English to Indian Languages Machine Transliteration System at NEWS 2010

Amitava Das[1], Tanik Saikh[2], Tapabrata Mondal[3], Asif Ekbal[4], Sivaji Bandyopadhyay[5]

Department of Computer Science and Engineering[1,2,3,5]
Jadavpur University

Department of Computational Linguistics[4]
University of Heidelberg

Transliteration Unit (TU)

Bengali, Hindi and Kannada and Tamil and English words divided into Transliteration Units (TUs):
English word TU: C*V*
where C represents a consonant and V represents a vowel.

Indian Languages Words TU: C+M?
where C represents a vowel or a consonant or a conjunct and M represents the vowel modifier or matra

Contextual information in the form of collocated TUs considered

Firdausi → Fi | r | dau| si
फ़ि | र | दौ | सी

Banphool → Ba | n | phoo | l
ब | न | फू | ल

Babooji → Ba | boo | ji
बा | बू | जी

English to Indian Languages Machine Transliteration

- Plausibility of transliteration from each English TU to various Indian Languages (ILs) candidate TUs calculated
  - ILs candidate TU chosen with **maximum transliteration probability**
  - Equivalent to choosing the most appropriate sense of a word in the source language to identify its representation in the target language

- System learns mappings automatically from the **bilingual training corpus**
  - **Learning guided by linguistic features**
    - **Output of mapping is a decision list classifier**
- Machine transliteration obtained by direct orthographic mapping or phonetic mapping
  - Equivalent ILs TU for each English TU in the input identified and placed in ranked order.

Transliteration Models

Orthographic Transliteration Models

- Trigram Model
- Joint Source-Channel Model (JSC)
- Modified Joint Source-Channel Model (MJSC)
- Improved Modified Joint Source-Channel Model (IMJSC)

International Phonetic Alphabet (IPA) Model

Trigram Model

Previous and next source TUs as the context

$$P(S \mid T) = \prod_{k=1}^{K} P(< s,t >_k \mid s_{k-1,} s_{k+1})$$

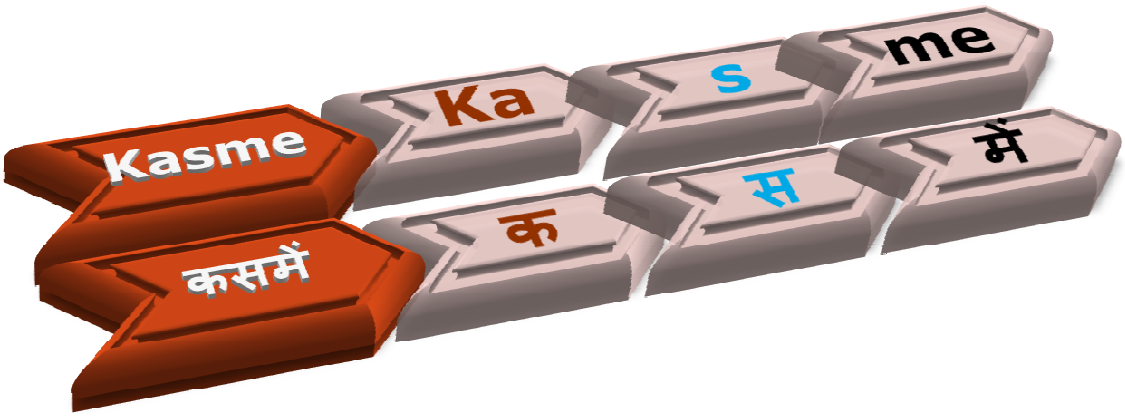
$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S \mid T)\}$$

Joint Source-Channel Model (JSC)

Previous TUs with reference to the current TUs in both the source (s) and the target sides (t) are considered as the context.

$$P(S|T) = \prod_{k=1}^{K} P(<s,t>_k | <s,t>_{k-1})$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S|T)\}$$

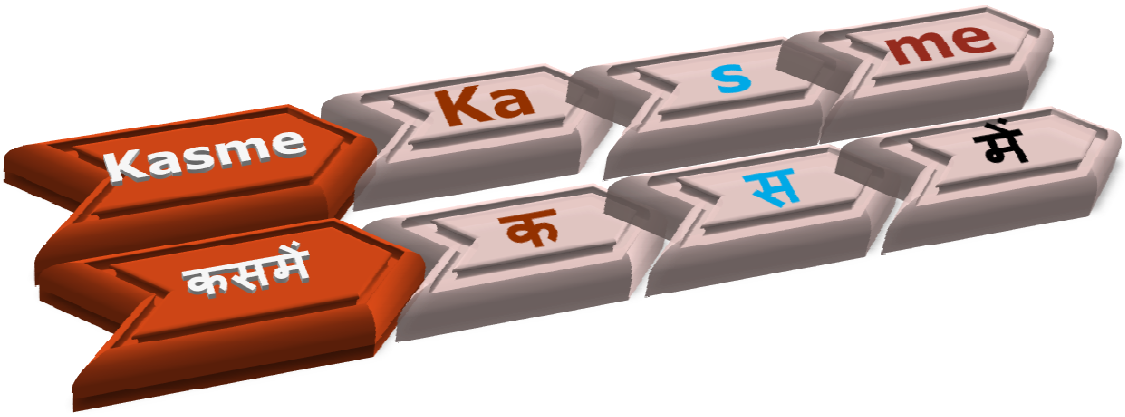


Modified Joint Source-Channel Model (MJSC)

the previous and the next TUs in the source and the previous target TU are considered as the context.

$$P(S \mid T) = \prod_{k=1}^{K} P(< s,t >_k \mid < s,t >_{k-1},\ s_{k+1})$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S \mid T)\}$$



Improved Modified Joint Source-Channel Model (IMJSC)

The previous two and the next TUs in the source and the previous target TU are considered as the context.

$$P(S \mid T) = \prod_{k=1}^{K} P(< s,t >_k \mid s_{k+1} < s,t >_{k-1,}\ s_{k+1})$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S \mid T)\}$$

International Phonetic Alphabet (IPA) Model

Large number of dictionary words in NEWS 2010 data set

	Training	Development	Test
Bengali	7.77%	5.14%	6.46%
Hindi	27.82%	15.80%	3.7%
Kannada	27.60%	14.63%	4.4%
Tamil	27.87%	17.31%	3.0%

IPA Symbol Set

Phoneme	Example	Translation
AA	*odd*	AA-D
AE	*at*	AE–T
AH	*hut*	HH-AH-T
AO	*ought*	AO-T
AW	*cow*	K-AW
AY	*hide*	HH-AY-D
B	*be*	B-IY
CH	*cheese*	CH-IY-Z
D	*dee*	D-IY
DH	*thee*	DH-IY
EH	*Ed*	EH-D
ER	*hurt*	HH-ER-T
EY	*ate*	EY-T
F	*fee*	F-IY
G	*green*	G R-IY-N
HH	*he*	HH-IY
IH	*it*	IH-T
IY	*eat*	IY-T
JH	*gee*	JH-IY
K	*key*	K-IY
L	*lee*	L-IY
M	*me*	M-IY
N	*knee*	N-IY
NG	*ping*	P-IH-NG
OW	*oat*	OW-T
OY	*toy*	T-OY
P	*pee*	P-IY
R	*read*	R-IY-D
S	*sea*	S-IY
SH	*she*	SH-IY
T	*tea*	T-IY
TH	*theta*	TH-EY-T-AH
UH	*hood*	HH-UH-D
UW	*two*	T-UW
V	*vee*	V-IY
W	*we*	W-IY
Y	*yield*	Y-IY-L-D
Z	*zee*	Z-IY
ZH	*seizure*	S-IY-ZH-ER

CRF Based IPA Model

- A pre-processing module checks whether a targeted source English word is a valid dictionary word or not.
- In the target side we use our TU segregation logic to get phoneme wise transliteration pattern.
- Modeled as a sequence labeling problem as transliteration pattern changes depending upon the contextual phonemes in source side and TUs in the target side.

Output Ranking

	Ranking Order				
Word Type	1	2	3	4	5
Dictionary	IPA	IMJSC	MJSC	JSC	TRI
Non-Dictionary	IMJSC	MJSC	JSC	TRI	-

Bengali Experimental Result

Parameters	Accuracy		
	BSR	BNSR1	BNSR2
Accuracy in top-1	0.232	0.369	0.430
Mean F-score	0.818	0.845	0.875
Mean Reciprocal Rank (MRR)	0.325	0.451	0.526
Mean Average Precision $(MAP)_{ref}$	0.232	0.369	0.430

- ✓ **BNSR2 has achieved the highest score among all the submitted Runs.**

Hindi
Experimental Result

Parameters	Accuracy		
	HSR	HNSR1	HNSR2
Accuracy in top-1	0.150	0.254	0.170
Mean F-score	0.714	0.752	0.739
Mean Reciprocal Rank (MRR)	0.308	0.369	0.314
Mean Average Precision $(MAP)_{ref}$	0.150	0.254	0.170

- ✓ **HNSR1 and HNSR2 runs are ranked as the 5th and 6th among all submitted Runs.**

Kannada Experimental Result

Parameters	Accuracy	
	KSR	KNSR1
Accuracy in top-1	0.056	0.055
Mean F-score	0.663	0.662
Mean Reciprocal Rank (MRR)	0.112	0.169
Mean Average Precision $(MAP)_{ref}$	0.056	0.055

Tamil Experimental Result

Parameters	Accuracy	
	TSR	TNSR1
Accuracy in top-1	0.013	0.082
Mean F-score	0.563	0.760
Mean Reciprocal Rank (MRR)	0.121	0.142
Mean Average Precision $(MAP)_{ref}$	0.013	0.082