SVM BASED PART OF SPEECH TAGGER FOR SINHALA LANGUAGE

Y. A. D. S. S. Wijerathna

A dissertation submitted in partial fulfilment of the requirement for
Master of Science degree in Advance Software Engineering

Department of Computing

Informatics Institute of Technology, Sri Lanka
in collaboration with
University of Westminster, UK

2020

Abstract

Natural language processing is providing the ability of understanding human language as it is spoken to a computer programme. It is a component of computer science, linguistics and artificial intelligence. Building NLP application is a difficult task because human speech is not always specific. NLP is a process of developing a system that can read text and translate between one human language and another. In performing NLP task part of Speech tagging is a basic requirement which needs to catered appropriately.

Part of Speech (POS) tagging is the task of labelling each word in a sentence with its appropriate syntactic category called part of speech. POS tagging is a very important pre-processing task for language processing activities.

Sinhala is the native language of the Sinhalese people who make up the largest ethnic group of Sri Lanka. Sinhala is a morphologically rich language in the Indic family. While this makes it harder to build an accurate tagger without a good morphological pre-processor, it provides a compelling reason for attempting to build a POS tagger for Sinhala. However, due to poverty in both linguistic and economic capital, Sinhala, in the perspective of Natural Language Processing tools and research, remains a resource-poor language.

This paper is an initiative to overcome above issue in lack of NLP tools for Sinhala language, here discusses the task of POS tagging for Sinhala language using Support Vector Machine (SVM). The POS tagger has been developed using a tag set of 30 POS tags, defined for the Sinhala languages by University of Moratuwa and used the SVMTool developed by Jesus Gimenez and Lluıs Marquez.

The accuracy of available Sinhala Part-Of-Speech taggers, which are based on Hidden Markov Models, still falls far behind state of the art. Our Support Vector Machine based tagger achieved an overall accuracy of 85.68% for known words.