PART OF SPEECH TAGGER FOR SINHALA LANGUAGE

A.J.P.M.P. Jayaweera

THESIS SUBMITED FOR THE DEGREE OF MASTER OF PHILOSOPHY
UNIVERSITY OF KELANIYA, KELANIYA, SRI LANKA
MARCH 2015

Abstract

This dissertation presents a stochastic based Part of Speech tagging method for Sinhala language. Part of Speech (POS) is a very vital topic in any Natural Language processing task that involves analyzing the construction of the language, behavior of the language and the dynamics of the language. This knowledge could be utilized in computational linguistics analysis and automation applications. The motivation behind the research was to fulfill the gaps which are existed at present in the research area of Natural Language Processing (NLP) and analysis of Sinhala language and giving a push to computational linguistics analysis of Natural Language processing research in Sinhala language.

Though Sinhala is a morphologically rich language, in which words are inflected with various grammatical features, tagging is very essential for further analysis of the language.

Our research is based on a statistical approach, in which the tagging process is done by computing the tag sequence probability and the word-likelihood probability from the given corpus, where the linguistic knowledge is automatically extracted from the annotated text.

Our effort was mainly focused on designing an architecture for the tagger and development of the tagger. The implementation of the tagger was based on a well-known stochastic model, known as Hidden Markov Model (HMM). The distinction between open class and closed class word categories together with syntactical features of the language were used to predict lexical categories of unknown words.

Simple Good-Turing algorithm and Witten-Bell discounting methods were used to resolve spare data issues.

The evaluation of the tagger was done by using the corpora and the tag set developed by the University of Colombo School of Computing (UCSC) in year 2005 under the PAN Localization Project. The model was tested against 90551 words, and 2754 sentences of Sinhala text corpus and the tagger could reach over 90% accuracy in the tagging process which shows a considerable success over previous works reported in 2004 and 2013. In 2004, a Hidden Markov Model based Part of Speech tagger was proposed using bigram model and reported only 60% of accuracy and in 2013 another Hidden Markov Model based approach was tried out and reported around 62% of accuracy. However, the overall accuracy of the tagger we implemented have shown more than 90%, a set of improvements are suggested in this dissertation mainly in the area of handling unknown words.

Eventhough these other research were carried out for Sinhala language, they are not available to use as tools for further language analysis of Sinhala language. So as an additional product of this work we have make the tagger that we implemented available as an on-line interface on web freely accessible to the public.