MSc Project Report

"A Parts of Speech Tagger for Divehi: A Step Towards Preserving and Enrichment of the Maldivian Language"

A Dissertation By

Mohamed Shamikh Rameez

Supervised By

Dr. Ruvan Weerasinghe

IIT Student ID: 2016249

RGU Student ID: 1614372

2018

A report submitted as part of the requirements for the degree of
MSc Big Data Analytics at Robert Gordon University, Aberdeen, Scotland

Abstract

Divehi is a language used solely by Maldivians. However most of the Maldivians are using English instead of Divehi and the use of Divehi language are becoming less day by day. There is doubt that the Maldivian language could become an endangered language in the next few years. Therefore, to prevent the extinction of the language, Divehi language needs to be compatible with the latest technologies and trends. Natural Language Processing helps in providing tasks such as translation and spell checkers which would encourage the usage of the language. NLP is mostly dependent on Parts of Speech Tagger which is not available currently for Divehi language

This research aims to implement a prediction system for automated tagging by using machine learning algorithm such as Support Vector Machine (SVM) and Conditional Random Field (CRF). The Parts of Speech Tagger would be a starting point for the researches on NLP domain in Maldives which leads to more usage of the Maldivian language. The research discusses the various resources available in Divehi and develops a corpus and two tag sets which can be used for NLP research.

Keywords:

Parts of Speech Tagger, Maldivian Language, Natural Language Processing, Support Vector Machine, Conditional Random Field