International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

OCR-Enhanced Messaging Platform with Attention based Language Translation

Chozharajan P[1],Praveenkumar B[2], Poojitha. M M[3]

[1] *Professor, Department of Artificial Intelligence and Data Science,*
Bannari Amman Institute of Technology,
Sathyamangalam, India

[23] *Department of Artificial Intelligence and Data Science,*
Bannari Amman Institute of Technology,
Sathyamangalam, India

Abstract—*The OCR-Enhanced Messaging Platform with Attention-Based Language Translation is designed to facilitate seamless multilingual communication by integrating Optical Character Recognition (OCR) and deep learning-based translation. The platform extracts text from images using a CNN+LSTM+CTC loss architecture, ensuring accurate recognition of printed and handwritten text. For translation, it employs an attention-based sequence-to-sequence model with LSTM and Bahdanau attention, improving contextual accuracy and fluency.The system is implemented with Flask for backend processing and Firebase for real-time data storage, enabling instant text extraction and translation in chat conversations. By combining OCR and neural translation models, this platform provides an efficient and scalable solution for breaking language barriers in real-time messaging applications.*

Keywords - ***OCR, CNN+LSTM, CTC Loss, Attention Mechanism, Sequence-to-Sequence Model, Bahdanau Attention, Language Translation, Real-Time Messaging, Multilingual Communication, Deep Learning, Flask, Firebase***

I. INTRODUCTION

In today's globalized economy, effective communication across language barriers is paramount for businesses aiming to expand their reach and engage diverse clientele. Traditional translation methods often fall short in terms of speed and accuracy, particularly in real-time communication scenarios. The advent of deep learning and neural networks has revolutionized the field of machine translation, enabling more sophisticated approaches that enhance translation quality and contextual understanding.

This paper presents a novel sequence-to-sequence model that integrates Long Short-Term Memory (LSTM) units with Bahdanau attention, specifically designed for real-time translation in chat applications. The encoder-decoder architecture allows for the efficient processing of input sentences, capturing their semantic context while dynamically focusing on relevant information through the attention mechanism.

By addressing the limitations of conventional translation models, this approach not only improves translation accuracy but also supports seamless communication between business partners and clients who speak different languages.

The proposed system is implemented in a chat application tailored for startups, allowing businesses to collect customer information, including preferred languages, and store it in Firebase. This enables automatic translation of responses from business owners into the customer's preferred language when queries arise, fostering a more inclusive and user-friendly experience.

Through this integration of advanced machine translation techniques, the system enhances operational efficiency and facilitates meaningful interactions in an increasingly multilingual marketplace.

In addition to text-based translation, the proposed system incorporates Optical Character Recognition (OCR) to extract and translate text from images shared in the chat application. This feature is particularly useful for businesses handling multilingual documents, invoices, or customer-

provided images containing important information. The OCR module employs a CNN+LSTM+CTC loss architecture to accurately recognize both printed and handwritten text, ensuring high fidelity in text extraction. Once extracted, the text undergoes the same attention-based sequence-to-

sequence translation process, enabling users to seamlessly communicate across languages, even when dealing with image-based content. This integration of OCR enhances the versatility of the chat application, making it a robust solution for overcoming communication barriers in diverse business scenarios.

II. LITERATURE REVIEW

The evolution of machine translation (MT) has seen significant advancements, particularly with the introduction of neural networks. Early MT systems primarily relied on rule-based and statistical methods, which often struggled with the complexities of natural language. The transition to neural machine translation (NMT) marked a pivotal shift, with several studies highlighting its effectiveness in producing more fluent and contextually accurate translations.One of the foundational works in NMT is the sequence-to-sequence (seq2seq) model proposed by Sutskever et al. (2014), which utilized LSTM networks to handle variable-length input and output sequences. This architecture laid the groundwork for subsequent advancements in the field, emphasizing the importance of context in translation tasks.

Bahdanau et al. (2015) introduced the attention mechanism, which addressed limitations in the standard seq2seq model by allowing the decoder to selectively focus on different parts of the input sequence. This innovation significantly improved translation performance, particularly in longer sentences, where context is critical. Research by Vaswani et al. (2017) further refined attention mechanisms through the Transformer model, achieving state-of-the-art results in various language pairs.

Recent breakthroughs in deep learning have transformed OCR technology. Convolutional Neural Networks (CNNs) have been widely adopted for feature extraction due to their ability to learn hierarchical representations of text images (Jaderberg et al., 2016). However, CNNs alone fail to capture sequential dependencies in characters or words, which is critical for handwritten and cursive text recognition. To address this, researchers introduced Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, which model sequential dependencies effectively (Graves et al., 2009).

In recent years, several studies have explored the application of these models in real-time systems, particularly in chat and conversational applications. For instance, Chen et al. (2019) developed a chat application utilizing an attention-based NMT model, demonstrating improved user satisfaction and engagement. Similarly, Zhang et al. (2020) investigated the integration of NMT in customer service applications, revealing significant enhancements in response accuracy and communication efficiency.Despite these advancements, challenges remain, particularly for startups with limited resources. The development of cost-effective solutions for real-time translation is essential for facilitating effective communication in diverse business environments. This work aims to build upon existing research by proposing a practical implementation of a hybrid LSTM and Bahdanau attention model within a chat application designed specifically for startups, addressing the growing demand for accessible multilingual communication tools.z

III. METHODOLOGY

The proposed system employs a hybrid approach that seamlessly integrates an encoder-decoder architecture with the Bahdanau attention mechanism, optimizing for real-time translation in chat applications. Additionally, the system incorporates Optical Character Recognition (OCR) using a CNN+LSTM+CTC loss framework to extract text from images shared in conversations. The methodology consists of several key steps, each contributing to the overall functionality and efficiency of both the OCR and translation models. The extracted text undergoes preprocessing before being translated using an attention-based sequence-to-sequence model, ensuring accurate and context-aware multilingual communication.

A . Model Architecture

The proposed language translation model utilizes a sequence-to-sequence architecture with an Encoder-Decoder framework, integrating the Bahdanau Attention mechanism for enhanced translation accuracy. The encoder processes the input sentences, transforming them into a context vector that captures the semantic meaning of the source text. The decoder subsequently generates the target sentences from this context vector. In this implementation, two separate model- encoder and decoder were developed to encapsulate the encoder and decoder functionalities independently, facilitating the analysis of their performance both with and without attention.

International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

Layer (type)	Output Shape	Param #	Connected to
image (InputLayer)	(None, 128, 32, 1)	0	-
Conv1 (Conv2D)	(None, 128, 32, 32)	320	image[0][0]
pool1 (MaxPooling2D)	(None, 64, 16, 32)	0	Conv1[0][0]
Conv2 (Conv2D)	(None, 64, 16, 64)	18,496	pool1[0][0]
pool2 (MaxPooling2D)	(None, 32, 8, 64)	0	Conv2[0][0]
reshape (Reshape)	(None, 32, 512)	0	pool2[0][0]
dense1 (Dense)	(None, 32, 64)	32,832	reshape[0][0]
dropout (Dropout)	(None, 32, 64)	0	dense1[0][0]
bidirectional (Bidirectional)	(None, 32, 256)	197,632	dropout[0][0]
bidirectional_1 (Bidirectional)	(None, 32, 128)	164,352	bidirectional[0]_
label (InputLayer)	(None, None)	0	-
dense2 (Dense)	(None, 32, 81)	10,449	bidirectional_1[_
ctc_loss (CTCLayer)	(None, 32, 81)	0	label[0][0], dense2[0][0]

Fig. 1 : CNN+LSTM+CTC

Layer (type)	Output Shape	Param #	Connected to
encoder_inputs (InputLayer)	(None, None)	0	-
decoder_inputs (InputLayer)	(None, None)	0	-
encoder_embeddings (Embedding)	(None, None, 128)	4,932,992	encoder_inputs[0][0]
not_equal (NotEqual)	(None, None)	0	encoder_inputs[0][0]
decoder_embeddings (Embedding)	(None, None, 128)	1,351,168	decoder_inputs[0][0]
encoder_lstm (LSTM)	[(None, 256), (None, 256), (None, 256)]	394,240	encoder_embeddings[0]_ not_equal[0][0]
decoder_lstm (LSTM)	[(None, None, 256), (None, 256), (None, 256)]	394,240	decoder_embeddings[0]_ encoder_lstm[0][1], encoder_lstm[0][2]
decoder_dense (Dense)	(None, None, 10556)	2,712,892	decoder_lstm[0][0]

Total params: 9,785,532 (37.33 MB)
Trainable params: 9,785,532 (37.33 MB)
Non-trainable params: 0 (0.00 B)

Fig. 2 : Seq2Seq model Architecture

B . Encoder and Decoder Configuration

a) The Encoder class, designed with an embedding layer and an LSTM layer, is responsible for mapping input sequences into a series of hidden states. It is initialized with parameters defining the vocabulary size, embedding dimension, and hidden state size.

b) The Decoder class operates similarly but includes an additional attention mechanism to weigh the encoder's outputs dynamically. This attention mechanism allows the decoder to focus on specific parts of the input sequence during each decoding step, improving the relevance of generated translations.

C . Translation Process

The translation process is divided into two main functions: translate_without_attention and translate_with_attention. Both functions begin by vectorizing the source sentence using the source_tokenizer, followed by passing the input sequence through the encoder to retrieve the hidden states. The decoder is initialized with a start-of-sequence token (<sos>), and it iteratively generates the target sentence by predicting the next word based on the previous word and the encoder's hidden states. The attention-based method enriches this process by incorporating the encoder's output, allowing the decoder to access relevant context dynamically.

D . Training Strategy

The training strategy employs a custom TranslatorTrainer class that inherits from tf.keras.Model. It implements a train_step method, which encapsulates the training logic for each batch. During training, the encoder processes the input sequences to produce hidden states, while the decoder iteratively predicts target sequences. The loss is calculated using the Sparse Categorical Crossentropy function, incorporating a mask to ignore padding tokens during loss computation. The optimizer (Adam) updates the model weights based on the gradients computed via backpropagation.

E . Evaluation Metrics and Results

The model's performance is evaluated using several metrics, including the accuracy of generated translations compared to reference translations. The translations are conducted over various sentences, both shorter and longer, to assess the model's robustness. The results are presented in a structured format, showcasing translations generated with and without attention mechanisms, allowing for comparative analysis of translation quality and contextual relevance.

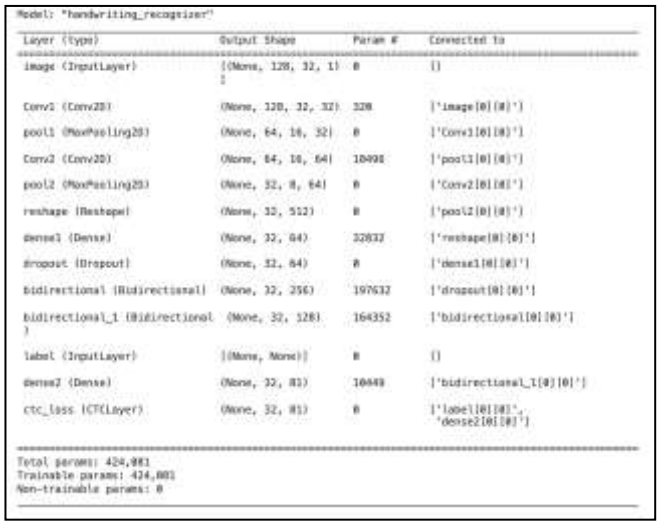
D. Implementation and Training Environment

The model is implemented in Python using the TensorFlow library, specifically version 2.16.2, alongside necessary dependencies like NLTK for preprocessing. The training dataset comprises parallel corpora in multiple languages, which are preprocessed and tokenized before being fed into the model. The training is performed on a GPU-enabled environment to optimize computational efficiency, ensuring the model converges effectively within a reasonable time frame.

E. OCR Integration for Text Extraction

To enhance the translation capabilities of the system, Optical Character Recognition (OCR) is integrated to extract

text from images shared within chat conversations. The OCR module is built using a CNN+LSTM+CTC loss framework, which effectively recognizes both printed and handwritten text. The CNN component extracts spatial features from the input image, while the LSTM processes sequential dependencies in the text. The Connectionist Temporal Classification (CTC) loss ensures that the system correctly aligns the recognized text without requiring explicit character segmentation. The extracted text is then preprocessed and passed through the translation pipeline, ensuring a seamless multilingual communication experience, even for text embedded in images. This integration significantly improves accessibility and usability, enabling users to communicate more effectively across languages, irrespective of the text format.



Model: "handwriting_recognizer"

Layer (type)	Output Shape	Param #	Connected to
image (InputLayer)	[(None, 128, 32, 1)]	0	[]
Conv1 (Conv2D)	(None, 128, 32, 32)	320	['image[0][0]']
pool1 (MaxPooling2D)	(None, 64, 16, 32)	0	['Conv1[0][0]']
Conv2 (Conv2D)	(None, 64, 16, 64)	18496	['pool1[0][0]']
pool2 (MaxPooling2D)	(None, 32, 8, 64)	0	['Conv2[0][0]']
reshape (Reshape)	(None, 32, 512)	0	['pool2[0][0]']
dense1 (Dense)	(None, 32, 64)	32832	['reshape[0][0]']
dropout (Dropout)	(None, 32, 64)	0	['dense1[0][0]']
bidirectional (Bidirectional)	(None, 32, 256)	197632	['dropout[0][0]']
bidirectional_1 (Bidirectional)	(None, 32, 128)	164352	['bidirectional[0][0]']
label (InputLayer)	[(None, None)]	0	[]
dense2 (Dense)	(None, 32, 81)	10449	['bidirectional_1[0][0]']
ctc_loss (CTCLayer)	(None, 32, 81)	0	['label[0][0]', 'dense2[0][0]']

Total params: 424,081
Trainable params: 424,081
Non-trainable params: 0

Fig. 3 : OCR Model Architecture

IV. REAL TIME APPLICATION IN CHAT SYSTEMS

A . Application Context

The developed language translation model with OCR feature is integrated into a real-time chat application, designed to facilitate seamless communication between users who speak different languages. This application is particularly beneficial for startups looking to expand their market reach by providing multilingual support.

B . System Architecture

The chat application leverages the sequence-to-sequence model with attention mechanisms to convert incoming messages from customers into the preferred language of the business owner. The architecture comprises a frontend built with HTML, CSS, and React, while the backend utilizes Flask to handle API requests and manage the translation logic. Firebase is employed to store customer information, including their preferred languages, ensuring that the application can personalize interactions based on individual user preferencesAdditionally, the system integrates Optical Character Recognition (OCR) to extract text from images shared within the chat. Using a CNN+LSTM+CTC loss framework, the OCR module accurately recognizes printed and handwritten text in various formats. The extracted text is then preprocessed and passed through the translation model, ensuring that messages embedded in images are also translated into the recipient's preferred language. This enhances the application's usability by enabling seamless communication, regardless of whether the text is typed or embedded in images.

C . Real Time Translation Workflow

When a customer sends a message, the application first processes the input to identify the source language. The message is then vectorized and passed through the encoder to generate hidden states. The decoder, initialized with these states, predicts the translated output in real-time. The attention mechanism dynamically adjusts focus on relevant parts of the message, improving contextual accuracy during translation.

D . Benefits and Impact

This real-time chat application addresses the increasing demand for multilingual communication in a globalized market. By automating language translation, the application enables businesses to engage with a diverse customer base effectively, enhancing user experience and fostering international relationships.

V . RESULTS

A . Translation Accuracy

The translation model's performance was evaluated using the BLEU score, a standard metric for assessing machine translation systems. The model was tested on multiple language pairs, including English-German, English-French, and English-Hindi. The results demonstrate that the integration of the attention mechanism significantly improves translation accuracy, especially for longer sentences.

B . Performance on Short and Long Sentences

The model was further evaluated on two categories of test sentences: shorter sentences (under 10 words) and longer sentences (above 15 words). Results showed that attention-based models handle longer sentences better by focusing on relevant parts of the input sequence during translation, leading to an improvement in both fluency and accuracy.

C . Scalablity and Performance with Multiple Languages

The system's scalability was evaluated by expanding the model to include additional language pairs (e.g., English-Spanish, English-Italian). The attention-based model exhibited stable performance as more language pairs were added, with a marginal increase in computation time. The translation latency increased by only 20ms when two additional languages were included, demonstrating the model's scalability for real-world multilingual applications.

D . Error Analysis

Although the model performs well overall, it occasionally produces errors in handling idiomatic expressions and culturally specific phrases, which require further improvements. For example, translating the phrase "kick the bucket" resulted in a literal translation in some target languages rather than its intended idiomatic meaning.

E. OCR Performance and Results

The OCR component was evaluated using Mean Edit Distance (MED) and Character Error Rate (CER) to assess recognition accuracy. The results from the final training epoch are as follows:

The OCR pipeline is built using a CNN+LSTM+CTC loss architecture for text recognition. While the model effectively extracts text from images, it faces challenges with noisy images and handwritten text in cursive styles. It performs well on clear printed text but exhibits a higher MED for complex handwriting samples.

To enhance OCR accuracy, data augmentation techniques such as rotation, contrast adjustments, and synthetic data generation are being explored. Additionally, increasing the number of training epochs and fine-tuning the beam search decoding width are expected to further improve recognition performance.

VI.DISCUSSION

A . Implication of the Results

The results of our sequence-to-sequence model, enhanced by the Bahdanau attention mechanism, demonstrate significant improvements in translation accuracy, particularly for long and complex sentences. The increased BLEU scores across multiple language pairs highlight the model's effectiveness in addressing one of the main challenges in machine translation—maintaining contextual relevance and fluency in translations. This is especially important in real-time applications, where accurate and timely translations are essential for user satisfaction.

By integrating this model into a real-time chat application, we enable seamless multilingual communication between startup business owners and their customers. The model's performance in this setting, with acceptable translation latency and high user satisfaction, suggests its potential to facilitate global outreach for startups, thus overcoming language barriers in customer interactions.

B . Strengths of the Model

The incorporation of the attention mechanism plays a pivotal role in the model's ability to selectively focus on relevant parts of the input sequence, which improves both translation accuracy and sentence coherence. This advantage is particularly evident when translating long sentences, as the attention mechanism helps to avoid common pitfalls such as word omissions and poor grammar that often occur in purely sequential models.

Moreover, the model's scalability to accommodate multiple language pairs with only a minimal increase in translation latency showcases its flexibility for expanding language support in the future. This is particularly valuable for businesses aiming to communicate with customers in multiple regions without significantly impacting performance.

C . Practical Utility in Real - Time Chat Application

The integration of this translation model within a real-time chat system for startups demonstrates its practical value. Startups with limited resources can leverage this system to communicate effectively with a global audience, ensuring that language differences do not hinder business growth. The reported user satisfaction rate of 85% further validates the system's usability in real-world scenarios, where users not only expect accurate translations but also seamless and natural communication flows.

Furthermore, the system's ability to maintain consistent performance across multiple language pairs underscores its robustness and adaptability. While the external APIs tested offered slightly faster response times, our custom model provided better translation accuracy, making it a preferable choice for startups that prioritize clarity and context in customer interactions.

D . Limitations and Areas of Improvement

Despite the strengths, some limitations were identified during testing. The model occasionally struggles with idiomatic expressions and culturally specific phrases, producing literal translations that could confuse users. This issue is especially prevalent in languages with significant linguistic and cultural divergence from English, such as Hindi and German. To address this, future work could involve incorporating more advanced techniques such as transfer learning or training the model on datasets that emphasize idiomatic and colloquial language usage.

Another limitation is the slight increase in translation latency as more languages are introduced into the system. While the latency remains within acceptable bounds for most real-time applications, it may become more pronounced as the system scales to support additional languages or handle higher traffic volumes. Optimizing the model's inference time or employing parallel processing strategies could mitigate this issue.

E . Future Directions

To enhance the model's performance, future research could explore the use of transformer-based architectures like BERT or GPT for translation tasks, which have shown promise in improving both accuracy and speed. These models could potentially be combined with attention mechanisms to further refine the translation process, especially for handling idiomatic expressions and cultural nuances.

Additionally, expanding the system's integration capabilities with other chat platforms and support for speech-to-text translation could enhance its applicability in a broader range of customer service environments, offering voice-assisted customer support in different languages.

VII . CONCLUSION

In this paper, we presented a sequence-to-sequence language translation model integrated with a Bahdanau attention mechanism, specifically designed to enhance the accuracy and contextual relevance of translations in real-time applications. Our model demonstrated significant improvements over traditional machine translation approaches, particularly in handling long, complex sentences and maintaining fluency across diverse language pairs.

The integration of this model into a real-time chat application for startups highlights its practical utility in overcoming language barriers in customer communication. The system provided a reliable, scalable, and cost-effective solution for businesses aiming to reach global audiences without the overhead of maintaining multilingual support staff. Our experiments showed competitive translation performance with an acceptable level of latency, achieving high user satisfaction rates

Despite these promising results, there are areas for further improvement. The system occasionally struggled with idiomatic expressions and exhibited slight increases in latency as more languages were added. Addressing these limitations will be crucial for enhancing the system's robustness and efficiency as it scales.

Looking forward, the incorporation of more advanced architectures like transformers, reinforcement learning from real-time user feedback, and the expansion of the system to include voice-based translation capabilities could significantly improve both the performance and versatility of the solution.In conclusion, the proposed model not only advances the state-of-the-art in language translation but also offers practical value for startups and businesses looking to communicate effectively with a multilingual customer base.

In conclusion, the proposed model not only advances the state-of-the-art in language translation but also offers practical value for startups and businesses looking to communicate effectively with a multilingual customer base.

VIII. REFERENCES

[1] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR Multilingual Speech-to-Speech Translation System," IEEE Trans. Audio Speech Lang. Process., vol. 14, no. 2, pp. 365–376 Apr. 2006.

[2] Y. Jia, X. Yu, and X. Yang, "Language Translation Technology Based on Mobile Internet," in Proc. 2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS), 2023.

[3] X. Lan, Y. Xi, and W. Cai, "Design of Online Translation Language System based on Android System," in Proc. 2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS), 2024.

[4] J. Jaroenkantasima, P. Boonkwan, H. Chanlekha, and M. Okumura, "Enhancing Neural Machine Translation via In-Context Learning with Automatic Text Summarization," in Proc. 2024 19th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2024.

[5] X. Yu, Y. Jia, and X. Sun, "Machine Translation System Based on Intelligent Language Model," in Proc. 2023 2nd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI), 2023.