Syllable-based Malay Word Stemmer

JunChoi Lee[1], Rosita Mohamad Othman[2], Nurul Zawiyah Mohamad[3]
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan
Sarawak, Malaysia
[1,2,3]{Jclee, morosita, mnzawiyah}@fit.unimas.my

Abstract— **Word stemmer is one of the basic and crucial text processing tools in any languages. Word stemmer is not only useful in morphological study but also play an important role in word level context analysis. Due to the existence of prefix, suffix, infix and a combination of affixes in Malay word, it raises the complexity of performing stemming to Malay word. An approach to stem Malay word using syllabification algorithm is introduced. This approach performs stemming through comparing syllable in the word thus reduces the parsing processes. The approach shows high practicality as it produces a very high accuracy in the evaluation.**

Keywords: *Stemmer, Malay Text, Syllabification, rule-based.*

I. INTRODUCTION

In natural language, a stem is the morphological base of a word to which affixes can be attached to form derivatives. Stemming is a technique used to find root words that are conflated or reduce morphological variants of words to a single index term.

Various stemming algorithms have been developed in a wide range of languages which to be used for different purposes. Malay serves as the most common language in Malaysia, a stemming algorithm for Malay word is very essential.

Stemmer plays an important role in text understanding. A word in root form can provide fundamental meaning for the particular word. Therefore there is a need for developing a usable word stemmer in term to further in Malay text understanding related research.

Current existing stemming processes started by analyzing the character pattern in a word, concatenate the portion of the word that signifies and lastly perform a morphological transformation of the word if needed.

All words are formed through a combination of one or more syllable. Therefore it is possible to perform stemming just by analyzing the syllable pattern in word rather than analyzing the word character-by-character. This study explores the possibility to perform the Malay word stemming through syllable structure.

This paper is organized as follows: the first section described the nature Malay word stemming process. Second section discussed previous related research on Malay word stemmer. The third section provides a step-by-step explanation on the proposed stemming method. Fourth section described the evaluation process and result of the evaluation. Fifth section discussed on the limitation of the proposed method based on the evaluation result. Sixth section mentions the future work that can expanded from the current study and the final section concludes the study for this paper.

II. RELATED WORKS

Many had studies the morphological structures of Malay word [4]. However, only few had researched on the stemming technique for Malay word. At first stemmed words were outputs of morphological analysis as Abdullah [6] applied in his Malay text retrieval system. A rule-based stemming algorithm is later developed by Othman [8]. The Rules Application Order approach is introduced for Malay word stemmer by [3], [7] combined N-gram string similarity and Rules Application Order in stemming process.

[1] introduced a Malay word stemming that utilized the morphological structures of Malay word and reduced the number of rules used in stemming. [11] implemented a porter-based Malay stemmer in 2006. Abdullah introduced the Rule Frequency Order Stemmer to improve the Rules Application Order [2]. The Rule Frequency Order stemmer was then enhanced by [9] using background knowledge of the word by referencing to a dictionary containing all affixed words. [10] implemented a Malay word stemmer called UniSZA with only 7 simplified stemming rules compare to the other.

From the previous work studied, it is noticeable that all the techniques used in the Malay word stemming process were based on character pattern in the word. The pattern is obtained either through character-by-character parsing or using N-gram pattern. Which in another word the smallest analysis unit used in the stemming process is a single character of the word structure.

Syllabification process can separate a word into different syllables. Therefore it raises an interest, whether the stemming process can be done directly at the syllable level compare to