Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 3- Issue 2, Paper 5, August 2020

NEURAL MACHINE TRANSLATION SYSTEM FOR MARATHI AND THAI

Poorvikha Ramesh Babu[1], Sasivimol Buakampan[2], Srividhya P[3], Dr. Radha Senthilkumar[4],
[1]*rpoorvikha@gmail.com*

Abstract*—Neural Machine Translation (NMT) is a technique that learns from provided dataset and predicts the likelihood of words. It is based on conditional probability of translating a given input to target language. The proposed model is a neural machine translation system for Marathi and Thai that contains encoder-decoder with Bahdanau attention mechanism. The performance of the system is compared with google translate using performance matrices like BLEU score and loss between training and testing phase.*

I. INTRODUCTION

Long Short-Term Memory (LSTM) is widely used in Natural Language Processing (NLP). In machine translation, using LSTM in encoder and decoder layers help the translation model is able to keep longer context. Sometimes text generated from the LSTM translation model can be meaningless since LSTM does not observe a whole sentence when performing translation tasks.

To overcome this problem, an attention layer is introduced. An attention layer lets the model observe the sentence even while translating so that the model can produce the translated sentence that has the same sentence structure as the target language.

In this paper, the proposed module combines the attention mechanism and Long Short-Term Memory (LSTM) together and generates a translation model. The generated translation model gets an input of text in English and generates the Marathi or Thai version of the input text.

The model uses Marathi and Thai to test how well the generated translation model can deal with different types of language structure because both languages have different language structure: in a sentence, Marathi words are separated by using spaces between each word, on the other hand Thai words are not separated.

II. LITERATURE SURVEY

Parth Shah and Vishvajit Bakrola have proposed "Neural Machine Translation System of Indic Languages-An attention based approach (2020) where they have taken Gujarati dataset and compared their performance with GNMT with evaluation matrices such as BLEU, perplexity and TER matrix. The encoder and decoder in their proposed model have two LSTM layers with 128 units of LSTM cells. Xu Tan, Yi Ren, Di He,TaoQin,Zhou Zhao and Tie-Yan Liu have proposed Multilingual Neural Machine Translation with Knowledge Distillation(2019), initially individual models are trained separately. The values are compared with the ones produced by the multilingual model. Multilingual model matches the output of individual models simultaneously through knowledge distillation.

Roee Aharoni, Melvin Johnson and Orhan Firat have proposed Massively Multilingual Neural

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 3- Issue 2, Paper 5, August 2020

Machine Translation(2019) where they compare the different setups for training models like low resource setting: many to one, one to many, many to many and similarly for high resource setting and study the trade-offs between translation and modelling decisions.

III. ARCHITECTURE

A. Data Cleaning

The data set is in raw format hence it needs to be transformed into appropriate format suitable to apply for machine learning algorithms.The raw text file contains the English sentences and the target sentences in two different columns. The English and target language sentences are separated into individual list.Then all the uppercase characters are converted to lower case,all the punctuation marksand all the unnecessary characters are removed. The data is then split into training and testing set.

Also, since seq-to-seq model require equal length input sequences, the input is padded with extra 0s

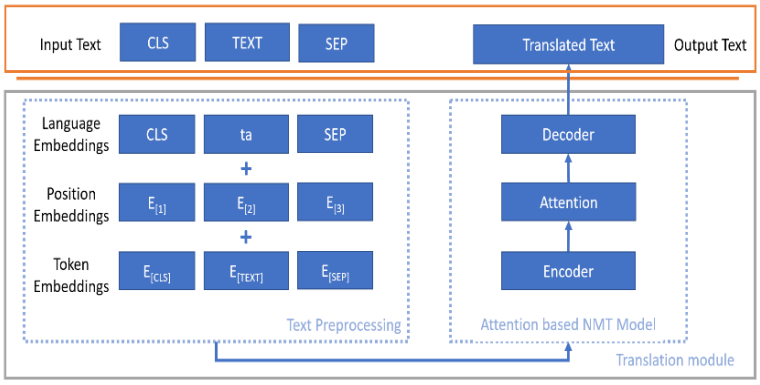

Fig. 1 Flow diagram for NMT translation module

The Attention mechanism used in this project is Bahdanau Attention mechanism. Traditional encoder-decoder model works well for small sentences but while translating large documents, attention mechanism is required in the middle.

The source language sentence is passed through the encoder and it is transformed into vector, the vector is then mapped to the target language using alignment matrix, for each translation the words are compared with different alignment matrix and the suitable translation is obtained.

IV. EVALUATION MATRICES

To compare the performance of the system two evaluation matrices have been considered.

A. BLEU (Bilingual Evaluation Understudy Score)

It uses the basic concept of n-gram precision to calculate the similarity between reference and generated sentence.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 3- Issue 2, Paper 5, August 2020

B. Loss difference between training and testing phase

Loss is calculated on the training set and validation set, its operation is based on how well the model is doing in the two sets. The value depicts how the model behaves after each iteration.

V. IMPLEMENTATION

A. Datasets

In order to work and test the NMT system with attention mechanism and get good result after training the model, large dataset is required. The English-Marathi dataset contains 38696 sentences and the English-Thai dataset contains 9000 sentences.The Marathi dataset is available in https://www.manythings.org/anki/ website and the Thai dataset is available in https://airesearch.in.th/releases/machine-translation-datasets/ website.

VI. RESULT AND DISCUSSION

In the proposed model the EarlyStopping was achieved at the 17th epoch for the Marathi dataset and 43rd epoch for the Thai data set.

Language	BLEU score of Google Translate (8 Layers)	BLEU score of proposed model (3 Layers)
Marathi	60.16	61.27
Thai	48.13	46.96

Table 1 BLEU score comparison between GNMT and proposed model for Marathi and Thai

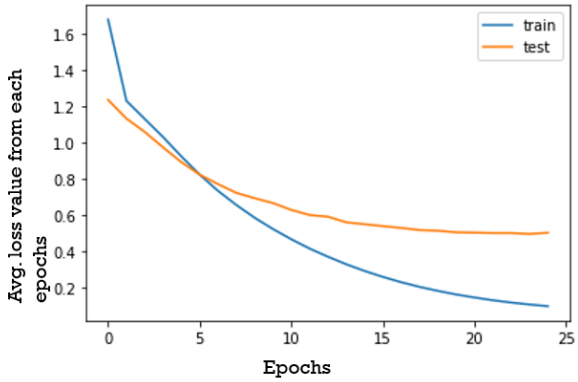

Fig. 2 Loss between training and testing phase for Marathi

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 3- Issue 2, Paper 5, August 2020

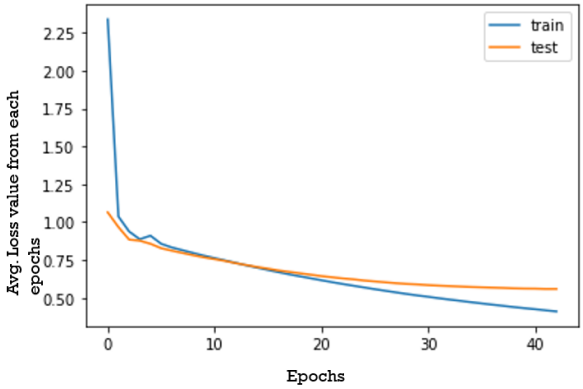

Fig. 3 Loss difference between training and testing phase for Thai

VII. CONCLUSION

There has been a huge surge in NMT (Neural Machine Translation) over the recent years. There have been some significance contributions in providing good accuracy with limited scope of applications. LSTM on its own does not give good results for long documents but when combined with attention mechanism it provides improved accuracy. The proposed model gives a BLEU score of 61.27 for the Marathi dataset and 46.96 for Thai dataset.

REFERENCES

[1] Roee Aharoni, Melvin Johnson and Orhan Firat,"Massively Multilingual Neural Machine Translation",2020.Avaliable:https://arxiv.org/abs/1903.00089

[2] Xu Tan,YichongLeng,JialeChen,YiRen,TaoQin,Tie-Yan Liu, "A study of Multilingual Neural Machine Translation",2019.Avaliable:Avaliable:https://arxiv.org/abs/1912.11625

[3] Raj Dabre,ChenhuiChu,AnoopKunchukuttan, "A Comprehensive Survey of multilingual Neural Machine Translation",2020.Availabe:https://arxiv.org/abs/2001.01115

[4] Xu Tan, Yi Ren, Di He,TaoQin,Zhou Zhao and Tie-Yan Liu,"Multilingual Neural Machine Tranlsation with Knowledge Distillation",2019.Available:https://arxiv.org/abs/1902.10461

[5] DzmitryBahdanau,KyungHyun Cho YoshuaBengio,"Neural Machine Translation by jointly learning to align and translate",2016.Available:https://arxiv.org/abs/1409.0473

[6] P. Shah and V. Bakrola, "Neural Machine Translation System of Indic Languages - An Attention based Approach," 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India, 2019, pp. 1-5, doi: 10.1109/ICACCP.2019.8882969

[7] WannaphongPhatthiyaphaibun, KorakotChaovavanich, CharinPolpanumas, ArthitSuriyawongkul, Lalita Lowphansirikul, &PattarawatChormai. (2016, Jun 27). PyThaiNLP: Thai Natural Language Processing in Python. Zenodo. http://doi.org/10.5281/zenodo.3519354

[8] Patel, H. (2020, June 15). Neural Machine Translation (NMT) with Attention Mechanism. Medium. https://towardsdatascience.com/neural-machine-translation-nmt-with-attention-mechanism-5e59b57bd2ac

[9] Neural machine translation with attention | TensorFlow Core. (2021, February 3). TensorFlow. https://www.tensorflow.org/tutorials/text/nmt_with_attention

[10] Brownie J. (2020, October 6). How to Develop a Neural Machine Translation System from Scratch. Machine Learning Mastery. https://machinelearningmastery.com/develop-neural-machine-translation-system-keras/

[11] Tran, T. (2020, December 8). Neural Machine Translation With Attention Mechanism. Trungtran.Io. https://trungtran.io/2019/03/29/neural-machine-translation-with-attention-mechanism/

[12] Wani, P. (2021, February 9). German To English Translator Using Keras and TensorFlow using LSTM model! Value ML. https://valueml.com/german-to-english-translator-using-keras-and-tensorflow-using-lstm-model/