Romanagari Detection in Twitter

Hrishikesh Terdalkar Shubhangi Agarwal

CS671: Natural Language Processing
Indian Institute of Technology Kanpur

Problem Definition

Romanagari is Devanagari-script-based-language written in Roman script. Given random collection of roman-script tweets, we want to find out tweets that are English-Hindi codemixed (or pure Hindi), tag the individual words as well as entire tweet with language prediction.

Challenges:

1. Twitter small-ish max. 140 character text, huge inflections.
2. Lack of clean or good annotated datasets for training and testing.

Data Collection

Used Datasets

- *Rovereto Twitter n-gram Corpus*: is an n-gram dataset. 42 million n-grams.[1]
- *NLTK tweet_samples*: English tweets collection, part of NLTK Corpora containing 20,000 tweets.
- *IITB Hindi Devanagari Corpus*: Devanagari script Hindi corpus containing around 1200 files.[2] It has roughly 220,000 lines (2.85 million words). We converted this to Roman script to use for training.

Collected Datasets

- *Hindi-English Tweets Corpus*: (Code-mixed)
  Using Twitter's REST API. **38,264 tweets** of rich code-mixed quality.
  skip-gram on 94 most-frequent Hindi words, 4,371 pairs, obtained **335,672 tweets** from this.
- *Social Media: gchat, WhatsApp, Facebook*: (Code-mixed)
  handpicked codemixed text from social media such as Google-talk, WhatsApp, Facebook.
  Overall 297 lines of Hindi and 390 lines of Marathi were collected.

Preprocessing

Tools:
various bash, awk, sed, grep, tr, python, js scripts, `SRILM Tools`, `NLTK Tools`

Cleaning and Statistics

Rovereto (RTC) corpus contains lot of noise. We only took n-grams that *do not* contain *any* special characters, and added up demographic information to obtain frequency of n-grams for $n = 1, 2, \ldots, 6$. This reduced total size of corpus from 250 GB to 1.2 GB.

Tweets Cleaning For tweets, removed duplicates, retweets and tweets containing URLs, accents. Also lower-cased the entire corpus. Replaced mentions by word "HANDLE".
Resulted in final corpus **59,287 Hindi** and **3187 English** tweets, tagged with start-end markers <s> and </s>.

Social Media (Handpicked)
passed it through `basic_cleaning`.
`tagger` script to tag 297 Hindi and 300 Marathi lines.
eg: `<s> <hi>bhaisaab itna mazaa kafi</hi> <en>time</en> <hi>baad aya</hi> <en>a lot of catching up</en> <hi>bhi ho gayi</hi> </s>`

IITB Hindi (Devanagari) For this large Devanagri corpus we ran `devToRom.js` using `node.js` and coverted it to Roman text (3 char-look-ahead character-level).

Data Statistics

N-grams in Training Set

	English	Hindi
1-grams	1,168,077	120,546
2-grams	10,644,439	998,300
3-grams	17,353,446	2,027,733
4-grams	14,007,551	65,186,143

Tagged Data

Language (type)	Count
English (lines)	3187 tweets
Hindi (lines)	3000 tweets
Hindi (words)	297 lines
Marathi (words)	390 lines

Soundex

Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English[3]
Evaluate Soundex's output, using FIRE 2013 data[4]
Over 30,000 transliteration pairs of (Roman Variation, Devanagari Word)
18,000 unique Hindi words, 6,500 words with 2 or more variations
Variations handled by Soundex: 53.6% (*all*), 57% (*top 1000*), 62.3% (*top 100*)

Simple Ngram - Model

q is query, with Soundex applied
$q = w_1, w_2, \ldots, w_n$

	Model v1	**Model v2**
$w\_score(w, L)$	$\max_n$ (ngram around $w \in q) \in L$	$\sum_n \sum_j$ n j is an ((ngram around w) $\in q) \in L$
$q\_score(q, L)$	$\sum_{w \in q}$ [$w\_score(w, L)$]^2	$\sum_{w \in q}$ [$w\_score(w, L)$]

SRILM Probabilistic Ngram-Model

Learning probabilities of words in vocab based on n-gram probabilities (*previous-context*)
Calculates backoff weights
Evaluation of the TestData gives the conditional probability of each word in the best ngram context found with its backoff weight.

$$p(B15200|A35200...) = [1gram]0[-1.0867]$$

3 zeroprobs, logprob= -81.4182 ppl= 33360.6 ppl1= 61561.1

Perplexity = Confusion
Less perplexity = more Confidence.

Results

Tweets Tagged using Various Models

	English		Hindi (codemix)	
	Correct	Wrong	Correct	Wrong
Simple v1	3165	22	352	2648
Simple v2	3176	11	97	2903
SRILM (ppl=20k)	2052	1135	1777	1223
SRILM (ppl=25k)	1920	1267	1936	1064
Wordratio(40,20)	2076	1111	2687	313
Wordratio(50,10)	890	2297	2921	79

Word-Ratio(M, N) Model

Based on word-tags output by SRILM Model

- $hi_{freq} = hi_{count}/total\_words$
- $en_{freq} = en_{count}/total\_words$
- **if** $(hi_{freq} + en_{freq}) < M$ then mark as Other
- **else if** $(hi_{freq} > N)$ then mark as Hindi
- **else** mark as English

Tested for $(40, 20)$ and $(50, 10)$

Conclusion and Future

- wordratio based sentence tagging works well with n-gram-probabilistic word tagging
- dependent on nature and statistics of data.
- soundex effective (resolves 53% variations on average, 62.3% for top 100)

Future

- soundex alternatives
- multiple datasets with different conditioned statistics
- "goodness measure" on models

References

[1] A. Herdagdelen. Rovereto twitter n-gram corpus. [Online]. Available: http://clic.cimec.unitn.it/amac/twitter_ngram/
[2] Iitb hindi corpus. [Online]. Available: http://www.cfilt.iitb.ac.in/Downloads.html
[3] Soundex indexing system. [Online]. Available: www.archives.gov/research/census/soundex.html
[4] Datasets of fire 2013. [Online]. Available: http://cse.iitkgp.ac.in/resgrp/cnerg/qa/fire13translit/
[5] U. Barman, A. Das, J. Wagner, and J. Foster, "Code mixing: A challenge for language identification in the language of social media," *EMNLP 2014*, p. 13, 2014.
[6] A. Das and B. Gambäck, "Identifying languages at the word level in code-mixed indian social media text," in *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*, 2014, pp. 169–178.
[7] S. Gella, K. Bali, and M. Choudhury, "'ye word kis lang ka hai bhai?' testing the limits of word level language identification."
[8] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit." in *INTERSPEECH*, 2002.
[9] Latin-to-roman transliteration. [Online]. Available: http://www.hindidevanagari.com/transliteration/
[10] Twitter package - nltk. [Online]. Available: http://www.nltk.org/howto/twitter.html

Acknowledgements

We are grateful to Prof. Amitabha Mukerjee and our senior M S Ram for valuable insights. We would like to also thank friends and family for help in data collection and tagging.