DETEKSI BAHASA UNTUK DOKUMEN TEKS BERBAHASA INDONESIA

Amir Hamzah[1)]
[1)]Jurusan Teknik Informatika IST AKPRIND Yogyakarta
Jl. Kalisahak No.28 Komp.Balapan 55222 Yogyakarta Telp (0274)-563029
e-mail : miramzah@yahoo.co.id

Abstract

In the multi language environment corpus such as Internet, the information retrieval system has faced difficulties that caused by the mixture of language document response of single query request that do not match the user need. One approach to handle this problem is by designing cross-language search engine. On the other hand this solution is no need for the user that only hoped the document answer only in one language such as Bahasa Indonesia. In the second case the solution is by designing search engine in certain language. In the construction of special language search engine in multi language environment, a critical step is language detection of the document being analyzed. This research was aimed to study comparison of several methods of language detection based on N-gram, i.e. unigram, bigram and trigram. Several news text documents in Bahasa Indonesia from 100 documents until 3000 document, two academic document collections of 88 and 450 documents and two abstract collection and full paper collection in English, each of those is 40 documents, were used as test collection. The results showed that unigram, bigram and trigram were good parameter to detect the language of documents. Among those methods, bigram was the best in time complexity and accuracy

Keywords: *Language detection, Search engine, N-gram*

1. PENDAHULUAN

Sumber informasi *on-line* yang berkembang pesat telah menyebabkan ledakan informasi. Saat ini diperkirakan ada sekitar 20 milyar dokumen terindeks dengan lebih dari 106 bahasa (www.google.com ,2009) dengan jutaan halaman web multi bahasa bertambah setiap hari. Volume dokumen yang besar menimbulkan kesulitan dalam organisasi, navigasi, temu kembali dan *summarization* (Xu et.al., 2003). Di samping itu koleksi dokumen teks hampir melibatkan seluruh bahasa di dunia (Adriani, 2002). Meningkatnya jumlah bahasa dalam *web* menambah kompleksitas problem dalam Sistem Temu Kembali Informasi (**STKI**).

Meningkatnya jumlah bahasa dalam *web* telah menjadi tantangan baru penelitian STKI. Saat ini masih sangat sedikit penelitian dibidang STKI yang berbasis bahasa Indonesia (Nazief, 2000; Vega, 2001). Menurut Asian et.al.(2004), Indonesia dengan jumlah penduduk diatas dua ratus juta saat ini sangat memerlukan penelitian di bidang STKI bahasa Indonesia.

Penelitian Hamzah (2009) tentang model *retrieval* berbasis konsep telah menghasilkan beberapa kesimpulan penting antara lain penerapan *clustering* dokumen dapat meningkatkan kinerja *retrieval* teks secara signifikan Diperlukan penelitian lanjutan untuk menyusun prototipe *search engine* untuk dokumen teks bahasa Indonesia. Penelitian lanjut tersebut menyangkut teknik mendeteksi bahasa (*language detector*), teknik menyusun *spider* (program pengamat situs) dan teknik menyusun *crawler* (program pen*download* otomatis), yang memerlukan tahapan penelitan berkelanjutan. Pada penelitian ini akan difokuskan pada langkah kritis dalam penyusunan mesin pencari dokumen bahasa Indonesia, yaitu langkah deteksi dokumen berbahasa Indonesia atau bukan. Penelitian difokuskan pada model deteksi menggunakan N-*gram*, yaitu *unigram*, *bigram* dan *trigram*.

Penelitian ini bertujuan untuk mengkaji metode-metode pendeteksian dokumen berbahasa Indonesia atau bukan. Secara lebih rincinya hal-hal yang ingin dilakukan dan menjadi tujuan dalam penelitian ini adalah mencari algoritma-algoritma deteksi dokumen yang efektif untuk mendeteksi suatu dokumen berbahasa Indonesia atau bukan. Algoritma yang akan dikaji adalah berbasis N-gram, yang menyangkut *Unigram, Bi-gram dan Tri-gram.*

Kebanyakan mesin pencari yang ditemukan adalah dari jenis *general search engine*, yaitu mesin pencari yang memandang web sebagai suatu korpus teks yang mencampurkan segala jenis teks. Pada kenyataannya seseorang mencari informasi tidak jarang untuk tujuan khusus, misalnya mencari referensi akademis penelitian, mencari berita tertentu dan lain-lain. Pada kasus terakhir ini *general search engine* memiliki kelemahan berkaitan dengan besarnya ukuran korpus yang menyebabkan jawaban tidak akurat, sehingga *special search engine* akan lebih menjanjikan.

Manfaat dan kontribusi dari hasil penelitian ini yang terpenting adalah untuk mendapatkan rekomendasi teknik deteksi bahasa seperti apakah yang paling efisien untuk mendeteksi dokumen itu berbahasa Indonesia atau bukan. Rekomendasi ini diharapkan dapat ditemukan dari pilihan tiga pendekatan *unigram*, *bi-gram* dan *tri-gram*. Metode ini akan sangat bermanfaat untuk menyeleksi secara otomatis dokumen berbahasa tertentu

ditengah melimpahnya koleksi dokumen multi-bahasa dalam koleksi internet. Metode deteksi bahasa juga sangat fital dalam perancangan mesin pencari khusus (*special search engine*).

2. TINJAUAN PUSTAKA

Deteksi bahasa menjadi isu penting semenjak makin banyaknya jenis bahasa yang terlibat dalam dokumen *online*. Salah satu contoh konkritnya adalah pustakawan yang berhubungan dengan banyak dokumen multi bahasa dan harus secara cepat menentukan bahasa apa yang digunakan dalam dokumen. Pada perancangan mesin pencari (*search engine*) dengan korpus khusus, salah satu kebutuhan awal sistem adalah kebutuhan untuk mendeteksi bahasa dari suatu dokumen yang akan digunakan (Vega and Bressan, 2000).

Menurut Sibun and Reynar (1996) sejumlah variasi fitur telah digunakan untuk kebutuhan deteksi bahasa antara lain kehadiran karakter khusus, kata khusus dan n-*gram* khusus. Pada deteksi otomatis dapat diterapkan analysis statistik seperti *discriminant analysis* atau *markov model*. Sibun anda Reynar (1996) mengajukan metode identifikasi bahasa dengan menggunakan ukuran *entropy*, yaitu pendekatan statistik probabilitas. *Entropy* relatif dari dua distribusi probabilitas menunjukkan besarnya tambahan informasi pada distribusi kedua dengan menggunakan kode optimal dari distribusi pertama. Jika distribusi pertama adalah distribusi *events* dari suatu jenis bahasa tertentu yang telah diketahui dan distribusi kedua adalah jenis distribusi *events* yang sama dari dokumen yang akan dilacak bahasanya maka keputusan apakah bahasa dari dokumen yang diuji dapat ditetapkan dengan mengamati perbedaan distribusi. Jika bahasa yang akan diuji banyak maka bahasa dari dokumen baru diputuskan berdasarkan nilai *entropy* minimal antar distribusi dokumen baru dengan distribusi bahasa-bahasa dalam tahap pelatihan. Events yang ditentukan distribusi probabilitas dapat berupa *unigram*, *bigram* atau *trigram*.

Bastrup and Popper (2003) mengajukan metode deteksi bahasa yang cukup sederhana dengan menyusun pohon keputusan dengan dasar distribusi unigram. Diasumsikan bahwa setiap bahasa akan memiliki keunikan distribusi *unigram*. Pada tahap *training* distribusi *unigram* tiap bahasa ditentukan. Selanjutnya pohon keputusan dibuat dengan setiap *node* dari pohon adalah *unigram* dan cabang menunjukkan nilai distribusinya sedangkan node akhir (*leaf*) adalah bahasa yang ditetapkan. Penggunaan pohon diawali dengan mula-mula distribusi unigram dari dokumen yang akan diuji ditentukan. Selanjutnya nilai unigram dari dokumen yang diuji digunakan untuk melacak bahasa yang akan ditetapkan. Dengan eksperimen menggunakan 10 bahasa-bahasa eropa metode ini memiliki akurasi antara 67% sampai 73%. Metode yang mirip dilakukan oleh Vega dan Bressan (2000) dengan menggunakan *trigram* untuk menetapkan apakah dokumen berbahasa Inndonesia atau bukan. Dengan *training* menggunakan 10.167 kata Indonesia metode ini mampu mendeteksi bahasa dengan presisi di atas 88%.

2.1 Pengertian Deteksi Bahasa

Deteksi bahasa (*language detection*), biasa juga disebut identifikasi bahasa (*language identification*) usaha untuk menentukan jenis bahasa secara otomatis (dengan program komputer) dari suatu teks atau dokumen berdasarkan kriteria-kriteria tertentu yang harus dipenuhi. Menurut Sibun and Reynar (1996) ada beberapa pertimbangan yang harus diperhatikan dalam upaya deteksi atau identifikasi bahasa dari suatu teks atau dokumen, yaitu :

1. **Tipe fitur** yang digunakan, apakah akan digunakan karakter, kata, n-*gram*. Apakah akan digunakan aturan linguistik seperti morfologi, *orthography* atau *capitalization*?
2. **Bentuk analisis**, apakah akan digunakan algoritma manual, setengah manual atau sepenuhnya otomatis.
3. **Bentuk *encoding***, apakah representasi kerja yang terbaik digunakan? Akurasi, robustness atau kecepatan. Apakah akan digunakan *encoding* menggunakan karakter, *simplified character* atau *shape of character*.
4. **Konstituensi pool bahasa,** apakah akan diterapkan pada semua bahasa yang berbasis *roman-alphabet* atau hanya sebagian bahasa yang akan diminati.
5. **Bentuk input**, apakah sistem akan menerima masukan berupa karakter teks, citra dari karakter teks atau keduanya.
6. **Ukuran teks**, apakah kita akan mengidentifikasi sebuah kata, sebuah kalimat, sebuah paragraph atau sebuah dokumen. Tidak jarang dijumpai kepentingan kita adalah mengidentifikasi bahasa dari dokumen utama dengan mengabaikan kemungkinan ada teks berbahasa lain dalam dokumen tersebut. Sistem yang baik mengidentifikasi pada potongan kecil teks maka apabila diterapkan pada suatu dokumen maka dapat ditempuh analisis dengan *random sampling* pada sebagian teks yang dianalisis secara lebih cepat.

2.2 Metode Deteksi Bahasa

Metode deteksi bahasa berpijak pada keberadaan fitur yang dipergunakan sebagai kriteria deteksi. Beberapa fitur yang telah dipergunakan antara lain adalah: kehadiran karakter khusus (Ziegler, 1991), kehadiran kata tertentu (Batchelder, 1992), kemunculan n-*gram* khusus (Souter et.al., 1994), bentuk kata tertentu (*shaped-word*) dalam citra karakter (Sibun and Spitz, 1994) dan frekuensi kemunculan n-*gram*.

Metode untuk deteksi juga melibatkan banyak pendekatan dan teknik, seperti pendekatan yang murni manual, semi otomatis sampai otomatis secara penuh. Teknik yang digunakan dapat berupa penerapan jaringan syaraf tiruan (Batchelder, 1992), penerapan sistem pakar (Ziegler, 1991), penerapan analisis diskriminan (Sibun and Spitz, 1994), penerapan model bahasa (Beesley, 1988), dan penerapan model Markov (Dunning, 1994).

N-gram

N-gram adalah potongan N-karakter yang diambilkan dari suatu string. Untuk mendapatkan N-gram yang utuh ditempuh dengan menambahkan blank pada awal dan akhir string. Misalnya suatu string "TEXT" setelah ditambah aal dan akhir dengan "_" sebagai pengganti blank akan didapat N-gram sebagai berikut :

Unigram : T,E,X,T
Bigram : _T, TE, EX,XT, dan T
Trigram : _TE,TEX,EXT, XT_ dan T_ _
Quadgram : _TEX, TEXT, EXT_, EX_ _, X_ _ _

Dapat disimpulkan bahwa untuk string berukuran n akan dimiliki n unigram dan n+1 bigram, n+1 trigram, n+1 quadgram dan seterusnya. Penggunaan N-gram untuk matching kata memiliki keuntungan sehingga dapat diterapkan pada recovery pada input karakter ASCII yang terkena noise, interpretasi kode pos, information retrieval dan berbagai aplikasi dalam pemrosesan bahasa alami.

Keuntungan N-gram dalam *matching* string adalah berdasarkan karakteristik N-gram sebagai bagian dari suatu string, sehingga kesalahan pada sebagian string hanya akan berakibat perbedaan pada sebagian N-gram. Sebagai contoh jika N-gram dari dua string dibandingkan, kemudian kita menghitung cacah N-gram yang sama dari dua string tersebut maka akan didapatkan nilai similaritas atau kemiripan dua string tersebut yang bersifat resistan terhadap kesalahan tekstual.

Kemiripan antara kata JOKO dengan JOKI (ada perbedaan 1 huruf), dapat diukur derajat kesamaan dengan cara menghitung berapa buah N-gram yang diambil dari dua kata tersebut yang bernilai sama, yaitu :

JOKO: _J, JO, OK,KO,O_ , JOKI : _J, JO, OK,KI, I_ kesamaan :3

Sementara antara kata JOKO dengan JONI (ada perbedaan 2 huruf), nilai kesamaan adalah :

JOKO: _J, JO, OK,KO,O_ , JONI : _J, JO, ON,NI,I_ kesamaan : 2

Sehingga dapat disimpulkan bahwa kemiripan atau kesamaan antara JOKO-JOKI dari pada antara JOKO-JONI.

Deteksi Bahasa dengan n-gram

Penggunaan n-*gram* untuk deteksi bahasa didasarkan pada anggapan bahwa pola sebaran n-*gram* dari suatu bahasa bersifat unik karena ini terkait dengan frekuensi penggunaan huruf, atau pasangan huruf baik itu vokal atau konsonan dari suatu bahasa yang umumnya berbeda dengan bahasa yang lain. Untuk unigram misalnya, yang jika dihitunmg frekuensinya adalah frekuensi keumunculan huruf dalam teks bahasa tertentu yang akan uniq untuk bahasa yang berbeda. Untuk teks bahasa Indonesia vokal a akan merupakan vokal yang frekuensi munculnya paling tinggi, sementara untuk bahasa inggris vokal e merupakan vokal yang frekuensinya paling tinggi. Demikian juga jika digunkan abi-gram dan tri-gram, keunikan pola n-gram dari suatu bahasa akan nampak lebih menonjol.

3. METODE PENELITIAN

Bahan Penelitian

Bahan Penelitian berupa koleksi dokumen teks yang terdiri dari:

a) Dokumen berbahasa Indonesia berupa koleksi dokumen berita seperti Tabel 1.

Tabel 1 Daftar koleksi dokumen teks berita

Nama Koleksi	Cacah dok	Cch Kata Unik	Rerata juml kata/ dok
Nws50.dok	50	2.860	354
Nws100.dok	100	4.385	368
Nws200.dok	200	6.634	372
Nws300.dok	300	8.471	373
Nws400.dok	400	10.152	388
Nws500.dok	500	11.636	385
Nws600.dok	600	13.432	388
Nws700.dok	700	14.800	385
Nws800.dok	800	15.751	410
Nws1009.dok	1009	18.259	425
Nws3000.dok	3000	35.282	397

b) Dokumen akademik terdiri: koleksi berbahasa Indonesia berupa abstrak seminar bidang teknik dan bidang teknologi informasi seperti Tabel 2, sedangkan koleksi berbahasa Inggris terdiri dari koleksi Abstract dan full paper berbahasa Inggris seperti Tabel 3

Tabel 2 Daftar dokumen akademik bahasa Indonesia

Nama Koleksi	Cacah dok	Cacah Kata	Rerata juml kata/dok
BahasaTes1.txt	88	17.257	196
BahasaTes2.txt	450	91.390	203

Tabel 3 Daftar dokumen teks akademik bahasa Inggris

Nama Koleksi	Cacah dok	Cacah Kata	Rerata juml kata/dok
EnglishCol1.txt	40	8.714	217
EnglishCol2.txt	40	348.638	8.939

Prosedur Penelitian

Perancangan Modul Language Detector

Modul untuk deteksi bahasa akan diimplementasikan menggunakan tiga pendekatan, yaitu deteksi bahasa dengan *unigram, bigram dan trigram.* Langkah-langkah deteksi adalah sebagai berikut :

Tahap pelatihan :

[1] Hitung frekuensi tiap *unigram* (a,b,c,...,z) dalam koleksi dokumen bahasa Indonesia yang dipilih sebagai *training set*.
[2] Tentukan probabilitas kemunculan untuk seluruh *unigram* :a,b,…,z
[3] Tentukan *profile* dokumen bahasa Indonesia dengan menetapkan nilai probabilitas tiap *unigram*.

Tahap pengujian dokumen baru :

1. Tentukan profil dari dokumen baru yang akan diuji bahasanya dengan menghitung nilai simQ sebagai berikut :

$$\text{simQ} = \sum_{i=a}^{z} \frac{(f_i - fR_i)^2}{fR_i} \qquad (1)$$

f_i : probabilitas unigram ke-i dari dokumen yang dideteksi
fR_i : probabilitas unigram ke-i dari dokumen *training set*
i : unigram ke i, yaitu a,b,c,…,z

2. Gunakan hasil profil simQ untuk mengevaluasi jenis dokumen. Jika *training set* dokumen berbahasa Indonesia dan dokumen yang diuji juga dokumen bahasa Indonesia maka nilai simQ akan cenderung kecil, tetapi jika dokumen yang diuji bukan bahasa Indonesia akan cenderung besar.
3. Tetapkan D dokumen berbahasa Indonesia jika : simQ(D)< T, dengan T adalah suatu nilai *Treshold* yang ditetapkan melalui eksperimen

Pada pengujian dengan pendekatan *bigram* dan *trigram* langkahnya adalah sebagai berikut :

Pelatihan :

[1] Tentukan statistik *bigram*/*trigram* pada dokumen *training set* bahasa Indonesia
[2] Hitung bobot untuk setiap *bigram*/*trigram* yang didapat dari pelatihan dengan rumus:

$$w_{i,d} = \begin{cases} \frac{freq_i}{N} & \text{jika trigram i ada dalam } \textit{training set} \\ -\frac{N_{i,d}}{\sum_j |word_{j,d}|} & \text{wm selainnya} \end{cases} \quad (2)$$

$w_{i,d}$ = bobot *bigram*/ *trigram* i pada dokumen d
$freq_i$ = frekuensi kemunulan *bigram*/*trigram* i dalam *training set*
N = banyaknya kata dalam *training set*
$N_{i,d}$ = banyaknya kata dalam dokumen d yang mengandung *bigram*/ *trigram* i
$word_{j,d}$ = kata ke-j dalam dokumen d yang mengandung *bigram*/*trigram* ke i
$|word_{j,d}|$ = panjang dari kata $word_{j,d}$
wm = adalah faktor modifikasi bobot

Pengujian dokumen baru *:*
Untuk menetapakan apakah dokumen baru **d** berbahasa Indonesia atau tidak, ditempuh langkah-langkah :
[1] Tentukan statistik *bigram*/*trigram* pada dokumen baru **d**
[2] hitung fungsi statistik h(**d**) dengan rumus :

$$h(\mathbf{d}) = \sum_i \left(\frac{f_{i,d}}{N_d} x w_{i,d} \right) \quad (3)$$

$f_{i,d}$ = frekuensi *bigram*/*trigram* i dalam dokumen **d**
N_d = Banyaknya kata dalam dokumen d
$w_{i,d}$ = dari persamaan (1)
[3] Tetapkan bahwa dokumen **d** adalah dokumen berbahasa Indonesia jika statisik memenuhi kriteria :
$h(\mathbf{d}) > q$
dengan nilai q adalah treshold yang bersifat *language dependent*.

4. HASIL DAN PEMBAHASAN

Deteksi Bahasa dengan *Unigram*

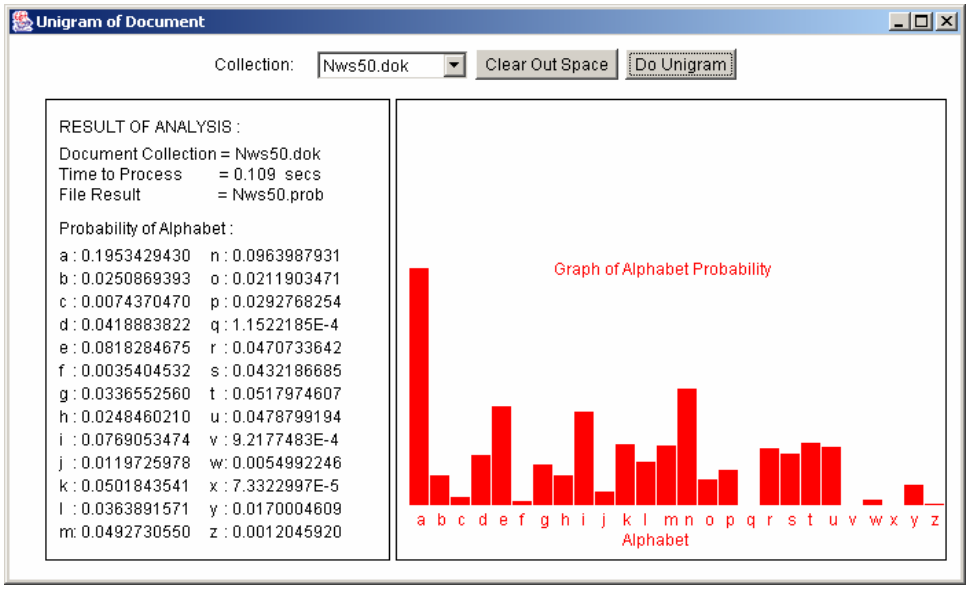
Unigram diperoleh dengan mencari frekuensi kemunculan huruf dalam suatu dokumen. Selanjutnya dari frekuensi kemunculan abjad tersebut dapat ditentukan probabilitas kemunculan abjad pada dokumen berbahasa Indonesia. Gambar 1 berikut menunjukkan contoh hasil analisis *unigram* untuk koleksi dokumen berita Nws50.dok

Dari gambar terlihat bahwa abjad 'a' menduduki frekuensi tertinggi, disusul abjad 'n', 'i' dan 'e'. Pada sisi lain abjad 'q', 'v', 'x' dan 'z' merupakan abjad-abjad yang paling sedikit muncul dalam koleksi. Pola ini ternyata bersifat tetap seperti dapat dibuktikan pada koleksi-koleksi dokumen yang lebih besar jumlah dokumennya. Gambar 2 menunjukkan profile frekuensi atau probabilitas kemunculan abjad pada koleksi dokumen berita Nws100.dok, Nws500.dok, Nws1009.dok dan Nws3000.dok yang menunjukkan konsistensi probabilitas tersebut.

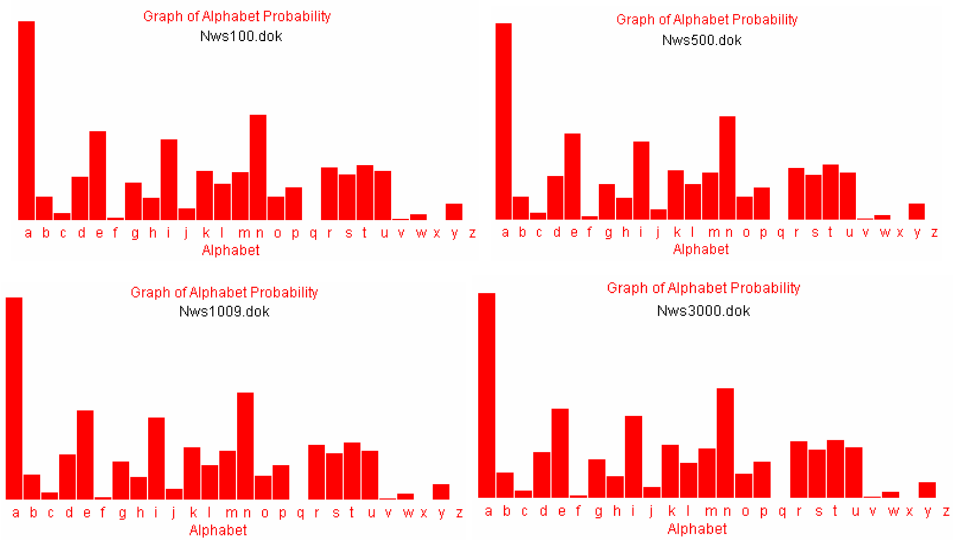
Probabilitas unigram sebagai cara deteksi bahasa

Profile kemunculan abjad pada bahasa yang berbeda ternyata memiliki profil probabilitas yang berbeda. Sebagai contoh jika koleksi dokumen berbahasa Inggris dianalisis ternyata akan memiliki profil yang berbeda. Gambar 3 menunjukkan profil probabilitas kemunculan abjad pada koleksi dokumen berbahasa Inggris EnglisCol1.txt dan EnglishCol2.txt yang merupakan koleksi 40 *abtract* ilmiah berbahasa Inggris dan 40 *full paper* berbahasa inggris. Terlihat dari gambar 3 bahwa frekuensi tertinggi adalah pada abjad 'e' disusul 't', 'i' dan 'a'.
Dengan menggunakan rumusan pada persamaan 1 diperoleh bahwa jika suatu dokumen berbahasa Indonesia dianalisis *unigram*-nya, maka akan memiliki profil yang mendekati profil pada Gambar 2 sehingga nilai similaritas probabilitasnya akan cenderung kecil. Dengan demikian jumlah kuadrat selisih similaritas probabilitas juga akan cenderung kecil, seperti dapat dilihat pada Gambar 4. Sebaliknya dengan profil bahasa Inggris nilai similaritas akna cenderung besar. Dengan demikian jika dapat ditetapkan suatu nilai treshld tertentu

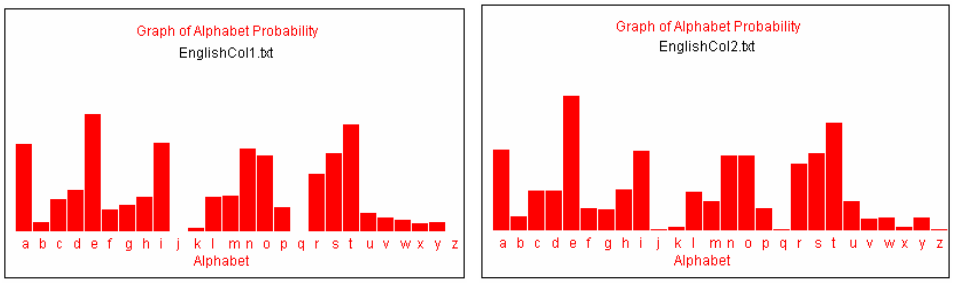
(misalnya ntuk unigram 0.5) maka dapat dipisahkan dokumen yang similaritasnya dibawah treshold adalah berbahasa Indonesia dan jika tidak berarti berbahsa Inggris.



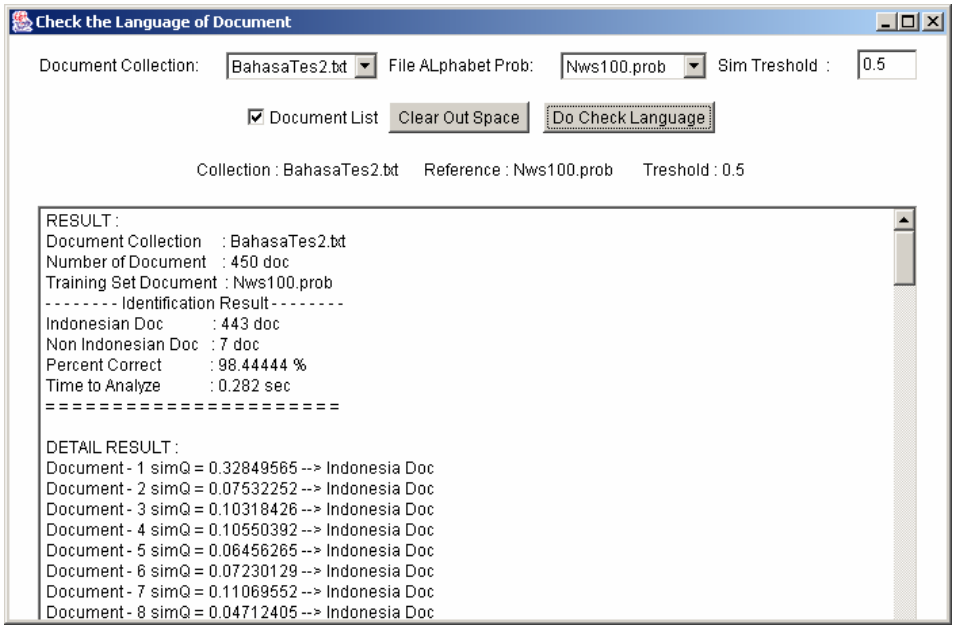
Gambar 1 Profile unigram koleksi dokumen berita bahasa Indonesia Nws50.dok



Gambar 2 Profile probabilitas abjad pada 4 koleksi



Gambar 3 Profile probabilitas abjad koleksi dokumen bahasa Inggris

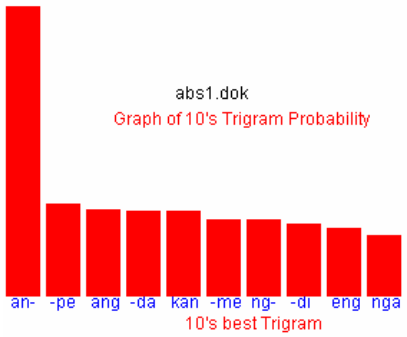


Gambar 4 Hasil identifikasi dokumen bahasa Indonesia

Probabilitas Bigram dan Trigram sebagai cara deteksi bahasa

Salah satu pola sebaran trigram dokumen bahasa Indonesia terlihat pada Gambar 5. Pola sebaran trigram untuk koleksi dokumen bahasa Inggris jika dilihat dari 10 trigram terbaik adalah "-th", "the", "he-","-in", "ion", "ing", "on-", "ng-", "of" dan "tio". Jika dibandingkan bigram atau trigram dari koleksi bahasa Inggris dan bigram atau trigram dari koleksi bahasa Indonesia, akan terlihat perbedaan kontras sebagaimana pada tabel 4 dan tabel 5.

Dengan memanfaatkan pola sebaran bigram atau trigram deteksi bahasa akan dapat dilakukan untuk membedakan apakah suatu dokumen berbahasa Indonesia atau berbahasa Inggris.



Gambar 5. Sebaran trigram dokumen bahasa Indonesia

Tabel 4. Perbandingan 10 *bigram* terbaik

Berita (Ind)	English
an, a-, n-, ng, i-, -d, er, **-m**, en, **ka**	**e-**, **s-**, **-t**, **in**, **th**, n-, **-a**, er, **he**, **-i**

Tabel 5. Perbandingan 10 *trigram* terbaik

Berita (Indonesia)	English
an-, -me,ang,kan,-pe,men,ng-,-di,-se,-ke	-th,the,he-,-in,ion,ing,on-,ng-,of,tio

Baik dengan bigram maupun trigram terlihat bawa pola n-gram akan sangat berbeda jika bahasa suatu dokumen berbeda. Dalam bigram hanya da dua bigram yang sama (er dan n-), sedangkan pada trigram tidak ada yang sama.

Tabel 6 Waktu deteksi dan akurasi unigram, bigram dan trigram

dok	Unigram		Bigram		Trigram	
	Acc	Time (s)	Acc	Time (s)	Acc	Time (s)
300	%99,7	0,360	%100	1,129	%100	6,437
400	%99,7	0,422	%100	1,891	%100	8,157
500	%99,8	0,453	%100	2,094	%100	9,063
1000	%99,9	0,890	%100	4,047	%100	18,61
3000	%99,3	2,340	%100	10,98	%100	42,183

Keterangan : Acc : Akurasi

Terlihat dari perbandingan *unigram*, *bigram* dan *trigram* dalam melakukan deteksi, yaitu akurasi akan semakin tinggi jika n semakin tinggi, akan tetapi waktu deteksi juga akan semakin lama. *Unigram* akan cenderung cepat dalam waktu deteksi dan akurasi cenderung lebih rendah dari *trigram*. Di satu sisi *bigram* dan *trigram* dengan akurasi 100% cenderung memiliki waktu deteksi yang lebih lama dari unigram. Perbandingan menunjukkan bahwa dengan bigram ternyata dapat diambil nilai tengah dalam pengertian akurasi tinggi (100%) tetapi waktu deteksi tidak terlalu lama. Dalam deteksi juga diperoleh hasil makin tinggi n makin kecil nilai *treshold*.

5. KESIMPULAN

Beberapa kesimpulan yang dapat ditarik dari penelitian ini adalah sebagai berikut :

1. Penggunaaan n-*gram* yang terdiri dari *unigram*, *bigram* dan *trigram* untuk melakukan identifikasi bahasa Indonesia terhadapa bahasa Inggris dari suatu dokumen berhasil dengan baik.
2. Penyusunan unigram memerlukan waktu paling cepat dibandingkan dengan penyusunan *bigram* atau *trigram*
3. Kemampuan identifikasi *trigram* dan *bigram* lebih baik dari *unigram*
4. Nilai *treshold* yang sebaiknya digunakan untuk unigram adalah 0,5 , untuk *bigram* 0,4 dan untuk *trigram* adalah 0,05.
5. Untuk hasil yang cukup akurat dengan waktu identifikasi tidak terlalu lama sebaiknya digunakan *bigram* dalam identifikasi bahasa.

6. DAFTAR PUSTAKA

Adriani, M, 2002, *Evaluating Indonesian Online Resources for Cross Language Information Retrieval*, SIGIR'2002, International Conference on Research and Development in Information Retrieval, Agustus 2002.

Asian, J., H. E. Williams, and S. M. M. Tahaghoghi, 2004, *Tesbed for Indonesian Text Retrieval*, 9th Australian Document Computing Symposiom, Melbourne December, 13 2004.

Bastrup, S. and C. Popper, 2003, *Language Detection Based on Unigram Analysis and Decision Trees*, www.citeseer.ist.psu.edu/bastrup03language.html

Batchelder, E.O., 1992, A Learnbing Experience: Training an Artificial Neural Network to Discriminate Languages. Unpublished Technical Report, 1992.

Hamzah, A., 2009, *Penerapan Clustering Dokumen untuk Meningkatkan Efektifitas Sistem Temu Kembali Informasi Dokumen Berbahasa Indonesia*, Disertasi Jurusan Teknik Elektro, Fakultas Teknik, Universitas Gadjah Mada, Yogyakarta.

Nazief, B., 2000, *Development of Computational Linguistic Research: a Challenge for Indonesia*", Computer Science Center, University of Indonesia.

Sibun, P. And Spits, A.L., 1994, *Language Determination: Natural Language Processing from Scanned Document Image*, Fuji Xerox, Palo Alto Laboratory.

Sibun, P. and J.C. Reynar, 1996, *Language Identification: Examining the Issues*, The 5th Symposium on Document Analysis and Information Retrieval, Las Vegas , Nevada ,U.S.A., pages: 125-135.

Vega, V.B., and S. Bressan, 2000, *Continuous-Learning Weighted-Trigram Approach for Indonesian Language Distinction: A Preliminary Study*, School of Computing, natinal University of Singapore.

Vega, V. B. , 2001, *Information Retrieval for the Indonesian Language*, Master's thesis, National University of Singapore.

Xu, W., X. Liu, and Y. Gong, 2003, *Document Clustering Based on Non-Negative Matrix Factorization*, SIGIR'03, 28 Juli-1 Agustus, Toronto, Canada.

Ziegler, D.V., 1992, The Automatic Identification of Languages Using Linguistic Recognition Signial, Dissertation, State University of New York at Baffalo.