Indian Language Computing

A Multilingual Perspective

Anoop Kunchukuttan

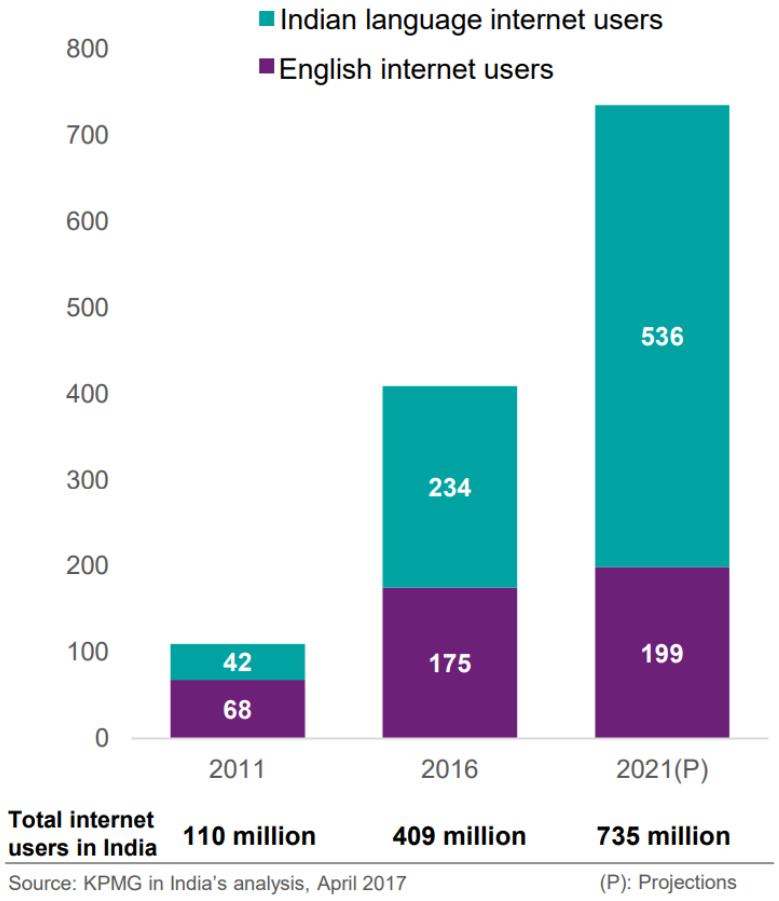
Senior Applied Researcher, Microsoft
Co-founder, AI4Bharat
anoop.kunchukuttan@gmail.com

Tamil Internet Conference 2021
4th Dec 2021

Usage and Diversity 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

Translation

Transliteration

Code-mix Processing

Entity Identification

Entity Linking

Information Extraction & Categorization

Digital payments

E-tailing

Online government services

Digital classifieds

Chat applications

Digital entertainment

Social media platforms

Digital news

Digital write-ups

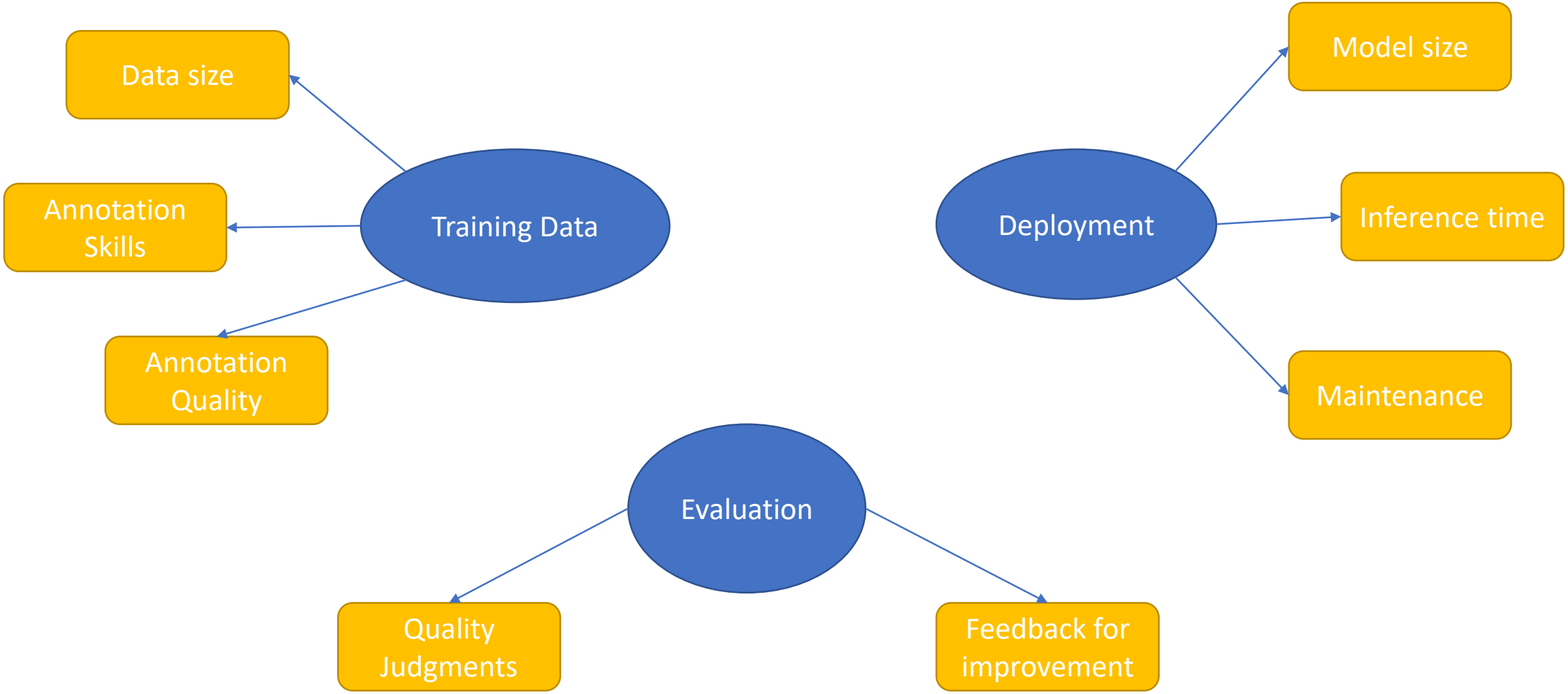
Search

Question & Answering

Recommenda tion

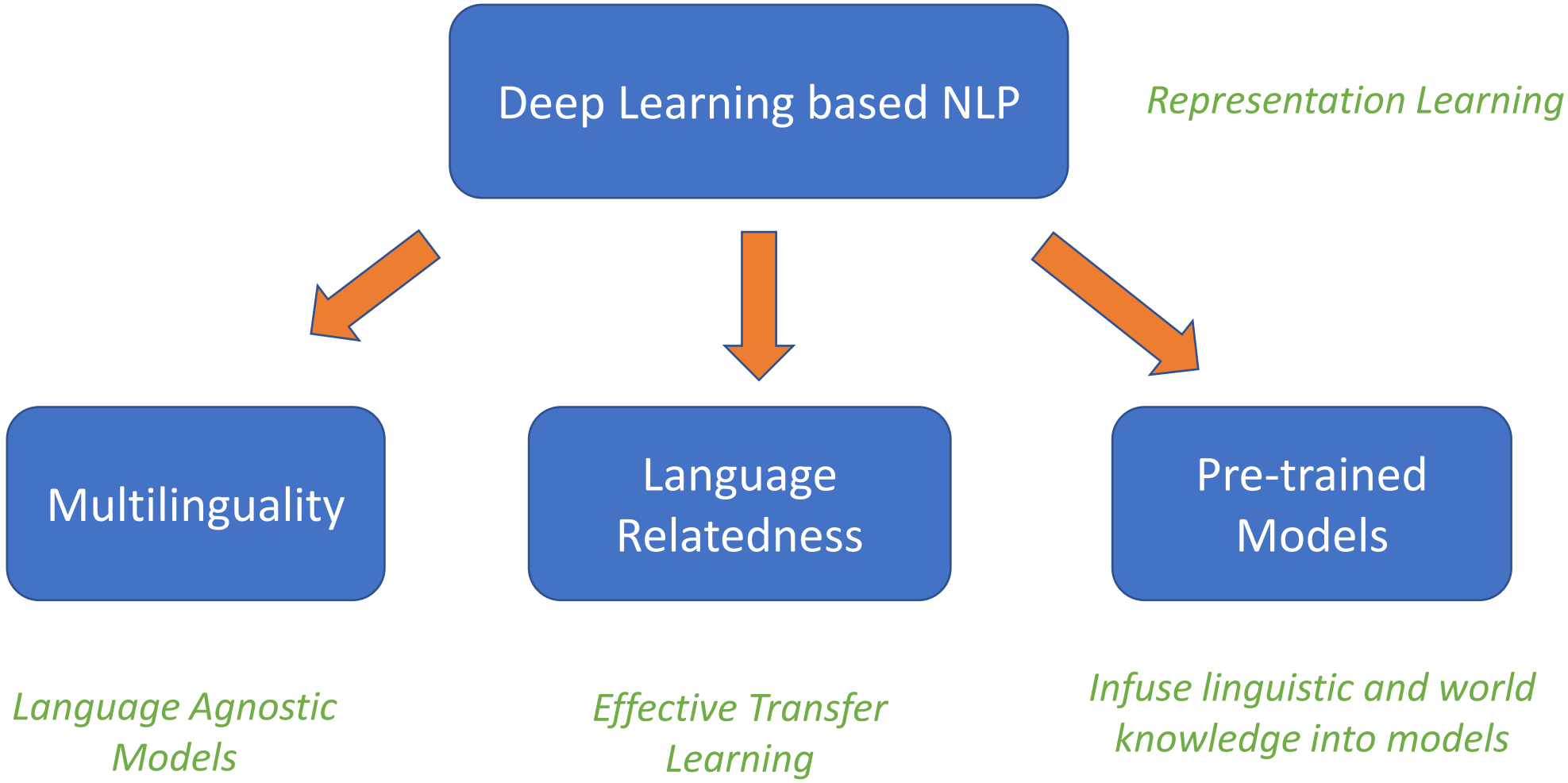
Applications requiring Indian language support

Scalability Challenges for NLP solutions



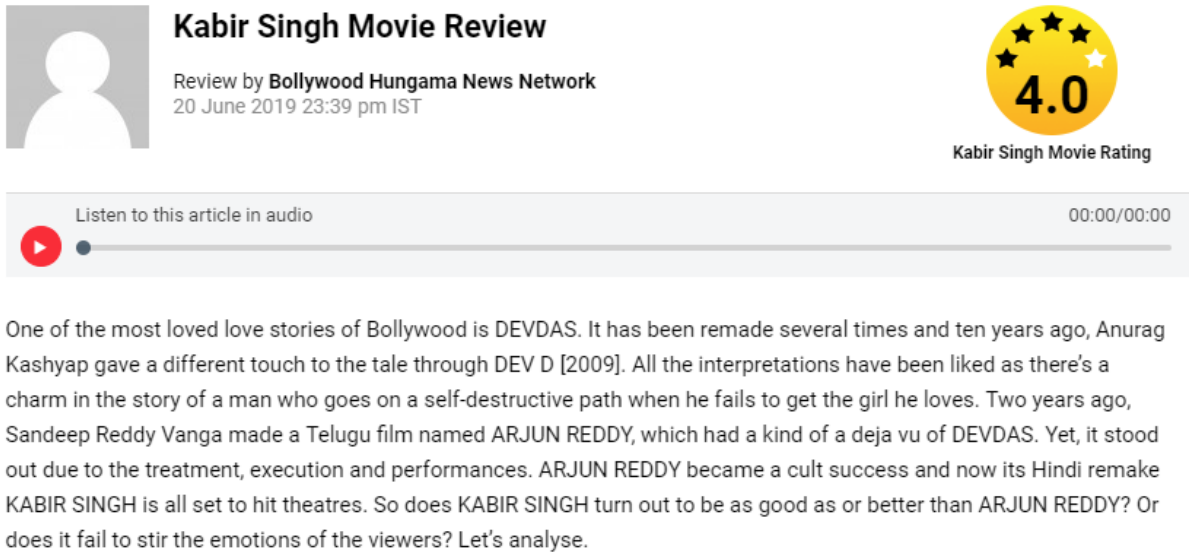
Effort and cost increase as languages increase

The Opportunity for Indian Language NLP



Representation Learning

Let us look at a simple NLP application – Sentiment Analysis



Kabir Singh Movie Review

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

4.0
Kabir Singh Movie Rating

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

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

?

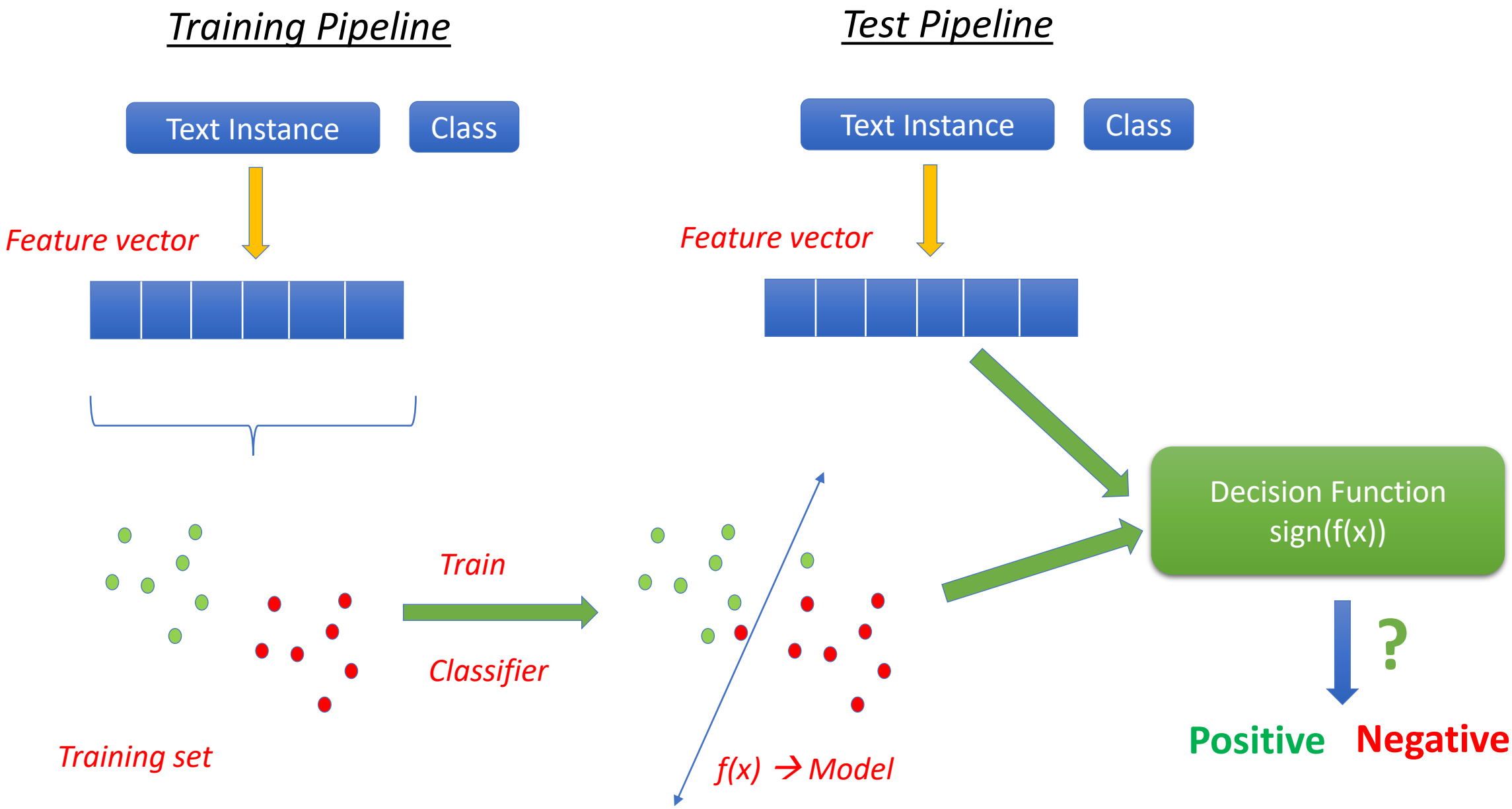
Positive

Negative

Neutral

An example of a text classification problem

A Machine Learning Pipeline for Text Classification



Simple Features

Bag-of-words (presence/absence)

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

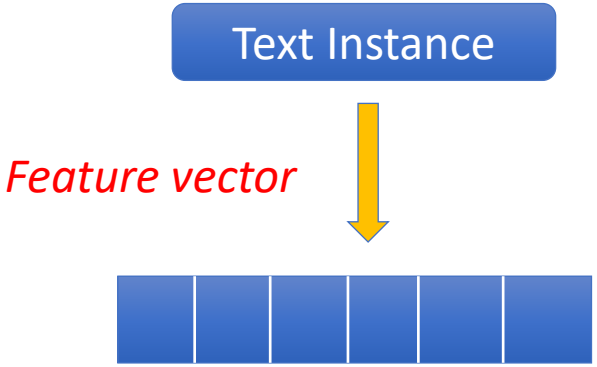
Large and sparse feature vector: size of vocabulary
Each feature is atomic → similarity between features, synonyms not captured

More features

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

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

Distributed Representations



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

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

Distributional Hypothesis

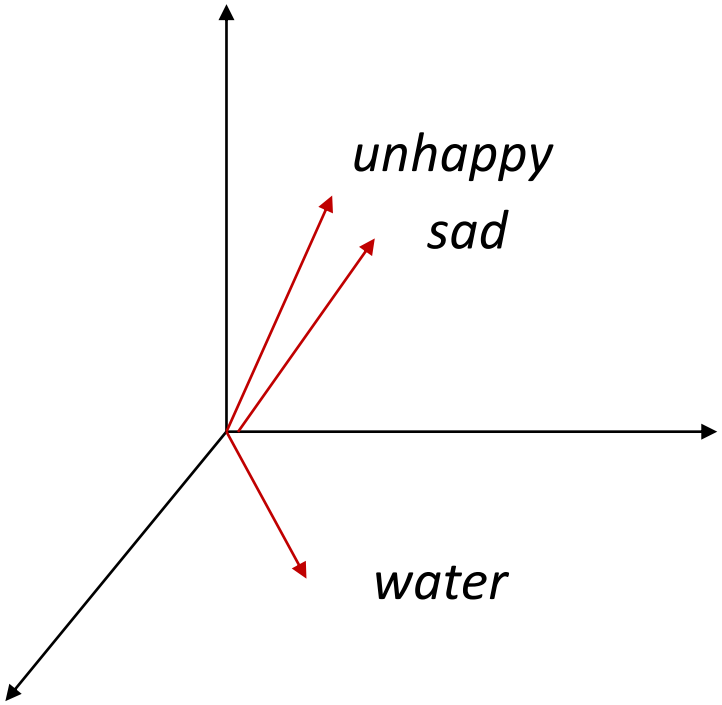
"A word us known by the company it keeps" *- Firth (1957)*

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

He is unhappy about the failure of the project

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

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

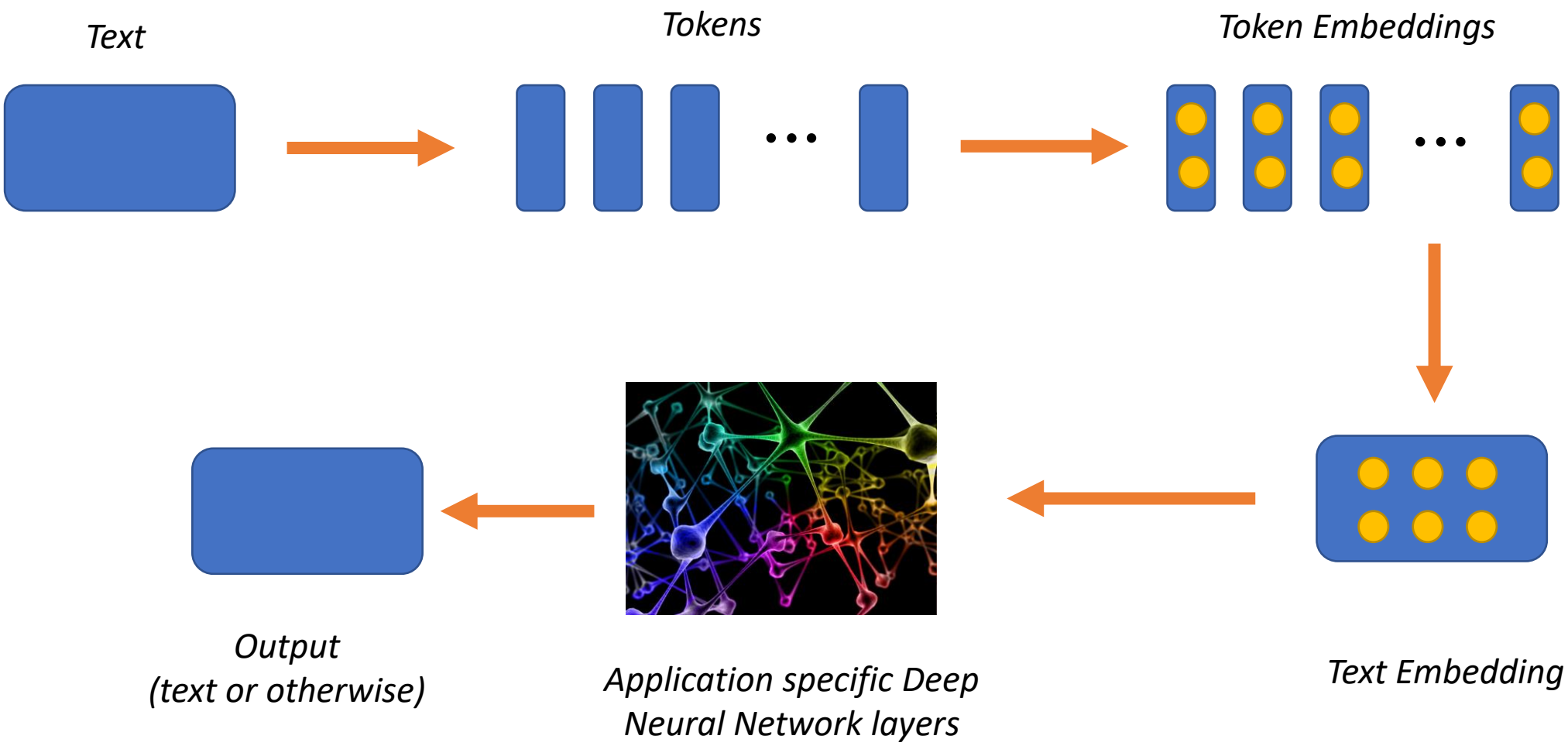


Similarity of words can be defined in terms of vector similarity: Cosine similarity, Euclidean distance, Mahalanobis distance

Similarity across languages

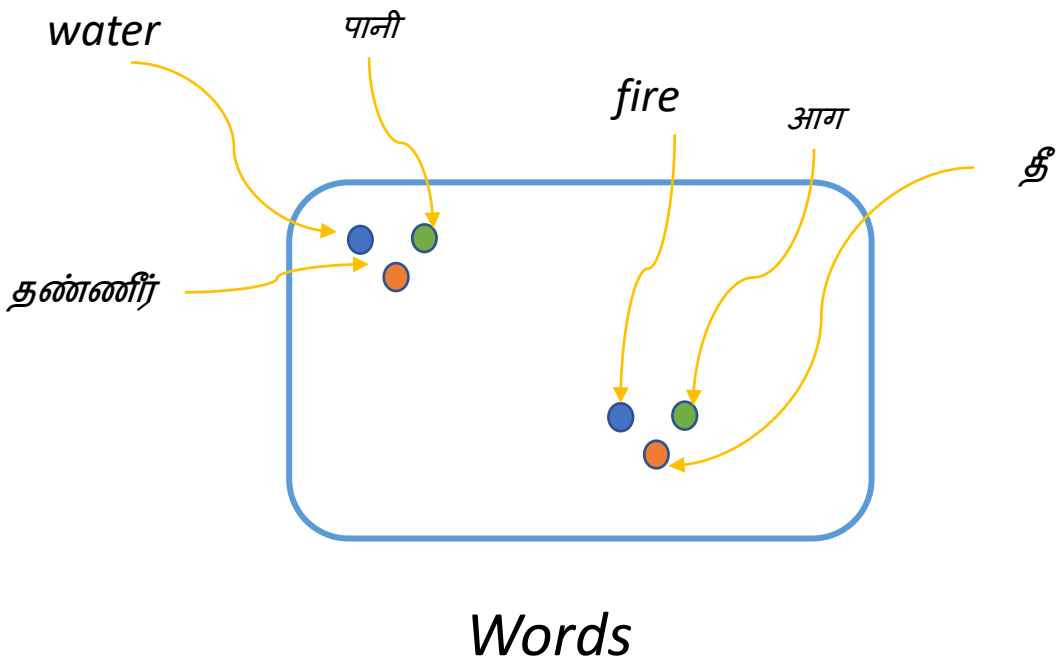
Contextual representation of words

A Typical Deep Learning NLP Pipeline

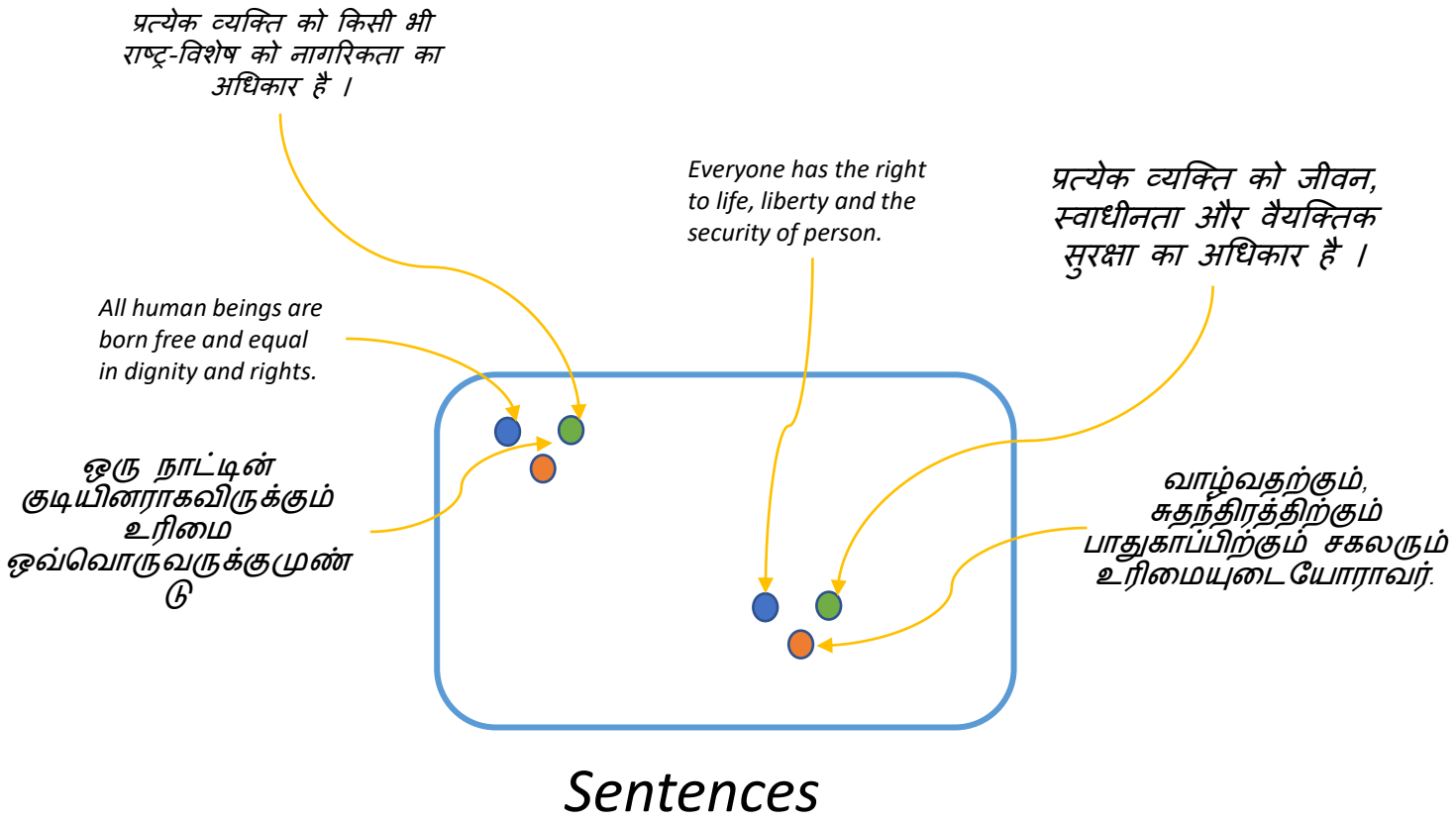


Multilinguality

Represent semantically similar language artifacts in the same vector space

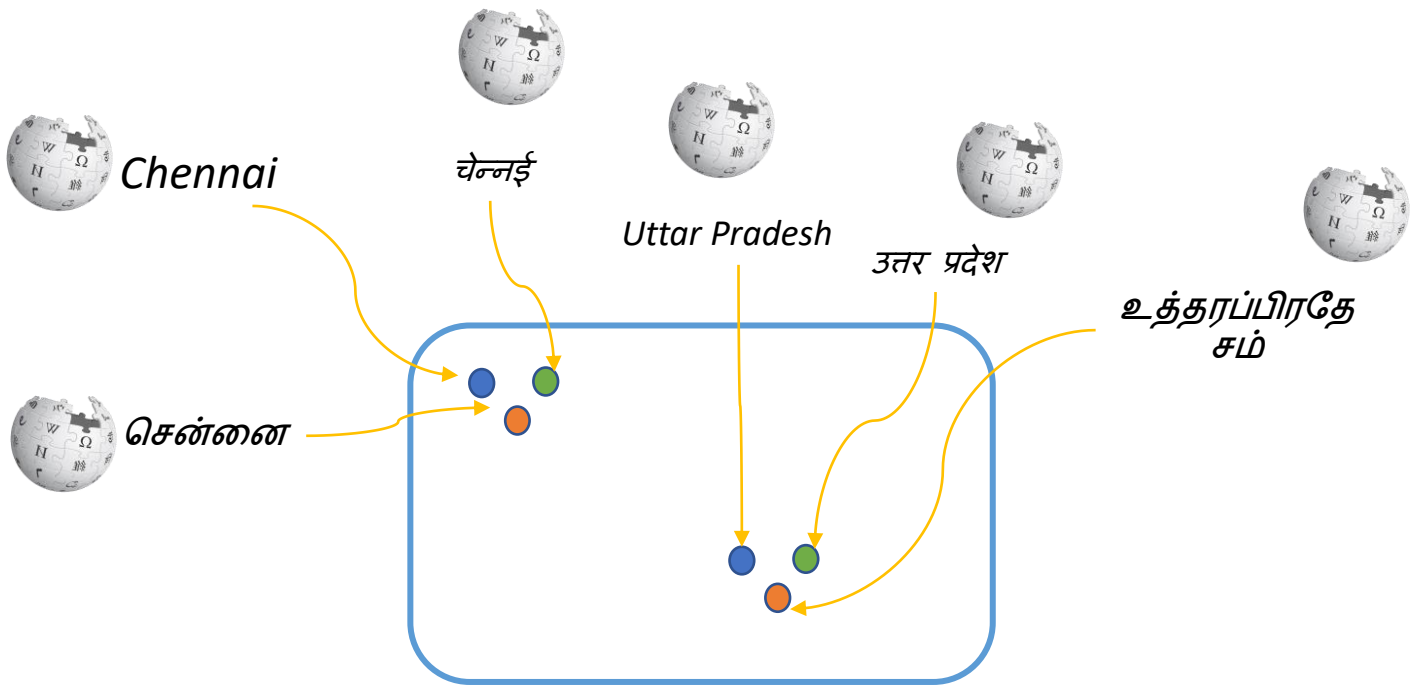


Represent semantically similar language artifacts in the same vector space



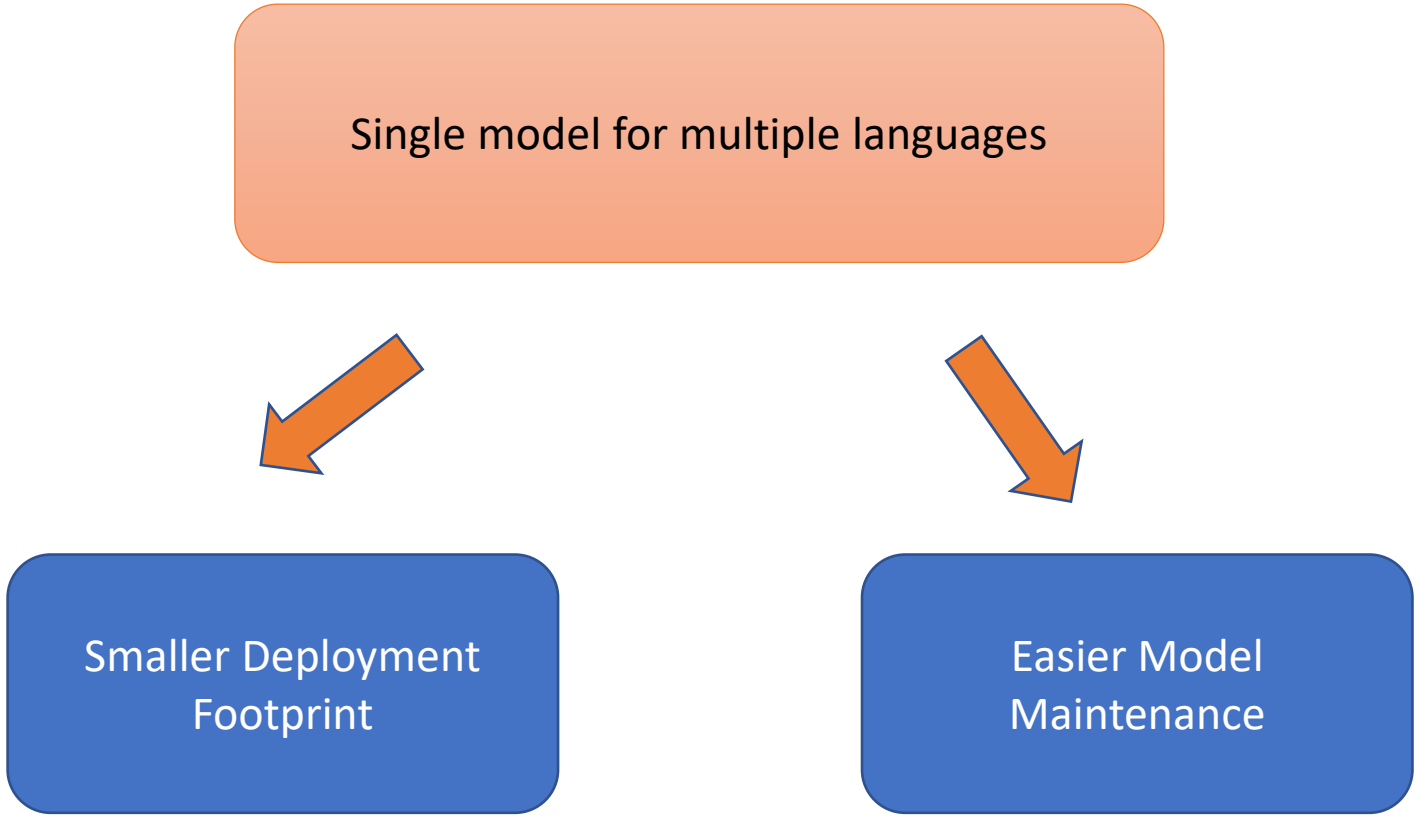
Sentences

Represent semantically similar language artifacts in the same vector space

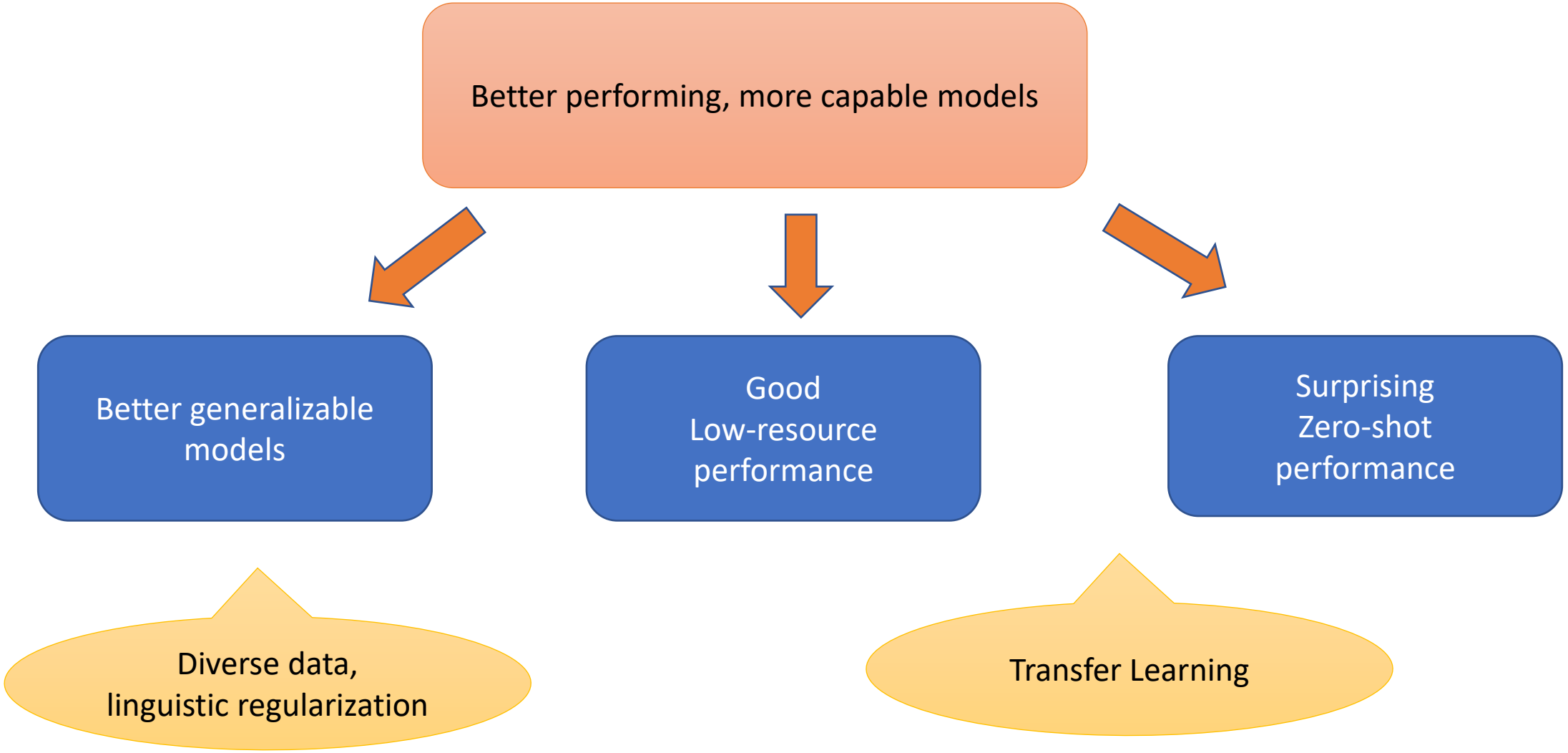


Documents

How does multilinguality help?

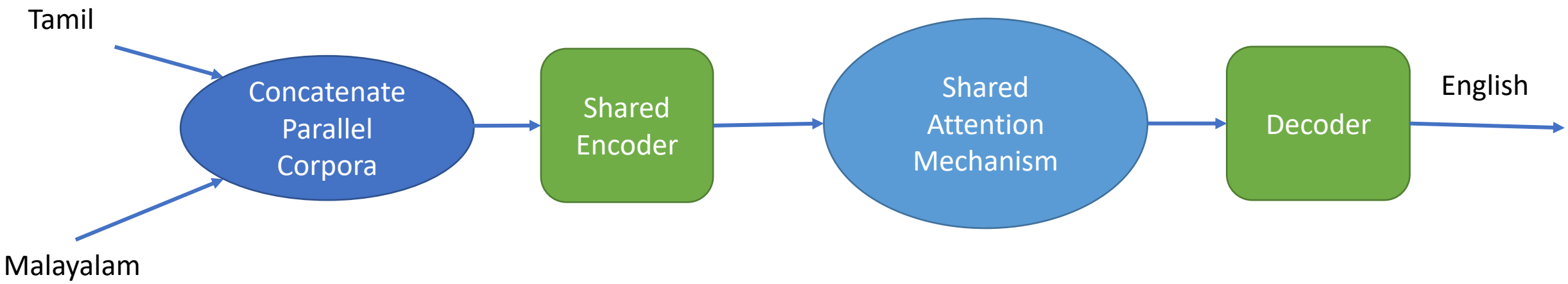


How does multilinguality help?



Multilingual Indian Language → en Translation Models

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



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

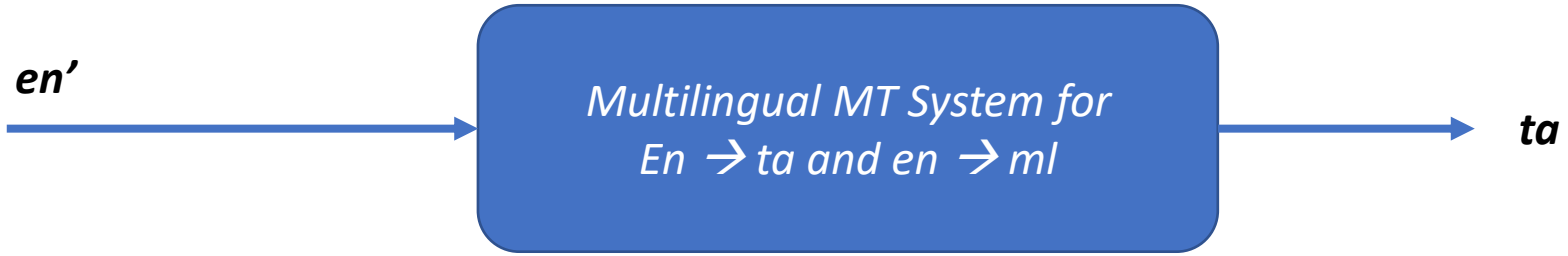
English → Indian Languages

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

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

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

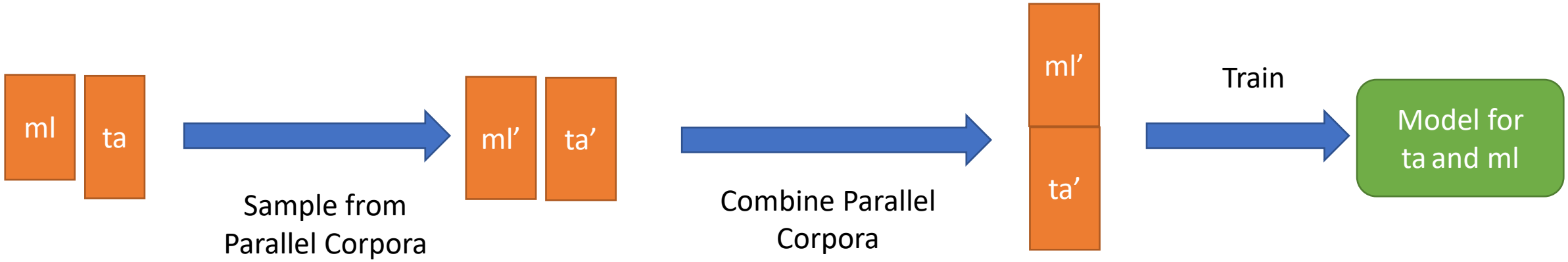
<u>Modified Input</u>: *France and Croatia will play the final on Sunday* *<ta>*



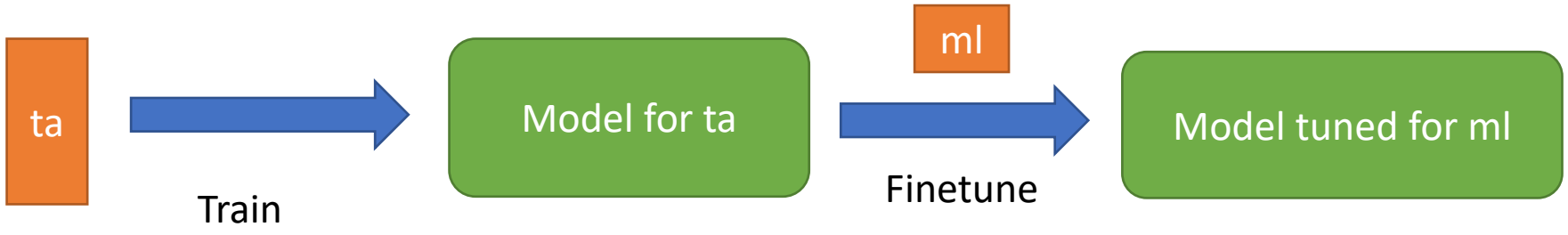
Still a challenging problem

Training Multilingual NMT systems

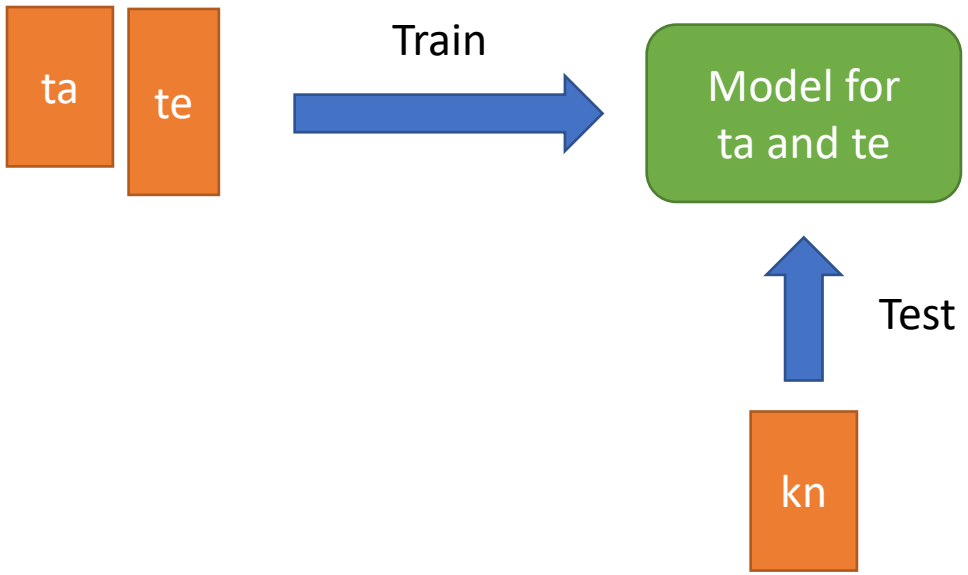
Joint Training



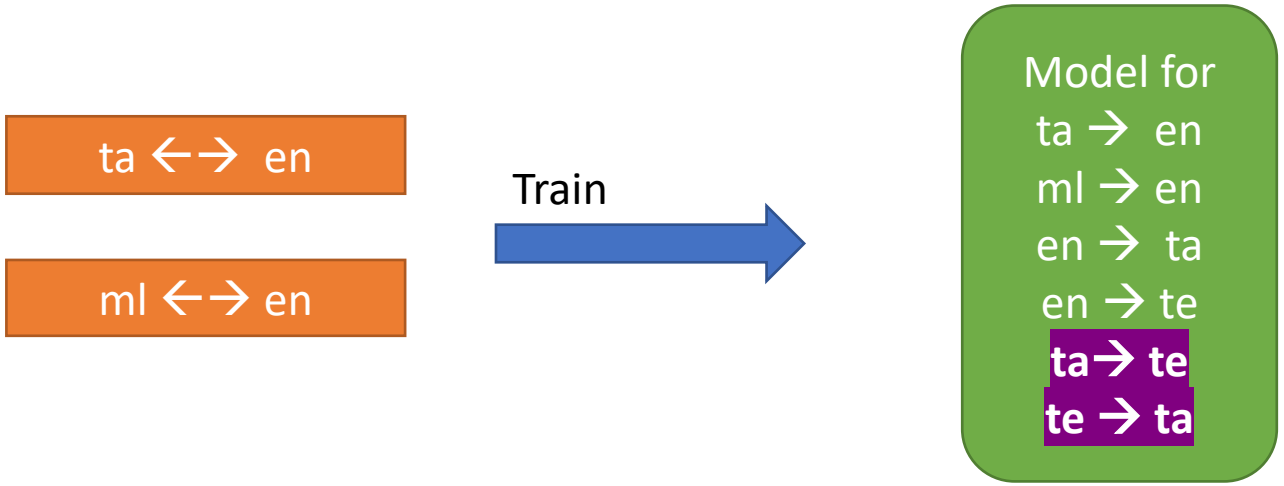
Transfer Learning



Zeroshot Translation into English



Zeroshot Translation between Indian languages



Language Relatedness

Why are Indian languages related?

Related Languages

Related by Genealogy

Related by Contact

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

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

<u>Linguistic Areas</u>
Indian Subcontinent, Standard Average European

(Trubetzkoy, 1923)

Related languages may not belong to the same language family!

Cognates & Borrowed words in Indian Languages

Indo-Aryan

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

Dravidian

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

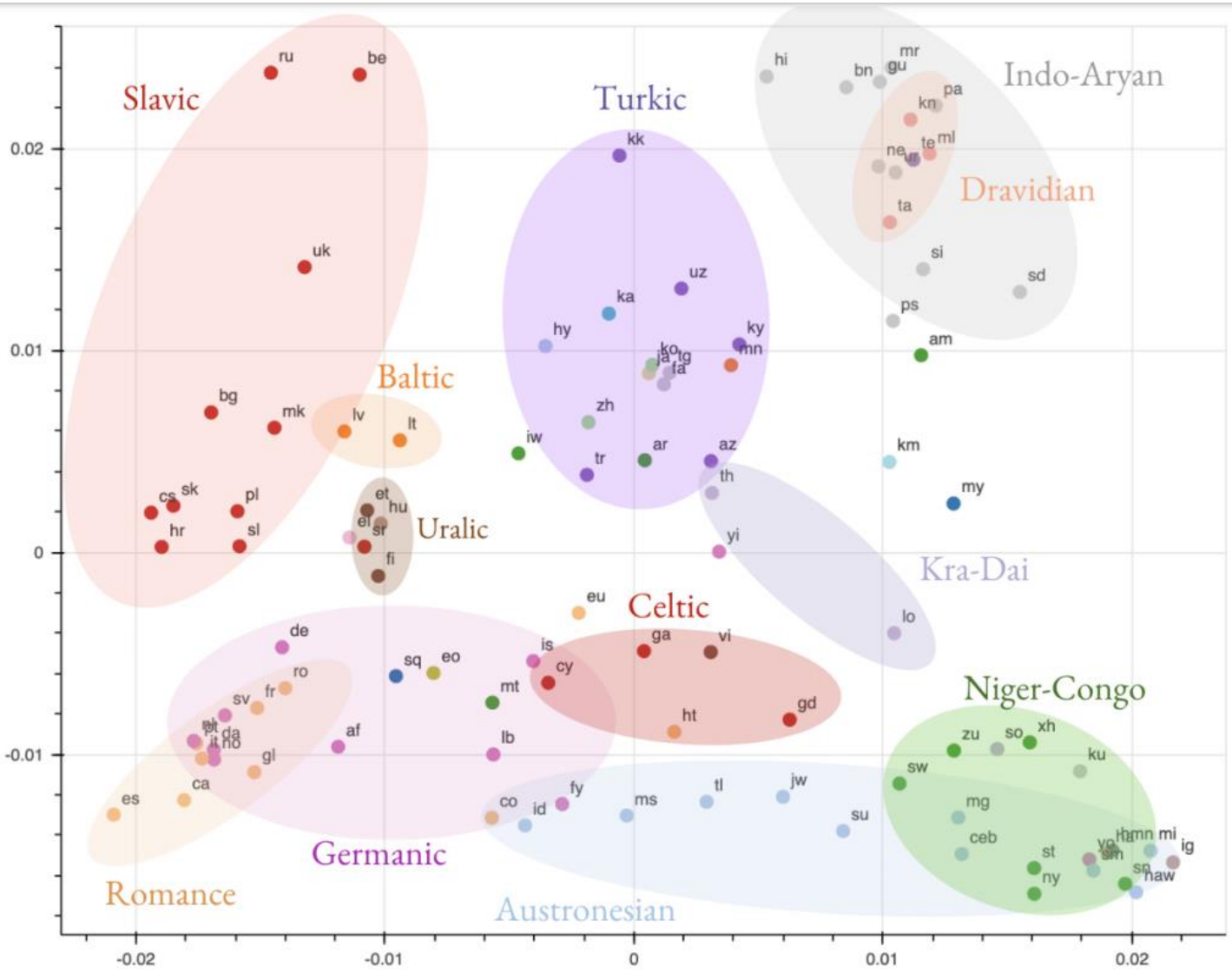
Indo-Aryan words in Dravidian languages

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

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

Source: Wikipedia and IndoWordNet

Transfer Learning works best for related languages



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

Motivation for:

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

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

Key Similarities between related languages

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

भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला

bhAratAcyA svAta.ntryadinAnimitta ameriketIla lOsa enjalsa shaharAta kAryakrama Ayojita karaNyAta AlA

Marathi

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

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

Marathi segmented

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

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

Hindi

Lexical: share significant vocabulary (cognates & loanwords)

Morphological: correspondence between suffixes/post-positions

Syntactic: share the same basic word order

Orthographic Similarity

Brahmi-derived Indic scripts are orthographically similar

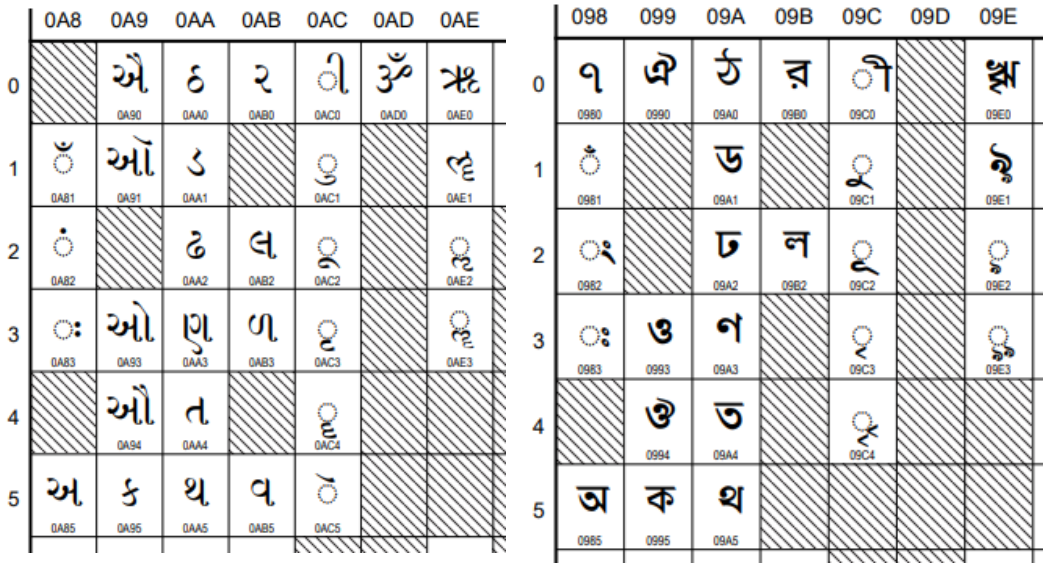
Devanagari	अ आ इ ई उ ऊ ऋ ऌ ऍ ऎ ए ऐ ऑ ऒ ओ औ क ख ग घ ङ च छ ज झ
Bengali	অ আ ই ঈ উ ঊ ঋ ঌ এ ঐ ও ঔ ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড
Gurmukhi	ਅ ਆ ਇ ੲ ਉ ਊ ਏ ਐ ਓ ਔ ਕ ਖ ਗ ਘ ਙ ਚ ਛ ਜ ਝ ਞ ਟ ਠ ਡ ਢ ਣ ਤ ਥ
Gujarati	અ આ ઇ ઈ ઉ ઊ ઋ ઌ એ ઐ ઑ ઓ ઔ ક ખ ગ ઘ ઙ ચ છ જ ઝ ઞ ટ ઠ
Oriya	ଅ ଆ ଇ ଈ ଉ ଊ ଋ ଌ ଏ ଐ ଓ ଔ କ ଖ ଗ ଘ ଙ ଚ ଛ ଜ ଝ ଞ ଟ ଠ ଡ ଢ ଣ
Tamil	அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஒ ஓ ஔ க ங ச ஜ ஞ ட ண த ந
Telugu	అ ఆ ఇ ఈ ఉ ఊ ఋ ఌ ఎ ఏ ఐ ఒ ఓ ఔ క ఖ గ ఘ ఙ చ ఛ జ ఝ
Kannada	ಅ ಆ ಇ ಈ ಉ ಊ ಋ ಌ ಎ ಏ ಐ ಒ ಓ ಔ ಕ ಖ ಗ ಘ ಙ ಚ ಛ ಜ ಝ ಞ
Malayalam	അ ആ ഇ ഈ ഉ ഊ ഋ ഌ എ ഏ ഐ ഒ ഓ ഔ ക ഖ ഗ ഘ

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

Script Conversion

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

unicode_codepoint(char) - Unicode_range_start(L_1) + Unicode_range_start(L_2)



केरला kerala

কেরলা કેરલા

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

Only a single mapping needed for Romanization too

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

A simple and powerful property to utilize

relatedness between Indian languages

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

Multilingual Transliteration

(Kunchukuttan, et al, 2018)

Pool training sets

Malayalam	കോഴിക്കോട്	kozhikode
Hindi	केरल	kerala
Kannada	ಬೆಂಗಳೂರು	bengaluru

Convert to a common script

Malayalam	कोऴिक्कोट्	kozhikode
Hindi	केरल	kerala
Kannada	बेंगळूरु	bengaluru

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

Example of Multi-task Learning

Similar tasks help each other

Zero-shot transliteration is possible

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

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

Traditionally organized as per sound phonetic principles

shows various symmetries

Useful for unsupervised transliteration

Primary vowels

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

Secondary vowels

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

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

1 2 3 4 5 6

Lexical Similarity

Lexical Similarity

(Words having similar form and meaning)

- *Cognates*

a common etymological origin

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

- *Named Entities*

do not change across languages

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

- *Loan Words*

borrowed without translation

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

- *Fixed Expressions/Idioms*

MWE with non-compositional semantics

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

Enables sharing of data across languages

Why it matters

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

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

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

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

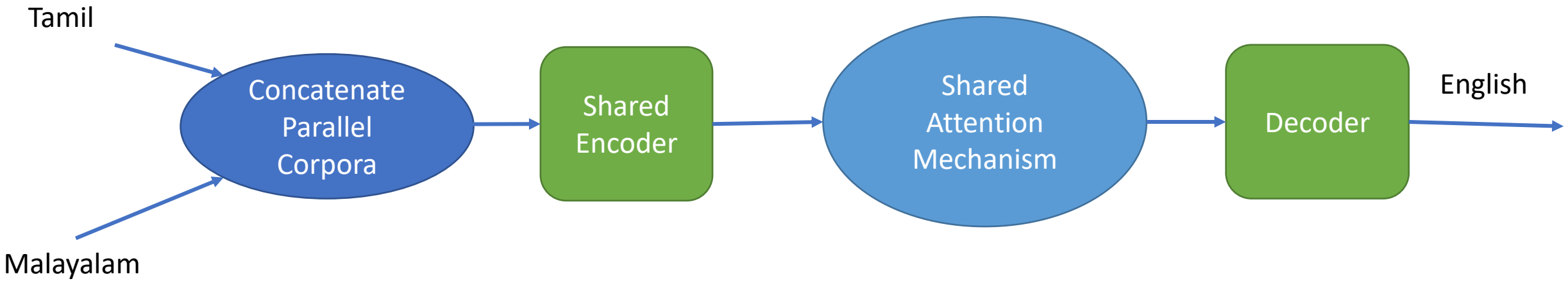
Lexical Overlap ➔ Representation overlap
Makes it easier for the model to learn

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

Multilingual Indian Language → en Translation Models

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

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



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

Syntactic Similarity

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

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

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

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

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

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

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

Source reordering for SMT

(Kunchukuttan et al., 2014)

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

Bahubali earned more than 1500 crore rupees at the boxoffice

Bahubali the boxoffice at 1500 crore rupees earned

बाहुबली ने बॉक्सऑफिस पर 1500 करोड रुपए कमाए

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

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

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

https://github.com/anoopkunchukuttan/cfilt_preorder

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

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

Tamil to English NMT with transfer-learning using Hindi

Pre-trained Models

Representation Learning

Automatic Feature Extraction
Continuous Space Representation
Numerical Optimization at disposal

Multilingual learning

Transfer Learning
Better generalizability across languages

Supervised data not sufficient

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

How do we know if the sentence is grammatically correct?

How do we know if the sentence makes sense?

These capabilities are important for generalization

Pre-trained Models

Task-independent models that know about language

Word Embeddings

Encoder Language Model for NLU

Decoder Language Model for NLG

Encoder-decoder Language Model for NLU+NLG

fastText

Google BERT

OpenAI GPT-3

FAIRSEQ

BART

Multilinguality

MUSE

mBERT

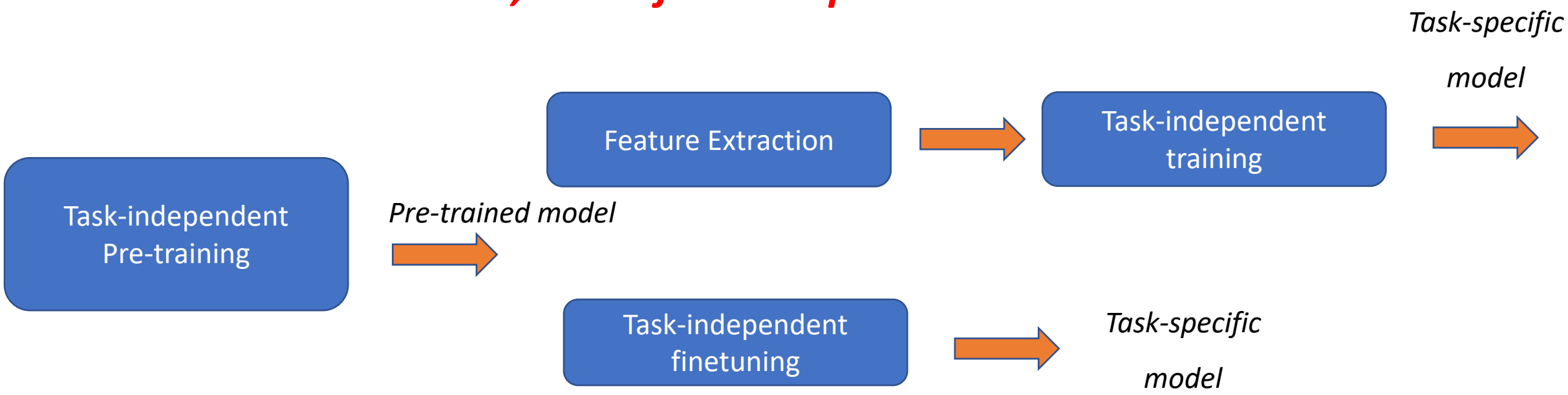
mBART

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

Language models are

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

Pre-train once, reuse for multiple downstream tasks



Only task-specific training: less data & less computation

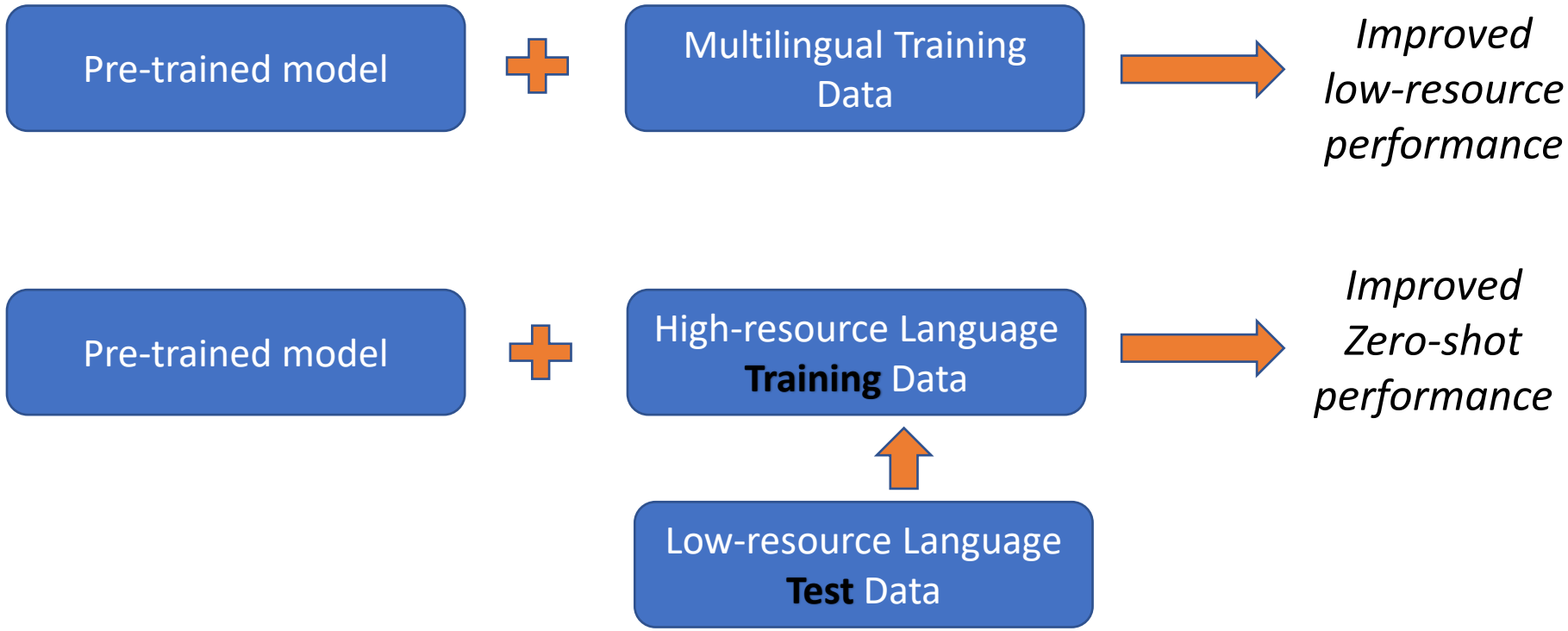
Language understanding for tasks like sentiment analysis, question answering, paraphrase detection

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

Multi-linguality and Pre-training are complementary

Language-family specific pre-trained model

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



Putting these ideas together into usable systems …

AI4Bharat

An IIT Madras Initiative

Mitesh M. Khapra

Associate Professor, IIT Madras
PhD, IIT Bombay
Areas - NLP, Deep Learning

Pratyush Kumar

Researcher Microsoft

Assistant Professor, IIT Madras
PhD, ETH Zürich
Areas - Deep Learning, Systems

Anoop Kunchukuttan

Researcher, Microsoft
PhD, IIT Bombay
Areas - NLP

+ many hard-working students and volunteers

https://indicnlp.ai4bharat.org

Mission Statement

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

THE APACHE® SOFTWARE FOUNDATION

We want to be the Apache for Indian Languages AI stack

What have we done so far?

IndicCorp

Corpora for 11 Indic languages

IndicGLUE

NLU Benchmarks for Indian languages

IndicBERT

Compact Language Models for NLU for Indian languages

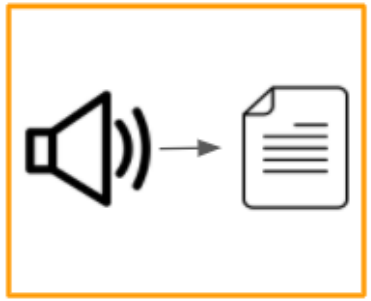
IndicBART

Compact Language Models for NLG for Indian languages

What have we done so far?

Samanantar

Parallel corpus, translation models between English & 11 Indic languages



IndicWav2Vec

State of the art ASR models for 9 Indian languages

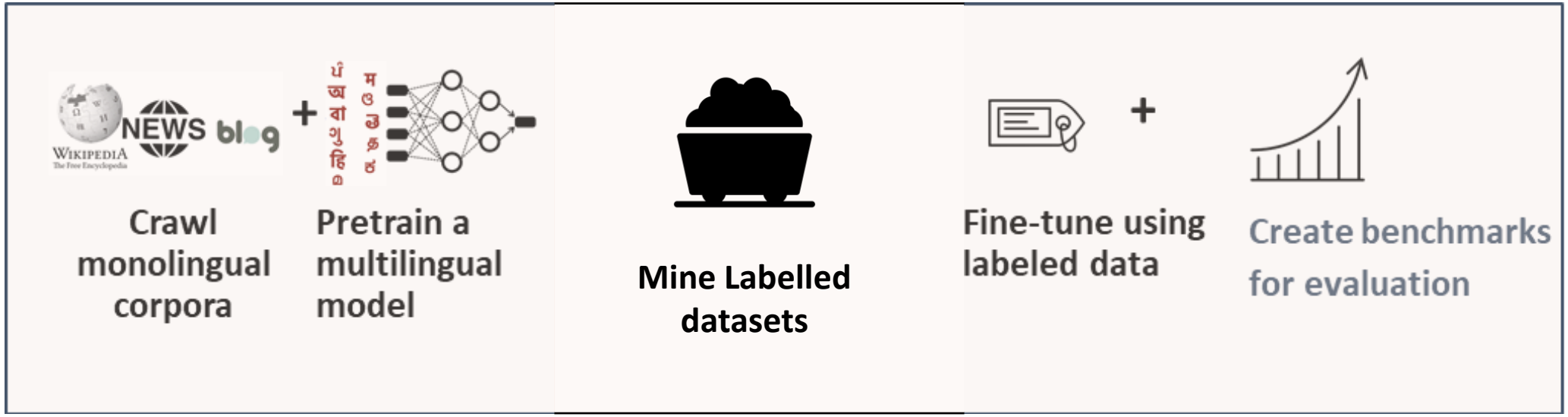
INCLUDE

Datasets and efficient models for isolated Indian Sign Language

Input Tools

Romanized keyboards for under-represented languages

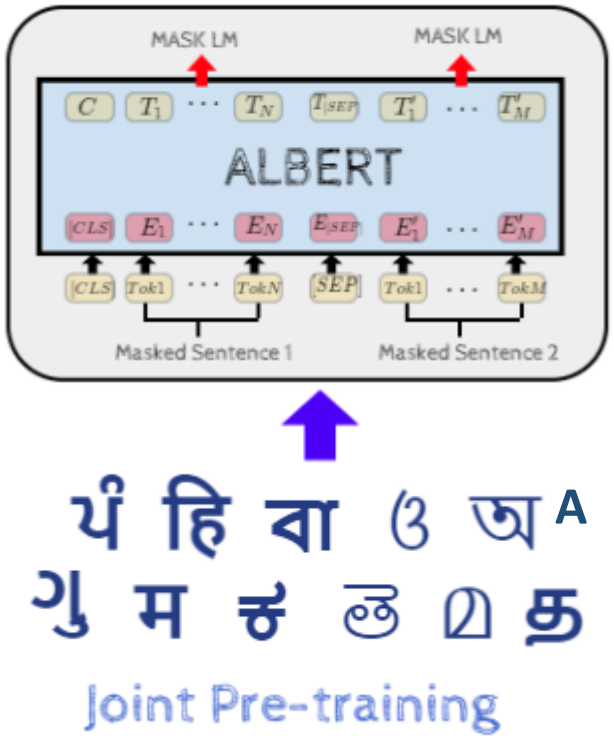
Our Approach



IndicBERT

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

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

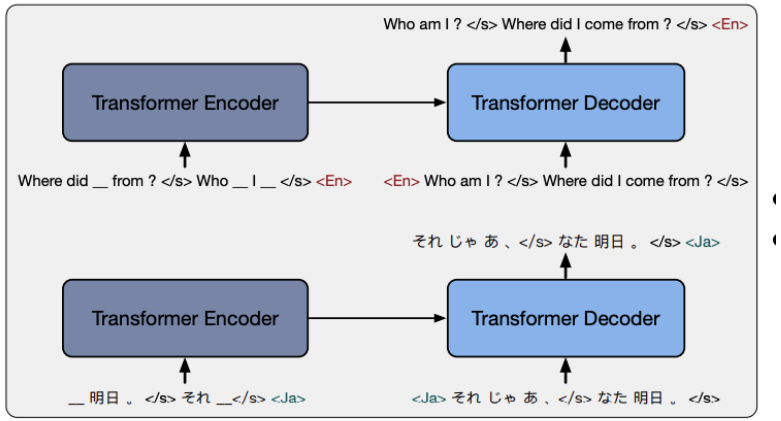


- Pre-trained Indic LM for NLU applications
- Large Indian language content (8B tokens)
  - 11 Indian languages
  - + Indian English content
- Multilingual Model
- Compact Model (~20m params)
- Competitive/better than mBERT/XLM-R
- Simplify fine-tune for your application
- 10k downloads per month on HuggingFace

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.

IndicBART

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



Multilingual Denoising Pre-Training (mBART)

- 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)
- Competitive with mBART50 for MT and summarization
- Simply fine-tune for your application

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

IndicTrans

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

https://indicnlp.ai4bharat.org/samanantar/

- Samanantar: Largest publicly available parallel corpus for Indian languages
  - English-Indian languages (11 language pairs, 49m sentence pairs)
  - Indian-Indian languages (110 language pairs, 80+ million sentence pairs)
- Large-scale mining of parallel corpora from web pages
- Multilingual Translation Model
  - State-of-the-art open-source model
- Deployed in the Supreme Court of India & Bangladesh

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. 2021.

Future Goals

for 22 languages

Full NLP stack
Text Generators
Inference Engines
Text Analysers
Input Tools

Translation

Dialog

Summarisation

....

QA

NLI

Paraphrase Detection

....

Named Entity Recognition

Sentiment Analysis

Topic Classification

Content Filters

....

Keyboards

Spell checkers

Standardise fonts

Summary

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

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.