Abstract

In this paper, we analyze the syntactic structure of Myanmar grammatical categories to be able to use in tagging Myanmar text with standard Part-of-Speech (POS) tags. In Myanmar lexicon,
all words are annotated with basic tags and these words can be called as stem words or root words. The Myanmar POS tagged corpus creation, which has been proposed in [11], used basic POS tagging for each word. Therefore, all words in this corpus have been tagged with only basic tags as in lexicon. For standard POS tagging, normalization step is needed to form more meaningful words and annotate some words with more appropriate finer POS tags and categories. The finer tags can be called as standard POS tags and these can be used to directly concatenate with English POS tags. These tags are very useful in Myanmar to English Machine Translation System. Hence, the main aim of this study is to develop the customized lexical rules in order to deduce finer or standard POS tag from basic POS tags combinations. By analyzing Myanmar grammatical categories, 27 rules are defined to normalize them. Evaluation has been made on a basic POS tagged corpus which contains 1000 basic POS tagged sentences and it yields full satisfaction for all words in these sentences.

References


Index Terms

Computer Science Natural Language Processing
**Keywords**

| Part-of-Speech tagging | Normalization of Grammatical Categories |