POS Tagging Bahasa Indonesia Dengan HMM dan Rule Based

Kathryn Widhiyanti[1]
kathrynwidhiyanti@gmail.com

Agus Harjoko[2]
aharjoko@ugm.ac.id

Abstract

The research conduct a Part of Speech Tagging (POS-tagging) for text in Indonesian language, supporting another process in digitising natural language e.g. Indonesian language text parsing. POS-tagging is an automated process of labelling word classes for certain word in sentences (Jurafsky and Martin, 2000). The escalated issue is how to acquire an accurate word class labelling in sentence domain.

The author would like to propose a method which combine Hidden Markov Model and Rule Based method. The expected outcome in this research is a better accurary in word class labelling, resulted by only using Hidden Markov Model. The labelling results –from Hidden Markov Model– are refined by validating with certain rule, composed by the used corpus automatically.

From the conducted research through some POST document, using Hidden Markov Model, produced 100% as the highest accurary for identical text within corpus. For different text within the referenced corpus, used words subjected in corpus, produced 92,2% for the highest accurary.

Keywords— *part of speech tagging, hidden markov model, rule based*

1. PENDAHULUAN

Part of Speech Tagging (POS-Tagging) adalah suatu proses yang memberikan label kelas kata secara otomatis pada suatu kata dalam kalimat (Jurafsky, 2000). Hasil dari *Part of Speech Tagging* (POS) ini sangat berpengaruh terhadap keluaran dari proses *Parsing* (Sukamto, 2009). Masalah yang muncul adalah bagaimana cara mendapatkan pelabelan kelas kata yang tepat dalam konteks kalimat.

Penelitian mengenai *Part of Speech Tagging* untuk teks berbahasa Indonesia juga sudah banyak dilakukan menggunakan berbagai macam metode dan hasil yang diperoleh juga sudah memiliki akurasi yang tinggi. Metode yang sudah pernah dilakukan antara lain Brill *tagger* dan memiliki kurasi 88% (Wicaksono & Purwarianti, 2010), *Conditional Random Field (CRF)* dan *Maximum Entropy Method* memiliki akurasi 97,57% (Pisceldo,

[1] *Magister Ilmu Komputer, Fakultas MIPA, Universitas Gadjah Mada, Yogyakarta*

[2] *Magister Ilmu Komputer, Fakultas MIPA, Universitas Gadjah Mada, Yogyakarta*

Andriani dan Manurung) dan *Hidden Markov Model* yang memiliki akurasi 96,5% (Wicaksono & Purwarianti, 2010).

Pada penelitian ini penulis mencoba suatu metode yaitu menggabungkan antara *Hidden Markov Model* dan *Rule Based* dengan harapan bisa menghasilkan pelabelan dengan tingkat akurasi yang lebih baik dari pelabelan kelas kata yang hanya menggunakan metode *Hidden Markov Model*.

2.LANDASAN TEORI

2.1 Kalimat

Kalimat adalah satuan bahasa terkecil, dalam wujud lisan atau tulisan, yang mengungkapkan pikiran yang utuh (Alwi,Dardjowidjojo, Lapoliwa & Moeliono, 2003). Kalimat dalam huruf Latin dimulai dengan huruf besar dan diakhiri dengan tanda titik (.), tanda tanya (?), atau tanda seru (!). Sementara, di dalam kalimat itu sendiri mungkin terdapat tanda baca yang lain seperti tanda koma (,), tanda titik dua (:), tanda titik koma (;), tanda sambung (-) atau spasi (Alwi,Dardjowidjojo, Lapoliwa & Moeliono, 2003).

2.2 Kelas Kata

Ada banyak kata yang terdapat di dalam suatu bahasa. Kata-kata ini dikategorikan ke dalam kelas-kelas tertentu dan menjadi posisi penting dalam deskripsi dan studi gramatika. Kelas kata kata adalah perangkat kata yang sedikit banyak berperilaku sintaksis sama (Kridalaksana, 2007). Dalam menentukan kelas kata, prinsip yang perlu dipegang ialah kenyataan bahwa kelas kata atau kategori kata adalah bagian dari sintaksis, jadi ciri-ciri tiap kata harus dijelaskan dari sudut sintaksis. Kelas kata tersebut adalah (Kridalaksana, 2007):

2.2.1Verba

Verba secara umum dikenal sebagai kata kerja. Contoh: *menyapu, memasak.*

2.2.2 Adjektiva

Adjektiva biasa dikenal sebagai kata sifat. Contoh: *adil, kesakitan.*

2.2.3 Nomina

Nomina biasa dikenal sebagai kata benda. Contoh: *meja, kursi.*

2.2.4 Pronomina

Pronomina merupakan kelas kata yang untuk menggantikan nomina. Contoh: *kami, dia, ibu(nya).*

2.2.5 Numeralia

Numeralia merupakan kelas kata untuk bilangan, dapat mendampingi nomina, dan atau mendampingi numeralia lain. Contoh: *1000, dua pertiga, beratus-ratus.*

2.2.6 Adverbia

Adverbia atau kata keterangan dapat mendampingi preposisi, adjektiva dan atau numeralia dalam konstruksi sintaksis. Contoh: *sebaiknya, boleh, jangan-jangan.*

2.2.7 Interogativa

Interogativa adalah kategori dalam kalimat interogatif yang berfungsi menggantikan sesuatu yang ingin diketahui oleh pembicara atau mengukuhkan apa yang telah diketahui oleh pembicara. Contoh: *apa.*

2.2.8 Demonstrativa

Demonstrativa adalah kategori yang berfungsi untuk menunjukkan sesuatu di dalam maupun diluar wacana. Contoh: *itu, begitu, demikian.*

2.2.9 Artikula

Artikula dalam bahasa Indonesia adalah kategori yang mendampingi nomina dasar. Contoh: *si pengemis, sang raja.*

2.2.10 Preposisi

Preposisi dikenal dengan kata depan. Contoh: *di rumah, ke pasar.*

2.2.11 Konjungsi

Konjungsi atau kata hubung berfungsi untuk meluaskan satuan yang lain dengan menghubungkan dua satuan atau lebih dalam konstruksi kalimat. Contoh: *agar, akan tetapi, bilamana.*

2.2.12 Kategori fatis

Kategori fatis adalah kategori yang bertugas memulai, mempertahankan, atau mengukuhkan komunikasi antara pembicara dan kawan bicara. Kelas kata ini biasanya terdapat dalam konteks dialog, atau wawancara bersambutan. Contoh: *ayo* dalam kata *ayo kita pergi.*

2.2.13 Interjeksi

Interjeksi adalah kategori yang bertugas mengungkapkan perasaan pembicara dan secara sintaksis tidak berhubungan dengan kata-kata lain dalam ujaran. Interjeksi bersifat ekstra kalimat dan selalu mendahului ujaran sebagai teriakan yang lepas atau berdiri sendiri. Contoh: *wahai, aduhai, astaga, alhamdullilah.*

2.3 *Hidden Markov Model*

Hidden Markov Model (HMM) adalah sebuah model statistik dari sebuah sistem yang melakukan perhitungan probabilitas dari suatu kejadian yang tidak dapat diamati berdasarkan kejadian yang dapat diamati (Jurafsky, 2000). Perhitungan probabilitas dilakukan dengan melihat kejadian-kejadian lain yang dapat diamati secara langsung.

Hidden Markov Model memiliki 2 macam bagian yaitu *observed state* dan *hidden state*. *Observed state* merupakan bagian yang dapat diamati secara langsung dan *hidden state* merupakan bagian yang tidak dapat diamati (Wibisono, Y. 2008). Pada kasus *Part of Speech Tagging*, urutan kelas kata tidak dapat diamati secara langsung sehingga dijadikan sebagai *hidden state* dan yang menjadi *observed state* adalah urutan kata-kata. Dari urutan kata-kata harus dicari urutan kelas kata yang paling tepat (Alwi,Dardjowidjojo, Lapoliwa & Moeliono, 2003).

Persamaan [1] merupakan persamaan dari *Hidden Markov Model* untuk kasus *Part of Speech Tagging*.

$$Tag_n = Max\ (P\ (word_i|tag_i)^*\ P(tag_i|tag_{i-1})) \quad [1]$$

dimana,

tag_n : kelas kata yang dicari

tag_i : kelas kata dari $word_i$ yang ada di corpus.

$Word_i$: kata yang dicari kelas katanya.

Tag_{i-1} : kelas kata sebelum kelas kata dari $word_i$ yang ada di corpus sebanyak 1

P : probabilitas

2.4 *Rule Based*

Metode *Rule Based* ini merupakan metode yang menggunakan aturan bahasa *(grammar)* untuk mendapatkan kelas kata pada suatu kata dalam suatu kalimat (Jurafsky, 2000). Metode *Rule Base* ini memiliki 2 arsitektur. Metode yang pertama adalah metode Rule base yang menggunakan kamus untuk menandai kata dengan kelas kata (leksikon). Metode yang kedua adalah menggunakan *disambiguation rule* secara manual yang nantinya diproses menjadi satu kelas kata saja untuk setiap kata (Jurafsky, 2000).

Ada juga penelitian *Part of Speech Tagging* yang menggunakan arsitektur leksikon dan *disambiguation rule* dikenal dengan *Engtwol tagger* [1]. Pada penelitian ini proses rule based diawali dengan memproses hasil dari *Hidden Markov Model* yang berupa kata berikut kelas katanya akan dipecah menjadi kalimat-kalimat dengan parameter titik, koma, tanda tanya dan tanda seru. Setelah itu kata akan diambil kelas katanya. Kemudian dari kelas kata pertama sampai terakhir akan dicocokan dengan *rule*(aturan) yang sudah ada di kamus aturan. Jika semua susunan *rule* dalam kalimat ada dalam kamus aturan, maka sistem akan menampilkan kata dengan kelas katanya sebagai output. Jika ditemukan perbedaan kelas kata dengan kelas kata dalam kamus, maka sistem akan memberi tanda pada kata tersebut dan menampilkan kelas kata yang lebih tepat dari kelas kata yang didapat dari proses *Hidden Markov Model.* Berdasarkan teori mengenai *rule based* (Jurafsky, 2000) ternyata untuk kasus pelabelan bahasa Indonesia mengalami sedikit kesulitan untuk

mendapatkan *disambiguation rule*. Dalam bahasa Indonesia aturan ambiguitas sangat banyak sehingga diperlukan penelitian khusus untuk memperoleh aturan ambiguitas ini. Pada penelitian ini mencoba mendapatkan aturan ambiguitas secara otomatisi yaitu dengan mengumpulkan susunan kelas kata dalam satu kalimat penuh. Dari kumpulan susunan kelas kata ini bisa dilakukan pengecekkan terhadap kelas kata yang dicari dalam kalimat. Langkah selanjutnya adalah memecah susunan kelas kata dalam kalimat dengan tujuan untuk mendapatkan susunan aturan kelas kata yang baru.

2.5 Analisis Data

Penelitian ini membutuhkan 2 macam data. Data yang pertama adalah data untuk pelatihan dan data yang kedua adalah data penguji. Data pelatihan berupa *corpus*. *Corpus* pada penelitian ini merupakan file dengan format teks (*.txt) yang berisikan kata-kata dalam susunan kalimat dan sudah diberikan pelabelan kelas katanya. *Corpus* yang digunakan dalam penelitian ini adalah *corpus* yang sudah pernah digunakan pada penelitian sebelumnya. *Corpus* pertama merupakan modifikasi *corpus* penelitian HMM Based POS (Wicaksono & Purwarianti, 2010). *Corpus* yang kedua adalah modifikasi *corpus* penelitian dengan metode CRF dan Maximum Entropy (Pisceldo, Andriani dan Manurung). Kelas kata yang dipakai pada penelitian ini ditunjukkan pada Tabel 1 dan Tabel 2.

Tabel 1.
Kelas kata Wicaksono dan Purwarianti (2010)

o	ag	Tag Name	Example
	P	Open Parenthesis	({[
	P	Close Parenthesis	)}]
	M	Slash	/
		Semicolon	;
		Colon	:
		Quotation	"
	, ?, !	Sentence terminator	., ?, !
	-, -	Dash	--, -
		Comma	,
0	J	Adjective	Kaya, manis
1	B	Adverb	Semetara, nanti
2	N	Common Noun	Mobil

Tabel 1. (lanjutan)
Kelas kata Wicaksono dan Purwarianti (2010)

o	ag	Tag Name	Example
3	NP	Proper Noun	Bekasi, Indonesia
4	NG	Genetive Noun	Bukunya
5	BI	Intranstive Verd	Pergi
6	BT	Transitif Verb	Membeli
7	N	Preposition	Di, ke, dari
8	D	Modal	Bisa
9	C	Coor-conjunction	Dan, atau, tetapi
0	C	Subor-conjunction	Jika, ketika
1	T	Determiner	Para, ini, itu
2	H	Interjection	Wah, aduh, Oi
3	DO	Ordinal Numerals	Pertama, kedua
4	DC	Colective Numerals	Bertiga
5	DP	Primary Numerals	Satu, dua
6	DI	Irregular Numerals	Beberapa
7	RP	Personal pronouns	Saya, kamu
8	P	WH-Pronouns	Apa, siapa
9	RL	Locative pronouns	Sini, situ,sana
0	RN	Number Pronouns	Kedua-duanya
1	EG	Negation	Bukan, tidak
2	YM	Symbol	@, #, %, $, ^
3	P	Particel	Pun, kah
4	W	Foreign Word	All, word

Tabel 2.
Kelas kata penelitian Pisceldo dkk.(2009)

o	ag	Tag Name	Example
	P	Open Parenthesis	({[
	P	Close Parenthesis	)}]
		Semicolon	:
		Colon	:
		Quotation	"
	, ?, !	Sentence terminator	., ?, !
	-, -	Dash	--, -
		Comma	,
	J	Adjective	Kaya, manis
0	B	Adverb	Semetara, nanti
1	NP	Proper Noun	Bekasi, Indonesia
2	NG	Genetive Noun	Bukunya
3	NU	Uncountabel nouns	Air, beras
4	NC	Countable nouns	Buku, rumah
5	BI	Intranstive Verd	Pergi
6	BT	Transitif Verb	Membeli
7	N	Preposition	Di, ke, dari
8	D	Modal	Bisa
9	C	Coor-conjunction	Dan, atau, tetapi
0	C	Subor-conjunction	Jika, ketika
1	T	Determiner	Para, ini, itu
2	H	Interjection	Wah, aduh, Oi
3	DO	Ordinal Numerals	Pertama, kedua
4	DC	Colective Numerals	Bertiga

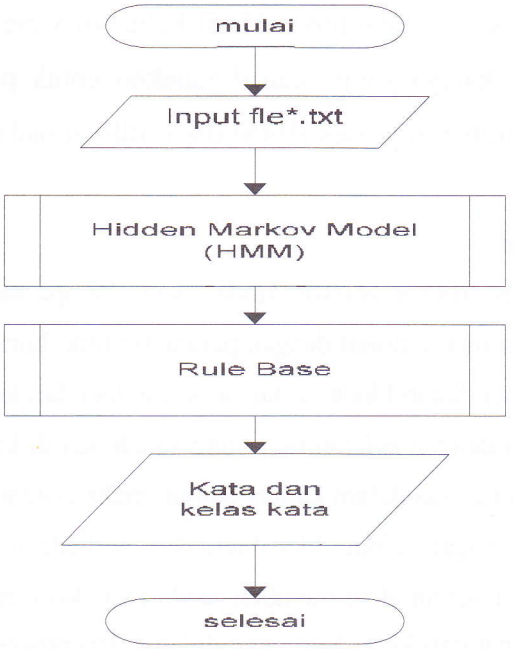
Tabel 2. (lanjutan)
Kelas kata penelitian Pisceldo dkk.(2009)

o	ag	Tag Name	Example
5	DP	Primary Numerals	Satu, dua
6	DI	Irregular Numerals	Beberapa
7	RP	Personal pronouns	Saya, kamu
8	DT	WH-Determiners	Apa, barangsiapa
		Locative pronouns	Sini, situ,sana
9	RL		
0	RN	Number Pronouns	Kedua-duanya
1	EG	Negation	Bukan, tidak
2	YM	Symbol	@, #, %, $, ^, &
3	P	Particel	Pun, kah
4	W	Foreign Word	All, word
5	P	WH-Pronouns	Apa, siapa

2.6 Analisis Proses

Sistem *Part of Speech Tagging* pada penelitian ini menggunakan dua metode yaitu *Hidden Markov Model* (HMM) dan *Rule Based.* Pada penelitian sebelumnya yang menggunakan *Hidden Markov Model* dan hasilnya sudah sangat baik yaitu memiliki tingkat keakuratan yaitu 96,50% dengan 94,5% merupakan kata-kata yang terdapat didalam corpus dan 80,4% merupakan kata-kata yang tidak dikenali atau tidak terdapat didalam corpus (Wicaksono & Purwarianti, 2010). Saat ini penulis mencoba menggabungkan *Hidden Markov Model* dengan *Rule Based* dengan tujuan mengetahui apakah penggabungan kedua metode tersebut akan mendapatkan hasil yang baik seperti metode lainnya pada penelitian-penelitian sebelumnya khususnya metode *Hidden Markov Model* (Wicaksono & Purwarianti, 2010). Proses dimulai dengan memberikan masukkan file dengan format teks (*.txt) terhadap sistem. Selanjutnya teks masukkan akan dicari kelas kata untuk setiap kata dengan *Metode Hidden Markov Model* yaitu dengan menghitung probabilitas masing-masing kelas kata. Langkah selanjutnya adalah hasil pelabelan dari metode Hidden Markov Model akan diperhalus dengan metode *rule based. Rule* (aturan) yang digunakan sebagai acuan pengecekkan aturan sudah disusun secara otomatis dari *corpus*. Setelah dilakukan

pengecekkan terhadap aturan hasil keluaran sistem adalah kata dan kelas kata dalam susunan kalimat. Garis besar langkah yang dilakukan pada sistem ini dapat dilihat pada Gambar 1.



Gambar 1. *Flowchart Part of Speech Tagging* Teks Bahasa Indonesia

2.6.1 Proses Penyusunan Aturan

Dalam penelitian ini, penulis memilih mengambil *rule* secara otomatis. Pengambilan *rule* ini menggunakan acuan *corpus* yang sudah dibuat oleh peneliti *Part of Speech Tagging* sebelumnya, dengan tujuan *rule* yang didapat merupakan *rule* benar.

Proses awal ini merupakan proses untuk mendapatkan aturan yang nanti digunakan untuk pengecekan aturan yang diperoleh dari *Hidden Markov Model*. Proses yang akan dilakukan pada bagian ini diawali dengan memecah teks dalam *corpus* menjadi kalimat-kalimat dengan parameter tanda baca titik (.), koma (,), tanda tanya (?), tanda seru (.), titik dua (:), tanda petik dua (“), tanda petik satu (‘), dash (-/--). Setelah itu label kelas kata yang mengikuti setiap kata akan diambil dan disimpan. Label kelas kata akan dipisah pisah kedalam frasa yang terdiri dari 1 kata, 2 kata , 3 kata dan seterusnya sampai satu kalimat penuh.

2.6.2 Proses *Hidden Markov Model*

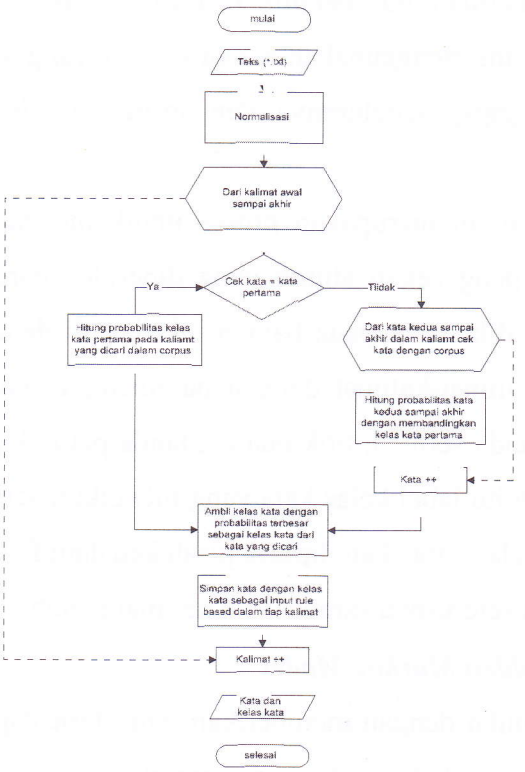
Proses dimulai dengan memberikan *input* terhadap sistem. Teks *input* akan dipecah kedalam suatu kalimat dengan parameter titik, koma, tanda tanya dan tanda seru. Kemudian setiap kata dalam kalimat akan dicari nilai probabilitas kelas

katanya terhadap kelas kata kata sebelumnya didalam *corpus*. Perhitungan probabilitas diawali dengan menghitung probabilitas kata pertama tanpa melihat kelas kata sebelumnya. Probabilitas kata kedua sampai terakhir akan dihitung dengan melihat kelas kata sebelumnya. Hasil keluaran yang dapat pada prose sini adalah kata dan kelas kataya yang akan digunakan untuk proses berikutnya yaitu proses *Rule Based*. Gambaran proses HMM dapat dilihat pada Gambar 2.

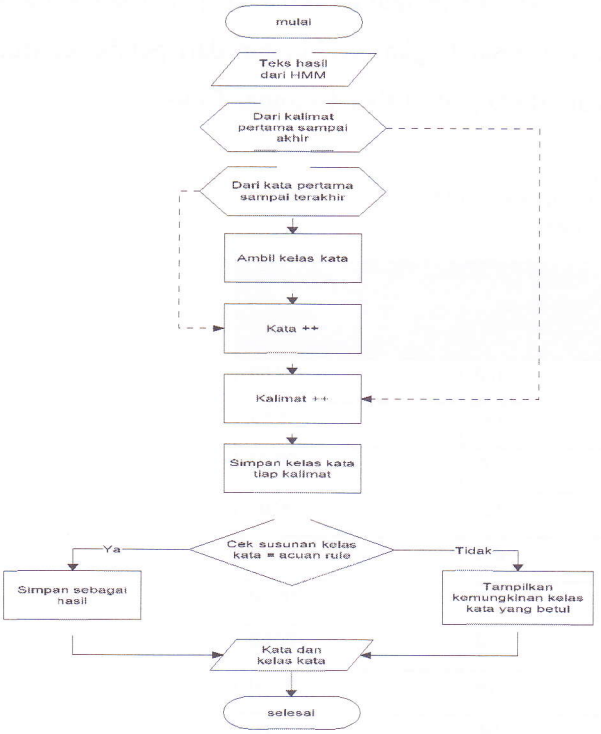
2.6.3 Proses *Rule Based*

Hasil dari proses *Hidden Markov Model* yang berupa kata berikut kelas katanya akan dipecah menjadi kalimat-kalimat dengan parameter titik, koma, tanda tanya dan tanda seru. Setelah itu kata akan diambil kelas katanya. Kemudian dari kelas kata pertama sampai terakhir akan dicocokan dengan *rule*(aturan) yang sudah ada di kamus aturan. Jika semua susunan *rule* dalam kalimat ada dalam kamus aturan, maka sistem akan menampilkan kata dengan kelas katanya sebagai output. Jika ditemukan perbedaan kelas kata dengan kelas kata dalam kamus, maka sistem akan memberi tanda pada kata tersebut dan menampilkan kelas kata yang lebih tepat dari kelas kata yang didapat dari proses *Hidden Markov Model.*

Gambaran kerja proses *Rule Based* ditunjukkan pada Gambar 3.



Gambar 2. *Flowchart* HMM



Gambar 3. Implementasi Rule Based

3.HASIL DAN PEMBAHASAN

3.1 Pengujian Penyusunan *Rule* (Aturan)

Percobaan yang dilakukan terhadap kedua corpus pada penelitian ini mendapat hasil yang baik. Kedua corpus dapat diproses menjadi susunan aturan yang nanti akan digunakan sebagai acuan susunan aturan. Tabel 3 menunjukkan hasil percobaab penyusunan *rule*. Setiap corpus yang diproes memiliki keluaran yang sama yaitu susunan aturan dengan kelas kata yang terdapat pada kelas kata masing-masing *corpus*.

Tabel 3.
Hasil percobaan pengambilan aturan

Jenis *Corpus*	Potongan teks *corpus*	Potongan susunan aturan
Corpus 1 : *Corpus* Wicaksono dan Purwarianti (2010) Jumlah kata yang dimiliki 10566. Jumlah aturan yagn diproleh 6922	Dia/PRP bangkit/VBI dari/IN keterpurukan/NN ./. Rani/NNP dan/CC Budi/NNP duduk/VBI di/IN bangku/NN ./. Mereka/PRP sedang/RB bersepeda/VBI	cc cc cdi cc cdi nn cc cdi nn nn cc cdi nn nn jj cc cdi nn nn jj vbi
Corpus 2 : *Corpus* Piceldo dkk. (2009) Jumlah kata yang dimiliki 10566. Jumlah aturan yagn diproleh 6922	Indeks/nn biaya/nnc tenaga/nnu kerja/nnu sektor/nnc swasta/jj secara/in keseluruhan/nn ,/,	cc cc cdi cc cdi rb

3.2 Pengujian Pelabelan Kelas Kata

Pada bagian ini dilakuka pengujian terhadap pelabelan kelas kata. Tujuannnya adalah mengetahui seberapa besar tingkat keakuratan dari pelabelan yang dilakukan. Hasil dari pengujian ini dapat dilihat pada Tabel 4 sampai Tabel 7.

Tabel 4.
Tabel hasil uji terhadap teks yang sama dengan teks pada *Corpus* 1

Teks Masukkan	Jumlah kata	Akurasi (%)
Teks CS1.1	100	99,00
Teks CS1.2	197	98,98
Teks CS1.3	268	100,00
Teks CS1.4	425	100,00
Teks CS1.5	293	99,65
Teks CS1.6	192	100,00
Teks CS1.7	202	100,00
Teks CS1.8	168	99,40
Teks CS1.9	220	100,00
Teks CS1.10	246	100,00
	Mean	99,70
	Standar deviasi	0,42

Tabel 5
Tabel hasil uji terhadap teks yang sama dengan teks pada *Corpus* 2

Teks Masukkan	Jumlah kata	Akurasi (%)
Teks CS2.1	165	98,79
Teks CS2.2	205	98,05
Teks CS2.3	300	97,00
Teks CS2.4	397	97,48
Teks CS2.5	260	96,41
Teks CS2.6	372	98,36
Teks CS2.7	336	96,61
Teks CS2.8	219	98,14
Teks CS2.9	316	99,39
Teks CS2.10	368	98,62
	Mean	97,89
	Standar deviasi	0,98

Tabel 6
Tabel hasil uji teks yang tidak sama dengan teks pada *Corpus* 1

Teks Masukkan	Jumlah kata	Akurasi (%)
Teks CT1.1	123	80,49
Teks CT1.2	214	62,15
Teks CT1.3	301	71,76
Teks CT1.4	431	72,85
Teks CT1.5	254	70,86
Teks CT1.6	259	67,18
Teks CT1.7	326	66,56
Teks CT1.8	202	70,29
Teks CT1.9	192	81,77
Teks CT1.10	231	84,41
	Mean	72,83
	Standar deviasi	7,22

Tabel 7.
Tabel hasil uji teks yang tidak sama dengan teks pada *Corpus* 2

Teks Masukkan	Jumlah kata	Akurasi (%)
Teks CT2.1	141	92.20
Teks CT2.2	213	90,14
Teks CT2.3	317	77,92
Teks CT2.4	443	89,39
Teks CT2.5	293	82,25
Teks CT2.6	167	86,82
Teks CT2.7	317	91,79
Teks CT2.8	348	91,37
Teks CT2.9	213	88,26
Teks CT2.10	215	89,76
	Mean	87,99
	Standar deviasi	4,58

Melihat hasil pada Tabel 4 sampai Tabel 7. Diketahui bahwa akurasi tertinggi dari POS *Tagging* untuk teks berbahasa Indonesia dengan HMM dan *Rule Based* yang diperoleh adalah 100%. Tabel 4 memiliki akurasi tertinggi 100% dengan mean sebesar 99,70% dan standar deviasi sebesar 0,42%. Pada Tabel 5, diperoleh akurasi tertinggi 99,39% dengan mean sebesar 97,89% dan standar deviasi sebesar 0,98%. Tabel 6 memiliki akurasi tertinggi 84,41%, mean 72,83% dan standar deviasi 7,22%. Hasil pada Tabel 7 diperoleh akurasi tertinggi 92,20% dengan mean sebesar 87,99% dan standar deviasi 4,58%.

Pengujian selanjutnya untuk pelabelan kata dilakukan Perbandingan POS *Tagging Hidden Markov Model* dengan POS *Tagging Hidden Markov Model* dan *Rule Based*. Pada bagian ini dilakukan uji coba dengan membandingkan antara POS *Tagging* yang menggunakan *Hidden Markov Model* saja dengan yang menggunakan *Hidden Markov Model* dan *Rule Based*.

Tabel 8.
Tabel perbandingan hasil uji terhadap teks yang sama dengan teks pada *Corpus* 1 antara HMM dengan HMM dan *Rule Based*

Teks Masukkan	Jumlah kata	Akurasi (%)	
		HMM	HMM dan Rule Based
Teks CT1.1	100	96	99
Teks CT1.2	197	97,46	98,98
Teks CT1.3	268	99,25	100
Teks CT1.4	425	99,29	100
	Mean	98,00	99.66
	Standar deviasi	1,58	0,59

Tabel 9.
Tabel perbandingan hasil uji terhadap teks yang sama dengan teks pada *Corpus* 2 antara HMM dengan HMM dan *Rule Based*

Teks Masukkan	Jumlah kata	Akurasi (%)	
		HMM	HMM dan Rule Based
Teks CT1.1	165	97,58	98,79
Teks CT1.2	205	98,05	98,05
Teks CT1.3	300	96	97
Teks CT1.4	397	96,22	97,48
	Mean	96,96	97,83
	Standar deviasi	1,01	0,77

Tabel 10.
Tabel perbandingan hasil uji terhadap teks yang tidak sama dengan teks pada *Corpus* 1 antara HMM dengan HMM dan Rule Based

Teks Masukkan	Jumlah kata	Akurasi (%)	
		HMM	HMM dan Rule Based
Teks CT1.1	123	80,49	80,49
Teks CT1.2	214	62,15	62,15
Teks CT1.3	301	71,76	71,76
Teks CT1.4	431	72,85	72,85
	Mean	71,81	71,81
	Standar deviasi	7,52	7,52

Tabel 11.
Tabel perbandingan hasil uji terhadap teks yang tidak sama dengan teks pada *Corpus* 2 antara HMM dengan HMM dan Rule Based

Teks Masukkan	Jumlah kata	Akurasi (%)	
		HMM	HMM dan Rule Based
Teks CT2.1	141	92.91	92.2
Teks CT2.2	213	90,14	90,14
Teks CT2.3	317	78,23	77,92
Teks CT2.4	397	89,62	89,39
	Mean	87,72	87,41
	Standar deviasi	6,49	6,44

Tabel 8 sampai Tabel 11 menunjukkan perbandingan antara POS Tagging dengan HMM saja dan POS Tagging dengan HMM dan Rule Based. Untuk teks masukkan yang sama dengan teks didalam corpus diketahui bahwa hasil POS Tangging dengan HMM dan *Rule Based* memiliki akurasi tertinggi 100%, sedangkan POS Tagging dengan HMM saja memiliki akurasi tertinggi 99,29%. Untuk teks masukkan yang tidak sama dengan teks dalam corpus, pada pengujian terhadap corpus 1 kedua metode memiliki hasil akurasi yang sama dan akurasi tertinggi pada percobaan ini adalah 80,48%. Sedangkan pada pengujian terhadap corpus 2 diperoleh hasil bahwa akurasi dengan HMM lebih tinggi dibanding dengan menggunakan HMM dan *Rule Based.* Pada penggunaan HMM saja diperoleh akurasi tertinggi 92,91% sedangkan pada penggunaan HMM dan *Rule Based* akurasi tertinggi yang diperoleh adalah 92,2%. Keakurasian yang sama atau bahkan turun terjadi karena perbaikan dengan hasil *rule based* tidak berjalan dengan baik. Ada beberapa kata yang tidak memiliki label (noTag) sehingga mempengaruhi pengecekkan ada proses *Rule Based.*

3.3 Perbandingan dengan *POS Tagging Hidden Markov Model*

Hasil perbandingan penelitian ini dengan penelitian sebelumnya yaitu penelitian POS *Tagging* yang dilakukan oleh (Wicaksono & Purwarianti, 2010) dapat dilihat pada Tabel 12.

Tabel 12.
Tabel perbedaan penelitian HMM *Based Part-of-Speech Tagger for* Bahasa Indonesia (Wicaksono & Purwarianti, 2010) dengan *Part of Speech Tagging* Teks Berbahasa Indonesia menggunakan Metode *Hidden Markov* dan *Rule Based*

Parameter pembanding	Wicaksono dan Purwarianti, (2010)	Kathryn Widhiyanti (2011)
Metode	Hidden Markov Model dengan pendekatan *N-Gram, Affiks-Tree* dan Leksikon	Hidden Markov Model dan *Rule Based.* *Rule Based* diperoleh dari *corpus* acuan

Tabel 12. (lanjutan)
Tabel perbedaan penelitian HMM *Based Part-of-Speech Tagger for* Bahasa Indonesia (Wicaksono & Purwarianti, 2010) dengan *Part of Speech Tagging* Teks Berbahasa Indonesia menggunakan Metode *Hidden Markov* dan *Rule Based*

Parameter pembanding	Wicaksono dan Purwarianti, (2010)	Kathryn Widhiyanti (2011)
Teks masukan	Sistem tidak bisa memproses teks masukkan yang memiliki tata cara penulisan tanda baca sesuai tata bahasa Indonesia.	Sistem bisa memproses teks masukkan yang memiliki tata cara penulisan tanda baca sesuai tata bahasa Indonesia.
Keakuratan pelabelan terhadap teks yang sama persis dengan corpus	Tingkat keakuratan tertinggi yang diperoleh dari pengujian adalah 100%	Tingkat keakuratan tertinggi yang diperoleh dari pengujiam adalah 100%
Keakuratan pelabelan terhadap teks yang tidak sama persis dengan corpus	Tingkat keakuratan tertinggi yang diperoleh dari pengujian adalah 100 %	Tingkat keakuratan tertinggi yang diperoleh dari pengujiam adalah 80,49%
Pelabelan terhadap kata yang tidak terdapat didalam corpus	Bisa memproses dan mendapatkan label yang seharusnya	Tidak bisa memproses, tetapi memberikan label noTag

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Kesimpulan dari penelitian ini adalah Pelabelan kelas kata terhadap teks berbahasa Indonesia menggunakan metode *Hidden Markov Model* dan *Rule Based* memiliki hasil keakuratan yang tinggi yaitu tertinggi 100% untuk teks yang ada didalam corpus. Jika dibandingkan dengan POS Tagging yang menggunakan HMM saja penggabungan 2 metode pada penelitian ini memberikah hasil yang lebih baik, akurasi tertinggi yang diperoleh adalah 100% untuk teks yang sama dengan corpus sedangkan POS dengan HMM saja memiliki akurasi tertinggi 99,29%. Jika dibandingkan dengan penelitian sebelumnya dengan metode HMM yang menggunakan tambahan pendekatan (Wicaksono & Purwarianti, 2010), penelitian ini masih sangat kurang dilihat dari belum bisa membedakan kata yang memiliki kelas kata ganda dan belum bisa memberikan pelabelan untuk kata yang tidak terdapat didalam corpus. Tetapi sistem ini sudah bisa memproses teks masukkan dengan tata cara penulisan yang benar. Penelitian ini memerlukan corpus yang besar agar bisa memberikan pelabelan yang lebih tepat

4.2 Saran

Saran yang dapat dilakukan pada penelitian berikutnya adalah perlu adanya pembuatan corpus yang lebih lengkap agar pelbelan yang dipeoleh bisa lebih tepat. Perlu dilakukan modifikasi langkah pengecekkan dan penambahan susunan aturan untuk penggecekkan. Ada baiknya jika dicoba suatu penelitian POS *Tagging* tanpa menggunakan corpus.

UCAPAN TERIMAKASIH

Terimakasih kepada Alfan Farizki Wicaksono dan Ayu Purwarianti yang telah mengijinkan untuk menggunakan corpus dan memakai sistem POS *Hidden Markov* sebagai sistem pembanding pada penelitian ini.

Daftar Pustaka

Alwi, H, Dardjowidjojo, S, Lapoliwa, H, Moeliono, A M. (2003). *Tata Bahasa Baku Bahasa Indonesia*. Balai Pustaka. Jakarta, Indonesia.

Jurafsky, D S. (2000). *Speech and Language Processing "An Introduction to Natural Language Processing, Cmputationak Linguistics, and Specch Reconition*. Prentice-Hall, Inc. New Jersey.

Kridalaksana, H. (2007). *Kelas Kata Dalam Bahasa Indonesia*. Ed.2. Gramedia. Jakarta.

Pisceldo, F. Andriani, M dan Manurung, R. *Probabilistic Part of Speech Tagging for* Bahasa Indonesia, Universitas Indonesia, Fakultas Ilmu Komputer.

Sukamto, R A. (2009). *Penguraian Bahasa Indonesia dengan Pengurai Collins*. Thesis. Program Magister Informatika. Institut Teknologi Bandung.

Wibisono, Y. *2008, Penggunaan Hidden Markov Model untuk Kompresi Kalimat. Tesis.* Program Magister Informatika. Institut Teknologi Bandung

Wicaksono, A F dan Purwarianti, A.(2010). *HMM Based Part-of-Speech Tagger for Bahasa Indonesia. Proceeding of the Fourth International MALINDO Workshop (MALINDO2010)*. Jakarta, Indonesia.