PERBANDINGAN DECISION TREE, MAXIMUM ENTROPY, DAN ASSOCIATION RULES PADA RESOLUSI KOREFERENSI UNTUK BAHASA INDONESIA

Astria Kurniawan Sumantri[1], Indra Budi[2], Heri Kurniawan[2]
[1,2,3] Fakultas Ilmu Komputer,Universitas Indonesia
Telp. 7863419 ext (3200), Fax. 7863415
Email: Aks41@ui.edu, indra@cs.ui.ac.id, herik@cs.ui.ac.id

ABSTRAKS
Metode Decision Tree, Maximum Entropy, dan Association Rules pernah diaplikasikan untuk melakukan penelitian resolusi koreferensi untuk Bahasa Inggris. Sedangkan penelitian resolusi koreferensi untuk Bahasa Indonesia baru dilakukan dengan menggunakan Association Rules. Penelitian ini dilakukan untuk membandingkan ketiga metode untuk melakukan resolusi koreferensi. Pengujian dilakukan terhadap 500 dokumen yang diambil dari sumber koran online.

Kata Kunci: resolusi koreferensi, Decision Tree, Maximum Entropy, Association Rules

1. PENDAHULUAN

Karena kesibukannya yang semakin bertambah, manusia tidak dapat membaca informasi dalam teks secara keseluruhan. Hal ini mendorong para peneliti untuk memperoleh informasi yang terkandung dalam teks secara otomatis. Hal ini dilakukan dengan cara menyajikan informasi yang terkandung dalam teks tersebut dengan lebih ringkas, tanpa mengurangi makna/isi informasi tersebut.

Sebuah sistem yang menggunakan teks sebagai masukan dan menghasilkan data terstruktur dengan format tertentu disebut sebagai sistem ekstraksi informasi. Beragam informasi yang dapat diambil dari teks masukan, contohnya antara lain: nama orang dan nama tempat. Untuk yang lebih canggih, skenario kejadian, dan bahkan hasi kesimpulan dari teks dapat diambil oleh sistem. Penelitian mengenai ekstraksi informasi ini cukup pesat berkembang setelah diadakannya *Message Understanding Conference (MUC)*. Menurut MUC, terdapat lima tahapan untuk mencapai sistem ekstraksi informasi yang utuh, antara lain: *named entity recognition* (pengenalan entitas bernama), *coreference resolution* (resolusi koreferensi), *template element construction*, *template relation construction* dan *scenario template production* (Chinchor, 1998).

Tugas dari resolusi koreferensi sendiri adalah untuk mengelompokkan satu atau lebih entitas bernama atau kata ganti yang merujuk kepada entitas bernama yang lain (MUC, 1998).

Misalkan terdapat kalimat:

> Irfan Fahmi$_1$ seorang sarjana psikologi. Dia$_2$ lulusan perguruan tinggi terkemuka di Kota Bandung. Semua orang di kampung sangat menghormati Irfan$_3$.

Pada kalimat di atas, frase "Irfan Fahmi$_1$" dirujuk oleh frase "dia$_2$" dan "Irfan$_3$". Resolusi koreferensi sangat penting untuk berbagai bidang dalam pemrosesan bahasa alami seperti ekstraksi informasi, *text summarization*, *question answering* dan lain-lain.

Penelitian ekstraksi informasi untuk bahasa asing telah banyak dilakukan, namun tidak demikian halnya untuk Bahasa Indonesia. Penelitian sistem ekstraksi informasi untuk Bahasa Indonesia telah sampai pada tahap *scenario template*, yang dilakukan oleh Marjuki dengan menggunakan *vector machine*. Budi, Markus, dan Wahyudi melakukan penelitian untuk mengenali entitas bernama dengan menggunakan metode yang berbeda. Penelitian pengenalan entitas bernama yang lain dilakukan oleh Wibowo dengan melanjutkan penelitian Budi dengan menambahkan kemampuan *multiple features*. Pengenalan *event extraction* dilakukan oleh Amin dengan menggunakan tiga pendekatan, yaitu *decision tree, neural network,* dan *association rules.* Resolusi koreferensi oleh Nasrullah dengan menggunakan metode *association rules* (Marjuki, 2006) (Budi, 2003) (Gatot, 2004) (Wibowo, 2005) (Amin, 2006) (Markus, 2007) (Nasrullah, 2005)

2. DT, ME, DAN AR

Dalam melaksanakan tugas resolusi koreferensi, terdapat dua pendekatan yang dilakukan, yaitu *knowledge engineering* dan *machine learning.* Metode *knowledge engineering* menggunakan aturan-aturan yang diperoleh dari pengamatan yang dilakukan oleh para ahli. Sedangkan metode *machine learning* memperoleh aturan-aturan tersebut secara otomatis. Metode-metode yang sering digunakan dalam pendekatan *machine learning* antara lain: *decision tree, maximum likelihood, association rules, support vector machine, maximum entropy,* dsb (Wu, 2002).

Metode *decision tree* pernah digunakan dalam melaksanakan tugas resolusi koreferensi dalam Bahasa Inggris. Metode ini menggunakan struktur

data *tree* dalam pegambilan keputusan. *Tree* dibangun dengan menggunakan algoritma C4.5 dengan menggunakan prinsip *information gain*, yaitu berapa banyak informasi yang benar yang dapat diperoleh dari dokumen pelatihan untuk suatu ciri tertentu. Dalam *information gain* ini dikenal adanya istilah *entropy*, yang merupakan derajat ketidakpastian dari suatu kondisi (Manning, 1999).

Entropy dituliskan dengan rumus:
$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$

Sedangkan rumus dari *information gain* sendiri adalah sebagai berikut:
$I = 1 - \Sigma H(p)$

Metode *maximum entropy* menggunakan statistika dalam prosesnya. Dokumen pelatihan yang dimasukkan ke dalam sistem akan digunakan untuk menciptakan suatu model melalui proses yang disebut *Generalized Iterative Scaling* (GIS). Resolusi koreferensi pada Bahasa Inggris dengan menggunakan metode *maximum entropy* pernah dilakukan oleh Denis dan Baldridge. Untuk Bahasa Indonesia, Markus membandingkan metode ini dengan metode *association rules* dalam penelitian untuk mengenali entitas bernama (Manning, 1999).

Untuk sejumlah fitur dan data pelatihan yang digunakan dalam penelitian, kita menghitung *conditional probability* untuk suatu keadaan *(y|x)* sebagai

$$P(y \mid x) = \frac{\prod_i \alpha_i^{f_i(x,y)}}{Z_\alpha(x)}$$

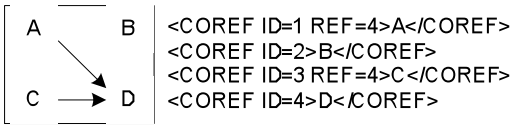
Di mana $Z_\alpha(x) = \sum_y \prod_i \alpha_i^{f_i(x,y)}$

Algoritma *Generalized Iterative Scaling* (GIS) digunakan untuk mencari nilai α_i untuk suatu fitur.

Association rules menggunakan aturan-aturan dalam mengenali apakah pasangan frase yang sedang diamati merupakan pasangan yang saling ekivalen atau tidak. Aturan-aturan tersebut diperoleh melalui tahap pelatihan dari sejumlah koleksi dokumen yang dilatih. *Association rules* merupakan sekumpulan aturan dengan setiap aturan dinyatakan dalam hubungan X Y (dibaca: jika X maka Y). Simbol X merepresentasikan ciri-ciri dari suatu pasangan frase, sedangkan Y untuk menyatakan pasangan frase tersebut saling ekivalen atau tidak. Nasrullah pernah melaksanakan tugas resolusi koreferensi dengan menggunakan metode ini dengan nilai *F-measure* tertinggi sebesar 79,68 (Nasrullah, 2005).

3. DOKUMEN PENELITIAN

Penelitian menggunakan 500 dokumen yang diambil dari www.kompas.com. 300 dokumen digunakan sebagai dokumen pelatihan sedangkan sisanya digunakan sebagai dokumen pengujian. Sebelum digunakan dalam penelitian, dokumen pelatihan terlebih dahulu ditandai pada entitas bernama, kata ganti, dan penanda bahwa kata ganti atau entitas bernama tersebut saling merujuk. Contoh pemberian tagnya ditunjukkan pada gambar berikut:



Gambar 1. Contoh relasi dan tag koreferensi

Pada dokumen pengujian, digunakan tag entitas bernama, seperti yang digunakan pada penelitian (Gatot, 2004).

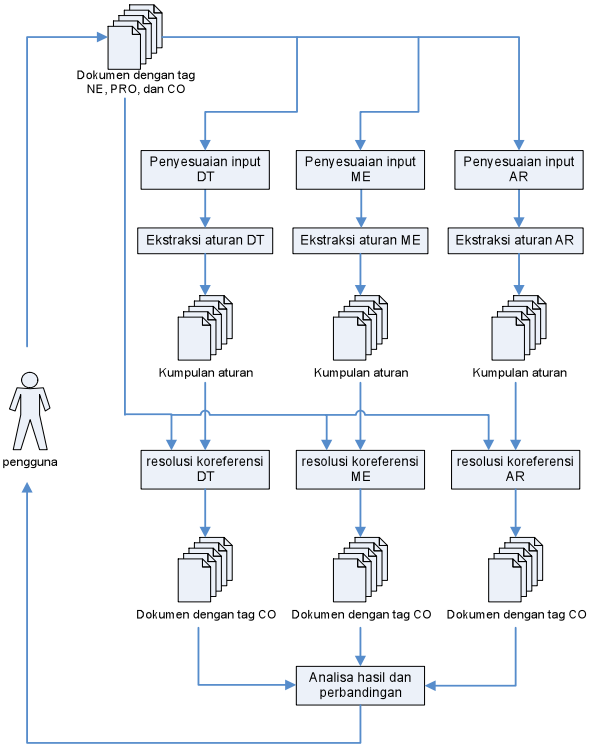
4. DESAIN DAN IMPLEMENTASI

Untuk penelitian menggunakan matode *decision tree*, penulis menggunakan *package* aplikasi yang telah tersedia di internet. Aplikasi tersebut dikembangkan oleh Jean-Marc François. Alasan penggunaan *package* tersebut adalah aplikasi tersebut menggunakan algoritma C4.5 yang banyak digunakan dalam penelitian bahasa asing. Dalam aplikasi itu pula disediakan mekanisme untuk mengetahui bentuk *tree* yang dihasilkan oleh dokumen penelitian. Analisa akan lebih mudah dilakukan dengan tree tersebut. Lisensi program ini adalah GPL (Francois, 2004).

Untuk metode *maximum entropy*, tools yang digunakan adalah tool yang disediakan oleh *opennlp*. Lisensi yang digunakan adalah bebas. Tool ini pernah beberapa kali digunakan dalam penelitian bahasa asing.

Hasil penelitian kedua tools di atas, lalu dibandingkan dengan hasil yang diberikan oleh sistem yang dikembangkan oleh (Nasrullah, 2005).

 ISSN: 1907-5022



Gambar 2. Alur penelitian

Secara umum, alur penelitian ditunjukkan pada gambar 2. Aturan-aturan diperoleh dari masing-masing metode, dengan memberikan dokumen pelatihan kepada sistem. Masing-masing metode akan menyimpan aturan-aturan yang diperoleh selama pelatihan ke dalam format yang spesifik. Jumlah dokumen pelatihan juga merupakan objek penelitian. Hal ini dimaksudkan untuk mengetahui berapa jumlah dokumen yang tepat untuk masing-masing metode.
Aturan-aturan tersebut digunakan kemudian untuk melakukan resolusi koreferensi. Pada tahap ini masukan sistem berupa teks-teks yang telah diberi tag entitas bernama seperti pada penelitian Wahyudi. Hasilnya adalah nilai yang direpresentasikan sebagai *recall, precision,* dan *f-measure* (Gatot, 2004)

5. FITUR-FITUR

Hubungan koreferensi direpresentasikan dalam fitur-fitur. Dalam penelitian ini digunakan delapan fitur yang membedakan. Fitur merupakan sebuah ciri yang ditentukan dalam penelitian.

Misalkan dalam kalimat berikut:

Adi_1 pergi ke sekolah bersama $Toni_2$, adiknya$_3$.

Dalam kalimat di atas terdapat tiga frase yang merupakan entitas bernama dan kata ganti, yaitu: Adi, Toni, dan –nya. Dalam penelitian, kesemua frase tersebut saling dibandingkan, sehingga terdapat 3 perbandingan sebagai berikut:

1. Frase "Adi_1" dan frase "$Toni_2$"
2. Frase "Adi_1" dan frase "-nya$_3$"
3. Frase "$Toni_2$" dan frase "-nya$_3$"

Dalam penelitian ini digunakan delapan fitur, yaitu:
1. Fitur *isEqualCharacters* bernilai *true* apabila kedua *markable* sama, dan bernilai *false* apabila sebaliknya.
2. Fitur *isEqualWithoutPunctuation* bernilai *true* apabila kedua *markable* yang telah dihilangkan semua tanda baca dan tidak menghiraukan huruf besar atau kecil adalah sama, dan bernilai *false* apabila sebaliknya.
3. Fitur *isAcronym* bernilai *true* apabila salah satu *markable* merupakan singkatan dari *markable* yang lain, dan bernilai *false* apabila sebaliknya.
4. Fitur *is1stPronoun* bernilai *true* apabila *markable* pertama merupakan kata ganti, dan bernilai *false* apabila sebaliknya.
5. Fitur *is2ndPronoun* bernilai *true* apabila *markable* pertama merupakan kata ganti, dan bernilai *false* apabila sebaliknya.
6. Fitur *isOnOneSentence* bernilai *true* apabila kedua markable berada dalam satu kalimat, bernilai *false* apabila sebaliknya.
7. Fitur *isMatchPartially* bernilai *true* apabila sebagian salah satu *markable* sama dengan keseluruhan *markable* yang lain, bernilai *false* jika sebaliknya.
8. Fitur *isSameNameClass* bernilai *true* apabila kedua markables memiliki nama kelas entitas yang sama. Misalkan keduanya memiliki nama kelas ORGANIZATION, maka fitur ini bernilai *true*.

Semua fitur di atas, kecuali fitur *isOnOneSentence* dan *isSameNameClass* telah digunakan sebelumnya dalam penelitian (Nasrullah, 2005).

Dari contoh perbandingan frase di atas, berikut adalah contoh penggunaan fitur-fitur yang digunakan dalam penelitian ini:
a. Pasangan frase "Adi_1" dan frase "$Toni_3$" memiliki fitur *isOnOneSentence* dan *isSameNameClass* yang bernilai *true.* Fitur selain itu bernilai *false.* Kedua frase "Adi_1" dan frase "$Toni_3$" tidak saling merujuk.
b. Pasangan frase "Adi_1" dan frase "-nya$_3$" memiliki fitur *is2ndPronoun*, *isOnOneSentence* yang bernilai *true.* Fitur selain itu bernilai *false.* Kedua frase ini saling merujuk.
c. Pasangan frase "$Toni_2$" dan frase "-nya$_3$" memiliki fitur *is2ndPronoun* dan *isOnOneSentence* yang bernilai *true*. Fitur selain itu bernilai *false*. Kedua frase itu tidak saling merujuk.

Masing-masing tool menghendaki format penulisan fitur-fitur yang berbeda-beda

6. HASIL DAN ANALISA TERHADAP JUMLAH DOKUMEN PENELITIAN

Untuk pengujian terhadap jumlah dokumen pelatihan, Tabel men1unjukkan nilai *f-measure* hasil pengujian.

Tabel 1. F-measure terhadap jumlah dokumen pelatihan

Jumlah Dokumen	*Decision Tree*	*Maximum Entropy*	*Association Rules*
10	75.87	75.28	81.27
20	80.73	80.62	81.31
30	80.71	80.71	81.32
40	80.76	80.71	81.24
50	71.27	80.65	81.33
60	71.12	80.7	80.75
70	71.12	80.65	80.77
80	71.12	80.65	80.77
90	71.12	80.65	81.36
100	71.12	80.65	81.36
110	71.12	80.65	81.31
120	71.12	80.65	81.36
130	71.12	80.65	81.36
140	71.12	80.65	81.36
150	71.04	80.65	81.36
160	71.06	80.65	81.36
170	71.06	80.65	81.36
180	71.06	80.65	81.36
190	71.06	80.65	81.36
200	71.06	80.65	81.36
210	71.06	80.65	81.36
220	71.06	80.65	81.36
230	71.06	80.65	81.36
240	71.06	80.65	81.36
250	71.12	80.65	81.36
260	71.12	80.65	81.36
270	71.12	80.65	81.36
280	71.12	80.65	81.36
290	71.12	80.65	81.36
300	71.12	80.65	81.36

Pada tabel 1 dapat dilihat, bahwa pada metode decision tree, nilai tertinggi *f-measure* dicapai saat dokumen pelatihan sejumlah 40 dokumen. Nilai *f-measure* semakin turun dan stabil ketika jumlah dokumen 60-300 dokumen. Sedangkan untuk *maximum entropy* dan *association rules*, keduanya memiliki perilaku yang hampir sama. Keduanya mencapai nilai tertinggi dan akhirnya stabil pada jumlah dokumen yang semakin banyak. *Maximum entropy* mencapai *f-measure* tertinggi dan stabil pada nilai 80.71 dicapai setidaknya pada jumlah dokumen 30. untuk *association rules, f-measure* tertinggi pada nilai 81.36, dicapai pada jumlah dokumen minimal 90 dokumen.

Pada *decision tree, tree* dibangun dengan menggunakan algoritma C4.5 dengan menggunakan *information gain* yang diperoleh dari dokumen-dokumen pelatihan. Pada jumlah dokumen pelatihan sedikit, kesalahan yang didapat dikarenakan masih terdapat kondisi yang belum terakomodasi di dalam *tree*, dikarenakan terlalu sedikit informasi yang dapat ditemukan dalam dokumen-dokumen pelatihan. Namun ketika jumlah dokumen semakin banyak, terdapat pula kesalahan yang dihasilkan oleh *tree* yang dibangun. Hal ini dikarenakan semakin banyak informasi yang didapat, maka semakin banyak *noise*/infomasi yang tidak perlu. Pada proses pembangunan *tree*, tidak dapat dibedakan mana informasi yang benar dan perlu dan mana informasi yang harus dibuang. Akibatnya *tree* yang dibangun justru memberikan hasil yang salah. Semakin banyak data yang digunakan, *f-measure* yang dihasilkan akan stabil. Pada penelitian, dengan formasi dokumen tertentu jumlah dokumen paling tepat adalah 40 dokumen. Namun hasil ini dapat berbeda jika formasi dokumen yang digunakan dirubah.

Pada *maximum entropy*, nilai tertinggi dicapai pada jumlah dokumen 30-40 dokumen. Selebihnya, *f-measure* yang dihasilkan stabil pada kisaran 80.65. Jika dilihat dari *f-measure*-nya, *maximum entropy* memiliki perilaku yang sama dengan *decision tree*. *Maximum entropy* menghendaki jumlah dokumen dan kombinasi tertentu untuk mencapai nilai *f-measure* tertinggi. Model GIS yang dihasilkan pada jumlah dokumen, merupakan mode yang paling baik untuk digunakan dalam melakukan resolusi koreferensi.

Association rules memberikan hasil yang berbeda dibandingkan dengan dua metode yang lain. Metode *association rules* memiliki hasil yang stabil dari dua metode sebelumnya. Dengan dokumen yang sedikit, metode ini mampu membuat aturan-aturan yang cukup untuk mengenali koreferensi. Saat jumlah dokumen 60-80, nilai *F-measure* yang diperoleh sempat menurun. Namun secara umum, hasil yang diberikan tetaplah yang terbaik dibandingkan dengan dua metode sebelumnya. Metode ini mencapai kestabilan saat jumlah dokumen yang digunakan untuk proses pelatihan sebanyak 130 dokumen.

7. HASIL DAN ANALISA TERHADAP KOMBINASI FITUR

Kombinasi fitur yang digunakan dalam pengujian adalah seperti ditunjukkan pada tabel 2.

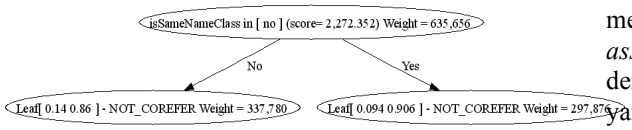
Kelompok	Fitur
Kelas kata (KK)	*Is1stPronoun, is2ndPronoun*
Kelas nama (KN)	*isSameNameClass*

Kesamaan karakter penyusun frase (KS)	*isEqualCharacter, isEqualWithoutPunctuation, isMatchPartially, isAcronym*
Letak atau posisinya di dokumen (L)	*isOnOneSentence*

Tabel 2. Kombinasi fitur dalam penelitian

Pada tahapan ini, pengujian langsung dilakukan terhadap 300 dokumen pelatihan dengan menggunakan satu atau lebih kombinasi fitur di atas.

Untuk *decision tree*, kombinasi fitur KS sudah cukup memberikan hasil tertinggi. Kombinasi fitur KS dengan jumlah dokumen 300 memberikan *tree* yang memiliki kemampuan paling besar untuk melakukan resolusi koreferensi. Kombinasi KN, KK, dan L tanpa KS tidak dapat digunakan untuk melakukan resolusi koreferensi. Jika menggunakan kombinasi fitur KN, berikut adalah *tree* yang dihasilkan dengan menggunakan 300 dokumen pelatihan.



Gambar 3. Contoh tree pada kombinasi KN

Tree tersebut memiliki node yang terlalu sedikit untuk digunakan dalam pengambilan keputusan.

Tabel 3. F-measure terhadap kombinasi fitur

Kombinasi Fitur	DT	ME	AR
KS	80.84	80.84	80.84
KK	0	0	0
KN	0	0	0
L	0	0	0
KS + KK	80.79	80.79	80.79
KS + KN	80.84	80.84	81.36
KS + L	80.84	80.77	80.84
KK + KN	0	0	0
KK + L	0	0	0
KN + L	0	0	0
KS + KK + KN	80.79	80.7	81.36
KS + KK + L	80.79	80.72	80.79
KS + KN + L	80.84	80.79	81.36
KK + KN + L	22.37	6.07	0
KS + KK + KN + L	71.12	80.65	81.36

Metode *maximum entropy* juga dapat dicapai dengan menggunakan kombinasi fitur KS. Sedangkan untuk *Association Rules*, hasil tertinggi dapat dicapai dengan menggunakan kombinasi fitur KS dan KN. Hasil yang dicapai *association rules* bahkan lebih tinggi dibandingkan dua metode lainnya.

8. KESIMPULAN

Penelitian ini menunjukkan, metode *decision tree* dan *maximum entropy* membutuhkan jumlah dokumen yang tertentu, sehingga hasil yang diberikan dapat maksimal. Jika jumlah dokumen kurang, maka terdapat kondisi yang tidak terakomodasi dalam sistem. Sedangkan jika jumlah dokumen yang digunakan terlalu banyak maka akan terdapat *noise* yang tidak perlu dipakai dalam membangun model. Sehingga hasil yang diberikan tidak maksimal. Untuk mencapai nilai maksimal ini, harus dilakukan pengujian dengan menggunakan jumlah dokumen yang bervariasi, sehingga bisa diketahui jumlah dokumen yang paling tepat untuk mencapai nilai maksimum. Sebaliknya, metode *association rules* memberikan hasil ang lebih bagus dengan jumlah dokumen yang lebih banyak. Namun yang perlu diperhatikan adalah jumlah dokumen minimal sehingga metode ini dapat memberikan hasil yang maksimal. Tidak perlu memberikan sebanyak mungkin dokumen pelatihan, karena nilai maksimal akan dicapai pada suatu jumlah dokumen tertentu.

Pada *decision tree* dan *maximum entropy*, kombinasi fitur kesamaan karakter penyusun kata merupakan fitur yang dominan. Penggunaan fitur ini mampu menghasilkan model yang terbaik yang dapat digunakan untuk melakukan resolusi koreferensi. Sedangkan pada *association rules*, kombinasi fitur kesamaan karakter penyusun frase dan kelas nama memberikan hasil yang maksimum.

PUSTAKA

Amin, M. F., (2006). Perbandingan Association Rules, Decision Tree dan Neural Network Untuk Event Extraction Pada Sistem Ekstraksi Informasi. Fakultas Ilmu Komputer Universitas Indonesia, Depok.

Budi, I. dan Bressan, S. (2003). Association Rules Mining for Name Entity Recognition. Proceeding of 2003 WISE Conference, Roma.

Chinchor, N., (1998). MUC-7 Information Extraction Task Definition. The MITRE Corporation and SAIC.

Francois, J. (2004). jaDTi - Decision Trees: a Java implementation [online]. Liege: run.montefiore.ulg.ac.be. Dari: http://www.run.montefiore.ulg.ac.be/~francois/software/jaDTi/;Internet; diakses 8 November 2007.

Markus. (2007). Pengenalan Entitas Bernama Menggunakan Metode Association Rules pada Dokumen Berbahasa Indonesia. Fakultas Ilmu Komputer Universitas Indonesia, Depok.

Manning, Christopher dan Schutze, H., (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Nasrullah. (2005). Resolusi Koreferensi Menggunakan Metode Association Rules. Fakultas Ilmu Komputer Universitas Indonesia, Depok.

MUC. (1998) MUC-7 Coreference task definition. Proceedings of the Seventh Message Understanding Conference (MUC-7).

Marjuki, K., (2006). Ekstraksi Person pada Sistem Event Extraction Bahasa Indonesia. Fakultas Ilmu Komputer Universitas Indonesia, Depok.

Wahyudi, G., (2004). Pengenalan Entitas Bernama Berdasarkan Informasi Kontekstual, Morfologi, dan Kelas Kata. Fakultas Ilmu Komputer Universitas Indonesia, Depok.

Wibowo, B. (2005). Pengenalan Entitas Bernama Menggunakan Metode Association Rules dengan Fitur Berganda (Association Rules with Multiple Features). Fakultas Ilmu Komputer Universitas Indonesia, Depok.

Wu, T., (2002). Theory and Applications in Information Extraction from Unstructured Text. Bethlehem: Lehigh University.