Machine Translation on Indigenous Indian Language

Summer Internship Report

Submitted By

Shubham Dikshit
iamsds123@gmail.com

Deepak Yadav
deepak842705@gmail.com

Under the guidance of:
Prof. Shi-Jim Yen
Department of Computer Science and Information Engineering,
National Dong Hwa University, Taiwan.

Abstract

In this paper we explore the strategies to incorporate the features of machine translation of indigenous languages mainly Bengali, Tamil, Telugu into English translation through various models to check which one is beneficial. All the experiments were done through Pytorch and all the data were collected through the web scraping and many websites. This paper is also the first attempt at complete sentence machine translations of the indigenous languages to English. The research and associate case studies are aimed to help promote the greater research responsibility in Machine Translation of the Indian Indigenous Languages by using different models and fine tuning the models for correctly translating them in Indian Languages using the input of sequence of words and generates an output sequence of the words using mainly Encoders and Decoders.

Keywords

Natural Language Processing, Deep Learning Algorithm, Languages Translation (Bangla, Telugu, and Tamil) in English, Sequence2Sequence Model, Pytorch

Contents

1 Introduction

Sequence to sequence network is a model in which the two recurrent neural networks work together to transform one sequence to another.An encoder network condenses an input sequence into a vector, and a decoder network unfolds that vector into a new sequence.There are many challenges in machine translation for Indian languages. For instance, (i) the size of parallel corpora and (ii) differences amongst languages, mainly the morphological richness and word order differences due to syntactical divergence are two of the major challenges. Indian languages (IL) suffer both of these problems, especially when they are being translated from English. There are only a few parallel corpora for English and Indian languages. To improve the model we use the attention mechanism, which leads the decoder to learn to focus over a specific range of input sequence. Tools for machine translation (MT) from English to certain Indian languages and from one Indian language to another are available; however, [3] such tools for MT from Indian language to English are very few. For Telugu, Bangla and Tamil there is only one known web based MT system available in the form of Google Translator A large number of experiments were conducted on the Google Translator to obtain the translation of various simple and moderately complex statements. The tool was successful in translating the documents to an appreciable extent, but could not provide expected results even for many simple sentences. MT to an Indian language from English is cumbersome, and MT to English from an Indian language is much more complex. The richness of Telugu language lies in the extremely large number of words representing different moods, expressions, contexts, etc. A more sophisticated method which is also a growing field used to address the issue of recognition of multiple phrases is with statistical and neural technique. In this translation of text from one language to another, there is no human involvement and it is the machine which performs the process of conversion. Though much work is being done on machine translation for foreign and Indian languages but apart from foreign languages most of works on Indian languages are limited to conventional machine translation techniques. There are three types of machine translation system-rules based, statistical and neural. Rule based is a conventional method which is a combination of language and grammar and the support of dictionaries. This work focuses on building an end to end machine translation pipeline. We have discussed multiple existing architectures and finally proposed a hybrid model to achieve a more powerful system for machine translation from English to Bangla, Telugu, and Tamil, etc.

2. Languages

The languages used in the data for the machine translation of the indigenous language translation is extracted through many sources

2.1 Tamil

Tamil is a Dravidian language predominantly spoken by the Tamil people of India and Sri Lanka, Tamil is an official language in the three countries India, Sri Lanka and Singapore. In India, it is an official language of the Indian state of Tamil Nadu and Union Territory of Puducherry. Tamil is spoken by significant minorities in the four other South Indian states of Karnataka, Kerala, Andhra Pradesh and Telangana, it is one of the 22 scheduled languages of India. Tamil is one of the longest-surviving classical languages in the world. A recorded Tamil literature has been documented for over 2000 years. It has more than 75 million native speakers in India and more than 6 million people speak it as the secondary language.

2.2 Telugu

Telugu is the most spoken Dravidian language. It is predominantly spoken in Andhra Pradesh, Telangana and Union Territories of Puducherry. It stands alongside Hindi and Bengali as one of the few languages with primary official language status in more than one Indian state. It is one of the six languages designated a classical language of India by the country's government. Telugu ranks fourth among the languages with the highest number of native speakers in India, with 6.7 percent and 15th in the Ethnologue list of most widely-spoken languages worldwide. Telugu has more than 82 million native speakers in India and more than 11 million people speaks it as a secondary language. It is the most widely spoken member of the Dravidian language family and one of the twenty-two scheduled languages of the Republic of India. It is also the fastest-growing language in the United States, where there is a large Telugu-speaking community.

2.3 Bengali

Bengali is also known by its endonym Bangla, is an Indo-Aryan language primarily spoken by the eastern part of the Indian subcontinent, native to Bangladesh and West Bengal. There are also a significant number of Bengali speakers in the states of Tripura, Assam. It is second most widely spoken in Indian after Hindi with more than 230 million native speakers in the country. Bengali has developed over the course of more than 1,300 years. Bengali literature, with its millennium-old literary history, has extensively developed since the Bengali Renaissance and is one of the most prolific and diverse literary traditions in Asia. The Bengali language movement from 1948 to 1956 demanding Bengali to be an official language of Pakistan fostered Bengali nationalism in East Bengal leading to the emergence of Bangladesh in 1971. In 1999, UNESCO recognised 21 February as International Mother Language Day in recognition of the language movement. The Bengali language is the quintessential element of Bengali identity and binds together a culturally diverse region.

3 Dataset

The first step in the implementation of any Deep Learning project is the investigation of DataSet datasets used for Machine Translation are from WMT (the website that is dedicated to research in statistical machine translation). Due to time constraints, the dataset contains small vocabulary. This facilitates the training in a reasonable time. The data located in a file path is loaded. The file contains English sentences with their Indian Indeginous language translations. Load the data from these files. Print the first two lines from each file.

```
Hello!   নমস্কার!
I see.   বুঝলাম।
I try.   আমি চেষ্টা করি।
Smile.   একটু হাসুন।
Smile.   একটু হাসো।
Attack!  আক্রমণ!
Get up.  ওঠো।
Get up.  উঠুন।
```

Fig.1. First few lines of English to Bengali from each File

We used the datasets obtained from EnTam V2.05 and Opus. The files are all in Unicode, to simplify we will turn Unicode characters to ASCII, make everything lowercase, and trim most punctuation. The complexity of vocabulary determines the complexity of the problem.
Similar to the character encoding used in the character-level RNN tutorials, we will be representing each word in a language as a one-hot vector, or giant vector of zeros except for a single one (at the index of the word). Compared to the dozens of characters that might exist in a language, there are many many more words, so the encoding vector is much larger

3.1 Seq2Seq Model

A Recurrent Neural Network, or RNN [1] [2], is a network that operates on a sequence and uses its own output as input for subsequent steps. A Sequence to Sequence network is a model consisting of two RNNs called the encoder and decoder. The encoder reads an input sequence and outputs a single vector, and the decoder reads that vector to produce an output sequence. The encoder of a seq2seq network is a RNN [1] that outputs some value for every word from the input sentence. For every input word the encoder outputs a vector and a hidden state, and uses the hidden state for the next input word. The decoder is another RNN that takes the encoder output vector(s) and outputs a sequence of words to create the translation. In the simplest seq2seq decoder [3] we use only the last output of the encoder. This last output is sometimes called the *context vector* as it encodes context from the entire sequence. This context vector is used as the initial hidden state of the decoder. At every step of decoding, the decoder is given an input token and hidden state. The initial input token is the start-of-string token, and the first hidden state is the context vector (the encoder's last hidden state).

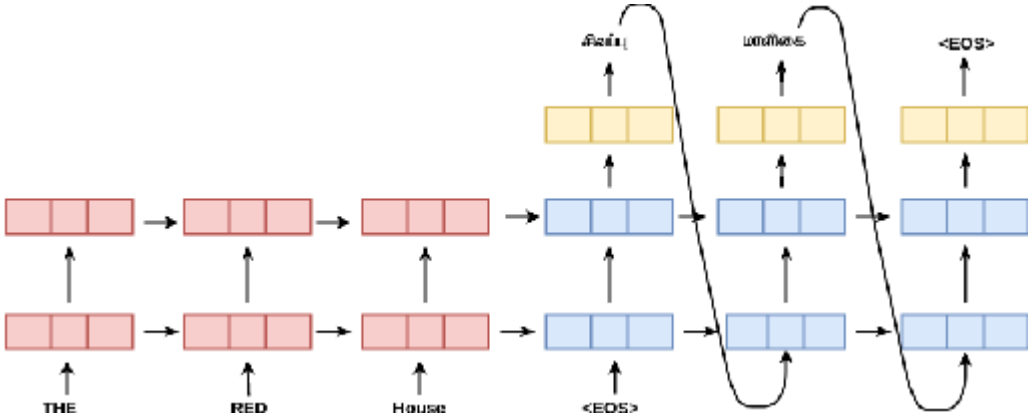

Figure 1: Seq2Seq architecture for English-Tamil [3].

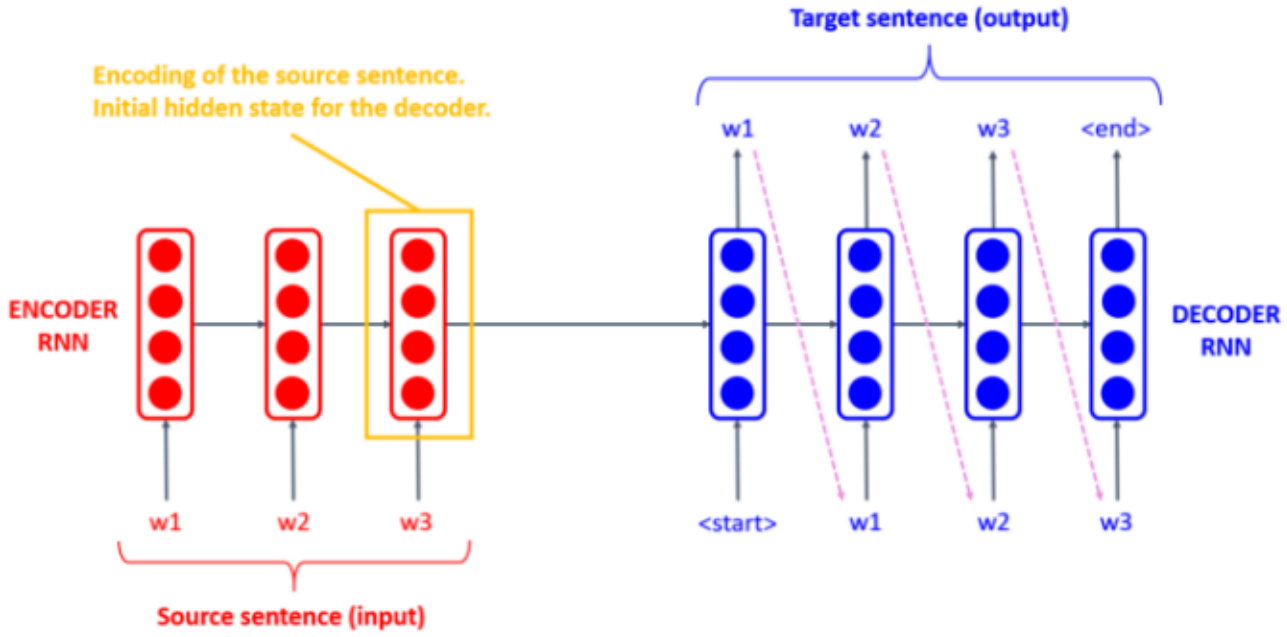

Figure 2. RNN [4] working Model (Input – Output)

3.1.1 The Attention Decoder

If only the context vector is passed between the encoder and decoder, that single vector carries the burden of encoding the entire sentence. Attention allows the decoder network to "focus" on a different part of the encoder's outputs for every step of the decoder's own outputs. First we calculate a set of *attention weights*. These will be multiplied by the encoder output vectors to create a weighted combination. The result should contain information about that specific part of the input sequence, and thus help the decoder choose the right output words. A useful property of the attention mechanism is its highly interpretable outputs. Because it is used to weight specific encoder outputs of the input sequence, we can imagine looking where the network is focused most at each time step.

In a basic encoder-decoder architecture, the encoder memorizes the whole sentence in terms of vector, and stores it in the final activation layer, then the decoder uses that vector to generate the target sentence. This architecture works quite well for small sentences, but for larger sentences, maybe longer than 30 or 40 words, the performance degrades. To overcome this problem attention mechanisms play an important role. The basic idea behind this is that each time, when the model predicts an output word, it only uses the parts of input where the most relevant information is concentrated instead of the whole sentence. In other words, it only pays attention to some weighted words. Many types of attention mechanisms are used in order to improvise the translation accuracy, but the multi-head self-attention overcomes most of the problems

4. Evaluation

We can compare the performance of any machine translation model by comparing it across various evaluation matrices. In this paper, [4] the following evaluation matrices are used for estimating the performance of our model. Evaluation is mostly the same as training, but there are no targets so we simply feed the decoder's predictions back to itself for each step. Every time it predicts a word we add it to the output string, and if it predicts the EOS token we stop there. We also store the decoder's attention outputs for display later. We evaluated the sentences randomly from the training set and printed out the input, target, and output to make some subjective quality judgements.

4.1. Evaluation Metric

BLEU uses the basic concepts of n-gram precision to calculate similarity between reference and generated sentence. It correlates highly with human expert review as it uses the average score of all results in the test dataset rather than providing the result of each sentence. BLEU score is a method to measure the difference between machine translation and human translation. The approach works by matching n-grams in result translation to n-grams in the reference text, where unigram is a unique token, bigram is a word pair and so on. A perfect match results in a score of 1.0 or 100%, Our proposed model is divided into mainly three different parts. Encoder, Decoder and Attention mechanism. Our encoder has two LSTM layers with 128 units of LSTM cells. This encoder will output encoded word embedding vectors. This embedding vector is provided as input to the decoder. Decoder also consists of two LSTM layers with 128 units of lstm cells. It will take an encoded vector and produce the output. Whenever any output is produced the value of the hidden state is compared with all input states to

derive weights for attention mechanism. Based on attention weights, context vector is calculated and it is given as additional input to decoder for generating context relevant translation based on previous outcomes [4].

4.2. Analysis

We conducted a survey with ten random sentences from our test data and accumulated the reviews of native Tamil speaking peoples. In comparison, it was found that our translation results were better in 60% cases.

Language	**Training**	**Testing**	**BLEU Score**
Bengali	20662	2000	4.07
Tamil	25782	4218	5.43
Telugu	21545	2100	4.97

5. Conclusion

In this report we created a sequence to sequence model using Pytorch on the indigenous languages. These features were obtained using the Attention Decoder. We first explored the dataset and turned it into encoders and decoders sized inputs. We then investigated the best way to incorporate the information from the model. We then got the BLEU scores of Tamil, Telugu and Bengali which were 5.43, 4.97 and 4.07 respectively and this happened due to low data for the languages. [3] This approach can be applied to the other Indian Languages as well. The common dilemma of transliteration and translation is to come to a decision where to transliterate, where to translate and where to perform both of them. Although Bangla and English are from the same family of languages having some features in common, they differ in many respects and aspects. These problems are due to the differences between the characteristics and properties of the languages concerned. So, in order to transliterate and translate from and to these languages has always been a very complicated job which necessitates the bilingual expertise to a great extent. It is also found that the lexical knowledge insufficiency; inadequate knowledge and practice of grammar; poor capability in phonetics and phonology; inadequate cultural background; inappropriate teaching atmosphere and methodology are the most important problems of transliteration and translation. Moreover, the influence of culture and religion is very strong in both the languages. As some phonetic problems show cultural aspect and background of language, a great care and attention should be employed. For transliteration, the basic knowledge in morphology, phonetics and phonology is required. Although lexical problems are greater in number, grammatical, stylistic and phonological problems are not marginal, specially for translation. However, anyone intending to carry out the task of transliteration or translation needs to be an expert in comparative linguistics. It is hoped that teachers, researchers, reader and students would benefit from this research work though the scope for further investigation has not been finished so far. More

comparative studies will result in providing the readers or learners with a more clear-cut knowledge about the two languages. So, to resolve the challenges of transliterating and translating Indigenous language into English or finding any straightforward means is in no way an easy task.

6. References

[1] Holger Schwenk, Yoshua Bengio," Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation", Arxiv: V3 ,3 Sep 2014.

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

[3] Md. Faruquzzaman Akan "Translation and Translation from Bengla to English: A P roblem Solving Approach".

[4] https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/