Support Vector Machine based Named Entity Recognition for Sinhala

P.S. Mallikarachchi*, S.A.S. Lorensuhewa and M.A.L. Kalyani

Department of Computer Science, University of Ruhuna, Matara, Sri Lanka
**Corresponding Author E-mail: mallikarachchi9926@usci.ruh.ac.lk, TP: +94719942728*

Named Entity Recognition (NER) can be defined as identifying Named Entities (NE) in human language and classifying them. A NER system is a major fundamental subtask that facilitates more complex tasks like automatic text summarization, question answering, etc. Today automated language tools are a more solved problem for resource-rich languages like English. But for Sinhala, which is a low resourced South Asian language, only a few prior works can be observed. Unfortunately, systems developed for the English language cannot be directly used for Indo-Aryan languages. Considering the attempts on Sinhala NER systems, it can be observed that only Conditional Random Fields (CRF) and Maximum Entropy (ME) were used. But for other low resourced Indo-Aryan languages, several other algorithms have been used and among them Support Vector Machines (SVM) have given more prominent results. In this paper, we present a novel NER system using SVM for the Sinhala language. Here we have only considered PER (person), LOC (location) and ORG (organization) tags. Since this is a data driven approach preprocessing of the training data is a crucial task. The most suitable format for the training data is word-per-line format (CONLL-2002). For a more extended classification task Beginning-Inside-Outside tagging scheme was followed increasing the total number of tags into 7. The dataset consisted of 100,000 tokens and the first we have observed that with size of the training data, performance is increasing. As the prior works have shown the effect of language features next we have observed the behavior of different feature combinations and figure out that gazetteers, clue words, word-length and Part-of-Speech features as the most effective for PER, LOC. Excluding the word-length from above mentioned features remaining are the best for ORG. Ultimately both sets of tags were able to prove the effect of gazetteers with SVM. Next we have set up the experiments to observe the impact of the word-length of 4,5,6,7. Lengths of 4 and 5 were best matched for the purpose of this work. As future work we have planned to experiment the influence of varying the kernels, context and degree while expanding the training data.

Keywords: SVM; NER; NLP; BIO