Boosting Phrase-based SMT with Unsupervised Morph-Analysis and Transliteration Mining

Anoop Kunchukuttan, Ratish Puduppully, Rajen Chatterjee, Abhijit Mishra,
Pushpak Bhattacharyya
{anoopk,ratishp,abhijitmishra,pb}@cse.iitb.ac.in,
rajen.k.chatterjee@gmail.com

Centre for Indian Language Technology
IIT Bombay

http://www.cfilt.iitb.ac.in/indic-translator

20 Dec 2014, ICON 2014

Motivation

- Scalability across language pairs
  - Minimize manual development of rules and resources
  - Explore unsupervised methods to exploit language and inter-language regularities
- Leverage shared characteristics of Indian languages
  - Common *abiguda* scripts derived from the *Brahmi* scripts
  - Shared vocabulary/cognates
  - Sentence structure
  - Morphological properties (at least within Indo-Aryan and Dravidian language families)
- Handle common divergences in a systematic way
  - Portable solutions which can be re-used across languages
  - *e.g.* Word order difference between English and Indian languages

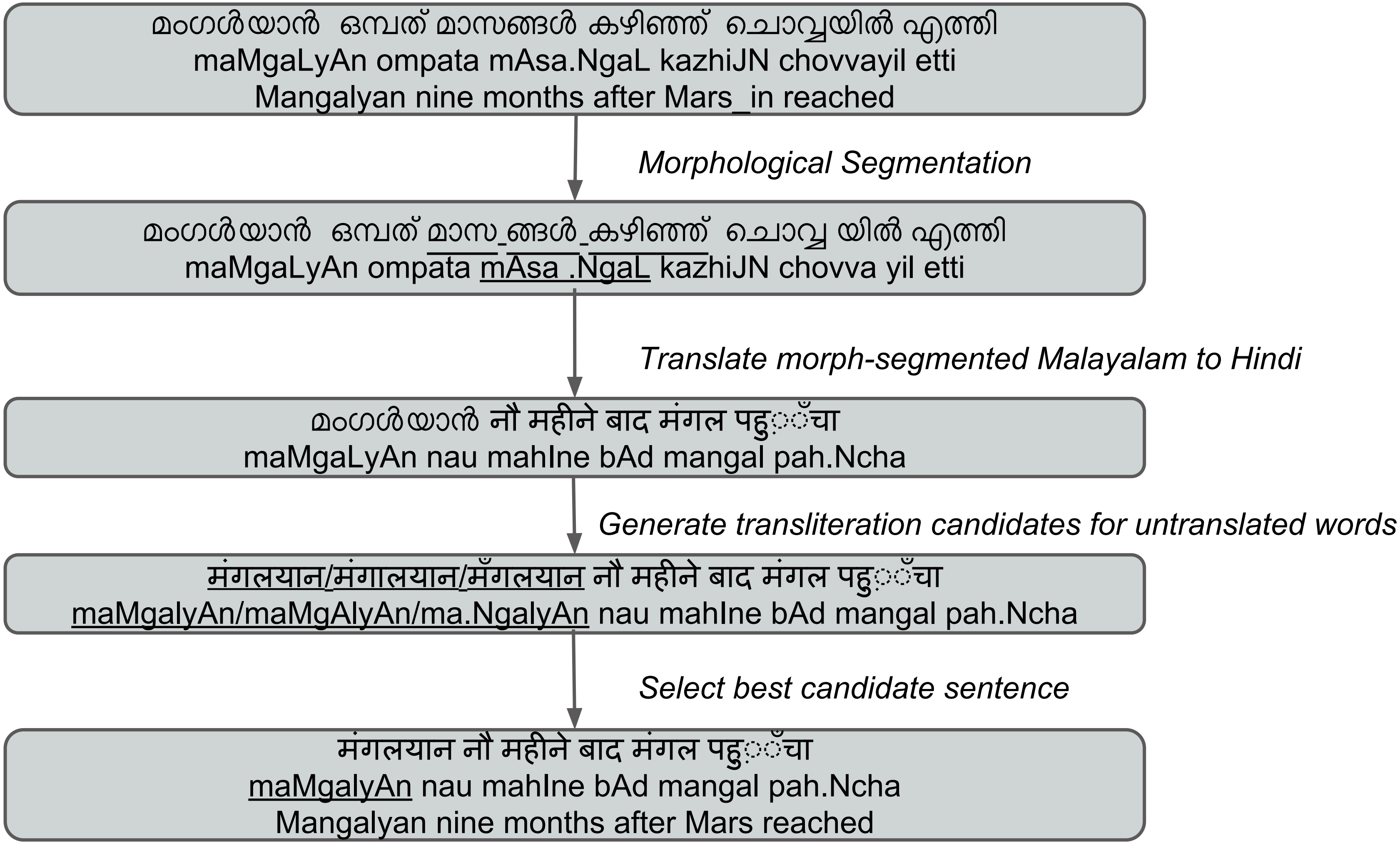
Address Key Limitations of Phrase-based SMT

- Morphological richness of Indian languages
  - Causes data sparsity, especially for agglutinative Dravidian languages
  - *अंगा + अंग + ◌ातून*   *(from every part of the body)*

    (aMgA + aMgA + tUn)
  - *जिल्हाध्यक्ष + पद + ◌ापर्यंत + च्या (till the post of District President)*

    (jilhAdhyakSh + pada + AparyaMt + chya)
- Named Entities, *Tatsam* words
  - Training corpus is small
  - Indian language share vocabulary: *tatsam* words, cognates, dialect continuum
  - Transliteration as Translation
  - *e.g. পারদর্শী (bn)   पारदर्शी (hi)*   (pArdarshI) *(transparency/foresight)*
- Structural divergence between English and Indian languages
  - Phrase based SMT lacks a good long-distance reordering model
  - SOV <-> SVO divergence
  - Prepositions become post-positions

Workflow
Indian Language to Hindi Translation



മംഗൾയാൻ ഒമ്പത് മാസങ്ങൾ കഴിഞ്ഞ് ചൊവ്വയിൽ എത്തി
maMgaLyAn ompata mAsa.NgaL kazhiJN chovvayil etti
Mangalyan nine months after Mars_in reached

Morphological Segmentation

മംഗൾയാൻ ഒമ്പത് മാസ_ങ്ങൾ_കഴിഞ്ഞ് ചൊവ്വ യിൽ എത്തി
maMgaLyAn ompata mAsa .NgaL kazhiJN chovva yil etti

Translate morph-segmented Malayalam to Hindi

മംഗൾയാൻ नौ महीने बाद मंगल पहु◌◌ँचा
maMgaLyAn nau mahIne bAd mangal pah.Ncha

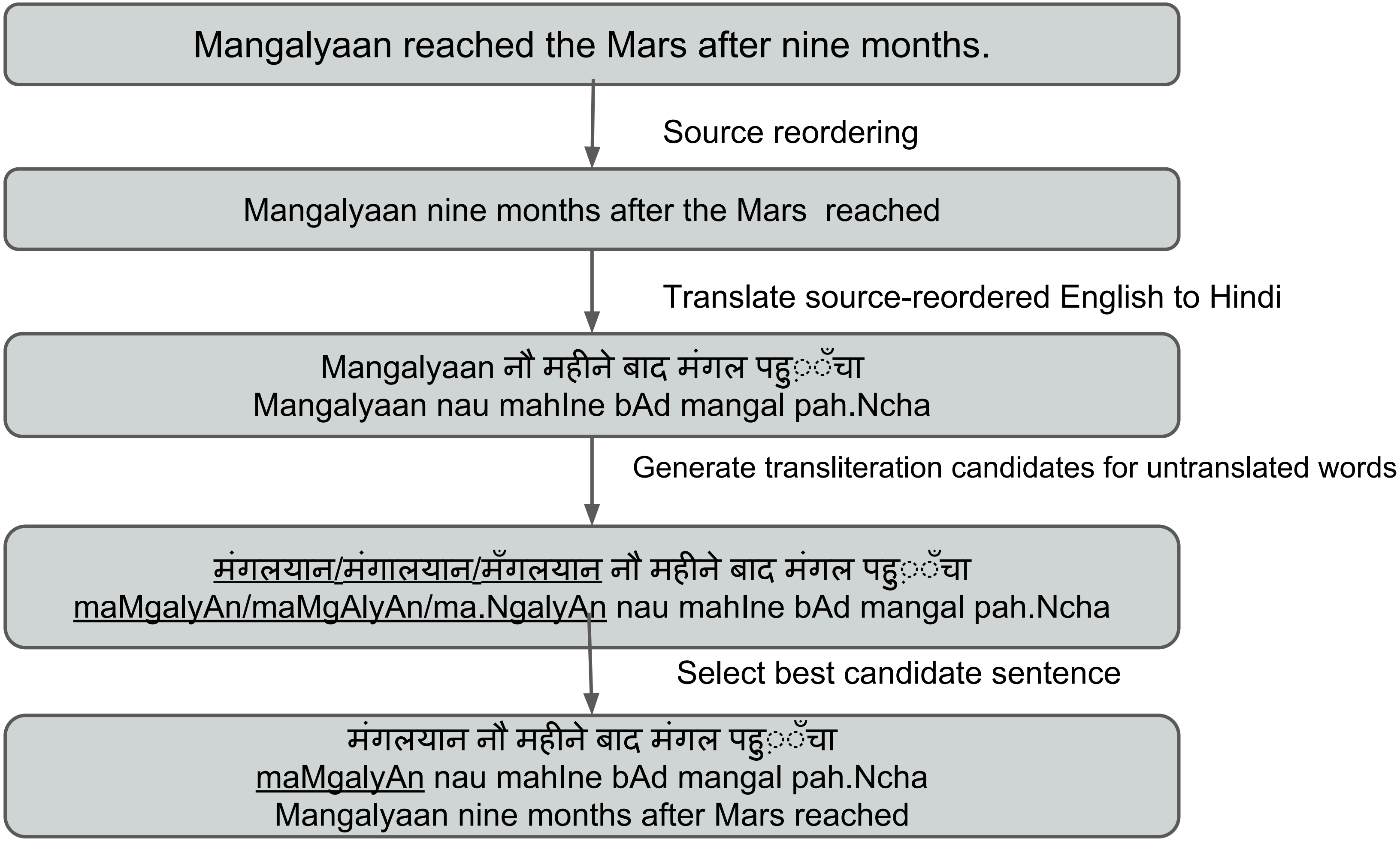
Generate transliteration candidates for untranslated words

मंगलयान/मंगालयान/मँगलयान नौ महीने बाद मंगल पहु◌◌ँचा
maMgalyAn/maMgAlyAn/ma.NgalyAn nau mahIne bAd mangal pah.Ncha

Select best candidate sentence

मंगलयान नौ महीने बाद मंगल पहु◌◌ँचा
maMgalyAn nau mahIne bAd mangal pah.Ncha
Mangalyan nine months after Mars reached

Workflow
English to Hindi Translation



Mangalyaan reached the Mars after nine months.

Source reordering

Mangalyaan nine months after the Mars reached

Translate source-reordered English to Hindi

Mangalyaan नौ महीने बाद मंगल पहु◌़◌ँचा
Mangalyaan nau mahIne bAd mangal pah.Ncha

Generate transliteration candidates for untranslated words

मंगलयान/मंगालयान/मँगलयान नौ महीने बाद मंगल पहु◌़◌ँचा
maMgalyAn/maMgAlyAn/ma.NgalyAn nau mahIne bAd mangal pah.Ncha

Select best candidate sentence

मंगलयान नौ महीने बाद मंगल पहु◌़◌ँचा
maMgalyAn nau mahIne bAd mangal pah.Ncha
Mangalyaan nine months after Mars reached

Unsupervised Morphological Segmentation

- Learn a segmentation model in an unsupervised setting given a list of words using the *Morfessor* method *[4]*
- Finds the lexicon (set of morphemes) such that the following objectives are met:
  - The likelihood of the tokens is maximized
  - The size of lexicon is minimized
  - Shorter morphemes are preferred
- *Frequency dampening*: did not use word frequency since it causes:
  - conservative segmentatation
  - reduction in boundary recall and F-1
- Given a new word, its segmentation can be computed using a generalization of the *Viterbi* algorithm

Examples: Morph-Segmentation (1)

Correct Segmentation

शरीर ाची	shariir aachii
फळ ांच्या	faL AMchyA
पदार्थ ांमध्ये	padarth AMmadhye

Missed Segmentation

सभामंडप ाचे	sabhAmaMdap Ache
महामस्तकाभिषेक	mahAmastakAbhiShek
सुरुवातीला	suruvAtiilA

Examples: Morph-Segmentation (2)

Aggressive Segmentation

पॅरा सिट ा मल ची	p.crA siT A mal chI
प्ले नेट ोरियम	ple neT oriyam
डिफ ि शंस ी	Dif i shaMs I
पर ं तू	par M tU
रोग ी	rog

Generally observed for named entities

Unsupervised Transliteration Mining

Learn a transliteration system using transliteration pairs mined from a parallel corpus [5]

Kailash Satyarthi won the Nobel Peace Prize for 2014

कैलाश सत्यार्थी ने २०१४ का नोबेल शांति पुरस्कार जीता

Kailash	कैलाश
Satyarthi	सत्यार्थी
won	जीता
Nobel	नोबेल
Peace	शांति
Prize	पुरस्कार
for	का
2014	२०१४

Align the words

Unsupervised Transliteration Mining

Learn a transliteration system using transliteration pairs mined from a parallel corpus [5]

Non-transliteration process

$$p_{ntr}(e, f) = \prod_{i=1}^{|e|} p_E(e_i) \prod_{i=1}^{|f|} p_F(f_i)$$

A generative model for the word pairs

Kailash	कैलाश
Satyarthi	सत्यार्थी
won	जीता
Nobel	नोबेल
Peace	शांति
Prize	पुरस्कार
for	का
2014	२०१४

Unsupervised Transliteration Mining

Learn a transliteration system using transliteration pairs mined from a parallel corpus [5]

Transliteration Process

$$p_{tr}(e,f) = \sum_{a \in Align(e,f)} \prod_{j=1}^{|a|} p(q_j)$$

Kailash	कैलाश
Satyarthi	सत्यार्थी
won	जीता
Nobel	नोबेल
Peace	शांति
Prize	पुरस्कार
for	का
2014	२०१४

Unsupervised Transliteration Mining

The transliteration mining model is an interpolation of both models

$$p(e, f) = (1 - \lambda)p_{tr}(e, f) + \lambda p_{ntr}(e, f)$$

λ is the prior probability of non-transliteration.

- Model parameters: λ and $p(q_j)$
- Estimated by maximum likelihood using the EM algorithm
- Word pairs for transliteration probability is greater are considered transliteration pairs

$$1 - \frac{\lambda p_2(e_i, f_i)}{p(e_i, f_i)} > 0.5$$

- F-scores of > 90% have been reported on en-hi transliteration mining task

Examples of Mined Pairs

Perfect Transliterations

syphilis	सिफिलिस
tandoori	तंदूरी
telephone	टेलिफोन
అంధేరీ	अंधेरी
అకబ్ర్	अकबर

Spelling variations

telephone	टेलीफोन/टेलिफोन
Belgaum	बेलगाँव/बेलगाम
फेब्रुवारी	फरवरी

Examples of Mined Pairs (2)

Sound Shifts

केरळ	(keraL)	केरल	(keral)
ஏரோபிக்ஸ்	(eropiks)	एरोबिक्स	(erobiks)
ஏரோபிக்ஸ்	(ka~Nkotari)	गंगोत्री	(gaMgotrI)

Cognates

अंधळेपणा	(aMdhLepaNa)	अंधेपन	(aMdhepan)
कसे	(kase)	कैसे	(kaise)
गाढव	(gaDhav)	गधा	(gadha)
பக்தர்கள்	(paktarkaL)	भक्तगण	(bhaktagaN)

Examples of Mined Pairs (3)

Inflectional variants

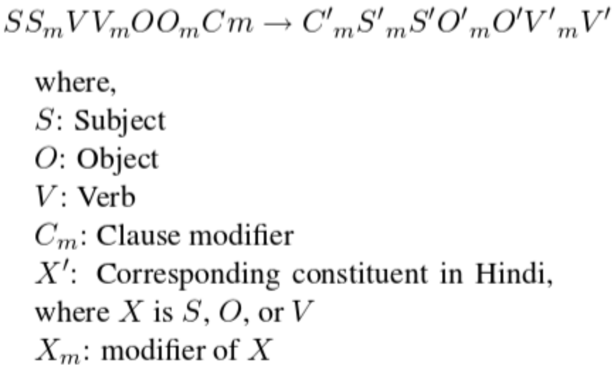
ஆகாயத்தின் (आकायत्तिऩ्)	आकाश
ஆகாயத்தில் (आकायत्तिल्)	आकाश
ஆகாயத்தை (आकायत्तै)	आकाश
खेळायला	खेलने
खेळायाला	खेलने
खेळाला	खेली
खेळावे	खेलें

Mistakes

Synonyms:	silent	शांत (shaMt)
Partial matches:	गर्भधारणा (garbhadharaNA)	गर्भावस्था (garbhAvasthA)

Source Reordering

- Significant structural divergence between English and Hindi
- Source Reordering improves PB-SMT:
  - Longer phrases can be learnt
  - Decoder cannot evaluate long distance reorderings by search in a small window
- Rule based reordering by applying transformation on English parse tree
  - works well for all target Indian languages *[1]*
- Basic Transformation



$$SS_mVV_mOO_mCm \rightarrow C'_mS'_mS'O'_mO'V'_mV'$$

where,
S: Subject
O: Object
V: Verb
C_m: Clause modifier
X': Corresponding constituent in Hindi, where X is S, O, or V
X_m: modifier of X

Experimental Details

Phrase based systems

- *Moses* baseline
- *grow-diag-final-end* heuristic
- Lexicalized Reordering
- MERT tuning

Morph Analyzers

- *Morfessor 2.0*
- Trained on Leipzig + ILCI monolingual corpora

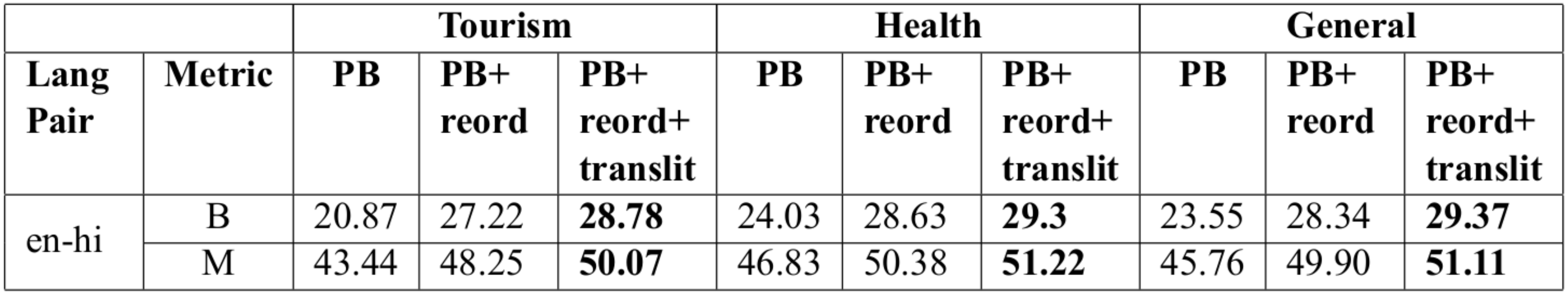
Language Model

- 5-gram model with Kneser-Ney smoothing
- 1.5 million sentences from ILCI+subset of WMT corpus

Evaluation Metrics

- BLEU (B)
- METEOR for Indian languages (M)
  - Stemming using *IndoWordNet* assisted stemmer *[7]*
  - Synonyms from *IndoWordNet [6]*

Results on devtest: en-hi

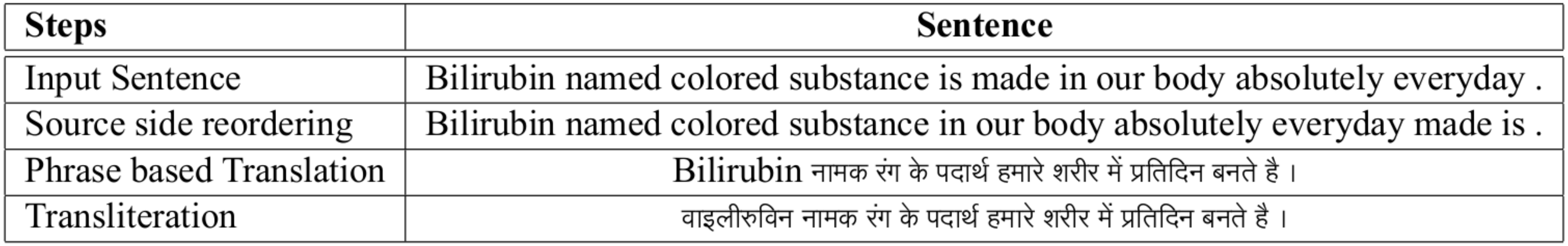


		Tourism			Health			General		
Lang Pair	**Metric**	**PB**	**PB+ reord**	**PB+ reord+ translit**	**PB**	**PB+ reord**	**PB+ reord+ translit**	**PB**	**PB+ reord**	**PB+ reord+ translit**
en-hi	B	20.87	27.22	**28.78**	24.03	28.63	**29.3**	23.55	28.34	**29.37**
	M	43.44	48.25	**50.07**	46.83	50.38	**51.22**	45.76	49.90	**51.11**

- Source reordering contributes to a major improvement
  - BLEU scores improve upto 30%
  - METEOR scrores improve upto 11%
- Transliteration post-editing contributes to improvement
  - BLUE and METEOR improvements of 5% and 3% respectively
  - Recall improvement of upto 2.6%
- Source Reordering helps phrase based SMT for structurally divergent languages
- The rules are portable to all target Indian languages

Examples

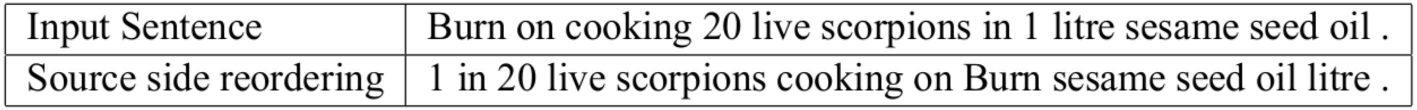
Source reordering helps improves word order



Steps	**Sentence**
Input Sentence	Bilirubin named colored substance is made in our body absolutely everyday .
Source side reordering	Bilirubin named colored substance in our body absolutely everyday made is .
Phrase based Translation	Bilirubin नामक रंग के पदार्थ हमारे शरीर में प्रतिदिन बनते है ।
Transliteration	वाइलीरुविन नामक रंग के पदार्थ हमारे शरीर में प्रतिदिन बनते है ।

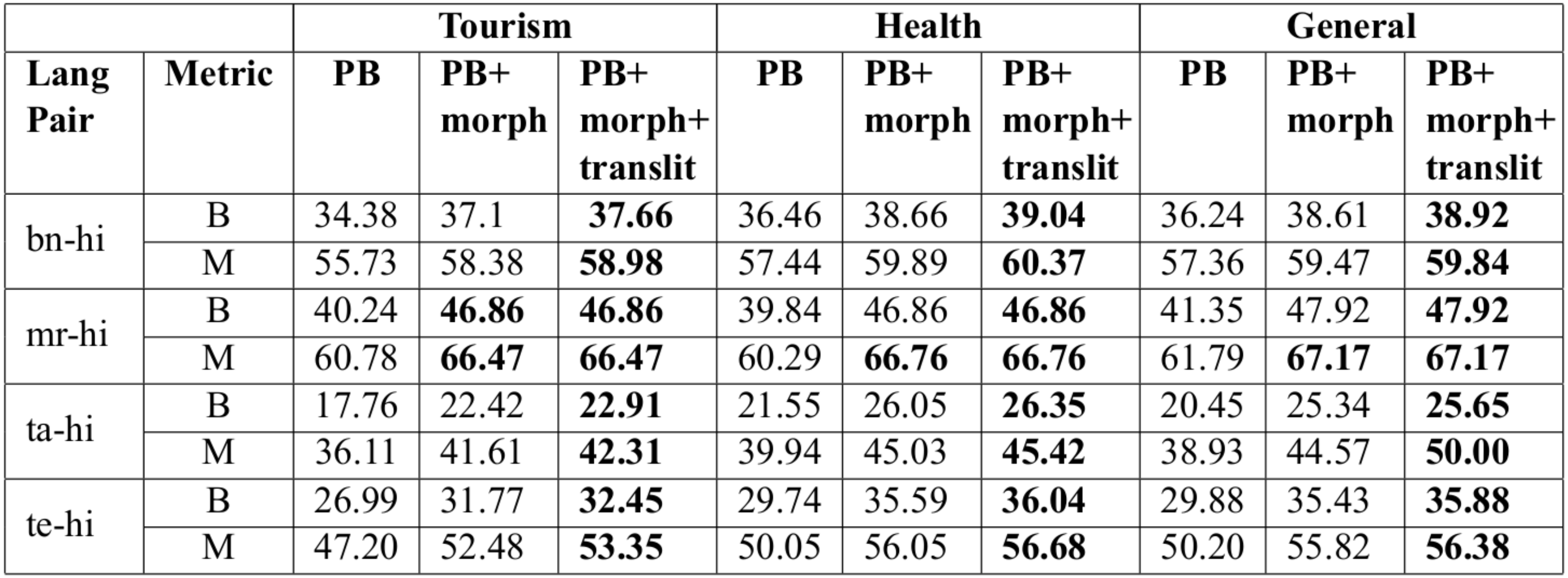
Reordering rules can generate wrong word order

In this example, no rules for imperative sentences cause reordering error



Input Sentence	Burn on cooking 20 live scorpions in 1 litre sesame seed oil .
Source side reordering	1 in 20 live scorpions cooking on Burn sesame seed oil litre .

Results on devtest: IL-hi



		Tourism			Health			General		
Lang Pair	**Metric**	**PB**	**PB+ morph**	**PB+ morph+ translit**	**PB**	**PB+ morph**	**PB+ morph+ translit**	**PB**	**PB+ morph**	**PB+ morph+ translit**
bn-hi	B	34.38	37.1	**37.66**	36.46	38.66	**39.04**	36.24	38.61	**38.92**
	M	55.73	58.38	**58.98**	57.44	59.89	**60.37**	57.36	59.47	**59.84**
mr-hi	B	40.24	**46.86**	**46.86**	39.84	46.86	**46.86**	41.35	47.92	**47.92**
	M	60.78	**66.47**	**66.47**	60.29	**66.76**	**66.76**	61.79	**67.17**	**67.17**
ta-hi	B	17.76	22.42	**22.91**	21.55	26.05	**26.35**	20.45	25.34	**25.65**
	M	36.11	41.61	**42.31**	39.94	45.03	**45.42**	38.93	44.57	**50.00**
te-hi	B	26.99	31.77	**32.45**	29.74	35.59	**36.04**	29.88	35.43	**35.88**
	M	47.20	52.48	**53.35**	50.05	56.05	**56.68**	50.20	55.82	**56.38**

- Source word segmentation significantly improves performance
  - For morphologically rich source like *ta*, improvements of upto 24% in BLEU
  - For comparatively poor source like *bn*, improvements of upto 6% in BLEU
  - Similar trends for METEOR score
- Transliteration post-editing marginally improves translation
  - BLEU scores improve by upto 1.2%
  - Recall improves by upto 1.4%

Examples

Morphological segmentation helps overcome data sparsity

Source	गौतम बुद्ध अभयारण्य कोडरमामध्ये वसलेले आहे जेथे चित्ता आणि वाघ आहेत .
Segmented	गौतम बुद्ध अभयारण्य कोडरमा_मध्ये वसलेल ◌े आहे जेथे चित्ता आणि वाघ आहेत .
Xlation: simple PBSMT	गौतम बुद्ध अभ्यारण्य कोडरमामध्ये स्थित है जहाँ चीता और बाघ हैं ।
Xlation: PBSMT + segmentation	गौतम बुद्ध अभ्यारण्य कोडरमा_में स्थित है जहाँ चीता और बाघ हैं ।

Aggressive segmentation results in deterioration of translation quality

Source	इक्ष्वाकु पुत्र राजा विशाल याला वैशाली राज्याचा संस्थापक मानले जाते .
Segmented	इ_क्ष_◌्वा_कु_पुत्र राजा विशाल याला वैशाली राज्य ◌ाचा संस्थापक मानले जाते .
Xlation: simple PBSMT	इक्ष्वाकु_पुत्र राजा विशाल इसे वैशाली राज्य का संस्थापक माना जाता है ।
Xlation: PBSMT + segmentation	सन_सफेद_◌्वा_विकृत पुत्र राजा विशाल इसे वैशाली राज्य का संस्थापक माना जाता है ।

Examples of transliteration post-editing

Named entity

अल्सर और खुले घाव न होना या मुँह के अंदर सफेद होना , <u>கோப்லேகியா</u> लगाई हो

alsar aur khule ghAv na honA yA mu.Nh ke andar safed honA, <u>koplekiyA</u> lagAI ho

अल्सर और खुले घाव न होना या मुँह के अंदर सफेद होना , <u>कोप्लेगिया</u> लगाई हो

alsar aur khule ghAv na honA yA mu.Nh ke andar safed honA, <u>koplegiyA</u> lagAI ho

Cognates

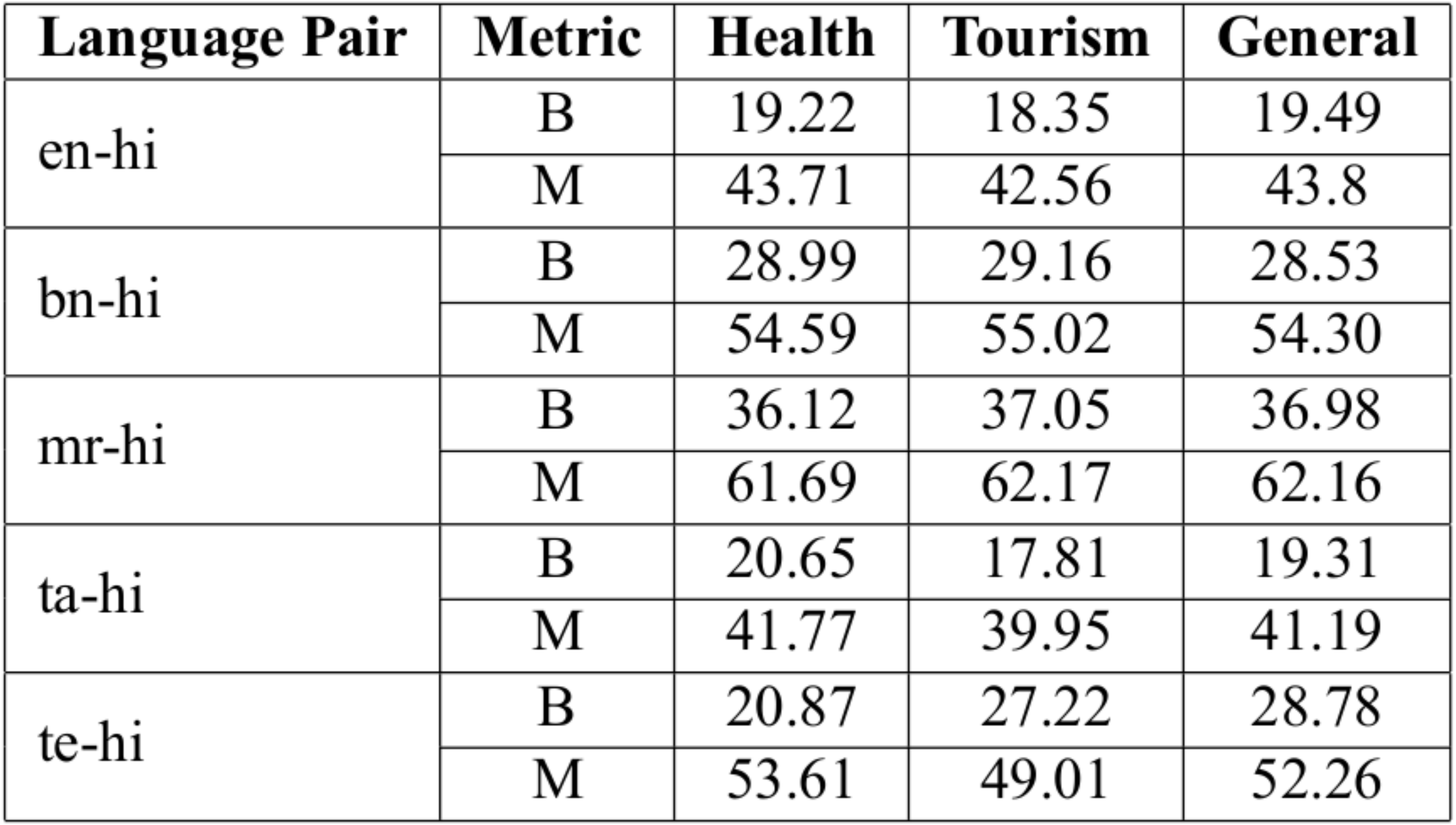
आजकल ऑपरेशन द्वारा **<u>पाऱदर्शी</u>** उसे मोड़ लाया गया

aajkal Apareshan dvArA <u>pAradarshI</u> use moD lAyA gayA

आजकल ऑपरेशन द्वारा <u>पारदर्शी</u> उसे मोड़ लाया गया

aajkal Apareshan dvArA <u>pAradarshI</u> use moD lAyA gayA

Results on official test set



Language Pair	Metric	Health	Tourism	General
en-hi	B	19.22	18.35	19.49
	M	43.71	42.56	43.8
bn-hi	B	28.99	29.16	28.53
	M	54.59	55.02	54.30
mr-hi	B	36.12	37.05	36.98
	M	61.69	62.17	62.16
ta-hi	B	20.65	17.81	19.31
	M	41.77	39.95	41.19
te-hi	B	20.87	27.22	28.78
	M	53.61	49.01	52.26

Conclusions

- Morphological segmentation of source language substantially improves translation quality
- Source side reordering helps in bridging the structural divergence between English and Indian languages
- ‘*Transliteration as translation*’ aids IL-IL SMT
- It is possible to scale to multiple language pairs by:
  - using unsupervised methods
  - leveraging shared characteristics of Indian languages

Future Work

- Combine hierarchical SMT with source reordering methods
- Multiple inputs to the decoder which can choose the best input:
  - segmented and non-segmented sentences
  - original and source-reordered sentences
- Handling morphologically complex target languages

Resources

- Word Segmentation Models
  - Python API
  - 10 languages
- Source Reordering Rules
  - Implements rules in *[2]*
- Transliteration Models
  - Moses based transliteration system
- METEOR for Hindi and Marathi (soon)

and more on:

http://www.cfilt.iitb.ac.in/static/download.html

References

1. Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, Pushpak Bhattacharyya. *Śata-Anuvādak: Tackling Multiway Translation of Indian Languages.* Language Resources and Evaluation Conference . 2014.
2. R. Ananthakrishnan, Jayprasad Hegde, Ritesh Shah, Pushpak Bhattacharyya and M. Sasikumar, *Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation*, International Joint Conference on NLP. 2008.
3. Raj Patel, Rohit Gupta, Prakash Pimpale, M. Sasikumar. *Reordering rules for English-Hindi SMT*. In Proceedings of the Second Workshop on Hybrid Approaches to Translation. 2013.
4. Virpioja, Sami, et al. *Morfessor 2.0: Python implementation and extensions for Morfessor Baseline*. Technical Report. 2013.
5. Nadir Durrani, Hieu Hoang, Philipp Koehn, and Hassan Sajjad. Integrating an unsupervised transliteration model into statistical machine translation. EACL 2014.
6. Pushpak Bhattacharyya. Indowordnet. In InProc. of LREC-10. 2010.
7. Pushpak Bhattacharyya, Ankit Bahuguna, Lavita Talukdar, and Bornali Phukan. Facilitating multi-lingual sense annotation: Human mediated lemmatizer. In Global WordNet Conference. 2014.

Thank You!