

APLIKASI SEGMENTASI TEKS DALAM BAHASA MANDARIN DENGAN METODE *RULE-BASED* DAN *STATISTICAL*

Rudy Adipranata¹⁾, Meliana Ongkowinoto²⁾, Rolly Intan³⁾

Jurusan Teknik Informatika, Fakultas Teknologi Industri, Universitas Kristen Petra, Surabaya
rudy@petra.ac.id¹⁾ rintan@petra.ac.id³⁾

ABSTRACT

Chinese Language is now widely used. Many applications have been developed to help the usage of Chinese Language in Information Technology (IT). One important phase in these applications are segmentation phase. In a Chinese Language sentence, there is no spacing between the words. In the segmentation phase, a Chinese Language sentence is segmented to become words in Chinese Language. In this research, we create an application which can segment the Language. The methods used for segmentation process are rule-based method, statistical method and the hybrid of them. The segmentable words are only in the form of unigram and bigram. The experiment result shows that the segmentation which used the hybrid method has a better result than the single method.

Keywords: *Chinese Word Segmentation, Rule Based Method, Statistical Method*

1. Pendahuluan

Selain bahasa Inggris, bahasa Mandarin adalah salah satu bahasa yang banyak digunakan di dunia. Sekitar seperlima dari penduduk dunia menggunakan bahasa Mandarin sebagai bahasa sehari-hari. Bahasa Mandarin adalah bahasa resmi di Tiongkok dan Taiwan. Di Singapura, bahasa Mandarin merupakan satu dari empat bahasa resmi. Selain itu, bahasa Mandarin merupakan salah satu dari enam bahasa resmi Perserikatan Bangsa Bangsa^[2]. Di Indonesia, bahasa Mandarin juga semakin berkembang dimulai sejak masa reformasi dimana budaya Tiongkok diperbolehkan untuk kembali berkembang. Sejak saat itu, bahasa Mandarin semakin dipakai secara luas dan muncul di berbagai lembaga untuk belajar bahasa Mandarin. Selain itu, beberapa bahasa Mandarin juga disisipkan sebagai salah satu mata pelajaran bahasa di samping bahasa Indonesia dan bahasa Inggris di sekolah.

Dengan semakin berkembangnya bahasa Mandarin, dalam dunia Informatika juga dibuat aplikasi-aplikasi yang berhubungan dengan bahasa tersebut. Antara lain adalah aplikasi untuk menerjemahkan, aplikasi *text to speech* dalam bahasa Mandarin dan *search engine* yang dapat digunakan untuk bahasa Mandarin. Dalam pembuatan aplikasi-aplikasi tersebut, terdapat satu tahap yang tidak dapat dilewatkan, yaitu tahap segmentasi. Yang dimaksud dengan segmentasi adalah membagi-bagi suatu kalimat menjadi kata-kata. Pada bahasa Mandarin tidak ada pemisah antara satu kata dengan kata lain seperti yang terdapat pada bahasa-bahasa Latin. Tidak adanya pemisah antar kata, menyebabkan berbagai macam kesulitan, antara lain adalah untuk aplikasi penerjemah, akan sulit menentukan arti dari suatu kalimat tanpa memisahkan dulu menjadi kata-kata. Pada aplikasi *text to speech*, untuk menentukan intonasi dari pembacaan suatu kalimat juga diperlukan pemisahan kata-kata yang ada dalam suatu kalimat. Oleh karena itu, segmentasi merupakan tahap yang penting dalam pembuatan aplikasi-aplikasi bahasa Mandarin.

2. Teori Penunjang

2.1 Struktur Bahasa Mandarin

Pada bahasa Mandarin, suatu kata dapat dibentuk dari satu atau lebih karakter. Kata yang dibentuk dari satu karakter disebut *unigram*, kata yang dibentuk dari dua disebut *bigram* dan seterusnya sampai *n-gram*. Sebagian besar dari kata-kata dalam bahasa Mandarin dibentuk dari satu dan dua karakter. Contoh pembagian kata yang dibentuk dari karakter-karakter Mandarin dengan C merupakan perwakilan dari satu karakter Mandarin dapat dilihat pada Tabel 1.

Tabel 1. Kata dalam Bahasa Mandarin

| Kalimat | $C_1C_2C_3C_4C_5C_6$ |
|----------------|--|
| <i>Unigram</i> | $C_1, C_2, C_3, C_4, C_5, C_6$ |
| <i>Bigram</i> | $C_1C_2, C_2C_3, C_3C_4, C_4C_5, C_5C_6$ |
| <i>Trigram</i> | $C_1C_2C_3, C_2C_3C_4, C_3C_4C_5, C_4C_5C_6$ |

2.2 Metode Segmentasi

Terdapat 2 macam metode untuk melakukan segmentasi dari kalimat berbahasa Mandarin, yaitu metode *rule-based* dan metode *statistical*.

2.1.1 Metode Rule-Based

Metode *rule-based* merupakan metode segmentasi kalimat dalam bahasa Mandarin dengan menggunakan aturan-aturan kata yang ada dalam kamus.

Contoh : 学习汉语 (xue xi han yu)

Kemungkinan-kemungkinan segmentasi yang dapat dihasilkan dari kalimat tersebut adalah sebagai berikut.

- 学习 | 汉语

- 学习 | 汉 | 语
- 学 | 习 汉 | 语
- 学 | 习 | 汉 语
- 学 | 习 | 汉 | 语

Dari kemungkinan-kemungkinan segmentasi di atas, dipilih kemungkinan pertama, karena kemungkinan pertama yang memiliki arti di kamus, yaitu belajar bahasa Mandarin. Kelemahan dari metode *rule-based* adalah keberhasilan untuk menghasilkan segmentasi yang benar dipengaruhi oleh banyaknya jumlah kata yang tersimpan di *database* (kamus). *Database* tidak mungkin menyimpan semua kata yang ada karena jika semua kata disimpan, jumlah kata yang ada sangatlah banyak. Selain itu dalam kehidupan sehari-hari, bahasa dapat berkembang dan tidak menutup kemungkinan terbentuk kata baru. Oleh karena itu hampir tidak mungkin semua kata yang ada dapat tersimpan dalam *database*.

2.2.1 Metode Statistical

Metode *statistical*^[1] merupakan metode segmentasi kalimat dalam bahasa Mandarin berdasarkan informasi statistik dari suatu bacaan dalam bahasa Mandarin. Pada metode *statistical* ini, terdapat dua tahap, yaitu tahap *training* dan tahap segmentasi. Pada tahap *training* yang dilakukan adalah menghitung probabilitas dari setiap karakter yang terdapat dalam data yang akan di-*training*. Untuk mendapatkan nilai probabilitas dari suatu karakter *C* dapat digunakan Rumus 1.

$$P(C) = \frac{\text{Frekuensi } C \text{ sebagai kata dalam training set}}{\text{Frekuensi } C \text{ dalam training set}} \quad (1)$$

Pada tahap *training* ini, frekuensi *C* sebagai kata dalam *training set* pada Rumus 1 didapat dari perhitungan jumlah munculnya *C* sebagai kata dari hasil segmentasi secara manual data yang sedang di-*training*. Pada tahap segmentasi, hasil segmentasi dari suatu *input* kalimat tersusun dari rangkaian kata C_i yang berpotensi menjadi kata, sehingga $\prod_i P(C_i)$

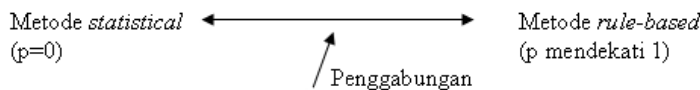
memiliki nilai terbesar. Kelemahan dari metode ini adalah keberhasilan dari segmentasi tergantung pada banyaknya jumlah data yang di-*training*.

2.2.2 Penggabungan Metode Rule-Based dan Metode Statistical

Pada dua metode sebelumnya, yaitu metode *rule-based* dan metode *statistical*, masing-masing memiliki beberapa kelemahan. Untuk mengatasi kelemahan-kelemahan tersebut, dilakukan penggabungan kedua metode dengan tujuan kedua metode tersebut dapat saling melengkapi^[1]. Penggabungan kedua metode ini seperti yang dilakukan manusia secara umumnya dalam melakukan proses segmentasi saat membaca suatu artikel. Pada waktu seseorang melakukan proses segmentasi suatu teks dalam bahasa Mandarin, orang tersebut akan melihat konteks teks tersebut dan menggunakan pengetahuannya tentang kata-kata yang ada. Selain itu, orang tersebut juga akan melihat frekuensi terbentuknya kata dalam bacaan tersebut. Dalam penggabungan kedua metode ini, terdapat beberapa prinsip sebagai berikut.

- Kata-kata yang tersimpan dalam *database* merupakan *background knowledge* dan informasi statistik merupakan *foreground knowledge*.
- Jika pada suatu kata terdapat informasi statistiknya, informasi statistik lebih diutamakan. Jika kata tersebut tidak terdapat informasi statistiknya, digunakan *default probability* yaitu suatu nilai probabilitas tertentu.

Penggabungan kedua metode ini sangatlah fleksibel. Jika nilai *default probability* bernilai 0, penggabungan metode ini tidak menghiraukan kata-kata yang tersimpan dalam *database* dan penggabungan metode ini cenderung menjadi metode *statistical*. Sebaliknya, jika nilai *default probability* bernilai besar yaitu mendekati 1, penggabungan metode ini menggunakan kata-kata yang tersimpan dalam *database* sebagai acuan utama dan penggabungan metode ini cenderung menjadi metode *rule-based*. Oleh karena itu, penggabungan kedua metode ini terdapat di antara metode *rule-based* dan metode *statistical* seperti yang terlihat pada Gambar 1.



Gambar 1. Perbandingan Tiga Buah Metode

Algoritma dari penggabungan kedua metode adalah sebagai berikut.

1. Menghitung probabilitas dari semua karakter yang ada.
2. Mencari kandidat-kandidat kata berdasarkan metode *rule-based*.
3. Setiap karakter pada kalimat yang di-*input*-kan berhubungan dengan semua kandidat kata dimulai dari karakter itu sendiri berikut dengan probabilitasnya.
4. Mengkombinasikan kandidat-kandidat kata. Kombinasi kata dengan probabilitas terbesar dipilih sebagai hasil.

2.3 Precision dan Recall

Untuk melakukan pengujian terhadap hasil segmentasi, digunakan pengukur standar *precision dan recall* dari *information retrieval*. *Precision* adalah perbandingan antara jumlah kata benar yang dihasilkan oleh sistem dan jumlah total kata yang dihasilkan oleh sistem. *Recall* adalah perbandingan antara jumlah kata benar yang dihasilkan oleh sistem dan jumlah kata

benar yang dihasilkan dari segmentasi secara manual, dengan asumsi bahwa segmentasi secara manual merupakan hasil segmentasi yang benar. Rumus dari *precision* dan *recall* didefinisikan pada Rumus 2 dan Rumus 3^[3].

$$precision = \frac{|A \cap B|}{|B|} \tag{2}$$

$$recall = \frac{|A \cap B|}{|A|} \tag{3}$$

dimana *A* merupakan kata yang benar dari segmentasi secara manual dan *B* merupakan kata yang dihasilkan oleh sistem.

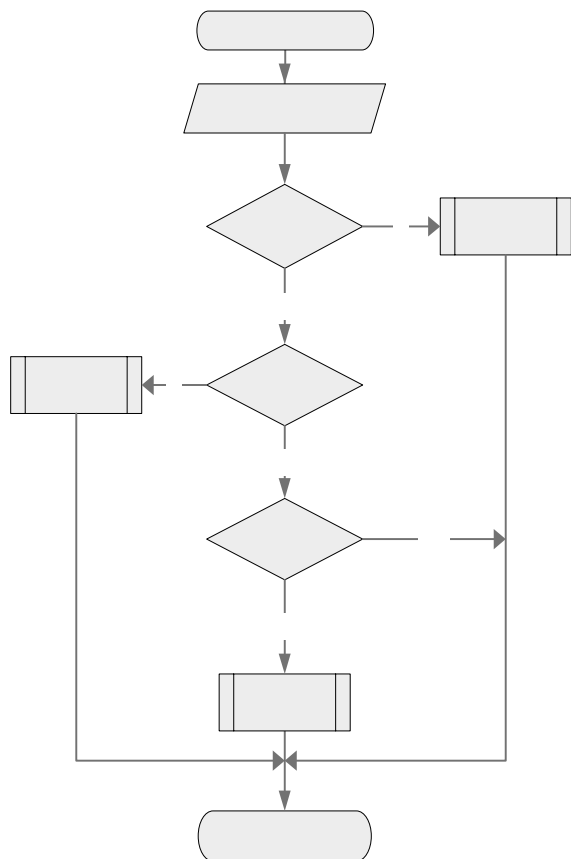
3. Perancangan Sistem

Secara garis besar, rencana kerja dari aplikasi ini ditunjukkan pada Gambar 2.

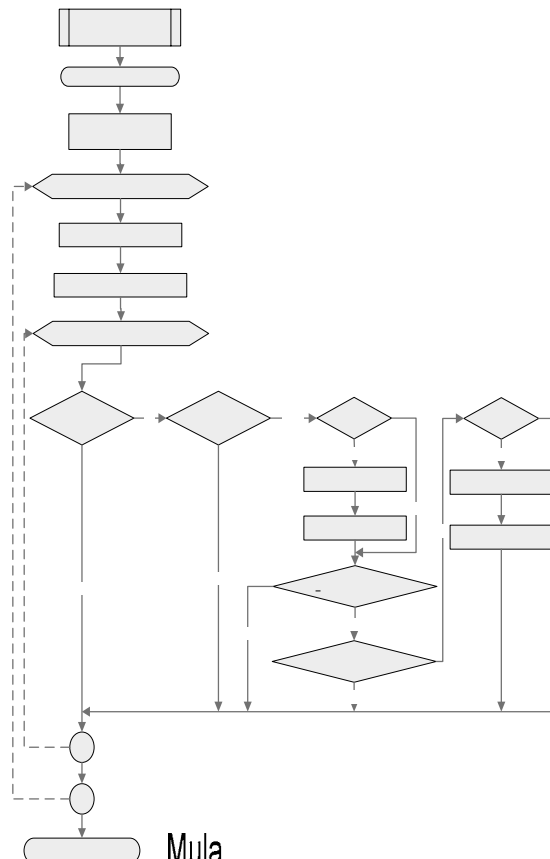
Pada awal dari aplikasi, pengguna diberi tiga menu pilihan, yaitu menu kata, menu *rule* kata dan menu segmentasi. Menu kata digunakan untuk memasukkan kata-kata dalam bahasa Mandarin. Menu *rule* kata digunakan untuk memasukkan *rule-rule* kata yang nantinya dipakai dalam segmentasi yang menggunakan metode *rule-based*. Menu segmentasi merupakan menu utama dari aplikasi ini. Menu segmentasi digunakan untuk melakukan segmentasi dari suatu *input* tertentu dengan menggunakan tiga macam pilihan metode, yaitu metode *rule-based*, metode *statistical* dan penggabungan antara metode *rule-based* dan metode *statistical*.

3.1 Segmentasi dengan Metode Rule-Based

Segmentasi menggunakan metode *rule-based* merupakan metode segmentasi berdasarkan kata-kata serta *rule* yang telah di-*input*-kan. Adapun diagram alir metode *rule-based* ini dapat dilihat pada Gambar 3.



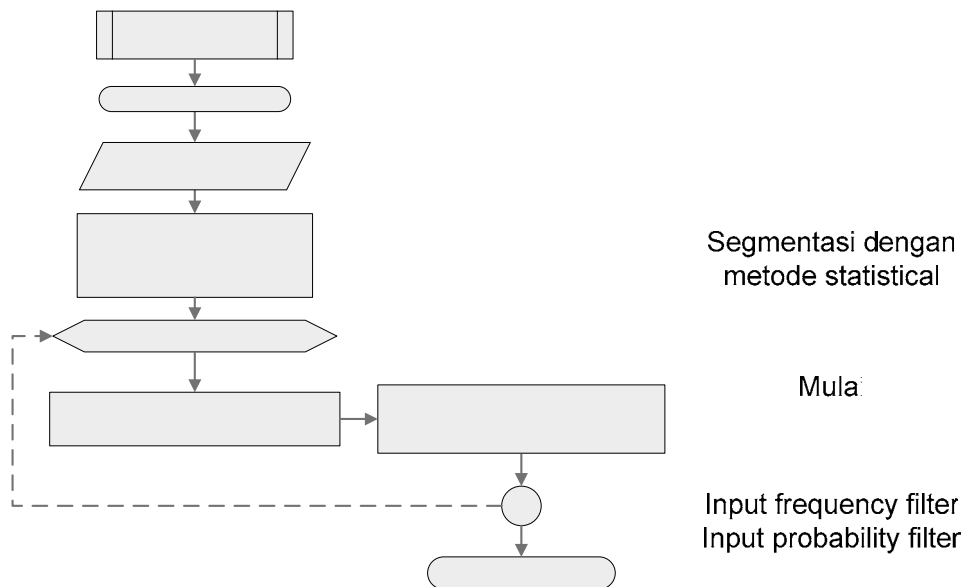
Gambar 2. Diagram Alir Aplikasi



Gambar 3. Diagram Alir Metode Rule-Based

3.2 Segmentasi dengan Metode Statistical

Segmentasi menggunakan metode *statistical* merupakan metode segmentasi yang menggunakan hasil perhitungan probabilitas dari munculnya suatu karakter menjadi kata *unigram* dan kata *bigram*. Adapun cara kerja metode ini dapat dilihat pada Gambar 4.



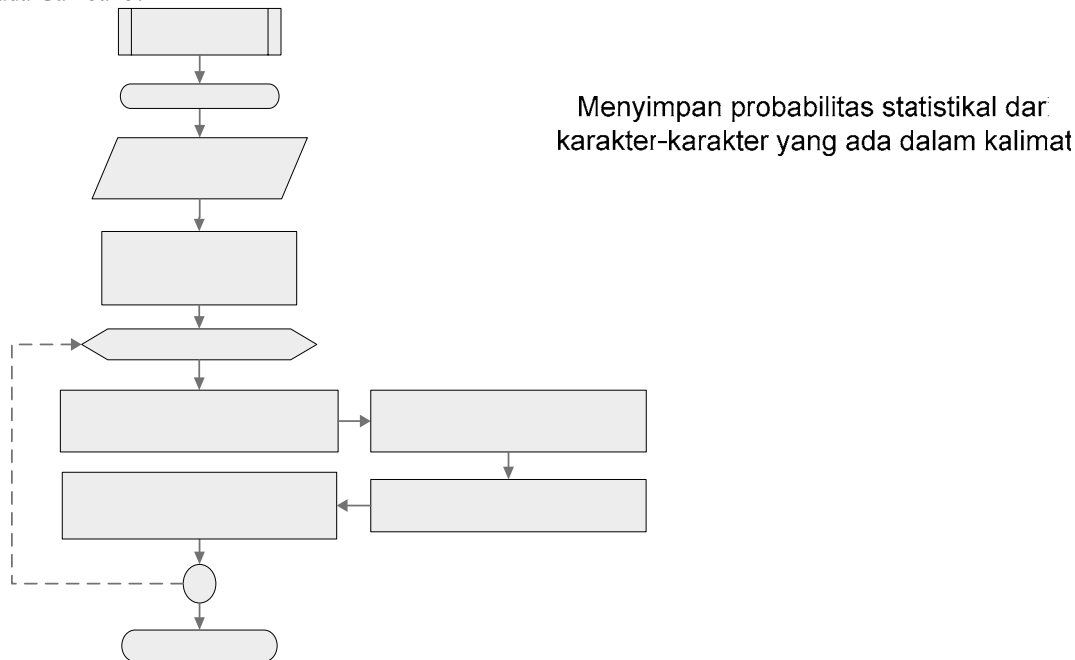
Gambar 4. Diagram Alir Metode *Statistical*

Pada saat metode ini dijalankan, pengguna diminta untuk memasukkan *frequency filter* dan *probability filter*. Proses selanjutnya adalah membagi artikel menjadi kalimat-kalimat dan sistem akan menghitung frekuensi serta probabilitas untuk setiap kemungkinan kata *unigram* dan kata *bigram*. Kemudian sistem akan melakukan proses segmentasi untuk setiap kalimat yang ada. Proses segmentasi dilakukan dengan membandingkan kemungkinan probabilitas dari karakter-karakter yang ada dalam kalimat tersebut.

Proses Kalimat
Hitung Unigram
Hitung Bigram
Proses Hitung Prob

3.3 Segmentasi dengan Penggabungan antara Metode *Rule-Based* dan Metode *Statistical*

Segmentasi menggunakan penggabungan antara metode *rule-based* dan metode *statistical*. Adapun cara kerja dari proses segmentasi ini dapat dilihat pada Gambar 5.



Gambar 5. Diagram Alir Penggabungan Metode *Rule-Based* dan Metode *Statistical*

Pengguna diminta untuk memasukkan *default probability*, *frequency filter* dan *probability filter*. Proses selanjutnya adalah membagi artikel menjadi kalimat-kalimat dan sistem akan menghitung frekuensi dan probabilitas untuk setiap kemungkinan kata *unigram* dan kata *bigram*. Kemudian sistem akan melakukan proses segmentasi untuk setiap kalimat yang ada dengan membandingkan kemungkinan probabilitas terbesar dari kata-kata dalam kalimat yang terdapat dalam *database*. Setelah itu dilakukan proses perbandingan kembali untuk kata-kata yang tidak ada dalam *database*.

4. Implementasi dan Hasil Pengujian

Hasil aplikasi dapat dilihat pada Gambar 6.



Gambar 6. Tampilan *Form* Utama

Jika pada menu utama dipilih menu *Word*, akan tampil *form* seperti pada Gambar 7 yang digunakan untuk memasukkan kata-kata ke *database* baik kata *unigram* ataupun kata *bigram* dalam bahasa Mandarin beserta dengan tipe kata tersebut, misal kata benda, kata kerja, kata sifat dan lain-lain.



Gambar 7. Tampilan *Form* *Word*

Pengujian dilakukan dengan menggunakan ketiga metode dan kemudian dilakukan perhitungan nilai *precision* dan *recall* dari masing-masing metode. Hasil pengujian untuk ketiga metode terdapat pada Tabel 2 hingga Tabel 4.

Tabel 2. *Precision* dan *Recall* Hasil Pengujian Metode *Rule-Based*

| <i>Precision</i> | <i>Recall</i> |
|------------------|---------------|
| 71.87% | 79.14% |

Tabel 3. *Precision* dan *Recall* Hasil Pengujian Metode *Statistical*

| <i>Frequency Filter</i> | <i>Probability Filter</i> | <i>Precision</i> | <i>Recall</i> |
|-------------------------|---------------------------|------------------|---------------|
| 0 | 0 | 56.14% | 49.08% |
| 0 | 0.5 | 56.14% | 49.08% |
| 0 | 1 | 55.79% | 48.77% |
| 0.5 | 0 | 56.14% | 49.08% |
| 0.5 | 0.5 | 56.14% | 49.08% |
| 0.5 | 1 | 56.14% | 49.08% |
| 1 | 0 | 56.14% | 49.08% |
| 1 | 0.5 | 56.14% | 49.08% |
| 1 | 1 | 56.14% | 49.08% |

Tabel 4. Precision dan Recall Hasil Pengujian Metode Rule-Based dan Statistical

| Frequency Filter | Probability Filter | Default Probability | Precision | Recall |
|------------------|--------------------|---------------------|-----------|--------|
| 0 | 0 | 0 | 80.90% | 88.34% |
| 0 | 0 | 0.5 | 69.87% | 82.52% |
| 0 | 0 | 1 | 63.34% | 77.91% |
| 0 | 0.5 | 0 | 80.90% | 88.34% |
| 0 | 0.5 | 0.5 | 69.87% | 82.52% |
| 0 | 0.5 | 1 | 63.34% | 77.91% |
| 0 | 1 | 0 | 80.90% | 88.34% |
| 0 | 1 | 0.5 | 69.87% | 82.52% |
| 0 | 1 | 1 | 63.34% | 77.91% |
| 0.5 | 0 | 0 | 81.41% | 88.65% |
| 0.5 | 0 | 0.5 | 81.41% | 88.65% |
| 0.5 | 0 | 1 | 81.41% | 88.65% |
| 0.5 | 0.5 | 0 | 81.41% | 88.65% |
| 0.5 | 0.5 | 0.5 | 81.41% | 88.65% |
| 0.5 | 0.5 | 1 | 81.41% | 88.65% |
| 0.5 | 1 | 0 | 81.41% | 88.65% |
| 0.5 | 1 | 0.5 | 81.41% | 88.65% |
| 0.5 | 1 | 1 | 81.41% | 88.65% |
| 1 | 0 | 0 | 81.41% | 88.65% |
| 1 | 0 | 0.5 | 81.41% | 88.65% |
| 1 | 0 | 1 | 81.41% | 88.65% |
| 1 | 0.5 | 0 | 81.41% | 88.65% |
| 1 | 0.5 | 0.5 | 81.41% | 88.65% |
| 1 | 0.5 | 1 | 81.41% | 88.65% |
| 1 | 1 | 0 | 81.41% | 88.65% |
| 1 | 1 | 0.5 | 81.41% | 88.65% |
| 1 | 1 | 1 | 81.41% | 88.65% |

Dari Tabel 2 hingga 4, terlihat bahwa nilai rata-rata *precision* dan *recall* dari hasil segmentasi dengan metode gabungan (*precision* 78.06%, *recall* 86.74%) lebih besar daripada hasil segmentasi dengan menggunakan metode *rule-based* (*precision* 71.87%, *recall* 79.14%) ataupun segmentasi dengan menggunakan metode *statistical* (*precision* 56.10%, *recall* 49.05%).

Selain pengujian dengan menggunakan ketiga jenis metode, dilakukan pula pengujian dengan menggunakan artikel lain yang mempunyai kemungkinan kata belum terdapat dalam *database*. Hasil dari pengujian ini dengan metode gabungan adalah rata-rata *precision* 45.14% dan rata-rata *recall* adalah 52.62%.

5. Kesimpulan

Dari hasil pengujian dapat ditarik kesimpulan bahwa hasil segmentasi dengan menggunakan penggabungan metode *rule-based* dan metode *statistical* mempunyai nilai *precision* ataupun *recall* yang lebih tinggi dibanding dengan menggunakan satu metode saja baik *rule-based* ataupun *statistical*. Segmentasi dengan menggunakan penggabungan kedua metode tetap bergantung pada kata-kata yang tersimpan dalam *database*. Hal ini dapat dilihat dari nilai *precision* dan *recall* yang lebih kecil saat menggunakan artikel lain yang mempunyai kemungkinan kata tidak terdapat pada *database*.

Daftar Pustaka

- [1] Nie, Jian Yun, Brisebois, M & Ren, Xiaobo. (1996). *On Chinese text retrieval*. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 225–233.
- [2] Chinese language. http://en.wikipedia.org/wiki/Chinese_language, diakses terakhir tanggal 7 November 2007.
- [3] Van Rijsbergen, C.V. (1979). *Information Retrieval*. London: Butterworth, <http://www.dcs.gla.ac.uk/Keith/Preface.html>, diakses terakhir tanggal 19 Mei 2008.