BAB 1
PENDAHULUAN

Bagian ini akan berfungsi sebagai pengantar tentang penelitian yang dilakukan penulis. Bab ini mencakup penjelasan tentang latar belakang pelaksanaan penelitian yang dilakukan penulis, permasalahan yang ingin diatasi, dan tujuan pelaksanaan Tugas Akhir tersebut. Termasuk di dalam bab ini juga metodologi yang dilakukan penulis dalam melakukan penelitian dan sistematika penulisan Laporan Tugas Akhir ini.

1.1. Latar Belakang

Bahasa telah menjadi medium berkomunikasi yang digunakan manusia semenjak manusia memasuki masa sejarah. Dengan bahasa, baik secara lisan maupun tulisan, manusia dapat menyampaikan berbagai hal satu sama lainnya. Sedemikian pentingnya peranan bahasa dalam kehidupan manusia sehingga telah banyak penelitian yang mengangkat bahasa sebagai subyeknya. Salah satu contoh yang paling sederhana adalah penelitian yang telah dilakukan manusia untuk mengelompokkan kata-kata dalam bahasa yang mereka gunakan ke dalam kelas-kelas kata. Kelas-kelas kata ini kemudian dinamakan *part of speech* (POS).

Part of speech (POS) adalah kategori linguistik dari kata yang didefinisikan berdasarkan sifat sintaktis, semantis, dan morfologis dari kata tersebut. Konsep POS pertama kali muncul pada antara abad ke-5 dan ke-6 sebelum Masehi saat Yaska, seorang ahli bahasa pada masa itu, mengelompokkan kata-kata dalam bahasa Sansekerta ke dalam empat kelas kata, yaitu kata benda, kata kerja, prefiks, dan partikel. Setelah terobosan ini, semakin banyak orang kemudian turut berusaha mengelompokkan kata-kata dalam bahasa mereka ke dalam kelas-kelas kata. Sebagai contohnya adalah Plato dan Aristoteles yang menyusun POS untuk bahasa Yunani.

Mengingat terdapat berbagai bahasa di dunia dan bahwa tiap bahasa tersebut memiliki sifat yang berbeda, maka dapat dimengerti bahwa terdapat banyak cara mengkategorikan kata. Sebagai contoh, dalam POS bahasa Inggris, terdapat beberapa jenis kata kerja yang dikategorikan berdasarkan *tense* sementara dalam Bahasa Indonesia, pengkategorian kata kerja semacam ini tidak ditemukan. Contoh lain adalah bahasa Jepang memiliki tiga kelas kata sifat sementara Bahasa Inggris hanya memiliki satu kelas kata sifat. Hal ini menyebabkan POS yang dianut suatu bahasa bisa jadi berbeda dengan bahasa lainnya. Bahkan mungkin terjadi adanya beberapa versi POS untuk satu bahasa, tergantung cara pandang yang dianut saat seseorang menyusun POS untuk bahasa tersebut.

Kegiatan pengelompokkan kata seperti telah dijelaskan di atas akan menghasilkan kelas-kelas kata yang masing-masing berisi kata-kata dengan sifat yang sama. Untuk memudahkan dalam membedakan kelas kata yang satu dengan yang lainnya, maka manusia pun memberi label untuk masing-masing kelas kata. Label yang diberikan diharapkan dapat menggambarkan sifat kata-kata yang terdapat dalam kelas kata yang bersangkutan. Sebagai contoh, kata dengan label 'NOUN' dalam bahasa Inggris menandakan kata tersebut termasuk ke dalam kelas kata 'kata benda'. Label ini kemudian disebut sebagai part *of speech tag* (POS *tag*).

Part of speech tag (POS *tag*) merupakan label (*tag*) yang diberikan pada tiap kata yang menyatakan POS dari kata tersebut. Dengan adanya POS *tag*, kita dapat mengetahui kelas dari suatu kata, termasuk juga sifat-sifat apa yang cenderung melekat pada kata dalam kelas tersebut. Informasi ini dapat menolong kita dalam menentukan makna suatu kata, menentukan aturan tata bahasa suatu kalimat, dan sebagainya. POS *tag* juga berperan besar dalam proses mengatasi keambiguan makna kata ataupun kalimat.

Proses memberikan label pada kata ini disebut *part of speech tagging* (POS *tagging*). Seperti telah dijelaskan di atas, keberadaan POS *tag* dapat sangat membantu manusia dalam hal pemrosesan bahasa sehingga POS *tagging* menjadi aktivitas yang penting sifatnya.

Secara tradisional, POS *tagging* dilakukan dengan cara manual, yaitu dengan bantuan satu atau beberapa linguis atau ahli bahasa memberikan *tag* yang bersesuaian untuk tiap kata pada suatu teks atau korpus. Namun pekerjaan ini sangat memakan waktu, tenaga, dan biaya. Hal ini menyebabkan timbulnya kebutuhan untuk adanya suatu aplikasi yang dapat melakukan POS *tagging* secara otomatis.

Perkembangan pemrosesan bahasa natural dalam kurun waktu belakangan ini juga menjadi pertimbangan mengapa aplikasi pelabelan kelas kata secara otomatis (*automatic part of speech tagger* atau POS *tagger* otomatis) menjadi kebutuhan yang mendesak. POS *tag* dapat menjadi dasar untuk melakukan Parsing, menolong dalam pemrosesan kueri dalam aplikasi *Question Answering*, dan memberikan bantuan pada berbagai aplikasi bidang pemrosesan bahasa natural lainnya. Sehingga aplikasi yang dapat melakukan POS *tagging* secara otomatis untuk setiap teks yang diberikan kepadanya, dengan catatan teks tersebut masih menggunakan kata-kata dengan atribut linguistik yang sama dengan yang dianut POS *tagger* tersebut, tentu akan sangat memberikan bantuan.

Berangkat dari hal ini, sudah banyak penelitian dilakukan dalam rangka pembuatan POS *tagger* yang dapat melakukan *tagging* dengan tingkat kesahahan yang relatif kecil. Namun hal ini tidaklah mudah. Manusia pun seringkali kesulitan dalam memutuskan apakah suatu kata termasuk dalam kelas kata yang satu atau yang lainnya, sehingga bagaimanakah caranya manusia dapat 'memerintahkan' mesin untuk secara tepat mengelompokkan kata-kata ke dalam kelas kata yang bersesuaian? Pertanyaan ini telah menjadi pendorong para peneliti selama bertahun-tahun untuk mencari sebuah cara yang mampu mengenali dan memberikan POS *tag* yang tepat untuk seiap kata yang diberkan kepadanya.

Berbagai metode kemudian diciptakan oleh para peneliti yang pada intinya bertujuan untuk menciptakan suatu POS *tagger* dengan tingkat akurasi lebih tinggi daripada yang dihasilkan metode lain sebelumnya. Pada awalnya, metode

yang paling banyak digunakan adalah metode Rule-Based, di mana para ahli bahasa secara manual menyusun kumpulan aturan bahasa yang kemudian disandikan ke dalam bahasa mesin. Namun metode ini masih dirasa kurang efisien karena masih membutuhkan tenaga dan biaya yang besar dalam proses penyusunan aturan. Terlebih lagi, manusa kesulitan untuk dapat menerjemahkan seluruh kemungkinan aturan tata bahasa yang ada ke dalam bahasa mesin sehingga POS *tagger* yang dibanngun dengan metode ini cenderung memiliki tingkat kesalahan lebih tinggi dibandingkan POS *tagger* yang dibangun dengan metode lainnya.

Berangkat dari hal ini, kemudian metode-metode statistik pun diciptakan, yang pada intinya menganut konsep probabilitas dalam menentukan POS *tag* dari suatu kata. Intinya, mesin diminta untuk memberi POS *tag* dari suatu kata berdasarkan suatu nilai probabilitas yang didapat secara otomatis dari proses pelatihan yang dilakukan mesin berdasarkan suatu teks berisi kata-kata yang sudah diberi POS *tag* (konsep *machine learning*). Metode-metode yang termasuk di alam kelompok ini di antaranya adalah metode Hidden Markov Model, metode Maximum Entropy Markov Model, dan metode Conditional Random Field. Metode statistik ini terbukti memberikan hasil yang baik jika disediakan cukup banyak teks untuk dijadikan bahan pelatihan bagi mesin. Namun para ahli bahasa merasa pendekatan dengan metode ini tidak sesuai dengan hakikat POS yang adalah atribut linguistik dari suatu kata. Metode statistik memang menghasilkan akurasi yang tinggi namun metode ini tidak dapat menjelaskan aturan apa yang dipakai shingga suatu kata diberikan suatu POS *tag* tertentu. Yang dapat diberikan oleh metode ini hanyalah suatu tabel statistik besar yang tidak menjelaskan sama sekali tentang konsep linguistik yang dipakai dalam poses *tagging*.

Metode Transformation Based Learning (TBL) kemudian diciptakan untuk dapat mengakomodasi baik kalangan ahli bahasa maupun komunitas ilmu pasti mengenai POS *tagging*. Metode ini juga menganut konsep ‘belajar’ dari teks. Teks berisi kata-kata yang telah diberi POS *tag* menjadi bahan pelatihan bagi mesin. Namun, berbeda dengan metode statistik yang membangun sebuah model

probabilistik besar berdasarkan teks pelatihan tersebut, TBL menciptakan suatu kumpulan aturan yang dapat secara mudah dibaca dan dimengerti manusia.

Berbagai aplikasi POS *tagger* pun sudah diciptakan demi kepentingan riset, ilmu pengetahuan, ataupun komersil. Beberapa di antaranya adalah TAGGIT oleh B. B. Greene dan G. M. Rubin [Greene & Rubin, 1971], POS *tagger* yang pertama kali dibangun dan mengimplementasikan teknik *Rule-Based*, Xerox, POS *tagger* yang dikembangkan oleh J. M. Kupiec [Kupiec et al., 1992] dengan teknik stokastik *Hidden Markov Model*, ataupun Brill *tagger*, POS *tagger* yang dikembangkan oleh Eric Brill dengan teknik *Transformation Based Learning* [Brill, 1992].

Aplikasi POS *tagger* pun sudah banyak dimodifikasi untuk digunakan pada bahasa selain bahasa Inggris. Contohnya adalah penggunaan metode dan aplikasi CRF dan TBL untuk aplikasi POS *tagger* bahasa Hindi, Telugu, dan Bengali [Avinesh & Karthik, 2007], penggunaan metode HMM untuk aplikasi POS *tagger* bahasa Perancis [Chanod & Tapanainen, 1995] dan bahasa Jerman [Brants, 2008], dan sebagainya. Namun di antara sekian banyaknya aplikasi yang ada, belum ada suatu aplikasi POS *tagger* yang bekerja untuk Bahasa Indonesia. Ketiadaan aplikasi POS *tagger* ini tentu menjadi penghambat untuk diciptakannya berbagai aplikasi pemrosesan bahasa natural yang lain.

Hal-hal seperti disebutkan di atas melatarbelakangi penulis untuk melakukan penelitian di bidang POS *tagging* untuk Bahasa Indonesia. Banyak aplikasi pemanfaatan Bahasa Indonesia yang dapat dibangun dengan adanya POS *tagger* untuk Bahasa Indonesia. Karena itulah penulis melakukan penelitian ini untuk menghasilkan suatu POS *tagger* yang dapat bekerja pada Bahasa Indonesia.

1.2. Permasalahan

Seperti telah disinggung pada subbab 1.1 Latar Belakang, belum ada aplikasi POS *tagger* untuk Bahasa Indonesia. Pemanfaatan metode-metode POS *tag* yang sudah ada pun tidak dapat langsung dilakukan terhadap Bahasa Indonesia. Bahasa

Indonesia memiliki berbagai kekhususan yang membedakannya dibandingkan dengan bahasa lain, misal bahasa Inggris. Sehingga aplikasi POS *tagger* yang sudah ada, yang sebagian besar memang dibangun dengan basis bahasa Inggris, tidak dapat langsung diaplikasikan pada Bahasa Indonesia tanpa modifikasi.

Belum adanya penelitian yang cukup banyak di bidang POS *tag* Bahasa Indonesia juga menjadi permasalahan tersendiri. Kurangnya penelitian ini menyebabkan penulis belum dapat memutuskan metode apakah yang paling cocok untuk diterapkan dalam Bahasa Indonesia. Kecocokan suatu metode untuk diterapkan pada suatu bahasa tentu sangat tergantung pada sifat dari bahasa tersebut. Namun, sekali lagi hal ini menjadi masalah karena memang belum ada definisi formal yang menerangkan mengenai sifat bahasa Indonesia. Belum ada penjelasan mengenai bagaimana ketergantungan kata-kata dalam kalimat Bahasa Indonesia, bagaimana struktur kalimat yang baku dalam Bahasa Indonesia, dan bahkan POS *tag* untuk Bahasa Indonesia yang baku dan diterima secara umum pun belum ada.

Berdasarkan penjelasan di atas, penulis kemudian membuat perumusan masalah yang berusaha untuk diatasi dengan penelitian ini. Permasalahan yang ingin diangkat penulis dalam penelitian ini dapat dijabarkan sebagai berikut:

1. Belum adanya suatu aplikasi *part of speech tagger* untuk Bahasa Indonesia karena aplikasi-aplikasi yang ada masih harus disesuaikan untuk dapat bekerja untuk Bahasa Indonesia.
2. Belum diketahuinya metode apa yang paling cocok untuk digunakan dalam pembuatan *part of speech tagger* Bahasa Indonesia.
3. Belum diketahuinya sifat ketergantungan struktural kata dalam kalimat Bahasa Indonesia.

1.3. Tujuan

Tujuan pelaksanaan penelitian yang dilakukan penulis secara garis besar adalah untuk menghasilkan suatu POS *tagger* untuk Bahasa Indonesia. Bersamaan juga dengan hal itu, penelitian ini juga bertujuan mengetahui konsep ketergantungan

struktural antar kata dalam Bahasa Indonesia dan fitur apa yang paling menentukan jenis POS *tag* dalam kata-kata Bahasa Indonesia. Kedua hal ini nantinya dapat dimanfaatkan dalam penelitian lebih lanjut di bidang POS *tagging* Bahasa Indonesia.

1.4. Metodologi Penelitian

Penelitian dilakukan penulis dengan menjalankan eksperimen. Hasil dari eksperimen kemudian diperbandingkan dan dianalisis untuk dapat menjawab permasalahan seperti telah dijelaskan pada subbab I.2. Permasalahan. Eksperimen ini dilakukan dengan satu mesin standar terhadap suatu korpus Bahasa Indonesia. Eksperimen juga dilakukan dengan melibatkan berbagai fitur dan berbagai proporsi pelatihan-pengujian.

1.5. Sistematika Penulisan

Laporan ini terdiri atas 5 bab dengan isi dari masing-masing bab adalah sebagai berikut:

1. Bab I Pendahuluan

   Bab ini akan menjadi pengantar. Dalam bab ini dibahas Latar Belakang, Permasalahan, Tujuan, Metodologi Penelitian, dan Sistematika Penulisan.
2. Bab II Landasan Teori

   Bab ini menjelaskan teori-teori yang diimplementasikan dalam penelitian ini. Teori yang dibahas termasuk teori mengenai *part of speech tag* secara general maupun teknik-teknik *part of speech tagging* yang ada. Turut dibahas juga pada bab ini hasil-hasil penelitian yang pernah dilakukan di bidang POS *tagging* sebagai bahan perbandingan bagi pembaca.
3. Bab III Implementasi

   Bab ini menjelaskan penelitian yang dilakukan penulis, apa saja yang dikerjakan, dengan aplikasi apa penelitian dikerjakan, bagaimana cara pengerjaannya, dan kesulitan apa yang dihadapi.
4. Bab IV Analisis

Bab ini memaparkan hasil dari tiap penelitian yang dilakukan penulis dan analisa penulis mengapa penelitian memberikan hasil tersebut.

5. Bab V Penutup

   Bab ini memuat kesimpulan yang dapat ditarik penulis berikut saran dari penulis terkait penelitian lebih lanjut di bidang ini.