Neural Machine Translation of Dravidian Languages to English

Peruri Sai Venkat
Dept. of CSE
National Institute of Technology
Puducherry

Dharavath Nithin
Dept. of CSE
National Institute of Technology
Puducherry

Abstract

In a globalized world where diverse cultures and languages converge, effective communication is often hindered by language barriers. This is the primary hurdle faced by many people when they travel to a new location where the language is unfamiliar. In this paper the Neural machine translation (NMT) system mainly focused on translation of three Dravidian languages spoken in southern region of India, i.e., Telugu (TE), Kannada (KN), and Malayalam (ML) to English (EN). As these Dravidian languages (TE, KN, ML) are low-resourced, a new parallel corpus of TE-EN, KN-EN, ML-EN was prepared by fine-tuning the existing datasets. Three deep learning models LSTM, Bi-LSTM and GRU are used for translation of Dravidian languages to English. On analysis of results, the limitation of this system showed that translation affected for lengthy sentences (sentence with more than 25 words) in all three parallel corpuses.

Keywords: Neural Machine Translation (NMT), RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory), Bi-LSTM (Bidirectional LSTM), GRU (Gated Recurrent Unit), Dravidian languages.

1 Introduction

Language translation is an important mechanism that connects linguistic groups and promotes global communication and understanding. In an increasingly interconnected world, the need for effective and efficient translation systems has never been greater. The proposed Neural machine translation (NMT) system is a powerful technology that addresses this need by translating the text of Dravidian languages to English text.

In this era of globalization and digitalization, machine translation has made amazing advancements, particularly in the context of widely spoken languages. However, there is an urgent need to extend the scope of machine translation by incorporating the world's rich linguistic diversity, especially languages that are underrepresented in technology-driven solutions. The perfect example of such languages are Dravidian languages. Telugu, Kannada, and Malayalam, primarily spoken in the southern region of India, are vibrant Dravidian languages with distinct linguistic characteristics and historical significance. Developing an efficient machine translation system for these languages not only fulfills an immediate need but also contributes to the broader mission of preserving linguistic heritage and promoting inclusivity in technology.

The Dravidian languages have a rich cultural heritage and are spoken by millions of people in India and around the world. While they have expanded in diverse domains, from literature to cinema, the accessibility of digital content and services in these languages has been relatively limited. This digital gap is due to the challenges of developing robust and accurate machine translation systems for these languages. Existing translation models often prioritize major world languages, leaving low-resourced languages underrepresented in the technology landscape.

The rest of the paper is structured as follows: Section 2 provides a literature review of related work in the field of machine translation. Methodology, outlining the data collection, preprocessing, and model development processes is discussed in Section 3. Section 4 presents the experimental setup, evaluation metrics, and the analysis of results on models' performance on the target languages. Finally, Section 5 concludes the paper and describes the directions for future research.

2 Literature Review

Sai Koneru et al.,(2021) proposed the Unsupervised Neural Machine Translation (UNMT) for translation of Dravidian languages (Kannada) to English and vice versa. This UNMT model performed well when source and target languages which are similar and monolingual data for both languages belong to the same domain. But in realistic scenarios these conditions are rarely fulfilled.

J.Sangeetha and S.Jothilakshmi (2017), developed a Speech-to-speech translation system which converts English speech to English text and then translated to Tamil and Malayalam text which is further converted to Tamil and Malayalam speech. ASR is used for speech recognition; TTS system, SVM and HMM are used for the development of the model, which English-Tamil translation got the accuracy of 85% and English-Malayalam translation got the accuracy of 83%.

Aditya Vyawahare et al.,(2022) focused on five language pairs Kannada-Tamil, Kannada-Telugu, Kannada-Malayalam, Kannada-Sanskrit, Kannada-Tulu which were trained using Seq2Seq models such as LSTM, BiLSTM, Conv2Seq; and pretrained models, transformers were also used. Therefore, LSTM performed well for Kn-Ml and Kn-Sn translations, BiLSTM performed well for Kn-Te, Transformer performed well for Kn-Tu and pretrained model performed well for Kn-Ta.

Bharathi Raja Chakravarthi et al.,(2019) proposed a Multilingual Multimodal Neural Machine Translation (MMNMT) for translating the closely related Dravidian languages to English. Phonetic transcriptions were also used along with the parallel corpus which improved the translation performance.

B. Premjith et al.,(2019) proposed an Encoder-Decoder network which consists of LSTM, BiLSTM neural networks and Attention mechanism is also used by considering the lengthy sentences as well. Four language pairs English-Tamil, English-Hindi, English-Malayalam, English-Punjabi in which English text is translated to Dravidian languages text where BiLSTM gave better performance among other models in all pairs.

Semi-supervised Neural Machine Translation (SMT), LSTM RNN and Transformer are the three models used for translation of English to Manipuri by Singh SM and Singh TD (2022). Results showed that SMT out-performed against supervised and mBART, achieved improvements of BLEU score by +4.5 and +1.2 respectively.

Unnikrishnan P et al.,(2010) proposed a Statistical Machine Translation (SMT) model for translation of English to Dravidian Languages (Kannada & Malayalam) which the dataset also includes the syntactic and morphological information to the corpus.

Santhanavijayan, A. et al.,(2020) proposed a hybrid system such as phrase-based translation, word-alignment and language model with transition probability computation which translates Hindi to Malayalam. This model yields average precision of 90.7% with Word Error Rate (WER) of 9.1.

Meetei LS et al.,(2023) proposed a multimodal machine translation which translates Hindi captions for images to English. The input for Encoders is image and its corresponding Hindi caption. They used VGG-19, a pre-trained CNN model which extracts numerical data from images. As a result, they achieved increment of BLEU score by 1.8.

Wanying Xie (2021) used two approaches for translation on four language pairs : English-Tamil, English-Telugu, English-Malayalam and Tamil-Telugu. One approach is multilingual translation and the other is back translation. Results showed that translation was performed well for English-Tamil, English-Telugu and Tamil -Telugu pairs compared to English-Malayalam on their proposed method.

Almost every paper which are reviewed, the translations are done from English to Dravidian languages but very less research is done in translation of Dravidian languages to English. Considering this point, our research is focused on translation of three language pairs: Telugu-English, Kannada-English and Malayalam-English and developed a system which is trained on a well fine-tuned dataset.

3 Proposed Methodology

3.1 Objectives

The primary goal of this paper is to bridge the digital gap by developing an efficient neural machine translation system for Telugu, Kannada, and Malayalam which will enable real-time translation of Dravidian languages text to English text and enhance the accessibility of digital content, for example generating live captions for

YouTube videos that are in these Dravidian languages and services for users of these Dravidian languages.

3.2 Data Collection

The parallel corpus for all three pairs TE-EN, KN-EN, ML-EN are collected from Kaggle and then fine-tuned all three parallel pairs using various existing translation systems like google translate so that the models are trained on accurate dataset. English sentences are same in all three parallel corpora.

Language Pair	Telugu - English	
No. of sentences	1,10,204	1,10,204
No. of words	60,537	34,883
Avg. words/sentence	8	10

Table 1: Telugu-English Language pair.

Language Pair	Kannada - English	
No. of sentences	1,10,204	1,10,204
No. of words	72,457	34,883
Avg. words/sentence	8	10

Table 2: Kannada-English Language pair.

Language Pair	Malayalam - English	
No. of sentences	1,10,204	1,10,204
No. of words	88,602	34,883
Avg. words/sentence	7	10

Table 3: Malayalam-English Language pair.

Table 1-3 refers to metadata of the collected dataset, which shows no. of parallel sentences, no. of unique words (vocabulary of the language) and average no. of words per sentence in each of the language pairs.

3.3 Data Preprocessing

As the raw data may contain the null values and duplicates, it affects the translation quality. So, the basic cleaning such as removal of null values, duplicates, special characters and punctations are done on all three parallel corpuses. For removing null values and duplicates, Pandas library was used and for removing punctations and special characters regex library was used.

3.4 Feature Selection

All unique words in each of the Dravidian languages (Telugu, Kannada and Malayalam) are converted to numerical data and selected as features for each model. These features are given as input for the model and trained, and same is translated to numerical data as output of the English unique words.

3.5 Model Selection

Text-to-text machine translation system for Telugu, Kannada, and Malayalam is implemented using three deep learning models: Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Gated Recurrent Unit (GRU). Each language pair is trained on all three models and the model which is outperformed among three is chosen for that particular language pair.

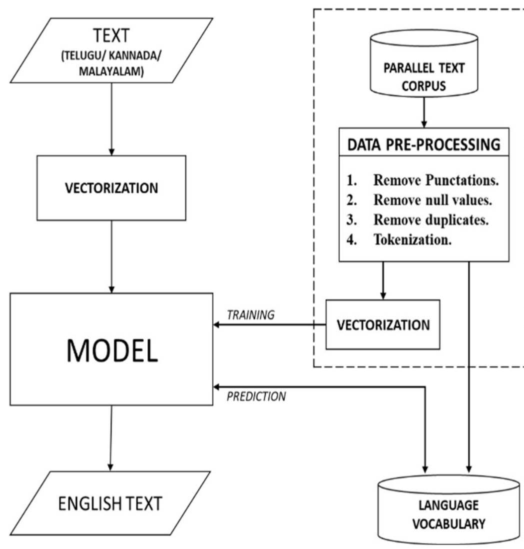

Figure 1: System architecture.

4 Results

4.1 Experimental Setup

All three parallel corpuses are taken and basic cleaning is performed using Pandas and Regex modules. From the dataset, each of the languages (TE, KN, ML, EN) were taken and all unique words of that language in the prepared dataset are considered as features and tokenized using Keras preprocessing text tokenizer, and then converted to numerical data which the deep learning models are trained. Each language pair is split into train data and test data each of 70% and 30% respectively. Then, train data of each language pairs TE-EN, KN-EN, and ML-EN are taken and trained on Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Gated Recurrent Unit (GRU).

LSTM (Long Short-Term Memory): It is a type of Recurrent Neural Network (RNN) which has a capability of capturing long-range dependencies in sequential data. The LSTM cell has gates: Forget gate, Input gate and Output gate which makes LSTM to learn and retain information over long sequences.

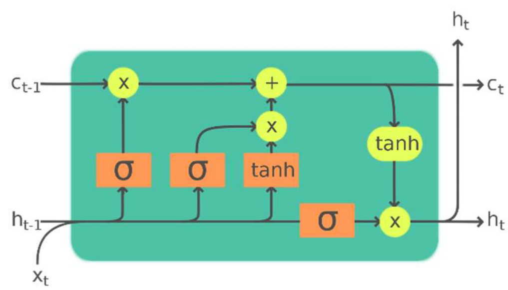

Figure 2: LSTM cell.

Equations for LSTM cell are:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$\tilde{c}_t = \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (5)$$

$$h_t = o_t \odot \sigma_h(c_t) \quad (6)$$

$$\sigma_x = \left(\frac{1}{1+e^{-x}}\right) \quad (7)$$

Here,

- $x_t \in \mathbb{R}$: input vector
- $f_t \in (0,1)^h$: forget gate's activation vector
- $i_t \in (0,1)^h$: input gate's activation vector
- $o_t \in (0,1)^h$: output gate's activation vector
- $h_t \in (-1,1)^h$: hidden state vector
- $\check{c}_t \in (-1,1)^h$: cell input activation vector
- $c_t \in \mathbb{R}^h$: cell state vector
- $W \in \mathbb{R}^{h \times d}$, $U \in \mathbb{R}^{h \times h}$: weight matrices
- $b \in \mathbb{R}^h$: bias vector parameters
- $\sigma(x)$: sigmoid activation function;

where superscripts d denotes no. of input features, h denotes no. of hidden parameters and initial values of c_0 & h_0 is 0, the operator ʘ represents element-wise product, subscript t indicates the time step.

Equations 1-7 of the LSTM cell shown in Figure 2 enables the model to selectively retain, discard, and update information from the source language to target language, facilitating the generation of accurate and contextually relevant translations.

BiLSTM (Bidirectional Long Short-Term Memory) : It is an extension of LSTM model in which it captures information from past and future long term dependencies between time steps of time series or sequential data. Figure 3 shows the architecture of BiLSTM demonstrating how the information is captured from both directions.

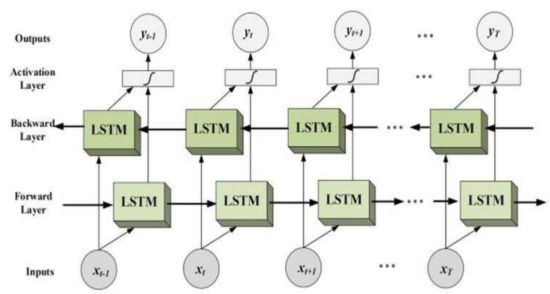

Figure 3: BiLSTM Neural Network.

GRU (Gated Recurrent Unit) : It is a type of RNN, similar to LSTM which has the ability to capture and remember information over long sequences and controls the flow of information in the network.

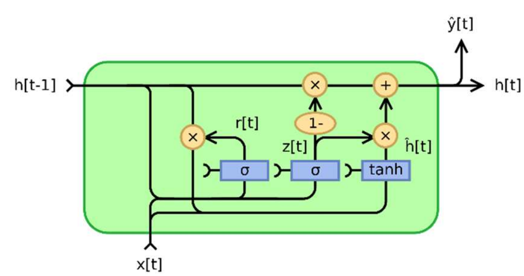

Figure 4: GRU Neural Network.

Equations for Gated Unit are:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (8)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (9)$$

$$\hat{h}_t = \phi(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \quad (10)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \quad (11)$$

Here,

- x_t : input vector
- h_t : output vector
- $\hat{h}_t$: candidate activation vector
- z_t : update gate vector
- r_t : reset gate vector
- W, U and b : parameter matrices and bias vector.

where initial values are $t = 0$ and $h_0 = 0$.

Equations 8-11 of the GRU cell shown in Figure 4 enables the model to efficiently update its hidden state, selectively retain relevant information, and handle long-range dependencies.

From the test data of each language pair, 2000 random samples are taken along with their machine translations which are translated by the proposed model to evaluate the model performance with BLEU score metric. As the test data is very large, machine translation of whole test dataset for all three language pairs on all three LSTM, BiLSTM and GRU models take long time, so only 2000 random samples are taken.

Training parameters of all three deep learning models LSTM, BiLSTM, and GRU for all three TE-EN, KN-EN, and ML-EN language pairs are shown in Table 4.

No. of hidden layers	2	
No. of Hidden Units in each layer	256/512/1024	
Word Embedding Size	TE-EN	Vocabulary size of TE
	KN-EN	Vocabulary size of KN
	ML-EN	Vocabulary size of ML
Batch Size	128	
No. of epochs	LSTM	60
	BiLSTM	30
	GRU	25
Optimization Method	Adaptive Moment Estimation (Adam)	

Table 4: Training Parameters.

Total time taken to train all three deep learning models LSTM, BiLSTM and GRU on all three language pairs TE-EN, KN-EN, ML-EN is approximately 15 hours, and trained of 5 GPUs.

4.2 Evaluation Metrics

To know the quality of translations, evaluation is done using *accuracy* metric and *BLEU score* metric. Based on the accuracy, no. of hidden units in each layer is considered to be 256 as best option in LSTM, BiLSTM as well as GRU. As deep learning models took long time for training, accuracy is recorded only for models with hidden units 256. Accuracy is calculated for whole test data of all three language pairs. BLEU score is evaluated using 2000 random samples from the whole test data. BLEU score is calculated by comparing the machine-translated sentences with set of reference sentences and a score between 0 to 1 is given, where 1 means machine-translated sentence matches the reference sentences. Corpus_bleu function is used for calculating BLEU score which is provided in NLTK library with equal weights of 0.25 for all 4 grams.

4.3 Analysis on Results

On observing the accuracy of the models, better translations are provided when the hidden units are considered as 256. Unexpected results are observed when hidden units are considered as 512 and 1024. Table 5 and Table 6 represents the accuracy results and BLEU score respectively.

For Telugu-English language pair, loss vs accuracy is plotted and shown below :

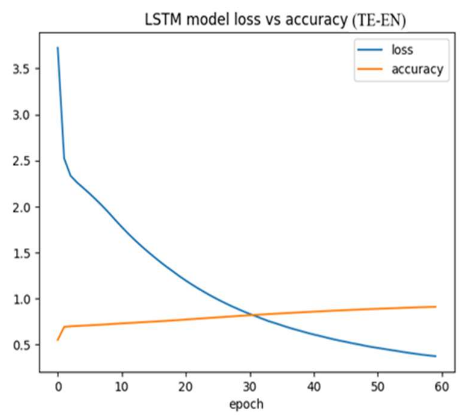

Figure 5: loss VS accuracy is plotted for LSTM model for TE-EN pair.

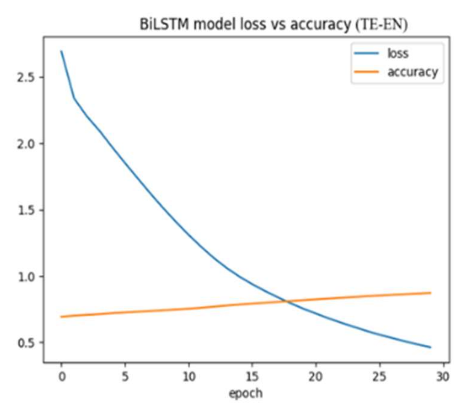

Figure 6: loss VS accuracy is plotted for BiLSTM model for TE-EN pair.

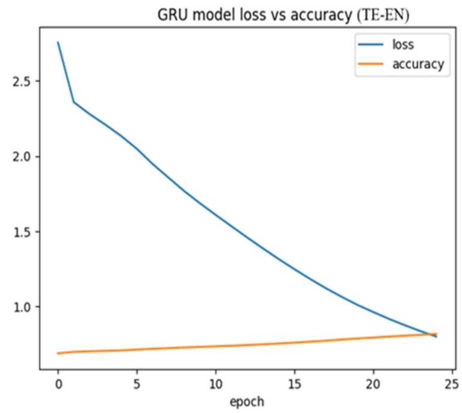

Figure 7: loss VS accuracy is plotted for GRU model for TE-EN pair.

For Kannada-English language pair, loss vs accuracy is plotted and shown below :

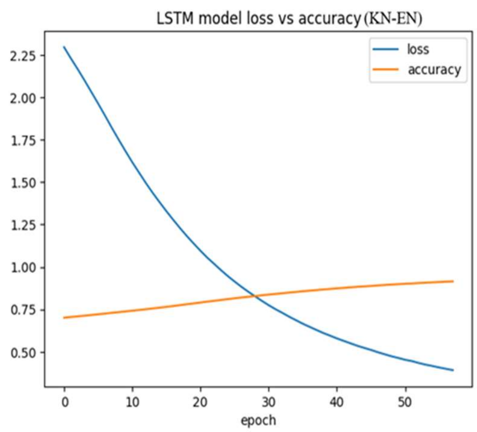

Figure 8: loss VS accuracy is plotted for LSTM model for KN-EN pair.

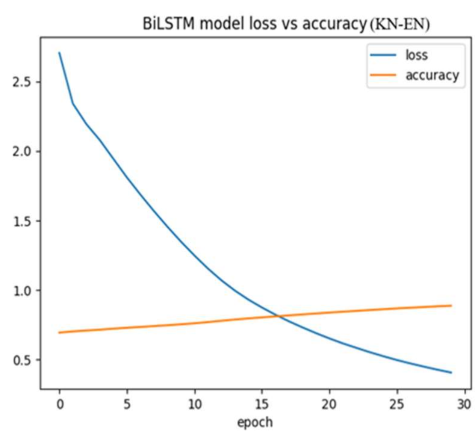

Figure 9: loss VS accuracy is plotted for BiLSTM model for KN-EN pair.

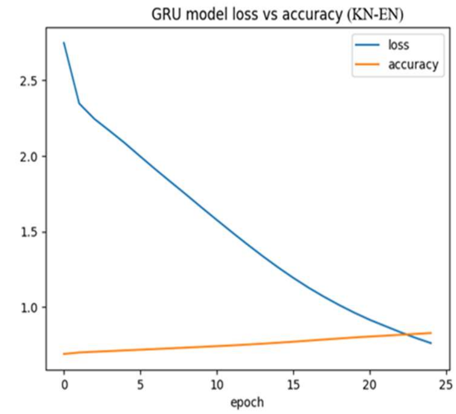

Figure 10: loss VS accuracy is plotted for GRU model for KN-EN pair.

For Malayalam-English language pair, loss vs accuracy is plotted and shown below :

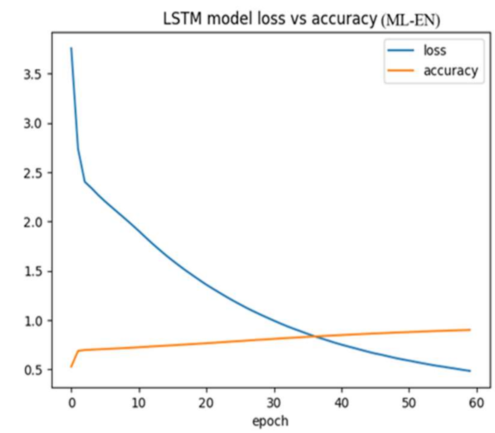

Figure 11: loss VS accuracy is plotted for LSTM model for ML-EN pair.

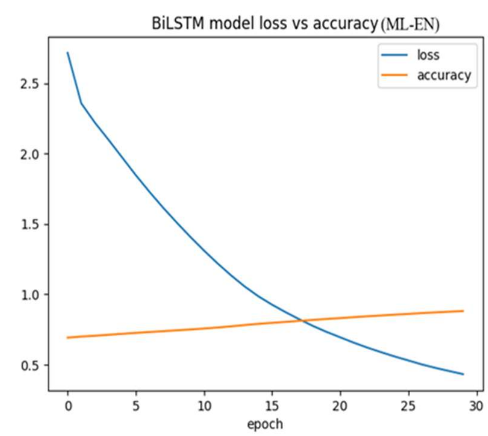

Figure 12: loss VS accuracy is plotted for BiLSTM model for ML-EN pair.

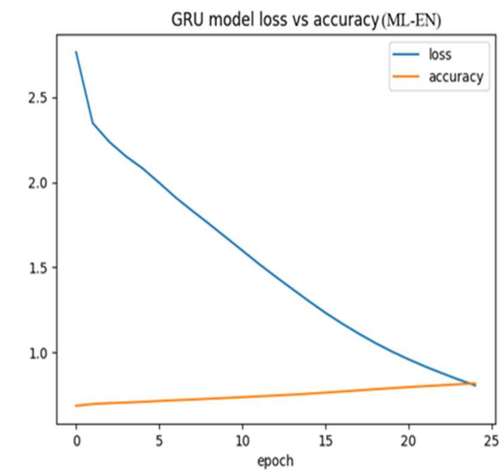

Figure 13: loss VS accuracy is plotted for GRU model for ML-EN pair.

MODEL	ACCURACY		
	TE-EN	KN-EN	ML-EN
LSTM	72.68%	71.71%	72.27%
BiLSTM	72.05%	72.78%	71.83%
GRU	72.39%	72.55%	71.92%

Table 5: Accuracy of all language pairs on all 3 models when hidden units=256 in each layer.

MODEL	BLEU SCORE		
	TE-EN	KN-EN	ML-EN
LSTM	0.599	0.328	0.602
BiLSTM	0.327	0.334	0.561
GRU	0.303	0.287	0.459

Table 6: BLEU score of all language pairs on all 3 models.

On observing the accuracy and BLEU score from Table 5 and Table 6 respectively; Telugu-English and Malayalam-English translations LSTM model performed well and Kannada-English translations BiLSTM performed well.

5 Conclusion and Future Work

In this paper, new dataset was prepared for Telugu-English, Kannada-English, and Malayalam-English language pairs and proposed a machine translation system, which translates Dravidian languages such as Telugu, Kannada, Malayalam to English using deep learning models. Due to the unavailability of the good corpus of Dravidian languages, it is still a challenging task to build a system where translations are accurate, and works for lengthy sentences. We contribute our prepared corpus which can be a valuable resource for further research and development in this field of Dravidian languages.

In the future, this research can be extended to other Dravidian languages as well as make translations from the speech of Dravidian languages and also train the models on a large dataset which includes various domains and translation output is also in speech of English. More advanced models, pre-trained models, and transformers can also be used for better results. The model described in this paper can be implemented to introduce the Dravidian languages to the digital world where translations are required, for example, many YouTube videos that are in Dravidian languages do not have auto-generated captions in English which is necessary.

References

Aditya Vyawahare, Rahul Tangsali, Aditya Mandke, Onkar Litake, and Dipali Kadam. 2022. PICT@DravidianLangTech-ACL2022: Neural Machine Translation on Dravidian Languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 177–183, Dublin, Ireland. Association for Computational Linguisticsm for English to Dravidian languages. *Applied Intelligence* 46, 534–550.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Bernardo Stearns, Arun Jayapal, Sridevy S, Mihael Arcan, Manel Zarrouk, and John P McCrae. 2019. Multilingual Multimodal Machine Translation for Dravidian Languages utilizing Phonetic Transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63, Dublin, Ireland. European Association for Machine Translation.

CH, Shih CC, Wang YC, Tsai RTH. 2022. Improvin g low-resource machine transliteration by using 3-way transfer learning. *Computer Speech & Language* 72:101283

Chandramma, P. Kumar Pareek, K. Swathi and P. Shetteppanavar, "An efficient machine translation model for Dravidian language," 2017 *2nd IEEE*

International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2017, pp. 2101-2105, doi: 10.1109/RTEICT.2017.8256970.

J.Sangeetha, S.Jothilakshmi. 2017. Speech translation system for English to Dravidian languages. *Applied Intelligence* 46, 534–550.

Luong, M.-T., and Manning, C. D. (2015). "Stanford neural machine translation systems for spoken language domains," *In Proceedings of the International Workshop on Spoken Language Translation*, 76–79.

Meetei LS, Singh SM, Singh A, Das R, Singh TD, Bandyopadhyay S. 2023. Hindi to English Multimodal Machine Translation on News Dataset in Low Resource Setting. Procedia Computer Science 218:2102–2109

Premjith, B., Kumar, M. Anand and Soman, K.P. "Neural Machine Translation System for English to Indian Language Translation Using MTIL Parallel Corpus" *Journal of Intelligent Systems*, vol. 28, no. 3, 2019, pp. 387-398. https://doi.org/10.1515/jisys-2019-2510

Raj Prajapati, Vedant Vijay Parikh, and Prasenjit Majumder. 2021. IRLAB-DAIICT@DravidianLangTech-EACL2021: Neural Machine Translation. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 262–265, Kyiv. Association for Computational Linguistics.

Sahinur Rahman Laskar, Bishwaraj Paul, Pankaj Dadure, Riyanka Manna, Partha Pakray, Sivaji Bandyopadhyay. 2023. English–Assamese neural machine translation using prior alignment and pre-trained language model. *Computer Speech & Language, 82:101524*

Sai Koneru, Danni Liu, and Jan Niehues. 2021. Unsupervised Machine Translation on Dravidian Languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 55–64, Kyiv. Association for Computational Linguistics.

Santhanavijayan, A., Naresh Kumar, D., Deepak, G. (2020). A Novel Hybridized Strategy for Machine Translation of Indian Languages. *In: Reddy, V., Prasad, V., Wang, J., Reddy, K. (eds) Soft Computing and Signal Processing. ICSCSP 2019. Advances in Intelligent Systems and Computing, vol 1118. Springer, Singapore.* https://doi.org/10.1007/978-981-15-2475-2_34

Singh SM, Singh TD.2022. Low resource machine translation of English–Manipuri: a semi-supervised approach. *Expert System Applications 209:118187*

Unnikrishnan P, Antony P, Soman K (2010) A novel approach for English to south Dravidian language statistical machine translation system. *Int J Computer Science Eng 2(08):2749–2759*

Wanying Xie. 2021. GX@DravidianLangTech-EACL2021: Multilingual Neural Machine Translation and Back-translation. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 146–153, Kyiv. Association for Computational Linguistics.