Understanding the Indian Languages: Challenges & Opportunities

A Language Diversity and Relatedness Perspective

Anoop Kunchukuttan

Machine Translation Group, Microsoft, Hyderabad

Atal FDP on Artificial Intelligence in Natural Language Processing, KIIT
18th October 2020

Outline

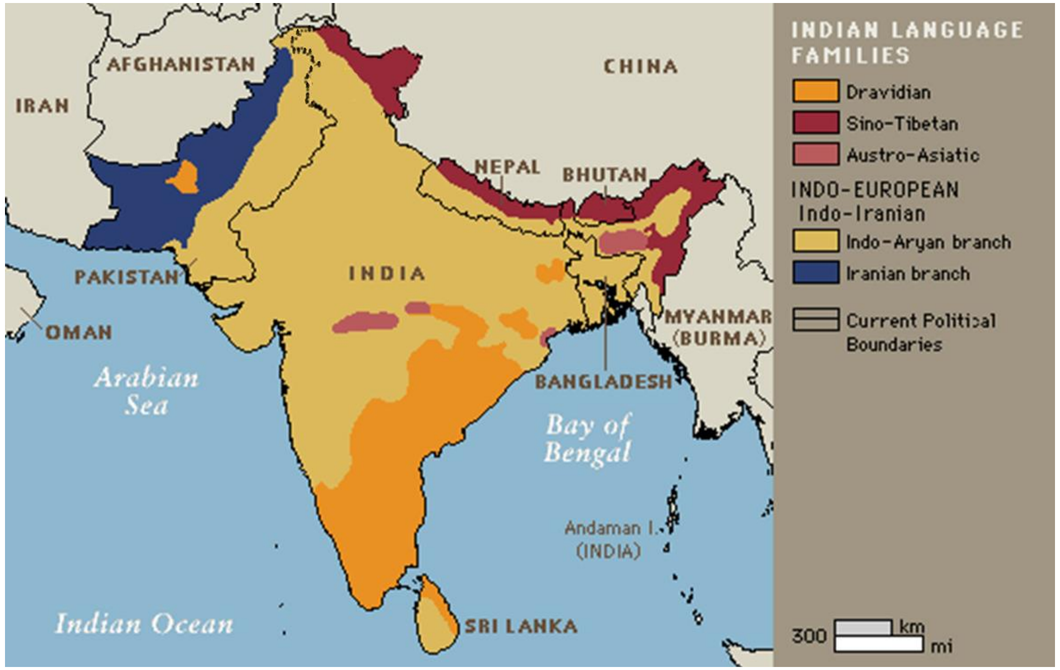
- **Introduction to Indian Languages**
- Opportunities & Challenges in Indic NLP
- Utilizing Relatedness between Indian Languages
- Getting Started with Indic NLP
    - IndicNLP Catalog
    - IndicNLP Library
    - IndicNLP Suite
- Summary

Diversity of Indian Languages

Highly multilingual country

Greenberg Diversity Index 0.9

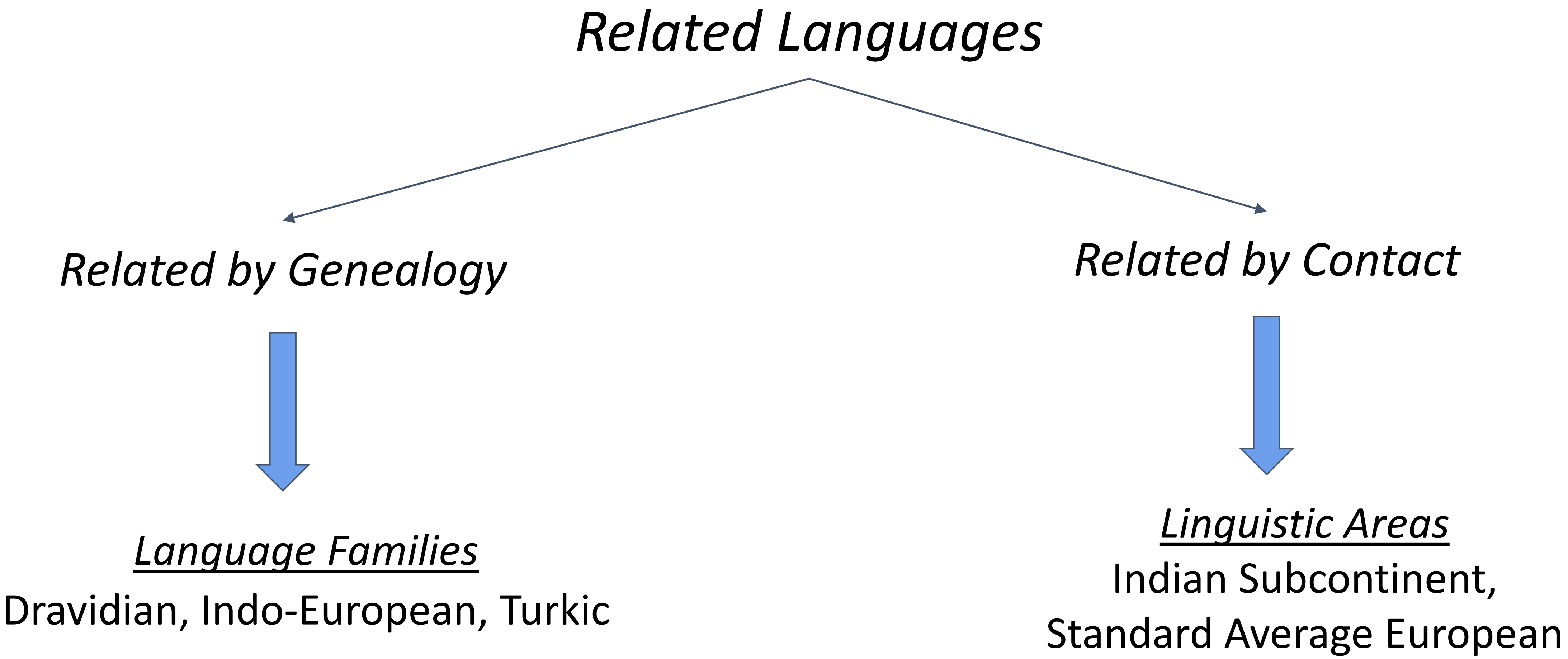
- *8 languages in the world's top 20 languages*
- *22 scheduled languages*
- *30 languages with more than 1 million speakers*
- *125 million English speakers*
- *1600 dialects*



Source: Quora

Sources: Wikipedia, Census of India 2011

There is also unity in Indian languages



Related Languages

Related by Genealogy

<u>Language Families</u>

Dravidian, Indo-European, Turkic

Related by Contact

<u>Linguistic Areas</u>

Indian Subcontinent,
Standard Average European

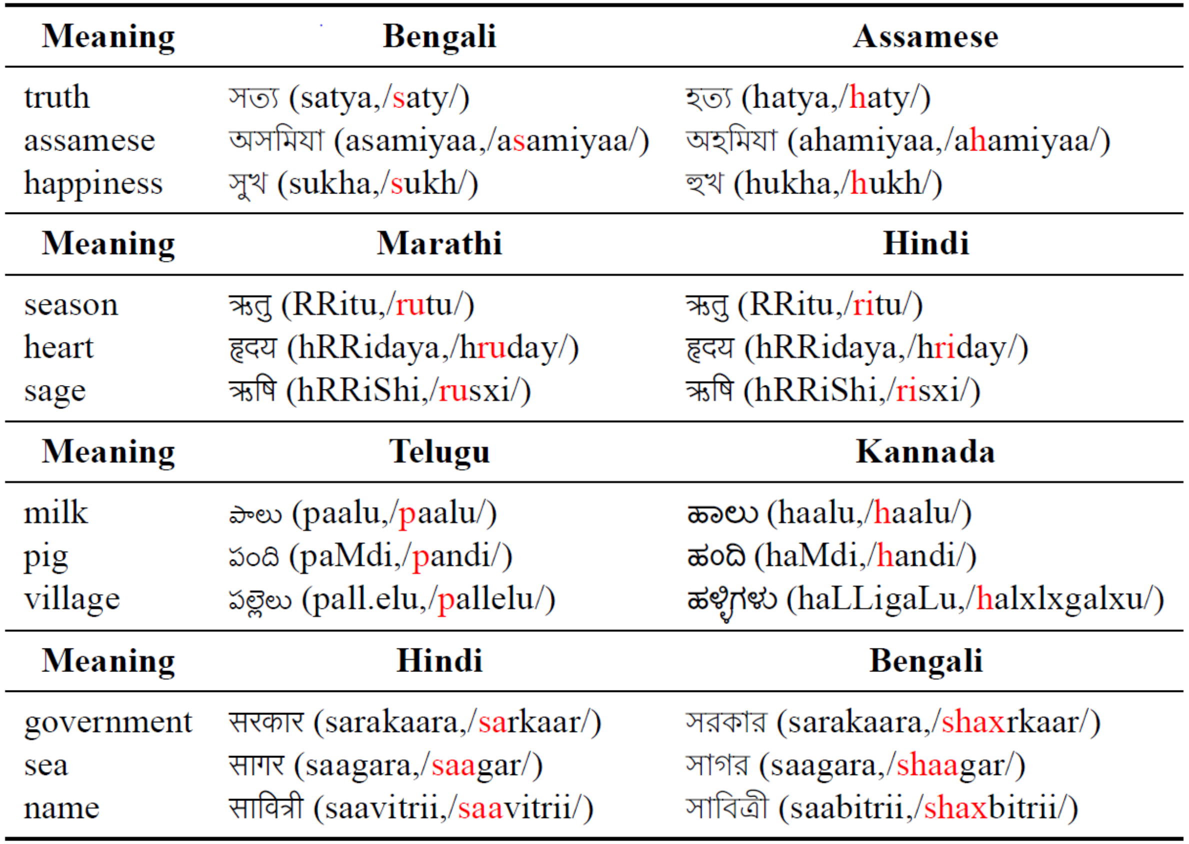
Related languages may not belong to the same language family!

Language Families

*Group of languages related through descent from a common ancestor, called the **proto-language** of that family*

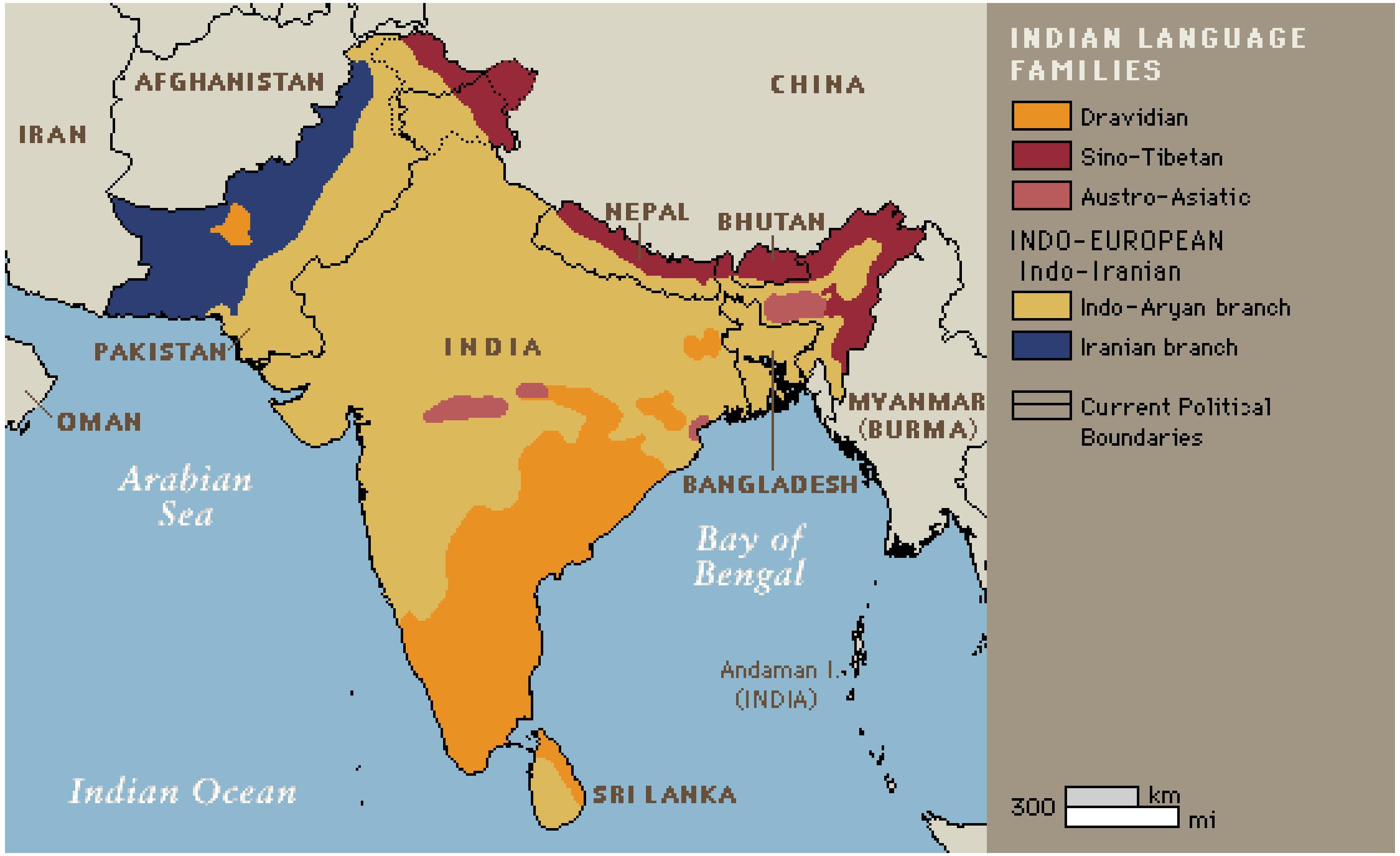
Regularity of sound change is the basis of studying genetic relationships

*These words are called **cognates***



Meaning	**Bengali**	**Assamese**
truth	সত্য (satya,/saty/)	হত্য (hatya,/haty/)
assamese	অসমিযা (asamiyaa,/asamiyaa/)	অহমিযা (ahamiyaa,/ahamiyaa/)
happiness	সুখ (sukha,/sukh/)	হুখ (hukha,/hukh/)
Meaning	**Marathi**	**Hindi**
season	ऋतु (RRitu,/rutu/)	ऋतु (RRitu,/ritu/)
heart	हृदय (hRRidaya,/hruday/)	हृदय (hRRidaya,/hriday/)
sage	ऋषि (hRRiShi,/rusxi/)	ऋषि (hRRiShi,/risxi/)
Meaning	**Telugu**	**Kannada**
milk	పాలు (paalu,/paalu/)	ಹಾಲು (haalu,/haalu/)
pig	పంది (paMdi,/pandi/)	ಹಂದಿ (haMdi,/handi/)
village	పల్లెలు (pall.elu,/pallelu/)	ಹಳ್ಳಿಗಳು (haLLigaLu,/halxlxgalxu/)
Meaning	**Hindi**	**Bengali**
government	सरकार (sarakaara,/sarkaar/)	সরকার (sarakaara,/shaxrkaar/)
sea	सागर (saagara,/saagar/)	সাগর (saagara,/shaagar/)
name	सावित्री (saavitrii,/saavitrii/)	সাবিত্রী (saabitrii,/shaxbitrii/)

Language Families in India



<u>4 major language families</u>

<u>Indo-Aryan</u>: *North India and Sri Lanka (branch of Indo-European)*

<u>Dravidian</u>: *South India & pockets in the North*

<u>Tibeto-Burman</u>: *North-East and along the Himalayan ranges*

<u>Austro-Asiatic</u>: *pockets in Central India, North-East, Nicobar Islands*

+

Andamanese family

Unknown language of the Sentinelese

Cognates & Borrowed words in Indian Languages

Indo-Aryan

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

Dravidian

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

Indo-Aryan words in Dravidian languages

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

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

Source: Wikipedia and IndoWordNet

Key Similarities between related languages

भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला
bhAratAcyA svAta.ntryadinAnimitta ameriketIla lOsa enjalsa shaharAta kAryakrama Ayojita karaNyAta AlA

Marathi

भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला
bhAratA cyA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA

Marathi segmented

भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया
bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsa shahara me.n kAryakrama Ayojita kiyA gayA

Hindi

Lexical: share significant vocabulary (cognates & loanwords)

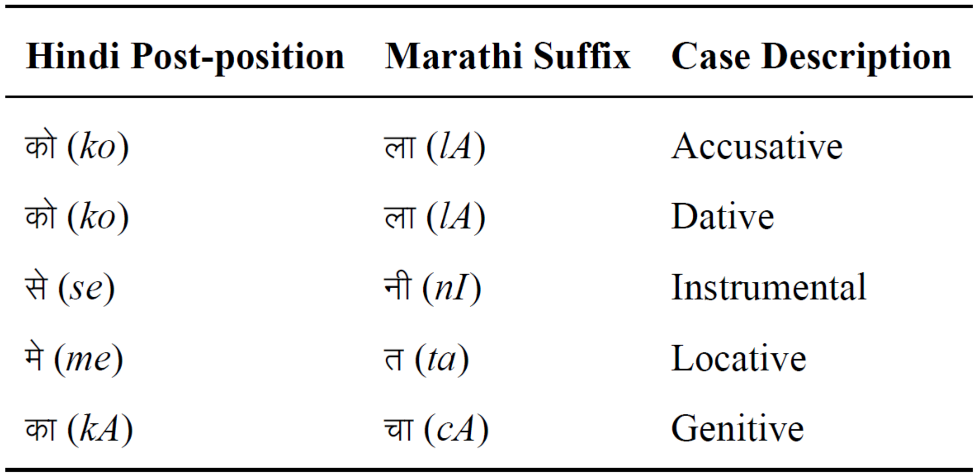
Morphological: correspondence between suffixes/post-positions

Syntactic: share the same basic word order

Morphological Similarity

- Inflectionally rich
- Sometimes agglutinative

  घरासमोरचा → घरा समोर चा
- Function words/suffixes
  - Largely 1-1 correspondence
- Similar case-marking systems



Hindi Post-position	Marathi Suffix	Case Description
को (*ko*)	ला (*lA*)	Accusative
को (*ko*)	ला (*lA*)	Dative
से (*se*)	नी (*nI*)	Instrumental
मे (*me*)	त (*ta*)	Locative
का (*kA*)	चा (*cA*)	Genitive

How similar are Indian Languages?

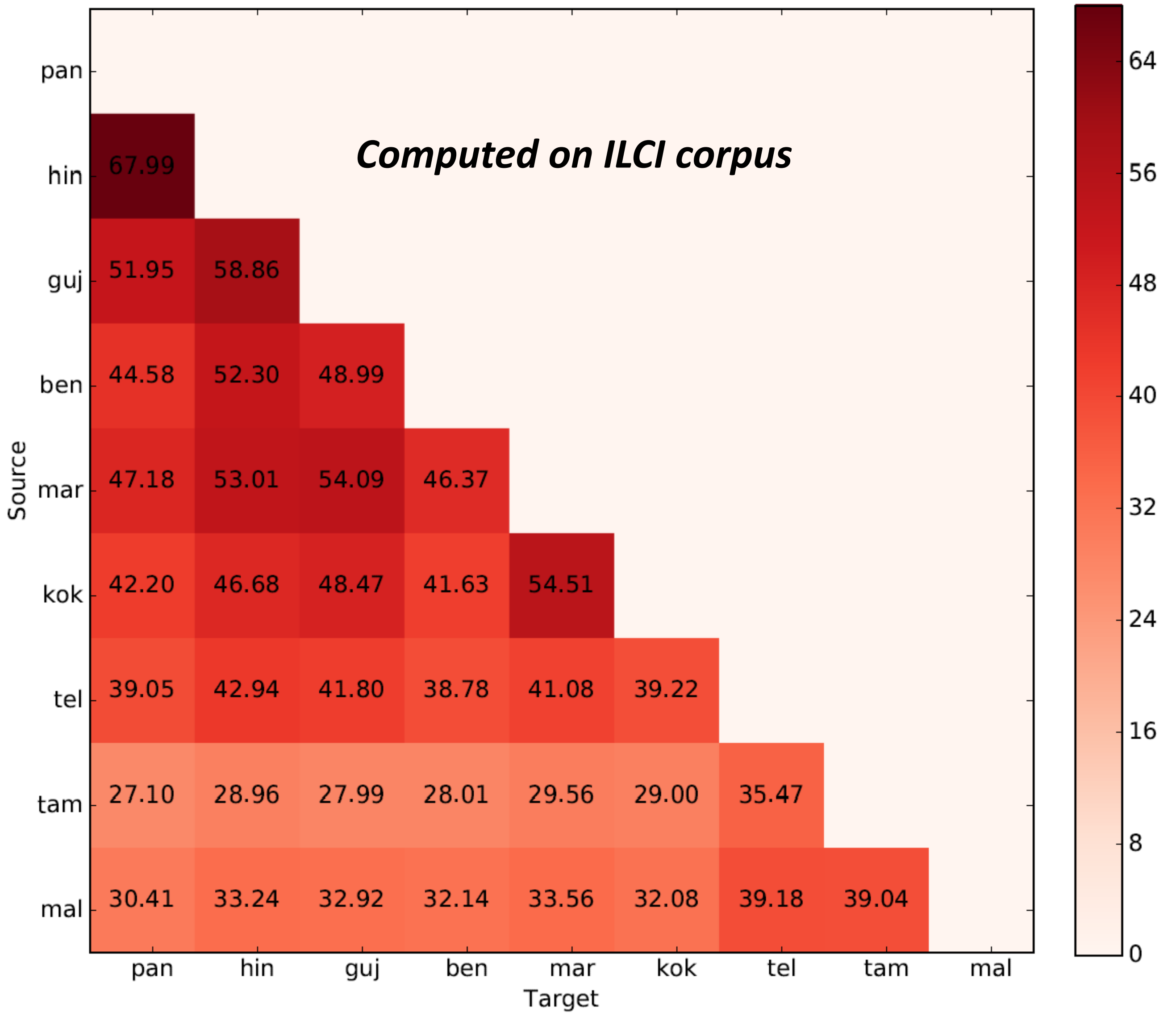
Estimate lexical similarity from parallel corpus

<u>Longest Common Subsequence Ratio (LCSR) for a sentence pair</u>

$$LCSR(s_1, s_2) = \frac{LCS(s_1, s_2)}{\max\big(len(s_1), len(s_2)\big)}$$

<u>LCSR for a language pair</u>

$$LCSR(L_1, L_2) = \frac{1}{|P(L_1, L_2)|} \sum_{\substack{(s_1,s_2)\in \\ P(L_1,L_2)}} LCSR(s_1, s_2)$$

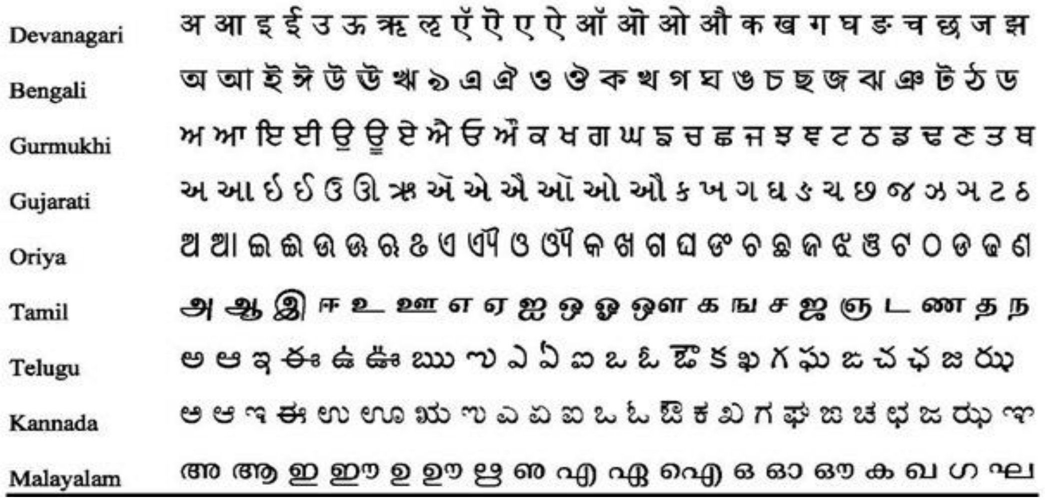


Computed on ILCI corpus

Source \ Target	pan	hin	guj	ben	mar	kok	tel	tam	mal
pan									
hin	67.99								
guj	51.95	58.86							
ben	44.58	52.30	48.99						
mar	47.18	53.01	54.09	46.37					
kok	42.20	46.68	48.47	41.63	54.51				
tel	39.05	42.94	41.80	38.78	41.08	39.22			
tam	27.10	28.96	27.99	28.01	29.56	29.00	35.47		
mal	30.41	33.24	32.92	32.14	33.56	32.08	39.18	39.04	

Anoop Kunchukuttan, Pushpak Bhattacharyya. *Utilizing Language Relatedness to improve SMT: A Case Study on Languages of the Indian Subcontinent*. eprint arXiv:2003.08925. 2020

Similarity of Indian Scripts



- Largely overlapping character set, but the visual rendering differs
- Traditional ordering of characters is same (*varnamala*)
- Dependent (*maatras*) and Independent vowels

***Abugida* scripts:**

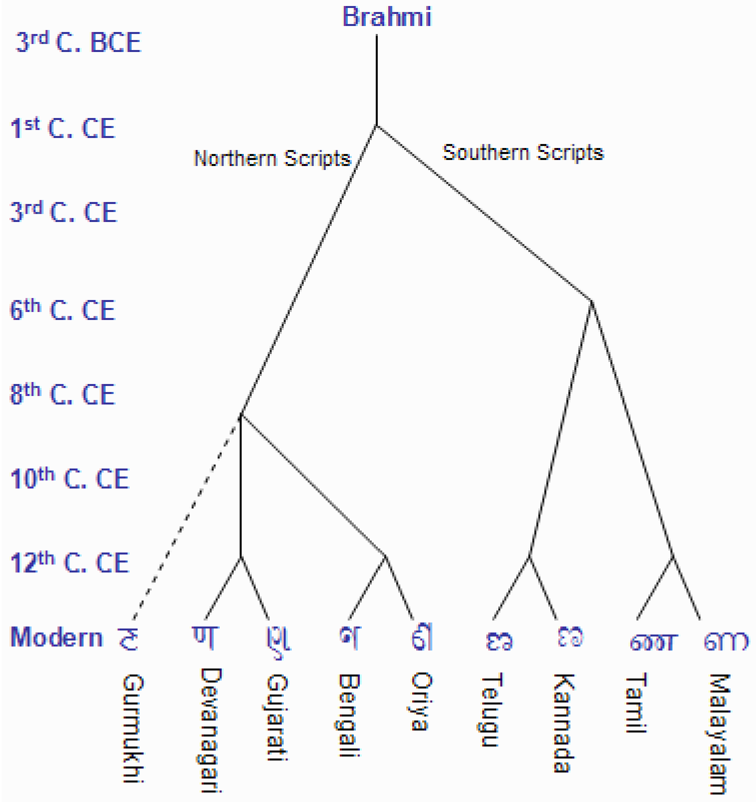
- primary consonants with secondary vowels diacritics (*maatras*)
- rarely found outside of the Brahmi family

- Consonant clusters (क्क,क्ष)
- Special symbols like:
  - *anusvaara* (nasalization), *visarga* (aspiration)
  - *halanta/pulli* (vowel suppression), nukta (Persian/Arabic sounds)
- Basic Unit is the akshar (a pseudo-syllable)

Origins

All major Indic scripts derived from the Brahmi script

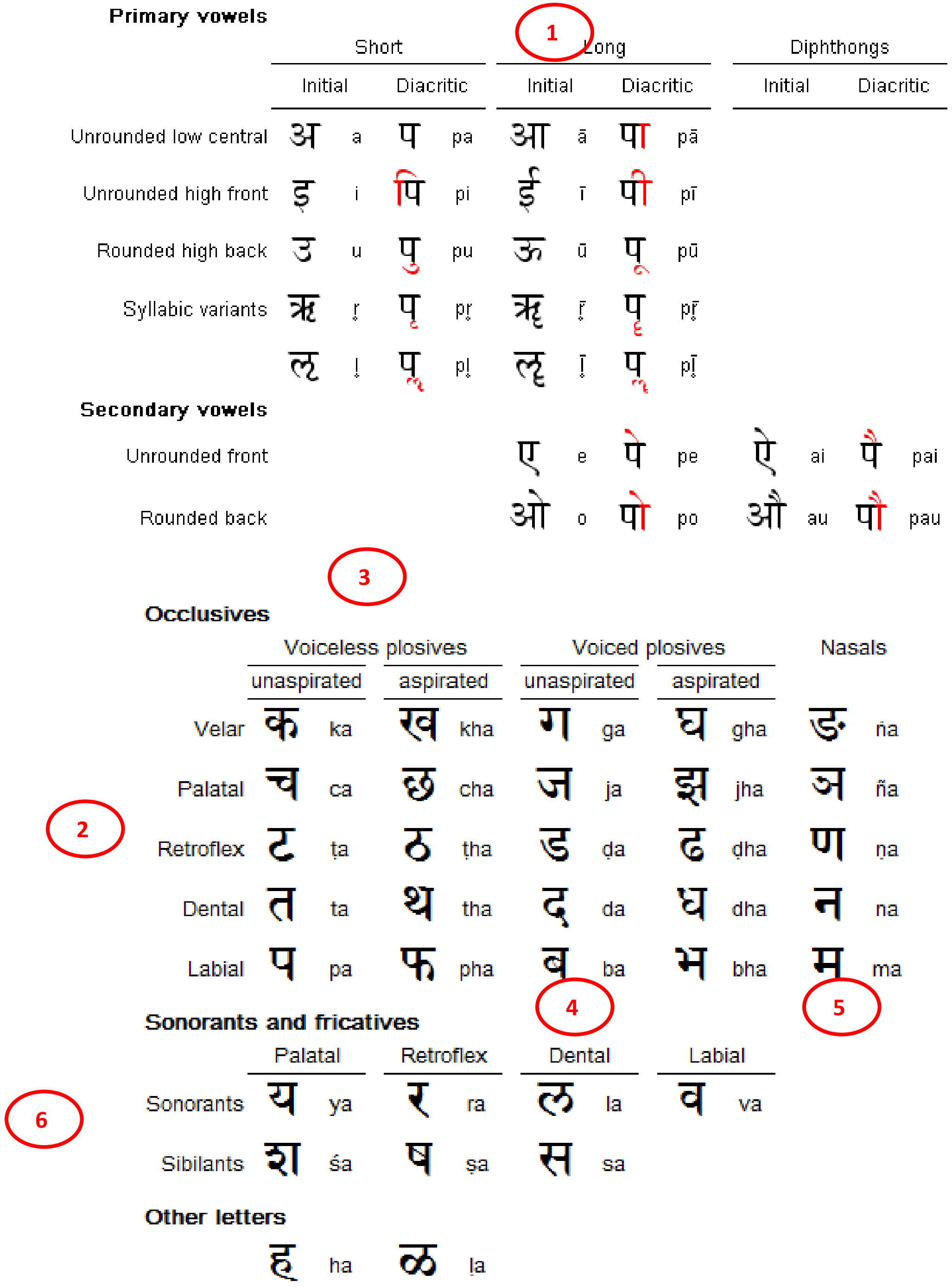
First seen in Ashoka’s edicts



- Same script used for multiple languages
  - Devanagari used for Sanskrit, Hindi, Marathi, Konkani, Nepali, Sindhi, etc.
  - Bangla script used for Assamese too
- Multiple scripts used for same language
  - Sanskrit traditionally written in all regional scripts
  - Punjabi: Gurumukhi & Shahmukhi, Sindhi: Devanagari & Persio-Arabic

Organized as per sound phonetic principles

shows various symmetries



Primary vowels

	Short		Long		Diphthongs	
	Initial	Diacritic	Initial	Diacritic	Initial	Diacritic
Unrounded low central	अ a	प pa	आ ā	पा pā		
Unrounded high front	इ i	पि pi	ई ī	पी pī		
Rounded high back	उ u	पु pu	ऊ ū	पू pū		
Syllabic variants	ऋ ṛ	पृ pṛ	ॠ ṝ	पॄ pṝ		
	ऌ ḷ	पॢ pḷ	ॡ ḹ	पॣ pḹ		

Secondary vowels

	Short		Long		Diphthongs	
Unrounded front			ए e	पे pe	ऐ ai	पै pai
Rounded back			ओ o	पो po	औ au	पौ pau

Occlusives

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

Sonorants and fricatives

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

Other letters

ह ha	ळ ḷa

Syllable as Basic Unit

akshara, the fundamental organizing principle of Indian scripts

(CONSONANT) + VOWEL

Examples: की (kI), प्रे (pre)

Hindi	पुस्तक	पु स्त क
Malayalam	പാലക്കാട് (पालक्काट्)	പാ ല ക്കാ ട് (पा ल क्का ट्)
Odia	ଉତ୍କଳ (उत्कळ)	ଉ ତ୍କ ଳ (उ त्क ळ)

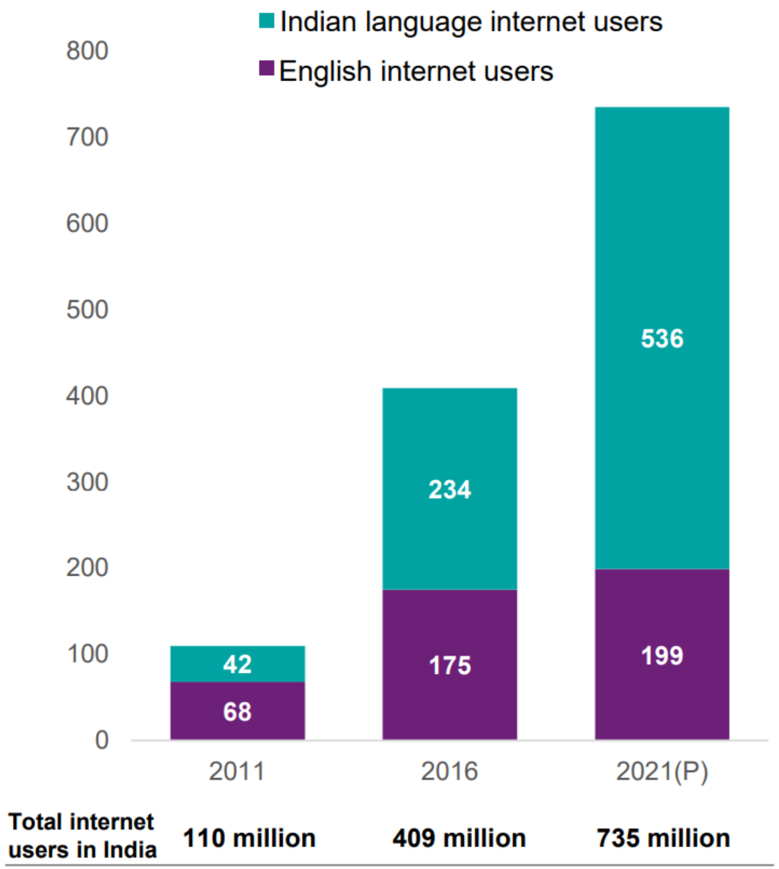
India as a linguistic area gives us robust reasons for writing a common or core grammar of many of the languages in contact

~ Anvita Abbi

Outline

- Introduction to Indian Languages
- **Opportunities & Challenges in Indic NLP**
- Utilizing Relatedness between Indian Languages
- Getting Started with Indic NLP
    - IndicNLP Catalog
    - IndicNLP Library
    - IndicNLP Suite
- Summary

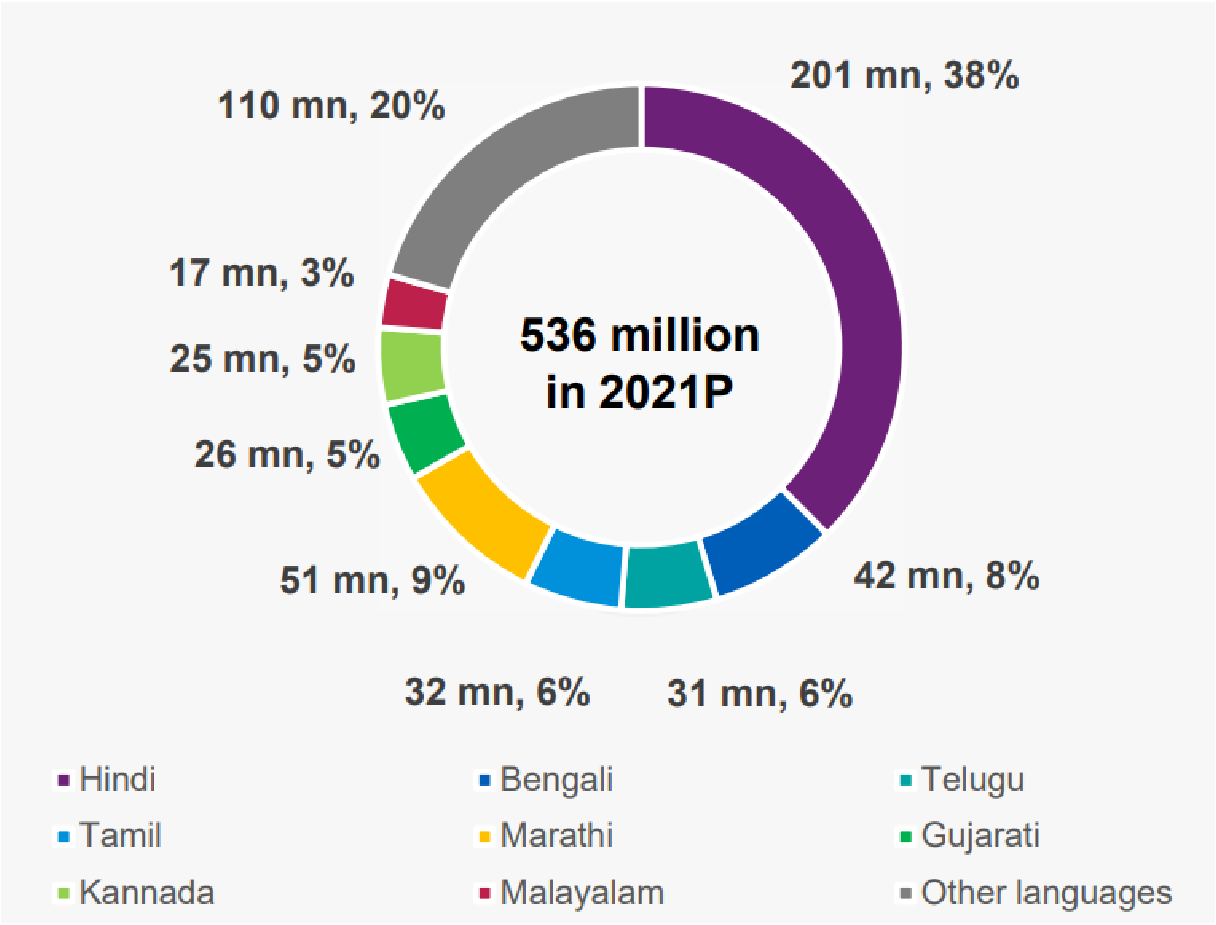
Indian Languages on the Internet



Source: KPMG in India's analysis, April 2017

(P): Projections

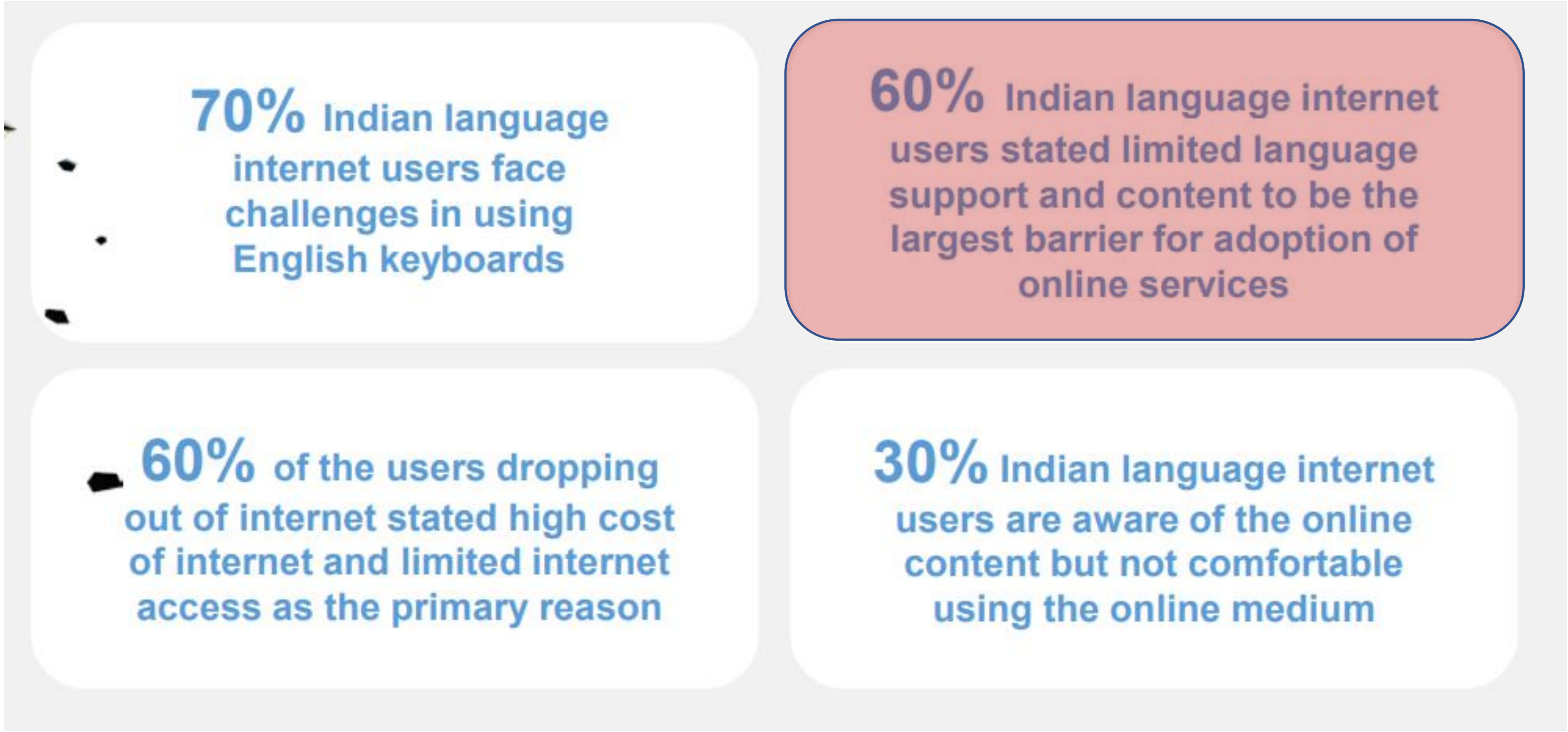
Internet User Base in India (in million)



Language Internet users 2021 projected (in million)

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

Challenges on language adoption on the Internet



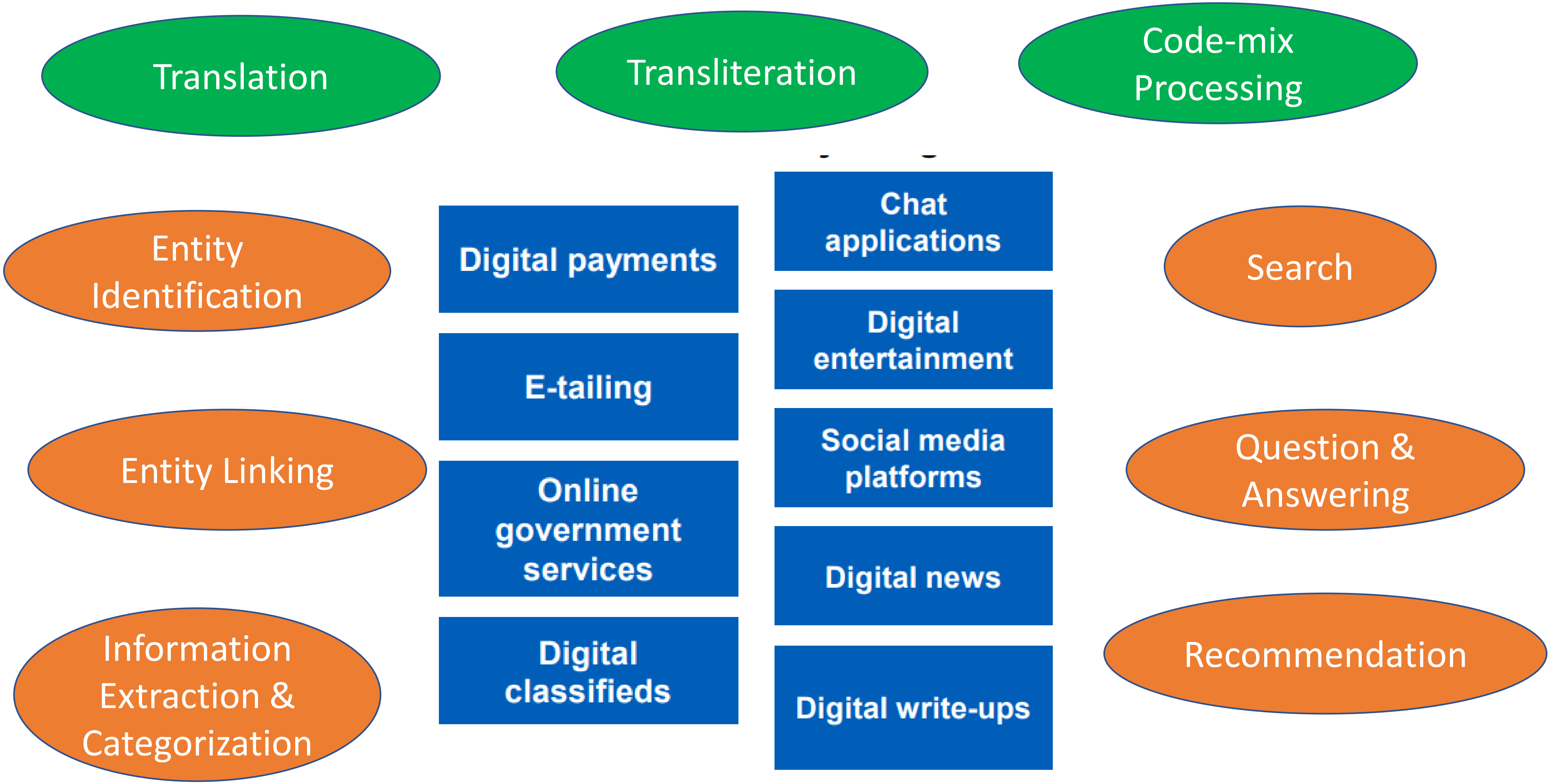
70% Indian language internet users face challenges in using English keyboards

60% Indian language internet users stated limited language support and content to be the largest barrier for adoption of online services

60% of the users dropping out of internet stated high cost of internet and limited internet access as the primary reason

30% Indian language internet users are aware of the online content but not comfortable using the online medium

How do we improve support for Indian languages?



Translation

Transliteration

Code-mix Processing

Entity Identification

Digital payments

Chat applications

Search

E-tailing

Digital entertainment

Entity Linking

Social media platforms

Online government services

Question & Answering

Digital news

Information Extraction & Categorization

Digital classifieds

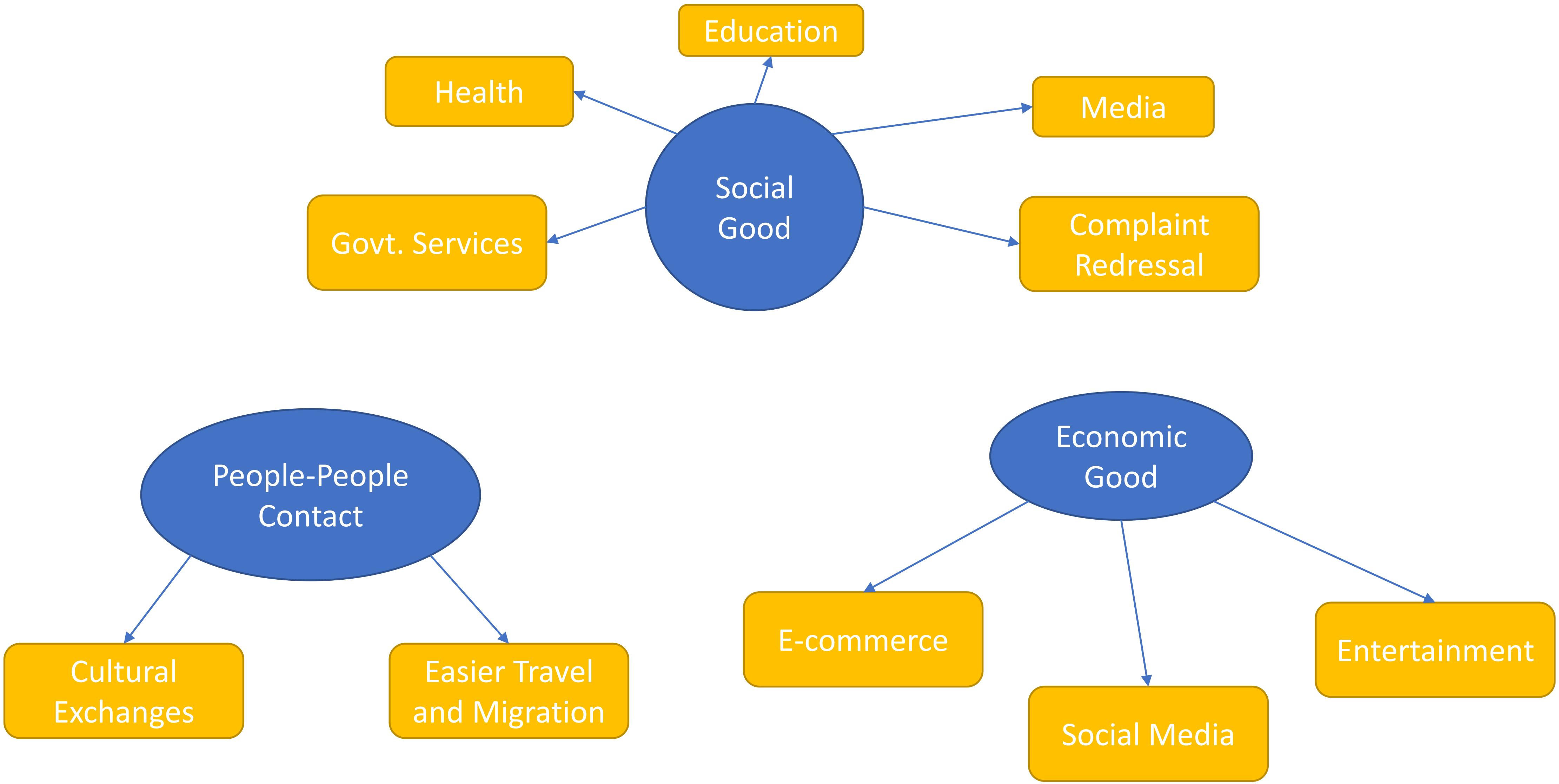
Recommendation

Digital write-ups

Applications requiring Indian language support

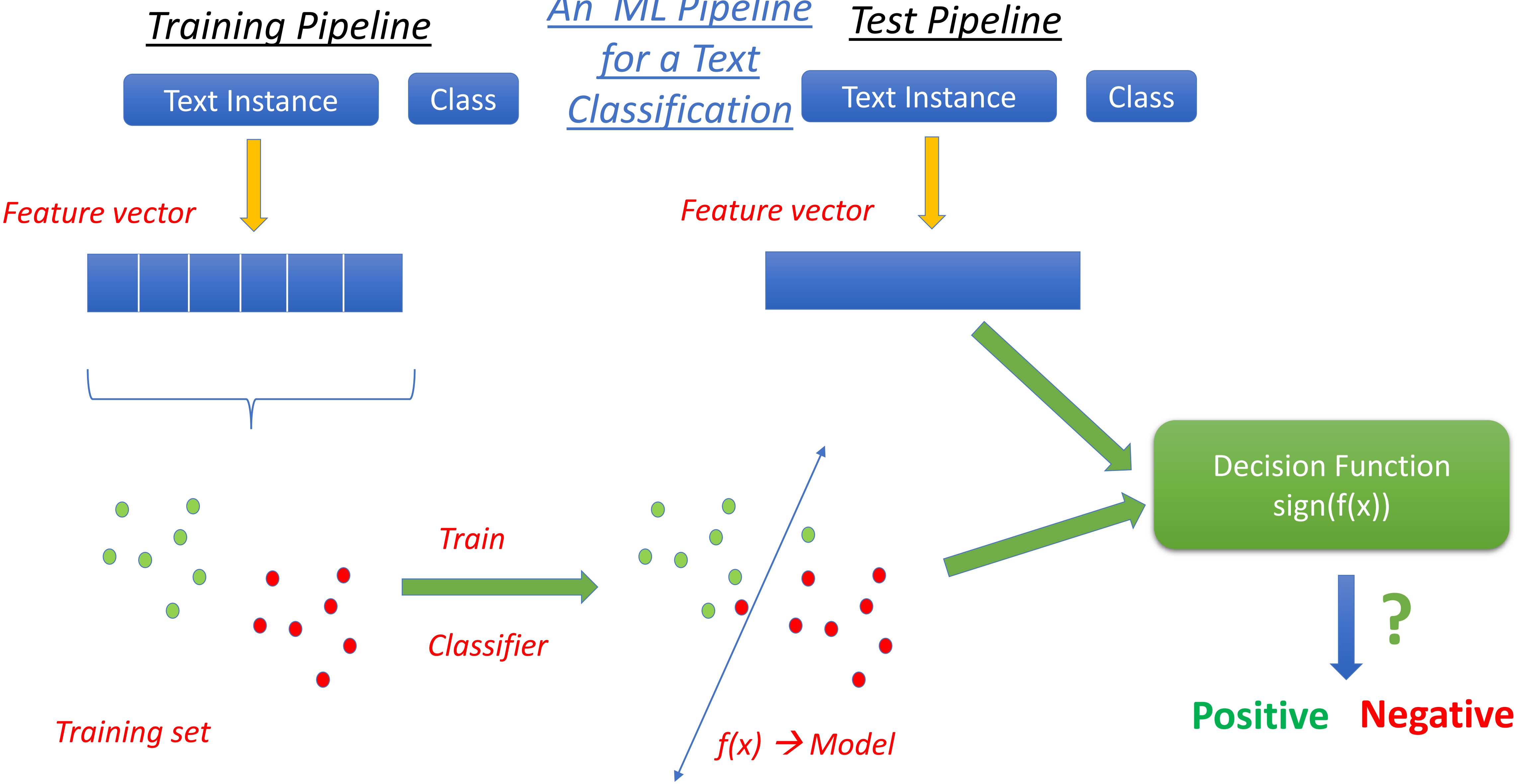
Addressing Multilinguality is important to maximizing impact of language technologies

Language Support *Cross-lingual Access*



Machine Learning is the dominant NLP Paradigm

An ML Pipeline for a Text Classification



Training Pipeline

Text Instance

Class

Feature vector

Training set

Train

Classifier

f(x) → Model

Test Pipeline

Text Instance

Class

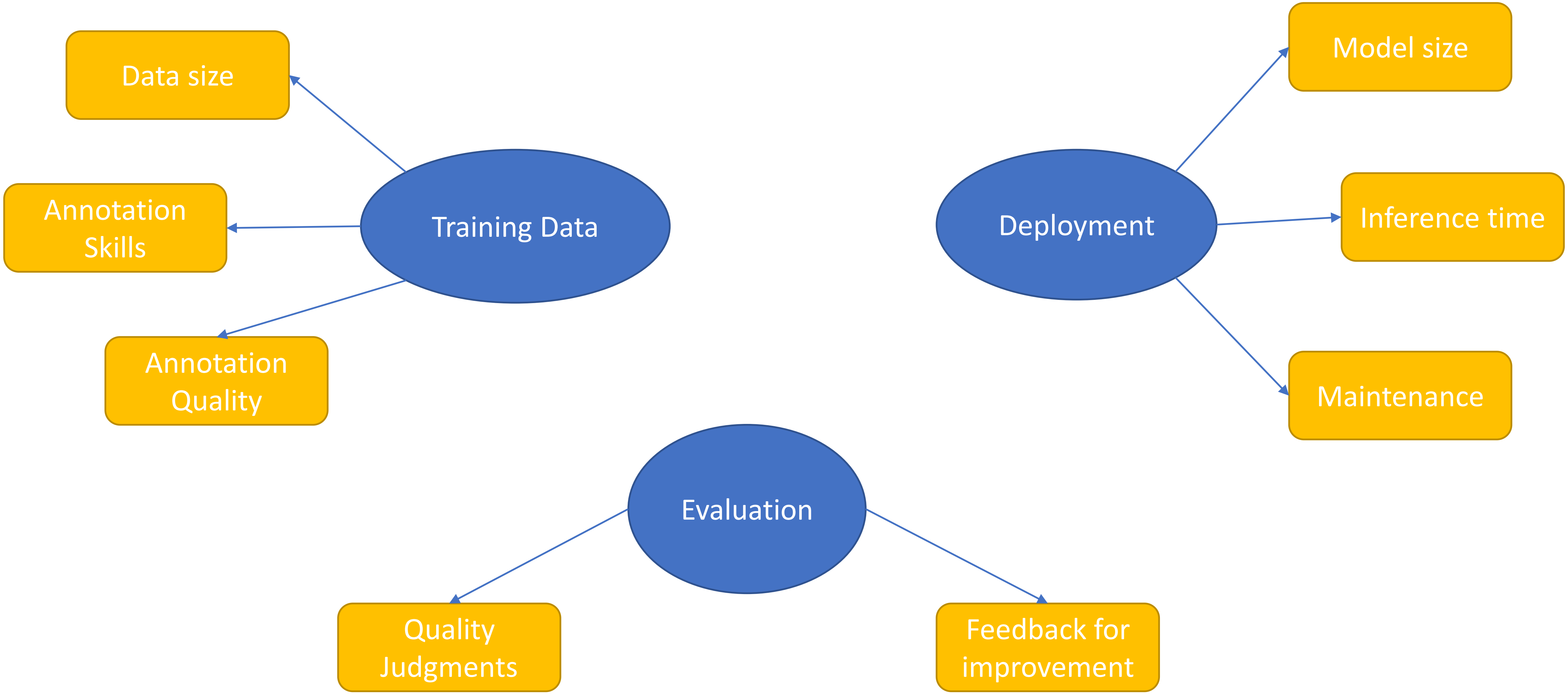
Feature vector

Decision Function
sign(f(x))

?

Positive **Negative**

Scalability Challenges for NLP solutions



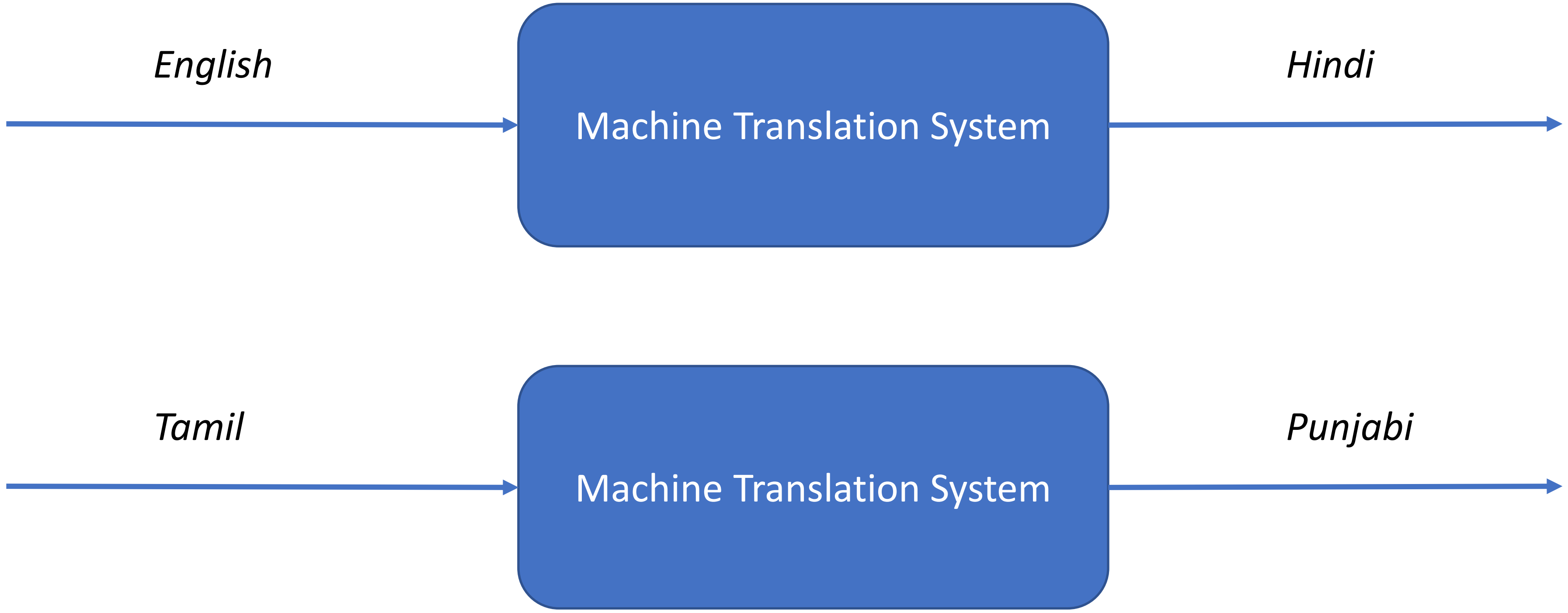
Effort and cost increase as languages increase

Need for a Unified Approach for Indic NLP

- *Can we share resources across languages?*
- *Can that also reduce effort & cost for deployment and maintenance?*
- *Can diversity of languages lead to better generalization?*

Can we utilize relatedness between Indian languages?

Broad Goal: Build NLP Applications that can work on different languages

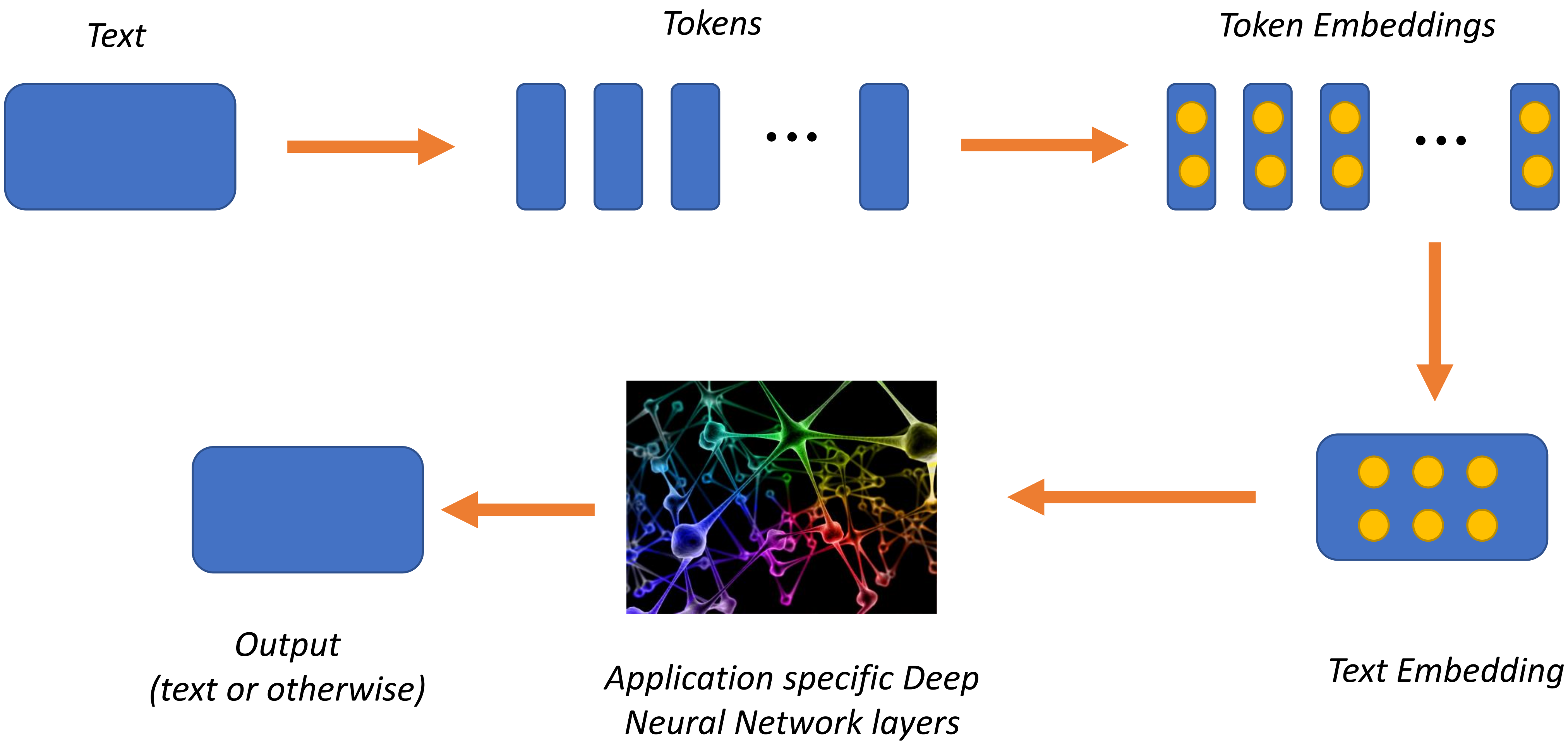


Can we improve English-Hindi translation using Tamil-Punjabi model?

Can we do English → Punjabi translation even if this data is not seen in training?

Can we train a single model for all translation pairs?

A Typical Deep Learning NLP Pipeline



Text

Tokens

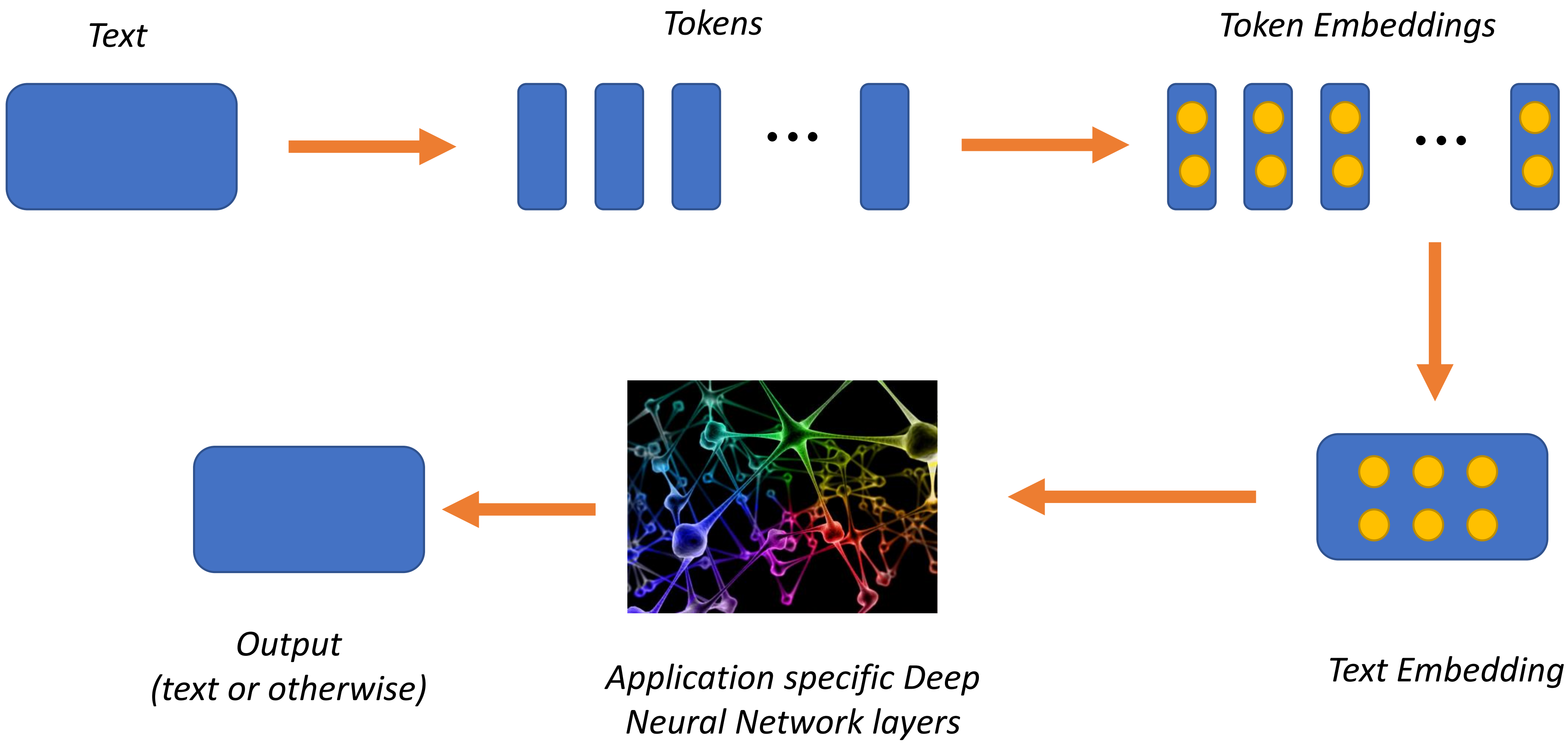
Token Embeddings

Text Embedding

Application specific Deep Neural Network layers

Output
(text or otherwise)

How do we transfer information across languages?



Text

Tokens

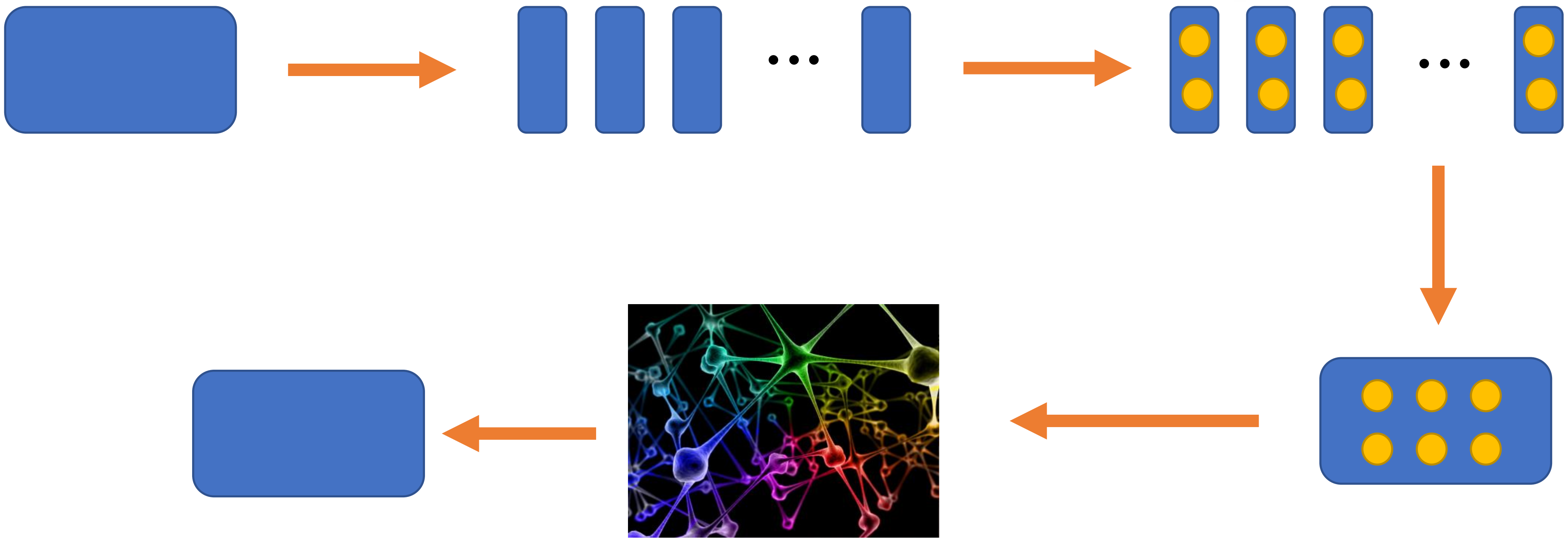
Token Embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

A Typical Multilingual NLP Pipeline



Text

Tokens

Token Embeddings

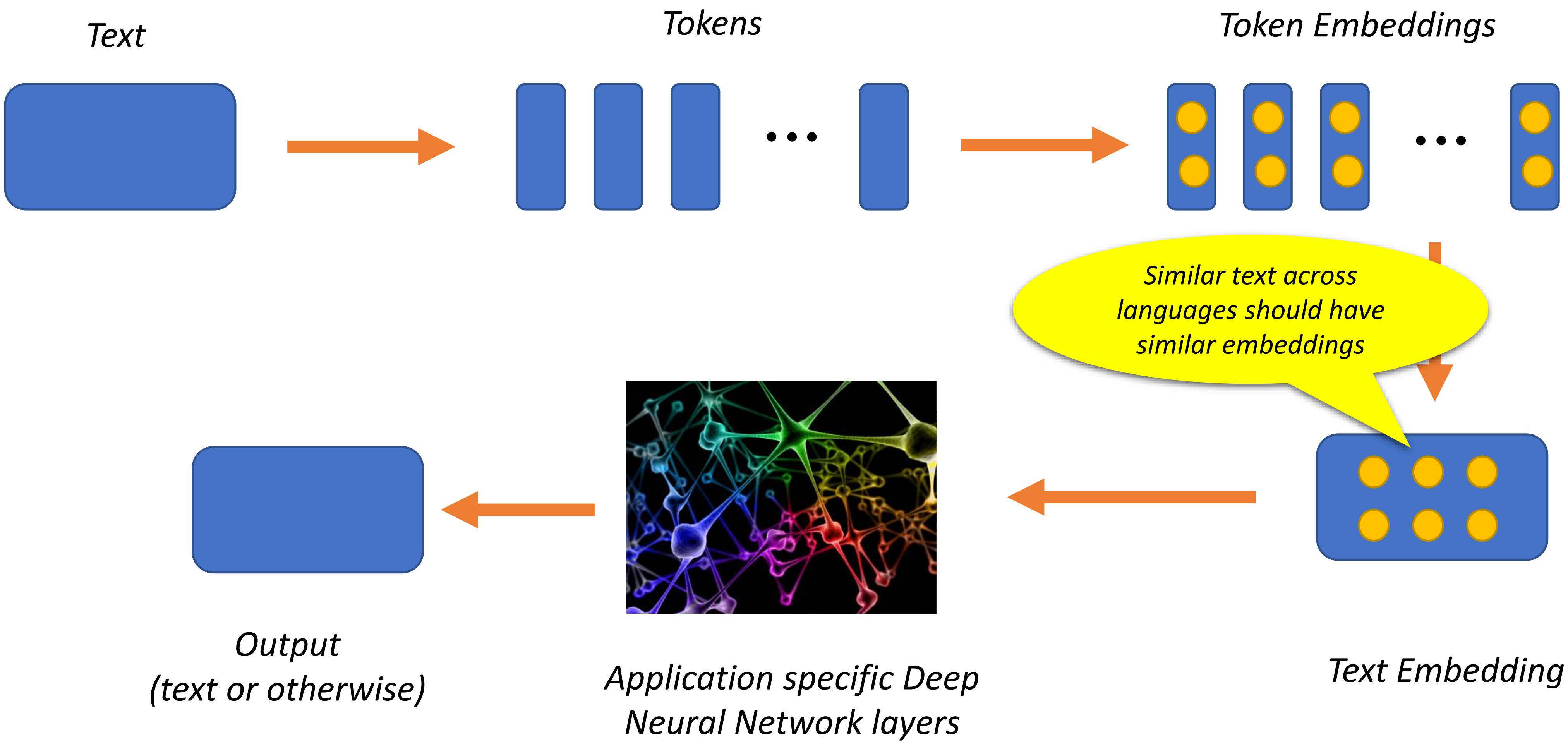
Similar tokens across languages should have similar embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

A Typical Multilingual NLP Pipeline



Text

Tokens

Token Embeddings

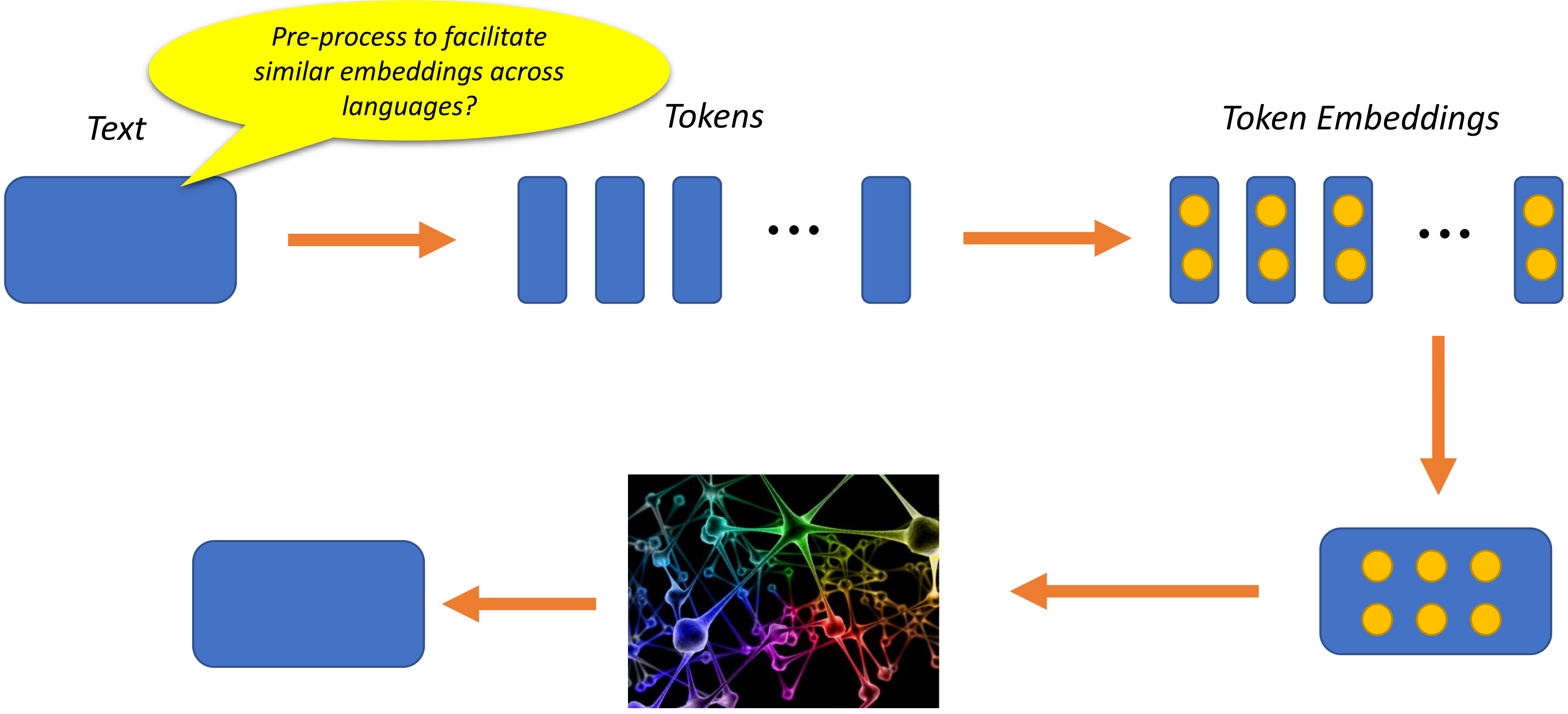
Similar text across languages should have similar embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

A Typical Multilingual NLP Pipeline



Pre-process to facilitate similar embeddings across languages?

Text

Tokens

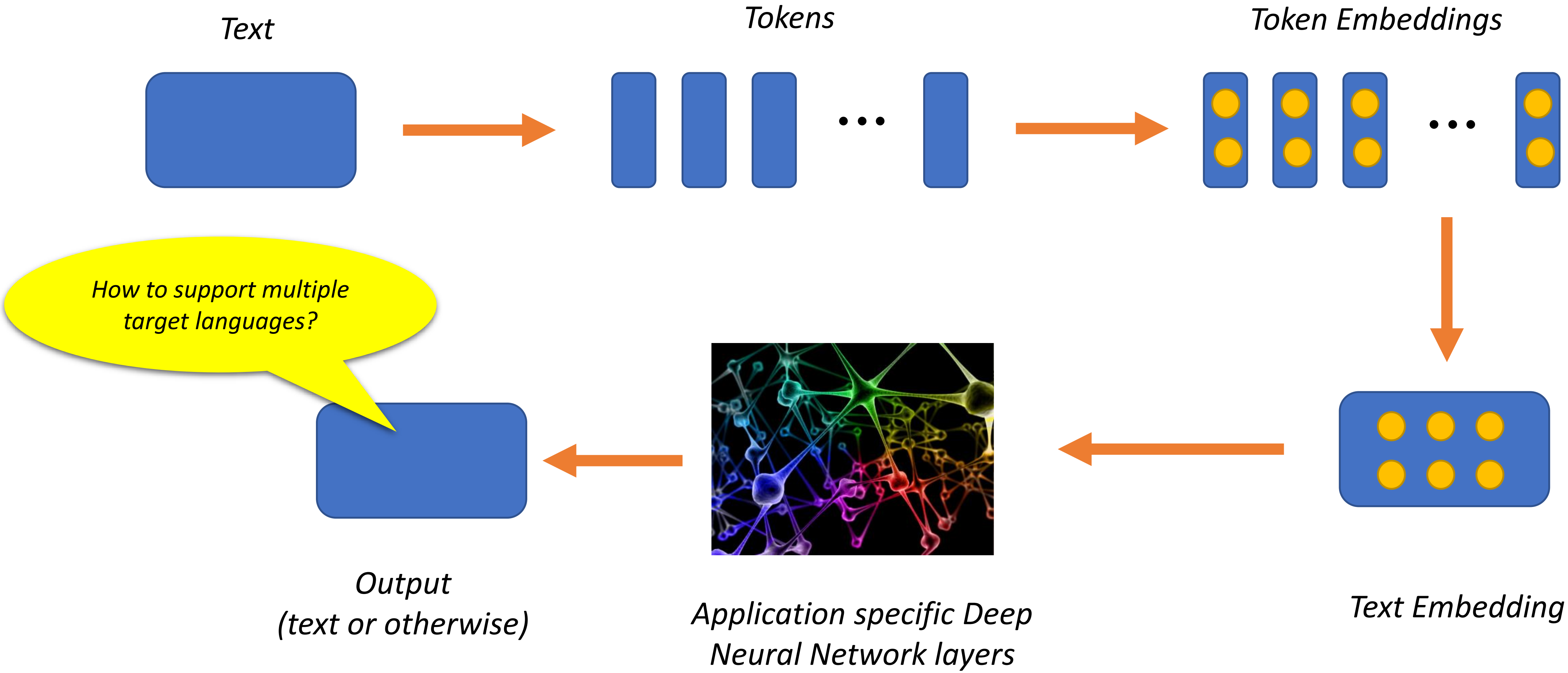
Token Embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

A Typical Multilingual NLP Pipeline



Text

Tokens

Token Embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

How to support multiple target languages?

Outline

- Introduction to Indian Languages
- Opportunities & Challenges in Indic NLP
- **Utilizing Relatedness between Indian Languages**
- Getting Started with Indic NLP
  - IndicNLP Catalog
  - IndicNLP Library
  - IndicNLP Suite
- Summary

Utilizing Relatedness between Indian Languages

Orthographic Similarity

Lexical Similarity

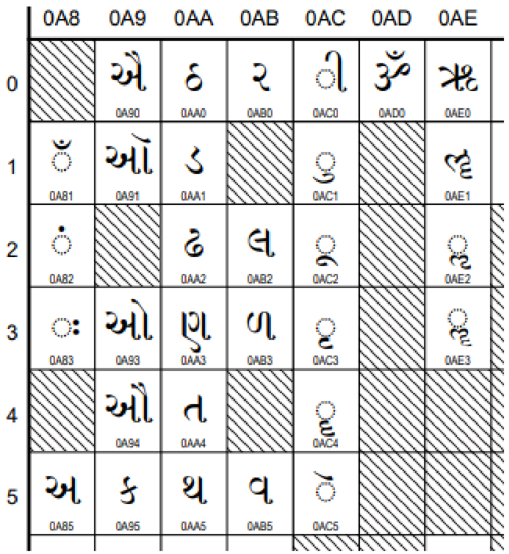
Syntactic Similarity

Utilizing Orthographic Similarity

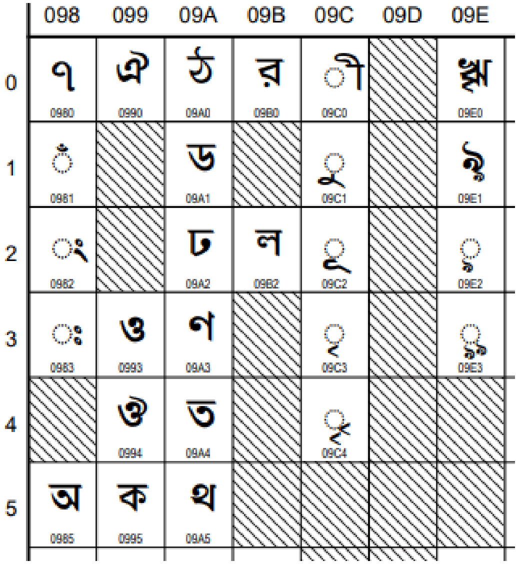
Script Conversion

- Read any script in any script
- Unicode standard enables consistent script conversion

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



	0A8	0A9	0AA	0AB	0AC	0AD	0AE
0	▨	ઐ 0A90	ઠ 0AA0	ર 0AB0	◌ી 0AC0	ૐ 0AD0	ૠ 0AE0
1	◌ઁ 0A81	ઑ 0A91	ડ 0AA1	▨	◌ુ 0AC1	▨	ૡ 0AE1
2	◌ં 0A82	▨	ઢ 0AA2	લ 0AB2	◌ૂ 0AC2	▨	◌ૢ 0AE2
3	◌ઃ 0A83	ઓ 0A93	ણ 0AA3	ળ 0AB3	◌ૃ 0AC3	▨	◌ૣ 0AE3
4	▨	ઔ 0A94	ત 0AA4	▨	◌ૄ 0AC4	▨	▨
5	અ 0A85	ક 0A95	થ 0AA5	વ 0AB5	◌ૅ 0AC5	▨	▨



	098	099	09A	09B	09C	09D	09E
0	ঀ 0980	ঐ 0990	ঠ 09A0	র 09B0	◌ী 09C0	▨	ৠ 09E0
1	◌ঁ 0981	▨	ড 09A1	▨	◌ু 09C1	▨	ৡ 09E1
2	◌ং 0982	▨	ঢ 09A2	ল 09B2	◌ূ 09C2	▨	◌ৢ 09E2
3	◌ঃ 0983	ও 0993	ণ 09A3	▨	◌ৃ 09C3	▨	◌ৣ 09E3
4	▨	ঔ 0994	ত 09A4	▨	◌ৄ 09C4	▨	▨
5	অ 0985	ক 0995	থ 09A5	▨	▨	▨	▨

केरला

কেরলা

કેરલા

Multilingual Transliteration

केरल	kerala

Hindi → English corpus

Bengali → English corpus

Telugu → English corpus

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

<u>Example of Multi-task Learning</u>

Similar tasks help each other

<u>Zero-shot transliteration is possible</u>

Perform Kannada → English transliteration even if network has not seen that data

Anoop Kunchukuttan, Mitesh Khapra, Gurneet Singh, Pushpak Bhattacharyya. *Leveraging Orthographic Similarity for Multilingual Neural Transliteration*. Transactions of Association of Computational Linguistics. 2018.

Concat training sets

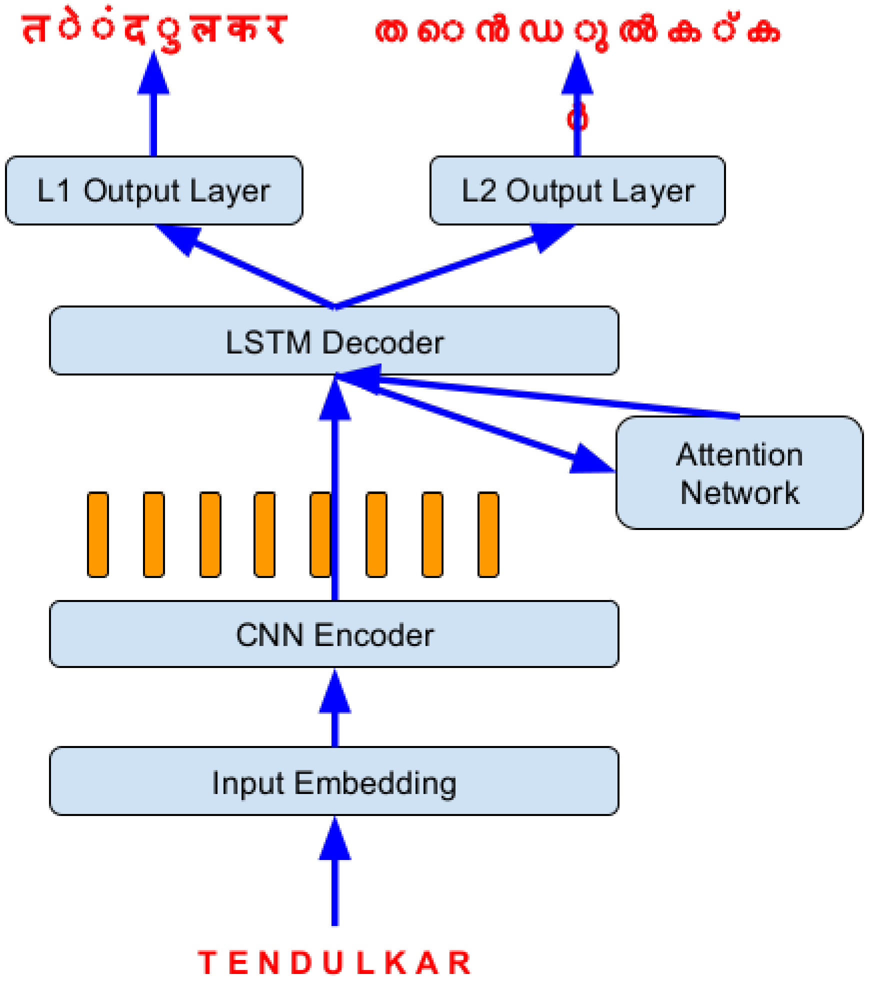
Malayalam	കോഴിക്കോട്	kozhikode
Hindi	केरल	kerala
Kannada	ಬೆಂಗಳೂರು	bengaluru

Convert to a common script

Malayalam	कोऴिक्कोट्	kozhikode
Hindi	केरल	kerala
Kannada	बेंगळूरु	bengaluru

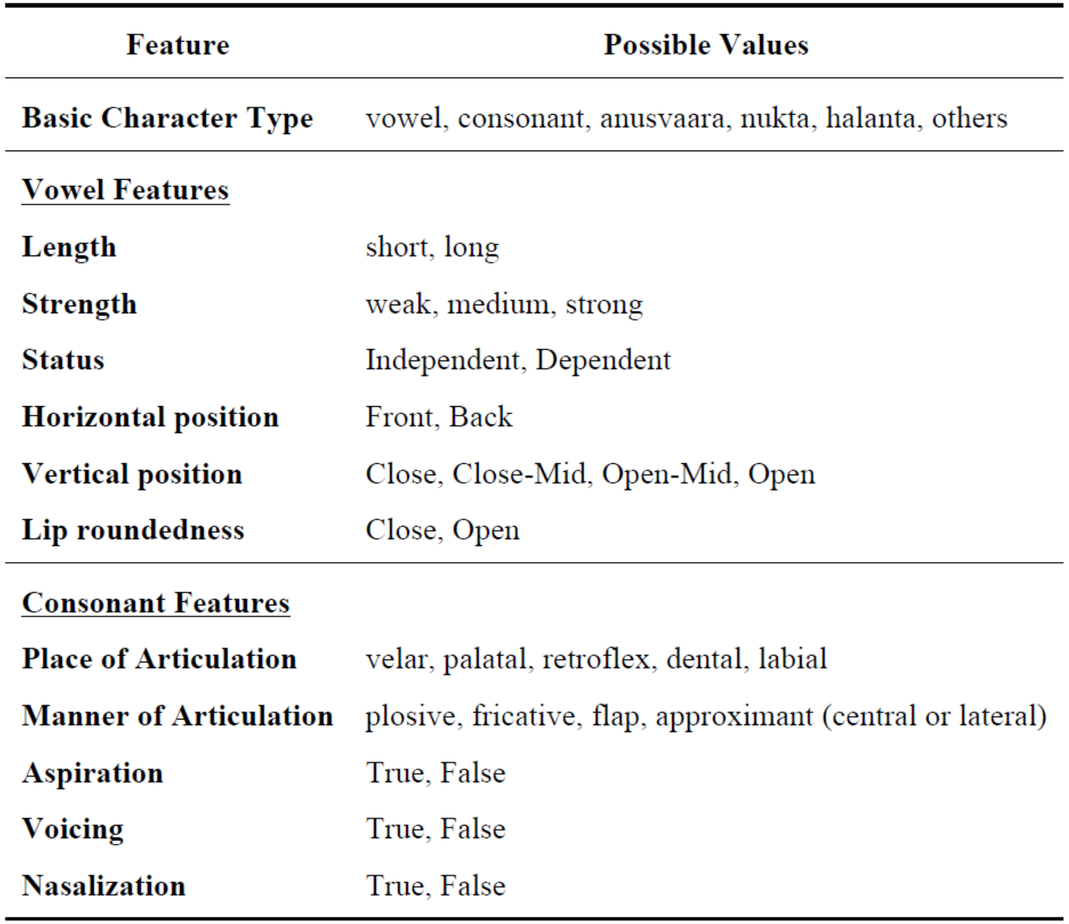
***Share network parameters** across languages*

Output layer for each target language

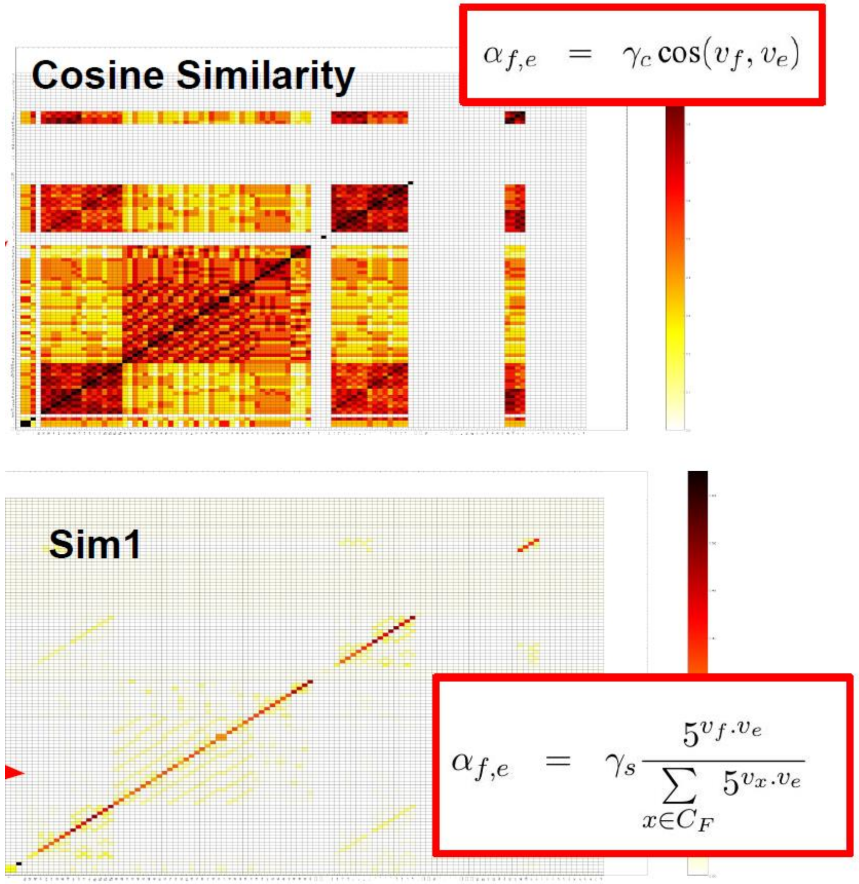


Unsupervised Transliteration

- Monolingual word lists (W_F and W_E)
- Phonetic Representations of words



Feature	Possible Values
Basic Character Type	vowel, consonant, anusvaara, nukta, halanta, others
<u>Vowel Features</u>	
Length	short, long
Strength	weak, medium, strong
Status	Independent, Dependent
Horizontal position	Front, Back
Vertical position	Close, Close-Mid, Open-Mid, Open
Lip roundedness	Close, Open
<u>Consonant Features</u>	
Place of Articulation	velar, palatal, retroflex, dental, labial
Manner of Articulation	plosive, fricative, flap, approximant (central or lateral)
Aspiration	True, False
Voicing	True, False
Nasalization	True, False



$$\alpha_{f,e} = \gamma_c \cos(v_f, v_e)$$

$$\alpha_{f,e} = \gamma_s \frac{5^{v_f.v_e}}{\sum_{x \in C_F} 5^{v_x.v_e}}$$

Use phonetic representation for parameter initialization and as parameter prior

Anoop Kunchukuttan, Pushpak Bhattacharyya, Mitesh Khapra. *Substring-based unsupervised transliteration with phonetic and contextual knowledge*. SIGNLL Conference on Computational Natural Language Learning. 2016.

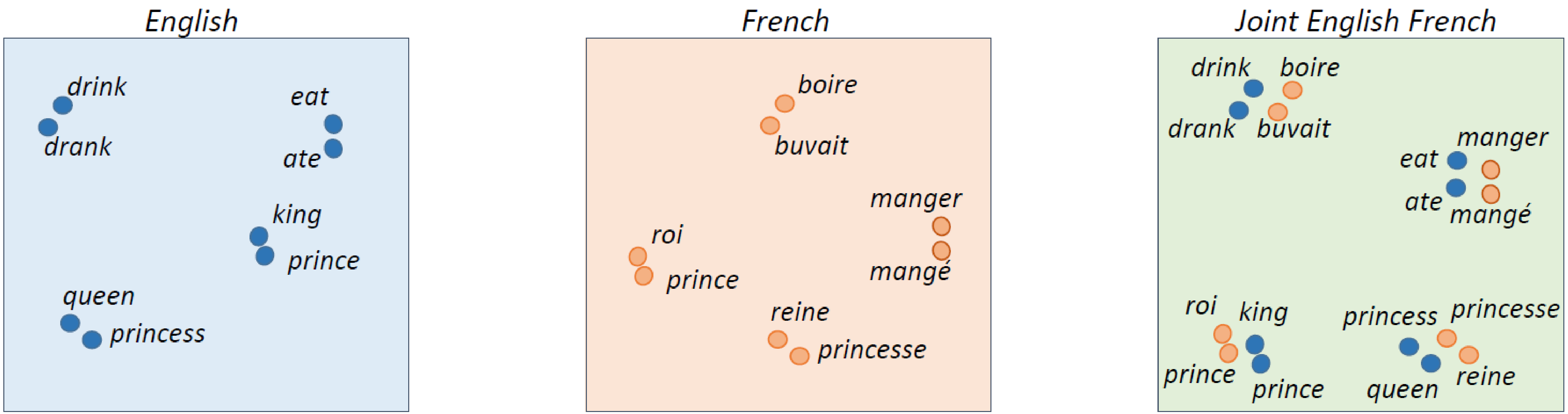
Utilizing Relatedness between Indian Languages

Orthographic Similarity

Lexical Similarity

Syntactic Similarity

Multilingual Word Embeddings



English

French

Joint English French

<u>Monolingual Word Representations</u>

(capture syntactic and semantic similarities between words)

<u>Multilingual Word Representations</u>

(capture syntactic and semantic similarities between words both within and across languages)

(Source: Khapra and Chandar, 2016)

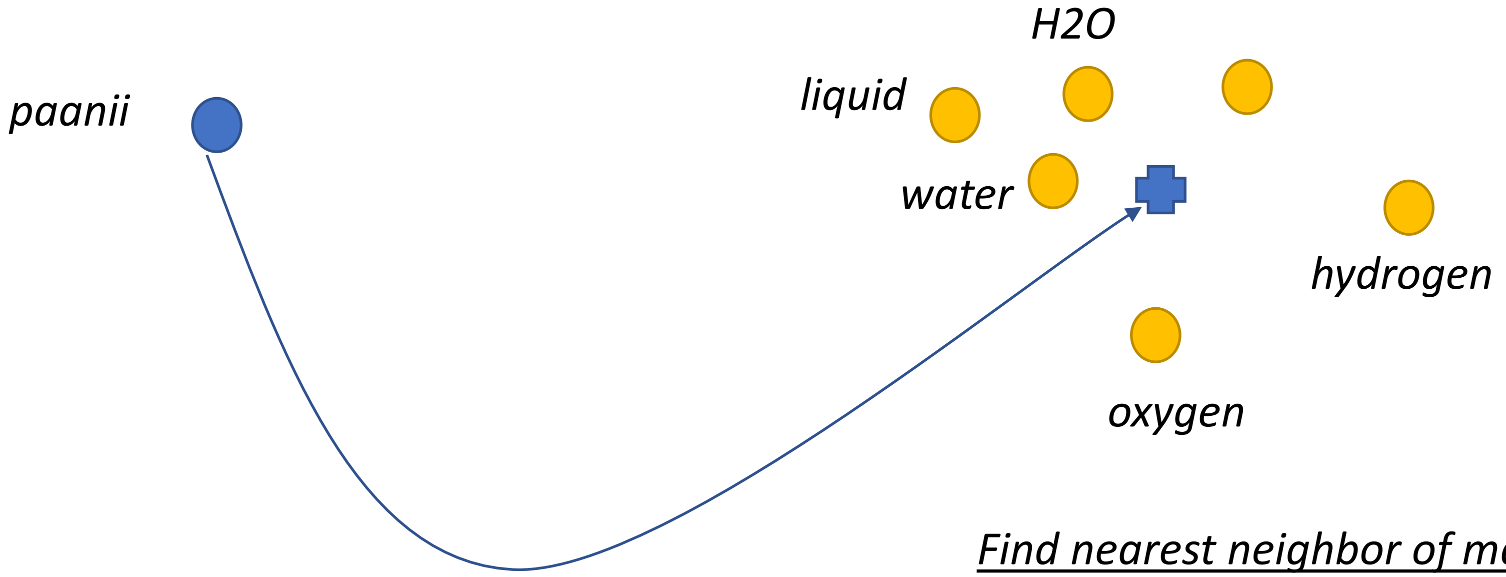
$$embed(y) = f(embed(x))$$

x, y are source and target words
$embed(w)$: embedding for word w

Bilingual Lexicon Induction

Given a mapping function and source/target words and embeddings:

Can we extract a bilingual dictionary?



y'=W(embed(paani))

Find nearest neighbor of mapped embedding

$\max_{y \in Y} \cos(embed(y), y')$ ➔ *water*

A standard intrinsic evaluation task for judging quality of cross-lingual embedding quality

The case of related languages

Concat

- Concat monolingual corpora and train embeddings
- Same words will have same embeddings
- Subword information in both languages considered by FastText

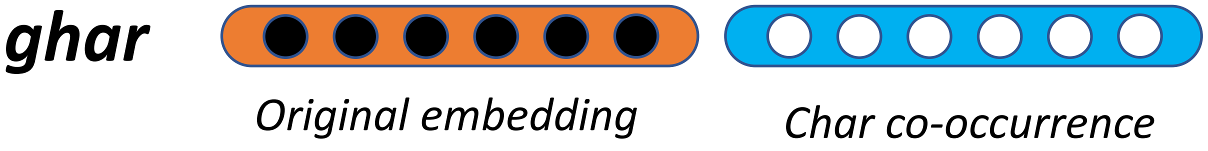
Identity

- For identical words, just assign corresponding embedding for word in other language

embedding(ghar,marathi) = embedding (ghar,hindi)

Enhanced embedding representation

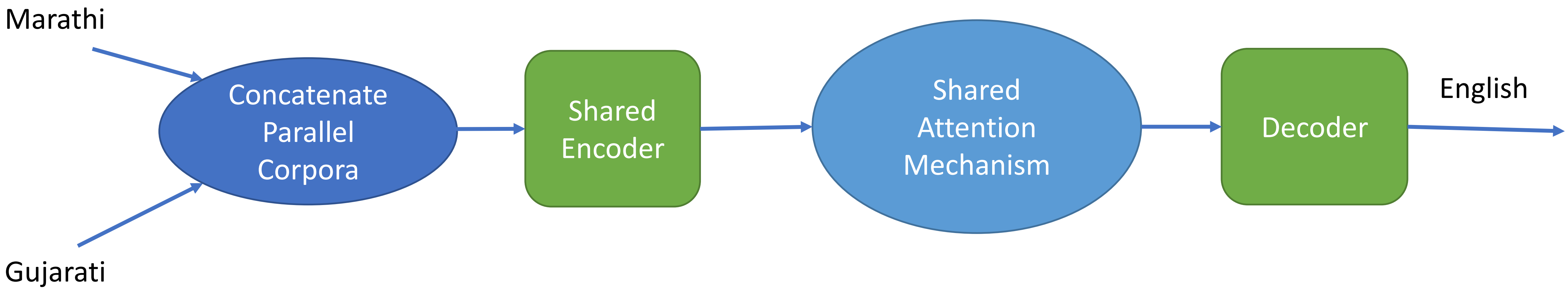
- Add features to monolingual embeddings to capture character occurrence
- Learn bilingual embeddings on these enhanced monolingual embeddings



Multilingual Neural Machine Translation

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

We want Gujarati → English translation → but little parallel corpus is available
We have lot of Marathi → English parallel corpus



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

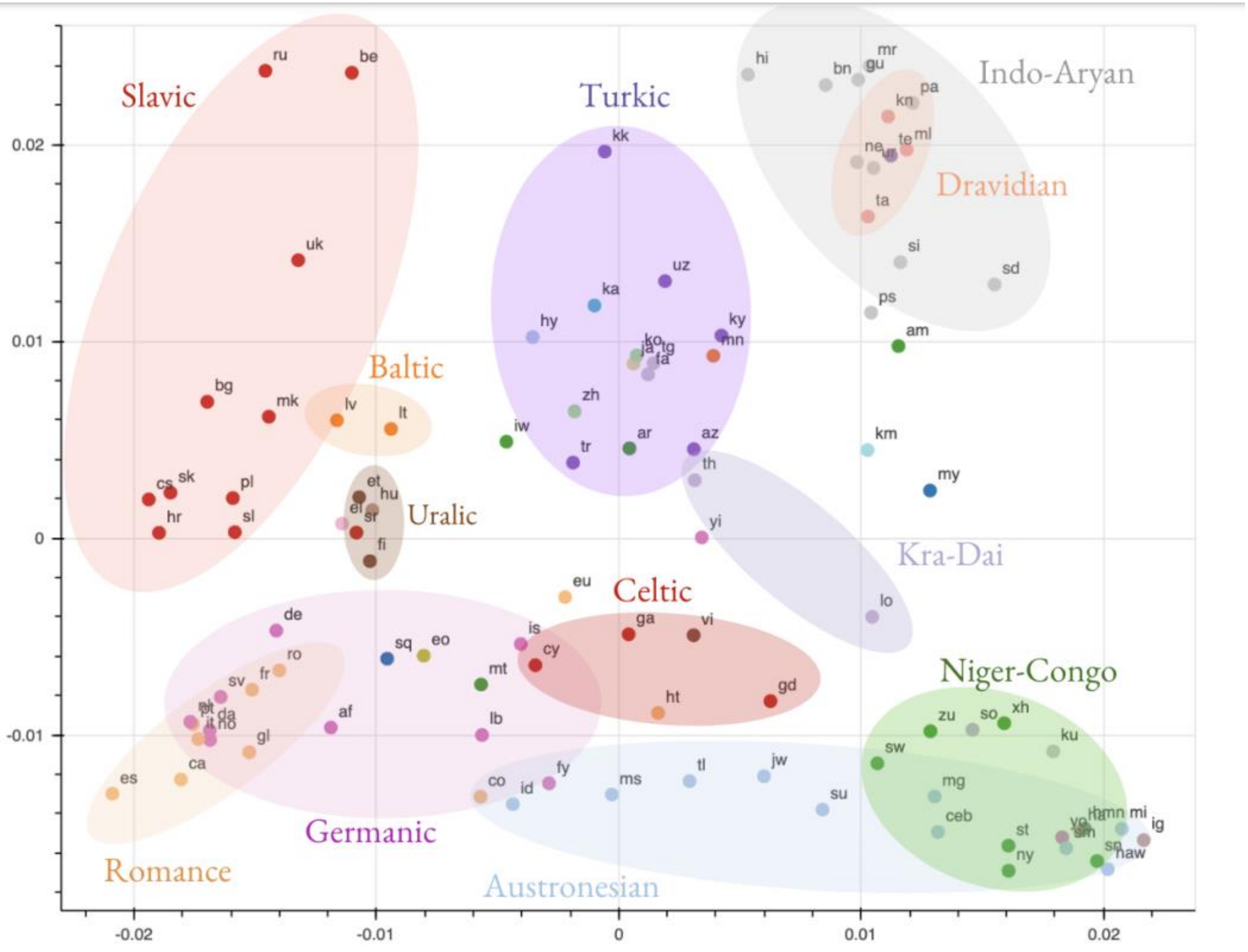
It is cold in Pune	पुण्यात थंड आहे
My home is near the market	माझा घर बाजाराजवळ आहे

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

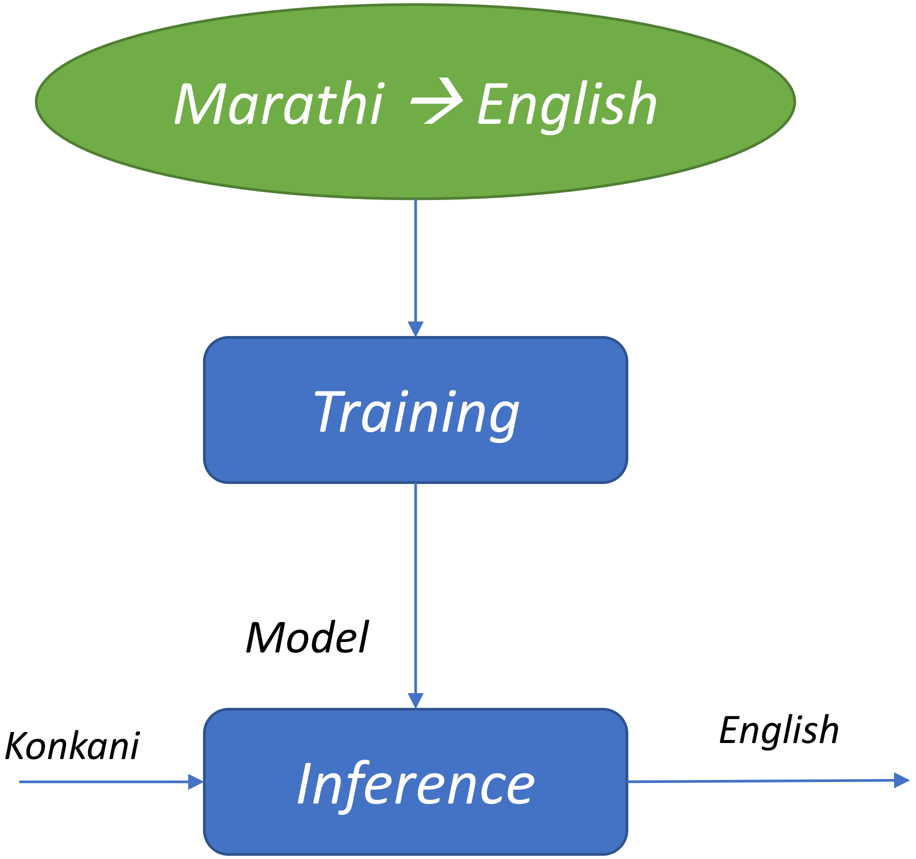
Transfer Learning works best for related languages



Encoder Representations cluster by language family

(Kudungta et al, 2019)

Zeroshot Translation



Marathi → English

Training

Model

Konkani

Inference

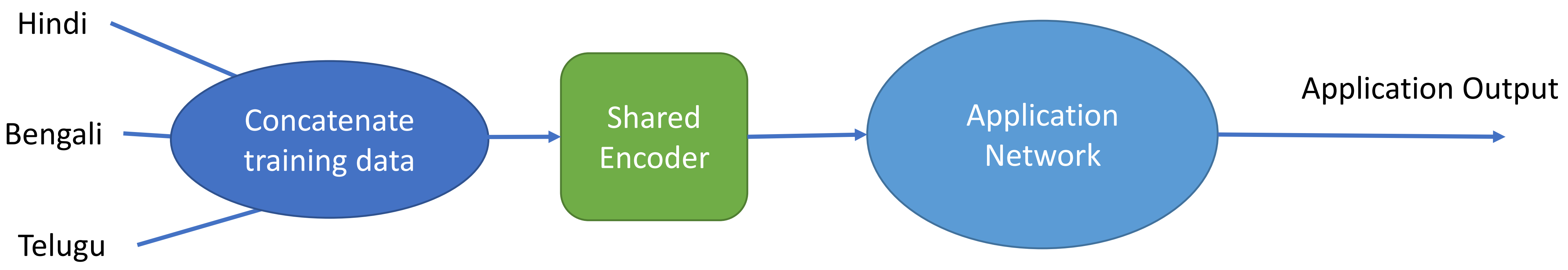
English

Subword-level Representation of Corpora

I am going home	हु घरे जव छू
It rained last week	छे_ ल्ला आठवडि_ या मा वर्सा_ द पाड्यो
It is cold in Pune	पुण्या त थंड आहे
My home is near the market	माझा घर बा_ जारा_ जवळ आहे

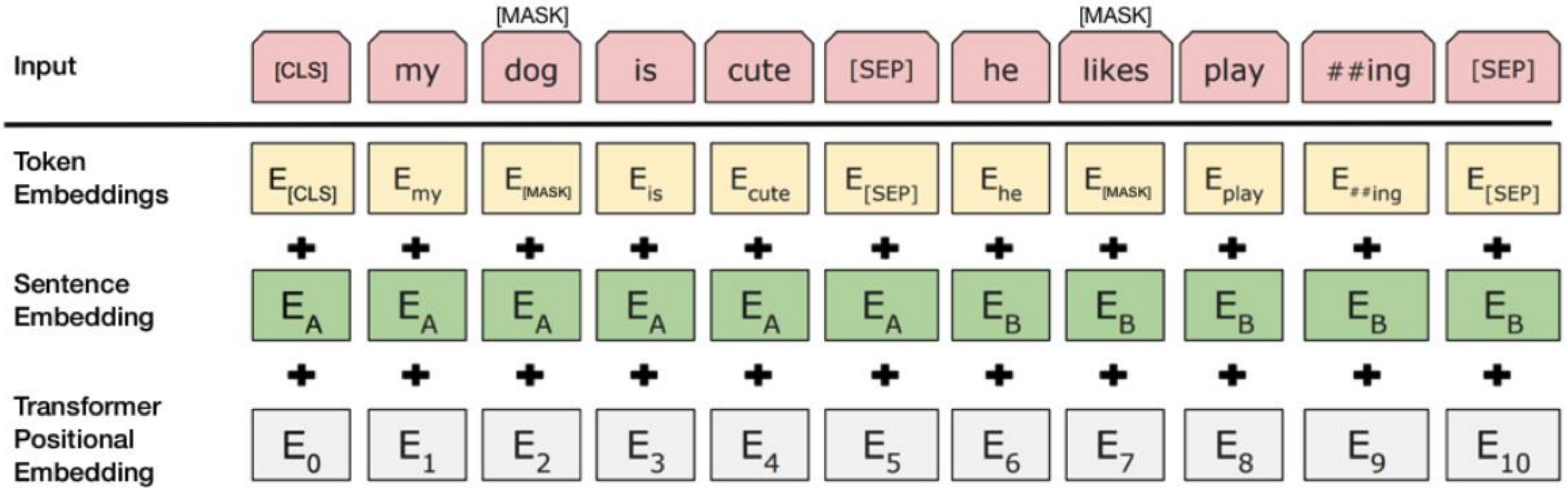
- Words don’t match exactly across languages: Subwords needed to utilize lexical similarity
- Possible Representations: Character, character n-grams, syllables, morph, Byte-Pair Encoded (BPE) Units
- BPE is very popular:
  - unsupervised segmentation, language-independent, identifies frequent substrings

How to make other NLP applications multilingual?



- Sentiment Analysis
- Named Entity Recognition

Multilingual BERT (Devlin et al., 2018)



Transformer encoder with masked LM objective – i.e. try to predict masked words

Concat data from all languages

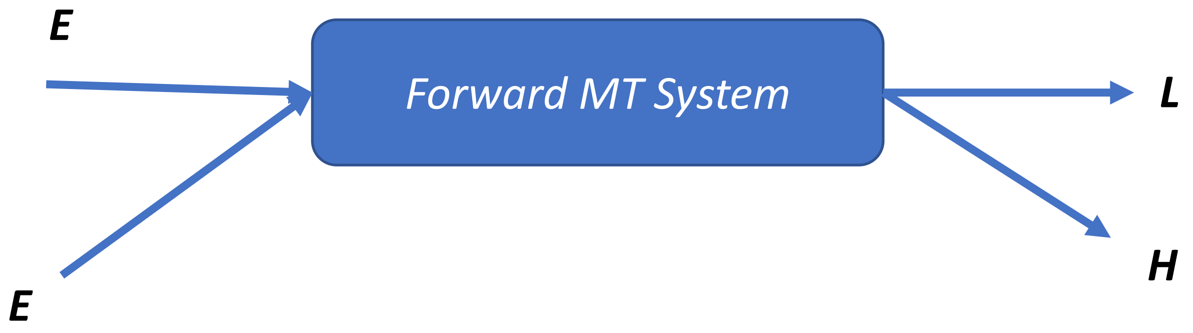
English → Indian Languages

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

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

Original Input: *France and Croatia will play the final on Sunday*

Modified Input: *France and Croatia will play the final on Sunday* *<hin>*



Still an open problem

Utilizing Relatedness between Indian Languages

Orthographic Similarity

Lexical Similarity

Syntactic Similarity

Source reordering for SMT

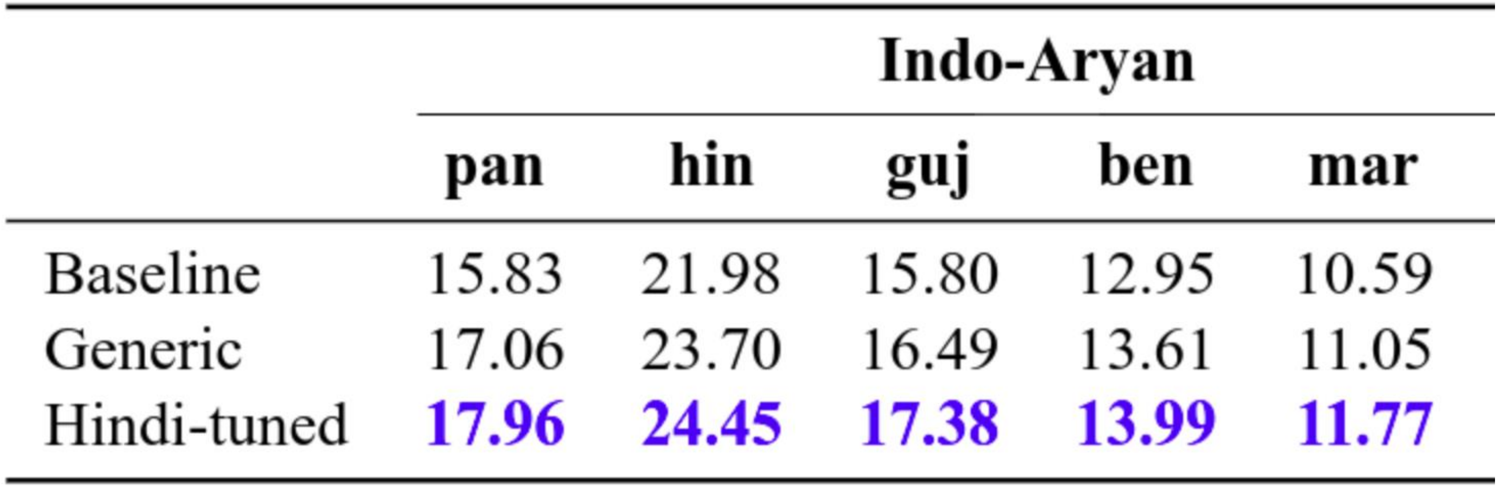
(Kunchukuttan et al., 2014)

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

Bahubali earned more than 1500 crore rupees at the boxoffice

Bahubali the boxoffice at 1500 crore rupees earned

बाहुबली ने बॉक्स ओफिस पर 1500 करोड रुपए कमाए



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

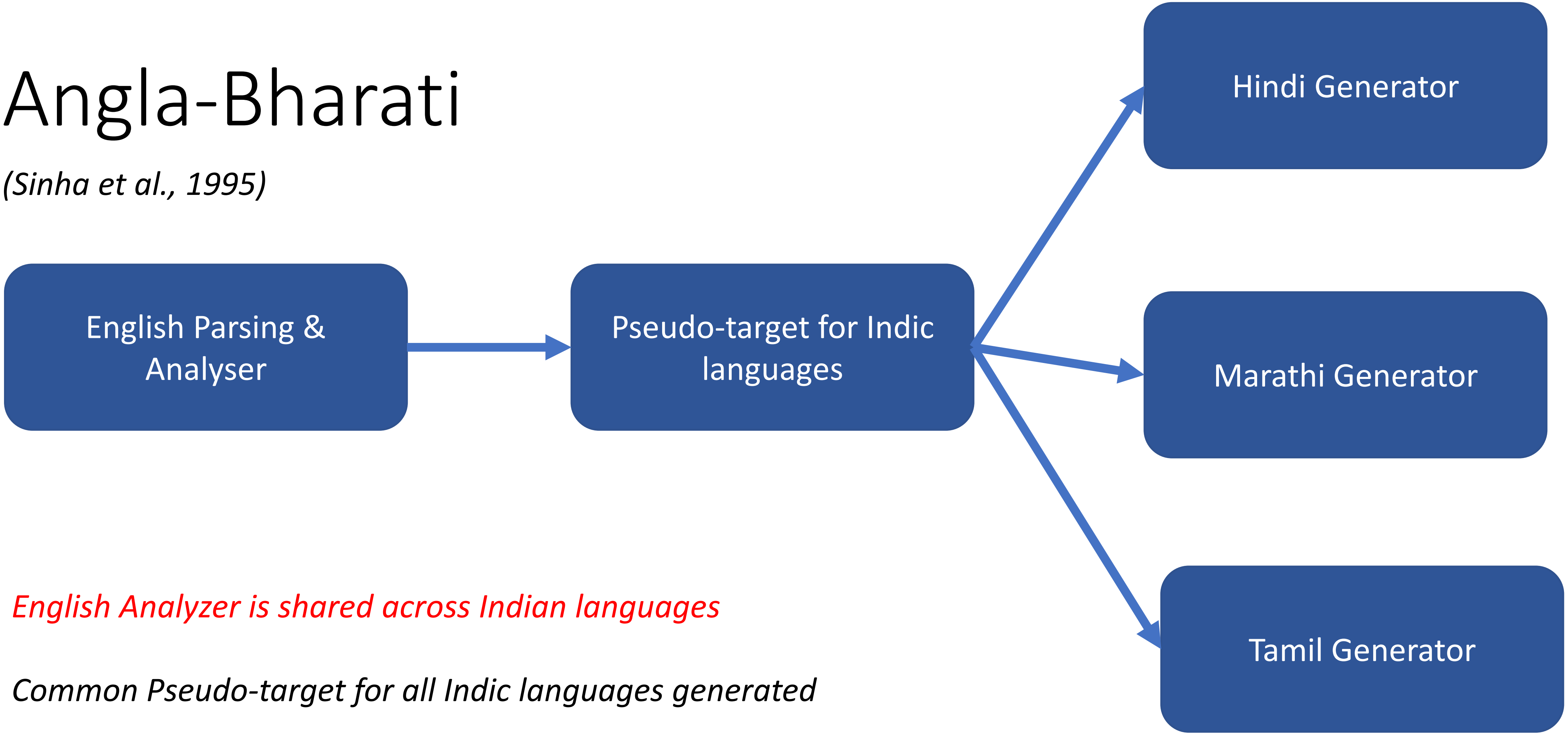
A common set of rules can be written for all Indian languages

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

https://github.com/anoopkunchukuttan/cfilt_preorder

Angla-Bharati

(Sinha et al., 1995)



English Analyzer is shared across Indian languages

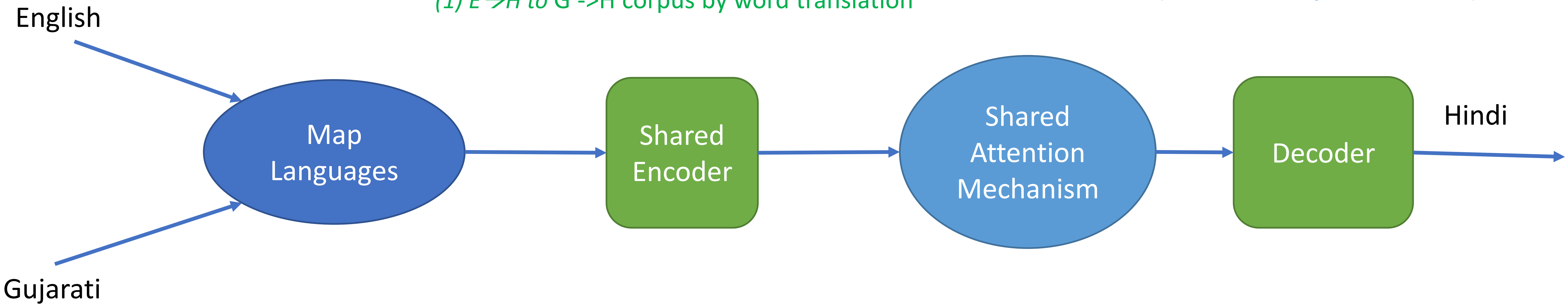
Common Pseudo-target for all Indic languages generated

Can generate specialized pseudo-target for language groups e.g. Indo-Aryan, Dravidian

Bridging Word-order Divergence for low-resource NMT

(Rudramurthy et al., 2019)

(1) E →H to G'->H corpus by word translation



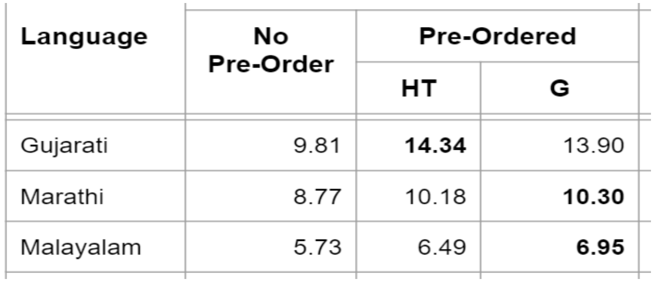
Little G →H corpus

(2) Train with G' → H

(3) Fine-tune with G' → H

Cannot ensure similar Gujarat and English words have similar representations

Solution: *Pre-order English sentence to match Gujarati word-order*



Language	No Pre-Order	Pre-Ordered	
		HT	G
Gujarati	9.81	**14.34**	13.90
Marathi	8.77	10.18	**10.30**
Malayalam	5.73	6.49	**6.95**

Exploiting syntactic similarity in IL-IL translation

Can reduce search choices and errors, improve decoding speed

RMT: No need to handle long-distance reordering.

- Anusaaraka *(Bharati et al. 2003)*
- Sampark *(Antes, 2010)*

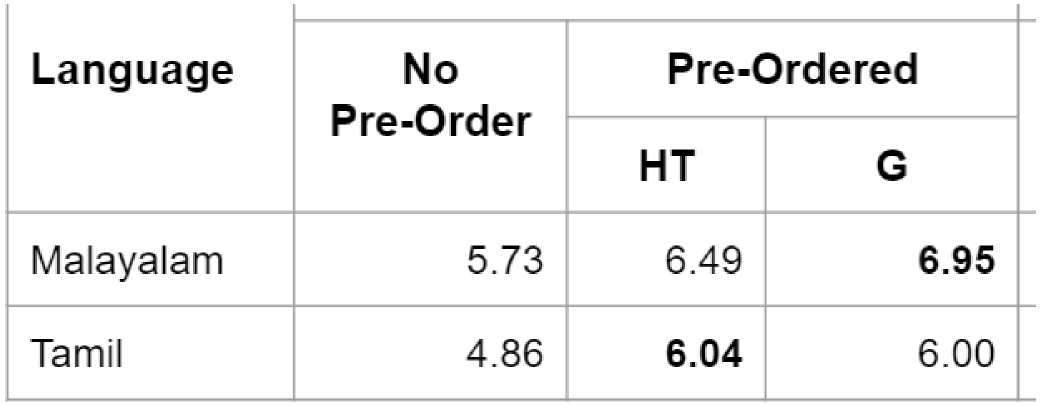
SMT: Monotonic Decoding, subword models.

NMT: Local attention between encoder and decoder. *(Luong et al., 2015)*

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

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

Tamil to English NMT with transfer-leaning using Hindi



Language	No Pre-Order	Pre-Ordered	
		HT	G
Malayalam	5.73	6.49	**6.95**
Tamil	4.86	**6.04**	6.00

Addressing syntactic divergence in NMT using Hindi-driven rules

Outline

- Introduction to Indian Languages
- Opportunities & Challenges in Indic NLP
- Utilizing Relatedness between Indian Languages
- **Getting Started with Indic NLP**
  - **IndicNLP Catalog**
  - IndicNLP Library
  - IndicNLP Suite
- Summary

Indic NLP Catalog

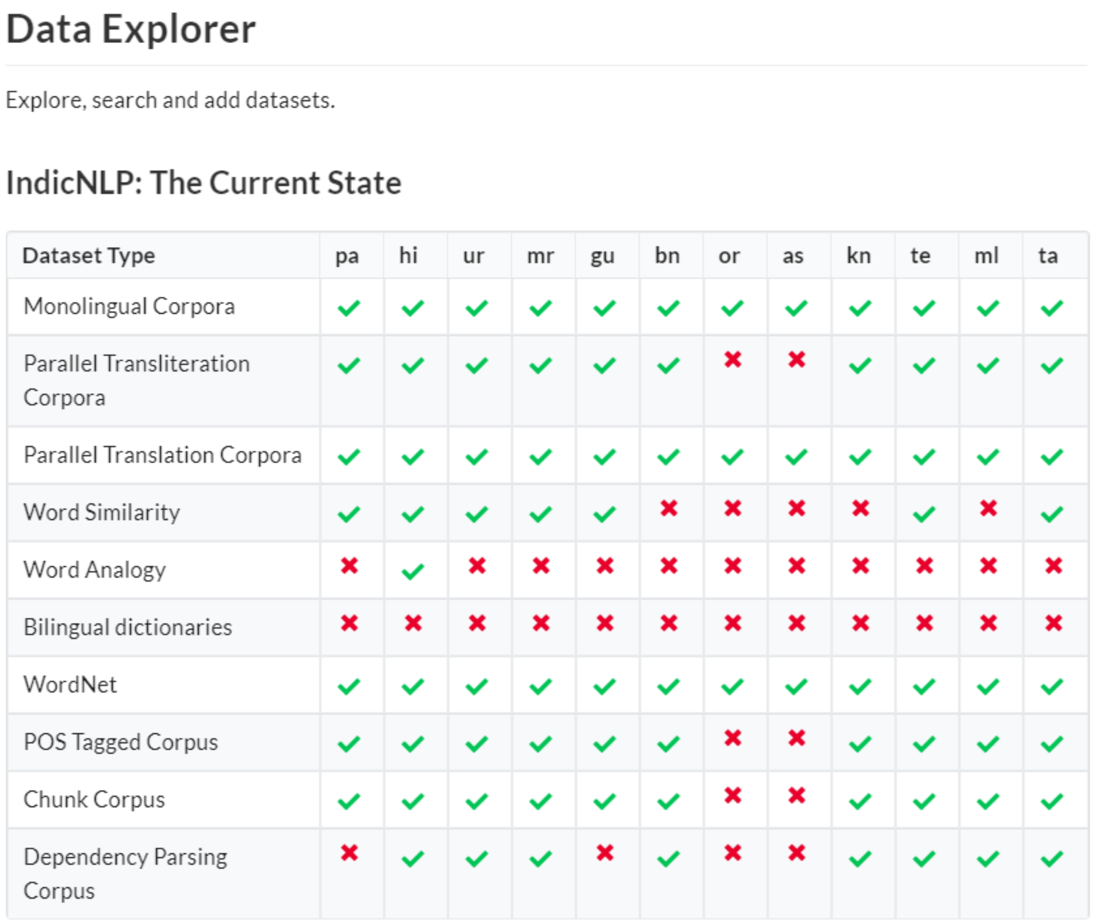
https://github.com/AI4Bharat/indicnlp_catalog

What datasets/libraries exist for Indian languages?

Where can I find these datasets?

What languages are supported?

https://indicnlp.ai4bharat.org/explorer

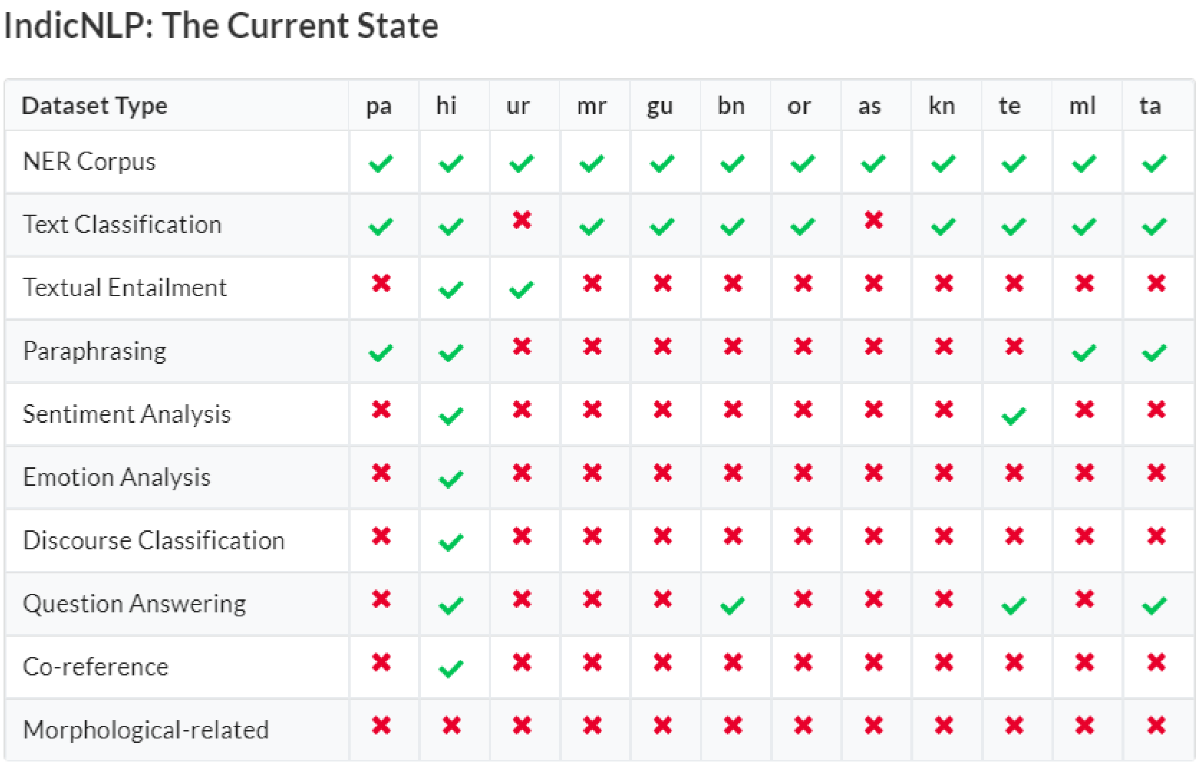


Data Explorer

Explore, search and add datasets.

IndicNLP: The Current State

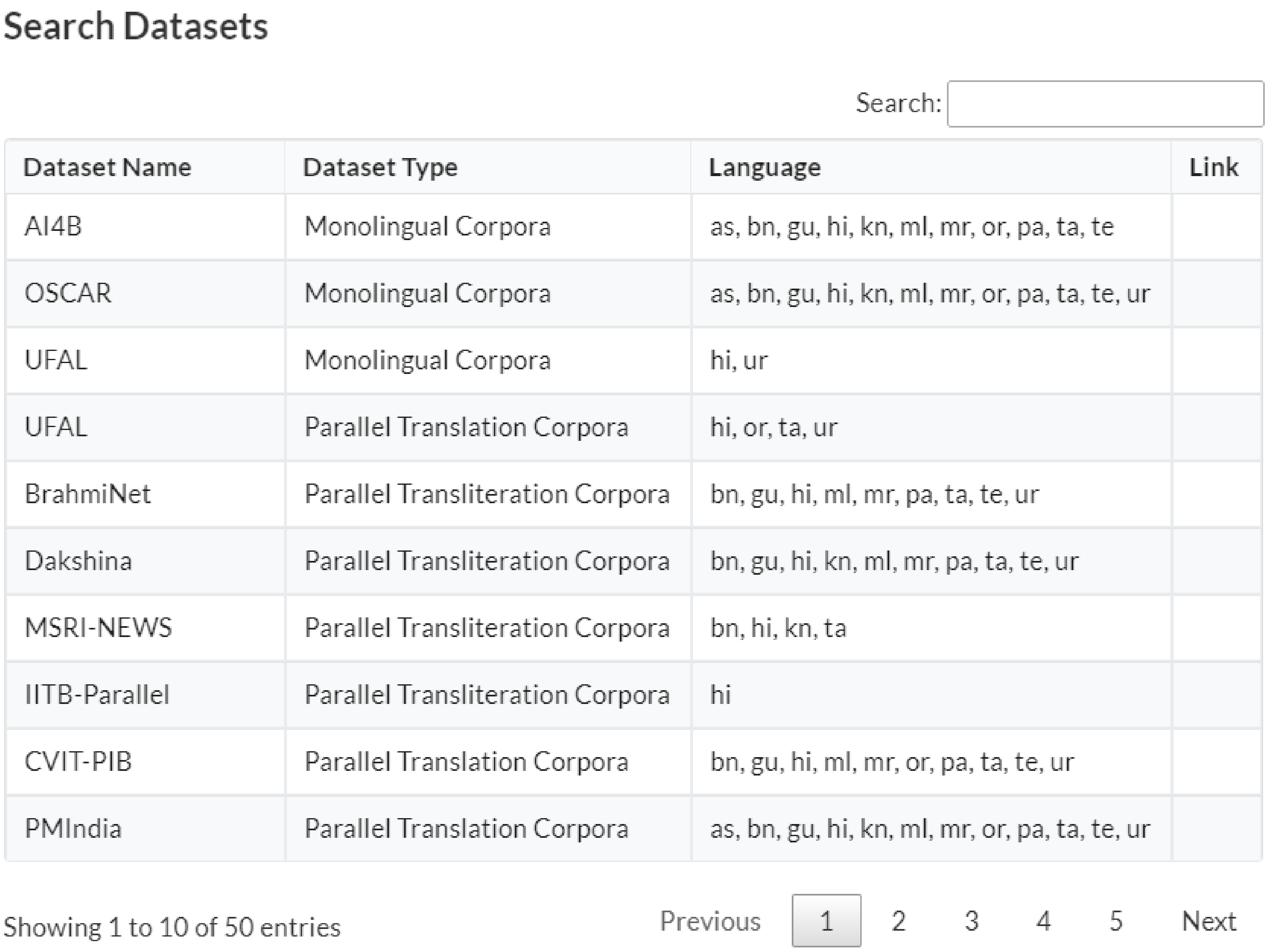
Dataset Type	pa	hi	ur	mr	gu	bn	or	as	kn	te	ml	ta
Monolingual Corpora	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Parallel Transliteration Corpora	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓
Parallel Translation Corpora	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Word Similarity	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✗	✓
Word Analogy	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Bilingual dictionaries	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
WordNet	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
POS Tagged Corpus	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓
Chunk Corpus	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓
Dependency Parsing Corpus	✗	✓	✓	✓	✗	✓	✗	✗	✓	✓	✓	✓



IndicNLP: The Current State

Dataset Type	pa	hi	ur	mr	gu	bn	or	as	kn	te	ml	ta
NER Corpus	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Text Classification	✓	✓	✗	✓	✓	✓	✓	✗	✓	✓	✓	✓
Textual Entailment	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Paraphrasing	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
Sentiment Analysis	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
Emotion Analysis	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Discourse Classification	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Question Answering	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓
Co-reference	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Morphological-related	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

https://indicnlp.ai4bharat.org/explorer/#search-datasets



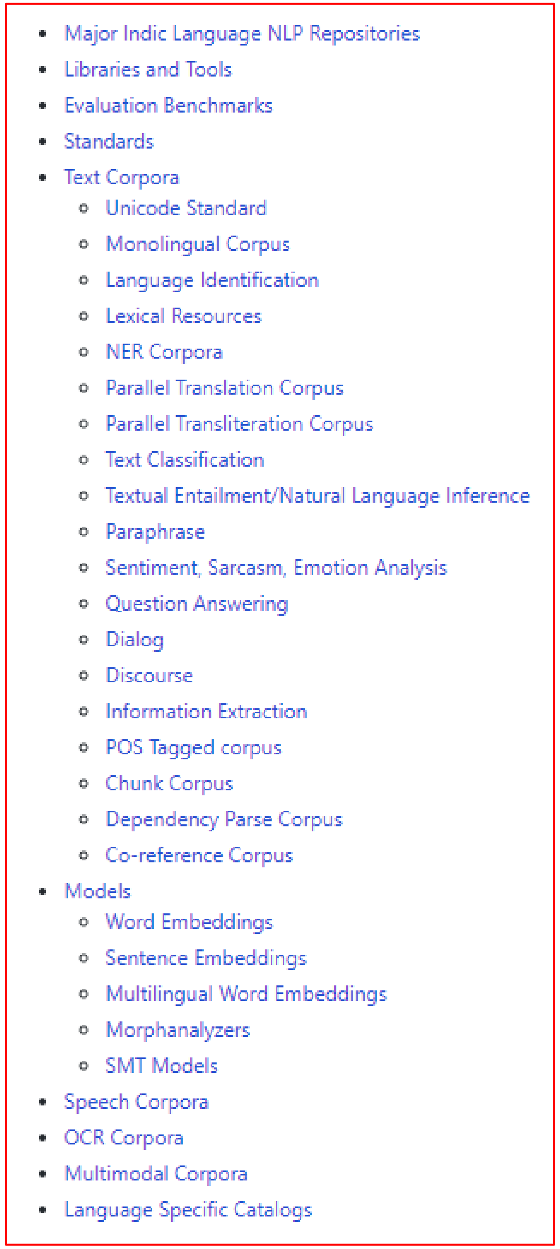
Search Datasets

Dataset Name	Dataset Type	Language	Link
AI4B	Monolingual Corpora	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te	
OSCAR	Monolingual Corpora	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te, ur	
UFAL	Monolingual Corpora	hi, ur	
UFAL	Parallel Translation Corpora	hi, or, ta, ur	
BrahmiNet	Parallel Transliteration Corpora	bn, gu, hi, ml, mr, pa, ta, te, ur	
Dakshina	Parallel Transliteration Corpora	bn, gu, hi, kn, ml, mr, pa, ta, te, ur	
MSRI-NEWS	Parallel Transliteration Corpora	bn, hi, kn, ta	
IITB-Parallel	Parallel Transliteration Corpora	hi	
CVIT-PIB	Parallel Translation Corpora	bn, gu, hi, ml, mr, or, pa, ta, te, ur	
PMIndia	Parallel Translation Corpora	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te, ur	

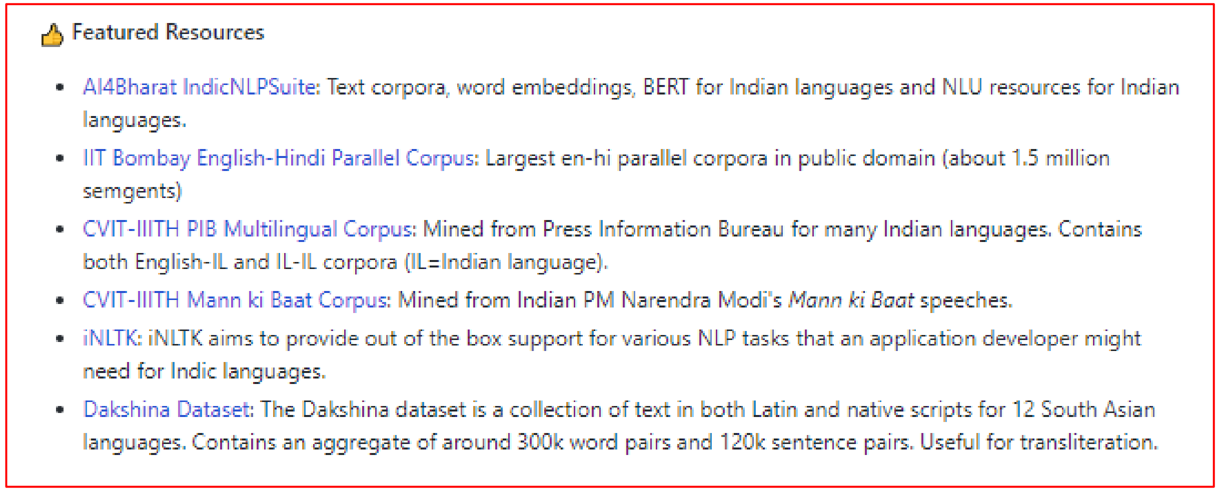
Showing 1 to 10 of 50 entries

The Detailed Catalog

https://github.com/AI4Bharat/indicnlp_catalog

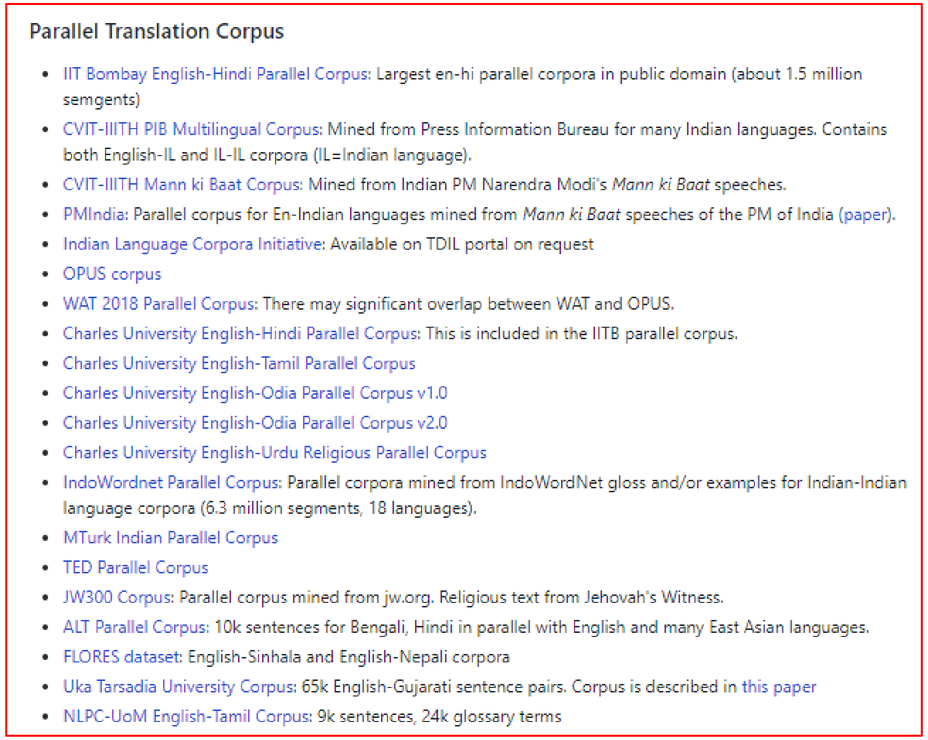


- Major Indic Language NLP Repositories
- Libraries and Tools
- Evaluation Benchmarks
- Standards
- Text Corpora
    - Unicode Standard
    - Monolingual Corpus
    - Language Identification
    - Lexical Resources
    - NER Corpora
    - Parallel Translation Corpus
    - Parallel Transliteration Corpus
    - 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

Evolving, collaborative catalog of Indian language NLP resources

Please add resources you know of and send a pull request

NLP Standards

Important to ensure sharing of data and annotations

Necessary to build multilingual NLP systems

- ***Unicode***: codifies Indic script commonalities
- ***Universal Dependencies****:* universal accepted tagset for many languages
- ***IndoWordNet:*** sense repository for Indian languages
- ***BIS POS Tag Set***: hierarchical tagset suitable for Indian languages

Outline

- Introduction to Indian Languages
- Opportunities & Challenges in Indic NLP
- Utilizing Relatedness between Indian Languages
- Getting Started with Indic NLP
  - IndicNLP Catalog
  - **IndicNLP Library**
  - IndicNLP Suite
- Summary

Indic NLP Library

https://github.com/anoopkunchukuttan/indic_nlp_library

- Utilize similarity between Indian languages for scaling to multiple Indian languages
- Design to support maximum number of Indian languages
- Modular and Extensible
- Easy of use:
  - Installation `pip install indic-nlp-library`
  - Consistent Use
  - Separation between code and data resources

Anoop Kunchukuttan. *The IndicNLP Library*. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf .2020.

Capabilities

Text Processing

- Text Normalizer
- Sentence Splitter
- Word Tokenizer
- Word Detokenizer

Word Segmentation

- Morphological Segmentation
- Syllabification

Script Processing

- Query Script Information
- Script Converter
- Romanization
- Indicization
- Acronym Transliterator
- Phonetic Similarity
- Lexical Similarity

Samples: https://nbviewer.jupyter.org/url/anoopkunchukuttan.github.io/indic_nlp_library/doc/indic_nlp_examples.ipynb

Language Support

Indo-Aryan			Dravidian
Assamese (as)	Marathi (mr)	Sindhi (sd)	Kannada (kn)
Bengali (bn)	Nepali (ne)	Sinhala (si)	Malayalam (ml)
Gujarati (gu)	Odia (or)	Sanskrit (sa)	Telugu (te)
Hindi (hi)	Punjabi (pa)	Konkani (kok/kK)	Tamil (ta)

	as	bn	gu	hi	mr	ne	or	pa	sd	si	sa	kok	kn	ml	te	ta
Text Processing	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
Morphological Segmentation	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓
Syllabification	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
Script Processing	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓

Working with Indian Language Text

- Use UTF-8 encoding
- Normalize Text
- For debugging:
  - Convert to some romanization script like ITRANS
  - Convert to some script you understand

Outline

- Introduction to Indian Languages
- Opportunities & Challenges in Indic NLP
- Utilizing Relatedness between Indian Languages
- Getting Started with Indic NLP
  - IndicNLP Catalog
  - IndicNLP Library
  - **IndicNLP Suite**
- Summary

Indic NLP Suite

https://indicnlp.ai4bharat.org

AI4Bharat indicnlp

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

Building Blocks for large-scale Indic NLP

Wide Coverage of Indian Languages

- 11 Indian languages and Indian English
- Indo-Aryan: Hindi, Punjabi, Gujarati, Bengali, Oriya, Assamese, Marathi
- Dravidian: Kannada, Telugu, Malayalam, Tamil

IndicCorp

Large-scale Monolingual corpora (8.8 billion tokens, 452 million sentences)

IndicFT

Pre-trained FastText-based word embeddings

IndicBERT

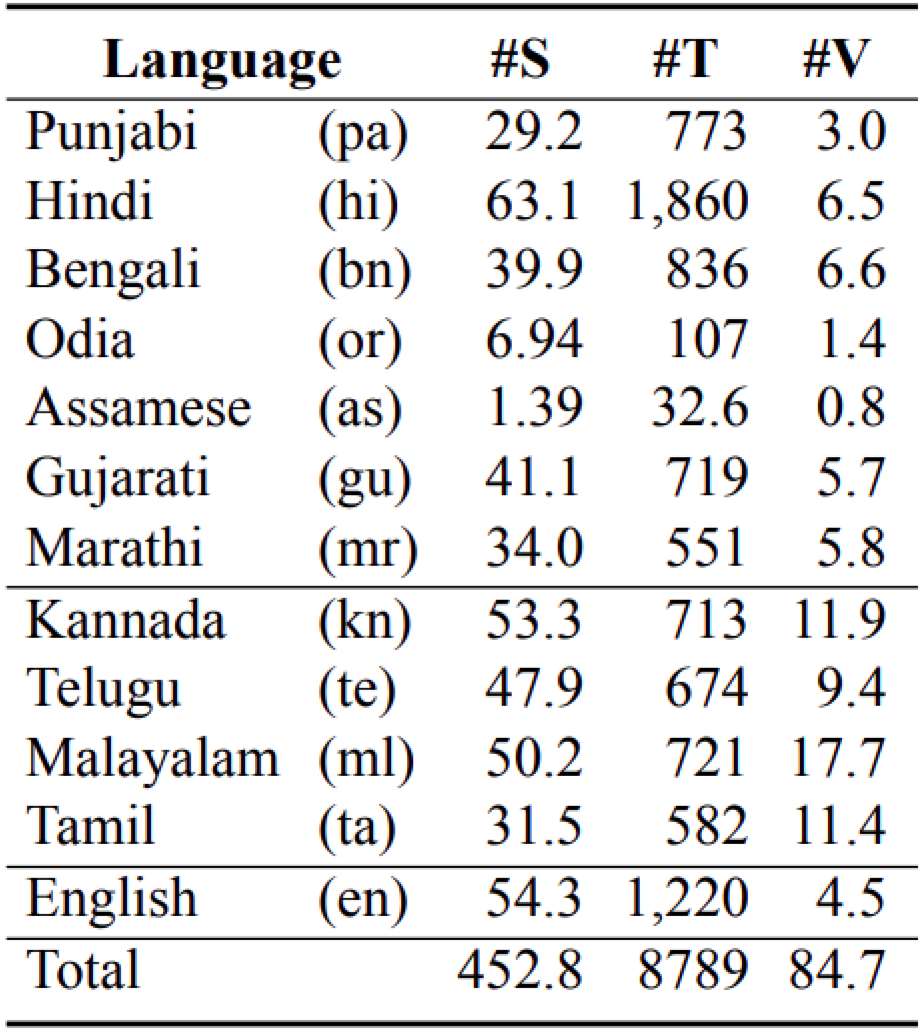
Pre-trained Transformer Language Model

IndicGLUE

NLU Evaluation benchmarks spanning many tasks

IndicCorp

https://indicnlp.ai4bharat.org/corpora



Language		#S	#T	#V
Punjabi	(pa)	29.2	773	3.0
Hindi	(hi)	63.1	1,860	6.5
Bengali	(bn)	39.9	836	6.6
Odia	(or)	6.94	107	1.4
Assamese	(as)	1.39	32.6	0.8
Gujarati	(gu)	41.1	719	5.7
Marathi	(mr)	34.0	551	5.8
Kannada	(kn)	53.3	713	11.9
Telugu	(te)	47.9	674	9.4
Malayalam	(ml)	50.2	721	17.7
Tamil	(ta)	31.5	582	11.4
English	(en)	54.3	1,220	4.5
Total		452.8	8789	84.7

- **500 million words for almost all languages**
  - Please suggest Odia sources!
- **Largest text corpus for Indian languages**
  - 47 times OSCAR corpus
  - 2x times CC100 corpus
- **English data sourced from Indian sources**
  - Representative data important for NLP
  - Named entities, topics are more relevant to Indian context
  - Easier alignment with Indic language corpora
- ***Covers news articles, magazines, blog posts, etc.***

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

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

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, 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

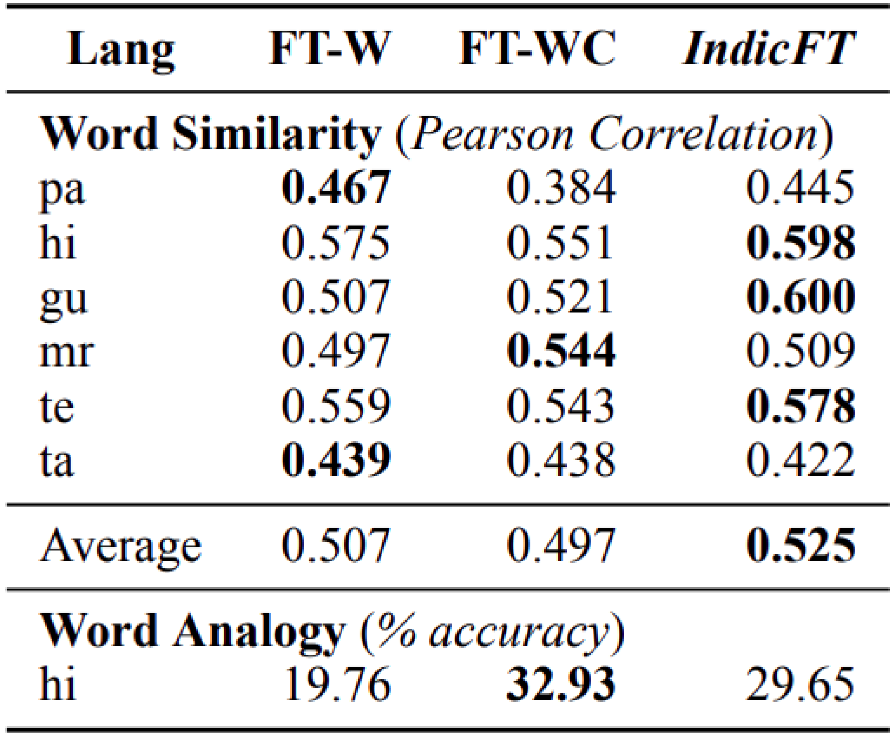
Task Type	Task	N	Languages
Classification	News Article Classification	10	bn, gu, hi, kn, ml, mr, or, pa, ta, te
	Headline Classification	4	gu, ml, mr, ta
	Sentiment Analysis	2	hi, te
	Discourse Mode Classification	1	hi
Diagnostics	Winograd Natural Language Inference	3	gu, hi, mr
	Choice of Plausible 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

Need to add more challenging tasks, cover more languages

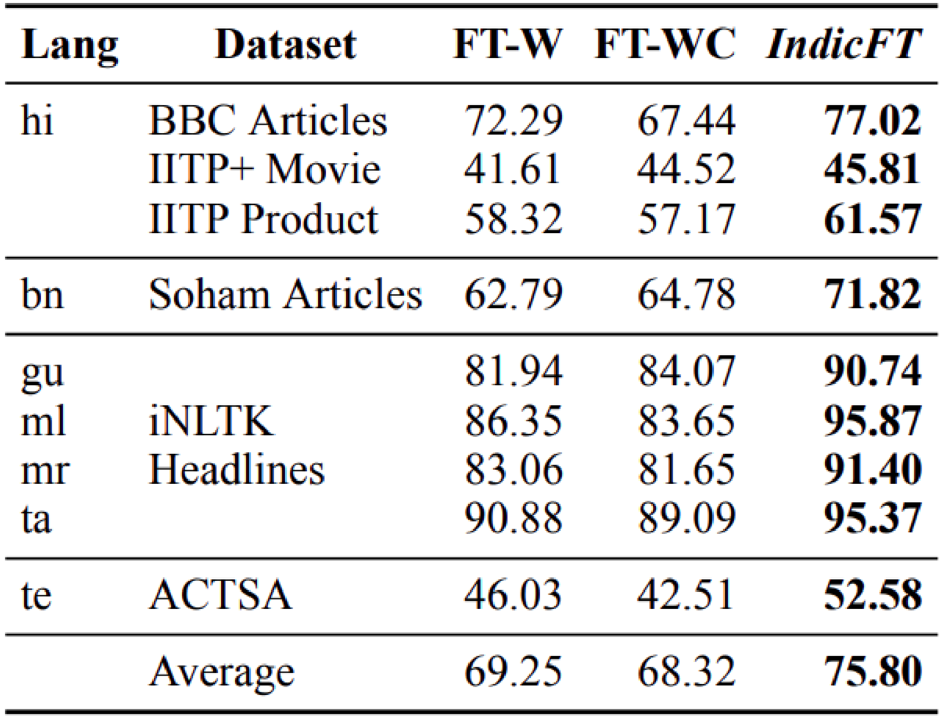
IndicFT

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 language model exclusively for Indian languages
- English supported, trained with Indian English content
- Multilingual model
- Compact Model
  - Based on the ALBERT model (a lightweight version of BERT)
  - Smaller number of parameters (10x fewer params compared to mBERT, XLM-R)
- Competitive/better than mBERT/XLM-R
- Simplify fine-tune for your application on Collab or simple GPU for a small time

Outline

- Introduction to Indian Languages
- Opportunities & Challenges in Indic NLP
- Utilizing Relatedness between Indian Languages
- Getting Started with Indic NLP
  - IndicNLP Catalog
  - IndicNLP Library
  - IndicNLP Suite
- **Summary**

Summary

- Utilizing language relatedness is important to scale NLP technologies to a large number of Indian languages.
- The orthographic similarity of Indian languages is a strong starting point for utilizing language relatedness.
- Contact as well as genetic relatedness are useful in the context of Indian languages.
- Multilingual pre-trained models trained on large corpora needed for transfer learning in NLU and NLG tasks.
- Efficient training and inference needed to experiment with more models that utilize language relatedness.

Thank You!

anoop.kunchukuttan@gmail.com

http://anoopk.in

References

1. Bharati, A., Chaitanya, V., Kulkarni, A. P., Sangal, R., & Rao, G. U. (2003). ANUSAARAKA: overcoming the language barrier in India. arXiv preprint cs/0308018.
2. Anthes, G. (2010). Automated translation of indian languages. Communications of the ACM, 53(1), 24-26.
3. Atreya, A., Chaudhari, S., Bhattacharyya, P., and Ramakrishnan, G. (2016). Value the vowels: Optimal transliteration unit selection for machine. In Unpublished, private communication with authors.
4. Basil Abraham, S Umesh and Neethu Mariam Joy. "Overcoming Data Sparsity in Acoustic Modeling of Low-Resource Language by Borrowing Data and Model Parameters from High-Resource Languages”, Interspeech, 2016.
5. Basil Abraham, Neethu Mariam Joy, Navneeth K and S Umesh. "A data-driven phoneme mapping technique using interpolation vectors of phone-cluster adaptive training." Spoken Language Technology Workshop (SLT), 2014.
6. Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In Annual meeting on Association for Computational Linguistics.
7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.
8. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
9. Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In Annual Meeting of the Association for Computational Linguistics.
10. Durrani, N., Sajjad, H., Fraser, A., and Schmid, H. (2010). Hindi-to-urdu machine translation through transliteration. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.
11. Emeneau, M. B. (1956). India as a Lingustic area. Language.
16. Firat, O., Cho, K., and Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In Conference of the North American Chapter of the Association for Computational Linguistics.
17. Jha, G. N. (2012). The TDIL program and the Indian Language Corpora Initiative. In Language Resources and Evaluation Conference.
18. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2016). Google’s multilingual neural machine translation system: Enabling zero-shot translation. arXiv preprint arXiv:1611.04558.
19. Kudugunta, S. R., Bapna, A., Caswell, I., Arivazhagan, N., & Firat, O. (2019). Investigating multilingual nmt representations at scale. arXiv preprint arXiv:1909.02197.
20. Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, Pratyush Kumar. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. arXiv preprint arXiv:2005.00085. 2020.
21. Anoop Kunchukuttan, Pushpak Bhattachyya. Utilizing Language Relatedness to improve Machine Translation: A Case Study on Languages of the Indian Subcontinent. arXiv preprint arXiv:2003.08925. 2020.

22. Rudramurthy V, Anoop Kunchukuttan, Pushpak Bhattacharyya. Addressing word-order Divergence in Multilingual Neural Machine Translation for extremely Low Resource Languages. NAACL. 2019.
23. Anoop Kunchukuttan, Mitesh Khapra, Gurneet Singh, Pushpak Bhattacharyya. *Leveraging Orthographic Similarity for Neural Machine Transliteration*. Transactions of the Association for Computational Linguistics**.** 2018
24. Anoop Kunchukuttan, Maulik Shah, Pradyot Prakash, Pushpak Bhattacharyya. *Utilizing Lexical Similarity between related, low resource languages for Pivot based SMT*. International Joint Conference on Natural Language Processing. 2017.
25. Anoop Kunchukuttan, Pushpak Bhattacharyya*. Learning variable length units for SMT between related languages via Byte Pair Encoding*. 1st Workshop on Subword and Character level models in NLP (SCLeM, collocated with EMNLP). 2017.
26. Anoop Kunchukuttan, Pushpak Bhattacharyya. *Orthographic Syllable as basic unit for SMT between Related Languages*. Conference on Empirical Methods in Natural Language Processing. 2016.
27. Anoop Kunchukuttan, Pushpak Bhattacharyya, Mitesh Khapra. *Substring-based unsupervised transliteration with phonetic and contextual knowledge*. SIGNLL Conference on Computational Natural Language Learning. 2016.
28. Anoop Kunchukuttan, Ratish Puduppully , Pushpak Bhattacharyya, *Brahmi-Net: A transliteration and script conversion system for languages of the Indian subcontinent* , Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies: System Demonstrations . 2015.
29. Rohit More, Anoop Kunchukuttan, Raj Dabre, Pushpak Bhattacharyya. *Augmenting Pivot based SMT with word segmentation*. International Conference on Natural Language Processing **(ICON 2015)**. 2015.
30. Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, Pushpak Bhattacharyya. *Shata-Anuvadak: Tackling Multiway Translation of Indian Languages* . Language and Resources and Evaluation Conference **(LREC 2014)**. 2014.
31. Kondrak, G. (2001). *Identifying cognates by phonetic and semantic similarity*. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (pp. 1-8). Association for Computational Linguistics.
32. Lee, J., Cho, K., and Hofmann, T. (2017). Fully Character-Level Neural Machine Translation without Explicit Segmentation. Transactions of the Association for Computational Linguistics.
33. Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.
34. Melamed, I. D. (1995). Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In Third Workshop on Very Large Corpora.

35. Nakov, P. and Tiedemann, J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2.
36. Nguyen, T. Q., & Chiang, D. (2017). Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. IJCNLP.
37. Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., & Ji, H. (2017, July). Cross-lingual name tagging and linking for 282 languages. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1946-1958).
38. Patel, R., Gupta, R., Pimpale, P., and Sasikumar, M. (2013). Reordering rules for English-Hindi SMT. In Proceedings of the Second Workshop on Hybrid Approaches to Translation.
39. Pourdamghani, N. and Knight, K. (2005). Deciphering related languages. In Empirical Methods in Natural Language Processing.
40. Ramanathan, A., Hegde, J., Shah, R., Bhattacharyya, P., and Sasikumar, M. (2008). Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In International Joint Conference on Natural Language Processing.
41. Ravi, S. and Knight, K. (2009). Learning phoneme mappings for transliteration without parallel data. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.
42. Rudramurthy, V., Khapra, M., Bhattacharyya, P., et al. (2016). Sharing network parameters for crosslingual named entity recognition. arXiv preprint arXiv:1607.00198.
43. Saha, A., Khapra, M. M., Chandar, S., Rajendran, J., and Cho, K. (2016). A correlational encoder decoder architecture for pivot based sequence generation.
44. Samudravijaya, Hema Murth. (2012). Indian Language Speech sound Label set. https://www.iitm.ac.in/donlab/tts/downloads/cls/cls_v2.1.6.pdf
45. Tanja Schultz and Alex Waibel. Experiments on cross-language acoustic modeling. In INTERSPEECH, pages 2721-2724, 2001.
46. Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, C.V. Jawahar (2020).A Multilingual Parallel Corpora Collection Effort for Indian Languages. LREC.
47. Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In ACL.
48. Sherif, T. and Kondrak, G. (2007). Substring-based transliteration. In Annual Meeting Association for Computational Linguistics.
49. Sinha, R. M. K., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R., & Jain, A. (1995, October). ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. In 1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century (Vol. 2, pp. 1609-1614). IEEE.
50. Ortiz Suárez, P. J., Sagot, B., & Romary, L. (2019). *Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures*.

51. Subbārāo, K. V. (2012). South Asian languages: A syntactic typology. Cambridge University Press.

52. Tao, T., Yoon, S.-Y., Fister, A., Sproat, R., and Zhai, C. (2006). Unsupervised named entity transliteration using temporal and phonetic correlation. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.

53. Tiedemann, J. (2009a). Character-based PBSMT for closely related languages. In Proceedings of the 13th Conference of the European Association for Machine Translation (EAMT 2009).

54. Trubetzkoy, N. (1928). Proposition 16. In Actes du premier congres international des linguistes à La Haye.

55. Vilar, D., Peter, J.-T., and Ney, H. (2007). Can we translate letters? In Proceedings of the Second Workshop on Statistical Machine Translation.

56. Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. EMNLP.