

Developer Manual for Tamil Typesetting

CQRL Bits LLP

S2, Shreya Shelters, 23, Vignesh Avenue Second Street

Selaiyur, Chennai 600073

Email: cqrlbits@gmail.com

Introduction

For Tamil language (unlike English language), the rules for breaking words have not been automated yet. We have implemented an SOTA (State of the Art) hyphenation algorithm using a robust Machine Learning framework using a corpus of hyphenated words with breaking points curated by language experts.

Method

We have used a training dataset of approximately 8,000 manually hyphenated words (with the help of late Crea Ramakrishnan) to train using a Javascript machine learning library Brain.js (<https://brain.js.org/#/>). The training was done offline using Node.js script and deployment was done in the client's side using a JSON model output.

For the command line implementation, we used Java-based Weka application to train a Hierarchical Bayesian Network model, and a test accuracy of 96% was achieved.

Deployment

The Javascript model can be deployed directly copying the static HTML files and the accompanying resource files. The command line webservice application is a Java module wrapped in a Perl script using standard IN and standard OUT Linux bash shell.

Deliverables

