iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37577

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Comparative Analysis of Language Translation and Detection System Using Machine Learning

Aishwarya R. Verma[1], Dr. R. R. Sedamkar[2]
[1]*M.E. Scholar, Computer Engineering, Thakur College of Engineering and Technology, Mumbai University*
[2]*Director of IQAC and HOD (Ph.D.), Thakur College of Engineering and Technology, Mumbai University*

Abstract: ***Words are the meaty component which can be expressed through speech, writing or signals. It is important that the actual message or meaning of the words sent must conveys the same meaning to the one receives. The evolution from manual language translator to the digital machine translation have helped us a lot for finding the exact meaning such that each word must give at least close to exact actual meaning. To make machine translator more human-friendly feeling, natural language processing (NLP) with machine learning (ML) can make the best combination. The main challenges in machine translated sentence can involve ambiguities, lexical divergence, syntactic, lexical mismatches, semantic issues, etc. which can be seen in grammar, spellings, punctuations, spaces, etc. After analysis on different algorithms, we have implemented a two different machine translator using two different Long Short-Term Memory (LSTM) approaches and performed the comparative study of the quality of the translated text based on their respective accuracy. We have used two different training approaches of encoding-decoding techniques using same datasets, which translates the source English text to the target Hindi text. To detect the text entered is English or Hindi language, we have used Sequential LSTM training model for which the analysis has been performed based on its accuracy. As the result, the first LSTM trained model is 84% accurate and the second LSTM trained model is 71% accurate in its translation from English to Hindi text, while the detection LSTM trained model is 78% accurate in detecting English text and 81% accurate in detecting Hindi text. This study has helped us to analyze the appropriate machine translation based on its accuracy.***

Keywords: Accuracy, Decoding, Machine Learning (ML), Detection System, Encoding, Long Short-Term Memory (LSTM), Machine Translation, Natural Language Processing (NLP), Sequential

I. INTRODUCTION

When we want or wish to learn a new language, we initially start searching for the dictionary books or websites where we can start learning basic new words, sentences and phrases. The most important thing here is to understand the exact meaning of the words and/or sentences of that new language. People also prefer to have a quick machine translator handy (through online/softwares) where it can be used quickly to find the meaning of the word or sentences. As per my previous study [1], all translators do not translate it correctly which results in improper knowledge transfer in the form of incorrect abbreviations, grammar with tenses, meanings, punctuations etc. The translators will try to translate the sentences or word based on its predictions. The translator must understand which words need to be translated and which words needs to be transliterated.

After studying different types of algorithms based on bilingual translators, we have designed two Long Short-Term Memory (LSTM) models, which are provided with different sets of training in encoding-decoding techniques with same datasets. This machine translator will translate English to Hindi word and sentences, taking care of the grammar, punctuations, spellings, tenses and meaning of the word and sentences. As we are training the model through LSTM, we have also designed language detection system, in which it can recognize whether the provided words and sentences is English or Hindi text. We have used Sequential LSTM model where we have loaded the same datasets used while translation. This translation-detection models have helped us in analyzing the predictions of the implemented translator which provides the close-to same meaning translated word and sentences when compared with the actual word and sentences.

The goal of this paper is to provide the comprehensive, comparative, self-contained analysis of deep learning LSTM models used for translation and detection system. These methodologies have helped in predicting the meaning of the word and sentences, based on which the translation is being performed to get the accurate translated word and sentences that conveys almost exact meaning as the source word or sentences.

©IJRASET: All Rights are Reserved

A. *Problem Definition*

A brief study of problems related to machine translation is related to the quality of the source text which is provided by the different users and the target text that provides you or user the translated text. It is very important that the meaning of the translated text must not change the meaning as provided by user while entering the source text. At one moment of time, the google translator, Bing translator could not be able to translate the given word/sentences, which sometimes land up in transliteration especially for some particular words. As I mentioned earlier, there are some situations in the rural area where the people are unable to understand while filling the important form. This can demotivate people for finding the 3rd person who know that language very well for each time for filling the form [1]. At times, some can't even recognize which language is used and how to read them.

B. *Motivation*

Machine translator can be trained in many ways such that the translated text can be predicted based on these following parameters: meaning, grammar and tense of the word or sentence, transliteration (especially in case of name of the persons, places etc.), feelings in punctuations, spellings etc. Google, Yahoo, Bing and other search engines translators uses the texts as data provided by us while searching on their search entry. Based on this, the training is performed which does not only help to identify what languages are been used, but also helps to find the more meanings in detail. For example, the word "good" can be expressed in a following way:

1) Good! I like it.
2) Is it a good one?
3) Good morning
4) I am good.

The actual meaning in Hindi for above can be as follows:

- बढ़िया है! मुझे ये पसंद आया ।
- क्या यह अच्छा है?
- शुभ प्रभात
- मैं ठीक हूँ।

As you can see from above example how the word "good" can be expressed in a different way for translated text.
This analysis in our day-to-day like motivates me to build such system, which can be trained in such a way that it should read the source text, understand the meaning, tense, emotion/situation of the sentences and gets translated without changing the meaning.

II. LITERATURE REVIEW

There are various research performed for Machine Translator and Detector systems. The 1st machine translator was invented in 1954, which was translating Russian to English sentences. There were different revolutions took place with different languages to make the translator more and more efficient in finding the meaning.
According to Sindhu and Sagar [17], Madura and Satish [13], the machine translation system can be classified as represented in fig. 1. As per the study[17][13], it had explained their each study towards each machine translation and found out the respective translators such as ANGLABHARTI MT for direct based MT which translates English and Punjabi to Hindi Language; MaTra system for Transfer based MT which was developed for translating news articles, annual reports and technical phrases; Universal Networking Language (UNL) for Interlingua based MT which were used for pictorial knowledge representation; VAASAANUBAADA for Example based MT for Bengali-Assamese language pair; ANUBHARTI for Hybrid based MT.
As per analysis performed by Md. Saidul Hoque Anik, Md. Adnanul Islam, A.B.M. Alim Al Islam [6] and Shamsun Nahar, Mohammad Nurul Huda, Md. Nur-E-Arefin, Mohammad Mahbubur Rahman [11]; their main focus while translating Bengali sentence to English sentence [6] and vice-versa [11] was 12 different tenses - present indefinite, continuous, perfect, perfect continuous; past indefinite, continuous, perfect, perfect continuous; future indefinite, continuous, perfect and perfect continuous.
The former used the Levenshtein distance and modified Levenshtein distance approach in which the root word given other form of verb then the weighted distance between a root word and another verb form is measured, and the minimum distance is calculated to convert the verb form into the root word by inserting, deleting or replacing the characters [11]. The latter worked on Corpus based approached, which were compared with Direct approach, Transfer approach and Google Translator [6]. As per Afsana Parveen Mukta, Al-Amin Mamun, Chaity Basak, Shamsun Nahar, Md. Faizul Huq Arif; they had worked on Rule-Based approach for phrase-based machine translation from English to Bangla [5].

©IJRASET: All Rights are Reserved

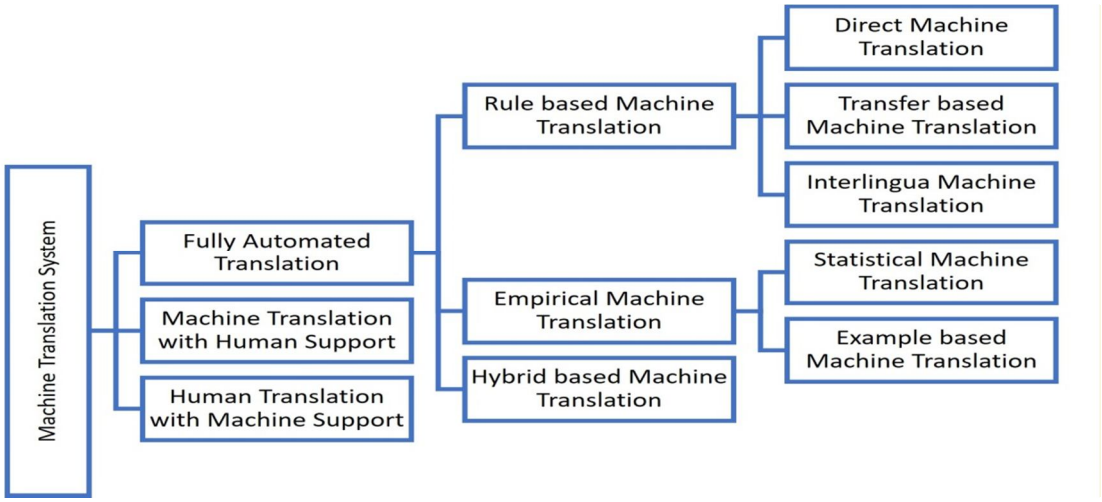


Fig. 1 Approaches to Machine Translation [13]

According to the analysis of Chandramma, Dr. Piyush kumar Pareek, Swathi K, Puneet Shetteppanavar; they implemented their machine translator in GIZA++ for providing the training in the form of small corpora for n-gram maintenance in each phrase table, which helped in the translation of Kannada text to Telegu text [14]. The direct translation system which accepts Marathi language, which is derived from Devnagri, focused on identifying part of speech and tokenization along with identifying English meaning for each word creating bilingual dictionary [7]. While the translation of Malayalam to English using Hybrid approach have provided good quality of translation as compared with rule-based translation and corpus-based translation [19] and the translation from English to Malayalam text focused on transfer approach which involved bilingual dictionary, transliteration and morphological inflection generation phase [20].

As per [21], statistical machine translation was the fastest one but its weak in its accuracy at the same time neural based network had higher accuracy but slow in computation process, so to consider the speed and accuracy both together they performed the comparison between RNN and statistical based network with n-gram model for English-Indonesian MT in which RNN obtains more excellent result which evaluation. According to [8][9], they had worked on multiligual translations based on neural machine translation, focusing on translation from Chinese to English and then English to German. [8] focused on DNNJM (Database Neural Network Joint Model) with PBT (Phrase-Based Translation) and HPBT (Hierarchical Phrase-Based Translation) for verifying their novel NN methology, while [9] focused on the translation from English to Chinese using PBT.

People are so busy in social media as a part of news, jokes, opinions etc. our thoughts can be expressed in many ways. People who can understand their mother tongue, they made the settings in their devices such that they can read and understand without affecting the meaning of the content. That's where the language translator comes into picture. Most of the websites and blogs are not supported with language translators, so if any naïve person who knows only English and suddenly lands up to the page which is in French or Spanish. The user may need to go every time to the website and check which language have been used. So, to atleast make the person aware the language, language detection is one of the important tools in picture too.

According to Omar, Mohammed, Kayes, Raza and Iqbal [22], it focuses on training the model using stochastic gradient descent (SGD) with unigram and bigram features, such that it could detect suspicious content in social media and blogs in Bengali texts. In [23], the training has been performed focusing on the word embedding to detect the language used in social media and blog. The main focus in this [23] was the reading of the transliterated text, comparing it with the text of the language and identifying which language are been used while conversation. While [24] focuses on Support Vector Machine (SVM) & Naïve Bayes (NB) classifiers in the detection of the Ekman's six types of basic emotions such happy, fear, anger, sadness, disgust and surprise in Punjabi text.

III.METHODOLOGY

Below are the methodologies for three LSTM models, in which 1st model is based on Forward LSTM with mask zero concept (which is also known as modified Forward LSTM) and 2nd model is Bi-directional LSTM model works on the concept of encoding-decoding for language translation, while 3rd model is Sequential LSTM which is trained based on the same dataset used for language detection.

©IJRASET: All Rights are Reserved

A. Forward LSTM with mask zero in Embedding Layer

The following are the steps involved while training the model:

1) *Step 1 - Data Input Layer:* We have used English and their respective Hindi meaningful sentences in csv format as a part of data input model. These statements are collected from different education sources, so that we can get meaningful translation.
2) *Step 2 - Data Preprocessing:* This step involves data preprocessing, which involves lowercasing of all the letters in English, removal of extra spaces, redundant data which are missing with either English or Hindi statement. Here we have prefixed "START_" and "_END" for every Hindi word/sentence for better reading. We have collected English and Hindi vocabulary for further training.
3) *Step 3 - Steps After Data Preprocessing:* To get the translation of the text, the maximum length for English and Hindi had been analyzed and set to 21. The number of encoder tokens and decoder tokens are taken based on the numbers of vocabulary of English and Hindi respectively, along with the increment by one respectively. These tokens and their indexes are then loaded into dictionaries.

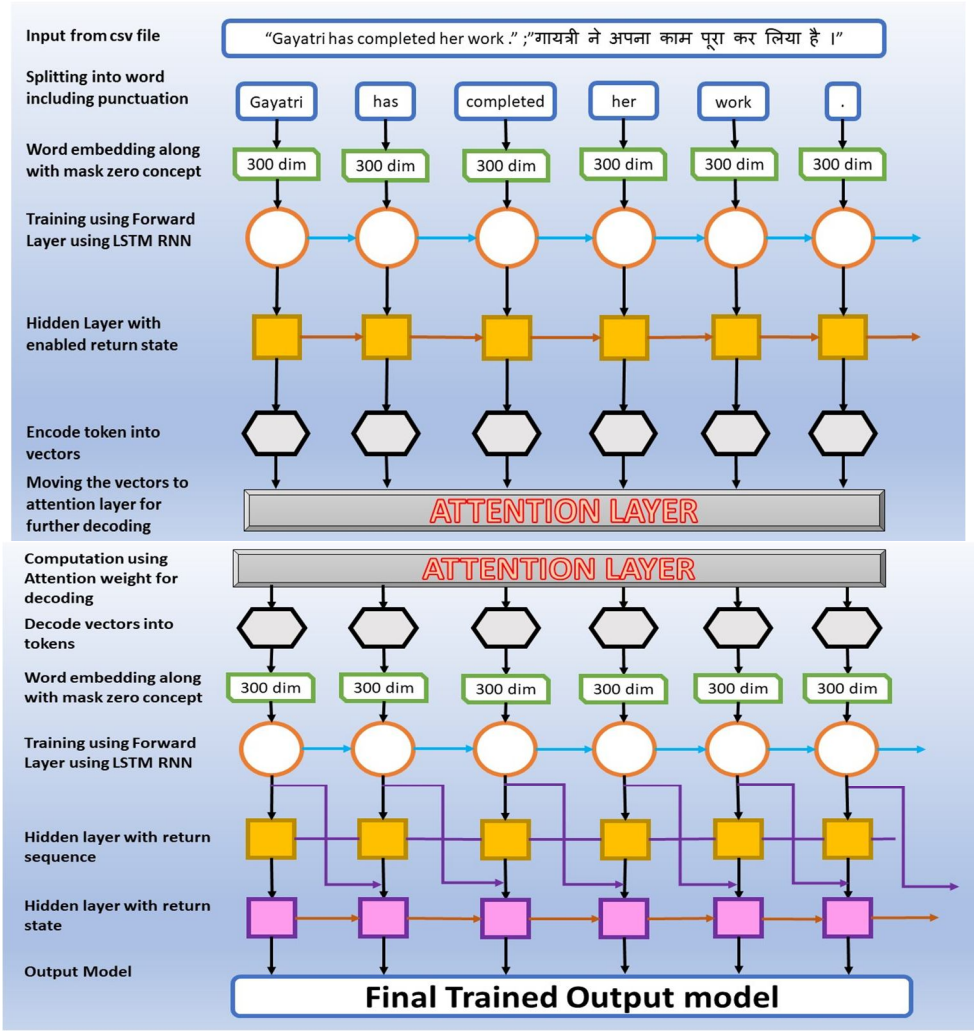


Fig 2 Flow of Forward LSTM with Mask Zero in embedding layer.

4) *Step 4: Classifying the Data for Training and Testing:* The data is divided in the ratio 4:1 for training and testing respectively. A batch of data had been created such that it can train the input sequences for encoder and decoder respectively, which involves splitting of source input text and target output text.
5) *Step 5 - Encoder for the source text and Decoder for Target Text:* For better word embedding, we have used 300 dims, enabled the mask zero and return state for further enhancement in encoder inputs. For decoder, we have enabled mask zero, return state and return sequence with 300 dims.

©IJRASET: All Rights are Reserved

6) *Step 6 - Attention Layer:* The tokenization of the context vector is calculated using the convex combination of the annotation vectors of source text, which is known as attention weights [3]. This computation of attention weights is based on the importance of source text that corresponds to the target text generated, which is calculated below as equation (1):

$$e_{i,j} = align(z_{i-1}, h_j), \qquad \forall j \in 1,2,\dots T, \forall j \in 1,2,\dots T' \quad \dots (1)[3]$$

where $e_{i,j}$ is the alignment score of j-th source word corresponding to i-th target word, z_{i-1} is the last state of the decoders, h_j is the annotation vector of the j-th source word and T, T' are the length of the source and the target text respectively [3].

The alignment scores are converted into probabilistic measures, which are known as attention weights ($\alpha_{i,j}$) which is calculated below as equation (2):

$$\alpha_{i,j} = \frac{exp(e_{i,j})}{\sum_{j'} \exp(e_{j',i})} \quad \dots (2)[3]$$

Using the annotation vector (c_i), the context vector is calculated below as equation (3):

$$c_i = \sum_{j=1}^{T} \alpha_{i,j} \times h_j \quad \dots (3)[3]$$

The next state decoders can be calculated as a non-linear function of context vector c_i, previous target word μ_{i-1} and decoders last state z_{i-1} , which is expressed in equation (4):

$$z_i = f(c_i, \mu_{i-1}, z_{i-1}) \quad \dots (4)[3]$$

The model had been trained as encoder inputs and decoder inputs which involves optimized compilations involving the parameter accuracy along with defining the validation loss. We have trained the model with the batch size 128 and epoch value as 100.

7) *Step 7 - Data Output Model:* To check whether model is working as expected, we are added a provision in backend so that the model performs the comparison of actual output text (provided by the model) and predicted output text (as per the mappings used in dataset), when user gives the input text at source.

B. Bi-Directional LSTM

The bi-directional LSTM involves the steps such as data input layer, data preprocessing and steps after data preprocessing which also includes classification of data into training and testing as mentioned in above methodology. The encoding for source text involves embedding layer of 300 dims with return state and backward training as this is bi-directional model. For decoding it into target text, embedding layer of 300 dims with return sequence and return state had been enabled.

The model had been as encoder inputs and decoder inputs which involves optimized compilations involving the parameter accuracy along with defining the validation loss, using the batch size of 128 and epochs 100. Based on these training the model is built for further output model testing.

C. Sequential LSTM as Language Detection

The following are the steps involved for training the model for Language detection, represented in fig.4 below:

1) *Step 1 - Data Input Model:* We had loaded the same dataset of English and Hindi in two different files. We have created list of characters and loaded in each dictionaries respectively.
2) *Step 2 - Classifying the Data for training and testing:* The data is divided in the ratio 4:1 for training and testing respectively. Here, as per training, the length of English and Hindi sentences is being used for getting sentence in next step.
3) *Step 3 - Get Sentences for Training and Testing:* In this case we have created two arrays respectively which takes sentence in 1st array and 2nd array is used for getting the next characters. This iteration will be performed till the end of the sentence. Here, the maximum length for the input is taken as 5 for training and testing purposes. This set of sentences and next string of sentences are stored in the respective 1st and 2nd arrays.

©IJRASET: All Rights are Reserved

4) *Step 4 - Get Vectors for Detection:* Here we define X and Y (X is for training and Y is for testing) and passes the parameters of 1st and 2nd array with 3rd parameter as the Boolean value for detecting language. This depends on number of sentences, characters, character indices and the upcoming next character. The shape of the English and Hindi word/sentence are analyzed and then it is tokenized into vectors accordingly.
5) *Step 5 - Building the Model:* The Sequential LSTM model is used for training. The batch size is 128, epoch value is 100 and input shape is used for building LSTM model. The sentences had been trained based on the parameter's accuracy and validation loss.
6) *Step 6 - Prediction of Input text Language:* After training the model, the model can predict the language given by the user, provided the user needs to enter the data from the dataset.

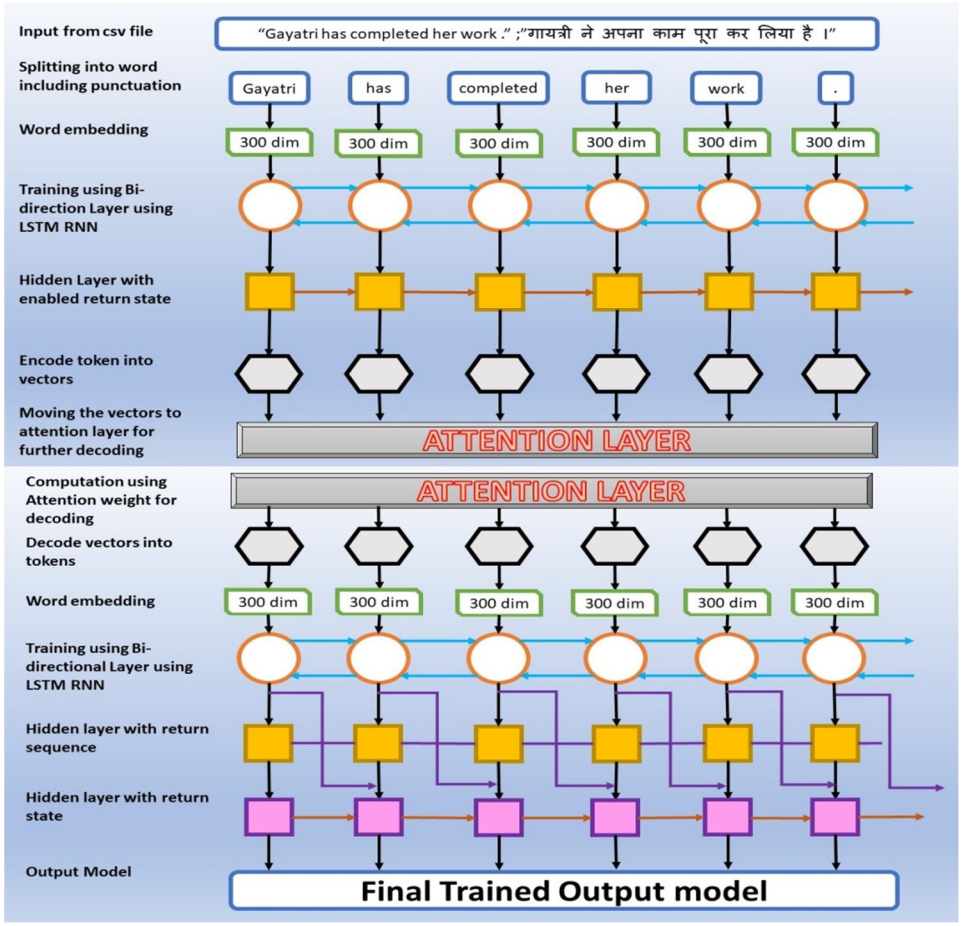


Fig. 3 Flow of Bi-Directional LSTM model.

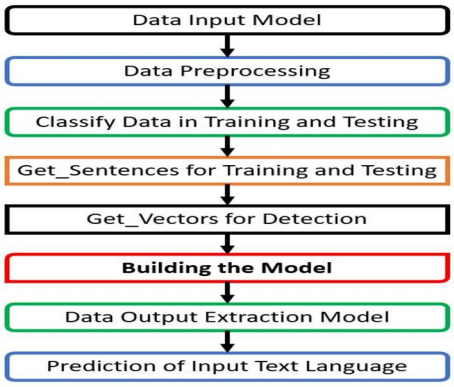


Fig. 4 Flow of Sequential LSTM model as Language Detection.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

IV. IMPLEMENTATION

These models had been trained and built using GoogleCollab (as we have trained the dataset around 40,000 records) and tested in Anaconda3 with Python 3.6, Tensorflow for LSTM models, sklearn for training and testing data split, Numpy for performing Linear Algebra, Pandas for data processing from CSV file, Tkinter for GUI. As per hardware details, RAM used is 8 GB with processor IntelCore i5 10th Generation.

V. RESULT AND ANALYSIS

Our research quality analysis is based on two parameters:

1) *Accuracy:* The accuracy is calculated based on training of model and its predicting ability for each word/sentence
2) *Validation Loss:* This loss is calculated on training and validation. Its interpretation is based on how well the model is doing in these two sets.

On performing different analysis using the words and statements with bi-directional model, our model had also been compared with Python standard module based on accuracy as parameter and the below are the overall analysis found:

A. Datasets

The dataset used is in CSV format in which we have added English word and sentence with their respective Hindi word and sentence respectively. This data is from different textbooks and education sites. The same dataset has been used while training model for language detector.

```
847  ted,You can't get away with this!,"तुम इससे बच नहीं सकते (सकती)!"
848  ted,I shall have to complain about you!,"मुझे तुम्हारी शिकायत करनी पड़ेगी!"
849  ted,Warm greetings,"नमस्ते"
850  ted,Warm greetings,"नमस्कार"
851  ted,Congratulations,"बधाई"
852  ted,Congratulations,"बधाइयां"
853  ted,Congratulations,"बहुत-बहुत बधाइयां"
854  ted,God bless you,"खुश रहो"
855  ted,Happy Birthday,"जन्मदिन की शुभकामनाएं"
856  ted,Happy New Year,"नया साल मुबारक"
857  ted,Happy New Year," नए वर्ष की बधाइयां"
858  ted,See you soon,"फिर जल्दी मिलेंगे"
859  ted,At the Bank,"बैंक में"
860  ted,I'd like to exchange one hundred dollars.,"मुझे सौ डॉलर के रुपए चाहिए।"
861  ted,Where should I present this cheque?,"यह चेक किस काउंटर पर देना होगा?"
862  ted,I want to open a new account. Please give me forms.,"मैं नया खाता खोलना चाहता (चाहती) हूं। फार्म चाहिए।"
863  ted,I want one-rupee notes.,"मुझे एक रुपए के नोट चाहिए।"
864  ted,I want to encash the traveller's cheque.,"मुझे ट्रैवलर चेक भुनाना है।"
865  ted,These signatures are mine.,"ये दस्तखत मेरे हैं।"
866  ted,I want to see the Agent.,"मुझे एजेंट से मिलना है।"
867  ted,Do you have a branch office of your bank at Agra?,"क्या आगरा में आपके बैंक की शाखा है?"
```

Fig 5. Dataset used for training and testing the model in CSV.

B. Training of Models

The training of the models is done using Google Collab. The fig. 6 and 7 represents the inner layer, hidden layer and LSTM trained output details for the Forward LSTM with mask zero model and Bi-Directional LSTM model.

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, None)]	0	
input_2 (InputLayer)	[(None, None)]	0	
embedding (Embedding)	(None, None, 300)	7406700	input_1[0][0]
embedding_1 (Embedding)	(None, None, 300)	8497200	input_2[0][0]
lstm (LSTM)	[(None, 300), (None,	721200	embedding[0][0]
lstm_1 (LSTM)	[(None, None, 300),	721200	embedding_1[0][0] lstm[0][1] lstm[0][2]
dense (Dense)	(None, None, 28324)	8525524	lstm_1[0][0]

```
Total params: 25,871,824
Trainable params: 25,871,824
Non-trainable params: 0
```

Fig. 6 Training Forward LSTM with mask zero in embedding layer.

©IJRASET: All Rights are Reserved

```
Layer (type)                 Output Shape              Param #     Connected to
==================================================================================================
input_3 (InputLayer)         [(None, None)]            0
__________________________________________________________________________________________________
input_5 (InputLayer)         [(None, None)]            0
__________________________________________________________________________________________________
embedding_2 (Embedding)      (None, None, 300)         7406700     input_3[0][0]
__________________________________________________________________________________________________
embedding_4 (Embedding)      (None, None, 300)         8497200     input_5[0][0]
__________________________________________________________________________________________________
lstm_2 (LSTM)                [(None, 300), (None,      721200      embedding_2[0][0]
__________________________________________________________________________________________________
lstm_4 (LSTM)                [(None, None, 300),       721200      embedding_4[0][0]
                                                                   lstm_2[0][1]
                                                                   lstm_2[0][2]
__________________________________________________________________________________________________
dense_2 (Dense)              (None, None, 28324)       8525524     lstm_4[0][0]
==================================================================================================
Total params: 25,871,824
Trainable params: 25,871,824
Non-trainable params: 0
__________________________________________________________________________________________________
```

Fig. 7 Training Bi-Directional LSTM model.

C. Results of Language Translator models

The table 1 represents the translation of each model as per input text. Here, model1 in the table represents the modified Forward LSTM i.e. with mask zero in embedding layer and model2 represents the Bi-Directional LSTM model.

Table I: Sentence Translation results

General Sentence	
Input English Sentence	Gayatri has completed her work.
Actual Hindi Sentence	गायत्री ने अपना काम पूरा कर लिया है ।
Predicted Hindi Sentence by Model1	गायत्री ने अपना काम पूरा कर लिया है ।
Predicted Hindi Sentence by Model2	गायत्री ने अपना **काम कर लिया** है ।
Predicted Hindi Sentence by Python Module	गायत्री ने अपना **काम पूरा किया** है।
Exclamation Sentence	
Input English Sentence	Hello! My name is Aishwarya.
Actual Hindi Sentence	नमस्ते! मेरा नाम ऐश्वर्या है।
Predicted Hindi Sentence by Model1	नमस्ते! मेरा नाम ऐश्वर्या है।
Predicted Hindi Sentence by Model2	नमस्ते! **सब**
Predicted Hindi Sentence by Python Module	नमस्कार! मेरा नाम ऐश्वर्या है।
Sentence with Numeric and punctuation	
Input English Sentence	Mehmud gajjni[971-1030] also declared himself as King.
Actual Hindi Sentence	महमूद ग़ज़नी [९७१-१०३०] ने अपने आप को तो सुल्तान भी घोषित कर दिया।
Predicted Hindi Sentence by Model1	महमूद ग़ज़नी [९७१-१०३०] ने अपने आप को तो **राजा** भी घोषित कर दिया।
Predicted Hindi Sentence by Model2	महमूद ग़ज़नी (971-1030) ने अपने आप को **राजा** घोषित किया।
Predicted Hindi Sentence by Python Module	महमूद गजनी [971-1030] ने भी खुद को **राजा** घोषित किया।
Interrogative Sentence	
Input English Sentence	What's the price of this shirt?
Actual Hindi Sentence	इस कमीज का दाम क्या है?
Predicted Hindi Sentence by Model1	इस कमीज की **कीमत** क्या है?
Predicted Hindi Sentence by Model2	इस **शर्ट** का **कीमत** क्या है ?
Predicted Hindi Sentence by Python Module	इस **शर्ट** की **कीमत** क्या है?

©IJRASET: All Rights are Reserved

In the table1 under exclamation sentence example, the model2 failed to translate the next statement. The overall expected translation sentence had been compared and highlighted accordingly. The fig. 8 represents the GUI model where we have tested our model.
The table 2 represents the parameter calculated for each model along with Python standard model used for translation.
The fig. 9 represents the comparative graphical representation of language model.

Table II: Parameter calculation and overall results of Language Translation model

Models with parameters	Accuracy	Validation Loss
Modified Forward LSTM	0.8428	0.422
Bi-directional LSTM	0.7157	0.522
Python standard model	0.5918	N/A

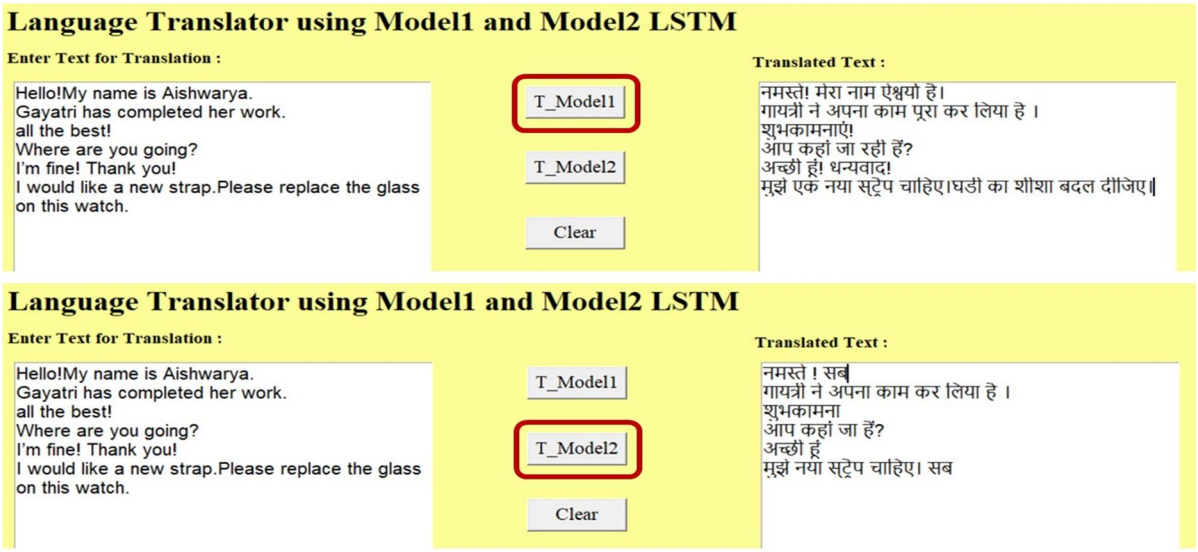


Fig 8. Language Translation using GUI.

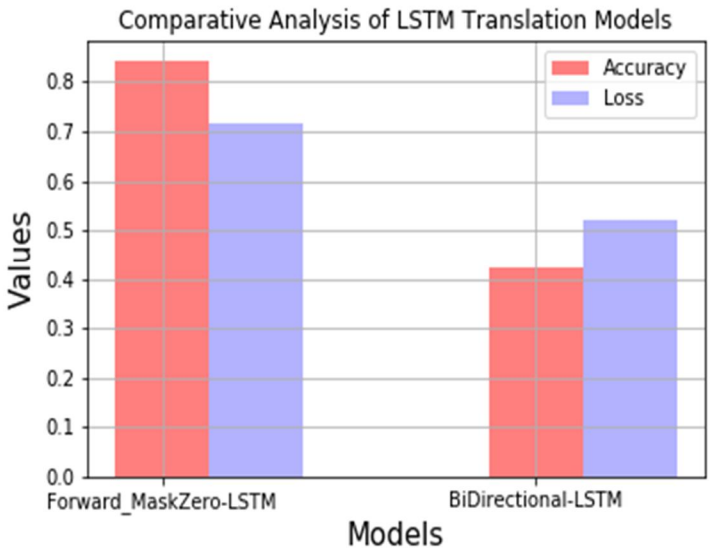


Fig 9. Comparative analysis of language translator models

©IJRASET: All Rights are Reserved

D. Result and Analysis of Language Detection Model

The table 3 represents the language detection according to model prediction and in GUI it is represented in fig. 10.
The table 4 represents the parameter calculated for each Sequential model for English and Hindi statement and the fig. 11 represents the comparative graphical representation of the detection model for each language.

Table III: Sentence Language Detection results

Input Sentence is English	
Input Sentence	My mother finally says to me 'I knew it!'
Predicted Language	English
Input Sentence is Hind	
Input Sentence	तालियाँ
Predicted Language	Hindi
Input Sentence is French	
Input Sentence	Merci
Predicted Language	Please enter English or Hindi Text
Input Sentence is Marathi	
Input Sentence	बाहुली
Predicted Language	Please enter English or Hindi Text

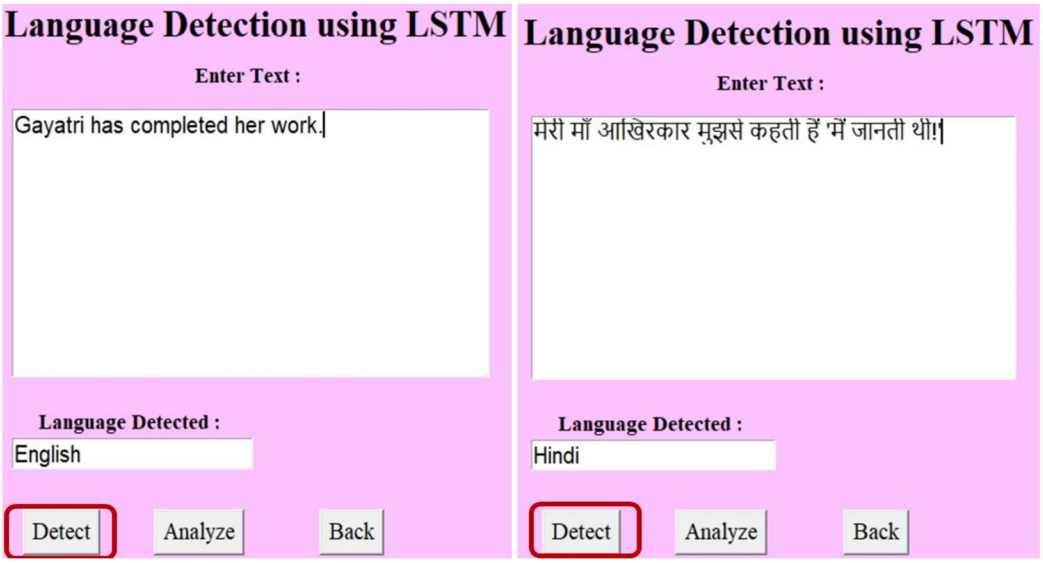


Fig 10: GUI representation of Language Detection model.

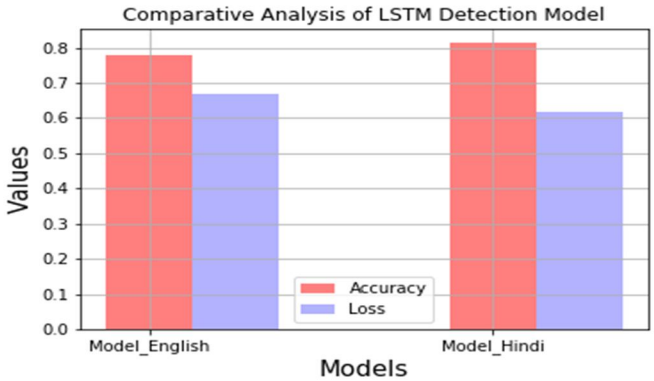


Fig 11. Comparative Analysis of LSTM Detection Model.

©IJRASET: All Rights are Reserved

Table IV: Parameter calculation and overall result of Language Detection model

Models with parameters	Accuracy	Validation Loss
English Detection model	0.7806	0.6685
Hindi Detection model	0.8147	0.6190

VI. DISCUSSION

It is important to retain the meaning of the sentence while translation, such that in future if the user is entering any sentence in the translator box, should provide the exact translated text that conveys the same meaning. India is full of diversity especially in culture, heritage, food, lifestyle and language. People are expressing their thoughts as a part of message in different languages, and to detect them it is important to identify in which language the message is written. The main application of this language translation and detection system is especially for social media and blogs where above nearby URL, it will display in which language the page is present, while if we want to translate for reading in our own comfortable language then we can select the desired language and allow to translate. For that, it is important to train the model in many different languages, taking care of the meaning should remain the same while translation.

We have used English and Hindi language to check the models if it is working fine as expected. We can train the model in more languages, provided that the language should be known by the implementer while implementing the model so that the translated text meaning can be checked and analyzed accordingly.

When our modified LSTM model were compared with another LSTM model, it was found that the implemented model is 13% more accurate than existing model, when trained along with punctuation. For language detection model, as each word is used for training sequentially, even in partial statement it was able to predict the language entered while testing the model.

Hence if it is trained using more dataset of different languages, then it can be used for different websites for better translation and detection model.

VII. CONCLUSIONS

Long-Short Term Memory (LSTM) model can be used for language translation and detection system. This implementation is purely based on how you train your model and apply the model as per need. We have focused on modified Forward LSTM model while we have compared with bi-directional LSTM model, which concludes that the former model performs better than the latter model. When we have trained the sequential LSTM using the same dataset, it is used as language detection model as each model is trained with each word and its subsequent word. Hence, it focuses on how the LSTM models can be used for as translation and detection system.

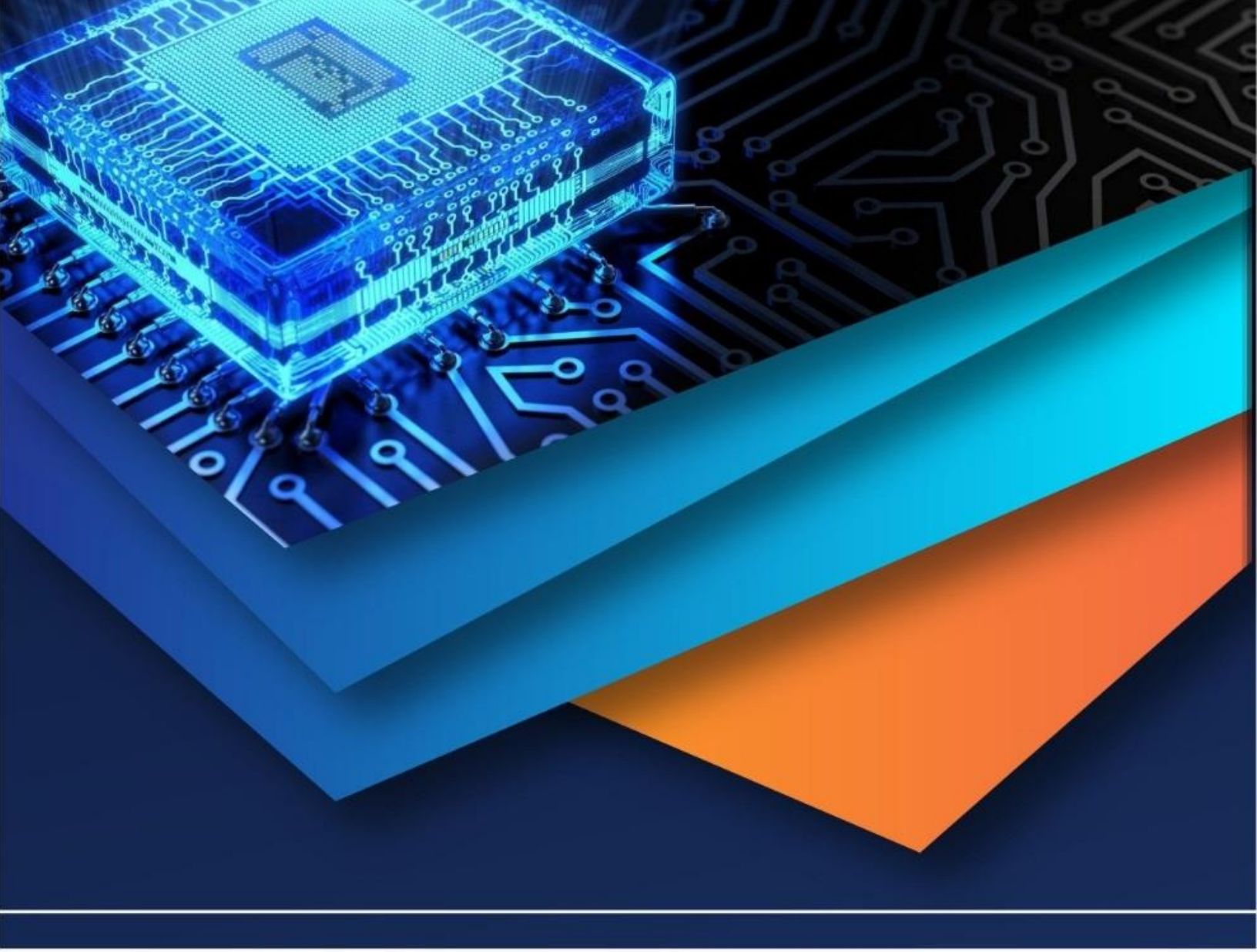
REFERENCES

[1] Aishwarya R. Verma, "Design and Development of Language Translation and Detection System using Machine Learning", International Conference on Intelligent Systems and Communication Networks (IC-ISCN 2019), pp. 303-305, 2019.
[2] Sandeep Saini, Vineet Sahula, "Neural Machine Translation for English to Hindi", 2018 Fourth International Conference on Information Retrieval and Knowledge Management.
[3] Sahinur Rahman Laskar, Abinash Dutta, Partha Pakray, Sivaji Bandyopadhyay, "Neural Machine Translation: English to Hindi", 2019 IEEE Conference on Information and Communication Technology (CICT).
[4] B. Premjith, M. Anand Kumar and K.P. Soman, "Neural Machine Translation System for English to Indian Language Translation Using MTIL Parallel Corpus", J. Intell. Syst. 2019; 28(3): pp. 387–398
[5] Afsana Parveen Mukta, Al-Amin Mamun, Chaity Basak, Shamsun Nahar, Md. Faizul Huq Arif, "A Phrase-Based Machine Translation from English to Bangla Using Rule-Based Approach", 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)
[6] Md. Saidul Hoque Anik, Md. Adnanul Islam, A.B.M. Alim Al Islam, "An Approach Towards Multilingual Translation Semantic-Based Identification and Root Word Analysis", 2018 5th International Conference on Networking, Systems and Security (NSysS)
[7] Jayshri A. Todase, Sushama Shelke, "Script Translation System for Devnagri To English", 2018 International Conference on Information, Communication, Engineering and Technology (ICICET)
[8] Kehai Chen, Tiejun Zhao, Muyun Yang, Lemao Liu, Akihiro Tamura, Rui Wang, Masao Utiyama, Eiichiro Sumita, "A Neural Approach to Source Dependence Based Context Model for Statistical Machine Translation", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 26, No. 2
[9] Jinsong Su, Jiali Zeng, Deyi Xiong, Yang Liu, Mingxuan Wang and Jun Xie, "A Hierarchy-to-Sequence Attention Neural Machine Translation Model", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 26, No. 3
[10] Carol Ebbert-Hübner, Clare Maas, "Can Translation Improve EFL Students' Grammartical Accuracy?", International Journal of English Language & Translation Studies. 5(4), pp. 191-202.
[11] Shamsun Nahar, Mohammad Nurul Huda, Md. Nur-E-Arefin, Mohammad Mahbubur Rahman, "Evaluation of Machine Translation Approaches to Translate English to Bengali", 20th International Conference of Computer and Information Technology (ICCIT), Dec-2017

©IJRASET: All Rights are Reserved

[12] Brenda Reyes Ayala, Jiangping Chen, "A Machine Learning Approach to Evaluating Translation Quality", JCDL'17: Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries.
[13] Madura Mandar Phadke, Dr. Satish R. Devane, "Multilingual Machine Translation: An Analytical Study", 2017 International Conference on Intelligent Computing and Control Systems (ICICCS 2017), Jun-2017
[14] Chandramma, Dr. Piyush kumar Pareek, Swathi K, Puneet Shetteppanavar, "An Efficient Machine Translation Model for Dravidian Language", 2nd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology (RTEICT), May-2017
[15] Omkar Dhariya, Shrikant Malviya, Uma Shanker Tiwary, "A Hybrid Approach for Hindi-English Machine Translation", 31st International Conference on Information Networking (ICOIN-2017), Jan-2017
[16] Gert De Sutter, Bert Cappelle, Orphée De Clercq, Rudy Loock, Koen Plevoets, "Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translation", Linguistica Antverpiensia New Series - Themes in Translation Studies, Vol. 16
[17] Sindhu D.V, Sagar B M, "Study on Machine Translation approaches for Indian Languages and their challenges", 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)
[18] Xinlian Hu, "Statistical Machine Translation Approach Based on Dependency-to-String Model", 2016 International Conference on Smart Grid and Electrical Automation
[19] Rosna P Haroon, Shaharban T A, "Malayalam Machine Translation Using Hybrid Approach", International Conference on Electrical, Electronic and Optimization Techniques (ICEEOT) – 2016
[20] Aasha V C, Amal Ganesh, "Machine Translation from English to Malayalam Using Transfer Approach", 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)
[21] Andi Hermanto, Teguh Bharata Adji, Noor Akhmad Setiawan, "Recurrent Neural Network Language Model for English-Indonesian Machine Translation: Experimental Study", 2015 International Conference on Science in Information Technology (ICSITech)
[22] Omar Sharif, Mohammed Moshiul Hoque, A. S. M. Kayes, Raza Nowrozy, Iqbal H. Sarker, "Detecting Suspicious Texts Using Machine Learning Techniques", Appl. Sci. 2020, 10, 6527; doi:10.3390/app10186527
[23] Inumella Chaitanya, Indeevar Madapakula, Subham Kumar Gupta, Thara S, "Word Level Language Identification in Code-Mixed Data using Word Embedding Methods for Indian Languages", IEEE 2018
[24] Sheeba Grover, Dr. Amandeep Verma, "Design for Emotion Detection of Punjabi Text using Hybrid Approach", 2016 International Conference on Inventive Computation Technologies (ICICT)

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)