Natural Language Processing for Indian Languages

A Language Relatedness Perspective

Anoop Kunchukuttan

Microsoft India Translation & Speech Group, Hyderabad

ankunchu@microsoft.com

5th Workshop on Indian Language Data: Resources and Evaluation

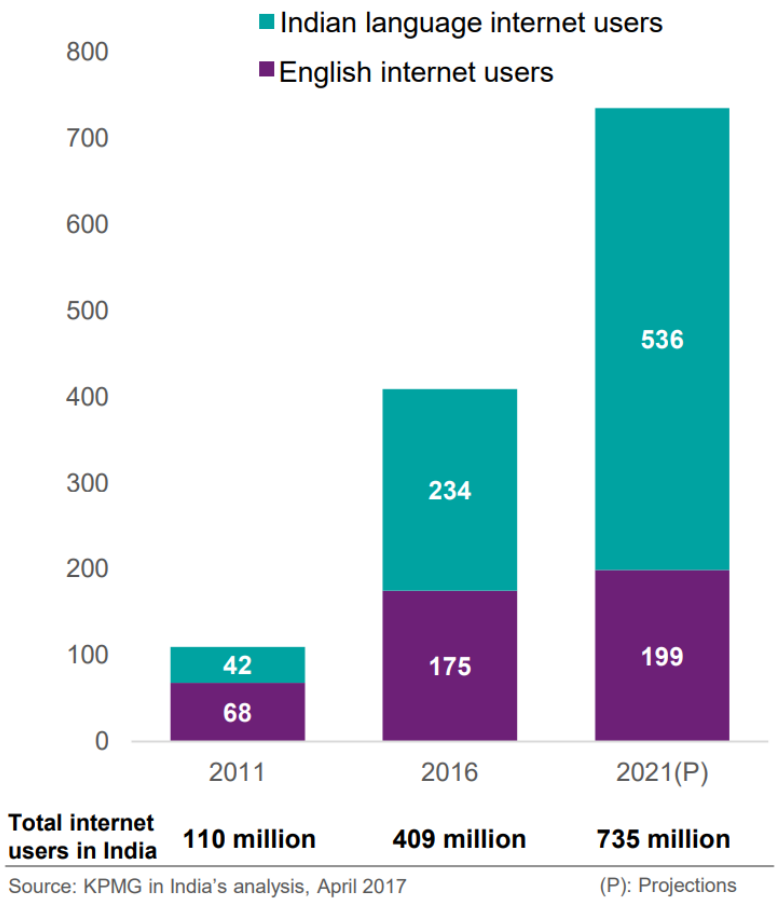
24th May 2020

Why Language Relatedness?

Usage and Diversity Indian Languages

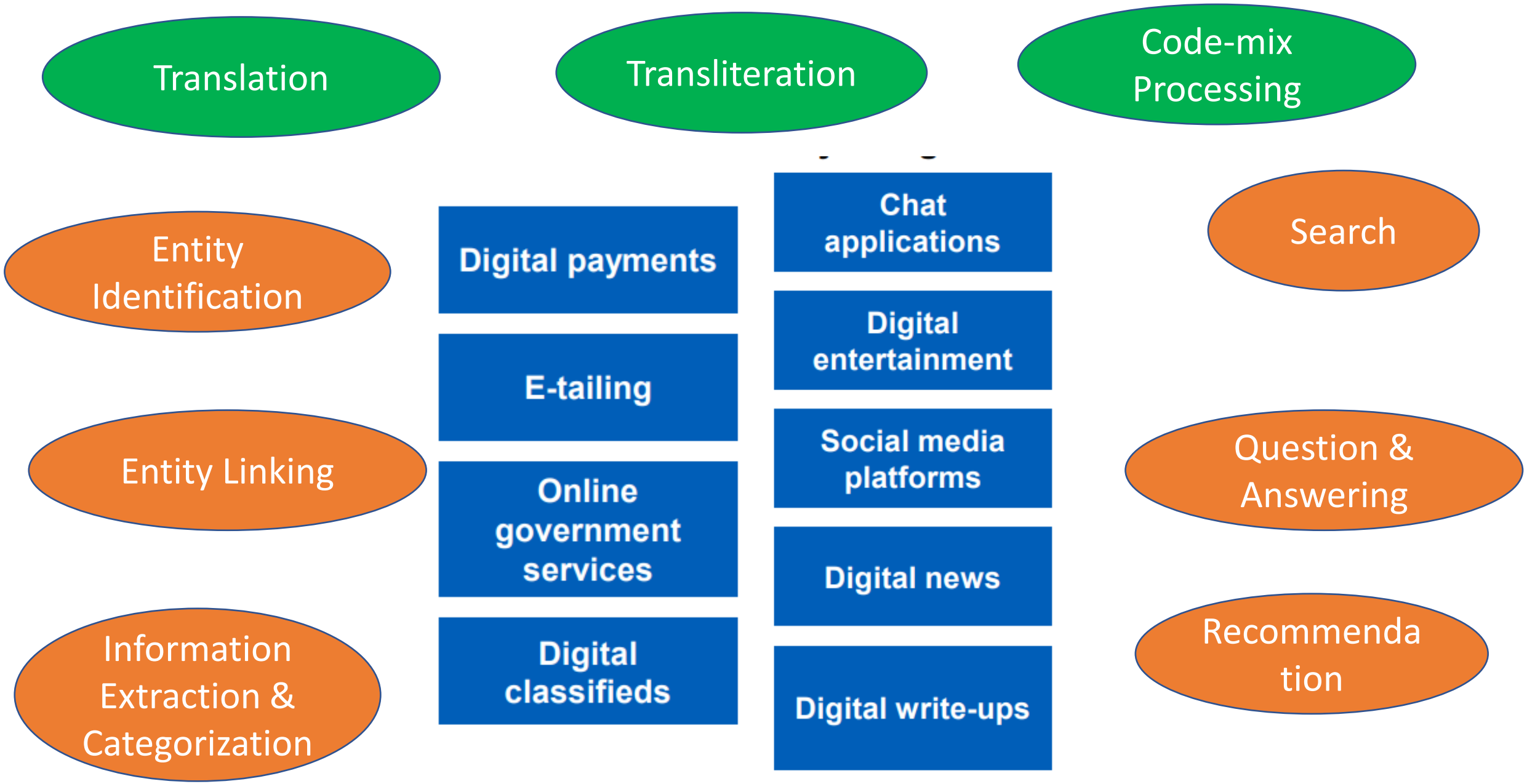
- *4 major language families*
- *22 scheduled languages*
- *125 million English speakers*
- *8 languages in the world's top 20 languages*
- *30 languages with more than 1 million speakers*

Sources: Wikipedia, Census of India 2011



Internet User Base in India (in million)

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



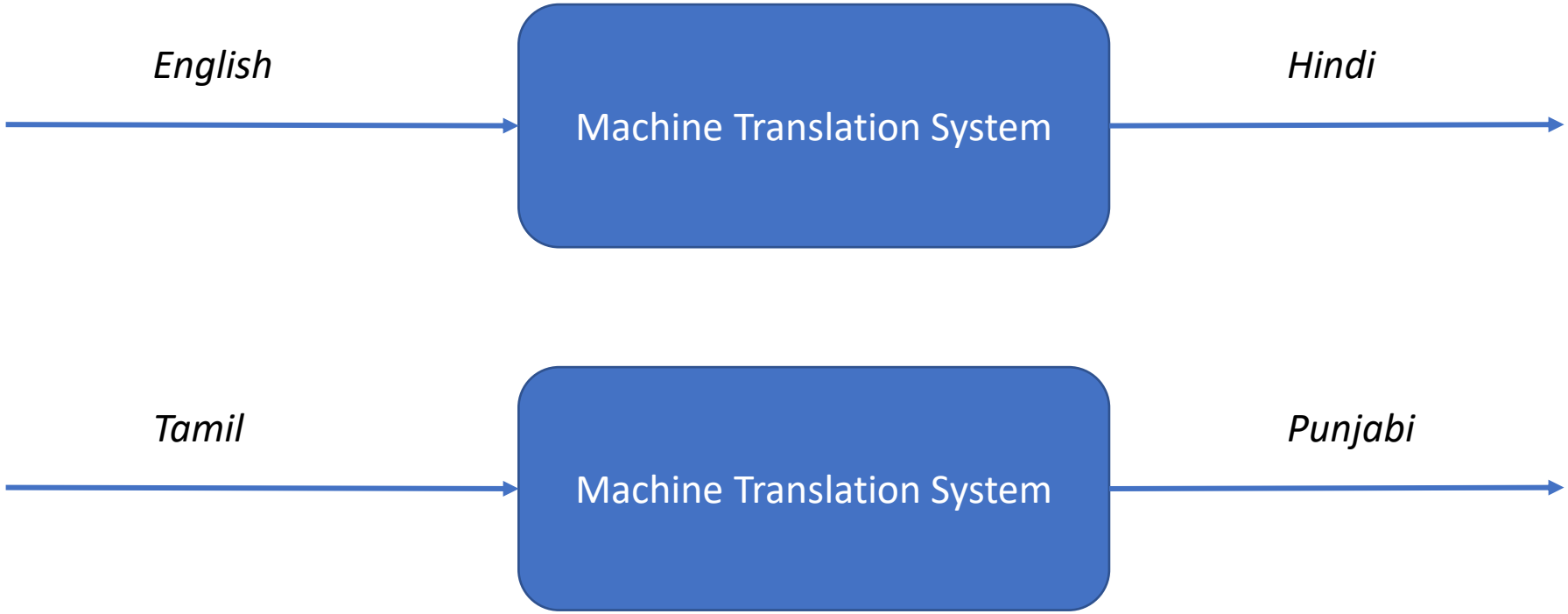
Applications requiring Indian language support

Scalability Challenges in ML solutions

- NLP requires human expertise → difficult and expensive to replicate for every language
  - Annotated data
  - Linguistic knowledge inputs
- Difficult to deploy and maintain systems for multiple languages

Expensive to create datasets for each language

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?

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?

What is 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

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

Orthographic Similarity

Devanagari	अ आ इ ई उ ऊ ऋ ऌ ऍ ऐ ए ऐ ऑ ओ 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 ਈ ਉ ਊ ਏ ਐ ਓ ਔ ਕ ਖ ਗ ਘ ਙ ਚ ਛ ਜ ਝ ਞ ਤ ਥ
Gujarati	અ આ ઇ ઈ ઉ ઊ ઋ ઍ 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 ఞ
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

Brahmi-derived Indic scripts are orthographically similar

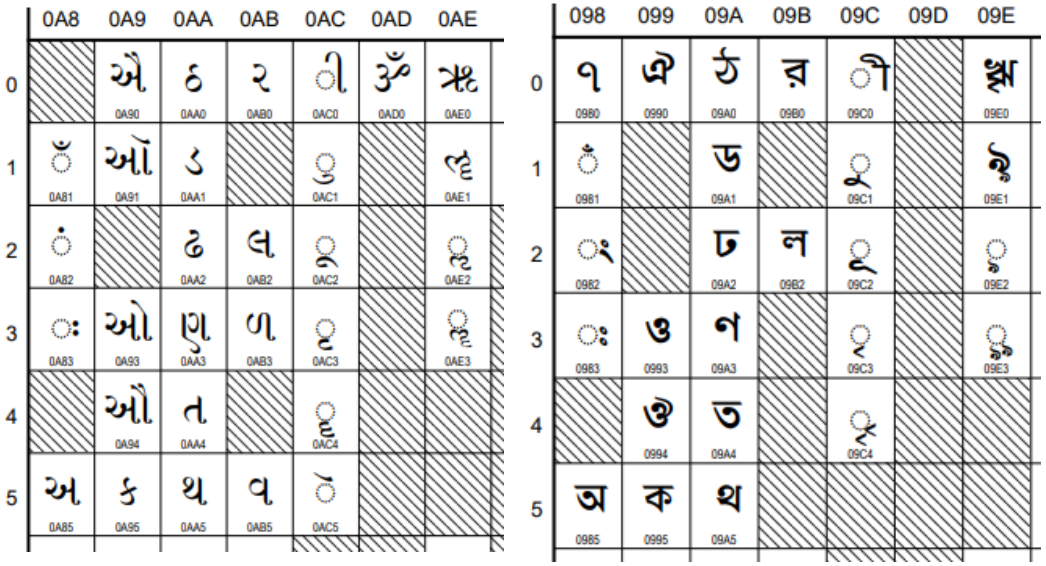
A simple and powerful property to utilize

relatedness between Indian languages

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 കേരള

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

Only a single mapping needed for Romanization too

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	ಬೆಂಗಳೂರು	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

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

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

Sl.No.	Label	IPA	Hindi	Marathi	Rajasthani	Gujarati	Odia
1	a	a	अ	अ	अ	અ	-
2	ax	ɔ	-	ऑ	-	ઓ	ଅ
3	aa	a:	आ	आ	आ	આ	ଆ
4	axx	ə	-	-	-	-	-
5	i	I, i	इ	इ	इ	ઇ	ଇ, ଈ
22	k	k	क	क	क	ક	କ
23	kh	k^h	ख	ख	ख	ખ	ଖ
24	g	g	ग	ग	ग	ગ	ଗ
25	gh	g^h	घ	घ	घ	ઘ	ଘ

Common set of phones and their mappings to Indic scripts can be defined

Useful for multilingual ASR, TTS, G2P
(Schultz et al 2001; Abraham et al, 2014, Abraham et al, 2016)

Phonetic Representation

- *Represent each Indic character as a feature vector*
- *Define a similarity measure based on the feature vector*
- *Could be used for transliteration, cognate identification, spelling correction, etc.*

(Kondrak, 2001; Kunchukuttan, et al., 2016)

Feature	Possible Values
Basic Character Type	vowel, consonant, anusvaara, nukta, halanta, others
Vowel Features	
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
Consonant Features	
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

Orthographic Syllable

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

Syllable as the basic transliteration unit
(Atreya et. al . 2015)

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

IndicNLP Library

Utilize similarity between Indian languages for scaling NLP applications to multiple Indian languages

- Text Normalizer
- Syllabification
- Query Script Information
- Phonetic Similarity
- Script Converter
- Romanization
- Indicization

https://github.com/anoopkunchukuttan/indic_nlp_library

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

How similar are Indian Languages?

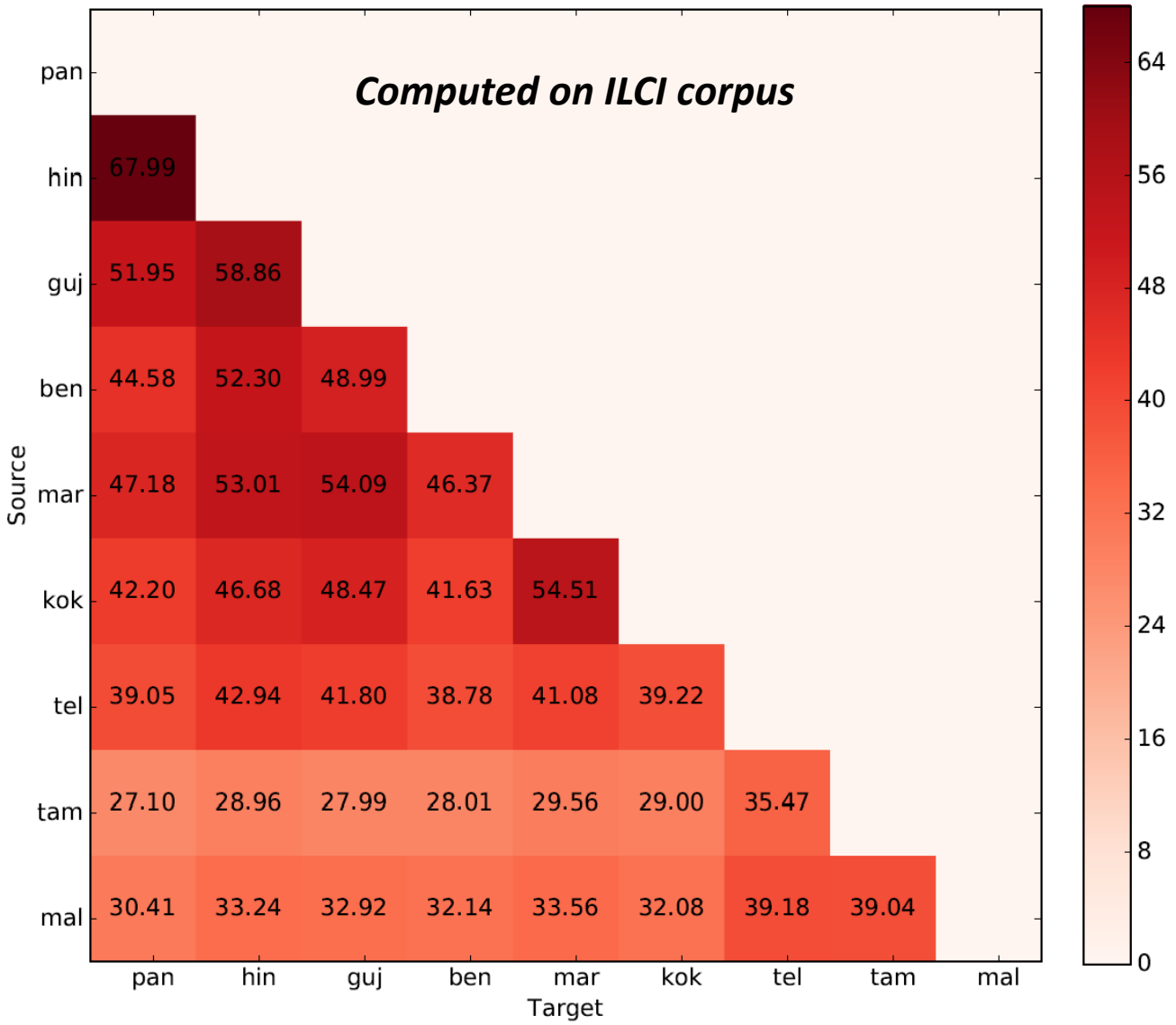
Estimate lexical similarity from parallel corpus

Longest Common Subsequence Ratio (LCSR) for a sentence pair

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

LCSR for a language pair

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



Indian-Indian Language subword-level MT

(Kunchukuttan & Bhattacharyya, 2016; Siripragada et al, 2020)

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

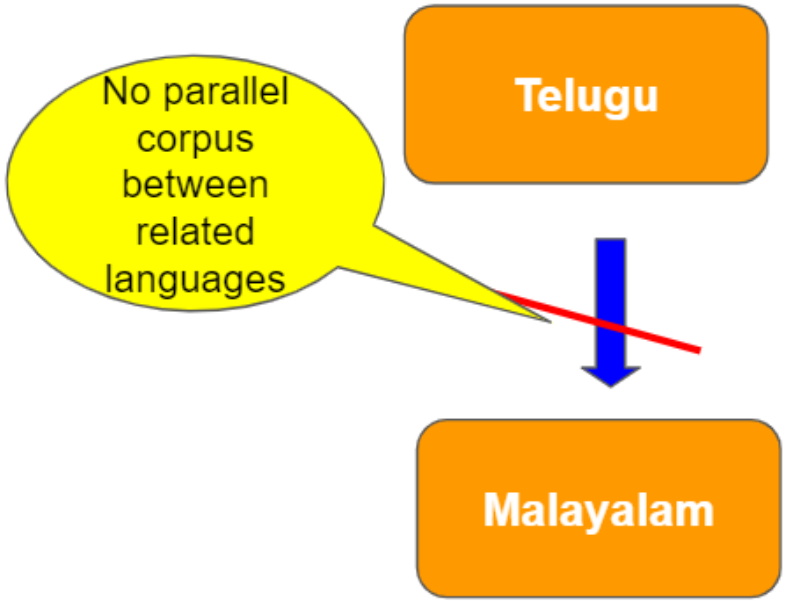
W: राजू , घराबाहेर जाऊ नको .
O: रा जू _ , _ घ रा बा हे र _ जा ऊ _ न को _ .

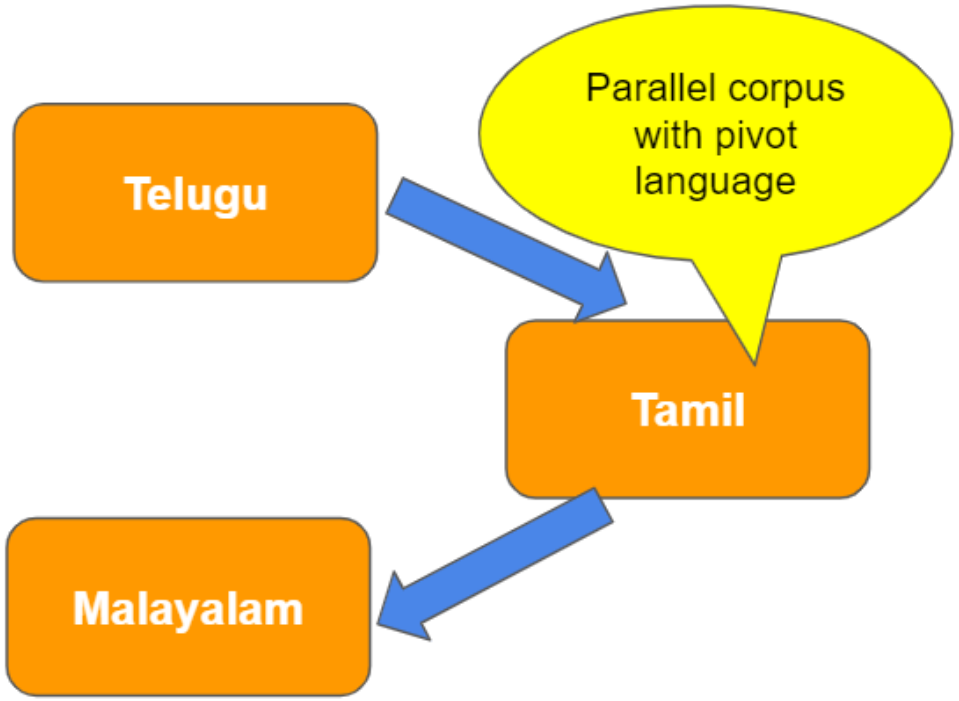
Fine vocab segmentations like BPE and SentencePiece are popular for NMT

- Can learn translation models with less data
- Balance between utilizing lexical similarity and word-level information

Pivot SMT between Indian Languages

(Kunchukuttan & Bhattacharyya, 2017)





Related languages ⇒ Use subword level translation units

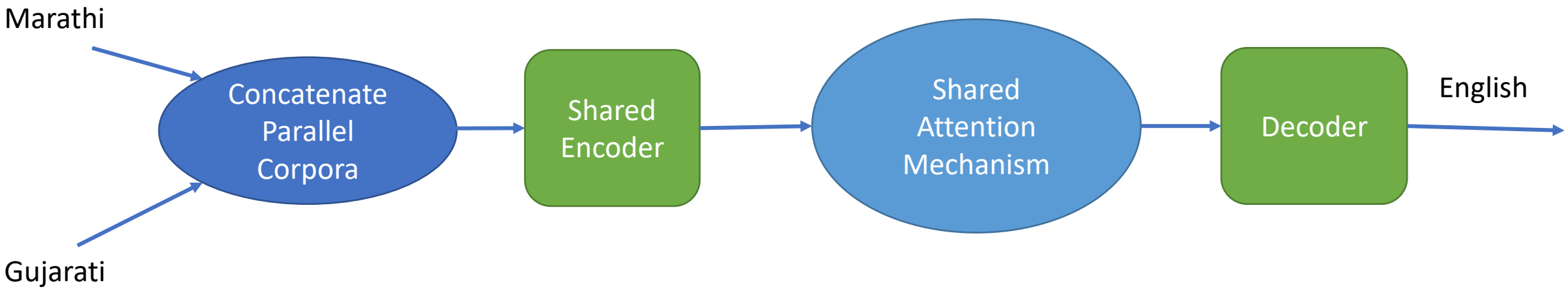
Translation through intermediate language ⇒ Use Pivot based SMT methods

Combine the two approaches

Transfer learning for En-IL NMT

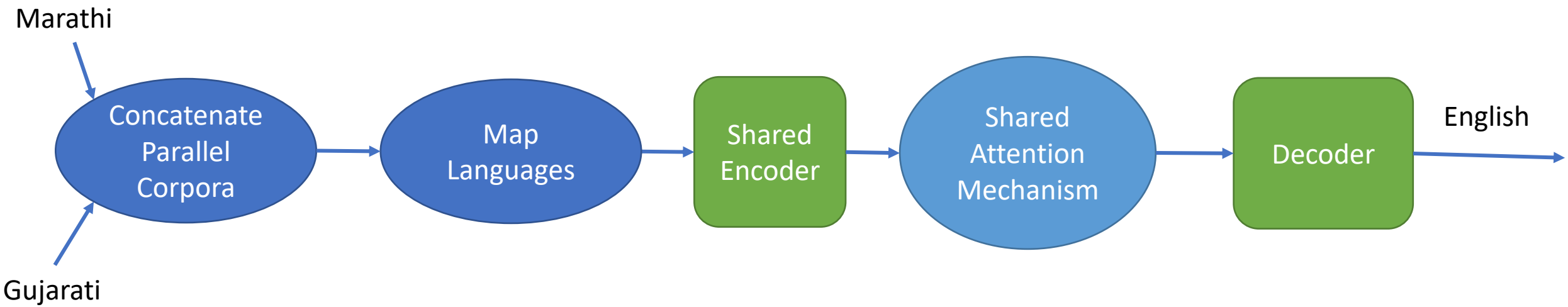
(Zoph et al., 2016; Nguyen et al., 2017; Lee et al., 2017; Dabre et al., 2018)

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



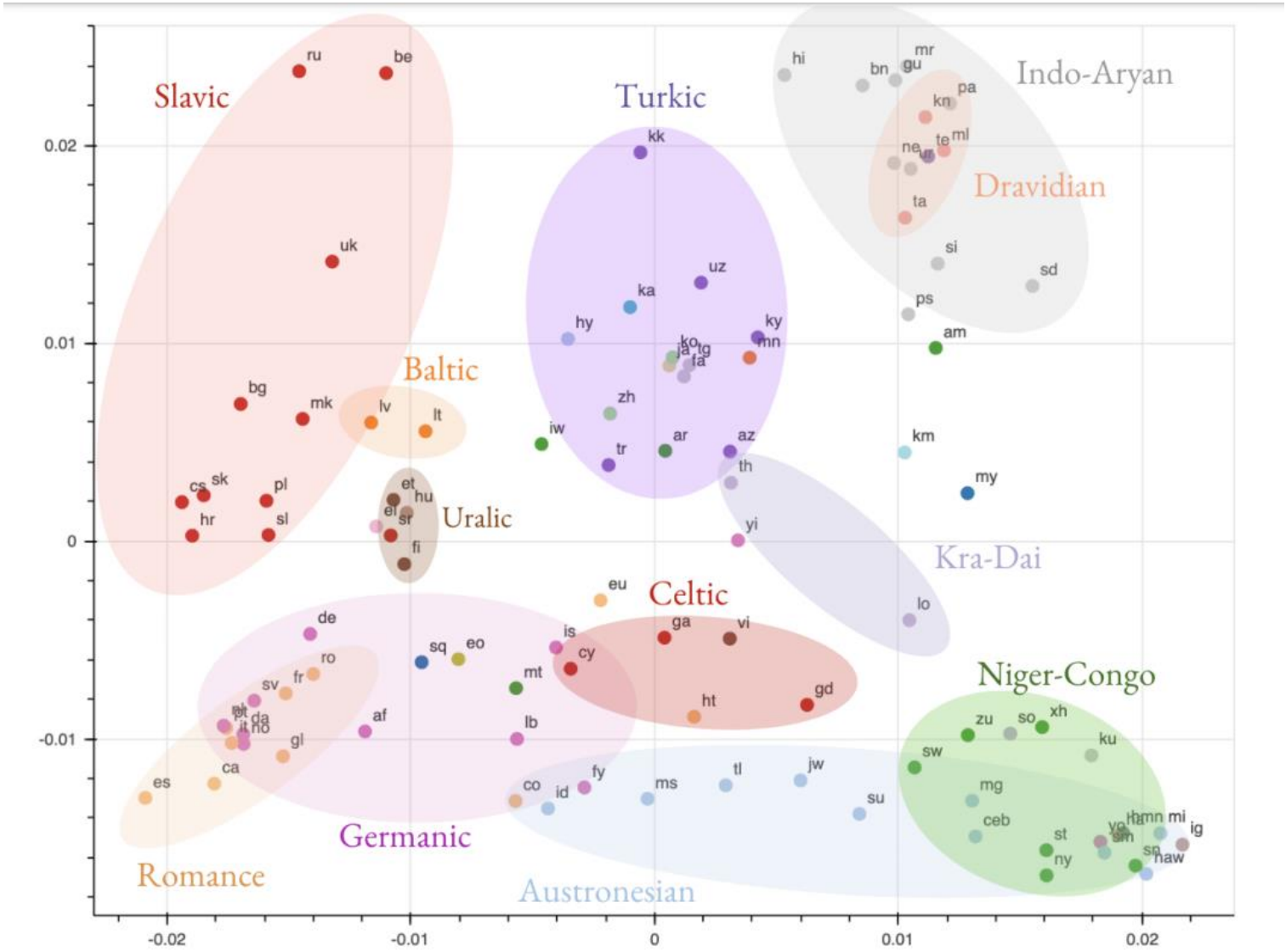
Train models at the subword-level

Make Indian Language Representations similar



- Script Conversion/Transliteration *(Dabre et al., 2018)*
- Word-by-word translation
- Word-by-word translation with rescoring using an LM

Transfer Learning works best for related languages



(Kudungta et al, 2019)

Encoder Representations cluster by language family

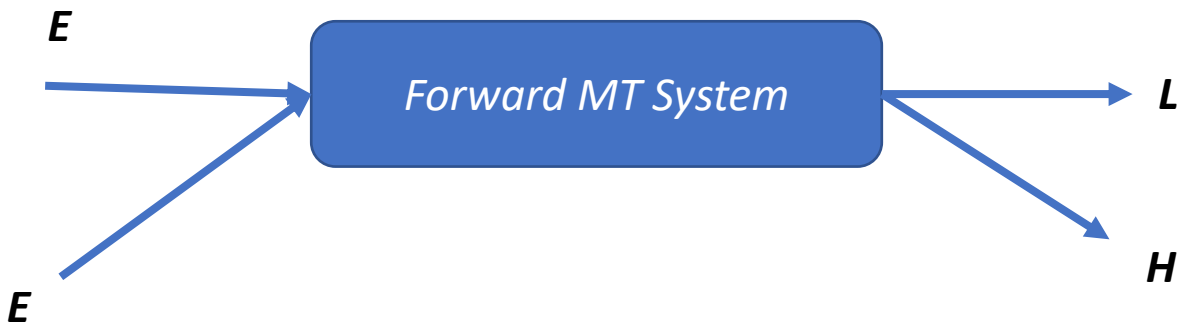
English → Indian Languages

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

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

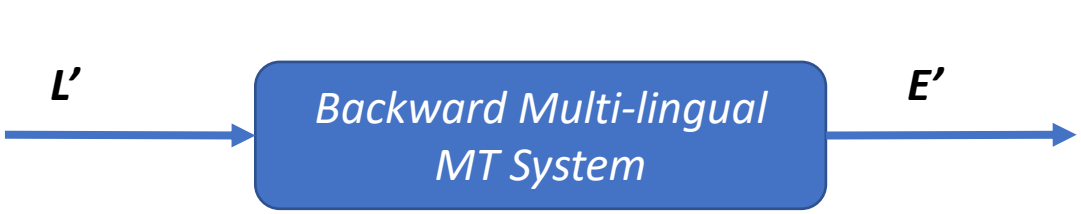
Original Input: *France and Croatia will play the final on Sunday*

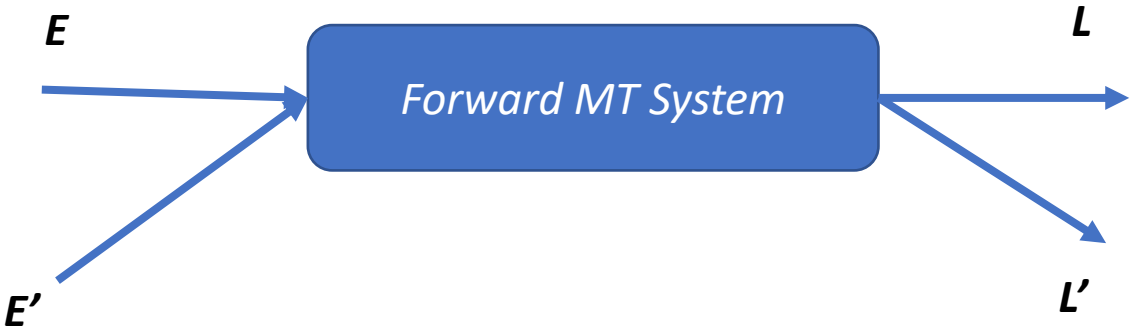
Modified Input: *France and Croatia will play the final on Sunday* *<hin>*



Still an open problem

Backtranslation via Multilingual Model



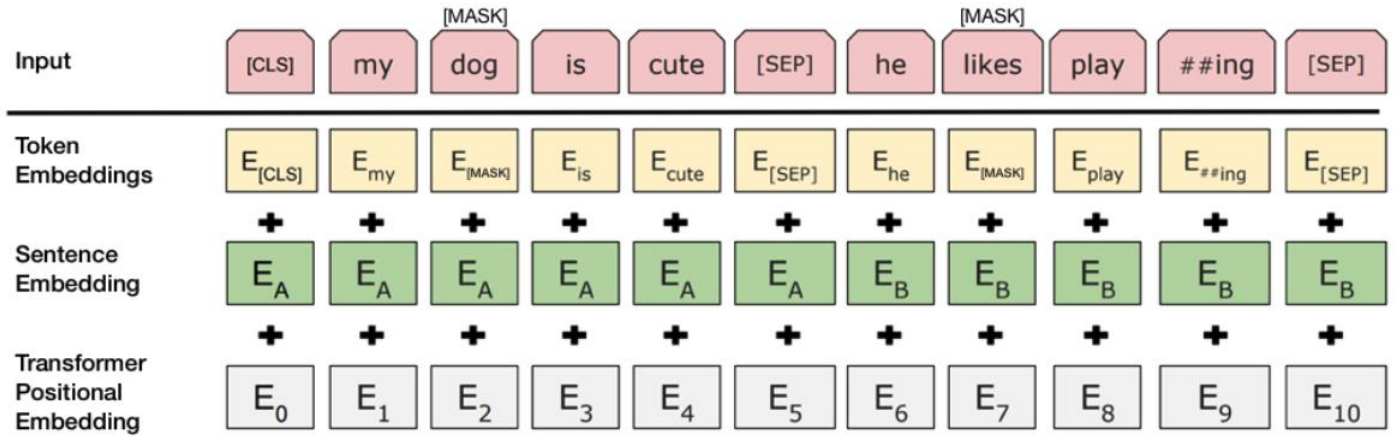


Experiment	BLEU
Baseline Bilingual	19.7
(2) Baseline Multilingual E→X	22.3
(2) + bilingual backtranslation	26.1
(2) + multilingual backtranslation	27.0

English → Spanish with English → French as helper pair

Multilingual Pre-trained NLU models

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



How can we explicitly model language relatedness?

How can language relatedness assumptions speedup pre-training?

- Multilingual BERT (Devlin et al., 2018) - Wikipedia
- XLM-R (Conneau et al , 2019) – CommonCrawl
- iNLTK - Wikipedia

Large-scale corpora and Evaluation sets

Some recent efforts

OSCAR Corpus: CommonCrawl *(Suarez et al, 2020)*

- https://oscar-corpus.com/

AI4Bharat Corpus: News websites *(Kunchukuttan et al, 2020)*

- https://github.com/ai4bharat-indicnlp/indicnlp_corpus

We also need Indian English content to overcome domain mismatches

Evaluation

- Few datasets for NLU tasks
- Mostly represent high resource languages like Hindi and Telugu
- Need datasets spanning all major languages
  - WikiAnnNER *(Pan et al, 2017)*
  - iNLTK News Headlines
  - AI4Bharat News Articles *(Kunchukuttan et al, 2020)*

https://github.com/ai4bharat-indicnlp/indicnlp_corpus

https://github.com/shantipriyap/Odia-NLP-Resource-Catalog

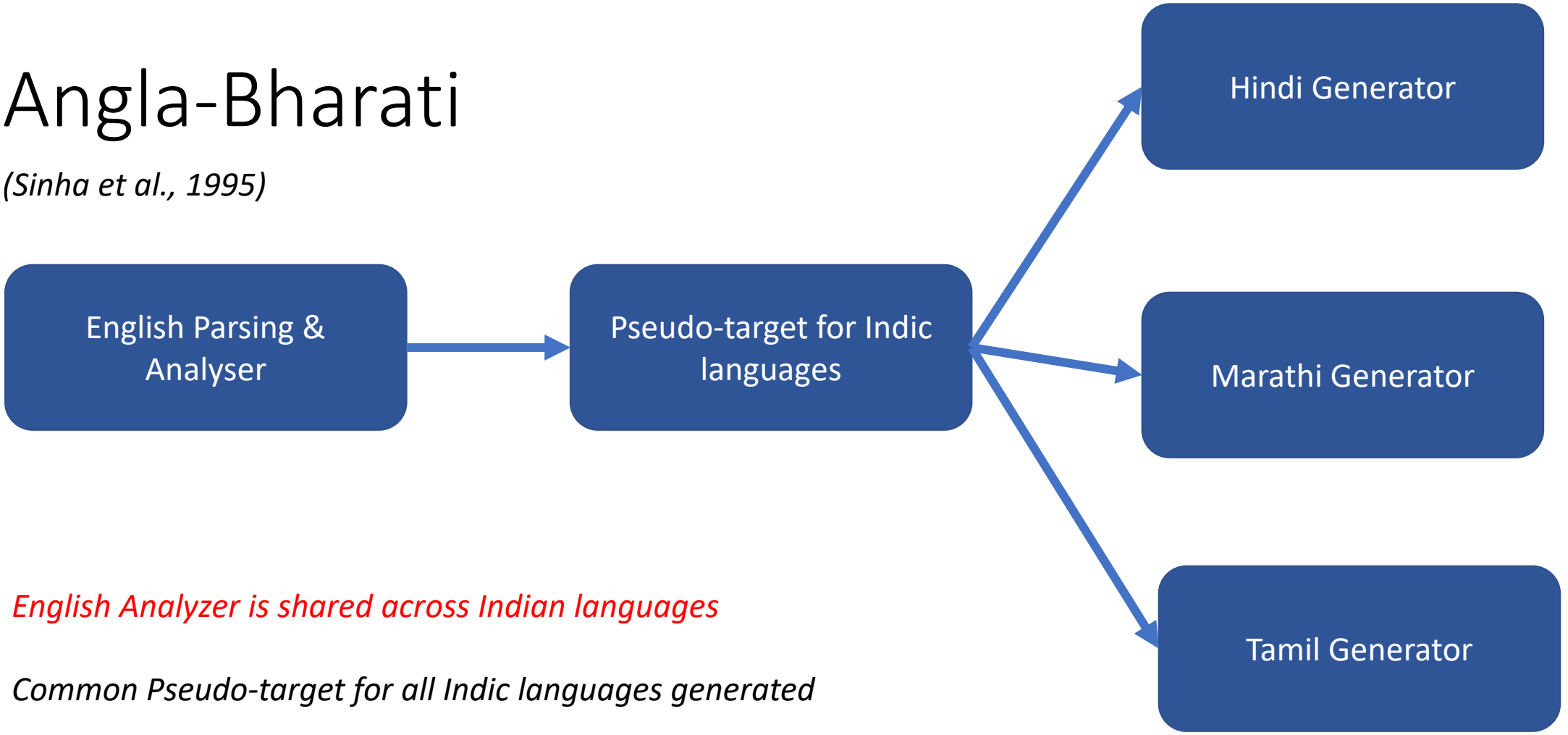
Syntactic Similarity

Syntactic Similarity between Indian languges

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

Angla-Bharati

(Sinha et al., 1995)



English Analyzer is shared across Indian languages

Common Pseudo-target for all Indic languages generated

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

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**

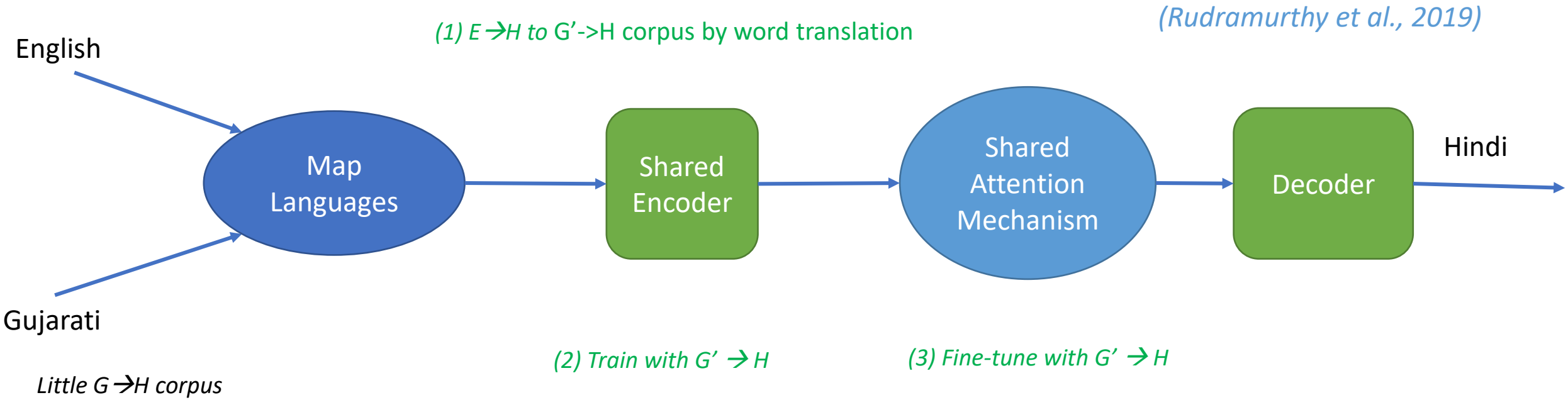
A common set of rules can be written for all Indian languages

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

https://github.com/anoopkunchukuttan/cfilt_preorder

Bridging Word-order Divergence for low-resource NMT

(Rudramurthy et al., 2019)



Cannot ensure similar Gujarat and English words have similar representations

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

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

Exploiting syntactic similarity in IL-IL translation

Can reduce search choices and errors, improve decoding speed

RMT: No need to handle long-distance reordering.

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

SMT: Monotonic Decoding, subword models.

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

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

	tel	tam	mal
Baseline	7.70	6.53	3.91
Generic	7.84	6.82	**4.05**
Hindi-tuned	**8.16**	**7.08**	4.02

Source reordering for SMT using Hindi-driven rules

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

Addressing syntactive divergence in NMT using Hindi-driven rules

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

Tamil to English NMT with transfer-leaning using Hindi

Summary

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

Thank You!

anoop.kunchukuttan@gmail.com

http://anoopk.in

References

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

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

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

51. Subbārāo, K. V. (2012). South Asian languages: A syntactic typology. Cambridge University Press.
52. Tao, T., Yoon, S.-Y., Fister, A., Sproat, R., and Zhai, C. (2006). Unsupervised named entity transliteration using temporal and phonetic correlation. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.
53. Tiedemann, J. (2009a). Character-based PBSMT for closely related languages. In Proceedings of the 13th Conference of the European Association for Machine Translation (EAMT 2009).
54. Trubetzkoy, N. (1928). Proposition 16. In Actes du premier congres international des linguistes à La Haye.
55. Vilar, D., Peter, J.-T., and Ney, H. (2007). Can we translate letters? In Proceedings of the Second Workshop on Statistical Machine Translation.
56. Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. EMNLP.