Implementation Of Lexical Chain Method With Word Sense Disambiguation For Health Article Summarization

Dika Muhammad Fazar

Informatics Engineering – Indonesia Computer University
Dipatiukur Street No. 112-114 Bandung
E-mail: dika.fazar@gmail.com

ABSTRAK

Text Summarization is a process for producing a summary of an article, but still have a picture or in other words do not eliminate the core information from the article. There are several methods for doing summary text, such as Method Lexical chain that has good performance for summary single text document, but the use of this chain Lexical method still has the disadvantage that it can not identify the ambiguous word in the formation of lexical chain. Therefore, to rectify these shortcomings Lexical Method chain comes with Word sense disambiguation. Methods of lexical chain with Word sense disambiguation is a method of development of the original method of lexical chain. From the research that the method of lexical chain with Word sense disambiguation can improve the weaknesses in the previous method because the method of lexical chain with Word sense disambiguation can identify meaning that ambiguous and produce summaries better accuracy.
Keywords: Text Summarization, lexical chain method, Word Sense disambiguation, lexical chain method with Word Sense disambiguation.

.

1. INTRODUCTION

Text Summarization is a process for producing a summary of an article, but still have a picture or in other words do not eliminate the core information from the article. There are several methods for doing summary text of journal informatics techniques entitled "Comparative Analysis summary Text Method Using Lexical chain and Methods TF-IDF with Extraction Phrases Top" stating the results of a comparative analysis of the two methods that the method Lexical chain that has a performance more good for summary text of a single document for the establishment of a chain of lexical in each statement so that the weight of the sentence for the extraction is determined by the chain's strongest does not depend on the title, but the use method lexical chain still has the disadvantage that it can not identify the ambiguous word in the chain formation lexical. Where the lexical chain has an important role later in the process of determining the weight of the sentence before generating the summary. Therefore if there is an ambiguity in the formation of lexical chain would be devastating to the weighting of the sentence and the summary results generated. In that study suggested that the method of Lexical chain comes with Word sense disambiguation [1].

Based on these descriptions, then in this research will improve the weaknesses in previous research that removes ambiguity by identifying the meaning of the word before the formation of lexical chain and to determine their effects on the summary results. Where the method to be used is the method Lexical chain with Word sense disambiguation for summary articles.

1.1 Text Summarization

Text Summarization atau summary 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 summary 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 Lexical Chain Method

Metode Lexical chain method is a method of forming a chain of lexical of words that relate to each other in the text. Lexical relationships can be formed on a whole word interrelated. Every word is generated from the noun filtering will be considered as a candidate word will then be sought relation to form a synset which is based on a dictionary that has been provided. Based synset of each candidate word

is formed, then do the matching of each candidate word and synset owned.
The semantic relationship between the two words can be formed if there is a match or similarity between synset that direpresntasikan by the word. Each lexical chains formed will have a specific chain length. To identify the value chain in each article be calculated by [2] using the formula (2.1) as follows:

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

Description:
a. Length: is the total of all constituent members of the chain of appearance in the text.
b. Distinctmember: the number of different members that make up the chain.
c. Homogenityindex: the value of the similarity of the members making up the chain.
Homogenityindex calculation used to show the degree of similarity of the members that make up the chain. Distinctmember divided by the length stated the level difference constituent members of the chain.

1.3 Word Sense Disambiguation

In Word sense disambiguation is a process of identifying the meaning of the words used in a particular sentence when the word has a different meaning [7].
In the computational grammar, meaning word sense disambiguation problem solving natural language processing, which consists of the identification process a word used in a sentence when the word has more than one meaning. A very disambiguation process requires a dictionary to determine the meaning of a word that has ambiguity. Word sense disambiguation has two variants, namely: all words and lexical sample. The meaning of all words is a program must perform disambiguation of all words, while samples lexical perform disambiguation of the words that were previously selected.

1.4 Noun Filtering

Noun filtering is a stage for an election noun (noun) from the collection of the results of preprocessing tokens are then considered as an input in shaping the lexical chain. Noun filtering can improve the performance of summary text for component type the subject of a sentence is usually a noun, verb and predicate noun that is considered the head of the sentence [2]. Additionally noun has more than semantic relation to the verb.

1.5 Algoitma Lesk

Lesk algorithm is an algorithm that is used to remove the ambiguity in the meaning of the word. Lesk algorithm is an algorithm to solve the problem of word sense disambiguation with the dictionary. This algorithm works by comparing the definition of the word berambigu by definition. The definition of the word neighbor by the dictionary definition [4].

1.6 Sentence Extractor

In the sentence phase extractor has stages scoring sentence and sentence extraction. Where the sentence scoring own goal to make the weighting of each sentence that compose an article based on the method used. Weighting phrases used in the method of lexical chain that is seen from the lexical chains that had been formed earlier. by [9] The weighting sentence is determined by the formula (2.2) as follows:

$$\text{score}(\text{sen}_j) = \sum_{i=1}^{m} \left(\frac{\text{score}(\text{ChainMember}_i, \text{Sen}_j)}{S_i} \right)$$

Where $\text{score}(\text{sen}_j)$ adalah is the value of a sentence j, score(ChainMember_i,Sen_j) is the value of ChainMember whose members appear in the sentence Senj, and m is the maximum number of chain formed on the article by Si is the number of sentences that contain one or all members ChainMemberi of lexical chains throughout the text.

Each sentence will be calculated weight, then do rerangking in descanding, who will then do the extraction process sentences on sentence extraction process. Where the purpose of sentence extraction is to find the words that will be extracted as a constituent of the summary results. The sentence will each be extracted from textual original article and reconstituted as a single summary. Sentence to be extracted selected based on the weight of each sentence and then sorted in order of sentence [9].

2. ISI PENELITIAN

2.1 Analysis

Systems analysis describes problems with a system and describe the requirements on system development summary articles. In any system, there are several processes that have their respective functions. Process flow for system summary article can be seen in Figure 9:

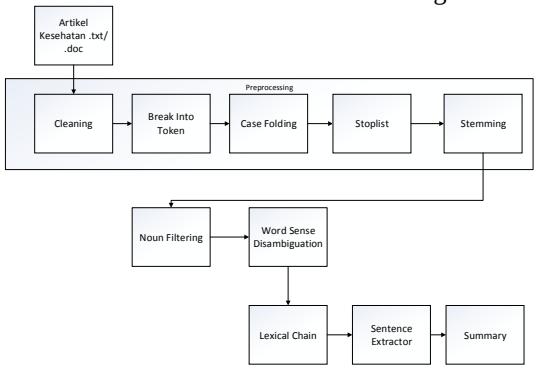


Figure 1 Process Flow

Documents such as articles of health as an input into the system prior to entry into the primary process, the feedback through the preprocessing stage first. Where the preprocessing phase consists of processes break into tokens, stoplist and stemming. The results of preprocessing are then entered into the main process in the system: noun filtering, word sense disambiguation, lexical chain and the last is the sentence extractor to produce an output in the form of an article summary (summary). For more details, here is a breakdown of the data input and each process.

1. Data Input

The input data required by the application of simulation are shaped article text with .txt or .doc format. which in this research data is taken from a collection of health articles.

2. Preprocessing

Preprocessing is the stage where before entering the main process in the system. Preprocessing done first in order to produce information that is ready for the next process. In this article summary system development preprocessing is performed as follows:

a. Cleaning

Stages of cleaning the characters that are not needed in the next process.

b. Break into token

Phase cutting input text by every word that constitute it.

c. Case folding

At this stage of folding Case This is the stage where all the letters will be converted to lowercase

d. Stoplist

Tokens that have been generated during the process Break into tokens and had to lowercase all the stages of folding case then processed again by dumping stoplist words included are the words that are considered not to have weight in the document. That is, these words are not significant in such documents like "adalah", "yang", "dari", and others. Results from stoplist disposal process that will be used for the next process is the process of stemming. Results disposal stoplist called non-stopword tokens.

e. Stemming

Non-stopword tokens from the stoplist be filtered back through the stemming process that is changing the basic constituent word being said. For example, the word "kelangsungan" converts into words "langsung", word ”merawat” converts into words”rawat”.

Examples of implementing preprocessing can be seen in Table 1.

Table 1 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 menerusk an kelangsun gan hidup setiap orang.	hati merupak an organ vital dari tubuh manusia dengan peran yang sangat penting untuk melengk api kinerja organ lain dalam menerus kan kelangsu ngan 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

In this stage of filtering noun token collection preprocessing results are then considered as an input in shaping the lexical chain. Filtering noun done by checking each word with a noun as a reference database to obtain a noun.

Examples of application noun filtering process can be seen in Table 2:

Table 2 Noun Fitering Process

No.	Preprocessing Results	Noun filtering Results
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

At this stage, identification of the meaning of words with lesk algorithm. Where lesk algorithm works by comparing the definition of the ambiguous word with its neighbors based on the definition of the word dictionary. The process of word sense disambiguation. Comparative Example ambiguous

sense of the word with the word neighbor can be seen in Table 3:

Tabel 3 Proses Word Sense Disambiguation

No.	ambiguous word	meaning of words
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

Comparative Word	Comparative meaning	value
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 makhluk insan orang	1
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

At this stage, the formation of a chain that describes the semantic relationship of words with other words and scoring with. Giving the value obtained from the formula (2.1) following:

Score(chain) = length * homogenityindex

Homogenityindex = 1 – (distinctmember/length)

Formation of lexical chains derived from the word candidates who had been elected from noun filtering process and has disambiguation on the word sense disambuguation. Lexical chain calculation results can be seen in Table 4.

Table 4 Lexical Chain

No.	Lexical Chain	Give the value	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

At this stage, scoring sentence and sentence extraction. Where the sentence scoring own goal to make the weighting of each sentence that compose an article based on the method used. Weighting phrases used in the method of lexical chain, namely in terms of the value chain of lexical been established.

Here is the formula (2.2) for weighting sentence:

$$\text{score}(\text{sen}_j) = \sum_{i=1}^{m} \left(\frac{\text{score}(\text{ChainMember}_i, \text{Sen}_j)}{S_i} \right)$$

extractor sentence calculation results can be seen in Table 5.

Table 5 Sentence Extractor

No.	Sentence	sentence scoring value
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

Sample results summary health articles, in which the analysis of the contents of the summary produced 50% of all sentences in the analysis. Thus the results of the summary amounts to three sentences:

Table 6 Article Summary Result

Article Summary Result
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

Based on research that has been done on the implementation of Lexical Chain Method with Word Sense disambiguation To Peringkasn Health Articles can be deduced as follows:

1. Metode Lexical Chain Method with Word Sense disambiguation can identify words that have ambiguity or double meaning, and can choose the meaning of a word that has ambiguity well.
2. Lexical Chain Method with Word Sense disambiguation has better accuracy than the original method of lexical chain in summary article.

REFERENCE

[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.