OCR and Translation from Sinhala to Tamil for Printed Documents

Odasara Karunachandra
Department of Computer Engineering
University of Peradeniya.
Sri Lanka.
odasaraik@gmail.com

Shazna Isthikar
Department of Computer Engineering
University of Peradeniya.
Sri Lanka.
shaznaisthikar@gmail.com

Mishel Rossmaree
Department of Computer Engineering
University of Peradeniya.
Sri Lanka.
mishelrossmaree@gmail.com

Roshan G. Ragel
Department of Computer Engineering
University of Peradeniya.
Sri Lanka.
roshanr@pdn.ac.lk

Asitha U. Bandaranayake
Department of Computer Engineering
University of Peradeniya.
Sri Lanka.
asithab@eng.pdn.ac.lk

Abstract
Sinhala and Tamil are the official national languages of Sri Lanka. While having a high literacy rate overall, a majority of the people are able to read only one language. This has created a significant language barrier, especially for the minority Tamil community. The need for an open platform that will allow Sinhala to Tamil translations, and vice-versa, is felt in every sphere of life. The implemented system takes Sinhala documents in different file formats as input and outputs the Tamil-translated text to the user. Optical Character Recognition (OCR) is used to identify the characters in the images, with Tesseract by Google used as the underlying technology. For the translation, the Google Translation API is used. Preprocessing techniques like binarization and canny are used to improve the accuracy, and text cleaning is used as the post-processing technique. Eventually, an accuracy rate of 86.34% for Optical Character Recognition (OCR) and an accuracy rate of 72.52% for translation are obtained, and a free and user-friendly web portal is launched as a tool to overcome the language gap between Sinhala and Tamil.

***Keywords* - Optical Character Recognition, OCR, Sinhala-Tamil, Translation**

I. Introduction

In a globally interconnected society, overcoming language barriers is crucial for successful communication and knowledge sharing. Language variance is a significant obstacle to absorbing and understanding information from sources. Technologies such as optical character recognition (OCR) and translation have become crucial tools in overcoming this language gap by making it possible to extract and translate text. The main goal of this research is to construct an OCR and translation system specifically designed for the Sinhala to Tamil language pair to overcome the language barrier and promote interlanguage communication.

Because of their unique linguistic characteristics and scripts, the Sinhala and Tamil languages are much harder to OCR and translate. The native tongue of Sri Lanka is Sinhala[9]. Tamil, another official language, is extensively spoken in the country's northern and eastern regions and has its own writing system. Numerous literary works, historical documents, and cultural artefacts are available in both languages. However, access to important information in documents and efficient communication has been limited due to insufficient OCR and translation technology for these languages.

Overcoming the language barrier is a challenging undertaking since it involves accurate character identification and understanding of the semantic and syntactic structures of the characters. Through developing an OCR and translation system for Tamil from Sinhala, this research aims to break down the language barrier and promote cross-language information retrieval.

II. Literature Review

A. OCR

OCR technology is a crucial component in digitising and automating printed documents. This section gives a thorough description and overview of OCR technology. OCR technology translates information from printed or handwritten documents into machine-editable electronic Forms [1]. It involves extracting characters, words, and textual information from scanned or photographed images of documents. OCR technology makes it easy to read written text and automates tasks like text extraction, data entry, and getting information from printed materials.

OCR systems typically comprise several fundamental components that work together to achieve accurate text recognition. These components include:

- Image Preprocessing

- Character Segmentation
- Feature Extraction
- Classification and Recognition

Tesseract OCR, which is an open-source OCR engine, will be a key tool in our research to pull text from photographs that are printed or written. Tesseract OCR can read more than 100 languages and recognize both in-line and character patterns. It's important to note that its latest version (4.0) includes LSTM Neural Network, which makes it better able to find and detect inputs of different sizes. We hope to improve our efforts to provide accurate and useful OCR results for printed Sinhala documents and develop language processing technology in this area by using Tesseract OCR.

In [2], the method uses image pre-processing techniques like grayscale conversion and binarization with local thresholding to improve and prepare scanned pictures. Based on tests with different letter sizes, the method is about 97% accurate. The success of using LSTM-based training on more than 20 legacy fonts to reduce character-level and word-level error rates, hence improving character recognition accuracy, is highlighted in [3]. The research significantly reduced Tamil and Sinhala character-level and word-level errors.

B. Techniques and algorithms used in OCR Systems

The successful identification and transformation of Sinhala characters in OCR systems depend on the efficient use of various methods and algorithms. This section presents an in-depth study of the fundamental methods and algorithms used in OCR systems, specifically focusing on research papers related to Sinhala OCR.

Various image preprocessing methods are used to improve the overall quality of input images prior to the extraction of text. Preprocessing methods are often used in image analysis, including noise reduction, image binarization, skew correction, and normalization [4]. The earliest steps include the execution of cleaning and de-skewing processes, which aim to eliminate artefacts and achieve horizontal alignment of the text [1].

The process of character segmentation involves the identification and subsequent separation of individual characters from the given input image. The addition of this phase is crucial for accurate character recognition and subsequent text processing. The images are resized to a predefined dimension and converted into binary format, facilitating further processing. The concept of the ”Universe of Discourse” (UOD) refers to a representation, often in the form of a matrix, that covers the skeletal structure of a character. This representation is designed to extract the area containing the character's picture effectively [1].

Converting the input data into several sets of features is referred to as feature extraction [5]. Feature extraction techniques capture characters' distinctive characteristics, including their shape, texture, and spatial arrangement. These features represent characters as numbers or symbols, facilitating their recognition and classification. Sinhala OCR research has explored different feature extraction methods, such as identifying line type segments. This method identifies the line-type segments, and based on this information, a feature vector is formed [5]. Another approach used in Sinhala OCR is utilising the Tesseract 4.0 OCR engine [6], which provides additional feature extraction capabilities using deep learning.

In the classification and recognition stage, extracted features are used to identify and interpret characters. Machine learning algorithms, such as Artificial Neural Networks (ANN) [1] and Hidden Markov Models (HMM) [7], have been widely employed in Sinhala OCR research for accurate character recognition.

The effective utilization of techniques and algorithms is essential in developing accurate and efficient Sinhala OCR systems. Character segmentation methods, such as connected component analysis, enable accurate separation of characters. Feature extraction techniques, including pixel-based features, stroke-based features, and structural features, capture the unique visual properties of Sinhala characters. Classification algorithms, such as SVM and ANN, are applied to recognize and interpret the extracted features accurately. By using these techniques and algorithms, Sinhala OCR systems can achieve high accuracy and efficiently convert Sinhala printed documents into machine-readable text.

C. Challenges of Sinhala OCR

The Sinhala writing system is characterized by a set of symbols representing consonants and their associated vowel sounds, which are used for the written representation of the Sinhala and Pali languages. The development of an OCR system is challenging due to the complex structure of the Sinhala script [6].

The challenges in Sinhala OCR arise from the complex structure of Sinhala scripts, which have developed from the ancient Brahmi scripts. Character recognition and differentiation get tough with around 18 vowels, 41 consonants, and 17 modifier symbols. Many modified characters have similar forms, making OCR devices difficult to use. The lack of research and resources dedicated to Sinhala character recognition complicates the development of good OCR systems. Traditional OCR methods fail to manage the complexities of Sinhala characters, making reliable identification difficult. Most research groups use computer science and neural networks to reduce the complexity of the Sinhala script [8].

To address these issues, a novel strategy integrating character geometric features (CGF) and artificial neural networks (ANN) has been presented. When compared to traditional approaches, this strategy achieves higher character recognition rates. Addressing these issues would need novel methodologies and more research in Sinhala OCR [1].

Overall, the challenges in Sinhala OCR arise from the complex script structure, the similarity of character shapes, the limited research and resources dedicated to the recognition of modified characters, and the limitations of current OCR technologies. Developing effective OCR systems for Sinhala requires innovative approaches and further research to improve recognition accuracy and broaden the scope of character recognition.

D. Improvements in OCR accuracy for Sinhala

Improving the accuracy of OCR systems for Sinhala and Tamil scripts is an important research topic. This section explores the advancements made in improving OCR accuracy, specifically for Sinhala scripts, drawing insights from relevant research papers.

Researchers have focused on developing language-specific preprocessing techniques to improve OCR accuracy for Sinhala scripts. For example, [1] proposed a method that leverages the unique characteristics of Sinhala, such as using character geometry features, to preprocess input images effectively. By incorporating language-specific preprocessing techniques, OCR accuracy can be significantly enhanced.

Deep learning techniques have shown promise in improving OCR accuracy. Researchers have explored the use of convolutional neural networks (CNNs) for character recognition. For instance, [4] proposed a deep learning-based OCR system for Sinhala, achieving an overall accuracy of 85.37% improvement in accuracy compared to traditional methods. Adopting deep learning approaches has significantly improved OCR accuracy for Sinhala and Tamil scripts.

Researchers have explored combining multiple OCR techniques to further enhance OCR accuracy. This involves integrating different algorithms, such as character segmentation methods, feature extraction techniques, and classification algorithms, to create a robust OCR system. For instance, Liyanage [4] proposed a hybrid approach that explores the application of contour-based segmentation to segment Sinhala characters and CNN to recognize the segmented characters for improving Sinhala scripts. By leveraging the strengths of multiple OCR techniques, accuracy improvements can be achieved.

Advancements in OCR accuracy for Sinhala and Tamil scripts have been achieved through language-specific preprocessing techniques, deep learning approaches, and combining multiple OCR techniques. These improvements have paved the way for more accurate and reliable OCR systems, contributing to enhanced document digitization, data extraction, and text analysis for Sinhala and Tamil languages.

E. Translating Low-Resourced Languages

In Neural Machine Translation (NMT), the availability of large parallel corpora plays a crucial role in achieving successful results. However, many languages, especially minority languages, lack such resources, posing a challenge for translation tasks. Over the years, several techniques have been proposed to address this issue. One common method is to add monolingual corpora by having a translation trained in the opposite way to translate monolingual data. This creates parallel words that aren't there and makes the corpus bigger [9]. The idea behind this is that getting large parallel corpora for two languages is hard, but it is easy to get large single corpora for each language individually.

Both source-side monolingual corpora [6] and target-side monolingual corpora [9] have been back-translated using this technique. This technique has considerably improved translation quality for low-resource languages by exploiting both parallel and monolingual corpora.In contrast to models trained on real parallel data, empirical research has found that models trained on synthetic data tend to ”forget” the right translation semantics [10]. As a result, a constraint must be placed on the monolingual data utilised to keep the translation accuracy.

The back-translation method's popularity is also a result of its ability to function with current network designs. As a result, attempts have been undertaken to improve the back-translator's quality, which by nature performs as a subpar MT system. A filtering method that prioritises the best back-translated synthetic sentences, improving translation quality and BLEU scores have been proposed by Imankulova et al. [5].

The back-translation technique has been incorporated, and by improving the back-translator's quality, researchers have made considerable strides in tackling the translation issues that low-resource languages face. These methods have improved translation capabilities by enabling the efficient use of both parallel and monolingual data.

F. Sinhala - Tamil Translation

Recent studies [11] in the Sinhala to Tamil translation field have made notable progress in resolving the complexity associated with languages that possess a high degree of morphological richness. Researchers have used advanced methodologies, such as statistical machine translation and unsupervised morphological modification to enhance the accuracy and quality of the translation. A strategy worth mentioning is using Morfessor [12], an unsupervised method for morphological change that was first introduced to analyze Sinhala morphology. The integration of Morfessor into the translation procedure has significantly improved the performance of Sinhala to Tamil translation, as shown by researchers.

G. Challenges of translation from Sinhala to Tamil

Translation between two languages is not an easy task. Since a translator needs to be fluent in both languages. Differences between cultural systems hinder translators as they translate texts such as idioms, proverbs, collocations, etc. Many people studying translation have discussed how cultural differences affect translation. Newmark [13] states that there are many linguistic problems that a translator has to deal with when translating, such as using the wrong word because the translator doesn't know how to write appropriately, using the wrong translation tools, translating literally, or not having enough background knowledge.

H. Evaluations of accuracy, performance and limitations

To effectively translate Sinhala to Tamil, one must possess an in-depth knowledge of the grammatical and syntactic differentiations between these two languages. In addition to the task of word translation, this procedure involves the identification of subtleties and cultural contexts embedded within the text. The translation output significantly affects the usability and efficiency of these systems in several fields, including content localization, cross-lingual information retrieval, and machine-assisted translation.

In [14], the focus is on a Two-dimensional Fourier Transform and Artificial Neural Networks-based system for recognizing characters in the Sinhala language that remain constant when rotated. Sinhala letters in different styles and sizes can be read with an accuracy of over 85.17%. This system uses a segmentation method based on the histogram. This method has an accuracy of over 70.14% for difficult Sinhala letters. Based on tests with different font sizes using picture pre-processing methods like grayscale conversion and binarization, the system in [2] was about 96.83% accurate. The suggested system, detailed in [5], uses image-capturing techniques, OCR with Tesseract, and the Google Translator API to translate text. It has methods for preprocessing mechanisms that improve the accuracy of OCR. The evaluation, which was done on a set of 123 sentences that people use every day, had a 96.75% accuracy rate and 91.06% of Tamil sentences and 88.62% of Sinhala sentences were successfully translated.

I. Summary

In conclusion, this literature review has highlighted the existing research in OCR and Translation. Despite the significant amount of study done in these fields separately, there is still a clear need for both research on OCR and Tamil to Sinhala Translation. The National Languages Processing Center (NLPC) at the University of Moratuwa has made strides in this domain by developing the SiTa Computer-Assisted Translation System for official document translation in Sinhala, Tamil, and English. However, the system, SiTa possesses challenges for some users, as it has limited accessibility, accepting only digital documents in Unicode .docx format. Our research aims to address this gap by developing a system that integrates OCR functionality to facilitate the input of scanned and PDF documents for translation from Sinhala to Tamil. Unlike the NLPC's SiTa system, our solution will be freely accessible to all users without limitations on the type of documents it can process.

Furthermore, Thiruthanigesan and Ragel [5] have done similar research. Still, their approach requires users to buy a specific device for OCR and Translation from Sinhala to Tamil, making it less accessible to individuals with financial constraints. In our endeavour, we aspire to create a user-friendly portal that will integrate both OCR and Translation services, making them readily available to everyone, regardless of their financial situation or the printed documents they wish to translate. We seek to develop language processing technology by filling this research gap and facilitating cross-lingual communication between Sinhala and Tamil speakers.

III. Methodology

The main objective of this research is to develop a web application system that can convert printed Sinhala documents to Tamil using OCR for various file formats, such as images, documents, PDFs, and texts. This system is open to the public and free of charge. The system consists of 5 main components: Web portal(Input and Output module), preprocessing module, Tesseract OCR, postprocessing module, and Google translate module. Any user may upload a file (image/pdf/text/document) to the system using the web portal. The request is processed using the Restful API. Text and document file types feature editable text, instantly translated using Google Translate and displayed on the web portal. PDFs are first converted into images and then preprocessed using the module to enhance image quality. The main preprocessing techniques used are Binarization and Canny.

- **Binarization:** This involves converting an image into a binary image that is black and white based on a threshold value.
- **Canny:** An edge detection operator used for text recognition.

The Tesseract OCR extracts the text from the preprocessed images. For text cleaning, we utilized a postprocessing module that eliminates excess spaces, special characters, and other unneeded characters from the text. The Sinhala text is then translated into Tamil using Google Translate, a free Google API that provides highly accurate translations. The web portal displays the final output of the Tamil translation and the intermediate Sinhala OCR.

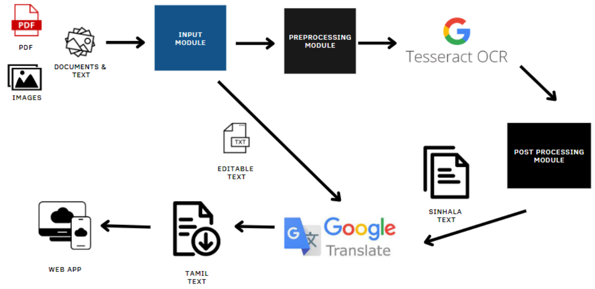

Fig. 1. The process of OCR and Translation from Sinhala to Tamil for Printed Documents

IV. Experiments

For this research, an experiment-based method of evaluation was used. The datasets used in the experiment mainly consisted of official documents like gazettes and circulars. The official government website included the original Sinhala documents and their Tamil versions. The original text was needed to compare with the system's OCR output and check the translation accuracy of the Tamil version of the documents. Informal language datasets like school textbooks and novels are incorporated into the dataset to address potential biases and represent a variety of linguistic patterns.

Text and document formats (.txt,.md,.log,.doc,.docx), pictures (.tif,.tiff,.jpg,.jpeg,.png), and PDFs (scanned and searchable, single or multi-page formats) are all included in the dataset. Hence, the experiment's input consisted of all the above-mentioned types.

The research hypothesis guiding our experiment is that by applying pre-processing and post-processing methods, OCR

Accuracy will improve, and as a result, the translation accuracy might also improve. In experimenting with calculating the Accuracy of OCR, there were three main phases.

1) **System using only Tesseract OCR:**
   In this initial phase, the system employs Tesseract OCR. This system is the base for comparing OCR accuracies.
2) **System with Tesseract OCR and Preprocessing:**
   To improve the quality of input data prior to OCR processing, Tesseract OCR is integrated with preprocessing techniques in the second phase.
3) **System with Tesseract OCR and Preprocessing and Post Processing:**
   Tesseract OCR is combined with preprocessing and postprocessing methods in the third and final stage.

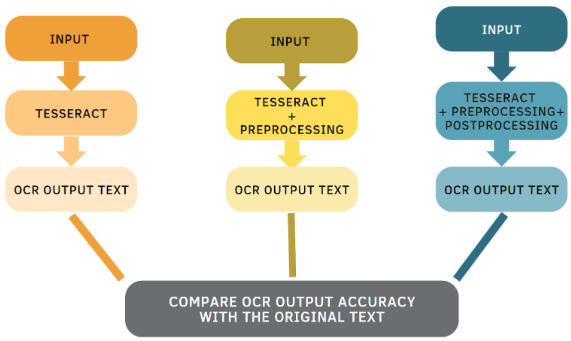

Fig. 2. Evaluation approach used to measure the OCR accuracy

The output generated from these experimental phases is systematically compared against the original text to assess accuracy. The following accuracy metrics will be used to quantify and measure accuracy. These metrics will offer an in-depth evaluation of how effectively the approaches have improved OCR's accuracy across a range of document formats and linguistic types.

$$\text{OCR ACCURACY} = \frac{\text{Number of words identified correctly}}{\text{Total number of words}} * 100\%$$

The OCR process's accuracy was calculated by implementing a Python script. The purpose of this script is to systematically compute accuracy metrics by comparing the output of each experimental phase with the original text. A manual verification procedure was carried out in addition to the automated assessment to guarantee the accuracy evaluation's reliability.

Using a thorough and methodical approach, the translation accuracy evaluation was conducted concurrently with the three stages of the OCR accuracy evaluation. In the system, we used Google Translation API to do the translation process. The main goal of the evaluation was to understand how improvements to OCR technology affect translation accuracy.

The experiment was conducted in three stages, which mirrored the format of the assessment of OCR accuracy:

1) **System using only Tesseract OCR:**
   OCR was used in the first stage to extract text, and then the Google Translation API was used for translation. The beginning point for translation accuracy was represented during this phase.
2) **System with Tesseract OCR and Preprocessing:**
   Before using the Google Translation API, the second phase combined preprocessing methods with Tesseract OCR. Improving the input data was the goal of this step in order to increase translation accuracy.
3) **System with Tesseract OCR and Preprocessing and Post Processing:**
   The last stage combined postprocessing techniques with Tesseract OCR to improve the translation process further.

Every document in our dataset had its original translated text available, so the output from each of these experimental phases produced by passing the OCR output through the Google Translation API was carefully compared to it.

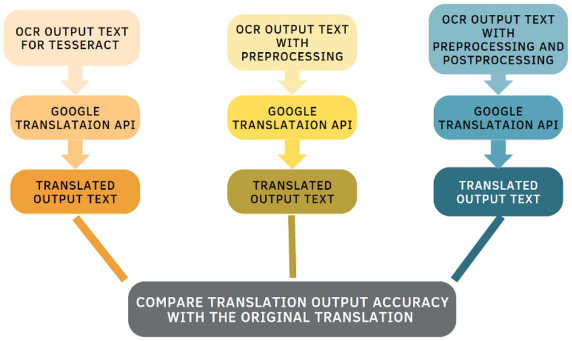

Fig. 3. Evaluation approach used to measure the Translation accuracy

Specific accuracy metrics were used to quantify and measure accuracy. It provides a side-by-side evaluation of the translation process's effectiveness across diverse document formats and linguistic styles.

$$\begin{array}{c}\text{TRANSLATION} \\ \text{ACCURACY}\end{array} = \frac{\text{Number of sentences translated correctly}}{\text{Total number of sentences}} * 100\%$$

Initially, the accuracy of the translation was evaluated using an online tool. However, this tool's limitations, like false negatives and a lack of synonym recognition, prevented it from producing reliable results. Afterwards, to guarantee accuracy in the assessment, we consulted some resource persons. This human-in-the-loop method addressed details that automated tools might miss, adding an extra layer of accuracy verification. Working with a domain expert enabled a more accurate and delicate evaluation of translation accuracy at every experiment stage.

V. Results and Discussion

The foundational framework used in the first phase of system development was Tesseract-based. However, it soon became clear that the Tesseract-based system fell short of the requirements. This caused the system to undergo iterative improvements through later approaches. To improve the input data before putting it through Tesseract OCR, the second strategy combined preprocessing techniques. Using postprocessing techniques, the third approach evolved on the changes presented in the second approach.

The main goal of these iterative methods was to gradually improve the system's accuracy, surpassing the constraints of the first Tessaract-based implementation. It is important to note that the system is evolving with improvements in OCR and translation.

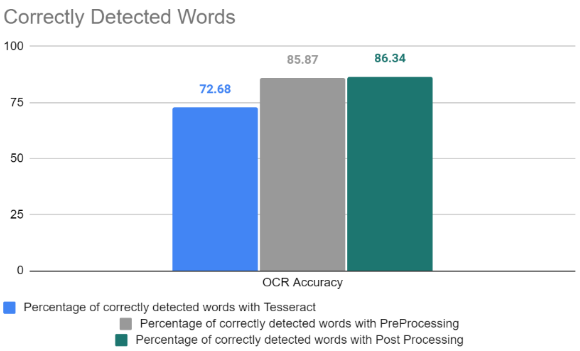

Fig. 4. Correctly detected words

The figure 04 illustrates the percentage of correctly detected sentences based on Equation 1. In sentences given through the Tesseract, nearly 72.68% of the sentences were converted without using any pre-processing mechanisms. However, it improved by using suggested pre-processing techniques for up to 85.87%. After adding post-processing techniques, OCR Accuracy went up to 86.34%. They correctly detected 86.34% of the text to take to the translation process.

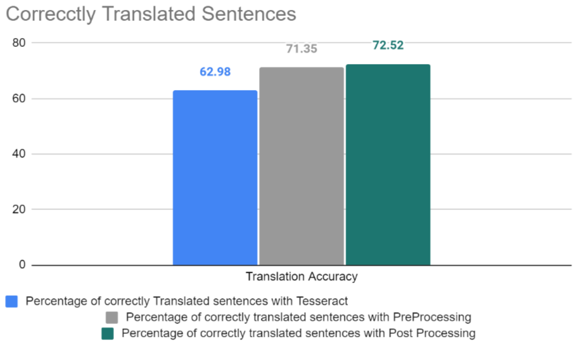

Fig. 5. Correctly translated sentences

Figure 05 illustrates the percentage of correctly translated sentences based on Equation 2 in which sentences are given through the Google API nearly 63% of the Sinhala sentences are translated properly without using any pre-processing mechanisms after using Canny and Binarization. The translation rate improved up to 71.35% of the Sinhala sentences translated proper manner. After removing unwanted characters in post-processing the translation accuracy too improves to 72.52%. The reason for the wrongly translated text of 17% of the Sinhala performance is the unavailability of the appropriate source words and symbols in the API and most of the grammatical mistakes in the APIs. The above result can be improved with the improvement of the APIs.

VI. Conclusion

At the end of the research, we were able to develop a tool that is accessible to a diverse user base, which enables the translation of Sinhala documents in different formats into editable Tamil texts, thereby overcoming linguistic barriers. This platform can enhance language understanding and promote information sharing between the two main cultural groups in Sri Lanka. The system's optical character recognition (OCR) accuracy rate was 86.34%, and its translation accuracy rate was 72.52%. These were observed after utilizing pre-processing and post-processing techniques. This tool may lead to further growth in the future, such as bringing out more reliability, accuracy, and modification of the tool to support other languages. As well as the research can be extended to overcome further related issues. A significant impact can be made in society through this as it allows people to get information and knowledge in their native language. It will interconnect the world and keep individuals informed and allied regardless of their linguistic aspects. In conclusion, this tool is a remarkable achievement as there was no platform available to the public in the case of translating Sinhala documents into Tamil.

References

[1] Premachandra C., Kimura T., Kawanaka H., Premachandra W., Waruna H., Premachandra H., ”Artificial Neural Network Based Sinhala Character Recognition.”, 2016 5th International Conference on Artificial Intelligence, Modelling and Simulation (pp.225-230) IEEE.

[2] Manisha U.K.D.N.a and Liyanage S.R., ”Sinhala Character Recognition using Tesseract OCR”, the 3rd International Conference on Advances in Computing and Technology, July 2018

[3] Charangan Vasantharajan, Laksika Tharmalingam, and Uthayasanker Thayasivam, ”Adapting the Tesseract Open-Source OCR Engine for Tamil and Sinhala Legacy Fonts and Creating a Parallel Corpus for Tamil-Sinhala-English”, International Conference on Asian Language Processing (IALP),2022

[4] Liyanage K.L.N.D “Improving Sinhala OCR using Deep Learning” 2018

[5] K. Thiruthanigesan, Roshan Ragel, ”Optical Character Translation Using Spectacles (OCTS),” 2019 IEEE 14th International Conference on Industrial and Information Systems (ICIIS), December 2019

[6] Anuradha, I., Liyanage, C., Wijayawardhana, H.,& Weerasinghe, R. (2020). Deep learning-based Sinhala Optical Character Recognition (OCR). In 2020 3rd International Conference on Advances in Computing and Technology (ICACT) IEEE.

[7] Hewavitharana, S., Fernando, H.C., Kodikara, N.D.: Off-line sinhala handwriting recognition using hidden Markov models. In: Indian Conference on ComputerVision, Graphics & Image Processing (ICVGIP), pp. 266–269 (2002)

[8] Rimas, Mohamad, Rohana Priyantha Thilakumara, and PriyarangaKoswatta. ”Optical character recognition for Sinhala language.” 2013 IEEE Global Humanitarian Technology J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[9] Wijayarathna, D., Dissanayake, D.,& Dahanayake, A. (2018). A robust approach for Sinhala character segmentation. 19th International Conference on Advances in ICT for Emerging Regions (ICTer), 36-41.

[10] Silva, S., Perera, H.,& Thilakarathne, P. J. (2019). Sinhala character recognition using Support Vector Machines. International Journal of Computer Science and Software Engineering, 8(10), 226-231.

[11] Pushpananda, R., Weerasinghe, R., Niranjan, M.: Statistical machine translation from and into morphologically rich and low-resourced languages (04 2015)

[12] Welgama, V., Weerasinghe, R., Niranjan, M.: Evaluating a machine learning approach to Sinhala morphological analysis (12 2013)

[13] Newmark. P (1988), a textbook of translation, New York.

[14] Gunaratne R., Jayaweera, S., Wijayasinghe N.,& Perera S., ”Enhancing Sinhala to Tamil Translation Using Tesseract OCR,” International Conference on Natural Language Processing, 2018