Sinhala Named Entity Recognition Model: Domain-Specific Classes in Sports

W.M.S.K.Wijesinghe[1] and Muditha Tissera[2]

[1]Department of Statistics and Computer Science, Faculty of Science, University of Kelaniya.
[2]Department of Software Engineering,Faculty of Computing & Technology, University of Kelaniya.
sithmini456@gmail.com, mudithat@kln.ac.lk

Abstract

Named Entity Recognition (NER) is one of the crucial and vital subtasks that must be solved in most Natural Language Processing (NLP) tasks. However, constructing a NER system for the Sinhala Language is challenging. Because it comes under the category of low-resource languages. Therefore, the proposed approach attempted designing a mechanism to identify specific named entities in the sports domain. Firstly, a domain-specific corpus was built using Sinhala sport e-News articles. Then a semi-automated, rule-based component named as 'Class_Label_Suggester' was built to annotate pre-defined named entities. After auto annotation, the outcome was further validated manually with a little effort. Finally, it was trained using the annotated data. Linear Perceptron, Stochastic Gradient Descent (SGD), Multinomial Naive Bayes (MNB), and Passive Aggressive classifiers were used to train the NER model. Though, the above Machine Learning (ML) algorithms showed approximately 98% accuracy, the MNB model demonstrated highest accuracy for the identified class labels of which, 99.76% for 'Ground', 99.53% for 'School', 98.55% for 'Tournament', and 97.87% for 'Other' classes. Additionally, high precision values of the above classes were 81% ,72%, 62%, and 98% respectively. An accurately annotated Sinhala dataset and the trained Sinhala NER model are main contributions of the study.

Keywords

Sinhala NLP, Semi-Automated NER, Class- Label-Suggester, Low Resource Language, Machine Learning, Sports Domain