MACHINE COMPREHENSION PLATFORM FOR SINHALA LANGUAGE

Chathuni Vidarshana Abhayasekara

A dissertation submitted in partial fulfilment of the requirements for the BEng (Hons) in Software Engineering degree

Department of Computing

Informatics Institute of Technology, Sri Lanka

In collaboration with

University of Westminster, UK

2021

Abstract

Along the time development in the natural language processing has helped a lot to improve the state -of-the-art performances on so many tasks regarding languages like English. But in Srilanka, Sinhala is the most used language, and it is used as the medium of documentation in almost all the activities. Due to the lack of research on Machine Comprehension for Sinhala, similar progress has not been achieved. Unlike English, Sinhala language does not have a collected benchmark large scale QA dataset or a pretrained language model which can be improved for Sinhala Machine Comprehension or a human baseline score for Question Answering as well.

This project is about exploring the possibilities of applying deep learning approaches for Sinhala Machine Comprehension. However, in this project state-of-the-art transformer models were used for the training of Machine Comprehension System on an artificial reading comprehension dataset which followed SQuAD2.0 structure. Furthermore, Sinhala Articles from Wikipedia were used in creating the dataset. Transfer learning is used for the implementation and system is capable of extracting answer from a given text.

Keywords:- Machine Comprehension, Deep Learning, NLP, Transformers