Improving Graph-Based Summarization with HTML Tag and Metadata Features

Dewi Wardani, Yuni Susanti, *Member, IAENG*

ABSTRACT—This research focused on automatic text summarization for web content (single-document). The method is by using the graph-based summarization method and improving it with unique features that only possessed by data object from webpages: HTML tag and metadata. Differences of given style into text, such as the using bold, underline, or italic style in typography will add some emphasizes to that portion of the document. This writing-emphasize indicates importance to the text. Additionally, metadata is "data about data". Metadata is usually made directly by a human to help indexing process by engines. Therefore, using metadata as a keyword to support this research was quite suitable because of that reason. The experimental studies have shown that the proposed method is quite promising. The scores in precision, recall and f-measure scores compared to the other techniques used by other state-of-the-art summarization systems in several kinds of data and testing.

***Index Terms*— automatic-text-summarization, graph-based summarization, HTML-tag, metadata, web-content summarization**

I. Introduction

TEXT summarization has been an important and challenging area almost over the past a decade [1] [2] and has continued to be a steady subject of research [3]. The increasing number of online information has necessitated intensive research in automatic text summarization field area, especially for the text on the web. According to existing methods, the text summarization can be categorized into two approaches; (i) extractive summarization and (ii) abstractive summarization. Extractive summaries are created by reusing portions (words, sentences, etc.) of the input text. While abstractive summaries are created by regenerating/reformulation the extracted content [4]. It sounds that the extractive summaries are easier to produce. The speed, the simplicity, the non-requirement of background knowledge and the domain independence are some of the features of extractive summarization [3].

The extractive summarization generates a summary from sentences that have the highest importance score in the document. The conventional method to determine sentence importance is vector product such as *tf*idf* (term frequency – inverse document frequency) or position weight parameters. One of the popular methods in text summarization is the graph-based summarization. This method has been proposed and successfully in its implementation [3] [5].

The trend in the text summarization shows that most of researchers have been used the graph-based summarization method to generate necessary text units from document [3] [5] [6]. Based on the techniques which were used, text summarization mainly on the extractive level [5]. Although the summary created by human usually abstractive, but sometimes extractive summarization could give a better result than abstractive ones.

This work focuses on the modification of the graph-based method for web-content summarization. Web content is composed of one kind of text called hypertext markup language (HTML) that mainly in the form of HTML tags and metadata. HTML tags have specific functions; <P> tag is used to indicate a paragraph, <H1> means heading and many others. Some tags that can be specially used to improve the result of summarization. They are used because its function that gives a unique style to the text e.g. bold, italic and underline tags. In typography, differences style that was carried into writing, such the use of bold, underline or italic style will add some emphasizes to that portion of text [7] [8] [9]. This writing-emphasize naturally indicates that the words which are tagged are important. Therefore, they can be used to support summarization when it determines the importance of the sentence. The same issue for metadata which are usually authored by human expert, therefore, the metadata can be guaranteed containing important information of webpage. They will be used to improve summarization. This is the novelty of the modified graph based summarization in this work.

We will compare our work to the results of existing tools. They are *GreatSum* [10] and *SweSum* [11]. *GreatSum* is an online summarization tool that based on the graph-based summarization. The method was used combining with Singular Value Decomposition. The other one is *SweSum* that used text-style factor (that is HTML tag bold), along with three other parameters (Position Score, Numerical Data and Keyword). *SweSum* combines all the parameters without special weighting. *SweSum* considers only to one feature, HTML tag.

The remaining sections are organized as follows: we briefly summarize the gap of research and our idea to fill the gap in Section 2. We explain our proposed approach in Section 3 and Section 4. Section 5 discusses that result of the experiment and finally the conclusion and future work in Section 6.

Manuscript received December 10th, 2018; revised November 2nd, 2019. This work was supported in part by the Department of Higher Education of Indonesia. Dewi Wardani is a researcher in Universitas Sebelas Maret, Indonesia (e-mail: dww_ok@uns.ac.id).

Yuni Susanti finished her PhD from the Computational Linguistics/ Natural Language Processing group, Computer Science Department, Tokyo Institute of Technology, Japan. Now, she is a researcher at Artificial Intelligence Lab, Fujitsu Laboratories (e-mail: susanti.yuni@fujitsu.com).

II. Related Work

Generally, to gain the score of a part of text, the summarization considers to these parameters [4];

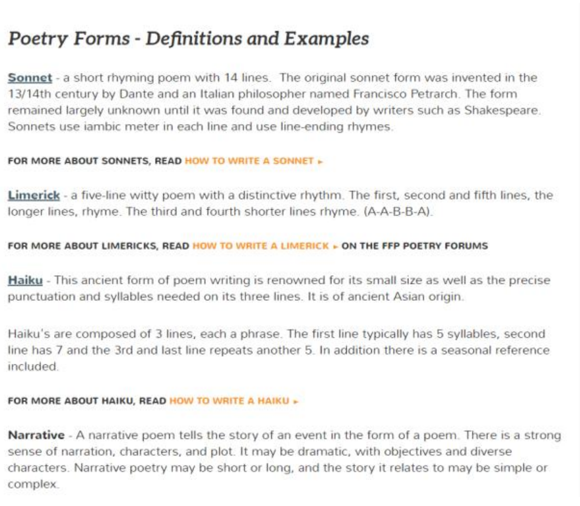
(a). Positional Criteria,
(b). Cue phrase Indicator Criteria,
(c). Word and phrase frequency criteria,
(d). Query and title overlap criteria,
(e). Cohesive or lexical connectedness criteria and
(f). Combination of various module scores.

The trend in the text summarization shows that most of the researches have been used graph-based summarization to generate necessary text parts from document [3] [5] [6]. Based on the previous methods, text summarization is mainly on the extractive level [5]. Even though the summary created by human usually abstractive, but sometimes extractive summarization could give a better result than abstractive ones.

III. Considered Problem and Idea

Most of the text summarization's researchers up to now are only using the pure content of document such as only consider the content of webpages [3] [5] [6]. As in the web summarization, it will unwrap the HTML tags to get the text content of webpages. It looks like that HTML tag is not a useful feature in summarization. Our main idea is to improve the graph-based summarization method by using the formatting or style of HTML tag and metadata of webpages. They are the additional features in this modified summarization. Figure 2 illustrates two features in webpage, HTML tag and metadata, which will be used in this approach as additional feature in the summarization process.

Some HTML tags have a function to give particular style in text, for example, bold tag (b) can provide bold appearance in the text. Font tag has several attributes that take care of font style including colour, size, etc. Concerning of typography, sometimes human-authors give



Poetry Forms - Definitions and Examples

Sonnet - a short rhyming poem with 14 lines. The original sonnet form was invented in the 13/14th century by Dante and an Italian philosopher named Francisco Petrarch. The form remained largely unknown until it was found and developed by writers such as Shakespeare. Sonnets use iambic meter in each line and use line-ending rhymes.

FOR MORE ABOUT SONNETS, READ HOW TO WRITE A SONNET ▸

Limerick - a five-line witty poem with a distinctive rhythm. The first, second and fifth lines, the longer lines, rhyme. The third and fourth shorter lines rhyme. (A-A-B-B-A).

FOR MORE ABOUT LIMERICKS, READ HOW TO WRITE A LIMERICK ▸ ON THE FFP POETRY FORUMS

Haiku - This ancient form of poem writing is renowned for its small size as well as the precise punctuation and syllables needed on its three lines. It is of ancient Asian origin.

Haiku's are composed of 3 lines, each a phrase. The first line typically has 5 syllables, second line has 7 and the 3rd and last line repeats another 5. In addition there is a seasonal reference included.

FOR MORE ABOUT HAIKU, READ HOW TO WRITE A HAIKU ▸

Narrative - A narrative poem tells the story of an event in the form of a poem. There is a strong sense of narration, characters, and plot. It may be dramatic, with objectives and diverse characters. Narrative poetry may be short or long, and the story it relates to may be simple or complex.

Fig. 1. The picture of the sample use of HTML tag in a webpage. The article is taken from "Poetry Forms - Definitions and Examples" https://www.familyfriendpoems.com/poems/other/

different style in texts to make them visible or eye-catching and to emphasize them [12]. Figure 1 shows the use of differences style in writing to give emphasize.

From the example in Figure 1, it can be seen clearly that bold style is used only in the part that mentions several genres or types of poetry. Additionally, because the article is about "Poetry Forms-Definitions and Examples", it is quite fair to conclude that these bold styles give emphasize to the text and successfully indicate the portion of the article that is important. We already have shown that several HTML tags indicate some important meaning to the portion of texts with those tags. Therefore, we use those tags to support the extractive summarization.

According to Stark [9], to emphasize in a text can be achieved with text formatting shown in Table I. In this research, we use only HTML tag bold, italic and underline because they are generally used in writing- emphasize [7] [9]. Metadata is known as a summary from an expert

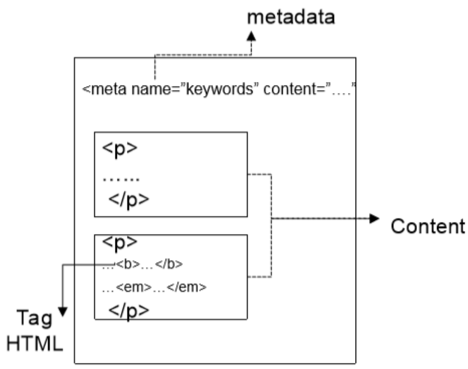


Fig. 2. The picture of an example of the metadata and tag HTML of a webpage which are considered in this approach

because human-author as the creator of the webpage write them and directly understand the main idea of the webpage. Hence, the content of metadata is a vital part instead as considered a feature to improve the summarization.

IV. Proposed Approach

Let us discuss the proposed approach, which improves the graph-based summarization method by using HTML tag and metadata and also the way to combine them.

A. *Continuous Sentence Rank (CSR)*

Graph-based summarization has been known as a promising approach for text summarization [13] [14] [15]. Continuous Sentence Rank [3] is the graph-based summarization method which we use in this research CSR is an enhancement of LexRank [5] with the addition of the discounting method and position weight into its formula. The next work of CSR can be found in managing the summarization for multi documents [16]. Graph based approach is also useful for abstractive summarization [17]. A graph based itself is useful in detecting proportion of data [18].

A document can be considered as a network of sentences that are related to each other. Erkan and Radev [5] made a hypothesis that the sentences that are similar to many other sentences in the document are more critical. They also considered the prestige side of those sentences. Assume it as a social network where a person with extensive communications or contacts with people in the organization is considered more important than a person with fewer connections. They also saw prestige side from each of those contacts; If the person has few contacts, but each of those contacts is highly placed. Thus, that person stands a chance to has some importance in the organization.

The similarity between two sentences x and y is determined by the cosine between the two-sentence vectors, modified by inverse document frequency. Although there exist several measures to evaluate the connectivity among the sentences, the cosine metric is found to be popular and superior to others [12]. *CSR* is given by the formula 1.

$$CSR(i) = \frac{d}{N} + gama * beta^{i-1} + (1-d) * \sum_{j \in S[i]} \frac{IdfModCos(i,j) * CSR(j)}{\sum_{k \in S[j]} IdfModCos(j,k)} \quad (1)$$

Where N be the total number of sentences in the document, d is damping factor while g or gama and b or beta be the parameters of position weight. Note that those parameters can be set to adjust which position needs to be given preference. This research offers choice to earlier sentences, with the consideration of the facts that most of the text document's main idea is at the beginning of the paragraph.

For the discounting method, we do not use that method in this research. The basic of this technique is when the process selects a sentence, the chance for repetition of information in the following sentence is minimized. The approach is made by immediately set corresponding row and column values of the adjacency matrix to zero. This technique is less suitable to be implemented because this research focuses on single-document summarization. In the single-document summarization, the chance for repetition of information in one document is not too big. It is different with multi-document summarization which is the repetition of information is more significant because it tends to center on more than one separate document on the same topic.

Position weight was added because the location of a sentence in a document plays a significant role in determining the importance of a sentence. This research will use position weight formula that gives preference for the beginning of sentences. With the consideration that much of the text document is deductive (the main idea/topic is at the beginning of the paragraphs, especially for news document). The formula to add a score for words which appear at the beginning of a sentence is as below:

$$P_{fi} = gama * beta^{i-1} \quad (2)$$

Gama and Beta are the parameters which the value sit between 0 and 1. Formula 2 will give more score to the sentences which appear at the beginning of the paragraph. Formula 3 will balance it, which also provide a rating in the sentence at the end of the section.

$$P_{fi} = gama * beta^{n-1} \quad (3)$$

B. HTML tag for Extractive Summarization

HTML tag score is calculated with consideration of our hypothesis that the sentences that contain more HTML tag emphasize (bold, italic, underline) in its words are more important in the document. According to the fact that those tags emphasize in the typography, as explained in section 3. *HTML Formula (HF)* is given by the formula 4.

$$HF[i] = \frac{Nht_i}{N_i} \quad (4)$$

where Nht_i refers to the number of words with HTML tag emphasize in the *i-th* sentence, while Ni is the total number of words in the *i-th* sentence. The result of formula 4 will give a score in the range 0-1. The formula four will provide a rating based on not only the number of tag HTML in a sentence but also consider the length of a sentence. As the example below,

S(i) = A big monkey appears in <u> Solo, Central Java </u>, today.

HTML tag <b> and <u> will be saved as an HTML property of the *i-th* sentence, and this process continues till each sentence in the document has *HF* score (*HF* score = 0, if there aren't any HTML tag emphasize in those sentences). Let see the other example below;

Sentence 1. **Things Fall Apart** is a book created by **Chinua Achebe.**
Sentence 2. **Chinua Achebe is Nigerian.**

With the bold formula, SweSum will return a higher score to the first sentence. It is because sentence 1 has more number of HTML tag than sentence 2 (Sentence 1 = 5; Sentence 2 = 4). Meanwhile, the proposed HTML tag formula will give a higher score to Sentence 2 (1) than Sentence 1 (0.5), because HTML tag is found in a whole sentence. Therefore, the score of Sentence 2 is maximum. It reflects that a sentence is more important if all words of it have an HTML tag than only a part of it.

C. Metadata for Extractive Summarization

Correspond to this research [19], an article's title is usually used as a query and this title method has shown better performance compared to other methods in general. This research was using special features taken from the dataset source, that is webpage metadata as the query. For this method, sentences are represented as a vector, with Boolean weighting that is given by formula 4. A vast number of unstructured data [20] makes the process of metadata will make the computation faster.

$$S_i = (w_{i1}, w_{i2} ... w_{ik}) \quad w_{ik} = \begin{cases} 1, if tf_{ik} > 0 \\ 0, otherwise \end{cases} \quad (5)$$

Where *wik* be the weight *k – th* word in *i– th* sentence, *tif* be the *TF* score *k – th* word in *i – th* sentence and S_i be the sentence vector. Then, *Metadata Formula (MF)* is defined as the similarity between sentence and metadata keywords is given by formula 6.

$$MF[i] = Similarity(S_i, TFM) = \sum_{k=1}^{n} w_{ik} w_{TFMk} \quad (6)$$

Where *MF[i]* or $Sim(S_i,TFM)$ be similarity score *i – th* with metadata keyword, *wik* is weight *k – th* word in *i– th* sentence, w_{TFMkf} be weight *k – th* word in metadata keyword. After all of the three algorithms presented above, each sentence will have *CSR*, *HF* and *MF* score. This score will be normalized first then integrate by simple linear combination to make each sentence's scores are between 0-1.

Formula 7 shows the normalization process

$$x_{normal} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad (7)$$

Where x be the original score will be normalized, x_{normal} be the normalized score, x_{min} be the minimum score in the set and x_{max} be the maximum score in the collection.
Therefore, the last score for each sentence is given by formula 8.

$$Stot[i] = CSR[i] + HF[i] + MF[i] \qquad (8)$$

V. Results and Analysis

A. Dataset and Evaluation Method

This research uses two kind of data: NIST DUC Dataset (DUC 2002) and real-webpage. Because the process needs data which is formed in webpages with HTML tag emphasize and metadata, so then authoring HTML tag and metadata is done to DUC data. Gold summary for data real-webpage was generated by human-expert judgements (English Education and English Literature students). The dataset for experimental is divided to a few types, official and non-official webpage, with and without HTML tag emphasize in writing, and a few variations of the compression ratio of summarization result (15%, 30%, 50%). Each experiment uses 50 – 60 data.

The experiments compare the result of the three evaluation scores (Precision-P, Recall-R and F-Measure-F) of the proposed method with other state-of-the-art summarization systems. Henceforth, the proposed method is called *CHMSum (CSR-HTML-Metadata Summarization)* while other systems used as the comparator is *GreatSum* and *SweSum*. Precision, Recall, and F-Measure scores were measured using the ROUGE Evaluation Toolkit [21].

B. Experimental Results

We conduct several experiments, as described in Table II. We intend to do several scenarios of experimental to make sure that *CHMSum* is a pretty solid approach. As the proposed method is a new idea; therefore, we need promising results in its beginning work.

Table 3 shows that the modification with the *CHMSum* against the official dataset website can improve performance as evidenced by the results of this test scenario, comparing to the outcome of *GreatSum*

Note: RG is the abbreviation of ROUGE

Table 4 shows that the *CHMSum* against the dataset non-official return worse scores compared to the result of *GreatSum*. We investigated that mostly non-official website (14 out of 20) do not have metadata. Therefore, we obtained MF scores are zero. Somehow, it influences the total score. Implicitly, we may conclude that if they have metadata, there is a bigger possibility to improve the scores.

Table V shows that official websites return better scores. This result makes the implicit conclusion from Table III stronger. The official website usually is written and are managed better than the non-official site. They are fully completed with HTML tags and metadata. As we know, metadata is useful as well in the searching issue. CHMSum considers calculating HTML tags and metadata into the calculation of summarization. Therefore, they improve the scores of the official-website.

Table VI shows that *CHMSum* slightly returns a better performance compare to the results of *SweSum*. Both approaches use HTML tags to improve the summarization.

Table VII shows worse scores compare to the results in Table VI in both approaches. It indicates that considering HTML tags improve the outcome of summarization. In experiment P2-B, the results show the same situation that *CHMSum* return better scores compare to *SweSum*. Table VI and Table VII also show that the calculation of metadata indeed improves the summarization.

CHMSum performs linear combinations of scores obtained from 3 algorithms (CSR, HF, and MF), so data without HTML tags (non-HTML data) will not have an HF score (HF score = 0). It shows that the effect of HF score is essential on the quality of the summary. The decrease of P, R, F value in non-HTML data, are the evidence as shown in the result of this test scenario, in Table VIII.

Table IX shows that *CHMSum* returns fair results against a different type of websites. It also concludes that *CHMSum* provides a pretty robust approach, as the results are appropriate for different kinds of website data (articles and news).

Table 10 shows that the results of *CHMSum* and *GreatSum* are also pretty fair. In a few scores, F of *GreatSum* returns slightly (below 0.001) higher scores than *CHMSum*. F-Measure (F) is a single measure that combines P and R. The F score will be high if there is a balance between the P and R values. In other words, the difference between the two scores is not too vast. The results of this study indicate that a few F score of *CHMSum* is lower than scores of *GreatSum* (on R-W and R-L). The smaller number may be due to the vast difference between R and P values in experiments against long text. The experimental results show that the R score is higher, reaching twice that of the P-value. The higher rating is because the operation of this test used a compression ratio of 50%. Thus, the extended data (> = 525 words) will give a much longer summary result than the gold- summary (average length of gold summary is ± 100 words). The vast difference in length between the summary of long text types with the gold summary resulted in the low P-value, which resulted in a slight decrease in the value of F. The short text (<= 250 words) on compression ratio 50% will produce a summary with an average length of ± 125 words. Not much different compared to the length of the gold summary (± 100 words).

Table 11 shows that the scores of *CHMSum* against short text are still slightly higher than the results of *GreatSum*. Table XII shows that *CHMSum* returns better results against short text compare to the long text. The experiment P4-C has done with a compression ratio of 50%. Hence, the long text will produce a relatively long summary result, too (depending on the length of the data). Meanwhile, the short text will provide a summary which is not much different from the gold summary. Thus, on short text, the scores are much better to compare to the scores of the long text.

GreatSum uses a graph-based (Singular Value Decomposition) algorithm, while *CHMSum* uses a graph-based algorithm (CSR) with the addition of HTML tagging and metadata (HF and MF) factors. Table XIII shows the results of the modification of the graph-based algorithm proved to improve performance at 15% compression ratio. Table XIV and Table XV show similar results like in Table XIII for the compression ratio of 30% and 50%.

The results of P8, in Table XVI shows that the experiment of *CHMSum* with the weight variation favoring HF (CSR: HF: MF = 20%: 50%: 30%) gives the best result compare to the other three weight variations. *CHMSum* uses a linear combination of the three algorithms, with a general implementation being no special weighting on each algorithm used. The results of this experiment give the lowest value on the weight variation favoring CSR. The following by the variety no favoring, then variation favoring MF and the highest is variation favoring HF. The weighting, variety favoring MF, where the HTML tag factor has given the weight more elevated than the other two algorithms. This experiment also shows that HF and MF scores can improve the quality of the result of the summary.

In overall, the results of the experiments have shown that *CHMSum* is a promising approach to improve the quality of summarization on single documents.

VI. Conclusions and Future Work

The modification of the graph-based summarization method with the addition of HTML tags and metadata features has been well-implemented and returns a better result. In general, compare to other methods used in other state-of-the-art summarization systems. HTML tag and metadata features are proven could increase precision, recall and f-measure scores which overall increased 0.05 – 0.1 on several kinds of data and testing. For future work, it can be considered to learn using this novel idea for multipage summarization.

TABLES

TABLE I
METHODS IN WRITING

Method	Usage
Italics	It gives a light emphasize. It uses for stressing of words.
Underline	It is more insistent. It works well for emphasizing a complete phrase.
Bold	It can be clearly visible when first look at the whole page. It creates tension as we read it.
Size	The bigger font stand out more and small fonts recede.
Uppercase	It is the written which equivalent of shouting. Avoiding it where possible, including use in headings.
Color	The brighter colors and the red color stands out more than others. The blue color and dark colors are more subtle.
Combinations	The combination of any of the above methods.

TABLE II
THE EXPERIMENTAL SCENARIO

Code	Description
P1	Type of Website P1-A. Official Website, CHMSum vs *GreatSum* P1-B. Non-Official Website, CHMSum vs *GreatSum* PI-C. Data Official vs Data Non-Official by implementing CHMSum
P2	Variety of Dataset – HTML tags P2-A. With HTML tags, CHMSum vs *SweSum* P2-B. Without HTML tags, CHMSum vs *SweSum* P2-C. With HTML tags vs without HTML tags by implementing CHMSum
P3	Variety of Dataset – topic of article by implementing CHMSum
P4	Variety of Dataset – the length of articles P4-A. Long article, CHMSum vs *GreatSum* P4-B. Short article, CHMSum vs *GreatSum* P4-C. Long article vs Short article by implementing CHMSum
P5	Compression Ratio – 15% CHMSum vs *GreatSum*
P6	Compression Ratio – 30% CHMSum vs *GreatSum*
P7	Compression Ratio – 50% CHMSum vs *GreatSum*
P8	Variation of weight by implementing CHMSum

TABLE III
THE EXPERIMENTAL RESULT OF P1-A

RG	**CHMSum**			**GreatSum**		
	P	R	F	P	R	F
R-1	0.3324	0.3575	0.3314	0.3864	0.1874	0.2354
R-2	0.2090	0.2258	0.2107	0.2628	0.1291	0.1602
R-L	0.3263	0.3502	0.3251	0.3741	0.1823	0.2285
R-W	0.2316	0.1485	0.1735	0.3114	0.0897	0.1321

TABLE IV
THE EXPERIMENTAL RESULT OF P1-B

RG	**CHMSum**			**GreatSum**		
	P	R	F	P	R	F
R-1	0.3324	0.3575	0.3314	0.3864	0.1874	0.2354
R-2	0.2090	0.2258	0.2107	0.2628	0.1291	0.1602
R-L	0.3263	0.3502	0.3251	0.3741	0.1823	0.2285
R-W	0.2316	0.1485	0.1735	0.3114	0.0897	0.1321

TABLE V
THE EXPERIMENTAL RESULT OF P1-C

RG	**Official Website**			**Non-Official**		
	P	R	F	P	R	F
R-1	0.4891	0.6645	0.5415	0.3324	0.3575	0.3314
R-2	0.3210	0.4364	0.3570	0.2090	0.2258	0.2107
R-L	0.4741	0.6435	0.5246	0.3263	0.3502	0.3251
R-W	0.3301	0.2965	0.2994	0.2316	0.1485	0.1735

TABLE VI
THE EXPERIMENTAL RESULT OF P2-A

RG	**CHMSum**			**SweSum**		
	P	R	F	P	R	F
R-1	0.4935	0.6442	0.5372	0.4679	0.5403	0.4549
R-2	0.3363	0.4411	0.3685	0.3236	0.3681	0.3114
R-L	0.4830	0.6300	0.5256	0.4615	0.5294	0.4473
R-W	0.3318	0.2875	0.2950	0.3289	0.2368	0.2481

TABLE VII
THE EXPERIMENTAL RESULT OF P2-B

RG	CHMSum			SweSum		
	P	R	F	P	R	F
R-1	0.4600	0.3913	0.3768	0.4168	0.3192	0.2883
R-2	0.2396	0.2064	0.1983	0.2095	0.2027	0.1694
R-L	0.4456	0.3753	0.3623	0.4084	0.3102	0.2808
R-W	0.2995	0.1659	0.1905	0.3187	0.1435	0.1609

TABLE VIII
THE EXPERIMENTAL RESULT OF P2-C

RG	With HTML tags			Without HTML tags		
	P	R	F	P	R	F
R-1	0.4935	0.6442	0.5372	0.4600	0.3913	0.3768
R-2	0.3363	0.4411	0.3685	0.2396	0.2064	0.1983
R-L	0.4830	0.6300	0.5256	0.4456	0.3753	0.3623
R-W	0.3318	0.2875	0.2950	0.2995	0.1659	0.1905

TABLE IX
THE EXPERIMENTAL RESULT OF P3

RG	News			Articles		
	P	R	F	P	R	F
R-1	0.4717	0.6746	0.5462	0.5364	0.6019	0.5544
R-2	0.2857	0.4112	0.3315	0.3901	0.4377	0.4031
R-L	0.4521	0.6476	0.5238	0.5264	0.5909	0.5440
R-W	0.3220	0.3032	0.3071	0.3489	0.2510	0.2847

TABLE X
THE EXPERIMENTAL RESULT OF P4-A

RG	CHMSum			GreatSum		
	P	R	F	P	R	F
R-1	0.2497	0.6518	0.3448	0.2445	0.6249	0.3398
R-2	0.1417	0.3486	0.1904	0.1397	0.3471	0.1914
R-L	0.2436	0.6327	0.3357	0.2400	0.6125	0.3333
R-W	0.1628	0.2818	0.1954	0.1624	0.2753	0.1962

TABLE XI
THE EXPERIMENTAL RESULT OF P4-B

RG	CHMSum			GreatSum		
	P	R	F	P	R	F
R-1	0.4642	0.6026	0.5203	0.3961	0.4602	0.4128
R-2	0.2818	0.3633	0.3151	0.2225	0.2599	0.2306
R-L	0.4439	0.5769	0.4978	0.3774	0.4425	0.3958
R-W	0.3222	0.2776	0.2957	0.2735	0.2121	0.2318

TABLE XII
THE EXPERIMENTAL RESULT OF P4-C

RG	Short text (<= 250 words)			Long text (> = 525 words)		
	P	R	F	P	R	F
R-1	0.4642	0.6026	0.5203	0.2497	0.6518	0.3448
R-2	0.2818	0.3633	0.3151	0.1417	0.3486	0.1904
R-L	0.4439	0.5769	0.4978	0.2436	0.6327	0.3357
R-W	0.3222	0.2776	0.2957	0.1628	0.2818	0.1954

TABLE XIII
THE EXPERIMENTAL RESULT OF P5

RG	CHMSum			GreatSum		
	P	R	F	P	R	F
R-1	0.5782	0.3032	0.3737	0.4649	0.2786	0.3358
R-2	0.3673	0.1792	0.2257	0.2439	0.1460	0.1759
R-L	0.5459	0.2862	0.3526	0.4468	0.2676	0.3226
R-W	0.3960	0.1361	0.1917	0.3106	0.1233	0.1705

TABLE XIV
THE EXPERIMENTAL RESULT OF P6

RG	CHMSum			GreatSum		
	P	R	F	P	R	F
R-1	0.4769	0.4713	0.4517	0.4176	0.3984	0.3906
R-2	0.2918	0.2737	0.2693	0.2268	0.2186	0.2130
R-L	0.4540	0.4487	0.4300	0.4031	0.3865	0.3784
R-W	0.3139	0.2038	0.2351	0.2824	0.1765	0.2080

TABLE XV
THE EXPERIMENTAL RESULT OF P7

RG	CHMSum			GreatSum		
	P	R	F	P	R	F
R-1	0.4067	0.6284	0.4755	0.3579	0.5267	0.4092
R-2	0.2530	0.3773	0.2918	0.1946	0.2893	0.2237
R-L	0.3928	0.6069	0.4592	0.3447	0.5093	0.3950
R-W	0.2731	0.2764	0.2634	0.2396	0.2323	0.2258

TABLE XVI
THE EXPERIMENTAL RESULT OF P8

RG	Variation no favoring			Variation favoring CSR		
	P	R	F	P	R	F
R-1	0.44933	0.46806	0.43736	0.42628	0.40628	0.39292
R-2	0.26192	0.25842	0.24842	0.23665	0.21028	0.21071
R-L	0.42932	0.44706	0.41778	0.40936	0.38888	0.37652
R-W	0.30026	0.20784	0.23395	0.28645	0.18090	0.20933

	Variation favoring HF			Variation favoring MF		
	P	R	F	P	R	F
R-1	0.44215	0.49577	0.44199	0.43392	0.49248	0.43751
R-2	0.25845	0.27787	0.25302	0.25457	0.27654	0.25118
R-L	0.42573	0.47675	0.42511	0.41792	0.47415	0.42108
R-W	0.29793	0.22222	0.24016	0.29431	0.22167	0.23948

REFERENCES

[1] H. P. Luhn, “The automatic creation of literature abstracts,” IBM J. Res. Dev., vol. 2, no. 2, pp. 159–165, 1958.

[2] H. P. Edmundson, “New methods in automatic extracting,” J. ACM JACM, vol. 16, no. 2, pp. 264–285, 1969.

[3] S. Hariharan and R. Srinivasan, “Enhancements to Graph based methods for Single Document Summarization,” Int. J. Eng. Technol., vol. 2, no. 1, 2010.

[4] E. Hovy and D. Marcu, “Automated text summarization,” Oxf. Handb. Comput. Linguist., pp. 583–598, 1998.

[5] G. Erkan and D. R. Radev, “LexRank: Graph-based lexical centrality as salience in text summarization,” J Artif Intell Res JAIR, vol. 22, pp. 457–479, 2004.

[6] M. Litvak and M. Last, “Graph-based keyword extraction for single-document summarization,” in Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, 2008, pp. 17–24.

[7] J. M. Levy, Take Command of Your Writing. Firebelle Production, 1998.

[8] D. Middleton, “Students struggle for words: Business schools put more emphasis on writing amid employer complaints,” Wall Str. J., p. B8, 2011.

[9] D. Stark, “Emphasis in Writing,” Changing Minds, 2010. .

[10] G. Team, GreatSummary 2008. 2008.

[11] H. Dalianis, SweSum-A Text Summarizer for Swedish (2000). Report No TRITA-NA-P0015, IPLab-174. Stockholm: Nada, KTH. Department of Computer and Systems Sciences. http://www. dsv. su. se/ hercules/papers/Textsumsummary. html.

[12] N. Allen, Working with words and images: New steps in an old dance. Greenwood Publishing Group Inc., 2002.

[13] M. Gambhir and V. Gupta, “Recent automatic text summarization techniques: a survey,” Artif. Intell. Rev., vol. 47, no. 1, pp. 1–66, 2017.

[14] H. Chen, H. Jin, and F. Zhao, “PSG: a two-layer graph model for document summarization,” Front. Comput. Sci., vol. 8, no. 1, pp. 119–130, 2014.

[15] N. Hassanlou, M. Shoaran, and A. Thomo, “Probabilistic graph summarization,” in International Conference on Web-Age Information Management, 2013, pp. 545–556.

[16] S. Hariharan, T. Ramkumar, and R. Srinivasan, “Enhanced graph based approach for multi document summarization.,” Int Arab J Inf Technol, vol. 10, no. 4, pp. 334–341, 2013.

[17] J. Tan, X. Wan, and J. Xiao, “Abstractive document summarization with a graph-based attentional neural model,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1171–1181.

[18] Nobuo Funabiki, Tana, Khin Khin Zaw, Nobuya Ishihara, and Wen-Chung Kao, "A Graph-based Blank Element Selection Algorithm for Fill-in-Blank Problems in Java Programming Learning Assistant System," IAENG International Journal of Computer Science, vol. 44, no.2, pp247-260, 2017

[19] S. Hariharan and R. Srinivasan, “A Comparison of Similarity Measures for Text Documents,” J. Inf. Knowl. Manag., vol. 7, no. 1, p. 1, 2008.

[20] Carlos Costa, and Maribel Yasmina Santos, "Big Data: State-of-the-art Concepts, Techniques, Technologies, Modeling Approaches and Research Challenges," IAENG International Journal of Computer Science, vol. 44, no.3, pp285-301, 2017

[21] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Proceedings of the workshop on text summarization branches out (WAS 2004), 2004, vol. 16.
.

Dewi Wardani is a member of IAENG since 2009. She received her doctoral degree from Johannes Kepler University in December 2016. She finished her master degree at National Cheng Kung University, ROC Taiwan. Now, she is a lecturer and a researcher at Informatics Department, Universitas Sebelas Maret, Surakarta, Central Java, Indonesia.

Her research interest is information system especially in knowledge management with semantic web approach. Her research area is included all aspects of it. Either fundamental layer, the middle layer or upper layer - use cases. She also has interest in language computation which supports the semantics issue in knowledge management.

Dr.techn. Dewi W. Wardani also is a member of IEEE, APTIKOM (the association of computer scientist in Indonesia). She is active as a presenter or a reviewer at several international conferences in Indonesia. She also has been being involved in several projects with some collaboration with the other researchers.

Yuni Susanti is a PhD student at Tokyo Institute of Technology, Japan. She is the third year of PhD student at the Computational Linguistics/ Natural Language Processing group, Computer Science Department, Tokyo Institute of Technology, Japan, under the supervision of Prof. Takenobu Tokunaga. She received her master degree from the same university. Her research interest is in the field of computational linguistics/ natural language processing.