Multilingual Learning and Mining Datasets for Building High-quality NLP Models

Anoop Kunchukuttan

Microsoft Translator, Hyderabad

Microsoft

AI4Bharat

IISER Bhopal, Apr 13 2022

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://indicnlp.ai4bharat.org/samanantar

Automatic conversion of text/speech from one natural language to another

Be the change you want to see in the world

वह परिवर्तन बनो जो संसार में देखना चाहते हो

Microsoft

Google Translate

Government: administrative requirements, education, security.

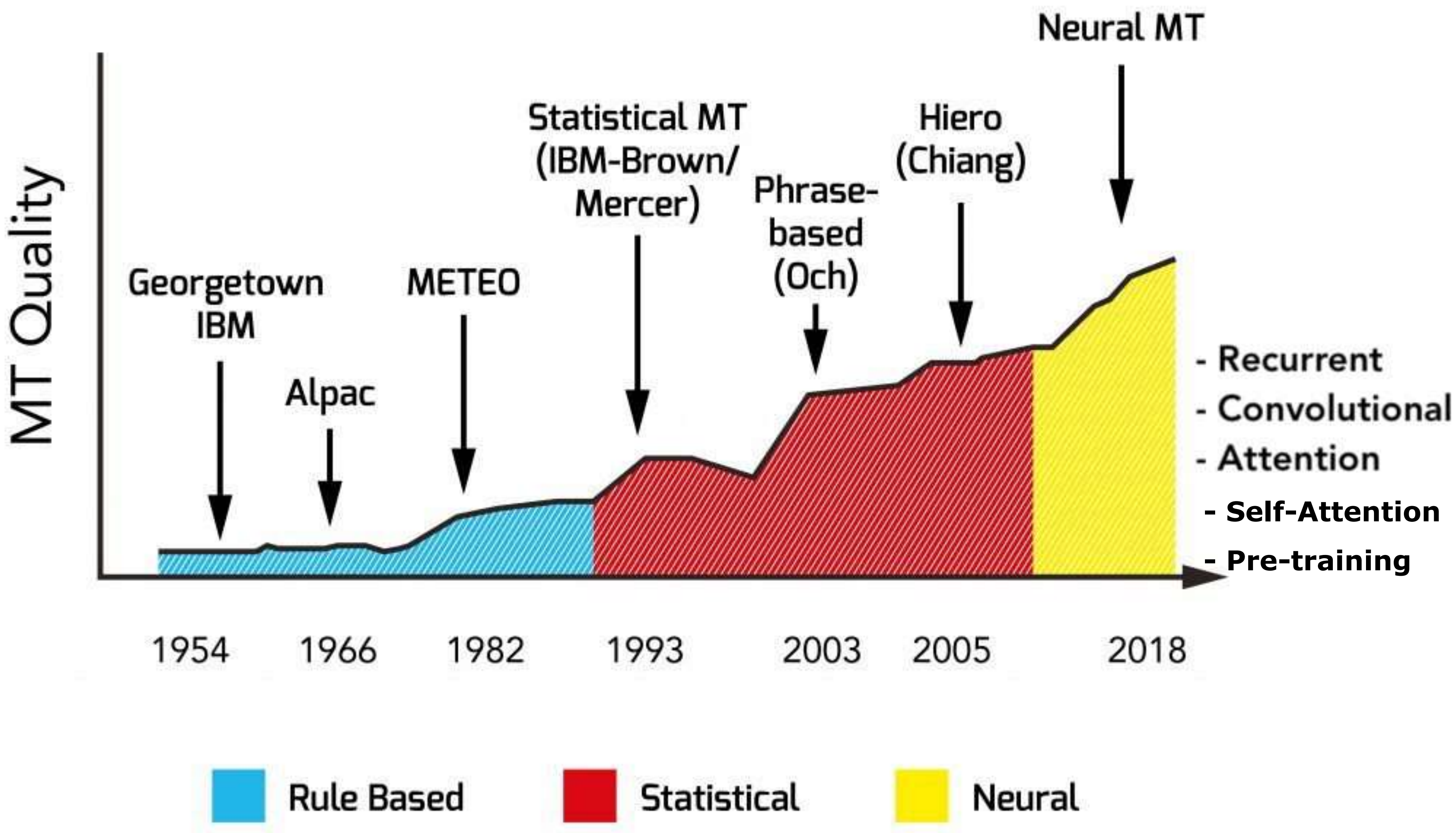
Enterprise: product manuals, customer support

Social: travel (signboards, food), entertainment (books, movies, videos)

Translation under the hood

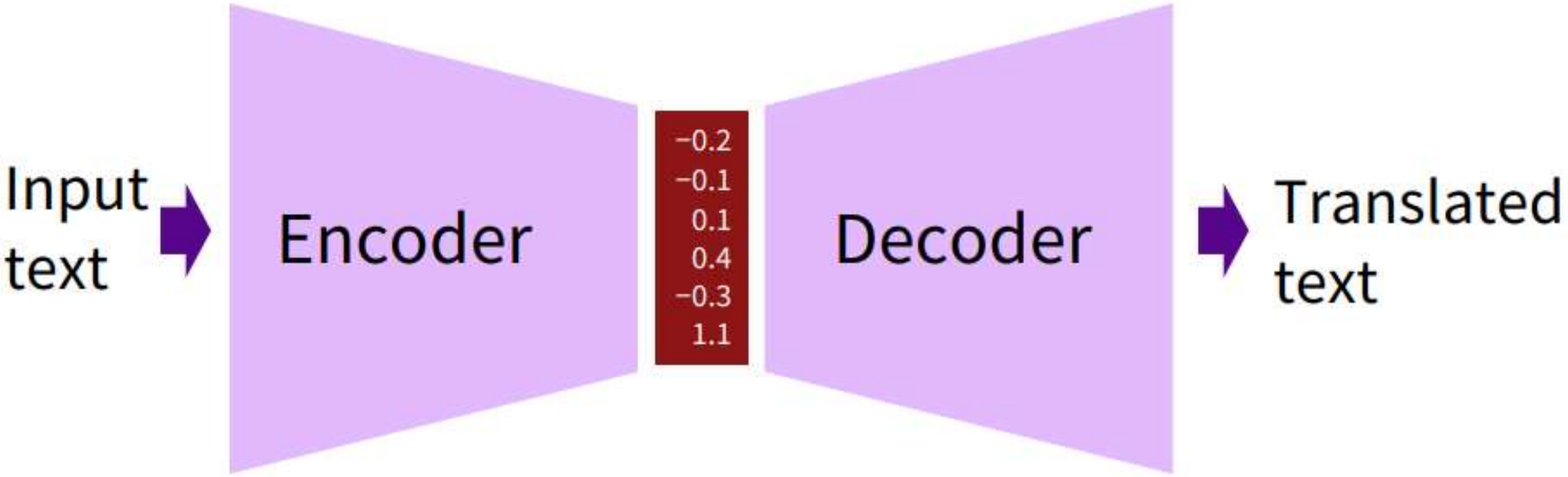
- Cross-lingual Search
- Cross-lingual Summarization
- Building multilingual dictionaries

Any multilingual NLP system will involve some kind of machine translation at some level



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

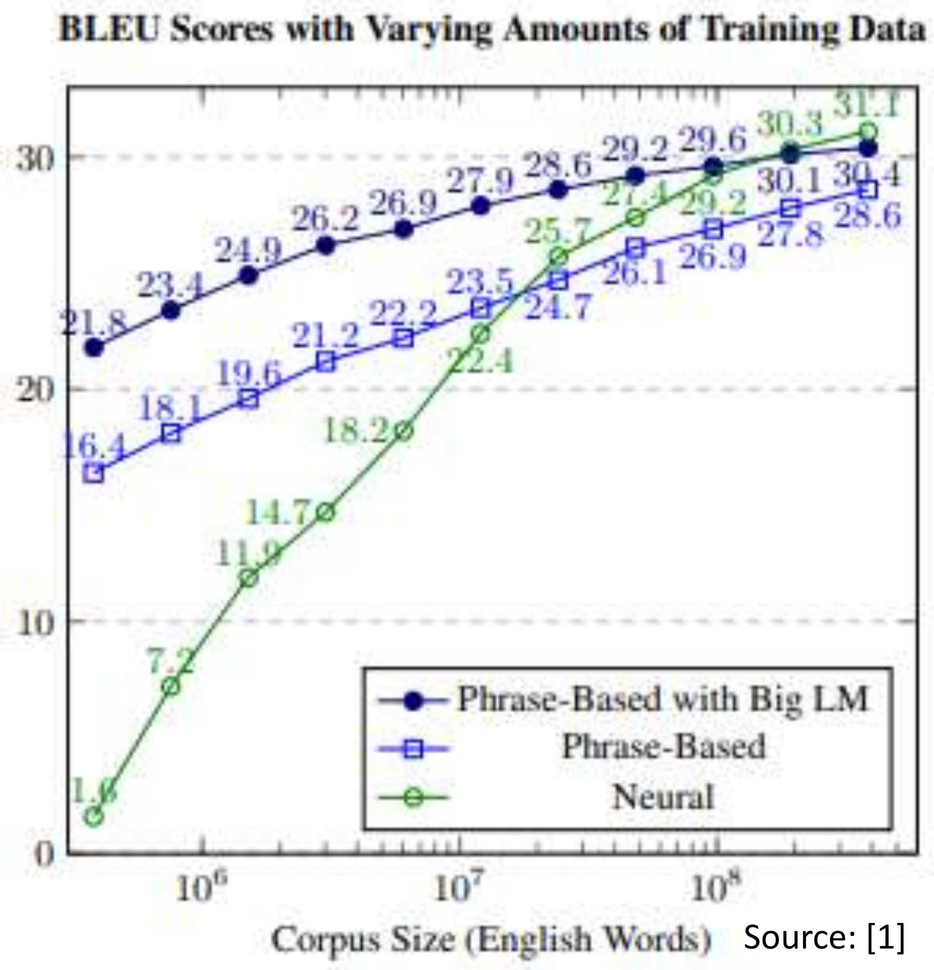
Transformer based encoder-decoder architectures are the de-facto standard for NMT today



Source: https://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf

Neural MT systems learn correspondences between words, phrases, etc. in context from paired translations

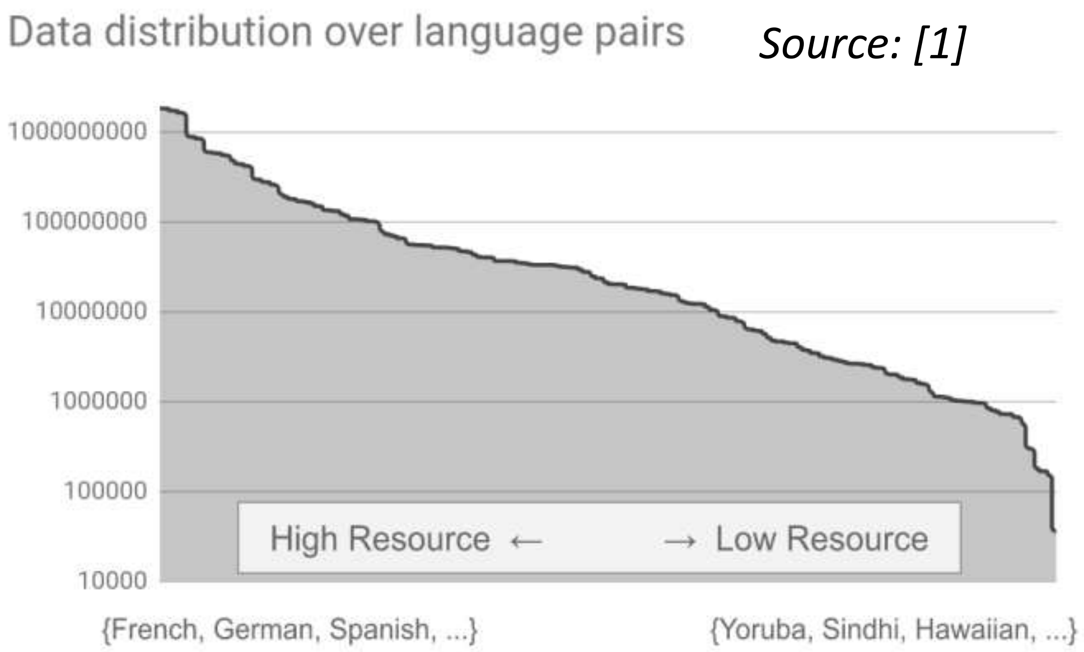
Sample Parallel Corpus	
A boy is sitting in the kitchen	एक लडका रसोई मे बैठा है
A boy is playing tennis	एक लडका टेनिस खेल रहा है
A boy is sitting on a round table	एक लडका एक गोल मेज पर बैठा है
Some men are watching tennis	कुछ आदमी टेनिस देख रहे है
A girl is holding a black book	एक लडकी ने एक काली किताब पकडी है
Two men are watching a movie	दो आदमी चलचित्र देख रहे है
A woman is reading a book	एक औरत एक किताब पढ रही है
A woman is sitting in a red car	एक औरत एक काले कार मे बैठी है



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.

Dataset Contributions

Parallel corpora for 11 Indian Languages + English

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

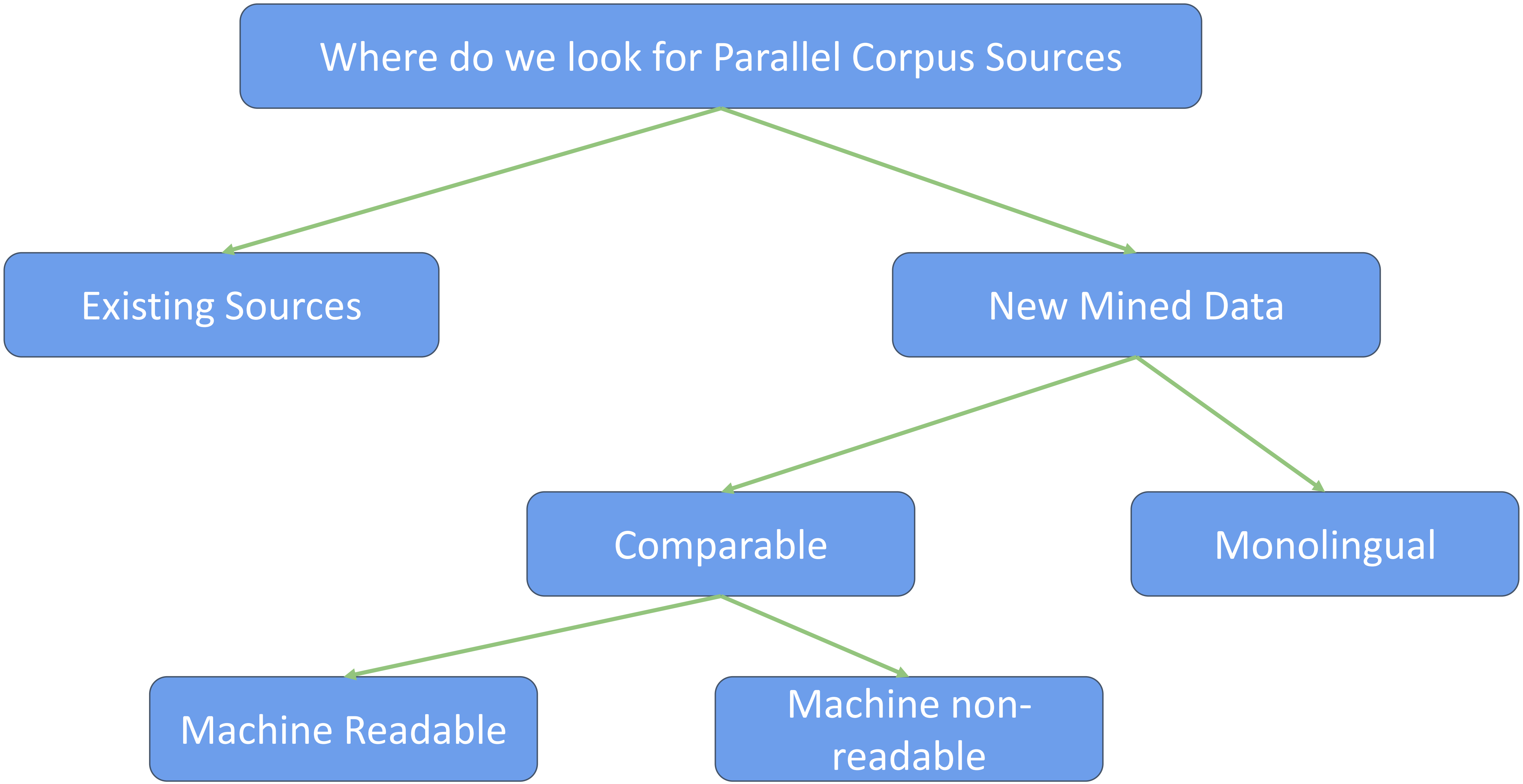
	#lang-pair	#sent-pair (million)
English-Indic languages	11	49.7
Indic-Indic languages	55	83.4

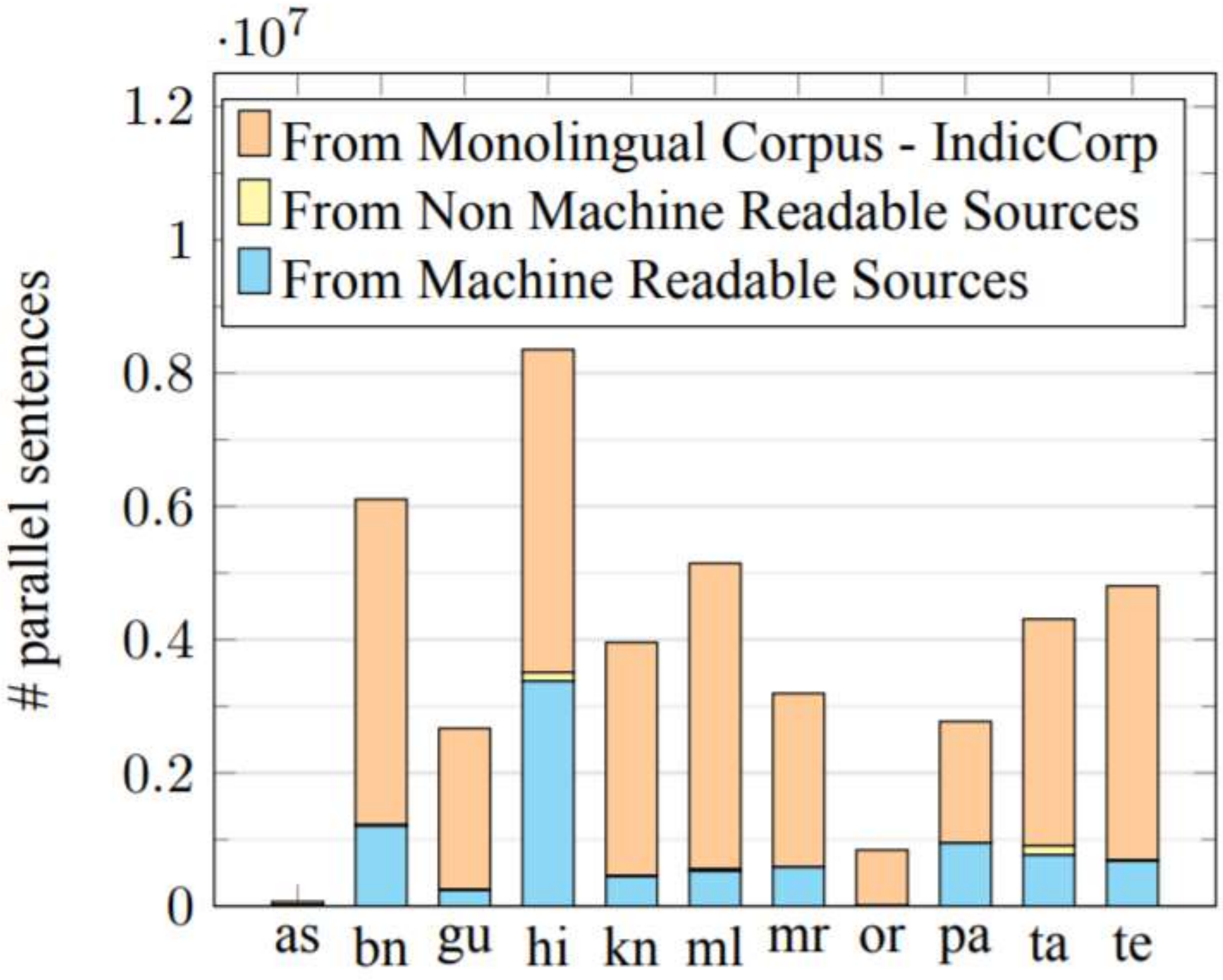
4x increase over existing corpora

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)


Where do we look for 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

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

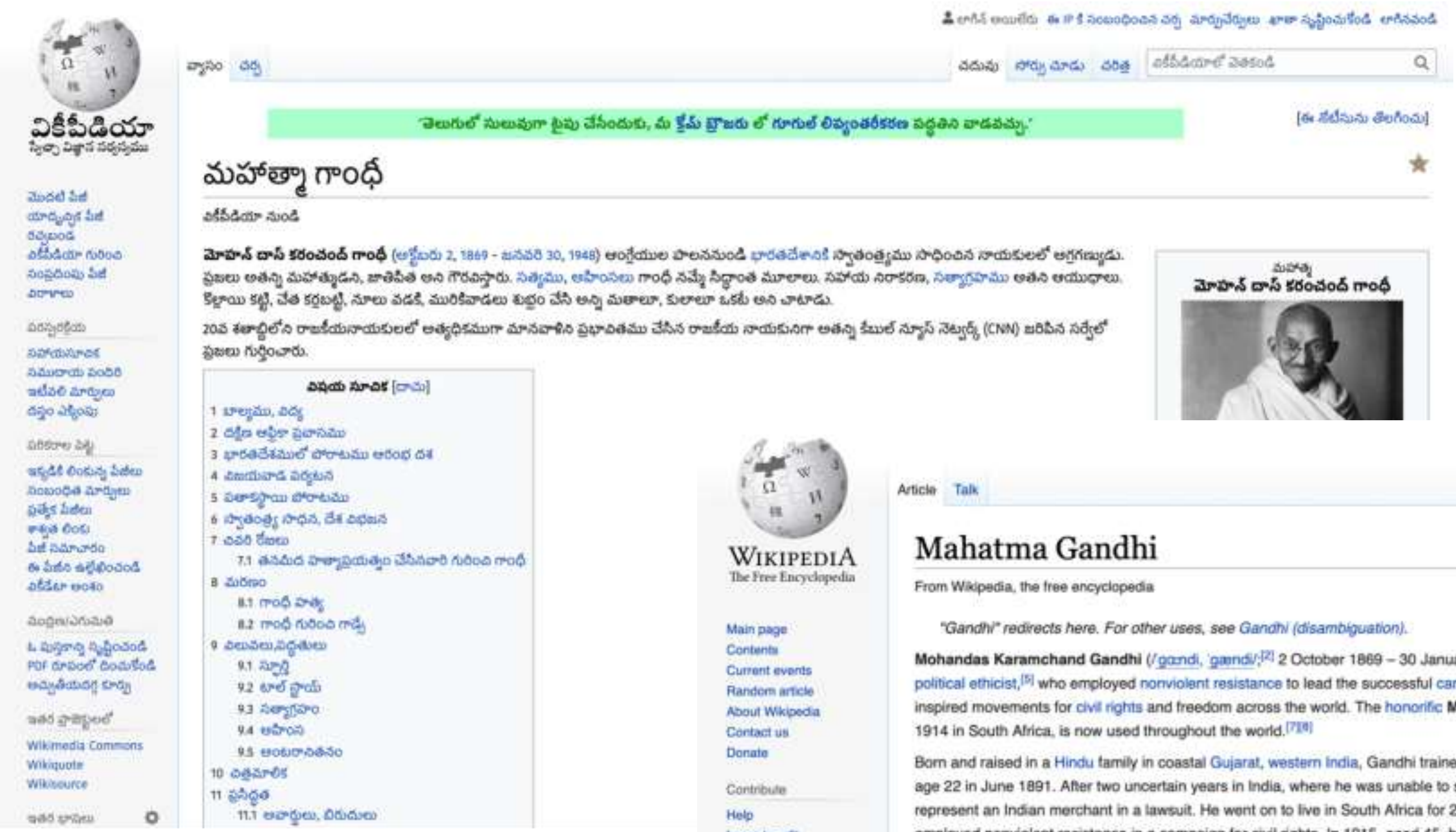
Going beyond comparable corpora

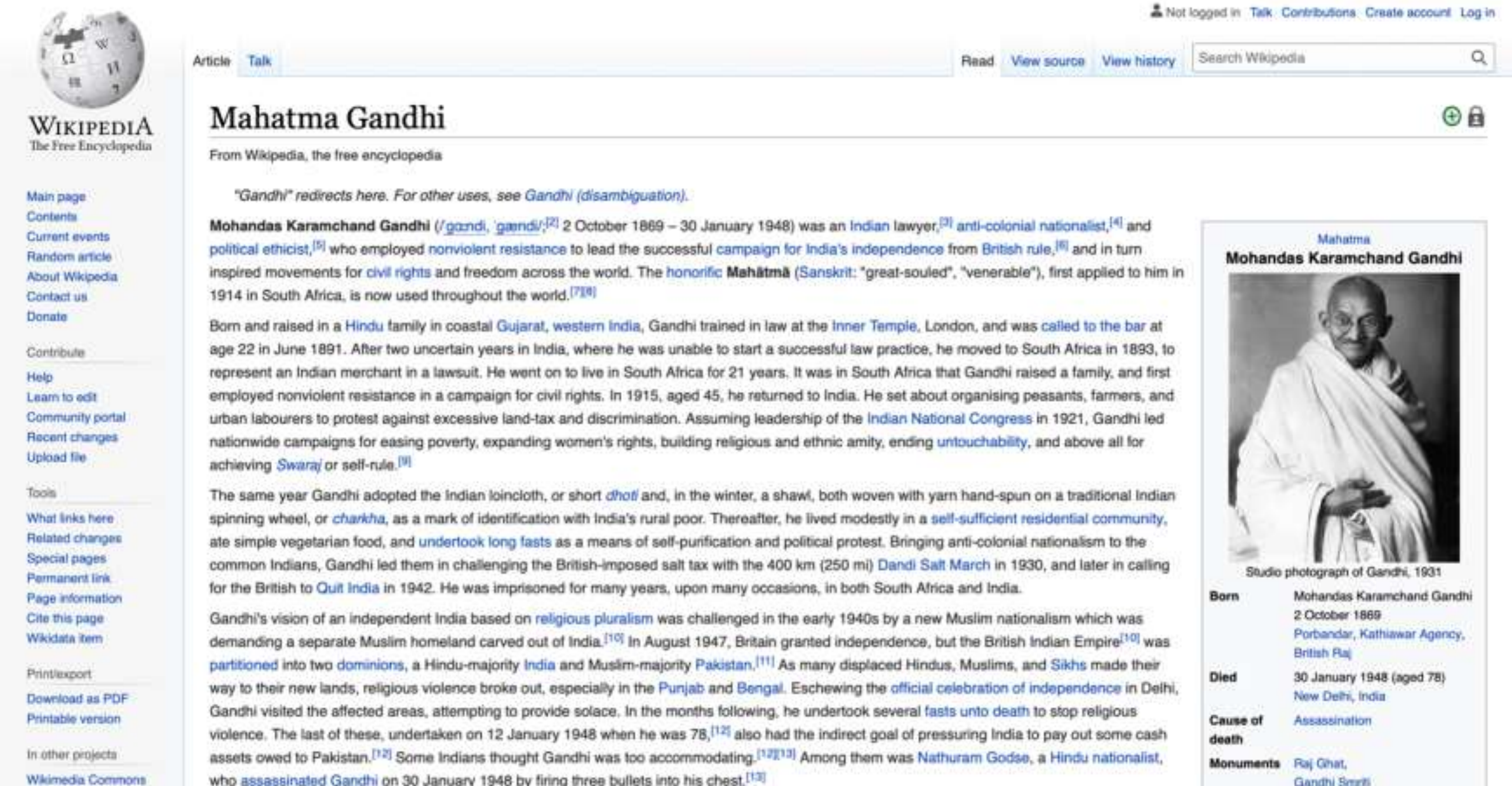
Audacious goal: can we mine parallel data from just large monolingual corpora

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

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

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


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.

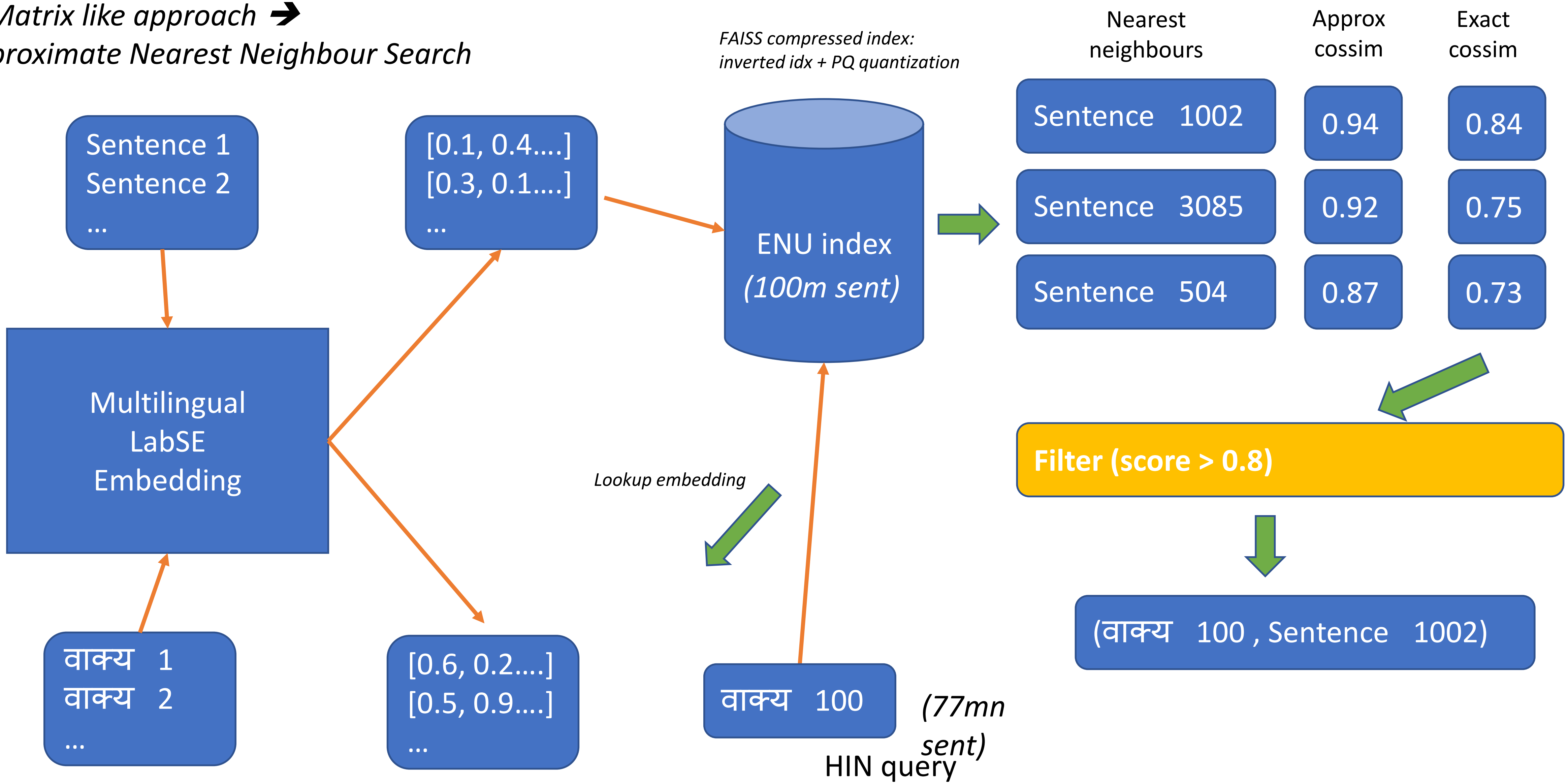
Gandhi's vision of an independent India based on religious pluralism was challenged in the early 1940s by a new Muslim nationalism which was demanding a separate Muslim homeland carved out of India.[10] In August 1947, Britain granted independence, but the British Indian Empire[10] was partitioned into 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

Parallel Corpus Mining from Monolingual Data

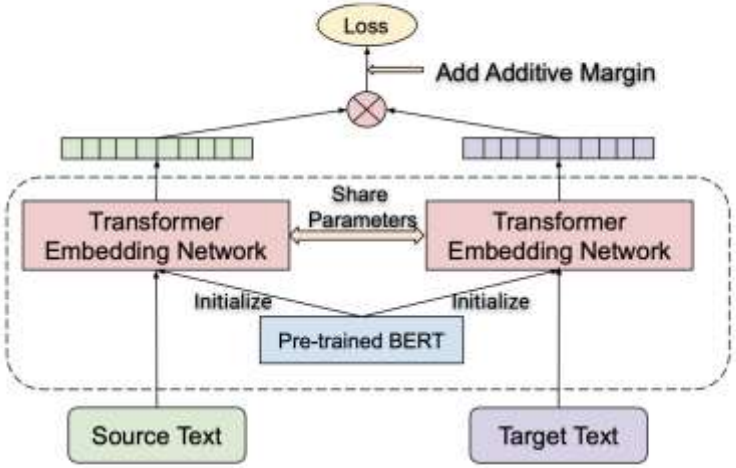
Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB. 2019. arXiv:1911.04944

CCMatrix like approach ➔
Approximate Nearest Neighbour Search



LaBSE Embedding

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



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

What helps scaling to large datasets

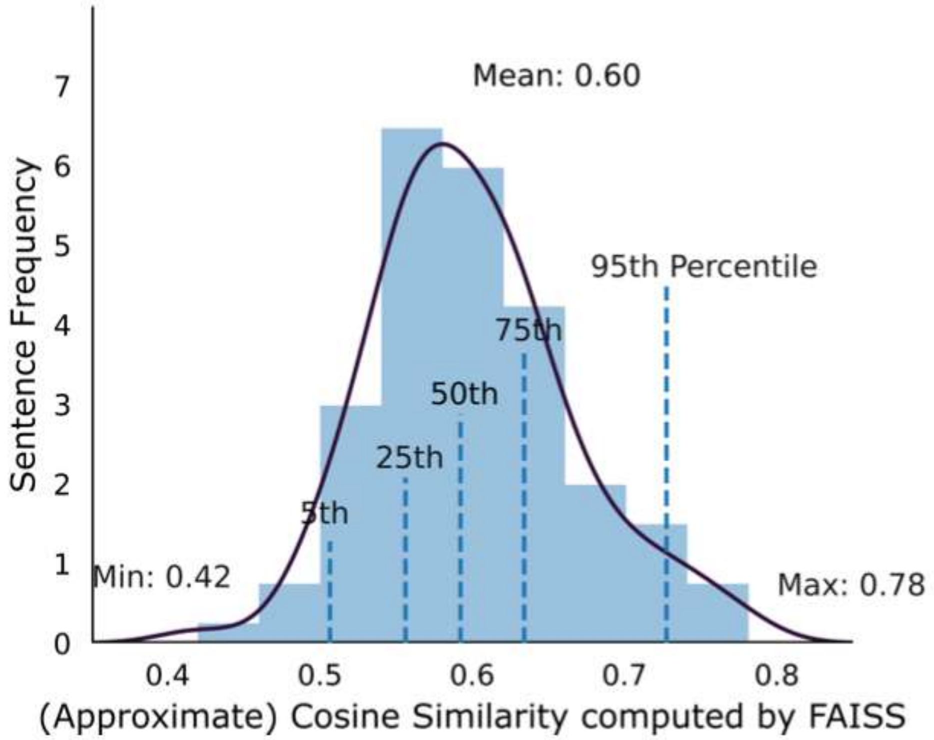
- Simple similarity metric (cosine similarity)
  - Distance from binary argument functions can’t scale (e.g. COMET score)
- Approximate nearest-neighbourhood search
- Compressed indexes to fit indices in GPU memory
  - 768d vector compressed from 3072 bytes to 72 bytes (+constant costs)
- Distributing 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/

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. We use a higher LAS of 0.8



Cross-lingual Semantic Textual Similarity dataset

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)

Decent Quality

Good Quality

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 मारुति 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.* SemEva.

Measuring the quality of the parallel corpora

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

Language	Annotation data		Sem
	# Bitext pairs	# Anno-tations	All accept
Assamese	689	1,973	3.48
Bengali	957	3,814	4.53
Gujarati	779	2,333	3.94
Hindi	1,277	4,679	4.38
Kannada	957	2,839	4.08
Malayalam	917	2,781	3.94
Marathi	779	2,324	4.14
Odia	500	1,497	3.97
Punjabi	689	2,265	4.16
Tamil	1,044	3,123	4.11
Telugu	951	2,968	4.51
Overall	9,570	30,596	4.17

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.

Other parallel corpora in Samanantar

Existing sources of parallel data

12m setence pairs

OPUS
IIIT-H PIB/Mann Ki Baat
IITB En-Hi

	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

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

myKhel CRICKET HOME | NEWS | FIXTURES | RESULTS | SERIES | TEAMS | ICC RANKINGS

MATCH: • HYD VS DEL - IN PLAY • CHE VS BAN - COMPLETED • PAK VS ZIM - COMPLETED • BAN VS SRL - COMPLETED • ZIM VS PAK - UPCOMING • + MORE

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

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]

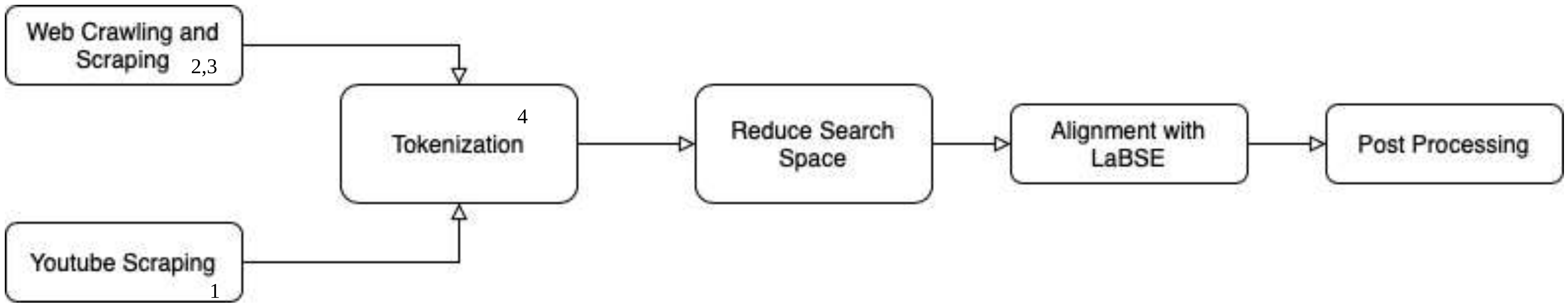
మ్యాచ్: • DEL VS HYD - IN PLAY • CHE VS BAN - పూర్తయింది • PAK VS ZIM - పూర్తయింది • BAN VS SRL - పూర్తయింది • ZIM VS PAK - రాబోయే • + మరి

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

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]

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/

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

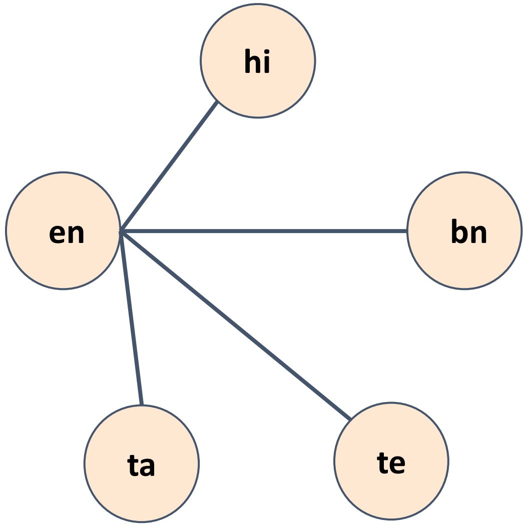
1. https://cloud.google.com/vision/docs/ocr
2. https://pypi.org/project/indic-nlp-library/

- 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.

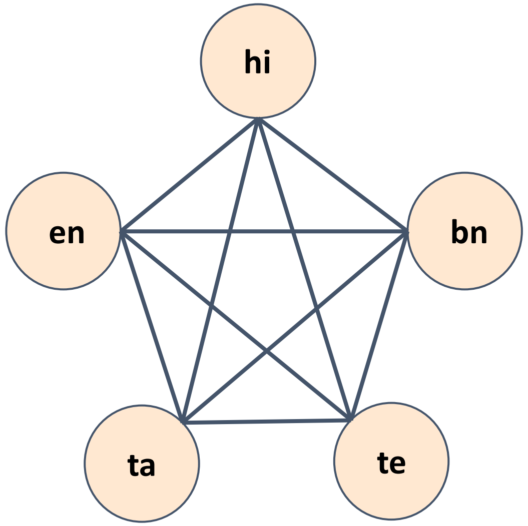
- **Tamil Nadu will be amongst India's most economically prosperous states by 2023, achieving a three-fold growth in per capita income (in real terms) over the next 7 years to be on par with the Upper Middle Income countries globally.**
- **Tamil Nadu will exhibit a highly inclusive growth pattern – 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.**

Mining between Indic Languages

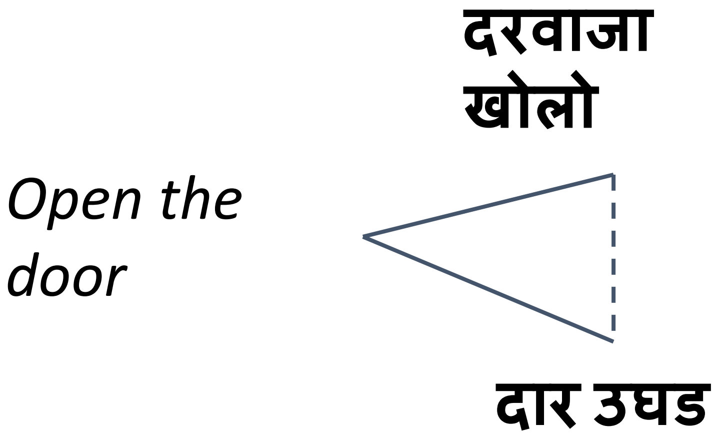
Mine Indic-Indic parallel corpora from English to Indic corpora



English-centric



Complete

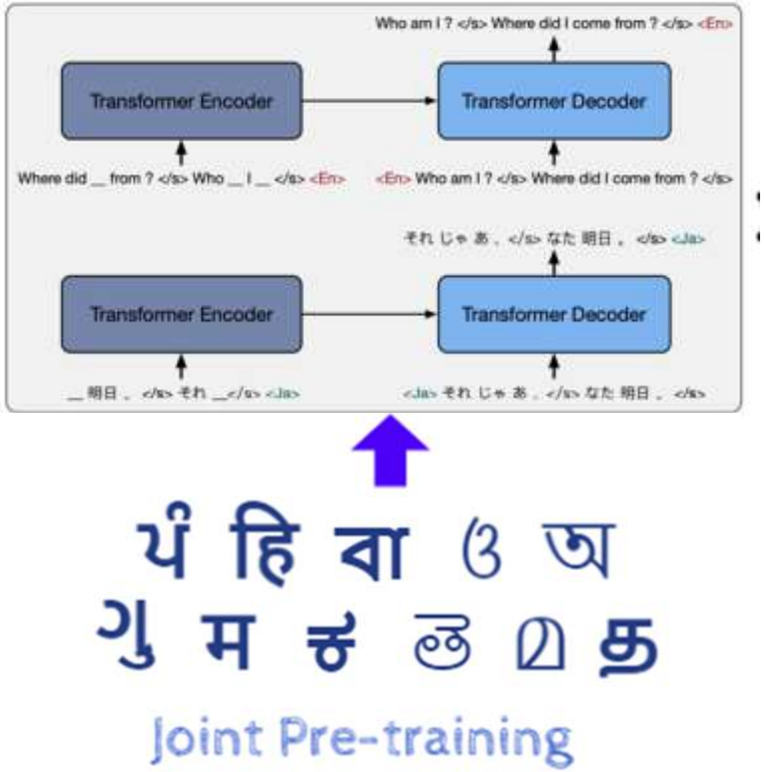


83.7 million sentence pairs for 55 language pairs

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

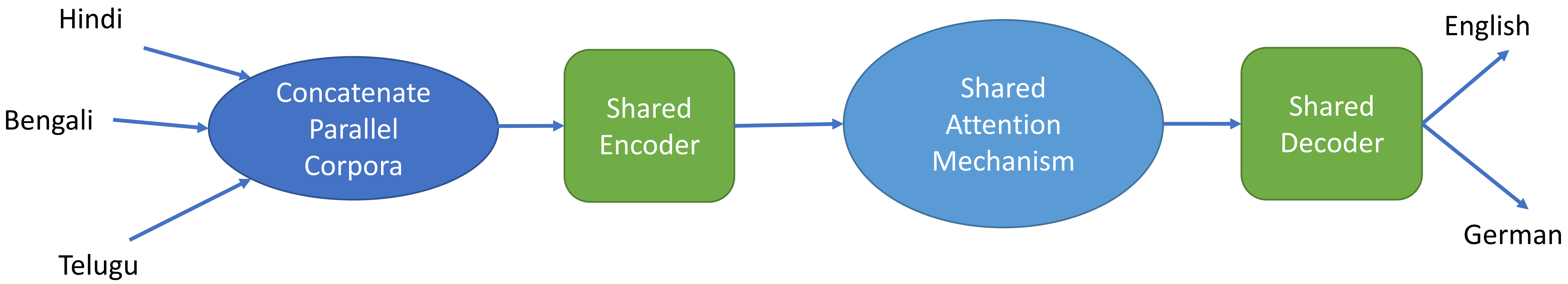


- Trained on Samanantar parallel corpus
- Multilingual Model (en→IL, IL→en, IL→IL)
- Single Script
- Input and output language tags
- Model size: (~430m params)

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.

Compact Multilingual NMT

(Johnson et al., 2017)



Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat et al. "Google's multilingual neural machine translation system: Enabling zero-shot translation." TACL (2017).

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

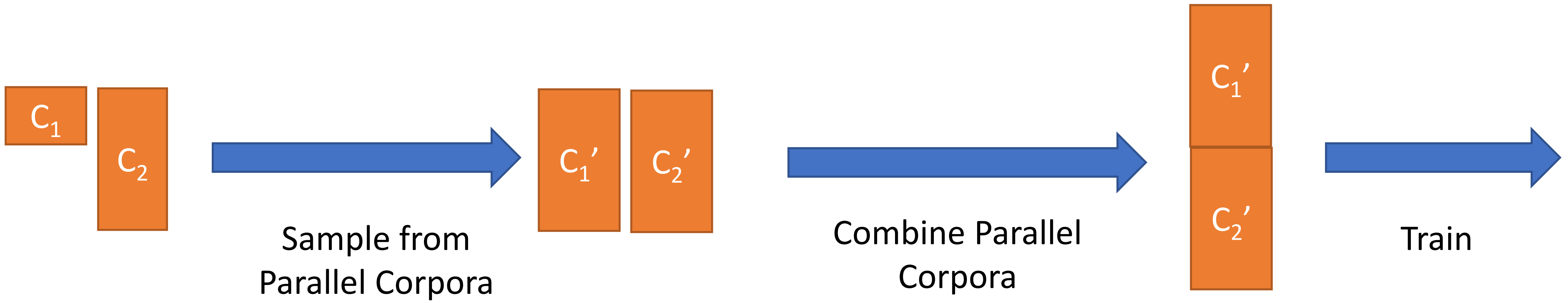
There is only one decoder, how do we generate multiple languages?

Language Tag Trick → Special token in input to indicate target language

Original Input: *मकर संक्रांति भगवान सूर्य के मकर में आने का पर्व है*

Modified Input: *मकर संक्रांति भगवान सूर्य के मकर में आने का पर्व है<eng>*

Joint Training



Key Results

Comparisons on WAT 2020, WAT2021, FLORES-101

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

	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	$IT^\dagger$	IT	GOOG	MSFT	CVIT	OPUS	mBART	$IT^\dagger$	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. $IT^\dagger$ 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

A Large-scale Evaluation of Neural Machine Transliteration for Indic Languages

Anoop Kunchukuttan

Siddharth Jain

Rahul Kejriwal

Microsoft

Microsoft India, Hyderabad

The 16th Conference of the European Chapter of the Association for Computational Linguistics. 19-23 April, 2021

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	എത്തനൂർ
(ettanUra)	(.ettanUr)

Indian languages are written in multiple scripts

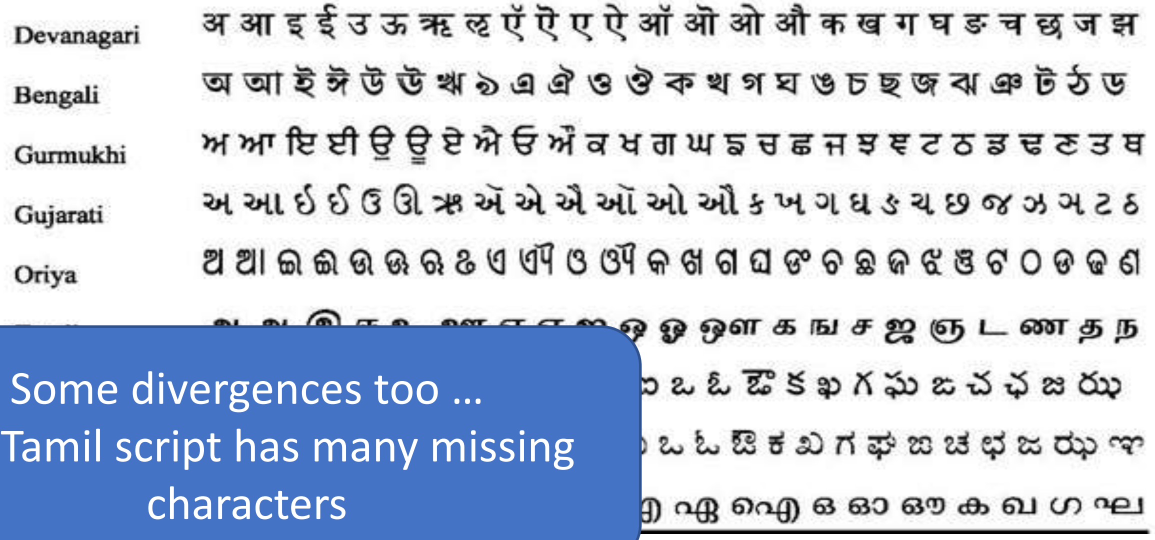
Brahmi, Persio-Arabic, Latin scripts

Brahmi scripts

- Abugida scripts
- Used to write major Indian languages
- 2 major language families
  - Indo-Aryan (IA): 6 languages
  - Dravidian (DR): 4 languages

English ←→ Indic transliteration

Shared characteristics



Some divergences too …
e.g. Tamil script has many missing characters

- Largely overlapping character set
- Highly overlapping phoneme sets
- Consistent grapheme-to-phoneme mapping

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., 2018b)

Mine Large-scale Transliteration Corpora

- *From parallel translation corpora*
- *From monolingual corpora*

Comprehensive analysis of multilingual transliteration models

- *Effect of language family*
- *Effect of script sharing*
- *Performance on Indian vs foreign names*

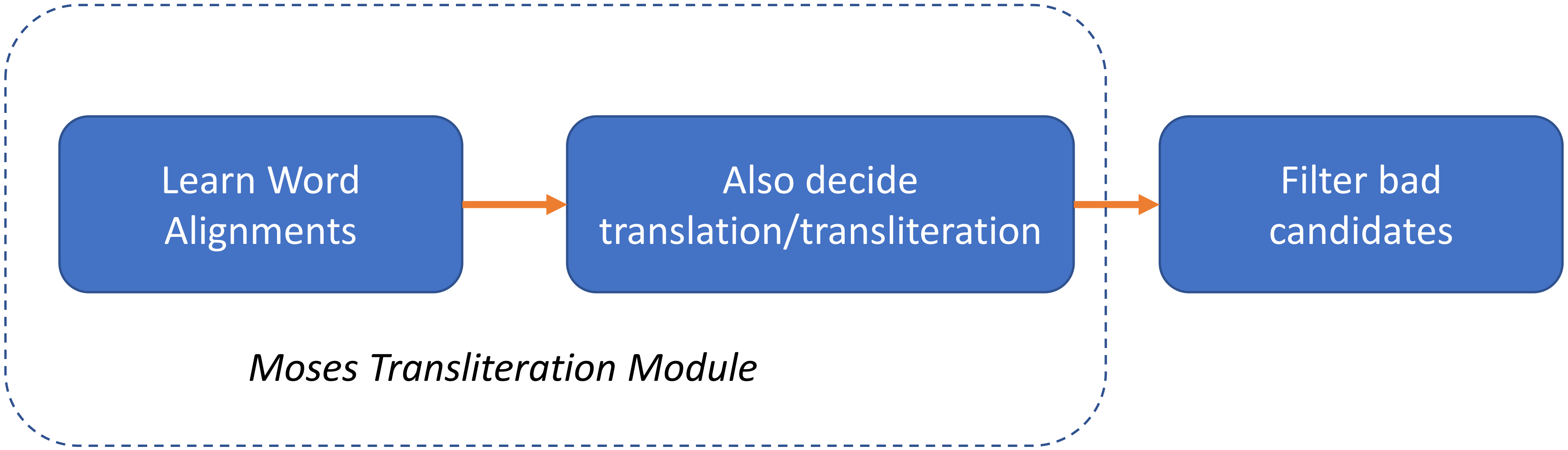
From Parallel Translation Corpora

(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

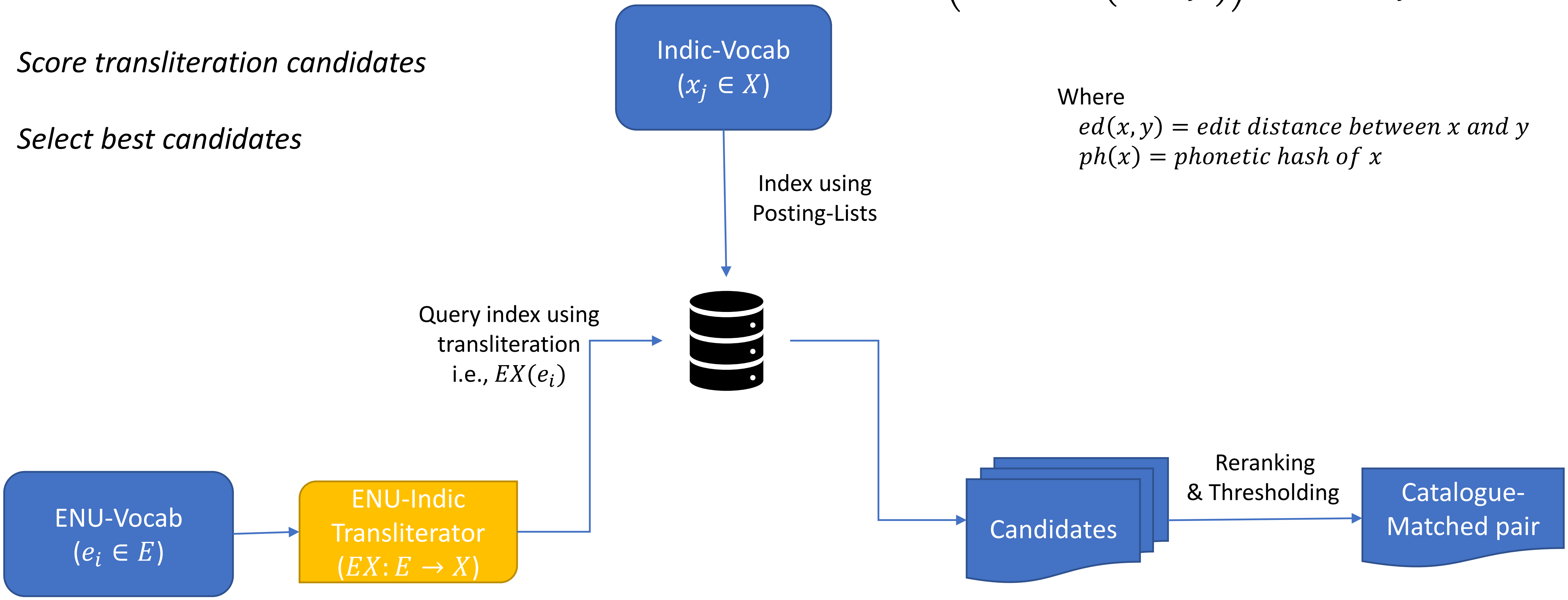
Reranking:

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

Where

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

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



Mined Dataset Statistics

__Data Sources__: Publicly available parallel translation corpora and monolingual corpora

Language	pa	hi	bn	or	gu	mr	kn	te	ml	ta
Word pair count ($\times$1000)	55.3	157.7	65.4	34.7	65.5	38.0	24.7	77.4	31.1	57.1
Mining Accuracy	81.2	NA	76.7	NA	93.0	89.0	87.1	86.2	82.3	77.9

Total Number of word pairs: 600k (373k from parallel and 339k from monolingual corpora)

Test Sets Composition

Set	Size
Foreign	928
Indian	572
Total	**1500**

Manually validated via crowdsourcing
Covers Indic and Foreign origin words

Multilingual Transliteration

Accuracy@1 reported on all slides

Model	IA	DR	IND
FOREIGN WORDS			
bilingual	49.54	50.18	49.8
multilingual	**60.26**	**56.7**	**58.83**
INDIAN WORDS			
bilingual	72.74	**67.45**	70.62
multilingual	**75.69**	66.86	**72.16**

X to E Transliteration

Model	IA	DR	IND
FOREIGN WORDS			
bilingual	73.96	70.00	72.37
multilingual	**78.35**	**74.43**	**76.78**
INDIAN WORDS			
bilingual	77.01	75.65	76.47
multilingual	**83.80**	**79.81**	**82.20**

E to X Transliteration

Transliteration into English:
Significant 20% improvement in accuracy for foreign words

Transliteration from English:
~6% improvement in foreign and Indian word transliteration accuracy

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	ஸூப்பர்சானிக்	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

Language Family Specific Training

Model	IA	DR	IND
FOREIGN WORDS			
all indic	60.26	**56.7**	58.83
by family	**61.97**	55.46	**59.37**
INDIAN WORDS			
all Indic	75.69	66.86	72.16
by family	**76.21**	**68.28**	**73.04**

X to E Transliteration

Model	IA	DR	IND
FOREIGN WORDS			
all Indic	**78.35**	74.43	76.78
by family	77.79	**75.38**	**76.83**
INDIAN WORDS			
all Indic	**83.80**	79.81	**82.20**
by family	83.10	**80.62**	82.11

E to X Transliteration

- *No major difference in training joint models*
- *Separate training benefits for transliteration into English for the case of Indian words*

Using the same script does not cause major degradation

Model	IA	DR	IND
FOREIGN WORDS			
different scripts	62.02	56.70	59.89
same script	61.97	55.46	59.37
+src tag	**62.16**	**57.24**	**60.19**
INDIAN WORDS			
different scripts	76.81	69.95	74.06
same script	76.21	68.28	73.04
+src tag	**76.86**	**70.19**	**74.19**

X to E Transliteration

Model	IA	DR	IND
FOREIGN WORDS			
different scripts	77.00	**76.54**	76.82
same script	**77.79**	75.38	**76.83**
INDIAN WORDS			
different scripts	81.41	79.46	80.63
same script	**83.10**	**80.62**	**82.11**

E to X Transliteration

Adding source language tags help, especially for languages with divergent spelling conventions

For Tamil, training with its character set alone helps improve accuracy

Model	ta-en		en-ta	
	foreign	indic	foreign	indic
hiscript	44.04	50.6	73.5	81.72
tascript	**47.37**	**53.3**	**78.8**	**83.9**

Summary

- Mined 600k transliteration pairs for 10 languages
  - From parallel translation and monolingual corpora
  - Covers Indian and foreign origin words
  - Manually validated testsets
- Recommendation for Indic transliteration
  - Multilingual model
  - Represent data in a single script
  - Separate models for Indo-Aryan and Dravidian languages
  - Adding source language tag
  - Tamil → represent data in Tamil script

AI4Bharat

An IIT Madras Initiative

https://indicnlp.ai4bharat.org

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.

Indian Language NLP

- Text
- Speech
- OCR
- Sign Language

Multimodal NLP

https://ai4bharat.org

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, mentors and volunteers

Mission Statement

Bring parity with English

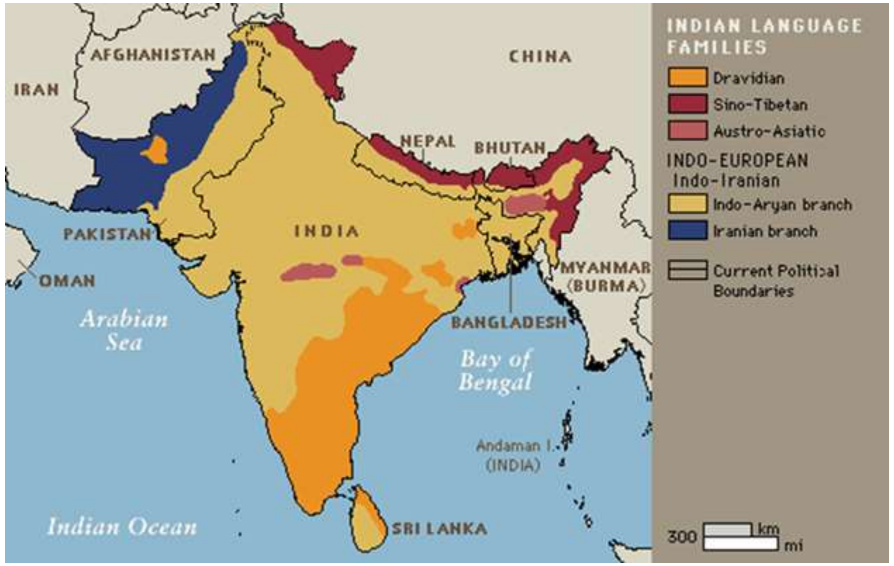
in AI tech for Indian languages

with open data and open source contributions

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

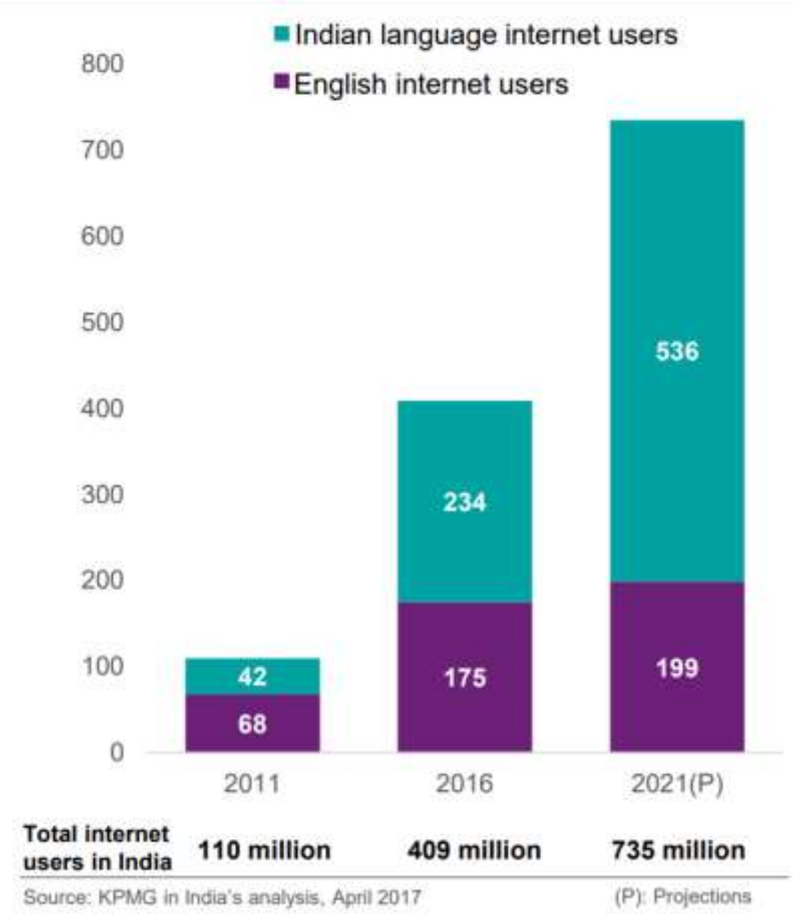
Where and Why do we need Indic NLP solutions today?

Usage and Diversity of Indian Languages



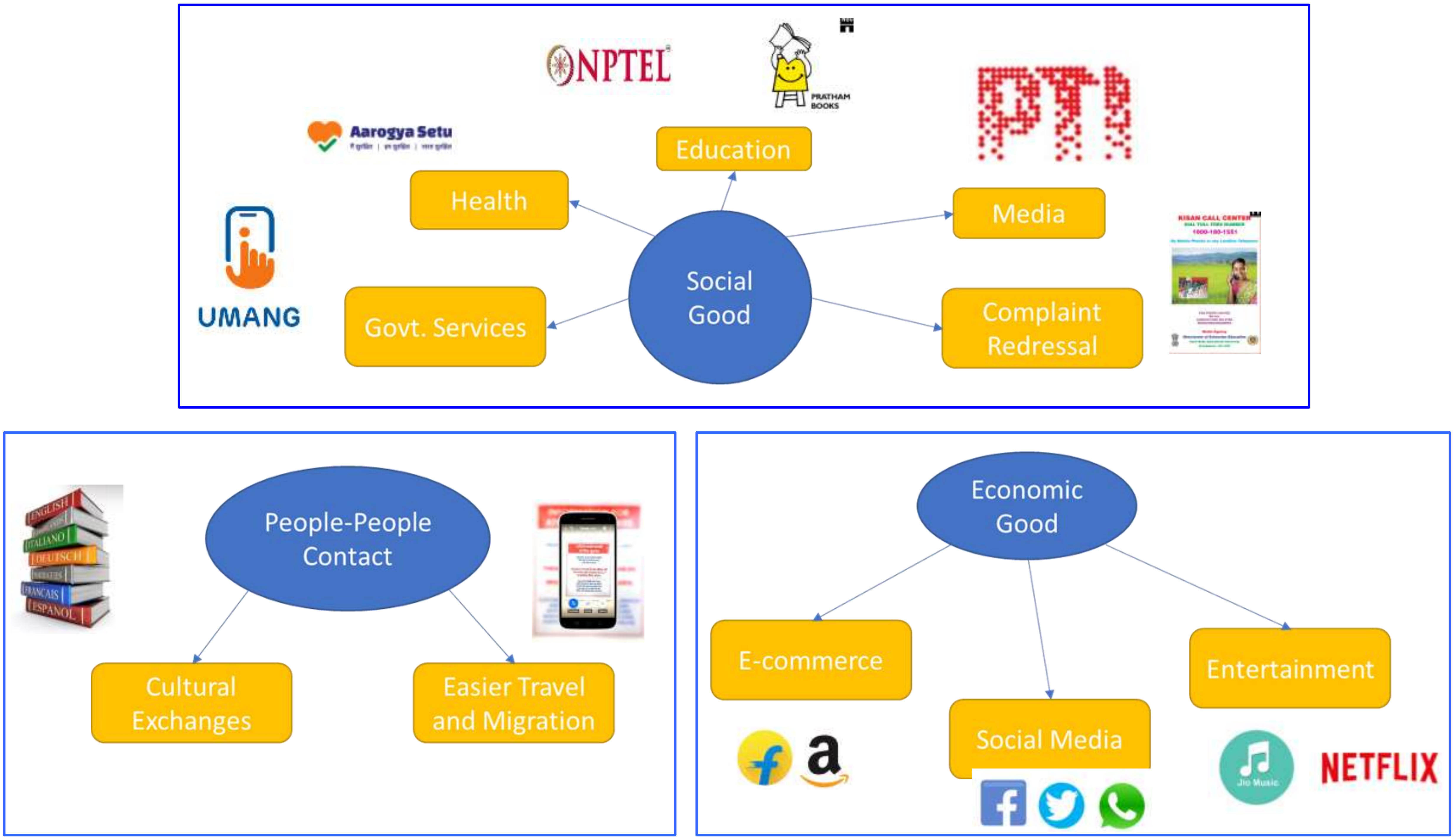
- *4 major language families*
- *22 scheduled languages*
- *125 million English speakers*
- *8 languages in the world's top 20 languages*
- *30 languages with more than 1 million speakers*

Sources: Wikipedia, Census of India 2011



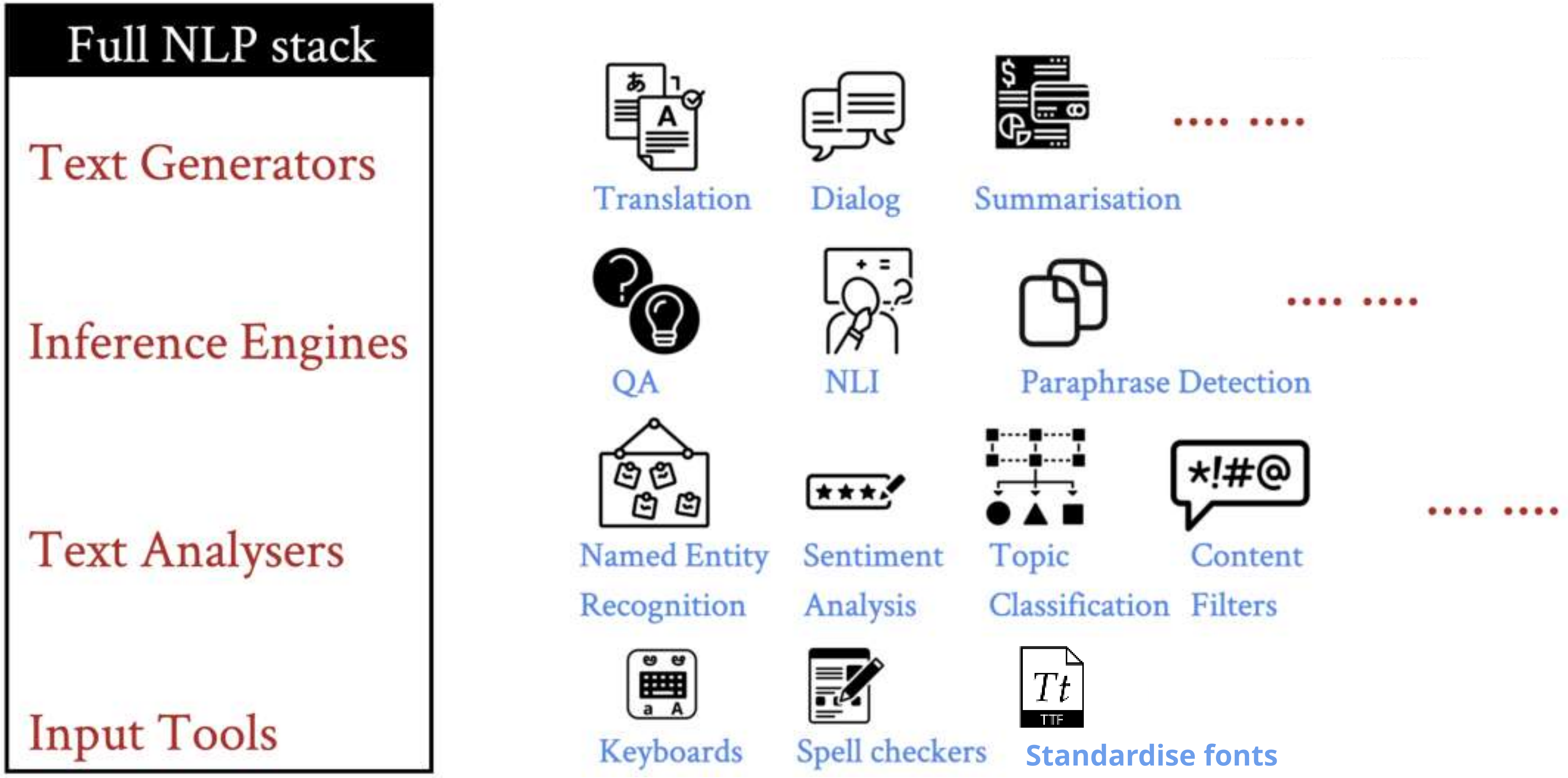
Internet User Base in India (in million)

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
People-People Contact
Cultural Exchanges
Easier Travel and Migration
Economic Good
E-commerce
Social Media
Entertainment
NETFLIX

Goal

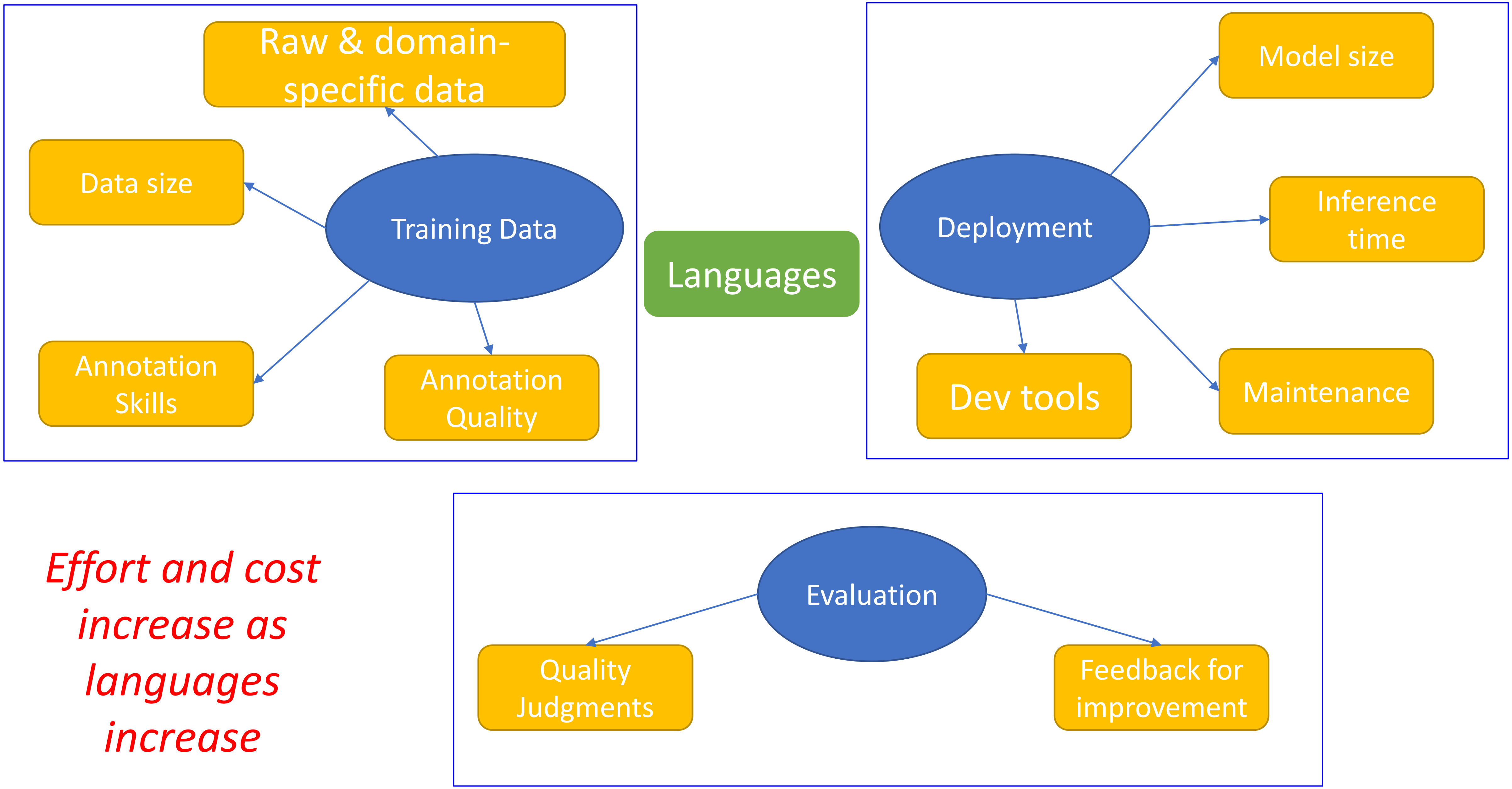
for 22 languages



We are faced with a huge data skew

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

Scalability Challenges for NLP solutions



Effort and cost increase as languages increase

What have we done so far?

What have we done so far?

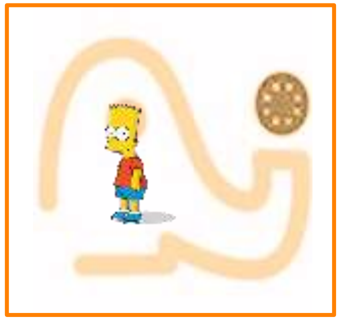
Basic Infrastructure: Raw corpora & core language models for 10+ Indian languages

IndicCorp

Large Monolingual corpora

IndicBERT

(masked LM)



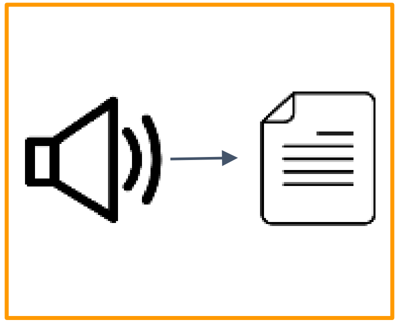
IndicBART

(seq2seq LM)

IndicFT

(word embeddings)

Compact pre-trained models for NLU & NLG



IndicWav2Vec

Raw-speech corpora

Pre-trained speech representations

What have we done so far?

Standard Evaluation Benchmarks

IGLUE

IndicGLUE *Benchmarks for Natural Language Understanding*

Datasets for tasks like article classification, COPA, WNLI, etc

INLG

Indic NLG Suite *Benchmarks for Natural Language Generation*

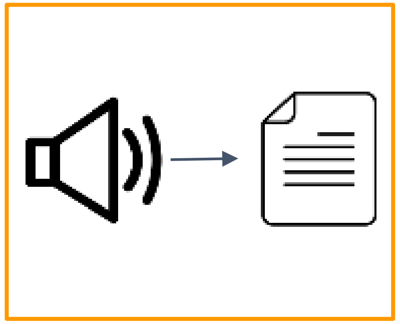
Datasets for tasks like headline generation, paraphrase generation, question generation, sentence summarization

What have we done so far?

Data and models for various foundational tasks

Samanantar

Parallel corpus, translation models between English & 11 Indic languages



IndicASR

ASR models for 9 Indian languages

Input Tools

Romanized keyboards for Indic languages

INCLUDE

Datasets and efficient models for isolated Indian Sign Language 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.

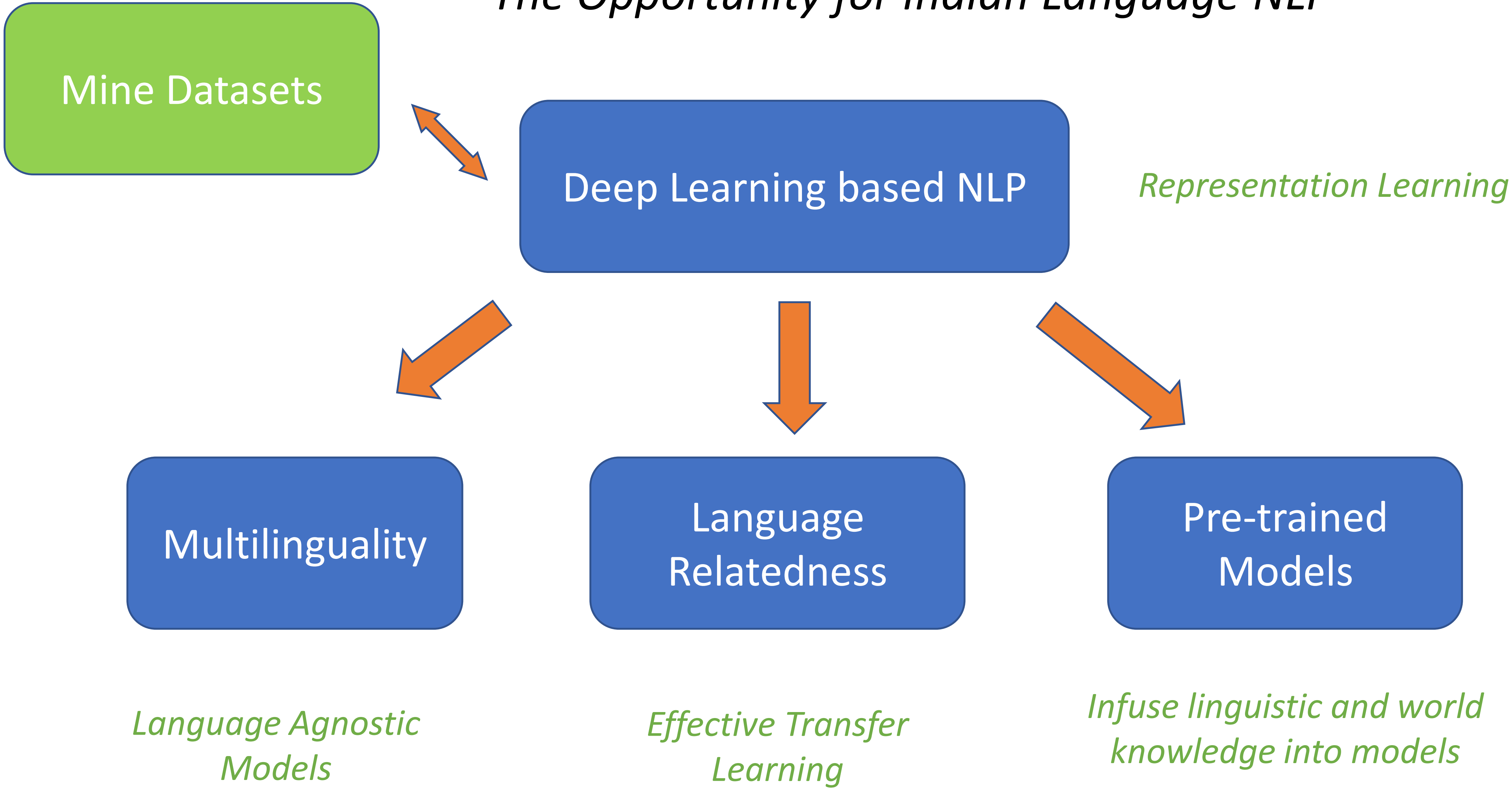
Parallel Translation Corpus

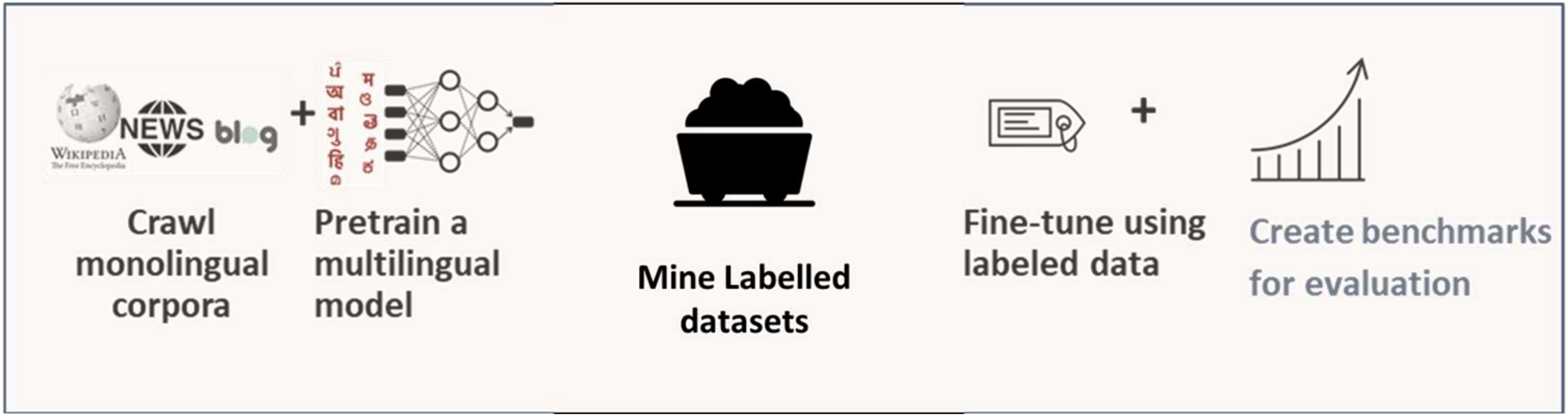
- IIT Bombay English-Hindi Parallel Corpus: Largest en-hi parallel corpora in public domain (about 1.5 million semgents)
- CVIT-IIITH PIB Multilingual Corpus: Mined from Press Information Bureau for many Indian languages. Contains both English-IL and IL-IL corpora (IL=Indian language).
- CVIT-IIITH Mann 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

What is our approach?

Our Technical Direction

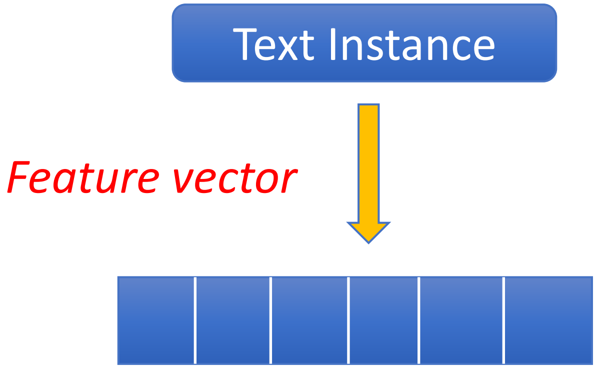
The Opportunity for Indian Language NLP





WIKIPEDIA
NEWS
blog
+
Crawl monolingual corpora
Pretrain a multilingual model
Mine Labelled datasets
+
Fine-tune using labeled data
Create benchmarks for evaluation

Distributed Representations



Replace traditional
high-dimensional, resource-heavy
document feature vector
with

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

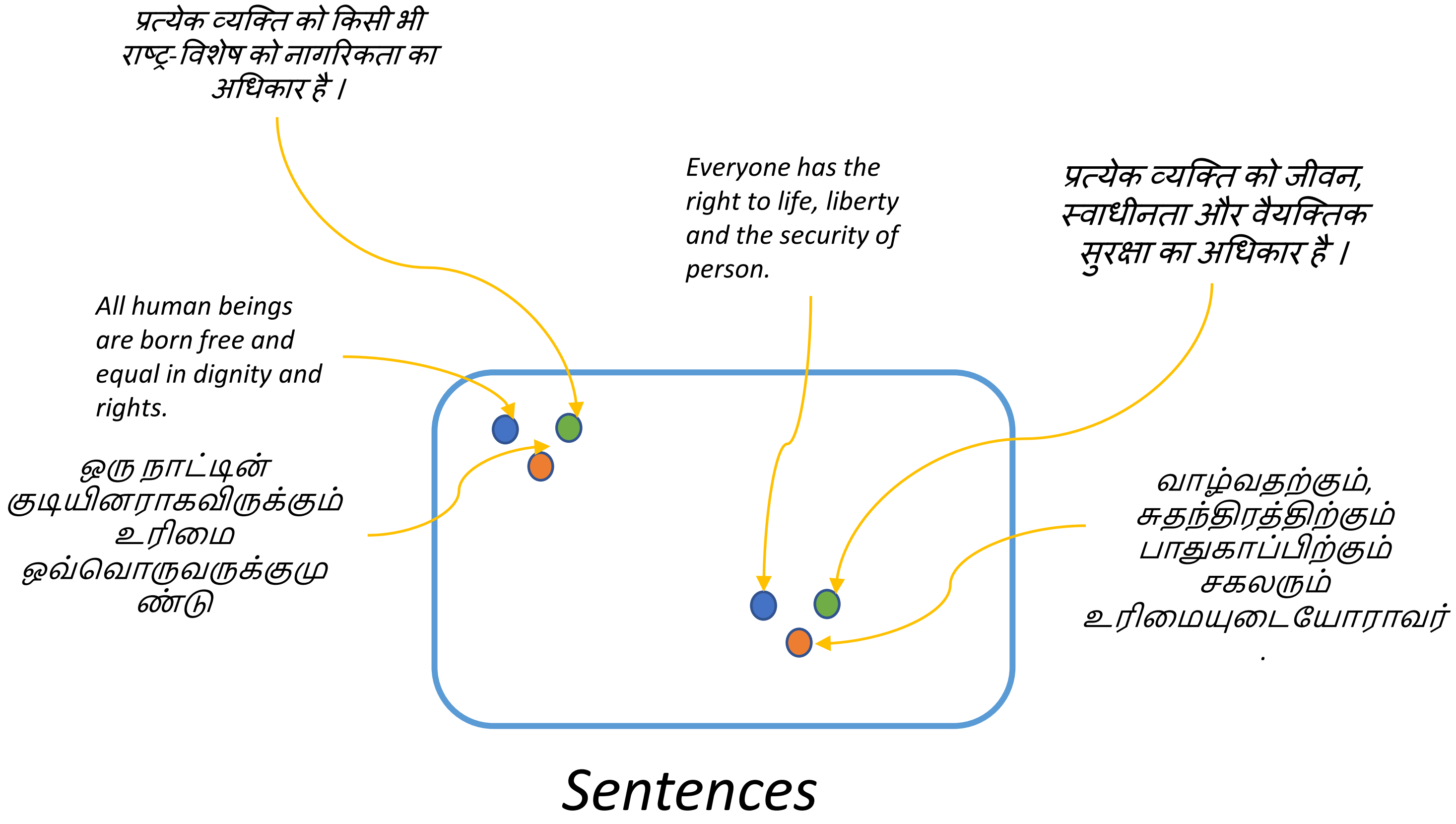
Distributional Hypothesis

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

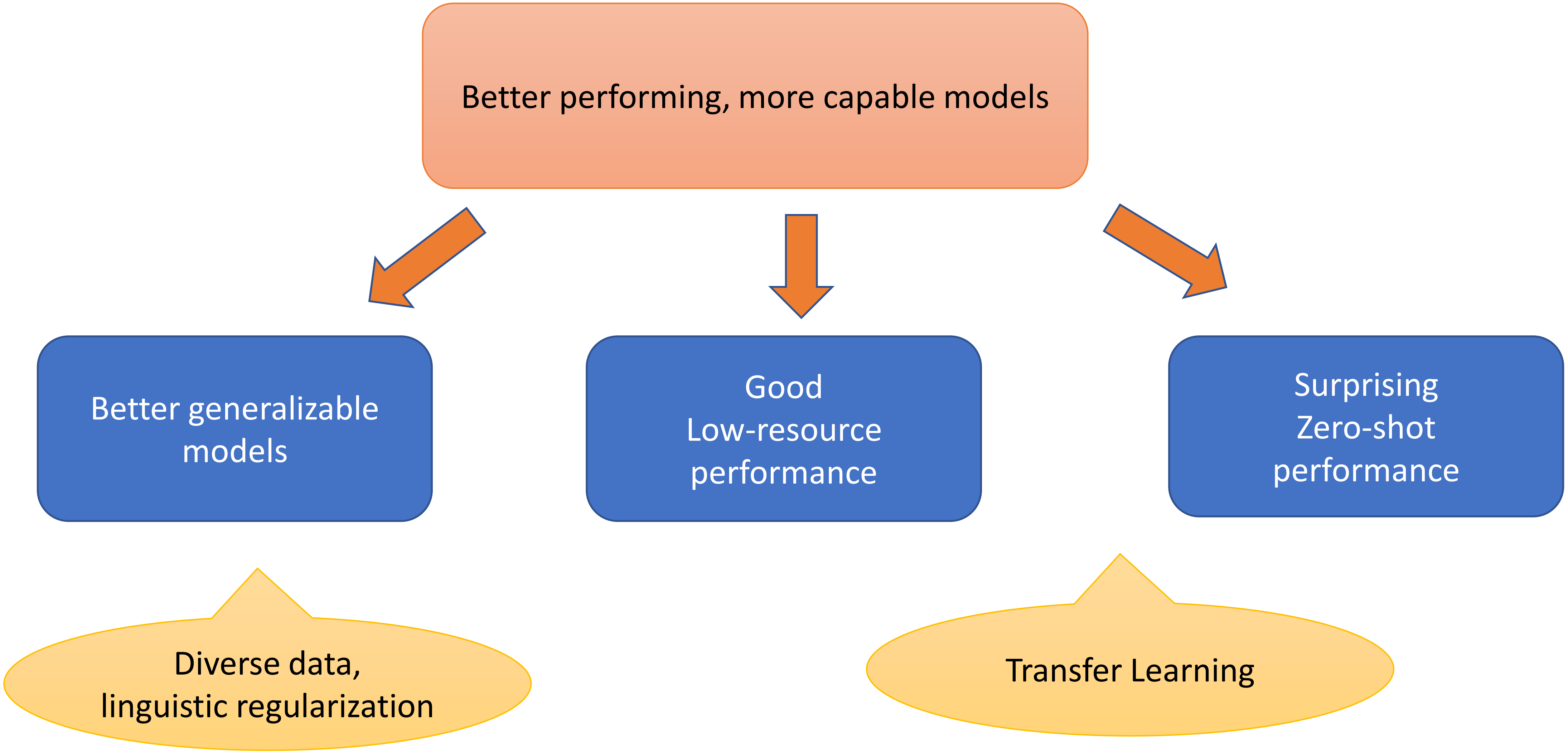
"Words that occur in similar contexts tend to have similar meanings"
- Turney and Pantel (2010)

What do multilingual models do?

Represent semantically similar language artifacts in the same vector space



How does multilinguality help?



Why are Indian languages related?

Related Languages

Related by Genealogy

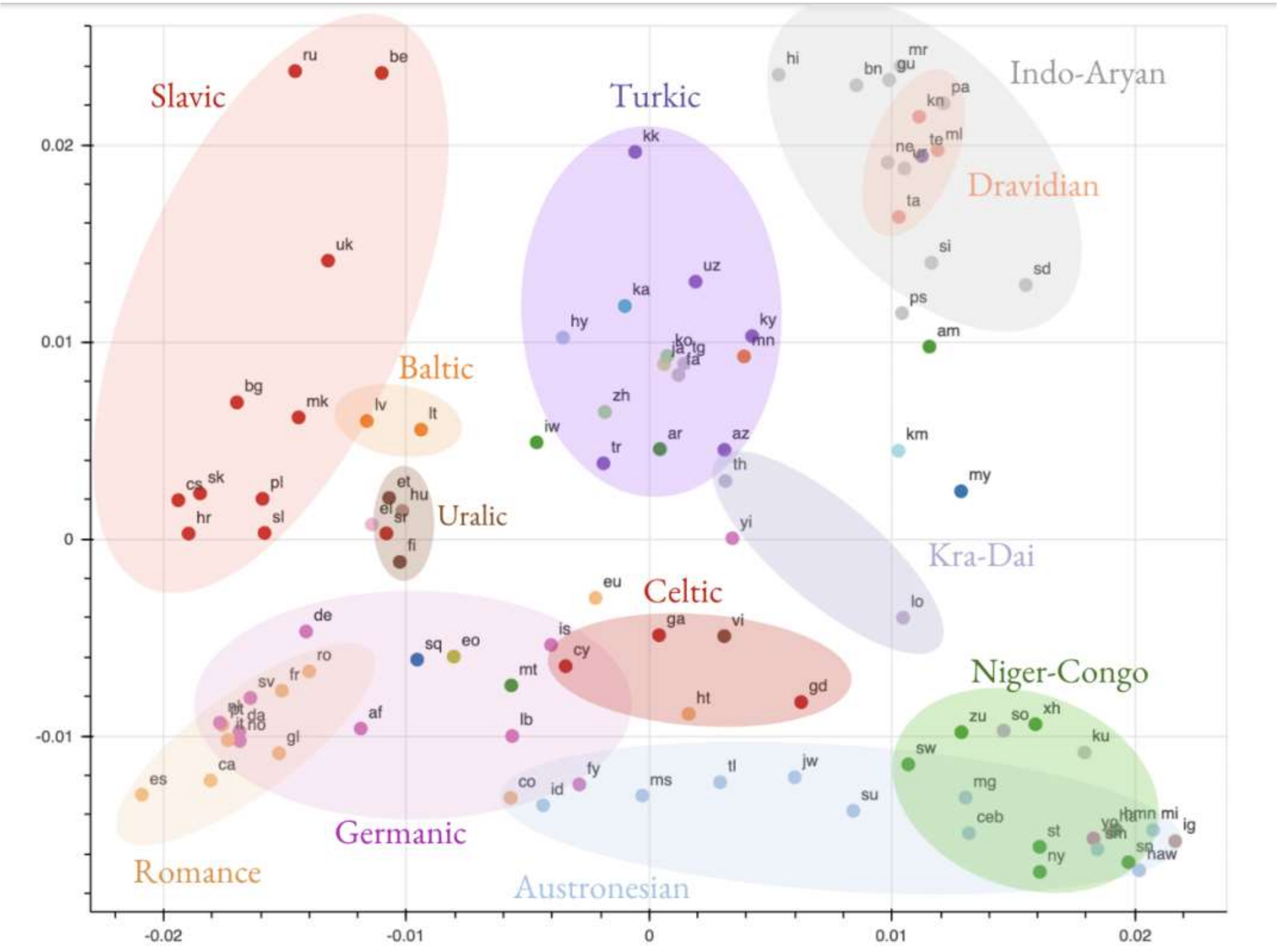
Related by Contact

<u>Language Families</u>
Dravidian, Indo-European

<u>Linguistic Areas</u>
Indian Subcontinent

Lexical, Syntactic & Orthographic similarities

How does language relatedness help?



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

Building multilingual systems systems specific to language families

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

How do pre-trained models help?

Supervised data not sufficient

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

How do we know if the sentence is grammatically correct?

How do we know if the sentence makes sense?

These capabilities are important for generalization

FAIRSEQ

BART

Task-independent models that know about language

Pre-train once, reuse for multiple downstream tasks

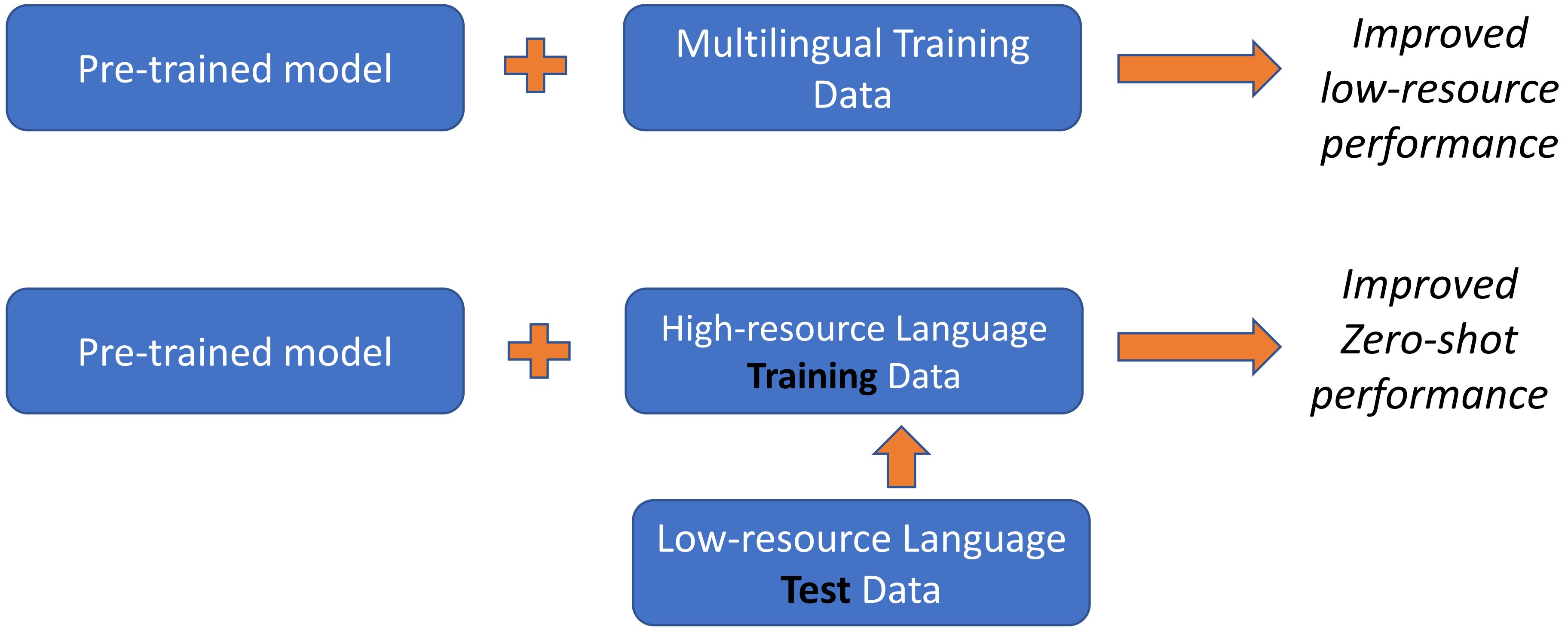
Task-independent Pre-training → *Pre-trained model* → Task-independent finetuning → *Task-specific model*

Only task-specific training: less data & less computation

Multi-linguality and Pre-training are complementary

Language-family specific pre-trained model

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



Some Projects

Models and Resources 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.* Arxiv pre-print 2203.05437. 2022.

Natural Language Understanding

Text Classification

Sentiment Analysis

Relation Extraction

Named Entity Recognition

Paraphrase Detection

Natural Language Inference

Natural Language Generation

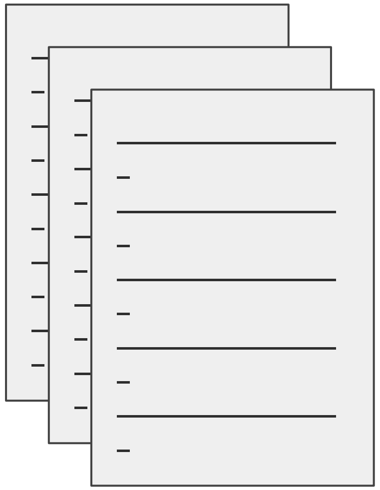
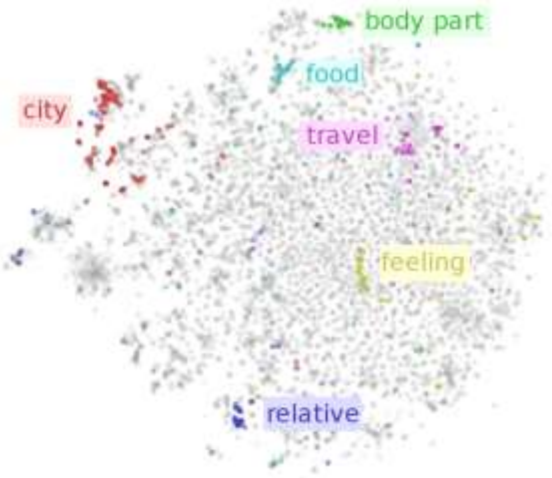
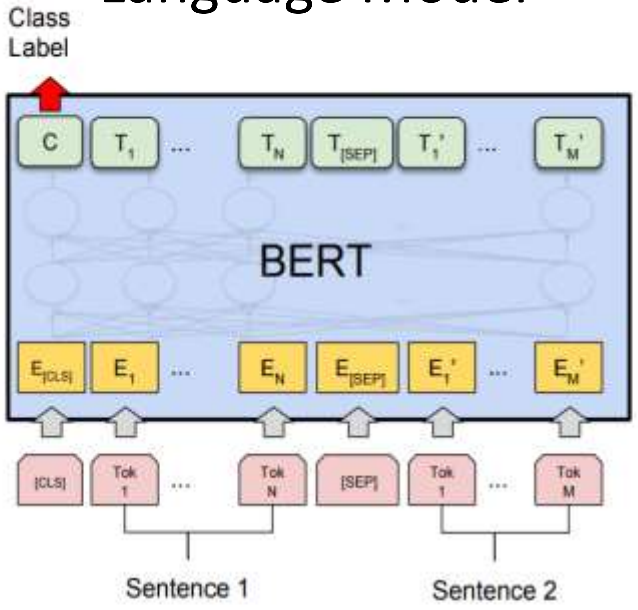
Abstractive Summarization

Paraphrase Generation

Machine Translation

Grammar Correction

IndicNLPSuite

Monolingual Corpora	Embeddings	Language Model	NLU Benchmark
			GLUE
IndicCorp	**IndicFT**	**IndicBERT**	**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

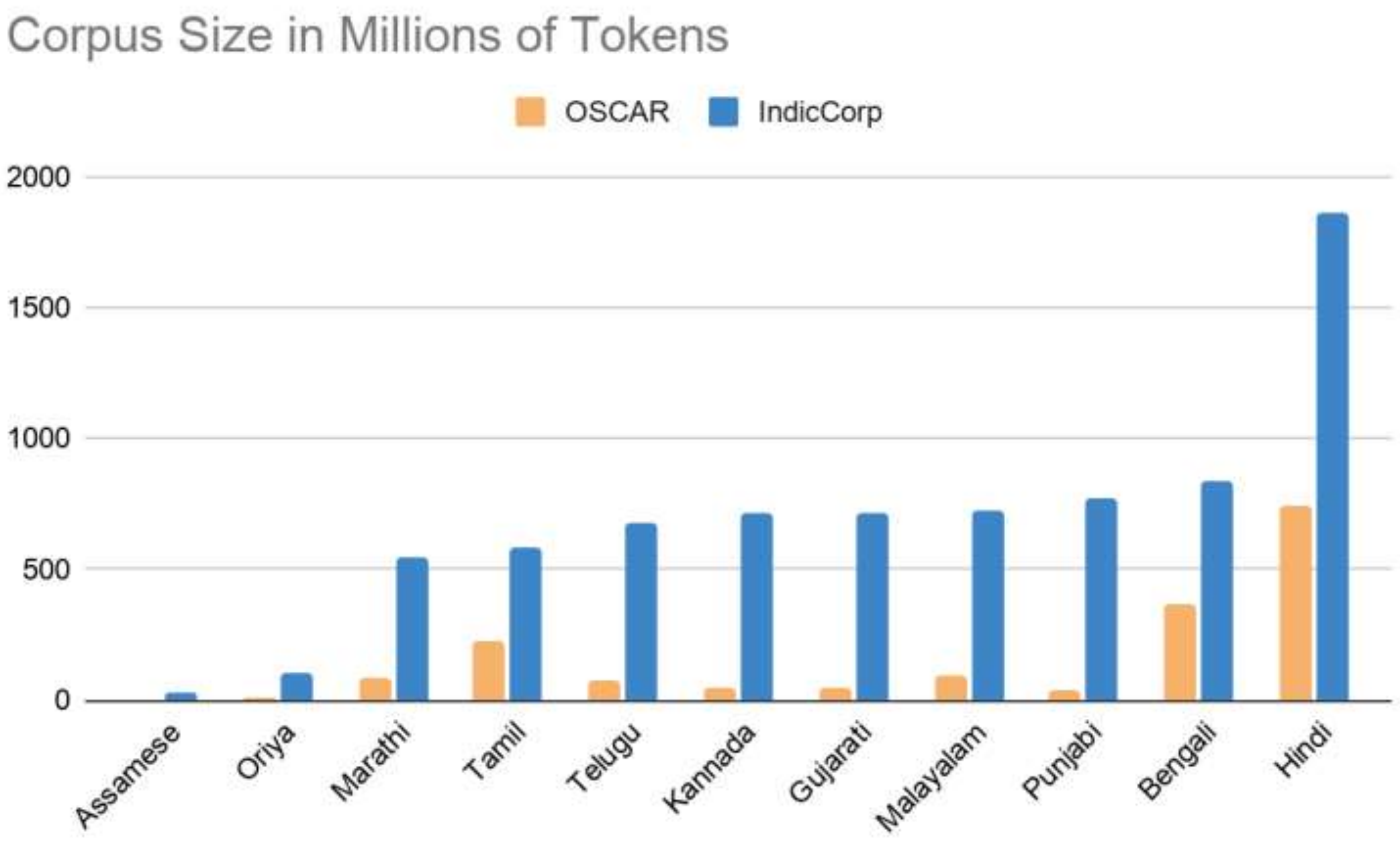
450M sentences

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

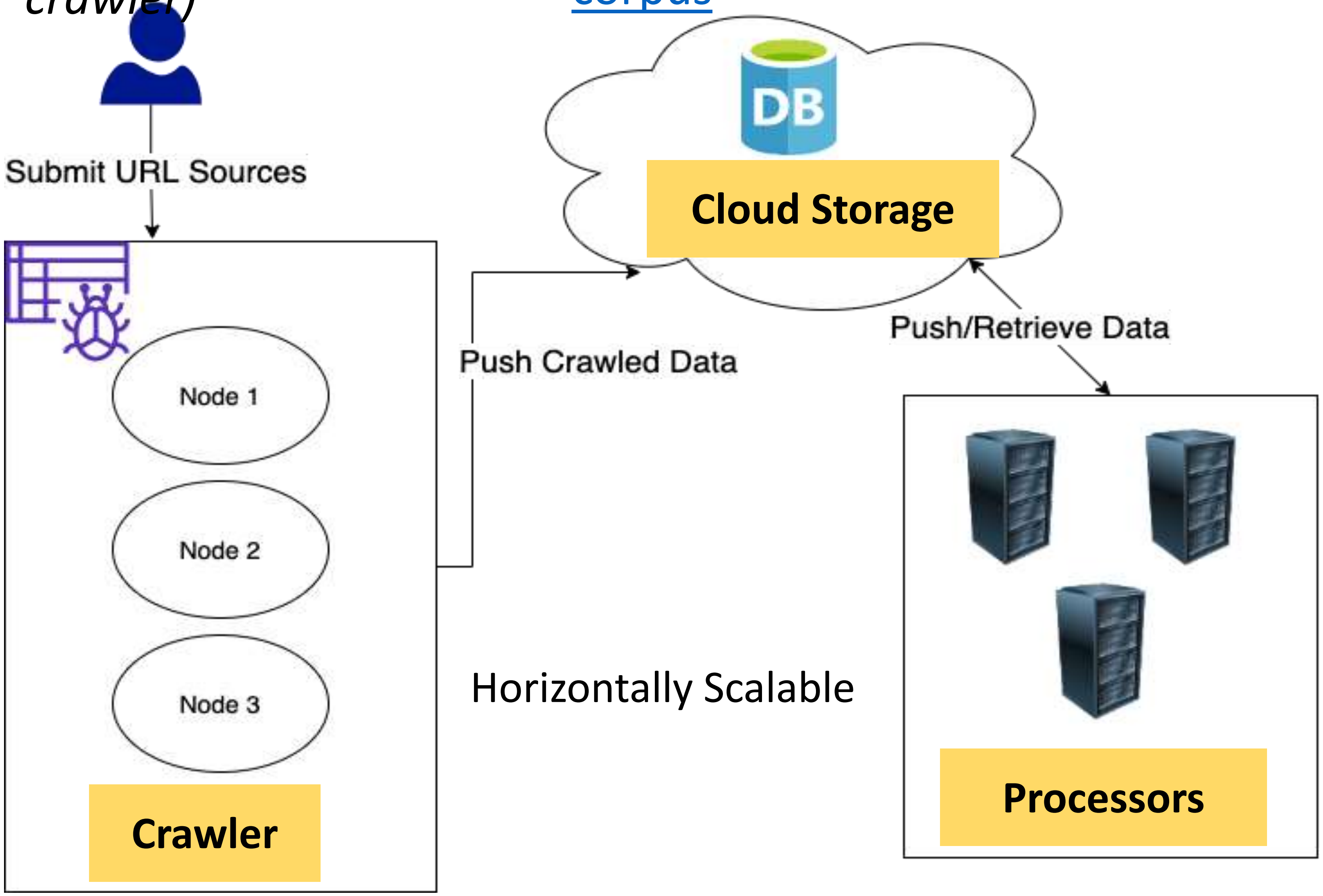
Webcorpus

(a scalable web crawler)

https://github.com/AI4Bharat/webcorpus

Distributed,
Multi-threaded

Dashboard



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

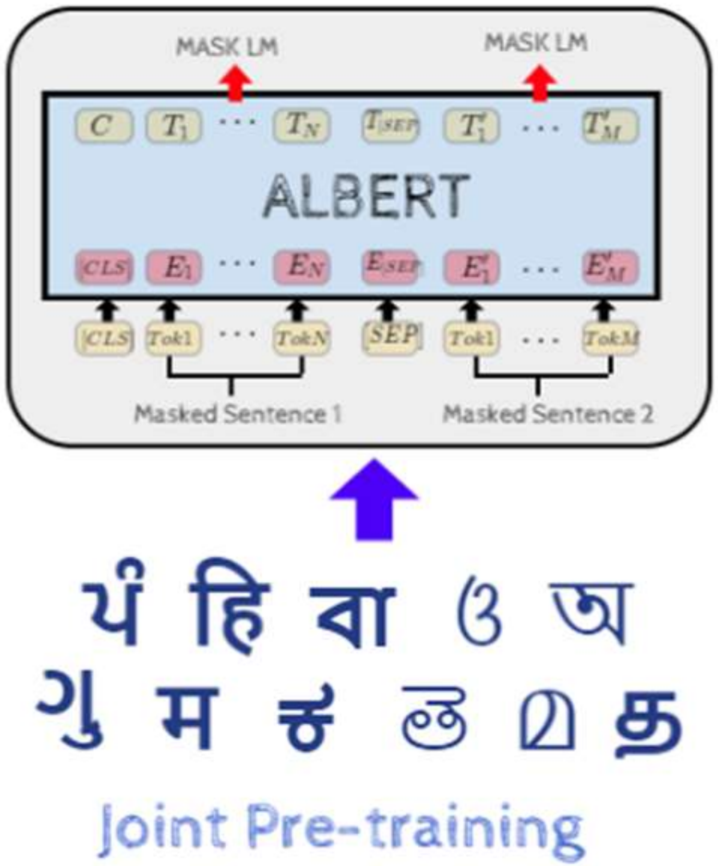
Lang	Dataset	FT-W	FT-WC	*IndicFT*
hi	BBC Articles	72.29	67.44	**77.02**
	IITP+ Movie	41.61	44.52	**45.81**
	IITP Product	58.32	57.17	**61.57**
bn	Soham Articles	62.79	64.78	**71.82**
gu	iNLTK Headlines	81.94	84.07	**90.74**
ml		86.35	83.65	**95.87**
mr		83.06	81.65	**91.40**
ta		90.88	89.09	**95.37**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

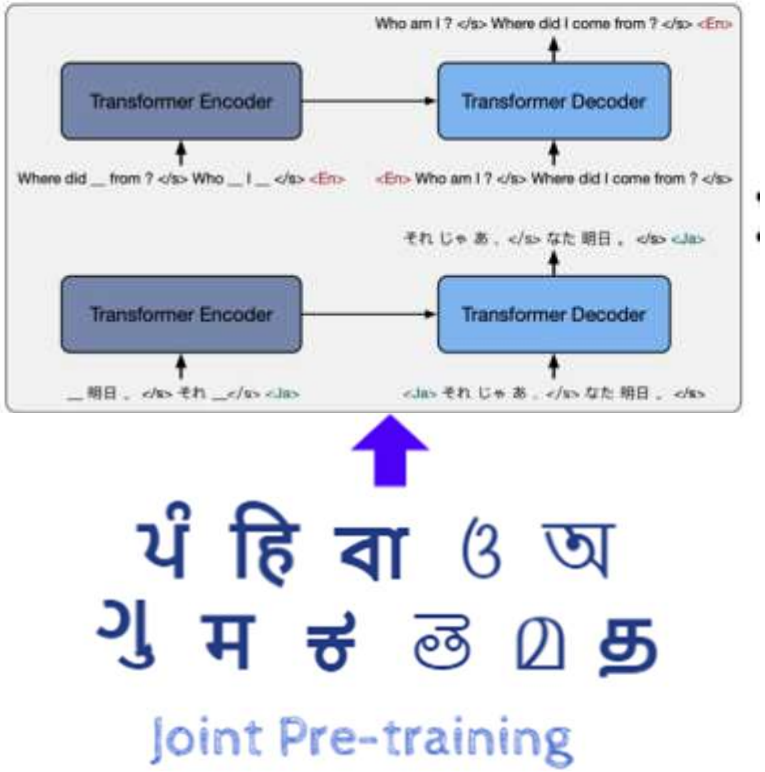
https://huggingface.co/ai4bharat/indic-bert



- 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-
https://huggingface.co/ai4bharat/Inc T



- 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.

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

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, [illegible] 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.

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: 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

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.* Arxiv pre-print 2203.05437. 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

Key Results

- Language group specific pre-trained models are better
    - Compact
    - Competitive with large global models like mBERT, mBART
- Multilingual fine-tuning and pre-training are useful
    - Particularly for low-resource languages

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

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

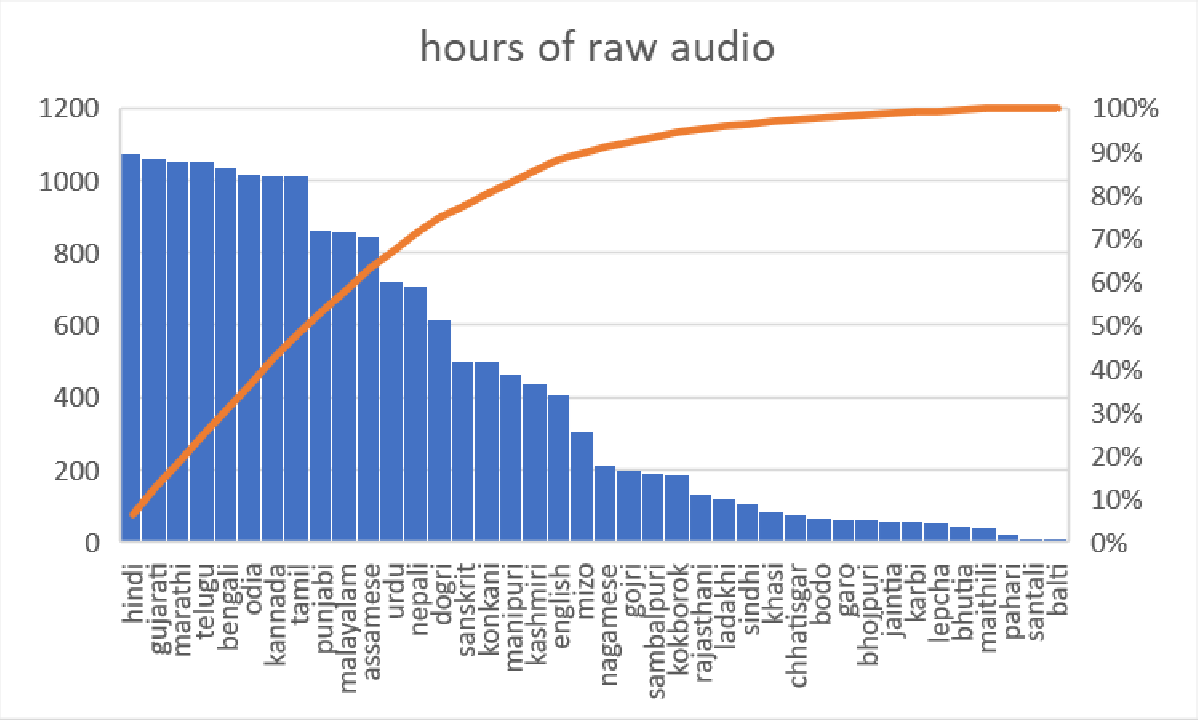
AI4Bharat, IITM, Microsoft, RBCDSAI,

AAAI 2022

Raw Speech Corpora

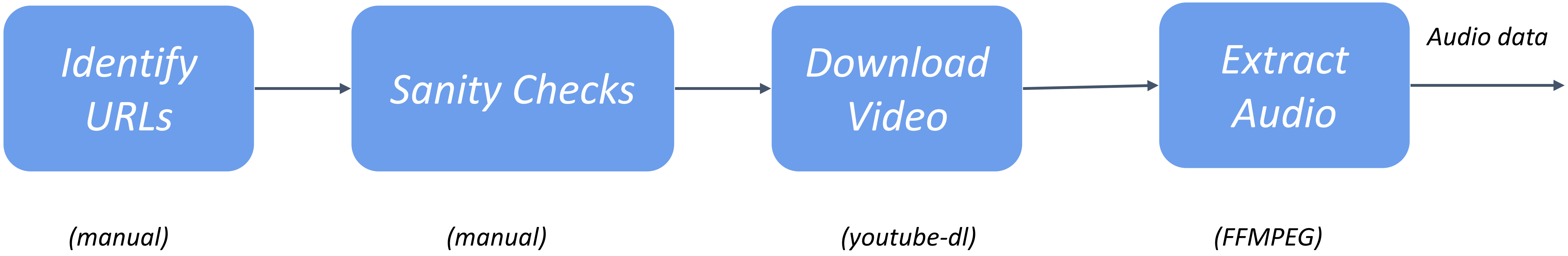
Sources: Youtube, NewsOnAir

- ~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

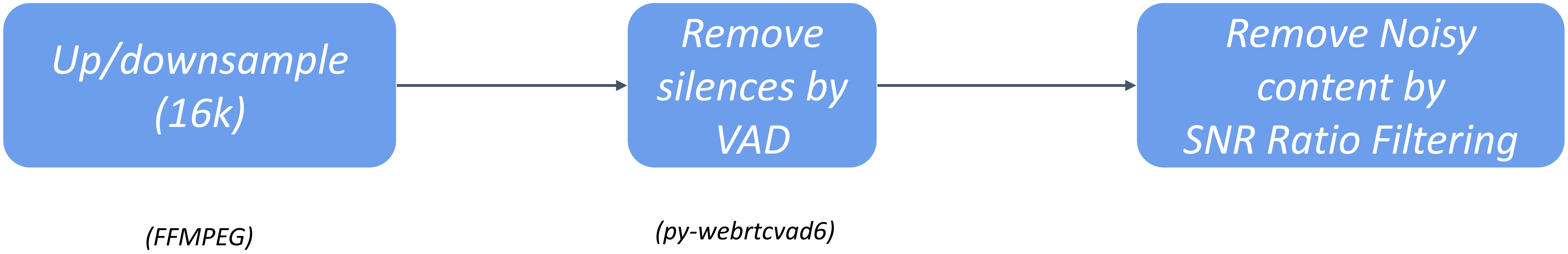


https://indicnlp.ai4bharat.org/indicwav2vec/

YouTube Data Extraction

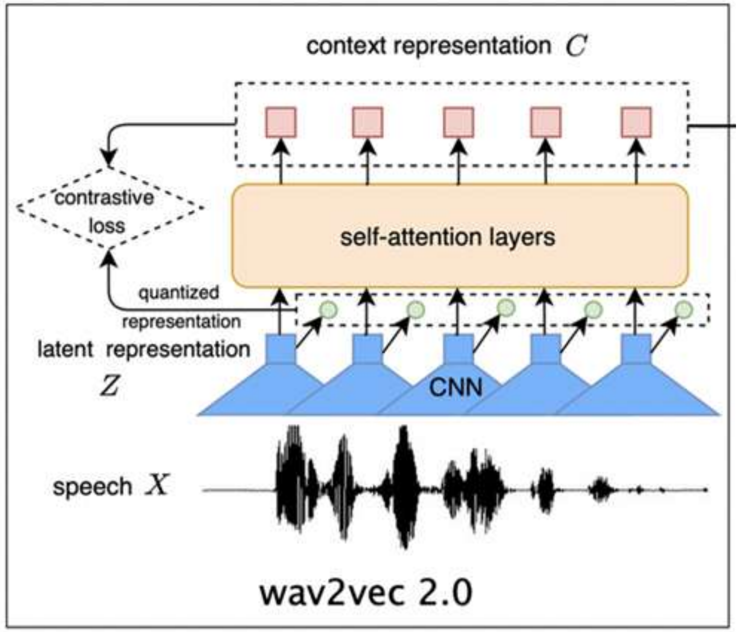


Audio Data Pre-processing



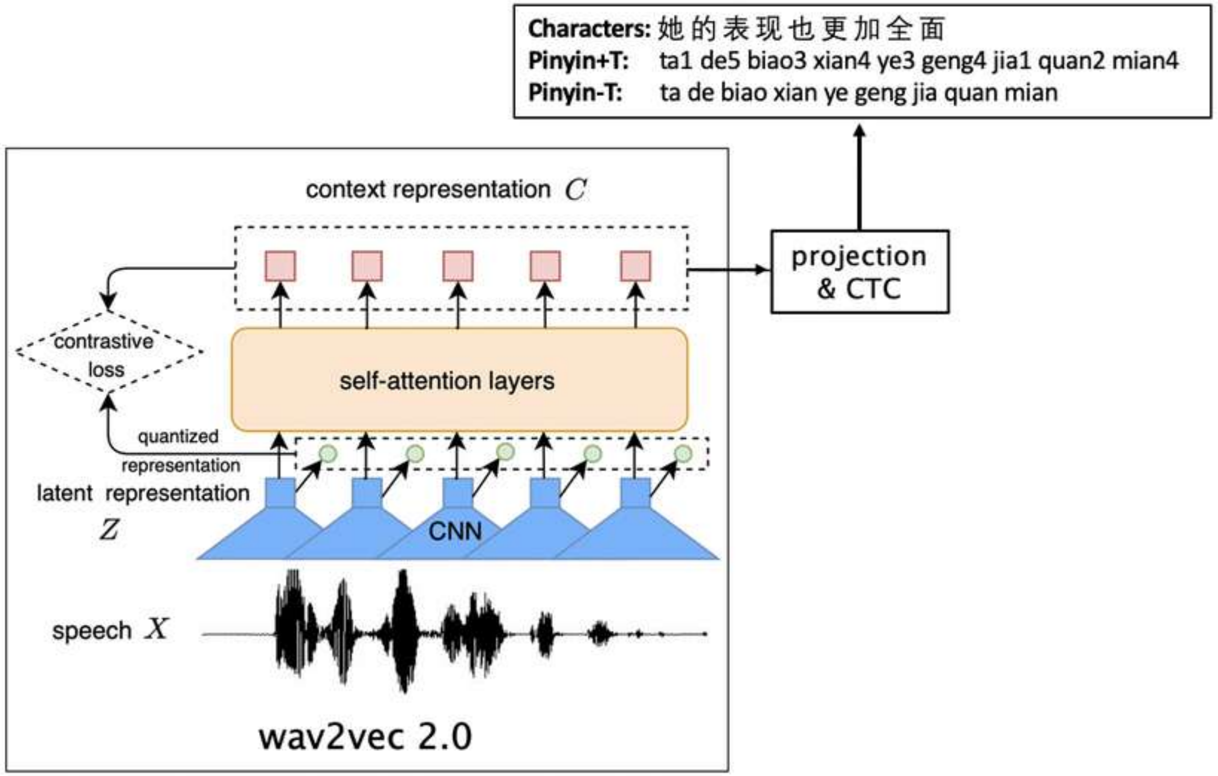
Unsupervised Pre-training

- 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



Decoding

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

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

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?

Thank You!