ISSN: No.1410 - 7066
Akreditasi: No. 48/DIKTI/Kep/2006

JURNAL PENELITIAN DAN PENGEMBANGAN TELEKOMUNIKASI

DESEMBER 2008 VOLUME 13 - NOMOR 2

INSTITUT TEKNOLOGI TELKOM

DIREKTORAT DUKUNGAN AKADEMIK
BIDANG PENELITIAN DAN PENGABDIAN MASYARAKAT
INSTITUT TEKNOLOGI TELKOM

EVALUASI KINERJA ALGORTIMA SUPPORT VECTOR MACHINE DALAM EKSTRAKSI INFORMASI KORPUS BERBAHASA INDONESIA

Kurnia Muludi[1], Kuspriyanto[2], Oerip S Santoso[3], dan Dwi H Widyantoro[4]

[1]Fakultas Pertanian, Universitas Lampung
[2,3,4]Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung
[1]kurnia@unila.ac.id, [2]kuspriyanto@yahoo.com, [3]pireo46@yahoo.com, [4]dwi@if.itb.ac.id

Abstrak

Perkembangan internet yang pesat menyebabkan perkembangan volume informasi tekstual yang luar biasa. Karenanya dibutuhkan *tool* dan metoda cerdas yang dapat mengakses konten sesuai kebutuhan. Salah satu metoda ekstraksi informasi yang berkembang adalah metoda yang menggunakan *Support Vector Machine* (SVM). Dalam tulisan ini dibahas kinerja SVM-GATE dalam ekstraksi informasi pada Korpus berbahasa Indonesia. Hasil percobaan menunjukkan dengan meningkatkan jumlah dokumen sampel dalam pembelajaran diperoleh peningkatan kinerja SVM GATE. Kinerja SVM terbaik diperoleh pada *margin tau*= 0,3 dan pada *window size*=4. Pada komposisi terbaik ini diperoleh *F-Measure* 49,64% (*Strict*) dan 58,45% (*Lenient*).

Kata kunci : Ekstraksi Informasi, *Support Vector Machine*, Korpus Bahasa Indonesia, NLP, GATE

Abstract

The rapid growth of internet causes the abundance of textual information. It is necessary to have smart tools and methods than can access text content as needed. One of the success methods is Support Vector Machine (SVM). In this paper we will discuss SVM-GATE performance in extracting information on Bahasa Indonesia corpus. The experimental results show that SVM-GATE performance increases as the training sample number grows. The best performace is shown when tau margin is 0.3 and window size is 4. In this composition *F-Measure* are 49.64% (*Strict*) and 58.45% (*Lenient*).

Keywords : Information Extraction, Support Vector Machine, Bahasa Indonesia Corpus, NLP, GATE

1. Pendahuluan

Dengan berkembangannya internet, volume informasi tekstual juga berkembang dengan sangat pesat. Saat ini teknologi *Information Retrieval* saja tidak mampu untuk mencukupi kebutuhan informasi yang spesifik karena teknologi ini hanya menyediakan informasi pada level koleksi dokumen. Pengembangan *tool* dan metoda cerdas yang dapat mengakses konten dokumen karenanya merupakan isu krusial dalam *Knowledge Management*.

Ekstraksi informasi merupakan proses pengambilan informasi mengenai *pre-specified events*, entitas atau *relationships* pada teks seperti artikel berita (*newswire*) dan halaman web. Banyak fokus riset ekstraksi informasi ditujukan pada *named entity recognition* yang merupakan *tags* mendasar. Secara umum *tags* ekstraksi informasi dapat dianggap sebagai *task* pengenalan entitas informasi pada teks. Ekstraksi informasi sangat berguna dalam banyak aplikasi seperti *business intelligence*, anotasi otomatis pada halaman web, dan *knowledge management*.

Ekstraksi informasi dapat didekati melalui pendekatan masalah klasifikasi dimana teks dipisahkan menjadi *token-token* dan dimasukkan dalam kelas yang sesuai. *Hidden Markov Models* merupakan metode populer untuk tugas tersebut, namun metode ini tidak dapat menangani *token* dengan *multiple attribute* [1].

Salah satu metode pembelajaran mesin yang sukses dalam ekstraksi informasi adalah *Support Vector Machine* (SVM), yang merupakan bagian dari algoritma *supervised machine learning*. Algortima ini telah mencapai kinerja *state-of-the-art* pada berbagai *classification task*, termasuk *named entity recognition*[3,4].

Klasifier SVM dapat menduga dimana suatu jenis *tag* berawal dan berakhir dalam teks. Klasifier ini dilatih dari teks yang telah dianotasi. Klasifier SVM digunakan untuk membedakan *item* suatu kelas terhadap kelas lainnya berdasarkan atribut-atribut *training examples*. Atribut-atribut ini disebut juga *features*. Masalah klasifikasi paling sederhana adalah membedakan contoh positif dan negatif suatu konsep. Dalam ekstraksi informasi permasalahannya adalah bagaimana menentukan posisi teks apakah merupakan awal suatu *tag* atau bukan dan akhir suatu *tag* atau bukan.

Dalam tulisan ini akan didiskusikan bagaimana kinerja algoritma SVM untuk ekstraksi informasi pada korpus berbahasa Indonesia dan akan ditampilkan hasil-hasil percobaan secara detail. Percobaan akan melihat pengaruh beberapa parameter SVM pada kinerja ekstraksi informasi. Kurva pembelajaran algoritma SVM juga akan

dievaluasi melalui percobaan bagaimana pengaruh jumlah dokumen contoh terhadap *F-measures*.

2. Penelitian Terkait

Pada sistem ekstraksi informasi berbasis SVM yang digunakan oleh Isozaki [4], empat buah klasifier dilatih menggunakan fungsi *sigmoid* untuk mentransfer keluaran SVM menjadi suatu probabilitas dan menerapkan algoritma *Viterbi* untuk menentukan sekuen label yang optimal untuk suatu kalimat. Sistem ini dievaluasi pada korpus berbahasa Jepang menggunakan *window size*=2. Hasil percobaan menunjukkan sistem ini mempunyai kinerja yang lebih baik dari sistem berbasis *Maximum Entropy* dan *Rule Learning*. Sistem ini juga menggambarkan implementasi yang efisien untuk kernel SVM kuadratik.

Mayfield [7] menerapkan SVM dengan pendekatan *lattice-based* pada *kernel* kubik untuk menghitung probabilitas transisi pada *lattice*. Dengan menggunakan *window size*=3 diperoleh hasil yang cukup memuaskan [5].

Sistem GATE-SVM merupakan varian dari SVM dengan *uneven margin*. Pada SVM biasa contoh positif dan negatif diperlakukan sama sedemikian rupa sehingga *margin hyperplane* ke contoh negatif sama dengan *margin* ke contoh negatif. Namun pada data training yang *imbalanced* dimana contoh positif jauh lebih sedikit maka SVM biasa tidak tepat merepresentasikan distribusi contoh positif sebenarnya. Karenanya, *margin* positif yang lebih besar dari margin negatif merupakan model SVM yang lebih baik. Li [6] memperkenalkan parameter *uneven margin* pada algoritma SVM. Parameter *uneven margin* merupakan rasio *margin* negatif terhadap *margin* positif. Dengan menggunakan parameter ini SVM mampu menangani *imbalanced data* lebih baik dari model SVM biasa.

3. Metode Percobaan

Secara formal jika diberikan himpunan pelatihan $Z = ((x_i,y_i)),\ldots,((x_m,y_m))$, dimana x_i adalah vektor input n-dimensi, dan y_i (= +1 atau -1) adalah label kelas, dan m adalah jumlah data pelatihan maka pada SVM dengan *uneven margin* diperoleh dengan memecahkan persoalan optimisasi kuadratik:

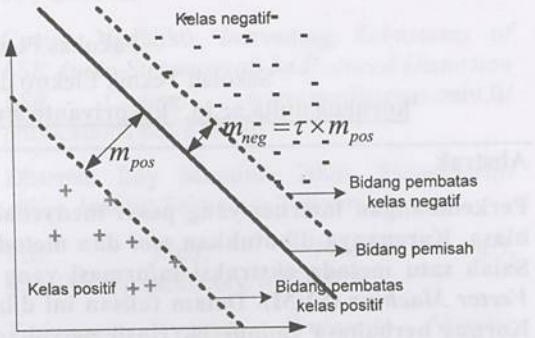
$$\min_{w,b,\xi} \langle w,w \rangle + C\sum_{i=1}^{m} \xi_i$$

$$\text{s.t } \langle w,x_i \rangle + \xi_i + b \geq 1 \quad \text{if } y_i = +1$$

$$\langle w,x_i \rangle - \xi_i + b \leq -\tau \quad \text{if } y_i = -1$$

$$\xi_i \geq 0 \quad \text{for } i = 1,\ldots,m$$

Dapat terlihat pada persamaan di atas terdapat penambahan parameter τ (*margin tau*). τ adalah rasio *margin* kelas negatif terhadap *margin* kelas positif, dan akan sama dengan 1 pada SVM standar. Pada *imbalanced* datasheet, digunakan *margin* yang lebih besar untuk kelas positif dibandingkan untuk kelas negatif, seperti yang dapat dilihat pada Gambar 1. Oleh karena itu, pada SVM dengan *uneven margin* nilai τ adalah $0<\tau<1$.



Gambar 1. Ilustrasi SVM dengan *Uneven Margin*

Dalam percobaan ini digunakan korpus berbahasa Indonesia yang berjumlah 60 buah teks yang diambil dari situs-situs surat kabar Media Indonesia (www.mediaindonesia.com), Kompas (www.kompas.com), Lampung Post (www.lampungpost.com), dan kantor berita Antara (www.antara.co.id). Berita-berita yang diambil adalah yang berkenaan dengan perkembangan harga komoditi sayur-mayur di kota-kota di Indonesia.

Pada pemerosesan awal, anotasi diberikan pada korpus sesuai dengan jenis informasi yang akan diekstrak. *Tag-tag* yang digunakan disini adalah *Tanggal, lokasi, komoditi,* dan *harga*. Contoh dokumen pada korpus berbahasa Indonesia yang telah dianotasi untuk pembelajaran dan evaluasinya disajikan pada Gambar 2.

Harga Sayur Mayur di Magelang Turun

Selasa, **<tanggal>**12 Agustus 2008**</tanggal>** | 20:16 WIB

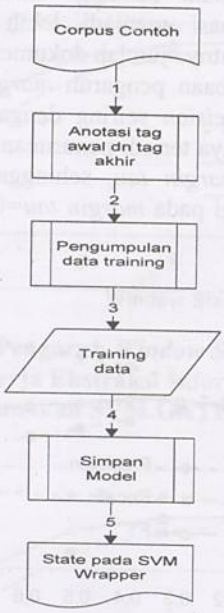
MAGELANG, SELASA - Harga sayur mayur di Kabupaten Magelang, kini turun secara signifikan. Pada berbagai jenis sayuran, penurunan harga terjadi bervariasi, mulai dari Rp 500 per kilogram (kg), hingga Rp 1.500 per kg.

Sumartini, salah seorang pedagang sayur di **<lokasi>**Pasar Muntilan**</lokasi>**, mengatakan, **<komoditi>** kacang panjang **</komoditi>** misalnya mengalami penurunan harga dari Rp 3.500 per kg menjadi **<harga>** Rp 2.500 per kg **</harga>**. Begitupun, harga **<komoditi>** seledri **</komoditi>**yang semula Rp 2.500 per kg, sekarang menjadi **<harga>** Rp 1.500 per kg **</harga>**. Untuk **<komoditi>** tomat **</komoditi>** dan **<komoditi>** wortel **</komoditi>**, masing-masing turun harga Rp 500 per kg, menjadi **<harga>** Rp 1.500 per kg **</harga>** dan **<harga>** Rp 1.000 per kg **</harga>.**

Gambar 2. Contoh Korpus Berbahasa Indonesia yang Digunakan.

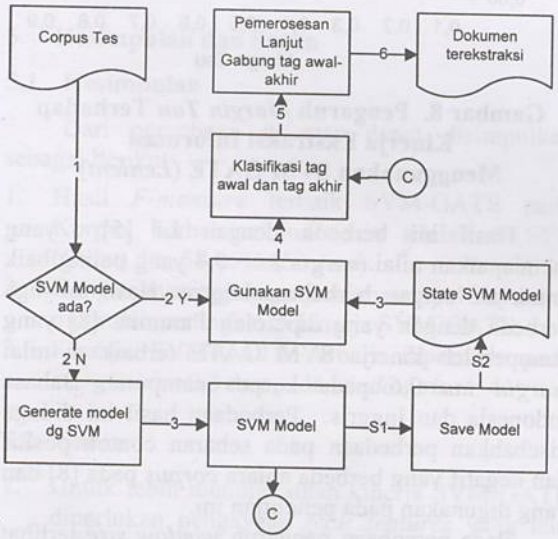
Dalam implementasi SVM untuk ekstraksi informasi digunakan tool GATE (*General*

Architecture for Texts Engineering)[3]. Pada proses pembelajaran mesin SVM pada GATE digunakan *SVM Light Wrapper* [2]. Proses tahap demi tahap ditunjukkan pada Gambar 3 dan Gambar 4.



Gambar 3. Bagan Alur Proses *Training* Data pada SVM-GATE

Gambar 3 menjelaskan langkah demi langkah proses *training* data untuk klasifier SVM. Korpus contoh pertama disimpan dalam format GATE dan dianotasi pada token yang sesuai dengan jenisnya (misal *komoditi*) dan dokumen ini digunakan sebagai asupan untuk membangun model SVM (dibutuhkan *tag* <*komoditi*> sebagai awal *tag* dan </*komoditi*> sebagai akhir *tag* pada obyek teks/*token* yang dimaksud). Model SVM yang dihasilkan kemudian disimpan pada *file* eksternal untuk digunakan kemudian.



Gambar 4. Bagan Alur Proses Ekstraksi dengan SVM-GATE

Pada percobaan ini untuk membangun *features vector* dari *token-token* pada SVM digunakan beberapa fitur NLP yaitu :

a. *Orthography* atau *Case*, yaitu penggunaan huruf besar dan huruf kecil oleh *token*.
b. Jenis *token*: kata, angka, simbol, atau tanda baca.
c. *Entity*, hasil keluaran modul *named entity recognition* standar yang dimiliki oleh GATE.

Adapun *window size* merupakan jumlah *token* sebelum dan setelah *token* sasaran yang digunakan sebagai masukan untuk SVM.

Untuk mengekstrak informasi pada dokumen baru, sistem membutuhkan model SVM yang dihasilkan pada proses pembelajaran. *SVM Wrapper* kemudian akan menganotasi teks sasaran dengan *tag* awal dan *tag* akhir sesuai model yang ada. Pada tahap berikutnya *tag* awal dan akhir digabungkan pada *token* yang sesuai (Gambar 4).

Untuk mengukur kinerja algoritma dalam mengekstraksi informasi digunakan *Precision*, *Recall* dan *F-Measure*.

$$precision = \frac{correct}{correct + falsePositive} \quad (1)$$

$$recall = \frac{correct}{correct + falseNegative} \quad (2)$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

Parameter *correct* adalah jumlah *slot* yang terisi dan benar, *falsePositive* adalah jumlah slot yang terisi namun salah, dan *falseNegative* adalah jumlah *slot* yang tidak terisi.

Tabel 1. Kinerja SVM GATE pada Korpus Berita Harga Sayur Mayur (*Strict*)

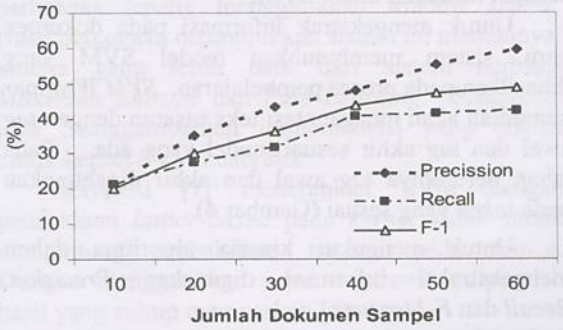
Jumlah Sampel	*Precision*	*Recall*	*F-1*
10	20,9	21,4	20,1
20	35,2	27,3	30,3
30	43,4	31,6	36,4
40	48,0	40,5	43,8
50	55,4	41,5	47,3
60	59,7	42,1	48,9

Untuk masalah *partial correct*, yaitu hasil prediksi ekstraksi hanya benar sebagian, di dalam rumus *precision, recall*, dan *F-measure (F-1)*, terdapat 3 cara pendekatan yaitu: (1) *Strict*, *partial correct* dianggap salah, yaitu baik sebagai *falsePositive* maupun *falseNegative*, (2) *Lenient*, *partial correct* dianggap benar, dan (3) *Average*, *partial correct* diberi bobot ½. Dalam tulisan ini akan digunakan pendekatan (1) dan (2). Untuk memperoleh hasil yang lebih baik, percobaan diulang sepuluh kali (*10-fold cross validatiion*)

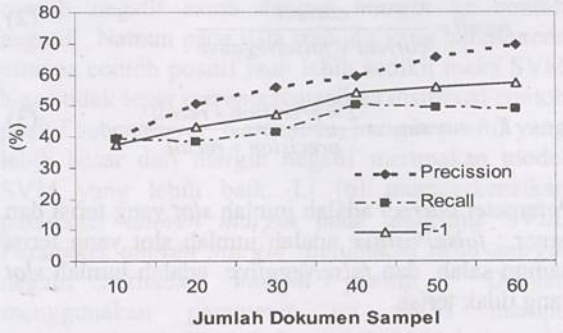
untuk tiap variasi asupan model yang diuji. Adapun model kernel SVM yang digunakan adalah *SVM Linear* dengan pembobotan *Reciprocal Weighting*. Sedangkan teknik klasifikasi yang dipilih adalah *one-against-all* [2].

4. Hasil dan Pembahasan

Hubungan jumlah dokumen sampel terhadap kinerja SVM-GATE dengan komposisi *margin tau*=0,4 dan *window size=3* dapat dilihat pada Gambar 5 dan Gambar 6.



Gambar 5. Hubungan Jumlah Dokumen Sampel dan *F-Measure (Strict)* pada Ekstraksi Informasi Menggunakan SVM GATE



Gambar 6. Hubungan Jumlah Dokumen Sampel dengan *F-Measure (Lenient)* pada Ekstraksi Informasi menggunakan SVM-GATE

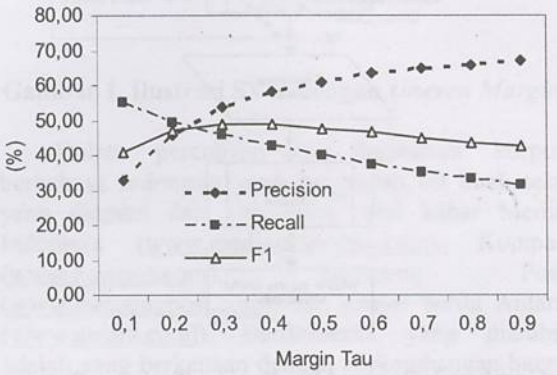
Dari Gambar 5 dan 6 terlihat bahwa dengan meningkatnya jumlah dokumen sampel yang digunakan dalam pembelajaran diperoleh secara umum indeks *Precision*, *Recall* dan *F-Measure* yang meningkat juga, kecuali *Recall* pendekatan *lenient* pada Tabel 2 yang cenderung stabil mulai sampel dokumen berjumlah 40 ke atas.

Tabel 2 Kinerja SVM GATE pada Korpus Berita Harga Sayur Mayur (*Lenient*)

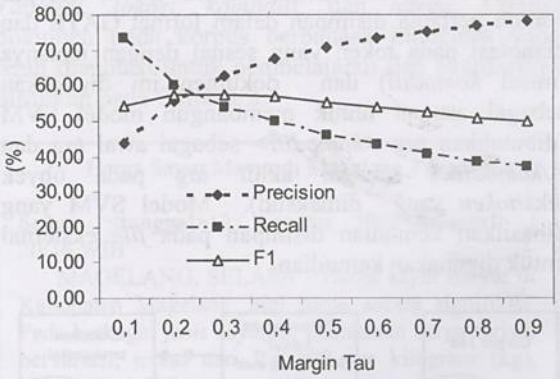
Jumlah Sampel	*Precision*	*Recall*	*F-1*
10	39,4	39,1	37,4
20	50,5	38,3	42,9
30	56,0	41,1	47,2
40	59,5	50,0	54,2
50	65,6	49,1	56,0
60	69,6	49,0	57,0

Hasil ini juga sama dengan yang diperoleh Li [5] pada *Job Corpus* untuk sampel kecil. Meningkatnya jumlah dokumen sampel dalam pembelajaran menyebabkan kualitas klasifier menjadi lebih baik sehingga kinerjanya dalam ekstraksi informasi menjadi lebih baik sejalan dengan meningkatnya jumlah dokumen sampel.

Pada percobaan pengaruh *margin tau* terlihat peningkatan *Precision* seiring dengan peningkatan *margin*. Sebaliknya terjadi penurunan *Recall* dengan meningkatnya *margin tau*, sehingga diperoleh *F-Measure* tertinggi pada *margin tau*=0,3. (Gambar 7 dan Gambar 8).



Gambar 7. Pengaruh *Margin Tau* Terhadap Kinerja Ekstraksi Informasi Menggunakan SVM GATE (*Strict*)

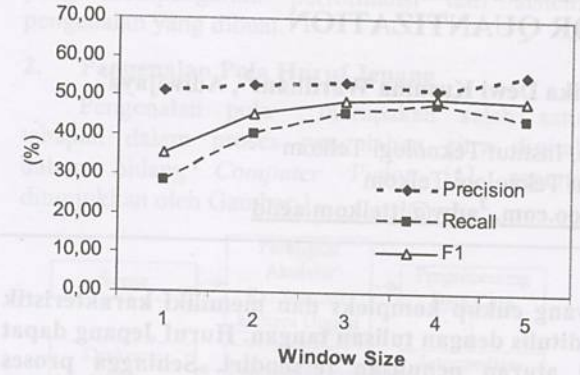


Gambar 8. Pengaruh *Margin Tau* Terhadap Kinerja Ekstraksi Informasi Menggunakan SVM GATE (*Lenient*)

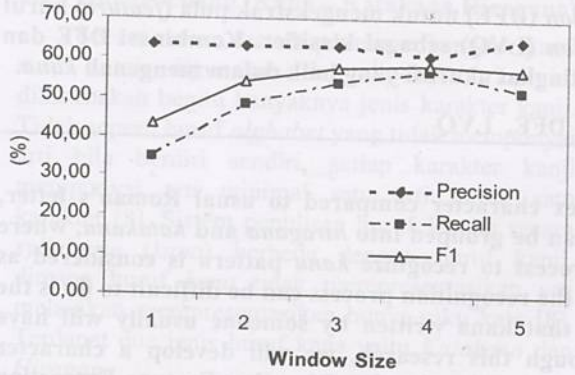
Hasil ini berbeda dengan Li [5] yang mendapatkan nilai *margin tau*=0,4 yang paling baik untuk *job corpus* berbahasa Inggris. Hasil ini juga berbeda dengan yang diperoleh Paramita [8] yang memperoleh kinerja SVM GATE terbaik di nilai *margin tau*=0,6 pada korpus campuran Bahasa Indonesia dan Inggris. Perbedaan hasil ini diduga disebabkan perbedaan pada sebaran contoh positif dan negatif yang berbeda antara *corpus* pada [8] dan yang digunakan pada penelitian ini.

Pada percobaan pengaruh *window size* terlihat *window size*=4 menunjukkan kinerja terbaik (Gambar 9 dan Gambar 10). Hasil ini berbeda

dengan hasil [5] dan [8] yang merekomendasikan *window size*=3.



Gambar 9. Pengaruh *Window Size* Terhadap Kinerja Ekstraksi Informasi Menggunakan SVM GATE (*Strict*)



Gambar 10. Pengaruh *Window Size* Terhadap Kinerja Ekstraksi Informasi Menggunakan SVM GATE (*Lenient*)

Dari hasil percobaan di atas, dapat diperoleh hasil *F-Measure* tertinggi dengan komposisi terbaik pada *window size*=4 dan *margin tau*=0,3 dengan nilai 49,64% (*Strict*) dan 58,45% (*Lenient*).

5. Kesimpulan dan Saran

5.1 Kesimpulan

Dari percobaan di atas dapat disimpulkan sebagai berikut:

1. Hasil *F-measure* terbaik SVM-GATE pada Korpus berbahasa Indonesia adalah 49,64% (*Strict*) dan 58,45% (*Lenient*).
2. Dengan meningkatnya jumlah sampel dokumen diperoleh peningkatan kinerja SVM-GATE.
3. Kinerja SVM-GATE terbaik diperoleh pada *margin tau*=0,3 dan pada *window size* = 4.

5.2 Saran

1. Untuk lebih meningkatkan kinerja SVM-GATE diperlukan pengayaan *NLP features* yang lain misalnya penggunaan *Part of Speech Tagger* untuk Bahasa Indonesia.
2. Perlu diteliti lebih lanjut apakah kinerja SVM-GATE tetap meningkat dengan meningkatnya jumlah dokumen sampel, dan bagaimana hubungan antara *window size* dengan karakter *Corpus*.

Daftar Pustaka

[1] Bouckaer, R.R., 2002, *Low level information extraction. A Bayesian network based approach*, In *Proc. TextML* 2002.

[2] Cunningham, Hamish et al., 2007, *Developing Language Processing Components with GATE Version 4 (a User Guide)*, The University of Sheffield 2001-2007, http://gate.ac.uk/sale/tao/

[3] Cunningham, Hamish, D. Maynard, K. Bontcheva, V. Tablan., 2002, *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), http://gate.ac.uk/sale/acl02/acl-main.pdf.

[4] Isozaki, H., Kazawa, H., 2002, *Efficient Support Vector Classifers for Named Entity Recognition*, In Proceedings of the 19th International Conference on Computational Linguistics(COLING'02), page 390-396, Taipei, Taiwan.

[5] Li Yaoyong, Kalina Bontcheva, dan Hamish Cunningham. 2005, *SVM Based Learning System For Information Extraction*. Sheffield Machine Learning Workshop, *Lecture Notes in Computer Science*, Springer Verlag.

[6] Li Y., Shawe-Taylor, J., 2003, *The SVM with uneven margins and Chinese document categorization*. In Proceedings of The 17th Pacific Asia Conference on Language, Information and Computation (PACLIC17), page 216-227, Singapore.

[7] Mayfeld, J., McNamee, P., Piatko, C., 2003, *Named entity recognition using hundreds of thousands of features*. InWalter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 184-187. Edmonton, Canada.

[8] Paramita, 2008, *Penerapan Support Vector Machine untuk Ekstraksi Informasi dari Dokumen Teks*. Tugas Akhir Program Studi Teknik Informatika STEI Institut Teknologi Bandung.