Ekstraksi Kata Kunci Pada Jurnal Ilmiah Berbahasa Indonesia Menggunakan Conditional Random Field (CRF) Model

Keyword Extraction On Indonesian Scientific Journal Using Conditional Random Field (CRF) Model

Riki Akbar[1], Ade Romadhony, S.T., M.T.[2], Bedy Purnama, S.Si., M.T.[3]

[1,2,3]Fakultas Informatika Institut Teknologi Telkom
Jl. Telekomunikasi, Bandung

[1]riki_akbar@rocketmail.com, [2]ade.romadhony@gmail.com, [3]bdp@gmail.com

---

Abstrak

Pesatnya perkembangan internet menyebabkan peningkatan jumlah dokumen ilmiah seperti jurnal-jurnal ilmiah. Besarnya jumlah dokumen menyebabkan pemrosesan informasi membutuhkan waktu yang lama. Dengan mengkestrak kata kunci yang merepresentasikan isi dokumen, pemrosesan keseluruhan informasi pada dokumen tidak lagi diperlukan. Namun, ekstraksi kata kunci secara manual tidak efektif dan efisien dari segi waktu maupun sumber daya. Otomasi ekstraksi kata kunci diperlukan untuk mengatasi keterbatasan tersebut.

Pada Tugas Akhir ini akan diimplementasikan *Conditional Random Field* (CRF) model dengan menganalogikan proses ekstraksi kata kunci sebagai proses pelabelan *term* pada dokumen. Pemodelan ini membutuhkan proses pelatihan untuk menghasilkan parameter pendukung fitur yang optimal. Pengujian dilakukan untuk mengetahui pengaruh jumlah fitur, pendekatan ekstraksi fitur, jumlah data latih, serta keterlibatan eliminasi *stopwords* terhadap performansi proses ekstraksi kata kunci. Pengujian dilakukan terhadap dua kelompok dokumen dengan tingkat homogenitas berbeda (dokumen kedokteran dan kesehatan masyarakat) untuk mengetahui performansi ekstraksi kata kunci pada kedua kelompok dokumen tersebut.

Hasil pengujian menunjukkan bahwa performansi ekstraksi kata kunci terbaik pada dokumen kedokteran memberikan nilai *precision* 0.3892, *recall* 0.714, dan *f-measure* 0.4648 sementara performansi ekstraksi kata kunci terbaik pada dokumen kesehatan masyarakat memberikan nilai *precision* 0.2833, *recall* 0.692, dan *f-measure* 0.3901. Penambahan jumlah fitur dan keterlibatan eliminasi *stopwords* menyebabkan penurunan performansi ekstraksi kata kunci.

Kata Kunci : Ekstraksi kata kunci, kata kunci, *Conditional Random Field* (CRF) Model

---

Abstract

The development of internet technology increases the number of scientific documents including scientific journals. The large quantity of these documents affects the time required to process information. By extracting keyword which represents the content of document observed, information processing on the entire document is no longer required. However, extracting keyword manually is both ineffective and inefficient as it is time and resource consuming. Therefore, automation of keyword extraction is applied to handle the drawbacks caused by manual keyword extraction.

In this final project, Conditional Random Field (CRF) model will be implemented to extract keywords from document by viewing the keyword extraction process as a sequence labelling process. This model requires training process to produce optimum feature supporting parameters. Testing process will be conducted to figure out the influences of factors such as number of features used, feature extraction approach used, the number of training documents, and the involvement of *stopwords* elimination on keyword extraction performance achieved by the system. The testing process will be conducted on two groups of document (medicine and public health documents) that have different rate of content homogeinity in order to figure out performance achieved on each document group.

The result shows that best performance of keyword extraction on medicine documents produces *precision* of 0.3892, *recall* of 0.714, and *f-measure* of 0.4648 while the best performance of keyword extraction on public health documents produces *precision* of 0.2833, *recall* of 0.692, and *f-measure* of 0.3901. The increase of the number of feature used and involvement of *stopwords* elimination produce lower performance of keyword extraction.

Keywords : Keyword Extraction, Keyword, *Conditional Random Field* (CRF) Model

---

1. PENDAHULUAN

1.1. LATAR BELAKANG MASALAH

Seiring dengan pesatnya perkembangan teknologi internet, peningkatan jumlah teks elektronik pun terjadi secara pesat. Begitu pun dengan perkembangan signifikan yang terjadi pada dokumen-dokumen ilmiah seperti jurnal ilmiah, tesis, disertasi, dan karya-karya ilmiah lainnya [17]. Oleh karena itu, kebutuhan akan pemrosesan informasi yang efektif dan efisien baik dari segi waktu maupun akurasi pemrosesan informasi pada dokumen-dokumen ilmiah tersebut juga meningkat.

Ekstraksi kata kunci merupakan salah satu cara untuk memproses informasi pada suatu dokumen teks. Kata kunci adalah kumpulan kata penting dalam sebuah dokumen yang memberikan gambaran mengenai isi dari dokumen tersebut [17]. Proses ekstraksi kata kunci pada dasarnya bertujuan untuk mendapatkan gambaran singkat mengenai isi suatu dokumen yang direpresentasikan oleh kumpulan kata kunci tanpa harus memproses dokumen tersebut secara keseluruhan.

Karakteristik kata kunci pada tiap dokumen berbeda-beda dan cenderung dipengaruhi beberapa aspek. Salah satu aspek tersebut ialah aspek linguistik yang melandasi penulisan dokumen tersebut. Kata kunci pada dokumen berbahasa Indonesia memiliki karakteristik tersendiri seperti halnya kata kunci pada dokumen berbahasa lain. Kata kunci pada bahasa Indonesia umumnya berada pada awal dan akhir paragraf [7]. Hal ini berbeda dengan kata kunci pada bahasa Turki yang umumnya berada pada bagian tengah paragraf [17]. Selain itu, perbedaan afiliasi suatu dokumen juga dapat mempengaruhi perbedaan karakteristik kata kunci. Sebagai contoh, dokumen-dokumen jurnal ilmiah yang berasal dari disiplin ilmu eksakta memungkinkan adanya kata kunci yang terdiri dari karakter numerik sebagaimana halnya dokumen-dokumen jurnal non eksakta memungkinkan keberadaan kata kunci yang merupakan sebuah frasa. Perbedaan-perbedaan karakteristik kata kunci inilah yang menjadi tantangan dalam proses ekstraksi kata kunci.

Conditional Random Field (CRF) merupakan model probabilistik yang memetakan distribusi probabilitas kondisional dari kumpulan keluaran berupa label yang mungkin untuk suatu sekuens data (observasi) [6]. Dengan memetakan distribusi probabilitas kondisional secara keseluruhan dibandingkan memetakan distribusi probabilitas kondisional pada tiap state, CRF dapat mengoptimalkan proses-proses yang berinteraksi dengan sekuens data. Hal ini membuat CRF banyak diimplementasikan pada proses-proses yang berkaitan dengan segmentasi, pengenalan pola, dan pelabelan sekuens data.

Proses ekstraksi kata kunci pada dasarnya dapat dipandang sebagai proses pelabelan pada *term-term* yang terdapat dalam suatu sekuens *term* yakni dokumen. Ekstraksi kata kunci dilakukan dengan mengekstrak *term-term* yang berlabel kata kunci. Dengan demikian CRF dapat diimplementasikan untuk memodelkan proses ekstraksi kata kunci dengan menganalogikan proses ekstraksi kata kunci sebagai proses pelabelan sekuens *term* pada suatu dokumen.

1.2. RUMUSAN MASALAH

Permasalahan pada Tugas Akhir ini dirumuskan ke dalam beberapa hal yakni:

1. Bagaimana model CRF dapat mengekstraksi kata kunci dari suatu dokumen teks?
2. Bagaimana menentukan kumpulan parameter pendukung optimal pada model CRF untuk dapat mengekstraksi kata kunci seakurat mungkin?
3. Bagaimana pengaruh jumlah data latih, pemilihan fitur, dan jumlah fitur yang dilibatkan terhadap akurasi ekstraksi kata kunci yang dihasilkan oleh model CRF?
4. Bagaimana pengaruh eliminasi *stopwords* terhadap performansi model CRF dalam mengekstraksi kata kunci?
5. Bagaimanakah pengaruh homogenitas kelompok dokumen terhadap performansi ekstraksi kata kunci dengan model CRF?

Adapun batasan masalah pada pengerjaan Tugas Akhir ini adalah sebagai berikut:

1. Model CRF yang digunakan adalah *Linear-chained CRF Model*.
2. Dokumen teks yang digunakan sebagai data latih dan data uji pada Tugas Akhir ini adalah abstrak jurnal ilmiah berbahasa Indonesia.
3. Jurnal Ilmiah yang digunakan pada Tugas Akhir ini ialah jurnal ilmiah pada bidang Kedokteran dan Kesehatan Masyarakat.
4. *Term* yang digunakan pada data latih dan data uji adalah kata.
5. Atribut *term* yang digunakan untuk mengekstraksi kata kunci dibatasi hanya pada atribut statistik *term* tersebut.

1.3. TUJUAN

Tujuan dari Tugas Akhir ini adalah sebagai berikut:

1. Mengimplementasikan CRF untuk mengekstraksi kata kunci pada jurnal ilmiah berbahasa Indonesia.
2. Menganalisis pengaruh jumlah data latih, pemilihan dan jumlah fitur yang digunakan, serta keterlibatan eliminasi *stopwords* terhadap akurasi ekstraksi kata kunci yang dihasilkan oleh model CRF.
3. Menganalisis performansi CRF dalam mengekstraksi kata kunci pada dokumen dengan tingkat homogenitas yang berbeda.

2. LANDASAN TEORI

2.1 Ekstraksi Kata Kunci

2.1.1 Karakteristik Ekstraksi Kata Kunci

Kata kunci adalah sekumpulan kata penting pada sebuah dokumen yang dapat memberikan gambaran mengenai isi dokumen tersebut [17]. Proses ekstraksi kata kunci pada dasarnya adalah proses identifikasi kumpulan kata yang terdapat pada dokumen sehingga didapatkan kumpulan kata kunci yang representatif dan dapat menggambarkan isi dokumen secara umum. Ekstraksi kata kunci secara manual menjadi tidak efisien apabila pemrosesan informasi dilakukan pada kumpulan dokumen dalam jumlah yang besar. Oleh karena itulah, diperlukan adanya otomasi proses ekstraksi kata kunci yang dilakukan secara sistematis dengan meminimalisir intervensi manusia pada proses ekstraksi kata kunci tersebut [10][17].

Proses Ekstraksi kata kunci dapat dilakukan melalui dua perspektif *learning* yakni *supervised learning* dan *unsupervised learning*. *Supervised learning* berarti proses belajar yang membutuhkan acuan pengetahuan mengenai informasi yang akan diproses [14]. Proses ekstraksi kata kunci pada *supervised learning* akan mempelajari karakteristik kata kunci melalui sekumpulan dokumen menjadi dokumen latih. *Term-term* yang terdapat pada dokumen latih ini dilengkapi dengan label observasi. Label ini akan membedakan apakah suatu *term* merupakan kata kunci atau bukan berdasarkan kumpulan kata kunci pakar pada dokumen latih. Label observasi ini akan menjadi informasi belajar bagi CRF untuk memaksimalkan kemiripan hasil pemodelan dengan dokumen latih yang dimodelkan. Hasil pemodelan terhadap kumpulan dokumen latih ini berupa kumpulan parameter pendukung fitur yang akan menjadi acuan pada saat mengidentifikasi apakah suatu kata merupakan kata kunci atau bukan pada proses pengujian. Adapun pada *unsupervised learning* tetap dilakukan proses belajar pada kumpulan dokumen latih hanya saja proses belajar pada dokumen tersebut tidak disertai dengan adanya kelas label observasi. Pada tugas akhir ini penulis melakukan penelitian ekstraksi kata kunci menggunakan pendekatan *supervised learning*.

2.1.2 Metode Ekstraksi Kata Kunci

Secara umum terdapat dua pendekatan yang dapat digunakan untuk mengidentifikasi kata kunci pada sebuah dokumen yakni *keyword assignment* dan *keyword extraction*. Pada pendekatan *keyword assignment*, terdapat sebuah kamus *term* yang digunakan sebagai acuan untuk memilih kata kunci dengan tujuan mencocokkan kumpulan kata-kata pada observasi dengan kamus *term* tersebut sehingga didapatkan kata-kata yang diklasifikasikan sebagai kata kunci. Adapun pada pendekatan *keyword extraction*, kata kunci dipilih berdasarkan relevansi kata tersebut yang didapat dari identifikasi dan analisis atribut dasar kata seperti frekuensi kemunculan kata.

Metode ekstraksi kata kunci saat ini dapat dikelompokkan ke dalam empat kategori yaitu [20][10]:

a. Statistik Sederhana
   Metode ini berkonsentrasi pada perhitungan terhadap fitur-fitur statistik dari dokumen teks seperti frekuensi kemunculan *term*, panjang *term*, dan posisi *term*. Informasi statistik tersebut digunakan untuk mengidentifikasi apakah suatu kata dapat digolongkan sebagai kata kunci atau bukan.
b. Linguistik
   Metode ini menggunakan fitur linguistik dari kata, kalimat, dan dokumen. Fitur linguistik yang menjadi pertimbangan antara lain seperti *part-of-speech*, struktur sintaktik, dan semantik.
c. *Machine Learning*
   Metode ini dapat dipandang sebagai sebuah metode *supervised learning* yang mengekstraksi kata kunci berdasarkan model yang dihasilkan

oleh proses belajar pada kumpulan data latih. Contoh dari metode ini antara lain Naive Bayes, SVM, dan Bagging.

d. Metode Gabungan
Metode ini merupakan kombinasi dari keseluruhan dan atau beberapa metode sebelumnya dan mengakuisisi kelebihan-kelebihan yang ada pada metode sebelumnya seperti informasi statistik, informasi linguistik, dan model yang dihasilkan oleh proses belajar pada kumpulan data latih.

2.2 Conditional Random Field (CRF)

Conditional Random Field merupakan metode probabilistik yang memformulasikan distribusi probabilitas kondisional untuk seluruh sekuens label yang mungkin bagi suatu sekuens data. CRF banyak digunakan pada proses-proses yang berkaitan dengan sekuens seperti segmentasi, *POS Tagging* dan pengenalan pola. Salah satu keunggulan CRF adalah keberhasilannya mengatasi ketergantungan asumsi yang tinggi pada *Hidden Markov Model* serta potensi bias label pada *Maximum Entropy Markov Model*. Secara matematis model distribusi probabilitas kondisional sekuens label bagi suatu sekuens data dengan CRF dapat ditulis sebagai

$$P(x|z) = \frac{1}{Z(z)}\exp\left(\sum_{t\in[1,T-1]}\sum_{k\in[1,K]} w_k f_k(x_t, z, t)\right) \quad (2.1)$$

Di mana *x* adalah sekuens label, *z* adalah sekuens data, *Z(z)* adalah fungsi normalisasi, $f_k(x_t, z, t)$ adalah fitur ke *k* untuk data ke-*t* pada sekuens data *z*, dan w_k adalah parameter pendukung untuk fitur ke-k.

Secara umum, pembentukan model CRF terbagi ke dalam tiga fase yakni fase Ekstraksi Fitur, *learning*, dan *decoding* [16].

2.2.1 Ekstraksi Fitur

Pada fase ini, seluruh *data pattern* diekstrak dari sekuens data yang diobservasi. *Data pattern* nantinya akan digunakan untuk membangun kumpulan fitur. Pada pembentukan CRF, fitur terbagi dua yaitu fitur *node* dan fitur transisi. Fitur *node* hanya memperhitungkan *data pattern* pada data yang dievaluasi tanpa memperhitungkan transisi atau label yang telah diberikan pada *term* sebelumnya untuk menentukan label yang paling sesuai. Adapun fitur transisi turut memperhitungkan label yang telah diberikan pada *term* sebelumnya untuk menentukan label yang paling sesuai dengan data saat ini. Implementasi CRF umumnya dapat menggunakan salah satu dari kedua fitur tersebut atau bahkan menggunakan keduanya. Pada Tugas akhir ini fitur yang digunakan adalah fitur *node* dikarenakan pelabelan kata kunci pada sekuens *term* untuk *term* pada posisi tertentu tidak dipengaruhi oleh label pada *term* sebelumnya tetapi lebih dipengaruhi fitur-fitur *node* yang dimiliki *term* tersebut.

2.2.2 Learning

Pada fase ini, CRF akan mempelajari data latih sebagai dasar untuk membangun model distribusi probabilitas kondisional yang paling optimal. Salah satu cara untuk membangun model yang representatif ialah dengan memaksimalkan kemiripan model yang dihasilkan dengan data latih yang dimodelkan dengan menggunakan instrumen yang disebut *log-likelihood*. Nilai maksimal *log-likelihood* dapat dicapai dengan cara menyelesaikan turunan pertamanya. Penyelesaian turunan pertama *log-likelihood* tidak diselesaikan secara linear melainkan menggunakan pemrograman dinamis. Turunan pertama log-likelihood ditulis sebagai

$$G_k^t = \sum_{t\epsilon[1,T]} (f_k(x_t', z, t) - \sum_{x_t} P_t(x_t|z) f_k(x_t, z, t)) - \frac{w_k^2}{\sigma^2} \quad (2.2)$$

dengan $f_k(x_t', z, t)$ adalah fitur dari data latih yang telah dilabeli dan $f_k(x_t, z, t)$ adalah fitur dari data latih yang akan diuji dengan label pada himpunan label yang mungkin ($x=\{x_1, x_2, \ldots, x_m\}$) .

Selanjutnya *log-likelihood* dan turunannya akan dihitung secara iteratif dan didapatkan himpunan $w^*=\{w_1^*, w_2^*, \ldots, w_n^*\}$ yang merupakan himpunan parameter pendukung fitur terbaik berdasarkan proses latih pada data latih tersebut. Himpunan parameter pendukung fitur inilah yang akan digunakan untuk melakukan proses pengujian terhadap sekumpulan data uji.

2.2.3 Decoding

Himpunan parameter pendukung fitur yang didapatkan dari proses learning kemudian akan digunakan untuk melakukan *decoding* terhadap sekumpulan data uji yang

belum diberi label guna menemukan sekuens label yang paling optimal untuk sekuens data uji secara keseluruhan. Ada banyak metode yang dapat digunakan untuk melakukan pelabelan pada sekuens data uji. Salah satu metode yang umum digunakan pada proses *decoding* adalah Algoritma Viterbi. Salah satu keuntungan penggunaan algoritma Viterbi ialah algoritma ini memang dirancang untuk pengujian sekuens data dengan cara menelusuri seluruh sekuens dan melakukan *backtracking* untuk melakukan proses pelabelan pada sekuens dengan melibatkan kalkulasi bobot *node* dan kalkulasi bobot *edge* sehingga mengoptimalkan akurasi pelabelan sekuens data [2].

Proses decoding diawali dengan kalkulasi fitur yang telah diekstrak dari sekuens data. Selanjutnya, dengan memanfaatkan kumpulan parameter pendukung yang dihasilkan proses pelatihan, maka akan dihitung nilai maximal forward pass yang kemudian akan disimpan pada suatu variabel *bookkeeper* $Y_t[X_t]$. Selanjutnya pada proses backtracking akan dilakukan penelusuran ulang terhadap sekuens data dan tiap data dalam sekuens data akan dilabeli mengacu pada label yang memiliki nilai probabilitas kondisional yang terbaik untuk tiap data dan dinyatakan sebagai

$$x_t^* = Y_t[x_{t+1}^*] \quad (2.3)$$

sehingga sekuens label optimal adalah $\{x_1^*, x_2^*, \dots, x_T^*\}$ untuk sekuens data uji sejumlah *T*.

2.3 Ekstraksi Kata Kunci Menggunakan CRF

Ekstraksi kata kunci dengan CRF menggunakan pendekatan *supervised learning* membutuhkan sekumpulan dokumen latih yang telah dilabeli berdasarkan informasi pakar untuk dimodelkan distribusi kondisionalnya dengan menggunakan CRF. Pelabelan yang digunakan untuk melabeli sekuens data tidak dibatasi jumlahnya maupun jenisnya. Akan tetapi, pada tugas akhir ini hanya digunakan dua label saja untuk mengenali sekuens data yaitu label ”1” yang menunjukkan kata kunci dan ”0” yang menunjukkan kata yang bukan merupakan kata kunci. Dengan dua label tersebut, CRF dapat mengekstraksi kata kunci tanpa membatasi ukuran kata dari kata kunci seperti batasan 1-*gram* (kata kunci yang terdiri dari satu kata), *bigram* (kata kunci yang terdiri dari dua kata), atau *trigram* (kata kunci yang terdiri dari tiga kata).

Ada dua fase utama pada proses ekstraksi kata kunci menggunakan model CRF yakni fase Training menggunakan metode *Stochastic Gradient* dan fase Decoding menggunakan algoritma Viterbi.

2.3.1 Fase Training Menggunakan Metode Stochastic Gradient

Untuk mendapatkan nilai parameter pendukung fitur w_k yang optimal, proses pelatihan dilakukan dengan metode *Stochastic Gradient* yang mengimplementasikan pemrograman dinamis dengan memanfaatkan prosedur *forward-backward pass* untuk menelusuri sekuens *term*. Penelusuran dengan *forward-backward pass* bertujuan untuk mendapatkan nilai dari seluruh probabilitas lokal label pada sekuens *term*. Untuk itu dibutuhkan variabel yang menyimpan penelusuran *forward-backward* pada tiap *term* yang dikunjungi yaitu *forward variable* ($\alpha_t[x_t]$) dan *backward variable* ($\beta_t[x_t]$). *Forward variable* dan *backward variable* akan digunakan pada proses perhitungan nilai probabilitas kondisional lokal untuk tiap label yang mungkin bagi term z_t pada sekuens *z*. Secara matematis probabilitas kondisional lokal dapat ditulis

$$P_t(x_t|z) = \lambda_t \alpha_t[x_t] \phi_t(x_t, z) \beta_t[x_t] \quad (2.4)$$

Di mana λ_t adalah faktor normalisasi untuk memastikan bahwa $\sum_{x_t} P_t(x_t|z) = 1$.

Selanjutnya proses *update* parameter akan dilakukan secara iteratif berdasarkan persamaan

$$w_k \leftarrow w_k + \omega G_k^z \quad (2.5)$$

Di mana ω adalah *learning rate* dengan $\omega \in [0.001, 0.1]$.

2.3.2 Fase Decoding Menggunakan Algoritma Viterbi

Fase selanjutnya adalah *decoding* yang bertujuan untuk melabeli *term-term* dalam sekuens *term* dengan label yang bersesuaian berdasarkan nilai probabilitas kondisionalnya. Pada fase ini terdapat 2 subafase utama yaitu *maximal forward pass* dan *backtracking*.

Maximal forward pass bertujuan untuk mencari nilai maksimal *forward variable*

tiap *term*. Nilai maksimal *forward variable* tiap *term* kemudian akan disimpan pada *maximal forward variable* $\alpha_t^{max}[x_t]$. Nilai $\alpha_t^{max}[x_t]$ pada tiap label yang mungkin akan dibandingkan dan label dengan nilai $\alpha_t^{max}[x_t]$ yang lebih baik akan disimpan pada sebuah variabel *bookkeeper* Y_t. Y_t nantinya akan menyimpan label yang paling mungkin untuk setiap *term* z_t pada sekuens *z*.

Setelah seluruh bookkeeper Y_t terisi, maka proses selanjutnya yang dilakukan adalah penelusuran ulang (*backtracking*) untuk melakukan proses decoding terhadap label yang telah diberikan pada tiap *term* z_t sehingga didapatkan label optimal x_t^* dengan persamaan (2.3). Selanjutnya *term-term* dengan label "1" yang merepresentasikan kata kunci akan diekstrak.

2.4 *Data Pattern* dan Fitur

2.4.1 Data Pattern

Data pattern adalah informasi statistik yang didapatkan dari kumpulan *term* yang diamati baik pada proses pelatihan maupun pengujian. *Data pattern* tidak secara eksplisit merepresentasikan karakteristik *term* sehingga *data pattern* bersifat statis. *Data pattern* direpresentasikan sebagai fungsi yang menangani informasi awal yang didapatkan dari sekuens *term* dan mengembalikan sebuah nilai baik dalam bentuk keluaran biner (0 atau 1) maupun keluaran riil ([0,1]).

2.4.2 Fitur

Fitur merupakan fungsi dari *data pattern* dan label yang mungkin bagi suatu *term*. Fitur menjadi parameter pembeda antara *term* yang merupakan kata kunci dan *term* yang bukan merupakan kata kunci. Fitur akan mengembalikan nilai berbeda untuk setiap masukan label yang berbeda meskipun *data pattern* yang menjadi masukan bernilai sama sehingga fitur lebih bersifat dinamis dan dapat dimanipulasi untuk mencapai akurasi ekstraksi yang lebih baik.

Pada tugas akhir ini ada 8 fitur yang digunakan yaitu antara lain:

1. Kemunculan bersama *term* pada dokumen
2. Posisi *term* pada judul
3. Pembobotan TF-IDF
4. Kemunculan bersama *term* pada judul
5. Kemunculan *term* pada awal paragraf
6. Kemunculan *term* pada akhir paragraf
7. Kemunculan pertama *term* pada dokumen
8. Keberadaan huruf capital pada *term*

2.5 *Stopwords*

Stopwords adalah kata umum yang frekuensi kemunculannya tinggi tetapi tidak memiliki makna yang signifikan. Contoh *stopwods* pada bahasa Indonesia adalah kata "di", "ke", "dan", dan "yang". Keberadaan *stopwords* pada proses terkait temu kembali informasi, optimasi pencarian serta proses-proses *text mining* lainnya kerap dieliminasi. Hal ini dikarenakan *stopwords* memiliki makna yang tidak signifikan dan frekuensi kemunculannya yang tinggi memperlambat proses-proses tersebut.

2.6 Evaluasi

Evaluasi pada proses-proses *text mining* seperti ekstraksi kata kunci umumnya menggunakan tiga instrumen yaitu *precision*, *recall*, dan *F-measure*. *Precision* menunjukkan rasio antara kata kunci yang terambil dan relevan dengan kata kunci yang terambil yang dihasilkan oleh sistem sedangkan *recall* menunjukkan rasio antara kata kunci yang terambil dan relevan dengan kata kunci relevan yang dihasilkan pakar. Adapun *F-measure* merupakan rataan harmonis dari *precision dan recall*. Secara matematis *precision*, *recall*, dan *F-measure* dapat ditulis sebagai

$$Precision = \frac{Retrieved\ Keywords \cap Relevant\ Keywords}{Retrieved\ Keywords} \quad (2.6)$$

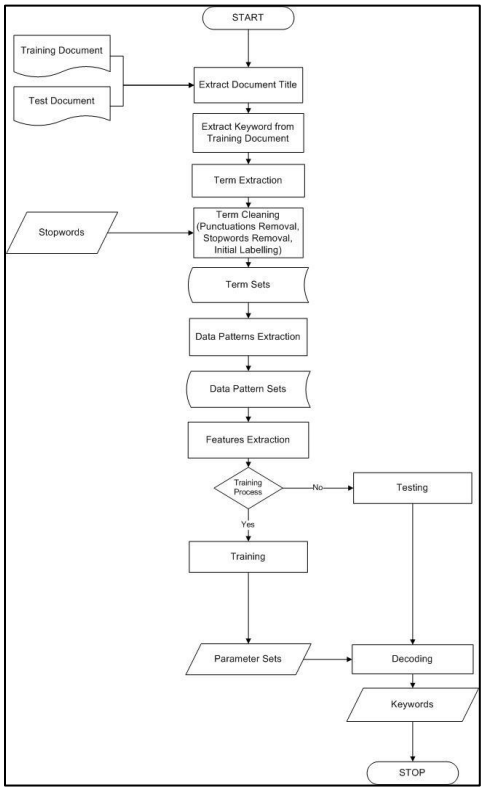
$$Recall = \frac{Retrieved\ Keywords \cap Relevant\ Keywords}{Relevant\ Keywords} \quad (2.7)$$

$$F - measure = 2.\frac{Precision\ .\ Recall}{Precision + Recall} \quad (2.8)$$

di mana *Retrieved Keywords* merupakan kata kunci yang dihasilkan oleh pemodelan CRF dan *Relevant Keywords* adalah kata kunci pakar yang terdapat pada dokumen uji. Pada Tugas Akhir ini, estimasi *error measures* yang digunakan adalah sebesar 30% merujuk pada penelitian-penelitian ekstraksi kata kunci sebelumnya [8]. Dengan demikian, pada Tugas Akhir ini berdasarkan persamaan (3.13),

persamaan (3.14), dan persamaan (3.15), nilai *precision*, *recall*, dan *f-measure* akan dikategorikan baik apabila memiliki nilai minimum masing-masing sebesar 0.7.

3. PERANCANGAN PERANGKAT LUNAK



Gambar 1 Gambaran Umum Sistem

Dalam pemodelan CRF, sekuens *term* akan diposisikan sebagai sekuens data yang akan dihubungkan dengan sekuens label yang bersesuaian berdasarkan kumpulan fitur yang diekstrak. Label merepresentasikan apakah suatu *term* merupakan kata kunci atau bukan. Fitur sendiri ditentukan berdasarkan *data pattern* yang diekstrak dari dokumen latih.

Kumpulan fitur selanjutnya akan dilatih dengan menggunakan metode *Stochastic Gradient* untuk memaksimalkan kemiripan hasil pemodelan dengan dokumen latih yang dimodelkan. Metode ini akan memperbaharui nilai parameter pendukung pada tiap iterasi sehingga akan didapatkan parameter yang paling optimal pada akhir iterasi. Proses pelatihan dilakukan dengan mengubah jumlah dokumen latih, jumlah fitur, memodifikasi dan mengkombinasi fitur, melibatkan eliminasi *stopwords*, serta melibatkan dokumen latih dengan tingkat homogenitas yang berbeda. Kumpulan parameter yang dihasilkan akan digunakan sebagai masukan dalam proses *decoding*.

Selanjutnya pada proses *decoding* akan dilakukan perhitungan nilai tiap fitur beserta parameter pendukung fitur yang dihasilkan oleh proses pelatihan sebelumnya sehingga didapatkan probabilitas kondisional tiap label yang mungkin untuk setiap *term*. Kemudian *term* akan dilabeli dengan label yang memiliki probabilitas kondisional yang lebih besar. Keluaran dari sistem adalah kumpulan *term* yang telah dilabeli sebagai kata kunci. Performansi sistem akan didapatkan dengan membandingkan kata kunci yang dihasilkan sistem dengan kata kunci pakar pada dokumen uji menggunakan parameter *Precision*, *Recall*, dan *F-measure*.

4. PENGUJIAN DAN ANALISIS

4.1 Pengujian Sistem

Pengujian sistem dilakukan untuk mengetahui pengaruh masing-masing elemen konfigurasi pada proses pelatihan dalam proses ekstraksi kata kunci. Pengujian sistem dilakukan pada dua kelompok dokumen yakni kelompok dokumen kedokteran dan kelompok dokumen kesehatan masyarakat dengan membandingkan kata kunci yang dihasilkan sistem dengan kata kunci pakar berdasarkan parameter pengukur performansi yakni *precision*, *recall*, dan *F-measure*.

4.2 Tujuan Pengujian

Tujuan dari dilakukannya pengujian sistem yaitu:

1. Menganalisis pengaruh jumlah dokumen latih, jumlah fitur dan pendekatan fitur yang digunakan, serta keterlibatan *stopwords* dalam proses ekstraksi kata kunci.
2. Menganalisis performansi ekstraksi kata kunci menggunakan model CRF pada kelompok dokumen yang memiliki tingkat homogenitas berbeda yakni pada kelompok dokumen kedokteran dan kelompok dokumen kesehatan masyarakat.
3. Menemukan konfigurasi latih yang memberikan performansi ekstraksi kata kunci yang paling baik.

4.3 Strategi Pengujian

Pengujian akan dilakukan terhadap 50 dokumen kedokteran dan kesehatan masyarakat yang tidak dilibatkan pada proses pelatihan. Ada empat skenario utama pengujian sistem yang akan dilakukan terhadap masing-masing kelompok

dokumen kedokteran dan kesehatan masyarakat yaitu:

1. Skenario 1 (tanpa eliminasi *stopwords* dan menggunakan pendekatan ekstraksi fitur *existence-based*)
2. Skenario 2 (tanpa eliminasi *stopwords* dan menggunakan pendekatan ekstraksi fitur *appearance-based*)
3. Skenario 3 (dengan eliminasi *stopwords* dan menggunakan pendekatan ekstraksi fitur *existence-based*)
4. Skenario 4 (dengan eliminasi *stopwords* dan menggunakan pendekatan ekstraksi fitur *appearance-based*)

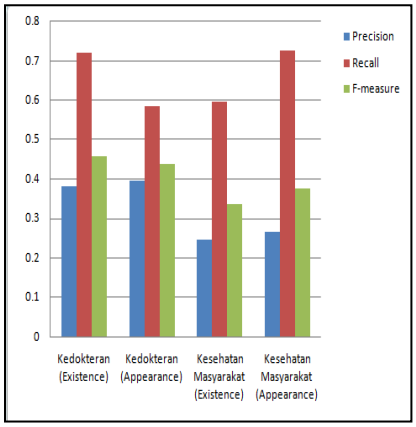
4.4 Analisis Hasil Pengujian

4.4.1 Analisis nilai parameter pendukung fitur

Nilai parameter yang dihasilkan menunjukkan kemampuan suatu fitur untuk membedakan apakah suatu *term* merupakan kata kunci atau non kata kunci. Berdasarkan Tabel 6 dan Tabel 7, nilai parameter pendukung untuk fitur kemunculan pada judul (w_2), fitur tf-idf (w_3), fitur kemunculan bersama pada judul (w_4), dan fitur kemunculan pada awal paragraf (w_5) relatif lebih tinggi dibandingkan dengan nilai parameter pendukung fitur lainnya. Ini berarti *term-term* kata kunci pakar pada proses pelatihan memiliki nilai relatif tinggi untuk fitur-fitur tersebut sementara *term-term* non kata kunci pakar memiliki nilai relatif rendah untuk fitur-fitur tersebut. Dengan kata lain, *term* kata kunci memiliki kecenderungan tinggi untuk muncul pada judul, memiliki kemunculan yang unik pada kumpulan dokumen, serta muncul pada awal paragraf.

Perbedaan nilai parameter pendukung fitur pun didapat dari proses pelatihan menggunakan pendekatan ekstraksi fitur yang berbeda. Umumnya nilai parameter pendukung fitur yang dihasilkan pelatihan menggunakan pendekatan ekstraksi fitur *appearance-based* lebih tinggi dibandingkan nilai parameter pendukung fitur yang dihasilkan pelatihan menggunakan *existence-based*. Ini dikarenakan pendekatan *appearance-based* memperhitungkan posisi kemunculan sehingga meskipun suatu *term* non kata kunci muncul pada bagian judul dan awal paragraf, *term* non kata kunci tersebut tetap akan mendapatkan nilai rendah untuk fitur-fitur bersesuaian apabila nilai posisi kemunculan *term* non kata kunci tersebut relatif rendah.

4.4.2 Analisis pengaruh pendekatan ekstraksi fitur



Gambar 2 Perbandingan Performansi Ekstraksi Kata Kunci dengan Pendekatan Ekstraksi Fitur Berbeda

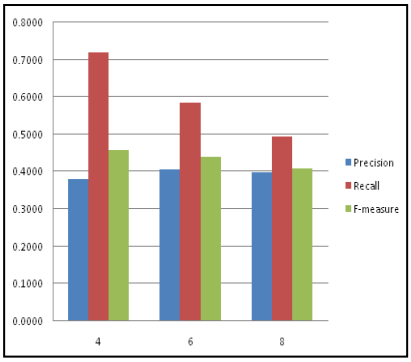
Berdasarkan nilai *f-measure* yang dicapai, performansi terbaik yang dicapai pada pengujian pada dokumen kedokteran didapatkan dengan menggunakan pendekatan *existence-based* sedangkan performansi terbaik yang dicapai pada pengujian pada dokumen kesehatan masyarakat didapatkan dengan menggunakan pendekatan *appearance-based*.

Penyebab perbedaan performansi tersebut adalah perbedaan tingkat keunikan *term-term* kata kunci berdasarkan sebaran nilai tiap *data pattern* serta kemunculan *term* pada kelompok dokumen kedokteran dan dokumen kesehatan masyarakat. Tingkat keunikan *term-term* kata kunci pada dokumen kedokteran relatif lebih tinggi dibandingkan

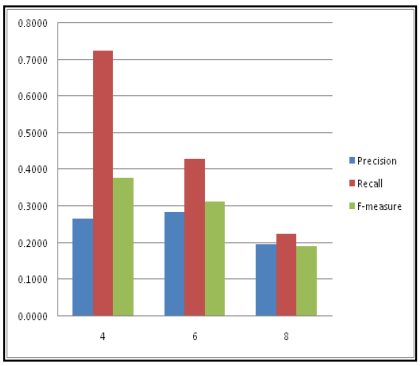
dengan keunikan *term-term* kata kunci pada dokumen kesehatan masyarakat. Dengan demikian apabila ekstraksi kata kunci pada dokumen kedokteran dilakukan dengan pendekatan *appearance-based* maka kata kunci yang dihasilkan cenderung *overfit*. Sebaliknya, apabila ekstraksi kata kunci pada dokumen kesehatan masyarakat dilakukan dengan pendekatan *existence-based*, kata kunci yang dihasilkan cenderung tidak optimal karena jumlah kata kunci yang dihasilkan terlalu banyak akibat dari tingkat kemiripan *term* kata kunci dan *term* non kata kunci relatif lebih tinggi dibandingkan pada dokumen kedokteran.

4.4.3 Analisis pengaruh jumlah fitur dan jumlah dokumen yang digunakan

Hasil pengujian dengan perubahan jumlah fitur yang diekstrak menunjukkan nilai *recall* untuk masing-masing kelompok dokumen cenderung berbanding terbalik dengan jumlah fitur. Hal ini dikarenakan penggunaan jumlah fitur yang semakin banyak akan semakin membatasi karakteristik *term-term* yang diduga sebagai kata kunci. Namun demikian, penggunaan fitur yang terlampau banyak dan tidak representatif justru akan mengurangi akurasi ekstraksi kata kunci karena tidak seluruh *term* kata kunci memiliki seluruh karakteristik yang bersesuaian dengan fitur. Dengan demikian, semakin bertambah fitur maka akan semakin berkurang jumlah kata kunci yang dihasilkan sistem yang sesuai dengan kata kunci pakar sehingga mengakibatkan nilai *recall* mengalami penurunan untuk tiap penambahan fitur.



Gambar 3 Perbandingan Performansi Rata-rata Berdasarkan Jumlah Fitur pada Dokumen Kedokteran

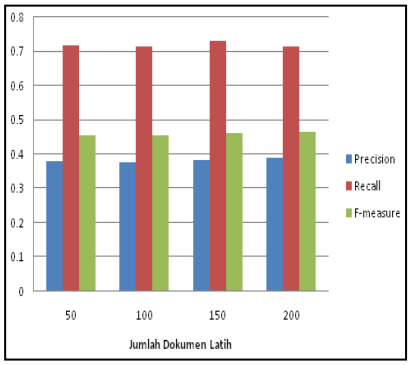


Gambar 4 Perbandingan Performansi Rata-rata Berdasarkan Jumlah Fitur pada Dokumen Kesehatan Masyarakat

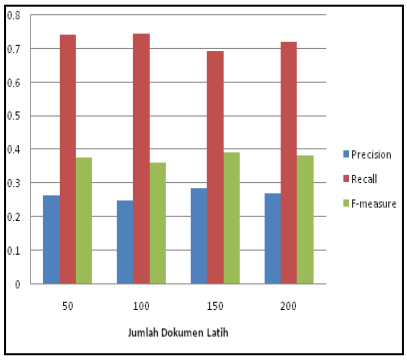
Dari hasil pengujian terhadap jumlah dokumen latih, tidak didapatkan suatu kecenderungan umum dari tiap-tiap parameter pengukur performansi baik pada pengujian terhadap dokumen kedokteran maupun dokumen kesehatan masyarakat. Perubahan nilai *f-measure* yang merepresentasikan performansi ekstraksi kata kunci cenderung fluktuatif dan tidak mengikuti suatu pola umum terkait dengan penambahan jumlah dokumen latih. Perubahan performansi lebih dipengaruhi oleh jumlah fitur yang digunakan.

Hal ini terjadi dikarenakan penambahan jumlah dokumen latih hanya berpengaruh kepada perubahan sebaran *term* yang ditinjau dari sebaran *pattern* yang diekstrak. Akan tetapi, perubahan sebaran ini pun tidak signifikan dikarenakan karakteristik *term* dari dokumen latih

tambahan secara umum memiliki kemiripan dengan karakteristik *term* dari dokumen latih yang telah ada sebelumnya. Dengan demikian, penambahan jumlah dokumen latih tidak berpengaruh signifikan pada perubahan performansi ekstraksi kata kunci yang dihasilkan baik pada kelompok dokumen dengan homogenitas tinggi maupun pada kelompok dokumen dengan homogenitas rendah.



Gambar 5 Perbandingan Performansi Terbaik Berdasarkan Jumlah Dokumen Latih pada Dokumen Kedokteran



Gambar 6 Perbandingan Performansi Terbaik Berdasarkan Jumlah Dokumen Latih pada Dokumen Kesehatan Masyarakat

4.4.4 Analisis Pengaruh Eliminasi Stopwords

Hasil pengujian pengaruh eliminasi *stopwords* menunjukkan bahwa terjadi penurunan performansi ekstraksi kata kunci yang ditinjau dari parameter *f-measure* antara pengujian dengan dan tanpa eliminasi *stopwords*. Penurunan performansi ekstraksi kata kunci dengan menggunakan eliminasi *stopwords* ini dikarenakan pengikutsertaan eliminasi *stopwords* menyebabkan sebaran *pattern* sekuens *term* menurun sehingga tingkat kemiripan karakteristik antara *term* kata kunci dan *term* non kata kunci menjadi lebih tinggi baik pada dokumen kedokteran maupun kesehatan masyarakat dibandingkan bila ekstraksi kata kunci tidak melibatkan eliminasi *stopwords*. Selain itu, dengan melibatkan eliminasi *stopwords*, pemodelan terhadap perbedaan karakteristik kata kunci *monogram* dan *n-gram* akan semakin tidak representatif dikarenakan *stopwords* yang menjadi *marker* dalam sekuens telah dieliminasi sehingga *term* kata kunci *monogram* dan *term* kata kunci *n-gram* akan muncul secara berdampingan. Hal ini akan menurunkan performansi pelabelan dikarenakan kata kunci *monogram* dan *n-gram* yang berdampingan akan dianggap sebagai satu kesatuan kata kunci sehingga jumlah kata kunci relevan yang dihasilkan akan mengalami penurunan yang berimbas pada penurunan nilai *recall*. Dengan demikian nilai *recall* terbaik yang dicapai dengan pengujian yang melibatkan eliminasi *stopwords* lebih rendah dibandingkan dengan nilai *recall* terbaik yang dicapai pengujian tanpa melibatkan eliminasi *stopwords*.

4.4.5 Analisis Kata Kunci Sistem

Pada pengujian terhadap kedua kelompok dokumen didapatkan bahwa homogenitas kelompok dokumen memengaruhi jumlah kata kunci rata-rata yang dihasilkan sistem. Pengujian pada kelompok dokumen yang lebih homogen yakni dokumen kesehatan masyarakat menghasilkan jumlah kata kunci rata-rata yang lebih tinggi. Hal ini dikarenakan pada dokumen homogen yakni kesehatan masyarakat, sebaran *pattern* lebih konvergen sehingga menghasilkan sebaran nilai masing-masing fitur yang tingkat keragamannya lebih rendah

dibandingkan dengan sebaran nilai masing-masing fitur pada dokumen kedokteran meskipun ekstraksi fitur dilakukan dengan pendekatan yang berbeda. Hal ini menyebabkan nilai fitur yang dimiliki oleh *term* kata kunci dan non kata kunci mengalami perbedaan nilai yang tipis sehingga lebih banyak term yang dilabeli sebagai kata kunci. Dengan demikian, jumlah kata kunci yang dihasilkan akan lebih sedikit dibandingkan dengan jumlah kata kunci yang dihasilkan pada pengujian terhadap dokumen kesehatan masyarakat.

Pembatasan jumlah kata kunci yang dihasilkan oleh sistem menghasilkan penurunan performansi ekstraksi kata kunci yang cukup signifikan. Hal ini disebabkan oleh banyaknya *term* kata kunci sistem yang relevan dengan kata kunci pakar yang memiliki nilai probabilitas kondisional label yang relatif rendah. Ini dipengaruhi oleh kurang representatifnya kata kunci pakar berdasarkan *pattern* yang diekstrak dari kata kunci pakar.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan yang didapatkan dari pelaksanaan Tugas Akhir ini ialah:

1. Pada pengujian terhadap Dokumen Kedokteran performansi optimal dicapai dengan nilai *precision* 0.3892, nilai *recall* 0.714, dan nilai *F-measure* 0.4648 melalui konfigurasi latih yang terdiri dari penggunaan 200 dokumen latih, 4 fitur, pendekatan ekstraksi fitur *existence-based*, dan tanpa melibatkan eliminasi *stopwords*.
2. Pada pengujian terhadap Dokumen Kesehatan Masyarakat performansi optimal dicapai dengan nilai *precision* 0.2833, nilai *recall* 0.692, dan nilai *F-measure* 0.3901 melalui konfigurasi latih yang terdiri dari penggunaan 150 dokumen latih, 4 fitur, pendekatan ekstraksi fitur *appearance-based*, dan tanpa melibatkan eliminasi *stopwords*.
3. Berdasarkan estimasi *error measures* sebesar 30%, performansi CRF dalam proses ekstraksi kata kunci secara keseluruhan belum dapat dikategorikan baik.
4. CRF menghasilkan performansi yang lebih baik untuk pengujian terhadap dokumen dengan homogenitas lebih rendah.
5. Penggunaan pendekatan ekstraksi fitur *existence-based* menghasilkan performansi terbaik pada pengujian terhadap kelompok dokumen dengan homogenitas rendah sementara pendekatan ekstraksi fitur *appearance-based* menghasilkan performansi terbaik pada pengujian terhadap kelompok dokumen dengan homogenitas tinggi.
6. Penambahan jumlah fitur menyebabkan penurunan performansi ekstraksi kata kunci. Pada pengujian terhadap kedua kelompok dokumen, penggunaan 4 fitur menghasilkan performansi yang paling baik.
7. Penambahan jumlah dokumen latih tidak memberikan pengaruh signifikan terhadap performansi ekstraksi kata kunci yang dihasillkan.
8. Pengujian dengan menggunakan eliminasi *stopwords* menghasilkan performansi ekstraksi kata kunci yang lebih rendah dibandingkan pengujian tanpa eliminasi *stopwords*.
9. Kendala utama yang ditemui pada proses pelatihan adalah tidak representatifnya kata kunci pakar. Tidak representatifnya kata kunci pakar mencakup ketidakmunculan kata kunci pakar pada bagian tertentu dokumen dan tingkat keunikan *pattern* yang rendah dari kata kunci pakar.

5.2 Saran

Saran yang penulis berikan untuk kelanjutan penelitian mengenai ekstraksi kata kunci dengan CRF model antara lain:

1. Peningkatan keragaman *pattern* yang diekstraksi dari dokumen latih dengan tidak hanya melibatkan atribut statistik tetapi juga atribut linguistik seperti relasi morfologis (struktur kata), sintaksis (struktur antar kata), dan semantik (pemaknaan).
2. Pengimplementasian metode latih lain seperti *Voted Perception*, *Conjugate Gradients*, dan metode lainnya untuk menghasilkan kumpulan parameter pendukung fitur yang optimal serta melakukan penelaahan terhadap pengaruh metode-metode latih tersebut

terhadap parameter pendukung fitur yang dihasilkan.

3. Pengujian dilakukan terhadap dokumen-dokumen dari disiplin ilmu yang berbeda dengan jangkauan homogenitas dokumen yang lebih luas.

DAFTAR PUSTAKA

[1] El-Khair, Ibrahim Abu. 2006. *Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study*. *International Journal of Computing & Information Sciences*.

[2] Forney, G.D. 1973.*The Viterbi Algorithm. Proceedings of the IEEE*.

[3] Gupta, Rahul. 2005. *Conditional Random Fields*. Mumbai. Indian Institute of Technology (IIT).

[4] Hulth, Anette. 2003. *Improved Automatic Keyword Extraction Given More Linguistic Knowledge*. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '03)*.

[5] Klinger, Roman, Katrin Tomanek. 2007. *Conditional Probabilistic Models and Conditional Random Fields*. Faculty of Computer Science Dortmund University of Technology.

[6] Lafferty, J., McCallum, A., Pereira, F. 2001. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann.

[7] Maharsi, Lisa. 2010. Ekstraksi Kata Kunci pada Dokumen Teks Menggunakan Metode Naive Bayes. Institut Teknologi Telkom. Bandung.

[8] Makhoul, John, Francis Kubala, Richard Schwartz, Ralph Weischedel. 1999. *Performance Measures for Information Extraction*. BBN Technologies, GTE Corp.

[9] Malouf, Robert. 2002. *A Comparison of Algorithms for Maximum Entropy Parameter Estimation*. *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, Taipei.

[10] Oelze, Iryna. 2009. *Automatic Keyword Extraction for Database Search*. Leibniz University. Hannover.

[11] Pinto, David, Andrew McCallum, Xing Wei, Bruce Croft. 2003. *Table Extraction Using Conditional Random Fields*. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*.

[12] Renz, Ingrid, Andrea Ficzay, Holger Hitzler. 2010. *Keyword Extraction for Text Characterization*. Information Mining DaimlerChrysler AG, Research and Technology.

[13] Surakhmad, Winarno. 2004. Pengantar Penelitian Ilmiah. Bandung: Tarsito.

[14] Suyanto. 2007. *Artificial Intelligence*. Penerbit Informatika. Bandung.

[15] Truyen, Tran The, Phung, Dinh. 2008. *A Tutorial on the Maths behind Conditional Random Fields for Sequential Labelling*. Curtin University of Technology.

[16] Truyen, Tran The, Phung, Dinh. 2008. *A Practitioner Guide to Conditional Random Fields for Sequential Labelling*. Curtin University of Technology.

[17] Uzun, Yasin. 2005. *Keyword Extraction Using Naïve Bayes*. Department of Computer Science, Bilkent University

[18] Wallach, Hanna M. 2004. *Conditional Random Fields: An Introduction*. *Technical Report MS-CIS-04-21*, University of Pennsylvania.

[19] Wartena, Christian, Rogier Brusse, Wout Slakhorst. 2010. *Keyword Extraction Using Co-occurrence*. *Proceedings of the 2010 Workshops on Database and Expert Systems Application*.

[20] Zhang, Chengzhi, Wang Huilin, Liu Yao, Wu Dan, Liao Yi, Wang Bo. 2008. *Automatic Keyword Extraction from Documents Using Conditional Random Fields*. Journal of Computational Information Systems.