Transliteration Challenges in Hinglish Language

Author: Hari Thapliyal {hari.prasad@vedavit-ps.com}
Reviewer: Prof. Anil Vuppala, IIIT Hyderabad, India
Date:14-Oct-20

Introduction

To understand the transliteration challenges let us understand some other important related terms and the background. To train a machine learning model we need data. Quality and volume of the data determines the performance of model. In NLP projects this data is the text from various books, journals, novels, researcher papers, newspapers, manuscripts etc. This text is called corpus.

We know computer need number to perform calculation and text cannot be used directly for doing any calculation. Therefore, before we start any modelling work on the a given text, we need to convert that text into numbers. A question is how to convert a text into number? Many researchers have proved that word frequency of a word in different context it occurred is that number which can represent the text. We will discuss this process of text to number conversion in following sections. To understand this let's see this example, there are two sentences. "CEO of the Apple is Tim Cook". "Apple should not be cooked for eating". In these two sentences word "Apple" and "Cook" has appeared but they are in difference context. Therefore, we cannot say apple or cook in both the sentence has same meaning. Due to this when these two words are converted into number they should have 4 numbers not two. Just because words have same spelling or have same pronunciation does not mean they are same.

Corpus is that huge volume of text which is used for creating number from the text. Larger the size of the text, more the variety of text better is the word representation we have in number format. If text size is small or text is only from a limited domain or country then this corpus cannot be used to create good number representation. To understand the need of variety of domain let's see three examples from general domain, banking domain and ethics domain. "I don't have interest in fiction novels". "Interest rate in India is better than US". "This is perfect example of conflict of interest". The word "Interest' has occurred in the different context. If we

don't use the text from the different domain for the number representation, then our word to number representation will be poor. To understand the need of variety of culture or countries observe this example. "I use cab to reach my office". "I use taxi to reach my office". In American English cab and British English taxi both are the same word. Therefore, meaning of both the sentences is same. For this reason, word to number conversion of both the words should be such that the numbers corresponding to both words is same, even if spelling and pronunciation of the words are different and both are English language.

What is Tokenization?

Breaking a text into words is called tokenization. For example, this sentence “I do not eat dinner” has 5 tokens [I, do, not, eat, dinner]. Before we convert text into numbers we need to understand morphing of the words. For example “I work for IBM”, “I worked for IBM”, “I am working for IBM” all the words [work, worked, working] have root in a word “work”. Meaning is changing little bit but primarily they are same. If we create 3 tokens for this and continue doing this for whole corpora, then number of tokens will be too many. If we create only one token of “work” then meaning of the sentence is lost. There are some other reasons but here we will limit our discussion and understand how creating token for each word is not good strategy. Thus, we can use subword to create tokens. Let us say we decide to create subwords using minimum 3 letters & max 5 letters then how many sub words possible from these words [work, worked, working]

work=[wor, ork, work]

worked = [wor, ork, rke, ked, work, orke, rked, worke, orked]

working = [wor, ork, rki, kin, ing, work, orki, rkin, king, worki, orkin, rking]

we can remove duplicate tokens and our final token list will contain [wor, rki, rke, ork, kin, ked, ing, work, rkin, rked, orki, orke, king, worki, worke, rking, orkin, orked] 18 tokens. In between we find some known words which were not originally part of our tokens. This will take care of out of vocabulary (OOV) problem.

What is Sequencing?

A process of assigning a unique number to a word or subword is called sequencing. Sequencing does not care about the frequency of the word, order in which they occur in the sentence and meaning of the word. For example, a sentence “I like chocolates” has three tokens and it can have sequence as following. I = 1, Like=2, Chocolates=3

What is Embedding?

Sequencing ignore the meaning, context, and frequency of the words. But to understand the meaning or to perform any NLP task it is extremely important to consider the meaning, context and frequency of the words. Embedding is a process in which word meaning is derived from the text and a word vector is created. There are different kind of embedding possible. Using the huge corpus when we generate word vector it is called word embedding. But when we generate sentence vector it is called sentence embedding. Sentence embedding can also be created using the words of the sentence or the context in which entire sentence is used. First method is more practical, lessor complex and need lessor resources. Second method need many fold voluminous corpus and resource to process the embedding. Similarity we can create vector for paragraphs, headings, chapters, books, and genre. Although we are not aware about the usage of these embeddings in the business or any product but it is logically possible.

Uniqueness of English Language

Before we understand the challenges lets us see the uniqueness of Hinglish. For that purpose we need to come out of bookish Hindi or Hindi of newspaper editorials. To make a good system which can perform NLU tasks on Hinglish language we need to peep in the language of social media, whatsapp msg, phone messages etc. Let us observer some Hinglish sentences on social media.

Sentence1: #Meherbaan ऐसे उबाऊ सोंग बनाकर हमपर मेहरबानी करे। धन्यवाद #BangBang

Sentence2: Breaking News रांची में आतंकी संगठन की धमकी 30 days में गांव छोड़ें वरना मार देंगे और लड़कियों को मुसलमान बना देंगे #इस्लामisReligionOfPiecing

Sentence3: उसने कहा मेरे प्यार में फ़ना हो जाओ मैंने कहा मेरे पास टाइम नहीं है दफा हो जाओ। 😉 😉 - feeling crazy

We can see people freely use different scripts, special letters, emoticons, words from different language. To perform any NLP task on this kind of text if we use embedding generated from pure Hindi corpus then it is not going to give good results. Therefore, it is important that we normalize this kind of text in some way. From scripting perspective either convert the complete sentence into Devanagari or convert the complete sentence into roman script. From language perspective convert all the English words into Hindi. Second work has lessor impact on the final result, but first work has more impact.

Some Challenges in Hinglish Transliteration

Before we use our corpus for training or predict the class of a given text, we need to transliterate our text into Devanagari. Let us see the challenges of Hinglish transliterations. Please make a note this is not an exhaustive list but the challenges faced by me in my NLP projects.

1. What to do when a text like #"SadStoriesOfTwitter" comes between Devanagari text? These are hashtags of combined words using camel casing
2. What to do when hashtag like #PulwamaAttack and #पुलवामाअटैक both are in the text. Meaning wise both are the same
3. What to do when a hashtag is having two scripts. #खूनीDemonetisation
4. What to do when a text like 10साल or 10Years
5. What to do when hashtag without camel case, space or "_". Like #baggapajjirockstar
6. When hashtag has an English, number, Devanagari hashtag like #Results2018लंदन
7. What to do when text is like 2019मोदीElections appear without the hashtag
8. What to do when words like "ReporterDiary" appear, they are not hashtag not Devanagari nor English without a split.
9. Slogans like VandeMataramBSF. Even after splitting this is not an English word but written in roman script. It is not hashtag either.
10. What to do when a text like "ED", "CM", "PMO" come between Devanagari text which already has "सीएम" "पीएम" (These are acronyms). With a normal transliteration technique, you can convert them correctly. But how to know which word in the corpus need this kind of special treatment.
11. When acronyms are with "." like C.M. P.M.O or "सी.एम." "पी.एम.". Note in Devanagari एम is two letters and then "." is appearing.
12. What to do when text like "Amit, Shah, Narendra, Modi, Hari, Thapliyal, Surya" come between Devanagari text. Devanagari text "अमित, शाह, नरेन्द्र, मोदी, हरि , थपलियाल, सूर्या" also exists. We cannot translate roman script name we need to transliterate them.

    We could not find a single transliteration library which can convert Roman written words to Devanagari words without any error. The reason is obvious, because we

choose different spelling for the same word. Without telling the system that Roman peer in Devanagari or reverse the word vector will not be representing the word correctly.

13. What to do when text like "orange, rice, work, good, excellent" come between Devanagari text? These are proper words from English. In the same text you will find transliterated words like "औरेंज राईस राइस वर्क गुड एक्सलेंट" or translated and scripted in Devanagari in "संतरा चावल काम अच्छा बहुत बढिया". For which word we should create embedding औरेंज or संतरा or Orange?
14. English word in Devanagari क्वारेंनटाईन and sometimes "Quarantine" also exists in the text. If we maintain them as different word, then word vectors will be different.
15. What to do when text like मानहानि, मुकदमा, नुकसान come between text. They belong to multiple Indian language but some of the library returns them as marathi or nepali language.
16. What to do when text like अमि, खाबो, पंगा, धश, कुडी come between text. They belong to different Indian language but sometimes used by bilinguals. Language identification fails! For translation, to Hindi, we need a bilingual dictionary between those languages and Hindi.
17. What to do when emoticons are combined with other roman and Devanagari letters to construct words. Like "👞खाएगा"
18. When emoticons are used to construct a sentence. And at different place you find different expressions, but both are same. 😘 दोस्तों = Love you friends. बहुत 🌞 है = Too 🌞 = It is too hot today.
19. How to tell system all are the same. They may be coming as a hashtag or without hashtag राहुल_गांधी = राहुल-गांधी = राहुल गांधी = राहुल.गांधी = राहुलगांधी = राहुल_गान्धी. In English spelling mistakes are taken seriously and in Hindi world more than spelling sound is important.
20. Acronym created by societies but not part of any formal communication like "LLTT" Looking London talking Tokyo.

21. Numeric Letters in Devanagari script. We frequently change digit between English and Devanagari. Sometimes they write 0123456789 other time they write ०१२३४५६७८९. The text which we are using for prediction or training need to have either roman or Devanagari text. This is mixed will affect our model results
22. Wrong spellings of Devanagari in Roman. In Devanagari, my surname is थपलियाल but in school when we have to fill our form for board exam, we need to write our name in Roman and that time we choose spelling of our name. For the same name, same Devanagari spelling and pronunciation my different cousins choose different spelling for this word. Possible variations are Thapliyal, Thapaliyal, Thapaliyaal, Thappliyaal, Thapliyaala, Thapaliyaala. In Hinglish people sometimes will write words in roman. This spelling problems create different vector for each variation.
23. Wrong spelling of English words in Hinglish sentence. Hinglish speakers don not care much about the spelling of English words. They use English words, write in Roman but with wrong spelling. Urope or Europe, America or Amrica or Amricaa, Merat or Merrut, sycologi or psychology. The main reason for this is Hindi speakers write the word the way they pronounce.
24. Splitting word with Samyukta akshara. How we want to split a word into letters. In English it is easy. “APPLE” => [A, P, P, L, E]. But in Devanagari how we want to split a word that need to be taken care. . "संयुक्त” => ['स', 'ं' , 'य्', 'ु', 'क्', 'त्', 'ा'] or ['सं', 'यु', 'क्त'] or ['सं', 'यु', 'क्', 'त']
25. In roman we have concept of uppercase letter and lowercase letters. So, when we find a word like “VoteForCleanIndia” it is easy to split using a simple logic. But when we find a word नशामुक्तभारत it is easy for a native speaker to split the word but using some computer program it is an overly complex to break this one word into meaning for 3 words.

Conclusion

Before selecting a corpus for training or creating a model we need to understand what kind of input text that model is going to see in production environment. If production time text is completely different from training time data then no matter which efficient algorithm, we use to build the model, we won’t be able to get the good results. Therefore, it is necessary to

consider above challenges and get a solution before expecting some good results from the model. We know the principle of garbage in garbage out. If embedding is not good, then NLP/NLU results will not be reliable.