Data Challenge: Differentiating Code-Borrowing from Code-Mixing[*]

IIIT-H submission report

AYUSHI PANDEY and BRIJ MOHAN LAL SRIVASTAVA[†], International Institute of Information Technology

The phenomenon of code-mixing assumes particular relevance in predominantly multilingual societies such as India, where medium of education and the mother tongue are usually different languages. Culminating perspectives from empirical linguistics and natural language processing, we propose a metric to rank the given set of words based on their variant fiborrowedfi or fimixedfi status. The 5-dimensional feature vector incorporates scores from three different models; phonological, orthographic, syllables models, the frequency of rhyming words, and the proximity of the language of the surrounding words. Automatic learning of this feature vector is conducted through stochastic averaging.

General Terms: code-mixing, borrowing, social media

ACM Reference format:
Ayushi Pandey and Brij Mohan Lal Srivastava. 2017. Data Challenge: Differentiating Code-Borrowing from Code-Mixing. 1, 1, Article 1 (January 2017), 4 pages.
DOI: 31.01.2017

1 INTRODUCTION

Bilingual speech communities recognize code-mixing and code-borrowing as predominant phenomena in conversational speech. Borrowed words are those that have been incorporated in the vocabulary of the monolingual speaker, usually having adapted the phonological and morphological patterns of the matrix language. Mixed words, on the other hand, belong largely in the repertoire of a bilingual speaker, often not reflecting the phonological/morphological influences of the language they are borrowed from. Distinguishing borrowed words from mixed ones reduces the computational cost of a search query, in selectively identifying monolingual or bilingual documents to search from. The proposed metric provides a 6-dimensional feature-vector comprising of the following feature-set.

- Phonological model with pronunciation variants
- Orthographic model
- Syllabification model for legitimate syllable sequence
- Prosodic model for legitimate prosodic structure
- Frequency of rhyming words

[*]

[†] Alphabetical order by last name. Authors have contributed equally to this work.

- Language proximity of neighbouring words

2 MODEL DESCRIPTION

This section provides a description of the 6-dimensional feature-set and the passing of this feature-set to the function of stochastic averaging. Each test word is transliterated to Devnagri font using litcm transliterator. (Bhat et al. 2015)

2.1 Pronunciation model

The pronunciation model is built using phoneme sequences associated with 15,500 words scraped from a monolingual corpus (Technologies 2015). Each phone sequence was given a weight obtained using max-normalized frequency from the corpus. Phoneme sequences were generated using a grapheme to phoneme (G2P) converter trained on $\tilde{7}000$ words using sequence-to-sequence (Sutskever et al. 2014) learning approach implemented in Tensorflow. The predictions were hand-corrected by the authors to generate a gold-standard pronouncing dictionary. We proposed a rule-based mapping for the set of likely phones that a particular phone can be confused with in case of code borrowing . Each input phone sequence was converted to its possible pronunciation variants according to these rules. The variant with the highest likelihood w.r.t. pronunciation model was selected. In addition to an independent score returned for the observed phoneme sequence, the pronouncing dictionary paved way for the following three modules.

2.2 Prosodic model

A borrowed word is more likely to have adapted to the legitimate syllable sequences and prosodic structure of the recipient language. For example, the word fiofficerfi has assumed the shape of “afsar”, resembling a prevalent prosodic structure in Hindi (“kamtar”, “aksar” etc). The stress patterns (implementation beyond the scope of this work) have also been decomposed into following the syllable-timed nature of the recipient Hindi . To obtain the prosodic structure of each word, the consonants were stripped and only the vowels were passed to the model. The prosodic structure for each word was used to create a prosodic model over those words, in the same way as for the phonological model described above. The log likelihood is computed from the model and passed to the feature vector.

2.3 Syllabification model

To learn the legitimate syllable sequence structure, the pronouncing dictionary was syllabified into syllable sequences of the form CV, VC, VC* and C*V. Consonants in the word-initial and word-final clusters were grouped together, but consonants of the word-medial clusters were divided between the syllables, for example “upkram” was syllabified as “up”, “kram”. The syllabifed form for each word was used to create a syllabification model over those words, in the same way as for the phonological and prosodic model described above. The log likelihood is computed from the model and passed to the feature vector.

2.4 Rhyming word frequency model

Following from a valid syllable and prosodic structure, the likelihood of a given word being borrowed or mixed may also depend on the frequency of similar sounding words in the recipient language. For a given test word then, the frequency of its rhyming words in a monolingual corpus may predict how likely the word is to be borrowed into the language. For example, words like “life”, “interview”, “price”, “link” etc do not have minimal or sub-minimal pairs in Hindi, and therefore can be assumed to rank lower in their borrowing likelihood. For a small set of words (like “state”), however, this feature may be an over-generalisation. Using this assumption, the probability score of partial and total rhymes of a given test word was computed w.r.t the monolingual corpus, and this feature was appended to the feature vector.

2.5 Orthographic model

Devnagri recognises a primarily one-to-one mapping between the orthography and the pronunciation of the word. Much of the phonological information is hence, likely to be contained in the orthography of the script. To obtain this information we calculate the log likelihood of character sequence of the monolingual corpus. This score is supplied to the feature vector as yet another component.

2.6 Language proximity model

Using language-identity information from the Twitter data supplied, we obtain the language tags of the immediate neighbours of a given test word for each of its occurences. Test words contained in a neighbourhood of Hindi are more likely to be borrowed than the ones surrounded by English. The frequency of monolingual neighbours is supplied as a feature for the 6-dimensional feature vector.

The final ranking score is obtained by computing the stochastic average of the probabilities independently returned by each of these 6 models. The stochastic average is represented by the following equation:

$$\begin{aligned} cbindex(w) = log10(argmax(p(eg2p(w)|\lambda_p))) + log10(p(dev(w)|\lambda_o))+ \\ log10(p(prosody(hg2p(dev(w)))|\lambda_{pro}))+ \\ log10(syll(hg2p(dev(w)))|\lambda_{syll}))+ \\ log10(p(rhyme(w)|\lambda_{rhyme}))+ \\ log10(p(data(w))) \end{aligned} \tag{1}$$

eg2p, *hg2p* are English and Hindi G2P functions, *dev* converts Roman characters to Devanagari, *prosody* strips consonants off the phoneme sequence, *syll* merges phonemes to obtain syllables, *rhyme* gives syllable rhyming score and *data* gives twitter probability of word w. $\lambda_p, \lambda_o, \lambda_{pro}, \lambda_{syll}, \lambda_{rhyme}$ are pronunciation, orthographic, prosodic, syllabic and rhyming models respectively.

3 CONCLUSIONS

As part of the IIIT-H submission to the ACM-IKDD Data Challenge, we propose a metric to evaluate a given set of words based on their borrowing likelihood. A feature-set deriving information from the phoneme sequence, prosodic structure, syllabic form and orthographic structure was proposed. Additionally, the frequency of rhyming words in the monolingual corpus, and the proximity of language in the neighbourhood were also proposed as features. The ranking of the 230 words was generated using stochastic averaging over the proposed feature vector.

REFERENCES

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. IIIT-H System Submission for FIRE2014 Shared Task on Transliterated Search. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE '14)*. ACM, New York, NY, USA, 48–53. DOI: http://dx.doi.org/10.1145/2824864.2824872

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *CoRR* abs/1409.3215 (2014). http://arxiv.org/abs/1409.3215

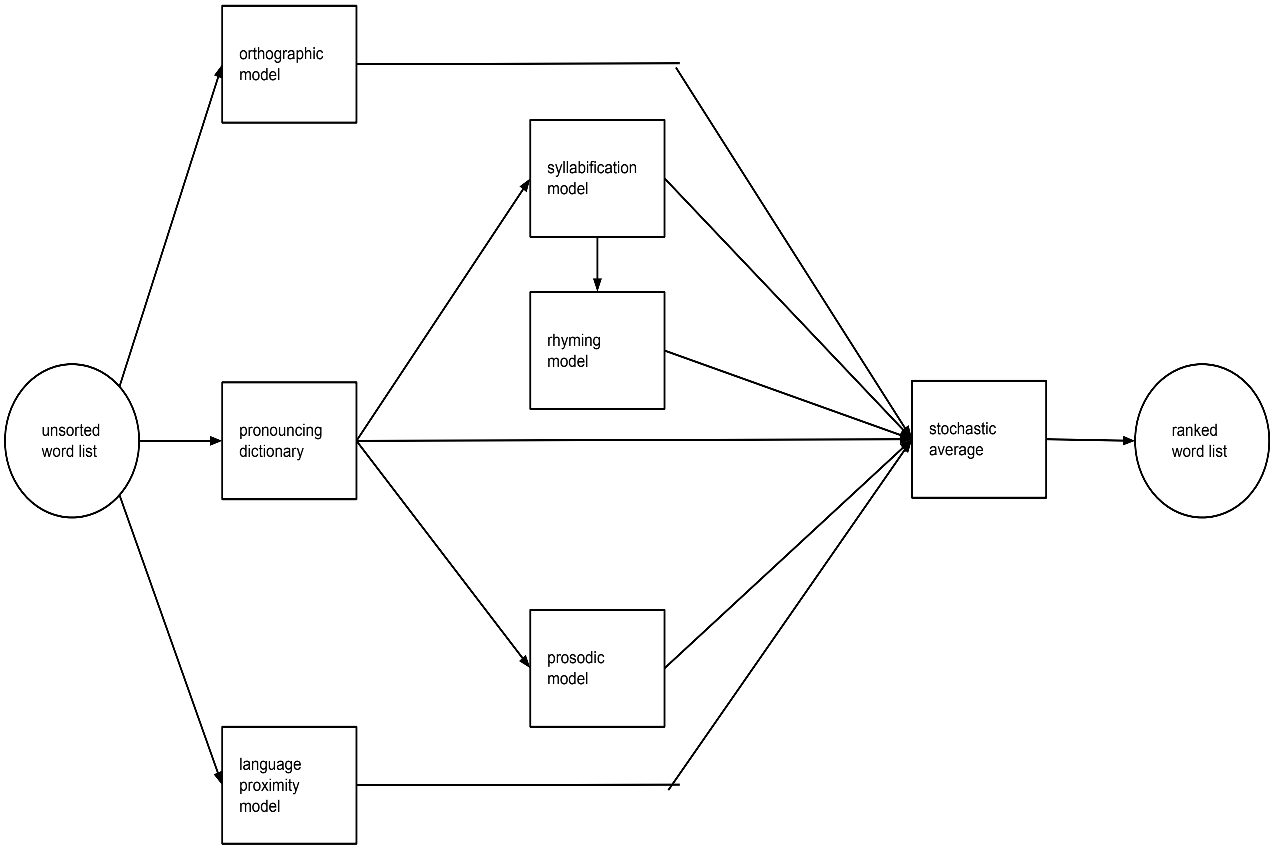

Fig. 1. Pipeline for the proposed metric: The transliterated unsorted word list is given to 6 different modules for log-likelihood and probability computation. The final ranking is based on the stochastic average compted over all scores.

White Planet Technologies. 2015. Hindi ki Duniya Essays in Hindi. (2015). http://www.hindikiduniya.com/