Romanagari Detection in Twitter
Hrishikesh Terdalkar  Shubhangi Agarwal

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.
- ITB Hindi Devanagari Corpus: Devanagari script Hindi corpus containing around 1,200 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, 36,264 tweets of rich code-mixed quality. Skip-grain on 94 most-frequent Hindi words, 4,971 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 are not 
- Special: gchat, WhatsApp, Facebook: (Code-mixed) 

Soundex
Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English.[14]

Learning probabilities of words in vocab based on n-gram probabilities (q word in n-gram around w ∈L)

Model v1
\[
\sum_{w \in L} \sum_{j} n_{j \in \text{models}} (\text{word tagged with n-gram-probabilistic word tagging})
\]

Model v2
\[
\sum_{w \in L} n_{j \in \text{models}} (\text{word tagged with n-gram-probabilistic word tagging})
\]

Results

<table>
<thead>
<tr>
<th>N-grams in Training Set</th>
<th>Tagged Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
</tr>
<tr>
<td>1-grams</td>
<td>148,007</td>
</tr>
<tr>
<td>2-grams</td>
<td>10,644,439</td>
</tr>
<tr>
<td>3-grams</td>
<td>17,353,446</td>
</tr>
<tr>
<td>4-grams</td>
<td>14,007,551</td>
</tr>
</tbody>
</table>

Language Type
- English (lines): 218 tweets
- Hindi (lines): 3440 tweets
- Hindi (words): 397 lines
- Marathi (words): 297 lines

Word-Ratio(M, N) Model

Based on word-tags output by SRILM Model

\[
\text{freq}(\text{word} | \text{context}) \text{ tag as English} / \text{freq}(\text{word} | \text{context}) \text{ tag as Hindi (or Other or Marathi)}
\]

Decision and Conclusion

Future
- soundex alternatives
- multiple datasets with different conditioned statistics
- “goodness measure” on models

References

Acknowledgements
We are grateful to Prof. Amitabh 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.

http://home.iitk.ac.in/~hrishit/cs671/project/ | hrisht, sagarwal | @cse.iitk