Attention based Neural Machine Translation of Indian Languages

Ghanta Swetha, Sreelatha Moturi

[1]*M.Tech Scholar, Department of CSE, RVR & JC College of Engineering, AP, India.*

[2]*Professor, Department of CSE, RVR & JC College of Engineering, AP, India.*

[1]*gswetha697@gmail.com*

Abstract

Machine translation is converting source text in one language to text in another language automatically. In recent years, Neural Machine Translation has emerged as a way of addressing this task and is proved to be more effective than rule-based and statistical techniques. In this paper, the researchers explored different configurations for setting up a Neural Machine Translation (NMT) System for languages Hindi and Telugu. Here, a sequence to sequence translator is built to convert English sentences to Hindi and Telugu sentences using various concepts like applying the encoder-decoder architecture and Bahdanau Attention mechanism. The performance of the developed NMT model is evaluated using the BLEU score metric.

Keywords: *Neural Machine Translation, sequence to sequence translation, Bahdanau Attention, Deep Learning, Machine Learning*

1. Introduction

One of the earliest goals for computers was the automated translation of text from one language to a different. Automatic translation with minimal human intervention is probably one of the foremost challenging AI tasks given the fluidity of human language. Classically, rule-based systems[1] were used for this task, which was replaced within the 1990s with statistical methods[2]. More recently, deep neural network models achieve state-of-the-art leads to a field that's aptly named NMT.

Given a large number of diverse languages in a country like India, the language translation is a key part of communication within various communities present in the country and worldwide. Machine translation helps us in this task by converting the text in the source language to the corresponding text in the target language. There is no single best translation and it is because of the nature and ambiguity of the dialect. It makes it more tricky and requires research in handling the challenges of the language.

It is particularly useful when large amounts of user-generated content need to be translated quickly. It also serves extremely in commercial sectors. For example,

machine translation makes it possible to quickly translate and review customer reviews, online comments, and social media posts thereby, increasing the depth of market study. Human translators tend to be more time consuming, exhausting, and more expensive.

2. Related Work

2.1. Word to Word translation

The conventional method followed was where each word was translated to its corresponding word in the translated language. However, this method was ineffective since the grammatical rules followed by different languages are not necessarily the same. Another issue, faced with this technique was the number of words in the output was the same as the number of words as the given input which does not have to be the case for all languages.

2.2. Rule-based

Rule-based MT (RBMT) relies on linguistic rules and bilingual dictionaries for each language pair wherein the rules capture the syntactic and semantic properties of a language. Rules are written based on linguistic knowledge gathered from the linguists.

2.3. Statistical

Statistical Machine Translation (SMT) uses a statistical model to generate translations and is based on the analysis of the bilingual corpus. The most important benefit of SMT over Rule-based Machine RBMT is that it does not require manual development of linguistic rules, which is quite costly.

2.4. Using RNNs

RNNs using the encoder-decoder architecture are used where variable length output from the input may be input. This is possible due to the ability of the model to encode the source text into an internal fixed-length representation called the context vector. The RNNs suffer from the long term dependency problem and to overcome the other variants like Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) are used.[3-5]

2.5. Seq2Seq without Attention mechanism

Encoder converts the input into a fixed-size vector and then the decoder predicts an output sequence[6]. It works fine for short sentences but fails when long sentences are considered, because it becomes difficult for the encoder to remember the entire

sequence into a fixed-sized vector and to compress all the contextual information from the sequence. From the observations, as the sentence length increases the model performance degrades.

2.6. NMT on Indian Languages

A lot of research work is done on NMT but it is limited to foreign languages like French, German and others[7]. One of the most significant works of neural machine translation on Indian languages is done by Revanuru et al[8]. They demonstrated that good translation accuracy can be achieved even by using a simple and shallow network. Himanshu et al[9] developed a MIDAS translator for English-Tamil translation. They used the Byte Pair Encoding (BPE) and word embeddings to overcome the Out Of Vocabulary (OOV) problem.

3. Model

3.1. Encoder

It reads the input sentence in source language and encodes the information as vectors which are referred as hidden states. If sequence to sequence model without attention is considered, then only the hidden state of the last RNN is taken and the rest are discarded. Here, using attention, these hidden states are not discarded but are used to calculate alignment scores and compute the context vector.

3.2. Decoder

A decoder task is to output a translation from the encoded vector. For that the decoder considers the hidden state of Encoder GRU cell as the initial state of Decoder GRU cell along with the <start> token. Decoder takes input from the previous state of decoder and present input from encoder followed by Attention.

3.3. Bahdanau Attention

Bahdanau attention[9] is also referred as additive attention because it uses the linear combination of encoder and decoder states. Using the encoder hidden states the alignment score is calculated, which is used to compute the context vector. At time ***t, t-1*** hidden state of the decoder is considered. The context vector is concatenated with hidden state of the decoder at t-1. So before the softmax function this concatenated vector goes inside the GRU.

4. Implementation

4.1. Dataset

For working with neural networks, a large training data is required, because the neural networks learning accuracy is dependent on experience. But for Indian languages, massive parallel corpora are not available. Despite the challenge, using the attention mechanism an efficient and simple encoder-decoder model is developed. IIT Bombay English-Hindi parallel corpus and manythings.org with 127607 sentence pairs. The English-Telugu parallel corpora is from OPUS and manythings.org with 155798 sentence pairs. The datasets used for training the model contains tab de-limited English sentences and their corresponding translations in Hindi and Telugu. This dataset covers almost all types of variability in sentences which are used on a daily basis.

4.2. Pre-processing

a) Add a start and end token to each sentence.
b) The dataset is made homogeneous by converting all upper-case letters to lower-case.
c) Clean the sentences by removing special characters and stop words using the regular expression library.
d) Create a word index and reverse word index by applying integer encoding for every token generated in the input language (English) and the target language (Hindi or Telugu).
e) Pad each sentence to a maximum length (30 in our case).

4.3. Dataset Preparation

a) Open the file, strip from beginning and end, then finally split the line when it sees line separator("\n").
b) Generate English-IN word pair separated by tab("\t"), where IN refers to Hindi or Telugu.
c) Apply pre-processing on the inputted text files.
d) Divide the dataset into train and validation sets using 80-20 split rule.

4.4. Training

To train the model, first pass the input through the encoder which returns the encoder output and the encoder hidden states. These along with the <start> token are passed to the decoder, which returns the predictions and the decoder hidden state. The hidden state of the decoder is passed back into the model and the predictions are used to calculate the loss. Here, teacher forcing technique is used, where the target word is passed as the next input to the decoder. The loss chosen was Sparse Categorical Cross entropy. The reason is it computes cross-entropy loss

when 2 or more labels are available. Adam Optimizer is used. Finally, calculate the gradients and apply it to the optimizer and backpropagate. The model used in this paper is depicted in Fig.1.

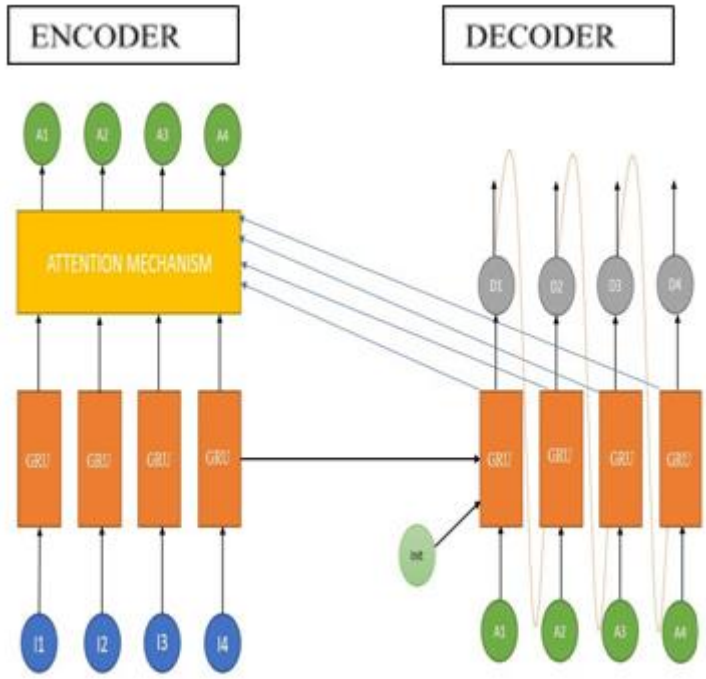

Fig.1. NMT with Attention model

4.5. Evaluation

The evaluation is similar to the training loop, except that the input to the decoder at each time step is its previous predictions along with the hidden state and the encoder output. When the model predicts the <end> token, the prediction is stopped. For every time step, the attention weights are stored.

4.6. Equations

a) Input:
$\{x_i = \text{source}_i , y_{i=1} \text{target}_i \}^N$

b) Encoder:
$h_t = \text{GRU}(h_{t-1}, x_t)$
$s_0 = h_T$

c) Decoder:

$$e_{jt}=V^T_{attn} \tanh(U_{attn}h_j + W_{attn}s_t)$$

$$\alpha_{jt}= \text{softmax}(e_{jt})$$

$$c_t = \sum_{j=1}^{T} \alpha_{jt} h_t$$

$s_t = \text{GRU}(s_{t-1}, [e(y'_{t-1}), c_t])$

$l_t = \text{softmax}(Vs_t + b)$

5. Experimental results

5.1. English-Telugu sample translations

Correct translations:

a) EN: I don't have time to talk right now.
   TE: నాకు ఇప్పుడు మాట్లాడేంత సమయం లేదు.
b) EN: Sitting down all day is bad for you.
   TE: రోజంతా కూర్చోవడం నీకు మంచిది కాదు.
c) EN: I don't know if I can do that
   TE: నేను అలా చేయగలనా అని నాకు తెలియదు.

Wrong translations:

a) EN: What is your favorite weather drink?
   TE: మీకు ఇష్టమైన రేడియో ఆహారం ఏమిటి?
b) EN: I can't let you go alone
   TE: నేను నిన్ను ఒంటరిగా ఉండాలని నేను వెళ్ళలేను.
c) EN: People always don't always tell the truth
   TE: ప్రజలు ఎప్పుడూ నిజం నిజం కాదు.

5.2. English-Hindi sample translations

Correct translations:

a) EN: I fixed the car yesterday.
   TE: मैंने गाड़ी को कल ठीक किया था।.
b) EN: He usually comes home late.
   TE: वह आमतौर पर घर देर से आता है।
c) EN: She said that she was happy.
   TE: उसने कहा कि वह खुश थी।

Wrong translations:

a) EN: My telephone is out of order.
   TE: मेरा फ़ोन लगा है।
b) EN: Is it possible to borrow money?
   TE: क्या पैसे उधार कोई फ़ायदा लिए जा सकते हैं?

c) EN: You owe me an apology for that.
TE: तुम्हे मुझसे माँगनी चाहिए।

5.3. BLEU score

This the most widely used method of automatic evaluation where n-gram precision with respect to reference translation is computed[11]. The proposed approach shows significant improvement in BLEU score as it is able to capture syntactic and semantic property of the sentences. The comparison of BLEU scores can be clearly seen in the Table 1 below.

Table 1. BLEU scores

Model	BLEU score for EN-TE	BLEU score for EN-HI
NMT without attention	18.45	16.09
NMT with attention	20.2	18.05

6. Future Work

The future work incorporates the use of other Indian languages for attention based neural machine translation. The research can also extend to using beam search strategy[12] or using a transformer[13] model which overcomes the attention mechanism drawbacks.

7. Conclusion

In this paper, a machine translation model is developed to translate English sentences to Indian languages like Hindi and Telugu. The parallel corpora considered are not massive but still pretty good translation accuracy is obtained for a limited dataset. This is because of the efficient encoder-decoder model with

attention. The attention mechanism also helped to deal with the long sentence translations. The BLEU scores obtained for Telugu and Hindi are 20.02 and 18.05 respectively.

Acknowledgments

The authors are greatly acknowledged to DST-FIST (Government of India) for funding to setting up the research computing facilities at RVR&JC College of Engineering.

References

[1] Ghosh, Siddhartha, Sujata Thamke, and Kalyani U.R.S. "Translation of Telugu-Marathi and Vice-Versa Using Rule Based Machine Translation." Academy and Industry Research Collaboration Center (AIRCC), ***(2014)****. pp. 1–13.*

[2] Cho, Kyunghyun et al. "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation." EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. Association for Computational Linguistics (ACL), ***(2014)****. pp. 1724–1734.*

[3] Cho, Kyunghyun et al. "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches." Association for Computational Linguistics (ACL), ***(2015)****. pp. 103–111.*

[4] Hochreiter, Sepp, and Jürgen Schmidhuber. "LSTM 1997." Neural Computation 9.8, November 15, 1997 ***(1997)****: pp. 1735–1780.*

[5] Dey, Rahul, and Fathi M. Salemt. "Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks." Midwest Symposium on Circuits and Systems. Vol. 2017-August. Institute of Electrical and Electronics Engineers Inc., ***(2017)****. pp. 1597–1600.*

[6] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to Sequence Learning with Neural Networks." Advances in Neural Information Processing Systems. Vol. 4. Neural information processing systems foundation, ***(2014)****. pp. 3104–3112.*

[7] Wu, Yonghui et al. "Google's NMT." ArXiv e-prints ***(2016)****: pp. 1–23.*

[8] Pathak, Amarnath, and Partha Pakray. "Neural Machine Translation for Indian Languages." Journal of Intelligent Systems 28.3 ***(2019)****: pp. 465–477.*

[9] Choudhary, Himanshu et al. "Neural Machine Translation for English-Tamil." Association for Computational Linguistics (ACL), ***(2019)****. pp. 770–775.*

[10] Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." 3rd International Conference on Learning Representations, ICLR ***(2015)*** *- Conference Track Proceedings.*

[11] Papineni, Kishore et al. "BLEU : A Method for Automatic Evaluation of Machine Translation." Computational Linguistics July ***(2002)****: pp. 311–318.*

[12] Freitag, Markus, and Yaser Al-Onaizan. "Beam Search Strategies for Neural Machine Translation." Association for Computational Linguistics (ACL), ***(2017)****. pp. 56–60.*

[13] Vaswani, Ashish et al. "Attention Is All You Need." Advances in Neural Information Processing Systems. Vol. 2017-December. Neural information processing systems foundation, ***(2017)****. pp. 5999–6009.*