POS Tagging of Hindi-English Code Mixed Text from Social Media

Ajita Singh, Amit Kanskar

Abstract: *Language is way of expressing ideas and feelings using movement, symbol and sounds; particular style of speaking and writing. Language is divided into two, spoken language and written language. Spoken language is a form of communication in which words derived from a large vocabulary (usually at 10.000) together with a diverse variety of names are uttered through or with the mouth, while written language is the representation of a language by means of a writing system. Hundreds of millions people in the world routinely use two or more languages in their daily lives (multilingual). Social media is the social interaction among people in which they treat, share information and ideas in virtual communities and networks. One of social media features that are updated any time by users is status. Through status, the user can inform all activity, news, opinions, exchange ideas, business, and so on. In addition, they also are able to comment or respond to the latest status of their fellow social media users. The user of the social media sometimes mixes and uses several languages to update their status or comment to their friends' status, for example when they chat with other people at facebook or web chat. Information retrieval deals with the issues of storing and retrieving information from all types of resources inlcuding social media which is very tough with regard to tokenizing and text processing.*

Keywords: Multilingual, virtual, vocabulary, tokenizing

1. Introduction

An essential prerequisite for any kind of automatic text processing is to be able to identify the language in which a specific segment is written. Longer documents tend to have fewer code-switching points, caused by loan words or author shifts where as shorter texts and status, messages in social media tend to have much more code switching. The code-mixing addressed here is more difficult and novel:. By the work of **Amitava Das** (*University of North Texas Norwegian University of Science and Technology Denton, Texas, USA*) we come to understand that social media text have phonetic text, transliterated text and also spellings created by the author at their own. Code switching and Mixing is under study since 1964 **(Gumperz, 1964; Auer, 1984; MyersScotton, 1993; Danet and Herring, 2007;Cardenas-Claros and Isharyanti, 2009)**but as it is researched we found that code mixing exists between each language inspite of our thought that English is the most used language but now it is not true in social text now a days. India as a country case of have several spoken languages and Hindi is our National Language so most people use it and English alternatively in the social media. Nevertheless, CM on social media has not been studied from a computational aspect. Moreover, social media content presents additional challenges due to contractions, non-standard spellings and non grammatical constructions. Furthermore, for languages written in scripts other than Roman, like Hindi, Bangla, Japanese, Chinese and Arabic, Roman transliterations are typically used for representing the words Since the earliest works on POS tagging use datasets which are either owned by the researchers themselves to confine to a particular research ground and the datasets are not readily available for download and experimentation. It poses various difficulties such as comparisons of the present works to the past works and the validation of the researchers' claims about the POS tagging results.

The above mentioned facts motivate us to design a new POS tagger for Hinglish which will accept a combination of English and Hindi text (typed in English or Hindi font) to produce a universal tagged output that can be directly used for other NLP applications.

Overview of Code Mixing in Social Media

Code–mixing involves mixing of:
A) Two languages
B) Two or more languages
C) Two or more diglossic varieties

There are two types of code–mixing:
a. Intra-sentential code–mixing
b. Inter-sentential code–mixing

Intra sentential code-mixing: Intra-sentential code–mixing takes place within sentence or clause or word boundaries with no apparent change in topic. An Intra-sentential code–mixing can occur in three processes:-

Firstly, noun insertion: This code mixing involves the mixing of noun in one code in a sentence which is another code.

Secondly, Verb insertion: This code mixing involves the mixing of verb in a sentence from one code and remaining parts are from another code.

And finally, Clause and sentence insertion: complex sentence with different clause of another code reflects this code-mixing.

Inter-sentential code-mixing: Inter sentential code-mixing takes place if one informant is asked to describe an incident she/he begins to describe it in her/his standard language and switches to his second language and then completes it in his local variety. Code-switching and code-mixing occur in all linguistic situations. These two occur frequently in multilingual, bilingual and even in monolingual countries.

Although during post Independence period (after 1947) English was considered as a foreign language in India and

was proposed that in due course of 15 years English will not be used. But in spite of tries by central Government in 1967 it was not met as several protest by states some of which don't use Hindi as their speaking language recognized English as their second language and now it is so in overall INDIA. Now it is considered as a general communication language. Officially English is recognized by the government as a second language, a language for the people pursuing higher education and social status. Even as a foreign language, English in India enjoys a special importance even in the educational policy by the government.

Officially English may not enjoy the status of an official language for communication, but it gradually became a very important language in INDIA. Though English is not used widely in interpersonal and social communication, as concerns academic and job placement, it is the most demanded and sought after language. English is very important in some areas- in judiciary at higher level, education and in international business. With the growth of globalization and technological advancement, the scope and intensity of communication in English is getting radically transformed. From urban educated cities of English or now in social context Hinglish became a tool for social advancement and development. **Spolsky (1998:45)** commented that, "…..it is common that people develop some knowledge and ability in a second language and so become bilingual.

2. Related Work

There are some Reasons for code-mixing
1) To draw the attention of others
2) To impress the others
3) The medium of education
4) To dominate other psychologically
5) Insufficient English words
6) To show off
7) To show smartness
8) To express feelings more easily and comfortably

There are various previous work been done in the field of code mixing in some of the various Indian Language such as Punjabi, Tamil, Bengali and also Hindi.

In this paper [1] we get noticed that POS taggers more or less receive 96+% performance on English news text with just about any method, with state-of-the-art systems going beyond the 97% point on the English Wall Street Journal corpus: Spoustová et al. (2009) report achieving an accuracy of 97.43% by combining rule-based and statistically induced taggers.

The first attempts at applying machine learning approaches to code-mixed language were by Solorio and Liu (2008a) who aimed to predict potential code alternation points, as a first step in the development of more accurate methods for processing code-mixed English-Spanish data. Only a few researchers have tried to tag code-mixed social media text: Solorio and Liu (2008b) addressed English-Spanish, while the **English-Hindi mix was previously discussed by Vyas et al. (2014).***According to this paper in future there are*

several possible avenues that could be further explored on NLP for code-mixed texts, for example, transliteration, utterance boundary detection, language identification, and parsing.

In this paper [2] the study of text is defined with basic details such as

Excerpt1: Example of code switching English/Spanish
A: The picture looks so cool.
B: Which picture?
A: The one you have in your messenger.
B: Ah…Si, me gusto mucho. (Ah…Yes, I liked it a lot.)
Excerpt 1 shows how participant B interacted in English during most of the conversation and suddenly switched into Spanish.

Excerpt2: Example of insertion (Spanish/English)
B: Pero bueno creo que basta con que incluya la pregunta de enhanced output más todas las demás.

(Well, I think it is enough if I just include the question of enhanced output.).

It describes that "**communication that takes place between human beings via the instrumentality of computers**" synchronous and asynchronous. Synchronous communication or interaction that takes place in real time via relay chats, chat rooms, instant messaging, voip sand tweets; asynchronous communication, or interaction that allows (Computer Mediated Communication) users to access the media at a different time includes emails, blogs, and wikis among others. *According to this paper in future "Further research comparing code switching and code mixing occurrences between genders and age groups is needed to better understand these phenomena of code mixing"*

In this paper [3] it describes a POS tagging system for code mixed social media text in Indian Languages. The features such as dictionary based information and some other word level features have been introduced into the HMM model. The paper have used only verb, pronoun and conjunctions in the dictionaries, so meta tag(XXXX) can take one three values: VERB, PNON and CONJ. The POS tagger have been tested for three types of data and results provided below

Table 1: The description of the data for various language pairs

Language	Total of sentences	
	Training data	Test data
Bengali-English	2837	1459
Hindi-English	729	377
Tamil-English	639	279

According to this paper in future "more number of Tags can be used to get a more accurate result on code mixing in a specialized language say Hindi-English"

In this paper [4] it describes the use of Fuzzy Logic to verify a specific form of Syntax in Natural Language processing. It

tries to give a percentage based on each sentence meaning and if it is 100 % then the statement is accepted otherwise rejected. The objective of a natural language understanding unit is to extract all the information from an utterance that is relevant for a specific application. A Fuzzy context-free grammar can be used to help specify the syntax of a programming language. In addition, if the grammar is designed carefully, much of the semantics of the language can be related to the rules of the grammar. This concept of detection can be used in this paper.

In this paper [5] it describes the issue of Tagging and parsing a text Modelling a linguistic structure is the primary task of a parser, which uses a set of rules and grammar. Parsers check the syntax, break it into smaller elements and align the words according to pre-defined rules. Most of the work done in the realm of Natural Language Parsing Systems focuses on a single language [1]. It describes the English sentence structure and the Hindi sentence structure both describing its capabilities.

The word order in English follows the **SVO** model, e.g., "A cat eats mice."Here, the sentence (S) consists of an initial noun phrase (NP) and a verb phrase (VP) as depicted in Figure 1.
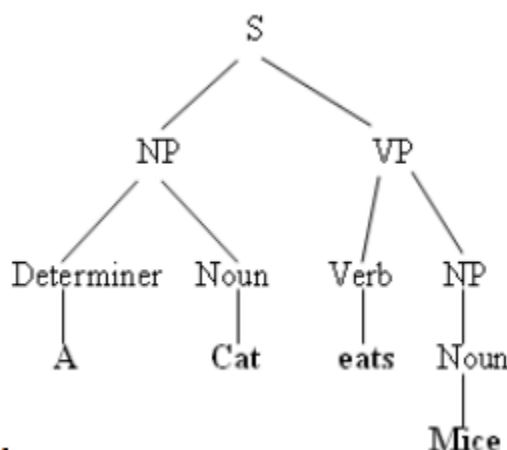

Figure 1: Parse tree of "A cat eats Mice"

In this example, "A Cat" is the Subject, "eats" is the Verb and "Mice" is the Object.

Hindi Syntax Model
The Hindi language has a free word order i.e. Subject Object Verb (**SOV**) e.g. "Billi chuhe khaati hai",
Noun (Subject) →Billi
Noun (Object) →Chuhe
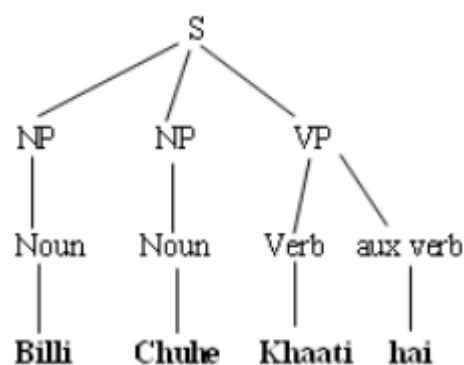Verb Phrase →Khaati hai

Figure 2, hence

Figure 2: Parse tree of "Billi Chuhe Khaati hai"

3. Proposed Work

In this section, we describe the mechanism of enhancing the method of detection of language of a word in code mixing and then using rule and dictionary based approach to get the Tags of the words in a sentence. The proposed work has taken into account both transliterated and phonetic text. Transliterated text is detected as it is in UTF format and then detected based on the phonetic info whether a word is in Hindi or it is in English as English words can't be trans literally used. A new dataset has been created, which contains Facebook posts and comments that exhibit code mixing between English and Hindi. Also present are some preliminary word-level language identification rules using this dataset.

Different techniques have been employed. For implementation a simple unsupervised dictionary-based approach, supervised word-level classification with and without contextual clues, and also sequence labelling using Conditional Random Fields.
Studies have previously indicated that the dictionary-based approach is somewhat minor to supervised classification and sequence labelling approach but can be better if dictionary is large. Also it is important to take contextual clues into consideration.

A simple language detection based heuristic needs to be employed where first the text can be divided into chunks of tokens belonging to a language, and then each chunk can then be categorized according to its language and further tagged by the POS tagger for that language. Language detection and transliteration text is tagged through an English monolingual tagger and then choosing one of the two tags for a word based on some heuristics that was detected by several language detection techniques. . *In our data the Hindi tokens are phonetically as well as Transliterally typed. As no such transliterated dictionary is, to our knowledge, available for Hindi, so the training set words as dictionaries have been implemented and taken into account. For words that have multiple meaning in training data (**ambiguous words**), selection of the majority tag is based on frequency, e.g. the word "**to**" will always be tagged as English. Our English dictionaries are those described in (**WORDNET**) and also the training set words. For others where no frequency information is available, considered it as a simple word list.*

4. Conclusion

As users of social media come from various different parts of the society so they tend to use their own set of words as well as own set of phoenetic Hinglish which needs to be more accurately detected to get the meaningful word from those tokens. This will increase the accuracy to more specific domains. This paper will not only enrich the Hinglish dictionary but will also help in getting more accurate POS tags as some major tags have been Enhanced which was not used earlier. We, no doubt need to study a lots of issues of phonetic words and consider their resolutions in future.

References

[1] Part-of-Speech Tagging for Code-Mixed English-Hindi Twitter and Facebook Chat Messages by Anupam Jamatia, IEEE 2013

[2] Mónica Stella Cárdenas-Claros, University of Melbourne on Code switching and code mixing in Internet chatting: between 'yes', 'ya', and 'si' a case study , ISSN1832-4215 Vol. 5 , The jalt callJournal 2009.

[3] Sarkar, K. and Gayen, V., 2012, November. A practical part-of-speech tagger for Bengali. In Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on (pp. 36-40). IEEE

[4] S.G.Kanakarddi and V.Ramaswamy,"Natural language parsing using Fuzzy Simple LR (FSLR) parser," in Proceedings of the IEEE international conference of Advance Computing Conference (IACC), Feb 21-22, 2014.

[5] T.V. Harsh Prasad and G Rama Krishna," Issues in parsing and POS tagging of hybrid language", in Proceedings of the IEEE international conference of Computational Intelligence and Cybernetics (Cybernetics Com) Feb 12-14, 2012.