Word Syllabification for Indonesian Language using Transformer

Muhammad Haykal Kamil
School of Computing
Telkom University
Bandung, Indonesia
kamilhaykal@student.telkomuniversity.ac.id

Suyanto Suyanto
School of Computing
Telkom University
Bandung, Indonesia
suyanto@telkomuniversity.ac.id

Mochammad Arif Bijaksana
School of Computing
Telkom University
Bandung, Indonesia
arifbijaksana@telkomuniversity.ac.id

Abstract

Syllabification is a process from word to a series of syllable. Syllabification can be used in Natural Language Processing (NLP) such as speech recognition, text-to-speech, rhyme detection, and many more. Syllabification of the Indonesian language will refers to the General Guidelines for Indonesian Language Spelling (PUEBI). The corpus for this research containing the main word and the syllable. The corpus for this research is using *"50k KBBI 5-k fold"* and combined with *"103k Named Entity 5-k fold"*. The evaluation for this model is using Word Error Rate (WER). WER of previous deep learning model for syllabification is still high with 3.75% WER. The objective of this research is to lower the WER for deep learning using Transformer and Syllable Tagging because it can accept long contextual dependency. The evaluation result of this model is 3.68% WER and can be used universally for Indonesian words because the margin between Formal Words and Named Entity Words is close with the average result. Thus, this model currently the better model for Indonesian syllabification deep learning model according from the average WER is lower than the other deep learning model.

Index Terms

syllabification, indonesian language, deep learning, transformer

I. Introduction

Language is an important part of communication between humans [1]. Indonesia has an official language, namely Indonesian Language as the identity of the country and nation [2]. Language has a component, namely words. The fragments of a word are called syllables [3].

Indonesian language syllabification is the method of breaking an Indonesian word into a series of syllables from that word [3]. For example, the word *"memporosi"* (pivoting) will be broken down to <mem-po-ros-i>. This research of syllabification can be interpreted as word segmentation at the surface level [4]. In Natural Language Processing (NLP), syllabification can be used for text-audiovisual, speech recognition, text-to-speech, rhyme detection, and many more [5], [6], [8]–[10]. Indonesian syllabification automation research itself has used methods such as FkNNC [7], rule-based [11], n-gram [10], BiLSTM-CNN-CRF [8], and ASnGT [10]. This research will use Transformer and `Syllable Tag` method to build syllabification deep learning model.

This study focuses on using Indonesian and will be using the datasets *"50k KBBI 5-k fold"* and combined with *"103k Named Entity 5-k fold"* both of which are Indonesian language syllabification datasets [8]. Evaluation of this syllabus can use Word Error Rate (WER) [9]. The lower the WER values, the better the model has been made. In the rule-based method, the WER value is only listed, which is equal to 2.9% [11]. The BiLSTM-CNN-CRF method has an average WER of 2.50% for the dataset *"50k KBBI 5-k fold"* and an average WER of 5.01% for the dataset *"103k Named Entity 5-k fold"* [8].

The problem found in previous research is that the WER in the deep learning method is still high, especially in Named Entity datasets which allow words that appear to have long words. Long words create long contextual dependencies because this research model will process words per character. The goal to be achieved from this research is to build a deep learning model that accepts long contextual dependencies to reduce WER values using Transformer and `Syllable Tag` method.

II. Related Works

A. Syllabification

Syllabication is a technique of breaking words into a series of syllables that build the word with various applicable rules [3], [11]. Syllables are units that make up a word. A syllable is made up of a series of vowels and consonants. Vowels used in Indonesian include 'a', 'i', 'u', 'e', and 'o'. In Indonesian language, vowels are the core of a syllable [3]. Single consonant letters are letters that exist in the alphabet except for vowels while combined consonants consist of "kh", "ng", "ny", and "sy". Diphthong is two series of vowels that are consist of 'ai', 'au', 'ei', and 'oi'. Syllables are an important component in the pronunciation system [3]. For example, the word *"merangkai"* (stringing) has syllables: <me-rang-kai>.