IPRC/16/192

Comparison of Part of Speech taggers for Sinhala Language

Manoj Jayaweera[1*], N. G. J. Dias[2]

Part of Speech (POS) tagging is an important tool for processing natural languages. It is one of the basic analytical model used in for many Natural language processing applications. It is the process of marking up a word in a corpus as corresponding to a particular part of speech like noun, verb, adjective and adverb. Automatic assignment of descriptors to the given tokens is called Tagging. The descriptor is called a tag. The tag may indicate one of the parts of speech category and the semantic information. So tagging is a kind of classification. The process of assigning one of the parts of speech to the given word is called parts of speech tagging. It is commonly referred to as POS tagging. In grammar, a part of speech (also known as word class, lexical class, or lexical category) is a linguistic category of words (or more precisely lexical items), which is generally defined by the syntactic or morphological behavior of the lexical item in the language. Each part of speech explains not what the word is, but how the word is used. In fact, the same word can be a noun in one sentence and a verb or adjective in another. In most of the natural languages in the world, noun and verb are common linguistic categories among others. Almost all languages have the lexical categories noun and verb, but beyond these there are significant variations in different languages. The significance of the part of speech for language processing is that it gives a significant amount of information about the word and its neighbours.

There are different approaches to the problem of assigning a part of speech tag to each word of a natural language sentence. The most widely used methods for English are the statistical methods that is Hidden Markov Model (HMM) based tagging and the rule based or transformation based methods. Subsequent researches add various modifications to these basic approaches to improve the performance of the taggers for English. In this paper we present a comparison of the different researches that was carried out of POS tagging for Sinhala language. For Sinhala language, there were 4 reported work for developing a POS tagger. In 2004, a HMM based POS tagger was proposed using bigram model and reported only 60% of accuracy. Another HMM based approach was tried out for Sinhala language in 2013 and reported a 62% of accuracy. In 2016, another research was reported 72% of accuracy which was a hybrid approach based on bi-gram HMM and rules based approach in predicting the relevant tag for unknown words. The tagger that we have developed is based on a trigram based HMM approach, which used the knowledge of distribution of words and parts of speech categories in predicting the relevant tag for unknown words. The Witten-Bell discounting technique was used for smoothing and our approach gave an accuracy of 91.50% with a corpus of 90551 annotated words.

Keywords: *Sinhala language, Natural Language Processing, Part of Speech tagging, Hidden Markov Model, Hybrid Tagging Approach*

[1] *mjayaweera@gmail.com

[2] University of Kelaniya, Sri Lanka