Jayaweera, A.J.P.M.P
Dias N.G.J.
PAPER

Evaluation of Stochastic Based Tagging Approach for Sinhala Language

A.J.P.M.P Jayaweera, Virtusa (Private) Limited, Colombo 9
N.G.J. Dias , Department of Statistics & Computer Science, University of Kelaniya

Part of Speech (POS) tagging is one of the fundamental and important steps of any Natural Language Processing (NLP) task, from speech recognition to machine translation, text to speech, spelling and grammar checking to language-based information retrieval on the Web, etc.Tagging is the process of assigning a part-of-speech or other lexical class marker to each word in a sentence based on its morphological and syntactical properties.
Sinhala is a morphologically complex and agglutinative language which has a lot of similar features to other South Asian Languages, such as Hindi, Tamil, Bengali, etc. In Sinhala language, words are inflected with various grammatical features; most words are postpositionally affixed to the root word.

Automatically assigning a tag to each word in a language like Sinhala is very complex. So the objective of this paper is to evaluate the Stochastic based tagging approach for Sinhala language, which uses statistical methods to assign tags to each word in a sentence. The approach discussed in the paper is based on a well known stochastic based tagging approach, the Hidden Markov Model (HMM) which selects the best tag sequence for a complete sentence rather than tagging word by word. The historical evidence shows that HMM based approach is a widely used tagging approach in other research studies carried out for other languages.

The tagger presented here takes a sentence, a tag set and a corpus as input and gives the tagged sentence as output. The tagging process is done by computing the tag sequence probability$P(t_i|t_{i-1})$ and a word-likelihood probability $P(w_i|t_i)$ from the given corpus, where the linguistic knowledge is automatically extracted from the annotated corpus. In this research, we have used the tagset and the corpus developed by UCSC/LRTL (2005) under PAN Localization Project. The current tagset consists of 29 morpho-syntactic tags. An algorithm is presented in this paper for implementing POS tagging system for Sinhala language. The evaluation was done by using a 14549 word tagged corpus. Testing was done with text extracted from different sources. The approach was evaluated, and produced tag sequences with accuracy between 80% - 97%. With the result obtained from this research, we could say Stochastic based tagging approach is well suited for the Sinhala language. But still there is much more research needed to optimize the accuracy of tagging the Sinhala language.