ISSN: 1391-8796
Proceedings of
4th Ruhuna International Science & Technology Conference
University of Ruhuna, Matara, Sri Lanka
January 26, 2017

A preliminary model of predictive text for Sinhala using N-gram Statistics

Chanaka K.M.R.*, Lorensuhewa S.A.S. and Kalyani M.A.L

Department of Computer Science, University of Ruhuna, Matara, Sri Lanka.

Most Sri Lankans use Sinhala text processing in their day to day activities. But, they feel it hard to type documents in Sinhala and also it takes more time and involves typing mistakes and therefore efficiency is low. Integration of word prediction facility helps the user to select words rather than typing the words repeatedly to reduce the number of required keystrokes, minimize mistakes and reduce time. The aim of this research is to explore the use of Natural Language Processing and Machine Learning techniques to assist Sinhala typing tasks by predicting the words.

We predict the next word to type from n-gram probabilistic model which involves bi-gram, tri-gram and a combination of bi-gram and tri-gram. This composite n-gram model includes both bi-gram and tri-gram, giving high priority to the tri-gram suggestions. The n-gram corpus is generated from Sinhala corpus collected from online Sinhala newspapers. A maximum prediction percentage of 41 was achieved for sports documents by using domain specific n-gram corpus of sports documents and obtained an 18.1% average keystroke reduction by using the prediction model. We tested with other news categories such as political, legal and local collected from local newspapers as well. According to our experimental results, composite n-gram model outperformed bi-gram and tri-gram word prediction models and the domain specific composite n-gram model performs better than the composite model created from a mixed corpus.

Our goal is to automatically cluster the document corpus and classify the edited text after entering certain amount of text and get the predictions from a relevant cluster dynamically to improve the accuracy at runtime, giving a more relevant prediction.

Keywords: *Word Prediction, Dynamic Text Prediction, N-Gram Model, NLP, Text Mining*

*Corresponding Author: ruwanpelwatta@gmail.com