Implementasi Metode *Lexical Chain* dengan *Word Sense Disambiguation* untuk Peringkasan Artikel Kesehatan

Dika Muhammad Fazar

Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
E-mail: dika.fazar@gmail.com

ABSTRAK

Text Summarization adalah suatu proses untuk menghasilkan ringkasan dari suatu artikel, tetapi tetap memiliki gambaran atau dengan kata lain tidak menghilangkan inti informasi dari artikel. Ada beberapa metode untuk melakukan peringkasan teks, diantaranya Metode *Lexical chain* yang memiliki performansi yang baik untuk peringkasan teks *single document*, namun penggunaan Metode *Lexical chain* ini masih memiliki kelemahan yaitu tidak bisa mengidentifikasi kata yang berambigu dalam pembentukan rantai leksikal. Oleh karena itu untuk memperbaiki kelemahan tersebut Metode *Lexical chain* dilengkapi dengan *Word sense disambiguation*. Metode *lexical chain* dengan *Word sense disambiguation* merupakan metode pengembangan dari metode *original lexical chain*. Dari hasil penelitian bahwa Metode *lexical chain* dengan *Word sense disambiguation* dapat memperbaiki kelemahan pada metode sebelumnya karena metode *lexical chain* dengan *Word sense disambiguation* dapat mengidentifikasi makna yang berambigu dan menghasilkan ringkasan yang lebih baik akurasinya.

Kata kunci: *Text Summarization*, Metode *lexical chain*, *Word Sense Disambiguation*, Metode *lexical chain dengan Word Sense Disambiguation.*

.

1. PENDAHULUAN

Text Summarization adalah suatu proses untuk menghasilkan ringkasan dari suatu artikel, tetapi tetap memiliki gambaran atau dengan kata lain tidak menghilangkan inti informasi dari artikel. Ada beberapa metode untuk melakukan peringkasan teks, dari jurnal teknik informatika dengan judul "Analisis Perbandingan Peringkasan Teks Menggunakan Metode *Lexical chain* dan Metode *TF-IDF* dengan Ekstraksi Frase Utama" yang menyatakan dalam hasil analisis perbandingan dua metode tersebut bahwa Metode *Lexical chain* yang memiliki performansi lebih baik untuk peringkasan teks *single document* karena adanya pembentukan rantai leksikal pada setiap kalimat sehingga bobot kalimat untuk di ekstraksi ditentukan oleh rantai terkuat tidak tergantung pada judul, namun penggunaan Metode *Lexical chain* ini masih memiliki kelemahan yaitu tidak bisa mengidentifikasi kata yang berambigu dalam pembentukan rantai leksikal. Dimana rantai leksikal tersebut memiliki peran penting dalam proses selanjutnya yaitu menentukan bobot kalimat sebelum menghasilkan ringkasan. Maka dari itu jika terjadi ambiguitas dalam pembentukan rantai leksikal akan sangat berpengaruh kepada pembobotan kalimat dan hasil ringkasan yang dihasilkan. Dalam penelitian tersebut disarankan agar Metode *Lexical chain* dilengkapi dengan *Word sense disambiguation* [1].

Berdasarkan uraian tersebut, maka dalam penelitian tugas akhir ini akan memperbaiki kelemahan dalam penelitian sebelumnya yaitu menghilangkan ambiguitas dengan mengidentifikasi makna kata sebelum pembentukan rantai leksikal dan untuk mengetahui pengaruhnya terhadap hasil ringkasan. Dimana metode yang akan digunakan adalah Metode *Lexical chain* dengan *Word sense disambiguation* untuk peringkasan artikel .

1.1 Text Summarization

Text Summarization atau peringkasan teks adalah sebuah proses untuk menghasilkan ringkasan (*summary*) dari suatu artikel tetapi tetap memiliki gambaran yang akurat dari isi suatu artikel. Tujuannya adalah mengambil sumber informasi atau garis besar informasi dengan mengutip sebagian besar isi yang penting dam menampilkan artikel dalam bentuk yang ringkas.

Text summarization atau peringkasan teks berdasarkan hasil ringkasan dapat dibagi menjadi dua, yaitu ringkasan ekstraksi dan abstraksi. Ringkasan ekstraksi adalah ringkasan yang dihasilkan berisi informasi yang mengutip langsung dari teks asli yang tidak mengalami perubahan susunan kata dari kalimat yang terekstrak. Sedangkan ringkasan abstraksi adalah ringkasan yang dihasilkan dibentuk dalam susunan kata yang berbeda dari inputan. Metode *lexical chain* adalah salah satu metode untuk menghasilkan ringkasan ekstraktif yang menghasilkan poin-poin penting atau informasi utama dari inputan [6]

1.2 Metode *Lexical Chain*

Metode *lexical chain* adalah suatu metode yang membentuk rantai leksikal dari kata-kata yang saling berelasi dalam teks. Hubungan leksikal dapat dibentuk pada keseluruhan kata yang saling berkaitan. Setiap kata yang dihasilkan dari *noun filtering* akan dianggap sebagai kandidat kata yang kemudian akan dicari relasinya dengan membentuk suatu *synset* yang berdasarkan kepada kamus yang telah tersedia. Berdasarkan *synset* dari setiap kandidat kata yang terbentuk, kemudian melakukan pencocokan dari setiap kandidat kata dan *synset* yang dimiliki.

Hubungan semantik antara dua kata dapat dibentuk jika terdapat kecocokan atau kesamaan antara *synset* yang direpresntasikan oleh kata tersebut. Setiap rantai leksikal yang dibentuk akan memiliki ukuran panjang rantai tertentu. Untuk mengidentifikasi nilai rantai pada setiap artikel dilakukan perhitungan berdasarkan [2] menggunakan rumus (2.1) sebagai berikut:

*Score(chain) = length * homogenityindex*

Homogenityindex = 1– (*distinctmember*/*length*)

Keterangan:

a. *Length*: merupakan jumlah total dari seluruh kemunculan anggota penyusun rantai dalam teks.
b. *Distinctmember*: banyaknya anggota yang berbeda yang menyusun rantai.
c. *Homogenityindex*: nilai kemiripan dari anggota-anggota penyusun rantai.

Perhitungan *Homogenityindex* digunakan untuk menunjukan tingkat kemiripan anggota-anggota yang membentuk rantai. *Distinctmember* dibagi dengan *length* menyatakan tingkat perbedaan anggota penyusun rantai.

1.3 Word Sense Disambiguation

Pada *Word sense disambiguation* adalah suatu proses mengidentifikasi makna kata yang digunakan dalam kalimat tertentu ketika kata memiliki sejumlah makna yang berbeda [7].

Dalam tata bahasa komputasi, arti *word sense disambiguation* adalah pemecahan masalah pengolahan bahasa alami, yang terdiri dari proses identifikasi kata yang digunakan dalam sebuah kalimat ketika kata tersebut memiliki makna lebih dari satu. Sebuah proses disambiguasi sangat memerlukan kamus untuk menentukan makna dari kata yang memiliki ambiguitas. *Word sense disambiguation* memiliki dua varian yaitu: semua kata dan sampel leksikal. Yang dimaksud semua kata adalah program harus melakukan disambiguasi terhadap semua kata, sedangkan sampel leksikal melakukan disambiguasi terhadap kata-kata yang sebelumnya dipilih.

1.4 Noun Filtering

Noun filtering merupakan tahapan untuk melakukan pemilihan kata benda (*noun*) dari kumpulan token hasil *preprocessing* yang kemudian dianggap sebagai masukan dalam membentuk rantai leksikal. *Noun filtering* dapat meningkatkan performansi dari peringkasan teks karena komponen subjek suatu kalimat biasanya bertipe kata benda, dan predikatnya menggunakan kata kerja sehingga kata benda dianggap *head* dari kalimat [2]. Selain itu kata benda lebih banyak memiliki relasi semantik dibanding dengan kata kerja.

1.5 Algoitma *Lesk*

Algoritma *lesk* adalah algoritma yang digunakan untuk menghilangkan ambiguitas makna kata. Algoritma *lesk* merupakan salah satu algoritma untuk menyelesaikan masalah *word sense disambiguation* dengan berbasis kamus. Algoritma ini bekerja dengan membandingkan definisi dari kata yang berambigu dengan definisi. Definisi dari kata tetangganya berdasarkan definisi kamus [4]

1.6 Sentence Extractor

Dalam tahap *sentence extractor* memiliki tahap *sentence scoring* dan *sentence extraction*. Dimana *sentence scoring* memiliki tujuan untuk melakukan pembobotan pada setiap kalimat yang menyusun artikel berdasarkan metode yang digunakan. Pembobotan kalimat yang digunakan pada metode *lexical chain* yaitu dilihat dari nilai rantai-rantai leksikal yang telah dibentuk sebelumnya. berdasarkan [9] Pembobotan kalimat ditentukan dengan rumus (2.2) sebagai berikut:

$$score(sen_j) = \sum_{i=1}^{m} \left(\frac{score(ChainMember_i, Sen_j)}{S_i} \right)$$

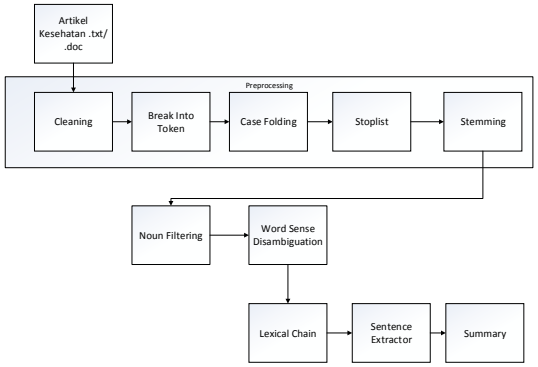
Dimana $score(sen_j)$ adalah nilai dari kalimat j, *score(ChainMember$_i$,Sen$_j$)* adalah nilai dari *ChainMember$_i$* yang anggotanya muncul pada kalimat *Sen$_j$*, dan m adalah jumlah maksimal rantai yang terbentuk pada artikel dengan S_i merupakan jumlah kalimat yang mengandung satu atau semua anggota dari rantai leksikal *ChainMember$_i$* diseluruh teks.

Setiap kalimat akan dihitung bobotnya, kemudian dilakukan *rerangking* secara *descanding*, yang kemudian akan dilakukan proses ekstraksi kalimat pada proses *sentence extraction*. Dimana tujuan dari *sentence extraction* adalah untuk menemukan kalimat yang akan di ekstrak sebagai penyusun hasil ringkasan. Kalimat tersebut masing-masing akan diekstraksi dari artikel tekstual yang asli dan disusun kembali sebagai satu kesatuan ringkasan. Kalimat yang akan diektraksi dipilih berdasarkan nilai bobot masing-masing kalimat kemudian diurutkan berdasarkan urutan kalimat [9].

2. ISI PENELITIAN

2.1 Analisis

Analisis sistem menjelaskan permasalahan pada suatu sistem dan mendeskripsikan kebutuhan pada pembangunan sistem peringkasan artikel. Pada sistem yang akan dibangun terdapat beberapa proses yang memiliki fungsi masing-masing. Alur proses untuk sistem peringkasan artikel dapat dilihat pada gambar 9:



Gambar 1 Alur Proses

Dokumen berupa artikel kesehatan sebagai masukan kedalam sistem sebelum masuk kedalam proses utama, masukan tersebut melalui tahap *preprocessing* terlebih dahulu. Dimana tahap *preprocessing* terdiri dari proses *break into token*, *stoplist* dan *stemming*. Hasil dari *preprocessing* tersebut kemudian masuk ke proses utama dalam sistem yaitu *noun filtering*, *word sense disambiguation*, *lexical chain* dan terkahir adalah *sentence extractor* untuk menghasilkan *output* berupa ringkasan artikel(*summary*). Untuk lebih jelasnya berikut adalah rincian dari data masukan dan masing-masing proses.

1. Data Masukan

Data masukan yang dibutuhkan oleh aplikasi simulasi ini adalah artikel berbentuk teks dengan format .txt atau .doc. dimana dalam penelitian ini data diambil dari kumpulan berbagai artikel kesehatan.

2. *Preprocessing*

Preprocessing adalah tahap dimana sebelum masuk ke proses utama dalam sistem. *Preprocessing* dilakukan terlebih dahulu agar menghasilkan informasi yang siap digunakan pada proses selanjutnya. Pada pembangunan sistem peringkasan artikel ini *preprocessing* yang dilakukan yaitu:

a. *Cleaning*

Tahap pembersihan karakter-karakter yang tidak diperlukan dalam proses selanjutnya.

b. *Break into token*

Tahap pemotongan teks input berdasarkan tiap kata yang menyusunnya.

c. *Case folding*

Pada tahap *Case folding* ini merupakan tahap dimana semua huruf akan diubah menjadi lowercase atau huruf kecil.

d. *Stoplist*

Token-token yang telah dihasilkan saat proses *Break into token* dan telah menjadi huruf kecil semua dalam tahap *case folding* kemudian diproses kembali dengan melakukan pembuangan kata yang termasuk *stoplist* yaitu kata-kata yang dianggap tidak memiliki bobot dalam dokumen. Artinya, kata-kata ini tidak memiliki makna penting dalam dokumen seperti "adalah", "yang", "dari", dan sebagainya. Hasil dari proses pembuangan *stoplist* inilah yang akan digunakan untuk proses selanjutnya yaitu proses *stemming*. Hasil pembuangan *stoplist* disebut dengan *non-stopword tokens*.

e. *Stemming*

Non-stopword tokens dari proses *stoplist* akan difilter kembali melalui proses *stemming* yaitu mengubah kata menjadi kata dasar pembentuknya. Misalnya, kata "kelangsungan" diubah menjadi kata "langsung", kata ”merawat” diubah menjadi kata ”rawat”.

Contoh penerapan preprocessing dapat diliihat pada tabel 1.

Tabel 1 Proses *Preprocessing*

No.	*Preprocessing*				
	Cleaning	***Break into token***	***Case folding***	***Stoplist***	***Stemming***
1	Hati merupakan organ vital dari tubuh manusia dengan peran yang sangat pentinguntuk melengkapi kinerja organ lain dalam meneruskan kelangsungan hidup setiap orang	Hati merupakan organ vital dari tubuh manusia dengan peran yang sangat penting untuk melengkapi kinerja organ lain dalam meneruskan kelangsungan hidup setiap orang.	hati merupakan organ vital dari tubuh manusia dengan peran yang sangat penting untuk melengkapi kinerja organ lain dalam meneruskan kelangsungan hidup setiap orang.	hati organ vital tubuh manusia peran sangat penting melengkapi kinerja organ meneruskan kelangsungan hidup orang	hati organ vital tubuh manusia peran sangat penting lengkap kinerja organ terus langsung hidup orang

3. *Noun filtering*

Pada tahapan ini dilakukan penyaringan kata benda (noun) dari kumpulan token hasil *preprocessing* yang kemudian dianggap sebagai input dalam membentuk rantai leksikal. Penyaringan kata benda

dilakukan dengan cara pengecekan setiap kata dengan database kata benda sebagai acuan untuk mendapatkan kata benda.
Contoh penerapan proses *noun filtering* dapat dilihat pada tabel 2:

Tabel 2 Proses *Noun Fitering*

No.	Hasil *Preprocessing*	Hasil *Noun filtering*
1	hati organ vital tubuh manusia peran penting lengkap kinerja organ terus langsung hidup orang	hati organ vital tubuh manusia peran kinerja organ terus hidup orang

4. *Word sense disambiguation*

Pada tahap ini dilakukan identifikasi makna kata dengan algoritma *lesk*. Dimana algoritma *lesk* bekerja dengan membandingkan definisi dari kata yang berambigu dengan definisI.definisi dari kata tetangganya berdasarkan definisi kamus. Proses *word sense disambiguation.* Contoh Perbandingan makna kata yang berambigu dengan kata tetangga dapat dilihat pada tabel 3:

Tabel 3 Proses *Word Sense Disambiguation*

No.	Kata Berambigu	Makna kata
1	Hati	[1]organ tubuh yang berwarna kemerah-merahan di bagian kanan atas rongga perut gunanya untuk mengambil sari-sari makanan di dalam darah dan menghasilkan empedu

Kata pembanding	Makna pembanding	bobot
organ	alat yang mempunyai tugas tertentu di dalam tubuh manusia	5
vital	organ tubuh yang sangat penting bagi kehidupan	3
tubuh	keseluruhan struktur organisme manusia atau binatang yang kelihatan dr bagian ujung kaki sampai ujung rambut	2
manusia	makhluk yang berakal budi mampu menguasai atau	1
	makhluk insan orang	
peran	pemain sandiwara	0
kinerja	kemampuan kerja	0
terus	tetap berlanjut atau tidak berhenti-henti	0
hidup	masih terus ada bergerak dan bekerja sebagaimana mestinya seperti manusia binatang atau tumbuhan	1
orang	makhluk hidup yang dapat mengembangkan kemampuan bertindak independen atau manusia	1

5. *Lexical chain*

Pada tahap ini dilakukan pembentukan rantai yang menggambarkan hubungan semantik dari suatu kata dengan kata yang lain dan pemberian nilai dengan. Pemberian nilai tersebut didapat dari rumus (2.1) berikut ini:

*Score(chain) = length * homogenityindex*
Homogenityindex = 1 – (*distinctmember*/*length*)

Pembentukan rantai leksikal didapat dari kandidat kata yang telah terpilih dari proses *noun filtering* serta telah didisambiguasi pada proses word sense disambuguation. Hasil perhitungan rantai leksikal dapat dilihat pada tabel 4.

Tabel 4 *Lexical Chain*

No.	Rantai leksikal	Pemberian Nilai	Score (chain)
1	hati organ vital tubuh manusia kinerja terus hidup orang	homo=1-(14/23)=0.39 score= 22*0.39=8.97	8.97
2	organ kelenjar tubuh ikut bawah cara hati manusia	Homo=1-(18/20)=0.10 Score=20*0.10	2
3	vital hati	Homo=1-(7/7)= 0 Score=7*0=0	0
4	atur metabolisme saring racun masuk tubuh bentuk antibodi hati	Homo=1-(10/16)=0.38 Score=16*0.38=6.08	6.08
5	pola hidup jadi organ tubuh jaga hati	Homo=1-(11/19)=0.42 Score=19*0.42	7.98

6. *Sentence extractor*

Pada tahap ini dilakukan sentence scoring dan sentence extraction. Dimana sentence scoring memiliki tujuan untuk melakukan pembobotan pada setiap kalimat yang menyusun artikel berdasarkan metode yang digunakan. Pembobotan kalimat yang digunakan pada metode *lexical chain* yaitu dilihat dari nilai rantai leksikal yang telah dibentuk.
Berikut adalah rumus (2.2) untuk pembobotan kalimat:

$$score(sen_j) = \sum_{i=1}^{m} \left(\frac{score(ChainMember_i, Sen_j)}{S_i} \right)$$

hasil perhitungan *sentence extractor* dapat dilihat pada tabel 5.

Tabel 5 *Sentence Extractor*

No.	Kalimat	Nilai sentence scoring
1	Hati merupakan organ vital dari tubuh manusia dengan peran yang sangat penting untuk melengkapi kinerja organ lain dalam meneruskan kelangsungan hidup setiap orang.	1.79
2	Sebagai organ kelenjar terbesar dalam tubuh membuat kita mau tidak mau harus merawatnya dengan baik, berikut dibawah ini cara merawat kesehatan organ hati pada manusia.	0.40
3	Vitalnya fungsi hati mengharuskan Anda untuk senantiasa merawatnya.	0
4	Mengatur metabolisme, menyaring racun yang masuk ke tubuh, hingga membentuk antibodi merupakan sebagian dari fungsi hati.	1.22
5	Pola hidup yang tidak sehat selalu menjadi momok bagi kesehatan organ tubuh, khususnya menjaga kesehatan hati.	1.60

Contoh hasil peringkasan artikel kesehatan, dimana pada analisis ini ringkasan yang dihasilkan 50% dari semua kalimat yang di analisis. Maka dari itu hasil dari ringkasan tersebut berjumlah 3 kalimat:

Tabel 6 Hasil Peringkasan Artikel

Hasil Ringkasan Kalimat
Hati merupakan organ vital dari tubuh manusia dengan peran yang sangat penting untuk melengkapi kinerja organ lain dalam meneruskan kelangsungan hidup setiap orang. Mengatur metabolisme, menyaring racun yang masuk ke tubuh, hingga membentuk antibodi merupakan sebagian dari fungsi hati. Pola hidup yang tidak sehat selalu menjadi momok bagi kesehatan organ tubuh, khususnya menjaga kesehatan hati.

3. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan mengenai Implementasi Metode Lexical Chain dengan Word Sense Disambiguation Untuk Peringkasn Artikel Kesehatan dapat ditarik kesimpulan sebagai berikut:

1. Metode Lexical Chain dengan Word Sense Disambiguation dapat mengidentifikasi kata yang memiliki ambiguitas atau makna ganda, dan dapat memilih makna dari kata yang memiliki ambiguitas dengan baik.
2. Metode Lexical Chain dengan Word Sense Disambiguation memiliki akurasi yang lebih baik daripada metode lexical chain original dalam peringkasan artikel.

DAFTAR PUSTAKA

[1] R. Y. Susanti, M. A. Bijaksana dan Y. F. A.w., "ANALISIS PERBANDINGAN PERINGKASAN TEKS MENGGUNAKAN METODE LEXICAL CHAIN DAN METODE TF-IDF DENGAN EKSTRAKSI FRASE UTAMA," Universitas Telkom, Bandung, 2009.

[2] R. Barzilay dan M. Elhadad, "Using Lexical Chains for Text," *Mathematics and Computer Science.*

[3] J. EKEDAHL dan K. GOLUB, "Word sense disambiguation using WordNet and the Lesk algorithm," Lund Univ, Sweden.

[4] S. Banerjee, "Adapting the Lesk Algorithm for Word Sense Disambiguation toWordNet," University of Minnesota, Duluth, Minnesota, 2002.

[5] P. D. Sugiyono, METODE PENELITIAN KANTITATIF, KUALITATIF, DAN KOMBINASI (MIXED METHODS), BANDUNG: ALFABETA, CV., 2013.

[6] R. A. García-Hernández, R. Montiel, Y. Ledeneva, E. Rendón, A. Gelbukh dan R. Cruz, "Text Summarization by Sentence Extraction Using Unsupervised Learning," Autonomous University of the State of Mexico, Mexico, 2008..

[7] R. Navigli, "Word sense disambiguation," *ACM Computing Surveys*, vol. 41, no. 2, 2009.

[8] N. Roberto, "Word sense disambiguation: A survey," 2009.

[9] M. Brunn, Y. Chali dan C. J. Pinchak, "Text Summarization Using Lexical Chains," Lethbridge, Alberta, Canada, Department of Mathematics and Computer Science University of Lethbridge.

[10] R. Miles dan K. Hamilton, "Learning UML 2.0," Sebastopol, O'Reilly Media, 2006.

[11] M. E. Khan, "Different Forms of Software Testing Techniques for Finding Error," *Journal of Computer Science*, vol. 7, pp. 11-16, 2010.

[12] P. P. Yanti, "Identifikasi Freycinetia Berbasis Citra Anatomi Epidermis," Institut Pertanian Bogor, Bogor, 2012.

[13] H. Jing, K. McKeown, R. Barzilay dan M. Elhadad, "Summarization Evaluation Methods," Dept. of Computer Science, New York, NY.

[14] M. M. Ardytha Luthfiarta, M. M. Junta Zeniarja dan M. Abu Salam, "INTEGRASI PERINGKAS DOKUMEN OTOMATIS DENGAN ALGORITMA LATENT SEMANTIC ANALYSIS (LSA) UNTUK PROSES CLUSTERING DOKUMEN," UNIVERSITAS DIAN NUSWANTORO, Semarang, 2013.

[15] C. Y. Lin, "A Package for Automatic Evaluation of Summaries," Information Sciences Institute University of Southern California, Marina del Rey, CA.

[16] P. P. YANTI, "IDENTIFIKASI FREYCINETIA BERBASIS CITRA ANATOMI EPIDERMIS DAUN MENGGUNAKAN K-NEAREST NEIGHBOR," INSTITUT PERTANIAN BOGOR, BOGOR, 2012.