Translation: Code-Mixed Language(Hinglish) to English

Dr. S.V. Kedar, Sakshi Bhangale, Kunal Deokar, Sahil Deshmukh and Parikshit Biradar

JSPM'S Rajarshi Shahu College Of Engineering,Pune,411033, Maharashtra, India

Article Info

Page Number: 159 - 167

Publication Issue:

Vol 71 No. 4 (2022)

Article History

***Article Received**: 25 March 2022*

***Revised**: 30 April 2022*

***Accepted**: 15 June 2022*

***Publication**: 19 August 2022*

Abstract

In a diverse country like India where most of the people are multilingualthe use of the pure language is decreasing. Interpretation of such mixed language becomes easy for human but complex for machine. There are many machine learning models which are trained for pure language translation but there is a research gap in the field of code-mixed language translation. In order to bridge this gap, the paper presents a model which uses Hinglish as a standalone language to be translated into English. In this paper, we have discussed the algorithm, technique and limitation of our system.

Keywords*:* Machine translation (MT); Code-mixing; Language Analysis; Hinglish; Corpus Based MT; and Rule based MT; Hybrid MT

1.Introduction

In the growing Era of technology, use of mixed language is getting normalized to a extend that even social media posts,speech,day-to-day communication are done in mixed language and use of pure language is decreasing.Hence there is lot of data available in mixed language which becomes difficult for the machine to interpret.There are lot many work done for translation of pure languages but now research need to focus on analyzing the content available in mixed languages.We have come up with the translation model for code-mixed language(Hinglish) in NLP.

The objective of this project is to translate Hinglish(Hindi+English) which is combination of Hindi and English language to pure English language.The proposed model uses Hinglish as standalone language which makes it a direct translator of code-mixed language to pure language.It helps in analyzing the content available in mixed languages.It will also bridge the gap between the interaction of machine and human making it more real.

1.1 Related Work

Lot of research is going on code-mixed content and in particular those involving language tagging.An ensemble model was created by Jhamtani et al. (2014) which was combination of two classi-fiers to form a LID mixed with Hindi-English code.Features like word frequency, modified edit distance,character n-grams were used by first classifier and second classifier used the output from previous one for current word as well as languages and pos tag for nearby words to give final tag. Rijhwani et al. (2017) came up with fully unsupervised language tagger which used arbitrary set of languages. About back-transcription, , Bilac and Tanaka (2004) proposed a hybrid approach.It combined phoneme, graphim and segmentation based modules.An architecture for bach-transliteration which uses SMT framework that is described in (Franz et al., 2003) was introduced by Luo and Lepage (2015).Ravishankar (2017) discussed a finite-sate system for back-translliteration of Marathi words to English.Sinha and Thakur (2005) worked on translation of English-Hindi code mixed to pure English from linguistic view by using morphologicalanalyzers but they did not do any depth evaluation.Dhar et al. 2018 also worked on translating code-mixed language using parallel corpus.

2. System Requirements

2.1. Software Requirements:

- OS - Windows 8 or above
- Any Code editor (VS Code)
- Libraries – Pickle, Numpy, Keras, Tensor Flow, NLTK

2.2. Hardware Requirements:

- RAM 4Gb or above
- Processor - i3 or above

3.Dataset preparation and pre-processing

Our model requires Hinglish-English sentence pairs.As we know for a efficient machine learning model quality and quantity of dataset play a vital role.Since Hinglish sentences dataset is not readily available we have created our own dataset.There are more than 10,506 english-hinglish pairs and still working on it.We used 9456 pairs for training and 1050 testing.

The dataset requires some pre-processing which includes:

- Punctuation removal to make it a plain sentence without any punctuation mark.
- Normalising of words to reduce its randomness.
- Separating the Hinglish-English pairs using similar symbol.

```
Tom turned down the offer.|Tom ne prastav ko thukra diya.
Tom unpacked his suitcase.|Tom ne suitcase khali kiya.
Tom used a legal loophole.|Tom ne apna kaanoonee khaamiyaan istamel kiya.
Tom used a legal loophole.|Tom ne apna kaanoonee khaamiyaan istamel kiya.
Tom used to be overweight.|Tom adhik vajandar hua karta tha.
Tom usually wears glasses.|Tom aksar chashme istamel karta hai.
Tom wanted an economy car.|Tom ko ek arthavyavastha kaar chahiye.
Tom wanted me to help him.|Tom ko mera madat karna tha.
Tom wanted to be a doctor.|Tom ko doctor hona tha.
Tom wanted to lose weight.|Tom ko vajan kaam karna tha.
Tom wanted to say goodbye.|Tom ko alvida kehna tha.
Tom wanted to say goodbye.|Tom ko alvida kehna tha.
Tom wanted to talk to you.|Tom ko mere se baat karna tha.
Tom wants Mary's approval.|Tom ko Mary ki anumodan chahiye.
Tom wants his money today.|Tom ko apna paise aaj chahiye.
Tom wants me to apologize.|Tom ko meri shama yachna chahiye tha.
Tom wants to donate money.|Tom ko paise daan karna hai.
Tom wants to dye his hair.|Tom ko apne baal dye karna hai.
Tom wants to go to Boston.|Tom ko Boston ko jana hai.
Tom wants to go to Boston.|Tom bostan jaana chaahata hai.
Tom wants to learn French.|Tom ko French shikana hai.
Tom wants to look younger.|Tom ko jaawan dikhana hai.
Tom was Mary's first love.|Mary ka pehla Pyaar Tom tha.
Tom was a little homesick.|Tom thoda ghar ke baahar rahane se khinn tha.
Tom was a prisoner of war.|Tom jang ka kaidee tha.
Tom was able to handle it.|Tom isse sambhaalane mein saksham tha
Tom was able to help Mary.|Tom Mary ko madat karne mein saksham tha.
Tom was acting on his own.|Tom apne dum par abhinay kar raha tha.
Tom was attracted to Mary.|Tom Mary ke prati aakarshit tha.
Tom was bitten by a cobra.|Tom ko cobra ne kaat liya tha.
```

Figure 1. Sample image of dataset

4.Proposed system

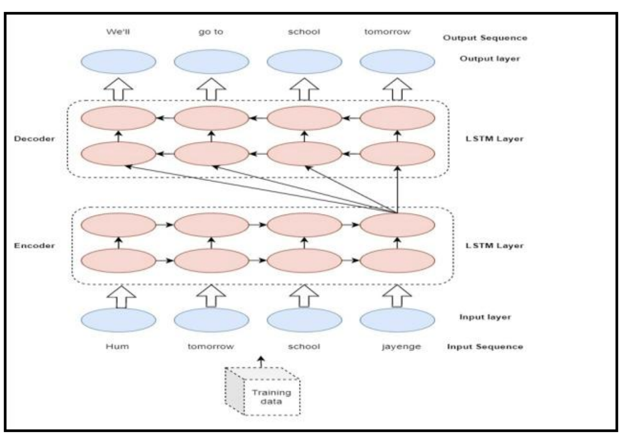

Figure 2. Architecture for mixed-code translation model.

ISSN: 2094-0343
2326-9865

Figure2. includes the architecture for the mixed-code translation model. The model is divided into three layers viz. input layer,LSTM layer and output layer.

4.1. Input layer

After preprocessing of data the cleaned string is given as a input.

4.2. LSTM

LSTM layer consists of encoder and decoder that are both stacks of residual attention blocks.The uniqueness of such encoder-decoder model is that such sttention blocks can process an input sequence i.e. X1:n for variable length n without showing repeating structure.

In order to solve a sequence to sequence problem we need to get an input sequence mapping X1:n to an output sequence Y1:m.

The encoder-decoder model defines conditional distribution of target vectors Y1:n when given input sequece X1:n:

$$p\ \theta\text{enc},\theta\text{dec}(Y1{:}m|X1{:}n)$$

The encoder part will then encode the input sequence into a sequence which is hidden states x1:n and thus mapping will be defined :

$$f\theta enc : X1{:}n \rightarrow X1{:}n$$

The decoder part will then define the conditional probability of target vector sequence y1:n when given the sequence of encoded hidden states X1:n :

$$p\theta dec(Y1{:}n|X1{:}n)$$

This distribution is factorised to a product of conditional probability distribution of target vector Yi given the encoded hiddedn states X1:n and also all previous target vectors Y0:i-1 :

$$p\theta dec(Y1{:}n|X1{:}n) = \prod_{i=1}^{n} p\theta dec(yi|Y0{:}i-1, X1{:}n)$$

The decoder will map the sequence of encoded hidden states X1:n and also previous target vectors Y0:i-1 to logit vector li which is then processed by softmax operation to define conditional probability $p\theta dec(yi|Y0{:}\,i-1, X1{:}\,n)$

After defining the conditional probability we can now auto-repeatedly generate output and thus mapping is defined on input sequence X1:n to output sequence Y1:m

5. Experimental Analysis

5.1 Calculating accuracy

The datset is splitted in 90:10 ratio.The accuracy of the model is caclculated using BLEU score.It is a number between 0 and 1 which measures quality of text translated by machine.With the available dataset the model gave BLEU score of 0.4 for testing which will eventually increase to 0.6-0.7 as quality and quantity of dataset will increase.

The mathematical details

Mathematically, the BLEU score is defined as:

$$\mathrm{BLEU} = \underbrace{\min\Big(1, \exp\big(1 - \frac{\text{reference-length}}{\text{output-length}}\big)\Big)}_{\text{brevity penalty}} \underbrace{\Big(\prod_{i=1}^{4} precision_i\Big)^{1/4}}_{\text{n-gram overlap}}$$

with

$$precision_i = \frac{\sum_{\text{snt} \in \text{Cand-Corpus}} \sum_{i \in \text{snt}} \min(m^i_{cand}, m^i_{ref})}{w^i_t = \sum_{\text{snt'} \in \text{Cand-Corpus}} \sum_{i' \in \text{snt'}} m^{i'}_{cand}}$$

where

- m^i_{cand} is the count of i-gram in candidate matching the reference translation
- m^i_{ref} is the count of i-gram in the reference translation
- w^i_t is the total number of i-grams in candidate translation

Figure 3. Mathematical Expression of BLEU score

6. Results

6.1.Train result

```
train
src=[voh jaanatee hai vah hamesha jaanatee hai], target=[she knows she always knows], predicted=[she knows knows knows knows]
src=[krpaya jo aapane kaha use likhen], target=[please write what you said], predicted=[why could what you you]
src=[munaapha bahut adhik tha], target=[the profits were very high], predicted=[the profits were very high]
src=[mujhe laga ki main tumhen samajh gaya], target=[i thought i understood you], predicted=[i thought i listened you]
src=[unhonne pichhale saal kyoto ka daura kiya], target=[he visited kyoto last year], predicted=[he visited kyoto last year]
src=[aap paris kab aae], target=[when did you come to paris], predicted=[when did you come to paris]
src=[mainne ise sveekaar karana seekh liya hai], target=[ive learned to accept that], predicted=[ive have to about that]
src=[aapko kis baat par garv hai], target=[what do you take pride in], predicted=[what do you like pride for]
src=[tom badbadaya], target=[tom grumbled], predicted=[tom grumbled]
src=[main jaanana chaahata hoon ki tom kee mrtyu kaise huee], target=[i want to know how tom died], predicted=[i want know how to tom tom]
BLEU-1: 0.636280
```

Figure 4. Training the Model

6.2. Test result

```
test
src=[main purushon ke kamare mein ja rahee hoon], target=[im going to the mens room], predicted=[i going to in the house]
src=[mujhe nahin lagata ki yah isake laayak hai], target=[i dont think its worth it], predicted=[i dont think it is it]
src=[mere paas maveshiyon ke sir hain], target=[i have head of cattle], predicted=[i have a of cattle]
src=[aapki pasandidarkhana konsi hain], target=[whats your favorite food], predicted=[what is this movie]
src=[mainne ek tattoo kee tasveer kheenchee], target=[i drew a picture of a pony], predicted=[i have a a a a hour]
src=[koee nahin jaanata ki vah kahaan rahata hai], target=[no one knows where he lives], predicted=[nobody knows what to is]
src=[main chaar baje tak intazaar karoonga], target=[ill wait till four oclock], predicted=[i snowed five four oclock]
src=[aapaka pasandeeda opera kaun sa hai], target=[whats your favorite opera], predicted=[whats your favorite opera]
src=[unhonne raat bhar kaam kiya], target=[he worked through the night], predicted=[he painted the the blue]
src=[mujhe tom ka chehara yaad nahin hai], target=[i dont remember toms face], predicted=[i dont feel tom in tom]
BLEU-1: 0.446557
```

Figure 5. Testing the model

7. Limitations

- As there is limited work in code-mixed language domain the dataset for Hinglish sentences are not readily available and needs to be created manually due to which minimal datais available which eventually hampers the accuracy of model.

- We could work to improve the performance of model to translate group of sentences or paragraph along with more grammatical aspects so that it will produce better results.
- The model is still in prototype phase which generates result for limited scope of sentences.In order to deploy it for real time translation the quality and quantity of dataset needs to be improved.

8. Future Research Direction

- Direct Machine approach translates word to word from SL to TL with basic analysis and less consideration of grammar. It works well only for small sentences. We could work to improve its performance to translate group of sentences or paragraph along with more grammatical aspects so that it will produce better results.
- All the approach mention above mostly work on pure language but for code mix language we need more systemic approach in machine translation also accuracy of the above mentioned approach varies drastically even if slight change in grammar or spelling nuance results in low accuracy.
- Now a day's code mixing has become very common phenomenon. In the Era of globalization where people from vivid background interact, combination of two or more languages while communicating is becoming a normalcy. As a human we can interpret it but for machine understanding such languages is quite difficult so we can work to improve accuracy of our system to analyse such combined languages e.g. Hinglish (English + Hindi).

9. Applications

- Chatbot: For more realistic and interactive communication.
- For various types of Security Purpose.
- Personal Assistant: personal assistant like Alexa, Google, etc.
- We can use this as a speech to text translation.
- Data analytics: To analyze social media post, comments, image, videos, blogs etc.

10.Conclusion

This implementation paper proposed the use of LSTM algorithm for translation of code-mixed language.It considered Hinglish as standalone language which is unique feature of the model.It will help in analysing the large chunk of data which cannot be accurately interpreted due to mixed words from different languages.It came up as improvement to various previously developed systems which involved use of intermediary language for translation of mixed languages.With improvement in dataset we will be able to increase the accuracy of the model to 0.6 -0.7 and deploy it for real time translation.

11.References

1. IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 2, September 2014 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784 www.IJCSI.org
2. Indonesian Journal of Electrical Engineering and Computer Science 1(1):182 DOI:10.11591/ijeecs.v1.i1.pp182-190
3. Peng L. A Survey of Machine Translation Methods. TELKOMNIKA Indonesian Journal of Electrical Engineering. 2013; 11(12): 7125-7130
4. Hutchins W.J, Somers H L. An introduction to machine translation. London: Academic Press.1992:
5. Slocum J. A survey of machine translation: its history, current status, and future prospects.Computational linguistics.1985;11(1):1-17
6. Antony P J. "Machine Translation Approaches and Survey for Indian Languages." International journal of Computational Linguistics and Chinese Language Processing. 2013; 18(1): 47-78
7. Peng L. A Survey of Machine Translation Methods. TELKOMNIKA Indonesian Journal of Electrical Engineering. 2013; 11(12): 7125-7130
8. Chéragui, Mohamed Amine. Theoretical Overview of Machine Translation. Proceedings ICWIT.2012;
9. Hutchins John. A new era in machine translation research. In Aslib proceedings. 1995; 47(10) 211-219
10. Tripath s , Sarkhel. K. Approaches to machine translations. Annals of Library and informationstudies.2010; 57: 388-393.
11. Ansary S. Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas. In 11th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt. 2011:

ISSN: 2094-0343
2326-9865

12. Hiroshi U , Meiying Z. Interlingua for multilingual machine translation. Proceedings of MT SummitIV, Kobe, Japan. 1993:157-169.
13. Juss`a, M, Farru´s M, Marin˜o.J, Fonollosa.J. study and comparison of rule- based and statistical Catalan-S panish MT systems Computing and Informatics. 2012; 31: 245–270
14. Saini Sandeep. Vineet Sahula. A Survey of Machine Translation Techniques and Systems for Indian Languages. In Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference.2015: 676-681.
15. Koehn P, Och J, Daniel Marcu. Statistical Phrase-Based Translation. Proceedings of HLT-NAACL,Edmonton, May-June 2003. Main Papers , 2003: 48-54.
16. MD Okpor. Machine translation approaches: issues and challenges. International Journal of Computer Science Issues (IJCSI), 11(5):159, 2014.
17. Béchara Hanna. Raphaël Rubino. Yifan He. Yanjun Ma. Josef van Genabith. An Evaluation of Statistical Post-Editing Systems Applied to RBMT and SMT Systems. In COLING. 2012; 21: 5-230.
18. Costa-Jussa Marta R. José AR Fonollosa. Latest trends in hybrid machine translation and its applications. Computer Speech & Language. 2015; 32(1): 3-10.