INFORMATICS INSTITUTE OF TECHNOLOGY
In collaboration with
The University of Westminster, UK

BEng/BSc (Hons) DEGREE PROGRAMME in
Software Engineering

Final Year Project 17/18

For

Semi-supervised Corpus based POS tagger for Sinhala

Dissertation by

2014159 Kumal Perera

Supervised by
Mr. Guhanathan Poravi

Abstract

Most of the rich morphological languages are being endangered due to the lack of resources and also since most of the countries are still being developed. It takes time to build up a status where a particular language has enough and more resources. It is found that 22 million people in Sri Lanka use Sinhala most of the time. Even though that much of people use the local language, not much priority is given to the obligation of building up technical libraries for it. Local language should be prioritized as it is our nationalistic obligation to hold the local culture. One finds it more comfortable in using their own language.

According to the research done so far, it is found that even though there have been a series of work done for the development of the local language for POS tagging, not much accuracy is found to help it grow. The whole purpose of the project **Psephology** is to come up with Part of Speech tags for Sinhala words. It has many uses in regard to implementing another system, and checking local grammar context as well. With the anticipation that this would overcome the chance of the local language being endangered, several approaches were proposed to uplift the performance and usability of the product and minimize the discrepancies.

All the libraries were implemented using the Python framework. Sub libraries such as NLTK, Scikit-learn and NumPy were also used. Implemented system was tested thoroughly under different conditions and the Lexicon system was evaluated by evaluators of various domains. Eventually, the test results attested that the analysis, design, implementation and documentation have been carried out in an effective and in an efficient manner.

Subject descriptors:

Natural Language Processing

Text classification

Keywords:

POS tagging, Machine Learning, Hidden Markov Model