IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304275 |

A Neural Machine Translation Method for the English-Telugu Language Pair Based on LSTM

T. Chandu Venkata Mani Prasanth, V. Tendul Pavan Kumar, K. Teja Venkata Vinesh Kumar, K. Siva Charan Bharath, Dr. Bh. V. S. R. K. Raju

Department of Information Technology, S.R.K.R. Engineering College, Bhimavaram, India

Professor, Department of Information Technology, S.R.K.R. Engineering College, Bhimavaram, India

ABSTRACT: This project focuses on building a language translator specifically tailored for translating between Telugu and English using Long Short-Term Memory (LSTM) neural networks. Leveraging a dataset containing parallel sentences in English and their corresponding Telugu translations, we employ a train-test split to prepare the data for model training and evaluation. The development process involves several stages, including data collection, preprocessing, model training, and evaluation. A substantial corpus of Telugu-English parallel text data is gathered, ensuring diverse linguistic patterns and vocabulary coverage. Preprocessing techniques such as tokenization, padding, and embedding are applied to prepare the data for model training. The core of the system comprises LSTM-based neural networks for both encoder and decoder components. The encoder network processes the input Telugu sentences, while the decoder generates the corresponding English translations. The training objective involves minimizing the translation loss using techniques like teacher forcing and gradient descent optimization. Through rigorous experimentation and fine-tuning, the proposed system achieves competitive performance in Telugu-English translation tasks. The results demonstrate the effectiveness of LSTM-based approaches in capturing linguistic nuances and producing high-quality translations. The developed translator holds significant potential for facilitating cross-lingual communication, fostering cultural exchange, and bridging language barriers in diverse contexts.

KEYWORD: Language Translator, LSTM, Telugu, English, Dataset, Training, Testing, Prediction

I. INTRODUCTION

Language translation plays a pivotal role in bridging communication barriers between individuals who speak different languages. In the digital age, the development of efficient and accurate translation systems has become increasingly important, enabling seamless communication across diverse linguistic communities. The technique of translating text from one language to another using software that incorporates both linguistic and computer information is known as machine translation, or MT. First, the machine translation (MT) system learns to translate text into the source language by merely using grammar to associate the meaning of words in the source language with the target language. However, because these methods were unable to capture the variety of sentence forms seen in the language, they did not yield satisfactory results. It takes a lot of time and competent artisans to translate between the two languages. To solve the shortcomings of legal-based systems, integrated translation techniques including mathematical translation (SMT) and neural translation (NMT) technologies have been introduced.[1]

The technique of translating text from one language to another using software by combining linguistic and computational expertise is known as machine translation, or MT. Initially, the machine translation (MT) system determines the translation of a text in the source language by merely using linguistic rules to associate the meaning of words in the source language with their target language. However, because these methods were unable to capture the different sentence structures found in the language, they did not produce satisfactory results. Those who are fluent in both languages are needed for this labor-intensive translation process. Then, in an effort to address the shortcomings of rule-based techniques, corpus-based translation techniques such as statistical machine translation (SMT) and neural machine translation (NMT) were created.[2]

Among the multitude of languages spoken worldwide, Telugu and English stand out as significant languages with large speaker populations and cultural influence. Hence, developing a language translator specifically tailored for translating

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304275 |

between Telugu and English holds great significance in this project, we aim to construct a language translator using Long Short-Term Memory (LSTM) neural networks, a powerful deep learning architecture renowned for its effectiveness in sequence-to-sequence tasks like translation. LSTM networks excel in capturing long-range dependencies and context in sequential data, making them well- suited for language translation tasks. The foundation of our language translator lies in a carefully curated dataset comprising parallel sentences in Telugu and their corresponding English translations. This dataset serves as the basis for training and evaluating the LSTM model. Through a rigorous process of data preprocessing, we split the dataset into training and testing subsets, ensuring that the model is trained on a diverse range of language patterns while also being adequately tested for generalization to unseen data. Deep A more modern method for machine translation is learning. Neural machine translation is a superior option for more accurate translation and offers greater performance than classical machine translation.[5] In conclusion, our study aims to create a trustworthy and effective Telugu-English translator by utilizing the potential of LSTM-based deep learning. Our aim is to promote understanding and connectedness in an increasingly interconnected world through the integration of artificial intelligence advancements and a dedication to cross-cultural communication.

II. EXISTING SYSTEMS

Several existing systems of language translation that utilize LSTM (Long Short-Term Memory) networks consist of:

1. Google Translate: LSTM layers are frequently used in neural machine translation (NMT) models, which are used by Google's translation service. Because LSTM networks can retain information over lengthy sequences, they are useful for translation jobs and are well-suited for handling sequential data such as language.
2. Microsoft Translator: Another neural machine translation tool is Microsoft Translator. For translation jobs, it probably uses LSTM or a related recurrent neural network (RNN) architecture.
3. DeepL: Known for producing translations of the highest caliber, DeepL is a well-liked translation service. To get its translation accuracy, it uses deep learning methods, potentially with LSTM networks.
4. OpenNMT: OpenNMT is a neural machine translation framework that is freely available. It offers machine translation implementations of several neural network topologies, including LSTM-based models.
5. Fairseq: Another open-source framework for machine translation and other sequence-to-sequence learning tasks is Fairseq. Among other designs for translation problems, it provides LSTM-based models.
6. Marian: Marian is a neural machine translation system that is quick and effective. It provides pre-trained models for several language pairs and supports LSTM-based architectures for translation jobs.

These systems train the mappings between source and target languages for translation purposes using LSTM networks, either solely or in conjunction with a broader neural network architecture.

Disadvantages of Existing Systems:

Contextual Ambiguity: Translators may struggle with ambiguous words or phrases that have different meanings based on context. This is particularly challenging for idioms, colloquial expressions, or words with multiple meanings.[1]
Rule-Based Translation Constraints: Translators cannot get other words which are not present in the dataset. Only some sentences get changed and some may not get changed.[4]

Time Consumption and Language Proficiency: This translation process is highly time-consuming and requires people who have proficiency in both languages. As a result, there will be lot of wastage of time and the person who don't know the other language mainly uses the translator but, in this person should be proficient in both the languages.[2]
Translation without Grammar Consideration: During translation grammar should be carefully observed. If the grammar is not carefully observed it may lead to wrong translation of sentences. [5]

Complexity Increase with Morphological Segmentation: If the model cannot translate large sentences into readable format, then it would be difficult for people to understand the translated text so the model should be able to build the large sentences also so that the person understands the translates text.[6]

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304275 |

III. PROPOSED SYSTEM

The proposed system aims to address the shortcomings of existing language translation tools by developing a robust Telugu-English translator using LSTM neural networks. By leveraging advanced deep learning techniques, the system will offer improved accuracy, enhanced fluency, and better contextual understanding in translations. It will feature a broader vocabulary coverage, ensuring accurate translations of diverse text types. Additionally, the system will be designed to work offline, eliminating the dependency on internet connectivity. Through these advancements, the proposed system seeks to provide users with a reliable and efficient tool for seamless communication between Telugu and English languages.

3.1 PROBLEM STATEMENT

Our proposed system aims to address the shortcomings of existing language translation tools by developing a robust Telugu-English translator using LSTM neural networks. By leveraging advanced deep learning techniques, our system will offer improved accuracy, enhanced fluency, and better contextual understanding in translations. It will feature a broader vocabulary coverage, ensuring accurate translations of diverse text types. Additionally, the system will be designed to work offline, eliminating the dependency on internet connectivity. Through these advancements, the proposed system seeks to provide users with a reliable and efficient tool for seamless communication between Telugu and English languages.

Language translators offer several advantages, making communication across different languages more accessible and efficient. Here are some key advantages of language translators. Language translation helps people understand each other better, making it easy to talk, share ideas, and connect with Telugu speakers. Translation tools allow more people to read and understand information in Telugu, so everyone can learn and know things, no matter what language they speak. Translating with tools saves time and effort, making it quick and simple to change words from English to Telugu, and making learning languages easier too.

IV. SYSTEM ARCHITECTURE

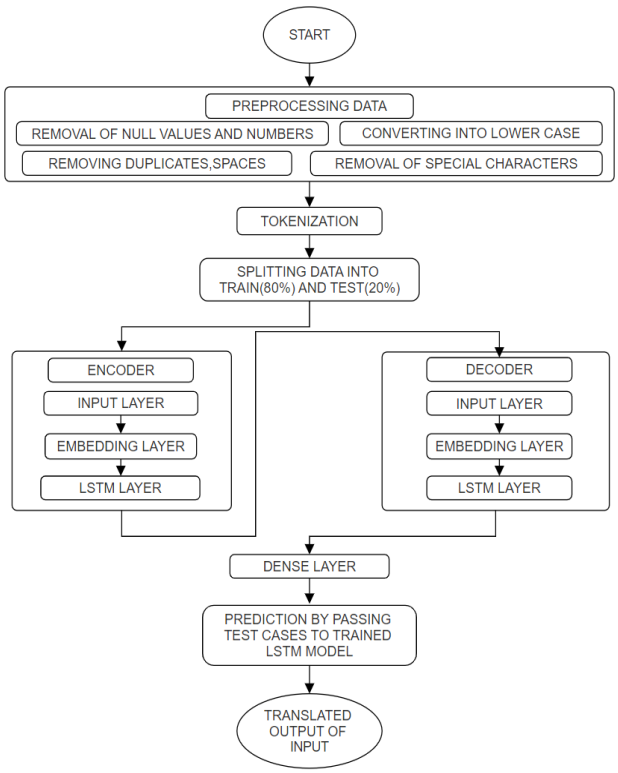

Fig-1: System architecture

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304275 |

As shown in the Fig-1, the user gives the input and this input goes through series of preprocessing steps like removal of special characters, removal of numbers, converting all the alphabets into lower case, adjusting spaces between two words, and this data is divided into training and testing data (80% for training and 20% for testing). The preprocessed data is fed into input layer. The data is then passed through the embedding layer, which converts input tokens into dense vectors. The embedded data is then processed through LSTM layers, which are designed to process sequential data. After the LSTM layers, the data is passed through dense layers, which process the LSTM output and produce the final prediction. The final output of the LSTM model represents the prediction.

4.1 LSTM ENCODER DECODER ARCHITECTURE

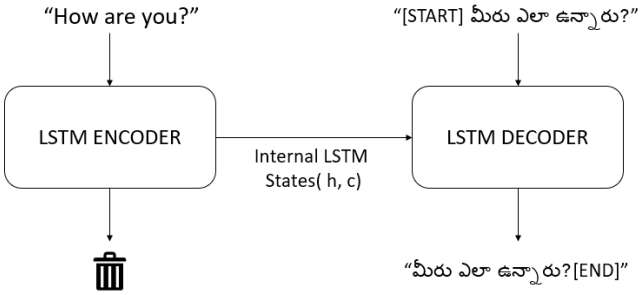

Fig-2: LSTM model

V. ALGORITHM

LONG SHORT-TERM MEMORY (LSTM)

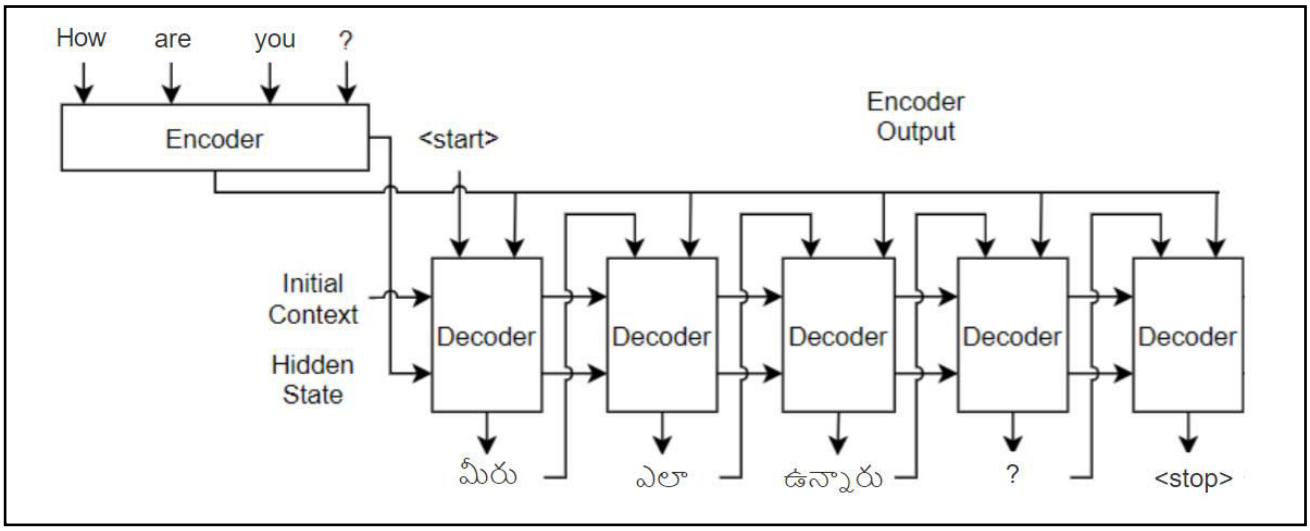

Fig-3: LSTM architecture

LSTM Encoder: The LSTM encoder processes the input sentence in the source language (English) word by word or token by token as shown in Fig-3. The LSTM encoder updates its hidden state based on the current input and the previous hidden state, effectively encoding the sequential information of the input sentence. Once the entire input sentence has been processed, the final hidden state of the LSTM encoder encapsulates the semantic information of the entire sentence in a fixed-length vector representation. This fixed-length vector serves as a context vector, capturing the essence of the input sentence, which is then passed to the decoder.

LSTM Decoder: As shown in Fig-3 the LSTM decoder takes the context vector (obtained from the LSTM encoder) and generates the target sentence in the desired language (Telugu) word by word. At the beginning of decoding, the

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304275 |

decoder is initialized with the context vector as its initial hidden state. Similar to the encoder, the decoder processes the target sentence step by step, generating one word/token at each time step. At each time step, the LSTM decoder takes in the current context vector, the word/token embedding from the previous time step and its own hidden state from the previous time step. The LSTM decoder updates its hidden state based on these inputs, and using its internal memory mechanism, it predicts the next word/token in the target language. This process continues until an end-of-sentence token is generated or a predefined maximum length is reached. The output sequence generated by the decoder constitutes the translated sentence in the target language.

VI. SYSTEM IMPLEMENTATION

6.1 DATA COLLECTION

The data used in this project is taken from Kaggle which contains of two columns that are having English corpuses followed by Telugu corpuses in the next column. The sample is shown in the Fig-4.

We need three cups.	మాకు మూడు కప్పులు అవసరం.
I warned Tom not to come here.	టామ్‌ను ఇక్కడికి రానివ్వమని హెచ్చరించాను.
You two may leave.	మీరిద్దరూ వెళ్ళవచ్చు.
He feels very happy.	అతను చాలా సంతోషంగా ఉన్నాడు.
Tom wasn't smiling when he entered the room.	గదిలోకి ప్రవేశించినప్పుడు టామ్ నవ్వలేదు.
What can it be?	అది ఏమిటి?

Fig-4: Data collected

6.2 DATA PREPROCESSING

Text Cleaning: Removing any unnecessary characters, punctuation, or special symbols that may not contribute to the translation process. This can include removing whitespace, punctuation marks, or any other non-alphanumeric characters.

Lowercasing: Converting all text to lowercase to ensure consistency and reduce the vocabulary size. This helps in improving the accuracy of the translation by treating uppercase and lowercase versions of words as the same.
Stopword Removal: Eliminating common words like "the", "is", "are", etc., which do not carry significant meaning in translation and can potentially hinder the accuracy of the translation model.

Normalization: Standardizing the text by converting numbers, dates, and other numeric expressions into a common format to facilitate better translation.

Lemmatization: Reducing words to their base or root form to handle different inflections and variations. This helps in reducing the vocabulary size and improving the consistency of the translation. Sentence Segmentation: Splitting the text into individual sentences, which allows for more granular translation and better handling of context.
Tokenization: Splitting the input text into individual words or tokens. This step is crucial for breaking down the text into smaller units for processing.

6.3 MODEL TRAINING

This step contains of training the LSTM model by using the following steps:
Forward Pass: Feed the input sequences into the encoder LSTM to generate the context vector. Initialize the decoder LSTM with the context vector and start generating the output sequence.

Backpropagation and Optimization: Calculate the loss between the predicted sequence and the ground truth sequence using a suitable loss function (e.g., categorical cross-entropy). Use backpropagation through time (BPTT) to compute the gradients and update the model parameters using an optimizer (e.g., Adam, RMSProp).'

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304275 |

VII. RESULTS AND DISCUSSIONS

7.1 Model Loss

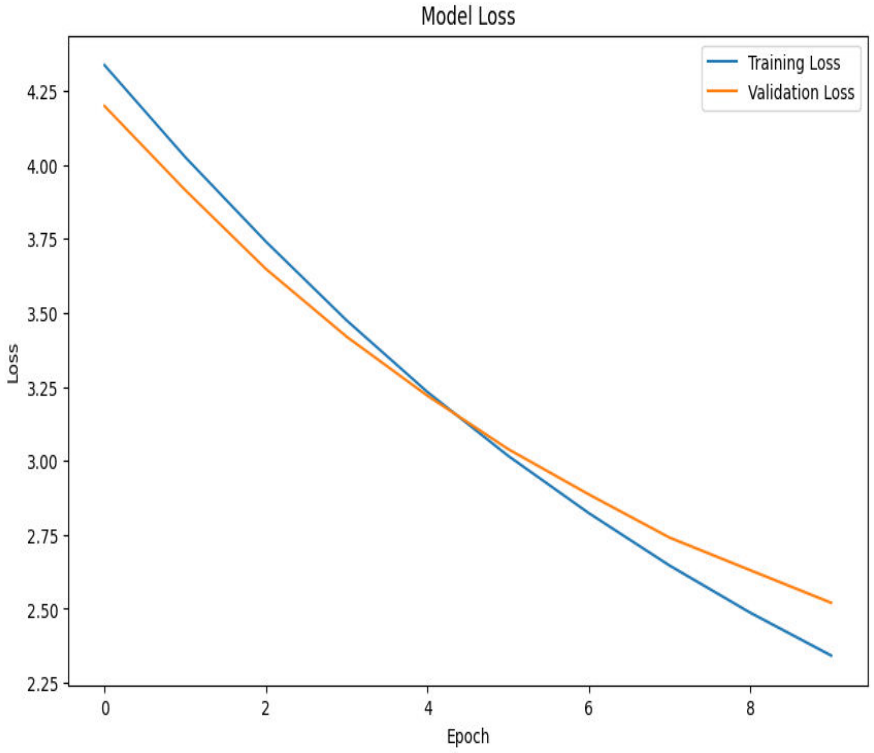

Fig-5: Model Loss

The x-axis of the graph in Fig-5 is labeled "Epoch" and it refers to the number of times that the entire training dataset is passed through the model. The y-axis is labeled "Loss" and it refers to how well the model is performing on a given set of data. In machine learning, loss is a way to measure how different the model's predictions are from the actual values. A lower loss indicates that the model is performing better.

Fig-5 shows that the training loss (the blue line) is generally decreasing over time. This means that the model is getting better at translating languages as it is trained on more data. The validation loss (the orange line) is also decreasing, but it is higher than the training loss. This is because the validation loss is measured on a separate dataset that the model has never seen before. This helps to ensure that the model is not simply memorizing the training data and that it can generalize to new data.

Overall, the graph in Fig-5 suggests that the language translation model is learning and improving over time.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304275 |

7.2 Sample Results

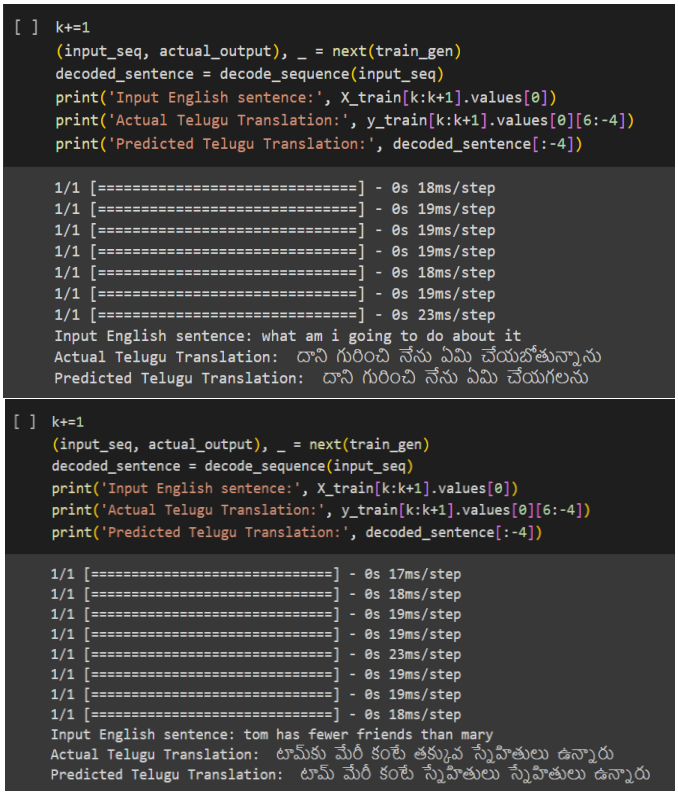

Fig-6: Results

As shown in Fig-6, the model trained resulted as a good translator but not that accurate since it should be trained on large dataset and have to be trained with a greater number of epochs so that it will result in good results, since it requires a large computer system to do this model, we have done this project to know how the language translator works with LSTM.

VIII. CONCLUSION

In conclusion, the project to create a Telugu-English translator using LSTM neural networks is a leap towards transcending language barriers. It reflects a harmonious blend of advanced computational linguistics and the desire to foster global communication. The initiative promises not only to facilitate understanding and share the rich tapestry of Telugu culture with the world but also to pave the way for innovations that could extend to other language pairs. With its user-friendly interface, it stands to serve a wide array of users, from students and educators to businesses and travelers. The project's scalable nature hints at a future where no language is left behind in the digital space, and its commitment to continuous improvement through user feedback ensures its growth and relevance over time. As we look ahead, the successful realization of this project has the potential to not only bridge linguistic gaps but also to unite communities, expand access to information, and enrich intercultural dialogue. Ultimately, this language translator is not merely a tool but a testament to the collaborative spirit of technology and humanity working in tandem towards a more interconnected world.

IX. FUTURE SCOPE

As we look ahead, the Telugu-English translation project stands on the cusp of significant advancements. Future enhancements will delve into the realms of advanced neural architectures like Transformer models, offering increased precision and accommodating the intricacies of language. A notable goal is the expansion into additional languages, broadening the horizon for linguistic inclusion. The project aspires to finely tune translations to reflect regional idioms

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304275 |

and cultural expressions, making conversations feel more natural and authentic. Anticipating the integration of speech-to-speech translation capabilities opens up new frontiers in real-time communication, breaking down barriers further. User experience will remain a central focus, with interfaces evolving through iterative design informed by user feedback. Specialized domain translation services are on the agenda, catering to sectors where technical accuracy is paramount. In the backdrop, the engine will be optimized for performance across various devices, ensuring accessibility and convenience. Enhancements in privacy safeguards will address the growing concerns over data security in digital interactions. The translation tool will also move towards supporting offline functionality, vital for uninterrupted access in connectivity-scarce regions. Educationally, the project has the potential to transform language learning, offering immersive, interactive tools tailored to the learner's journey. User profiles could bring a level of personalization previously unseen in translation tools, further smoothing the path of communication. Lastly, community-driven improvements will tap into the collective wisdom of users, creating a dynamic and ever-improving translation resource. The future of this project is a tapestry woven with technological innovation, aimed at creating a world where language is no longer a barrier but a bridge to a richer, more connected human experience.

REFERENCES

[1] Aman Sharma and Mr. Vibhor Sharma, Language translation using machine learning, International Research Journal of modernization in Engineering Technology and Science. Volume:03/Issue:06/June-2021.
[2] B.Premjith, M. Anand Kumar and K.P.Soman, Neural Machine Translation System for English to Indian Language Translation using MTIL Parallel Corpus, March 20, 2019.
[3] Middi Venkata Sai Rishit, Middi Appala Raju and Tanvir Ahmed Harris, Machine Translation using Natural Language Processing, MATEC Web of Conferences 277, 02004(2019).
[4] Syed Abdul Basit Andrabiand Abdul Wahid, Machine Translation System using Deep Learning for English to Urdu, Computational Intelligence and Neuroscience/2022.
[5] Shashi Pal Singh, Ajai Kumar, Hemant Darbari, Lenali Singh Anshika Rastogi and Shikha Jain, Machine Translation using Deep Learning, Conference Paper July 2017.
[6] K Hans (SSN College Of Engineering), Milton R S (SSN College Of Engineering), Improving the performance of Neural Machine Translation involving Morphologically rich languages, 8 JAN 2017.
[7] Lu Yang1, Da Chen1 and Wenxue Wu, Applications Research of Machine Learning Algorithm in Translation System, 6th International Conference on Machinery, Materials and Computing Technology (ICMMCT 2018).
[8] Arvinder Singh, Ninad Bhase, Manav Jain, and Tushar Ghorpade, Machine Translation Systems for English Captions to Hindi Language using Deep Learning, ITM Web of Conferences 44, 03004 (2022)
[9] Isaac O. Eleș emoyo, Machine Translation System for Numeral in English text to Yoruba language, Ife Journal of Information and Communication Technology, Volume 6, 2022. 23
[10] Keerthi Lingam E. Ramalakshmi and Srujana Inturi, English to Telugu rule-based Machine translation system, International Journal of Computer Applications (0975- 8887), Volume 101- No.2, September 2014.
[11] P. Sujatha and D. Lalitha Bhaskari, Sentence wise Telugu to English Translation of Vemana Sathakam using LSTM, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details