Tamil OCR Post Processing

By

Group 15

2013103047	Sai Mageshvar C S
2013103550	Mathan Pandi P
2013103564	Sidharth Yatish M

CS8612 – Creativity and Innovative Project

Submitted to:

Dr. Rajeswari Sridhar

Asst. Professor (Sr. Grade), Department of Computer Science and Engineering,

Anna University, Chennai - 25

Phone: (O) 044 - 22358838

94450 – 01236

Summary:

Tamil OCR post processing aims at correcting the errors produced by a Tamil language OCR. Documents stored as papers are in need to be digitized for future reference. This can be obtained by scanned copies of such documents. There is a need of editable text from this image file so that the text can be reproduced somewhere else. This can be obtained by process of OCR (Optical Character Recognition) that extracts text from an image. The text so obtained will contain lot of errors arising from the bad quality of image and false character prediction by the OCR. Manual correction of such errors are cumbersome for very huge documents.

The text must be processed so as to correct the errors. This must be done in reference to the context so as to preserve the originality. English documents yield high accurate results. Tamil language being one of the oldest and morphologically rich makes it difficult to obtain much accuracy. There is a wide set of characters with close resemblance of several characters. In addition to character level difficulty, the partial free word order nature of the language makes it difficult to correct the errors at sentence level.

In projects done so far, a bigram model is used to predict the next word based on the context of the sentence. Errors in formation of words are corrected using Tamil grammar rules. Words are split into root and suffixes and possible root words are listed. From this the appropriate root word is chosen by constructing a bigram model. The accuracy of the project from a single OCR software (Ponvizhi) was 91.75%.

This project aims at correcting such errors with high accuracy. This can be further extended to trigram model to achieve higher accuracy in prediction. Increasing the set of suffixes can also help in obtaining accurate result. Moreover, instead of obtaining results from a single OCR software, outputs of multiple OCR software are used to help in improving the accuracy.