NEURAL MACHINE TRANSLATION MODELS FOR ENGLISH -TELUGU TRANSLATION

Chandana Enugala

***Abstract*—Language is a medium of communication between people and even the systems. However, there are diverse cultural people having their unique native language. This has created a wide range of languages in use in various parts of world. And a standard translator tool is required to communicate between two different languages. With growing technology and modern techniques, translation can be data driven using machine translation methods. Neural machine translation (NMT) has shown significant promise in the translation of one language into another when compared with previous methods of machine translation. This paper presents word-level sequence-to-sequence neural machine translation using an attention mechanism and long-short term encoder decoder model for translating English to Telugu. From our experiments we found that a BLEU score of 0.48 was achievable on the NMT with attention model.**

***Index Terms*—Nature language processing, machine translation, deep learning, neural networks, attention mechanism.**

I. INTRODUCTION

In early days of digitalization English used to be a prominent medium of communicating data over the Web. However, as the time passed data is consumed in the local language and the need for translating the data to the common language has become more vital to caters the need of users. Particularly, in India with diverse languages being used this is more crucial. Various translation techniques are used to translate these diverse languages into the desired languages using methods such as probabilistic algorithms, by teaching machine based on hand-written text [1]. There are various constraints such as cost and accuracy of these techniques in translating the information from source language to desired language. Languages can be categorized into two types, Having rich morphology and simple morphology. English is considerably simple language having morphology of average complexity. However, Telugu is a morphologically rich language. Telugu has a wide variety of text and grammar consisting of 56 unique sounds across 3 genders. This makes the task of translating telugu very challenging. There are various machine translation techniques that can be used.

One common method is Example Based Machine Translation. In this method a complex long sentence is broken down into smaller parts/phrases and arranged combined based on the target requirement to form the desired sentence [2].

Another method is Statistical Machine Translation which is a less robust technique but it is still reliable. When there are multiple translating outcomes for a given sentence into target language, this technique picks the outcome which has higher probability [3].

One of the best techniques is Neural Machine Translation. In this technique one large neural network is build and it is configured to increase the performance efficiency. This technique is taking a prominent place in the translation technique.

Another technique Rule Based Machine which is one of the oldest Technique. This is a manual process where the rules are defined by human.[13] It's a costly process as it involves handwritten rules of both source and target language, but it performs well.

If the performance of a single model is not satisfactory it is better to use a Hybrid Machine Technique. It is technique which is combination of one or more above mentioned techniques. It is more optimum techniques where the best results can be attained at a better cost. This also offsets the hindrances of one specific process by the other process [1].

II. LITERATURE REVIEW

Machine Translation is the process of turning a source language into the target language. Translation is said to be successful when words from one language are translated into another without losing their original meaning. Early work on MT began in the middle of the 1950s [3], and as computational limits became more trivial, it advanced steadily until the 1990s. Several strategies, including "rule-based interpretation, knowledge-based interpretation, corpus-based interpretation, hybrid interpretation, and statistical machine translation," have been put forth during that time to achieve more accurate Chinese interpretation (SMT) [3]. In recent years, neural networks have become widely used in machine interpretation. The most popular machine interpretation method today, known as "Neural Machine Translation, or NMT," uses a neural system. Even as recent years have seen an increasing number of NMT studies, very few of those have been on Indian dialects[3]. The NMT approach to Indic languages, specifically bilingual machine interpretation, has not yielded good results in previous research. The Angla Bharati-I machine translation system was created by IIT Kanpur researchers in 1991 [4][5]. An adaptive general-purpose translation system was created particularly to translate from English to Hindi. The transfer-based machine translation method was used by CDAC to create the machine translation system MANTRA [4] in the year 1999. The system has been designed to operate with English-Gujarati, English-Hindi, English-Bengali, and English-Telugu data pairings.

In this paper, we used English-Telugu sentences pairs for applying to two proposed models. One is neural machine translation with attention model and sequence-to-sequence encoder decoder model. Both methods have been evaluated using the BLEU score.

III. METHODOLOGY

Dataset: Dataset contains 155798 English-Telugu-Bilingual-Sentence-Pairs. For our dataset we obtained bilingual English-telugu pairs compiled by scionoftech.

A. Preprocessing

The source(English) and target (telugu) sentences are separated by unwanted characters "++++$+++", done splitting based on the pattern. Converted all the letters of source language sentences to lower case. As a part of preprocessing, removed extra spaces before and after each word by stripping. The target sentences were also padded at start and end of the sentence which helps the decoder to understand the beginning and finish of sentences. The longest possible sentences from the source and the destination, as well as all the unique terms, are stored.

B. Models Used

As mentioned earlier, we used two neural machine translation methods. Lstm sequence-to-sequence modeling was used as a starting point and compared with a sophisticated NMT attention model.

C. Word level Seq-seq encoder decoder model

Encoder and decoder are both considered stages in NMT. The standard RNN will often be utilized for the encoder and decoder stages. The longer-range dependencies in the source phase will provide a challenge for the conventional RNN. Long short-term memory (LSTM) is used in encoder and decoder instead of conventional RNN in the proposed framework. In the context of long-range dependencies, LSTM will offer higher accuracy than standard RNN.[10]

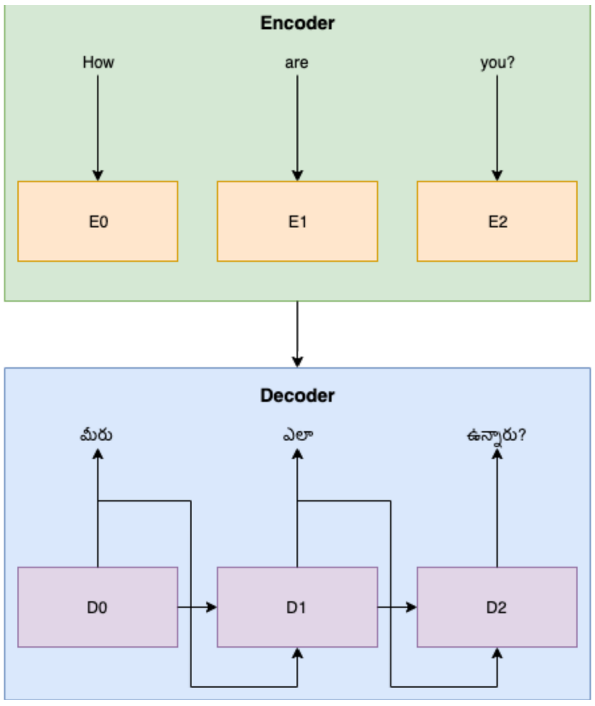

Fig. 1. Encoder and decoder model

[7]

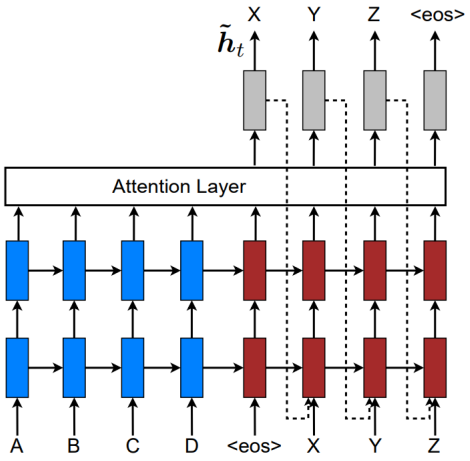

Fig. 2. attention based NMT

[8]

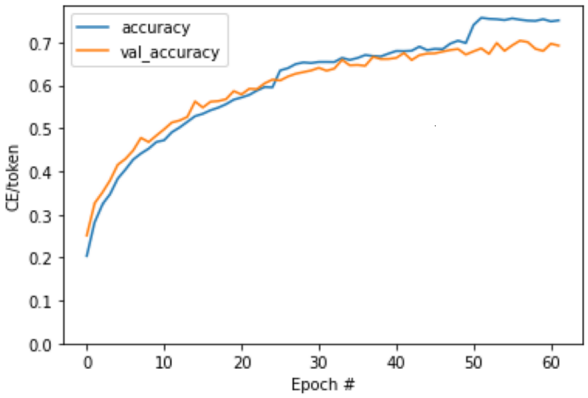

Fig. 3. accuracy graph of attention model

D. NMT with attention model

Encoder: The encoder's objective is to transform the context sequence into a series of vectors the decoder may use to anticipate the output for each timestep.

In the encoder layer, we The embedding layer converts tokens to vectors, The rnn layer consists of bidirectional layer and Gru layer with merge mode set to sum.

The decoder may access the data that the encoder has extracted from the attention layer. The decoder's job is to create predictions for the next token at each location in the target sequence, thus it computes a vector from the complete context sequence and adds that to the decoder's output. In the decoder, we have an embedding layer and a unidirectional RNN and an attention layer.[9]

IV. RESULTS

We used bleu score as an evaluation metric for both the models. It is a core evaluation measure in machine translation where the n-gram of source is matched with n-gram of the

```
english_source Don't you know her name?
actual_telugu ఆమె పేరు మీకు తెలియదా?

predTelData ఆమె పేరు మీకు తెలుసా?
=========================
english_source Tom brought Mary to our party.
actual_telugu టామ్ మేరీని మా పార్టీకి తీసుకువచ్చాడు.

predTelData టామ్ మేరీ మా పార్టీకి తెచ్చాడు.
=========================
english_source I'm sorry. I forgot about this.
actual_telugu నన్ను క్షమించండి.

predTelData నేను క్షమించండి, నేను దీని గురించి మర్చిపోయాను.
```

Fig. 4. predictions from the model

reference sentence without giving priority to positioning of the words. The method that performed best is word level sequence-to-sequence attention model which yielded a BLEU score of 0.48 and lstm-seq-seq model without attention layer yielded a bleu score of 0.385.

V. CONCLUSION AND FUTURE WORK

Neural machine translation overcomes the challenges of conventional machine translation methods. Recent NMT models, such as Seq-to-Seq attention based, have produced translation from one language to another efficiently. Questions still remain over whether the model will work effectively in real-world settings and more testing is required. Overall, Translating Dravidian languages is challenging task We still need to put in a lot of work to produce a decent translation result. For Future work we recommend raising the bleu score by performing hyper-parameter tuning and experiment with different models.

REFERENCES

[1] Din, U. M. ud. (2020, January 12). Urdu-English machine transliteration using Neural Networks. arXiv.org. Retrieved December 10, 2022, from https://arxiv.org/abs/2001.05296

[2] M. Zafar and A. Masood, "Interactive english to urdu machine translation using example-based approach," International Journal on Computer Science and Engineering

[3] A. Lopez, "Statistical machine translation," ACM Computing Surveys (CSUR), vol. 40, p. 8, 2008.

[4] P. Sheridan, "Research in language translation on the ibm type 701", IBM Technical Newsletter, vol. 9, pp. 5-24, 1955.

[5] S. K. Dwivedi and P. P. Sukhadeve, "Machine translation system in indian perspec

[6] Effective preprocessing based neural machine translation for English to (n.d.). Retrieved December 10, 2022, from https://www.researchgate.net/publication/352033065_Effective_preprocessing_based_neural_machine_translation_for_English_to_Telugu_cross-language_information_retrieval

[7] Scionoftech. (2019, December 25). Neural_Machine_Translation_English_Telugu/English_to_telugu_encoder_decoder.ipynb at master · scionoftech/neural_machine_translation_english_telugu. GitHub. Retrieved

[8] Fig Luong, M.-T., Pham, H., Manning, C. D. (2015, September 20). Effective approaches to attention-based neural machine translation. arXiv.org. Retrieved December 10, 2022

[9] Neural machine translation with attention text tensorflow. TensorFlow. (n.d.). Retrieved December 10, 2022,

[10] Neural machine translation for Tamil-telugu pair - researchgate. (n.d.). Retrieved December 11, 2022

[11] S. B. Sitender, "Survey of indian machine translation systems," IJCST, vol. 3, no. 1, 2012

[12]] N. Jayanthi, A. Lakshmi, C. S. K. Raju and B. Swathi, "Dual Translation of International and Indian Regional Language using Recent Machine Translation," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 682-686, doi: 10.1109/ICISS49785.2020.9316016

[13] S. Mall and U. C. Jaiswal, "Survey: Machine Translation for Indian Language," International Journal of Applied Engineering Research, vol. 13, pp. 202-209, 2018