A Literature Review and Discussion of Malay Rule-based Affix Elimination Algorithms

Abstrak

Stemming is one of the techniques in natural language processing that is used to reduce a word to its root. Information retrieval and knowledge management can further be improved by improving the stemming process. There are four strategies that are being used widely in stemming that includes table lookup, rule-based affix elimination, successor variety and n-gram. How-ever, not all of these strategies are being applied in Malay stemming algorithm. The well-known strategy used in stemming Malay text documents is called a rule-based affix elimination algorithm. In this paper, several Malay stemming algorithms will be discussed such as Othman's algorithm, Sembok's algorithm, Idris's algorithm, Rule Frequency Order Stemmer and Mangalam's algorithm. This paper also discusses some of the improvements made by researchers based on previous Malay stemming algorithm and this provides the current trend of Malay stemming algorithm. Different morphologies rules also being applied in different Malay stemming algorithms. Based on this review paper, it can be concluded that there are a lot of works related to the arrangement of the mor-phologies rules are conducted. However, this stemming process can still be im-proved by applying certain background knowledge such as root words diction-aries that can be used for checking the word during the process of eliminating affix words.