IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204110 |

# Conversion of Transliterated English Words to Regional Language Words using Transformer Neural Networks

**Akshara U J, Harriharan M K, Siva Alagesh S.**

UG Student, Dept. of IT, Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur, India

UG Student, Dept. of IT, Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur, India

Assistant Professor, Dept. of IT, Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur, India

**ABSTRACT:** The widespread use of Transliterated English words, blending English with regional languages, poses a significant challenge for effective digital communication in regional language contexts. Current translation systems struggle with this mix due to limited data and inherent Machine Translation constraints. To address this, our research proposes leveraging neural translation systems based on the Transformer architecture. By gathering bilingual datasets from online sources containing Transliterated English words and their regional language equivalents, we aim to seamlessly convert these words to their native forms. This approach promises to advance the digitalization of regional languages, offering a robust means for broader audiences to access regional content. By strengthening regional languages' digital presence, our initiative not only champions linguistic diversity but also underscores their importance in the evolving digital landscape.

**KEYWORDS**: Neural translation, Transformer architecture, bilingual datasets.

## I. INTRODUCTION

In the realm of digital communication, the coexistence of multiple languages presents both opportunities and challenges. One such linguistic phenomenon that has gained prominence in recent years is Transliterated English, a fusion of English and other regional languages frequently observed in social media discourse and online interactions. This blending of languages reflects the dynamic nature of communication in multicultural contexts but also poses significant obstacles for effective language processing and digitalization, particularly in regional language domains. Traditional translation systems often struggle to accurately interpret Transliterated English due to its unique linguistic characteristics and the scarcity of adequate training data. As a result, the seamless conversion of Transliterated English expressions into proper regional language poses a formidable challenge, hindering the accessibility and comprehension of regional content online. To address this challenge, our research proposes a novel approach leveraging advanced neural translation systems based on the Transformer architecture. By harnessing the power of machine learning and natural language processing techniques, we aim to develop a robust translation tool capable of accurately translating Transliterated English expressions into regional expressions (Tamil). Key to the success of this endeavor is the acquisition of bilingual datasets comprising Transliterated English text alongside its corresponding regional counterparts. Through careful curation and analysis of such datasets from online sources, we seek to train our neural translation model to understand and effectively process Transliterated English. The implications of this research extend beyond mere language translation; they encompass the broader goal of enhancing the digital presence and accessibility of regional languages. By providing a reliable mechanism for converting Transliterated English into regional language, our initiative seeks to democratize access to regional content, empowering broader audiences to engage with and appreciate the richness of the regional languages and culture in the digital sphere. Through this research, we aim to not only address the immediate challenges posed by Transliterated English in digital communication but also to underscore the importance of linguistic diversity and cultural heritage in the evolving digital landscape.

## II. RELATED WORK

### Adding visual attention into encoder-decoder model for multi-modal machine translation

The proposed work delves into the realm of multi-modal machine translation, emphasizing the integration of visual attention within the encoder-decoder model. In light of the evolving landscape of translation technology, this literature

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204110 |

review critically assesses the current methodologies for multi-modal translation, highlighting the pivotal role of visual information in refining translation outputs. Existing studies underscore the significance of capturing the intricate interactions between visual and textual features, with a growing recognition of the influence of global context provided by visual data. In this context, Chun XU's proposed approach, elucidated in the "Adding visual attention into encoder-decoder model for multi-modal machine translation," presents a pioneering method to concurrently incorporate visual information into both the encoder and decoder units, thereby facilitating a comprehensive understanding of the dynamics between image-sentence pairs and advancing the accuracy of multi-modal translation systems.

**Neural Machine Translation with Deep Attention**

This proposed work offers a comprehensive exploration of deep attention mechanisms within the context of neural machine translation. This literature review contextualizes the evolving landscape of neural machine translation techniques, emphasizing the crucial role of attention mechanisms in enhancing translation accuracy. Previous studies have demonstrated the efficacy of incorporating attention mechanisms to improve the alignment of source and target language sequences, thereby bolstering the overall translation performance. Zhang et al.'s research makes a significant contribution by delving into the nuances of deep attention mechanisms, shedding light on their potential to facilitate a more nuanced understanding of contextual relationships, ultimately enriching the translation process within the framework of neural machine translation.

**Beyond English-Centric Multilingual Machine Translation**

This proposed work critically examines the current landscape of multilingual machine translation, challenging the traditional English-centric approach. This literature review situates their work within the broader context of multilingual translation methodologies, emphasizing the growing recognition of the limitations imposed by an English-centric perspective. Previous research endeavors have highlighted the need for a more inclusive and diversified approach that encompasses the linguistic complexities and nuances inherent in non-English languages. Fan et al.'s research represents a significant step toward addressing these limitations by advocating for a broader, more inclusive framework that acknowledges the importance of non-English languages, thereby enriching the multilingual machine translation paradigm.

**Attention is all you need**

This proposed work marks a pivotal contribution to the field of natural language processing and machine learning. This literature review contextualizes their research within the broader landscape of attention mechanisms in neural networks, emphasizing the transformative role of attention mechanisms in revolutionizing various tasks, including machine translation and language understanding. Previous studies have underscored the significance of attention mechanisms in facilitating improved context awareness and information processing, thereby enhancing the efficiency and accuracy of neural network-based models. Vaswani et al.'s groundbreaking research represents a significant milestone in this trajectory, showcasing the potential of attention-based architectures as a fundamental building block in the development of sophisticated and high-performing neural network models.

## III. PROPOSED ARCHITECTURE

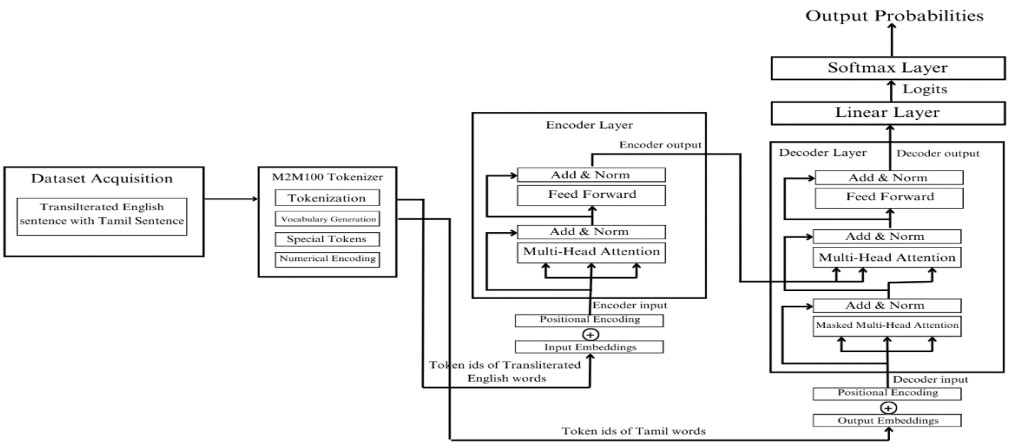


**Figure 1: Proposed architecture for the Conversion of transliterated English words to regional language words**

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204110 |

### 3.1 Dataset Acquisition

Acquiring the dataset posed a challenge due to the absence of readily available resources. The scarcity of existing datasets for this purpose necessitated a comprehensive approach to gathering data from various sources. Utilizing Python libraries like Beautiful Soup, we scraped the dataset from lyric websites hosting transliterated English content and regional language content. Each regional song lyric was meticulously matched with its transliterated English equivalent. The alignment ensured that every transliterated English sentence had a corresponding regional counterpart. To refine the dataset, we filtered out lengthy sentences and those containing symbols or emojis. This process was crucial for extracting meaningful content. By facilitating the seamless conversion of transliterated English expressions into pure regional expressions, it played a pivotal role in advancing the digitalization of the regional language. This initiative not only addressed the immediate challenge of dataset scarcity but also laid the groundwork for future research and development in this domain, fostering a deeper understanding of language dynamics in digital communication.

### 3.2 Dataset Preprocessing

Split the text into individual words or sub words (tokens). This step ensures that each word or sub word is treated as a separate unit during training. For languages like English, simple word-level tokenization may suffice. However, for languages like Tamil or other regional languages, sub word tokenization using techniques like Byte Pair Encoding (BPE) might be more effective due to their complex morphology. Ensure that all sequences have the same length by padding shorter sequences with a special token to match the length of the longest sequence in the dataset. This step is necessary for efficient batch processing during training. Build vocabularies for both the source (Transliterated English) and target (Tamil) languages. This step requires assigning a distinct integer index to each token. Unknown words can be assigned a special <UNK> token. Convert the tokenized sequences into numerical format by replacing each token with its corresponding index in the vocabulary. Divide the dataset into training, validation, and testing sets to assess the model's performance accurately. On our case, 80% of the data is used for training, 10% for validation, and the remaining portion for testing. Group the numerical sequences into batches, which are then fed into the model for training. Batching improves training efficiency by allowing parallel processing of multiple sequences. Since Transformer models do not inherently understand the order of the input sequences, positional encodings are added to the input embeddings to provide information about the position of each token in the sequence.

### 3.3 Language Translation using Transformer Architecture (Seq2Seq)

The proposed architecture is a four Encoder Layer and six Decoder Layer Each consisting of eight Heads attention units in it. The dataset is preprocessed in its initial stage. The source text and target text are padded. Then the Text are tokenized and Converted to a normal number. Later these numbers are converted to Tensors such that to be processed with Calculations. After Creating attention_mask for padding_id the data is ready to be fed into the model. These data are Batched together for Parallel Processing. These Batch are also used for updating gradients.

As an initial step Tokenized tensors choose their weights with the model dimension from the Embedding Layer. Then These Embedded Weights are added to Positional Encoded Input Tensors. These Input Tensors then fed into an Encoder Layer Which is a Layer that has Three Linear Layers (QUERY, KEY, VALUE). Input Tensors are matrix multiplied with QUERY, KEY and Then the output value received from that is added to the matrix multiplication of that value vector. In Between there will be a Layer Normalization Layer and Residual Connections that would help in building a strong attention mechanism. These Methods are repeated for Four times as there are Four Encoder Layers. Atlast these Embeddings are connected to a Feed Forward network with model dimension. The output Embedding that was exhibited from the Feed Forward network is called Memory.

Decoder Layer is the same as Encoder Layer. First the Target Inputs are Passed to the Embedding Layer and added to the positional encoding. Then with an attention mask the attention unit handles the value in sequence. By repeating this six times as it is a six layered decoder with 8 head attention and a Linear Layer with the model dimensions we would derive a Layer with Batch size and vocab size. Then the last layer with batch size and vocab size is passed to a softmax layer to make everything to a probability. Then the index of the maximum probability vector is considered as the output while evaluating and inferencing otherwise the probability index of the actual output is taken and then log was used to calculate its loss. AdamW optimizer is used to track the gradients and update it with the learning rate.

The Loss function used is Cross entropy Loss with Ignored padding index. Finallyafter Training the model for 2 Epoch the model accuracy is measured using BLEU metrics. BLEU (Bilingual Evaluation Understudy) is a metric used for evaluating the quality of machine-translated text by comparing it to one or more reference translations. It is widely used in the field of natural language processing and machine translation. BLEU assesses the likeness between the output

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204110 |

generated by a machine and reference translations by considering n-grams, which are consecutive sequences of n words. It is calculated based on the precision of matching n-grams between the candidate translation and the reference translation.

### 3.3.1 Formulas used

**Scaled Dot-Product Attention:**

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

This formula calculates the attention scores between the query ($Q$) and key ($K$) matrices, scales them by the square root of the dimensionality of the keys ($d_k$), applies a softmax function to obtain attention weights, and then multiplies these weights with the value ($V$) matrix to produce the attention output.

**Positional Encoding:**

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

This formula generates positional encodings to provide information about the position of each token in the sequence, where $pos$ is the position and $i$ is the dimension and $d_{model}$ is the dimensionality of the model.

## IV. RESULT

For the evaluation of the Transformer model's predictive capabilities, a specific input text, "agara mudhala eluthellaam aadhi bagavan mudhatrea ulagu," was introduced for inference. Notably, this particular sample was not included in either the training or validation datasets utilized during the model's training phase. Despite its absence from the training data, the model successfully generated and presented a predicted output text.

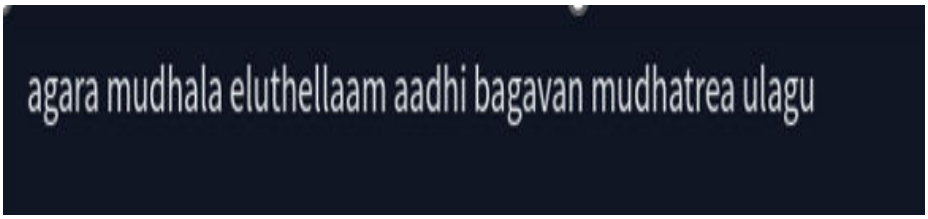


**Figure 2: Input text to the model (Transliterated English)**

To gauge the accuracy of this prediction, the BLEU score—a standard metric for evaluating machine-generated text—was employed. By comparing the model's predicted output against the actual text it was intended to generate, quantitative insights into the model's performance were obtained. In this instance, the intended text, "அகர முதல எழுத்தெல்லாம் ஆதி பகவன் முதற்றே உலகு," differed slightly from the model's output, "அகர முதல எலுத்தெல்லாம் ஆதி பாகவன் முதற்றீ உலகு."

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204110 |

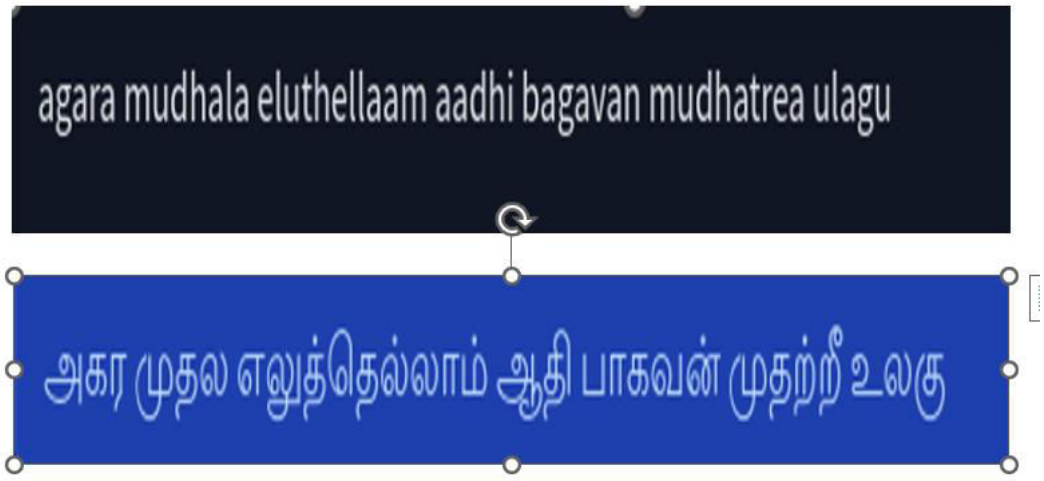


**Figure 3: Output text generated by the model (Tamil)**

This comparison facilitated a quantitative assessment of the model's predictive accuracy, providing valuable insights into its performance in generating the target language text. By evaluating the model against the ground truth that is the actual text it was expected to produce, the reliability and effectiveness of the model were verified. Moreover, this evaluation process serves as a guiding framework for further improvements in training methodologies and parameter fine-tuning, ultimately enhancing the model's predictive capabilities across various language tasks and domains.

## V. CONCLUSION

In conclusion, this research highlights the pressing need for an effective translation system to bridge the gap between the widespread usage of Transliterated English and the preservation of the regional language in the digital sphere. The study underscores the limitations of existing translation systems in addressing the unique challenges posed by the hybrid nature of Transliterated English. By leveraging the advanced capabilities of the Transformer architecture, this research offers a promising solution to this complex linguistic conundrum. Through the systematic collection of data from various online sources containing both Transliterated English and regional language (Tamil), the proposed neural translation system demonstrates the potential to facilitate seamless and accurate conversions of Transliterated English content into native Language (Tamil).

Furthermore, the successful implementation of this solution is expected to pave the way for the enhanced digitalization of the regional languages. This not only enables a wider audience to engage with and understand regional content but also serves as a crucial step in reinforcing the language's significance and relevance in the ever-evolving digital landscape. By promoting the accessibility and comprehensibility of regional content online, this research strives to contribute to the preservation and promotion of the regional language in the digital age, fostering a deeper connection between regional speakers and their cultural heritage.

## REFERENCES

1. Chun XU , Fang Chen c , Xiayang Shi b ,Zhengqing Yu b “Adding visual attention into encoder-decoder model for multi-modal machine translation”, Volume 11, Issue 2, June 2023, doi.org/10.1016/j.jer.2023.100077
2. B. Zhang, D. Xiong and J. Su, "Neural Machine Translation with Deep Attention," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 1, pp. 154-163, 1 Jan. 2020, doi: 10.1109/TPAMI.2018.2876404.
3. Angela Fan, Ahmed El-Kishky, Holger Schwenk, Shruti Bhosale, Zhiyi Ma, “Beyond English-Centric Multilingual Machine Translation” (Oct 2020), doi.org/10.48550/arXiv.2010.11125
4. Ashish Vaswani, Aidan N. Gomez, Illia Polosukhin, Jakob Uszkoreit, Llion Jones, , Łukasz Kaiser, Niki Parmar, Noam Shazeer, “ Attention is all you need”. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, doi.org/10.48550/arXiv.1706.03762
5. Alinejad, M., Sarkar, A., Siahbani, “A.: Prediction Improves Simultaneous Neural Machine Translation.” In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3022–3027. Association for Computational Linguistics, Brussels, Belgium, 2018, dx.doi.org/10.18653/v1/D18-1337

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204110 |

6. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," CoRR, vol. abs/1409.3215, 2014.
7. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. of ICLR, 2014.
8. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention based neural machine translation," in Proc. of EMNLP, September 2015, pp. 1412–1421.
9. Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Coverage-based neural machinetranslation," CoRR, vol. abs/1601.04811, 2016.
10. J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," Transactions of the Association for Computational Linguistics, vol. 4, pp. 371–383, 2016.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details