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

Abstrak

Tokenisasi merupakan sebuah konsep yang mencakup proses sederhana dimana urutan teks dipecah menjadi bagian-bagian yang lebih kecil atau token dan kemudian dimasukkan sebagai input ke dalam model Natural language processing (NLP), atau proses model yang lebih kompleks seperti menerapkan pengetahuan dunia Deep Learning (DL). Tokenisasi akan lebih rumit ketika berhadapan dengan kasus semua kata dikelompokkan menjadi satu token atau tanpa pemisah dan kesalahan dalam tipografi. Paper ini mengusulkan model unsupervised tokenization menggunakan subword unit tokenizer dan detokenizer representasi oleh neural network, implementasi algoritma Byte Pair Encoding (BPE) dan Unigram Language Model. Selain itu, mengeksploitasi sentencePiece, model segmentasi pada kalimat dapat dilatih tanpa spasi. Ekpsperimen menggunakan bahasa Indonesia menghasilkan akurasi 54.6% dan 87.0% untuk Byte Pair Encoding (BPE) dan Unigram Languange Model, masing-masing.

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