DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4847

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

WHY DO PEOPLE CODE SWITCH?A BRIEF ANALYSIS OF CHALLENGES FACED IN SOCIAL MEDIA TEXT

Neetika
Assistant Professor, MCA Department
Punjabi University College of Engineering & Management
Rampura Phul, Bathinda, Punjab

Sumanpreet Kaur
Assistant Professor, ECE Department
Punjabi University College of Engineering & Management
Rampura Phul, Bathinda, Punjab

Dr. Sunita Rani
Assistant Professor, Applied Sciences Department
Punjabi University College of Engineering & Management
Rampura Phul, Bathinda, Punjab

Abstract: In a bilingual or multilingual society people interact with each other using more than one language. Due to various language factors a situation arises which is known as code mixing or code switching. The capacity of code mixing increases with the number of languages learned by the user. It is commonly used in every area whether marketing, advertising, film industry, teaching etc. Nowadays social media has become the strongest platform for communication. There are numerous applications like Facebook, Twitter, Whatsapp, Skype and Instagram etc. which are being used for communication. The main focus of paper is on understanding of code mixed social media text in English and Punjabi.

Keywords: Romanized text, code mixing, code switching, language identification, normalization

I. INTRODUCTION

The revolutionary induce of social media has given a new perspective to cultural enhancements as people of different cultures can understand and communicate with one another in a light manner. Human being intelligible has great adaptability and skill to communicate with one another either through gestures or on some linguistic agreements. Worldwide English is still by far the most popular language in SMC, its dominance is receding**.** Only half of the tweets were found in English from 62 million tweets. They have developed an automatic language detection algorithm to identify the top 10 most popular languages on Twitter [1].

In India there are about 500 spoken languages (over 1600 including different dialects) and 22 languages being declared officially. English and Hindi are commonly used for communication, but English dominates as other languages being used in Romanized form as few people use UNICODE. Such text is called as Romanized Text. Code Mixing commonly prevails over short geospatial distances. India comprises of people from various regions, sub divisions, religions, castes etc. that's why code mixing so commonly used in India.

Language identification is an essential prerequisite for automatic text processing. It has been considered as almost a solved problem for monolingual text in which n-gram approaches, character encoding detection or stop word lists can reach up to 100% accuracy. There has been a strong insight into various models and tokenization used and deduced that LID becomes difficult with no. of languages and reduction of training data and length of text [2]. There are two main challenges faced in Language Identification. First one is Language annotation. When the mixed languages are closely related either linguistically or it becomes difficult to decide the language ID of a particular token. For English-Hindi language e.g. "Sat shri akal, my dear friends". It is difficult to decide language of sat (means greeting in Punjabi or verb form of sit or Saturday in short). Second main challenge of Language Identification is mixing of two languages inside a word. Such mixed words are treated differently among researchers. e.g. the word confusa is a mixture of English and Punjabi language.

Text normalization is the task of standardizing text that deviates from some agreed-upon (or canonical) form. In Social media text people write their native languages in Roman (English) Script rather than using their conventional scripts. Machine translation is considered consuming and phonetic translation occupies too much memory. So a balance between two can be used to solve the problem. Various issues that need to be addressed during normalization are slangs, acronyms and short forms, omission of punctuation marks or stylish use of punctuation marks, phonetic spelling, misspelling etc. Repeating letters or punctuation for emphasizing and emotional expressions is used e.g.. "goooodmorniiing". Some people use phonetic spelling in a generalized way or to reflect a local accent. e.g. "wuz up bro"(what is up brother). Substituting phonetically similar letters like phone as (fon) and substituting numbers for letters like 4get" (forget), "2morrow" (tomorrow), and "b4" (before)are frequently used in social media. Slang abbreviations which abbreviate multi-word expressions like LMS (Like My Status), rofl (rolling on floor laughing) etc. are also common.

Social media content has been assigned with many names like bad language, noisy or informal. People don't always use UNICODE they mix multiple languages.

II. SOCIAL MEDIA TEXT

A. Bilingualism/Multilingualism

Indians are strongly bilingual or multilingual in Social Media Content due to “The Three Language Formula” implemented by Central Advisory Board on Education in 1956. The Board (CABE) devised the three-language formula in its 23rd meeting held in 1956 with a view to removing inequalities among the languages of India [3].

Bilingualism example: “So I thought maybe *tu mainu yaad kar rahi hoyengi*” (English and Punjabi are being used).
Multilingualism example: guys’ wuz up? Chalo *ghumne chalen* Punjabi *mele ch*” (English, Hindi and Punjabi are being used).

B. Borrowing

Borrowing means inserting a foreign language word into native language. It has been difficult to define when borrowings/ Anglicism ends stop and code-mixing starts [4]. Code Switching and Borrowing are found to be universally related [5]. According to her Matrix Language Frame (MLF) Model, CS occurs everywhere within a frame which is set by the matrix language. There are many cases in which there is no clear distinction between borrowing, anglicism, code mixing or code switching. e.g.. One such case is as follows:
e.g.. “*sare* artists *nu bulayaa hai*”(all artists have been called),
e.g.. “*sare* artist *kal aaonge*”(all artists will come tomorrow)

C. Code Switching

Code Switching means starting with one language and then switching onto other. In context to the work while communicating one starts with English language and then switches to Punjabi language. e.g. “I was going to the market, suddenly *ik cycle wale ne picchhon takkar mari”*. Other way, one starts with Punjabi language and then switches to English language. e.g.. “Rohit *ne roti khadhi* and he started vomiting”. Code switching has been categorized as inter- vs. intra-sentential (outside or inside sentence or clause boundaries), intra-word vs. tag switching (within a word, at morpheme boundary, or by inserting a tag phrase or word from one language into another). There are four types of code switching. In Intra-sentential Switching change occurs within a clause or sentence boundary, within a clause or sentence, with no interruptions, hesitations or pauses. In Inter-sentential Switching change occurs at a clause or sentence boundary, where each clause or sentence is in one language or the other. It is most often between fluent bilingual speakers and it becomes difficult to define matrix language and embedded language. In Intra-word switching change occurs when change is within a word boundary. It is similar to Word Level Code Mixing. In Tag-switching or Extra-sentential Switching change occurs when certain set phrases in one language are inserted into an utterance from another language, similar to intra-code sentential switching e.g.. insertion of words like but(*per*) , I mean(*matlab*) , then(phir), hana(isn’t it) etc.

There is an exceptional case with 129 intra-sentential language switches for spoken Spanish-English corpus [6]. It has been declared that Inter-Sentential Code Switching was dependent upon social motivations [7]. It has been investigated that in facebook posts inter-sentential switching (59%) predominates over intra-sentential (33%) and tag switching (8%). Also 45% of the switching was due to lexical needs, 40% for talking about a particular topic, and 5% for content clarification [8].

D. Code Mixing

Code Mixing means using two languages in an utterance or mixing words of two languages. In Intra-utterance Code Mixing, mixing takes place in a single sentence or utterance as the speaker is proficient in both languages. e.g. “*Us kudi de* cardigan *da* color *kinna* nice *hai”*. *In* Inter-utterance Code Mixing, mixing occurs when one changes from one language to another in the same conversation/utterance. e.g.“*main apne parivar naal haan* and mi husband”. Word Level Code Mixing is the smallest unit of code mixing. It captures intra-word code mixing and includes cases where code mixing has occurred within a single word. e.g. *samosas, bigdofy, computeran* etc.

III. CHALLENGES IN SOCIAL MEDIA TEXT

Automatic understanding of social media text becomes a difficult task as multilingual / bilingual speaker frequently switch between languages due to a number of reasons like emerging new tends, globalization, style of writing and resulting in wrong spellings or interpretations. The main reasons sorted out for Code Mixing are for role identification , style identification, inability of expression, impact & effective speech, done deliberately to exclude a person from a conversation, choice of topic, ethnic identity, emotional arousal, the topic decides the language, showing off or showing kinship.

A. High Percentage of spelling errors

In communication when one keeps typing at a fast pace or due to auto fill feature in smart phones/ tabs /laptops the chances of typing error increases. Some words are commonly written like media (as mcdia), accept (as acept), flour (as flore), save (as savc) etc.

B. Shortening of words

In tweets or sms people try to write in a very short manner by reducing the number of words and creatively write text in a new manner. The variation in abbreviation of words or phrase depends on user. e. g. gr8, gud nyt, nyc, tc, pic, admin, ppl, lvly, 2moro etc.

C. Prolonged words used intentionally

When users in excitement/emotions express their feelings by repeating the characters within the words they use words like goooood, cutieeeeeeeee, hahaha, soooooooooooooo, rocxxxx, wah wah etc.

D. Phonetic Misspelling (language specific)

Romanized text is spelled phonetically (local language based) but others can understand the meaning of sentence in the same context as initiated. Meaning is well understood by other people who understand the language. e.g.. even if one writes “nahin” as nhi or “phone” as fone.

Phonetic Similarity of Spellings Units
Due to phonetic typing some words share the same surface. e.g.. “PhD *de* Seminar *te* ***main*** *confuse ho gaya si*” (main means I in Punjabi and main means important in English)

E. Multiword Tokens and Abbreviations

Users use multiword tokens in place of multiple words and some common abbreviations. There is a rapid increase in number of multiword tokens day by day. Some examples are "OMG" (oh my God), "RIP" (rest in peace), "LOFR" (Laughing on the Flour Rolling), PM, BJP, US etc. Many abbreviations have different meanings for multiple words at different places. e.g. NLP (as Natural Language Processing or Neuro-Linguistic Programming).

F. Meta tags, URL tags, Hash tags

Some people on social media creatively use emoticons, Meta tags, URL tags, Hash tags in Social Media Text. "SireDaBrand#, #IsOutNow#, #WaheGuruMeharRakhan#, #NeedYourBlessings". Also "ATM *wich cash hai URL* ----- cashnocash.com".

G. Use of song titles, movies and advertisements

Since decades Code Mixing has gained existence in movies, songs, advertisements in order to increase their importance. The main intention behind Code Mixing here is to gain attention of people of particular area. Even TV commercials have started using code mixing to a great extent. e.g.. "Carry on *jatta*", "*Ladki* beautiful *kar gayi .chul*" , Coca Cola, "*Thanda matlab* coca cola", Vicco Turmeric, "Vicco Turmeric, *Nahi* Cosmetic".

H. Ambiguous words (same words across languages)

Several English words are so frequent in Punjabi that they have become an indispensible part of Punjabi language. e.g.. Bus, truck, cycle, scooter, T.V., radio, video, cigarette, diploma, flashback, furniture etc. words like college, department, university, vote, radio, marketing, traffic, school etc. have become a part of language because they have no Punjabi equivalents.

I. Named Entities

Often many names of persons match with movies names, places or objects and such words need to be handled. Such as:
"Mainu nahin lagda Apple *nawan tab release karega"* (brand or fruit).

J. Reduplication

It is quite interesting fact. English and Punjabi both include some reduplicated words. In Punjabi people also use some words two times or with a modification. Such words have become slangs. E.g.. cha-chu, dance-vance, super-duper, kam-vam , balle-balle, chotte-chotte, tap-tap etc.

K. Some Special Cases

There are some special cases which are very difficult to be handled which will hamper the accuracy of the system.e.g.. "just hune" , "center vich *rakho", "gol* round" etc.

IV. CONCLUSION

The basic tasks of NLP related to Code Mixing are normalization, POS Tagging, parsing, language modeling, language identification, machine translation, and automatic speech recognition. Automatic understanding of code mixed Social Media Text can be enhanced by performing all above tasks. The data need to be collected from facebook and whatsapp and Twitter messages, posts, comments etc. Also API Twitter can be used to filter tweets from different users. Automatic understanding of social media content has been one of the strong areas of NLP. Researchers use simple dictionary method or machine learning techniques such as Naïve Bayes, SVM, and CRF, HMM, etc. The main advantage of using dictionary based approach is that annotation becomes easy and full length dictionaries are more preferable to most frequent word list and moreover normalization dictionaries have prove to be a boon for normalization. But the main drawback of using dictionaries is that dictionaries need to be updated again and again and they don't contain distorted words.

V. REFERENCES

[1] L. Hong, G. Convertino, and E. H. Chi, "Language Matters In Twitter: A Large Scale Study," Proc. AAAI weblogs and social media (ICWSM 2011) , The AAAI Press, July 2011, pp. 518–521.

[2] T.Baldwin, M.Lui, "Language identification: The long and the short of the matter," Proc. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,June 2010, pp. 229-237.

[3] R. Meganathan, "Language Policy in Education and the Role of English in India: From Library Language to Language of Empowerment." H. Coleman (Ed.), Dreams and realities: Developing countries and the English language,pp. 2–31. London: The British Council.,2011.

[4] B. Alex,"Automatic detection of English inclusions in mixed-lingual data with an application to parsing." PhD diss., University of Edinburgh, 2007.

[5] C. Myers-Scotton,"Constructing the frame in intrasentential codeswitching." Multilingua-Journal of Cross-Cultural and Interlanguage Communication 11, no. 1 ,1992, pp. 101-128.

[6] T.Solorio, and Y. Liu,"Learning to predict code-switching points." Proc Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 973-981.

[7] H.K. San, "Chinese-English code-switching in blogs by Macao young people." Master's Thesis, The University of Edinburgh, August, 2009.

[8] T.Hidayat, "An analysis of code switching used by facebookers (a case study in a social network site)" , BA Thesis, English Education Study Program, College of Teaching and Education (STKIP), Indonesia, October 2012.