IMPLEMENTASI *BRILL TAGGER* UNTUK MEMBERIKAN POS-*TAGGING* PADA DOKUMEN BAHASA INDONESIA

(The implementation of Brill Tagger for Post-tagging on Indonesian Documents)

Viny Christanti M.*, Jeanny Pragantha, Endah Purnamasari

Fakultas Teknologi Informasi Jurusan Teknik In*form*atika
Universitas Tarumanagara – Jakarta
***viny@untar.ac.id**

Abstrak

Part-of-speech tagging atau POS-*Tagging* merupakan kegiatan pemberian label kelas kata pada suatu kata sehingga akan diketahui keterangan dari masing-masing kata. Tujuan dari perancangan program aplikasi ini adalah untuk merancang sebuah sistem yang dapat membantu proses POS-*Tagging* terhadap dokumen Bahasa Indonesia dengan mengimplementasikan program *Brill Tagger*. Hasil yang diperoleh menunjukkan bahwa dari 11.411 kata (20 dokumen berita) yang digunakan pada proses *testing*, sebanyak 154 kata mengalami pemberian *tag*ging yang tidak tepat sedangkan 11.257 kata diberi label tepat sesuai dengan kelas kata yang seharusnya. Hal ini menunjukkan bahwa program aplikasi POS-*Tagging* dengan implementasi *Brill Tagger* mencapai nilai keakuratan 98,65%. Akurasi meningkat menjadi 99.75% setelah mengalami penyesuaian aturan leksikal dan kontekstual.

Kata Kunci: *Brill Tagger,* pemrosesan bahasa alami, *part-of-speech tagging, rule based, transformation based learning*

Abstract

Part-of-speech (POS) tagging is the process of marking up a word in a text. The aim of this program application is to design a system that is able to proceed POS-Tagging in Indonesian documents by implementing Brill Tagger program. The results showed that of 11.411 words (20 news) used in testing process, 154 words underwent incorrect tagging and 11.257 words were properly labeled according to their part of speech. This indicated that the accuracy of POS-Tagging application program which implemented Brill Tagger Program was 98.65%. The accuracy became 99.75 % after being adapted with lexical and contextual rules.

Keywords: *Brill Tagger, natural language processing, part-of-speech tagging, rule based, transformation based learning*

1. PENDAHULUAN

Seiring dengan perkembangan teknologi, setiap orang dituntut untuk dapat memanfaatkan perkembangan tersebut dalam kehidupan sehari-hari. Perkembangan teknologi mencakup semua aspek kehidupan, salah satunya dalam bidang bahasa. Bahasa memiliki peranan yang sangat penting dalam pertukaran informasi dan atau menerima informasi. Membaca adalah salah satu cara untuk pertukaran informasi. Dalam membaca diperlukan pengetahuan tentang tata bahasa untuk mendapatkan informasi yang tepat. Salah satu pemanfaatan teknologi dalam bidang bahasa adalah adanya program POS

(*Part-Of-Speech*) *Tagging*. *Part-of-speech tagging* adalah kegiatan pemberian label kelas kata pada suatu kata [1]. Seperti *saya makan nasi* menjadi *saya/KG makan/V nasi/NN*. Dimana label KG = kata ganti, V = kata kerja, NN = kata benda. Pemberian *tag* kata pada dokumen atau kalimat dapat bermanfaat untuk berbagai hal, antara lain *information retrieval, language generator, information extraction, question answering, speech recognition, intelligent tutoring system, parser, summarization* dan *machine translation* [2].

Beberapa perancangan sistem pemberian POS-*Tagging* untuk bahasa Indonesia yang sudah dilakukan adalah *Probabilistic Part-of-speech tagging for* Bahasa Indonesia oleh Femphy Pisceldo, Mirna Adriani, dan Ruli Manurung dengan metode *Conditional Random Fields* (CRF) dan *Maximum Entropy*. Hasil akurasi dari sistem ini adalah Indonesia sebesar 95,19% dengan menggunakan 37 *tagset* [3]. Pada tahun 2004 pengembangan *part of speech tagger* untuk Bahasa Indonesia dilakukan berdasarkan metode *Conditional Random Fields* (CRF) dan *Transformation Based Learning* oleh Triastuti Chandrawati [2]. Sementara itu, Freddy Kurniawan dkk, menerapkan program *Stanford* POS *Tagger* untuk melakukan *Tagging* pada kalimat berbahasa Indonesia dengan menggunakan 29 jenis *tag* (29 jenis kelas kata) [4]. Hasil keakuratan *tagging* bahasa Indonesia dengan *Stanford* POS *Tagger* sebesar 80%. Pada tahun 2010, Alfan Farizki Wicaksono dan Ayu Purwarianti menggunakan metode *Hidden Markov Model* (HMM) [5]. Namun hasil dari setiap penelitian tersebut masih memiliki ketidaktepatan dalam pemberian kelas kata dan sulit untuk dilatih kembali karena tidak dilengkapi dengan tampilan antarmuka.

Part-of-speech tagging untuk setiap bahasa memiliki aturan yang berbeda, misalnya antara bahasa Indonesia dan bahasa Inggris. Aplikasi POS-*Tagging* yang sudah ada sebagian besar dibangun dengan bahasa Inggris dan tidak dapat langsung diaplikasikan untuk bahasa Indonesia [6]. *Brill Tagger* adalah salah satu program pemberian kelas kata yang berdasarkan pada peraturan (*Rule Based*) dan transformasi, yaitu metode yang menggunakan aturan berupa tata bahasa yang diinduksi langsung dari pelatihan *corpus* tanpa adanya campur tangan manusia atau ahli pengetahuan [7]. *Brill Tagger* sudah dilatih untuk berbagai macam bahasa, seperti Belanda, Norwegia, Denmark, Inggris, dan lain sebagainya [8].

Berdasarkan hasil penelitian yang telah ada, tingkat keakuratan program *Brill Tagger* untuk bahasa Hungaria adalah sebesar 98,8% [7]. Hal ini menunjukkan bahwa *Brill Tagger* dapat menghasilkan tingkat akurasi yang cukup tinggi untuk bahasa selain bahasa Inggris. Keunggulan dari program *Brill Tagger* adalah aturan leksikal dan kontekstual yang akan digunakan dapat disesuaikan dengan bahasa yang akan di-*tagging* [9].

Beberapa keunggulan dan hasil akurasi yang cukup baik dalam proses pemberian kelas kata dari *Brill Tagger* merupakan dasar dalam implementasi *Brill Tagger* untuk bahasa Indonesia. Permasalahan yang muncul dalam perancangan ini adalah bagaimana sistem mengimplementasikan POS-*Tagging* dengan *Brill Tagger* untuk mendapatkan hasil *tagging* yang benar pada dokumen bahasa Indonesia, bagaimana sistem akan melakukan proses pelatihan dokumen, dan bagaimana sistem akan menerapkan aturan leksikal dan kontekstual pada dokumen.

Tujuan implementasi *Brill Tagger* untuk bahasa Indonesia ini adalah merancang sebuah sistem untuk melakukan proses *tagging* bahasa Indonesia berdasarkan algoritma *Brill Tagger*, mempermudah proses pemberian POS-*Tagging* terhadap kata, menemukan *tag* yang tepat dari suatu kata dan memperbaiki hasil aturan leksikal dan kontekstual dari *Brill Tagger* agar sesuai dengan aturan bahasa Indonesia. Perlu dilakukan penyesuaian pada setiap penggunaan *Brill Tagger* untuk bahasa yang berbeda sehingga pada implementasi ini dibutuhkan pelatihan dan penyesuaian aturan yang dihasilkan dari *Brill Tagger*. Kegunaan dari perancangan ini adalah untuk menciptakan sebuah program yang mengimplementasikan program *Brill Tagger* untuk melakukan POS-*tagging* yang tepat

untuk bahasa Indonesia dengan menerapkan aturan leksikal dan kontekstual.

Perbedaan pada perancangan ini adalah dilakukannya peninjauan ulang terhadap hasil *output* dari proses *Brill Tagger* secara keseluruhan. Peninjauan ulang ini dilakukan untuk melihat apakah aturan yang diperoleh dari *Brill Tagger* sudah sesuai dengan tata bahasa Indonesia. Apabila belum sesuai maka dilakukan perubahan agar menghasilkan hasil yang lebih sesuai dengan isi dari bahasa Indonesia.

2. *PART-OF-SPEECH TAGGING*

POS (*Part-Of-Speech*) *Tagging* adalah proses memberi label pada setiap kata dalam kalimat dengan POS atau *tag* yang sesuai untuk kata tersebut [10]. *Tagging* dapat dimanfaatkan pada aplikasi bahasa alami lainnya, seperti sistem tanya jawab, informasi ekstraksi. Beberapa penggunaan POS-*Tagging* adalah untuk menghapus perbedaan yang tidak relevan, menghapus ambiguitas, membantu *stemming*, dan membantu pencarian kata benda. POS-*Tagging* dapat dilakukan secara manual maupun otomatis. POS-*Tagging* dilakukan secara manual dengan menggunakan bantuan satu atau beberapa ahli bahasa yang memberikan *tag* yang bersesuaian untuk tiap kata pada suatu teks atau korpus [2]. POS-*Tagging* secara otomatis dilakukan dengan menggunakan metode matematika atau lainnya.

Beberapa metode yang digunakan dalam POS-*Tagging* adalah metode statistik yang terdiri dari *Hidden Markov Model*, *Maximum Entropy,* dan *Conditional Random Field*. Metode *Rule Based* adalah metode yang secara manual menyusun kumpulan aturan bahasa yang kemudian disandikan ke dalam bahasa mesin. Namun, metode ini kurang efisien karena membutuhkan tenaga dan biaya yang besar dalam proses penyusunan aturan. Metode *Transformation Based Learning* adalah metode yang menganut konsep belajar dari teks. Teks berisi kata-kata yang telah diberi *tagging* secara manual menjadi bahan pelatihan bagi mesin.

Berbeda dengan metode statistik yang membangun sebuah model probabilistik, *Transformation Based Learning* menciptakan suatu kumpulan aturan yang dapat dengan mudah dibaca dan dimengerti oleh manusia [2]. Tingkat keakuratan sebuah *tagger* dipengaruhi oleh beberapa faktor, yaitu jumlah data yang digunakan saat *training*, jumlah *tagset* yang digunakan lebih besar, perbedaan antara *corpus* (teks) yang digunakan pada saat *training* dengan saat menggunakan aplikasi, serta jumlah *unknown words* (kata yang tidak dikenali) [3].

Beberapa POS *tagset* dalam bahasa Inggris, diantaranya *Penn Treebank tagset* yang hanya membedakan kata-kata ke dalam 45 jenis *tag*, *Brown Corpus tagset* menggunakan 87 jenis *tag*, dan *Lancaster* C7 menggunakan 145 jenis *tag*. Ada beberapa jenis *tagset* untuk bahasa Indonesia, yaitu 35 jenis *tag*, 21 jenis *tag*, 37 jenis *tag*, dan 29 jenis *tag*. Perbedaan penggunaan *tagset* untuk bahasa Indonesia dan bahasa Inggris dikarenakan tidak semua *tagset* dapat diimplementasikan dari bahasa Inggris ke bahasa Indonesia.

2.1 *Brill Tagger*

Brill Tagger merupakan program *tagging* yang diperkenalkan oleh Eric Brill pada tahun 1992. Program ini didasarkan pada peraturan atau transformasi, yaitu tata bahasa yang diinduksi langsung dari pelatihan *corpus* tanpa adanya campur tangan manusia atau ahli pengetahuan [10]. *Brill Tagger* melakukan pemberian anotasi pada *corpus* dengan tiga cara, yaitu *lexicon* yang merupakan daftar semua kata dan jenis *tag*-nya, *lexical* yang dilakukan dengan melihat morfologi (perubahan-perubahan bentuk kata) dari setiap jenis kata, dan *contextual* yang dilakukan dengan melihat konteks dari kata (dua kata sebelum dan dua kata sesudah kata tersebut). *Brill Tagger* merupakan salah satu metode *Rule Based* yang memperoleh aturan berdasarkan *Transformation Based*

Learning merupakan metode yang sangat kompetitif dibandingkan dengan metode stokastik. Terdapat pula kelemahan dalam metode *Rule Based tagger* ini karena tidak adanya algoritma pelatihan tanpa pengawasan yang telah disajikan untuk mempelajari aturan secara otomatis tanpa adanya catatan manual *corpus* [11].

Brill Tagger dengan menggunakan metode *Rule Based* sudah dilatih untuk berbagai macam bahasa, seperti Belanda, Norwegia, Denmark, Inggris, dan lain sebagainya [7]. Dalam bahasa Inggris telah tersedia *Rule Based tagger*-nya, tetapi untuk bahasa Indonesia belum tersedia. Oleh karena itu, pada perancangan ini akan dilakukan proses *training* terlebih dahulu sehingga terbentuk aturan seperti aturan leksikal sebagai *Rule Based tagger* yang digunakan untuk bahasa Indonesia. Algoritma *Brill Tagger* terdiri dari [9]:

1) Proses Inisialisasi.
   - *Known words* (di dalam kosakata): menentukan *tag* yang paling sering diberikan ke suatu bentuk kata.
   - *Unknown words* (di luar kosakata):
     - Kata benda umum (NNP) jika diawali dengan huruf besar dan kata benda lainnya (NN) jika sebaliknya.
     - Mempelajari dan menebak aturan dasar yang sama seperti aturan kontekstual.
2) Fase Pembelajaran
   - Pengulangan dalam menghitung nilai kesalahan dari setiap calon aturan (perbedaan antara jumlah kesalahan sebelum dan sesudah menerapkan aturan).
   - Pilih aturan yang terbaik (skor yang lebih tinggi).
   - Tambahkan dalam perangkat aturan dan diterapkan pada teks.
   - Ulangi sampai tidak ada aturan yang memiliki skor di atas ambang tertentu atau yang telah diberikan (jika ambang yang dipilih adalah nol, yang dapat mengakibatkan *over fitting* berlebihan).

Aturan bahasa alami yang diterapkan oleh *Brill Tagger* adalah [8]:

- Aturan leksikal adalah makna yang sebenarnya, makna yang sesuai dengan hasil observasi alat indera, makna apa adanya, atau makna yang ada di dalam kamus. Contoh untuk bahasa Indonesia adalah kata "apel" memiliki dua arti:

  Ibu membeli **apel**
  Ayah mengikuti **apel** pagi setiap hari senin

  "apel" pada kalimat pertama bermakna buah apel.
  "apel" pada kalimat kedua bermakna pertemuan, upacara.
- Aturan kontekstual adalah makna kata yang berada di dalam konteks suatu kalimat. Contoh untuk bahasa Indonesia adalah:
  - Kata "kepala" pada kalimat di bawah ini memiliki arti kepala sesungguhnya.

    Rambut di **kepala** kakek belum ada yang putih
  - Kata "kepala" pada kalimat di bawah ini memiliki arti pemimpin.

    Bapak Hasan adalah seorang **kepala** sekolah

dimana aturan leksikal digunakan sebagai inisialisasi dan aturan kontekstual digunakan untuk memperbaiki *tag*.

Secara singkat algoritma *Brill Tagger* meliputi proses inisialisasi yang terdiri dari *known words* dan *unknown words*, serta fase pembelajaran yang terdiri dari pengulangan dalam menghitung nilai kesalahan dari setiap calon aturan, memilih aturan yang terbaik, menambahkan perangkat aturan dan diterapkan pada teks, mengulangi sampai tidak ada aturan yang memiliki nilai di atas ambang tertentu atau yang telah diberikan.

2.2 Rancangan

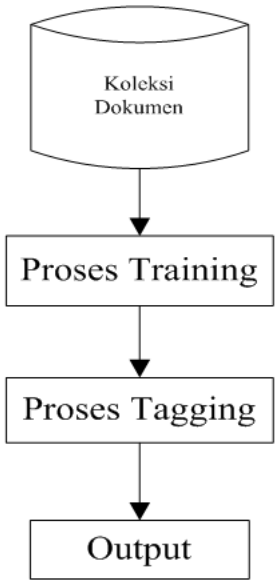
Program aplikasi yang dirancang bertujuan untuk memberikan *tagging* secara otomatis pada suatu kata sehingga dapat diketahui *tag* yang tepat untuk kata tersebut. Data yang di-*input* pada aplikasi ini berupa dokumen artikel berita yang telah diubah ke

dalam bentuk teks dalam *form*at .txt. Pada perancangan ini, diimplementasikan program *Brill Tagger* serta diterapkan aturan leksikal dan kontekstual yang sesuai dengan kaidah bahasa Indonesia. Data merupakan komponen yang paling penting dalam perancangan ini karena data berperan sebagai *input* untuk pelatihan dan pengujian pada program aplikasi yang dibuat. Data yang diproses oleh program adalah dokumen artikel berita berbahasa Indonesia yang telah diubah ke dalam *form*at teks.

Berikut ini adalah tahapan dalam implementasi program *Brill Tagger* untuk bahasa Indonesia:

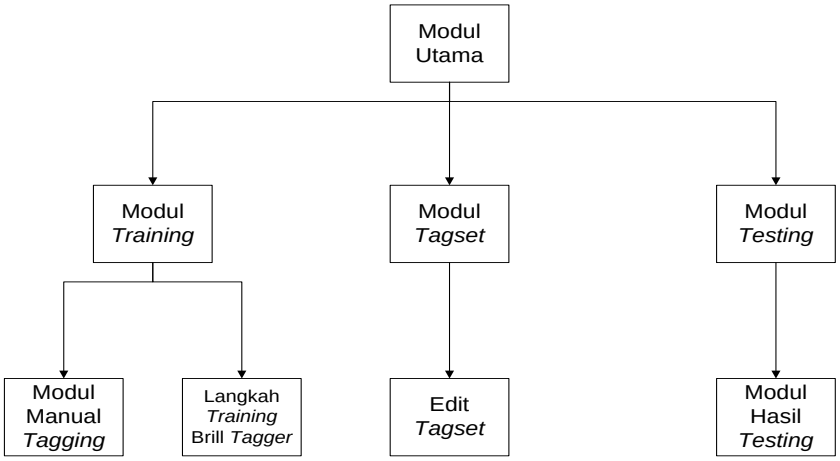
1) Menyiapkan POS *tagset* untuk bahasa Indonesia.
2) Menyiapkan dokumen *training* yang sudah diberi anotasi atau *tag* secara manual.
3) Melakukan proses *training* pada dokumen yang sudah diberi *tag* secara manual dengan menggunakan *Brill Tagger*.
4) Dalam proses *training* akan didapatkan aturan leksikal untuk bahasa Indonesia.
5) Memperbaiki aturan leksikal dan kontekstual untuk bahasa Indonesia yang diperoleh dari tahap *training*.
6) Melakukan proses *testing* pada korpus yang belum diberi *tag* secara manual dengan menggunakan *Brill Tagger*.

Setelah seluruh proses *Brill Tagger* dilakukan maka langkah selanjutnya adalah peninjauan ulang terhadap hasil *output* dari proses *Brill Tagger* secara keseluruhan. Peninjauan ulang ini dilakukan untuk melihat apakah aturan yang diperoleh dari *Brill Tagger* sudah sesuai dengan tata bahasa Indonesia. Apabila belum sesuai maka dilakukan perubahan agar menghasilkan hasil yang lebih sesuai dengan isi dari Bahasa Indonesia. Kemudian aturan yang sudah disesuaikan yang akan menjadi *tagger* untuk pemberian kelas kata. Gambar proses perancangan program POS-*Tagging* untuk bahasa Indonesia dapat dilihat pada Gambar 1.



Gambar 1. Proses perancangan program POS-*Tagging*

Perancangan diagram hirarki bertujuan untuk mendapatkan gambaran mengenai modul yang dibuat. Tampilan pertama dalam program aplikasi ini adalah menu utama yang menampilkan empat tombol menu yang mengarah ke modul-modul yang dapat dipilih pengguna, yaitu modul *training*, modul *tagset*, modul *testing*, dan modul *exit*.



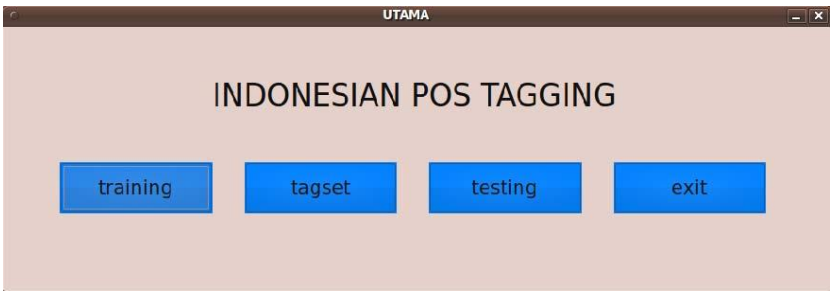
Gambar 2. Rancangan diagram hirarki

3. HASIL PENGUJIAN

Pembuatan sistem diawali dengan membuat rancangan sistem yang digunakan. Setelah itu dilakukan tahap pembuatan program aplikasi yang dimulai dari pembuatan GUI (*graphical user interface*) sampai dengan pengujian hasil dan evaluasi hasil POS-*tagging* dari program yang dirancang. Spesifikasi dari perangkat keras yang akan digunakan dalam perancangan program aplikasi ini adalah satu set komputer dengan *Processor* Intel(R) Core(TM)2 Duo T5550 1.83 Ghz dan Memori RAM 512 MB. Perangkat lunak yang akan digunakan adalah Sistem Operasi Linux Ubuntu 9.10, Gambas 2, Java, GCC dan Active Perl 5.12.2.

Program aplikasi dibuat dengan menggunakan bahasa pemrograman yang berbasis GUI (*graphical user interface*). Tahap-tahap dalam membuat program adalah:

1) *Form* Utama (Gambar 3)
   Form utama merupakan *form* awal pada program yang berhubungan dengan *form* lainnya. *Form* utama selalu muncul di awal pada saat program mulai dijalankan. Setelah *Form* utama selesai dibuat, tambahkan *button training*, *button testing*, *button tagset*, dan *button exit*. *Button training* berfungsi untuk menuju *form training*. *Button tagset* berfungsi untuk menuju *form tagset*. *Button testing* berfungsi untuk menuju *form testing*, dan *button exit* untuk keluar dari program.

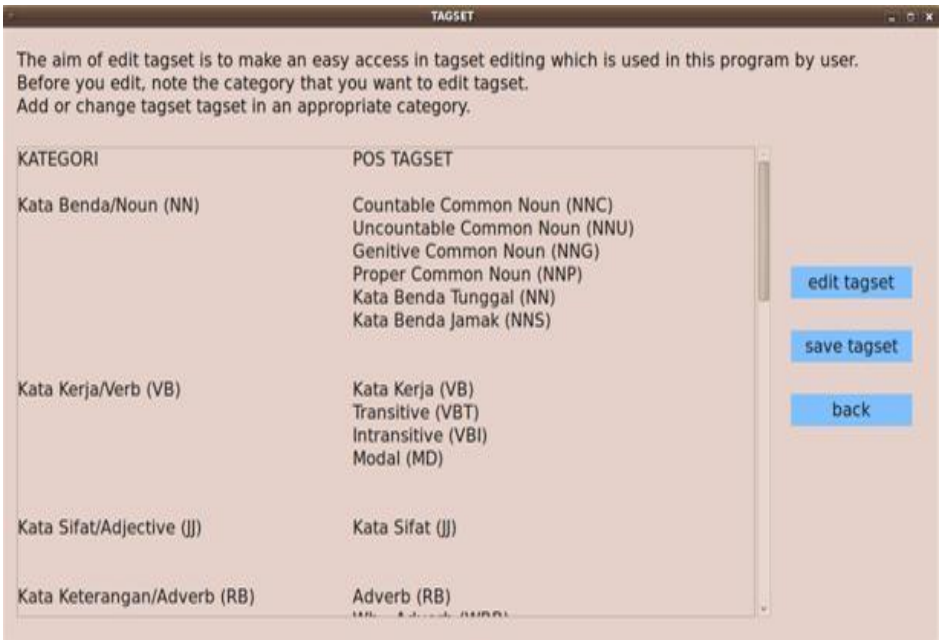


Gambar 3. *Form* utama

2) *Form Training* (Gambar 4)
   Pada *form* ini terdapat 25 *button* yang terbagi dalam dua grup, yaitu grup proses *training* dan grup hasil *training*. Pada grup proses *training* terdiri dari 10 *button* dan 11 *button* pada grup hasil *training* yang mengarah kepada hasil dari *training* pada grup proses *training*.

Gambar 4. *Form training*

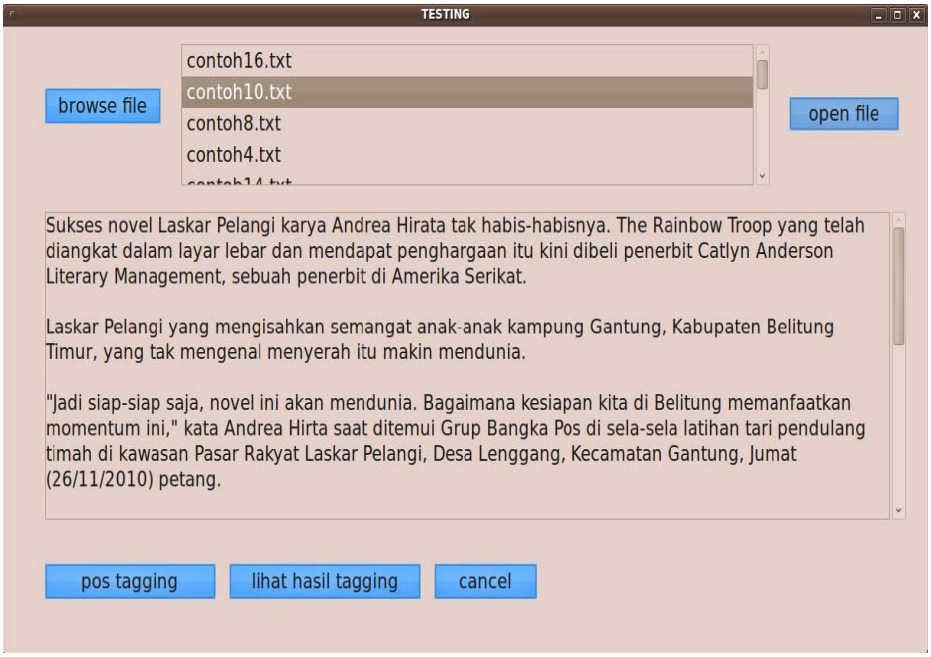
3) *Form Tagset* (Gambar 5)
Form ini berisi *textareatagset* yang memuat *tagset* yang digunakan pada perancangan. *Button edit tagset* yang berfungsi untuk meng-*edit* daftar *tagset* yang sudah ada. *Button save* untuk menyimpan *tagset* yang sudah diubah, serta *button back* untuk kembali ke *form* utama.



Gambar 5. *Form tagset*

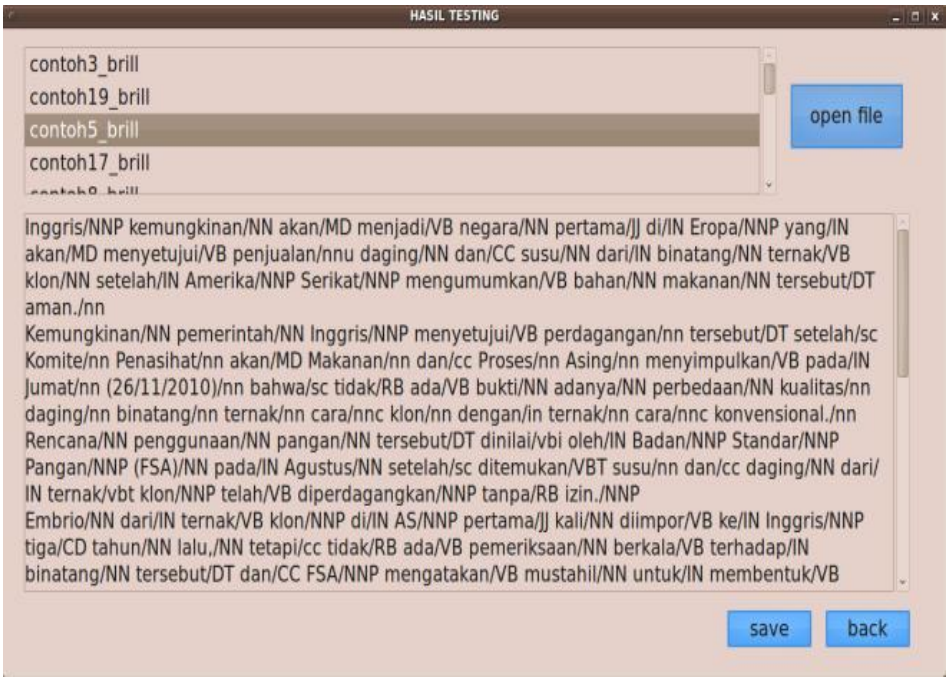
4) *Form Testing* (Gambar 6)
Form proses *testing* berisi *button browse* untuk mem-*browse* dokumen yang akan di-*tagging*. *File* yang telah di-*browse* ini diletakkan pada *listboxdokumen* yang dapat dilihat isinya pada *textareadokumen* dengan menekan *button open*. *Button* lakukan POS-*tagging* berfungsi untuk melakukan pemberian *tag* atau *testing* pada dokumen

yang sudah di-*browse* sebelumnya. *Button* lihat hasil *testing* untuk menuju *form* hasil *testing*. *Button cancel* untuk membatalkan proses *testing* dan kembali ke *form* utama.



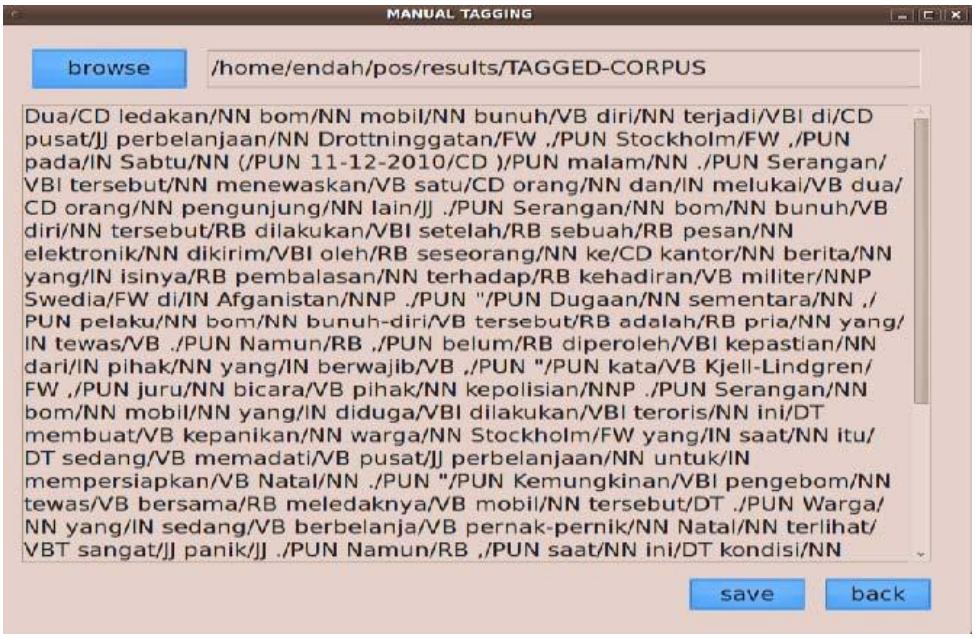
Gambar 6. *Form testing*

5) *Form* Hasil *Testing* (Gambar 7)
Form ini terdiri dari *listbox* yang berisi daftar hasil *tagging*, dan *textareahasil* yang memuat isi dari dokumen yang sudah diberi *tag*. *Button back* untuk kembali ke *form* utama dan *button save* untuk menyimpan hasil POS-*tagging*.



Gambar 7. Hasil *testing*

6) *Form Manual Tagging* (Gambar 8)
Form ini berisi *button browse* untuk memilih *file* yang akan diberi *tag* secara manual, *textbox1* sebagai tempat meletakkan *directory file*, *textarea1* untuk *user* melakukan proses pemberian *tagging* secara manual. Terdapat pula *button save* dan *button back*.



Gambar 8. *Manual tagging*

Jumlah dokumen yang digunakan pada perancangan ini adalah sebanyak 100 dokumen artikel berita, terdiri atas 80 dokumen artikel untuk proses *training* dan 20 dokumen artikel untuk proses pengujian. Sementara itu, untuk jumlah kata yang diproses pada sebuah dokumen, tidak dibatasi. Tahap-tahap dalam pengujian sistem, antara lain:

1) Mengumpulkan dokumen artikel yang digunakan sebagai bahan pengujian program.
2) Melakukan pengujian terhadap setiap modul dan tombol untuk mengecek apakah semua modul dan tombol yang terdapat pada program berjalan dengan baik sesuai dengan fungsinya masing-masing.
3) Melakukan proses *training* pada *Brill Tagger* dengan menggunakan data *training* kemudian diperoleh *lexical rule*, serta *contextrule* yang dapat digunakan untuk proses *testing* selanjutnya.

Pengujian keseluruhan terhadap aplikasi ini dilakukan dengan menjalankan *form-form* yang tersedia, yaitu *form* utama, *form training*, *form tagset*, *form testing*, *form* hasil *testing* dan *form manual tagging*. Pengujian terhadap seluruh *form* dapat dikatakan berhasil karena seluruh *form* berjalan sebagaimana mestinya. Semua menu dan tombol dalam masing-masing *form* dapat menjalankan fungsinya dengan baik**.**

Setelah dilakukan pengujian terhadap *form-form* yang ada, maka dilakukan pengujian terhadap hasil pemberian POS-*tagging*. Pengujian dilakukan terhadap 20 dokumen dan diketahui bahwa jumlah seluruh kata yang digunakan adalah sebanyak 11.411 kata. Tabel 1 adalah hasil pengujian pemberian *tagging* pada setiap dokumen. Pada tabel 1 dapat terlihat bahwa jumlah kata yang salah diberi *tagging* relatif sedikit.

Tabel 1. Hasil pemberian POS-*tagging* pada 20 dokumen *testing*

Dokumen ke-	Jumlah seluruh kata	Jumlah kata yang salah	Jumlah kata yang benar	Persentase
1	199	4	195	97.99
2	349	6	343	98.28
3	354	4	350	98.87
4	1095	8	1087	99.27
5	501	6	495	98.80
6	1226	11	1215	99.10
7	355	7	348	98.03
8	1020	9	1011	99.12
9	586	11	575	98.12
10	474	5	469	98.95
11	372	10	362	97.31
12	494	6	488	98.79
13	452	8	444	98.23
14	852	9	843	98.94
15	272	12	260	95.59
16	410	6	404	98.54
17	836	12	824	98.56
18	444	9	435	97.97
19	471	4	467	99.15
20	649	7	642	98.92
Total	11.411	154	11.257	98.65

Secara keseluruhan dapat dihitung bahwa total hasil *tagging* yang benar adalah sebanyak 11.257 kata dan hasil *tagging* yang tidak benar sebanyak 154 kata. Hal ini didapat dengan cara membandingkan secara manual hasil *tagging* dari masing-masing dokumen artikel dan hasil *tagging* dengan proses pemberian *tagging* secara manual. Hasil akurasi program ini adalah 98,65% untuk ketepatan pemberian POS-*tagging* terhadap bahasa Indonesia dengan implementasi *Brill Tagger*.

Dua ledakan bom mobil bunuh diri terjadi di pusat perbelanjaan Drottninggatan, Stockholm, pada Sabtu (11/12/2010) malam. Serangan tersebut menewaskan satu orang dan melukai dua orang pengunjung lain. Serangan bom bunuh diri tersebut dilakukan setelah sebuah pesan elektronik dikirim oleh seseorang ke kantor berita yang isinya pembalasan terhadap kehadiran militer Swedia di Afganistan. "Dugaan sementara, pelaku bom bunuh diri tersebut adalah pria yang tewas. Namun, belum diperoleh kepastian dari pihak yang berwajib," kata Kjell Lindgren, juru bicara pihak kepolisian. Serangan bom mobil yang diduga dilakukan teroris ini membuat kepanikan warga Stockholm yang saat itu sedang memadati pusat perbelanjaan untuk mempersiapkan Natal. "Kemungkinan pengebom tewas bersama meledaknya mobil tersebut. Warga yang sedang berbelanja pernak-pernik Natal terlihat sangat panik. Namun, saat ini kondisi telah terkendali," kata Petra Sjolander, pihak keamanan setempat. Serangan tersebut diduga dilakukan teroris, demikian dikatakan Menteri Luar Negeri Swedia Carl Bildt dalam sebuah pesan yang dikirim dari akun Twitter-nya. "Kebanyakan pebisnis khawatir tentang serangan teroris di pusat bisnis kota Stockholm," tulis Bildt. "Meski dinilai gagal, tapi ini benar-benar sebuah bencana," katanya menambahkan.

Gambar 9. Contoh dokumen yang belum diberi *tag*

Dua/NN ledakan/VB bom/NN mobil/NN bunuh/NN diri/VB terjadi/VB di/IN pusat/NNP perbelanjaan/NN Drottninggatan,/nn Stockholm,/nn pada/IN Sabtu/NNP (11/12/2010)/CD malam./nn Serangan/NNP tersebut/DT menewaskan/VB satu/CD orang/NN dan/CC melukai/VB dua/CD orang/NN pengunjung/NN lain./nn Serangan/NNP bom/NN bunuh/NN diri/VB tersebut/DT dilakukan/VB setelah/IN sebuah/NN pesan/NN elektronik/NN dikirim/VB oleh/IN seseorang/NN ke/IN kantor/NN berita/nnc yang/IN isinya/NN pembalasan/NN terhadap/IN kehadiran/NN militer/JJ Swedia/NNP di/IN Afganistan./nn "Dugaan/NN sementara,/nn pelaku/NN bom/NN bunuh/NN diri/VB tersebut/DT adalah/VB pria/NN yang/IN tewas./nn Namun,/nn belum/NEG diperoleh/VB kepastian/NN dari/IN pihak/NN yang/IN **berwajib,"/nn** kata/NN Kjell/NNP Lindgren,/nn juru/NN bicara/NN pihak/NN **kepolisian./nn** Serangan/NNP bom/NN mobil/NN yang/IN diduga/VB dilakukan/VB teroris/VB ini/DT membuat/vbt kepanikan/VB warga/NN Stockholm/NNP yang/IN saat/IN itu/DT sedang/NN memadati/VB pusat/NNP perbelanjaan/NN untuk/IN mempersiapkan/VB Natal./nn **"Kemungkinan/NN** pengebom/NN tewas/VB bersama/VB meledaknya/VB mobil/NN tersebut./nn Warga/NNP yang/IN sedang/NN berbelanja/VB pernak-pernik/NN Natal/nn terlihat/vbi sangat/RB panik./nn Namun,/nn saat/IN ini/DT kondisi/NN telah/VB terkendali,"/nn kata/NN Petra/NNP Sjolander,/nn pihak/NN keamanan/NN **setempat./nn** Serangan/NNP tersebut/DT diduga/VB dilakukan/VB teroris,/nn demikian/NN dikatakan/VB Menteri/NNP Luar/NNP Negeri/NNP Swedia/NNP Carl/NNP Bildt/NNP dalam/IN sebuah/NN pesan/NN yang/IN dikirim/VB dari/IN akun/NN **Twitter-nya./nn "Kebanyakan/VB** pebisnis/NN khawatir/NN tentang/NN serangan/NN teroris/VB di/IN pusat/NNP bisnis/nn kota/nnc **Stockholm,"/nn** tulis/NN Bildt./nn "Meski/NN dinilai/vbi **gagal,/nn** tapi/IN ini/DT benar-benar/RB sebuah/NN bencana,"/nn katanya/NN menambahkan/VB

Gambar 10. Contoh dokumen yang telah diberi *tagging* oleh program

Pada Gambar 9 dapat dilihat contoh dokumen yang belum diberi *tag*. Dokumen tersebut menjadi salah satu *input* pada proses pengujian dokumen. Hasil dari dokumen yang telah diberi *tagging* oleh sistem dapat dilihat pada Gambar 10. Kelas kata diberikan satu kata per kata dan diletakkan langsung disebelah kata tersebut dan dipisahkan oleh tanda "/".

Apabila hasil pengujian dibandingkan antara dokumen hasil *tagging* manual dengan *tagg*ing dari aplikasi maka terdapat 12 ketidaktepatan pemberian *tagging*. Tabel 2 adalah perbandingan antara dokumen yang diberi *tagging* secara manual dan otomatis oleh program. Kesalahan pemberian kelas kata terjadi pada saat tanda baca muncul langsung bersama kata tanpa tanda spasi sehingga kata tersebut dianggap sebagai PUN (*punctuation*). Namun secara keseluruhan jumlah dan jenis kesalahan tidak terlalu signifikan.

Tabel 2. Contoh perbandingan dokumen yang telah diberi *tagging* secara manual dan otomatis oleh program

***Tagging* secara manual**	***Tagging* oleh program aplikasi**
Dua/NN ledakan/VB bom/NN mobil/NN bunuh/NN diri/VB terjadi/VB di/IN pusat/NNP perbelanjaan/NN Drottninggatan,/nn Stockholm,/nn pada/IN Sabtu/NNP (11/12/2010)/CD malam./nn Serangan/NNP tersebut/DT menewaskan/VB satu/CD orang/NN dan/CC melukai/VB dua/CD orang/NN pengunjung/NN lain./nn Serangan/NNP bom/NN bunuh/NN diri/VB tersebut/DT dilakukan/VB setelah/IN sebuah/NN pesan/NN elektronik/NN dikirim/VB oleh/IN seseorang/NN ke/IN kantor/NN berita/nnc yang/IN isinya/NN pembalasan/NN terhadap/IN kehadiran/NN militer/JJ Swedia/NNP di/IN Afganistan./nn "Dugaan/NN sementara,/nn pelaku/NN bom/NN bunuh/NN diri/VB tersebut/DT adalah/VB pria/NN yang/IN tewas./nn Namun,/nn belum/NEG diperoleh/VB kepastian/NN dari/IN pihak/NN yang/IN berwajib **,/PUN "/PUN** kata/NN Kjell/NNP Lindgren/NNP juru/NN bicara/NN pihak/NN **kepolisian./NNP** Serangan/NNP bom/NN mobil/NN yang/IN diduga/VB dilakukan/VB teroris/VB ini/DT membuat/vbt kepanikan/VB warga/NN Stockholm/NNP yang/IN saat/IN itu/DT sedang/NN memadati/VB pusat/NNP perbelanjaan/NN untuk/IN mempersiapkan/VB Natal./nn **"/PUN Kemungkinan/NN** pengebom/NN tewas/VB bersama/VB meledaknya/VB mobil/NN tersebut./nn Warga/NNP yang/IN sedang/NN berbelanja/VB pernak-pernik/NN Natal/nn terlihat/vbi sangat/RB panik./nn Namun,/nn saat/IN ini/DT kondisi/NN telah/VB terkendali,"/nn kata/NN Petra/NNP Sjolander,/nn pihak/NN keamanan/NN setempat/NN **./PUN** Serangan/VB tersebut/DT diduga/VB dilakukan/VB teroris,/nn demikian/NN dikatakan/VB Menteri/NNP Luar/NNP Negeri/NNP Swedia/NNP Carl/NNP Bildt/NNP dalam/IN sebuah/NN pesan/NN yang/IN dikirim/VB dari/IN akun/NN Twitter-nya/RB **"/PUN** Kebanyakan/VB pebisnis/NN khawatir/NN tentang/NN serangan/NN teroris/VB di/IN pusat/NNP bisnis/nn kota/nnc Stockholm **,/PUN "/PUN** tulis/NN Bildt./nn **"/PUN** Meski/NN dinilai/vbi **gagal/JJ ,/PUN** tapi/IN ini/DT benar-benar/RB sebuah/NN bencana,"/nn katanya/NN menambahkan/VB	Dua/NN ledakan/VB bom/NN mobil/NN bunuh/NN diri/VB terjadi/VB di/IN pusat/NNP perbelanjaan/NN Drottninggatan,/nn Stockholm,/nn pada/IN Sabtu/NNP (11/12/2010)/CD malam./nn Serangan/NNP tersebut/DT menewaskan/VB satu/CD orang/NN dan/CC melukai/VB dua/CD orang/NN pengunjung/NN lain./nn Serangan/NNP bom/NN bunuh/NN diri/VB tersebut/DT dilakukan/VB setelah/IN sebuah/NN pesan/NN elektronik/NN dikirim/VB oleh/IN seseorang/NN ke/IN kantor/NN berita/nnc yang/IN isinya/NN pembalasan/NN terhadap/IN kehadiran/NN militer/JJ Swedia/NNP di/IN Afganistan./nn "Dugaan/NN sementara,/nn pelaku/NN bom/NN bunuh/NN diri/VB tersebut/DT adalah/VB pria/NN yang/IN tewas./nn Namun,/nn belum/NEG diperoleh/VB kepastian/NN dari/IN pihak/NN yang/IN **berwajib,"/nn** kata/NN Kjell/NNP Lindgren,/nn juru/NN bicara/NN pihak/NN **kepolisian./nn** Serangan/NNP bom/NN mobil/NN yang/IN diduga/VB dilakukan/VB teroris/VB ini/DT membuat/vbt kepanikan/VB warga/NN Stockholm/NNP yang/IN saat/IN itu/DT sedang/NN memadati/VB pusat/NNP perbelanjaan/NN untuk/IN mempersiapkan/VB Natal./nn **"Kemungkinan/NN** pengebom/NN tewas/VB bersama/VB meledaknya/VB mobil/NN tersebut./nn Warga/NNP yang/IN sedang/NN berbelanja/VB pernak-pernik/NN Natal/nn terlihat/vbi sangat/RB panik./nn Namun,/nn saat/IN ini/DT kondisi/NN telah/VB terkendali,"/nn kata/NN Petra/NNP Sjolander,/nn pihak/NN keamanan/NN **setempat./nn** Serangan/NNP tersebut/DT diduga/VB dilakukan/VB teroris,/nn demikian/NN dikatakan/VB Menteri/NNP Luar/NNP Negeri/NNP Swedia/NNP Carl/NNP Bildt/NNP dalam/IN sebuah/NN pesan/NN yang/IN dikirim/VB dari/IN akun/NN **Twitter-nya./nn "Kebanyakan/VB** pebisnis/NN khawatir/NN tentang/NN serangan/NN teroris/VB di/IN pusat/NNP bisnis/nn kota/nnc **Stockholm,"/nn** tulis/NN Bildt./nn "Meski/NN dinilai/vbi **gagal,/nn** tapi/IN ini/DT benar-benar/RB sebuah/NN bencana,"/nn katanya/NN menambahkan/VB

4. PEMBAHASAN HASIL PENGUJIAN

Penerapan aturan leksikal Bahasa Indonesia dapat dilihat pada contoh kalimat berikut:

> **Ia berusaha meledakkan sebuah van sarat bahan peledak yang diparkir di dekat lokasi.**

Kalimat di atas akan mendapatkan *tag* secara manual sebagai berikut:

> **Ia/PRP berusaha/VB meledakkan/NN sebuah/NN van/NN sarat/NN bahan/NN peledak/NN yang/SC diparkir/NN di/IN dekat/JJ lokasi/NN ./.**

Sedangkan kalimat hasil pemberian *tagging* dengan program adalah:

> **Ia/PRP berusaha/VB meledakkan/NN sebuah/NN van/NN sarat/NN bahan/NN peledak/NN yang/SC diparkir/VB di/IN dekat/JJ lokasi/NN ./.**

Kata "diparkir" mengalami perubahan *tag* dari "diparkir/NN" menjadi "diparkir/VB" karena penerapan aturan leksikal:

> **di haspref 2 VB 3.5**

Aturan tersebut berarti kata yang memiliki dua huruf awalan di-, maka diberi *tag* VB. Dapat dilihat, bahwa kata "diparkir" berasal dari kata dasar "parker" yang memiliki awalan di-. Berdasarkan aturan leksikal di atas, kata "diparkir" mengalami perubahan *tag* karena memiliki awalan di-.

Penerapan aturan kontekstual Bahasa Indonesia dapat diketahui dari hasil pengujian implementasi *Brill Tagger* untuk Bahasa Indonesia ini seperti pada contoh kalimat berikut:

> **Departemen Kehakiman AS mengidentifikasi tersangka utama pelaku bom bunuh diri.**

Kalimat di atas akan mendapatkan *tag* secara manual sebagai berikut:

> **Departemen/NN Kehakiman/NN AS/NNP mengidentifikasi/VB tersangka/VBT utama/JJ pelaku/NN bom/NN bunuh/VB diri/NN ./.**

Sedangkan dari hasil pemberian *tag*ging dengan program, diperoleh hasil:

> **Departemen/NN Kehakiman/NN AS/NNP mengidentifikasi/VBT tersangka/VBT utama/JJ pelaku/NN bom/NN bunuh/VB diri/NN ./.**

Perubahan *tag* pada kata mengidentifikasi/VB menjadi mengidentifikasi/VBT disebabkan karena mengalami aturan kontekstual "**VB VBT PREV1OR2OR3*TAG* NN".**

Aturan tersebut memiliki arti kata yang memiliki *tag* VB akan berubah *tag*-nya menjadi VBT jika 1 atau 2 atau 3 kata sebelumnya memiliki *tag* NN. Apabila diperhatikan dua kata sebelum kata "mengidentifikasi" adalah kata kehakiman yang

memiliki *tag* NN. Ini berarti kata "mengidentifikasi" mengalami perubahan *tag* karena penerapan aturan kontekstual.

Setelah diperhatikan beberapa aturan yang dihasilkan dari proses *Transformation Based Learning* pada tahap pelatihan, ditemukan beberapa aturan leksikal yang tidak sesuai dengan Bahasa Indonesia seperti "**aya hassuf 3 JJ 2.5**". Dalam Bahasa Indonesia tidak ada aturan yang menyebutkan bahwa terdapat kata yang memiliki akhiran "aya". Kesalahan lain yang muncul adalah adanya aturan kontekstual "**NNP NN PREV1OR2OR3*TAG*.**". Seharusnya di dalam bahasa Indonesia titik tidak memberikan pengaruh apapun dalam perubahan bentuk kelas kata.

Pada pengujian pertama akurasi hasil penggunaan *Brill Tagger* untuk bahasa Indonesia adalah 98,65%. Ketidaksesuaian aturan leksikal dan kontekstual yang dihasilkan oleh algoritma *Brill Tagger* menimbulkan kesalahan dalam pemberian *tag*ging sehingga setelah disesuaikan dengan aturan Bahasa Indonesia diperoleh hasil 99,75%. Jumlah kata yang salah diberi kelas kata berkurang dari 157 kata menjadi 28 kata.

Secara keseluruhan waktu pemberian *tagging* pada tahap *testing* relatif singkat yaitu ± 0.2 detik. Pada penelitian ini tidak dilakukan penelitian secara khusus mengenai waktu proses pemberian *tagging*. Proses yang memakan waktu lebih lama adalah proses *training* dimana peneliti harus memberikan *tagging* secara manual terlebih dahulu.

5. KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan terhadap program aplikasi, hasil *testing*, serta respon dari *user*, maka dapat ditarik kesimpulan sebagai berikut:

- Pengujian yang dilakukan terhadap hasil pemberian POS-*tagging* dengan program aplikasi menghasilkan sebuah kesimpulan bahwa pemberian POS-*tagging* ditentukan oleh data *training*. Banyaknya kosakata pada data *training* dibutuhkan untuk memperkecil kesalahan. Makin banyak kosakata yang dimiliki, program aplikasi dapat semakin belajar untuk memberikan POS-*tagging* pada proses *testing*.
- Program *Brill Tagger* dapat diimplementasikan dengan baik untuk memberikan POS-*tagging* pada dokumen bahasa Indonesia. Hasil akurasi penggunaan *Brill Tagger* untuk bahasa Indonesia adalah sebesar 98,65 %.
- Aturan leksikal dan kontekstual yang dihasilkan oleh *Brill Tagger* sudah dapat digunakan sebagai aturan untuk memberikan kelas kata untuk kalimat bahasa Indonesia. Namun masih terdapat kesalahan aturan sehingga menghasilkan kelas kata yang tidak tepat. Setelah memperbaiki aturan leksikal dan kontekstual maka hasil akurasi meningkat menjadi 99,75%.
- Penerapan aturan leksikal dan kontekstual yang tepat dapat menghasilkan POS-*Tagging* yang lebih akurat.

REFERENSI

[1]. Arman, Arry Akhmad, "*Teknologi Pemrosesan Bahasa Alami sebagai Teknologi Kunci untuk Meningkatkan Cara Interaksi antara Manusia dengan Mesin"*, diakses dari http://www.itb.ac.id/focus/focus_file/Pidato Ilmiah pada Sidang Terbuka PMB 2004.pdf, 2004.

[2]. Chandrawati, Triastuti, "*Pengembangan Part Of Speech Tagger untuk Bahasa Indonesia Berdasarkan Metode Conditional Random Fields dan Transformation Based"*, Fakultas Ilmu Komputer Universitas Indonesia (Skripsi tidak dipublikasikan), Depok, UI, 2008.

[3]. Adriani, Mirna, Hisar M. Manurung dan Femphy, Pisceldo. "*Probabilistic Part-of-speech tagging for Bahasa Indonesia"*, dalam MALINDO 2009: "*Third International Workshop on Malay and Indonesian Language Engineering"*, 2009.

[4]. K. Freddy, P. F. Tania, P. Endah dan S. Stephen, "*Menerapkan Program Stanford* POS *Tagger untuk Melakukan Tagging pada Kalimat Berbahasa Indonesia"*, Fakultas Teknologi Informasi Universitas Tarumanagara, 2010.

[5]. W. Alfan Farizki dan P. Ayu, "*HMM Based Part of Speech Tagger for Bahasa Indonesia"*, dalam MALINDO 2010: "*4th International Workshop on Malay and Indonesian Language Engineering"*, 2010.

[6]. Brants, Thorsten, *Natural Language Processing in Information Retrieval*, Google.Inc, diakses dari www.cnts.ua.ac.be/clin2003/proc/03Brants.pdf., 2003.

[7]. Megyesi, B. 1999. Brill's POS *Tagger with Extended Lexical Templates for Hungarian*. dalam *Proceedings of the Workshop (W01) on Machine Learning in Human Language Technology*, ACAI'99, Chania, Crete, Greece July 5 - July 16, 1999, pp. 22-28.

[8]. Brill, Eric, "*A Simple Rule-Based Part of Speech Tagger"*, dalam proceedings of the third conference on Applied natural language processing ANLC '92, P. 152-155.

[9]. Sari, Syandra, H. Herika, A. Mirna dan B. Bressan, "*Part-of-speech Tagging Using Transformation-Based Error Driven Learning"*. dalam ADD-3 Workshop, Bangkok, 2008.

[10]. D. Manning, Christoper. "*Foundations of Statistical Natural Language Processing*, MIT Press", Juni 1999.

[11]. Smeaton, Alan F., "*Natural Language Processing & Information Retrieval"*, Second European Summer School in Information Retrieval (ESSIR'95), Glasgow, Scotland, September 1995**.**