Implementation of Cross-Lingual Neural Machine Translation for COVID-19 Treatment Using Bahdanau Attention Mechanism

Dr.M.Dhanalakshmi[1], Deshapaga Sindhuja[2]
[1] *Professor of IT, Jawaharlal Nehru Technological University, Hyderabad, Hyderabad, India*
[2] *Masters Student of CNIS, University College of Engineering, Science and Technology Hyderabad, JNTUH, Hyderabad, India*

***Abstract*— Early on, computers were tasked with the challenge of translating text from one language to another without human intervention. In this paper, we implemented the system which handles spontaneous freeform text-to-text translation. Machine translation in difficult situations with limited data and linguistic resources for Telugu and Hindi languages, as well as the requirement to rapidly build skills for new languages, are among the hurdles. Here our methods for developing a Bahdanau Attention mechanism capable of translating between languages on handheld devices with constrained memory and computational power. We discuss our approach, challenges, and achievements in creating a functional model within these limitations.". Bahdanau attention mechanism is a sophisticated approach that can be applied to various domains, including medical treatment information such as that for COVID-19. We have also implemented a text-to-speech conversion system for the target languages using Tacotron 2 model.**

***Index Terms*— Artificial Intelligence, Bahdanau Attention mechanism, Tacotron 2 model.**

I. INTRODUCTION

The COVID-19 pandemic has highlighted the need for effective communication across language barriers, especially in healthcare. Cross-lingual neural machine translation (NMT) is crucial for disseminating treatment information to diverse populations. Previously, rule-based systems[1] were utilised for this work, but statistical methods[2] were introduced in the 1990s. Recent advancements in deep neural networks have led to cutting-edge results in machine translation. There are 1599 languages spoken in India, with 22 main languages. Thirty languages worldwide boast over a million speakers, while 22 hold official language status. Remarkably, the average person can use between eleven and twenty-five thousand words daily. It's most useful when a huge volume of user-generated content needs to be translated quickly. It is also very useful in the commercial sector. Machine translation, for example, allows you to easily and rapidly translate and assess social media posts, online comments, and customer reviews, allowing you to expand the scope of your market research. Human translators appear to take longer, be more exhausting, and cost more. The previous common generic methods of machine translation are: word to word translation, rule-based Machine Translation(RBMT), Statistical Machine Translation (SMT), using Recurrent neural networks (RNN), Seq-to-seq without Attention mechanism, and Seq-to-seq with Attention mechanism.

Converting text into human-like speech has been a primary hurdle in the field of natural language processing. Over the previous few years, TTS research has made significant progress, with various separate sub-systems of a comprehensive TTS system considerably enhanced. Incorporating advancements from previous models like Tacotron and WaveNet, Tacotron 2 represents the next generation of text-to-speech technology. Utilizing the Bahdanau attention mechanism, this approach enhances translation accuracy by dynamically focusing on relevant parts of input sentences. This method not only improves translation quality but also ensures that critical health information is accessible globally, fostering better understanding and management of COVID-19 treatments across different languages.

II. LITERATURE SURVEY

The existing works [1] Ghosh and et al. proposed a model where, when translating from one Indian language to another. This study revealed that early research did not pay enough attention to the treatment of local proverbs/idioms. This article focuses on the translations of the two most widely spoken Indian languages, Marathi and Telugu. Proverbs and idioms have received special attention in these languages.

[2] Cho et al. analyzed neural machine translation using RNN Encoder-Decoder and closed recursive convolutional neural networks. Their findings indicate that these models excel at translating short sentences without uncommon words but struggle with longer sentences containing unfamiliar vocabulary. Notably, the study suggests that closed recursive convolutional networks can automatically learn the grammatical structure of phrases. The suggested Closed Recursive Convolutional Network automatically learns the grammatical structure of phrases, according to this research.

[3] Sutskever et al. introduced a flexible end-to-end method for learning sequences with minimal structural assumptions. This approach involves encoding input sequences into a fixed-length vector using a multi-layer LSTM, followed by decoding the target sequence from this vector using another deep LSTM. A significant finding of this work is that the LSTM achieved a BLEU score of 34.8 on the English-to-French translation task from the WMT'14 dataset. despite the LSTM BLEU score falling for terms outside the lexicon.

[4] Isaac Elias and et al. proposed parallel Tacotron2, a non-autoregressive neural text-to-speech model that uses a completely differentiable period model and doesn't require a regulated period input, is described in this white paper. The duration model is based on a one-of-a-kind Soft Dynamic Time Warping based attention and iterative reconstruction loss mechanism. This model can automatically recall marker frame alignment and duration. According to the findings, the parallel Tacotron 2 exceeds the subjective spontaneity baseline in numerous multi-speaker tests. The ability to control the duration has also been proven.

Previous works in cross-lingual neural machine translation for healthcare have faced several limitations. One major issue is limited language coverage, with models often focusing on widely spoken languages while neglecting those less common, restricting accessibility in diverse regions. Additionally, these models frequently struggle with accurately interpreting nuanced medical terminology and context, leading to potential misunderstandings. The scarcity of high-quality, domain-specific bilingual datasets further hampers translation performance and accuracy. Moreover, earlier models were computationally intensive, posing challenges for deployment in resource-constrained settings. Static attention mechanisms in traditional models failed to dynamically adapt to the varying importance of sentence components, affecting translation quality. Lastly, many models lacked robustness, resulting in inconsistent performance across different contexts and scenarios.

III. METHODOLOGY

Firstly, the system architecture of the proposed system is as shown in the fig.1 and fig.2. The architecture of the Bahdanau attention mechanism contains majorly two parts one is encoder section and decoder section. Similarly, Tacotron-2 system contains convolution and vocoding.

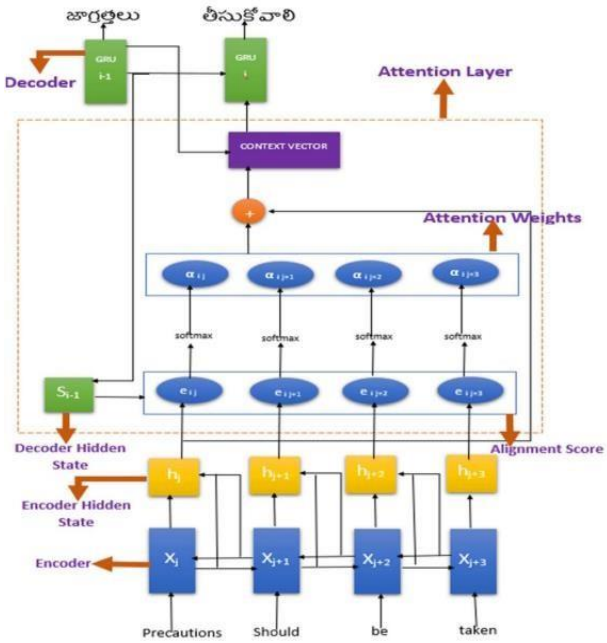

Fig.1 System Architecture of Bahdanau attention mechanism

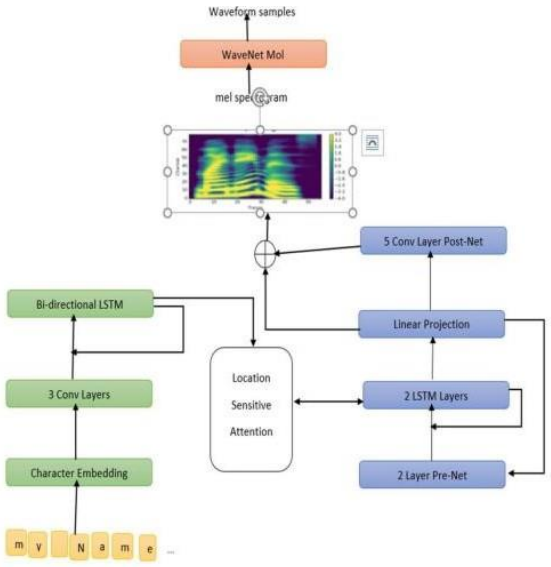

Fig.2 System Architecture of Tacotron -2

A. Dataset Description

Dataset is taken from the kaggle [1]. Dataset size is 42.1MB. Dataset contains around 70% of training data and 30% of testing data. It contains English to Telugu and vice-versa, English to Hindi and vice-versa, and Telugu to Hindi vice-versa sequence to sequence translation data. The total number of English to Telugu and vice-versa seq-to-seq are of 5615 that is 1.7MB. The total number of English to Hindi and vice-versa seq-to-seq are of 127608 that is 38.8MB. The total number of Hindi to Telugu and vice-versa seq-to-seq are of 3413 that is 1.6MB of the total memory.

Seq-to-Seq language	Count
English to Hindi and vice-versa	5615
English to Telugu and vice-versa	127608
Telugu to Hindi and vice-versa	3413

Fig.3 Used dataset details

The sentence length of the sequences present in dataset can be varied from minimum of 2 words to maximum of 21 words.

B. Model Implementation

Text-to-Text translation system using Bahdanau Attention Mechanism:

For the given input sequence, the output should also be a sequence and hence RNN has present. Any NMT model uses an encoder-decoder architecture which consists of RNNs to implement the translation system. An encoder takes a source phrase and turns it into a fixed-length vector. A Decoder produces the Encoded Vector's translation (output sentence). The Encoder-Decoder system is jointly trained to maximize the probability of generating a correct translation given a corresponding source sentence. A limitation of the encoder-decoder architecture is its reliance on the previous hidden state for predicting the next output word. This sequential processing restricts the decoder's access to the entire encoded input, leading to a potential degradation of the encoder's information over time. Thus attention is necessary.

Attention Mechanism:

Sequence-to-sequence models often benefit from attention mechanisms, which allow the model to prioritize specific parts of the input sequence when generating each output element. This is essential as the translation of a word can be influenced by its context within the sentence. In Bahdanau Attention Mechanism, the system works by learning to ‘align and translate’ jointly. 10 When the NMT model outputs a translated word as an output, it will soft-search for a set of points in the source sentence at the encoder's end, looking for the positions with the highest concentration of relevant information. It's similar to choosing the terms that make the most sense in the final translation.

Bahdanau attention mechanism:

Bahdanau's attention mechanism, often referred to as "additive attention," calculates alignment scores by combining information from both the encoder and decoder hidden states. This learned alignment helps the model determine which parts of the input are most relevant for generating each output token.. [9]. traditional sequence-to-sequence models that rely solely on the final encoder hidden state, attention mechanisms incorporate information from all encoder hidden states when constructing the context vector. This allows the decoder to focus on different parts of the input sequence at each output step Attention mechanisms employ a feed-forward network to calculate alignment scores between input and output sequences. By assigning higher weights to more relevant input elements, these mechanisms enable the model to focus on crucial information within the source sequence. The generated context vector, informed by these alignment scores and previous output words, is used to predict the next target word.. The Bahdanau attention mechanism is a sequence-to-

sequence model comprising an encoder, decoder, and attention layer. The attention layer encompasses components such as alignment weights, context vectors, and alignment scores. Alignment scores quantify the relevance between input positions and output positions. These scores are computed by comparing the encoder's hidden states with the previous decoder's hidden state, enabling the model to focus on pertinent input regions when predicting the next output word.

$X_{ab}=y(p_{a-1},h_a)$; Alignment score

Rather than relying on a fixed-length representation of the entire input sequence, the decoder in an attention-based model can dynamically focus on different parts of the source sentence when generating each output word. To achieve this, the model calculates an alignment vector, matching the length of the input sequence, for every output prediction. A soft max activation 11 k=1 function is used on the alignment scores to create the attention weights. The attention weights are created using a soft max activation function on the alignment scores

$\alpha^{ab}{}_{Tx}= exp(xab)/\Sigma_{k=1}\ (xbk)$; Attention Weight

Soft max activation function computes the probabilities of the input word for selection as target word, whose arithmetic sum will be 1 This assigns a weight to each word in the input sequence, indicating its relative importance for generating the current output.

Context Vector:

The context vector is computed as a weighted combination of the encoder's hidden states, where the weights are derived from the attention layer. This vector serves as input to the decoder for generating the final output. (h1,h2, …,h_x) of the encoder computes the context vector c_i, it maps to the input sentence.

$d_a=\Sigma_{a=l}{}^{Tx}\ \alpha_{ab}h_b$; Context Vector

Predicting the target word:

The decoder employs Context vector (d_b) to forecast the target word, which includes Decoder output from the previous time step (y_{b-1}) and previous decoder's hidden state (s_{b-1}).

$P_b=f(p_{b-1},d_b,y_{b-1})$

Text-to-Speech Conversion (Model-Tacotron2):

Tacotron 2 is a cutting-edge AI-powered speech synthesis technology that transforms text into speech. The neural network architecture of Tacotron 2 generates voice (audio) from text (script). The system is composed of two primary components: A recurrent neural network and a convolutional neural network. Given an input text sequence, state-of-the-art sequence-to-sequence models with attention mechanisms generate a series of mel spectrogram frames. Subsequently, a modified WaveNet model converts these mel spectrogram frames into raw audio waveform samples..[11]

Tacotron 2 essentially combines two neural networks: one to transform text into a visual representation of sound (a spectrogram), and another to convert this visual representation back into actual audio. WaveNet serves as the generative model responsible for transforming the synthesized spectrograms into raw audio waveforms. Griffin-Lim technique is used for phase estimation in the vocoding process, followed by Tacetron's inverse short-time Fourier transform. In comparison to WaveNet, the Griffin-Lim algorithm delivers lesser audio fidelity and character is tic noises. As a result, WaveNet is used instead of Griffin-Lim

To enhance audio quality, Tacotron 2 employs WaveNet instead of the traditional Griffin-Lim algorithm for converting spectrograms into waveforms. Renowned for its superior audio output, WaveNet, as used in Google Assistant, analyzes linguistic features, phoneme durations, and pitch at a frame rate of 5 milliseconds. Unlike previous models that relied on complex linguistic and acoustic data, Tacotron 2 is trained on simpler voice samples and corresponding text transcripts. Tacotron-2 can also tell the difference between heteronyms and pronounce them correctly based on their context. [12].

IV. EXPERIMENTAL RESULTS

The trained model is verified among the languages Telugu, English and Hindi, which yielded the following desirable outputs. Following are our output snips:

Class Encoder: Defines a class called Encoder which will be responsible for encoding input sequences.

Initialize (vocab_size, embedding_dim, enc_units, batch_size):This is the constructor for the Encoder class.It takes the following arguments: vocab_size:

The size of the vocabulary (number of unique words). batch_size: The batch size used during training.It initializes the following attributes: self.batch_size: Stores the batch size. self.enc_units: Stores the number of GRU units(Gated Recurrent Unit). self.embedding: Creates an embedding layer using tf.keras.layers.Embedding. This layer will convert integer-encoded words into dense vectors. self.gru: Creates a Gated Recurrent Unit (GRU) layer using tf.keras.layers.GRU. This layer will process the sequence of word embeddings. Call (x, hidden) This method defines the forward pass of the encoder.

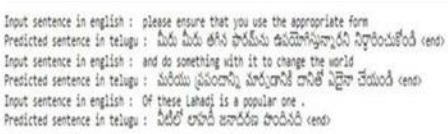

Fig. 5 English to Telugu Translation Output

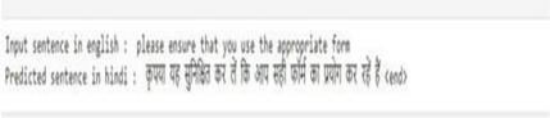

Fig.6 English to Hindi Translation Output

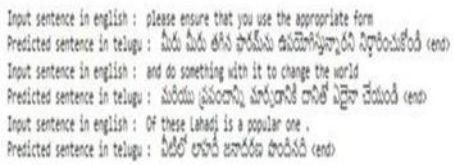

Fig.7 Hindi to Telugu Translation Output

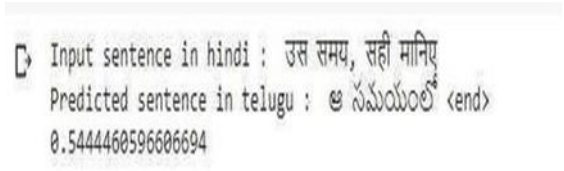

Fig.8 Hindi to English Translation Output

Fig.9 Telugu to Hindi Translation Output

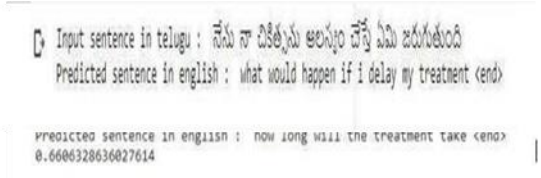

Fig.10 Telugu to English Translation Output

Following snippet is the result and audio generation snippet:

Fig.11 Audio generation

BLEU is a metric used to assess the quality of a generated sentence compared to a reference sentence. A perfect match between the generated and reference text yields a maximum score of 1.0, while a complete mismatch results in a score of 0.0.

The following tables show the results of the models bahdanau attention mechanism and Tacotron 2.The bahdanau attention mechanism model BLEU scores are as mentioned in the table 1. When English is the source language then the target languages Telugu and Hindi gives a score of 0.585 and 0.632. When Telugu is the source language then the target languages English and Hindi gives a score of 0.525 and 0.543. When Hindi is the source language then the target languages Telugu and English gives a score of 0.597 and 0.385.

SOURCE LANGUAGES \ TARGET LANGUAGES	ENGLISH	HINDI	TELUGU
ENGLISH	1.00(Ideal)	0.585	0.632
HINDI	0.385	1.00(Ideal)	0.597
TELUGU	0.525	0.543	1.00(Ideal)

Table.1 BLEU scores for cross-lingual machine translation using bahdanau attention mechanism

	English	Hindi	Telugu
Tacotron2	63.23%	61.56%	60.21%
WaveNet	56.69%	53.24%	50.02%

Table.2 TTS–Tacotron 2 model Accuracy in comparison with Wavenet model (normalised cross-correlation)

The Tacotron 2 model gave an accuracy score of 63.23%, 61.56% and 60.21% for English, Hindi and

Telugu respectively. And the corresponding accuracy scores of the languages in other model that is WaveNet are compared and are shown in the table 2.

V.CONCLUSION

Implementing cross-lingual NMT for COVID-19 treatment using the Encoder-Decoder model Bahdanau attention mechanism can significantly aid in disseminating crucial medical information globally. This project requires careful planning, high-quality data, and robust evaluation to ensure effectiveness and reliability. This can probably be attributed to the limited size of the vocabulary and dataset size, small feed forward network and diverse vocabulary and grammatical rules within the limited dataset with less number of possible number of eligible permutations of vocabulary. Glove encoding works better for English while one hot encoding yields better results for local languages in feature vector encoding probably due to extensive trained glove vectors for English and its grammatical rules. The TTS system accuracy is less probably due to difference in pronunciation and voice resulting in major difference in DFT values in the spectra in calculation of accuracy.

REFERENCE

[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. AssociationforComputationalLinguistics(ACL),(2015).pp.103–111.

[4] Hoch Reiter, 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,Yonghuietal."Google'sNMT."ArXive-prints(2016):pp.1–23.

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

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

[10] "English-Hindi Neural Machine Translation-LSTM Seq2Seq and ConvS2S", byGaurav Tiwari, Arushi Sharma, Aman Sahotra and Rajiv Kapoor published in 2020 International Conference on Communication and Signal Processing (ICCSP). 21

[11] "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions" by Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Che, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A.Saurous, Yannis Agiomyrgiannakis, Yonghui Wu , accepted to ICASSP 2018

[12] "Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling" by Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, RJ SkerryRyan, Yonghui Wu