Query Word Labeling and Back Transliteration for Indian Languages: MSRI Shared task system description

Spandana Gella[1,2], Jatin Sharma[1], Kalika Bali[1]

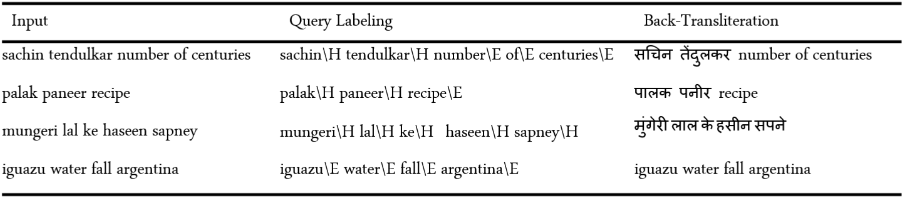
[1]Microsoft Research, India [2]University of Melbourne, Australia

December 9, 2013

Microsoft Research

SubTask1: Query Word Labeling

Many Indian languages esp. in social media is written using romanized script



Input	Query Labeling	Back-Transliteration
sachin tendulkar number of centuries	sachin\H tendulkar\H number\E of\E centuries\E	सचिन तेंदुलकर number of centuries
palak paneer recipe	palak\H paneer\H recipe\E	पालक पनीर recipe
mungeri lal ke haseen sapney	mungeri\H lal\H ke\H haseen\H sapney\H	मुंगेरी लाल के हसीन सपने
iguazu water fall argentina	iguazu\E water\E fall\E argentina\E	iguazu water fall argentina

Table: Shared Task description in two separate steps of query labeling and back transliteration

Our Methodology

- Word level language identification
  - based on character n-gram features learned from wordlists extracted from monolingual corpus (King and "Abney, 2013)
  - Adding context switch probability to indirectly learn the language sequence patterns
  - Frequency based filtering
- Back-Transliteration
  - Hash based mapping between source and target languages (Kumar and Udupa, 2011)
  - Use indic character mapping to create training data in poor-resource languages

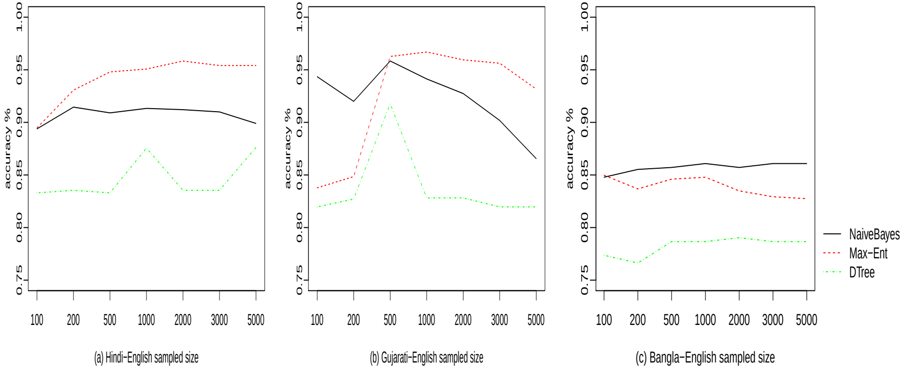
Microsoft Research

Terminology, Datasets and Tools

- Character n-gram features: hello :’h’,’e’,..,’o’,’he’,el’..,’hel’..,’hell’,’ello’,’hello’
- Training resources: Word lists (from Leipzig Corpus, Anandbazar Patrika), word frequencies and transliterated pairs given as part of shared task
- Training size from 100 - 5000 words (Always <=546 for Gujarati)
- (McCallum, 2002) for learning classifiers, MSRI Name Search Tool for Transliteration

Microsoft Research

Word label prediction based on n-gram features

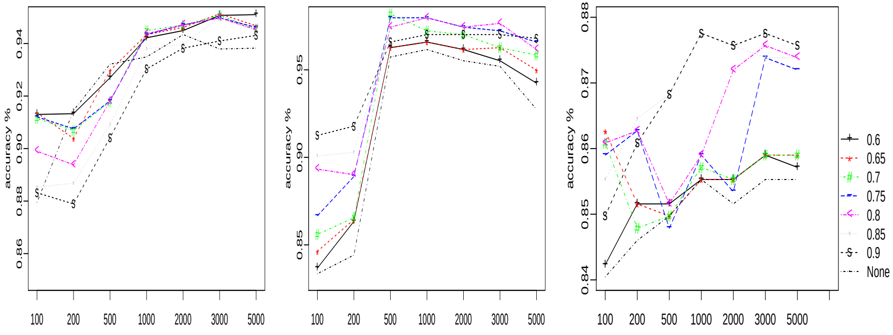


(a) Hindi (b) Gujarati (c) Bangla

Figure: Learning curves for maximum entropy, naive Bayes and decision tree on word labeling for Hindi, Gujarati and Bangla language on development data

Microsoft Research

Adding context-switch probability



(a) Hindi - Maxent

(b) Gujarati - Maxent

(c) Bangla - Naive

Figure: Learning curves with varying context switch probabilities

Language Identification Errors

Type	Romanized	Predicted	Reference
Short Words	i; ve	H; E	E; H
Ambiguous Words	the; ate	E; E	H; H
Erroneous Words	emosal	H	E
Mixed Numerals Words	zara2; duwan2	E; E	H; H

Table: Annotation Errors

Back Transliteration

- MSRI Name Search Tool, built based on n-gram based feature hashing
- Used indic character mapping between Hindi-Bangla and Hindi-Gujarati
- All 3 systems for Gujarati and Bangla uses indic character mapping

Microsoft Research

Test set Results

System	Hindi			Gujarati			Bangla		
	LA	TF	TQM	LA	TF	TQM	LA	TF	TQM
MSRI-1	0.9823	0.8127	0.1940	0.9614	0.4711	**0.0800**	0.9259	0.4914	**0.0100**
MSRI-2	**0.9848**	**0.8130**	**0.1980**	**0.9755**	**0.4803**	0.0733	**0.9499**	0.5033	**0.0100**
MSRI-3	0.9826	0.8101	0.1860	0.9661	0.4748	0.0667	0.9459	**0.5137**	**0.0100**
Maximum	**0.9848**	**0.8130**	**0.1980**	**0.9755**	**0.4803**	**0.0800**	**0.9499**	**0.5137**	**0.0100**
Median	0.9540	0.4160	0.0290	0.9661	0.4748	0.0733	0.9359	0.4973	0.0100

Table: Language labeling analysis on submitted runs in all three languages, along with maximum and median scores. Our runs which had maximum scores are presented in **bold**. LA - Labeling Accuracy, TF- Transliteration F-score, TQM - % of queries that had exact labeling and transliteration

Transliteration Error Analysis

Type	Romanized	Predicted	Reference
Erroneous Latin Source	hau\H; utari\G; banglae\B	हाउ; ઉતારી; বংগলে	है; ઉતારો; বাংলায়
Multiple Candidates	kali\H; vidhi\G; par\B	काली; વિધી; পর	कली; વિધિ; পার
Multiple Transcriptions	tanhai\H; barbadi\G	तनहाई; બર્બાદી	तन्हाई; બરબાદી
Merged Words	gayazamana\H; hradayama\G; saralikaraner\B	घनश्याम; હમદર્દ; সর্রক্ষেণ	गयाज़माना; હૃદયમાં; সরলীকরণের
Plural Words	neendo\H; mandiro\G	नींदों, મંદિરોં	नींदो; મંદિરો
Distorted Words	mauja\H	मौज	मौजा
Language Specific	paani\G; kolkatar\B	પાની; কোলকাতার	પાણી; কলকাতার
Lexicon Coverage	chaudavi\H	चढ़ाव	चौदवी
Vowel Error	Gai\B; bali\B	গাঈ; বালী	গাই; বালি
Errors in Training Set	bijuriya\H; nahi\G	बिजुरिय; નહિ	बिजुरिया; નહીં
Miscellaneous	bina\H	बिना	सिवा

Table: Transliteration Errors

Microsoft Research

Summary

- Contributions:
  - Using context switch probability increases the performance of language labeling in code-mixed language.
  - Cross-language character mapping to increase transliteration accuracy - promising direction for resource-poor languages
- Future Work:
  - Extending it to text with spelling variations (covering text normalization)
  - Working on multiple languages esp. poor resource languages by exploiting resources from related languages

Microsoft Research

Questions?

Microsoft Research

Bibliography I

King, B. and "Abney, S. (2013). Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of NAACL-HLT*, pages 1110–1119.

Kumar, S. and Udupa, R. (2011). Learning hash functions for cross-view similarity search. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two*, pages 1360–1365. AAAI Press.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.

Microsoft Research