Abstract 23

A Comprehensive Part of Speech (POS) Tag Set for Sinhala Language

Nimasha Dilshani[1], Sandareka Fernando[2], Surangika Ranathunga[3]
Associate Prof. Sanath Jayasena[4], Prof. Gihan Dias[5]

[1]Department of Computer Science and Engineering, University of Moratuwa
nimasha.dilshani.161@gmail.com
[2] Department of Computer Science and Engineering, University of Moratuwa
sandarekaf@cse.mrt.ac.lk
[3]Department of Computer Science and Engineering, University of Moratuwa
surangika@cse.mrt.ac.lk
[4]Department of Computer Science and Engineering, University of Moratuwa
sanath@cse.mrt.ac.lk
[5]Department of Computer Science and Engineering, University of Moratuwa
gihan@cse.mrt.ac.lk

Sinhala, which belongs to Indo-Aryan language family, is a morphologically complex language. Most of the features of the words are postpositionally affixed to the root word. Thus, well-developed Part of Speech (POS) tag sets for languages such as English cannot be easily adopted to create a POS tag set for Sinhala. Moreover, currently available Sinhala POS tag sets have many limitations such as the unavailability of tags for certain words. The objective of the research is to overcome and to identify ambiguities and limitations of the present POS tag sets for Sinhala language, and to develop a comprehensive multi-level tag set for Sinhala language. The new tag set was designed after a thorough evaluation of different types of corpora such as news articles and official government letters, and as well as an analysis of the existing POS tag set for Sinhala. This new tag set consists of 148 tags and is organized into 3 levels. Thus, it covers most of the word classes and inflection based grammatical variations of the Sinhala language. The ultimate purpose of developing this tag set is to implement an automatic POS tagger, which is an essential tool in implementing Natural Language Processing Applications. To train the automatic POS tagger, a corpus of 300000 words has been POS annotated manually using this tag set. This tag set produced an overall accuracy of 84.68% and it bypasses the other Sinhala POS taggers. However, this annotation is done only up to level 2 in the tag set. Annotating at level 3 has the potential to introduce many ambiguities to the manual annotation process, due to the large number of POS tags. Thus this opens up new research avenues to investigate on the use of inflectional morphological features of Sinhala language, in order to determine the POS tag of a word at the third level.

Key words: Lexical, Morphology, Natural Language Processing (NLP), Parts of Speech (POS), Sinhala