

# An Analysis of POS Tagging English-Hindi Code Mixed Text

Gajala Praveen<sup>1</sup>, Danish Raza Rizvi<sup>2</sup>

<sup>1</sup>Jamia Millia Islamia, New Delhi

<sup>2</sup>Jamia Millia Islamia, New Delhi

(E-mail: [praweengajala@gmail.com](mailto:praweengajala@gmail.com), [drizvi@jmi.ac.in](mailto:drizvi@jmi.ac.in) )

*Abstract*— Social media in today's world possess enormous amount of data. This data is used by various companies for advertisement to suitable group, declaring promotions, etc. But the problem starts in bilingual or multilingual populations where a lot of people tend to use multiple languages in the same sentence. Now analysis of such a text unravels a whole new field of study. Language is way of expressing themselves by movement, symbol and sounds; particular style of speaking and writing. Language is divided into two types namely: spoken language and written language. Hundreds of millions people in the world routinely use two or more languages in their daily lives (multilingual). Information retrieval explains storing and retrieving information from all types of resources including social media which is very tough with regard to tokenizing and text processing. We report our work on Hindi language mixed with English. In this paper we have described our approaches to the Parts of Speech (POS) tagging techniques and exploited for this task. Statistical techniques have been used to POS tag the mixed language text. The data is collected from social media text like messages of chat content. The work is performed on automatic tagged corpora in three phases: language identification, back-transliteration and POS tagging. We show results of specific Trigrams 'n' Tags (TNT) tagger for Hindi language and statistical Hidden Markov Model (HMM) Technique for English language.

*Keywords*— *Bilingual code-mixed text, tokenizing, POS tagging*

## I. INTRODUCTION

One of the fundamental steps of any natural language processing system is Parts of Speech (POS) tagging. A set of word is fed as an input with tagset and word with tagged is obtained as output. There are various POS tagger has been developed for tagged accurately of different language. There are many rules are available for easily tagging the corpora of English languages. A tagger plays an important roles for accurate tagging the text with their tagset of different languages and it passes through the flowing process such as of Chunking, Parsing, Morphological analysis etc. A tagger facilitates in the process of annotated tagset creation. Although automatic POS tagging is a well-defined research paradigm

even there are various efforts in literature for these two Indian languages. In natural language processing, Parts of Speech (POS) tagging is associated with every word in the sentence a lexical tag POS tagging is very useful, because it is usually the first step of many practical tasks.eg: speech synthesis, grammatical parsing and information extraction. There are various technique are used for tagging the text, mainly it is categorized as supervised and unsupervised. In supervised tagging the training data is already tagged on the other hand in unsupervised POS tagging, tagged training data is not required. There are various techniques used in unsupervised model for generate the tagset and generate the transformation rules. In Supervised POS technique, we obtain the probability of tagged word. Now-a-days a machines learning approach is used in NLP system to tagging the corpora which gives better results. If we used Machine learning approaches then taggers can be constructed within less time, and learning curve increases sharply. There are various research has been done using machine learning approaches over POS tagging model. Recently it is common to obtain the tagged output corpora using machine learning based tagger for the different language. If we train any tagger then we explore the base of supervised, unsupervised and semi-supervised approaches in the construction of POS tagger. In the field of NLP many task are uncertain and ambiguity is found at different level of NLP transforming task. As for example the one word having more than one POS tag. The accurate tag of the word based on use of those words in the sentence and also the relationship between the words who assert exact meaning of the word after that accordingly tagged. The text contains various POS ambiguity of the word which is obtained after determining the sentence. The presence of words in the sentence having different tags like the word cat and making may be a noun or verb. Similarly the word can be a preposition, an adjective or an adverb. POS tagging also recognizes the uncertainty by choosing correct tag of particular word in the sentence. This also based on the classification of the POS tagging task. POS tagging mainly based on the use of the words in the given context. It is also called as tagging based on grammatical rule and disambiguation of word. Before tagging it is also observe the connectivity between the words in the phrase, corpora or text. Due to the machine learning approach, POS tagging achieved a most significance in the NLP system. Part-of-Speech (POS) Tagging is the primary step in the development

of any NLP Application. It is the process which assigns POS labels to words supplied in the text. The first step of POS tagging is language identification which is important task in NLP system and performance based on the achieved high accuracy on this step. Analysis of grammatically syntax is done after the language identification. It is the process of assigning the category of language in every word in the text. The POS tagging of Code-mixing text is based on the various things like grammatical mistakes, spelling variation and back-transliteration. The combination of more than two languages is known as code-mixing language. Code mixed language is most frequently used in social media chat and easily writing the short form of the word which refers to the ambiguity or spelling mistakes, in this condition difficult to tag accurately. Making NLP methods for social media text (SMT) has recently received significant attention. Most of the research on SMT till date is concentrated on English therefore making technologies for other languages are as par necessity. Rapid growth in social media instigated enormous possibilities for information extraction research but those emergences would have to face several challenges due to the terse nature of the SMT. India is a nation of languages. India is a land of many languages. There are 500 languages are spoken by the people in India. Among the 500 languages, 30 languages are spoken by the 1 million people. Hindi is the widely spoken language and 4<sup>th</sup> worldwide in terms of first language speaker. Generally the code mixed language is used in social media for conversation to each-other. Hindi is most spoken language in the country. Mostly the highly educated persons have spoken English, Hindi or it may be mixed in their daily life. English is international language and highly communicated language in various places between the educated and uneducated people. social media text having more ambiguous words due to short form of word spelling to express itself, like TX for thanks, K for ok. The Hindi speaker used more than one language including English to write their chat in social media. Even phonetic typing and creative Romanization are added challenges for Indian social media. Therefore making NLP techniques for Indian SMT is far more challenging than English. We have noticed that monolingual Unicode tweets have relatively lower wordplay or spelling errors, therefore empirical question rises how different/difficult this task is than the general (like NEWS) text POS tagging. To answer this question our rationale is tweets are syntactically very different due to the 140-character length restriction. Moreover URL, hashtags, emoticons and unnecessary symbols made this text genre very different from formal text. Even to establish our rationale we have reported performances of general purpose POS system on our tweet data. Most of the Indians and many other Non-English speakers across the world do not use to one language to express themselves. People generally use more than one language including English to write the chat content. English still is the principal language for social media communications, but this kind of multilingual content is growing and calls for the development of language technologies for languages other than English. If we observe twitter and facebook feeds of Indians, it's full of frequent

code-mixing. It's not a surprise given the diverse linguistic culture across India. But this possess additional difficulties for automatic Indian social media text processing. POS tagging of English text are now a quite matured filed in NLP and a lot of work is in progress for English social media text. Many people are utterance code mixed language in their daily life which is combination of more than one language. The mixed content created on social media platforms can be called as (CMST). Code-mixing leads to presence of more than one language in the text and its social nature adds all the complexities mentioned above. Additionally, CMST is being generated at an enormous scale and there is a need to create special NLP tools for it as traditional NLP social content which ensures they will perform poorly on CMST. Mixing of languages is called code mixing. Code mixing occurs due to various reasons. According to a work by [1], "An analysis of code switching used by facebookers: a case study in a social network site", While explaining something, for better clarification of the audience, to make the audience more clear about the topic, code switching is used. A bit older work by [2] said that strong emotional arousal also increases code mixing frequency. As social media contains valuable information, due to the presence of above mentioned type no proper tools that deals with this type of data. The primary reason behind this limitation is due to proper corpus acquisition and there have not been any. This project proposes a model that POS tags the code mixed text which can be used for various tasks in Natural Language Processing. The first step of any NLP text is to recognize the language which is used for written the text. In case of larger data then it contains less code switching on the other hand in smaller dataset having large code switching points. The recognition of code-mixing is not easy: By 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 but as it is researched we found that code mixing exists between each language in spite of our thought that English is the most used language but it is not true in social text now a days. India as a country is case of have several spoken languages and Hindi is our National Language so most people use it and English alternatively in the social media. Part-of-Speech Tagging is a primary necessities of many Natural Language Processing system. It is fundamental step of Natural Language Processing, which has been applicable in various fields like: information extraction, speech recognition, semantic processing, building parse trees, Dialog systems, parsing, machine translation, disambiguate homonyms, text-to-speech processing, natural language parsing and information retrieval system.

The rest of the paper is organized as follows. We present related works that has been done in the part-of-speech tagging in Section 2. In Section 3, we discuss the dataset. The models and experiment has been described in Section 4. The conclusion and future work have been presented in Section 5.

## II. RELATED WORK

Previous work is a body of a text whose main purpose to analysis the prevailing knowledge including searching, as well as analytical and technological improvement of a particular topic. The main aim of previous work is research question, trying to recognize, choose and gather all high-quality research proof and dispute. Parts-of-speech tagging on Indian Social Media Code-Mixed Text is a very incipient research problem in the research of natural language processing (NLP). Indian NLP researchers are working on various issues of Code-Mixed corpora. In this digital era, nowadays inescapable social media (viz. e-mails, tweets, chat, discussion forum, comments, and blogs/microblogs etc.) are part of communication and the 'netizens' are highly innovative and collective to produce text with the help of different language which is mostly utterance by the people represent experience examined. It observed that linguistic forms and its application of verbal terms are highly often seen in the Twitter messages compare to the chats, which is most discussion and gives less minor. The most precedent research in the field of SMT, concentrated mainly on tweeter message due to easy availability while the talky behavior of written text to express themselves in code-mixing language. There has not been much work done in terms of POS tagging of code mixed text. We came across related paper [3]. They used word level language identification using a logistic classifier and to take into account the context they calculate context switching. The first efforts at applying machine learning come for more than one language by [7], which is mainly focus for calculating the hidden repeated points as a first process in the development of most progress of procedure processing of CMST Spanish-English data. An approach was developed by [8], for parts-of-speech tagging of code mixed language of English-Spanish corpora through assembled the conversation and discussion between the three persons and interpreting the record-keeping of spokesperson. They uses the rule based method for tagging the text of the English as well as Spanish mono-lingual tagger and select one tag among the two output tag which is applicable for: The confidence score of POS tagger, The lemma of the words and The language of the word which is identified by various language recognition methods. This method were enlarge the substructure by observing the output of the two singular tagger for two language of tagged word and also observe the various characteristics like language labels, confidence scores also word of those language. There are various methods such as SVM, logic Boost, J48 and Naive bayes were used for their POS tagging performance. The tagging based on machine-learning method was most significant and obtained a word level accurate tagging accuracy approximately 93.5 %, which is 4% improvement of rule based approach of tagging method. Recently [9] explained POS tagging of code-mixed text of English and Hindi language whose main focus to tag the Hindi language text. Some Random Forest based pilot approaches were used in their experimental setup on 400 code-mixed text (all Romanized) from Facebook and Twitter. They achieved 63.5% word level tagging accuracy. On the other hand, if the

Hindi tweets written in Devanagari then authors achieved 87% accuracy on Hindi language text. Hence, it represents the toughness for back-transliteration in code-mixed text. In paper [9], experimental analysis has been performed on POS tagging by gathering information based text of Hindi-English mixed text from social media. Gupta et al. (2014) proposed deep learning methods in the context of code-mixed information retrieval which is recognize the language of the word in code mixed text. More recently, an initial effort is done by [10] which is based on POS tagging related to code-mixed text of two languages Hindi and English. In code-mixed text, many things are required to obtained better accuracy such as ungrammatical text-correction, spelling correction, back-transliteration. The authors generated a text with multilevel annotations. In paper [11], introduced a technique which find out the language category and group them into a single language after that apply the POS tagger on separate chunk. It applied the Twitter POS tagger on English chunk and CRF based Hindi POS tagger on Hindi chunk of the language. They obtained approximately 79% accuracy. The paper [12] demonstrated progress on English based Twitter POS tagger. If we were used the un-supervised tagging method on word features, they obtained maximum accuracy of 90%. Firstly, language identification is done using a simple language detection based heuristic after that the words of same language is group them and apply the POS tagger on each chunk of the language. Language Identification and transliteration of particular language were done by [11]. There were three experiments done on different sets who predict the effects of language identification, transliteration and the POS tagging accuracy. If the language identification and back-transliteration were done automatically then we obtained a POS tagging accuracy of word level around 79.02%. It is also focus the hardness of the problem rather than the importance of correctly language identification and transliteration for POS tagging of text which gives 15% increment among the previous cases. Clearly, it shows POS tagging for code mixed text is difficult to be tagged. However the first attempt is done to tag the social media text is single text like English only. It has improved the passion to tag the word of language other than the English text.

## III. DATASET

Data is collected from the amitavadas.com and <http://tinyurl.com/oewsyx7> which is the chat text of facebook and twitter. We used these data in our experimental setup. So after tagging, the structure looks like this:

word / Language (E/H) / POS tag Example: kolkata /H/NOUN  
kaa/H/ADP charm/E/NOUN aur/H/CONJ busy/E/ADJ  
life/E/NOUN mujhe/H/PRON behad/H/ADJ pasand/H/VERB  
hai/H/VERB

Therefore, the tags are"" separated, the words are space separated and the sentences are line separated.

#### IV. EXPERIMENTAL ANALYSIS AND RESULTS

POS tagging of English-Hindi code-mixed data requires language identification and back-transliteration of the text. To understand the usage of normalization and test our normalization module, we worked on a code-mixed sentence, and it returns tagged word. Due to the complexities discovered in the data, annotation guidelines play an important role in our dataset preparation process. Here, we explain these guidelines for the Language Identification and Normalization processes. These guidelines were given to the annotators who manually tagged the data.

##### A. Language Identification

According to [14], we categorized the process of language identification into two classes hi and en. Each word contains a tag with its two classes hi and en language. Words belonging to bilingual spoke-person would identify and marked it Hindi or English. Language Identification is the first process to recognize the language of every word. All word is tagged with the particular labels such as E or H.

##### B. Transliteration/Normalization

Transliteration is the second process to obtain the correct form of word. If language identification is done and each word is labelled with their language then transliteration is performed only Hindi language. If word is identified as Hindi then it must be back-transliterated to Devanagari script, so that any Hindi POS tagger can be used. On the Hindi chunks we used Google API for back-transliteration. If word is identified as English then transliteration is not done. The Words belonging to Hindi language marked with label 'hi' in roman script is back-transliterated to Devanagari script of Hindi language in their original form. Words belong to English language is marked with 'en' are kept as it is and there are no back-transliteration is performed.

##### C. Part-of-speech

Every word is with its POS labels. For tagging we use universal POS tagsets because it is valid for both Hindi and English language. After the language identification we take consecutive English and Hindi words and group them. On the Hindi chunks we used the Google API for back-transliteration. This gave us the Hindi text in Devanagari. Now we are ready to do our main task. We took each sentences and splitted them into contiguous fragments of words called as chunks. Therefore all the words that corresponds to a chunk have same language either English (E) or Hindi (H) but not the combination. Then we applied TNT based Hindi POS tagger on the Hindi chunks. Similarly we applied the HMM on the English chunks. As we are using two different tagging method, they have different tagsets. The HMM POS tagger has its own POS tagset. The TNT based Hindi POS tagger has ILPOST tagset [1]. Therefore these POS tags remain conserved across languages and hence to ensure uniformity,

we mapped these POS tagsets to the Universal POS tagset [6] which has 12 POS tags. For testing the performance of our system, we developed a text corpus of 2805 words. Result describes the accuracies obtained from the following equation. The accuracy of this module was computed by the following equation.

$$\text{Accuracy} = \frac{n(\text{no of words tagged correctly})}{n(\text{Total no of words})} * 100$$

Where n(X) represents the count of X.

Tagger	language	Accuracy
HMM tagger	En Acc.	91.54
TNT tagger	Hi Acc.	51.75
	Total Acc.	57.98

**Table 1: POS Tagging accuracies for the different model**

Table 1 gives the POS tagging accuracies (in %). It provides what percentage of correctly POS tagged word in the entire text. In the case of Hindi chunk, obtained low accuracies than the English chunk due to the transliteration of Hindi chunk. There are various words having spelling error and it is not correct transliteration.

$$\text{Precision}(P) = \frac{\text{No. of correct POS tags assigned by the system}}{\text{No. of POS tags assigned by the system}}$$

$$\text{Recall}(R) = \frac{\text{No. of correct POS tags assigned by the system}}{\text{No. of POS tags assigned by the system}}$$

	precision	Recall
<b>En</b>	91.54%	91.54%
<b>Hi</b>	51.75%	50.23%

**Table 2: precision and recall of the module**


#### II. CONCLUSION AND FUTURE WORK

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. 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 proven 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.

The paper has aimed to put the spotlight on the issues that make code-mixed text challenging for language processing. We report work on collecting, annotating, and measuring the complexity of code-mixed English-Hindi social media text (Twitter and Facebook Posts). In this paper, we have focused on creating tools for enabling further research on Hindi-English code mixed social media text. The language identification and normalization systems follow supervised machine learning and report final accuracies of 91.54% and 51.75% for our dataset, respectively. We have also developed a complete shallow parsing pipeline, which consists of a POS tagging system and a shallow parsing system, in addition to the language identification and normalization systems. To the best of our knowledge, this system is the first of its kind. We have released this system online and also provided a public API to access it .A dataset of code-mixed Hindi-English words has also been released, to further facilitate research in this direction. This dataset has been annotated by our language identification and normalization systems. The final errors in the dataset due to the inaccuracy of the systems were manually corrected. To summarize this chapter, we have released a dataset which contains code-mixed Hindi-English social media text. It consists of 2805 words, and each word is annotated with its language and standardized form. We also discussed some issues in language identification and word normalization, and compared the effectiveness of noisy channel techniques over English and Hindi normalization.

## REFERENCES

- [1] Hidayat, T. (2008). An analysis of code switching used by facebookers.
- [2] Dewaele, J. (2010). *Emotions in multiple languages*. Springer.
- [3] Vyas, Y., Gella, S., Sharma, J., Bali, K., & Choudhury, M. (2014). Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 974-979).
- [4] Gella, S., Sharma, J., & Bali, K. (2013). Query word labeling and back transliteration for indian languages: Shared task system description. *FIRE Working Notes*, 3.
- [5] Baskaran, S., Bali, K., Bhattacharya, T., Bhattacharyya, P., & Jha, G. N. (2008). A common parts-of-speech tagset framework for indian languages. In *In Proc. of LREC 2008*.
- [6] Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.
- [7] Solorio, T., & Liu, Y. (2008, October). Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 973-981). Association for Computational Linguistics.
- [8] Solorio, T., & Liu, Y. (2008, October). Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1051-1060). Association for Computational Linguistics.
- [9] Jamatia, A., Gambäck, B., & Das, A. (2015). Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 239-248).
- [10] Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). "I am borrowing ya mixing?" An Analysis of English-Hindi Code Mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 116-126).
- [11] Vyas, Y., Gella, S., Sharma, J., Bali, K., & Choudhury, M. (2014). Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 974-979).
- [12] Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 380-390).
- [13] Das, A., & Gambäck, B. (2015). Code-mixing in social media text: the last language identification frontier?.
- [14] Barman, U., Das, A., Wagner, J., & Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching* (pp. 13-23).

	<p><b>Miss. Gajala Praveen</b> is M. Tech student in Department of Computer Engineering, Jamia Millia Islamia, New Delhi. Her area of research include Natural Language Processing .</p>		<p><b>Mr. Danish Raza Rizvi</b> is the assistant professor of Department of computer Engineering, Jamia Millia Islamia, New Delhi. His area of research include Natural Language Processing, Network Security, Cryptography and Steganography.</p>
--	--	--	--