Indic NLP: A Multilinguality and Language Relatedness Perspective

Anoop Kunchukuttan

Machine Translation Group, Microsoft, Hyderabad

ankunchu@microsoft.com

Microsoft

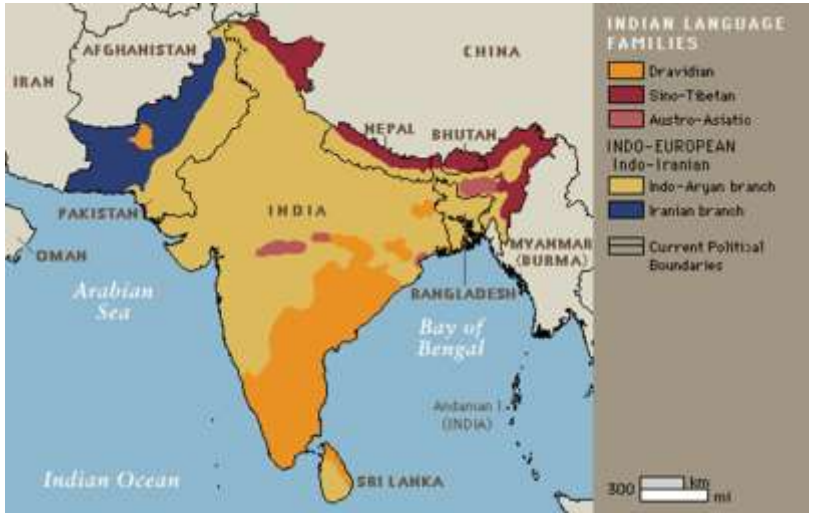
Joint work with

AI4Bharat

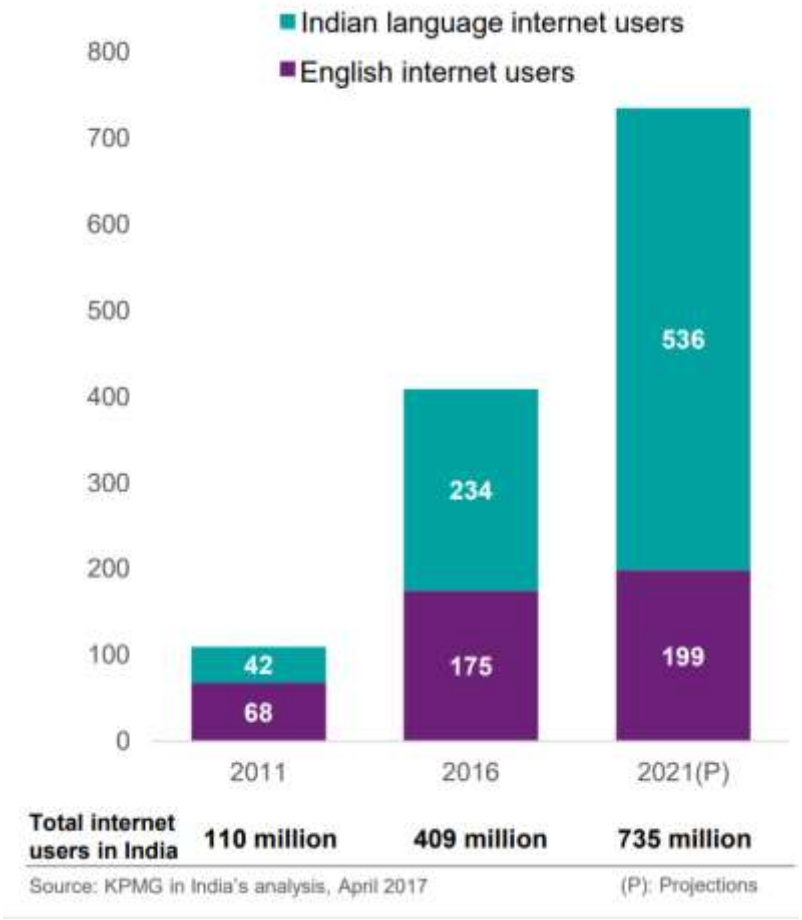
CFILT

Vaibhav Summit, NLP Session
19th October 2020

Multi-linguality – A matter of fact in India



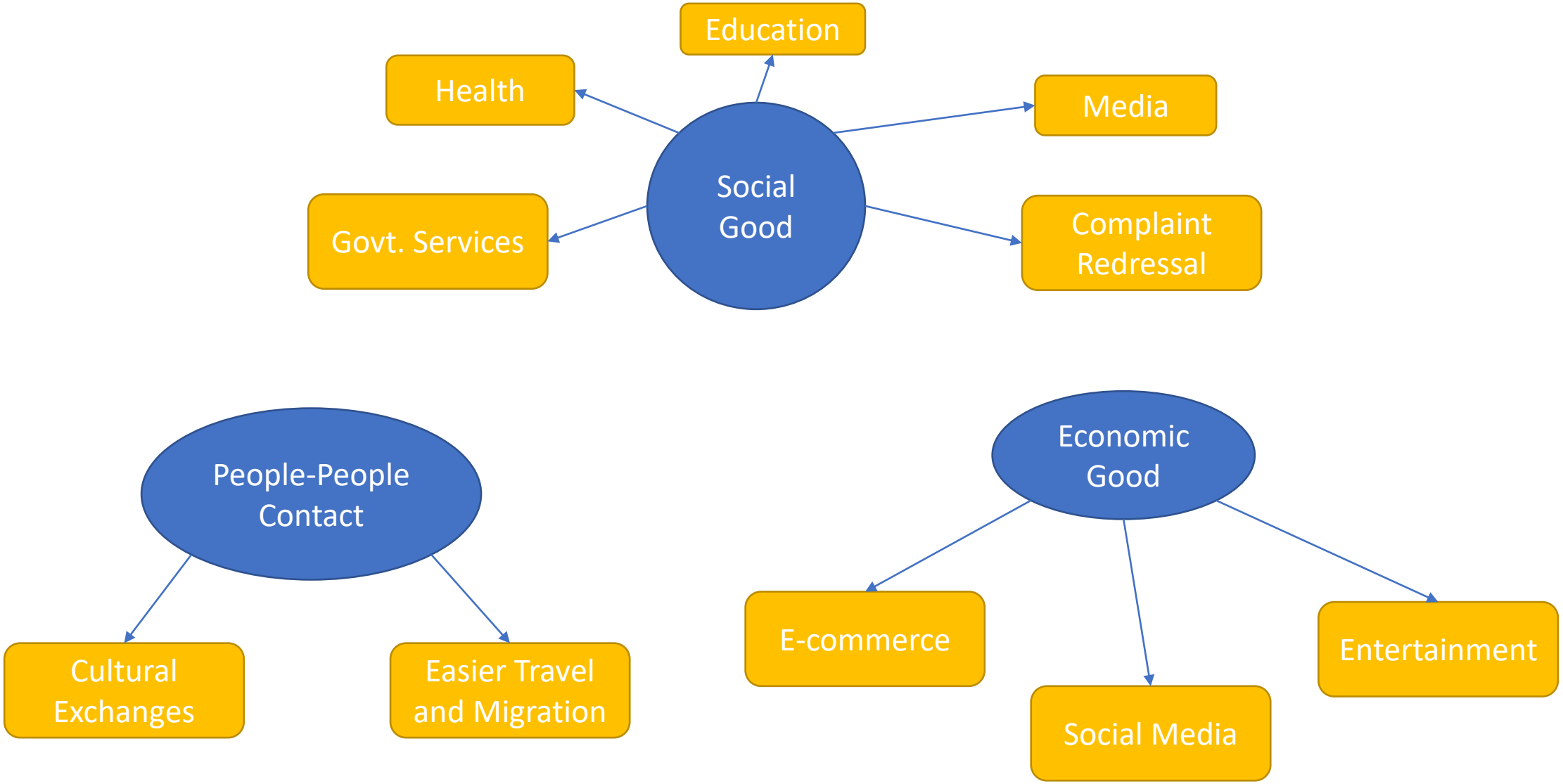
- *22 scheduled languages*
- *English as lingua franca: 125 million speakers*
- *8 languages in the world's top 20 languages*
- *30 languages with more than 1 million speakers*



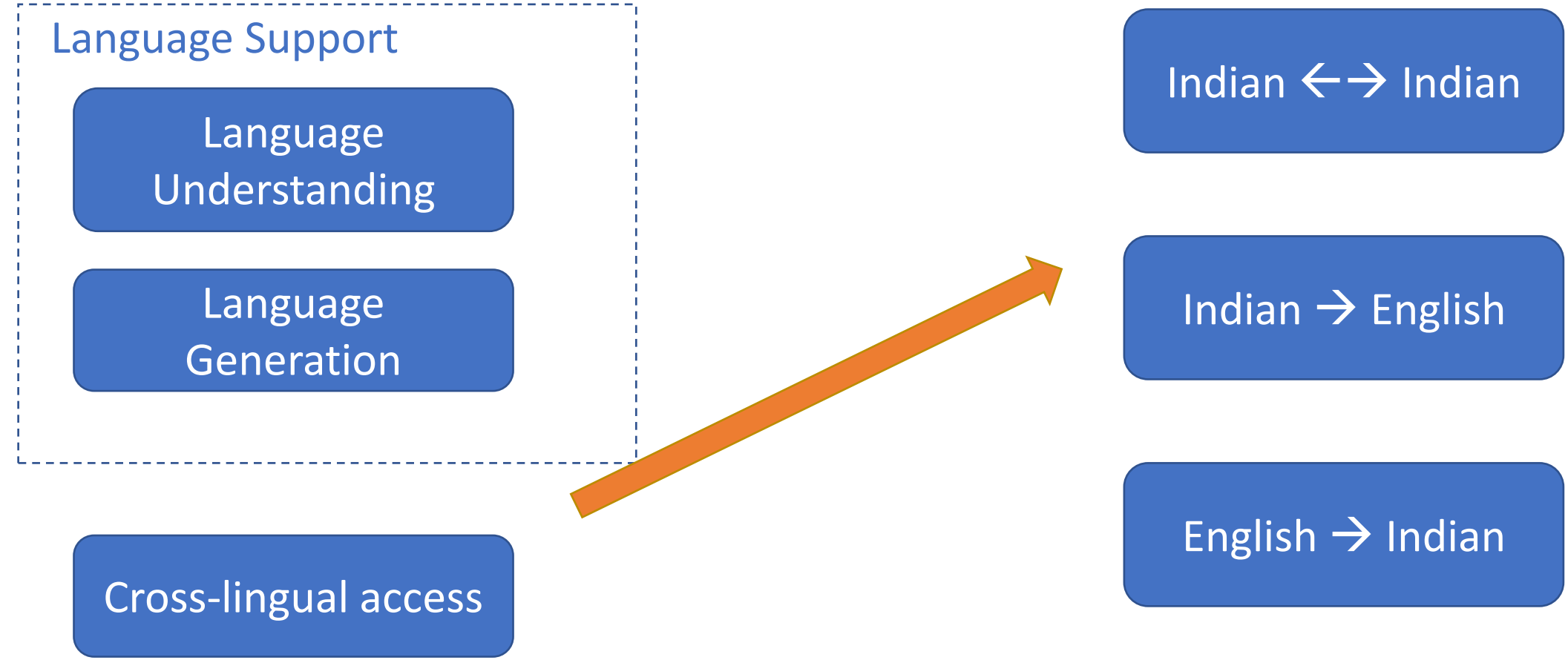
Internet User Base in India (in million)

Sources: Wikipedia; Census of India 2011; Indian Languages: Defining India's Internet KPMG-Google Report 2017

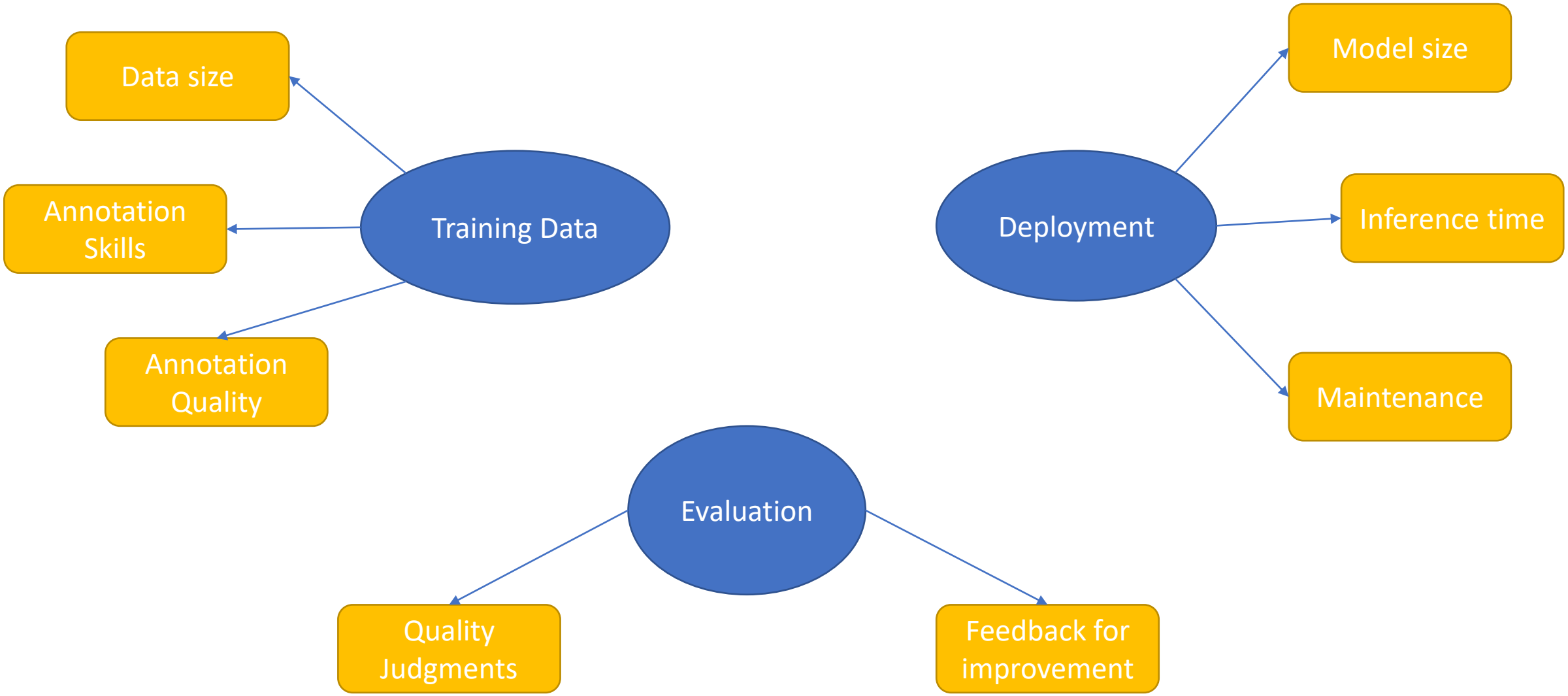
Addressing Multilinguality is important to maximizing impact of language technologies



Multilingual NLP Scenarios



Scalability Challenges for NLP solutions



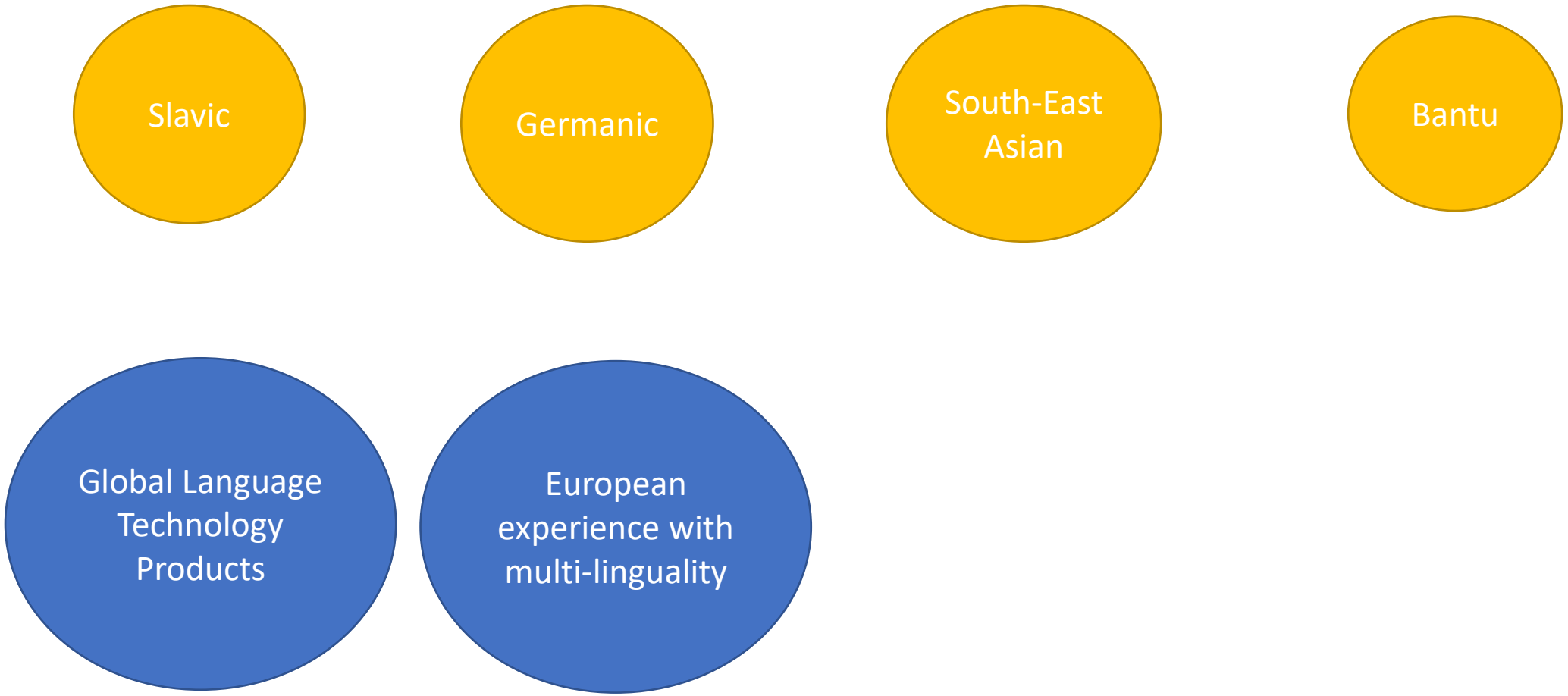
Effort and cost increase as languages increase

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?

Multilingual challenges are not uniquely Indian → India is a microcosm of the world



How are Indian languages related?

Related Languages

Related by Genealogy

Dravidian, Indo-Aryan,
Tibeto-Burman, Austro-Asiatic

English	Vedic Sanskrit	Hindi	Punjabi	Gujarati	Marathi	Odia	Bengali
bread	Rotika	chapātī, roṭī	roṭi	paū, roṭlā	chapāti, poli, bhākarī	pauruṭi	(pau-)ruṭi

Related by Contact

Indian Subcontinent as a linguistic area

Sanskrit word	Language	Loanword	English
cakram	Tamil	cakkaram	wheel
matsyah	Telugu	matsyalu	fish

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

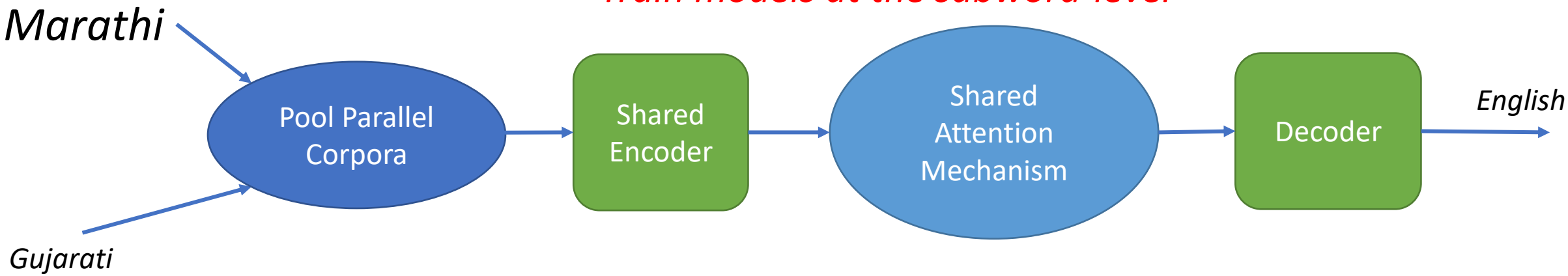
Lexical: share significant vocabulary (cognates & loanwords)

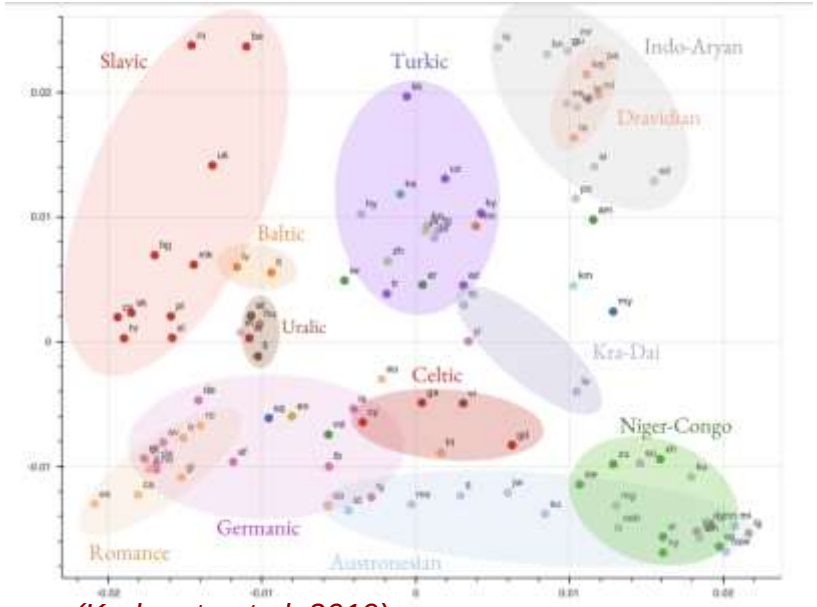
Morphological: correspondence between suffixes/post-positions

Syntactic: share the same basic word order

Transfer Learning Recipe

Train models at the subword-level





(Kudungta et al, 2019)

Encoder Representations cluster by language family

Transfer Learning works best for related languages

https://www.microsoft.com/en-us/translator/blog/2020/04/15/microsoft-adds-five-languages-of-india-to-microsoft-translator

Moving beyond the simple transfer learning paradigm

Can we better utilize the similarities between Indian languages?

Similarity in Scripts

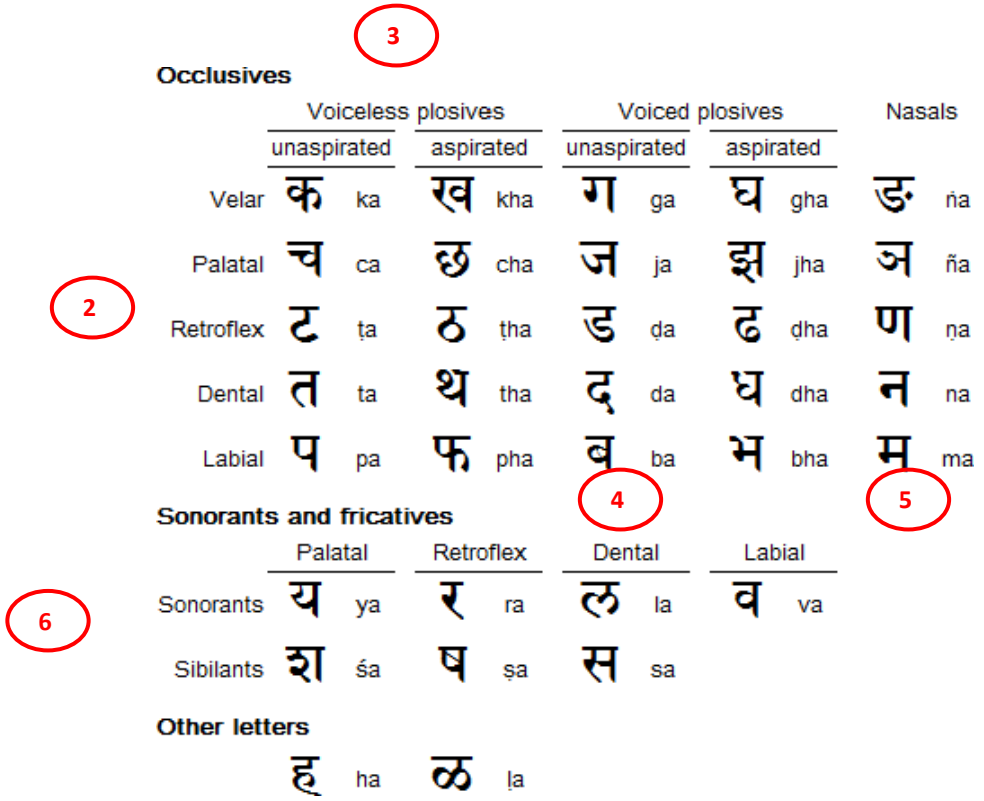
Multilingual transliteration

Convert to a common script & Pool

Malayalam	കോഴിക്കോട്	कोऴिक्कोट्	kozhikode
Hindi	केरल	केरल	kerala
Kannada	ಬೆಂಗಳೂರು	बेंगळूरु	bengaluru

Make inputs more similar, reduce vocabulary size

Anoop Kunchukuttan, Mitesh Khapra, Gurneet Singh, Pushpak Bhattacharyya. *Leveraging Orthographic Similarity for Multilingual Neural Transliteration*. TACL. 2018.



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

Scientific Design of scripts enables feature representation of characters/sounds

- *Unsupervised Transliteration*
  - *Initialization and priors on Character transliteration probabilities*
- *Cognate Identification*

Anoop Kunchukuttan, Pushpak Bhattacharyya, Mitesh Khapra. *Substring-based unsupervised transliteration with phonetic and contextual knowledge*. CoNLL. 2016.

Word Segmentation using Aksharas

akshara, the fundamental organizing principle of Indian scripts

(CONSONANT) + VOWEL

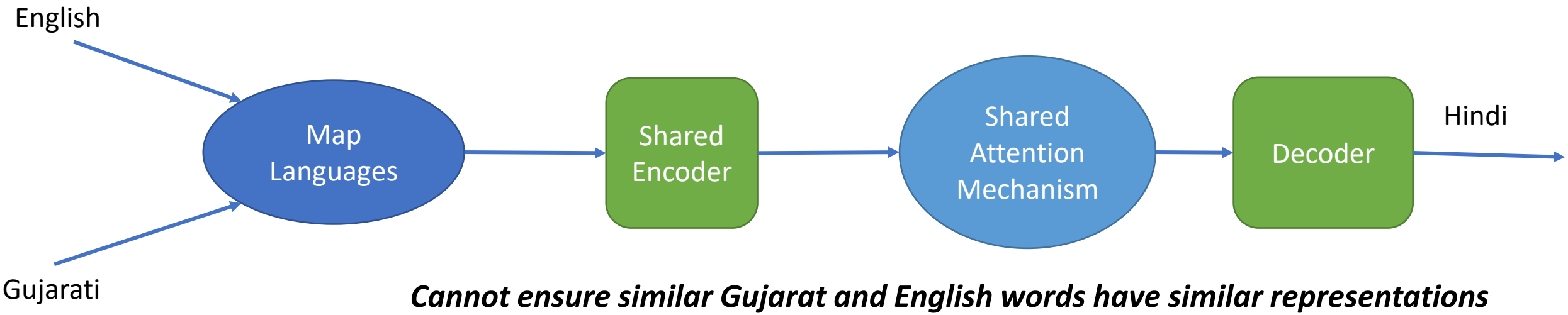
Examples: की (kI), प्रे (pre)

Hindi	Kannada	English
वि द्या ल य	ವಿ ದ್ಯಾ ಲ ಯ	vi dya lay
अ र्जु न	ಅ ರ್ಜು ನ	a rju n

Useful as basic units for transliteration and translation

Anoop Kunchukuttan, Pushpak Bhattacharyya. *Orthographic Syllable as basic unit for SMT between Related Languages*. EMNLP. 2016.

Transfer from English to Indian languages



Little G→H corpus

Cannot ensure similar Gujarat and English words have similar representations

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

Syntactic divergence can be overcome with shared rules

1. Rudra Murthy V, Anoop Kunchukuttan, Pushpak Bhattacharyya. *Addressing word-order Divergence in Multilingual Neural Machine Translation for extremely Low Resource Languages*. NAACL. 2019.
2. Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, Pushpak Bhattacharyya. *Shata-Anuvadak: Tackling Multiway Translation of Indian Languages* . Language and Resources and Evaluation Conference. 2014.

Transfer between contact languages

Tamil to English NMT with transfer-learning using Hindi

Addressing syntactic divergence in NMT using Hindi-driven rules

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

1. Vikrant Goyal, Anoop Kunchukuttan, Rahul Kejriwal, Siddharth Jain, Amit Bhagwat. *Contact Relatedness can help improve multilingual NMT: Microsoft STCI-MT @ WMT20*. WMT 2020. 2020.
2. Rudra Murthy V, Anoop Kunchukuttan, Pushpak Bhattacharyya. *Addressing word-order Divergence in Multilingual Neural Machine Translation for extremely Low Resource Languages*. NAACL. 2019.

Putting these ideas together into usable systems …

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)

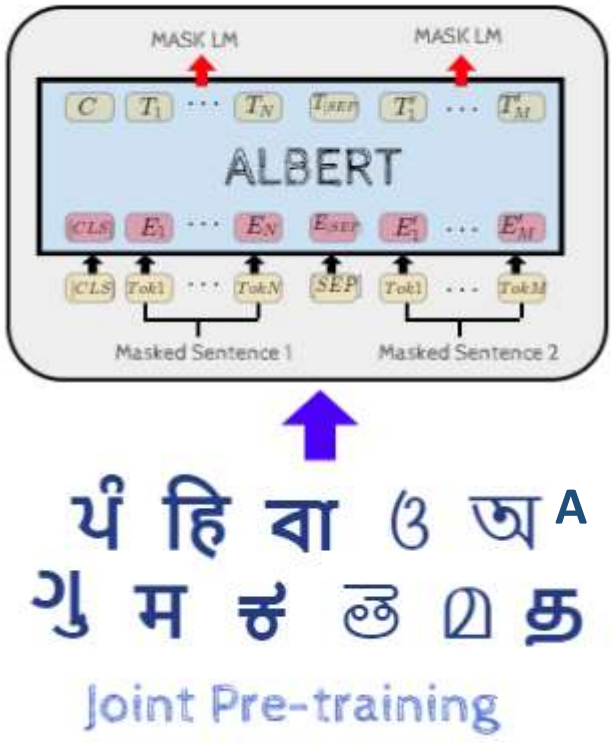
	as	bn	gu	hi	mr	ne	or	pa	sd	si	sa	kok	kn	ml	te	ta
Text Processing	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
Morphological Segmentation	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓
Syllabification	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
Script Processing	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓

Future: add support for multilingual pre-trained embeddings, fundamental tools like POS, Dependency parsing, NER, etc.

IndicBERT

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

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



- 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

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.

Multilingual Approaches are important for language technologies to scale and make social impact

The field is nascent, there are many directions to explore

- *Better representation methods to utilize relatedness*
- *Bridging typological divergence between English and Indian languages*
- *Utilizing relatedness for generation tasks*
- *Cross-lingual Evaluation benchmarks*

Thank You!

http://anoopk.in