1. Pendahuluan

Latar Belakang

Tokenisasi adalah hal yang sangat penting dalam pemrosesan text seperti sentimen analisis, deteksi topik, dan spam filtering[6]. Dalam klasifikasi text, representasi kalimat dapat diperhitungkan berdasarkan token penyusun kalimat yang secara khusus, sebuah kalimat terlebih dahulu diubah menjadi unit-unit yang lebih kecil bermakna seperti karakter, kata-kata dan sub kata[11]. Kemudian token dapat dimodelkan pada neural network seperti Convolutional Neural Network (CNN) [3], atau Neural Machine Translation (NMT)[10].

Bahasa Indonesia memiliki kekayaan kosakata yang memadai sebagai sarana pikir, ekspresi, dan komunikasi di berbagai bidang kehidupan, terdiri dari sekitar 85.000 jumlah entri, 41.250 lema, 48.250 sublema, serta peribahasa sebanyak 2.036[1]. Dalam bahasa Indonesia menggunakan alfabet latin[2] sama seperti bahasa Inggris dapat di lihat dengan white space yang merupakan indikator yang baik pada segmentasi kata. Namun, tokenisasi akan menjadi masalah apabila dalam bahasa yang tidak tersegmentasi seperti pada kalimat yang tidak memiliki spasi dan kesalahan dalam tipografi. Berbagai kesalahan dalam penulisan bahasa Indonesia yang paling sering terjadi adalah penggunaan kata depan dan kata penghubung. Misalnya, kata ”di jalan” sering kali ditulis dengan kata ”dijalan” atau pada kata ”ke rumah” sering kali ditulis menjadi kata ”kerumah”. Padahal, untuk setiap keterangan yang merujuk tempat ataupun waktu, penulisannya dipisahkan dengan kata induknya.

Pada penelitian ini, kami mengeksplorasi tokenisasi yang mengacu pada pemisahan tanda baca dan membagi token menjadi kata atau subword menggunakan model *SentencePiece* sebagai unsupervised tokenizer dan detokenizer pada neural network-based text generation systems. Implementasi *SentencePiece* menggunakan dua algoritma yaitu, Byte Pair Encoding (BPE) dan Unigram Languange Model[5]. Oleh karena itu, tokenisasi yang benar dapat membantu meningkatkan kualitas pada pemrosesan text khususnya pada segmentasi.

Topik dan Batasannya

Dari latar belakang yang sudah dipaparkan, topik serta batasan yang diangkat dalam penelitian ini sebagai berikut :

1. Kamus Data

   Batasan kamus data yang diangkat berupa kumpulan kosakata berbagai artikel berita dalam bahasa Indonesia di susun menjadi sebuah dataset. Diambil dari riset Information and Language Processing Systems (ILPS) [9]. Membutuhkan setidaknya lebih dari 3 juta kata dalam dokumen agar dapat meningkatkan kualitas proses pelatihan dalam tokenisasi.

2. Input dan Output

   Inputan dari sistem berupa kalimat bahasa Indonesia menggunakan huruf kecil tanpa spasi yang akan diprediksi peluang penebakan kata yang benar berdasarkan dataset yang sudah ada. Output dari sistem berupa hasil subword tokenisasi pada kalimat. Contoh input dan output dapat dilihat pada tabel 1:

Tabel 1. Contoh Input dan Output Sistem

Input Kata	Gold Standar	Output
pemerintaharusbertindak	[’pemerintah’,’harus’,’bertindak’]	’pemerintah’,’harus’,’bertindak’
penanganansangatlambat	[’penanganan’,’sangat’,’lambat’]	’penanganan’,’sangat’,’lambat
sedangmelakukanrapat	[’sedang’,’melakukan’,’rapat’]	’sedang’, ’melakukan’,’rapat’

3. Rencana pengujian Sistem

   Proses pengujian menggunakan model implementasi SentencePiece berbasis neural network menggunakan algoritma Byte Pair Encoding (BPE) dan Unigram Language Model pada bahasa Indonesia. Adapun hasil evaluasi menggunakan akurasi measure diantaranya adalah Precision, Recall, F-Measure untuk mencapai hasil yang optimal.

Tujuan

Adapun tujuan dari tugas akhir ini adalah untuk mengimplementasikan model SentencePiece pada bahasa Indonesia. Menggunakan sebuah algoritma Byte Pair Encoding (BPE) dan Unigram Language Model pada kalimat tanpa spasi kemudian memisahkan tanda baca dan mambagi token menjadi unit subword untuk mendapatkan hasil tokenisasi yang optimal. Mengevaluasi algoritma BPE dan Unigram LM dalam pelatihan ini.

Organisasi Tulisan

Susunan dari tulisan ini adalah sebagai berikut : Bagian 2 membahas tentang teori, studi dan literatur terkait dengan tulisan ini. Bagian 3 akan dijelaskan sistem yang akan dibangun. Bagian 4 akan terdapat hasil dan evaluasi dari sistem yang di bangun. Dan Bagian 5 akan dijelaskan kesimpulan.