Polylingo - A Short Utterance Based Automatic Sinhala Language Identification & Translation Tool

Aysha Manal Arafath

A dissertation submitted in partial fulfilment of the requirements for

Bachelor of Science (Honours) degree in Computer Science

Department of Computing

Informatics Institute of Technology

In Collaboration With

University of Westminster, UK

2020

Abstract

Language Identification (LI) has become a popular research field in the past couple of years. It is the process of identifying the language spoken from an audio recording. Researches have been done using different approaches to increase the accuracy of the system. Language identification also play an important role in systems such as the Automatic Speech Recognition Systems (ASR). Hence, it has many uses to it. However, most of the researches in this field focuses on the commonly used languages and languages which are low resourced tend to get left behind from these benefits. Sinhala language which is spoken by over 16 million people is still considered low resource, as efforts are not made to do research in this field and make resources public. Despite certain researches been done in the text field of Sinhala, there are no corpuses available publicly for research to be done in Sinhala speech.

PolyLingo is an approach to automatically identify Sinhala language and translate it to other languages. A create a clean dataset will be built and made publicly available in order to aid for future researches in the field of Sinhala speech. Bidirectional Long Short Term Memory (LSTM) will be used in order to automatically identify the language within a short time frame.

Keywords: Speech language identification, Natural language processing, Sinhala language, Bidirectional LSTM