Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 5): pp. 7214-7219
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.05.813

A Translator for Indian sign boards to English using Tesseract and seq2seq model

Sharada K.A, Associate Professor, HKBKCE, Dept. of CSE, VTU, INDIA, sharadaa1234@gmail.com
Suma.T, Associate Professor, SVCE, Dept of CSE, INDIA, tsumamurthy.cs@gmail.com
Deepak N.R, Professor, HKBKCE, Dept. of CSE, VTU, INDIA, deepaknrgowda@gmail.com
Abdul Majid, Professor, Dr SMCE, Dept. of CSE, VTU, INDIA, dr.majid.wahab@gmail.com

Abstract- Language translator for Indian language to English have been developed and it have proven to a challenging domain due to large combination of character in india scripts such as Tamil, Kannada and Hindi. In this paper we propose a system which captures Indian printed character and translates it into English, we have discussed the various method and machine learning model that was used to build this system with a accuracy of 87%.In addition the paper also includes the project .A Generic sensor Frame for the webOSplatform.A centralized daemon to control and access all the sensors that are connected to the webOS device.

Keywords: OCR, Seq2Seq, Tesseract and Sensor Framework

I. INTRODUCTION

In India, which is a diverse country, having a total of 28 states and 8 union territories and about 122 languages spoken across the country among which there are 22 scheduled languages. In addition, the Government of India has awarded the distinction of classical language to Tamil , Odia, Malayalam, Kannada, Telugu and Sanskrit, which have a rich heritage and independent nature. This languages can be classified under the major languages of Indian which is used by more than half the population of the country. It is observed that people face the problem of commutation when moving from one part of the country to the other, to overcome this issue there is need for a translator which converts one language to another. In today's date, there are few number of machine learning models and software's that cater to this issue. A simple example would be the google translate which can translate one language to the other. Such applications have been developed to translate speech from one language to another. Each language is different from the other and some of the differences and importance of language are listed below. Tamil is ancient languages dating back to the 5th century, It is one of the Dravidian language family, being the official language of Tamil Nadu, India, It is also an official language in Sri Lanka and Singapore. It has a substantial number of speaker in South Africa, Fiji, Mauritius and Malaysia. The Tamil Alphabets has 12 vowels and 18 consonants the combination of this lead to a total of 216 character and 20 numeric characters. Kannada is also one of classical language due to its rich heritage dating back to the 5Th century, it is also part of the Dravidian languages, It is the official language of the state Karnataka. It is the second oldest among the Dravidian languages. The Kannada scripts has 13 vowels and it has 2 type of consonants structured and unstructured, there are 25 structured consonants and 9 unstructured consonants. Hindi is the official language of India and the state Delhi, it is one among the 22 scheduled language, it has a rich history dating back to the 7th century .it is one of the Indo-Aryan languages. It is written using Devanagari script. It consists of 16 vowels, 33 consonants and 10 numeric characters.

II. EXISTING WORK

This project uses a com bination of machine learning model to achieve the result. The technology that have been used in this project are Tesseract OCR Engine and the Sequence to Sequence model. The research that has been carried out in this Domain, we have spilt the review in two sections and are as follows:study about different technique to develop OCR.Study about different language translation technique.The Tesseract OCR engine is an open-source software and is been widely used. Liyanage et al.[1] has proposed a OCR which can recognized printed Tamil characters of different font and size using the Tesseract OCR engine.The authors have made a dataset to address this application. The dataset is a collection all Tamil character vowels, consonants, Consonants with touching modifiers, Consonants with pulli Ligatures, Ligatures with touching modifiers Non-touching modifiers and Punctuation. The dataset was made in different sizes and the font. thenneUni ,Sundaram and Akshar was used ,this fonts pro duce better results frotamilocr.The system was evaluated by taking 20 from 20 ancient books.it contained a total of 14031 characters. The existing tesseracts was able to recognize 10103 characters and the fine tuned tesseract was able to recognize 11398 and able to archive 81% accurate results.

In Smiths[2] study, he was able to compare the performance of the two versions of the tesseract engine on printed roman scripts. The tesseract engine improvement with its predecessors is shown. the tesseract version 2 was improved by making the OCR in stages.at first, the borderlines were extracted using image processing, and the characters are separated and made as blobs. each character is recognized one at a time resulting in a stream of words and it was observed to have significant improvement when compared to its predecessor. Yamakawa and Yoshiura [3] in there approach have used Tessract OCR engine, detect spam mails, while many researchers have made an effort to increase the efficiency of the tesseract engine ,This paper used the tesseract to identify mail filter, it doesn't not include any fine tuning and the tesseract is sued as it is. The text data output is analysed to find patterns and inappropriate content. they have trained the model to recognize spam content even if the background are changed. Mishra et al.[4] ,This paper proposes an methodology to increase the accuracy of the tesseract OCR for Hindi language. The model is split into stage first being the training data generation where the following steps are carried out. Smart Hindi database selection, Training image generation, Box file generation, Train file generation, Character set file generation, Font properties selection, Feature extraction, Clustering, Dictionary data preparation, Post processing ambiguity removal and Training data compaction and the second stage Shirorekha Chopping Algorithm is sued for pre processing the image.

The proposed system has an accuracy of 94%.The tesseract engine is widely used in the recognition of hand written characters, Rak- shit and Basu [5] have developed a system which recognize roman scripts and where able to achieve better accuracy for specific handwritten text. Tesseract is trained to recognize handwritten character by individuals it addresses both isolated text and free-flow texts. The dataset was created using the bb Tesseract tool and labeled accordingly. this dataset was used to train the new Tesseract language .experiment was conducted by taking three users from whom the dataset was created, it was observed the Tesseract accuracy was good for the specific individuals but dropped when new individuals hand- written was checked.

Sagar et al[6]. they have attempted to develop up Kannada OCR by using a ternary search to predict the word in the text document. The following steps was car- ried out increase the accuracy of the OCR , segmentation, Character Recognition and Postprocessing modules, however segmentation is a problem as Kannada characters are curved and hence a challenge. Whereas Sanjeev and Sudhakar samuel[7] have developed a system based on wavelet transformation and 2 layer classifier based on neural network.In the domain of language translation, there has been significant work that has been accomplished, Devi et al.[8] in their approach have developed a machine learning based solution which can translate a finite Tamil text to Hindi ,have made an system which converts Tamil verbs to Hindi as verb transfer is challenging in language translation have proposed a method to overcome this issue. this solution is attempted only for a finite verb set consisting of one to one and one to many verb transfer sets. this system consists of 8 module which used machine learning and rule based approach's for verb translation. each of this module focus on the problems of verb transfer the first mod ule being the morphological analyzer where the problem of finite and infinite verbs are focused the second addressing person's number and Gender, the third module negative and positive verb forms ,following is verb chunks., using this modules individually to address each issue. while Thenmozhi et al.[9] have built a verb phrase translator using the seq2seq deep learning model to translate English to Tamil and Hindi to Tamil. in this study an the authors have proposed a methodology which uses the seq2seq model for phrase translation from English to Tamil and Hindi to Tamil. to activate this, they have followed the following steps. extract the Hindi and English along with the Tamil input sequences from the datatset. the dataset is split into training and testing dataset. the vocabularies for English, Hindi and Tamil sequence are determined i.e. each word is given a unique id. A encoder-decoder model is build using deep neural network. with attention model for selecting the right sequence of words. then the sequence in the test dataset is used to find the accuracy of the developed model. the accuracy was increase by 10% for English to Tamil and 16.84%.Similarly using seq2seq model Cho et al.[10] have proposed a RNN Encoder-Decoder model which can translate a sequence phrase to it equivalent.

III. PROPOSED SYSTEM

The purpose of this research is to develop OCR translator which converts Indic sign board to English equivalent using machine learning models. This is achieved by by using a combinations of machine learning model.The tesseract engine is used as a OCR to recognize the language and provide a machine encoded text for each language and the seq2seq model translates the Indic language Sequence to English.The schema of Reseach is as depicted in the Figure 1 .This represents the different stages of the project and how it has been carried out.In this proposed system the Tesseract OCR engine and the Seq2Seq model is used, data is required to to train and fine tune this model. The Tesseract OCR engine is already capable of recognizing indian language ,This project mainly focus on 3 languages Tamil, Kannada and Hindi. To improve the accuracy of the Tesseract Engine ,some training is required. The Seq2Seq model has to be trained with the acquired dataset for all 3 Languages. Both these models are tested individual for accuracy . once a agreeable level of accuracy is obtained from both the model .The model is integrated and tested as a whole.

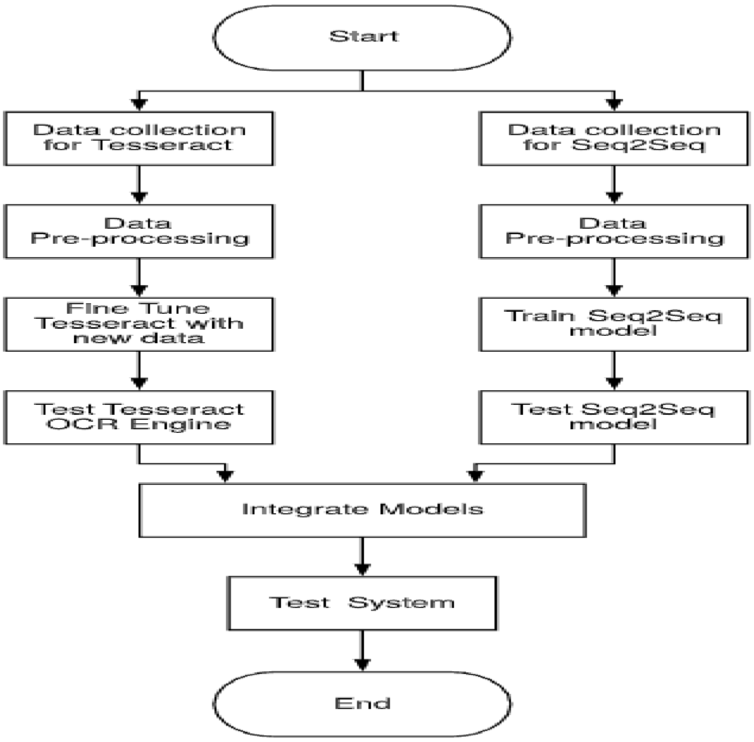
Start
Data collection for Tesseract
Data collection for Seq2Seq
Data Pre-processing
Data Pre-processing
Fine Tune Tesseract with new data
Train Seq2Seq model
Test Tesseract OCR Engine
Test Seq2Seq model
Integrate Models
Test System
End

Figure 1: Sequence Diagram

To recognize printed Tamil characters for our OCR. The Tesseract OCR engine has to be fine-tuned for various fonts as in our approach . we aim to recognize printed Tamil characters on name boards, we have obtained a dataset of printed character, it consists of 124 Tamil characters, each character is captured in 100 different fonts and size as shown in fig 1. all this images are reprocessed .we will use this image to train the tesseract model for character recognition. The dataset has 124 folder each folder represents the character in the folder and is named after the character, in each folder the different fonts are addressed by the number 1-100. Fig 1.Tamil character dataset .The tesseract engine can recognize Hindi and Kannada characters and have a reasonable level of accuracy. Hence we have to use the existing OCR engine, Testing for this dataset was done by creating three folders .Corresponding to Tamil, Hindi and Kannada, each folder containing 100 images of printed sentences in each language and tested at ran- dom.The sequence to sequence model is used for text translation, this model has been adopted by various research, as it is proven to provide accurate results. Since we aim to translate from Tamil to English, we need a source text in Indian language and the target text in English. We have created the dataset by using the google translate API to convert the Indian language sentence to English. As the study is aimed at translating Indian sign boards and name boards to its English equivalent, we have the dataset as per this requirement. The accuracy of the sequence to sequence model can be improved by having a dynamic dataset. the text present in the image .it is as depicted in the FIGURE 2.

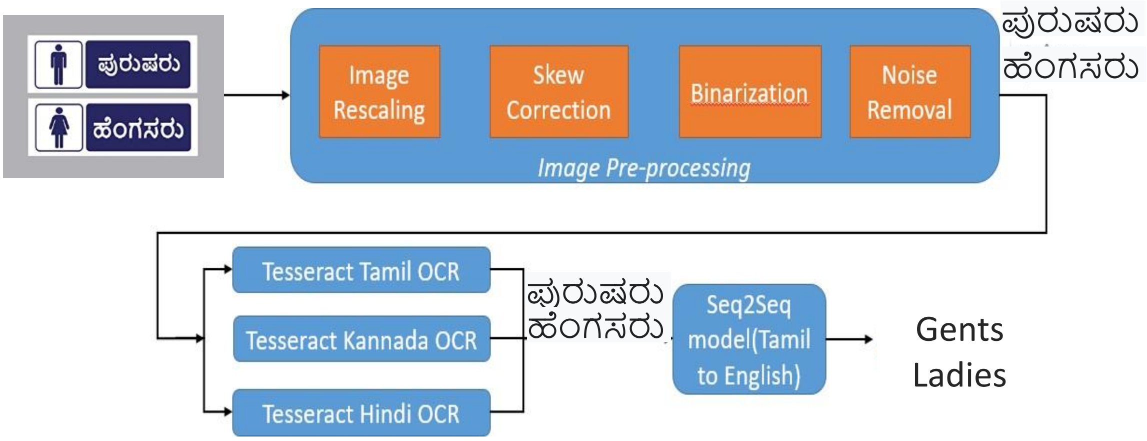

Figure 2: Sequence Diagram

IV. RESULTS AND DISCUSSIONS

In this approach we have three main components image pre-processing, Tesseract OCR engine and the sequence to sequence model for Tamil to English translation. Image pre-processing is a crucial step as it can increase the efficiency of the OCR engine, the processed image is given as input to the Tesseract engine, and the result is Tamil text , this result is then feed to the sequence to sequence model, This model will provide the English equivalent of tamil.

A. Image Pre-processing:

The OCR is dependent on various factors, the image quality is one of the key factors which improves the accuracy of the output, factors such as noise, angle and image play a vital role in the output and hence it is required to pre-process the image to get better results. The steps in pre-processing are :-

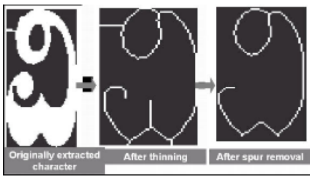

Figure 3: Image Pre-processing

We are-scale each input image to at least 300 DPL .Most images captured are not straight, hence we have to perform skew correction, In this process we need to identify the text. Find the angle and rotate the image to the correct skew. Converting the image to black and white, this is an essential process as our dataset is also Binarized. Raw images always contain noise , removing them will improve the performance of the OCR engine.

B. Tesseract OCR Engine

Tesseract is an open source project owned by google, it is widely used OCR , the new version 4.0 is capable of recognizing 100 languages, however if we need the tesseract to recognize handwritten character we need to train it explicitly, in our system we are interested in printed Tamil text and hence we have trained the tesseract engine to recognize different Tamil character in different text and font. The tesseract engine first identifies the blobs i.e. the words in the image then each blob is then decomposed to the corresponding words. It also checks, if there any missing words or misspelled words and corrects it. This is the reason why the tesseract engine is accurate than other OCR engine available. The loss of accuracy is the high complexity in the Indic scripts. The various possible combination in the scripts is hard to predict, The accuracy achieved 90%.

C. Sequence to Sequence model

The Indian language to English Translation is a sequence to sequence modelling problem. The input language has a sequence and the output represents the English equivalent sequence, it is a many to many sequence problem as there N number of combination in framing a sentence. A simple example is represented in the Figure 3, as in the figure the sequence of words used are different in Indian languages and English. To solve this problem, we used the sequence to sequence model[11], It is a deep learning model .it has a encoder and decoder, both of this are neural network, which are combined into a single giant network. As shown in fig 4 .

The sequence of words are converted into vectors and each vector is identified uniquely. The encoder takes this sequence of vectors as input and understand the sequence of input and creates an intermediate representation of the sequence, the decoder is responsible to understand the sequence and provide the best probable sequence of character.

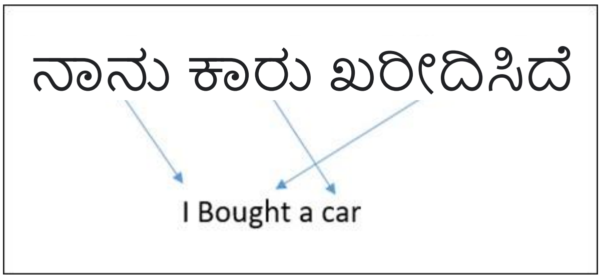

Figure 4: Sequence in Tamil and English

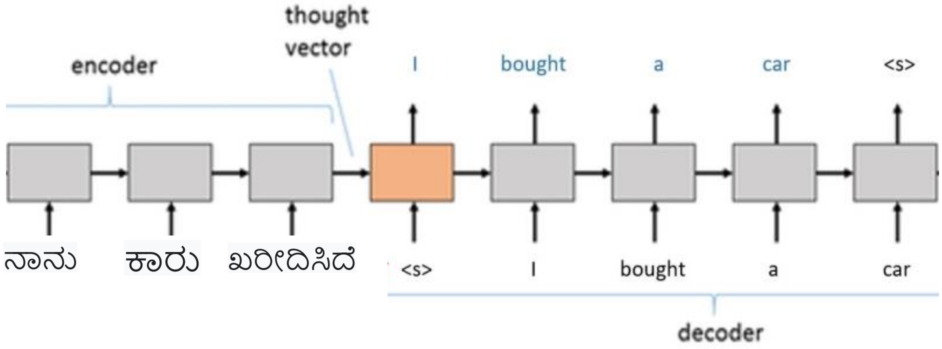

Figure 5: Sequence to Sequence Model

However it is not wise to use the sequence to sequence model directly, the model has to be refined to produce better results. It requires a means of selecting the best sequence of words for all the possible options to achieve better performance, in this study we have used the Beam search method. Each word in the sequence is generated by the decoder network, the network choose the word with the highest probability with respect to the previous sequence of words. The decoder network has to choose the best probable word in the sequence every time, this is one approach, but the beam search method takes the probability of next k words, the next word is choose by taking the highest combined probability. The accuracy of the sequence to sequence model can be further enhanced by having wide combination in the dataset, for this study we have only taken 5200 different sequences for training and 2000 sequences for testing the model.

V. CONCLUSION

In this paper we have proposed a system, which translates Indian signboard to English, Indic language OCR have always been a challenge as they contain a verity combination as compared to the Other languages such as English. In previous approaches Tamil OCR's have been developed by fine tuning the tesseract engine as per the application requirement and have had acceptable results. In the proposed system, we have trained the tesseract with 100 different and of different size for Tamil and used the existing OCR engine for Hindi and Tamil.. We were able to increase the accuracy of the tesseract engine. In the proposed method we have used the sequence to sequence model and have further increased the accuracy by using the beam search model. We have used minimal data to train both model ,this system could be further enhanced by have a wider variety of fronts for the tesseract engine and huge dataset for the sequence to sequence model with various combination of sentences. In this paper we have not approached the problem of recognizing Tamil, Kannada and Hindi character in different angles, this feature could further enhance the reliability of the system.

REFERENCES

[1] C. Liyanage, T. Nadungodage, R. Weerasinghe, L. Technology, and S. Lanka, "De- veloping a commercial grade Tamil OCR for recognizing font and size independent text," pp. 130–134, 2015.

[2] R. Smith, "An Overview of the Tesseract OCR Engine," 2005.

[3] D. Yamakawa, "Applying Tesseract-OCR to Detection of Image Spam Mails," vol. 1.

[4] N. Mishra, C. Patvardhan, C. Vasantha Lakshmi, and S. Singh, "Shirorekha Chop- ping Integrated Tesseract OCR Engine for Enhanced Hindi Language Recognition," Int. J. Comput. Appl., vol. 39, no. 6, pp. 19–23, Feb. 2012.

[5] S. Rakshit and S. Basu, "Development of a multi-user handwriting recognition sys- tem using Tesseract open source OCR engine," pp. 1–6, 2009.

[6] B. M. Sagar, G. Shobha, and P. Ramakanth Kumar, "Complete Kannada Optical Character Recognition with Syntactical Analysis of the Script," in Proceedings of the 2008 International Conference on Computing, Communication and Networking, ICCCN 2008, 2008, pp. 1–4.

[7] "An OCR System for Printed Kannada Text Using Two - Stage Multi-network Classification Approach Employing Wavelet Features. International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)10.1109/iccima.2007.191." https://ieeexplore.ieee.org/document/4426720. [Ac- cessed: 14-Apr-2020].

[8] S. L. Devi, P. Pralayankar, S. Menaka, T. Bakiyavathi, V. S. R. R, and V. Kavitha, "Verb Transfer in a Tamil to Hindi Machine Translation System," pp. 269–272, 2010.

[9] D. Thenmozhi, B. S. Kumar, and C. Aravindan, "Deep Learning Approach to English-Tamil and Hindi-Tamil Verb Phrase Translations," pp. 1–9.

[10] B. Van Merri and C. S. Fellow, "Learning Phrase Representations using RNN Encoder – Decoder for Statistical Machine Translation," pp. 1724–1734, 2014.

[11] I. Sutskever, "Sequence to Sequence Learning with Neural Networks," pp. 1–9.