Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766422

ANCIENT TEXT TRANSLATOR

[1]Aayush Rathi,[2] Abhishek Yadav, [3]Bhawna Mallick,
[4] Alpika Suman, [5] Aarushi Arya Thakur
Meerut Institute of Engineering and Technology, Meerut, U.P., India
{aayush.rathi.cs.2019, abhishek.yadav.cs.2019, bhawna.mallick, alpika.suman.cs.2019, aarushi.arya.cs.2019 } @miet.ac.in

ABSTRACT
Language is the only way to transmit with the community from different extent. To be able to clearly convey our message to another person we need to have a good hold of their language, in the modern days, this barrier of language can be reduced or even eliminated by using a text translator. This Ancient Text Translator can be helpful in the following conditions-
To collect and analyze the ancient script images available in the palm leaf literature format,To pre-process i.e by removing the stop word, various punctuations and finally converting all the upper case into lower case, To translate the recognizable script using vectorization and Long Short Term Memory (LSTM) modeling into a current recognizable format i.e English,To verify and validate (V&V) the developed ancient script text translation system by performing comparative analysis with the various existing system.
When we visit any monument and encounter any guide stone or memorial stone having engraved text of ancient language. When we're reading any piece of work of that language.The main aim is for translating the provided ancient scripts to the desired language i.e 'English',For this, we would require a text mapping from that language to the target language.
Keywords – LSTM, V&V, Memorial Stone, Palm Leaf literature format, Text mapping

1. Introduction:

Google Translator , a website which has a variety of language that can translate into another language instantly. It has made the world a nice place by increasing communication with different background people when language is a barrier to them. It not only changed the level of communication but it also increased the business a lot(through advertising). Earlier on , the language was the biggest hurdle for communication or conversation, but the problem has disappeared slowly after the entry of language translators. Even with the help Google Translator can not translate the ancient language whose data is not present in a structured way. Our language translator will interpret the structure of the sentence in the source language and will generate a translation of it into the language the user is translating to. The Source Language for this Project is 'Sundanese' and the target language is 'English'. We will be using sequence-to-sequence learning, it is very strong technique which is used to solve many kinds of problems. After that, we will use neural machine translation for further translation.

2. Literature Review

Currently much of the research has been done which is related to this topic and also published their paper related to the work.

Mozhden Gheini ,Jonathon May in the year 2019 done this work by using the approach of word segmentation algorithm, sequence to sequence mapping , LSTM by using 19 different language pair and get 96.90% accuracy as a result.

Adam Geitgey in the year 2017 uses the approach of language translation with deep learning and the magic of sequences by using multi-lingual word corpus astheir dataset with the accuracy of 75.42%.

Thomas Tracy in the year 2017 uses the approach of language translation with RNNsusing linguistic word corpus as dataset for work with the accuracy of 74.77%.

Ajay Kumar,Shashi Pal Singh in the year 2013 uses the approach of MACHINE CONVEYING USING DEEP LEARNING by using English- hindi words corpus with the accuracy of 94.20%.

3. Design and Implementation of the system

Input data will include text in some ancient language and the task will be to convert the current language into desired language which is 'English'.

The training dataset will include two text files

1- The text file in an ancient language.

2- The English translation of that text file.

The testing dataset will include only the text in the ancient language.

We will reach our goal by following these steps:

(a) Preprocess Text Data

Firstly, the dataset in such a manner in which protect the Unicode Sundanese char. This would include the following steps.

- Discard all the non printable chars.
- Discard all the punctuation chars.
- Formalise the case to lowercase.
- Discard any existing token that are special symbols.

(b) Split_Text

Better dataset of model translation, In which we solve the hitch little bit to dramatically minimize the size of model needed, the training time needed to apt the model.

We will split the text into various different batches before starting with the actual training of the data. We split the dataset into trains and test the set of model training and assessment.

(c) Data_Preprocessing:

Already there is a bunch of sentences and we train something fast, then cut-off the dataset to short and easy sentences. Then the maximum length will be of 10 words.

The entire process of produce the data is:

- Scan text file and break into lines, then break lines into pairs
- Formalise text, which riddle by length and content
- Make word lists from the sentences in pairs

(d) Model_Building

Encode the Unicode as input sequences and English sentences as the desired sequences. It applied to both the training dataset ,test dataset.
It is started by defining the model architecture for further process:

We will use LSTM layer and an embedding layer on the encoder side.

We will use one more LSTM layer and followed by dense layer on the decoder side.

For training of the model, 70% of the data will be used and the left 30% of the data will be used for evaluation.

At last, we upload the saved dataset and make the predictions on hidden dataset. These predictions are the sequences of number. Convert predictions into text (English).

(e) Encoder

An RNN is the encoder of order to order network which gives an output value for each word from the inserted sentence. Encoder built vector of words for each input char and it contain situation and that is a hidden situation , it helps to predict the next word.

(f) Decoder

Decoder is basically a RNN which feed with the output provided by encoder and again gives an output of vector of a sequence of chars to make the translation possible.

In decoder, we are only allowed to use the rearmost output of encoder, and this ouput vector is called context vector and it encode from the complete order. Context vector is basically used in the state of decoder which is hidden.

(g) Testing and Deployment: Testing will be done on a dataset that would be completely unseen for the model

Encoder:

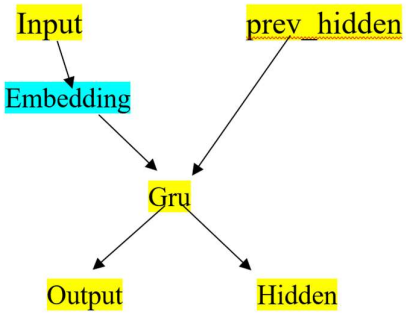


Figure 1

Decoder:

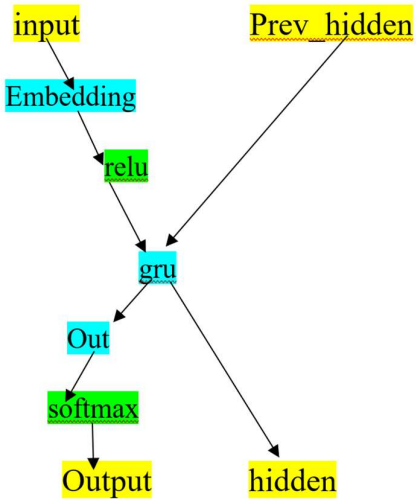
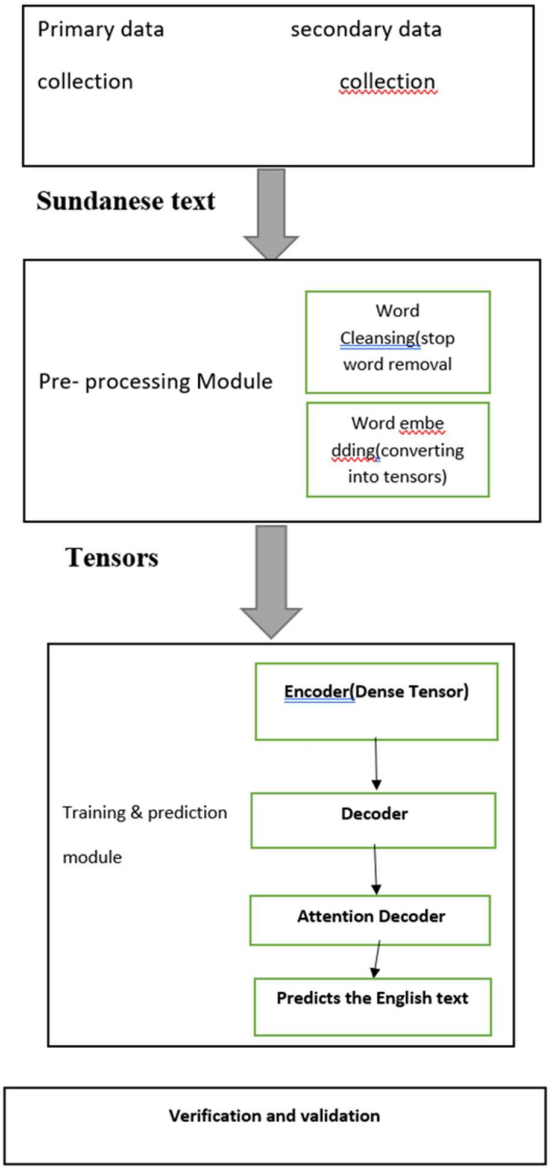


Figure 2

4. MODEL ARCHITECTURE



5. Test Designed

A total of 60s Sentences were taken for the testing of the model. Some of them are listed below-

- ningali anjeun waktos salajengna
- kunaon kunaon urang persia kuno
- pangéran sanés émut guruna
- tapi kuring hoyong janten pejuang
- tugas bumi sapertos cuci

6. Training of Dataset

The Translator was trained for a dataset of 60 pairs of Sundanese- English Sentences, and the model was trained for 45000 epochs, then the encoder and decoder model was saved. Given below is a Snapshot of the training curve.

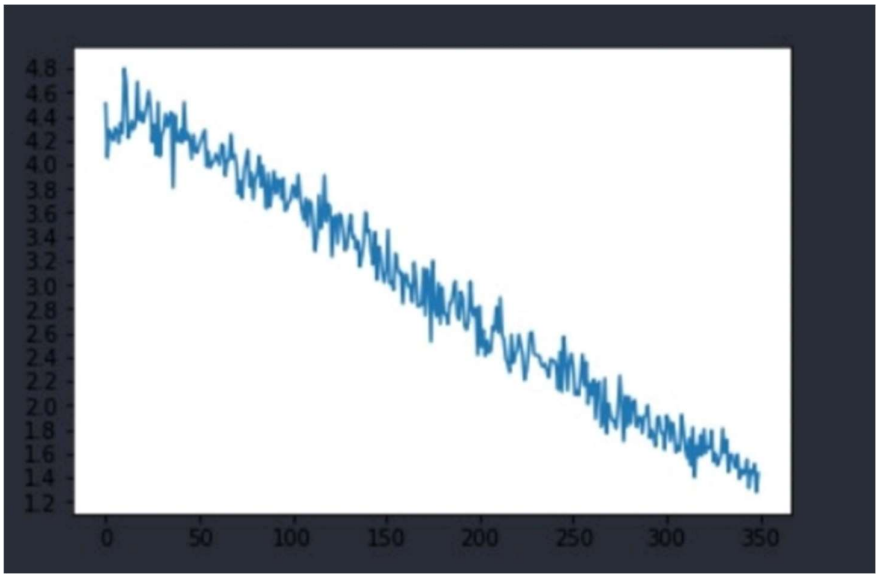


Figure 3

7. Performance Evaluation

The performance of our Ancient Text translator which translates Sundanese Language to the English Language was measured using two techniques

(i) Sentence BLEU score (ii) Word Error Rate

On Comparing with the existing state of the art Sundanese to English translator which is “imtranslator.net “, it was found that our translator gave better results with the sentences involving nouns in them.

The overall BLEU score was slightly less than the BLEU score of imtranslator.net but surprisingly the word error rate was better.

Word error rate of imtranslator.net = 0.25

Word error rate of our translator (sun trans) = 0.23

Sentence blue of imtranslator.net = 0.75

Sentence blue of our translator (Sun Trans) = 0.72

REFERENCES

[1] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, "Sequence to Sequence Learning with Neural Networks", Google.

[2] Jiajun Zhang And Chengqing Zong, "Deep Neural Network in Machine Translation", Institute of Automation, Chinese Academy of Sciences.
[3] Siddiqui T. And Tiwary U.S., "Natural Language Processing and Information Retrieval", Oxford University press

[4] Desika (Natural Language Understanding System

[5] "Information on Anusaaraka system", Available *http://anusaaraka.iiit.ac.in/*, Accessed Nov 2021.

[6] Noha Adly and Sameh Al Ansary. Evaluation of Arabic Machine Translation System based on the Universal Networking Language. H.Horack et al. (Eds.):NLDB , LNCS 5723,pp.243-257,2019. Springer-Verlag Berlin Heidelberg
.
[7] Nisheeth Joshi, Hemant Darbari, Iti Mathur. Human and Automatic Evaluation of English-Hindi Machine Translation Systems. Advances in Computer Science, Engineering and Applications. Advances in Intelligent and Soft Computing Series, Vol. 166, pp 423-432, Springer Verlang. (2020)

[8] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc VLe, Mohammad Norouzi, Wolfgang Macherey,Maxim Krikun, Yuan Cao, Qin Gao, KlausMacherey, et al. 2019. Google's neural ma-chine translation system: Bridging the gap between human and machine translation. arXiv preprintarXiv:1609.08144.

[9] Daniel Schlegel, "Deep Machine Learning on Gpu", University of Heidelber-Ziti, 12 January 2019.

[10] Holger Schwenk, Yoshua Bengio," Learning Phrase Representations Using Rnn Encoder–Decoder for Statistical Machine Translation", Arxiv: V3 .