NLP for Social Media

Lecture 4: Processing Indic Language Social Media Content

Monojit Choudhury

Microsoft Research Lab, *monojitc@microsoft.com*

Indic Language specific phenomena

Non-standard spellings	2mrw → tomorrow	Spelling Normalization
Non-standard grammar	even i want to → Even I want to do this.	Grammar correction
Language mixing	Kothakar\B Master\E chef\E contest\E ?	Language Detection
Transliteration	Kothakar → কোথাকার	Machine Transliteration
Emoticons, Tags, mentions, slangs	abae → ?, :P → ?, @Mallar → ?, ??? → ?	Special Treatment

What will we learn?

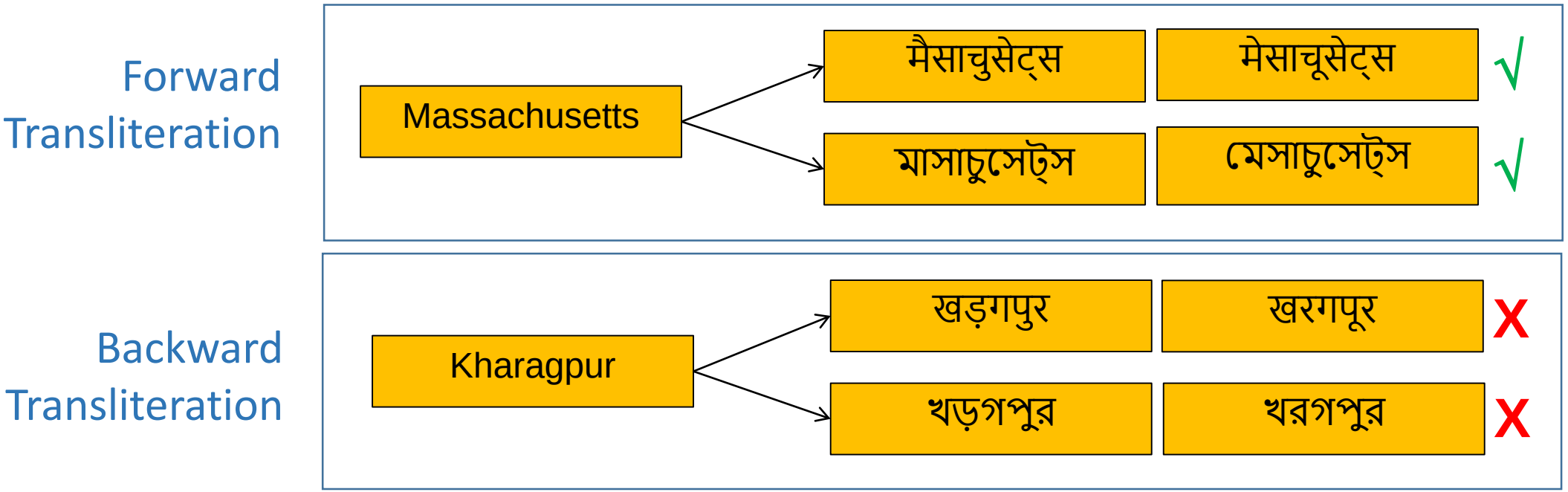
- Transliteration in SM: What, why and how much
- Noisy Channel models for transliteration
- Word embedding for transliteration

What will we learn?

- Transliteration in SM: What, why and how much
- Noisy Channel models for transliteration
- Word embedding for transliteration
- Language Detection

Transliteration

The process of “loosely” representing the sound (pronunciation) of the words of a language in a script of another language.



Forward Transliteration

- Named Entities
  - People: *Faxian* (also *Fa-hsien, Fa-hien*)
  - Places: Calcutta, Kolkata
  - Organization: आईआईटी
  - Movies/Song/Book Title: *Muddu manse, Yelaa satyam*

- Technical Terms: इंटरनेट, क्वांटम मैकेनिक्स

Regularly features in all languages.
Should be handled during Machine Translation

A Transliterated World Wide Web

This weekend's cooking experiment was to make a traditional kerala dish called "Kappa Puzhukku" (Tapioca Pudding). Grand success.

Like · Comment · Share

16 people like this.

View all 15 comments

Athula Balachandran Prashanth Mohan: Awesome 🙂 Koode kazhikkan nalla kudam puli itta meen curry undakkiyo? 😃

November 12 at 2:19am · Like

Sudha A Prakash Prashanth Mohan - wow...looks stunning and am sure must've been really tasty - need 'kanju' along with it!

November 12 at 6:47pm · Like

Ushan Ananthanarayanan This puzhukku seems to be without coconut ?!?!

November 12 at 9:47pm · Like

Prashanth Mohan Athula, Kooda kazhikkan oru ugran split pea soup undaki. Adhu rendum aazhappo thanna US standards ni sadhyai aazhilae.

A Transliterated World Wide Web

Song Lyrics

SS

Ami je jalsaghare- Lyrics and Song-

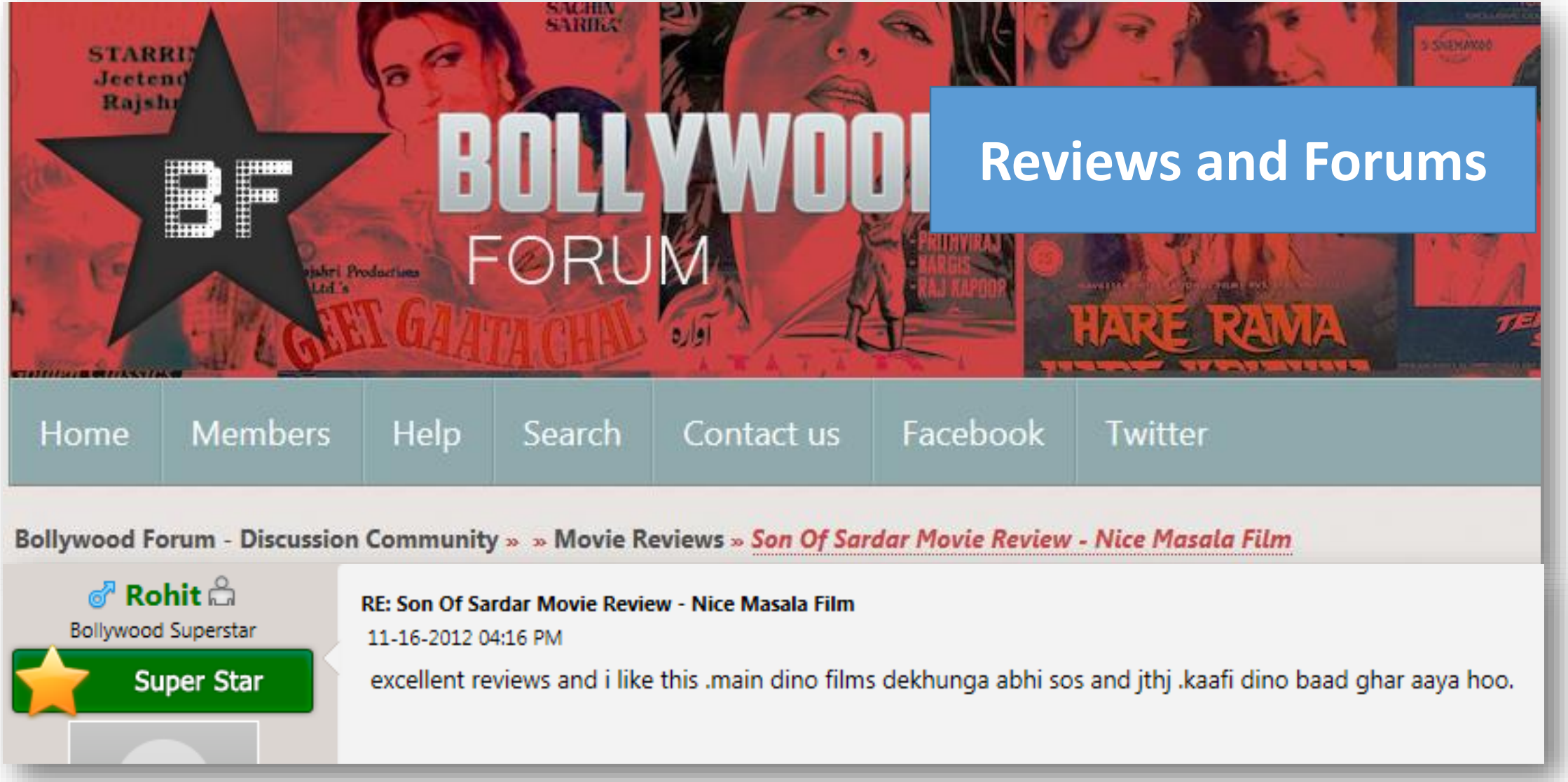
Posted on September 26, 2009 by Gaanwala

Sur : Anil Bagchi
Gayok : Manna Dey
Cinema : ANTONY FIRINGEE

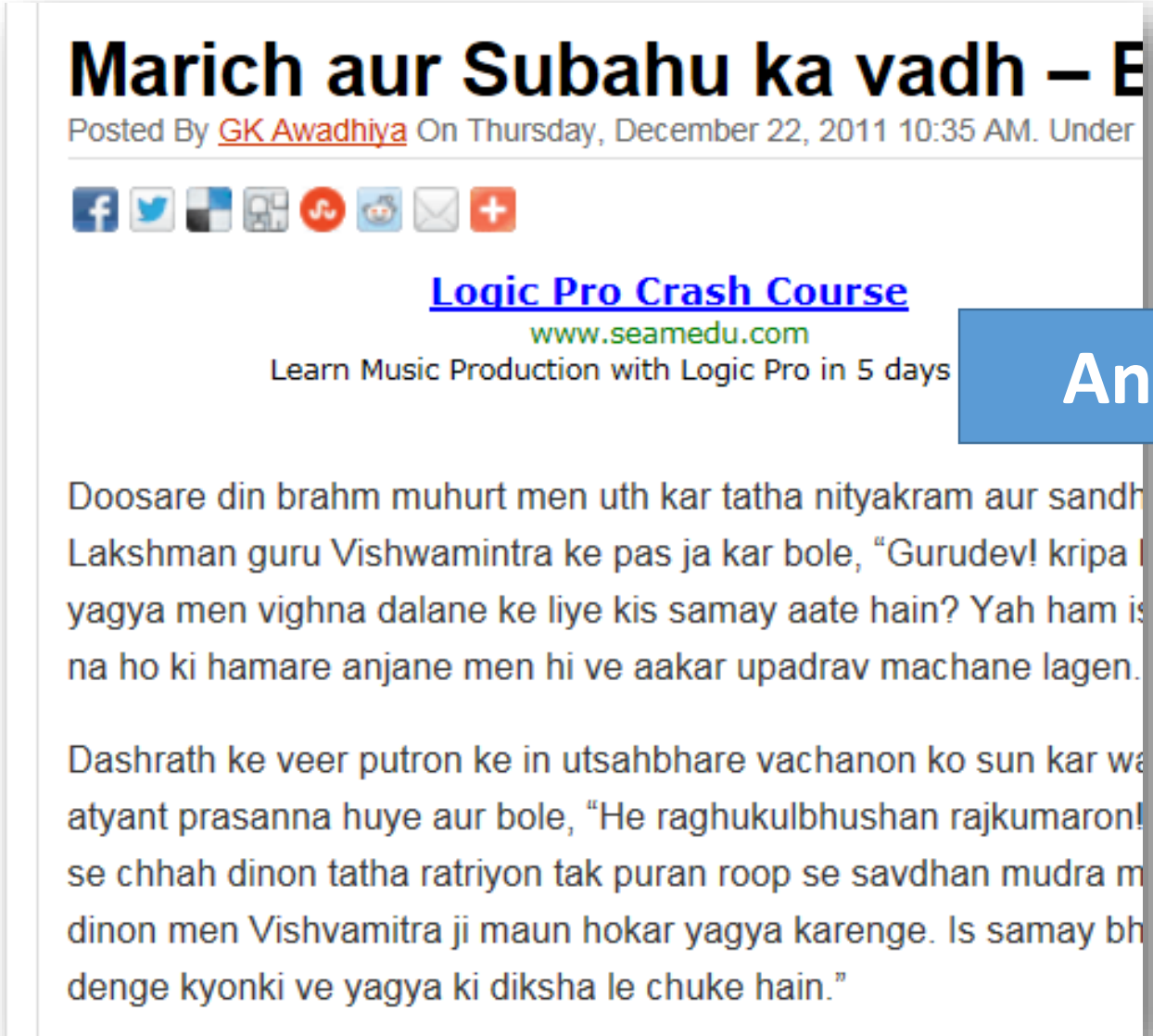
[ami je jalsaghare beloari jhaar.](2)
ami je jalsaghare.
[nishi furale keho chaye na aamaye jaani go aar.](2)

ami je jalsaghare.

A Transliterated World Wide Web



A Transliterated World Wide Web



Marich aur Subahu ka vadh – E

Posted By GK Awadhiya On Thursday, December 22, 2011 10:35 AM. Under

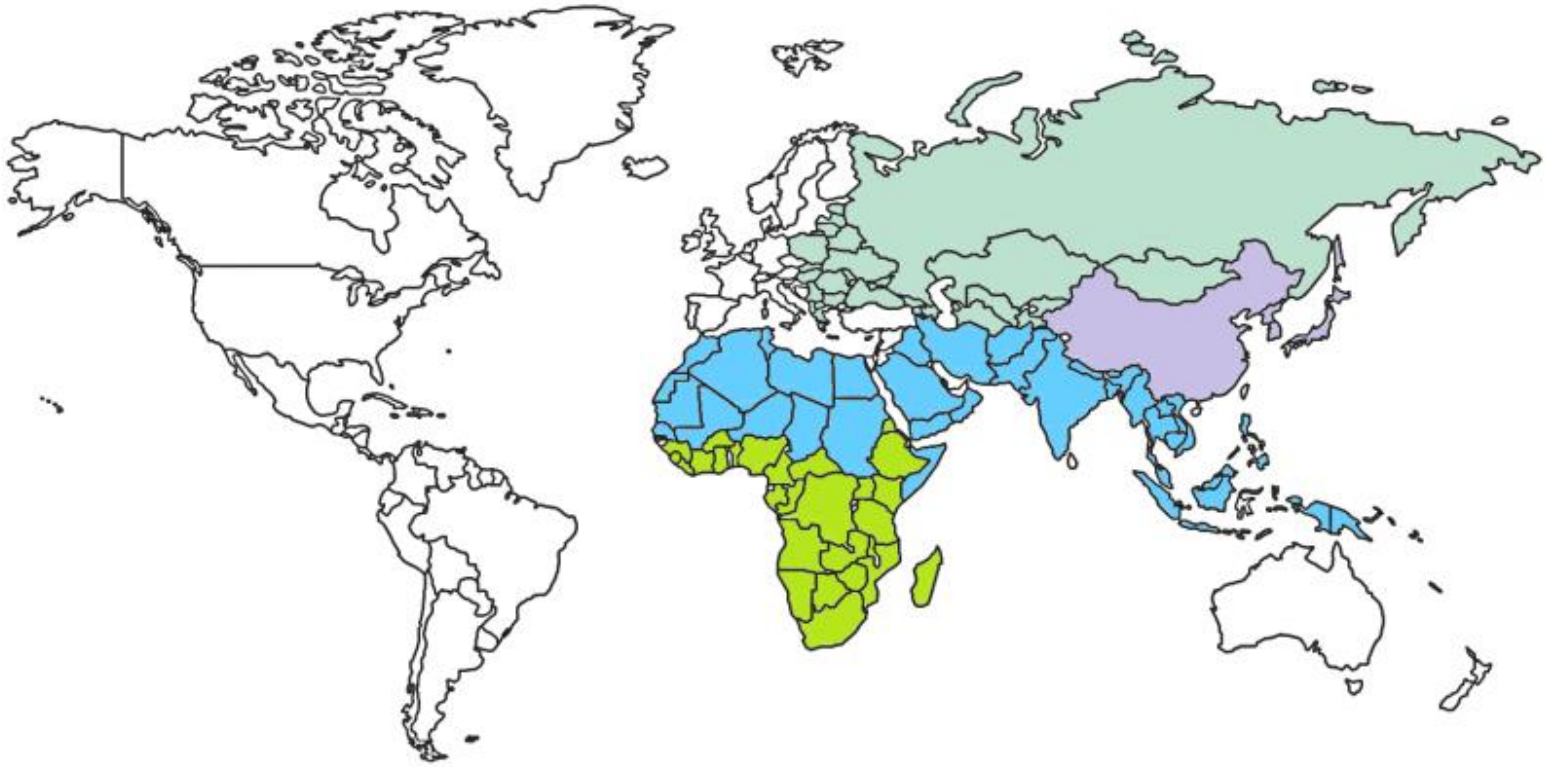
Doosare din brahm muhurt men uth kar tatha nityakram aur sandh
Lakshman guru Vishwamintra ke pas ja kar bole, “Gurudev! kripa
yagya men vighna dalane ke liye kis samay aate hain? Yah ham is
na ho ki hamare anjane men hi ve aakar upadrav machane lagen.

Dashrath ke veer putron ke in utsahbhare vachanon ko sun kar wa
atyant prasanna huye aur bole, “He raghukulbhushan rajkumaron!
se chhah dinon tatha ratriyon tak puran roop se savdhan mudra m
dinon men Vishvamitra ji maun hokar yagya karenge. Is samay bh
denge kyonki ve yagya ki diksha le chuke hain.”

And lot more

Languages and Scripts

- Arabic (Saudi Arabia, UAE, Egypt, Morocco,…)
- Persian
- Indian sub-continental languages (IL & Dzongkha, Nepalese, Sinhala)
- Thai
- Cyrillic (Russian, Ukrainian)
- Chinese, Japanese, Korean (rare)



Roman Transliteration for IL

- Used extensively in CMC
  - Chats, SMS, Emails
- Reasons
  - Lack of standard IL input mechanisms
  - Familiarity with the QWERTY keyboard
  - Familiarity with English for the Indian Internet user

Examples

অভ্যেস

- abhyes
- abhyesh
- abbhes
- abhes
- abhesh
- abhess
- abhyas
- obhesh
- ovesh
- abhyash
- avesh
- avyas
- obbhes
- obbhesh
- obbhyash
- obbhyesh
- obbyesh
- obhes

সমষ্টি

- samashti
- samosti
- smoshti
- somoshthi
- somoshti
- somosti samarti
- samasti
- samosthi
- somosthi

লাভ

- laabh
- labh
- laab
- lab
- lav
- luv

What will we learn?

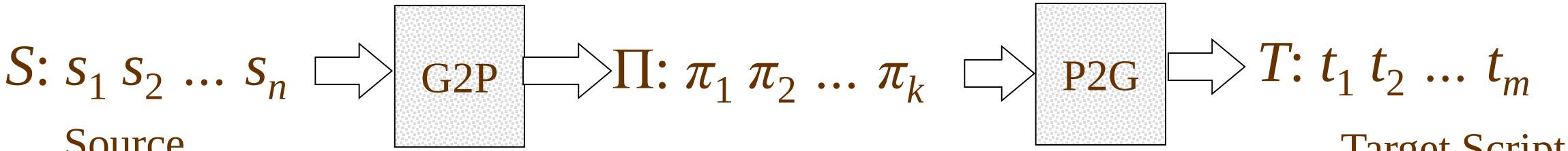
- Transliteration in SM: What, why and how much
- Noisy Channel models for transliteration
- Word embedding for transliteration
- Language Detection

The Basic Noisy Channel Model

$$S^* = \delta(T) = \underset{S}{argmax}\ Pr(S|T)$$

$$= \underset{S}{argmax}\ Pr(T|S)Pr(S)$$

Modified Noisy Channel Model



$$S^* = \delta(T) = \underset{S}{argmax}\ Pr(S|T)$$

$$= \underset{S}{argmax}\ Pr(T|S)Pr(S)$$

$$= \underset{S,\,\Pi}{argmax}\ Pr(T|\Pi)Pr(\Pi|S)Pr(S)$$

Estimating the channel model

$$S^* = argmax\ Pr(T|\Pi)Pr(\Pi|S)Pr(S)$$

Bayesian Approach:

$$S^* = argmax_S \sum_{\substack{all\ possible \\ phoneme\ strings: \Pi}} P(T|\Pi)P(\Pi|S)$$

Maximum Likelihood/frequentist Approach:

$$S^* = argmaxSargmax_{\Pi} P(T|\Pi)P(\Pi|S)$$

How to represent phonemes?

What data do we need to learn the probabilities?

Do we really need pronunciation data?

What will we learn?

- Transliteration in SM: What, why and how much
- Noisy Channel models for transliteration
- Word embedding for transliteration
- Language Detection

Geometry of words

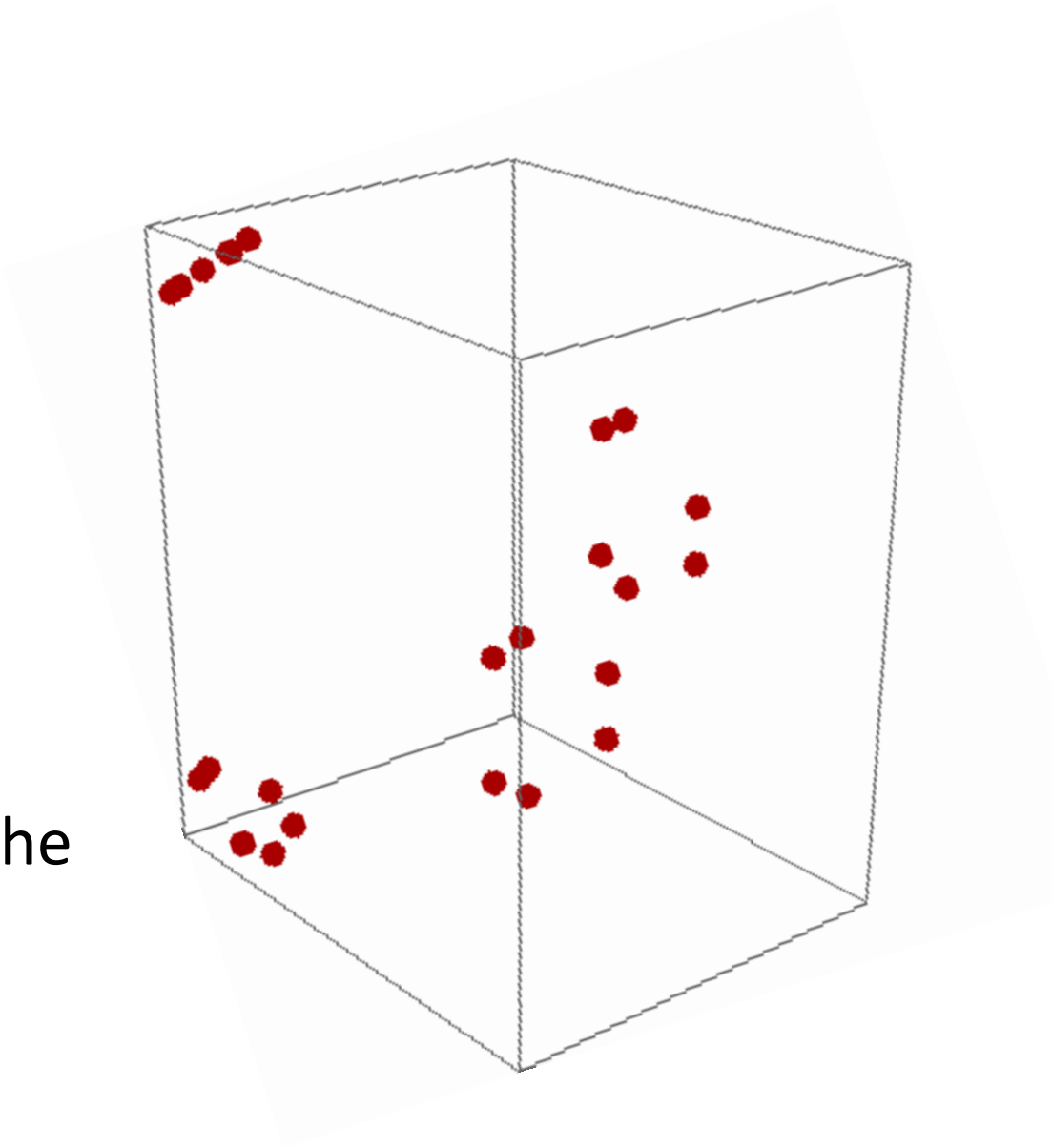
Words can be mapped to abstract n-dimensional spaces.

$$f: \mathbf{L} \rightarrow \mathbf{R}^n$$

$$f(w) = \{x_1, x_2, x_3, \ldots, x_n\}$$

Objective:

The distance between two words in the abstract space is an indicator of their transliterational similarity.



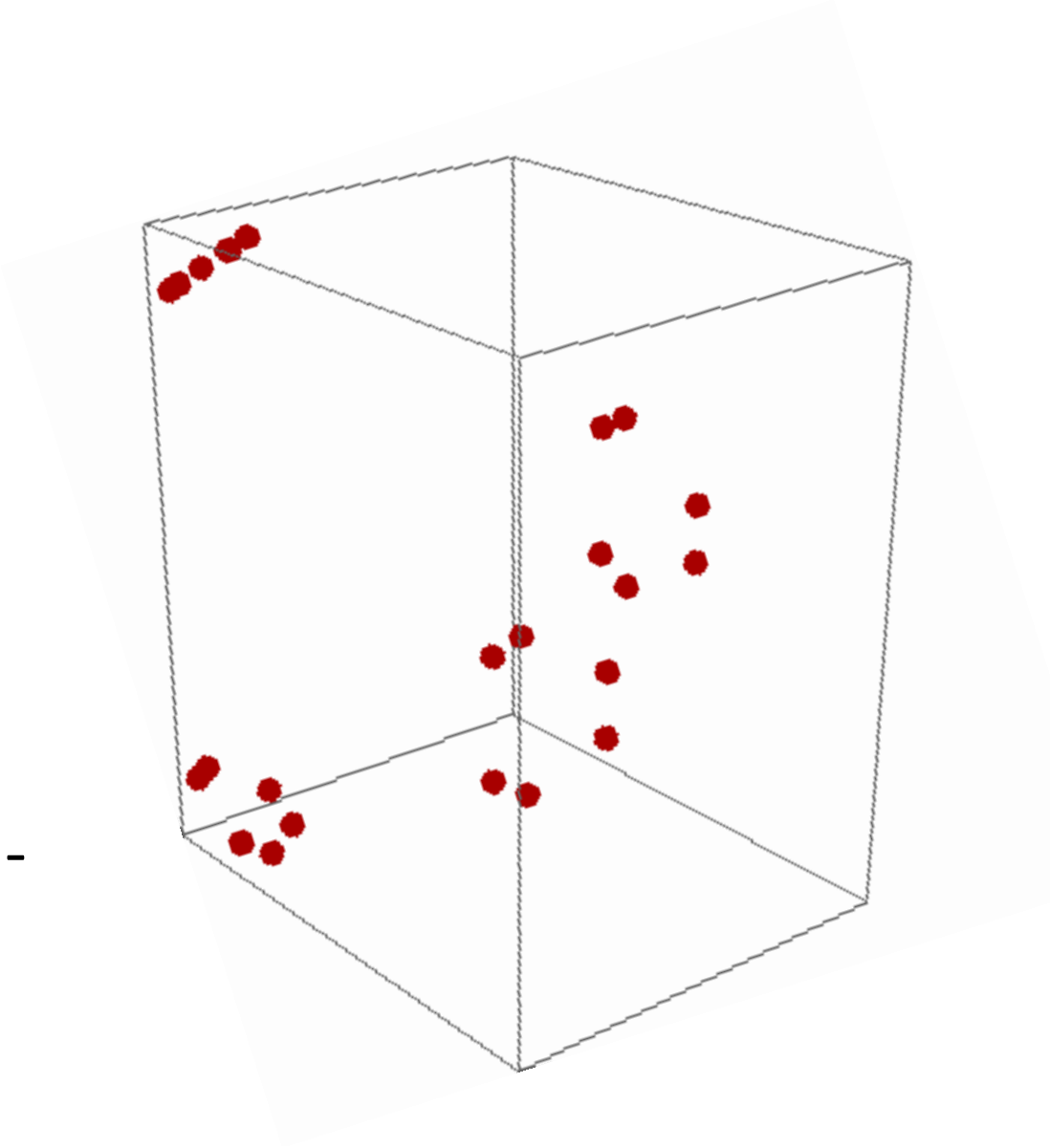
Geometry of words

$$f(w') - f(w) = \sum (x'_i - x_i)^2$$

Euclidian distance, cosine distance, …

The function f

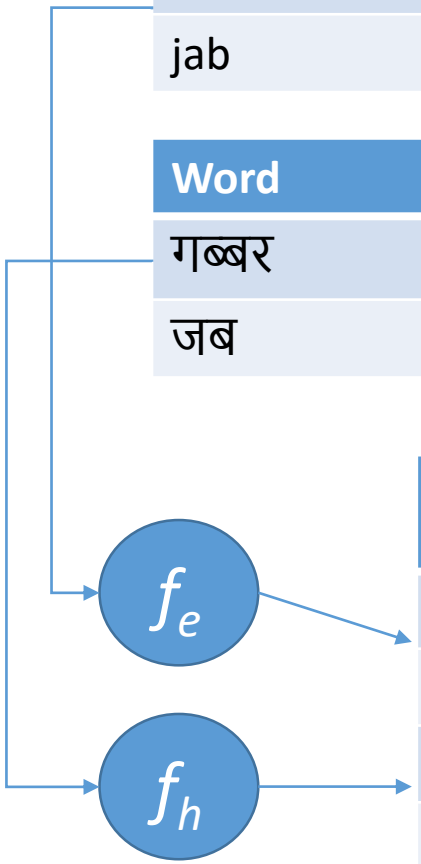
- is called an embedding of the words.
- can be learnt from the data
- commonly used features: character n-grams



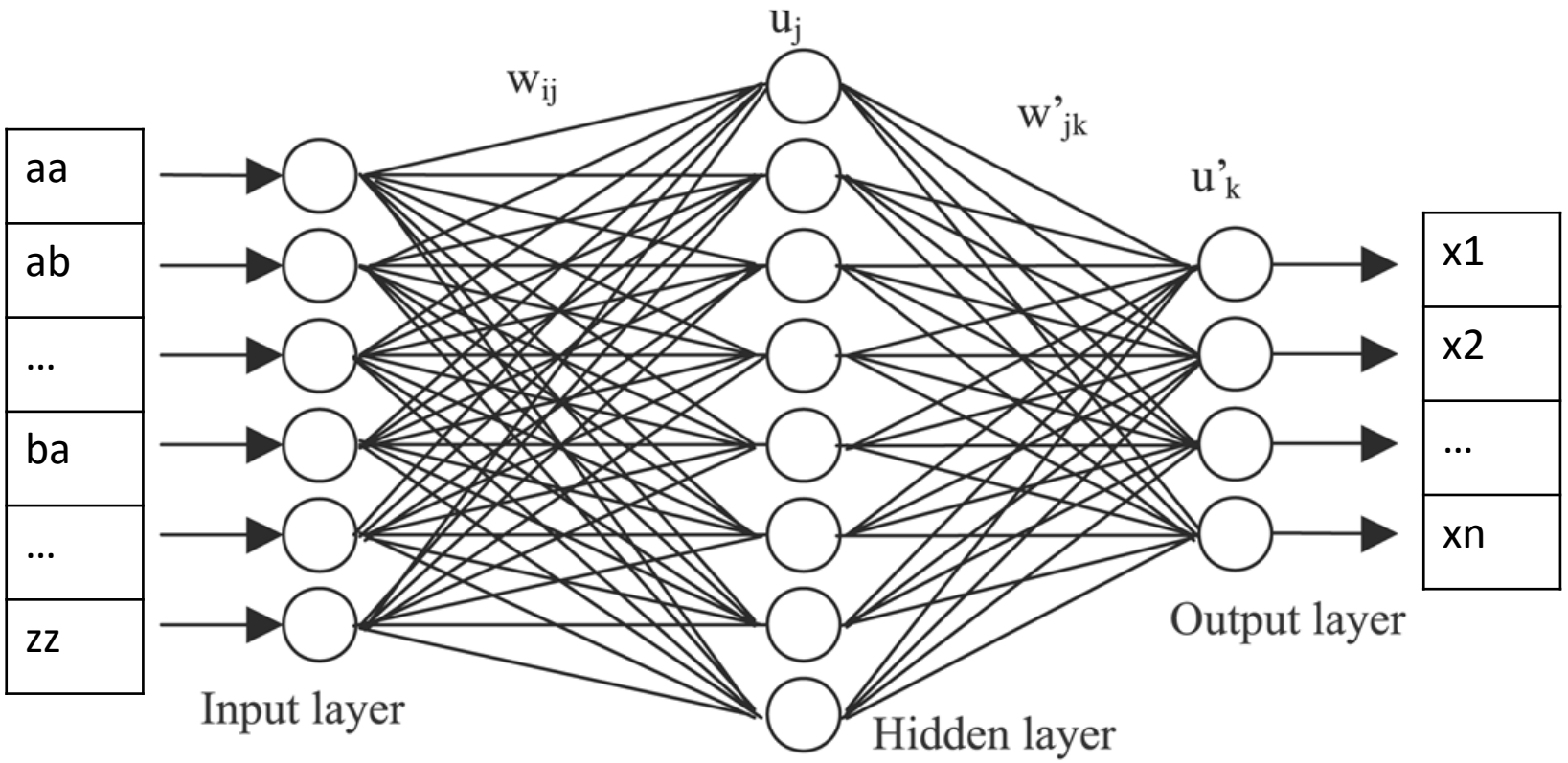
Character bigram based embedding

Word	aa	ab	…	ba	bb	…	zy	zz
babbar	0	1	…	2	1	…	0	0
jab	0	1	…	0	0	…	0	0

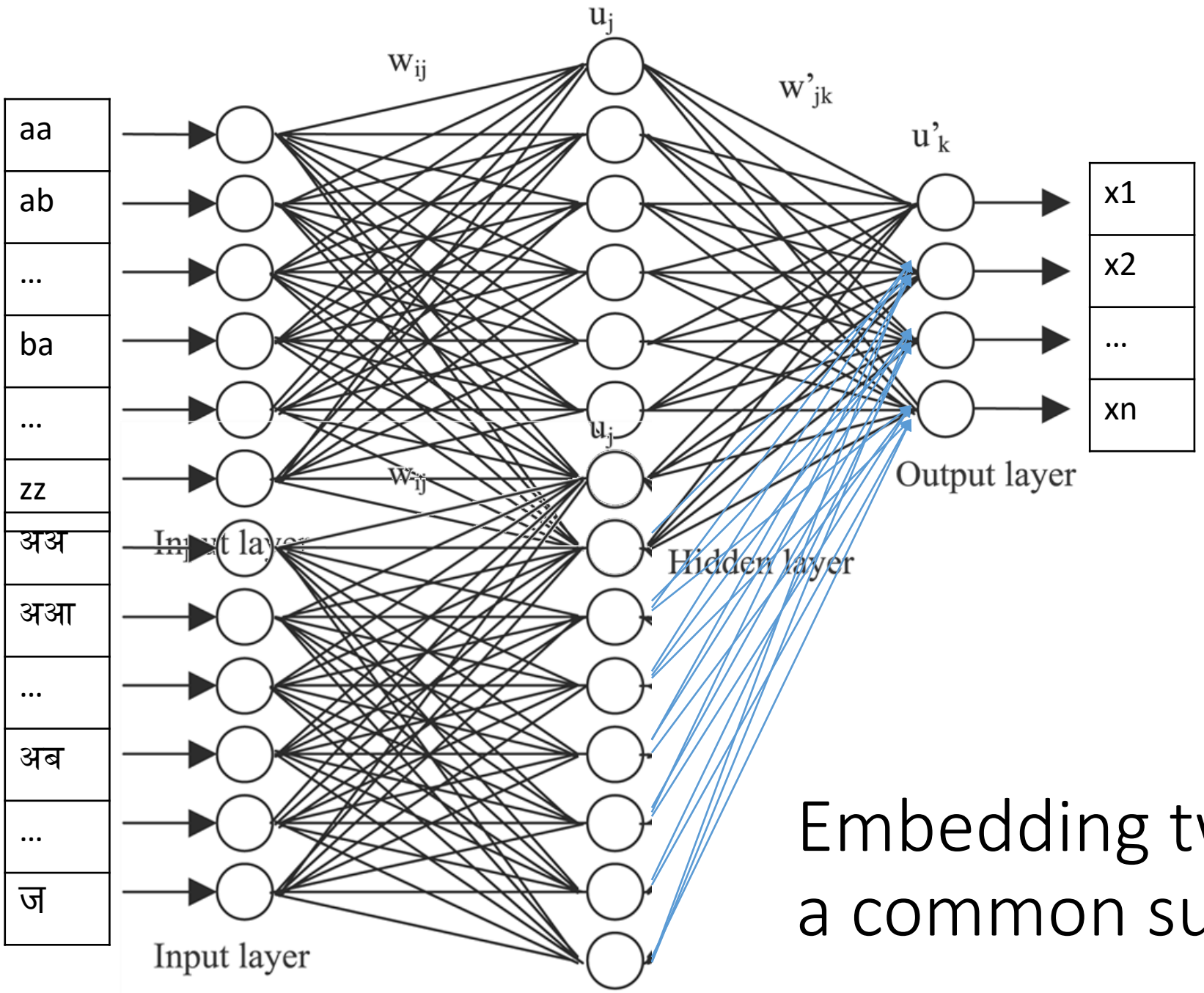
Word	अअ	अआ	…	अब	ज	…	ग	गा
गब्बर	0	0	…	1	0	…	1	0
जब	0	0	…	1	1	…	0	0



word	x1	x2	x3	x4	…	x_{n-1}	x_n
babbar	0.23	0.00	0.12	0.92	…	0.00	0.00
jab	0.00	0.72	0.5	0.02	…	0.25	0.00
गब्बर	0.00	0.05	0.31	0.90	…	0.00	0.00
जब	0.00	0.65	0.55	0.00	…	0.42	0.00



Learning to embed



Embedding two scripts to a common sub-space

Suggested Readings & References

NC based Transliteration:

- Knight, Kevin, and Jonathan Graehl. "Machine transliteration." *Computational Linguistics* 24.4 (1998): 599-612.
- Sarvnaz Karimi, Falk Scholer and Andrew Turpin, Machine Transliteration Survey. In ACM Computing Surveys (CSUR), Volume 43 Issue 3, April 2011

Word Embedding for Transliteration:

- Gupta, Parth, et al. "Query expansion for mixed-script information retrieval." *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014.
- Udupa, Raghavendra, and Mitesh M. Khapra. "Transliteration equivalence using canonical correlation analysis." *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2010. 75-86.

References

APIs and Web Services for Indic Transliteration:

Google API, Yahoo API, Quillpad

Surveys:

- P. J. Antony and K. P. Soman, Machine Transliteration for Indian Languages: A Literature Survey. In *International Journal of Scientific & Engineering Research*, Volume 2, Issue 12, December-2011

Datasets:

- Kanika Gupta and Monojit Choudhury and Kalika Bali, Mining Hindi-English Transliteration Pairs from Online Hindi Lyrics, In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

FIRE Shared Task on Mixed Script IR

http://research.microsoft.com/en-us/events/fire13_st_on_transliteratedsearch/fire15st.aspx

Training/Dev data release: 11th Aug 2015

Registration closes: 31st Aug 2015

Test Set release: 28th Sep 2015

Submit Run: 12th Oct 2015