

Natural Language Processing for Indian Languages

A Language Relatedness Perspective

Anoop Kunchukuttan

Microsoft AI & Research

Machine Translation & Speech Group, Hyderabad

ankunchu@microsoft.com



AI Deep Dive Workshop at NASSCOM DSAI-CoE, 30th August 2019

Outline

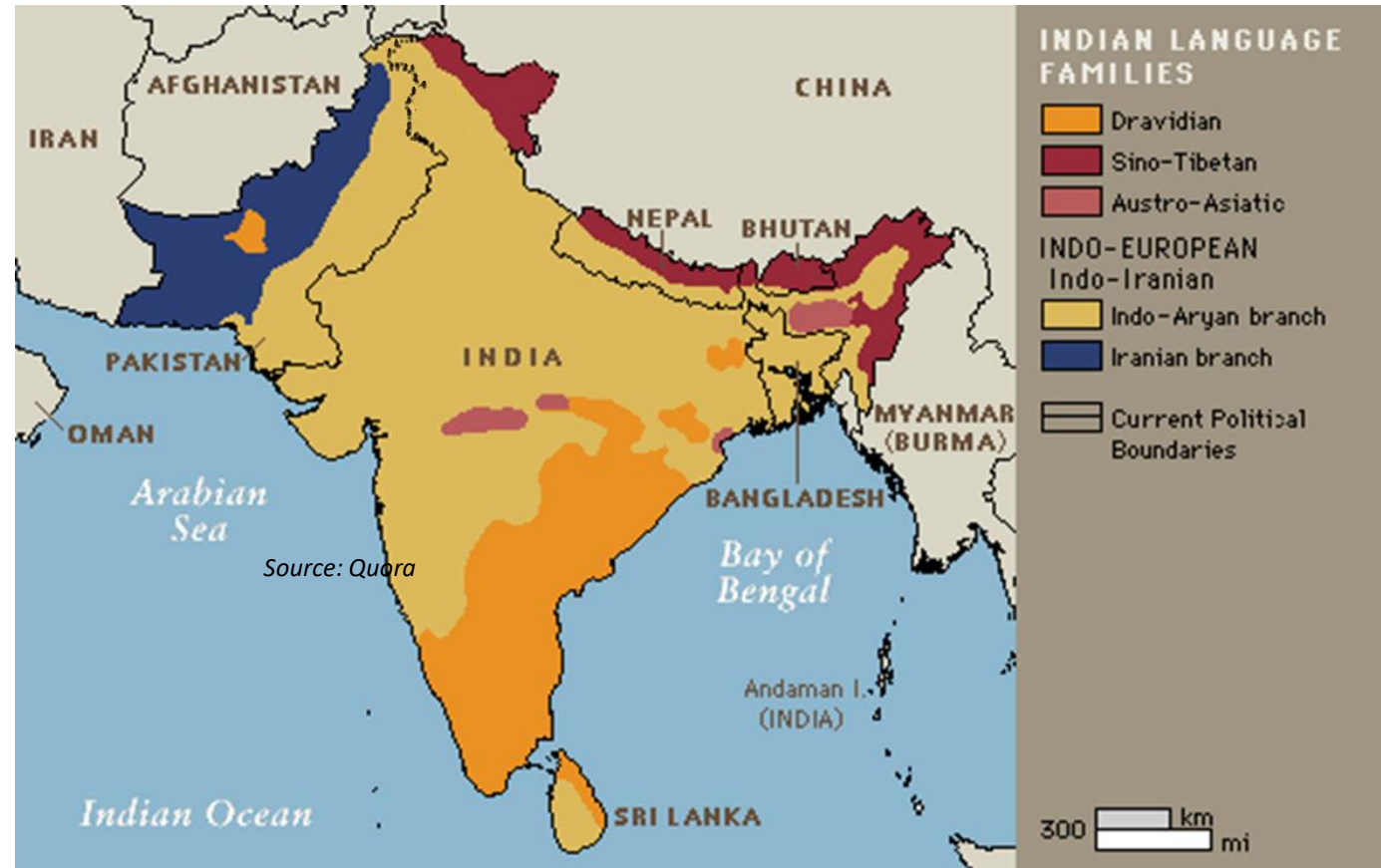
- **Motivation**
- Relatedness between Indian Languages
- Utilizing Relatedness between Indian Languages
- IndicNLP Library
- Datasets, Services and Standards
- Summary

Diversity of Indian Languages

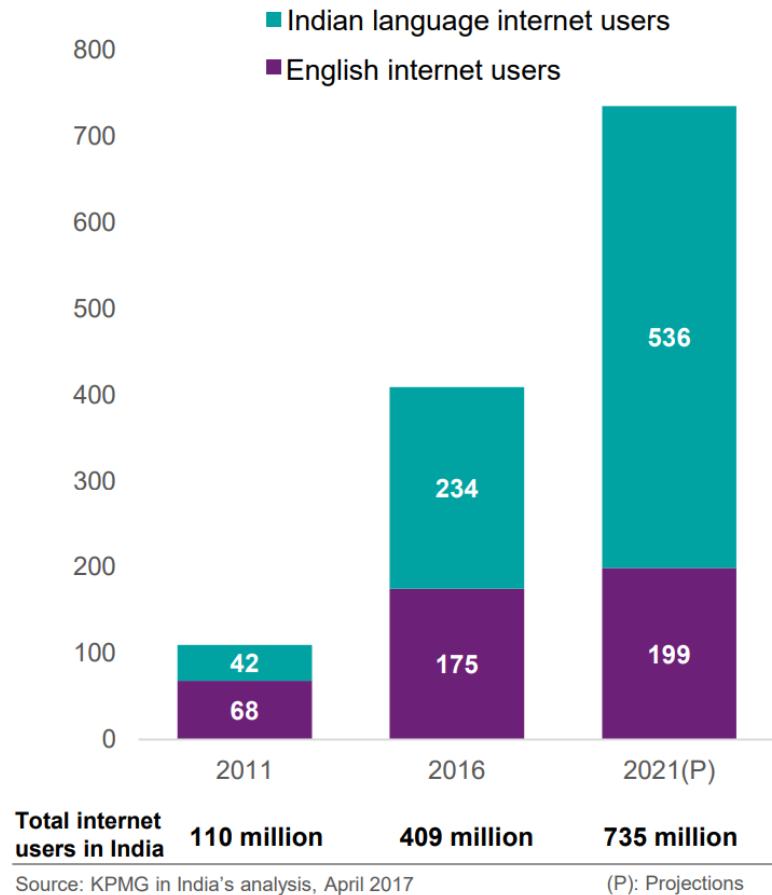
Highly multilingual country

Greenberg Diversity Index 0.9

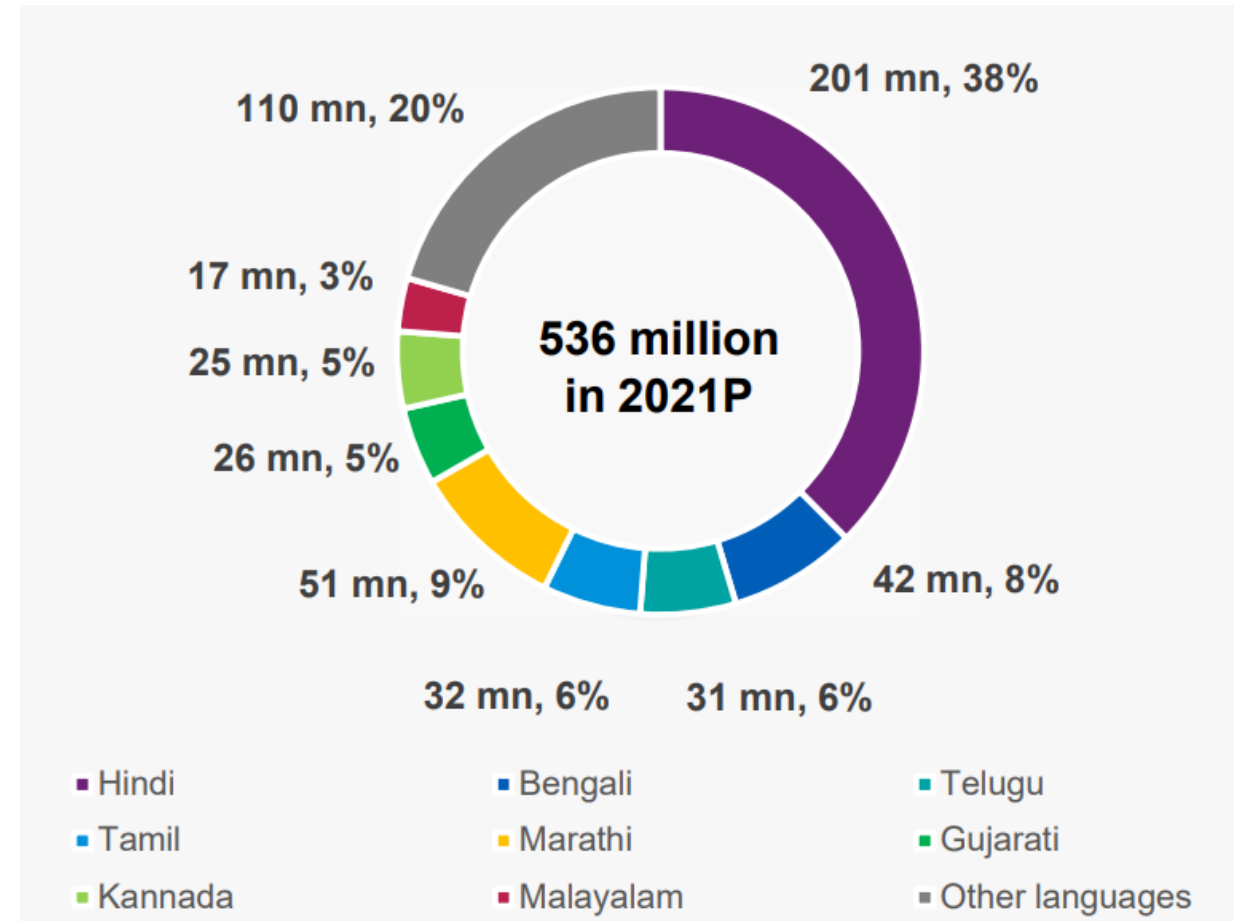
- 4 major language families
- 1600 dialects
- 22 scheduled languages
- 125 million English speakers
- 8 languages in the world's top 20 languages
- 11 languages with more than 25 million speakers
- 30 languages with more than 1 million speakers



Indian Languages on the Internet



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

Major Applications requiring Indian language support

Digital payments

E-tailing

Online government services

Digital classifieds

Chat applications

Digital entertainment

Social media platforms

Digital news

Digital write-ups

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?

Improving Indian Language Support

Applications and websites which support rich experiences:

Search

Recommendation

Translation

Question &
Answering

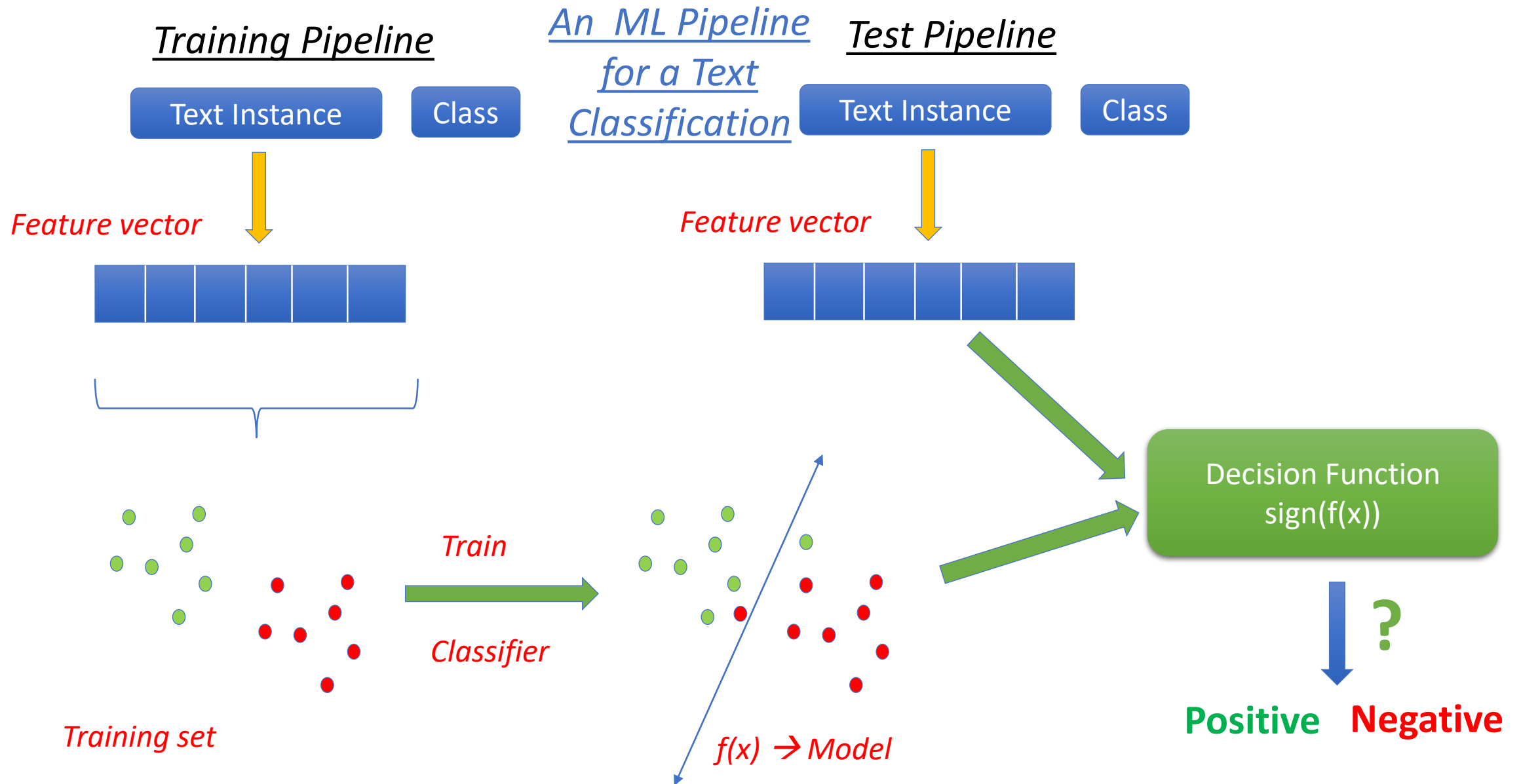
Information
Extraction &
Categorization

Transliteration

Entity
Identification

Entity Linking

Machine Learning is the dominant NLP Paradigm



Scalability Challenges in ML solutions

- NLP requires human expertise → difficult and expensive to replicate for every language
 - Annotated data
 - Linguistic knowledge inputs
- Expense cannot be justified for all languages
- Difficult to deploy and maintain systems for multiple languages

Let's look at examples of different kinds of annotations ...

Monolingual Corpora – easy to collect

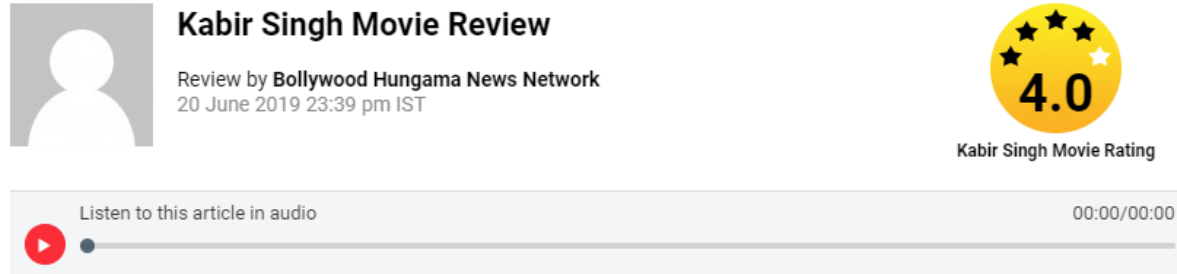
ನವದೆಹಲಿ: ಆರ್ಥಿಕ ಬೆಳವಣಿಗೆ ಕುಂಠಿತಗೊಳ್ಳುವುದನ್ನು ತಡೆಯಲು ಕೇಂದ್ರ ಸರ್ಕಾರವು ಹಲವು ಕ್ರಮಗಳಿಗೆ ಮುಂದಾಗಿದೆ. ಒಂದೇ ಬ್ರ್ಯಾಂಡ್ ಚಿಲ್ಲರೆ ಮಾರಾಟ, ಡಿಜಿಟಲ್ ಮಾಧ್ಯಮ, ಕಲ್ಲಿದ್ದಲು ಗಣಿಗಾರಿಕೆ ಮತ್ತು ತಯಾರಿಕೆ ಸೇರಿದಂತೆ ವಿವಿಧ ಕ್ಷೇತ್ರಗಳಲ್ಲಿ ವಿದೇಶಿ ನೇರ ಹೂಡಿಕೆಯ (ಎಫ್‌ಡಿಐ) ನಿಯಮಗಳನ್ನು ಸಡಿಲಿಸಿದೆ.

ಪ್ರಧಾನಿ ನರೇಂದ್ರ ಮೋದಿ ಅವರ ಅಧ್ಯಕ್ಷತೆಯಲ್ಲಿ ಬುಧವಾರ ನಡೆದ ಕೇಂದ್ರ ಸಚಿವ ಸಂಪುಟ ಸಭೆಯು ಹೂಡಿಕೆದಾರ ಸ್ನೇಹಿಯಾದ ಹಲವು ಕ್ರಮಗಳನ್ನು ಕೈಗೊಂಡಿದೆ. ಏಕ ಬ್ರ್ಯಾಂಡ್ ಚಿಲ್ಲರೆ ಮಾರಾಟ ಮಳಿಗೆಗಳಲ್ಲಿ ಶೇ 30ರಷ್ಟು ಭಾರತೀಯ ಸರಕುಗಳು ಇರಬೇಕು ಎಂಬ ನಿಯಮವನ್ನು ಸಡಿಲ ಮಾಡಲಾಗಿದೆ. ಹಾಗೆಯೇ, ಗುತ್ತಿಗೆ ತಯಾರಿಕೆ ಹಾಗೂ ಕಲ್ಲಿದ್ದಲು ಗಣಿಗಾರಿಕೆಯಲ್ಲಿ ನೂರರಷ್ಟು ಎಫ್‌ಡಿಐಗೂ ಅವಕಾಶ ನೀಡಲಾಗಿದೆ.

ಒಂದೇ ಬ್ರ್ಯಾಂಡ್ ಚಿಲ್ಲರೆ ಮಾರಾಟ ಮಳಿಗೆಗಳಲ್ಲಿ ಪ್ರತಿ ವರ್ಷ ಶೇ 30ರಷ್ಟು ಭಾರತೀಯ ಸರಕು ಮಾರಾಟವಾಗಬೇಕು ಎಂಬ ನಿಯಮ ಇದೆ. ಆದರೆ, ಐದು ವರ್ಷದ ವಹಿವಾಟು ಸರಾಸರಿಯಲ್ಲಿ ಶೇ 30ರಷ್ಟು ದೇಶೀಯ ಸರಕುಗಳು ಮಾರಾಟವಾದರೆ ಸಾಕು ಎಂದು ನಿಯಮವನ್ನು ಸರಳಗೊಳಿಸಲಾಗಿದೆ. ಹಾಗೆಯೇ, ಶೇ 30ರಷ್ಟು ಭಾರತೀಯ ಸರಕನ್ನು ರಫ್ತು ಮಾಡುವುದಕ್ಕೂ ಅವಕಾಶ ಕೊಡಲಾಗಿದೆ.

Digital Content available varies by languages

Sentiment Analysis - Simple Annotation



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

*Named Entity Annotation –
More time consuming, but does not require a lot of expertise*

Deutsche Telekom yesterday delayed the demerger of its T-Mobile wireless business, which owns UK network One2One, until next year owing to the volatility of the stock markets.

Deutsche Telekom was among a number of European telecom companies looking to demerge or float wireless operations this year. BT is expected to demerge its wireless business in the autumn.

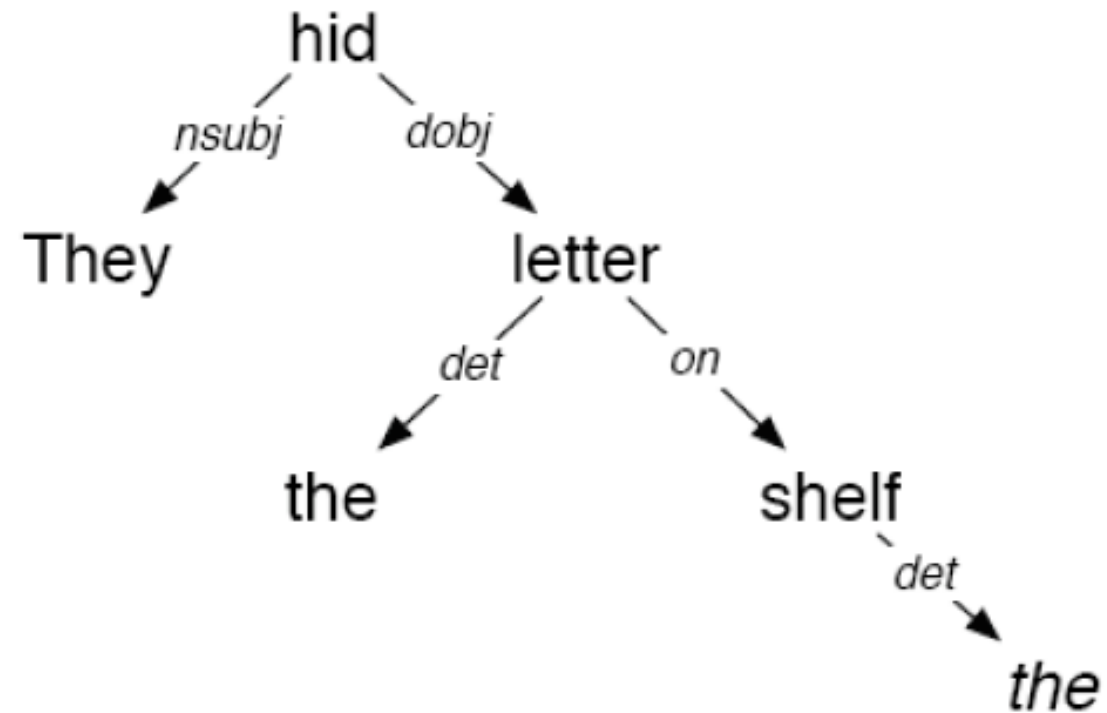
However Deutsche Telekom, still partly state-owned, warned yesterday that investor appetite for hi-tech shares is still too low. Analysts consider T-Mobile, which has 60m subscribers across Europe, to be one of Deutsche Telekom's most valuable assets and welcomed the decision to delay.

- Date
- FirstPerson
- JobTitle
- Location
- Lookup
- Money
- Organization
- Person

Parallel Corpora – large requirement, needs good language skills

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

Parse Tree – needs good linguistic expertise



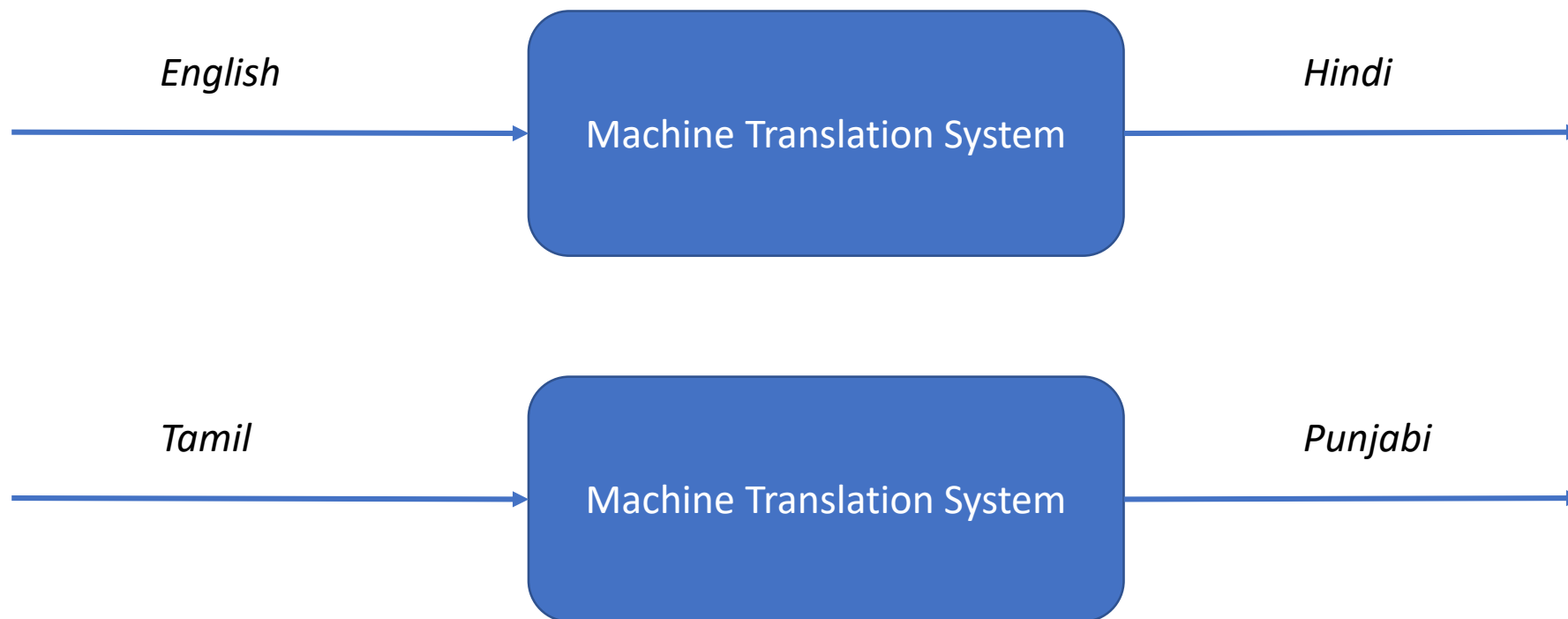
They hid the letter on the shelf

Need for a Unified Approach for Indic NLP

Expensive to create datasets for each language

- Can we utilize resources developed for some languages for other languages?
- Can diverse input from different languages lead to better generalization?
- Can we support multiple languages with reduced effort & cost for deployment and maintenance?
- Can we use unsupervised data sources?
- ***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?

Linguistic
Underpinnings of
Relatedness

Standards

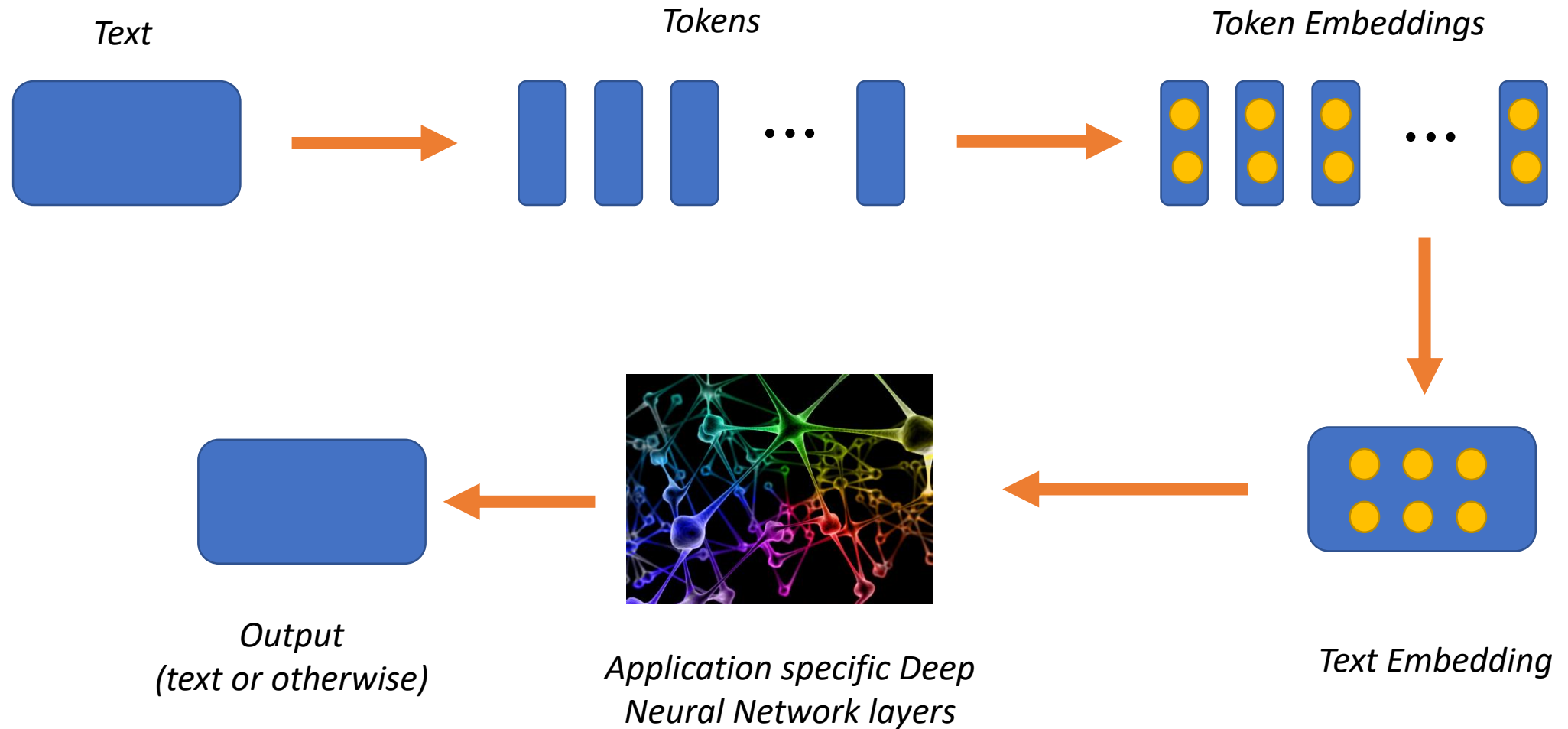
**Unified
Approaches to
Indic NLP**

Algorithms &
Methods

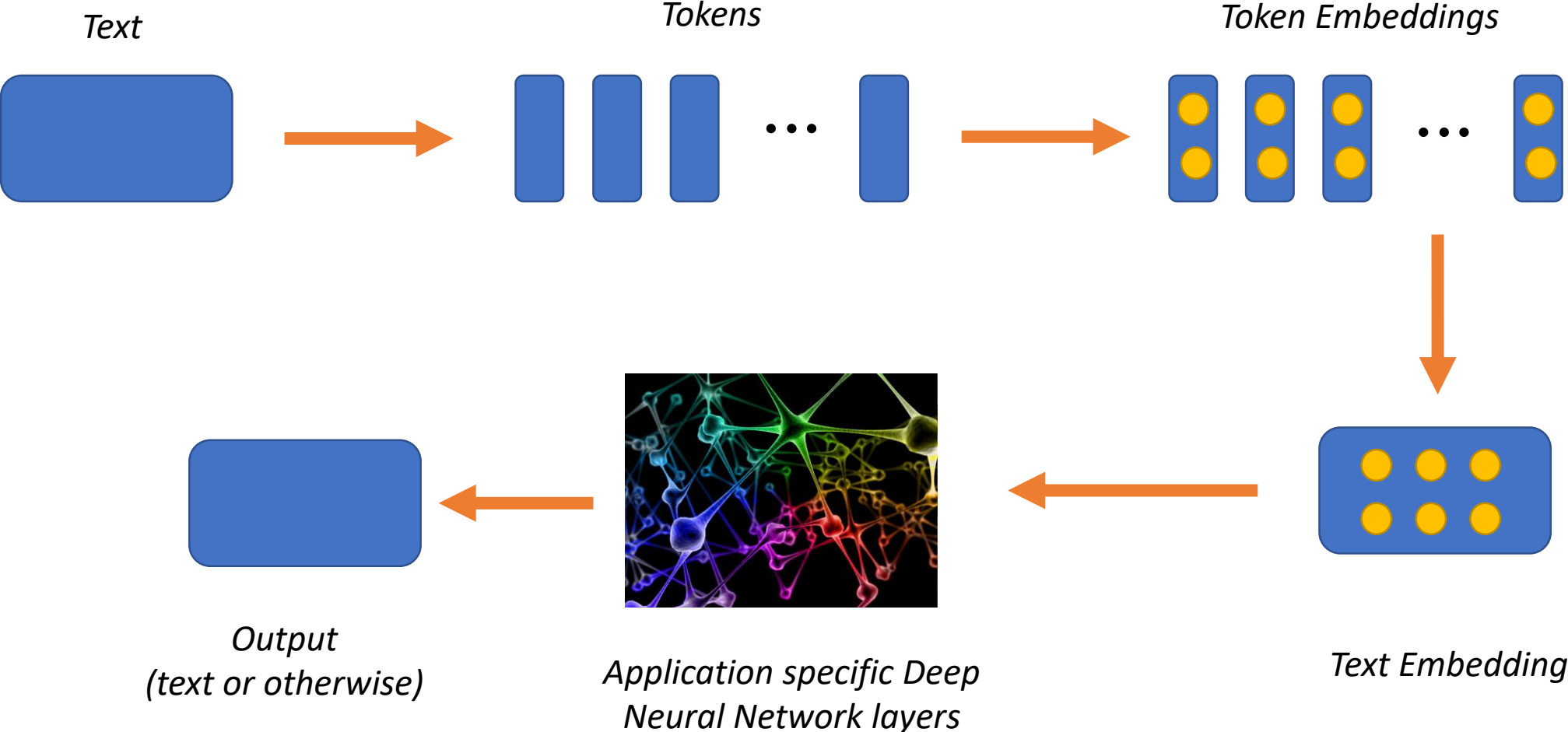
Services/APIs

Datasets

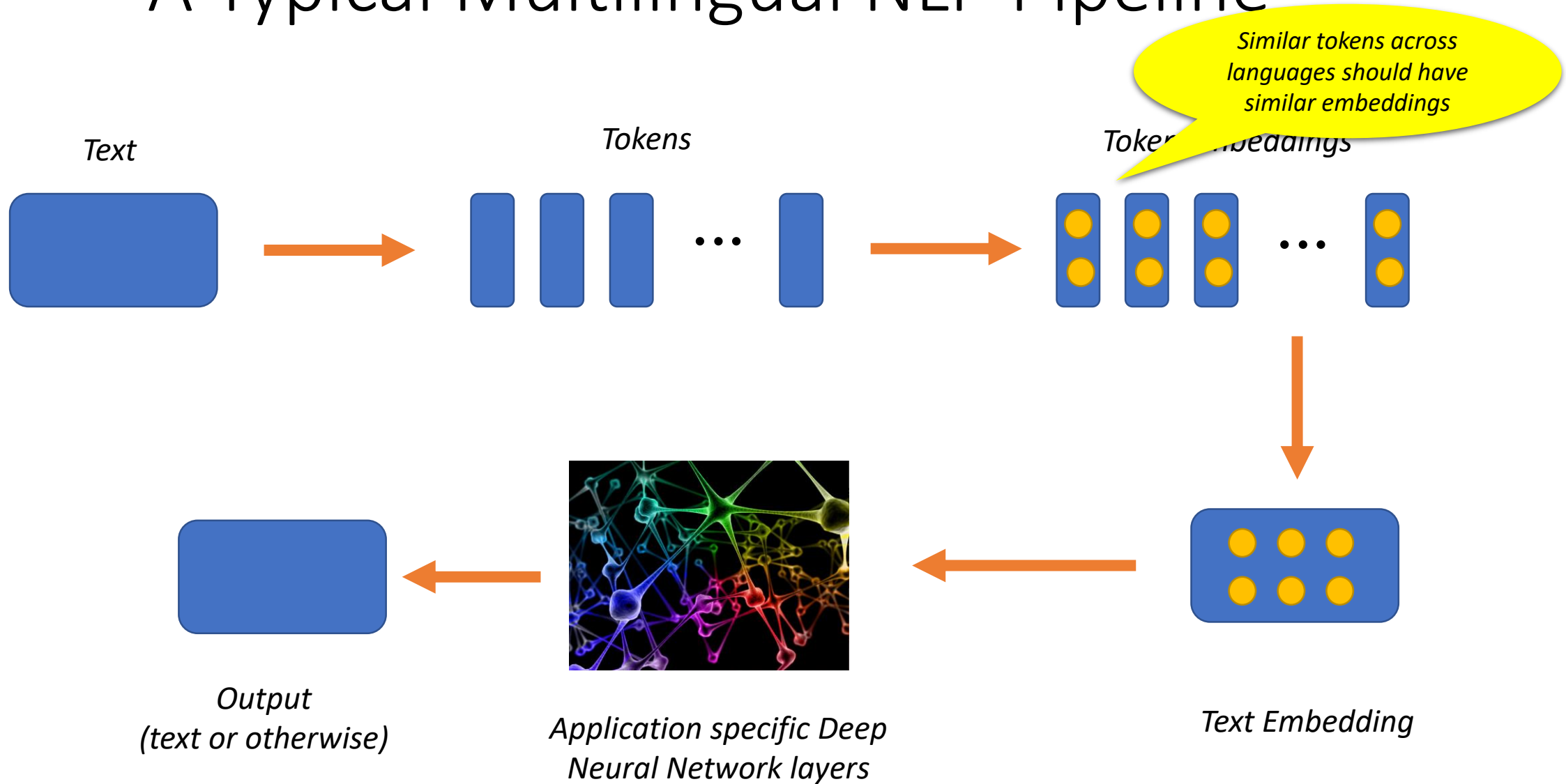
A Typical Deep Learning NLP Pipeline



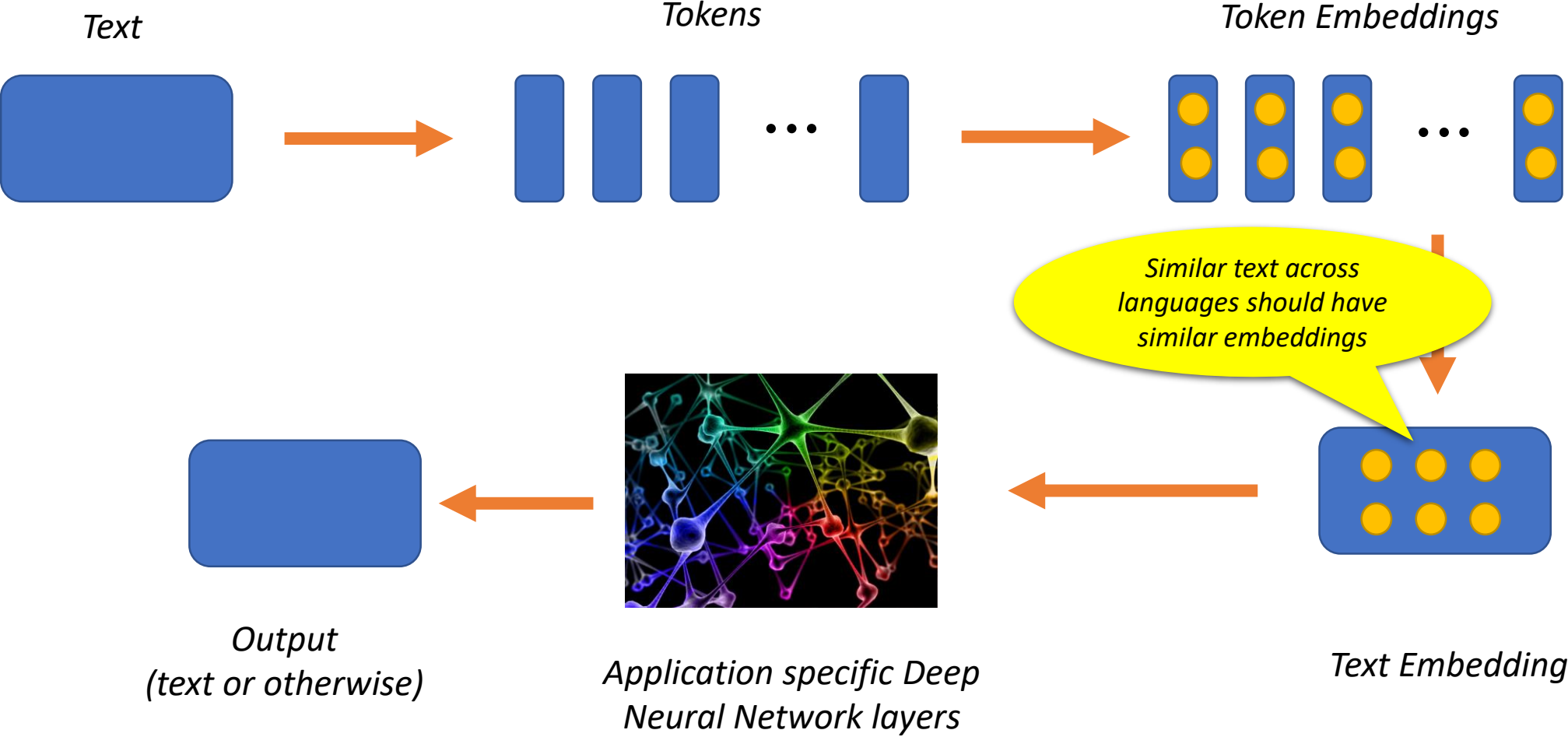
How do we transfer information across languages?



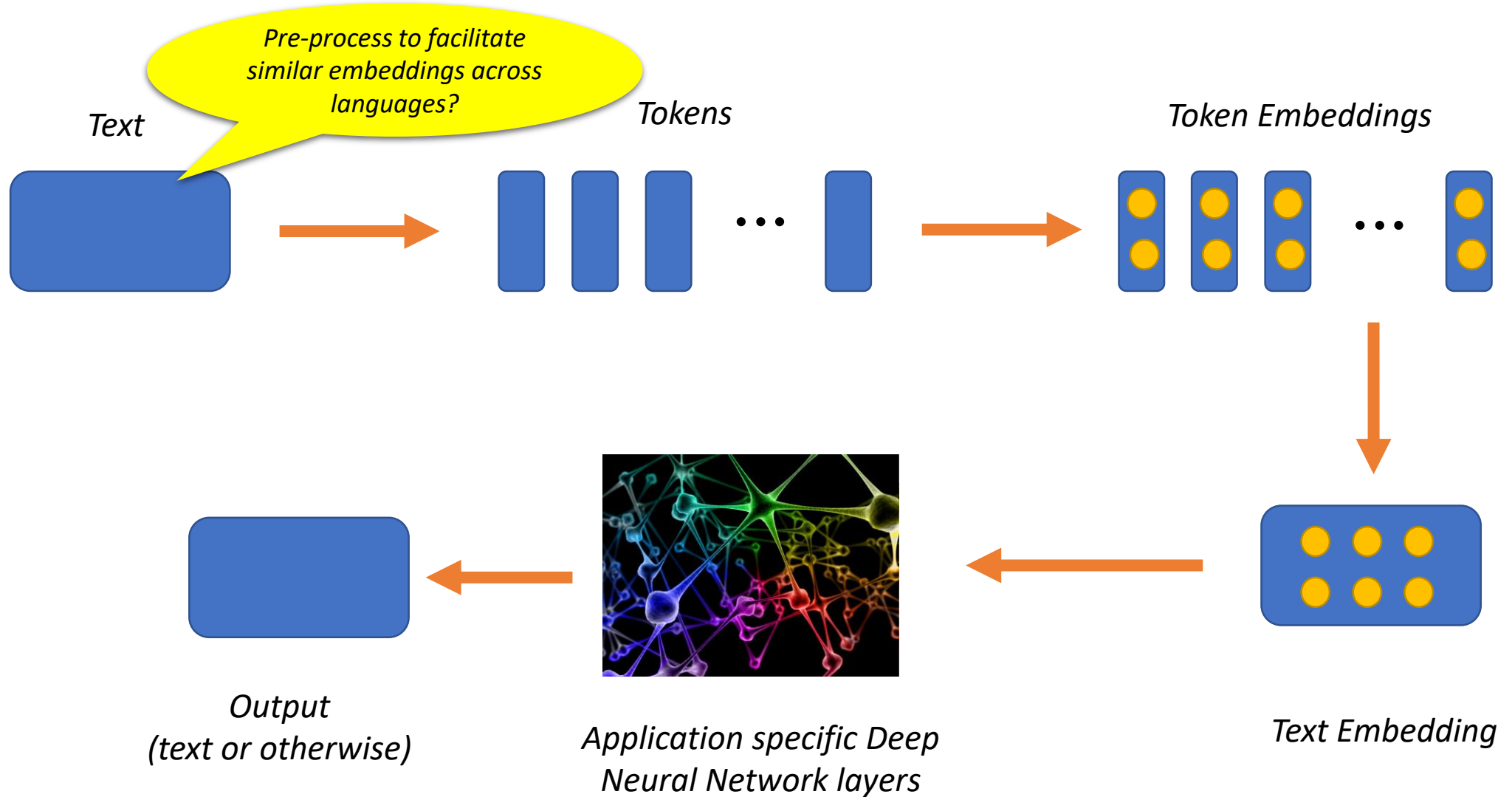
A Typical Multilingual NLP Pipeline



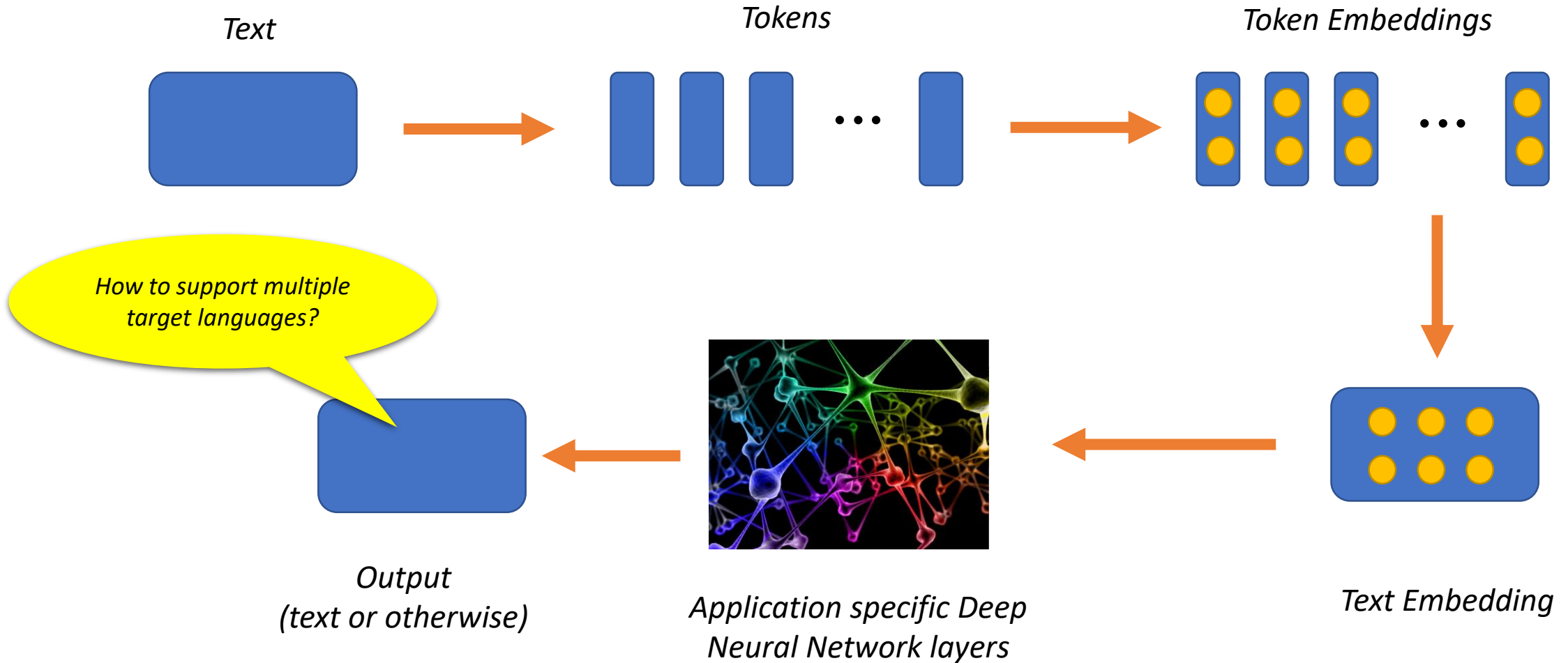
A Typical Multilingual NLP Pipeline



A Typical Multilingual NLP Pipeline



A Typical Multilingual NLP Pipeline



Outline

- Motivation
- **Relatedness between Indian Languages**
- Utilizing Relatedness between Indian Languages
- IndicNLP Library
- Datasets, Services and Standards
- Summary

Relatedness between Indian Languages

Why are Indian languages related?

Related Languages

```
graph TD; A[Related Languages] --> B[Related by Genealogy]; A --> C[Related by Contact]; B --> D[Language Families]; D --- E[Dravidian, Indo-European, Turkic]; E --- F["(Jones, Rasmus, Verner, 18th & 19th centuries, Raymond ed. (2005))"]; C --> G[Linguistic Areas]; G --- H[Indian Subcontinent, Standard Average European]; H --- I["(Trubetzkoy, 1923)"]; J["Related languages may not belong to the same language family!"]
```

Related by Genealogy



Language Families

Dravidian, Indo-European, Turkic

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

Related by Contact



Linguistic Areas

Indian Subcontinent,
Standard Average European

(Trubetzkoy, 1923)

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

	Sanskrit	Greek	Latin
'father'	<i>pitā</i>	<i>patēr</i>	<i>pater</i>
'foot'	<i>pad-</i>	<i>pod-</i>	<i>ped-</i>
'blood'	<i>krūra-</i>	<i>kreas</i>	<i>cruor</i>
'three'	<i>trayah</i>	<i>treis</i>	<i>trēs</i>
'that'	<i>tad</i>	<i>to</i>	<i>-tud</i>

Basis of classification

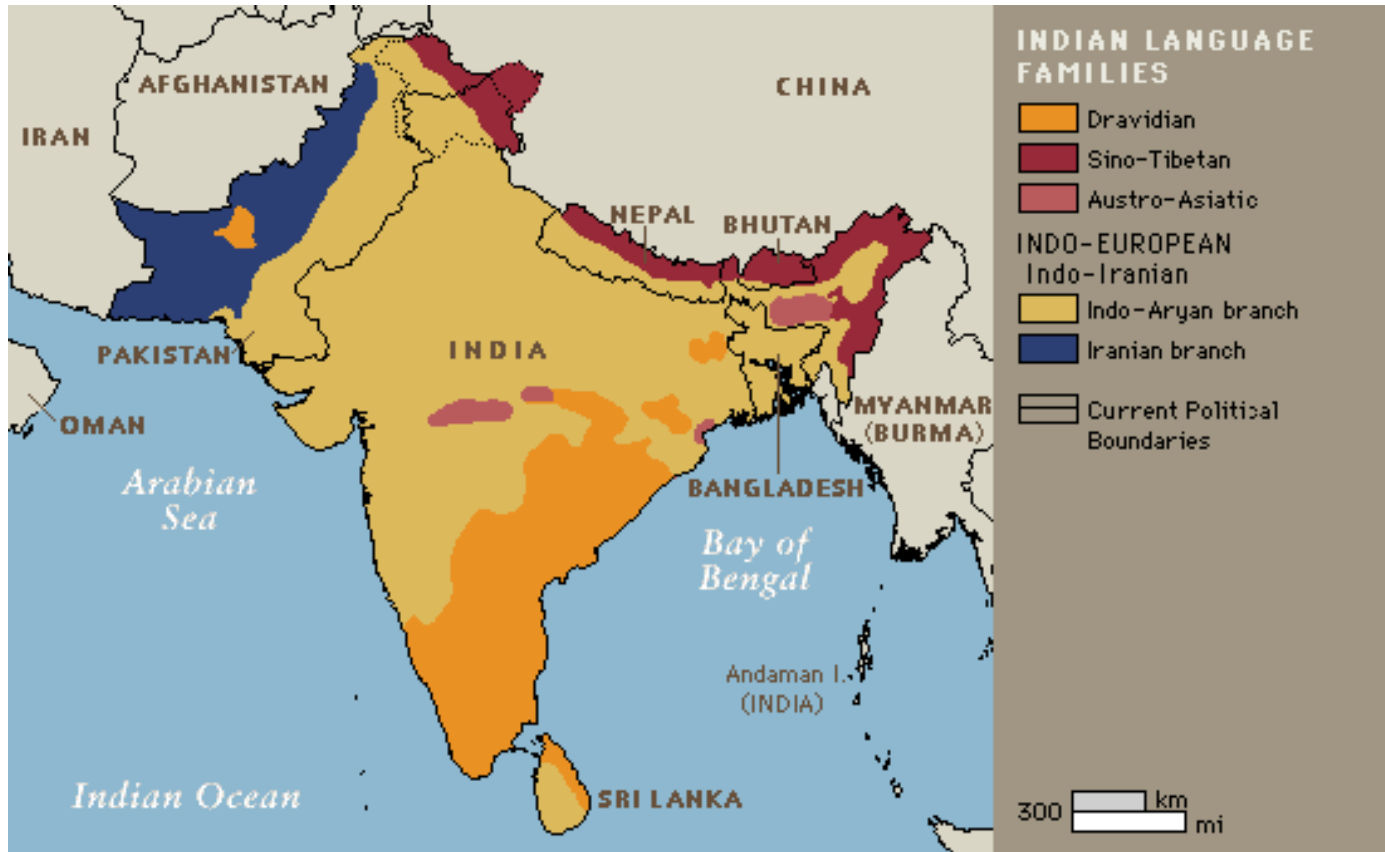
Regularity of sound change is the basis of studying genetic relationships

MEANING	LATIN	PORTUGUESE ²	CASTILIAN	ITALIAN	ROMANIAN
'eight'	<i>octo</i> /'okto:/□	<i>oito</i> /'ojtu/□	<i>ocho</i> /'otʃo/□	<i>otto</i> /'otto/□	<i>opt</i> /'opt/□
'milk'	<i>lactem</i> /'laktẽ/□	<i>leite</i> /'lɛjtə/□	<i>leche</i> /'letʃe/□	<i>latte</i> /'latte/□	<i>lapte</i> /'lapte/□
'fact'	<i>factum</i> /'faktũ/□	<i>feito</i> /'fɛjtu/□	<i>hecho</i> /'etʃo/□	<i>fatto</i> /'fatto/□	<i>fapt</i> /'fapt/□

Source: Eifring & Theil (2005)

*These words are called **cognates***

Language Families in India



4 major language families

Indo-Aryan: North India and Sri Lanka (branch of Indo-European)

Dravidian: South India & pockets in the North

Tibeto-Burman: North-East and along the Himalayan ranges

Austro-Asiatic: pockets in Central India, North-East, Nicobar Islands

Plus

Andamanese family

Unknown language of the Sentinelese

Cognates 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
language	bhāshā, vāNī	bhāshā, zabān	boli, zabān, pasha	bhāshā	bhāshā	bhāsā	bhasha
ten	Dasha	Das	das, daha	das	dahā	dasa	dôsh

Dravidian

English	Tamil	Malayalam	Kannada	Telugu
fruit	pazham , kanni	pazha.n , phala.n	haNNu , phala	pa.nDu , phala.n
fish	mInn	matsya.n , mln, mlna.n	mInu , matsya , jalavAsi, mlna	cepalu , matsyalu , jalaba.ndhu
hunger	paci	vishapp , udarArtti , kShutt , pashi	hasivu, hasiv.e,	Akali
language	pAShai, m.ozhi	bhASha , m.ozhi	bhASh.e	bhAShA , paluku
ten	pattu	patt,dasha.m,dashaka.m	hattu	padi

Source: Wikipedia and
IndoWordNet

Linguistic Area (*Sprachbund*)

- To the layperson, Dravidian & Indo-Aryan languages would seem closer to each other than English & Indo-Aryan
- **Linguistic Area:** A group of languages (at least 3) that have common structural features due to geographical proximity and language contact
(Thomason 2000)
- Not all features may be shared by all languages in the linguistic area

Examples of linguistic areas:

- **Indian Subcontinent** (*Emeneau, 1956; Subbarao, 2012*)
- Balkans

Borrowed Words

Indo-Aryan words in Dravidian languages

Most classical languages borrow heavily from Sanskrit

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

Dravidian words in Indo-Aryan languages

- A matter of great debate
- Could probably be of Munda origin also
- See writings of Kuiper, Witzel, Zvelebil, Burrow, etc.
- Proposal of Dravidian borrowing even in early Rg Vedic texts

Borrowed Syntactic Features

Retroflex Sounds in Indo-Aryan Languages: ट ठ ड ढ ण

- Found in Indo-Aryan, Dravidian and Munda language families
- Not found in Indo-European languages outside the Indo-Aryan branch

Echo Words: Generally means *etcetera* or *things like this*

hi: *cAya-vAya*, **te:** *pull-gull*, **ta** *v.elai-k.elai*

- Standard feature in all Dravidian languages
- Not found in Indo-European languages outside the Indo-Aryan branch

SOV word order in Munda languages

- Their Mon-Khmer cousins have SVO word order
- Munda language were originally SVO, but have become SOV over time

Similarities between Indian languages

Key Similarities between related languages

भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला

bhAratAcyA svAta.ntryadinAnimitta ameriketIla IOsa enjalsa shaharAta kAryakrama Ayojita karaNyAta AIA

Marathi

भारता च्या स्वातंत्र्य दिना निमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला

bhAratA cyA svAta.ntrya dinA nimitta amerike tIla IOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AIA

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

Lexical Similarity

(Words having similar **form** and **meaning**)

- *Cognates*

a common etymological origin

<i>roTI (hi)</i>	<i>roTIA (pa)</i>	<i>bread</i>
<i>bhai (hi)</i>	<i>bhAU (mr)</i>	<i>brother</i>

- *Loan Words*

borrowed without translation

<i>matsya (sa)</i>	<i>matsyalu (te)</i>	<i>fish</i>
<i>pazha.m (ta)</i>	<i>phala (hi)</i>	<i>fruit</i>

- *Named Entities*

do not change across languages

<i>mu.mbal (hi)</i>	<i>mu.mbal (pa)</i>	<i>mu.mbal (pa)</i>
<i>keral (hi)</i>	<i>k.eraLA (ml)</i>	<i>keraL (mr)</i>

- *Fixed Expressions/Idioms*

MWE with non-compositional semantics

<i>dAla me.n kuCha kAlA honA</i>	<i>(hi)</i>	<i>Something fishy</i>
<i>dALa mA kAlka kALu hovu</i>	<i>(gu)</i>	

Enables sharing of data across languages

But, be warned of

False Friends: Similar spelling ; different meaning

- *Different origin: pAnl (hi) [water] → panl (ml) [fever]*
- *Semantic shift: bala means hair (hi, frequent sense) and baLa means child (mr)*

Short words:

jaLa ← → jAla

How similar are Indian Languages?

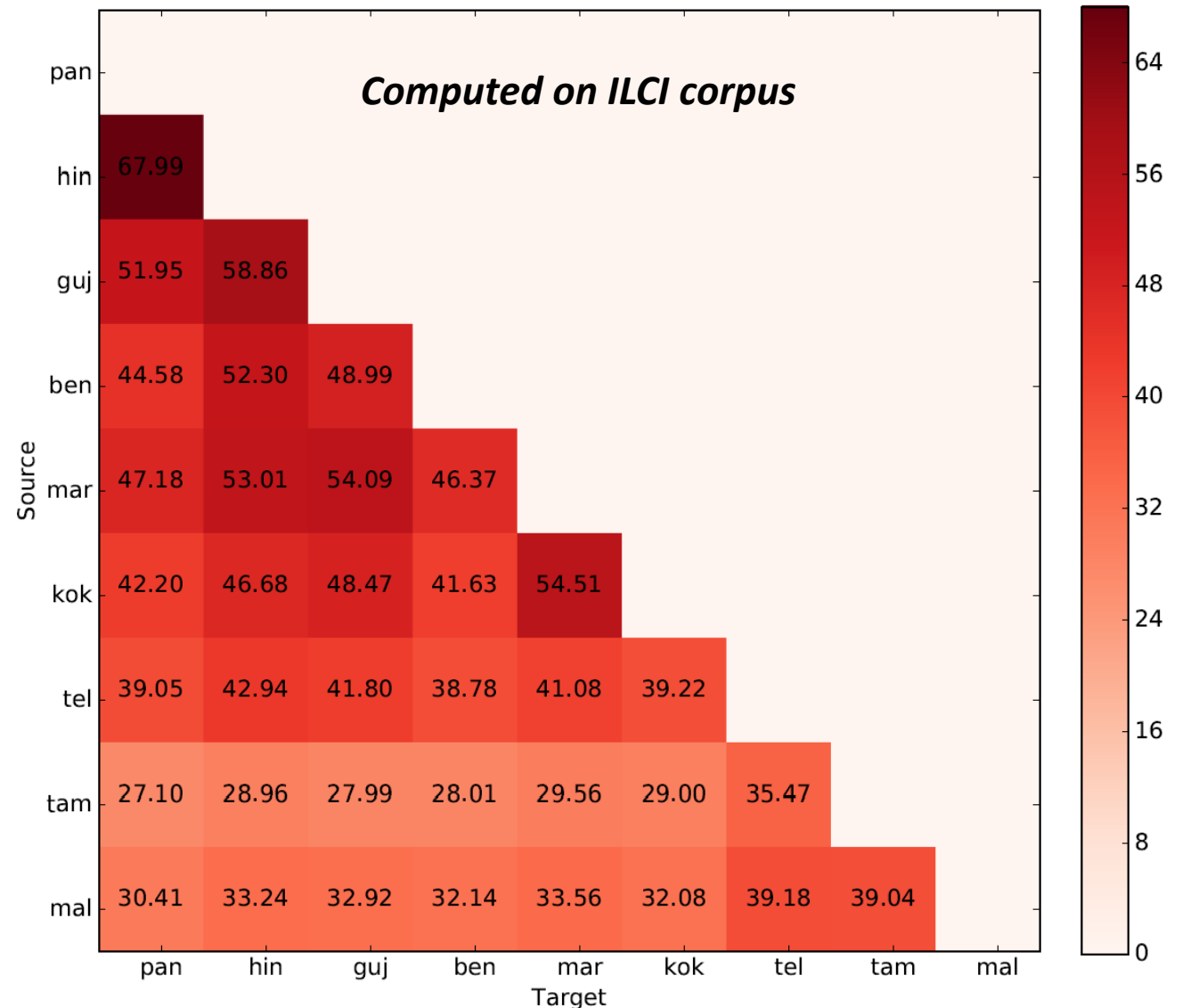
Estimate lexical similarity from parallel corpus

Longest Common Subsequence Ratio (LCSR)
for a sentence pair

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

LCSR for a language pair

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



Syntactic Similarity

- Almost all Indian languages has SOV word order
- SOV word order determines relative order between:
 - Noun-adposition
 - Noun-genitive
 - Noun-Relative clause
 - Verb-Auxiliary
- Word order plays a very important role in most NLP applications
 - Language Modelling
 - Machine Translation
- Relatively Free Word Order

Morphological Similarity

- Inflectionally rich
- Sometimes agglutinative
- Function words with largely 1-1 correspondence
- Similar internal word structure and compositional semantics
- Similar case-marking systems

Hindi Post-position	Marathi Suffix	Case Description
को (<i>ko</i>)	ला (<i>la</i>)	Accusative
को (<i>ko</i>)	ला (<i>la</i>)	Dative
से (<i>se</i>)	नी (<i>ni</i>)	Instrumental
में (<i>me</i>)	त (<i>ta</i>)	Locative
का (<i>ka</i>)	चा (<i>ca</i>)	Genitive

Orthographic Similarity

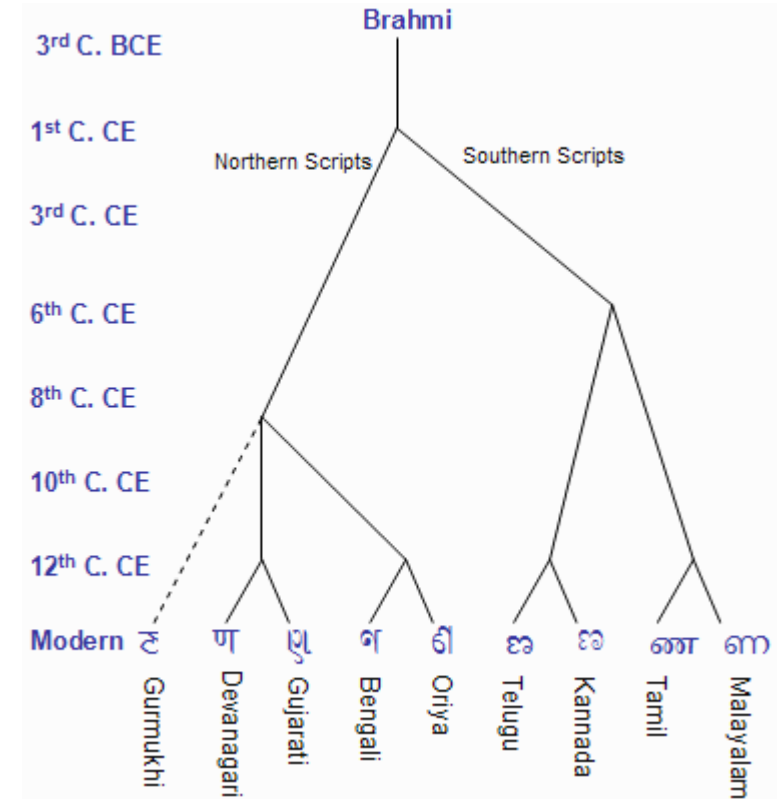
- *highly overlapping phoneme sets*
- *mutually compatible orthographic systems*
- *similar grapheme to phoneme mappings*

Indic Scripts



*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

Common characteristics

Devanagari	अ आ इ ई उ ऊ ऋ ॠ एँ ऐ ए ऐ आँ औ ओ औ क ख ग घ ङ च छ ज झ
Bengali	অ আ ই ঐ উ ঊ ঋ ৠ এ ঐ ও ঔ ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড
Gurmukhi	ਅ ਆ ਇ ਈ ਉ ਊ ਏ ਐ ਓ ਔ ਕ ਖ ਗ ਘ ਙ ਚ ਛ ਜ ਝ ਵ ਟ ਠ ਡ ਢ ਣ ਤ ਥ
Gujarati	અ આ ઇ ઈ ઉ ઊ ઋ ઋ ઌ ઌ ઐ ઐ ઔ ઔ ક ખ ગ ઘ ઙ ઙ ઙ ઙ ઙ ઙ ઙ ઙ ઙ ઙ
Oriya	ଅ ଥା ଇ ଈ ଊ ଋ ଠ ଣ ଡ ଢ ଣ ଟ ଠ ଡ ଢ ଣ ଟ ଠ ଡ ଢ ଣ ଟ ଠ ଡ ଢ ଣ
Tamil	அ ஆ இ ஐ உ ஊ ஏ ஏ ஐ ஒ ஓ ஔ க ங ச ங ஞ ட ண த ந
Telugu	అ ఆ ఇ ఈ ఉ ఊ ఋ ఠ ణ డ ఢ ణ ట ఠ డ ఢ ణ ట ఠ డ ఢ ణ
Kannada	ಅ ಆ ಇ ಈ ಉ ಊ ಯ ಋ ಏ ಏ ಐ ಒ ಓ ಔ ಕ ಖ ಗ ಘ ಜ ಚ ಛ ಜ ರು ಣ
Malayalam	അ അ ഇ ഇയ ഉ ഉയ ള ണ ഏ ഏ ഐ ഐ ഓ ഓ ഔ ഔ ക ഖ ഗ ങ

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)

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

Syllable as Basic Unit

akshara, the fundamental organizing principle of Indian scripts

(CONSONANT) + VOWEL

Examples: की (ki), प्रे (pre)

Pseudo-Syllable

True Syllable ⇒ Onset, Nucleus and Coda

Orthographic Syllable ⇒ Onset, Nucleus

Primary vowels

	Short		1 Long		Diphthongs			
	Initial	Diacritic	Initial	Diacritic	Initial	Diacritic		
Unrounded low central	अ	a	पा	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	ऌ	ḷ	पृ	pṛ
	ऌ	ḷ	पृ	pṛ	ऍ	ḥ	पृ	pṛ

Secondary vowels

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

Organized as per sound phonetic principles

shows various symmetries

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
2 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
	6 Sonorants	य	रा	ल
Sibilants	श	ष	स	

Other letters

ह	ha	ळ	ḷa
---	----	---	----

Benefits due to script design

- Common design and standardization enables easy conversion from one script to another
- Makes exploiting lexical similarity possible
- Phonetic scripts: helps capture similarity between characters

	0A8	0A9	0AA	0AB	0AC	0AD	0AE
0		ॐ	ॐ	ॐ	ॐ	ॐ	ॐ
1	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ
2	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ
3	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ
4	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ
5	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ

	098	099	09A	09B	09C	09D	09E
0	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ
1	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ
2	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ
3	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ
4	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ
5	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ	ॐ

Feature	Possible Values
Type	Unused (0), Vowel modifier (1), Nukta (2), Halant (3), Vowel (4), Consonant (5), Number (6), Punctuation (7)
Height (vowels)	Front (1), Mid (2), Back (3)
Length	Short (1), Medium (2), Long (3)
Svar1	Low (1), Lower Middle (2), Upper Middle (3), Lower High (4), High (5)
Svar2	Samvrit (1), Ardh-Samvrit (2) Ardh-Vivrit (3), Vivrit (4)
Sthaan (place)	Dvayoshthya (1), Dantoshthya (2), Dantya (3), Varstya (4), Talavya (5) Murdhanya (6), Komal-Talavya (7), Jivhaa-Muliya (8), Svaryantramukhi (9)
Prayatna (manner)	Sparsha (1), Nasikya (2), Parshvika (3), Prakampi (4), Sangharshi (5), Ardh-Svar (6)

Source: Singh, 2006

Useful for natural language processing: transliteration, speech recognition, text-to-speech

The Periodic Table and Indic Scripts

Dmitri Mendeleev is said to have been inspired by the two-dimensional organization of Indic scripts to create the periodic table

<http://swarajyamag.com/ideas/sanskrit-and-mendeleevs-periodic-table-of-elements/>

The Full List of Mendeleev's Predictions with their Sanskrit Names

<i>Mendeleev's Given Name</i>	<i>Modern Name</i>
<i>Eka-aluminium</i>	Gallium
<i>Eka-boron</i>	Scandium
<i>Eka-silicon</i>	Germanium
<i>Eka-manganese</i>	Technetium
<i>Tri-manganese</i>	Rhenium
<i>Dvi-tellurium</i>	Polonium
<i>Dvi-caesium</i>	Francium
<i>Eka-tantalum</i>	Protactinium

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

- Motivation
- Relatedness between Indian Languages
- **Utilizing Relatedness between Indian Languages**
- IndicNLP Library
- Datasets, Services and Standards
- 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

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

	0A8	0A9	0AA	0AB	0AC	0AD	0AE
0	ঐ	ঔ	ঠ	ড	ণ	ত	থ
1	ঐ	ঔ	ঠ	ড	ণ	ত	থ
2	ঐ	ঔ	ঠ	ড	ণ	ত	থ
3	ঐ	ঔ	ঠ	ড	ণ	ত	থ
4	ঐ	ঔ	ঠ	ড	ণ	ত	থ
5	ঐ	ঔ	ঠ	ড	ণ	ত	থ

	098	099	09A	09B	09C	09D	09E
0	৐	ঐ	ঔ	ঠ	ড	ণ	ত
1	ঐ	ঔ	ঠ	ড	ণ	ত	থ
2	ঐ	ঔ	ঠ	ড	ণ	ত	থ
3	ঐ	ঔ	ঠ	ড	ণ	ত	থ
4	ঐ	ঔ	ঠ	ড	ণ	ত	থ
5	ঐ	ঔ	ঠ	ড	ণ	ত	থ

केरला

কেরলা

కేరలా

Multilingual Acronym Generation

Simple application of script conversion

Need to build Latin to Indic script mappings only once

ACL

ए सी एल

এ সী এল

ಎ ಸೀ ಎಲ

Multilingual Transliteration

Hindi → English corpus

Bengali → English corpus

Telugu → English corpus

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

Concat training sets

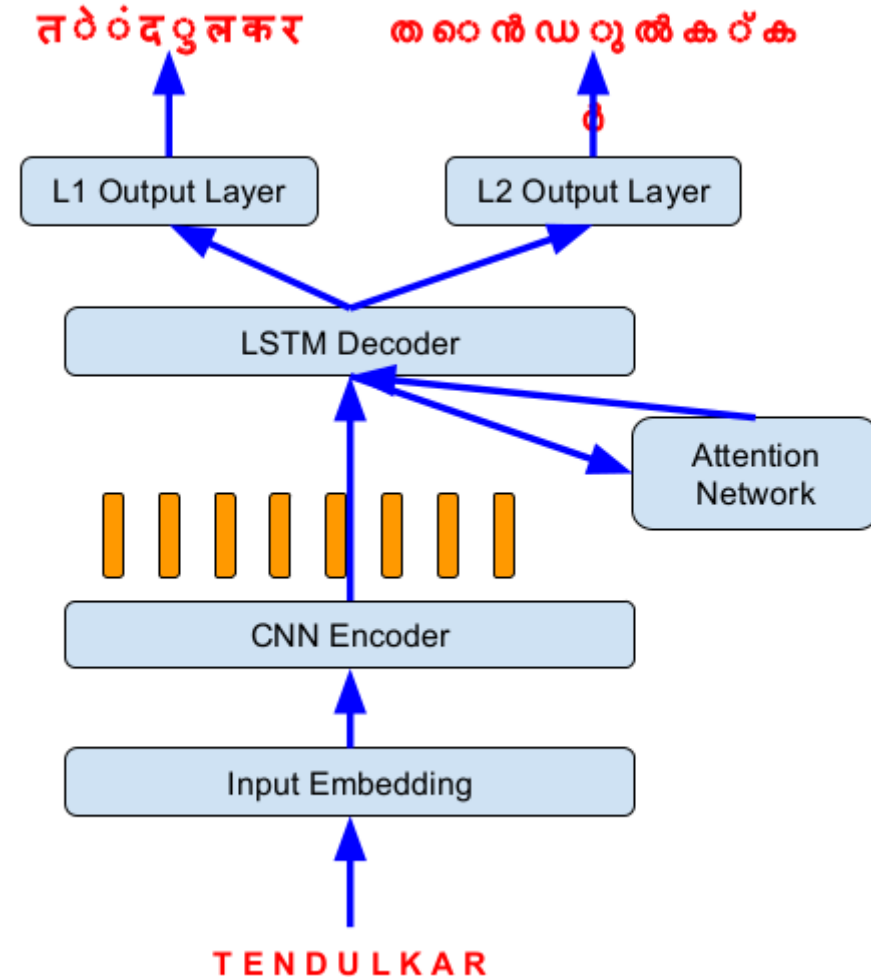
Malayalam	കോഴിക്കോട്	kozhikode
Hindi	केरल	kerala
Kannada	ಬೆಂಗಳೂರು	bengaluru

Convert to a common script

Malayalam	कोळिक्कोट्	kozhikode
Hindi	केरल	kerala
Kannada	बेंगळूरु	bengaluru

Share network parameters across languages

Output layer for each target language



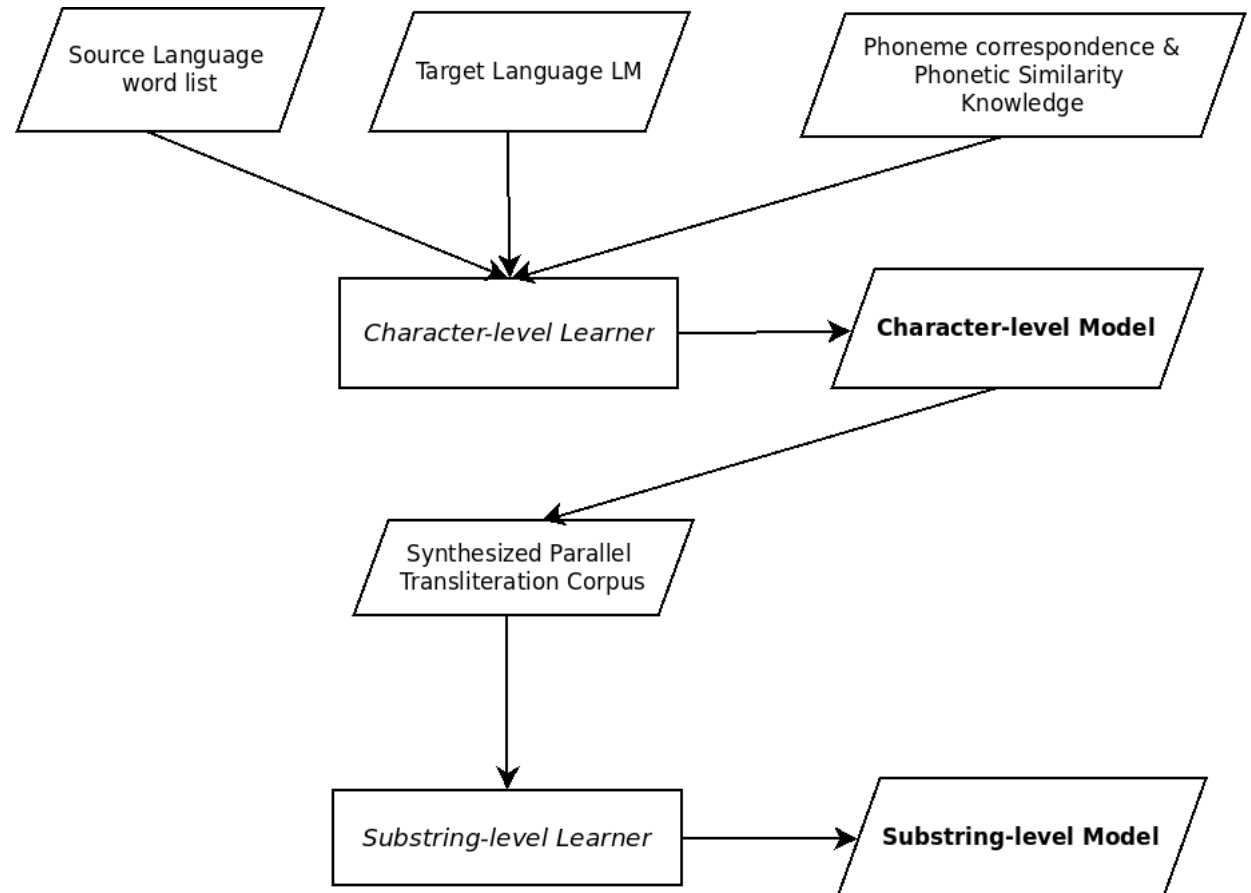
Unsupervised Transliteration

Learn:

Transliteration model (T_x) for source language (F) to target language (E)

Inputs:

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

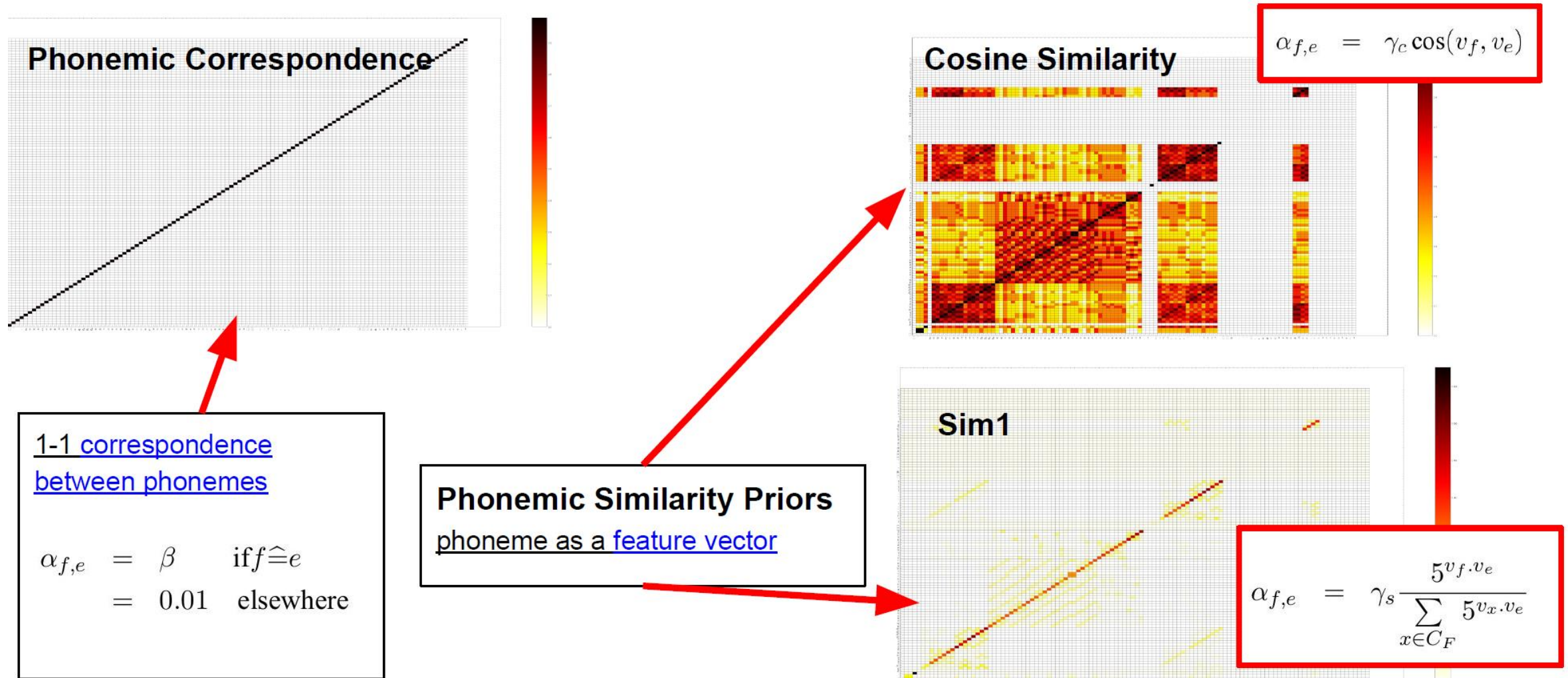


- *Represent each Indic character as a feature vector*
- *Define a similarity measure based on the feature vector*

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

Linguistically-informed phonetic priors

Priors capture how similar two characters are (use Dirichlet Priors)



Effect of linguistic priors

Method	ben-hin			hin-kan			kan-hin			tam-kan		
	A_1	F_1	A_{10}	A_1	F_1	A_{10}	A_1	F_1	A_{10}	A_1	F_1	A_{10}
Ravi & Knight (2009)	12.72	68.95	18.94	0.00	44.76	0.07	0.20	48.84	0.54	0.00	44.46	0.27
Rule-based	16.13	74.60	16.13	13.75	79.67	13.75	12.90	79.29	12.90	10.25	68.49	10.25
<i>Phonemic Correspondence Initialization + Prior:</i>												
Correspondence	18.27	75.50	27.04	12.53	77.32	17.89	27.69	81.06	43.55	13.49	69.85	29.06
Cosine	17.74	75.09	26.57	11.38	75.08	18.09	17.54	77.69	32.86	13.21	69.44	26.64
Sim1	18.07	75.25	29.05	11.72	75.61	20.26	19.69	78.18	37.84	13.55	69.74	28.19

- Rule based and use of linguistic priors outperforms Ravi & Knight's (2009) model
- Significant increase in top-1 accuracy over rule-based
- Good top-10 accuracy, which rule-based cannot provide

Syllable-based Transliteration

(Atreya, et al 2015)

*Syllable as the basic
transliteration unit*

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

Transliteration Accuracies

	pa	as	bn	hi	gu	mr	te	kn	ml	ta
pa		CS:77.50 VS:82.50	CS:89.80 VS:93.70	CS:96.80 VS:98.60	CS:90.30 VS:89.50	CS:77.80 VS:78.90	CS:95.70 VS:97.90	CS:96.80 VS:98.40	CS:96.90 VS:98.50	CS:98.50 VS:98.30
as	CS:73.10 VS:83.10		CS:82.58 VS:86.89	CS:76.30 VS:85.90	CS:74.30 VS:84.80	CS:71.00 VS:80.60	CS:71.40 VS:81.70	CS:73.80 VS:85.20	CS:69.00 VS:78.40	-
bn	CS:90.30 VS:93.10	CS:78.60 VS:87.70		CS:97.40 VS:97.80	CS:90.40 VS:93.80	CS:68.20 VS:80.60	CS:96.20 VS:96.90	CS:95.50 VS:97.00	CS:98.40 VS:98.20	CS:97.70 VS:98.00
hi	CS:86.40 VS:87.60	CS:79.30 VS:84.80	CS:79.70 VS:88.30		CS:81.20 VS:88.00	CS:72.77 VS:82.88	CS:95.70 VS:96.50	CS:93.30 VS:93.60	CS:95.40 VS:96.70	CS:95.60 VS:95.80
gu	CS:89.30 VS:88.80	CS:83.00 VS:87.00	CS:84.10 VS:91.20	CS:98.70 VS:99.00		CS:81.60 VS:83.00	CS:97.00 VS:97.00	CS:95.70 VS:96.70	CS:98.00 VS:98.40	CS:98.00 VS:98.20
mr	CS:78.70 VS:79.90	CS:79.40 VS:88.60	CS:75.40 VS:84.40	CS:66.87 VS:75.88	CS:77.40 VS:81.40		CS:67.00 VS:74.60	CS:74.90 VS:78.60	CS:69.20 VS:73.90	-
te	CS:97.40 VS:98.40	CS:75.20 VS:79.80	CS:96.40 VS:98.10	CS:99.20 VS:99.30	CS:97.60 VS:98.20	CS:70.10 VS:76.90		CS:98.70 VS:98.80	CS:99.00 VS:97.70	CS:98.50 VS:98.80
kn	CS:97.60 VS:98.40	CS:76.40 VS:81.30	CS:94.60 VS:97.40	CS:98.50 VS:98.90	CS:96.20 VS:96.80	CS:71.50 VS:79.60	CS:99.20 VS:99.60		CS:99.50 VS:99.90	CS:98.90 VS:99.30
ml	CS:99.00 VS:99.10	CS:72.20 VS:77.70	CS:99.60 VS:99.60	CS:99.10 VS:99.30	CS:98.40 VS:99.00	CS:71.80 VS:77.70	CS:98.90 VS:99.40	CS:99.80 VS:99.90		CS:97.20 VS:97.90
ta	CS:84.10 VS:94.30	-	CS:86.20 VS:95.30	CS:86.80 VS:95.50	CS:86.70 VS:96.60	-	CS:86.50 VS:96.60	CS:86.90 VS:96.20	CS:85.70 VS:95.90	

Syllable-level transliteration (VS) outperforms character-level transliteration (CS)

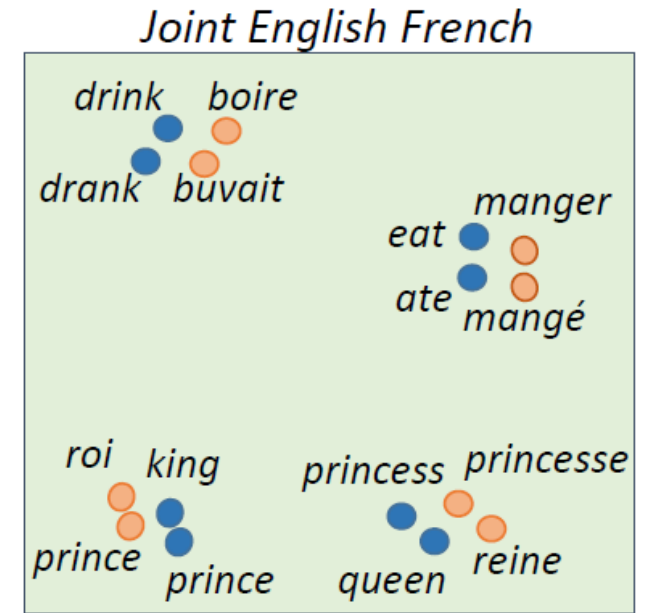
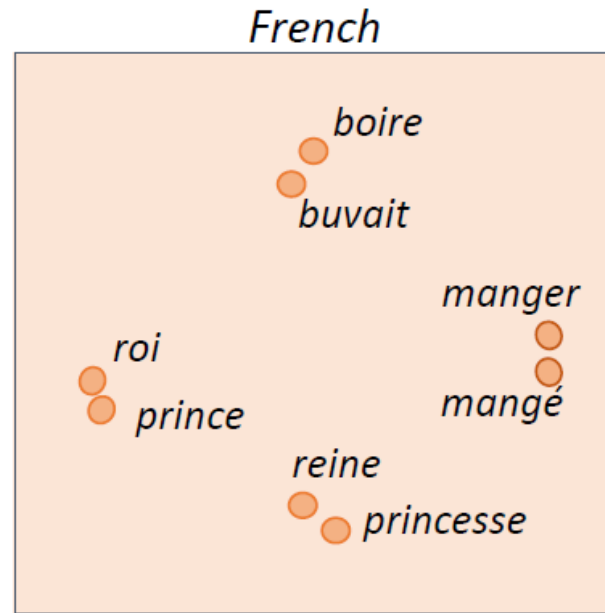
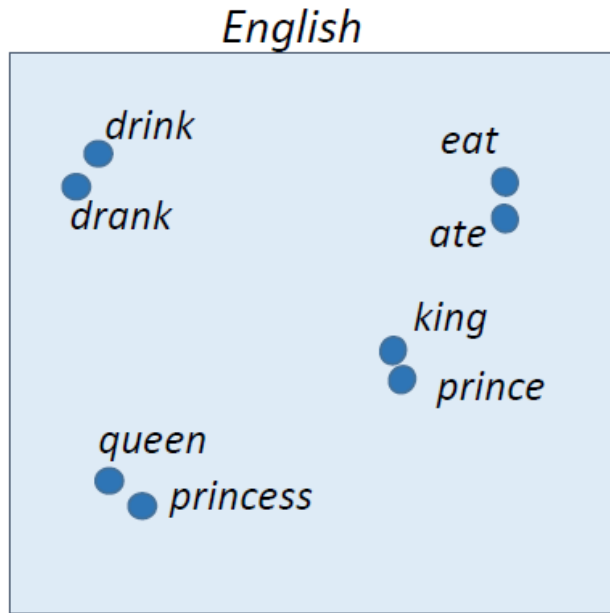
Utilizing Relatedness between Indian Languages

Orthographic Similarity

Lexical Similarity

Syntactic Similarity

Multilingual Word Embeddings



Monolingual Word Representations

(capture syntactic and semantic similarities between words)

Multilingual Word Representations

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

$$\text{embed}(y) = f(\text{embed}(x))$$

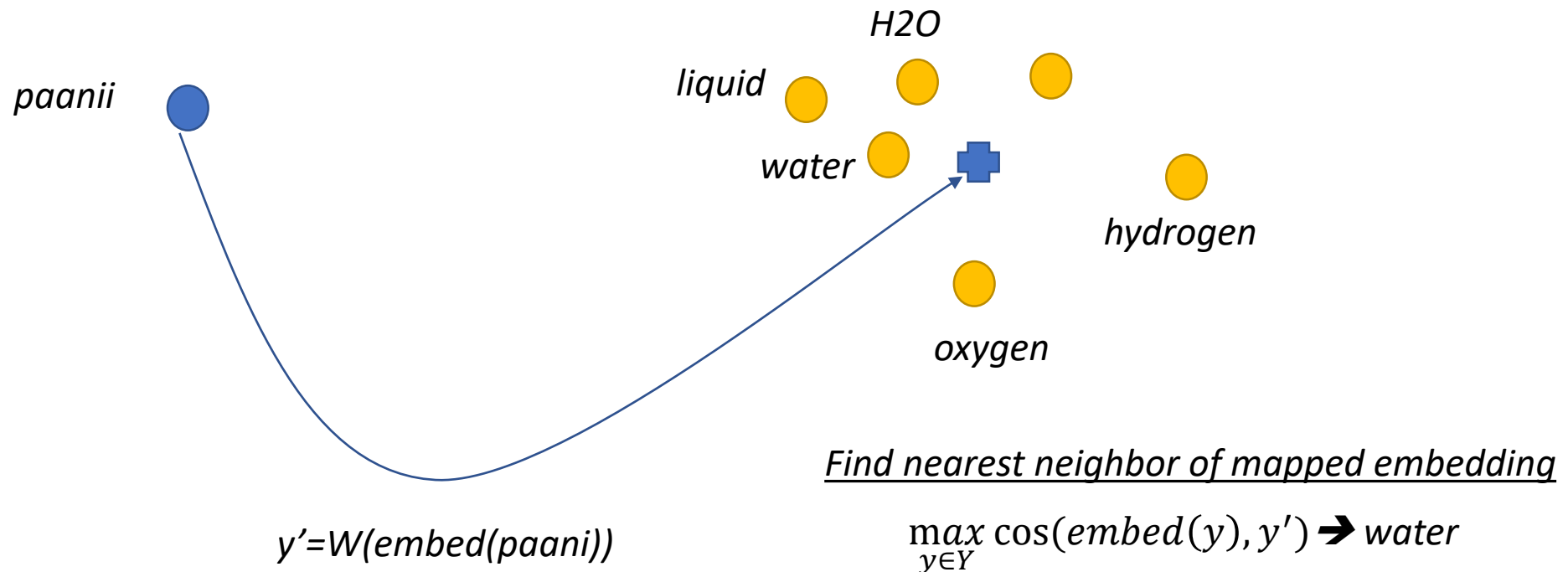
x, y are source and target words
 $\text{embed}(w)$: embedding for word w

(Source: Khapra and Chandar, 2016)

Bilingual Lexicon Induction

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

Can we extract a bilingual dictionary?



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



Evaluation

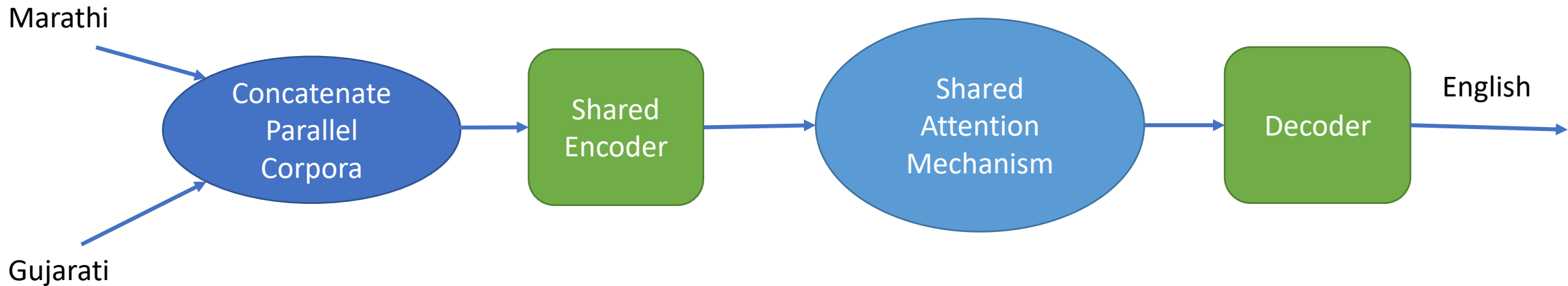
Method	En-German	En-Italian	En-Finnish
Baseline (B)	40.27	39.40	26.47
B + identity (I)	51.73	44.07	42.63
B + enhanced (E)	50.33	48.40	29.63
B + I + E	55.40	47.13	43.54

Precision@1

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	हू घरे जव छू
It rained last week	छेल्ला आठवडिया मा वर्साद पाड्यो

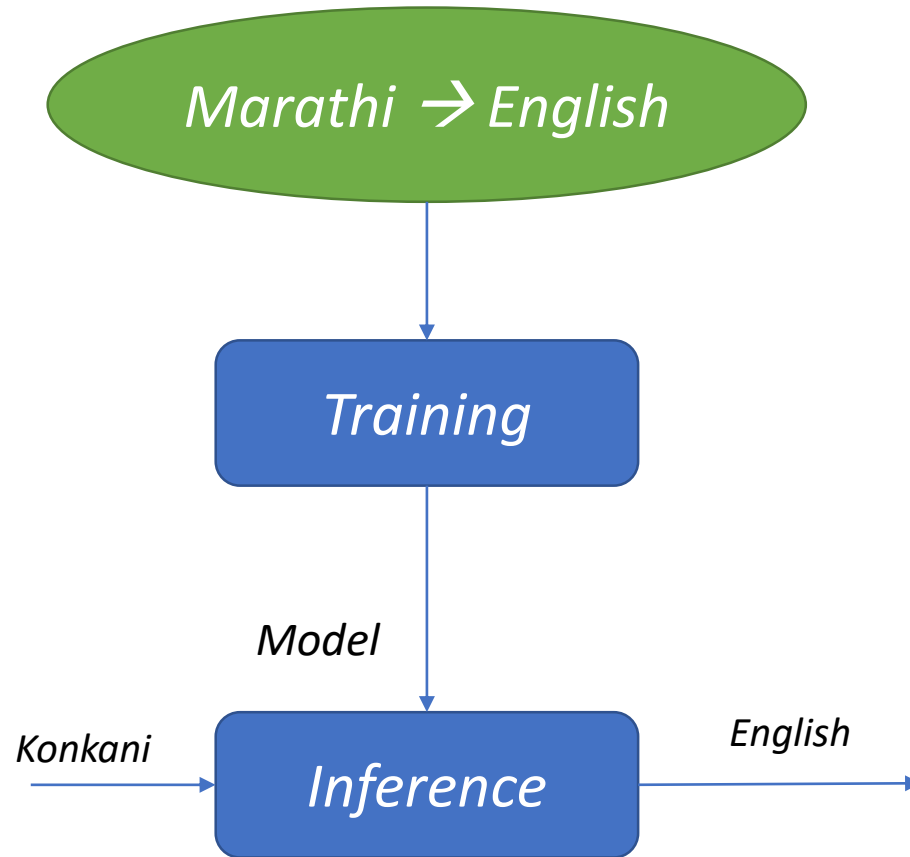
It is cold in Pune	पुण्यात थंड आहे
My home is near the market	माझा घर बाजाराजवळ आहे

Convert Script

Concat Corpora

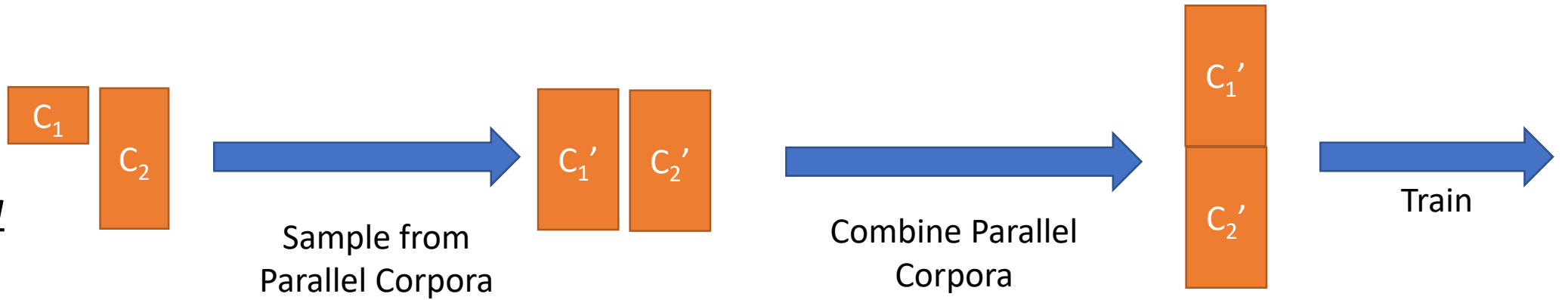
I am going home	हू घरे जव छू
It rained last week	छेल्ला आठवडिया मा वर्साद पाड्यो
It is cold in Pune	पुण्यात थंड आहे
My home is near the market	माझा घर बाजाराजवळ आहे

Zeroshot Translation

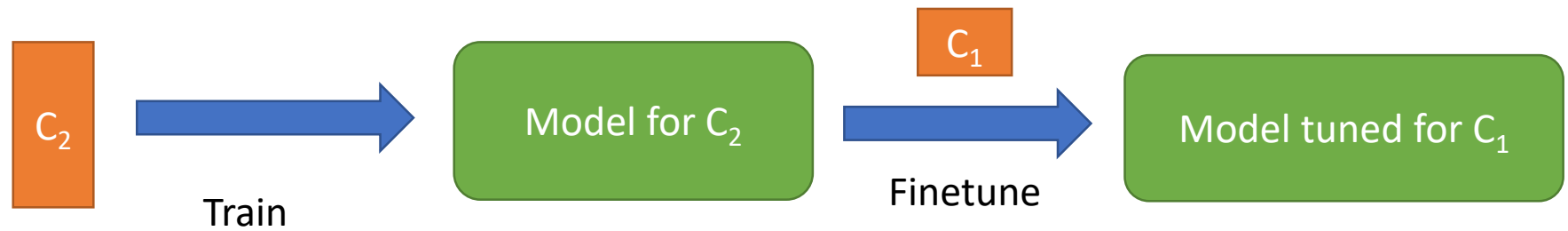


Training Multilingual NMT systems

Method 1 Joint Training



Method 2 Fine-tuning



Subword-level Representation of Corpora

I am going home	ह _ु 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

Backtranslation with a high-resource language

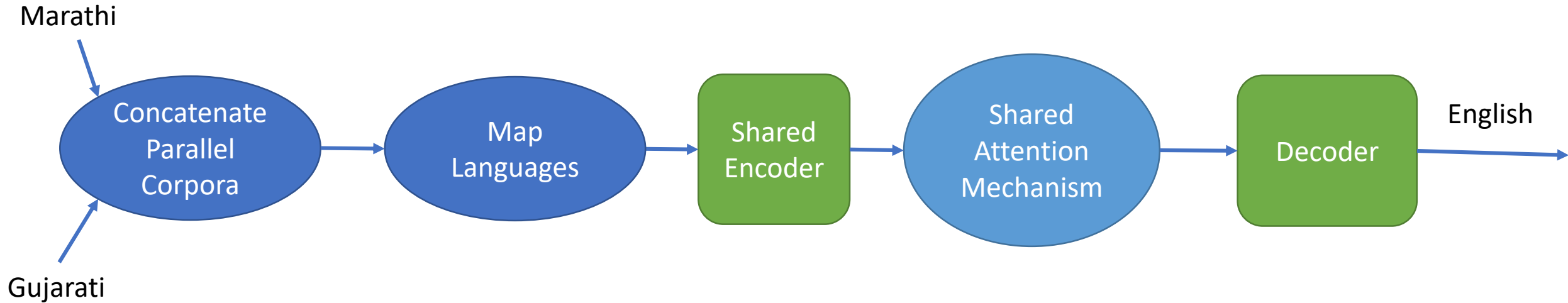


Standard backtranslation



Modified backtranslation

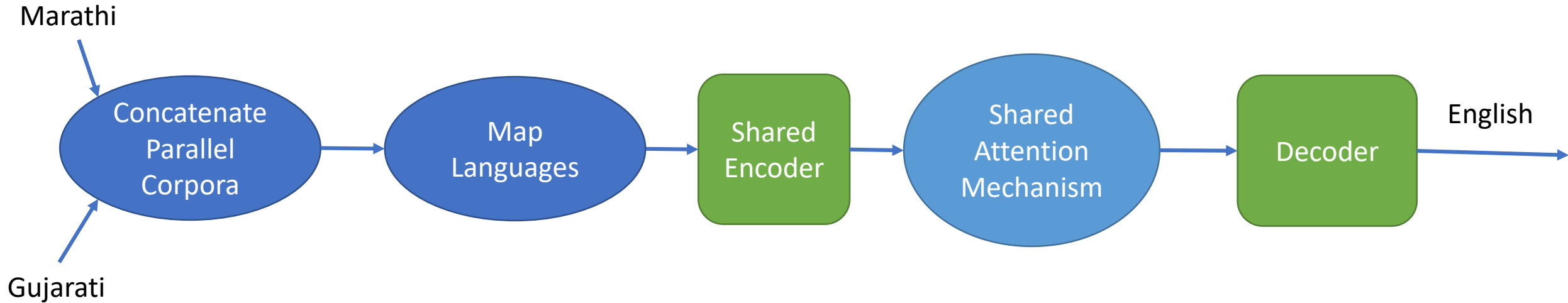
Make Indian Language Representations similar



Surface form approaches

- Transliteration
- Word-by-word translation
- Word-by-word translation with beam search

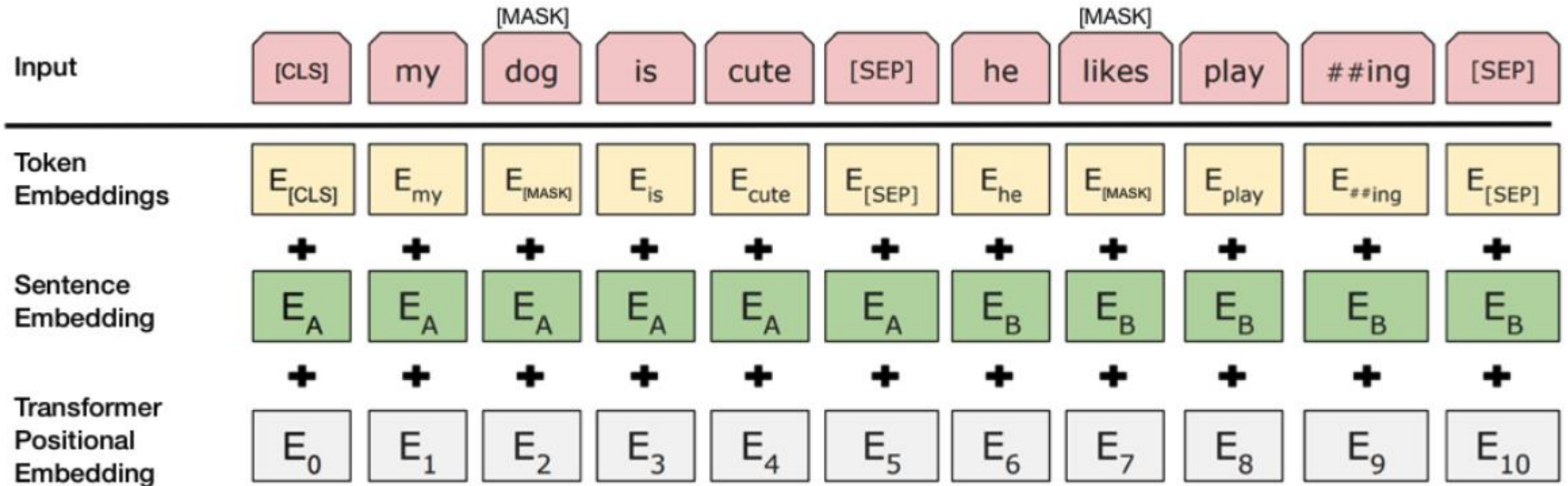
Make Indian Language Representations similar



Multilingual Embedding approaches

- Multilingual Word Embeddings
- Multilingual Sentence Embeddings

Multilingual BERT (Devlin et al., 2018)

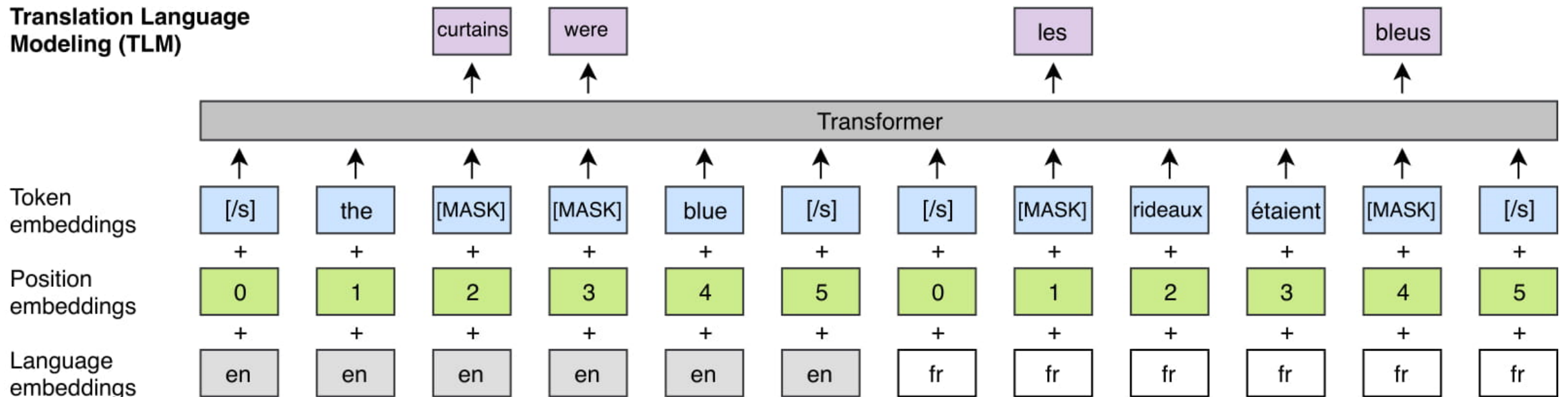


Transformer encoder with masked LM objective – i.e. try to predict masked words
Concat data from all languages

Cross-lingual Language Model Pre-training

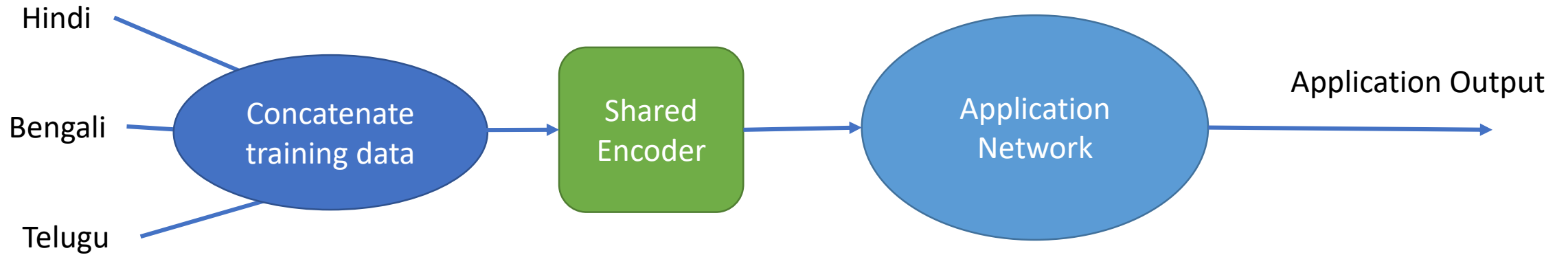
(Lample & Conneau, 2019)

Translation Language Modeling (TLM)



- Variant of BERT that adds a translation objective
- Needs parallel corpus

How to make other NLP applications multilingual?



- Sentiment Analysis
- Named Entity Recognition

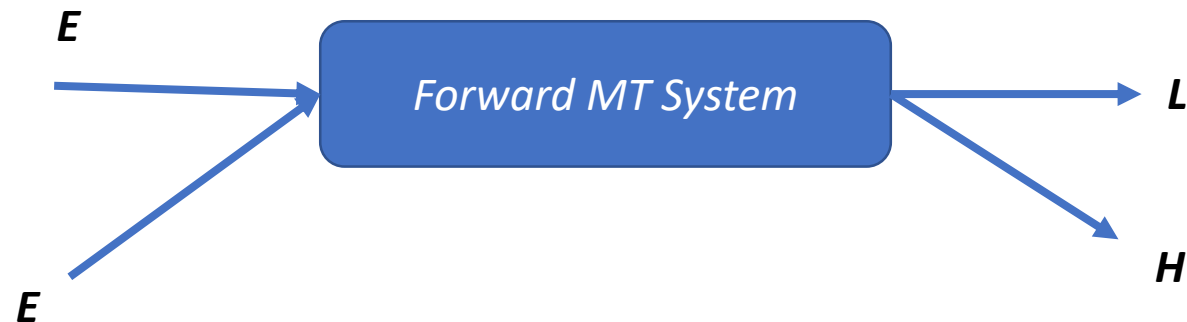
English → Indian Languages

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

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

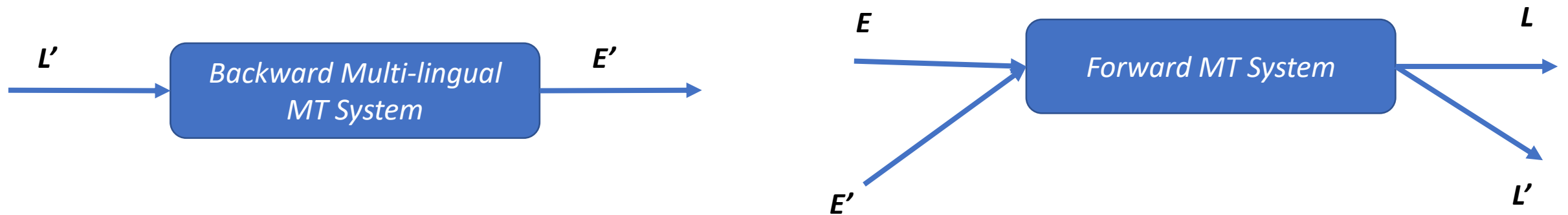
Original Input: *France and Croatia will play the final on Sunday*

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



Still an open problem

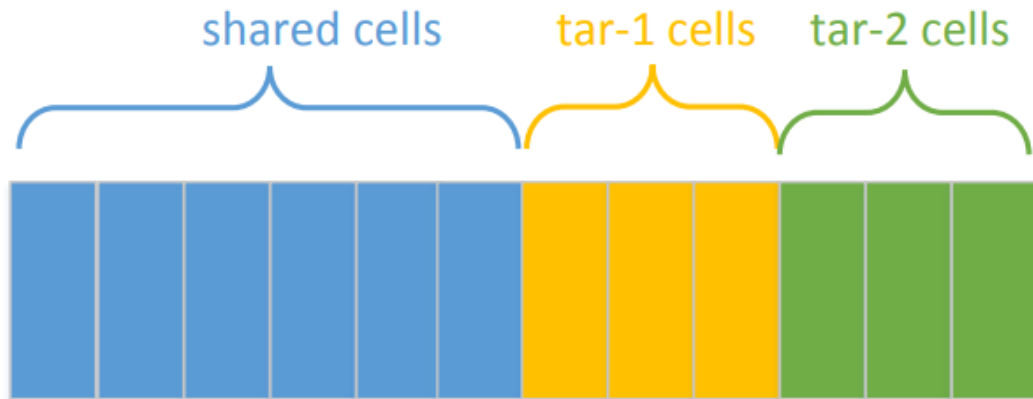
Backtranslation via Multilingual Model



Experiment	BLEU
Baseline Bilingual	19.7
(2) Baseline Multilingual $E \rightarrow X$	22.3
(2) + bilingual backtranslation	26.1
(2) + multilingual backtranslation	27.0

English \rightarrow Spanish with English \rightarrow French as helper pair

Shared and Private Decoder Hidden Units



- Shared units allow learning common features
- Language-dependent layers capture language specific information

Experiment	English-Chinese	English-German	English-French
Baseline Bilingual	44.23	27.84	41.50
(2) Baseline Multilingual $E \rightarrow X$	44.30	26.78	41.56
(2) + shared/private units	45.25	27.11	41.98

Indian-Indian Language MT

- Syllable as basic translation unit
- Balance between utilizing lexical similarity and word-level information

Basic Unit	Symbol	Example	Transliteration
Word	W	घरासमोरचा	gharAsamoracA
Morph Segment	M	घरा समोर चा	gharA samora cA
Orthographic Syllable	O	घ रा स मो र चा	gha rA sa mo racA
Character unigram	C	घ र ा स म ो र च ा	gha r A sa m o ra c A

something that is in front of home: ghara=home, samora=front, cA=of

Various translation units for a Marathi word

W: राजू , घराबाहेर जाऊ नको .

O: रा जू _ , _ घ रा बा हे र _ जा ऊ _ न को _ .

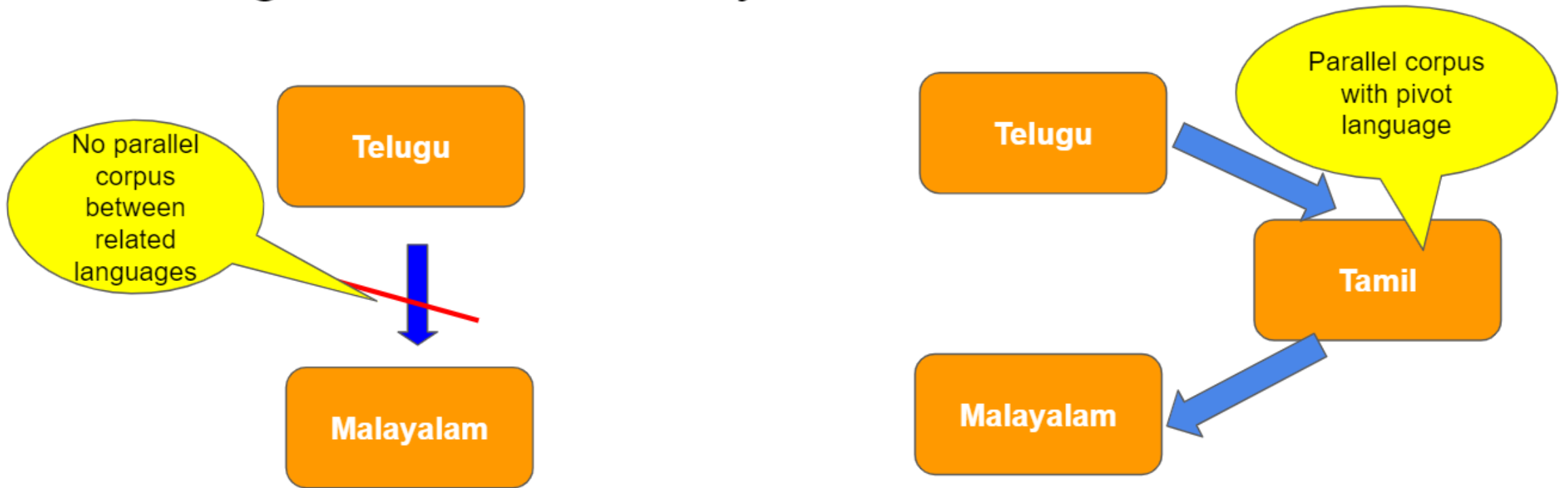
Results

Families	Lang Pair	Sim	W	W _x	M	M _x	C	O
IA-IA	ben-hin	52.30	31.23	32.79	32.17	32.32	27.95	33.46
	pan-hin	67.99	68.96	71.71	71.29	71.42	71.26	72.51
	kok-mar	54.51	21.39	21.90	22.81	22.82	19.83	23.53
DR-DR	mal-tam	39.04	6.52	7.01	7.61	7.65	4.50	7.86
	tel-mal	39.18	6.62	6.94	7.86	7.89	6.00	8.51
IA-DR	hin-mal	33.24	8.49	8.77	9.23	9.26	6.28	10.45
DR-IA	mal-hin	33.24	15.23	16.26	17.08	17.30	12.33	18.50

Results in % BLEU (*Sim: Lexical Similarity [LCSR], IA: Indo-Aryan, DR: Dravidian*)

- Substantial improvement over char-based (46%)
- Significant improvement over strong baselines: W_x (10%) & M_x (5%)
- Improvement when languages don't belong to same family (contact exists)
- More beneficial when languages are morphologically rich

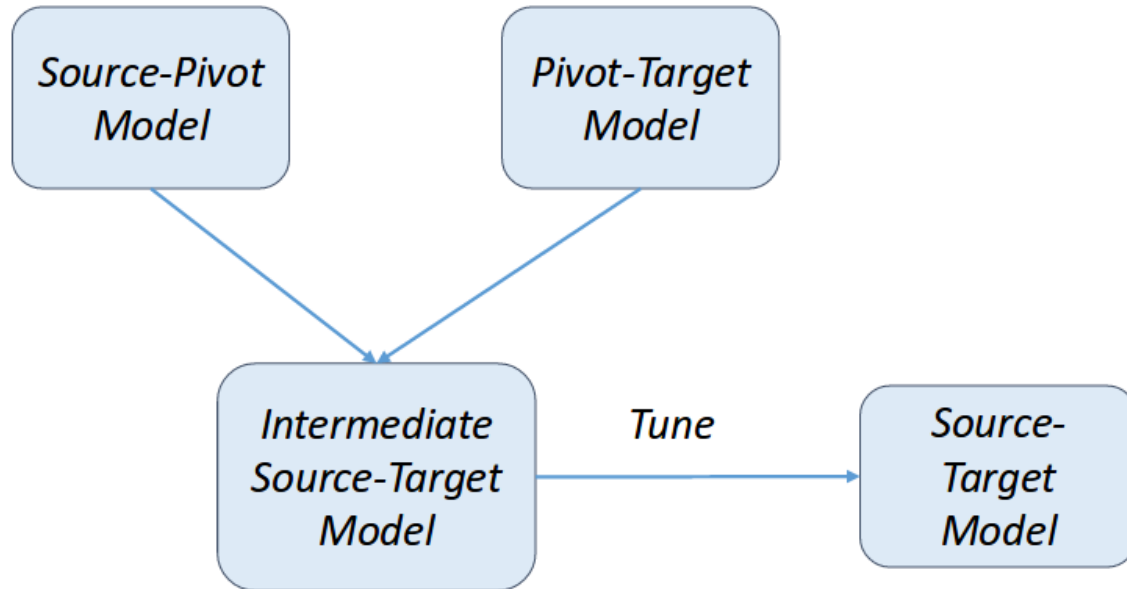
Pivot Translation for Related Languages



Related languages \Rightarrow Use subword level translation units

Translation through intermediate language \Rightarrow Use Pivot based SMT methods

Combine the two approaches



src-pivot phrase table

A	X	0.4	0.4
B	X	0.6	0.8
B	Y	0.8	0.9
C	Y	0.2	0.1

X	P	0.5	0.4
Y	P	0.5	0.6
Y	Q	1.0	1.0
Z	R	1.0	1.0

pivot-tgt phrase table

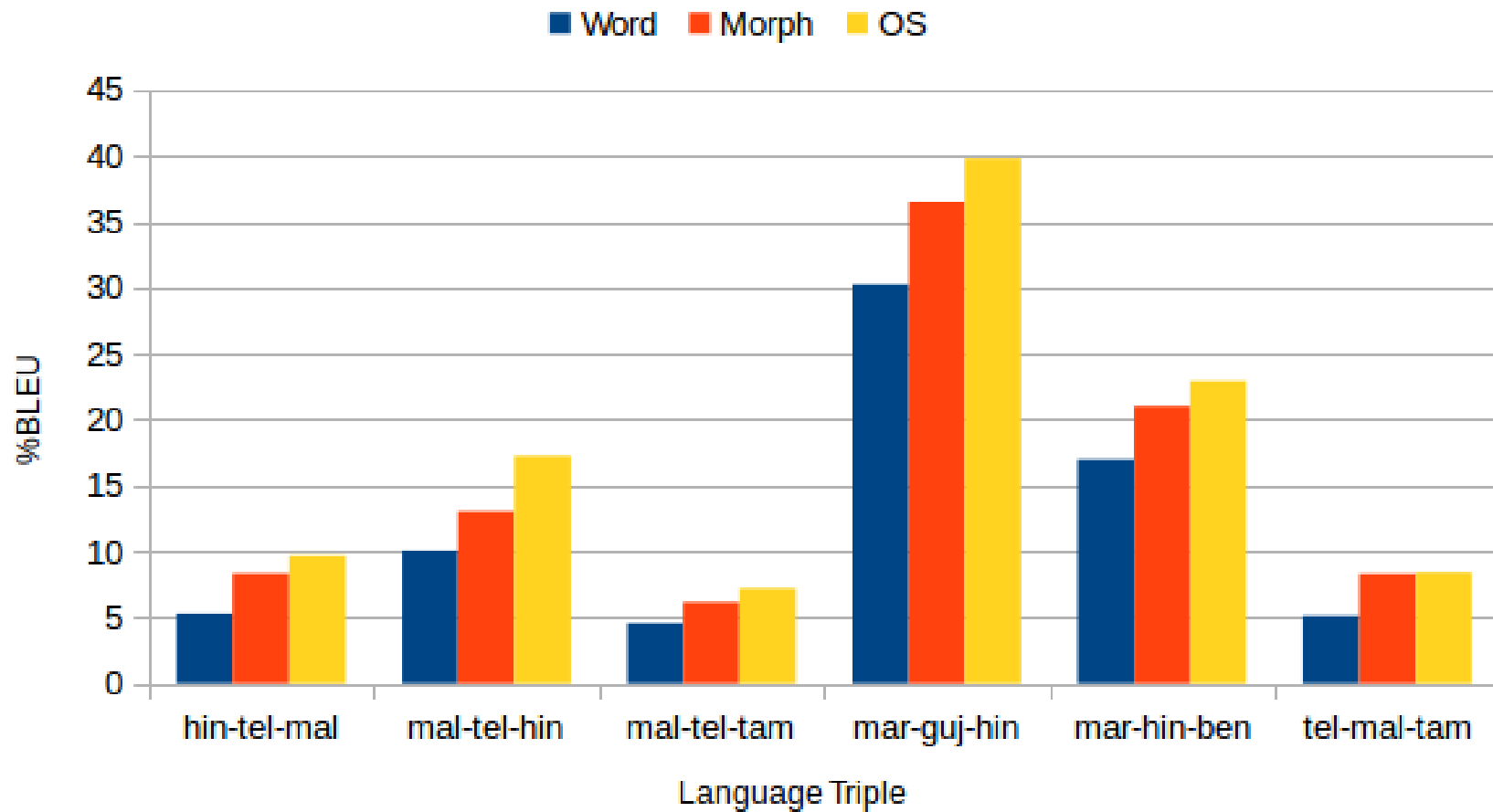


Word-level join results in sparse table

A	P	?	?
B	P	?	?
B	Q	?	?
C	Q	?	?
C	P	?	?

Why is syllable-based pivot model better?

- The underlying source-pivot and pivot-target models are better
- Data loss during join is minimized with subword representation



OS level pivot system outperforms other units

- *~60% improvement over word level*
- *~15% improvement over morph level*

Utilizing Relatedness between Indian Languages

Orthographic Similarity

Lexical Similarity

Syntactic Similarity

Use Source Re-ordering for Phrase-based SMT

(Kunchukuttan et al., 2014)

Phrase based MT is not good at learning word ordering

Solution: Let's help PB-SMT with some preprocessing of the input

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

Let's take an example

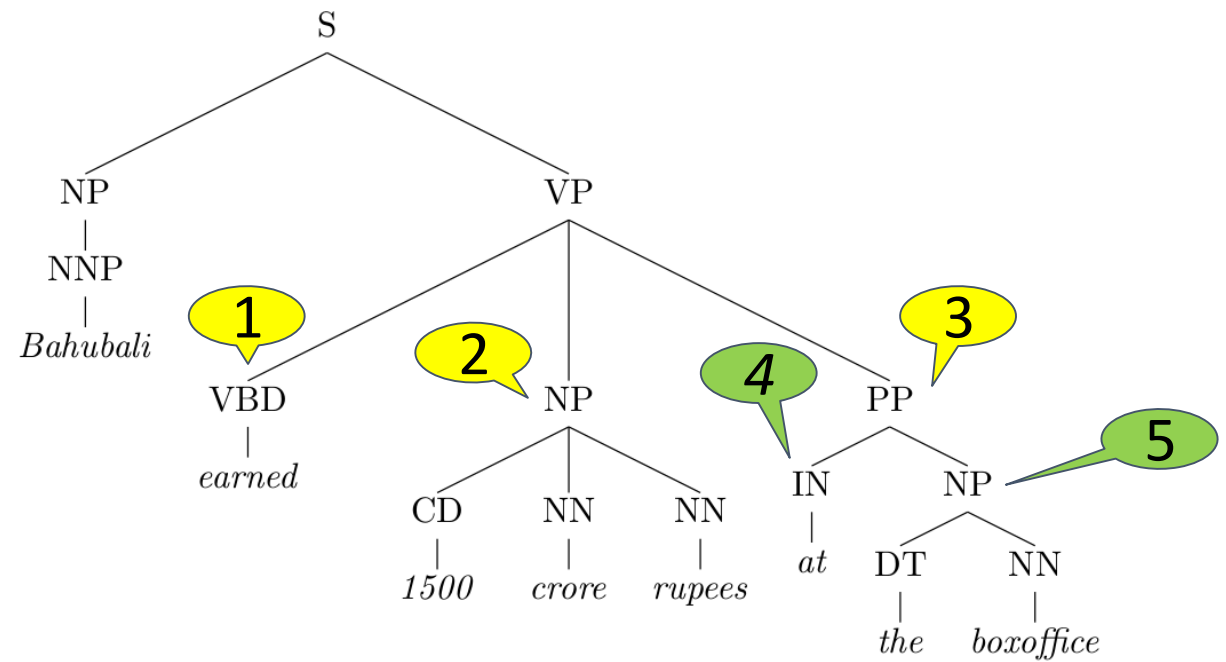
Bahubali earned more than 1500 crore rupee sat the boxoffice

Parse the sentence to understand its syntactic structure

Apply rules to transform the tree

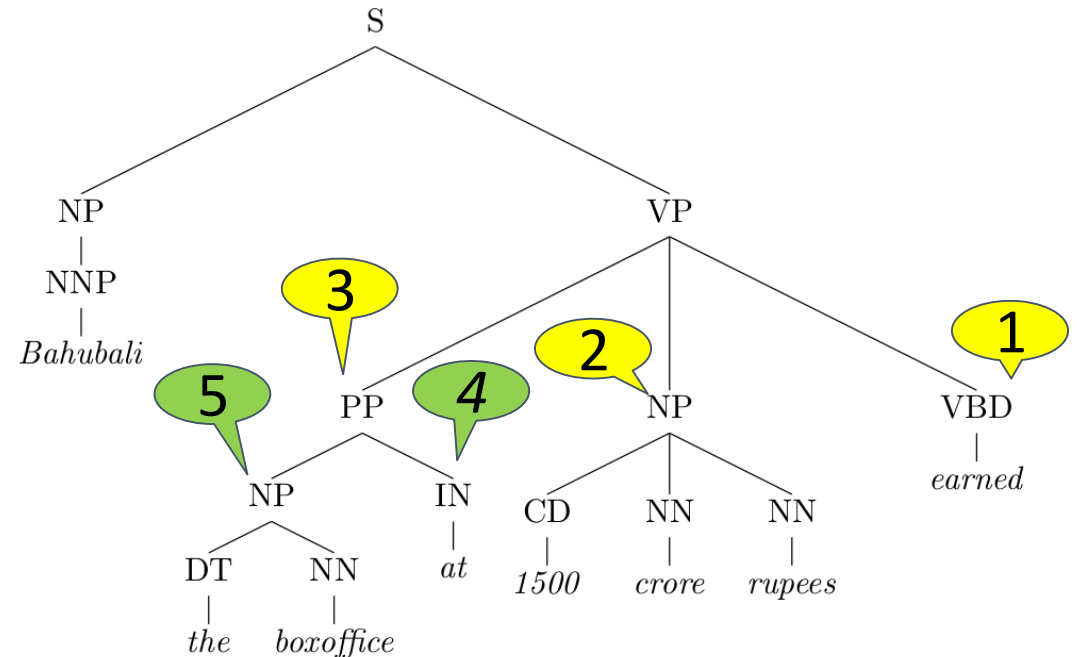
VP → VBD NP PP ⇒ VP → PP NP VBD

PP → IN NP ⇒ PP → NP IN



The new input to the machine translation system is:
Bahubali the boxoffice at 1500 crore rupees earned

Now we can translate with little reordering:
बाहुबली ने बॉक्सओफिस पर 1500 करोड रुपए कमाए



Can we reuse English-Hindi rules for English-Indian languages?

All Indian languages have the same basic word order

	Indo-Aryan						Dravidian		
	pan	hin	guj	ben	mar	kok	tel	tam	mal
Baseline	15.83	21.98	15.80	12.95	10.59	11.07	7.70	6.53	3.91
Generic	17.06	23.70	16.49	13.61	11.05	11.76	7.84	6.82	4.05
Hindi-tuned	17.96	24.45	17.38	13.99	11.77	12.37	8.16	7.08	4.02

(Kunchukuttan et al., 2014)

Generic reordering (*Ramanathan et al 2008*)

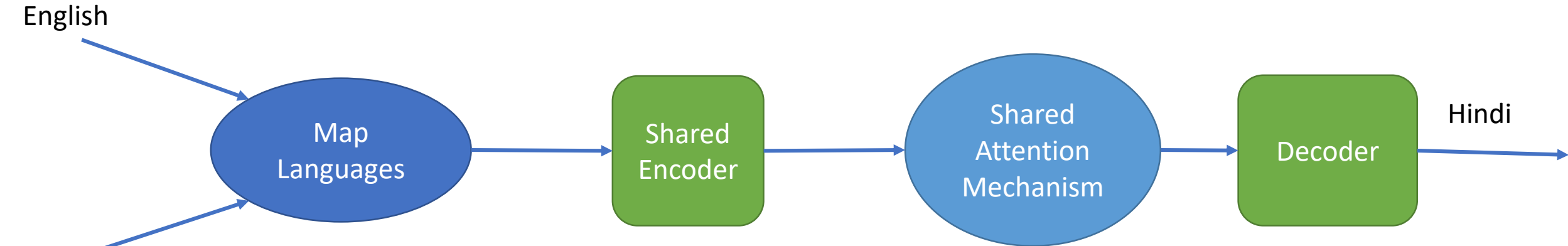
Basic reordering transformation for English → Indian language translation

Hindi-tuned reordering (*Patel et al 2013*)

Improvement over the basic rules by analyzing English → Hindi translation output

Bridging Word-order Divergence for low-resource NMT

(Rudramurthy et al., 2019)



Gujarati

Translate Gujarati \rightarrow Hindi

Given: English \rightarrow Hindi parallel corpus ($E \rightarrow H$)

Little Gujarati \rightarrow Hindi parallel corpus ($G \rightarrow H$)

- Translate English to Gujarati word-by-word \rightarrow $G' \rightarrow H$ corpus
- Train the $G' \rightarrow H$ corpus
- Fine-tune on small $G \rightarrow H$ corpus

Problem: Difference in Gujarati-English word order

Cannot ensure similar Gujarat and English words have similar representations

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

Same rules work for all Indic languages

Language	BLEU			LeBLEU (%)		
	No Pre-Order	Pre-Ordered		No Pre-Order	Pre-Ordered	
		HT	G		HT	G
Bengali	6.72	8.83	9.19	37.10	41.50	42.01
Gujarati	9.81	14.34	13.90	43.21	47.36	47.60
Marathi	8.77	10.18	10.30	40.21	41.49	42.22
Malayalam	5.73	6.49	6.95	33.27	33.69	35.09
Tamil	4.86	6.04	6.00	29.38	30.77	31.33

Outline

- Motivation
- Relatedness between Indian Languages
- Utilizing Relatedness between Indian Languages
- **IndicNLP Library**
- Datasets, Services and Standards
- Summary

Indic NLP Library

https://github.com/anoopkunchukuttan/indic_nlp_library

Design Principles

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

Capabilities

- Text Normalizer
- Sentence Splitter
- Word Tokenizer
- Word Detokenizer
- Word Segmenter
- Syllabification
- Query Script Information
- Phonetic Similarity
- Script Converter
- Romanization
- Indicization
- Transliteration
- Acronym Transliterator
- Statistical Machine Translation
- Lexical Similarity

Language Support

	Indo-Aryan			Dravidian	Others
Assamese (asm)	Marathi (mar)	Sindhi (snd)	Kannada (kan)	English (eng)	
Bengali (ben)	Nepali (nep)	Sinhala (sin)	Malayalam (mal)		
Gujarati (guj)	Odia (ori)	Sanskrit (san)	Telugu (tel)		
Hindi/Urdu (hin/urd)	Punjabi (pan)	Konkani (kok)	Tamil (tam)		

Tasks

Monolingual	Indo-Aryan													Dravidian			
	san	hin	urd	pan	nep	snd	asm	ben	ori	guj	mar	kok	sin	kan	tel	tam	mal
Script Information	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
Normalization	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
Tokenization	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Word segmentation	✗	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✓	✓	✓
Romanization (ITRANS)	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ITRANS to Script	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Bilingual

- **Script Conversion:** Amongst the above mentioned languages, except Urdu and English
- **Transliteration:** Amongst the 18 above mentioned languages
- **Translation:** Amongst these 10 languages: (hin, urd, pan, ben, guj, mar, kok, sin, kan, tel, tam, mal) + English

Library Initialization

```
# The path to the local git repo for Indic NLP Library  
INDIC_NLP_LIB_HOME="/data/t-ankunc/installs/indic_nlp_library_py3"  
  
# The path to the local git repo for Indic NLP Resources  
INDIC_NLP_RESOURCES="/data/t-ankunc/installs/indic_nlp_resources"
```

```
import sys  
sys.path.append('{}/src'.format(INDIC_NLP_LIB_HOME))
```

```
from indicnlp import common  
common.set_resources_path(INDIC_NLP_RESOURCES)
```

```
from indicnlp import loader  
loader.load()
```

Outline

- Motivation
- Relatedness between Indian Languages
- Utilizing Relatedness between Indian Languages
- IndicNLP Library
- **Datasets, Services and Standards**
- Summary

Indic Standards & Datasets

Enable sharing of data and annotations

Standards

Important to ensure sharing of data and annotations

Necessary to build multilingual NLP systems

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

Catalog of Indian Language NLP Resources

https://github.com/indicnlpweb/indicnlp_catalog

Evolving, collaborative catalog of Indian language NLP resources

Please add resources you know of and send a pull request

Commercial Offerings for Indian Languages

	Microsoft	Google	Amazon
Translation	Yes	Yes	Yes
Transliteration	Yes	No	No
Information Extraction	No	No	No

Information Extraction includes entity recognition, intent recognition, sentiment analysis, relation extraction, POS tagging, syntactic parsing, etc.

Outline

- Motivation
- Relatedness between Indian Languages
- Utilizing Relatedness between Indian Languages
- IndicNLP Library
- Datasets, Services and Standards
- **Summary**

Thank You!

<http://www.cse.iitb.ac.in/~anoopk>