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

Abstrak

***Multi-word Expression* (MWE) tokenizer merupakan mesin untuk melakukan tokenisasi lebih dari dua kata, yang dapat digunakan untuk melakukan identifikasi kata majemuk. Pada tugas akhir ini dilakukan pembangunan mesin tersebut berbahasa Indonesia dengan metode berbasis aturan (*rule based*) berdasarkan pola kata majemuk dengan menggunakan tiga metode POS *tagger* yaitu, *Conditional Random Fields* (CRF) *tagger*, Bigram *tagger*, dan *Classifier Based tagger* dengan data latih sebanyak 226.328 kata dan data uji sebanyak 1.865 kata, lalu setelah melakukan uji coba dan evaluasi hasil, akurasi yang didapatkan dengan CRF *tagger* sebesar 77.97%, total kata yang didapat 295 kata kandidat kata majemuk, jumlah benar 230 kata dan jumlah salah 65 kata, lalu dengan Bigram *tagger* akurasi yang didapat sebesar 86,80%, total kata yang didapat sebanyak 144 kata kandidat kata majemuk, jumlah benar 125 kata dan jumlah salah 19 kata, dan yang terakhir menggunakan *Classifier Based tagger* akurasi yang didapat sebesar 82,13%, total kata yang didapat 235 kata kandidat kata majemuk, jumlah benar 193 kata dan jumlah salah 42 kata, jadi, jika menggunakan Bigram *tagger*, jumlah kata yang didapatkan sedikit tetapi akurasi yang didapatkan tinggi, sedangkan jika menggunakan CRF *tagger*, jumlah kata yang didapatkan banyak tetapi akurasi yang didapatkan rendah.**

Kata kunci: ***Multi-word Expression, Tokenizer, Rule Based, Conditional Random Fileds Tagger, Bigram Tagger, Classifier Based Tagger***