Machine Translation for Related Languages

Anoop Kunchukuttan

Microsoft AI and Research

Center for Indian Language Technology
Indian Institute of Technology Bombay

Microsoft Academic Accelerator (AXLE), 15th May 2018

What is Machine Translation?

Automatic conversion of text/speech from one natural language to another

Be the change you want to see in the world

वह परिवर्तन बनो जो संसार में देखना चाहते हो

Related Languages

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!

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

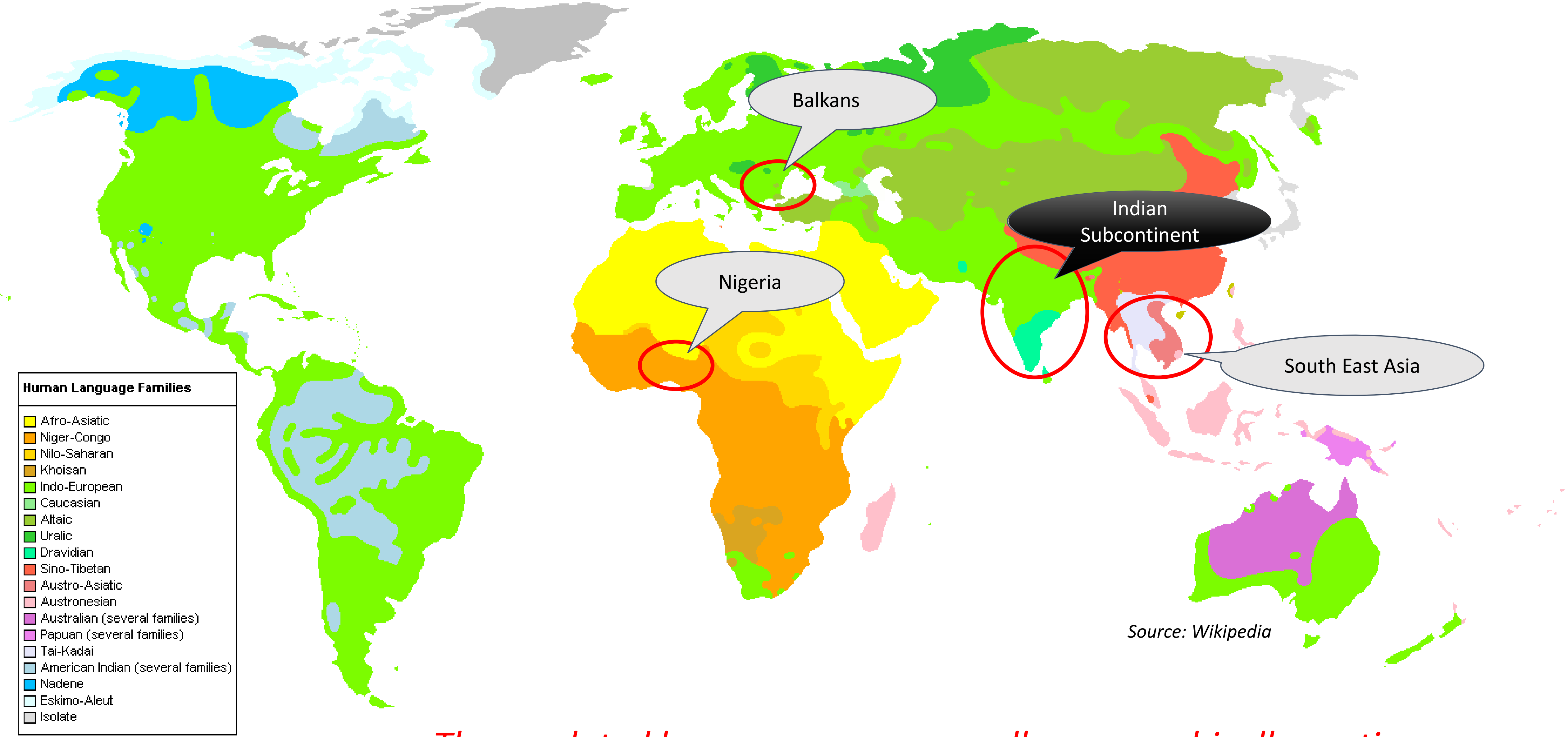
Hindi

Lexical: share significant vocabulary (cognates & loanwords)

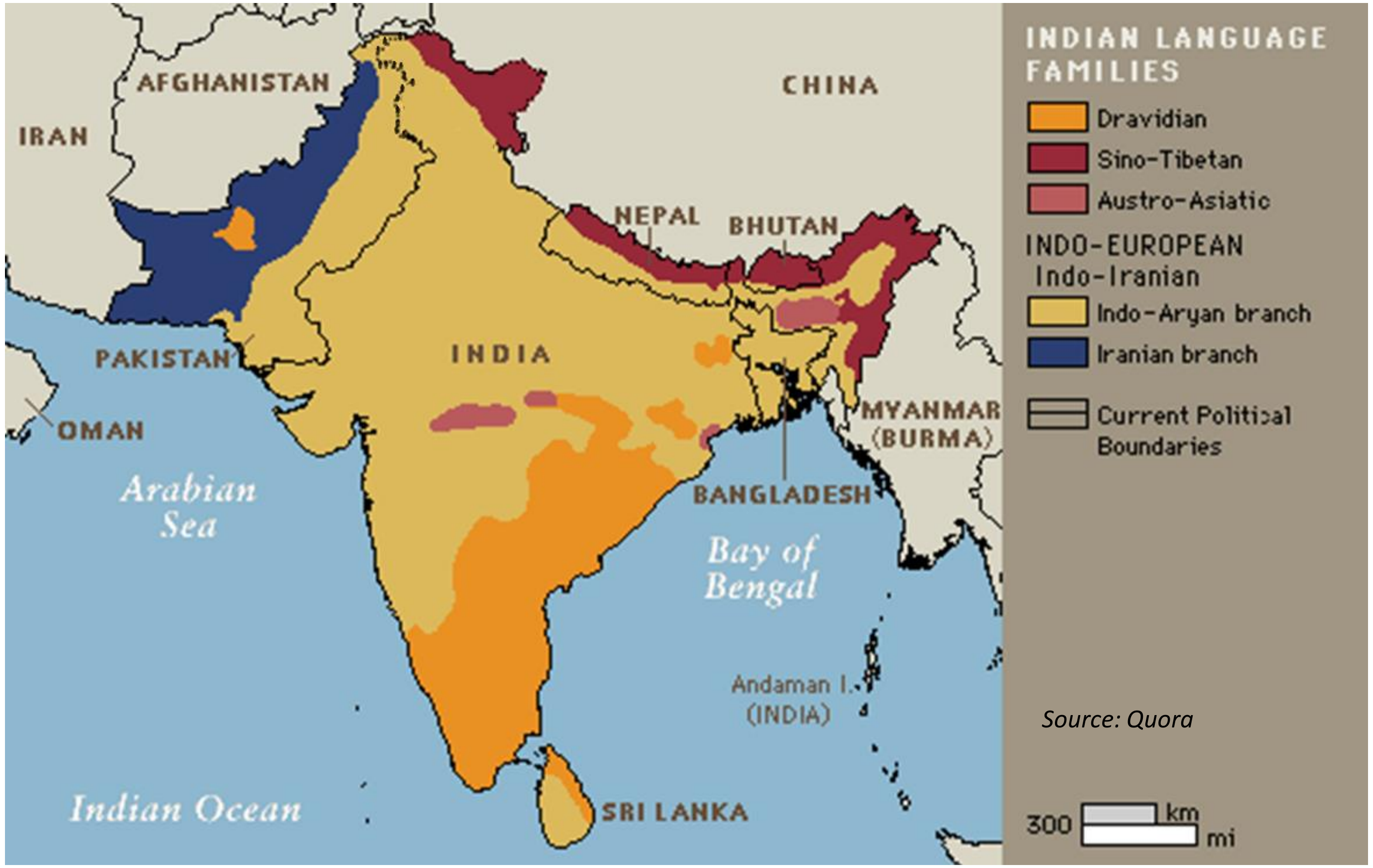
Morphological: correspondence between suffixes/post-positions

Syntactic: share the same basic word order

Why are we interested in such related languages?



These related languages are generally geographically contiguous



- 5 language families (+ 2 to 3 on the Andaman & Nicobar Islands)
- 22 scheduled languages
- 11 languages with more than 25 million speakers
- Highly multilingual country

Naturally, lot of communication between such languages (government, social, business needs)

⇓

Most translation requirements also involves related languages

Between related languages

Hindi-Malayalam

Marathi-Bengali

Czech-Slovak

Related languages ⇐⇒ Link languages

Kannada,Gujarati ⇒ English

English ⇒ Tamil,Telugu

We want to be able to handle a large number of such languages

e.g. 30+ languages with a speaker population of 1 million + in the Indian subcontinent

Is vanilla Statistical Machine Translation not sufficient?

Parallel Corpus	
A boy is **sitting** in the kitchen	एक लडका रसोई मे **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 है**
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 है**
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 मे **बैठा** है

Machine Learning

* Learn word/phrase alignments
* Learning to reorder

Let's begin with a simplified view of Statistical Machine Translation (SMT)!!

Co-occurrence is the dominant learning signal

Makes SMT language independent

Hence very popular →

- Over-reliance on co-occurrence alone increases parallel corpus requirements
- Problem is grave for agglutinative languages
  - *e.g. Marathi, Dravidian languages*
  - *घरासमोरचा* ➔ *घर* + *समोर* + *चा*
- Language-specific learning signals are ignored

Aren’t “*language independent*” Statistical/Neural Machine Translation methods sufficient?

- Implicit assumptions increase need for:

  (1) Parallel Corpora (2) Linguistic Resources (3) Language specific processing

- *‘Limited language independence’* can be achieved between some languages if we can make assumptions that hold across all these languages

- Related languages can serve as a good level of abstraction to utilize linguistic regularities:
  - Reduce parallel corpora
  - Reduce linguistic resource requirements
  - Better Generalization

Utilizing Lexical Similarity for Subword-level translation

Kunchukuttan & Bhattacharyya, EMNLP (2016)

Kunchukuttan & Bhattacharyya, SCLeM (2017)

Lexically Similar Languages

*(Many 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.mbal (hi)	mu.mbal (pa)	mu.mbal (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*

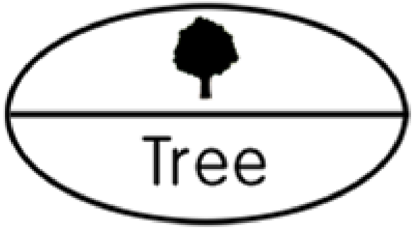
MWE with non-compositional semantics

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

Why do we use word-level translation?

- MT learns mappings between meaning bearing linguistic units ➔ *Words and Morphemes*

- Why? ⇒ Fundamental principle of linguistics
  - *Arbitrariness of a word's form and meaning (Saussure, 1916)*



- Is the mapping between forms of similar words across languages arbitrary?
  - *Probably true in the most general case*
  - *Not true for related languages due to lexical similarity*

Utilize lexical similarity between related languages: Sub-word level transformations

Related Work

Transliterate unknown words *[Durrani, etal. (2010), Nakov & Tiedemann (2012)]*

(a) Primarily used to handle proper nouns *(b) Limited use of lexical similarity*

स्वातंत्र्य → स्वतंत्रता

Translation of shared lexically similar words can be seen as kind of transliteration

Is there a better translation unit?

Character Level Translation *[Vilar, etal. (2007), Tiedemann (2009)]*

Limited context of character level representation

Character n-gram ⇒ increase in data sparsity

Limited benefit ….

… just for closely related languages

Macedonian - Bulgarian, Hindi-Punjabi, etc.

Orthographic Syllable *(Kunchukuttan & Bhattacharyya, EMNLP 2016)*

(CONSONANT) + VOWEL

Examples: ca, cae, coo, cra, की (kI), प्रे (pre)

अभिमान ➔ अ भि मा न

<u>Pseudo-Syllable</u>

True Syllable ⇒ Onset, Nucleus and Coda

Orthographic Syllable ⇒ Onset, Nucleus

- Generalization of ***akshara***, the fundamental organizing principle of Indian scripts
- Linguistically motivated, variable length unit
- *Number of syllables in a language is finite*
- Used successfully in transliteration

Byte Pair Encoded (BPE) Unit

(Kunchukuttan & Bhattacharyya, SCLeM 2017)

- *There may be frequent subsequences in text other than syllables*
- *Herdan-Heap Law* $\Rightarrow$ *Syllables are not sufficient*
- *These subsequences may not be valid linguistic units*
- *But they represent statistically important patterns in text*

How do we identify such frequent patterns?

Byte Pair Encoding (Sennrich et al, 2016), Wordpieces (Wu et al, 2016), Huffman encoding based units (Chitnis & DeNero, 2015)

Byte Pair Encoded (BPE) Unit

Byte Pair Encoding is a compression technique (Gage, 1994)

Number of BPE merge operations=3
Vocab: A B C D E F

P_1=AD P_2=EE P_3=P_1D

Words to encode	Iterations 1	2	3	4
BADD	BADD	BP_1D	BP_1D	BP_3
FAD	FAD	FP_1	FP_1	FP_1
FEEDE	FEEDE	FEEDE	FP_2DE	FP_2DE
ADDEEF	ADDEEF	P_1DEEF	P_1DP_2F	P_3P_2F

Data-dependent segmentation

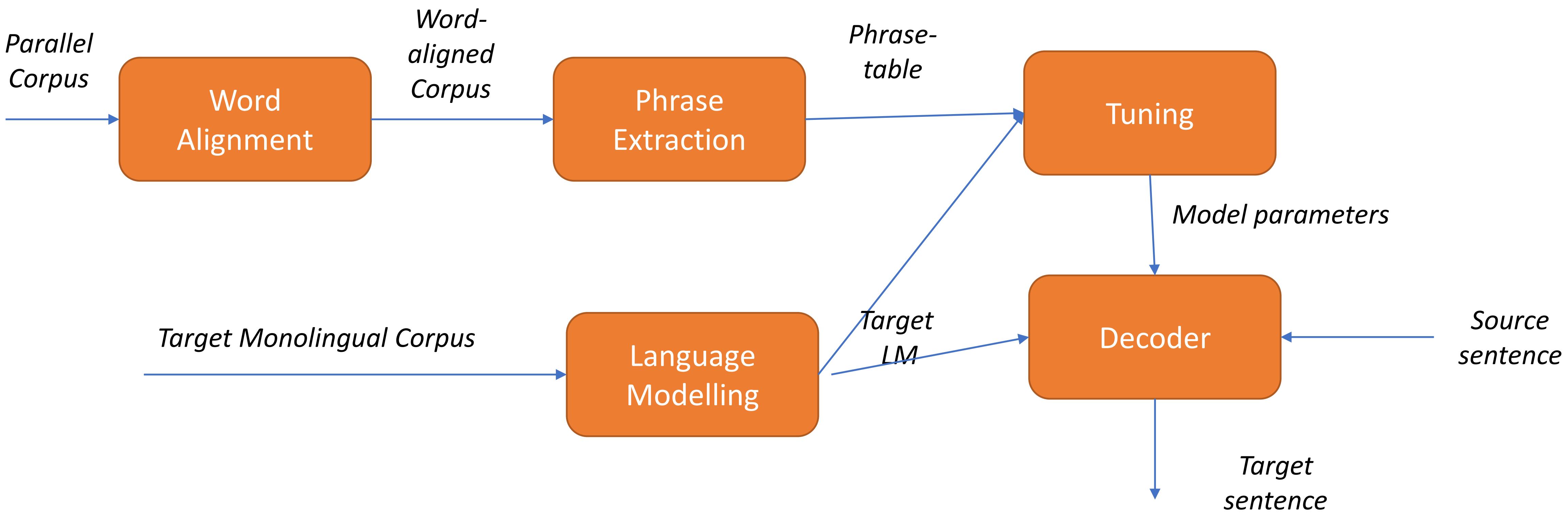
- Inspired from compression theory
- MDL Principle *(Rissansen, 1978)* ⇒ Select segmentation which maximizes data likelihood

Example of various translation units

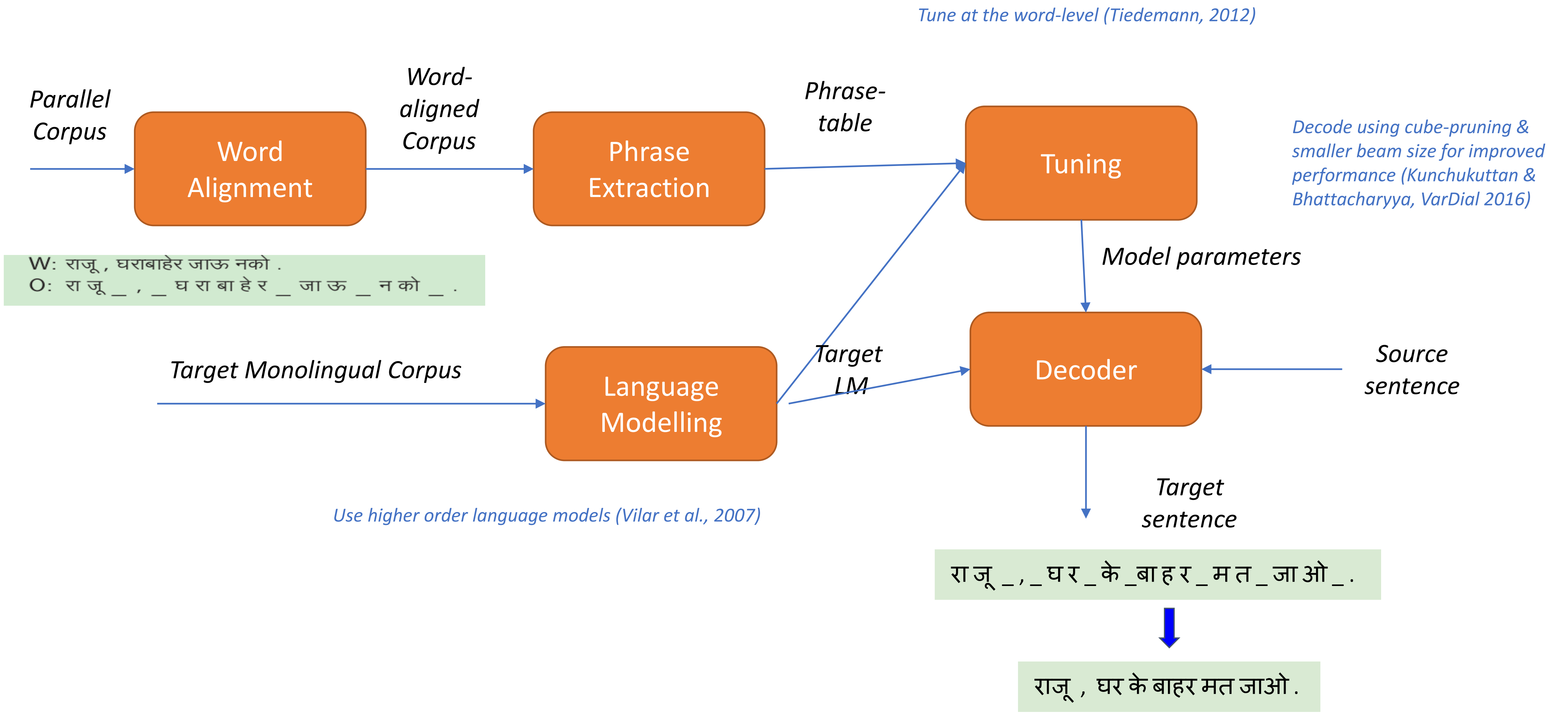
Basic Unit	Symbol	Example	Transliteration
Word	W	घरासमोरचा	gharAsamoracA
Morph Segment	M	घरा समोर चा	gharA samora cA
Orthographic Syllable	O	घ रा स मो र चा	gha rA sa mo racA
Character unigram	C	घ र ◌ा स म ◌ो र च ◌ा	gha r A sa m o ra c A
something that is in front of home:		ghara=home, samora=front, cA=of	

Various translation units for a Marathi word

Typical SMT Pipeline



Adapting SMT for subword-level translation



Comparison of subword level units

	OS	**BPE**
Unit	pseudo-syllable	frequent char sequence
Motivation	Linguistic ⇒ approximate syllable	Statistical ⇒ Minimum Description Length
Length	Variable length	Variable length
Vocab size	Some mutiple of |char_set|	Some mutiple of |char_set|
OOV	Few	No
Extraction	Rule-based	Data-oriented
Script	Should use vowels	Any

Experiments: Language Pairs & Datasets

have vowels

Language Family		Type of writing system	
Dravidian	mal,tam,tel	Alphabet	dan[1],swe[1],may[1] ind[1],buc[2],mac[2]
Indo-Aryan	hin,urd,ben kok,mar,pan	Abugida	mal,tam,tel,hin ben,kok,mar,pan
Slavic	bul,mac		
Germanic	dan,swe	Syllabic	kor
Polynesian	may,ind	Logographic	jpn
Altaic	jpn,kor	Abjad	urd

1: Latin
2: Cyrillic

doesn't have vowels

6 language groups, 17 languages, 5 types of writing systems, 11 writing systems

Datasets: **ILCI corpus** (for Indian languages, ~50k), **OPUS corpus** (non-Indic languages, ~150k)

Results for languages using abugida and alphabetic scripts

Src-Tgt	Char	Word	Morph	OS	BPE
ben-hin	27.95	32.47	32.17	**33.54**	33.22
pan-hin	71.26	70.07	71.29	**72.41**	72.22
kok-mar	19.83	21.30	22.81	23.43	**23.63**
mal-tam	4.50	6.38	7.61	7.84	**8.67**†
tel-mal	6.00	6.78	7.86	8.50	**8.79**
hin-mal	6.28	8.55	9.23	10.46	**10.73**
mal-hin	12.33	15.18	17.08	18.44	**20.54**
bul-mac	20.61	21.20	-	**21.95**	21.73
dan-swe	35.36	35.13	-	35.46	**35.77**
may-ind	60.50	**61.33**	-	60.79	59.54†

- Substantial improvement over char-level model (27% & 32% for OS and BPE resp.)
- Char-level model is competitive only when languages are very closely related
  - else even word outperforms char
- Significant improvement over word and morph level baselines (11-14% and 5-10% resp)
- Improvement even when languages don't belong to same family (contact exists)
- More beneficial when languages are morphologically rich
- BPE slightly better than OS (2.5%)
  - not statistically significant

Comparison with post-processing using transliteration

Language Pair	Word$_X$	Morph$_X$	OS	BPE
ben-hin	32.79	32.32	**33.54**	33.22
pan-hin	71.71	71.42	**72.41**	72.22
kok-mar	21.9	22.82	23.43	**23.63**
mal-tam	7.01	7.65	7.84	**8.67**
tel-mal	6.94	7.89	8.50	**8.79**
hin-mal	8.77	9.26	10.46	**10.73**
mal-hin	16.26	17.3	18.44	**20.54**

Significant improvement over strong baselines: Word$_X$ (10%) & Morph$_X$ (5%)

Results for languages using non-vowel scripts

Src-Tgt	Char	Word	Morph	BPE
urd-hin	52.57	55.12	52.87	**55.55**
ben-urd	18.16	27.06	27.31	**28.06**
urd-mal	3.13	6.49	7.05	**8.44**
mal-urd	8.90	13.22	15.30	**18.48**
kor-jpn	8.51	9.90	-	**10.23**
jpn-kor	8.17	8.44	-	**9.02**

- Orthographic syllables cannot be used
- BPE units outperform both word and morph units. Over word based:
  - 18% improvement for Urdu pairs
  - 6% improvement for kor-jpn pairs
- More improvement when morphologically rich languages are involved

BPE works well for non-vowel scripts also

Some Illustrations from Hindi-Malayalam translation

	English	Hindi	Malayalam
Translates cognates	time	samaya	samaya.m
Translates non-cognates	door	darvAzA	vatila
Translates morphological suffixes	ago	pahale	munpe
False friends can cause problems	**chintA**	**worry**	**thought**

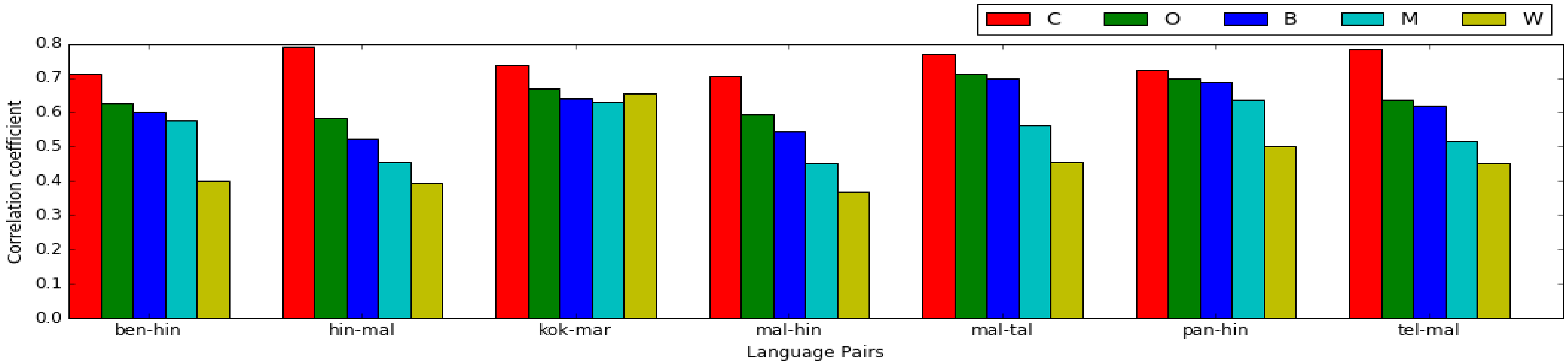
Why do OS and BPE outperform other units?

Reduction in vocabulary size

- Addresses Data Sparsity
- Ability to learn Diverse Lexical Mappings
- Judicious use of Lexical Similarity

	hin	mar	mal
OS	tI, stha	mA, nA	kka, nI
Suffix	ke, me.m	ChyA, madhIla	unnu, .e∼Nkill.m
Word	paryaTaka, athavA	prAchIna, aneka	bhakShaN.m, yAtra

Judicious use of Lexical Similarity



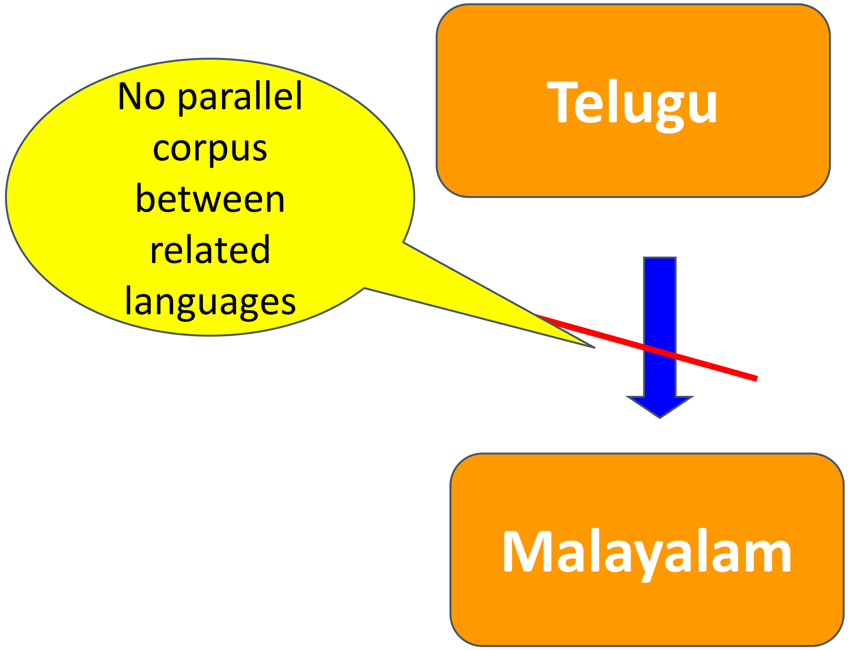
Pearson's correlation coefficient between translation accuracy & lexical similarity (sentence level using LCSR)

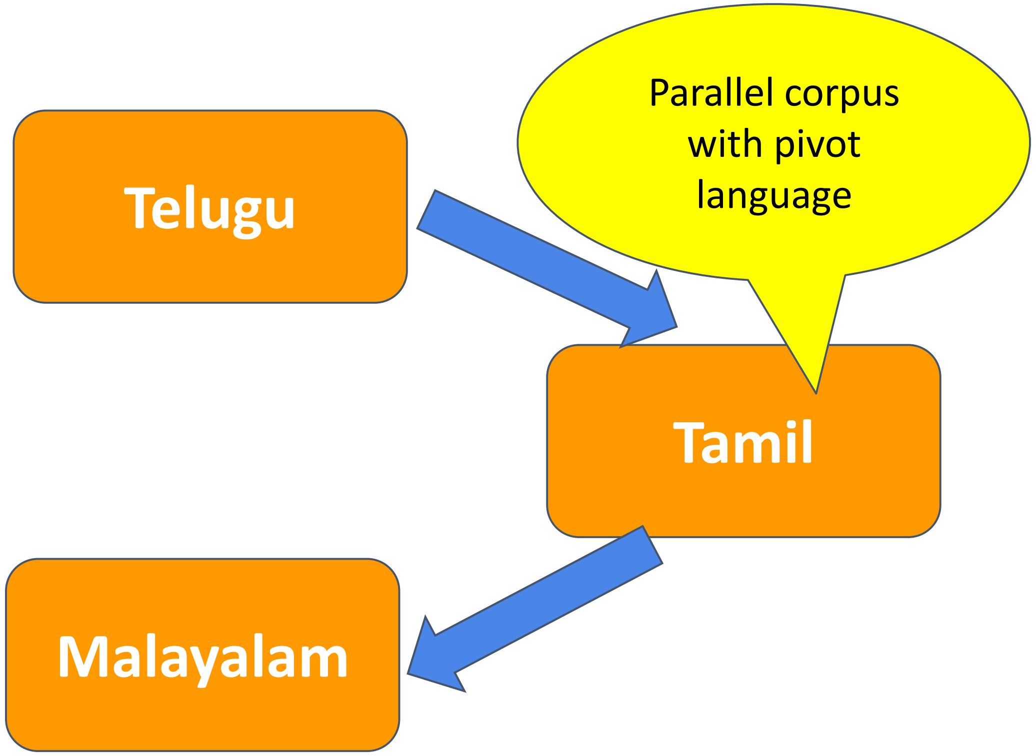
1. Morph and Word doesn't sufficiently utilize lexical similarity
   - Word level is least correlated
   - Morph level output is less correlated than BPE or OS
2. Character level performance highly correlated with lexical similarity
   - Little context for translation ⇒ learns character transliterations
3. OS & BPE strike a balance between using lexical similarity and word-level information

Utilizing Lexical Similarity between related, low resource languages for Pivot based SMT

Kunchukuttan et al., IJCNLP (2017)

Utilizing Lexical Similarity for Pivot-based SMT





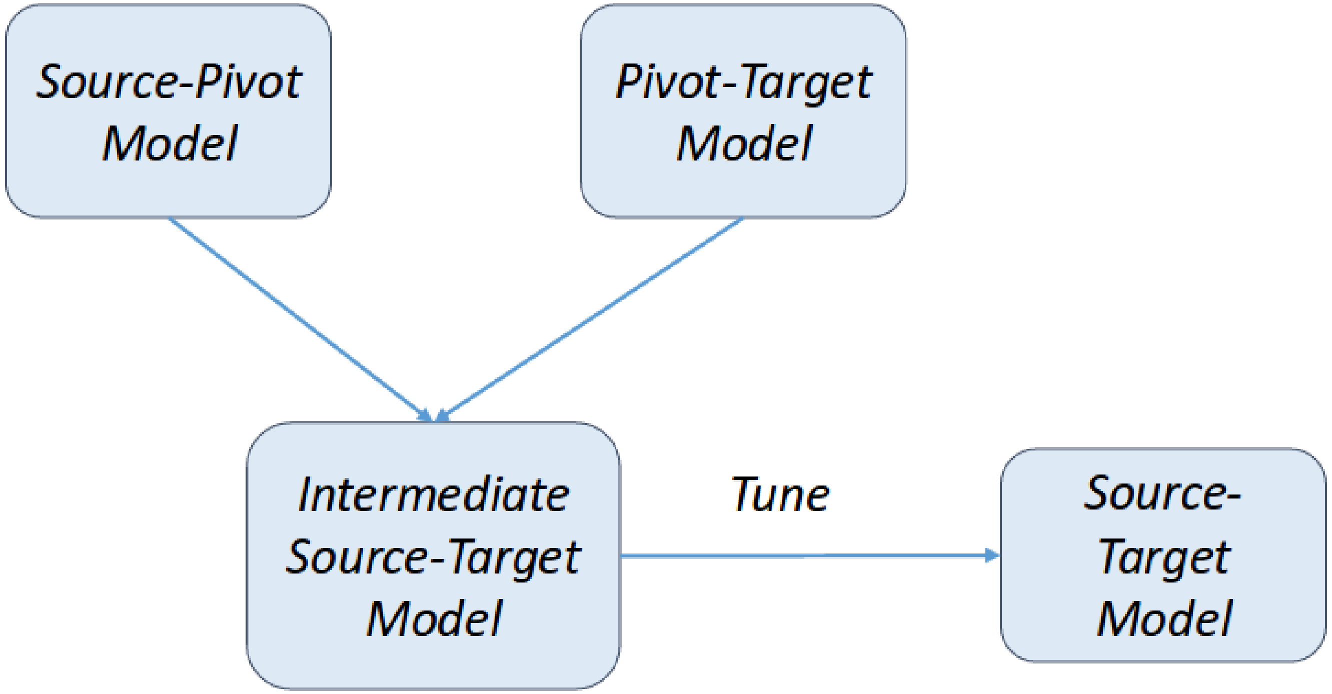
Related languages ⇒ Use subword level translation units

Translation through intermediate language ⇒ Use Pivot based SMT methods

Our work brings together these two strands of research

Triangulation of Pivot Tables *(Utiyama & Isahara,2007; Wu & Wang, 2007)*

Pivot related to source & target ⇒ subword level



src-pivot phrase table

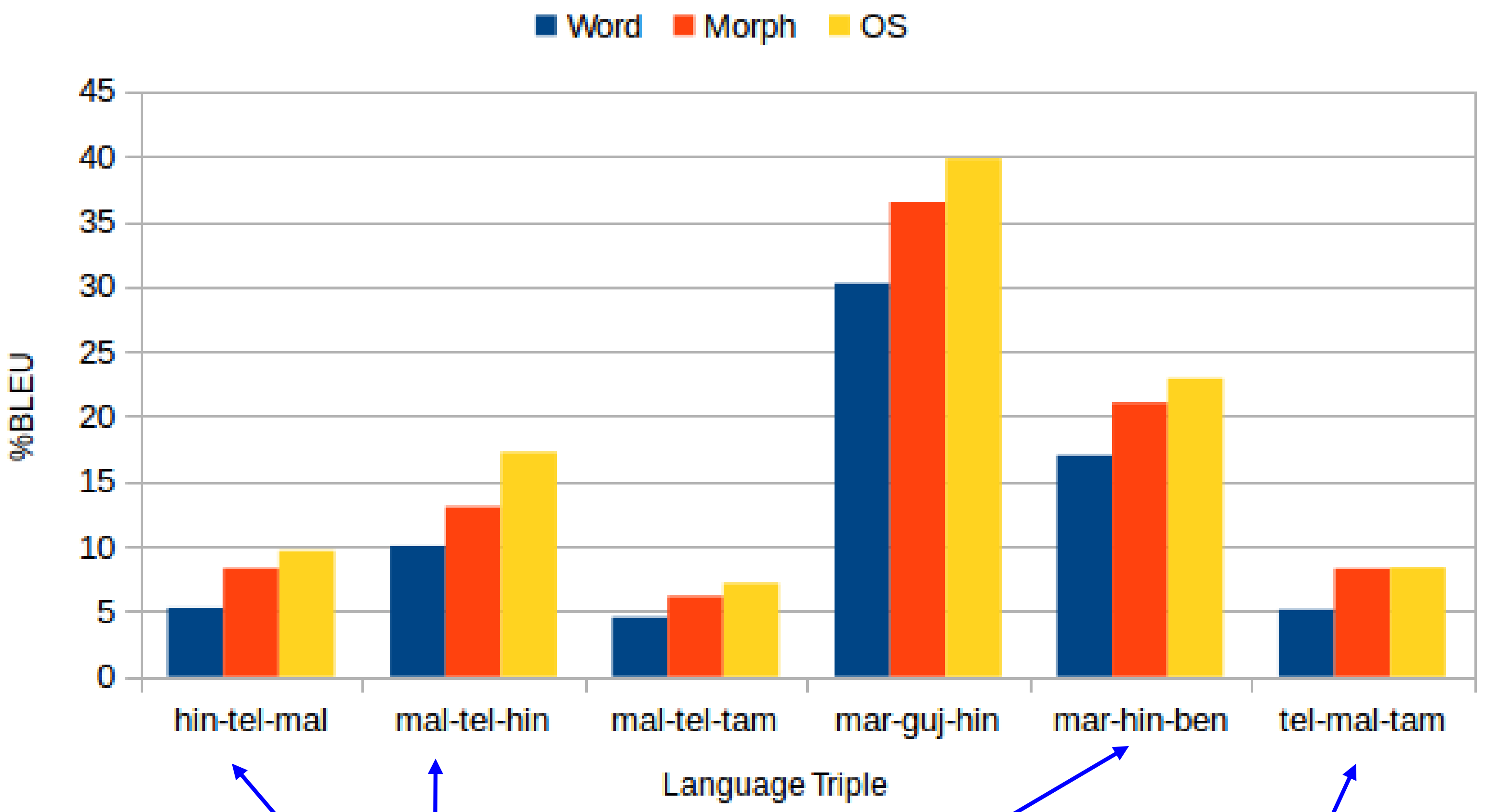
A	X	0.4	0.4
B	X	0.6	0.8
B	Y	0.8	0.9
C	Y	0.2	0.1

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

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

Comparison of translation units for pivot SMT



Indo-Aryan

Indo-Aryan ⇆ Dravidian

Dravidian

OS level pivot system outperforms other units

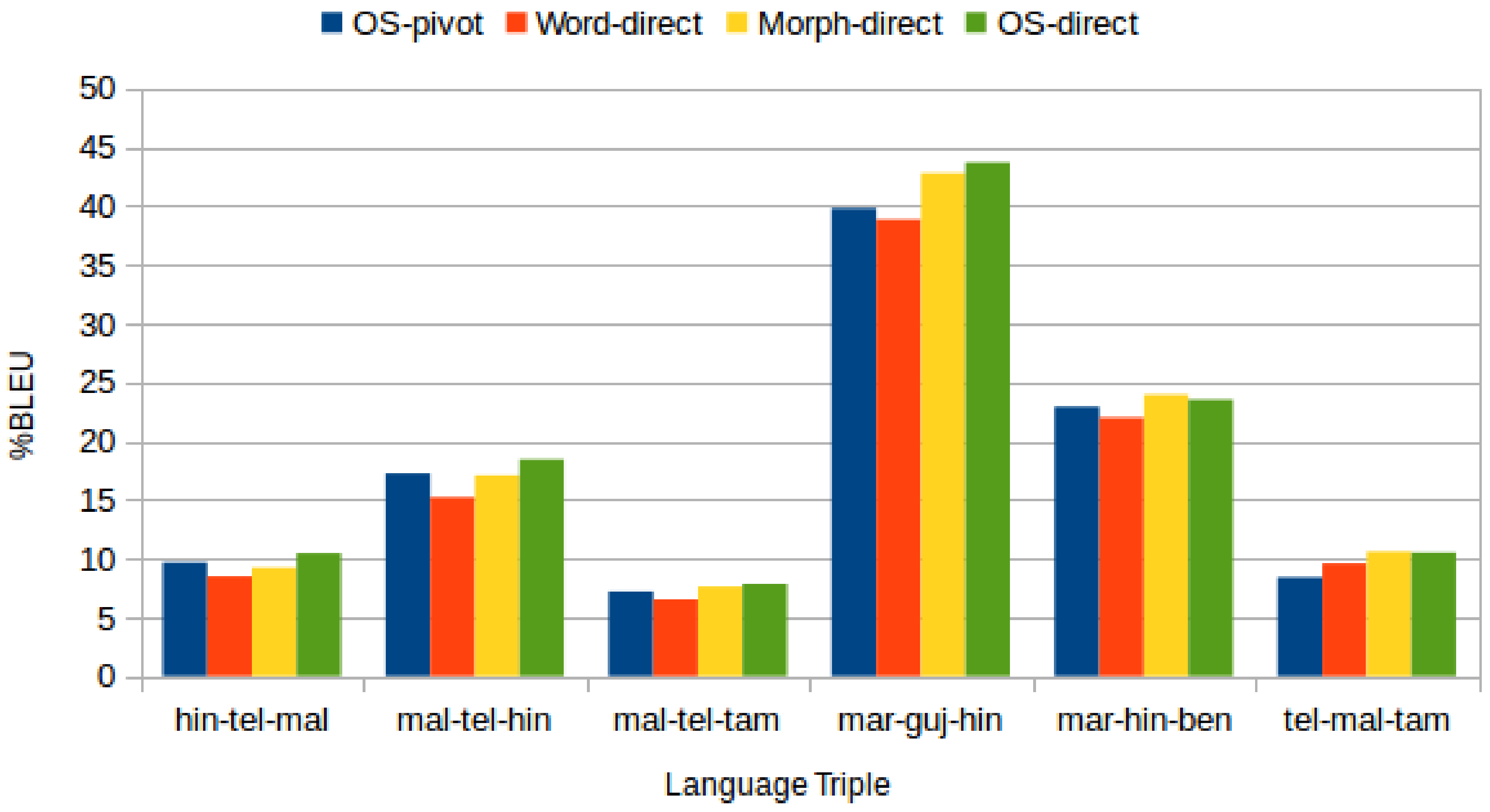
- *~60% improvement over word level*
- *~15% improvement over morph level*

Why is OS level pivot better?

Better direct source-pivot & pivot-target translation systems

Triangulation at OS level faces lesser data sparsity

Comparison of OS level pivot with direct models



Better than word level direct model
(~5% improvement)

Competitive with direct morph and OS level models
(~95 and 90% respectively of the direct system scores)

OS level system is competitive with the best word and morph level direct systems

Can multiple pivot languages do better?

Model	mar-ben	mal-hin
best pivot	22.92 (*hin*)	17.52 (*tel*)
direct	23.53	18.44
all pivots	23.69	19.12
direct+all pivots	24.41	19.44

Pivots used for ...
mar-ben: guj, pan, hin
mal-hin : tel, mar, guj

Combining multiple pivot systems can outperform direct systems also

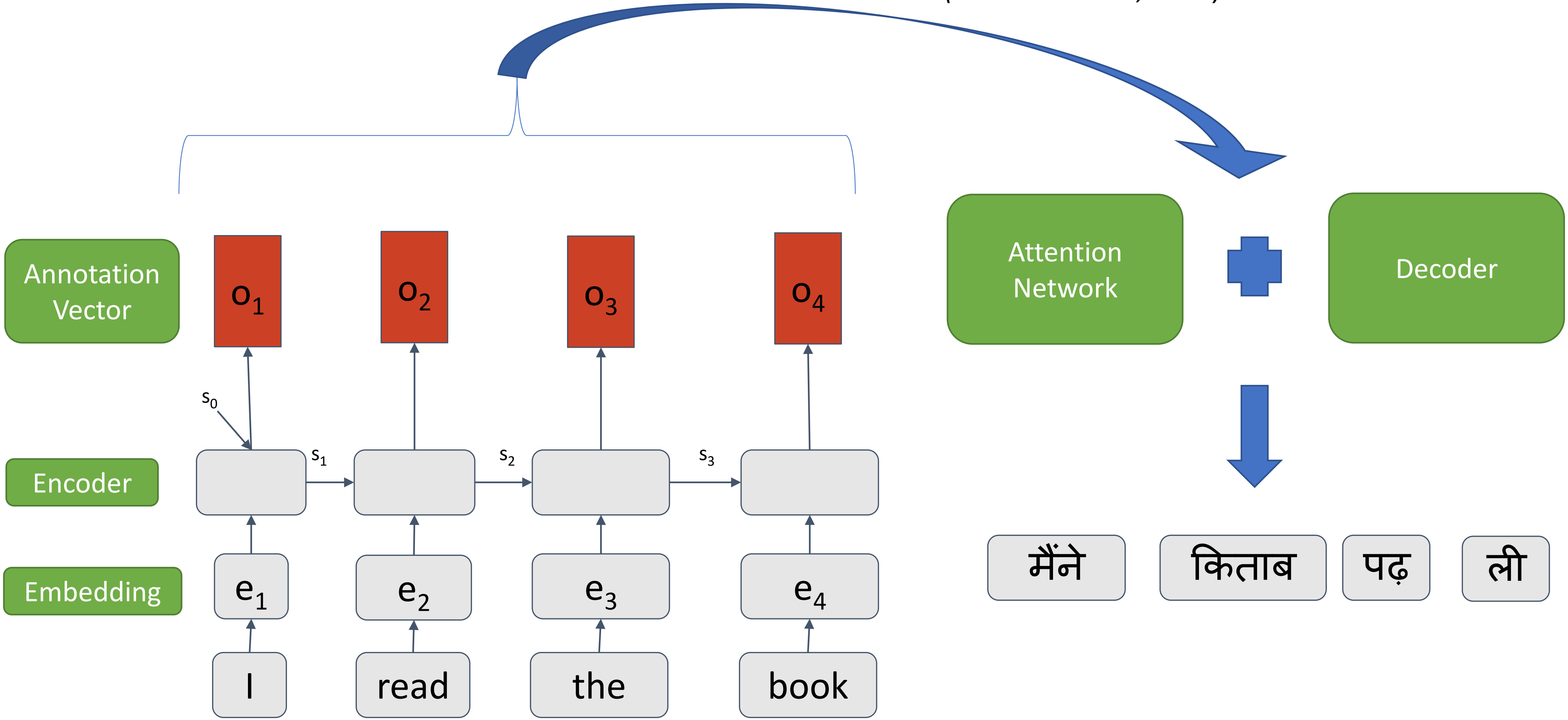
This cannot be achieved with word/morph level pivoting

Linear Interpolation with equal weights used to combine phrase tables

Multilingual Neural Machine Translation

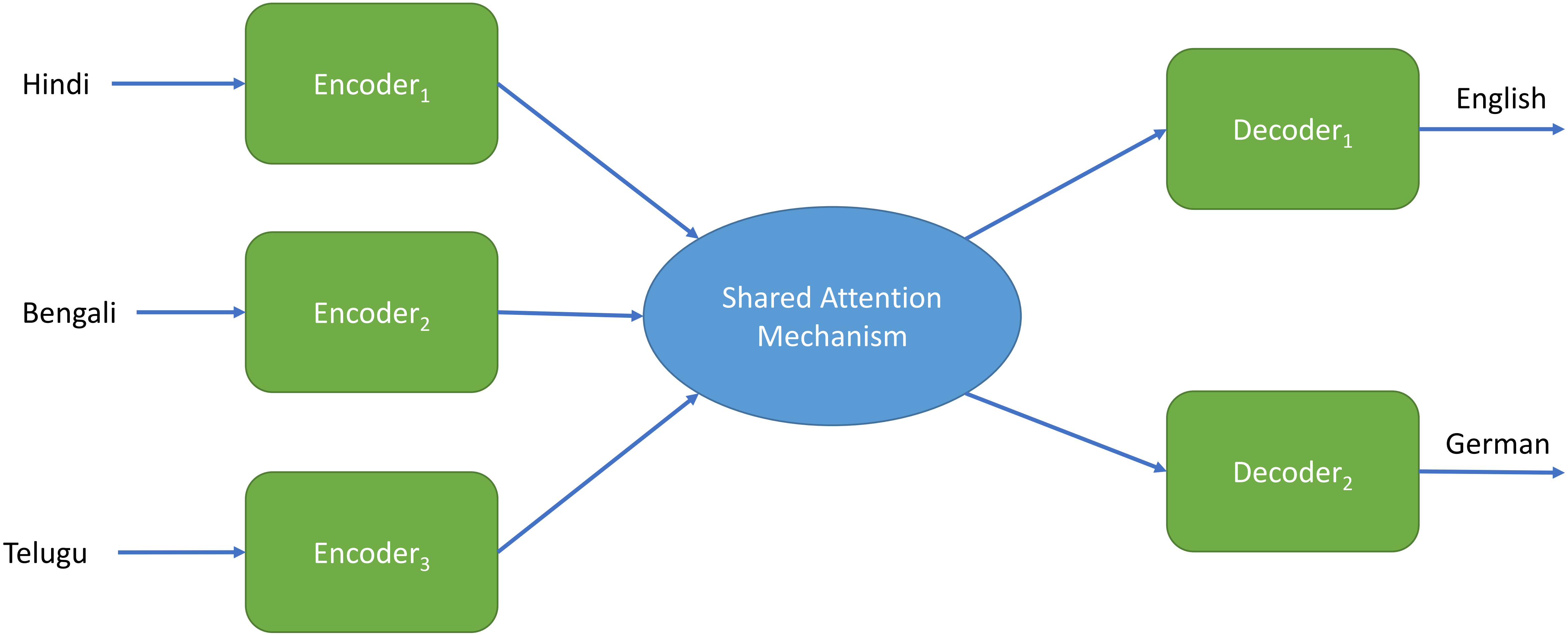
Embed - Encode - Attend - Decode Paradigm

(Bahdanau et al, 2015)



Multilingual Neural Translation

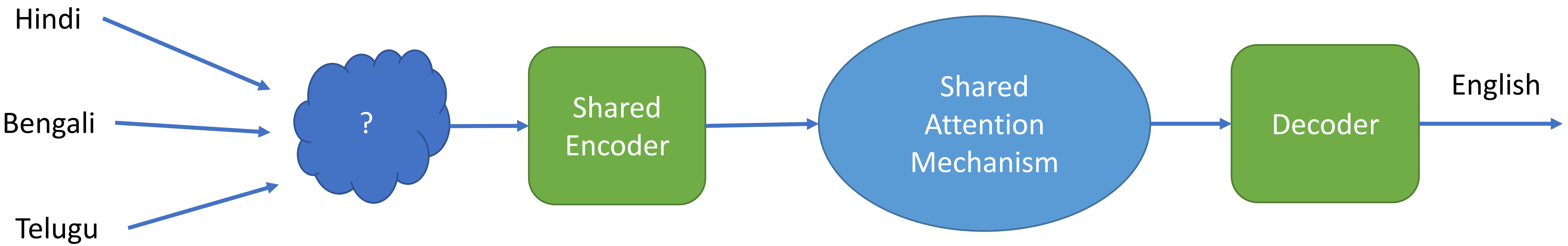
(Firat et al., 2016; Johnson et al., 2017)



Translate unseen language pairs ➔ Zeroshot Translation

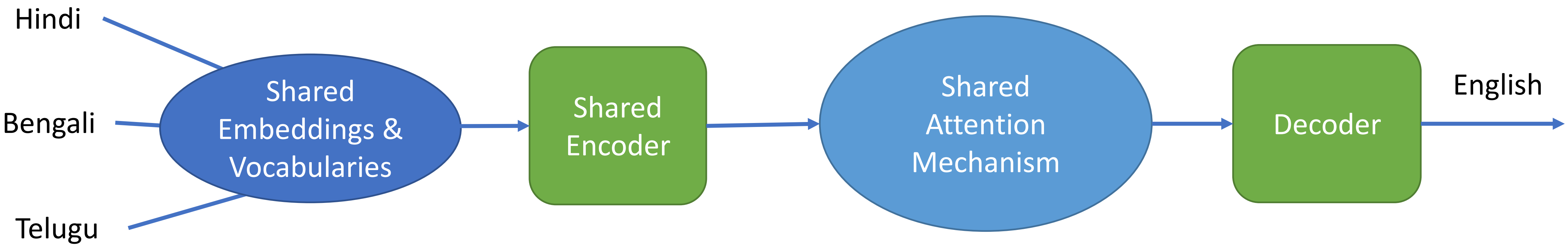
Shared Encoder

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



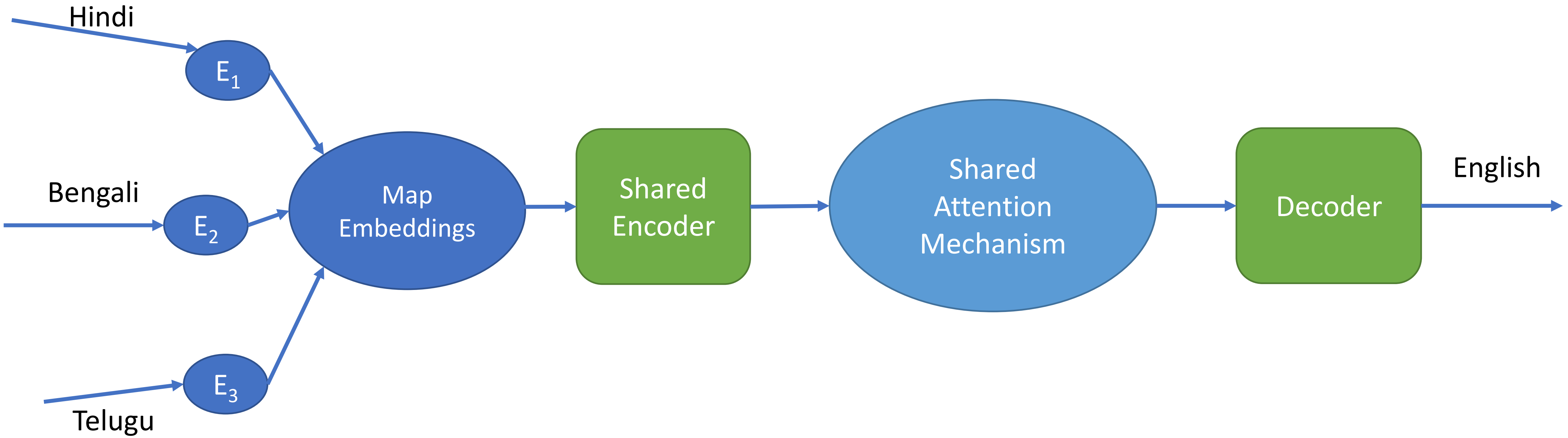
Shared Encoder

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



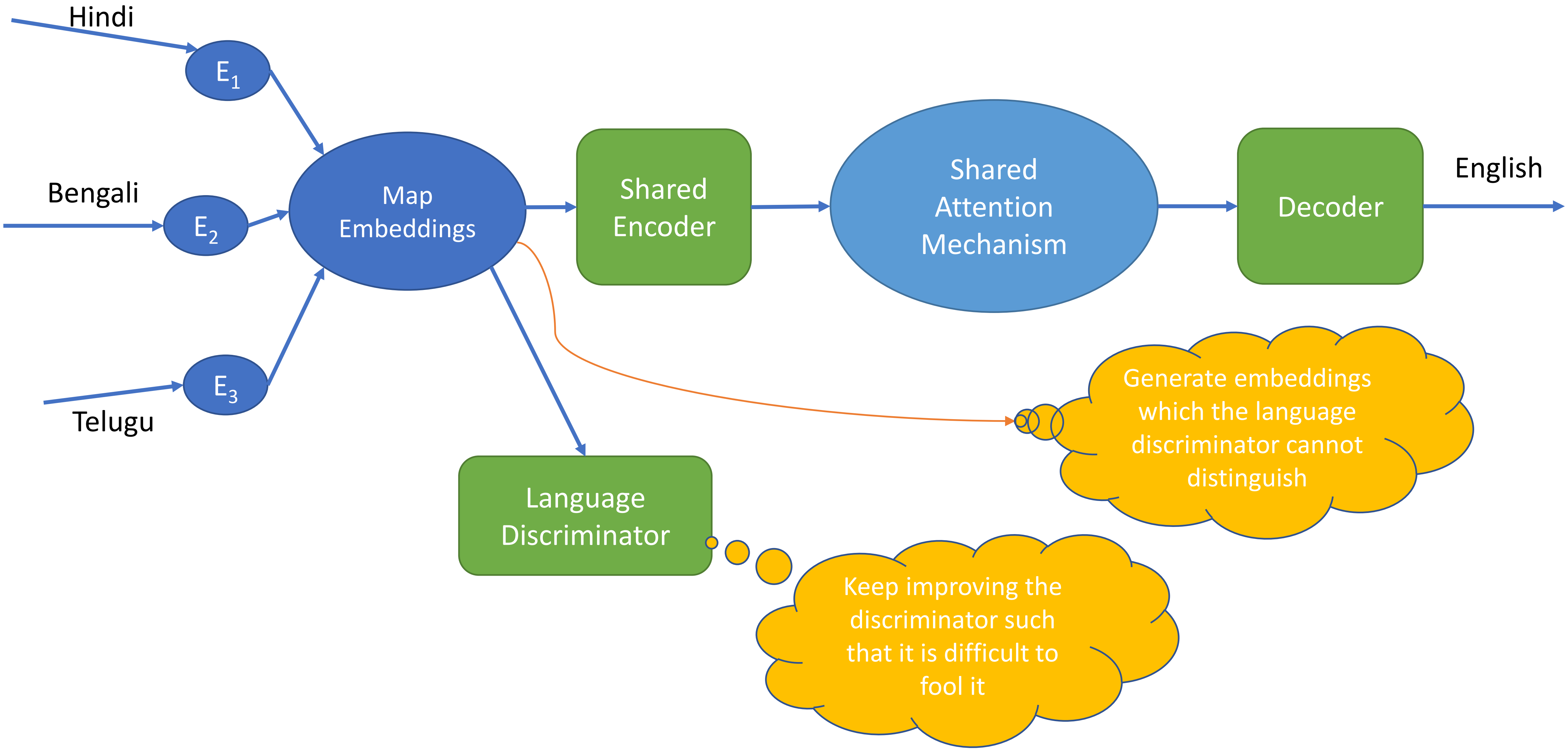
Shared Encoder

(Gu et al., 2018)

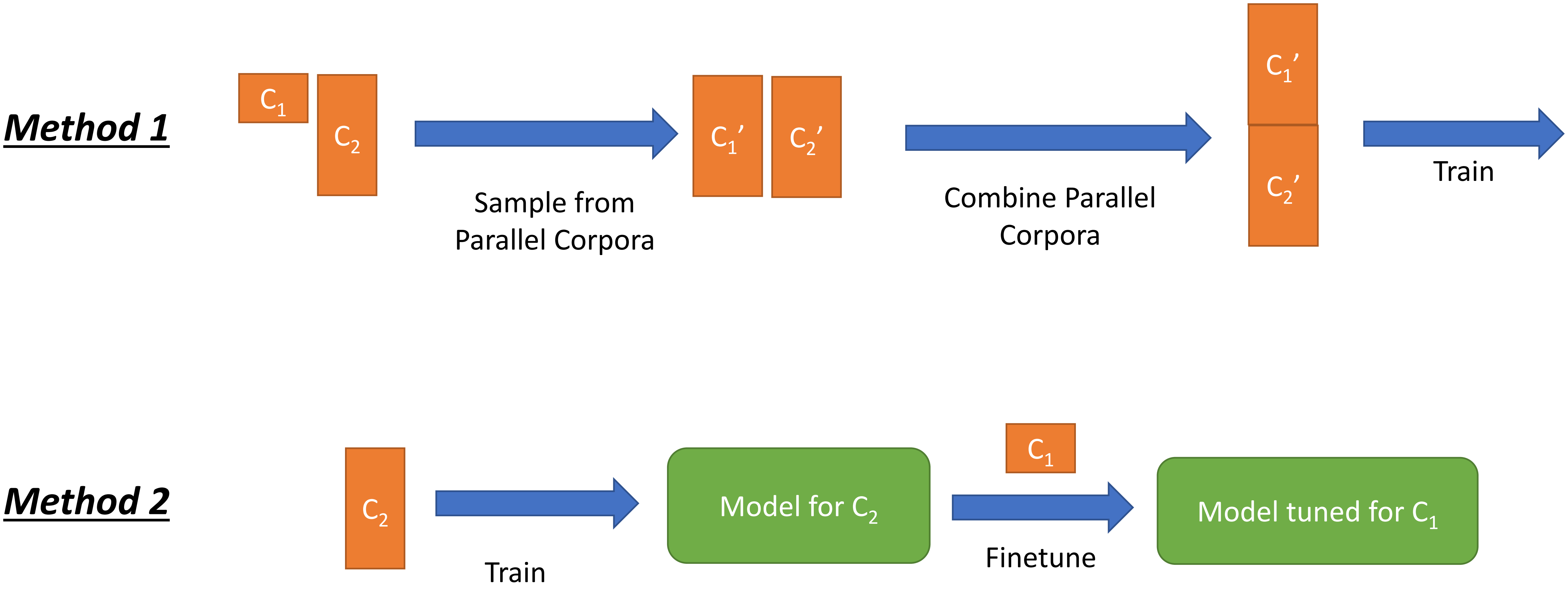


Shared Encoder with Adversarial Training

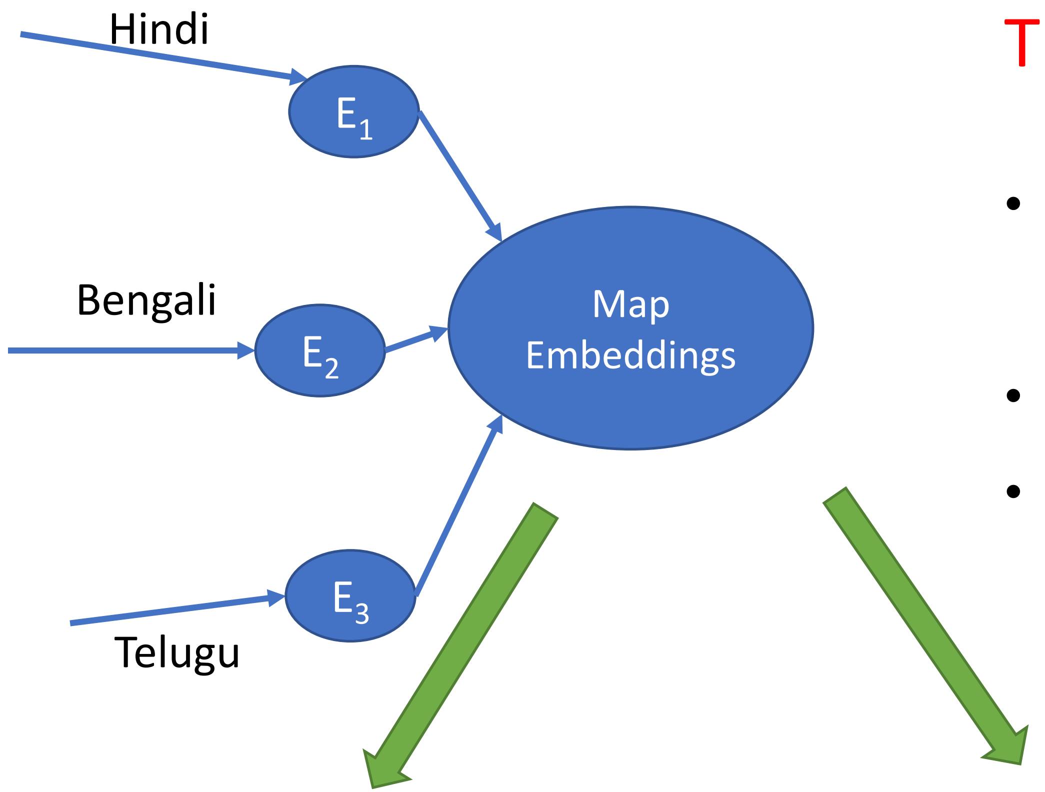
(Joty et al., 2017)



Training Multilingual NMT systems



Learning Multilingual mappings/embeddings



The key to multilingual learning

- Text Classification (sentiment analysis, Question matching)
- Sequence Tagging (POS, NER, etc.)
- Sequence to Sequence Learning (Machine Translation, Transliteration, etc)

Offline Mapping of embeddings

$e_2 = A_{21}e_1$

Joint training for multilingual embeddings

$A_1e_1=A_2e_2$

Needs parallel corpora or bilingual dictionaries

Summary and Future Directions

- Related Languages serve as an important level of abstraction for building MT systems
- Utilizing lexical similarity can reduce parallel corpus requirements
- Combining lexical similarity and multilingual learning can provide significant improvements in translation quality
- Advances in Transfer Learning and Adversarial Learning are interesting direction for improving multilingual learning
- Learning good multilingual embeddings efficiently can help make NLP applications multilingual

Multilingual data, code for Indian languages

http://www.cfilt.iitb.ac.in

https://www.cse.iitb.ac.in/~anoopk

Thank you!

Work with Prof. Pushpak Bhattacharyya, Prof. Mitesh Khapra, Abhijit Mishra, Ratish Puduppully, Rajen Chatterjee, Ritesh Shah, Maulik Shah, Pradyot Prakash, Gurneet Singh, Raj Dabre, Rohit More

References

1. Abbi, A. (2012). Languages of india and india and as a linguistic area. http://www.andamanese.net/LanguagesofIndiaandIndiaasalinguisticarea.pdf. Retrieved November 15, 2015.
2. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. ICLR 2015.
3. Caruana, R. (1997). Multitask learning. Machine learning.
4. De Saussure, F. (1916). Course in general linguistics. Columbia University Press.
5. 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.
6. 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.
7. Emeneau, M. B. (1956). India as a Lingustic area. Language.
8. Finch, A., Liu, L., Wang, X., and Sumita, E. (2015). Neural network transduction models in transliteration generation. In Proceedings of the Fifth Named Entities Workshop (NEWS).
9. 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.
10. Gillick, D., Brunk, C., Vinyals, O., and Subramanya, A. (2016). Multilingual language processing from bytes. NAACL.
11. Gispert, A. D. and Marino, J. B. (2006). Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In In Proc. of 5th International Conference on Language Resources and Evaluation (LREC).
12. Gordon, R. G., Grimes, B. F., et al. (2005). Ethnologue: Languages of the world, volume 15. SIL International Dallas, TX.
13. Gu, J., Hassan, H., Devlin, J., & Li, V. O. (2018). Universal neural machine translation for extremely low resource languages. NAACL.
14. Jha, G. N. (2012). The TDIL program and the Indian Language Corpora Initiative. In Language Resources and Evaluation Conference.
15. 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.
16. Joty, S., Nakov, P., Màrquez, L., & Jaradat, I. (2017). Cross-language Learning with Adversarial Neural Networks: Application to Community Question Answering. CoNLL.

References

17. Kunchukuttan, A., & Bhattacharyya, P. (2016). Orthographic syllable as basic unit for smt between related languages. EMNLP.
18. Kunchukuttan, A., & Bhattacharyya, P. (2016). Faster decoding for subword level Phrase-based SMT between related languages. VarDIAL.
19. Kunchukuttan, A., & Bhattacharyya, P. (2017). Learning variable length units for SMT between related languages via Byte Pair Encoding. SCLeM.
20. Kunchukuttan, A., Shah, M., Prakash, P., & Bhattacharyya, P. (2017). Utilizing Lexical Similarity between Related, Low-resource Languages for Pivot-based SMT. IJCNLP.
21. Lee, J., Cho, K., and Hofmann, T. (2017). Fully Character-Level Neural Machine Translation without Explicit Segmentation. Transactions of the Association for Computational Linguistics.
22. 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.
23. Nguyen, T. Q., & Chiang, D. (2017). Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. IJCNLP.
24. Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In ACL.
25. Subbārāo, K. V. (2012). South Asian languages: A syntactic typology. Cambridge University Press.
26. 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).
27. Tiedemann, J. (2009b). News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In Recent Advances in Natural Language Processing.
28. Tiedemann, J. and Nakov, P. (2013). Analyzing the use of character-level translation with sparse and noisy datasets. In Recent Advances in Natural Language Processing.
29. Trubetzkoy, N. (1928). Proposition 16. In Actes du premier congres international des linguistes à La Haye.
30. Utiyama, M. and Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In HLT-NAACL, pages 484–491.
31. Vilar, D., Peter, J.-T., and Ney, H. (2007). Can we translate letters? In Proceedings of the Second Workshop on Statistical Machine Translation.
32. Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM.

References

33. Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. Machine Translation, 21(3):165–181.
34. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., and Norouzi, M. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. ArXiv e-prints: abs/1609.08144.
35. Yang, Z., Salakhutdinov, R., and Cohen, W. (2016). Multi-task cross-lingual sequence tagging from scratch. arXiv preprint arXiv:1603.06270.
36. Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. EMNLP.

Extra Slides

Building a subword level translation system

Pre-processing:
Segment the corpus

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

Use higher order language models *(Vilar et al., 2007)*

Tune at the word-level *(Tiedemann, 2012)*

Decode using cube-pruning & smaller beam size for improved performance
(Kunchukuttan & Bhattacharyya, VarDial 2016)

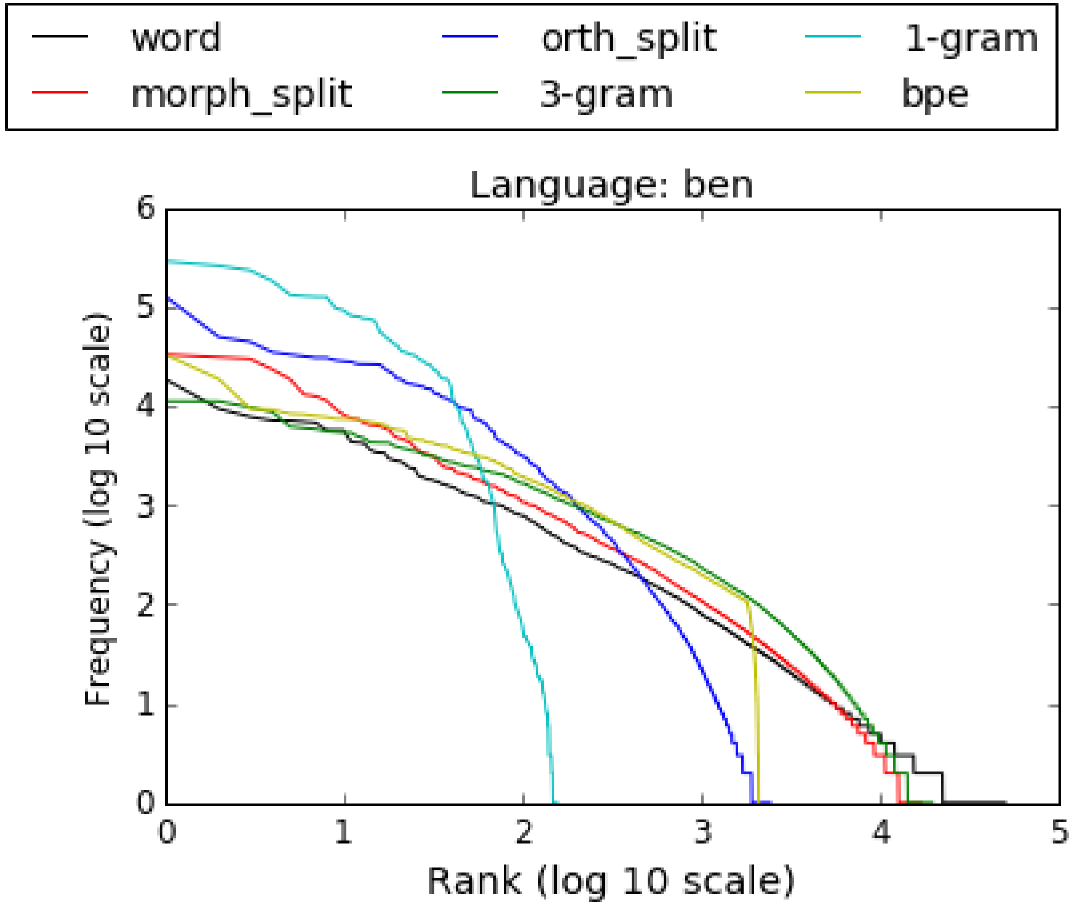
De-segment translation output

रा जू _ , _ घ र _ के _बा ह र _ म त _ जा ओ _ .

↓

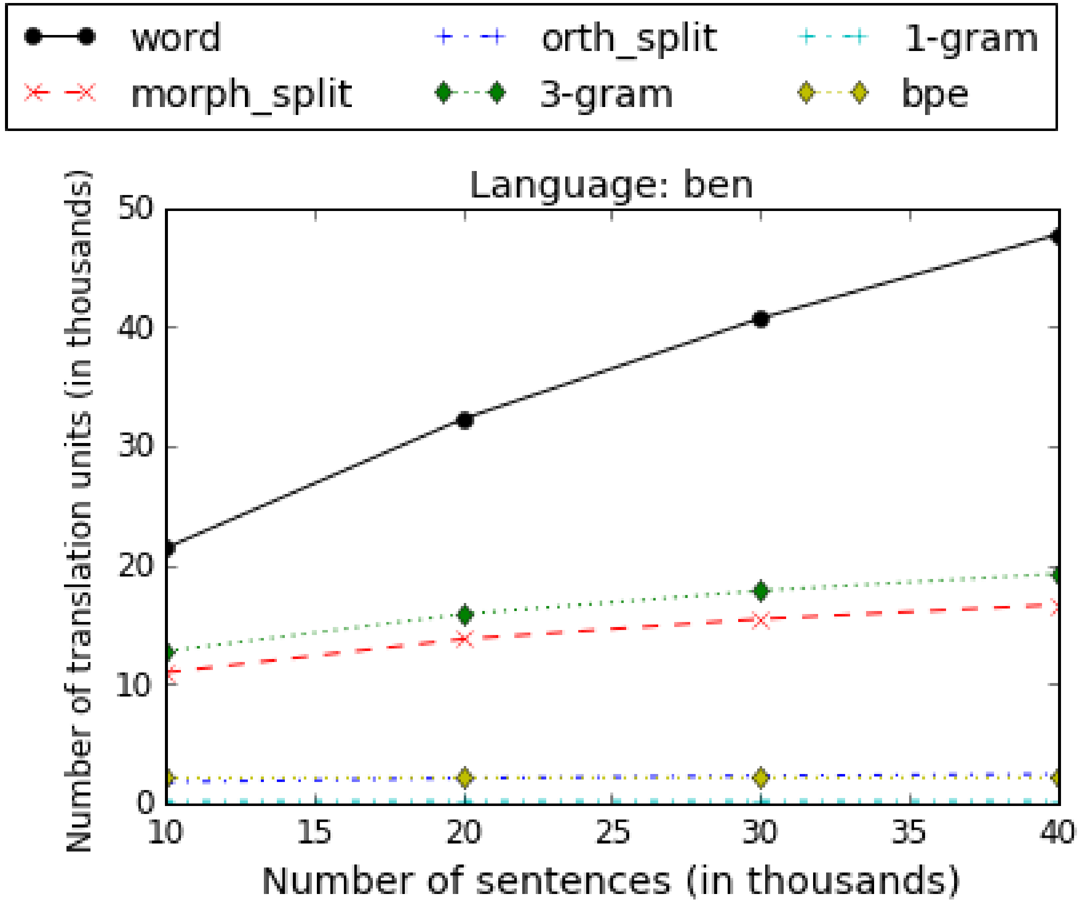
राजू , घर के बाहर मत जाओ .

Zipf's Law



Character 1-gram and OS don't follow a "strong" Zipf's Law

Herdan-Heap Law



Character 1-gram, OS and BPE don't follow a Herdan-Heap Law

Manning et al (2008)

Note: BPE vocab size is fixed

Address Data Sparsity

- Reduction in vocabulary size
- Explain improvement compared with word and morph units

Ability to learn Diverse Lexical Mappings

	hin	**mar**	**mal**
OS	tI, stha	mA, nA	kka, nI
Suffix	ke, me.m	ChyA, madhIla	unnu, .e∼Nkill.m
Word	paryaTaka, athavA	prAchIna, aneka	bhakShaN.m, yAtra

- *Using BPE, different types of translation units can be learnt*
- *The vocabulary size can be chosen as per the corpus size*
- *Non-linguistic mappings as well*

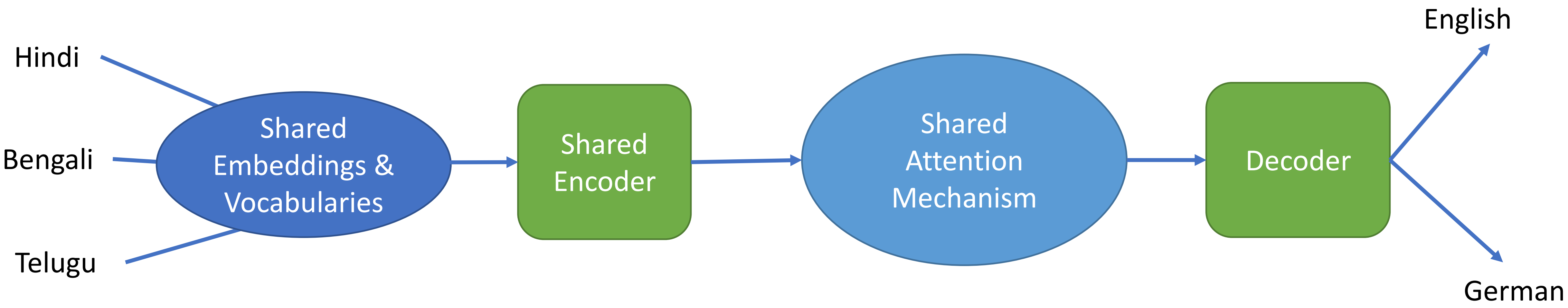
Additional Observations for Subword Translation

- Just a small vocabulary needed for translation
- Improving decoding speed: use a small beam size
- Particularly beneficial for more synthetic languages
- Robust to domain changes & works with small parallel corpora

Additional Findings for Pivot+Subword

- These results also hold for:
  - Transfer-based Pivot SMT
  - BPE translation units
- We see improvements in cross-domain translation also using subword units
- Pivot language closer to the target seems to be better
  (suggested by *Paul et al (2013)*)

Shared Decoder



Shared Embeddings for the target languages

Single Output layer

Target language is indicated by a special lang-id token in the input sequence