BAB I
PENDAHULUAN

1.1 Latar Belakang

Bahasa Indonesia adalah bahasa resmi negara Indonesa yang secara luas digunakan sebagai alat komunikasi sehari-hari oleh lebih dari 222 juta orang. Dengan lebih dari 742 bahasa daerah yang berbeda, Bahasa Indonesia merupakan bahasa pemersatu bagi penduduk Indonesia [1]. Sehingga memiliki peralatan untuk penelitian *Natural Language Processing* (NLP) yang tersedia untuk masyarakat luas menjadi penting.

Label *Part-of-Speech* (POS) adalah label kategori kelas kata yang berupa kata kerja (*verb*), kata benda (*noun*), kata sifat (*adjectives*), kata keterangan (*adverb*) dan seterusnya pada tiap kata dalam suatu kalimat. POS *tagging* (pelabelan kelas kata) merupakan salah satu bagian yang sangat penting dalam aplikasi NLP seperti *Speech Recognition*, *Question Answering* dan *Informarion Retrieval*. Melakukan palabelan POS secara manual membutuhkan waktu yang lama dan biaya yang mahal karena harus memerlukan ahli bahasa. Oleh karena itu mengembangkan POS *tagging* secara otomatis merupakan kebutuhan yang mendesak.

POS *tagging* telah secara luas dipelajari dan dikembangkan untuk bahasa Indonesia. Beberapa pendekatan telah digunakan untuk mengembangkan POS *tagging* diantaranya adalah pendekatan *Statistic-Based* [2,3,4], pendekatan *Rule-Based* [5] dan pendekatan *Transformasion-Based learning* [6]. Salah satu metode POS *tagging* yang telah dikembangkan dan menghasilkan keakuratan yang tinggi adalah POS *tagging* dengan pendekatan berbasis statistik (*Statistic-Based*) menggunakan metode *Hidden Markov Model* (HMM) dikombinasikan dengan metode lain [7]. HMM sendiri merupakan pengembangan dari Markov Model yang mengasumsikan bahwa kata secara probabilitas bergantung hanya pada

kategori POS dua kata sebelumnya.

Salah satu POS *tagging* untuk bahasa Indonesia yang menerapkan pendekatan statistik dengan metode HMM adalah IPOSTAgger, yang dikembangkan oleh Wicaksono dan Purwarianti [3]. Dengan menggunakan HMM bigram (urutan pertama) dan HMM trigram (urutan kedua) sebagai model dasar IPOSTAgger juga menerapkan beberapa metode lain seperti Jelinek-Mercer s*moothing,* Affix Tree (pohon prefix – suffix)*,* Lexicon (kamus) dari KBBI-Kateglo dan Succeding POS tag. Metode-metode tersebut diuji untuk mengetahui konfigurasi yang menghasilkan nilai keakuratan terbaik. Adapun konfigurasi terbaik yang didapatkan kombinasi metode HMM trigram, Affix tree dan Lexicon.

Masalah utama dalam POS *tagging* antara lain kata ambigu dan kata *Out-of-Vocabulary* (OOV) [8]. Kata ambigu merupakan kata yang memiliki sifat berbeda jika ditempakan pada konteks yang berbeda. Sedangkan kata OOV merupakan kata yang ada dalam korpus uji namun tidak ada dalam korpus latih, hal ini akan menyebabkan masalah *sparse* data. Sistem morfologi bahasa Indonesia cukup rumit, termasuk diantaranya afiksasi yang menjadi salah satu sumber dari masalah kata OOV. Bahasa Indonesia menggunakan banyak kata imbuhan untuk membuat kata jadian. Penggunaan prefik, sufiks, infiks, atau kombinasinya dapat merubah label POS dan makna dari suatu kata. Kata kerja dapat menjadi kata benda, kata keterangan maupun kata sifat. Salah satu bagian dari afiksasi adalah pengklitikan. Fenomena pengklitikan (proklitik, enklitik) sangat sering terjadi dalam bahasa sehari-hari. Kata berimbuhan yang ketambahan klitik akan menjadi kata ambigu, contohnya kata *kumengirimkanmu* (kata benda nama diri) yang terdiri dari kata *mengirimkan* (kata kerja transitif) ketambahan proklitik *ku* (kata benda) dan enklitik *mu* (kata benda). Sehingga diperlukannya proses pengolahan berupa pemotongan pada kata berklitik. Penentuan kombinasi afiks ataupun pemotongan klitik memerlukan analisis morfologi terlebih dahulu sehingga tidak menimbulkan kesalahan pemberian label POS atau kesalahan pemotongan klitik yang akan mengurangi tingkat keakuratan POS *tagger*.

Penerapan analisis morfologi dapat membantu pemberian kategori kelas kata karena dapat diketahui unsur-unsur pembuat kata tersebut. Salah satu metode

analisis morfologi adalah Affix tree yang diadaptasi oleh IPOSTAgger. Namun Affix tree hanya dapat melakukan pencocokan pola tidak memberikan informasi morfologi lebih jauh. Sistem yang menerapkan analisis morfologi untuk bahasa Indonesia yang dapat menangani afiksasi dan pengklitikan salah satunya adalah penganalisis morfologi (Morphology Analyzer) MorphInd [9].

Penelitian ini menerapkan analisis morfologi pada POS *tagging* untuk mengatasi masalah kata ambigu dan kata OOV yang banyak disebabkan oleh kata berimbuhan. Sistem MorphInd diterapkan untuk pelabelan POS pada kata OOV dan membantu pemotongan klitik pada tahap *preprocessing*. Selanjutnya tahap HMM *tagging* dilakukan menggunakan IPOSTAgger, konfigurasi model terbaik akan dibandingkan untuk mengetahui nilai keakuratan tertinggi pada korpus uji yang telah disiapkan.

Berdasarkan analisis tersebut, penelitian ini mengambil judul "Penerapan analisis morfologi untuk penanganan kata berimbuhan pada POS *Tagger* bahasa Indonesia berbasis statistik".

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang, masalah yang ada pada POS *tagging* seperti penanganan kata OOV dan kata ambigu sangat penting untuk meningkatkan keakuratan pada model POS *tagging*. Salah satu masalah dalam POS *tagging* bahasa Indonesia adalah kata imbuhan. Hal ini dapat diatasi dengan menerapkan analisis morfologi pada POS *tagging*. Dengan demikian rumusan dari masalah tersebut adalah, "Bagaimana menerapkan analisis morfologi untuk penanganan kata berimbuhan dalam POS *tagging* bahasa Indonesia berbasis statistik".

1.3 Batasan Masalah

Untuk menyelesaikan permasalahan yang ada, diperlukan adanya batasan yang dapat mencakup kajian yang berhubungan dengan masalah tersebut sehingga

penyelesaian tidak menyimpang dari masalah. Adapun batasannya adalah sebagai berikut:

1. Penelitian diterapkan pada sistem POS *tagging* IPOSTAgger.
2. Penelitian menggunaan MorphInd untuk penganalisis morfologi bahasa Indonesia.
3. Korpus latih dan korpus uji menggunakan bahasa Indonesia dan merupakan korpus yang telah ditentukan.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Mengimplementasikan analisis morfologi pada POS *tagging* berbasis statistik.
2. Menentukan konfigurasi model POS *tagging* dengan penerapan analisis morfologi yang dapat meningkatkan keakuratan POS *tagger*.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat bermanfaat bagi:

1. Universitas Dian Nuswantoro
   (a) Penelitian ini dapat menjadi tinjauan pustaka baru untuk studi pada bidang NLP khususnya POS *tagging*.
2. Masyarakat Umum
   (a) Dapat diterapkan untuk alat *preprocessing* pada aplikasi-aplikasi NLP.
   (b) Dapat dimanfaatkan kembali untuk penelitian tentang analisis morfologi dalam komputasi linguistik, POS *tagging* ataupun penelitian NLP selanjutnya.