Machine Translation

Santosh Cheruku
1113852
Masters in Comp. Sci.
Lakehead University
scheruk1@lakeheadu.ca

Veera Venkata Nagendra Murthy Chandu
1093805
Masters in Comp. Sci.
Lakehead University
chanduv@lakeheadu.ca

Naga Teja Guttikonda
1104370
Masters in Comp. Sci.
Lakehead University
nguttiko@lakeheadu.ca

Tirth Desai
0889915
Masters in Comp. Sci.
Lakehead University
tdesai1@lakeheadu.ca

***Abstract*—Abstract-Machine translation (MT) plays a significant role in helping linguists, sociologists, computer scientists, etc. to translate it into some other natural language through manipulating natural language. And this demand has grown exponentially over the past couple of years, given the tremendous exchange of information with different regional languages across different regions. Machine translation faces various problems, some of which are: a) Not all words in one language have corresponding words in another language. b) Two languages may have entirely different systems. c) Words may have more than one meaning. Because of these challenges, MT has been an active area of research for more than five decades, along with many others. In the past, numerous methods have been proposed which either aim to improve the quality of the translations they produce, or to test the robustness of these systems by testing their output in many different languages. In this literature review, we address statistical approaches (throughout particular word-based and phrase-based approaches) and neural approaches that have gained widespread popularity in multiple major languages due to their state-of - the-art outcomes. Sequence to Sequence model is used so as to do the process of execution with the Hindi ENcorp dataset.**

***Index Terms*—LSTM, Sequence-to-Sequence model, Corpus, Decoder, Statistical Machine Translation**

I. Introduction

Machine Translation is a sub-field of computational linguistics which aims to automatically translate text from one language to another using a computer device. Petr Petrovich Troyanskii was the first person to formally introduce machine translation to the best of our knowledge. Petr approached the Academy of Sciences in 1939 with plans for mechanical translation, but these ideas were never focused on barring preliminary discussions. Starting in the late 1980s, computational machine translation gained prominence and numerous word-based and phrase-based techniques were introduced that needed little to no linguistic knowledge. With the introduction of deep neural networks in 2012, a significant area of research has become the application of these neural networks in machine translation systems. Recently researchers reported using neural machine translation to achieve human parity on automated Chinese to English news translation. Although early machine translation systems were used mainly for the translation of scientific and technical documents, contemporary applications are extensive.

Computer translation for Indian languages faces many challenges. For example, I the size of parallel corpora and (ii) language differences, mainly morphological richness and difference in word order due to syntactic divergence are two of the major challenges. Indian languages (IL) both suffer from these problems, especially when translated from English. In addition, both word order and morphological complexity of Indian languages such as Tamil vary from English. English for example has Subject-Verb- Object (SVO) while Tamil has Subject-Object-Verb (SOV). English is also a fusional language whereas Tamil is an agglutinative language. While syntactic differences contribute to the translation model difficulties, morphological differences contribute to data sparsity. Corpus-based approach uses a large sized parallel corpus in the form of raw data. This raw data contains text with their respective translations. These corpora are used to acquire knowledge for translation. A corpus-based approach divides itself into two sub types: (i) statistical machine translation (SMT) and (ii) example-based machine translation (EBMT) (Somers, 2003). SMT2 generates its translation on the basis of statistical models. It depends on the combination of language model as well as translation model with a decoding algorithm.State-of - the-art machine translation (MT) systems, including both phrase-based statistical translation approaches and neural network-based translation approaches that have recently appeared, rely heavily on coordinated corporate parallel training. These parallel data, however, are expensive to obtain in practise and are therefore generally restricted in scope, which can restrict the related research and applications.

II. Literature Review

In the 1970s, the primary research emphasis was on Rule-based Machine Translation (RBMT). Those systems fall into one of the following three categories: Direct systems (this map input sentence directly to the output sentence), Transfer RBMT systems (these use morphological and syntactic analysis to translate sentences), and Interlingual RBMT systems (these translated the input sentence into an abstract representation and mapped the abstract representation to the final output). One such research is by Carbonell et al. in 1978 in the Interlingual RBMT method. The suggested solution transforms text by:

1) Translating the source text into a language-free conceptual representation,

2) Through that representation with details implied in the source text,

3) Converting that increased representation into the target

$$P(o_1, o_2, ..., o_n) = \prod_{n=1}^{n} P(o_i / o_{i-(n-i)}, ...o_{i-1})$$

Fig. 1. Formulae for Markov Chain

language.

The authors argue that translation requires a detailed understanding of the source text which semantic rules are inadequate to capture and therefore need to be supplemented with detailed domain knowledge.

A simple method for translating multiple languages using a single model, using multilingual data to boost NMT in all languages. The approach does not need to modify the conventional architecture of the NMT model. Instead we add an artificial token to the input sequence to signify the target language needed, a simple modification to the data only. The explicit bridging is used where we translate given text to an intermediate language, and it is further translated to target language without any parallel data. Training model is applied to sentence of each triple which is translated to two languages and it takes 9,987 steps corresponding to the context vectors. Tensor Flow embedding projector is used to map the context vector to the more accessible 3D space. The largest experiment is made where 12 language pairs are merged and only obtain low translation quality in one language among them.

A. *Statistical Machine Translation System for Indian Languages*

Statistical machine translation is used here for translating the given sentence from one language to another language. The system is trained by using the English and Telugu parallel corpus where the translation mainly depends on the quality of corpus. Data Preparation is carried out by tokenizing the corpus, Filtering out the sentences of long length, Conversion of given corpus into lower case. Even though SMT can translate the given sentence in plenty ways it mainly consists of Language Model, Translation Model and decoder.

Language model: It calculates the probability p(e) for the target language where it helps to achieve the adequacy and fluency for the translated text. LM helps us mainly for two purposes namely word order and word choice. Word order help us to find which word should precede the sequence of words in a perfect way. In Word choice there will be set of words for the translated text where we will choose the most accurate word. Probability in LM is calculated by using the n-gram model for estimating the contiguity of words. Decomposition of probability is performed by using the Markov chain rule as below:

Generally, the probability for the given sentence can be calculated by making the class conditional probability for each word. By n-gram the probability calculation of sentence is made very simple by estimating the probability of word by using its neighbours. Sometimes n-gram may be missed then its probability becomes zero to solve this we use phenomenon called count smoothing where it will add 1 to count of all possible occurrences. Translation model (TM): The conditional probability p(t/e) is estimated by taking the parallel corpus of both source and target languages. Here the translation is done for each word separately at word level rather than going at sentence level. Inside TM we use phrase-based translation model we divide the given sentence into set of phrases and each phrase is translated individually. The best translation level is reached by obtaining the maximum probability distribution inside TM. When distortion occurs in phrase reordering, we consider the distance for better position of words.

t argmax P(t/e) P(t) max = t

Fig. 2. Formulae for Markov Chain

III. APPROACHES FOR MACHINE TRANSLATION

Decoder: Finally, it is used to maximize the probability of translation by using the maximum likelihood parameter. Inside search space we consider all the possible translations and pick the best among them. Here the system is tested with the corpus of ten lakh words of text corpus from both English and Telugu.The adequate results has been obtained but still the accuracy will be high if we choose the large corpus for training. So, we can add more amount of data to the existing corpus and retrain it to for further improvement in the results. 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), 22-24 December, 2017

Various approaches used for machine translation are:

Direct Approach: Here the input text is directly converted to target language output text without any intermediate stages. It is based on morphological analysis and identification of constituents in between them.

Example: You are playing cricket

1. Morphological Analysis: You playing Present Continuous Cricket

2. Identify constituents: You, playing Present Continuous, cricket

3. Reorder them according to the target language: You, cricket, playing Present Continuous

Corpus Based Approach: Here two parallel corpora are present in both source and target language where the alignment of sentences is done. Next, we match the fragments against the parallel corpus and then the translation principle is applied. The main use of this approach is finding out senses of words and phrases in various contexts in a quick manner. It takes lot of time for the computation and the texts in corpora may cause problems in analysis

Transfer Based Approach: In this there are three phases namely analysis, transfer, and generation. Analysis involves finding sentence structure and constituents of the given

sentence
Ex: I eat chocolates
Here words are I, eat, chocolates

Sentence Structure: [subject] [verb] [object]

In transfer stage we apply the transformation to source language parse tree to obtain structure of target language. By using the transfer-based approach we obtain the high-quality translations. The drawback of this approach is sometimes it is not possible to show the word meaning properly and some words will be getting missed.

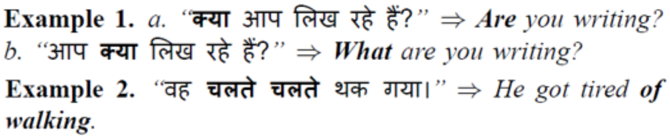
Here corpus-based approach provides the better results than the google translator and other implemented methods. The main aim of this work is to find out the best method which make its accuracy rate higher. Further work needs to be done detecting multiple Bengali meaning for an single English word.

A Hybrid approach for Hindi-English Machine Translation

An expanded combined approach of phrase-based statistical machine translation (SMT), example-based MT (EBMT) and rule-based MT (RBMT) is proposed in this paper to build a novel hybrid data-driven MT system capable of outperforming the baseline SMT, EBMT, and RBMT systems from which it is derived. In short, after having a set of partial candidate translations provided by EBMT and SMT subsystems, the proposed hybrid MT method is driven by the rule-based MT. Previous work has shown that EBMT systems are capable of outperforming phrase based SMT systems and the RBMT method has the power to produce more accurate results, both structurally and morphologically. This hybrid approach improves the fluidity, precision and grammatical accuracy that improve the quality of a machine translation system. A comparison of the hybrid machine translation (HTM) model being proposed with renowned translators i.e. Google, BING and Babylonian which indicates that the proposed model works better on both ambiguous and idiom-composed sentences than others.

A computer translator, at the basic level, literally consents by substituting word to word from source language to target language. But only the word replacement would not be able to deliver desired results since it doesn't know about the target language's semantic and syntactic constraints. Several freely available Hindi-English machine translation systems already exist, such as Google Translator, MS-Bing, and Babylon. Such systems are built according to various approaches i.e. Rule Based Machine Translation (RBMT), Example Based Machine Translation (EBMT), Statistical SMT. Yet not all of them are effective in addressing word sense challenges.

RBMT systems do the translation on the basis of rules discovered by linguists that say how terms, word sequences or any other structure from the source language would be translated into target language. The two systems EBMT and SMT automatically extract the rules themselves instead from the manually formed parallel corpora between source and target language. That's why they are called data-driven approaches. A new way of implementing the hybrid approach for machine translation (HMT) has been discussed in this paper that utilize the strength of EBMT, RBMT and SMT.



Example 1. *a.* "**क्या** आप लिख रहे हैं?" ⇒ ***Are*** *you writing?*
b. "आप **क्या** लिख रहे हैं?" ⇒ ***What*** *are you writing?*
Example 2. "वह **चलते चलते** थक गया।" ⇒ *He got tired **of walking**.*

Fig. 3. Translation of the text

The main challenge in machine translation (MT) is to recognise an intrinsic difference in translation between source and target languages.

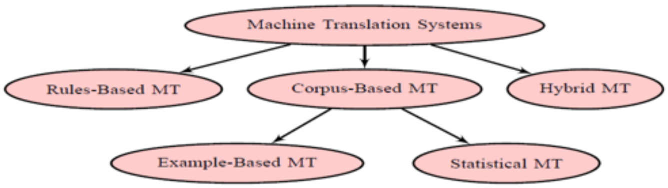


Figure 1. Various approaches of Machine Translations

Overall proposed approach has been implemented in a sequence of four primary steps: 1) Segmentation, 2) Translation, 3) POS Tagging and 4) Rearrangement. In figure 2, the flowchart depicts the working relation among multiple steps for translating a Hindi sentence to English "Vikas did development.". A inputted sentence has to go through all these steps being transformed from one form to another and at last translated to corresponding English sentence.

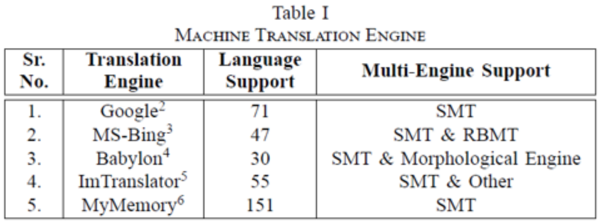


Table I
MACHINE TRANSLATION ENGINE

Sr. No.	Translation Engine	Language Support	Multi-Engine Support
1.	Google[2]	71	SMT
2.	MS-Bing[3]	47	SMT & RBMT
3.	Babylon[4]	30	SMT & Morphological Engine
4.	ImTranslator[5]	55	SMT & Other
5.	MyMemory[6]	151	SMT

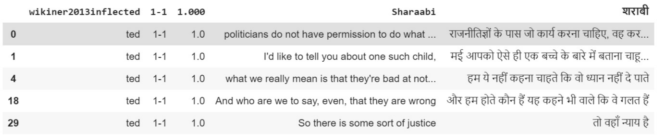
Figure. Machine Translation engine

Segmentation: Segmentation is done by finding first that all possible sub-parts in the sentence belong to the Hindi-English parallel database. The remaining parts of the sentence would later be divided into terms. Output would have collection of phrases, simple sentences, and words at the end of this point.

Proper Noun identification: According to POS, proper noun denotes a name used for an individual person, place, or organization. The problem is solved through incorporating some rules based on the possible contextual morphological information required to denote a word as pronoun.

Tagging: The Hindi-English parallel dictionary contains the tag for every English word. The system determines the right grammatical structure for the sentence, based on the tag assigned, and rearranges the words to create a grammatically correct sentence.

Translation: All the segments whether words or partial simple sentences are translated individually. Words are translated



	wikiner2013inflected	1-1	1.000	Sharaabi	शराबी
0	ted	1-1	1.0	politicians do not have permission to do what ...	राजनीतिज्ञों के पास जो कार्य करना चाहिए, वह कर...
1	ted	1-1	1.0	I'd like to tell you about one such child,	मई आपको ऐसे ही एक बच्चे के बारे में बताना चाहू...
4	ted	1-1	1.0	what we really mean is that they're bad at not...	हम ये नहीं कहना चाहते कि वो ध्यान नहीं दे पाते
18	ted	1-1	1.0	And who are we to say, even, that they are wrong	और हम होते कौन हैं यह कहने भी वाले कि वे गलत हैं
29	ted	1-1	1.0	So there is some sort of justice	तो वहाँ न्याय है

Fig. 4. Data Description

based on its assigned POS tag referring to parallel Hindi-English dictionary. If the Hindi word exist in the dictionary, the corresponding English word will be retrieved and tagged accordingly. On the other hand, if the word denoted as proper noun, it would be transformed to English by the Hindi-English transliteration. The words which represent many English words in translation, will be selected through learned SMT.

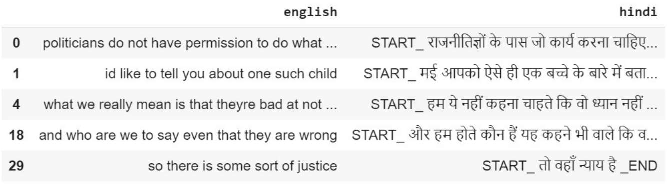
IV. Architecture of the Model

Encoder -Decoder Architecture:
An encoder-decoder architecture is a neural network design pattern mainly in natural language processing.
Encoder is a stack of several recurrent units (LSTM or GRU cells for better performance) where each one accepts a single element of the input sequence, gathers and propagates information for that element moving forward. Encoder vector is the final hidden state created from the model encoder section. This vector attempts to encapsulate the information for all elements of the input to help the decoder make predictions that are accurate.
Decoder: A stack of several repeating units where each predicts a y t output at phase t of a time. Each recurrent unit accepts a hidden state from the previous unit and produces and produces its own hidden state as well as production. In inference mode, i.e. when unknown input sequences are to be decoded, we go through a slightly different process:
1) Encode the input sequence into state vectors.
2) Begin with a size 1 target sequence (only the begin-of-sequence character).
3) Feed the decoder with the state vectors and 1-char target sequence to make predictions for the next character.
4) Use those predictions to analyse the next character (we simply use argmax).
5) Append the sampled character to the destination sequence
6) Repeat until the end-of-sequence character is created or the character limit is reached.
Model configuration: Epochs – 100 , Batch size : 128

V. Implementation

A. Data Description:

A) Dataset name : Hindi Encorp 0.5 B) No of rows in the dataset are 273,000 rows are initially there in the dataset.
C) In total 17 different sources are used to retrieve the data for the dataset.
D) Only TEDx talks source is used from the dataset for the process of execution.



	english	hindi
0	politicians do not have permission to do what ...	START_ राजनीतिज्ञों के पास जो कार्य करना चाहिए...
1	id like to tell you about one such child	START_ मई आपको ऐसे ही एक बच्चे के बारे में बता...
4	what we really mean is that theyre bad at not ...	START_ हम ये नहीं कहना चाहते कि वो ध्यान नहीं ...
18	and who are we to say even that they are wrong	START_ और हम होते कौन हैं यह कहने भी वाले कि व...
29	so there is some sort of justice	START_ तो वहाँ न्याय है _END

Fig. 5. Data Description

B. Preprocessing:

A) The dataset and its features are initialised from the process of execution. Moreover on the dataset process and the feature extraction. From the given dataset and sentences in dataset, all the upper case alphabets are changed into the lower case alphabets for better understanding of the dataset by the system for the better execution.
B) Removal of the digital values from the dataset is done like 0,1......n, from all the sentences.
C) Removal of the punctuation is done in the process execution of the cleaning of the dataset.
D) Empty spaces are removed for the better understanding of the phrases.
E) Removal of the spaces at the Initial start and the End of all the sentences.

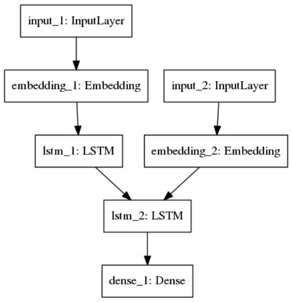
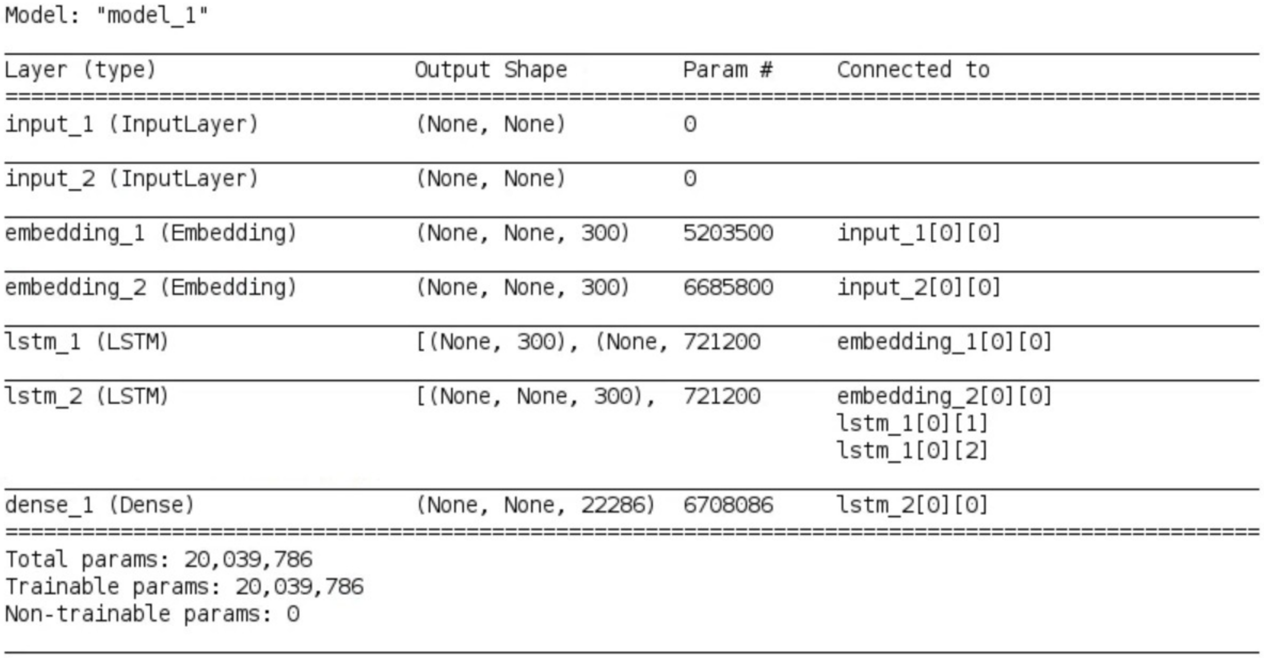


Fig. 6. Model Architecture

C. Target Output Language

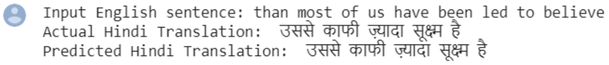
A) ”START” and ”END” is given at the starting and the ending of the sentence in the whole dataset.
B) Selection of the Unique words for the vocabulary from the both English and the Hindi language sentences.
C) Length of the each sentence is checked and the and the data is stored in the storage format for the future use.
D) If the length of the sentence is ¡20 then the sentence is taken into the consideration and the next part is removed from the dataset and is not furthur used used for the process.
E) For the Tokenization, unique words are assigned with the numerical values at all the instances and send for the dataset for all the unique words that was selected for the vocabulary.
F) Whole data that was prodeced till now is shuffled and the training is done on that part of the dataset.
G) Training and Testing of the dataset is split into 0.8 and 0.2 ratio.



```
Model: "model_1"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            (None, None)         0
__________________________________________________________________________________________________
input_2 (InputLayer)            (None, None)         0
__________________________________________________________________________________________________
embedding_1 (Embedding)         (None, None, 300)    5203500     input_1[0][0]
__________________________________________________________________________________________________
embedding_2 (Embedding)         (None, None, 300)    6685800     input_2[0][0]
__________________________________________________________________________________________________
lstm_1 (LSTM)                   [(None, 300), (None, 721200      embedding_1[0][0]
__________________________________________________________________________________________________
lstm_2 (LSTM)                   [(None, None, 300),  721200      embedding_2[0][0]
                                                                 lstm_1[0][1]
                                                                 lstm_1[0][2]
__________________________________________________________________________________________________
dense_1 (Dense)                 (None, None, 22286)  6708086     lstm_2[0][0]
==================================================================================================
Total params: 20,039,786
Trainable params: 20,039,786
Non-trainable params: 0
__________________________________________________________________________________________________
```

Fig. 7. Model Configuration

VI. Result



```
Input English sentence: than most of us have been led to believe
Actual Hindi Translation:  उससे काफी ज़्यादा सूक्ष्म है
Predicted Hindi Translation:  उससे काफी ज़्यादा सूक्ष्म है
```

Fig. 8. Result produced

A. Model used:

```
# Encoder
encoder_inputs = Input(shape=(None,))
enc_emb = Embedding(num_encoder_tokens,
    latent_dim, mask_zero =
    True)(encoder_inputs)
encoder_lstm = LSTM(latent_dim,
    return_state=True)
encoder_outputs, state_h, state_c =
    encoder_lstm(enc_emb)
encoder_states = [state_h, state_c]
decoder_inputs = Input(shape=(None,))
dec_emb_layer = Embedding(num_decoder_tokens,
    latent_dim, mask_zero = True)
dec_emb = dec_emb_layer(decoder_inputs)

decoder_lstm = LSTM(latent_dim,
    return_sequences=True, return_state=True)
decoder_outputs, _, _ = decoder_lstm(dec_emb,
initial_state=encoder_states)
decoder_dense = Dense(num_decoder_tokens,
    activation='softmax')
decoder_outputs =
    decoder_dense(decoder_outputs)
model = Model([encoder_inputs,
    decoder_inputs], decoder_outputs)
```

VII. Conclusion

Therefore we implemented a machine translation system with the Sequence to Sequence model using LSTM. In future, with bigger Corpus and a bigger configuration of the system there can be a better output with a better accuracy.

References

[1] B.N.V Narasimha Raju, M S V S Bhadri Raju, Statistical Machine Translation System for Indian Languages, IEEE 6th International Advanced Computing Conference, IEEE Access, 2016

[2] Akshat Joshi, Kinal Mehta, Neha Gupta, Varun Kannadi Valloli, INdian Language Transliteration using Deep learning, 2018 IEEE Recent Advances in Intelligent Computational Systems, IEEE 2018

[3] Omkar Dhariya, Shrikant Malviya and Uma Shanker Tiwary, "A Hybrid Approach For Hindi-English Machine Translation", 2017 International Conference on Information Networking (ICOIN), IEEE, 2017

[4] B N V Narasimha Raju et.al, "Translation Approaches in Cross Language Information Retrieval," In Proceedings of International Conference on Computer and Communications Technologies (ICCCT), publisher is IEEE, 2014.

[5] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Greg Corrado, Macduff Hughes, Jeffrey Dean, Martin Wattenberg,' Transactions of the Association for Computational Linguistics", vol. 5, pp. 339–351, 2017.

Contribution of the students :
Santosh Cheruku (1113852) - 30%
Veera Venkata Nagendra Murthy Chandu (1093805) - 30%
Guttikonda Nagateja (1104370) - 30%
Tirth Desai (0889915) - 10%