

Automated Identification of Compound Nouns in Sinhala Using Natural Language Processing

A.D.D.C.M Mahindarathna* and R.S. Wickrama Arachchi

Department of Information and Communication Technology, Faculty of Technology, University of Ruhuna, Karagoda Uyangoda, Kamburupitiya, Matara, Sri Lanka.

* madhubhashana99@gmail.com

Abstract

This research proposes an automated system for the identification and classification of compound nouns in Sinhala, one of the most complex and crucial features of the language. In Sinhala compound nouns, a prefix is often combined with meaningful stems to convey specific semantic and grammatical roles. As part of this, two major types of Sinhala compound nouns will be addressed in this research, namely *Avya Samasa* and *Dakarartha Samasa*. This study aims to build an effective NLP system for correctly identifying the types of compound nouns. The system studies collocations such as "ගස් ද වැල් ද" forming "ගස්වැල්" and "පරසිරිත්" from "පර" and "සිරිත්". The methodology to be followed is creating a dataset of 6,000 Sinhala compound nouns by using 43 frequent Sinhala prefixes, together with POS tagging to ensure that the stems are valid nouns, verbs, or adjectives. Various machine learning algorithms were applied; these include Random Forest, Support Vector Machines (SVM), and Naive Bayes. The highest accuracy attained was by Naive Bayes at 75%. POS tagging plays a very important role in verifying the grammatical integrity of stems, and filtering out non-compound words when the stem is not meaningful. Results show that this system enhances the classification of Sinhala compound nouns and decreases the complexity of identifying valid compound forms. However, due to the limited NLP resources in Sinhala, as well as the morphological complexity of the language itself, some problems are encountered. The system is of relevance both for linguistic research and teaching, to learners of Sinhala for a better comprehension of compound formations. Future studies may be done to expand the dataset and refine machine learning techniques to increase the accuracy. The limitations of the paper include dependence on the quality of the dataset, and it is challenging to handle all minute differences in the linguistic subtleties in handling different types of compound nouns, where deep learning models may come in handy.

Keywords: *Sinhala language, NLP, compound nouns, POS tagging, Naive Bayes*