Hunting Elusive English in Hinglish and Benglish Text: Unfolding Challenges and Remedies

Subhash Chandra, Bibekananda Kundu and Sanjay Kumar Choudhury
Language Technology, ICT & Services
Centre for Development of Advanced Computing (CDAC), Kolkata, India
E-mail: {subhash.chandra, bibekananda.kundu, sanjay.choudhury}@cdac.in

Abstract

Language mixing is highly observed in Computer Mediated Informal Communication (CMIC) and arising new challenges for Natural Language Processing. In this paper we have analyzed the reasons of language mixing and its characteristics. The paper focuses on the mixed language called Benglish and Hinglish which are actually fusion of English with Bangla and Hindi language. The major goal of this research is to propose a methodology for extracting English words written in Bangla script from Benglish Text. A hybrid approach combining rule based and statistical methods has been proposed here. When tested on 9152 Benglish sentences containing 13795 unique mixed words collected from CMIC, the proposed approach yielded an accuracy of 95.96% comparatively higher than 83.67% and 54.70% achieved by rule based and statistical approach respectively.

1 Introduction

A city dweller naturally wonders to hear the conversation of youth of cities because of their mixed conversational language. From educational institute to entertainment media, the mixing of languages is propelling every day. In linguistics, this phenomenon is formally known as Code-Mixing (CM) and Code-Switching (CS). These terms are also used interchangeably in the relevant literature (Bhatt, 1997).

Fusion of English with Indian languages Bangla and Hindi evolves new mixed languages that are known as Benglish (Kundu and Chandra, 2012) and Hinglish (Sinha and Thakur, 2005) simultaneously. In popular radio channel often we hear a Benglish sentence like: 'লিসেনার্স আপনাদের জন্য আমরা এখন যে সংগীত টেক্স করেছি, তা হল নির্দোহ কিউকু একটি মোমালিক সংগীত যেটা গেয়েছেন...' [ITRANS: ‘lisenaarsa aapanadera jany aamaraa ekhana Je sa.nga-TaTa ple karabo, taa hala riYeli ki_uTa ekaTa romaantika sa.nga JeTaTa geYechhena...’] [English: Listeners, right now we will play a song for you that is really a cute romantic song sung by...]. With the invention of new products, apps and services, new foreign words are getting assimilated in native languages that cannot be replaced with any similar words in that native language. Certain words like- ‘e-mail’, ‘SMS’, ‘Chat’, ‘Bus’, ‘Car’, etc. are always used as it is (but with transliterated form in native script). Mixing of languages is highly observed in Computer Mediated Informal Communication (CMIC). CMIC follows its own language and culture (Thorne, 2008).

Our research primarily focuses on Benglish and Hinglish languages due to popularity of Bangla and Hindi.

2 Overview

CM refers to the mixing of various linguistic units (morphemes, words, modifiers, phrases and clauses) primarily from two participating grammatical systems within a sentence (Bhatia and Ritchie, 1996). Now we are formally defining the mixed language. Let L(M) is a language of mixed sentence, L(P) is the primary language and L(S) is the secondary language. G(P) is the primary grammar and G(S) is the secondary grammar. A ‘mixed’ sentence S is not a sentence of either L(P) or L(S) but contains lexical items from both L(P) and L(S).

Most studies in CM of Indian languages have not been done yet. The most significant work on language mixing for Indian languages has been initiated by Joshi (1982). Thereafter, only a few

---

1 'where’s the party, yaar?’, ‘bahut tension hai bhuat’, ‘adjust kiiye’, ‘koi seat hai kya?’ etc. are the examples of Hinglish sentences used by youngsters.

2 http://www.ethnologue.com/statistics/size
like Kapoor and Gupta (1991), Bhatt (1997), Sinha and Thakur (2005), Kanthimathi (2009), Sridhar (2009), Bhattacharja (2010), Das and Bandypadhyay (2011), Kundu and Chandra (2012), etc. are worthy to mention.

2.1 Linguistic Patterns in Code-mixing

After manual inspection and statistical analysis of Benglish text (Kundu and Chandra, 2012), we observed following CM patterns. Similar patterns are also observed in Hinglish text (Sinha and Thakur, 2005; Goyal et al., 2003):

i. English root words written in mixed language text using Roman script. For example, ‘friendship’, ‘project’ etc.

ii. English root words appeared in mixed language text with L(P) morphological suffixes and written in transliterated form in L(P) script, for example [ITRANS: phrenDake, kaanTriTaake] etc. are the mixed words while फ्रेंड, कान्तिता [English: friend, Country] are English words agglutinated with Bangla suffix क, टाक [ITRANS: ke, Taake] respectively.

iii. In mixed sentence most English root verbs appear followed by light verb of L(P). In Benglish, English root verbs appear followed by the Bangla root verb ह [ITRANS: ha] and कर [ITRANS: kara] with Bangla verbal inflation. For example:

Benglish: लेखारा शेषे इंकारी पोटी कोन।

ITRANS: lekhaara sheShe koDaTi peShTa karuna.

English: Paste the code at the end of write up.

Similar phenomenon can also be observed in Hinglish (Goyal et al., 2003; Sinha and Thakur, 2005).

iv. When English words or phrases get mixed in Hindi or Bangla sentences, they maintain the syntactic structure of L(P). Figure 1 illustrates this with an example of Hinglish sentence ‘गवमᱸ᭛ट गवमᱸ᭛ट गवमᱸ᭛ट गवमᱸ᭛ट ने जच᭹᭛ट, िच᭨ᮟेन, िमड-डे, मील, सरकारी िव᳒ालयきっかけ मᱶ िलये’ [ITRANS: gavarmeNTa ne ciDrena ke liye skUla meM miDa-De mlla sTArTa kiya’] [English: ‘Government has started mid-day meal for children in school.’] as shown in Figure 1. Hindi words corresponding to its English words shown in dotted box.

v. Introduction of foreign words might change the gender specificity of the subject and/or object of a particular sentence. For example in Hinglish sentence, सरकारी िव᳒ालय滢 in L(P) script, for example गवमᱸ᭛ट गवमᱸ᭛ट गवमᱸ᭛ट गवमᱸ᭛ट  ने जच᭹᭛ट िलये [ITRANS: saraKArI vidyAlayoM meM shikShakoM kl selarI nahl mill ] [English: Teachers did not receive salary in the government schools.]

vi. In the government schools [ITRANS: selarI] [English: Salary] is used as feminine gender while same word when used in pure Hindi (केवल [ITRANS: vetana]) then their gender is Masculine.

vi. English pronouns, articles/determinants, prepositions, quantifiers, possessives etc. generally do not mixed in Hindi (Sinha and Thakur, 2005) and Bangla text (Kundu and Chandra, 2012). In the following Hinglish example, # marked words never mixed with Hindi: आई # ने माई कालेज की कैटीन इन टेबल आई िव᳒ालय िकया। [ITRANS: A# ne mAi# kA.cleja kl kaiMTIIna ina# Tebala A.cn# lanca kiyA.] [English: I have taken lunch on the table of my college canteen]. However, following mixing is allowable like: मैं # मेरे कालेज की कैटीन में टेबल पर िव᳒ालय िकया। [ITRANS: maine mere kA.cleja kl kaiMTIIna meM Tebala para lanca kiyA.] [English: I have taken lunch on the table of my college canteen]. Similar CM patterns are also found in mixed texts of other Indian languages (Joshi, 1982; Kanthimathi, 2009; Sridhar, 2009).
3 Challenges in Computational Analysis of Code Mixed Sentences

3.1 Challenges in Machine Translation
The mixed language poses a new challenge to Machine Translation (MT) system (Sinha and Thakur, 2005). Detection of foreign words is essential for good quality MT where given input is a mixed sentence. Preprocessing is required for transforming the mixed code to its source language before feeding it to the MT system (Sinha and Thakur, 2005). Table 1 shows that better translation can be achieved after preprocessing of Hinglish and Benglish sentences before sending to Google Translation System (http://translate.google.co.in) and Bing Translation system (http://www.bing.com/translator). This result clearly shows how the performance of a MT system is affected by CM.

3.2 Challenges in Part of Speech (POS) Tagging
Most foreign words used in mixed texts are Out-of-Vocabulary words of the source language, which increase difficulties during POS tagging (Zhao et al., 2012). The standard POS tags for Indian languages (Bharati et al., 2006) and English (Marcus et al., 2004) are different. Now, foreign words in mixed sentences are mostly tagged as Unknown Words (UNK) which is not helpful for syntactic and semantic analysis of the entire mixed text.

3.3 Challenges in Information Retrieval
Generally, an interactive Information Retrieval (IR) system includes query construction and relevant document searching using the given text query (Qu et al., 2000). IR from documents of mixed language potentially increases difficulty. Quarry in mixed language also poses difficulties in Cross-Lingual Information Retrieval (CLIR) where quarry need to be translated into one of the language before providing to CLIR system for getting relevant information.

3.4 Ambiguities in Mixed Words
However, detection of foreign words/phrases is very crucial in NLP. Word sense ambiguity is prevalent in almost all natural languages, where large number of words in any given language carrying more than one meaning (Banea and Mihalcea, 2011). It is observed that some specific English words after transliteration create valid Bangla or Hindi words which are shown in Table 2 and 3. Therefore, such ambiguities pose challenges in English word detection in Hinglish and Benglish text. Moreover, transliteration of English words sometimes produce multiple meanings when mixed in Hindi and Bangla sentence.

4 Methodology
We have proposed a hybrid approach combining rule based and statistical based for detection of English words (written in Bangla script) in Benglish text collected from CMIC. Initially we have applied our methodology on Benglish text. The proposed methodology can also be applied on Hinglish text after modification of some rule-patterns. After manual introspection of the sentences of CMIC, we have extracted some linguistic patterns for detection of English words in Benglish text.
The score of a root word is estimated using the following formula:

\[ \phi_1 = \lambda_1 \ast EPLM_{Score}(Ph) + \lambda_2 \ast EGLM_{Score}(Gr) \]

\[ \phi_2 = \lambda_1 \ast BPLM_{Score}(Ph) + \lambda_2 \ast BGLM_{Score}(Gr) \]

Where \( \phi_1 \) and \( \phi_2 \) are the score of a root word being English word and Bangla word respectively. 

\( EPLM_{Score} \) and \( EGLM_{Score} \) estimates the joint probability using trigram language model of the phoneme sequences (Kundu and Chandra, 2012; Basu et al., 2009) and grapheme sequences of a given root word being English word respectively. Similarly, \( BPLM_{Score} \) and \( BGLM_{Score} \) are defined as above. \( Ph \) represents phoneme sequenc-
es and $Gr$ represents grapheme sequences of the given root word. $\lambda_1$ and $\lambda_2$ represents the weights given to individual language model and their values always lies between 0 to 1.

<table>
<thead>
<tr>
<th>English Word Detection Models</th>
<th>Rule Based</th>
<th>Statistical Based</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique words in test data</td>
<td>13795</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique English words in Bangla script</td>
<td>3246</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique words in roman script</td>
<td>822</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique Bangla words</td>
<td>9727</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positive (TP)</td>
<td>1136</td>
<td>2698</td>
<td>2276</td>
</tr>
<tr>
<td>False Positive (FP)</td>
<td>8</td>
<td>5328</td>
<td>5</td>
</tr>
<tr>
<td>True Negative (TN)</td>
<td>9719</td>
<td>4399</td>
<td>9722</td>
</tr>
<tr>
<td>False Negative (FN)</td>
<td>2110</td>
<td>548</td>
<td>500</td>
</tr>
<tr>
<td>Accuracy</td>
<td>83.67%</td>
<td>54.70%</td>
<td>95.96%</td>
</tr>
</tbody>
</table>

Table 7: Evaluation results of English word detection models

![Figure 2: Accuracies of rule based, statistical based and Hybrid Models](image_url)

5 Results and Discussion

The proposed methodology has been evaluated on a corpus of 9152 sentences with 13795 unique words collected from social networking websites, blogs, e-newspapers, online tutorials etc. This corpus contains 822 numbers (5.95%) of unique English words written in Roman form (e.g. Armature, Tool, Photoshop etc.) and 3246 numbers (23.53%) of unique English words written in Bangla script (Transliterated form e.g. ক্লিক, সিলিক্ট, কাউটআউট, টেক্সট, ইফেক্ট etc. [ITRANS: klika, silekTa, kaaTa_aa_uTa, TeksaTa, iphekTa], [English: Click, select, cut-out, text, effect]). The detection of English words written in Roman form is not a difficult task. One can easily detect it using a regular expression like “[a-zA-Z]+”. The main challenge of our research was to detect the English words written in Bangla script. Initially a statistical model, reported in Kundu and Chandra (2012), has been applied on the test data to detect such words in Bengali text. Thereafter, our proposed hybrid model has been used to see the improvement.

Experimental result shown in Table 7 and Figure 2, revealed that the proposed hybrid approach detected English words from these Bengali sentences with 95.96% accuracy. Applying only the Rule based and Statistical models, the accuracies are 83.67% and 54.70% respectively on the same data. In this table, TP means number of detected English words which are actually English. FP means number of Bangla words wrongly detected as English. TN means number of detected Bangla words which are actually Bangla. Similarly, FN means number of English words which are wrongly detected as Bangla.

It is observed that some specific English words written in Bangla script create appropriate Bangla words as shown in Table 2. Therefore, such ambiguities pose challenges in English word detection in Bengali text. It is also observed that there exists some English words that are adopted (Mostafa and Jamila, 2012) in Bangla language as it is like মাইনা, লোকা, পুলিশ, সুপারিটি etc. [ITRANS: glasha, lokaal, pulisha, propaarTi] [English: glass, local, police, property]. Therefore, we need to devise a novel methodology that can automatically classify such English adopted words from rest of the English words mixed in the Bengali text. Frequency based measure may be helpful in such situation. The most frequent English words written in Bengali text can be considered as adopted English words. However, more investigation is required to conclude this hypothesis.

6 Conclusion and Future Direction

This paper presents a brief overview on CM and discussed the possible reasons of mixing of languages. Linguistic patterns in mixed languages have been discussed briefly. A hybrid approach
for automatic detection of English words in Benglish text has also been discussed here. Challenges involve in NLP due to CM have been illustrated. As a future work, we would like to use ensemble classifier (Opitz and Maclin, 1999) combining CRF (Lafferty et. al., 2001) and SVM (Cortes and Vapnik, 1995) to detect ambiguous words (as shown in Table 2 and 3). A study need to be carried out to find out minute linguistic features and contextual evidences to resolve such ambiguities at the time of detecting English words in Benglish and Hinglish text. We also interested to investigate the possibilities to classify the adopted English words from other English words mixed in Benglish text.

References


