Unknown Words Analysis in POS tagging of Sinhala Language

A.J.P.M.P. Jayaweera[#1], N.G.J. Dias[*2]

[#] *Virtusa Pvt. Ltd.*
No 752, Dr. Danister De Silva Mawatha, Colombo 09, Sri Lanka
[1]mjayaweera@gmail.com

[*]*Department of Statistics & Computer Science, University of Kelaniya*
Kelaniya, Sri Lanka
[2]ngjdias@kln.ac.lk

***Keywords*—Part of Speech tagging, Morphology, Unknown words, Sinhala Language.**

Appearance of unknown words is one of the frequently occurring problems facing in part of speech (POS) tagging process, i.e., the words that appear in sentences, but are not contained within the training corpus. New words are continually coined to the language, and people will often use words that are parsing, the system may not expect. This problem get worse when NLP systems are used for more and more on-line computer applications. New words are continually entering the language, Acronyms and proper names are created very often and new nouns and verbs are adding to the language in a surprising rate. So it is impossible to train the tagger for every possible word in the language. So unknown words are non-negligible in POS tagging. Therefore, in order to build a complete tagger, tagger must be incurred with some knowledge of suggesting the tag for an unknown word.

Previous techniques reported for other languages such as English, have mostly utilize the guessing rules to analyse the word features by looking at leading and trailing characters. Most of them employ the analysis of trailing characters and other features such as capitalization and hyphenation. Some of them use more morphologically oriented word features such as suffixes, prefixes, and character lengths. The guessing rules are usually use knowledge of morphology of the language. But agglutinative language presents more serious problems with unknown words, unlike English. Since Sinhala is also a complex, morphologically rich and agglutinative language, information about morphology or how word is spelled is very difficult to use in unknown word prediction algorithms. So in our research, the important source of information that is used the distribution of words and parts of speech.

The distinction between closed class and open class words should help to refine the possibilities for unknown words, then syntactic knowledge can be used to aid in the analysis of unknown words sentence structure, which can be a strong clue for the possible POS tagging of an unknown word. Closed classes are those that have relatively fixed membership, they are generally function words: which tend to be very short, occur frequently, and play an important role in grammar. By contrast, open class is the type that lager numbers of words are belongs in any language, and new words are continually coined or borrowed from other languages. Hence, improvement of the algorithm could be done by incurring knowledge of distinction between closed class and open class words. Syntactic knowledge can be used to aid in the analysis of unknown words sentence structure. Then suggest corresponding POS tag and calculate the trigram using Hidden Markov Model (HMM) for unknown words. The algorithm is to pretend that each unknown word is ambiguous among all open class parts of speech tags, with equal probability. Then the tagger computes the tag sequence probability and maximum likelihood probabilities rely on text corpus and suggest the proper tag for that word.

The evaluation of the system was mainly driven by training the system using Sinhala text corpus that comprised of 2754 sentences and 90551 words annotated with corresponding POS tag. To evaluate the performance of the tagger, two gold standard test sets were created using the text corpus, one test set with only known words and the other one with unknown words. The tagger evaluated by comparing the tagged output with the Gold standard test set. The accuracy was calculated using number of correct tags proposed by the system and total number of words in the sentence/s.

The performance of the tagger is recorded in Table I, for three different versions of progressively upgraded tagging mechanisms. Version 1 is the simplest form of the tagger that performs well only with known words. Version 2 is a somewhat upgraded version that treats all unknown words as nouns and suggest NNN (Common Noun Neuter) to each new words encountered in the tagging process. Version 3 uses statistical technique to guess the best tag for unknown words by considering the context of surrounding words. Based on the performance of the Version 3, we can conclude that, considering the distinction between closed and open classes in guessing POS tags for unknown words is proved to be useful and an effective way for guessing parts of speech for unknown words in Sinhala language.

TABLE I

Approach	Tagging Approach	Performance of the Tagger
Version 1	Only with known words	91.30%
Version 2	All unknown words considered as NNN	89.73%
Version 3	Consider distinction between closed class and open class	91.50%