Mining Datasets at scale for Building High-quality NLP Models

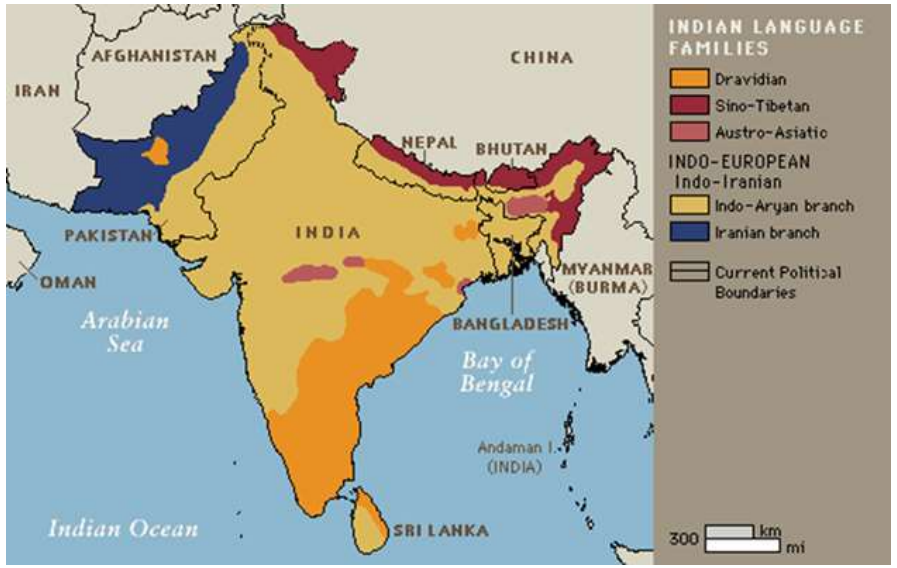
Anoop Kunchukuttan

Microsoft Translator, Hyderabad

AI4Bharat

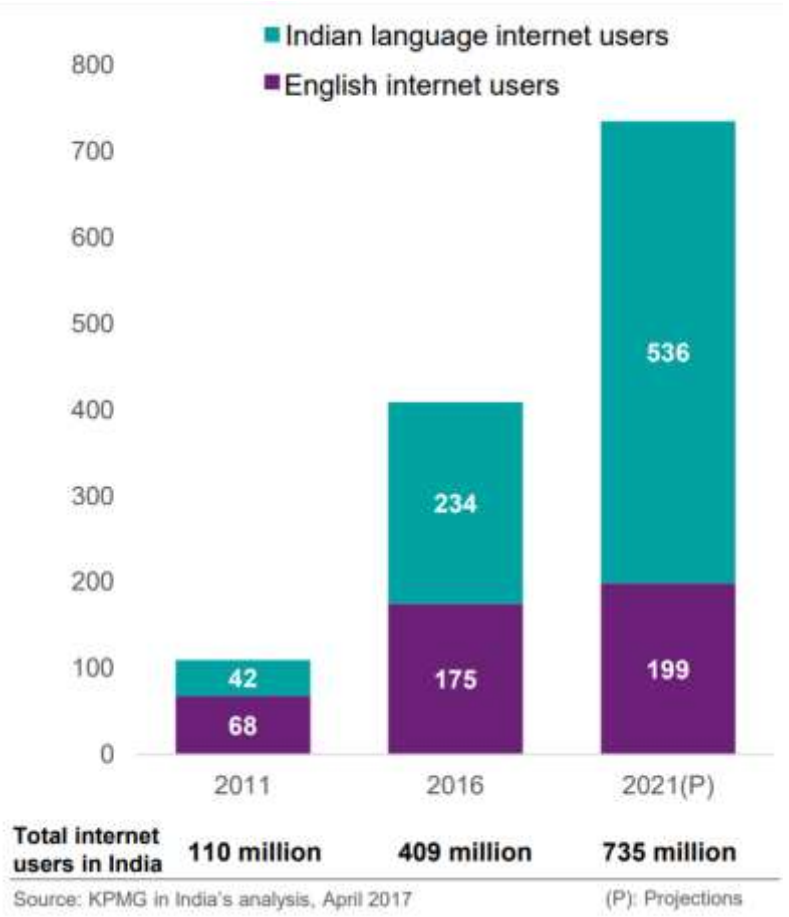
AI Department Day, IIT Hyderabad, 24 January 2023

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*

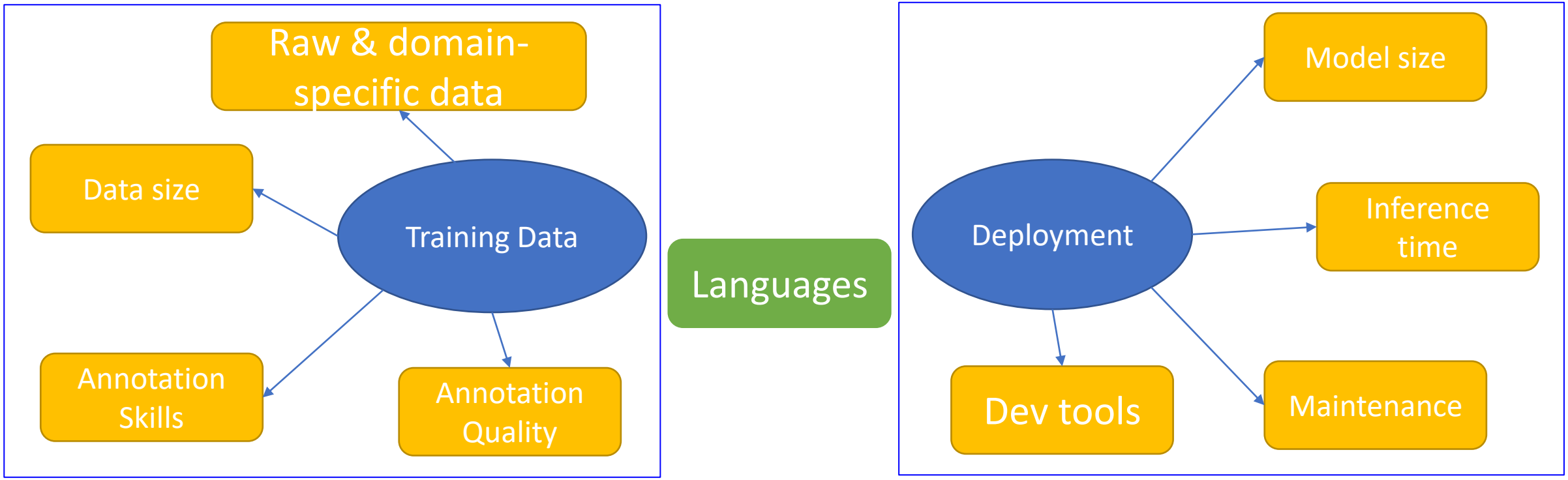
Sources: Wikipedia, Census of India 2011



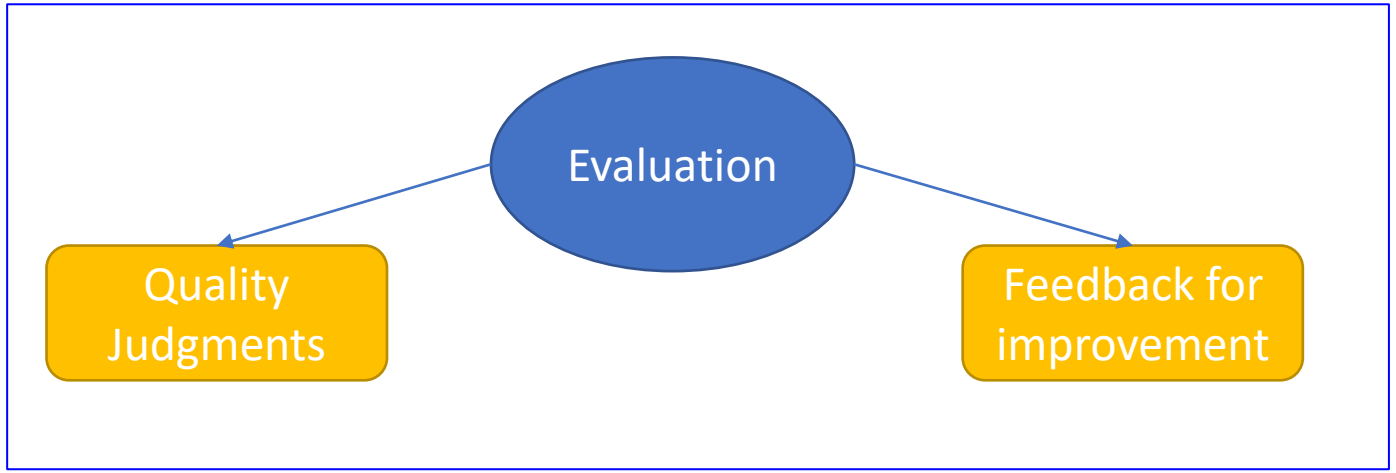
Internet User Base in India (in million)

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

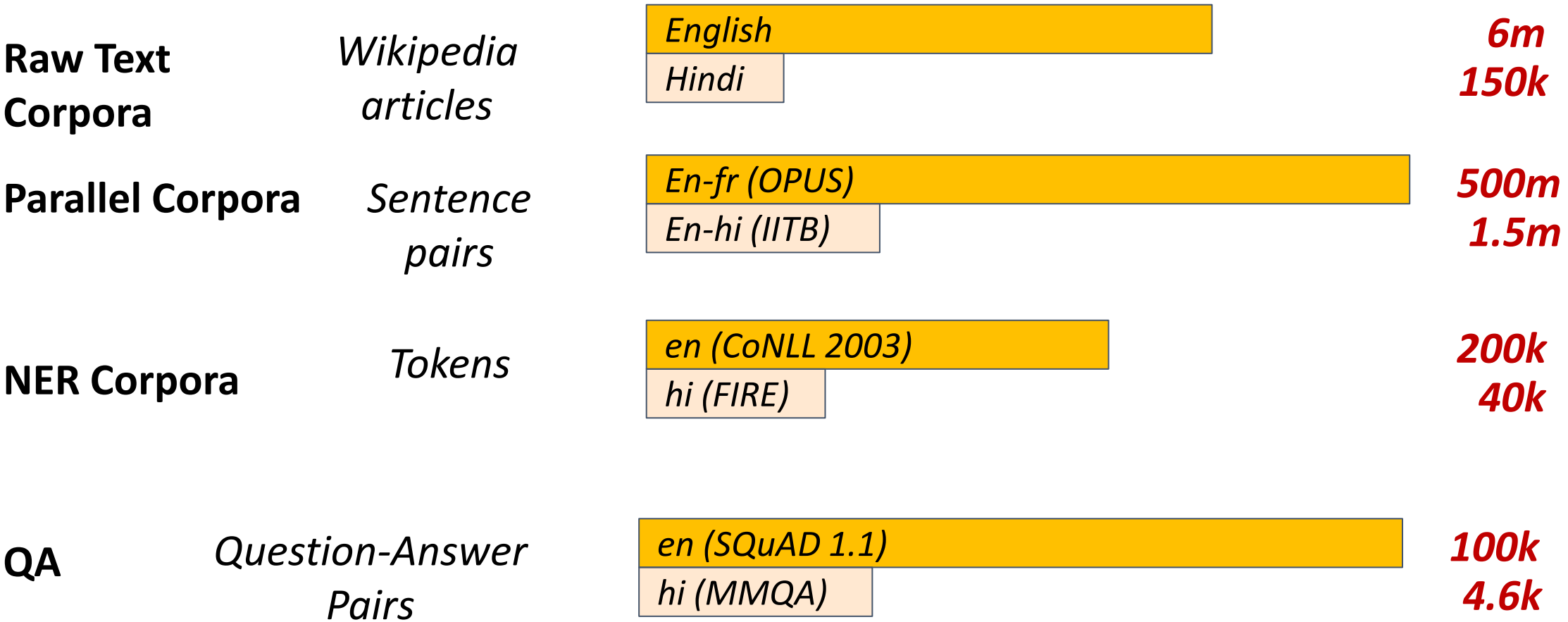
Scalability Challenges for NLP solutions



Effort and cost increase as languages increase



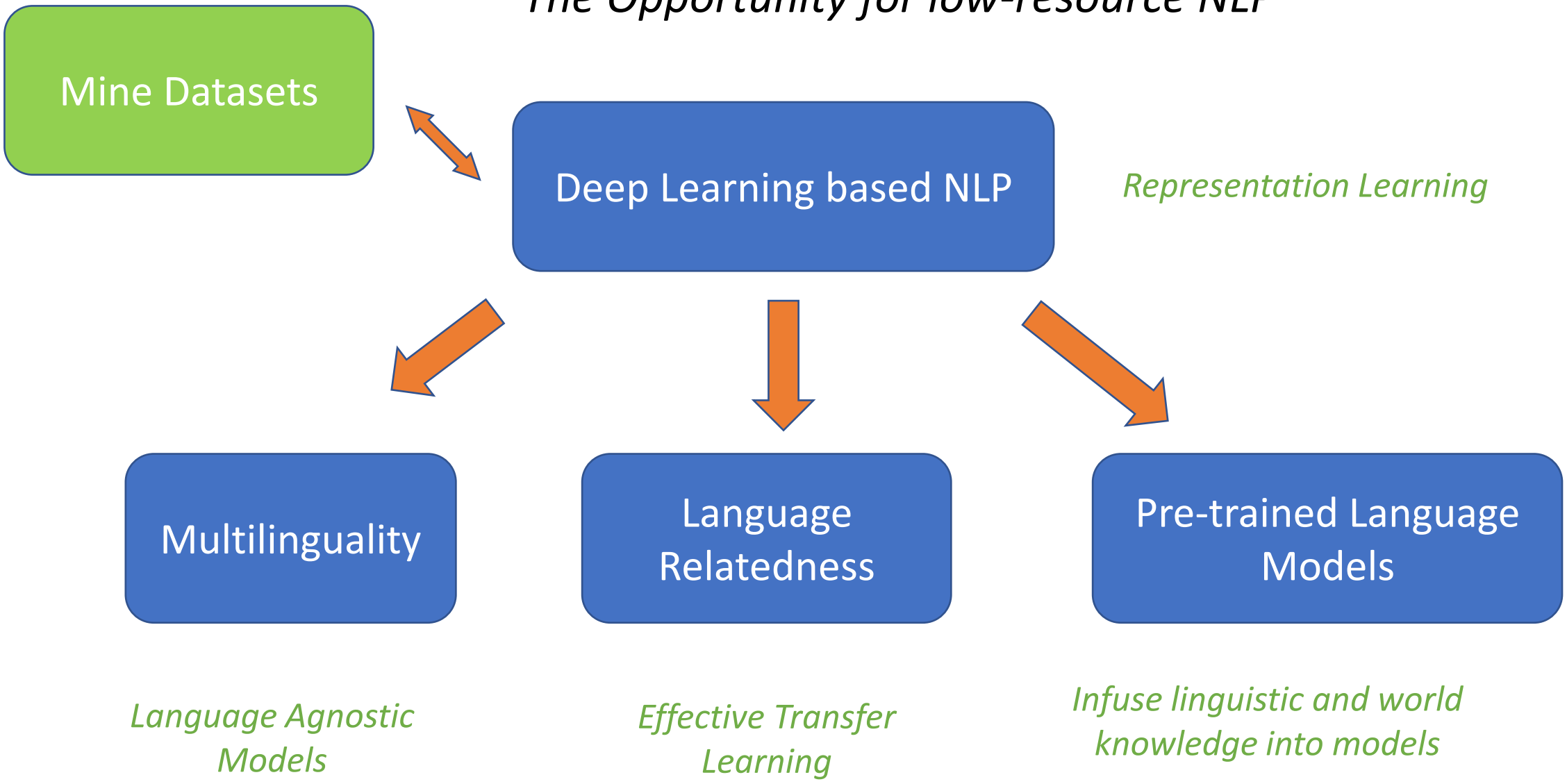
We are faced with a huge data skew



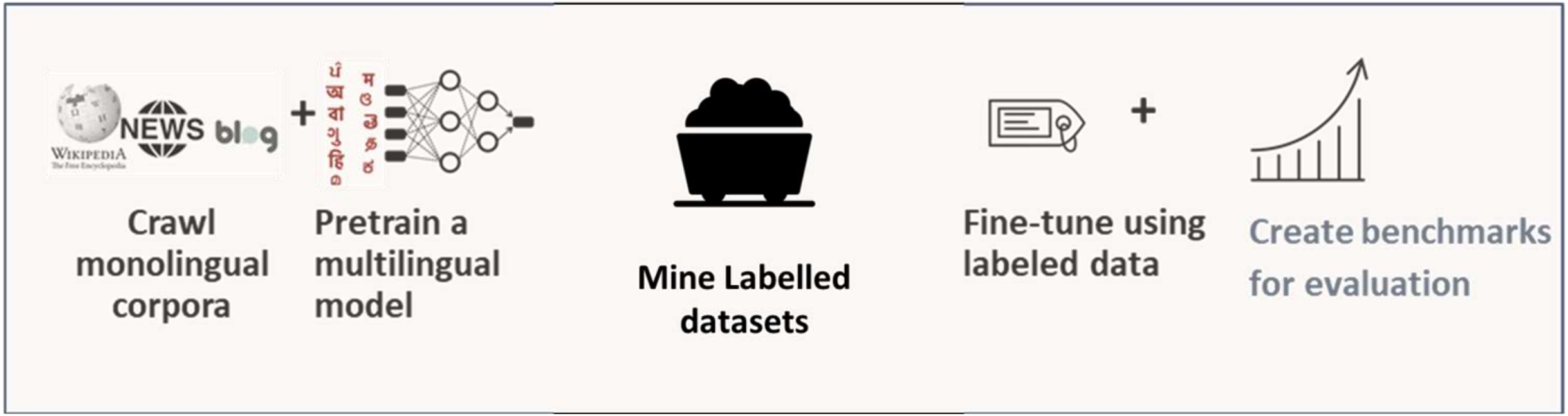
How do we approach this problem?

Our Technical Direction

The Opportunity for low-resource NLP



The “Recipe” for Language Scalability



Our Contributions

NLP Infrastructure: Raw corpora & language models

Data and models for various foundational tasks

Standard Evaluation Benchmarks

https://ai4bharat.iitm.ac.in/datasets

https://ai4bharat.iitm.ac.in/models

NLP Infrastructure: Raw corpora & language models

IndicCorp

Large Monolingual corpora

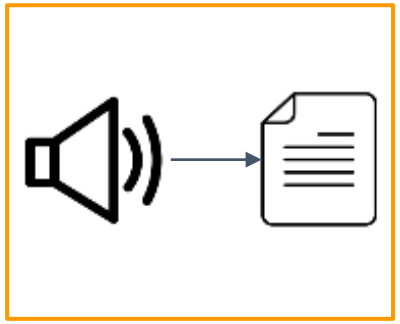
20B tokens, 24 languages

IndicBERT

(masked LM)

IndicBART

(seq2seq LM)



Dhwani

Raw speech corpora

(17k hours, 40 languages)

IndicFT

(word embeddings)

IndicWav2Vec

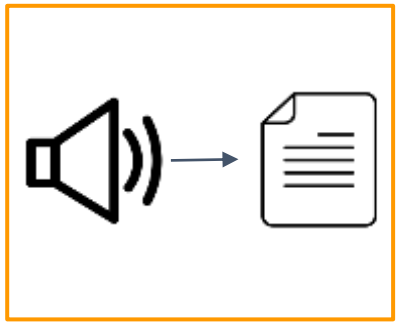
Pre-trained speech representations

Compact pre-trained models for NLU & NLG

Data and models for various foundational tasks

Samanantar

Parallel corpus, translation models between English & 11 Indic languages

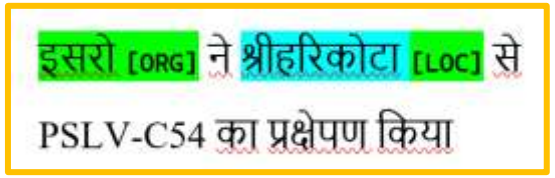


Shrutilipi & KathBath

ASR datasets & models for 12 Indian languages

Aksharantar

Transliteration Models & datasets for 20 Indic languages



Naamapadam

Datasets and models for Named Entity Recognition in 11 Indian languages

Standard Evaluation Benchmarks

IndicGLUE *In-language Benchmarks for Natural Language Understanding*

IndicXTREME *Cross-lingual Benchmarks for Natural Language Understanding*

Datasets for tasks like question answering, paraphrase detection, sentiment analysis, article classification, COPA, WNLI, etc

Indic NLG Suite *Benchmarks for Natural Language Generation*

Datasets for tasks like headline generation, paraphrase generation, question generation, sentence summarization

Indic SUPERB *Benchmarks for Speech Language Understanding*

Datasets for tasks like Automatic Speech Recognition, speaker verification, speaker identification (mono/multi), language identification, Query By Example, and keyword spotting

Mining Resources and building Models for Indian Language NLU and NLG

1. 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*. EMNLP-Findings. 2020.
2. 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*. ACL-Findings. 2022.
3. Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, Pratyush Kumar. *IndicNLG Suite: Multilingual Datasets for Diverse NLG Tasks in Indic Languages.* EMNLP. 2022.
4. Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating Multilingual Inference for Indian Languages. EMNLP 2022.
5. Doddapaneni, Sumanth, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. *IndicXTREME: A Multi-Task Benchmark For Evaluating Indic Languages.* arXiv preprint arXiv:2212.05409. 2022.

IndicCorp

https://ai4bharat.iitm.ac.in/corpora

23 Indic languages

(+Indian English)

20 B tokens

1.1 B sentences

General domain

1000+ Sources

Data filtering, offensive text removal

	Wikipedia	CC-100	mC4	IndicCorp
#Indic lang.	20	12	15	23
#Indic lang. tokens	0.2B	5.0B	20.2B[3]	14.4B
Verified source URLs	✓	✗	✗	✓

Crawled with the WebCorpus framework
https://github.com/AI4Bharat/webcorpus

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

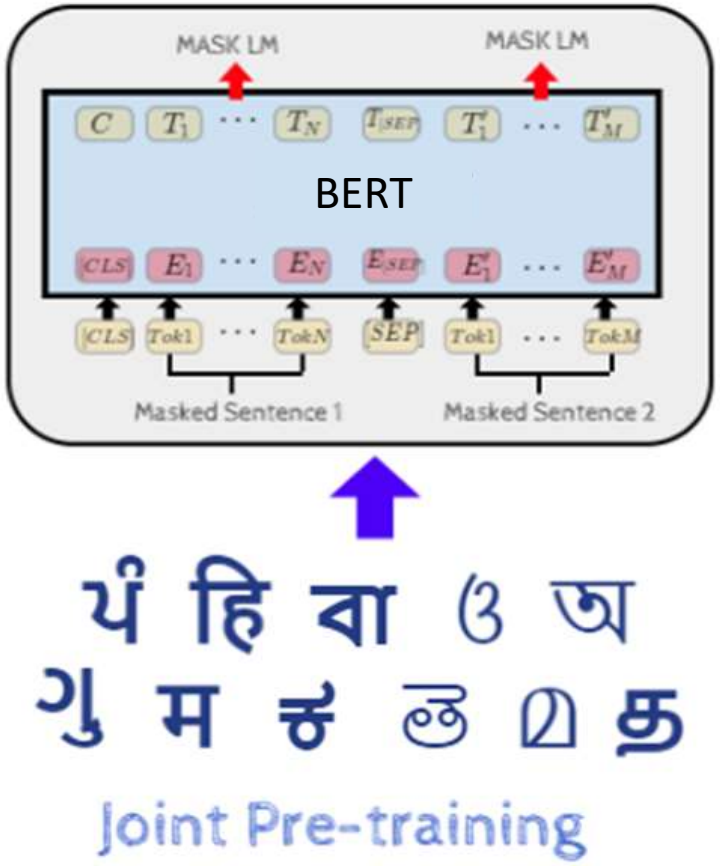
Language Generation

Benchmark Datasets

IndicBERT

https://ai4bharat.iitm.ac.in/bertv2

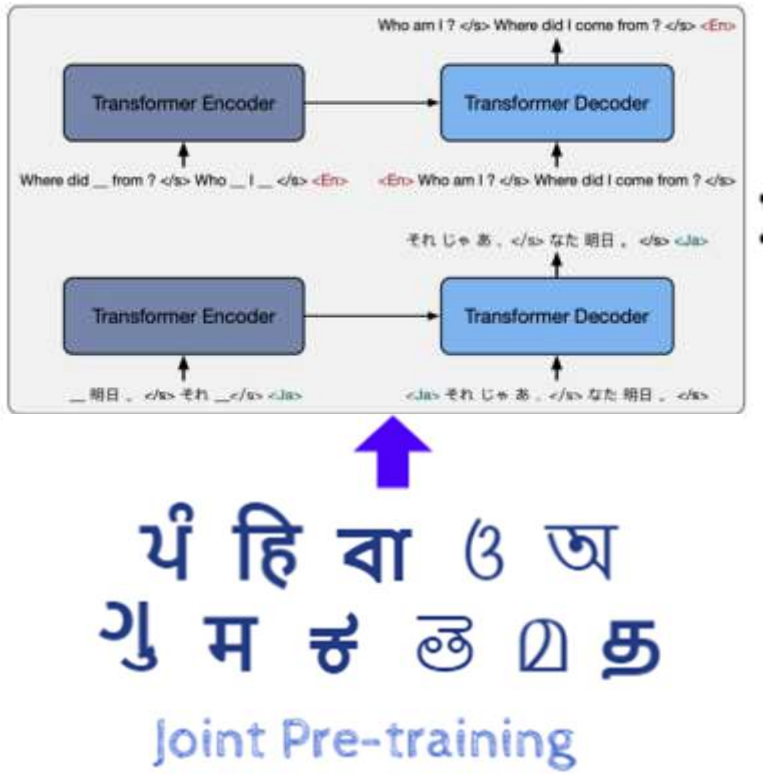
https://huggingface.co/ai4bharat/IndicBERTv2-MLM-only



- Pre-trained Indic LM for NLU applications
- Large Indian language content
  - 23 Indian languages
  - + Indian English content
- Available in MLM/TLM variants
- Multilingual Model
- Better than mBERT/XLM-R/MuRIL on IndicXTREME
- Simplify fine-tune for your application

IndicBART

https://indicnlp.ai4bharat.org/indic-bart
https://huggingface.co/ai4bharat/IndicBART



- 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.* Findings of ACL. 2022.

Key Results

- Language group specific pre-trained models are better
  - Compact
  - Competitive with large massively multilingual models like mBERT, mBART
  - Flexibility in curation of content
- Multilingual fine-tuning and pre-training are useful
  - Particularly for low-resource languages

IndicGLUE *(Indic General Language Understanding Evaluation Benchmark – In Language)*

New tasks

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

Difficult tasks

Span all languages

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.

Careful Negative Sampling

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

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: 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

https://indianexpress.com/article/sports/cricket/ipl2021-Mayank-agarwal

Indic NLG Suite *(Datasets for Indian language generation tasks)*

Dataset	Languages	Communicative Intent	Input Type	Total Size
Biography Generation	as, bn, hi, kn, ml, or, pa, ta, te	One-sentence biographies	key-value pairs	55K
Headline Generation	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te	News article headlines	news article	1.43M
Sentence Summarization	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te	Compacted sentence with same meaning	sentence	431K
Paraphrase Generation	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te	Synonymous sentence	sentence	5.57M
Question Generation	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te	Question leading to answer given context	context-answer pairs	1.08M

Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, Pratyush Kumar. *IndicNLG Suite: Multilingual Datasets for Diverse NLG Tasks in Indic Languages.* EMNLP. 2022.

Biography Generation

कैप्टन मनोज कुमार पांडेय परमवीर चक्र	
जन्म	25 जून 1975 सीतापुर, उत्तर प्रदेश.
देहांत	3 जुलाई 1999 (उम्र 24) कारगिल युद्ध के दौरान बटालिक सेक्टर, कारगिल, जम्मू और कश्मीर
निष्ठा	भारत
सेवा/ शाखा	भारतीय सेना
उपाधि	कैप्टन, भारतीय सेना
दस्ता	1/11 गोरखा राइफल्स
युद्ध/ झड़पें	कारगिल युद्ध ऑपरेशन विजय
सम्मान	परमवीर चक्र

कैप्टन **मनोज कुमार पांडेय** भारतीय सेना के अधिकारी थे जिन्हें सन १९९९ के कारगिल युद्ध में असाधारण वीरता के लिए मरणोपरांत भारत के सर्वोच्च वीरता पदक परमवीर चक्र से सम्मानित किया गया।

Paraphrase Generation

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

Input

दिल्ली यूनिवर्सिटी देश की प्रसिद्ध यूनिवर्सिटी में से एक है

Output

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

Innovative methods for mining task-specific datasets

Samanantar

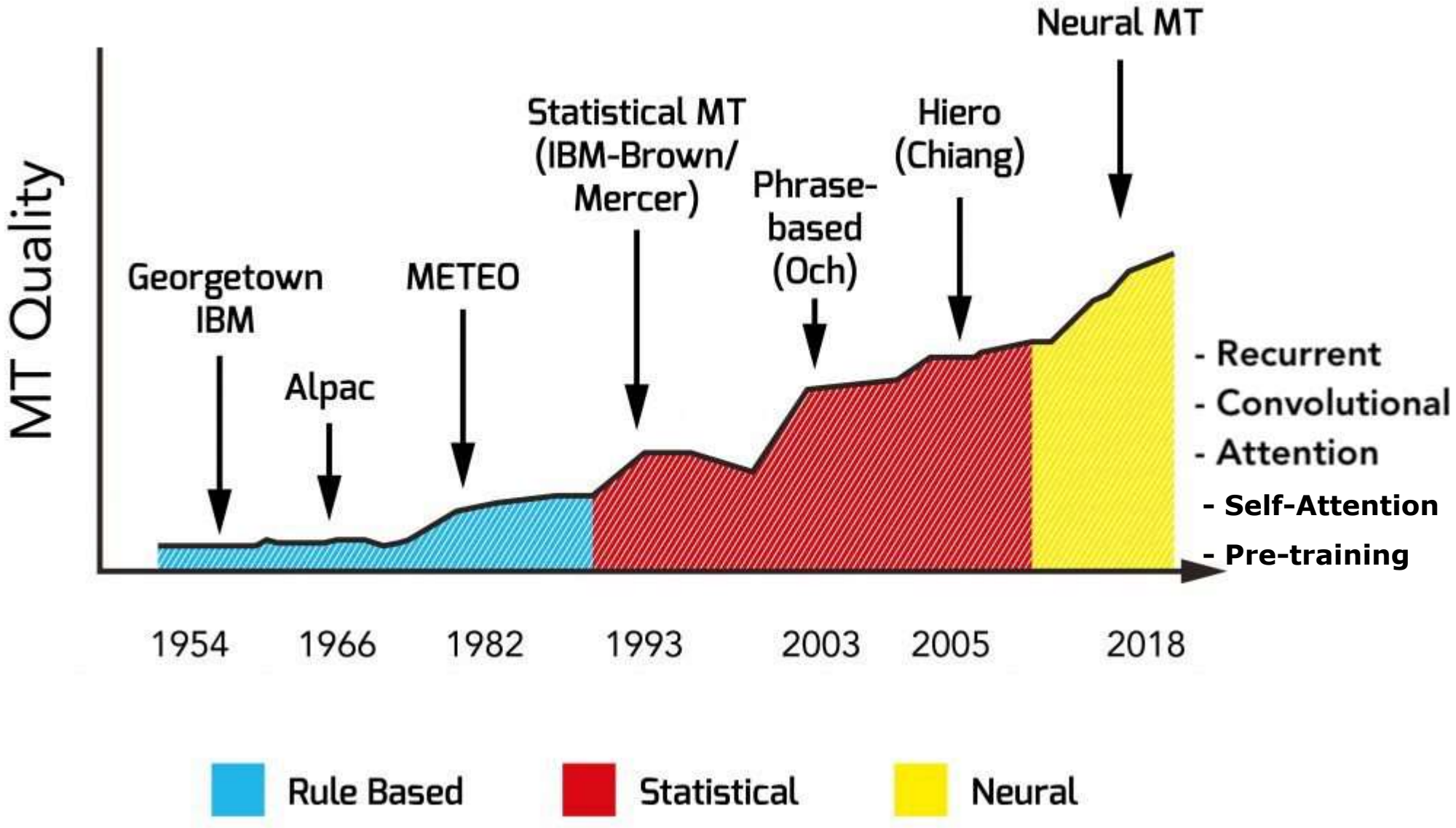
The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages

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

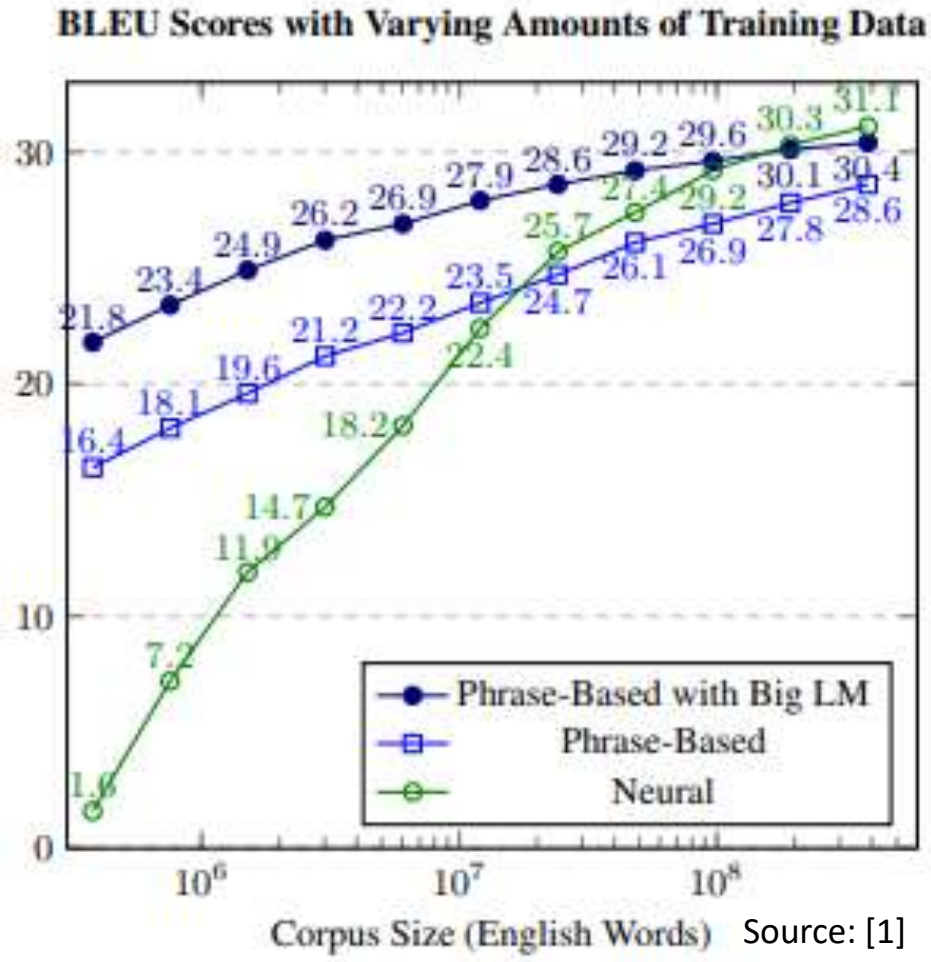
AI4Bharat, EkStep, IITM, Microsoft, RBCDSAI, Tarento

TACL 2022

https://ai4bharat.iitm.ac.in/samanantar



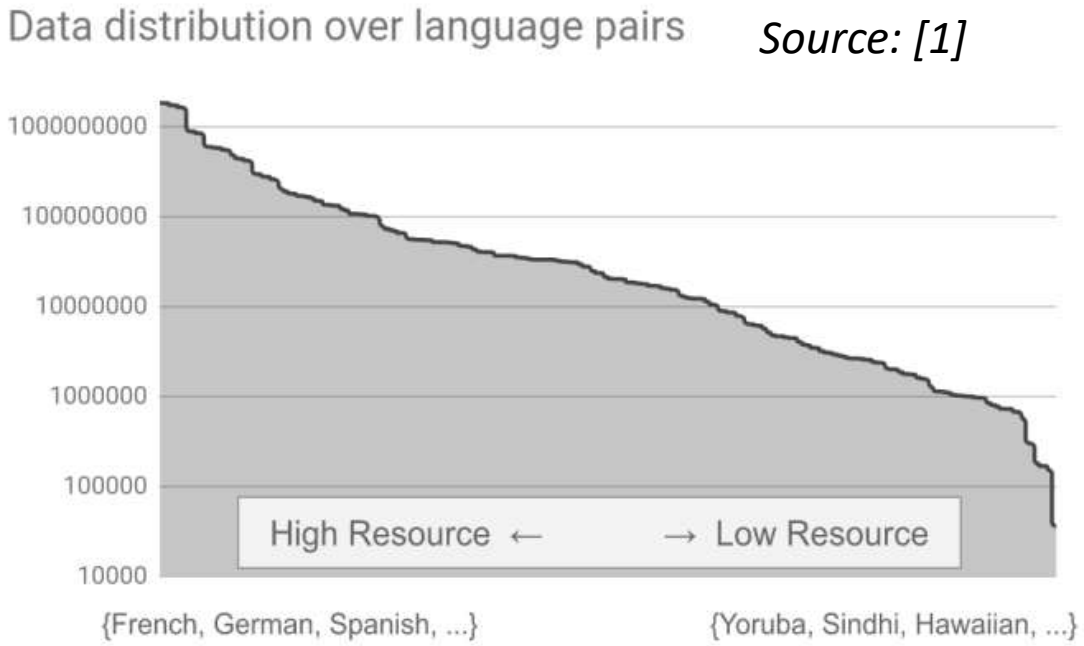
Adapted from https://iconictranslation.com/what-we-do/neural-machine-translation/



Source: [1]

Translation Quality improves with increasing parallel corpus size

1. Philipp Koehn, Rebecca Knowles. Six Challenges for Neural Machine Translation. W-NMT. 2017.



Source: [1]

Availability of parallel corpora varies widely across languages

Publicly available parallel corpora for Indian languages was very small

bn	gu	hi	kn	ml	mr	or	pa	ta	te	Grand Total
1,302,737	517,901	3,069,364	396,852	1,142,011	621,328	252,160	518,499	1,354,152	457,402	9,632,406

WAT 2021 shared task corpus stats (number of sentence pairs) Source: [2]

1. Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, Yonghui Wu. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. 2019. https://arxiv.org/abs/1907.05019.
2. Nakazawa, Toshiaki, et al. "Overview of the 8th workshop on Asian translation." *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*. 2021.

Samanantar Parallel Corpora

Parallel corpora for 11 Indian Languages + English

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

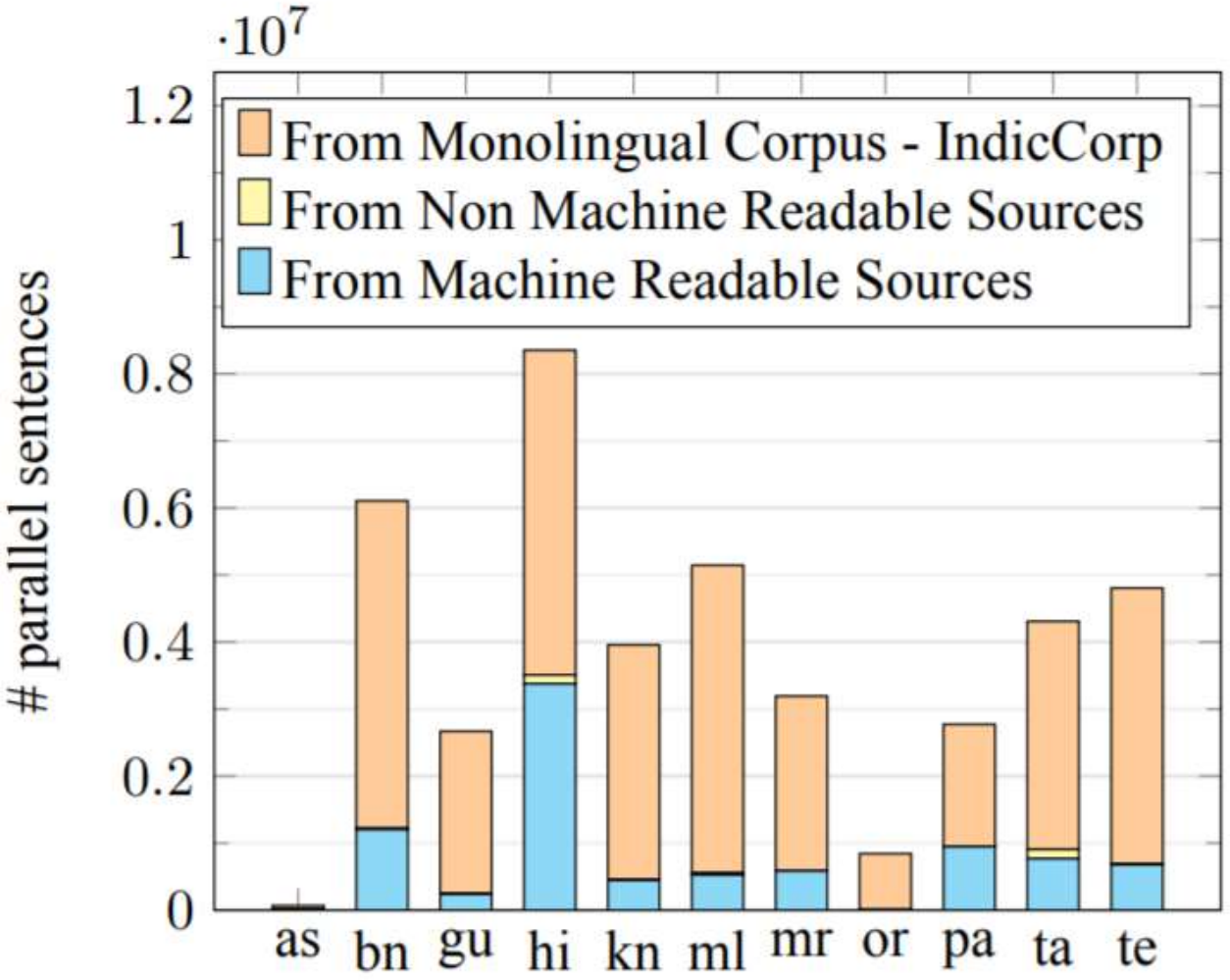
	#lang-pair	#sent-pair (million)
English-Indic languages	11	49.7
Indic-Indic languages	55	83.4

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)



Mining from monolingual corpora is the largest contributor to Samanantar

Going beyond comparable corpora

Discovering parallel sources is non-trivial

Not necessarily Regular URL patterns across websites	
https://zeenews.india.com/news/india/pm-modis-jk-visit-on-diwali-as-it-happened_1488741.html	https://zeenews.india.com/hindi/india/pm-narendra-modi-meets-soldiers-in-jk-wishes-happy-diwali-from-siachen/236490
Parallel content can exist across different domains	
https://english.jagran.com/india/sorry-state-of-affairs-chief-justice-nv-ramana-on-lack-of-debate-in-parliament-10030745	https://hindi.theprint.in/india/its-a-sorry-state-of-affairs-in-parliament-there-is-no-clarity-in-laws-cji-ramana-says/233719
Sometimes, it is difficult to say that the websites are parallel	
https://nagalandpage.com/sunil-chhetri-overtakes-messi	https://newswing.com/charismatic-striker-chhetri-overtakes-messi-just-one-step-behind-all-time-top-10/261946

Going beyond comparable corpora

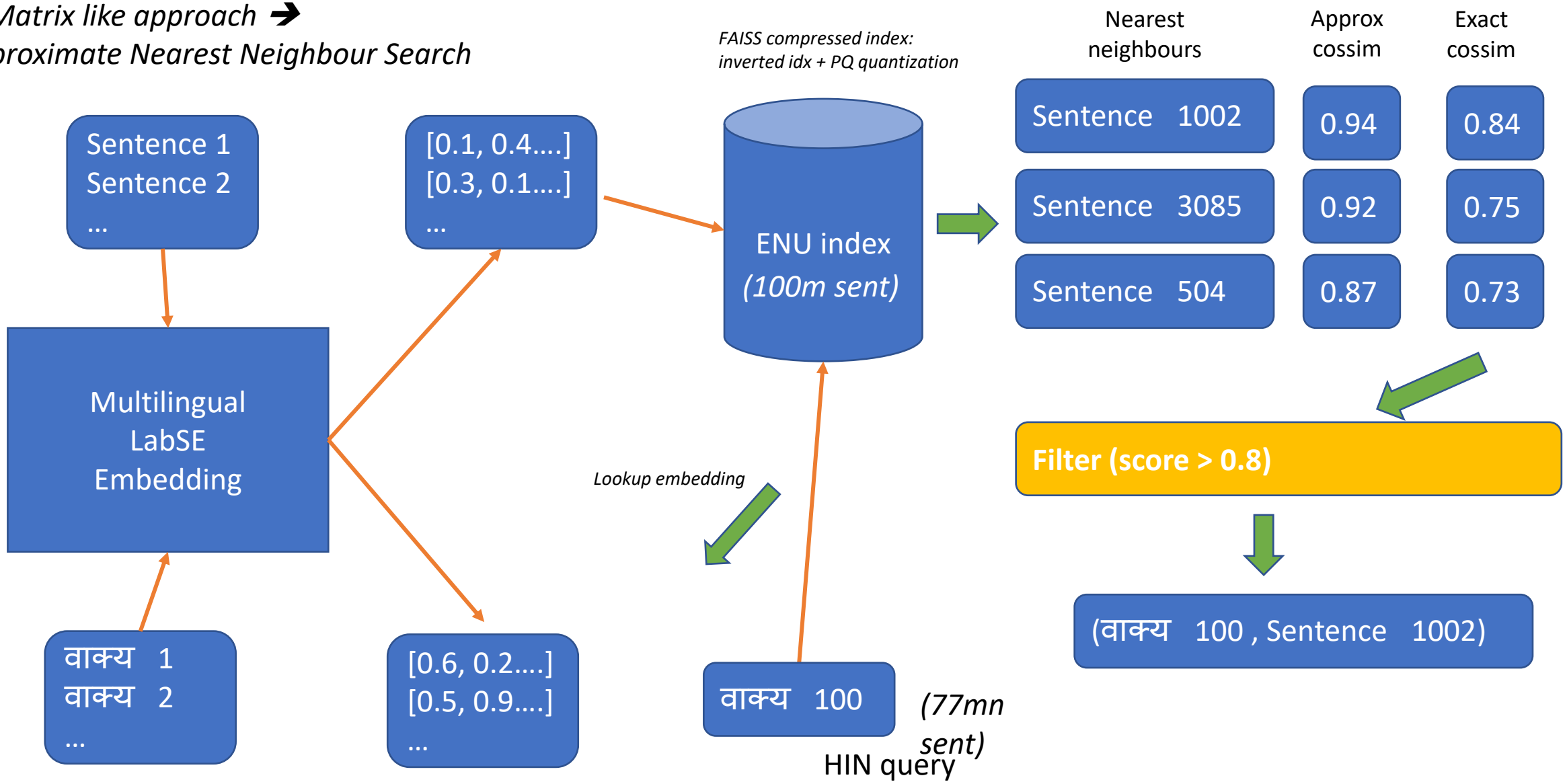
Audacious goal: can we mine parallel data from just large monolingual corpora

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

Parallel Corpus Mining from Monolingual Data

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

CCMatrix like approach ➔
Approximate Nearest Neighbour Search



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
  - 768d vector compressed from 3072 bytes to 72 bytes (+constant costs)
- Distributing indices over multiple GPUs
- Searching over multiple indices *(to speed up searches)*

FAISS - Billion-scale similarity search with GPUs, Jeff Johnson, Matthijs Douze, Hervé Jégou, ArXiv 2017

https://www.pinecone.io/learn/product-quantization/

Qualitative Analysis of the parallel corpus

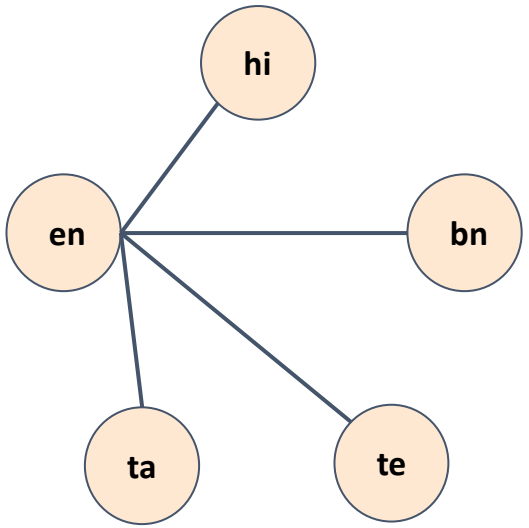
10000 samples manually evaluated using 30+ annotators across 11 languages
Using SemEval-1 guidelines for cross-lingual semantic textual similarity
Available for **cross-lingual STS** studies (https://storage.googleapis.com/samanantar-public/human_annotations.tsv)

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

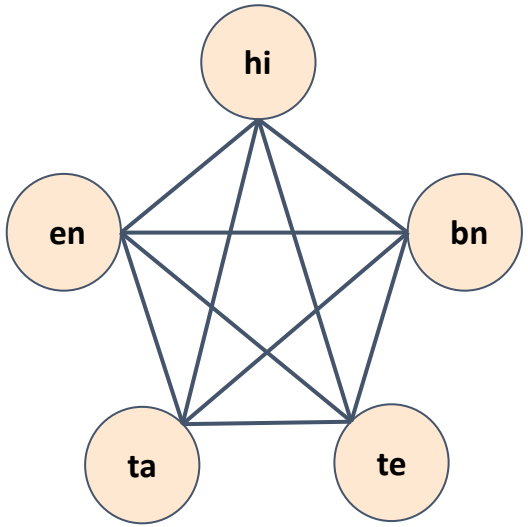
Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. *SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation.* SemEva.

Mining between Indic Languages

Mine Indic-Indic parallel corpora from English to Indic corpora

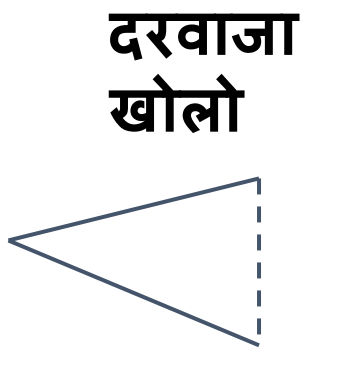


English-centric



Complete

Open the door

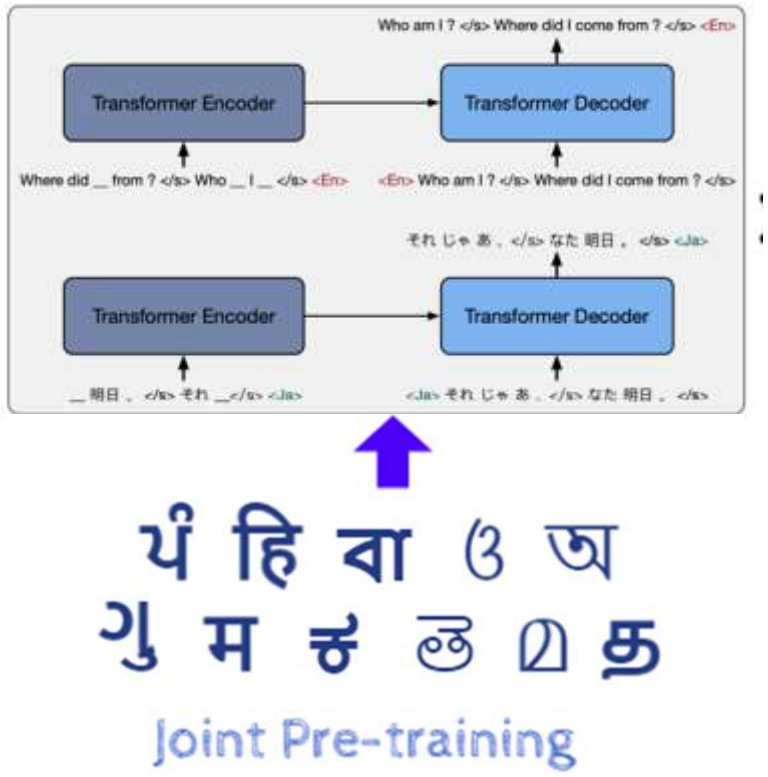


83.7 million sentence pairs for 55 language pairs

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://ai4bharat.iitm.ac.in/indic-trans



- Trained on Samanantar parallel corpus
- Multilingual Model (en→IL, IL→en, IL→IL)
- Single Script
- Input and output language tags
- Model size: (~430m params)
- Best performing open-source model for Indian languages

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.

Combine Corpora from different languages

(Nguyen and Chang, 2017)

I am going home	હુ ઘરે જવ છૂ
It rained last week	છેલ્લા આઠવડિયા મા વર્સાદ પાડ્યો

It is cold in Pune	पुण्यात थंड आहे
My home is near the market	माझा घर बाजाराजवळ आहे

Convert Script

Concat Corpora

I am going home	हु घरे जव छू
It rained last week	छेल्ला आठवडिया मा वर्साद पाड्यो
It is cold in Pune	पुण्यात थंड आहे
My home is near the market	माझा घर बाजाराजवळ आहे

Mining Named Entity Datasets

Mhaske, Arnav, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy V, and Anoop Kunchukuttan. "Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages." *arXiv preprint arXiv:2212.10168* (2022).

Summary

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 9 Indic languages
  - Bn, Hi, Kn, Ml, Mr, Ta, Te, Gu, Pa

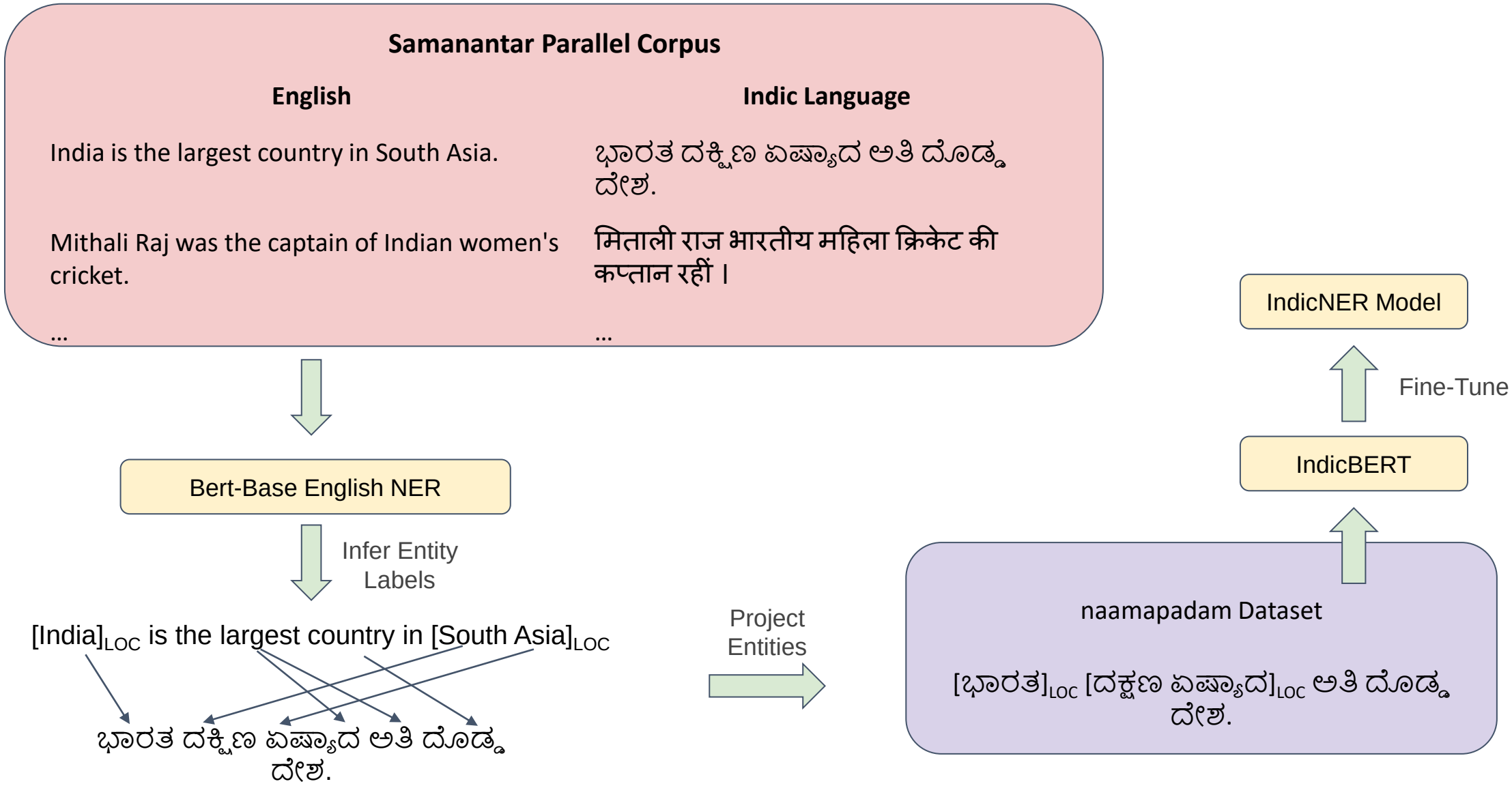
Multilingual IndicNER model

- 11 Indic languages (As, Bn, Gu, Hi, Kn, Ml, Mr, Or, Pa, Ta, Te)

(Model) https://huggingface.co/ai4bharat/IndicNER

(Dataset) https://huggingface.co/datasets/ai4bharat/naamapadam

Naamapadam Dataset and IndicNER Model



Possible to mine large datasets

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

	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te
Naamapadam	5.0K	1.6M	769.3K	2.2M	658K	1.0M	735.0K	190.0K	880.2K	745.2K	751.1K
WikiANN	218	12K	264	7.3K	220	13K	7.3K	265	211	19.7K	2.4K
FIRE-2014	-	6.1K	-	3.5K	-	4.2K	-	-	-	3.2K	-
CFILT	-	-	-	262.1K	-	-	4.8K	-	-	-	-
MultiCoNER	-	9.9K	-	10.5K	-	-	-	-	-	-	-
MahaNER	-	-	-	-	-	-	16K	-	-	-	-
AsNER$^{\phi}$	6K	-	-	-	-	-	-	-	-	-	-

Accurate projections (>70 F1-Score compared with human annotations)

bn	gu	hi	kn	ml	mr	ta	te	Average
82.11	69.77	**90.32**	70.22	**69.83**	**76.51**	**70.09**	77.70	**75.82**

Testsets were created by volunteers

High annotator agreement on this task

Language		Token-level Cohen's Kappa
Bengali	bn	83.28
Gujarati	gu	80.85
Hindi	hi	80.90
Kannada	kn	74.06
Malayalam	ml	69.58
Marathi	mr	78.03
Punjabi	pa	70.19
Tamil	ta	71.74
Telugu	te	89.98

Results

Language	Naamapadam	FIRE-2014	WikiANN	MultiCoNER	CFILT	MahaNER
bn	81.02 ± 0.40	35.68 ± 3.96	51.67 ± 1.24	26.12 ± 1.96	-	-
gu	80.59 ± 0.57	-	0.11 ± 0.12	-	-	-
hi	82.69 ± 0.45	47.23 ± 0.92	59.84 ± 1.25	41.85 ± 2.34	75.71 ± 0.67	-
kn	80.33 ± 0.60	-	2.73 ± 1.47	-	-	-
ml	81.49 ± 0.15	58.51 ± 1.13	62.59 ± 0.32	-	-	-
mr	81.37 ± 0.29	-	62.37 ± 1.12	-	58.41 ± 0.62	71.45 ± 1.44
pa	71.51 ± 0.59	-	0.7 ± 0.37	-	-	-
ta	73.36 ± 0.56	44.89 ± 0.94	49.15 ± 1.17	-	-	-
te	82.49 ± 0.60	-	49.28 ± 2.17	-	-	-

Table 8: Comparison of models trained on different datasets and evaluated on Naamapadam-test set (F1 score).

mBERT model fine-tuned on Naamapadam train outperforms models fine-tuned on existing datasets

Better than zeroshot NER

	PER	LOC	ORG	Overall
bn	77.63	84.29	73.25	80.06
gu	81.14	88.65	67.63	80.83
hi	82.31	89.37	74.03	83.27
kn	78.16	87.29	73.12	81.28
ml	84.49	87.85	61.49	81.67
mr	83.70	88.66	66.33	81.88
pa	76.26	77.95	55.68	72.08
ta	76.01	83.09	58.73	74.48
te	84.38	84.77	70.92	81.90
as	75.00	54.55	57.14	62.50
or	41.78	21.40	13.39	26.42

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

Transliteration Mining

Anoop Kunchukuttan, Siddharth Jain, Rahulk Kejriwal. *A Large-scale Evaluation of Neural Machine Transliteration for Indic Languages. EACL 2021.*

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra. *Aksharantar: Towards building open transliteration tools for the next billion users. Arxiv pre-prnt 2205.03018 .* 2022.

What is transliteration?

Transliteration

“conversion of text from one script to another such that (i) it is phonetically equivalent to the source name and (ii) it matches the user intuition on its equivalence wrt the source text”

Ethanur

एत्तनूर	ഏത്തനൂർ
(ettanUra)	(.ettanUr)

Useful for

- Romanized input
- Romanized search, translation, etc

Related Work

- Small datasets
  - MSR-NEWS (Banchs et al., 2015)
  - BrahmiNet (Kunchukuttan et al., 2015)
  - Dakshina (Roark et al., 2020)
  - Others (Kunchukuttan et al., 2018b; Gupta et al. 2012; Khapra et al., 2014)
- Most dataset span few languages
- Lack of comprehensive testsets
  - Limited analysis of foreign/India word performance
- Limited work on multilingual/joint transliteration (Kunchukuttan et al., 2018, 2021)

Mine Large-scale Transliteration Corpora

- *From parallel translation corpora*
- *From monolingual corpora*
- *Obtain transliterations from human judgments*

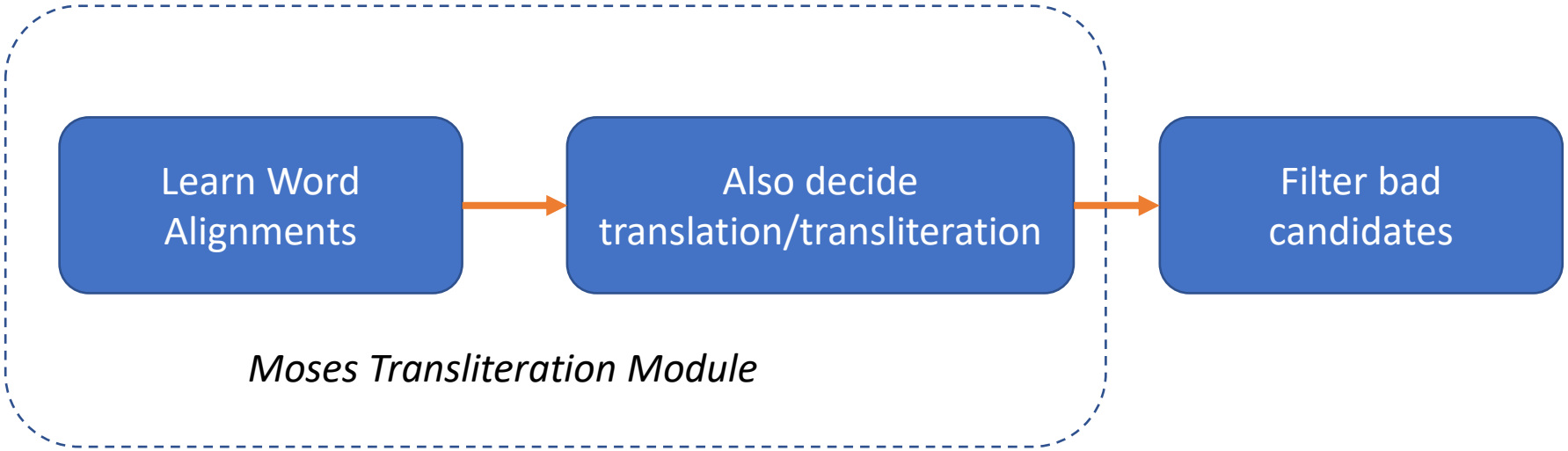
From Parallel Translation Corpora

(Sajjad et al., 2012; Durrani et al., 2014)

A boy is sitting in the kitchen	एक लडका रसोई मे बैठा है
A boy is sitting on a round table	एक लडका एक गोल मेज पर बैठा है
Rafale aircrafts arrived in Ambala	राफेल विमान अंबाला पहुंचे
Rafale is manufactured in France	राफेल फ्रांस में निर्मित होता है

Word alignment probability is a linear interpolation of a transliteration model (p_1) and non-transliteration model (p_2) .

$$p(e,f) = (1-\lambda)\, p_1(e,f) + \lambda\, p_2(e,f)$$



Score thresholding, soundex matches and morphological variant elimination

From Monolingual Corpora

From AI4Bharat-IndicNLP Corpus (Kunchukuttan et al., 2020)

Train an initial transliteration model

Score transliteration candidates

Select best candidates

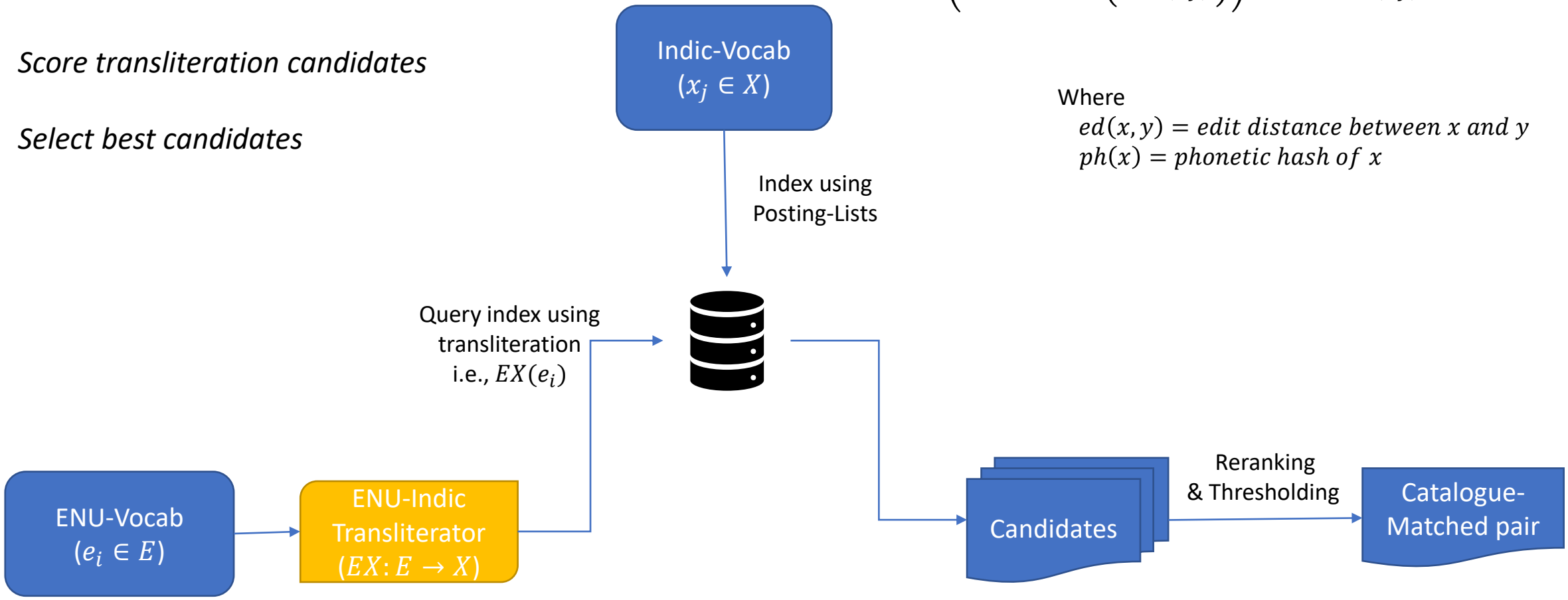
Reranking:

$$s(e_i, x_j) = ed\left(e_i, XE(x_j)\right) + ed\left(x_j, EX(e_i)\right) + ed\left(ph(e_i), ph\left(XE(x_j)\right)\right) + ed(ph(x_j), ph(EX(e_i)))$$

Where

$ed(x, y) = edit\ distance\ between\ x\ and\ y$

$ph(x) = phonetic\ hash\ of\ x$



Collection from Expert Judges

- Karya: Crowdsourced platform
- 68 annotators from across the country
- Quality Control
- Automatic Validation Checker

Useful to capture native words, rare words and words in low-resource languages

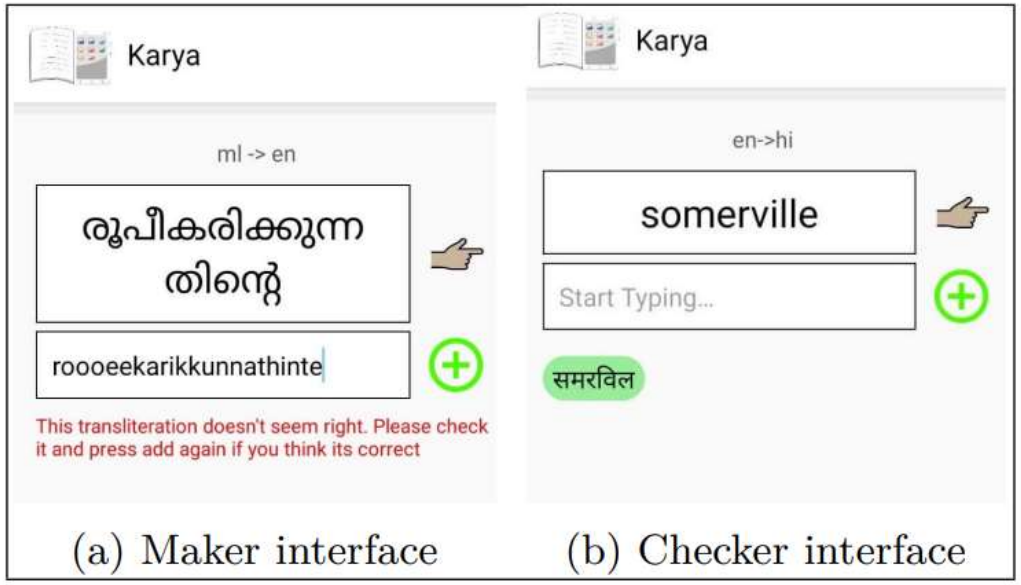


Figure 1: Annotation UI in the *Karya* app.

Aksharantar Dataset Statistics

Data Sources: Publicly available parallel translation corpora and monolingual corpora

- Training: **26 million** transliteration pairs from 21 Indic languages
- Test: 103k word pairs from 19 Indic languages covering native words and named entities

Dataset	asm	ben	guj	hin	kan	kok	mai	mal	mar	pan	san	tam
IndicCorp	0.91	0.93	0.91	0.97	0.98	0.99	0.91	0.94	0.97	0.95	0.78	0.80
Samanantar	0.93	0.92	0.84	0.76	0.80	-	-	0.80	0.90	0.86	0.84	0.80
Average	0.92	0.93	0.88	0.86	0.88	0.99	0.91	0.87	0.94	0.90	0.81	0.80

Accuracy of mined data as per human judgment

Lang	Tot
asm	217
ben	1,337
brx	44
guj	1,236
hin	1,522
kan	3,010
kas	64
kok	702
mai	370
mal	4,195
mni	16
mar	1,594
nep	2,458
ori	398
pan	611
san	1,881
snd	82
sin	37
tam	3,301
tel	2,521
urd	748
Tot	26,345

Per-language Training statistics (in thousands)

IndicXlit

https://ai4bharat.iitm.ac.in/indic-trans

- Trained on Aksharantar parallel transliteration corpus
- Multilingual Model (en→IL, IL→en)
- Significantly improves performance over existing datasets like Dakshina

Examples of improvement with multilingual training

lang	src_word	src_word_itrans	tgt_ref_word	bilingual	multilingual
hi	ब्राउज़र	brauzara	browser	brouser	browser
hi	क्लैश	kliisha	clash	klash	clash
hi	अरेबिया	arebiyaa	arabia	arebiya	arabia
ml	ബ്രിഗേഡ്	briged	brigade	bregade	brigade
ml	ഫൗണ്ടേഷൻ	fouNteShan	foundation	fountation	foundation
ml	പ്ലേഹൗസ്	plehaus	playhouse	plehouse	playhouse
ta	ஸுப்பர்சானிக்	supparchaanik	supersonic	suppersanic	supersonic
ta	எக்ஸ்பிளோரர்	.eksipLorar	explorer	exflorer	explorer

Multilingual model generates more canonical spellings

Lesser confusion in generation of characters for underspecified Tamil script

Summary

- Large scale datasets are critical to performance of NLP systems
- Need to harness publicly available datasets and make them available in the public domain
- Innovative ways to mining datasets will help drive progress for many NLP tasks
- Multilingual learning & self-supervised learning can help low-resource languages benefit from high resource languages
- We need to engage the community for the long tail of languages
- High quality testsets need to be created with human inputs
- **Food for thought**: How do we adapt to the world of large language models for generative AI?

Acknowledgements

Mentors, collaborators, students from

Microsoft, AI4Bharat, IIT Bombay, IIT Madras, NICT Japan, A* Singapore, MILA, IBM India, Tarento Technologies

Support from Microsoft, EkStep Foundation, Nilekani Philanthrophies

Thank You!

https://anoopkunchukuttan.gitlab.io/