Neural Machine Translation for English to Hindi Using GRU

P Shalu
PG Scholar, Dept. of Computer Science
Government Engineering College Idukki
Kerala,India
shaluashraf234@gmail.com

Meera M
Asst. Professor, Dept. of Computer Science
Government Engineering College Idukki
Kerala,India
meeram@gecidukki.ac.in

***Abstract*—Language translation helps people to communicate, share information and establish a worldwide relationship. Neural Machine Translation helps to build the performances since it translates text from one language into another. This paper generates a summary with a headline and also compares three Neural Machine Translation models based on different Techniques for English-Hindi language pairwise: Sequence Architecture with both encoder and decoder (1) Long Short Term Memory (2) Bidirectional Long Short Term Memory (BI-LSTM) Conditional Random Field (CRF) and (3) Gated recurrent units (GRUs) with attention mechanism applied in three models. The comparison showed that GRU is better in performances than LSTM and Bi-LSTM CRF.**

***Index Terms*—Neural machine translation, Text summarization, Natural language processing, Bidirectional Long Short Term Memory (Bi-LSTM), CRF, Gated Recurrent Unit.**

I. INTRODUCTION

Natural language processing (NLP) is a branch of computer science and artificial intelligence that focuses on creating systems that allow computers to interact with people who speak a variety of languages. Summarization is the subtopic of Natural language processing. Summarization is the method of breaking down a large volume of text into smaller chunks. There are two styles of summarization: Abstractive and Extractive. Extractive summarization generates a summary from phrases, terms, sentences, and other elements of the input text document, while abstractive summarization necessitates a deeper comprehension and rationale of the text and generates a summary without using the same words or sentences as the input text.

Neural Abstractive text Summarization used sequences to sequence models which used more. Nowadays so many different techniques are proposed for better performance and improved versions of sequences to sequences model which is capable of challenges. Automatic or neural machine translation comes out the most challenging AI tasks for the human language. Deep learning becomes advanced resulting in a variety of different research fields and application domains where here use it. Aim to provide a comprehensive empirical evaluation of different deep learning models for text language generation in this paper. The aim is to see whether Bi-LSTM like neural networks have something in common with natural language, which humans generate most frequently. As a consequence, this paper focuses on abstractive summarization methods to translate English to Hindi summarization and generate a Headline and also offers an overview of some of the most widely used techniques as well as some of the most recent applications focused on them.

In this section it present the dataset, the word embedding models with their configurations and neural network configurations that are utilized in this study.

A. Sequence to Sequence model

One of the best and simple Abstractive Text summarization is the sequence to sequence model[11]. The Encoder is inserting the input text one word at a time then it will pass through an embedding layer that transforms the word into distributed representation. This will combine using a multilayer neural network which contains a hidden layer generated after inserting the previous word for the first word in the text. The decoder takes as input from a hidden layer which is generated after insert in the last word of input text. The decoder will decode to generate the text summaries using a softmax and attention mechanism.

B. Bidirectional-Long Short Term Memory

Bidirectional recurrent neural networks allow both forward and backward information about the sequences. The input of bidirectional in two ways: Past to Future and from Future to Past. Compare to LSTM, Bi-LSTM shows very good results since it understands the context better.

C. Gated Recurrent Unit (GRUs)

Gated Recurrent Unit is an improved version of standard recurrent neural networks which solve the problem of vanishing gradient problem. GRUs uses an update gate and the reset gate. Depending on the two gates, decide what information will be getting in output.

The paper is split as follows: section 2 presents the related work on this field. Section 3 demonstrates the methodology. Section 4 demonstrates the results, compares the different methods between them, and discusses the findings. Finally, section 5 concludes the paper.

II. RELATED WORKS

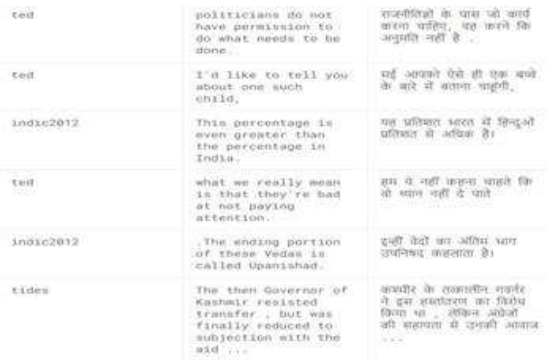
Abu Kaisar Mohammad Masum[1], In this paper author introduced Abstractive method of text summarization with sequence to sequence RNNs[12] where it translate English to English text summary using the encoder and decoder with LSTM. The main limitation is summary will be correct only for short text. In case the long text summary will be incorrect. Haijun Zhang[2], In this paper, the author introduced Understanding Subtitles by the character-Level Sequence-to-Sequence Learning. The author has analyzed and compared the performances in English to Chinese subtitle translation and it's embedded an RNN into the encoder-decoder approach for generating the character level sequence representation. It can also be improved by GRU in the language model of the encoder.Konstantin Lopyrev[3], The author of this paper implemented Recurrent Neural Networks for Generating News Headlines. For creating new headlines using text, [13] an encoder-decoder recurrent neural network with LSTM units and an attention mechanism was used. For the limited number of neurons that the attention weights are calculated using a simple attention mechanism.Jianpeng Cheng, Mirella Lapata[4], In this paper, the author has proposed Neural Summarization by Extracting Sentences and Words where data-driven approach based on neural networks and continuous sentence features. The general framework which used in single document summarization is composed of a hierarchical document encoder and the attention mechanism based on an extractor. Mahmood Yousefi-Azar , Len Hamey[5], In this paper, the author introduced Text summarization using unsupervised deep learning, the method [14] of extractive query-oriented single document summarization using a deep auto-encoder to find feature space from the term-frequency input. The main limitation is the computational cost of training. Shayak Chakraborty[6], In this paper, the author introduced the Study of Dependency on the number of LSTM units for Character-based Text Generation models where it increases LSTM cells and also increases the semantic relationship between the characters. The limitation is small corpus language character-based text generation is not good. The solution is a Neural network with an average number of LSTM cells. Sandeep Saini[7], In this paper, the author introduced Neural Machine Translation for English to Hindi requires a very less amount dataset for training. It exhibits satisfactory translation for a few thousand training sentences as well. The main limitation is not good in sentences using smaller data sets. Shashi Pal Singh[8], In this paper author, introduced Bilingual Automatic Text Summarization Using Unsupervised Deep Learning. Here is analyzed and compared the performances in two languages Hindi and English using an unsupervised deep learning approach. It extracts the eleven features from each sentence of the document and generates the feature matrix that is passed through the Restricted Boltzmann Machine. The main limitation is it will work only in multiple document summarization. Shengli Haitao Huang Tongxiao Ruan[9], In this paper the author introduced Abstractive text summarization using LSTM-CNN based deep learning. Here extractive text summarization model is concerned with syntactic structure and the abstractive text summarization model is concerned with semantics. Su Zhao, Encong Deng[10], In this paper the author introduced Generating summary using the sequence to sequence model. This paper deals with the problem existing in generating abstracts. The limitation is the need to add more weight parameters and also increase training time.

III. METHODOLOGY

This section presents the dataset, the word embedding models with their configurations, and neural network configurations that are utilized in this study.

A. Dataset

The Neural machine translation for the English to Hindi dataset is available at https://www.clarin.eu/resource-families/parallel-corpora where the site contains different language datasets. The dataset consists of 124318 English sentences and 97662 Hindi sentences. Tensor Flow framework has been used for the implementation task for the performances. Figure1 shows the English-Hindi-Truncated-corpus datasets. The dataset will be split the data into Training, Validation, and Testing purposes.



ted	politicians do not have permission to do what needs to be done.	राजनीतिज्ञों के पास जो कार्य करना चाहिए, वह करने कि अनुमति नहीं है .
ted	I'd like to tell you about one such child,	मई आपको ऐसे ही एक बच्चे के बारे में बताना चाहूंगी,
indic2012	This percentage is even greater than the percentage in India.	यह प्रतिशत भारत में हिन्दुओं प्रतिशत से अधिक है।
ted	what we really mean is that they're bad at not paying attention.	हम ये नहीं कहना चाहते कि वो ध्यान नहीं दे पाते
indic2012	.The ending portion of these Vedas is called Upanishad.	इनमें वेदों का अंतिम भाग उपनिषद कहलाता है।
tides	The then Governor of Kashmir resisted transfer , but was finally reduced to subjection with the aid ...	कश्मीर के तत्कालीन गवर्नर ने इस अंतरण का विरोध किया था , लेकिन अंतत: वो सहायता से उनकी आधीन ...

Fig. 1. Dataset

The Figure 2 shows the System Architecture consists of Encoder parts, Bahdanau Attention Mechanism and the decoder which translate English to an respective Hindi Translation using an GRU.

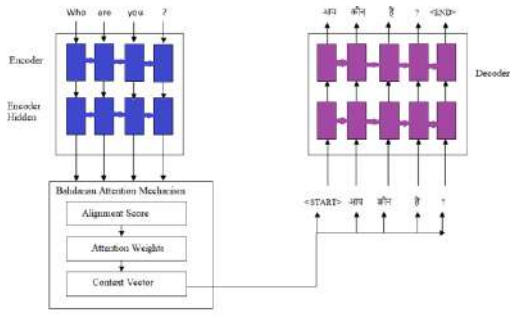


Fig. 2. System Architecture

B. Preprocessing

To boost the system's training performance, the first step is preprocessing which prepares and cleans the dataset and reduces the noise of the dataset. This task included the conversion of all tweets into lowercase, removing special characters, remove stop words, contraction mapping, removing whitespaces and tabs. The second step is tokenization which means that a function that breaks down a sentence into words. Build Tokenized for the review and summary. Among 80% are going to be for the training stage and 20% for the prediction stage are going to be taken for it.

C. Training of Neural Machine Translation

The goal is to translate English - Hindi language and generate a summary with a headline. Deep Learning's recent success in Natural Language Processing has inspired using DL to translate the language. While compare's the performances with three neural machine translations based on different Techniques for English-Hindi language pair-wise: Long Short term memory (LSTM), Bidirectional Long short term memory(Bi-LSTM) -Conditional random field(CRF), and the Gated Recurrent Unit(GRUs).

3.1 Word Embedding: The word embedding models used in Word2Vec, and GloVe. The Word2Vec model is used to create 25-dimensional word vectors that support the dataset described before. The Word2Vec was done by using the CBOW model. Additionally, words that appeared less than five times were discarded. Finally, the utmost skip length between words was set to 10. The encoder vector is the final hidden state from the encoder part. The vector will encapsulate all the elements in the information to help the decoder to make predictions accurately.

3.2 Bahdanua Attention Mechanism

In the way of passing the input sequence to the encoder, a secret state/output will be created for each information passed in. Rather than utilizing just the secret state/output at the last time step, it forwards every one of the secret states /output created by the encoder to the next step. After the process, it will calculate the alignment score of each encoder output with decoder input and hidden state/output. Here the alignment scores for Bahdanau Attention are calculated using the previous decoder hidden state and encoder hidden states. The alignment vector will give the weight to the encoder's output. The encoder's hidden state and its alignment score will be multiplied to get a context vector. The context vector will decide the final output of the decoder. The Prediction stage contains Preprocessing, Tokenization, and three models that are encoder inferences, Attention Inferences, Decoder inferences, and generate the summary. New Text will be preprocessed and get tokenize. To create representation gives to encoder Inferences, Attention Inferences, and the Decoder inferences. The most used one will be selected and new text will be updated but the update will be done up to a given limit.

IV. EVALUATION RESULT

ROUGE means Recall-Oriented Understudy for Gisting Evaluation. It is essentially a set of metrics for evaluating automatic summarization of texts and also for machine translation. It works by automatically produced a summary or translation by a set of reference summaries. ROUGE indicators for precision and recall objectively demonstrate that the summary result and original IP documents have high similarity and are consistent. The equation for Precision and Recall are given below and performances shown in table and chart:

$$\text{Precision} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

TABLE I
EVALUATION RESULT

Methods	Precision	Recall
LSTM	0.72	0.69
Bi-LSTM	0.81	0.77
CRF	0.82	0.79
BILSTM-CRF	0.87	0.83
GRU	0.92	0.90

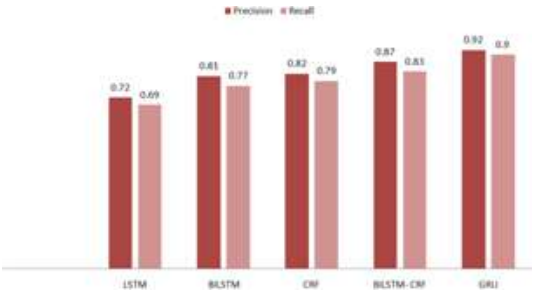


Fig. 3. Flow Chart

V. CONCLUSION

A modern Neural Machine translation offers a way for researchers to translate one source language to a target language which helps around the world. In this paper, proposed to apply word embedding, which trains a large number of the dataset and based on it trains the deep neural network with GRU. Here is a Bahdanau mechanism used for better usage. NLP and Machine learning techniques help to generate the summarization and Headline. Here also compare the different methods and contain better performances. As a future work with different attention mechanisms and also a method that will improve the accuracy of the model.

REFERENCES

[1] Abu Kaisar Mohammad Masum; Sheikh Abujar; Md Ashraful Islam Talukder; A.K.M. Shahariar Azad Rabby; Syed Akhter Hossain, "Abstractive method of text summarization with sequence to sequence RNNs", 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

[2] Haijun Zhang, Jingxuan Li, Yuzhu Ji, and Heng Yue," Understanding Subtitles by Character-Level Sequence-to-Sequence Learning", IEEE Transactions on Industrial Informatics (Volume: 13, Issue: 2, April 2017).

[3] H. W. Lin and M. Tegmark,"Critical Behavior from Deep Dynamics: A Hidden Dimension in Natural Language," arXiv preprint arXiv:1606.06737, 2016.

[4] Jianpeng Cheng ,Mirella Lapata, "Neural Summarization by Extracting Sentences and Words", ACL2016 conference paper with appendix, Computation and Language (cs.CL).

[5] Mahmood Yousefi-Azar, Len Hamey "Text summarization using unsupervised deep learning", Expert Systems with Applications, 2017 – Elsevier.

[6] Shayak Chakraborty; Jayanta Banik; Shubham Addhya; Debraj Chatterjee , "Study of Dependency on number of LSTM units for Character based Text Generation models", 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)

[7] Sandeep Saini, Sandeep Saini, "Neural Machine Translation for English to Hindi", 2018 Fourth International Conference on Information Retrieval and Knowledge Management.

[8] Shashi Pal Singh, Ajai Kumar, Abhilasha Mangal, Shikha Singhal, "Bilingual Automatic Text Summarization Using Unsupervised Deep Learning", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016.

[9] Shengli Song Haitao Huang Tongxiao Ruan , "Abstractive text summarization using LSTM-CNN based deep learning", Multimedia Tools and Applications volume 78, pages857–875(2019), Springer.

[10] Su Zhao, Encong Deng, Mengfan Liao, Wei Liu , Weiming Mao , "Generating summary using sequence to sequence model", 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC 2020).

[11] https://machinelearningmastery.com/encoder-decoder-models-text-summarization-keras/

[12] https://meta-guide.com/data/data-processing/summarization/text-summarization-2019

[13] https://www.researchgate.net/publication/286302083 Generating News Headlines with Recurrent Neural Network

[14] https://medium.com/analytics-vidhya/bengali-abstractive-text-summarization-using-s