Identifikasi Kata Majemuk Bahasa Indonesia

Fikri Haykal[1], Arie Ardiyanti Suryani[2], Sri Widowati[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]fikrihaykal@students.telkomuniversity.ac.id, [2]ardiyanti@telkomuniversity.ac.id,
[3]sriwidowati@telkomuniversity.ac.id

Abstract

Multi-word Expression (MWE) tokenizer is a machine to tokenize more than two words, which can be used to identify compound words. In this final project, the construction of the machine in Indonesian with a rule-based method based on compound word patterns using three POS tagger methods, namely, Conditional Random Fields (CRF) tagger, Bigram tagger, and Classifier Based tagger with as many as 226,328 training data. The word and test data were 1,865 words, then after testing and evaluating the results, the accuracy obtained with the CRF tagger was 77.97%, the total words obtained were 295 compound word candidate words, 230 correct words and 65 wrong words, then with Bigram The accuracy tagger obtained is 86.80%, the total words obtained are 144 compound word candidate words, 125 correct words and 19 incorrect words, and the last one using Classifier Based tagger, the accuracy obtained is 82.13%, the total words used There are 235 candidate compound words, 193 correct words and 42 incorrect words, so, if you use Bigram tagger, you get less words but the accuracy you get is high, whereas if you use the CRF tagger, you get more words but the accuracy you get is low.

Keywords: Multi-word Expression, Tokenizer, Rule Based, Conditional Random Fields Tagger, Bigram Tagger, Classifier Based Tagger