# Peringkasan Dokumen Berbahasa Inggris Menggunakan Sebaran *Local Sentence*

**Aminul Wahib[1], Agus Zainal Arifin[2], Diana Purwitasari [3]**
Jurusan Teknik Informatika, Institut Teknologi Sepuluh Nopember
Kampus ITS Keputih, Sukolilo, Surabaya 60111, Jawa Timur
*E-mail:* [1]wahib13@mhs.if.its.ac.id

Masuk: 11 Juli 2015; Direvisi: 31 Juli 2015; Diterima: 31 Juli 2015

***Abstract.*** *The number of digital documents grows very rapidly causing time waste in searching and reading the information. To overcome these problems, many document summary methods are developed to find important or key sentences from the source document. This study proposes a new strategy in summarizing English document by using local sentence distribution method to find and dig up hidden important sentence from the source document in an effort to improve quality of the summaries. Experiments are conducted on dataset DUC 2004 task 2. Measurement ROUGE-1 and ROUGE-2 are employed as a performance evaluation of the proposed method with sentence information density and sentence cluster keyword (SIDeKiCK). The experiment shows that the proposed method has better performance with an average achievement ROUGE-1 0.398, an increase of 1.5% compared to SIDeKiCK method and ROUGE-2 0.12, an increase 13% compared to SIDeKiCK method.*
***Keywords:*** *Summarize Document, Important Sentences, Distribution of Local Sentence, ROUGE.*

***Abstrak.*** *Jumlah dokumen digital yang berkembang sangat pesat menyebabkan banyaknya waktu terbuang dalam mencari dan membaca informasi. Untuk mengatasi permasalahan tersebut banyak dikembangkan metode peringkasan dokumen yang diharapkan mampu menemukan kalimat-kalimat penting dari dokumen sumber. Penelitian ini mengajukan strategi baru peringkasan dokumen berbahasa inggris menggunakan metode sebaran local sentence untuk mencari dan menggali kalimat penting yang tersembunyi dalam dokumen sumber sebagai upaya untuk meningkatkan kualitas hasil ringkasan. Uji coba dilakukan terhadap dataset task 2 DUC 2004. Pengukuran ROUGE-1 dan ROUGE-2 digunakan sebagai evaluasi performa metode yang diusulkan dengan metode lain yaitu metode sentence information density dan kata kunci cluster kalimat (SIDeKiCK). Hasil ujicoba didapatkan bahwa metode yang diusulkan memiliki performa lebih baik dengan capaian rata-rata ROUGE-1 0,398, meningkat 1,5% dibanding metode SIDeKiCK dan ROUGE-2 0,12 meningkat 13% dibanding metode SIDeKiCK.*
***Kata Kunci:*** *Peringkasan Dokumen, Kalimat Penting, Sebaran Local Sentence, ROUGE.*

## 1. Pendahuluan

Jumlah dokumen digital mengalami peningkatan yang sangat pesat sehingga banyak memunculkan permasalahan baru dalam menggali dan memperoleh informasi secara cepat dan akurat. Jumlah dokumen yang terus berkembang menyebabkan para penggali informasi harus meluangkan waktu ekstra dalam mencari dan membaca informasi. Permasalahan lain yang muncul adalah besarnya potensi kehilangan informasi penting yang ada pada dokumen tersebut. Para peneliti mencoba menyelesaikan permasalahan ini dengan melakukan pengembangan metode dalam peringkasan dokumen.

Peringkasan dokumen adalah proses mengambil teks dari sebuah dokumen, menggali dan menyajikan informasi penting bagi user atau aplikasi dalam bentuk rangkuman yang singkat dan padat (Kogilavani, dkk., 2010). Peringkasan dokumen dapat menjadi solusi bagi setiap

orang yang tidak memiliki banyak waktu dan sedang membutuhkan informasi penting dalam tumpukan dokumen yang terus berkembang.

Ringkasan dokumen yang baik adalah ringkasan yang mampu mencakup (*coverage*) sebanyak mungkin konsep-konsep penting (*salient*) yang ada pada dokumen sumber (Ouyang, dkk., 2013). *Coverage* dan *saliency* adalah masalah utama dalam metode peringkasan dimana strategi pemilihan kalimat menjadi sangat penting karena harus mampu memilih kalimat-kalimat utama dan terhindar dari redundansi sehingga mampu mencakup banyak konsep (Suputra, dkk., 2013).

Beberapa penelitian (Sarkar, 2009; He, dkk., 2008; Suputra, dkk., 2013) telah mengembangkan metode pemilihan kalimat penting untuk menangani masalah *coverage* dan *salient*. Salah satu metode yang cukup baik adalah metode *sentence information density* dan kata kunci *cluster* kalimat (*SIDeKiCK*) yang diusulkan oleh Suputra, dkk. (2013). Menurut Suputra, dkk. (2013) kalimat penting adalah kalimat yang memiliki kepadatan informasi kalimat (*sentence information density*) dan memiliki banyak kata kunci *cluster* kalimat. Kepadatan informasi kalimat dapat digali dengan pendekatan *positional text graph* dan kata kunci *cluster* kalimat dapat digali menggunakan pendekatan metode *TF.IDF* (Suputra, dkk.,2013). Namun pendekatan *positional text graph* pada kondisi terdapat beberapa kalimat dengan bobot yang hampir sama, sulit untuk menentukan kalimat penting (Kruengkrai, dkk., 2003). Sedangkan metode kata kunci *cluster* kalimat yang diperoleh dengan konsep *TF.IDF* tidak mampu memberikan bobot kata kunci *cluster* secara maksimal (Tian, dkk.,2011).

Menurut Tian, dkk. (2011) kata penting adalah kata tersebar. Kata tersebar pada sebuah dokumen lebih mencerminkan topik dari dokumen tersebut. Kata tersebar tidak sama dengan *frekuensi* kemunculan kata. Kata yang memiliki *frekuensi* kemunculan kata sama dapat memiliki bobot yang berbeda karena perbedaan sebarannya. Jika diaplikasikan dalam *cluster* kalimat kata penting adalah kata tersebar dalam *cluster*, maka kalimat penting seharusnya dapat dihitung dengan mengkalkulasi sebaran kalimat pada *cluster* tersebut.

Penelitian ini mengusulkan strategi baru pembobotan kalimat untuk peringkasan dokumen berbahasa inggris menggunakan metode sebaran *local sentence*. Sebaran *local sentence* digali berdasarkan sebaran unsur kata pembentuk kalimat dalam sebuah *cluster* kalimat. Pemilihan kalimat penting dengan metode tersebut diharapkan mampu meningkatkan kualitas hasil ringkasan.

## 2. Penelitian Terkait

Metode *centroid-based* (Randev, dkk., 2004) adalah salah satu metode yang pupoler dalam peringkasan dokumen secara *extractive*. *Centroid-based Summarization* (CBS) diimplementasikan pada MEAD menggunakan konsep *topic detection* dengan memberikan skor terhadap kalimat berdasarkan *sentence-level* dan *inter-sentence features*. Fitur-fitur yang diekstraksi adalah fitur *cluster centroid value*, *sentence position value*, dan *fist sentence overlap*. Pada paper tersebut juga diajukan sebuah metode baru yaitu *Cross-Sentence Informational Subsumption* (CSIS) yang digunakan untuk mengantisipasi redundansi pada kalimat-kalimat ringkasan. Kalimat-kalimat penyusun ringkasan diurutkan berdasarkan urutan (kronologi) kemunculan kalimat pada suatu dokumen.

Kombinasi metode dalam peringkasan multi-dokumen berdasarkan fokus *query* telah dilakukan untuk mendapatkan kalimat-kalimat utama pada suatu kumpulan dokumen (He, dkk., 2008). Fitur-fitur yang diajukan yaitu *query-similarity*, *skip Bi-gram co-occurrence with query* dan fitur *information density* yang digali dari *positional text graph*. *Query-similarity* dan *skip Bi-gram co-occurrence with query* adalah fitur yang fokus pada similaritas kalimat dengan *query* sedangkan fitur *information density* merupakan fitur yang mampu menggambarkan kepadatan informasi yang ada pada suatu kalimat. Pada penelitian tersebut juga digunakan konsep *Maximal Marginal Relevance* (MMR) (Carbonell, dkk., 1998) untuk mengantisipasi kalimat-kalimat redundan pada proses penyusunan ringkasan.

Kombinasi metode yang digunakan oleh Suputra, dkk. (2013) digunakan untuk memilih kalimat repesentatif pada setiap *cluster* kalimat. Setiap *cluster* kalimat akan dipilih satu kalimat

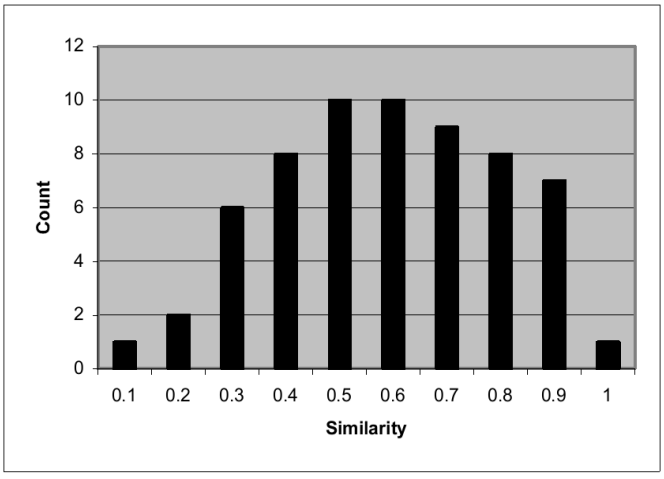
yang memiliki bobot tertinggi berdasarkan perhitungan kepadatan informasi kalimat (*sentence information density*) dan kata kunci *cluster* kalimat. Kalimat-kalimat penting terpilih akan menjadi kalimat ringkasan yang tersusun berdasarkan kualitas atau urutan bobot *cluster* kalimat. Penelitian ini memakai framework penelitian yang hampir sama dengan penelitian Sarkar (2009).

## 3. Kajian Pustaka

### 3.1. *Clustering* Kalimat

*Clustering* kalimat adalah bagian yang penting dalam sistem peringkasan otomatis karena setiap topik dalam set dokumen harus diidentifikasi secara tepat untuk menemukan *similarity* dan *dissimilarity* yang ada dalam dokumen sehingga menjamin *good coverage* (Sarkar, 2009). Jika kalimat-kalimat dikelompokkan ke dalam sejumlah *cluster* yang telah ditentukan, *cluster* mungkin tidak koheren karena beberapa kalimat bisa saja terpaksa menjadi salah satu anggota *cluster* meskipun seharusnya tidak. *Cluster-cluster* tidak koheren mungkin mengandung unit-unit teks yang terduplikasi pada *cluster* yang berbeda dan menyebabkan pemilihan kalimat menjadi redundan untuk ringkasan. Sebaliknya, jika *cluster* sangat ketat, sebagian besar *cluster* menjadi *singletons*. Dengan demikian, harus dipilih metode *clustering* yang menjamin koherensi *cluster* dan meminimalkan jarak anggota antar-*cluster*. Pada penelitian ini digunakan *Similarity based Histogram Clustering*(*SHC*) yang diadopsi dari penelitian Sarkar (2009).

Salah satu cara untuk mendapatkan derajat koherensi yang tinggi dalam *cluster* adalah mempertahankan derajat *similarity* antar anggota tetap tinggi. Ilustrasi tersebut dapat dilihat pada Gambar 1, dimana setiap pasangan kalimat digambarkan dengan *bin*. bin adalah total jumlah dari nilai pasangan *similarity* sesuai dengan interval dari *bin* tersebut. Untuk menjaga agar tidak terjadi pemilihan kalimat secara *redundant* dalam ringkasan harus hati-hati dalam menjaga setiap *cluster* agar tetap berkaitan (koheren). Jika ada anggota baru yang masuk dan anggota tersebut dapat menurunkan *similarity* dalam *cluster* maka lebih baik anggota baru dikeluarkan dari *cluster* tersebut. Untuk kriteria sebuah anggota baru dapat masuk ke dalam *cluster* maka nilai *threshold* digunakan sebagai tolak ukur. Hanya anggota yang memenuhi *threshold* yang telah ditentukan dapat masuk kedalam *cluster*.hal ini dilakukan agar koherensi antar anggota *cluster* tetap optimal.




**Gambar 1. Distribusi *similarity* setiap pasangan anggota *cluster*** (Sarkar, 2009)

Setiap elemen (kalimat) pada tiap *cluster* sebelum dapat bergabung dengan suatu *cluster* diukur terlebih dahulu tingkat kemiripannya menggunakan metode *uni-gram matching-based similarity measure* seperti pada Persamaan (1). Persamaan (1) menunjukkan *similarity* antar kalimat dihitung berdasarkan kata-kata yang sesuai antara kalimat s ke-*i* dan kalimat s ke-*j* $|s_i| \cap |s_j|$ dibagi dengan jumlah panjang kalimat *s* ke–*i* dan s ke-*j* $|s_i| + |s_j|$.

Jika $n$ adalah jumlah dari kalimat pada suatu *cluster*, maka jumlah dari pasangan kalimat yang ada pada *cluster* tersebut adalah $m$ dimana $m=n(n+1)/2$ dan $Sim=\{sim_1, sim_2, sim_3, \ldots, sim_m\}$ adalah kumpulan dari pasangan *similarity* antar kalimat sejumlah $m$. *Similarity histogram* dari *cluster* dinotasikan dengan $H=\{h_1, h_2, h_3,..., h_{nb}\}$. Fungsi untuk menghitung $h_i$ ditunjukkan pada Persamaan (2). Persamaan (2) menunjukkan jumlah *similarity* pasang setiap kalimat pada *bin* ke-i $h_i$ dalam *cluster* tertentu diperoleh dengan menjumlahkan *similarity* pasangan setiap kalimat pada *bin* ke-*i* $sim_j$ pada *cluster* tersebut dengan batas bawah nilai *similarity* pada *bin* ke-*i* $sim_{li}$ dan batas atas nilai *similarity bin* ke-*i* $sim_{ui}$.

*Histogram ratio* (*HR*) dari suatu *cluster* dapat dihitung dengan Persamaan (3) dan (4). Persamaan (3) menunujukkan *histogram ratio* sebuah *cluster* dapat dihitung dengan menghitung seluruh jumlah *similarity* pasangan kalimat $h_i$ yang lolos *threshold* $S_T$ dibagi dengan seluruh jumlah *similarity* pasangan kalimat pada keseluruhan *bin* $n_b$.

$$sim(s_i, s_j) = \frac{(2*|s_i| \cap |s_j|)}{|s_i| + |s_j|} \quad (1)$$

$$h_i = count(sim_j) \quad (2)$$

$$HR = \frac{\sum_{i=T}^{n_b} h_i}{\sum_{j=1}^{n_b} h_j} \quad (3)$$

$$T = \lfloor S_T * n_b \rfloor \quad (4)$$

### 3.2. Pembobotan *Cluster* Kalimat

Salah satu kelemahan pada tahap *clustering* kalimat dengan *similarity based histogram clustering* (*SHC*) adalah tidak diketahui jumlah *cluster* kalimat yang akan terbentuk. Untuk itu pengurutan *cluster* dapat digunakan sebagai solusi menentukan kriteria *cluster* yang layak dijadikan bagian dari proses penyusunan ringkasan. Pengurutan *cluster* kalimat diperlukan untuk memberikan prioritas terhadap *cluster* yang mengandung kekayaan informasi dalamnya. *Cluster* yang memiliki kekayaan informasi seharusnya diberikan prioritas yang lebih tinggi dalam peringkasan dokumen. untuk itu perlu adanya mekanisme yang dapat mengatur dan mengukur tingkat prioritas *cluster*.

Sarkar (2009) mengusulkan metode *cluster importance* untuk melakukan pengurutan *cluster* kalimat dengan cara *descending*. Pengurutan *cluster* dapat digunakan sebagai solusi menentukan kriteria *cluster* yang layak dijadikan bagian dari proses penyusunan ringkasan, yaitu dengan menguji setiap kata yang ada pada *cluster* berdasarkan nilai *threshold* $\theta$. Sebuah bobot *cluster* ke-*j* *Weight*($c_j$) yang memiliki jumlah *frekuensi* suatu kata $w$ (*count*($w$)) memenuhi *threshold* $\theta$ maka kata tersebut adalah kata *frequent*. Bobot kata $w$ dihitung berdasarkan *frekuensi* dari seluruh kata pada dokumen input. Perhitungan bobot *cluster* hanya dilakukan mengacu pada banyaknya kata $w$ yang *frequent* yang dimiliki oleh *cluster* tertentu. Pengurutan *cluster* berdasarkan bobot *cluster importance* dihitung dengan Persamaan (5).

$$Weight(c_j) = \sum_{w \in c_j} \log(1 + count(w)) \quad (5)$$

### 3.3. Metode Sebaran Kata

Dalam penelitian yang dilakukan oleh Tian, dkk. (2011) ditemukan permasalahan bahwasanya sebuah *term* dalam sebuah dokumen yang tersebar dalam beberapa paragraf seharusnya memiliki nilai yang lebih tinggi jika dibanding dengan *term* yang memiliki *frekuensi* yang sama namun hanya tersebar dalam bagian (paragraf) tertentu, karena sebuah *term* yang

tersebar dalam dokumen lebih mempresentasikan topik dari dokumen tersebut, untuk lebih memahami permasalahan ini perhatikan contoh kalimat pada Tabel 1. Contoh sebaran kata pada Tabel 1 jika dihitung bobot kata-katanya menggunakan *frekuensi* kemunculan kata maka hasilnya dapat dilihat pada Tabel 2.

**Tabel 1. Contoh sebaran kata**

| |
|---|
| KPK sedang dirundung masalah yang pelik, setelah menetapankan calon kapolri sebagai tersangka beberapa pemimpinnya dilaporkan ke bareskrim dalam kasus kriminal. |
| Banyak pihak berpendapat bahwa pimpinan KPK sedang dikriminalisasi oleh berbagai pihak yang tidak senang dengan pemberantasan korupsi. |
| Bareskrim juga terkesan ikut melemahkan KPK dengan merespon cepat laporan masyarakat ke bareskrim. |

**Tabel 2. Bobot kata berdasarkan *frekuensi* kemunculan kata**

| Term | Frekuensi |
|---|---|
| Kpk | 3 |
| Bareskrim | 3 |
| Kriminal | 2 |
| Pihak | 2 |
| masalah | 1 |
| kapolri | 1 dst |

**Tabel 3. Perbedaan bobot menggunakan *TF* dan Sebaran Kata**

| Term | TF | $W_{d\text{-}l}$ |
|---|---|---|
| Kpk | 3 | 0,913 |
| bareskrim | 3 | 0,305 |
| kriminal | 2 | 0,691 |
| Pihak | 2 | 0,155 |
| masalah | 1 | 0,297 |
| Kapolri | 1 | 0.297 |

Pembobotan *TF* (*term frequency*) pada Tabel 2 menunjukkan bahwa *term kpk* dan *term bareskrim* akan memiliki nilai bobot yang sama, sedangkan *term kpk* sendiri lebih mewakili isi dari seluruh dokumen jika dibandingkan dengan *term bareskrim*, karena *term kpk* lebih tersebar luas disetiap bagian(paragraf). Begitu juga dengan *term kriminal* dan *term pihak* yang seharusnya memiliki bobot yang berbeda karena memiliki sebaran yang berbeda. Tian, dkk. (2011) mengusulkan sebuah metode baru sebaran kata menggunakan pendekatan *Chi-Square Test Statistic*, sehingga jika dihitung dengan metode sebaran kata diperoleh bobot baru dokumen tersebut seperti pada Tabel 3. Persamaan 6 menunjukkan metode sebaran kata yang diusulkan dalam penelitian tersebut.

Untuk menghitung sebaran kata dalam sebuah dokumen ($W_{d\text{-}l}$) ke-*j* dipengaruhi oleh nilai *p* yang menunjukkan banyaknya paragraph yang mengadung kata ke-*j*, nilai *P* menunjukkan jumlah paragraph dalam dokumen tersebut, $v_i$ menunujukkan frekuensi kemunculan kata ke-*j* pada paragraph ke-*i*, *n* menunjukkan fekuensi kata ke-*j* dalam dokumen dan $r_i$ adalah peluang sebaran kata dalam paragraph ke-*i* yang diperoleh dari jumlah kata berbeda pembentuk paragraph ke-*i* ($C_i$) dibagi dengan jumlah $C_i$ dalam dokumen tersebut.

$$W_{d-l} = \log_2(1+\frac{\log_2(1+\frac{p}{P})}{1+\sum_{i=1}^{m}\frac{(v_i - nr_i)^2}{nr_i}} \quad (6)$$

### 3.4. Evaluasi Ringkasan

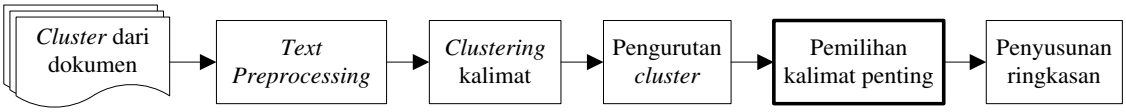
Salah satu metode untuk mengukur hasil ringkasan adalah *ROUGE*. *ROUGE* mengukur kualitas hasil ringkasan dengan menghitung unit-unit yang *overlap* seperti *N-gram*, urutan kata dan pasangan-pasangan kata antara ringkasan kandidat dan ringkasan sebagai referensi *ROUGE*

sangat efektif digunakan untuk mengevaluasi peringkasan dokumen (Lin, 2004). Perhitungan *ROUGE-N* yang diadopsi dari perhitungan Lin (2004) ditunjukkan pada Persamaan (6), dimana *N* menunjukkan panjang dari *N-gram* yang digunakan, $Count_{match}(N\text{-}gram)$ adalah jumlah maksimum dari *N-gram* yang muncul pada ringkasan kandidat dan ringkasan sebagai referensi, *Count(N-gram)* adalah jumlah dari *N-gram* pada ringkasan sebagai referensi.

$$ROUGE-N = \frac{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count_{match}(N-gram)}{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count(N-gram)} \quad (7)$$

## 4. Metode Penelitian

Metode yang digunakan dalam penelitian ini mengadopsi metode peringkasan dokumen dari penelitian Sarkar (2009). Tahapan dalam penelitian ini ditunjukkan pada Gambar 2.




**Gambar 2. Metode peringkasan dokumen berbahasa inggris**

### 4.1. Data Penelitian

Penelitian ini menggunakan data *task 2 DUC* (*Document Understanding Conference*) *2004*. Dataset task 2 DUC 2004 merupakan kumpulan dokumen berita dalam bahasa Inggris dari *Associated Press* dan *New York Times* yang terdiri dari atas 50 kelompok dokumen dan setiap kelompok dokumen memiliki 10 dokumen berita. Dokumen-dokumen berita yang terdapat pada *task 2 DUC 2004* adalah dokumen dengan format XML. *Dataset DUC 2004* tersedia di: http://www-nlpir.nist.gov/projects/duc/data/2004_data.html. Tabel 4 menunjukkan salah satu contoh format data DUC 2004 task 2 yang digunakan dalam penelitian ini.

**Tabel 4. Format data DUC 2004 task 2 2004**

```
<DOC>
<DOCNO> APW19981101.0843 </DOCNO>
<DOCTYPE> NEWS </DOCTYPE>
<TXTTYPE> NEWSWIRE </TXTTYPE>
<TEXT>
<P>
In Honduras, at least 231 deaths have been blamed on Mitch, the National Emergency
Commission said Saturday. El Salvador where 140 people died in flash floods _ declared a
state of emergency Saturday, as did Guatemala, where 21 people died when floods swept
away their homes. Mexico reported one death from Mitch last Monday. In the Caribbean,the
U.S. Coast Guard widened a search for a tourist schooner with 31 people aboard that
hasn't been heard from since Tuesday. By late Sunday, Mitch's winds, once near 180 mph
(290 kph), had dropped to near 30 mph (50 kph), and the storm now classified as a
tropical depression was near Tapachula, on Mexico's southern Pacific coast near the
Guatemalan border. Mitch was moving west at 8 mph (13 kph) and was dissipating but
threatened to strengthen again if it moved back out to sea.
</P>
</TEXT>
</DOC>
```

### 4.2. *Text Preprocessing*

Sebelum dilakukan *clustering* kalimat, kumpulan dokumen akan diproses melalui tahap *preprocesing* yang meliputi proses *tokenizing*, *stopword removal* dan *stemming*. *Tokenizing* adalah proses pemenggalan kata-kata sehingga setiap kata dapat berdiri sendiri. *Stopword removal* dilakukan untuk menghapus kata kunci yang tidak layak untuk digunakan, seperti kata sambung, kata depan, kata ganti dls. Sedangkan *stemming* adalah proses untuk memperoleh kata dasar dari setiap kata. Proses *tokenizing* dalam penelitian ini dilakukan menggunakan *Stanford Natural Language Processing,* proses *stopword removal* menggunakan kamus *stoplist* dan proses *stemming* menggunakan *Library English Porter Stemmer*.

### 4.3. *Clustering* Kalimat

Setelah tahap *text preprocessing* dokumen, kalimat yang diperoleh akan dikelompokkan menggunakan metode *Similarity based histogram clustering* (*SHC*). Tujuan pengelompokan kalimat adalah untuk menghindari redudansi pada kalimat ringkasan. Kalimat yang sudah terkelompok berdasarkan nilai kemiripannya akan dihitung bobotnya menggunakan metode sebaran *local sentence*. Kalimat yang memiliki bobot paling tinggi pada tiap *cluster* kalimat akan mewakili kelompoknya sebagai kalimat ringkasan. Pengelompokan dengan metode ini dilakukan dengan mengukur nilai *similarity* antar kalimat. Fungsi *similarity* yang digunakan adalah *uni-gram matching-based similarity* sesuai dengan Persamaan (1). Suatu kalimat dapat masuk kedalam suatu *cluster* apabila kalimat tersebut memiliki kriteria dari *cluster* tersebut. Namun, jika kalimat tidak memiliki kriteria pada semua *cluster* yang ada, maka *cluster* baru akan terbentuk. Metode *SHC* untuk *clustering* kalimat yang digunakan dalam penelitian ini mengadopsi dari penelitian (Sarkar, 2009). Penjelasan detail dari fase *clustering* kalimat menggunakan *SHC* dijelaskan pada kajian pustaka sub bab 3.1.

### 4.4. Pengururan *Cluster*

Salah satu kelemahan pada tahap *clustering* kalimat dengan *similarity based histogram clustering* (*SHC*) adalah tidak diketahui jumlah *cluster* kalimat yang akan terbentuk. Untuk itu pengurutan *cluster* dapat digunakan sebagai solusi menentukan kriteria *cluster* yang layak dijadikan bagian dari proses penyusunan ringkasan. *Cluster-cluster* kalimat yang ada akan dihitung nilai bobotnya menggunakan metode *cluster importance* sesuai Persamaan (5) dan hasil ringkasan akan diurutkan secara *descending*.

### 4.5. Pemilihan Kalimat Penting dan Kontribusi Penelitian

Tahap pemilihan kalimat penting adalah tahap untuk memilih kalimat representatif *cluster* menggunakan metode sebaran *local sentence*. Tahap ini juga merupakan kontribusi dari penelitian yang dilakukan yaitu mengusulkan strategi baru pembobotan kalimat menggunakan sebaran *local sentence*. Metode sebaran *local sentence* diusulkan dalam penelitian ini terinspirasi dari penelitian Tian, dkk. (2013) yang mengklaim metode sebaran kata yang diusulkan mampu memperbaiki metode *TF.IDF* yang sudah terkenal cukup baik. Metode sebaran *local sentence* dibentuk melalui proses menghitung peluang sebaran, menghitung total sebaran, menghitung perluasan sebaran, menghitung bobot komponen kalimat dan menghitung bobot sebaran *local sentence*.

Misal sebuah *cluster* mengandung *i* kalimat, maka $|S_{ij}|$ merepresentasikan kalimat ke-*i* mengandung sejumlah komponen kalimat ke-*j* yang berbeda. Setiap komponen kalimat ke-*j* dalam kalimat, jika komponen kalimat tersebut adalah komponen kalimat yang sama dan tersebar maka peluang komponen kalimat dalam kalimat ke-*i* dihitung menggunakan teori K. Pearson adalah sesuai Persamaan (8).

Peluang sebaran $r_{ij}$ diperoleh dari jumlah komponen kalimat berbeda penyusun kalimat *S* ke-*i* pada *cluster* ke-*k* $| s_{ik} |_{dt}$ dibagi dengan jumlah $| s_{ik} |_{dt}$ pada *cluster* ke-*k* $| c_k |$.Jumlah perbedaan antara *frekuensi* komponen kalimat dengan *frekuensi* sebaran komponen kalimat ke-*j* dalam kalimat ke-*i* dapat dihitung menggunakan *chi-square test statistics* sehingga diperoleh sebaran komponen kalimat seragam ke-*j* dalam *cluster* ke-*k* seperti pada Persamaan (9). Sebaran komponen kalimat $\chi^2_{jk}$ diperoleh dari jumlah kuadrat perbedaan antar *frekuensi* komponen kalimat $v_{ij}$ dengan *frekuensi* sebaran komponen kalimat ke-*j* dalam *cluster* ke-*k* $n_{jk}r_{ij}$ dibagi dengan *frekuensi* sebaran komponen kalimat ke-*j* dalam *cluster* ke-*k* tersebut. Persamaan (9) semakin kecil nilai $\chi^2_{jk}$ menunjukkan bahwa komponen kalimat ke-*j* semakin mendekati sebaran maksimal yang mana nilai tersebut bertentang dengan hubungan bobot dan sebaran kata yang memiliki korelasi positif non linier (Tian, dkk.,2011) sehingga diperoleh Persamaan (10).

Bobot komponen kalimat ke-*j* dalam *cluster* ke-*k* tersebar $U_{jk}$ berbanding terbailik dengan sebaran komponen kalimat tersebut $\chi^2_{jk}$. Untuk memperoleh perhitungan bobot

komponen kalimat tersebar ke-*j* dalam *cluster* ke-*k* secara optimal dilakukan perluasan perhitungan sehingga diperoleh Persamaan (11). Perluasan sebaran komponen kalimat ke-*j* pada *cluster* ke-*k* $St_{jk}$ diperoleh dari varibel jumlah kalimat yang mengandung kata ke-*j* pada *cluster* ke-*k* $P_{jk}$ dengan variabel jumlah seluruh kalimat pada *cluster* ke-*k* $P_k$. Sehingga bobot komponen kalimat ke-*j* dalam sebuah *cluster* ke-*k* dapat dihitung dengan Persamaan (12).

Bobot komponen kalimat lokal ke-*j* pada *cluster* ke-*k* $Wt_{l,jk}$ akan membentuk bobot sebaran *local sentence* $W_{ls}(s_{ik})$ dengan menjumlahkan seluruh komponen kalimat pembentuk kalimat *s* ke-*i* pada *cluster* ke-*k* dibagi dengan banyak komponen kalimat pembentuk kalimat *s* ke-*i* pada *cluster* ke-*k* $|s_{ik}|$ seperti pada Persamaan (13). Persamaan (13) merupakan metode sebaran *local sentence* yang akan digunakan untuk mencari dan mengkalkulasi kalimat tersebar dalam *cluster* kalimat. Kalimat tersebar dengan bobot tertinggi dari masing-masing *cluster* akan digunakan sebagai kalimat penyusun ringkasan.

$$r_{ij} = \frac{|s_{ik}|_{dt}}{|c_k|} \tag{8}$$

$$\chi^2_{jk} = \sum_{j=1}^{|c_k|_{dt}} \frac{(v_{ij} - n_{jk} r_{ij})^2}{n_{jk} r_{ij}} \tag{9}$$

$$U_{jk} = \frac{1}{1 + \chi^2_{jk}} \tag{10}$$

$$St_{jk} = \log_2(1 + \frac{p_{jk}}{P_k}) \tag{11}$$

$$Wt_{l,jk} = \log_2(1 + U_{jk} * St_{jk}) \tag{12}$$

$$W_{ls}(s_{ik}) = \frac{1}{|s_{ik}|} \sum_{Wt_{l,jk} \in s_{ik}} Wt_{l,jk} \tag{13}$$

### 4.6. Penyusunan Ringkasan

Ringkasan akan disusun berdasarkan urutan *cluster* kalimat yang telah diperoleh pada tahap pengurutan *cluster*. Setiap *cluster* yang telah berisi beberapa kalimat yang mirip satu sama lain akan diseleksi menggunakan metode sebaran *local sentence* dan kalimat yang memiliki bobot paling tinggi akan dipilih untuk mewakili *cluster*-nya dalam menyusun ringkasan. Jumlah kalimat penyusun ringkasan adalah sama dengan jumlah *cluster* kalimat yang terbentuk.

## 5. Hasil Uji Coba dan Pembahasan

Uji coba dilakukan menggunakan metode sebaran *local sentence* dan dibandingkan dengan metode *SIDeKiCK* yang diusulkan oleh Suputra, dkk. (2013). Parameter uji coba yang digunakan untuk *clustering kalimat* adalah $HR_{min}$=0.7, $\varepsilon$=0,3, *similarity threshold* ($S_T$)=0,4, parameter *cluster ordering* adalah $\theta$ =10 dan parameter khusus metode *SIDeKiCK* $\alpha$=0,4, $\lambda$=0,2 dimana parameter-parameter tersebut adalah parameter terbaik yang direkomendasikan pada penelitian Suputra, dkk. (2013). Dataset yang digunakan untuk uji coba merupakan data *DUC 2004 task 2*.

Untuk evaluasi hasil ringkasan digunakan metode *ROUGE*-1 dan *ROUGE*-2. *ROUGE*-1 merupakan pengukuran dengan konsep *unigram matching*, dimana pengukuran ini dihitung berdasarkan jumlah setiap satu kata (*unigram*) yang sesuai antara ringkasan hasil sistem dengan ringkasan referensi yang dibuat manual oleh pakar. Sedangkan untuk *ROUGE*-2 dihitung berdasarkan jumlah setiap dua pasang kata (*bigram*) yang sesuai antara ringkasan hasil sistem dengan ringkasan referensi yang dibuat oleh pakar, dimana pada kondisi terbaik nilai *ROUGE*-1 dan *ROUGE*-2 maksimal adalah 1.

Tabel 5 menunjukkan bahwa metode sebaran *local sentence* memiliki rata-rata nilai *ROUGE* lebih tinggi jika dibandingkan dengan metode *SIDeKiCK*, baik pada pengujian

*ROUGE*-1 dan *ROUGE*-2. Nilai rata-rata yang diperoleh pada pengujian *ROUGE*-1 dengan menggunakan metode sebaran *local sentence* adalah 0,398, pada metode *SIDeKiCK* didapatkan hasil rata-rata 0,392, artinya metode sebaran *local sentence* lebih baik atau terjadi peningkatan sebesar 1,5% dibandingkan dengan metode *SIDeKiCK*. Pada skema pengujian menggunakan *ROUGE*-2 diperoleh rata-rata nilai *ROUGE*-2 pada sebaran *local sentence* adalah 0,12 metode *SIDeKiCK* 0,1063, artinya metode sebaran *local sentence* lebih baik atau meningkat sebesar 13% dibandingkan dengan metode *SIDeKiCK*. Hasil uji coba menunjukkan bahwa metode sebaran *local sentence* memiliki performa lebih baik jika dibandingkan dengan metode *SIDeKiCK*.

**Tabel 5. Hasil uji coba metode Sebaran *Local Sentence* vs SIDeKiCK**

| Metode Peringkasan | ROUGE-1 | ROUGE-2 |
|---|---|---|
| *Cluster*ing kalimat (*SHC*) + *Cluster Ordering* + Sebaran *local sentence* | 0,398 | 0,12 |
| *Cluster*ing kalimat (*SHC*) + *Cluster Ordering* + SIDeKiCK | 0,392 | 0,106 |

## 6. Kesimpulan

Penelitian ini mengusulkan metode baru pembobotan kalimat penting pada peringkasan dokumen berbahasa inggris. Hasil pengujian menunjukkan metode yang diusulkan (metode *local sentence*) memberikan nilai evaluasi *ROUGE-1* dan *ROUGE-2* lebih baik dibandingkan dengan metode *SIDeKiCK* dengan perolehan *ROUGE-1* 0,398 dan *ROUGE-2* 0,12. Skema pengujian *ROUGE-1* metode *local sentence* mengalami peningkatan 1,5% dibandingkan dengan metode *SIDeKiCK* dan skema pengujian menggunakan *ROUGE-2* metode *local sentence* mengalami peningkatan 13% dibanding dengan metode *SIDeKiCK*.

## Referensi

Carbonell, J., & Goldstein, J. 1998. *The use of MMR, Diversity-Based Reranking for Reordering Documents Andproducing Summaries*. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Eds: Moffat, A. dan Zobel, J., ACM, Melbourne, Australia, hal. 335–336.

He, T., Li F., Shao, W., Chen, J., & Ma, L. 2008. *A New Feature-Fusion Sentence Selecting Strategy for Query-Focused Multi-document Summarization*. Proceeding of International Conference Advance Language Processing and Web Information Technology. Eds: Ock C., dkk., University of Normal, Wuhan, China, hal. 81-86.

Kogilavani, A. & Balasubramani, P. 2010. *Clustering and Feature Sprecific Sentence Extraction Based Summarization of Multiple Documents*. International Journal of Computer Science & Information Technology (IJCSIT). Vol. 2, No. 4, hal. 99-111.

Kruengkrai, C. & Jaruskulchai, C.2003. *Generic Text Summarization Using Local and Global Properties of Sentences*. Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03), IEEE Computer Society Washington DC, Halifax, Canada, hal. 201-206.

Lin, C. Y. 2004. *ROUGE: a Package for Automatic Evaluation of Summaries*. In Proceedings of Workshop on Text Summarization Brances Out. Eds: Moens, M. F. dan Szpakowicz, S., Association for Computational Linguistics, Barcelona, hal. 74-81.

Ouyang, Y., Li W., Zhang R., Li S., & Lu Q. 2013. *A Progressive Sentence Selection Strategy for Document Summarization*. Journal of information Precessing and Management. Vol. 49, Issue 1, hal. 213-221.

Randev, D. R., Jing, H., Stys, M., & Tam, D. 2004. *Centroid-Based Summarization of Multiple Documents.* Journal Information Processing and Management: an International Journal, Vol. 40 Issue 6, hal. 919-938.

Sarkar, K. 2009. *Sentence Clustering-based Summarization of Multiple Text Documents*. International Journal of Computing Science and Communication Technologies. Vol. 2, No. 1, hal. 325-335.

Suputra H. G. I., Arifin Z. A., & Yuniarti A. 2013. *Strategi Pemilihan Kalimat pada Peringkasan Multi-Dokumen Berdasarkan Metode Clustering Kalimat*, Master Thesis of Informatics Engineering ITS.

Tian, X. & Chai Y. 2011. *An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm*. Journal of Software, Vol. 6, No.3, hal 413-420.