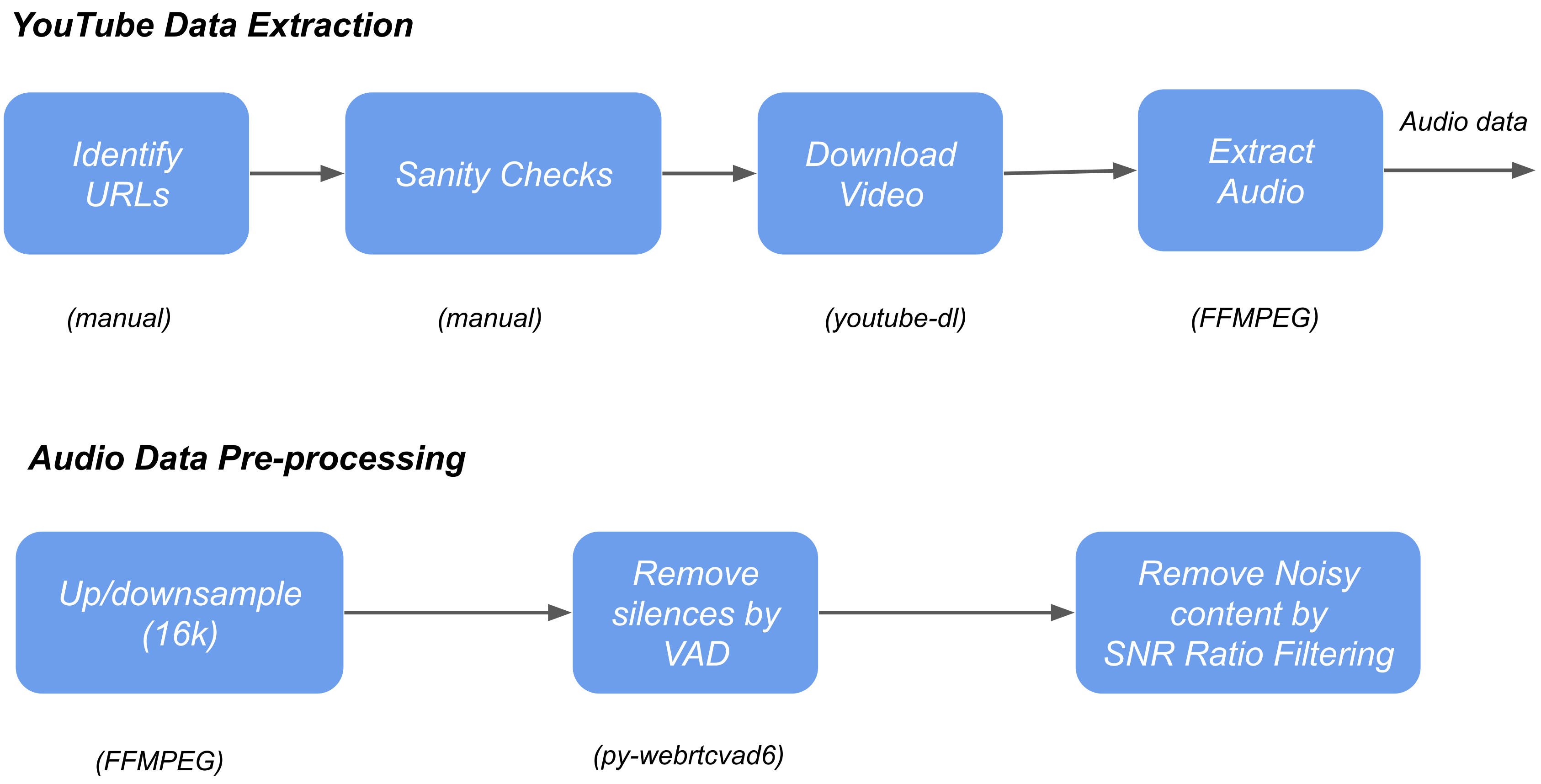
- ~17,000 hrs
- 40 languages
  - All 22 languages in the 8th Schedule
  - Balanced across languages
- 4 language families
- Speaker/channel diversity
- No background noise
- Predominantly target language

https://indicnlp.ai4bharat.org/indicwav2vec/

Sources: Youtube, NewsOnAir

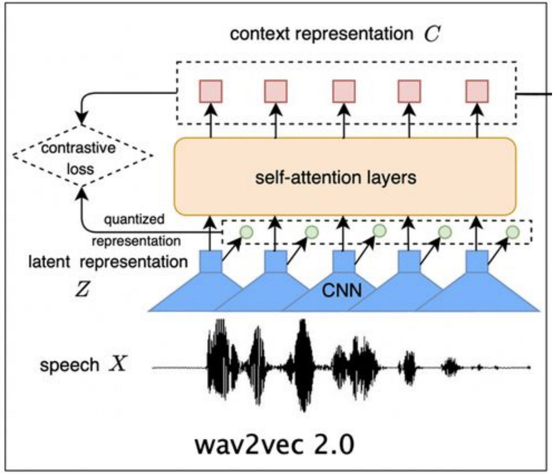


***youtube**: Content licensed under CC-BY*


YouTube Data Extraction
Identify URLs
(manual)
Sanity Checks
(manual)
Download Video
(youtube-dl)
Extract Audio
(FFMPEG)
Audio data
Audio Data Pre-processing
Up/downsample (16k)
(FFMPEG)
Remove silences by VAD
(py-webrtcvad6)
Remove Noisy content by SNR Ratio Filtering

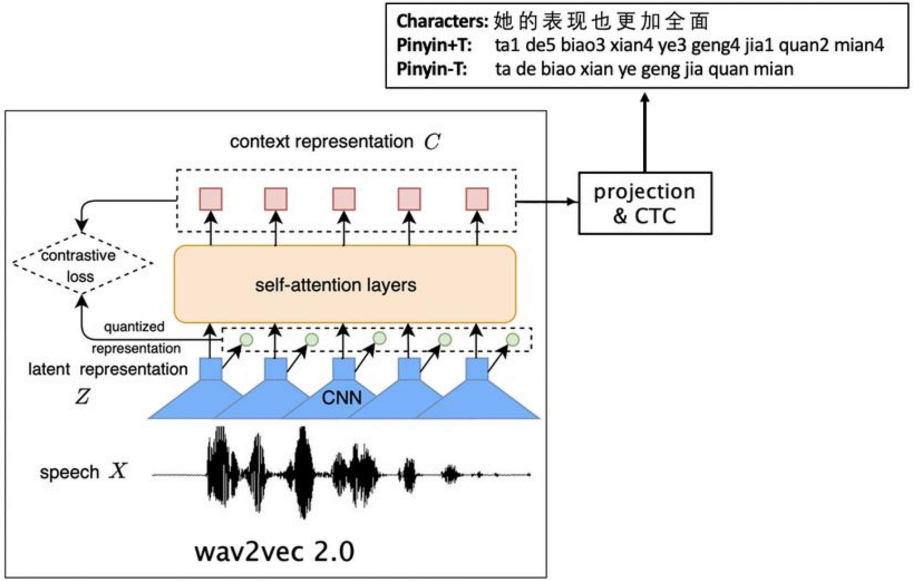
Unsupervised Pre-training

- Follows Wav2Vec 2.0 architecture
- Inspired by BERT pre-training in NLP
- Quantization to learn discrete targets for semi-supervised learning
- Masking + contrastive loss
- Temperature sampling to address data imbalance
- **Initialize** with English wav2vec 2.0
- Model variants:
    - BASE (95m)
    - LARGE (317m)



Finetuning

- Add Linear Projection head
- CTC Loss
- SpecAugment for data augmentation
- Finetune all params except feature encoder



<u>Decoding</u>

LM: 6-gram trained on IndicCorp
Lexicon-based beam search decoder (Flashlight)

$$\mathbf{y}^* = \mathrm{argmax}_{\mathbf{y}} \log p_{AM}(\mathbf{y}) + \alpha \log p_{LM}(\mathbf{y}) + \beta|\mathbf{y}|$$

Key Results and Observations - I

- Pretraining significantly improves the performance on benchmark datasets.
- Our pretraining data has **more diversity, better distribution** of data across languages
    - Result - It **generalises better** for languages not seen during pretraining.
- The LARGE model consistently outperforms the BASE model.
- Starting with **English wav2vec checkpoint** saves compute resources
- The **Language Model** plays an important role.
    - Especially when limited training data is available
- **Finetuning data size**: very small data size (~1hr) not sufficient
    - unlike results on English Wav2Vec: Pre-training size? Language characteristics?

Key Results and Observations - II

	MSR			MUCS						OpenSLR		
	gu	ta	te	gu	hi	mr	or	ta	te	bn	ne	si
M0: No pretraining	46.0	37.5	35.5	53.2	48.1	87.1	73.4	44.8	44.7	36.0	78.8	37.0
M1: IndicW2V$_b$ (EkStep data)	23.4	24.0	25.8	30.3	18.0	26.5	28.7	29.9	33.2	19.7	14.4	31.0
M2: IndicW2V$_b$ (our data)	22.8	23.7	24.9	29.4	17.8	24.3	27.2	29.3	31.9	18.1	13.8	24.3
M3: IndicW2V$_l$ (our data)	20.5	22.1	22.9	26.2	16.0	19.3	25.6	27.3	29.3	16.6	11.9	24.8
M4: + LM$_{small}$	16.6	14.9	14.4	18.0	16.3	14.8	19.0	25.4	22.4	14.3	13.0	**18.6**
M5: + LM$_{large}$	**11.7**	**13.6**	**11.0**	17.2	14.7	13.8	**17.2**	25.0	20.5	13.6	13.6	-
M6: + augmented lexicon	12.3	15.1	12.4	14.8	10.5	12.2	21.9	20.0	15.2	10.6	9.7	-
M7: + Rescoring	11.9	14.8	12.0	**14.3**	**9.5**	**11.7**	20.6	**19.5**	**15.1**	**10.5**	**9.4**	-

Table 3: Comparison of different choices for pretraining, fine-tuning, and decoding. IndicW2V$_b$ and IndicW2V$_l$ refer to our base and LARGE models respectively. LM$_{small}$ refers to the language model trained using transcripts from the training and validation data and LM$_{large}$ refers to the one trained using IndicCorp in addition to the transcripts.

SOTA results on ASR Benchmarks

	hi	gu	mr	or	ta	te
Baseline	27.45	25.98	20.41	31.28	35.82	29.35
CSTR	14.33	20.59	15.79	25.34	23.16	21.88
BSA	16.59	21.30	15.65	17.81	28.59	25.37
EM	17.54	20.11	20.15	19.99	28.52	26.08
EkStep	12.24	30.65	39.74	27.10	27.20	22.43
Uniphore	22.79	22.79	14.9	29.55	**18.8**	28.69
Lottery	17.81	23.62	58.78	17.74	30.69	27.67
IIITH	31.11	26.94	33.8	37.19	35.03	17.00
M5:	14.7	17.2	13.8	**17.2**	25.0	20.5
M6:	**10.5**	**14.8**	**12.2**	21.9	20.0	**15.2**

Table 8: Comparison of our best models (M5, M6) with the the top performers from the MUCS 2021 leaderboard. Individual best models from the leaderboard are underlined.

	gu	ta	te
Baseline (Srivastava et al., 2018)	19.8	19.4	22.6
Jilebi (Pulugundla et al., 2018)	14.0	13.9	14.7
Cogknit (Fathima et al., 2018)	17.7	16.0	17.1
CSALT-LEAP (Srivastava et al., 2018)	-	16.3	17.6
ISI-Billa (Billa, 2018)	19.3	19.6	20.9
MTL-SOL (Sailor and Hain, 2020)	18.4	16.3	18.6
Reed (Sen et al., 2021)	16.1	19.9	20.2
CNN + Context temporal features (Sen et al., 2020)	18.4	24.3	25.2
EkStep model*	19.5	22.1	21.9
M5:	**11.7**	**13.6**	**11.0**
M6:	12.3	15.1	12.4

Table 9: Comparison of our best models (M5, M6) with the the top performers from the MSR 2018 leaderboard as well as other recent state of the art methods.

	bn	ne	si
Baseline (Shetty and Umesh, 2021)	17.9	12.9	21.8
Ekstep model*	15.2	13.8	20.0
M4:	14.3	13.0	**18.6**
M6:	**10.6**	**9.7**	-

Table 10: Comparison of our best models (M4, M6) with state-of-the-art results reported in the literature. * The Ekstep model was fine-tuned by us.

Future Possibilities

- Increase pre-training corpus size
- Collect supervised training data
- Combining supervised and unsupervised objectives
- Create benchmark testsets

INCLUDE

Indian Sign Language (ISL)

People – Data – Models

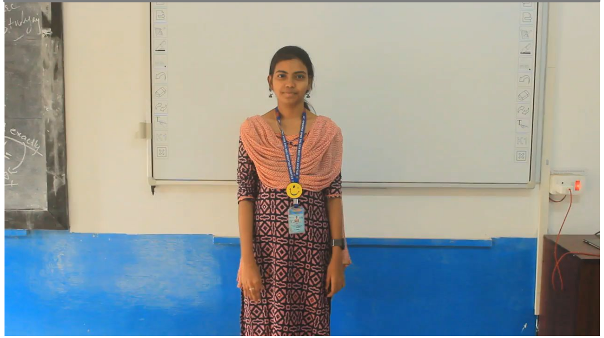
with Advaith, Gokul, Prem, Mohit, Vivek, Manohar, Mitesh, Roshni

NISH

We are here because, we care

Microsoft Philanthropies

Hello

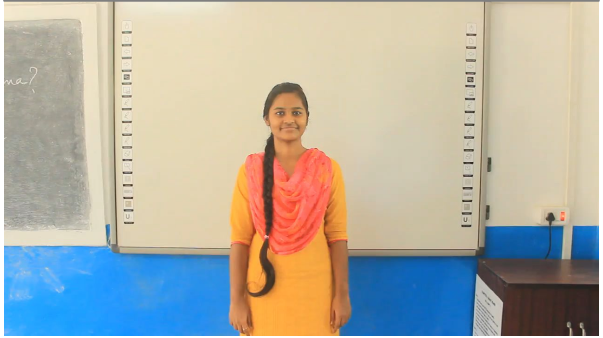


Cat

Dad

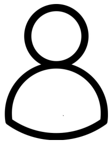
Summer

Long



Thank you

Context



> 5 million

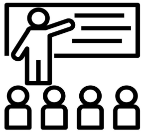
Size of the Deaf community in India

Indian Sign Language (ISL)

Primary means of communication for the community

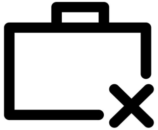
Data Poverty

No large publicly available dataset on ISL



Education

<1% has formal education in ISL



Unemployment

76% within the Deaf community

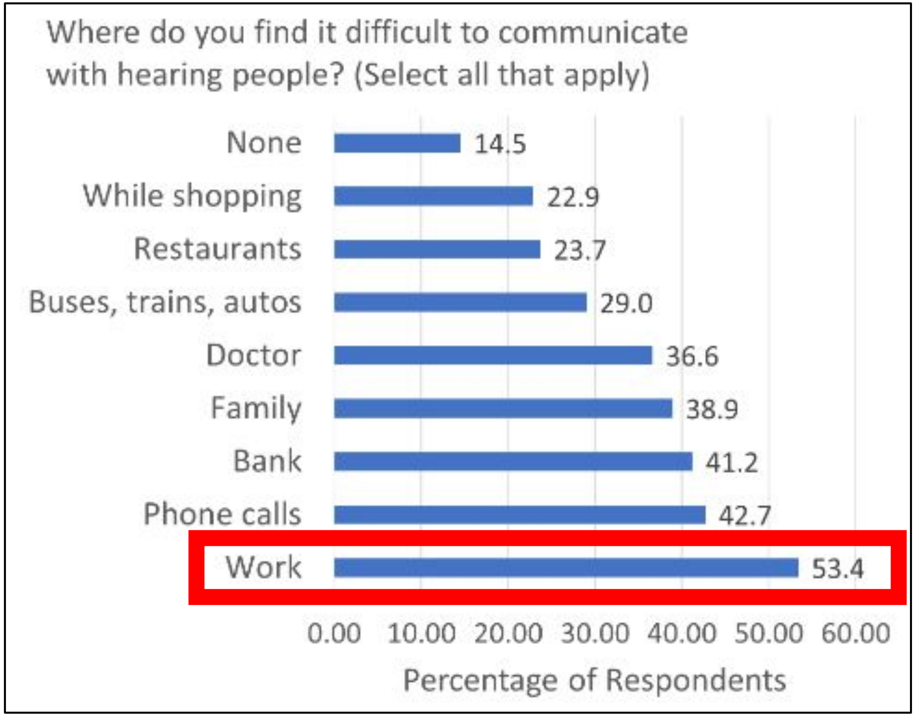
People

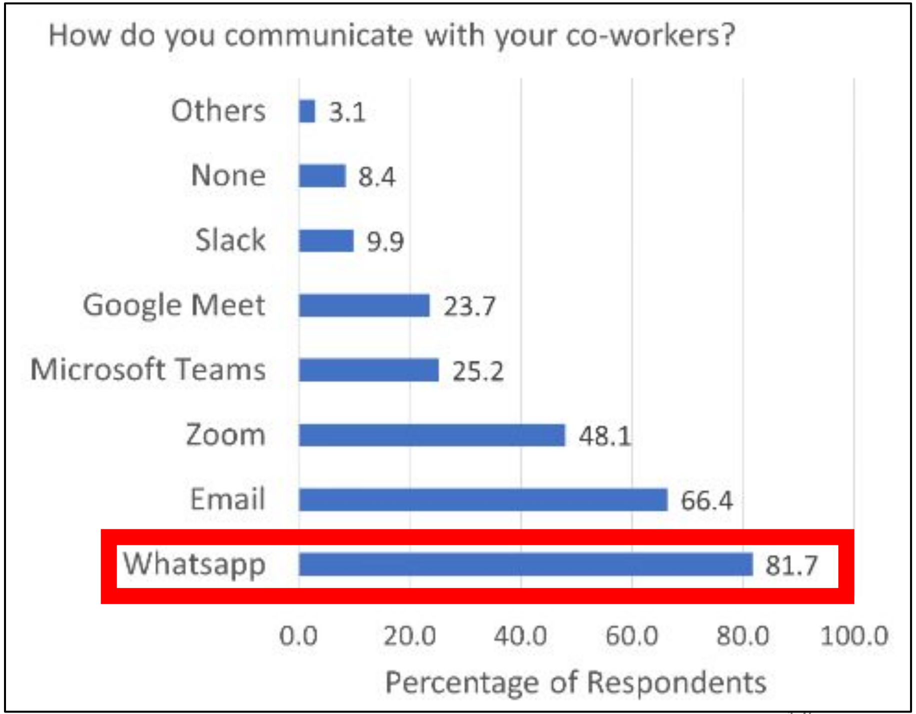
First large-scale survey of challenges of Indian Deaf

Method

Online forms
131 participants

1:1 interviews with interpreter
15 participants





Survey findings

ISL Diversity

Stigma

Grammatical issues

Challenges at workplace

Issues with tools

Survey findings

ISL Diversity

Stigma

Grammatical issues

Challenges at workplace

Issues with tools

“The sign for ‘marriage’ is shown by a mangalsutra necklace in Chennai, while in Hyderabad it is shown by the holding of hands.”

Survey findings

ISL Diversity

Stigma

Grammatical issues

Challenges at workplace

Issues with tools

"*I go start my exercise walking now it*"

instead of

"*I shall now start my walking exercise*".

Survey findings

ISL Diversity

Stigma

Grammatical issues

Challenges at workplace

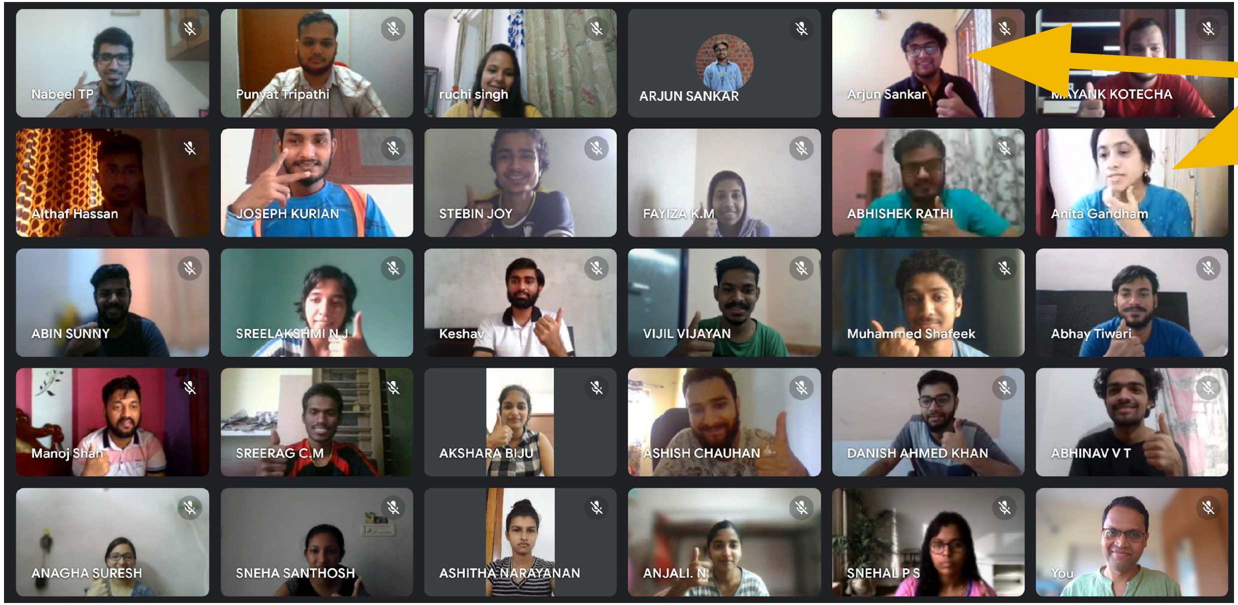
Issues with tools

"Now I want to go to the next level... But my manager is not able to understand what I'm trying to say. If I have to write and ask about the promotion, the management team will be asking questions... I am afraid that they will give me lecture on it, so its better to not talk about it"

Data

Collaborative data collection

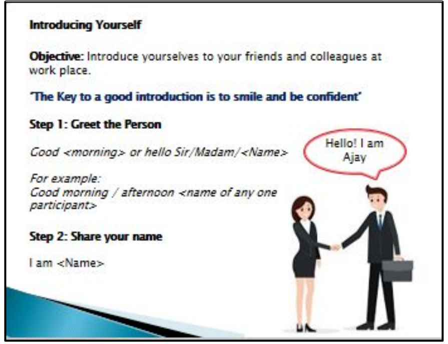
Collect data while teaching a course @ National Institute of Speech and Hearing



Our instructors

English for workplace course (8 weeks)

1. Introducing yourself

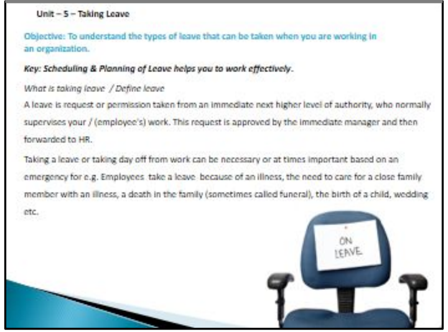


2. Knowing about job

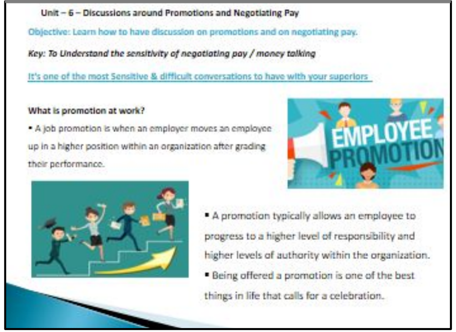
3. Talking to HR

4. Meetings

5. Taking leave



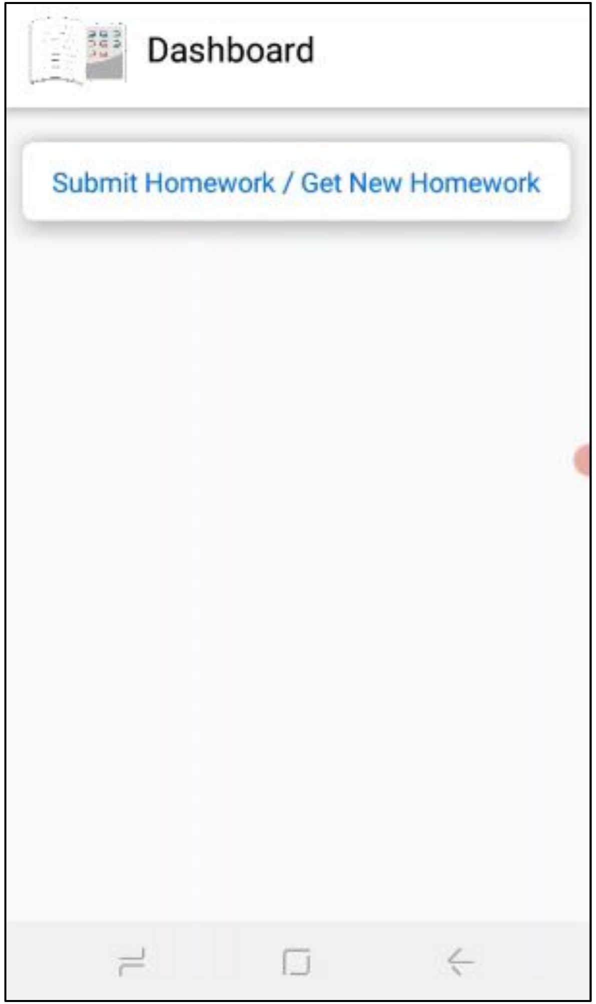
6. Promotions



7. Interviews

8. Examination

Written exam
and
signing test



Karya app

7
app versions

4603
sentences

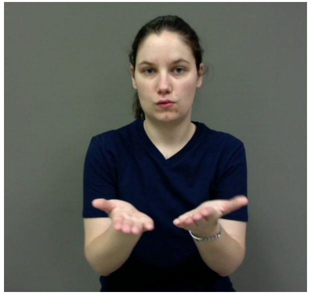
29
contributors

Crucial tool for online education for Deaf given low penetration of formal education

Models

🤲 OpenHands Library

Open-source pose-based efficient models



American
WLASL2000

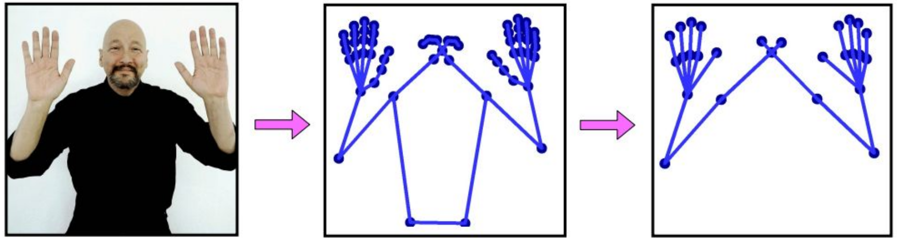
Chinese
DEVISIGN_L

Argentinian
LSA64

Indian
INCLUDE

Turkish
AUTSL

Greek
GSL



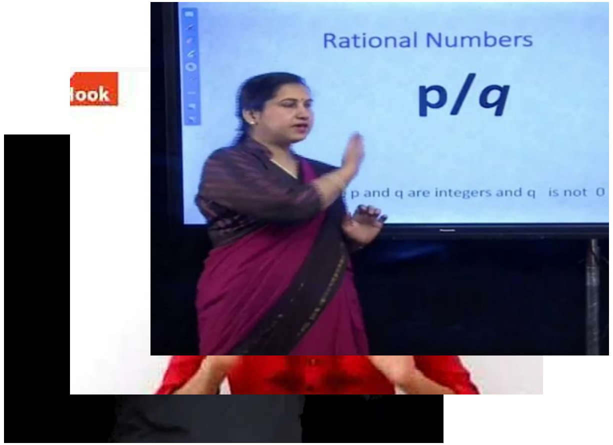
On CPU real-time Transformer-based networks achieving SOTA results on most languages

🙌 OpenHands Github repo

Self-supervised learning for sign language

Pre-training trick

With unlabeled video learn what sign language “looks like”
Then fine-tune on smaller amount of labelled data



1,129 hours
of ISL videos

91.2 to 94.7
accuracy on INCLUDE

Additional benefits for pre-training

Dataset (Language)	Videos per word	No pretraining	Pretraining on ISL
INCLUDE (Indian)	Full – Avg 17	91.2	94.7
	10	79.7	86.3
	5	45	57.4
	3	15.2	35.4

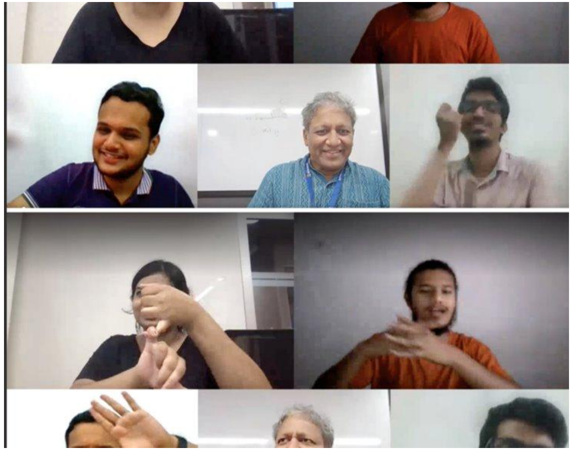
Pretraining is particularly effective when there are fewer examples per label

Pretraining in ISL is effectively transferring to other languages

What's next

People

Ludic Design



Joyful and inclusive play
between the Hearing and Deaf

Data

Curating labelled sign language
data from the web
Crowdsourcing 1,000 hours of
labelled data using Karya

Models

Pretrained multilingual models
Open-source pose-based models
for continuous sign language

"... we must be second to none in the application of advanced technologies to the real problems of man and society."

– Vikram Sarabhai

Research Roadmap

DL Modeling

- Model distillation
- Model compression
- Fast inference
- Training with noisy data

Multilinguality

- Multilingual Generation
- Mixture of Experts
- Curriculum Learning
- Romanized/code-mixed input

Self-supervised Learning

- Unsupervised + Supervised objectives
- Utilizing parallel data
- Low monolingual data scenarios

Multimodal Modeling

- Speech understanding
- Speech/Image Translation
- Multi-task modeling

We would love to engage with the community

Help build the IndicNLP Catalog

IndicNLP Catalog

Evolving, collaborative catalog of Indian language NLP resources

Please add resources you know of and send a pull request

- Major Indic Language NLP Repositories
- Libraries and Tools
- Evaluation Benchmarks
- Standards
- Text Corpora
  - Unicode Standard
  - Monolingual Corpus
  - Language Identification
  - Lexical Resources
  - NER Corpora
  - Parallel Translation Corpus
  - Parallel Transliteration Corpus
  - Text Classification
  - Textual Entailment/Natural Language Inference
  - Paraphrase
  - Sentiment, Sarcasm, Emotion Analysis
  - Question Answering
  - Dialog
  - Discourse
  - Information Extraction
  - POS Tagged corpus
  - Chunk Corpus
  - Dependency Parse Corpus
  - Co-reference Corpus
- Models
  - Word Embeddings
  - Sentence Embeddings
  - Multilingual Word Embeddings
  - Morphanalyzers
  - SMT Models
- Speech Corpora
- OCR Corpora
- Multimodal Corpora
- Language Specific Catalogs

Featured Resources

- AI4Bharat IndicNLPSuite: Text corpora, word embeddings, BERT for Indian languages and NLU resources for Indian languages.
- IIT Bombay English-Hindi Parallel Corpus: Largest en-hi parallel corpora in public domain (about 1.5 million semgents)
- CVIT-IIITH PIB Multilingual Corpus: Mined from Press Information Bureau for many Indian languages. Contains both English-IL and IL-IL corpora (IL=Indian language).
- CVIT-IIITH Mann ki Baat Corpus: Mined from Indian PM Narendra Modi's *Mann ki Baat* speeches.
- iNLTK: iNLTK aims to provide out of the box support for various NLP tasks that an application developer might need for Indic languages.
- Dakshina Dataset: The Dakshina dataset is a collection of text in both Latin and native scripts for 12 South Asian languages. Contains an aggregate of around 300k word pairs and 120k sentence pairs. Useful for transliteration.

Parallel Translation Corpus

- IIT Bombay English-Hindi Parallel Corpus: Largest en-hi parallel corpora in public domain (about 1.5 million semgents)
- CVIT-IIITH PIB Multilingual Corpus: Mined from Press Information Bureau for many Indian languages. Contains both English-IL and IL-IL corpora (IL=Indian language).
- CVIT-IIITH Mann ki Baat Corpus: Mined from Indian PM Narendra Modi's *Mann ki Baat* speeches.
- PMIndia: Parallel corpus for En-Indian languages mined from *Mann ki Baat* speeches of the PM of India (paper).
- Indian Language Corpora Initiative: Available on TDIL portal on request
- OPUS corpus
- WAT 2018 Parallel Corpus: There may significant overlap between WAT and OPUS.
- Charles University English-Hindi Parallel Corpus: This is included in the IITB parallel corpus.
- Charles University English-Tamil Parallel Corpus
- Charles University English-Odia Parallel Corpus v1.0
- Charles University English-Odia Parallel Corpus v2.0
- Charles University English-Urdu Religious Parallel Corpus
- IndoWordnet Parallel Corpus: Parallel corpora mined from IndoWordNet gloss and/or examples for Indian-Indian language corpora (6.3 million segments, 18 languages).
- MTurk Indian Parallel Corpus
- TED Parallel Corpus
- JW300 Corpus: Parallel corpus mined from jw.org. Religious text from Jehovah's Witness.
- ALT Parallel Corpus: 10k sentences for Bengali, Hindi in parallel with English and many East Asian languages.
- FLORES dataset: English-Sinhala and English-Nepali corpora
- Uka Tarsadia University Corpus: 65k English-Gujarati sentence pairs. Corpus is described in this paper
- NLPC-UoM English-Tamil Corpus: 9k sentences, 24k glossary terms

https://github.com/AI4Bharat/indicnlp_catalog

We would love to engage with the community

Help build the IndicNLP Catalog

Feedback/ feature-requests on models/datasets

Discovering datasources

Educate us on important usecases

Resources we have created so far

Indic BERT - https://indicnlp.ai4bharat.org/indic-bert/
Indic monolingual corpus - https://indicnlp.ai4bharat.org/corpora
IndicNLP suite - https://indicnlp.ai4bharat.org/home/
Samanantar bitext corpus -
https://storage.googleapis.com/samanantar-public/V0.3/source_wise_splits.zip
Translation models - https://github.com/AI4Bharat/indicTrans
ASR dataset and models - https://indicnlp.ai4bharat.org/indicwav2vec/
INCLUDE dataset - https://zenodo.org/record/4010759
OpenHands sign language models - https://github.com/AI4Bharat/OpenHands

Thank you!

Website: https://indicnlp.ai4bharat.org
Github: https://github.com/AI4Bharat

Contact: miteshk@cse.iitm.ac.in
pratyush.k.panda@gmail.com
anoop.kunchukuttan@gmail.com