Romanagari Detection in Twitter

Hrishikesh Terdalkar - Shubhangi Agarwal

Problem:

Romanagari is writing Devanagari-script-based-language in Roman script. We often see examples of these in social media such as twitter. e.g. "*kya kar rahe ho*".
Most NLP techniques deal with only English text, and thus in majority of twitter based studies, such tweets have to be discarded as noise.
We intend to tackle this problem, by attempting detection and translation of such tweets. Targeted languages are Hindi and Marathi.

Approach:

The main task is that of "back-transliteration". That is, going back to Devnagari script from Romanagari text. After which, translation is mere formality.

Steps:

- ❏ Create dictionary of Romanagari words, by handling words at phoneme level
- ❏ Consider common inflections of Romanagari lexemes
- ❏ Collect / Obtain a twitter-dataset containing Romanagari text
- ❏ Attempt the back transliteration followed by translation
    - ❏ for this work by Barman, Utsab, et al 2013 [2] can be used as reference point

Dataset Collection:

Since we are working on twitter, dataset will be (most likely) collected with help of existing tweet-collection software.
We choose *python programming language*.
One of the notable tools available is,

- ❏ **Tweepy**: https://github.com/computermacgyver/twitter-python

Details about usage at:

- ➔ http://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/
- ➔ http://stats.seandolinar.com/collecting-twitter-data-using-a-python-stream-listener/

References:

1. Barman, Utsab, et al. "**Code Mixing: A Challenge for Language Identification in the Language of Social Media.**" *EMNLP 2014* (2014): 13.
2. Gella, Spandana, Jatin Sharma, and Kalika Bali. "**Query word labeling and back transliteration for indian languages: Shared task system description.**" *FIRE Working Notes* (2013).
3. Gella, Spandana, Kalika Bali, and Monojit Choudhury. "**"ye word kis lang ka hai bhai?" Testing the Limits of Word level Language Identification.**"
4. Das, Amitava, and Björn Gambäck. "**Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text**." *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*. 2014.
5. Han, Bo, and Timothy Baldwin. "Lexical normalisation of short text messages: Makn sens a# twitter." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011.
6. Proceedings of Social india 2014