PENGENALAN ENTITAS BERNAMA BERDASARKAN INFORMASI KONTEKSTUAL, MORFOLOGI DAN KELAS KATA

Gatot Wahyudi dan Indra Budi

Fakultas Ilmu Komputer Universitas Indonesia, Kampus Baru UI Depok 16424
email: gatot100@mhs.cs.ui.ac.id, indra@cs.ui.ac.id

ABSTRAK

Paper ini mendeskripsikan sistem pengenalan entitas bernama pada teks berbahasa Indonesia yang disebut dengan InNER (*Indonesian Named Entity Recognizer*). InNER dikembangkan dengan pendekatan *knowledge engineering* dengan membangun aturan-aturan (*rules*) berdasarkan informasi kontekstual, leksikal, dan morfologi. Aturan dibuat dari hasil observasi pola entitas bernama yang muncul pada teks dokumen. Hasil eksperimen menunjukkan bahwa sistem InNER memberikan *recall* sebesar 63,43% dan *precision* sebesar 71,84%.

Kata kunci: InNER, ekstraksi informasi, entitas nama, *rules*, *knowledge engineering*.

Makalah diterima [28 Februari 2004]. Revisi akhir [5 April 2004].

1. PENDAHULUAN

Named entity recognition atau pengenalan entitas bernama adalah salah satu tugas dalam ekstraksi informasi (*information extraction*) yang berkaitan dengan pengenalan dan pengklasifikasian entitas bernama yang terdapat dalam teks [10]. Entitas bernama yang dikenali dalam pengenalan entitas bernama digolongkan dalam tiga kategori sebagai berikut [9]:

1. Entitas Nama: meliputi pengenalan nama orang (PERSON), nama organisasi (ORGANIZATION), nama lokasi (LOCATION)
2. Ekspresi Waktu: meliputi tanggal (DATE), jam (TIME), dan durasi (DURATION)
3. Ekspresi Bilangan: meliputi ekspresi moneter (MONEY), persentase (PERCENTAGE), ukuran (MEASURE), dan kardinal (CARDINAL).

Pengenalan entitas bernama dilakukan terhadap teks dokumen yang merupakan sumber dari informasi yang akan diambil. Misalnya dari teks berikut:

Presiden Habibie bertemu dengan Prof. Amien Rais di Jakarta kemarin.

Sistem pengenalan entitas nama akan mengidentifikasi istilah (*term*) **Habibie** dan **Amien Rais** sebagai nama orang sedangkan istilah **Jakarta** sebagai nama lokasi. Pengenalan ini dapat dilakukan berdasarkan sejumlah fitur dari istilah seperti fitur morfologi, kalimat (kontekstual), teks dan sintaknya atau dapat juga berdasarkan informasi leksikal berdasarkan kamus seperti kamus kemiripan istilah (*thesaurus*) dan kamus kata (*dictionary*).

Berbagai penelitian sudah dilakukan untuk melakukan pengenalan entitas bernama pada berbagai bahasa. Penelitian terhadap berbagai bahasa ini dilakukan karena ciri leksikal, ciri fisik, dan kontekstual berbagai bahasa tersebut dapat berbeda sehingga menyebabkan perbedaan pada teknik dan metode yang digunakan. Suatu teknik yang efektif untuk suatu bahasa belum tentu efektif untuk bahasa yang lain.

Penelitian tentang pengenalan entitas bernama untuk dokumen berbahasa Indonesia masih belum banyak dilakukan. Karena itu penelitian ini mencoba mengembangkan sistem yang dapat melakukan pengenalan entitas bernama pada teks berbahasa Indonesia yang disebut dengan *Indonesian Named Entity Recognizer* (InNER). Melalui paper ini dilaporkan hasil penelitian tentang sistem InNER yang dikembangkan dengan pendekatan *knowledge engineering*. Sistem InNER melakukan pengenalan entitas bernama berdasarkan aturan-aturan yang dibuat dari kombinasi informasi kontekstual, leksikal, dan morfologi. Entitas bernama yang dikenali oleh InNER adalah nama orang (*person*), nama organisasi (*organization*), dan nama lokasi (*location*).

Selanjutnya paper ini disusun dengan sistematika sebagai berikut. Bagian kedua menjelaskan sistem pengenalan entitas bernama, arsitektur dan aturan yang digunakan pada InNER dijelaskan pada bagian ketiga. Bagian keempat menjelaskan eksperimen awal yang dilakukan serta kesimpulan dijelaskan pada bagian akhir paper ini.

2. SISTEM PENGENALAN ENTITAS BERNAMA

Terdapat tiga pendekatan besar dalam pengembangan sistem pengenalan entitas bernama yaitu pendekatan manual (*knowledge engineering*), otomatis (*machine learning*), dan *hybrid*. Pendekatan manual merupakan pengembangan sistem berdasarkan model dan teknik "*hand craft*" untuk mengenali kelas atau tipe entitas bernama. Secara umum model tersebut terdiri dari sejumlah aturan-aturan menggunakan *grammar* (mis. *part of speech*), sintaks (mis. istilah yang mendahului), dan fitur morfologi (mis. penggunaan huruf kapital) yang dikombinasikan dengan kamus kata dan *thesaurus*. Contoh aturan yang dapat dibuat

penggunaan huruf kapital) yang dikombinasikan dengan kamus kata dan *thesaurus*. Contoh aturan yang dapat dibuat untuk melakukan pengenalan entitas bernama bertipe nama orang adalah "Jika terdapat *proper noun* (kata benda yang berawalan huruf kapital) yang didahului oleh titel/gelar maka *proper noun* tersebut adalah nama orang".

Pada pendekatan manual ini, Appelt mengenalkan sebuah sistem pengindentifikasian nama berdasarkan ekspresi regular yang ditulis secara manual [2]. Iwanska [7] menggunakan sumber daya yang lebih luas seperti *gazetteers*, dan *white and yellow pages*. Untuk tujuan yang sama, Morgan menggunakan analisa linguistik yang lebih kompleks [11]. Sistem-sistem yang dikembangkan tersebut diaplikasikan pada dokumen berbahasa Inggris.

Pengenalan entitas bernama secara otomatis dilakukan dengan menggunakan teknik *machine learning*. Untuk dokumen berbahasa Inggris terdapat beberapa teknik yang telah dikembangkan. Bikel menggunakan metode berdasarkan *hidden markov model* (HMM) dalam sistemnya yang dikenal dengan nama Nymble [3]. Borthwick [1] mendeskripsikan sistem pengenalan entitas bernama yang dibangun berdasarkan kerangka *maximum entropy*. Sistemnya menggunakan berbagai sumber pengetahuan seperti morfologi, leksikal, fitur *section* dan kamus data. Selanjutnya, Chieu [5] menggunakan *maximum entropy* dan mengkombinasikan fitur-fitur lokal tersebut dengan fitur global dalam teks seperti singkatan nama. Pada bahasa Jepang, Sekine menggunakan *decision tree* [12] untuk pengenalan entitas bernama. Selain itu, Gokhan Tur, Hakkani-Tur, dan Oflazer melakukan penelitian *name tagging* atau pengenalan nama menggunakan informasi leksikal, kontekstual, dan morfologi [4]. Mereka mengembangkan sistem untuk mengenali nama orang, nama organisasi, dan nama lokasi dari teks berbahasa Turki.

Penelitian pengenalan entitas bernama pada teks berbahasa Indonesia pernah dilakukan oleh Budi dan S. Bressan [6]. Mereka menggunakan aturan asosiasi (*association rules*) untuk melakukan pengenalan entitas nama.

Pendekatan terakhir dalam pengembangan sistem pengenalan entitas nama adalah pendekatan *hybrid* yang mengkombinasikan pendekatan manual dan otomatis. LTG (HCRC Language Technology Group) merupakan salah satu sistem pengenalan entitas nama yang dikembangkan dengan pendekatan *hybrid*. Sistem ini menggabungkan *rules* dan metode *maximum entropy* untuk melakukan pengenalan entitas nama pada teks berbahasa Inggris [8].

3. INDONESIAN NAMED ENTITY RECOGNIZER (InNER)

Indonesian Named Entity Recognizer merupakan sistem pengenalan entitas bernama pada teks berbahasa Indonesia yang dikembangkan menggunakan pendekatan *knowledge engineering*, yaitu dengan membangun aturan-aturan (*rules*) pengenalan entitas bernama berdasarkan hasil observasi pada dokumen. Dokumen observasi yang digunakan merupakan kumpulan berita pada koran berbahasa Indonesia. Aturan-aturan sistem dibuat dengan mempelajari pola entitas bernama yang terdapat pada kalimat. Selanjutnya aturan dibuat berdasarkan informasi kontekstual, leksikal, dan informasi morfologi yang dimiliki oleh token entitas benama pada dokumen observasi.

Informasi kontekstual mencakup informasi konteks kalimat yang dapat dimanfaatkan untuk mengenali nama atau entitas. Misalnya, dalam konteks kalimat yang mengandung kata titel atau gelar seperti kata "Prof.", maka secara umum kata berikutnya dapat digolongkan sebagai nama orang. Sebagai contoh adalah pada kalimat berikut:

Prof. Habibie berkunjung ke Surabaya

Pada kalimat di atas, kata "Habibie" akan ditandai sebagai nama orang karena dalam konteks kalimat tersebut kata "Habibie" diawali dengan kata Prof.".

Informasi leksikal yang digunakan mencakup *string* token dan kelas kata token. Informasi leksikal suatu token ditentukan dengan menggunakan kamus kata bahasa Indonesia. Contoh informasi leksikal, misalnya, token "jeruk" adalah token dengan *string* "jeruk" dan mempunyai kelas kata sebagai kata benda.

Informasi morfologi diperoleh dengan menganalisa ciri fisik token. Analisa ciri fisik token bertujuan untuk mengetahui karakter-karakter pembentuk token. Misalnya, token "saya" terdiri dari huruf kecil semua, token "6100" adalah token yang terdiri dari angka, token "7.500" terdiri dari titik dan angka, dan sebagainya.

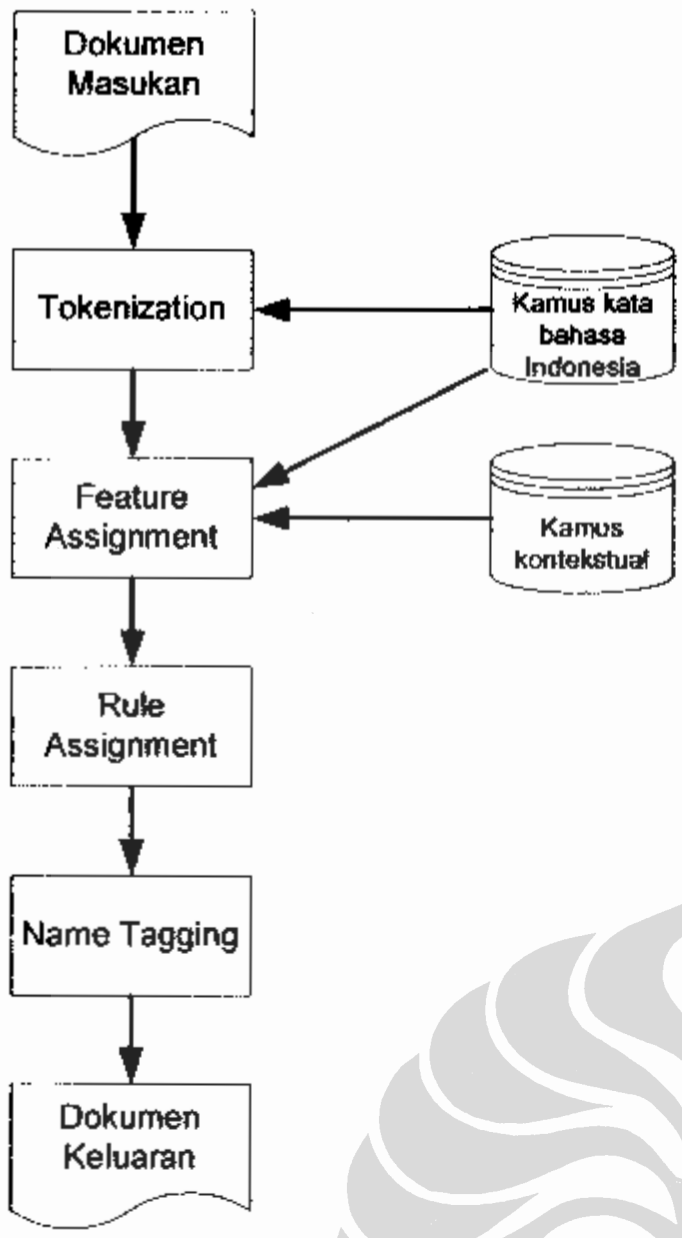
Berdasarkan informasi kontekstual, leksikal, morfologi tersebut maka dapat dibuat aturan-aturan sistem yang digunakan untuk melakukan pengenalan entitas bernama. Contoh aturan yang dibangun antara lain:

> **Jika** terdapat token titel orang misalnya "Prof." dan diikuti token yang diawali oleh huruf besar **maka** kata yang mengikuti tersebut akan ditandai sebagai nama orang.
>
> **Jika** terdapat kata depan misalnya "di" dan diikuti oleh kata yang berawalan huruf besar **maka** kata yang mengikuti tersebut ditandai sebagai nama lokasi.

Proses pengenalan entitas bernama pada sistem InNER dilakukan dengan mencocokkan setiap token dari suatu kalimat dengan setiap aturan yang sudah diproduksi untuk menentukan apakah token tersebut dapat diklasifikasikan sebagai suatu entitas nama atau bukan.

3.1. Arsitektur Sistem

Sistem pengenal entitas bernama pada teks berbahasa Indonesia (InNER) mempunyai empat proses utama. Alur proses sistem dapat dilihat pada Gambar 1.



Gambar 1. Arsitektur Sistem InNER

Sistem menerima dokumen masukan berupa teks kalimat berbahasa Indonesia, misalnya:

Presiden Habibie bertemu dengan Prof. Amien Rais di Jakarta kemarin.

Pada proses tokenisasi, kalimat tersebut diuraikan menjadi token-token penyusunnya. Penguraian kalimat kedalam token-token dilakukan dengan menggunakan bantuan kamus kata bahasa Indonesia. Setelah token-token diperoleh, sistem menganalisa fitur-fitur yang dimiliki setiap token menggunakan bantuan kamus kata dan kamus kontekstual atau kamus pemicu yang berisi kata-kata kontekstual entitas bernama. Selanjutnya pada proses *rule assignment* sistem menerapkan aturan-aturan pada setiap token untuk menentukan tipe entitas bernama dari token. Berdasarkan hasil *rule assignment* tersebut proses *name tagging* memberikan *tag* entitas bernama yang sesuai. Keluaran dari sistem InNER adalah dokumen yang merupakan dokumen masukan yang entitas bernama di dalamnya sudah ditandai dengan *tag* entitas bernama yang sesuai, yaitu:

Presiden <ENAMEX TYPE="PERSON">Habibie</ENAMEX> bertemu dengan Prof. <ENAMEX TYPE="PERSON">Amien Rais</ENAMEX> di <ENAMEX TYPE="LOCATION">Jakarta</ENAMEX> kemarin.

3.2. *Tokenization*

Proses tokenisasi (*tokenization*) meliputi proses pembacaan file teks masukan yang dilakukan per kalimat. Setiap kalimat yang dibaca kemudian diuraikan menjadi token-token penyusunnya. Proses inilah yang merupakan proses inti tokenisasi. Pada proses tokenisasi dilakukan penyimpanan informasi dasar token yaitu *string* dari token dan jenis token. Pemisah antar token adalah karakter spasi sehingga spasi tidak dijadikan sebagai token. Karakter khusus yang menjadi pemisah token sekaligus menjadi token sendiri antara lain tanda baca seperti karakter titik, koma, tanda kurung, tanda seru, tanda tanya, serta simbol seperti tanda dolar dan persen.

Sebagai contoh, kalimat berikut:

Ketua MPR, Amien Rais pergi ke Bandung kemarin (24/4).

Proses tokenisasi terhadap kalimat di atas memberikan hasil seperti ditunjukkan pada Tabel 1.

3.3. *Feature Assigment*

Pada proses ini, setiap token hasil proses tokenisasi dianalisa untuk mengetahui fitur-fitur apa saja yang dimiliki oleh token tersebut. Fitur-fitur yang dianalisa meliputi fitur kontekstual, fitur leksikal, dan fitur morfologi. Fitur kontekstual dianalisa menggunakan bantuan kamus kontekstual atau kamus pemicu entitas bernama. Fitur leksikal dianalisa menggunakan bantuan kamus kata bahasa Indonesia sedangkan fitur morfologi dianalisa berdasarjan karakter-karakter pembentuk token. Daftar fitur yang digunakan pada penelitian ini selengkapnya dapat dilihat pada Tabel 2, Tabel 3 dan Tabel 4.

Tabel 1. Contoh Hasil Tokenisasi

String Token	Jenis Token
Ketua	WORD
MPR	WORD
,	OPUNC
Amien	WORD
Rais	WORD
pergi	WORD
ke	WORD
Bandung	WORD
kemarin	WORD
(	SPUNC
24/4	WORD
)	EPUNC

Tabel 2. Fitur-fitur Kontekstual

Fitur	Keterangan	Contoh
PPRE	kata yang mendahului nama orang	Dr., Pak, K.H.,
PMID	kata yang merupakan nama tengah dari nama orang	bin, van
PSUF	kata yang mengikuti nama orang	SKom, SH
PTIT	kata jabatan yang mendahului nama orang	Menristek, Mendagri
OPRE	kata yang menjadi awal nama organisasi	PT., Universitas
OSUF	kata yang menjadi akhir dari nama organisasi	Ltd., Company
OPOS	kata jabatan yang muncul sebelum nama organisasi	Ketua
OCON	kontekstual lain pada nama organisasi	Muktamar, Rakernas
LPRE	awal dari nama lokasi	Kota, Propinsi
LSUF	akhir dari nama lokasi	Utara, City
LLDR	jabatan yang mendahului nama lokasi	Gubernur, Walikota
POLP	kata depan yang mendahului nama orang	oleh, untuk
LOPP	kata depan yang mendahului nama lokasi	di, ke, dari
DAY	nama hari	Senin, Sabtu
MONTH	nama bulan	April, Mei

Tabel 3. Fitur-fitur Leksikal

Fitur	Arti	Contoh
ART	kata artikula	Si, Sang
ADJ	kata sifat	indah, baik
ADV	kata keterangan	telah, kemarin
AUX	kata bantu	harus
C	kata penghubung	dan, atau, lalu
DEF	kata definisi	merupakan
NOUN	kata benda	rumah, gedung
NOUNP	kata benda yang merujuk manusia	ayah, ibu
NUM	bilangan	satu, dua
MODAL	modal	akan
OOV	kata tidak terdapat dalam kamus kata	
PAR	partikel	kah, pun
PREP	kata depan	di, ke, dari
PRO	pronomina	saya, beliau
VACT	kata kerja aktif	menuduh
VPAS	kata kerja pasif	dituduh
VERB	kata kerja	pergi, tidur

Tabel 4. Fitur-fitur Morfologi

Fitur	Contoh
TitleCase	Soedirman
UpperCase	KPU
LowerCase	menuntut
MixedCase	LeIP
CapStart	LeIP, Muhammad
CharDigit	P3K
Digit	2004
DigitSlash	17/5
Numeric	20,5; 17.500,00
NumStr	satu, tujuh, lima
Roman	VII, XI
TimeForm	17:05, 19.30

3.4. *Rule Assigment*

Pada proses *rule assignment*, setiap token yang sudah dilengkapi informasi fitur lalu digolongkan ke suatu entitas bernama tertentu. Penggolongan ini ditentukan berdasarkan aturan-aturan (*rules*) yang sudah dibangun. Setiap token akan diuji dengan setiap aturan sistem untuk menentukan aturan mana yang akan menggolongkan token tersebut ke suatu entitas bernama tertentu.

Aturan-aturan pengenalan entitas bernama pada sistem InNER dibangun dengan cara mengobservasi pola entitas bernama yang muncul pada dokumen observasi. Bentuk aturan atau *rule* adalah aturan bersyarat. Sebuah aturan menguji apakah suatu token memenuhi syarat yang ditentukan sehingga token tersebut dapat digolongkan ke suatu entitas bernama tertentu. Syarat yang terdapat pada suatu aturan menunjukkan informasi atau fitur-fitur yang harus dipenuhi oleh suatu token agar dapat digolongkan ke suatu entitas bernama tertentu. Contoh aturan yang dibangun adalah sebagai berikut:

Jika token yang sedang diperiksa adalah **ORGP dan** token sesudahnya berjenis **WORD dan** berbentuk **TitleCase**
maka token yang sedang diperiksa digolongkan sebagai **ORGANIZATION dan** token sesudahnya juga digolongkan sebagai **ORGANIZATION**

Tabel 5 menunjukkan hasil proses *rule assignment* terhadap token-token hasil proses tokenisasi dan *feature assignment* sebelumnya. *String* kosong ("") pada kolom jenis entitas bernama Tabel 5 menunjukkan bahwa token tidak termasuk entitas bernama apapun. *String* **ORGANIZATION** menunjukkan bahwa token merupakan bagian dari entitas bernama dengan tipe nama organisasi. *String* **PERSON** menunjukkan bahwa token merupakan bagian dari nama orang sedangkan *string* **LOCATION** menunjukkan bahwa token merupakan bagian dari nama lokasi.

Tabel 5. Hasil *Rule Assignment*

Token	Jenis Entitas Bernama
Ketua	"
MPR	ORGANIZATION
,	"
Amien	PERSON
Rais	PERSON
pergi	"
ke	"
Bandung	LOCATION
kemarin	"
(	"
24/4	"
)	"

3.5. *Name Tagging*

Proses terakhir pada sistem InNER adalah proses pengelompokkan deret token yang mempunyai jenis atau tipe entitas bernama yang sama. Selanjutnya kelompok token tersebut ditandai dengan *tag* pembuka entitas bernama sebelum token pertama dan ditandai dengan *tag* penutup entitas bernama setelah token terakhir.

Bentuk *tag* yang digunakan untuk menandai entitas bernama dalam penelitian ini mengacu pada bentuk *tag* standar *name entity recognition* [2]. Bentuk *tag* untuk entitas nama (nama orang, nama organisasi, nama lokasi) adalah sebagai berikut:

<ENAMEX TYPE="N">Frase Nama</ENAMEX>

N yang terdapat dalam *tag* di atas menyatakan tipe dari entitas nama yang dapat berupa **PERSON, ORGANIZATION** atau **LOCATION. Frase Nama** menyatakan frase yang ditandai sebagai entitas nama. Contoh penggunaan *tag* tersebut adalah sebagai berikut:

<ENAMEX TYPE="PERSON">Nur Laila </ENAMEX>

Berdasarkan aturan *tagging* di atas, mengacu pada contoh dari hasil tahap *rule assignment* sebelumnya, maka diperoleh hasil proses *name tagging* sebagai berikut:

Ketua <ENAMEX TYPE="ORGANIZATION">MPR</ENAMEX>, <ENAMEX TYPE="PERSON">Amien Rais</ENAMEX> pergi ke <ENAMEX TYPE="LOCATION">Bandung</ENAMEX> kemarin (24/4)

4. EKSPERIMEN

Bagian ini menjelaskan eksperimen sistem InNER dalam melakukan pengenalan entitas bernama pada teks berbahasa Indonesia. Pengenalan yang dilakukan sistem InNER meliputi pengenalan nama orang (*person*), nama organisasi (*organization*), dan nama lokasi (*location*).

4.1. Dokumen Observasi dan Uji Coba

Dokumen observasi digunakan untuk mengambil kata-kata kontekstual yang akan disimpan dalam kamus pemicu entitas bernama. Selain itu dokumen observasi juga digunakan untuk menemukan pola entitas bernama yang dijadikan dasar dalam pembuatan aturan penegnalan sistem. Dokumen observasi terdiri dari 802 kalimat yang mengandung 559 nama orang, 853 nama organisasi, dan 418 nama lokasi. Dokumen uji coba terdiri dari 1.258 kalimat yang mengandung 801 nama orang, 1.031 nama organisasi, dan 297 nama lokasi.

Dokumen observasi dan uji coba yang digunakan dalam penelitian merupakan kumpulan artikel koran berbahasa Indonesia versi *online* (www.kompas.com, www.republika.co.id).

4.2. Evaluasi Kinerja

Evaluasi dilakukan untuk mengetahui kinerja sistem melalui nilai parameter kinerjanya. Proses evaluasi sistem mengacu pada proses evaluasi sistem pada MUC (*Message Understanding Conference*).

Nilai parameter kinerja sistem dihitung dengan melibatkan nilai-nilai berikut:

- ***Correct***: jumlah pengenalan bernilai benar yang dilakukan sistem
- ***Partial***: jumlah pengenalan kurang tepat yang dilakukan sistem
- ***Actual***: jumlah entitas nama yang seharusnya dapat dikenali sistem
- ***Possible***: jumlah keseluruhan pengenalan yang dilakukan sistem

Berdasarkan nilai-nilai tersebut dapat dihitung nilai parameter kinerja sistem yaitu *recall* dan *precision*. Rumusan kedua parameter kinerja sistem tersebut adalah sebagai berikut:

$$Recall = \frac{Correct + 0{,}5*Partial}{Possible}$$

$$Precision = \frac{Correct + 0{,}5*Partial}{Actual}$$

4.3. Hasil Eksperimen dan Analisa

Dalam eksperimen dibuat empat jenis aturan pengenalan entitas bernama yang digunakan dalam sistem InNER yaitu sebagai berikut:

1. Aturan Kontekstual: merupakan aturan pengenalan entitas bernama yang dibuat berdasarkan fitur kontekstual saja.

2. Aturan KL: merupakan aturan pengenalan entitas bernama yang mengkombinasikan fitur kontekstual dan leksikal (kelas kata).
3. Aturan KM: merupakan aturan pengenalan entitas bernama yang mengkombinasikan fitur kontekstual dan fitur morfologi.
4. Aturan KMK: merupakan aturan pengenalan entitas bernama yang mengkombinasikan fitur kontekstual, morfologi, dan leksikal.

Aturan Kontekstual merupakan cara dasar pengenalan entitas bernama yang dilakukan oleh sistem InNER. Aturan-aturan lainnya dibuat untuk mengetahui peningkatan kinerja sistem jika dilakukan penambahan fitur leksikal dan morfologi. Hasil pengenalan entitas bernama oleh sistem InNER dengan masing-masing aturan tersebut dapat dilihat pada Tabel 6 sedangkan peningkatan kinerja yang diperoleh dengan penambahan fitur dapat dilihat pada Tabel 7.

Tabel 6. Hasil Kinerja Sistem

Aturan	*Recall*	*Precision*
Kontekstual	34,50%	33,52%
KL	46,81%	49,80%
KM	47,91%	70,30%
KMK	63,43%	71,84%

Tabel 7. Peningkatan Kinerja Sistem Akibat Penambahan Fitur dalam Aturan

Aturan	Peningkatan	
	Recall	*Precision*
KL	12,31%	16,28%
KM	13,41%	36,78%
KMK	28,93%	38,32%

Dari tabel 6 dapat dilihat bahwa sistem dengan aturan yang hanya menggunakan fitur kontekstual memberikan kinerja paling rendah yaitu dengan *recall* 34,50% dan *precision* 33,52%. Kombinasi fitur kontekstual dan leksikal dalam aturan KK mampu memberikan peningkatan *recall* sebesar 12,31% dan peningkatan *precision* sebesar 16,28%.. Aturan KM yang merupakan kombinasi fitur kontekstual dan morfologi memberikan kinerja lebih baik yaitu dengan peningkatan *recall* sebesar 13,41% dan peningkatan *precision* sebesar 36,78%. Kombinasi fitur kontekstual, leksikal, dan morfologi dalam aturan KLM memberikan peningkatan *recall* sebesar 28,93% dan peningkatan *precision* sebesar 38,32%, Sistem dengan aturan KLM memberikan kinerja terbaik dengan *recall* 63,43% dan *precision* 71,84%.

Perbedaan kinerja masing-masing aturan dapat dilihat dari perbedaan hasil pengenalan oleh aturan-aturan tersebut terhadap kalimat berikut ini:

Ikatan Mahasiswa Peduli Papua mendesak Megawati bersikap tegas tentang masa depan Papua.

Dengan masukan berupa kalimat di atas, sistem seharusnya dapat menganali frase *Ikatan Mahasiswa Peduli Papua* sebagai nama organisasi, token *Megawati* sebagai nama orang, dan token *Papua* sebagai nama lokasi.

Pada penggunaan aturan Kontekstual, pengenalan entitas bernama hanya dapat dilakukan berdasarkan token yang menjadi kontekstual dan satu token sebelum atau sesudahnya. Karena itu aturan Kontekstual hanya dapat mengenali frase *Ikatan Mahasiswa* sebagai nama organisasi karena token *Ikatan* merupakan kata kontekstual nama organisasi. Aturan ini juga tidak dapat mengenali token *Megawati* dan *Papua* sebagai nama orang dan nama lokasi karena token tersebut tidak memiliki informasi kontekstual.

Penambahan fitur leksikal pada aturan KL memungkinkan token *Megawati* dapat dikenali sebagai nama orang dengan memanfaatkan posisinya yang didahului oleh kata kerja aktif. Namun, aturan KL juga tidak dapat mengenali token *Papua* sebagai nama lokasi karena tidak mengandung informasi leksikal misalnya didahului oleh kata depan. Aturan KL juga hanya dapat mengenali frase *Ikatan Mahasiswa* sebagai nama organisasi. Frase *Peduli Papua* tidak dikenali karena token *Peduli* merupakan kata kerja sedangkan nama organisasi jarang mengandung kata kerja.

Kombinasi fitur kontekstual dan morfologi pada aturan KM ternyata memberikan pengenalan nama organisasi yang lebih tepat pada frase *Ikatan Mahasiswa Peduli Papua*. Namun, aturan KM juga tidak dapat mengenali token *Megawati* dan *Papua* sebagai nama orang dan nama lokasi karena token tersebut tidak mengandung informasi kontekstual yang dapat digunakan untuk menentukan tipe entitas bernama.

Berdasarkan penjelasan di atas dapat diketahui bahwa penggunaan fitur leksikal memungkinkan token *Megawati* dikenali sebagai nama orang sedangkan penggunaan fitur morfologi memungkinkan pengenalan yang tepat pada frase *Ikatan Mahasiswa Peduli Papua*. Aturan KLM berhasil menggabungkan kelebihan yang dimiliki oleh dua fitur tersebut sehingga aturan KLM dapat mengenali dua jenis entitas bernama tersebut dengan tepat. Namun, aturan KLM juga belum dapat mengenali token *Papua* sebagai nama lokasi karena tidak adanya informasi kontekstual, leksikal maupun informasi morfologi yang dapat digunakan untuk menentukan jenis entitas bernama dari frase *Papua* tesebut.

Berdasarkan penjelasan di atas, sistem InNER melakukan pengenalan entitas bernama berdasarkan informasi kontekstual. Kemudian dengan menambahkan fitur leksikal diperoleh peningkatan *recall* sistem. Peningkatan *recall* ini diperoleh karena fitur leksikal dapat dimanfaatkan untuk mengenali entitas bernama yang tidak memiliki informasi kontekstual. Misalnya, pada frase *di Jakarta* maka token *Jakarta* dapat dikenali sebagai nama lokasi karena didahului oleh token yang memiliki informasi leksikal berupa kata depan. Penambahan fitur morfologi memberikan peningkatan *precision* karena sistem dapat mengenali frase entitas bernama lebih tepat dengan

memperhatikan kesamaan fitur morfologi. Misalnya, jika terdapat token yang mengikuti nama organisasi dan token tersebut memiliki ciri morfologi yang sama maka token tersebut akan dikenali sebagai bagian nama organisasi juga. Dengan menggabungkan ketiga fitur yang mempunyai kelebihan yang berbeda tersebut maka diperoleh sistem InNER dengan kinerja terbaik.

5. KESIMPULAN

Berdasarkan hasil eksperimen, sistem InNER dapat melakukan pengenalan entitas bernama pada teks berbahasa Indonesia dengan *recall* sebesar 63,43% dan *precision* sebesar 71,84%. Kinerja terbaik sistem InNER tersebut diperoleh dengan aturan yang mengkombinasikan fitur kontekstual, leksikal, dan morfologi.

Dari hasil eksperimen juga dapat diketahui bahwa kinerja sistem InNER masih belum maksimal. Hal ini kemungkinan disebabkan oleh keragaman pola entitas bernama yang muncul dalam teks dokumen yang juga seringkali tidak sesuai dengan tata bahasa baku sehingga terdapat pola-pola tertentu tidak dapat dikenali oleh aturan sistem. Hal ini kemungkinan dapat diatasi dengan menambahkan aturan-aturan yang menangani pola-pola tersebut. Di samping itu, karena sistem InNER menggunakan informasi kontekstual sebagai dasar pengenalan maka kinerja sistem akan dipengaruhi oleh jumlah koleksi kata kontekstual yang digunakan.

Sistem InNER yang dikembangkan dalam penelitian ini diharapkan dapat digunakan untuk membantu menemukan informasi penting dari suatu dokumen secara otomatis misalnya mengambil informasi nama orang, nama organisasi, dan nama lokasi yang terlibat pada suatu kejadian atau *event*. Selain itu sistem ini juga dapat digunakan untuk mendukung tugas selanjutnya dalam ekstraksi informasi yaitu tugas *coreference recognition*, *template element*, *template relation*, dan *scenario template*.

REFERENSI

[1] A. Borthwick, et. al., Exploiting diverse knowledge sources via maximum entropy in named entity recognition, *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, Canada, 1998.

[2] D. Appelt dan et. al., SRI International FASTUS system MUC-6 test results and analysis, *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, NIST, Morgan-Kaufmann Publisher, Columbia, 1995.

[3] D. Bikel, dan et. al., NYMBLE: A High Performance Learning Name-Finder, *Proceeding of the fifth Conference on Applied Natural Language Processing*, pp 194-201, 1997.

[4] G. Tur, D. Z. Hakkani-Tur, dan K. Oflazer, Name Tagging Using Lexical, Contextual, and Morphological Information, *Workshop on Information Extraction Meets Corpus Linguistics LREC-2000, 2nd International Conf. Language Resources and Evaluation*, Athens, Greece, 31 May - 2 June 2000.

[5] H.L. Chieu dan Hwee Tou Ng,. Named Entity Recognition: A Maximum Entropy Approach Using Global Information, *Proceedings of the 19th International Conference on Computational Linguistics*, 2002.

[6] I. Budi dan S. Bressan, Association Rules Mining for Name Entity Recognition, *Proceeding of 2003 WISE Conference*, Roma, 2003.

[7] L. Iwanska dan et. al, Wayne state university: Description of the UNO natural language processing system as used for MUC-6, *Proceeding of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD, NIST, Morgan-Kaufmann Publishers, 1995.

[8] A. Mikheev, C. Grover, dan M. Moen, Description of the LTG System Used for MUC-7, *Proceeding of the MUC-7*.

[9] N. Chinchor dan et. al, 1999 Named Entity Recognition Task Definition Version 1.4, The MITRE Corporation and SAIC, 1999.

[10] R. Grishman, Information Extraction: Techniques and Challenges, *Lecture Notes in Computer Science Vol. 1299*, Springer-Verlag, 1997.

[11] R. Morgan dan et. al., University of durham: Description of the LOLITA system as used for MUC-6, *Proceeding of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD, NIST, Morgan-Kaufmann Publishers, 1995.

[12] S. Sekine, R. Grishman and H. Shinnou, A Decision Tree Method for Finding and Classifying Names in Japanese Texts, *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, Canada, 1998.