Perbandingan Algoritma Sentencepiece BPE dan Unigram Pada Tokenisasi Artikel Bahasa Indonesia

Triwidyastuti Jamaluddin[1], **Moch Arif Bijaksana**[2], **Ibnu Asror**[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]triwidyastuti@student.telkomuniversity.ac.id, [2]arifbijaksana@telkomuniversity.ac.id,
[3]iasror@telkomuniversity.ac.id

Abstract

Tokenization is a concept that includes a simple process where the sequence of the text is split up into smaller parts or tokens and then entered as input into the model of natural language processing (NLP), or more complex process models such as applying the world knowledge of Deep Learning (DL). Tokenization will be more complicated when dealing with cases where all words are grouped into a single token or without separators and errors in typography. This paper proposes a model unsupervised tokenization using subword tokenizer and detokenize representation by neural networks, implementation of algorithm Byte Pair Encoding (BPE) and Unigram Language Model. Moreover, exploiting sentencePiece, the segmentation model of sentences can be trained without spaces. Experiments using the Indonesian language resulted in 54.6% and 87.0% in accuracy for Byte Pair Encoding (BPE) and Unigram Language Model, respectively.

Keywords: ***SentencePiece*****,** ***Subword Tokenizer*****,** ***Byte Pair Encoding*** **(BPE),** ***Unigram Language Model*****.**