Multi-document Summarization by using TextRank and Maximal Marginal Relevance for Text in Bahasa Indonesia

Dani Gunawan, Siti Hazizah Harahap, Romi Fadillah Rahmat
Department of Information Technology, Universitas Sumatera Utara, Indonesia
danigunawan@usu.ac.id

***Abstract*—The text summarizer reduces unnecessary information by selecting the important sentences. In multi-document summarization, there is a possibility that two or more important sentences share similar information. Including those sentences to the summary result will cause redundant information. This research aims to reduce similar sentences from multi-document that share similar information to obtain a more concise text summary. In order to accomplish the objective, this research uses the combination of several online news articles, divided into six groups. The combined articles are pre-processed to produce a clean text. After obtaining the clean text, this research utilizes the TextRank algorithm to extract the important sentences by using the similarity measurement. This process yields the summarized text. However, the summarized text is still containing similar sentences. The next process is calculating Maximal Marginal Relevance (MMR) to reduce similar sentences. The result of this process is the final text summary. The evaluation uses ROUGE-1 and ROUGE-2 with the average F-score is 0.5103 and 0.4257, respectively.**

***Index Terms*—multi-document text summarization, text summarization, textrank, maximal marginal relevance**

I. Introduction

Online news is a part of digital life. Today, there are a ton of online news publishers provide various kind of news. Some topics receive very hot attention, but some other topics do not. There is an issue in the hot topics articles. Some news publishers only write an article based on previously published news and only add a little part of the new idea to it. The advantage for the publishers is that they have a piece of new news to be published.

On the other hand, Internet users have to read the same news repeatedly. The text summarization is an alternative to overcome this issue. The text summarization works by extracting the important sentences [1]. There are two methods of text summarization, namely extractive and abstractive [2]. Both methods extract important sentences from the article. The extractive method combines all the important sentence to be a summary result. Conversely, the abstractive method rebuilds new sentences from all the extracted sentences.

Previous research performed automatic text summarization by using TextTeaser [3], which was applied to single documents only. They utilize TextRank summarization result as the benchmark despite using human-generated text summary. They argue that TextRank can be implemented in any language because the determination of the important sentences does not depend on the meaning of the words. The previous study has shown that the TextRank yields a good result on text summarization [4].

Several studies discuss multi-document summarization. For example, research by Yong-dong [5] used Hierarchical Topic to create multi-document summaries. The hierarchical topic represents the whole picture of the text. Another research summarizes multi-document by using Query-focused summarization using hypergraph-based ranking [6]. It uses input queries as a benchmark to summarize the text. Meanwhile, another research [7] used the combination of Wordnet-based as abstractive summarization and title word extraction based as extractive summarization for multi-document summarization.

Research about multi-documents text summarization for news articles in Bahasa Indonesia considers sentence structure information, (subject, predicate, object or complement) [8]. They propose several processes such as extracting important information using dependency tree, sentence clustering by using DBSCAN, fusing sentences, select the sentences using Maximal Marginal Relevance (MMR). The results show that there are many rooms of improvement for the multi-documents summarization.

The research about multi-document summarization focuses on reducing redundant sentences. In order to provide a solution, we propose a combination of TextRank and Maximal Marginal Relevance (MMR). The TextRank is expected to find the important sentences, and the Maximal Marginal Relevance is expected to reduce the redundant sentences.

II. Automatic Text Summarization

One of the automatic text summarization that uses extraction method is the TextRank algorithm. It works by assigning the rank to the graph [4]. The process in TextRank algorithm represents document content into the graph. Every single sentence from the pre-processing stage will be turned into the vertex. The number 1-7 in Fig. 1 are the vertices which represent sentences in a text. The similarity of each pair of sentences will be the edge. The edges are illustrated as a line between two vertices as shown in Fig. 1. Vertex 7 has no line that connects to another vertex. This means that vertex 7 has no similar word with the other vertices.

The similarity measurement yields the score that represents the relation between two sentences. The equation 1 is the

978-1-7281-4880-9/19/$31.00

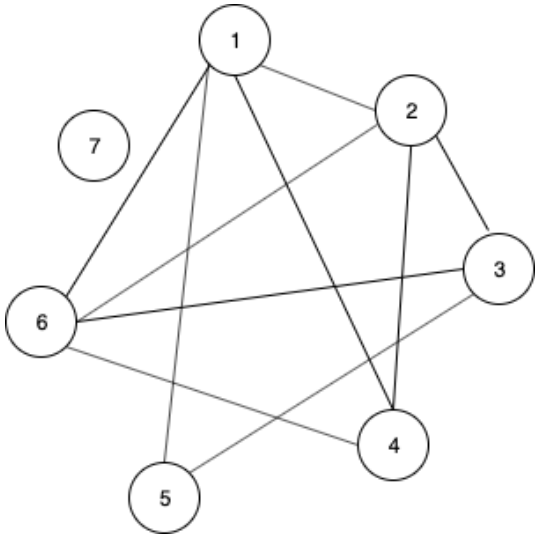

Fig. 1. Vertices and Edges

formula to calculate the similarity measurement. W_k is the number of overlapped words between two sentences, meanwhile $log(|S_x|) + log(|S_y|)$ represents the length of the words in the vertex.

$$Similarity(S_x, S_y) = \frac{|\{W_k | W_k \in S_x \& W_k \in S_y\}|}{log(|S_x|) + log(|S_y|)} \quad (1)$$

The TextRank algorithm has a particular calculation for graph weighting, as shown in equation 2. The d variable is a damping factor which has a value between 0 and 1. $In(V_i)$ means the set of vertices that points to V_i (predecessors). Meanwhile, $Out(V_i)$ means the set of vertices, which is pointed out by vertex V_i (successors).

$$S_i = (1 - d) + d * \sum_{j \epsilon In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \quad (2)$$

According to previous research [3], the TextRank algorithm has several excellences. It does not require any data training to summarize a text (unsupervised). Also, the TextRank algorithm is a language-independent which does not require grammar understanding to summarize the text. The TextRank works by utilizing the available text in the corpus.

III. Research Methodology

A. Data

This research focused on summarizing online news articles. We collected the article from two famous online news publishers according to Alexa Rank (www.alexa.com), namely Tribunnews (www.tribunnews.com) and Detik (detik.com). The articles are categorized to general news, economics, entertainment, criminals, sports, and politics. We manually selected similar articles from the same topic published in the different sources. The number of articles after the selection process is thirty. These articles are divided into six categories. Each category consists of five articles. Each article consists of 200 to 700 words.

B. Pre-processing

Multi-document summarization works by combining the news article digest of several similar articles in the same topic. We obtained the article digests from combined articles by selecting the important sentences of each document. These important sentences will form a summarized text.

The multi-document summarization process requires the cleaned texts. The article in an online newspaper usually consists of several items besides the main text. Fig. 2 shows a snippet text with an advertisement. In order to obtain a cleaned text, we manually remove unrelated contents such as advertisements, links, and images for each article. As illustrated in Fig. 3, the input is the combination of more than one similar articles. In this case, the combined text consists of more than one cleaned text. The text combination aims to facilitate the process of calculating the sentences similarity for the future process.

Next, we prepared the combined text by applying some tasks. First, we remove the punctuation, such as "!" (exclamation), "?" (question mark), "." (period), "()" (brackets), "{}" (curly brackets). Second, we remove the token that has less significant meaning, such as "I", "you", "are", "is", "am", and so on. We use common stopword list to remove the token. This process is called stopword removal. This process will reduce the dimension of the combined text. Processing the smaller dimension will accelerate the time to produce a summarized text.

After removing the less significant token from the combined text, we will convert all the token to its root word. This process is necessary in order to find similar words for each sentence. We calculate the similarity measurement for the sentences based on the same words in sentences. In order to obtain

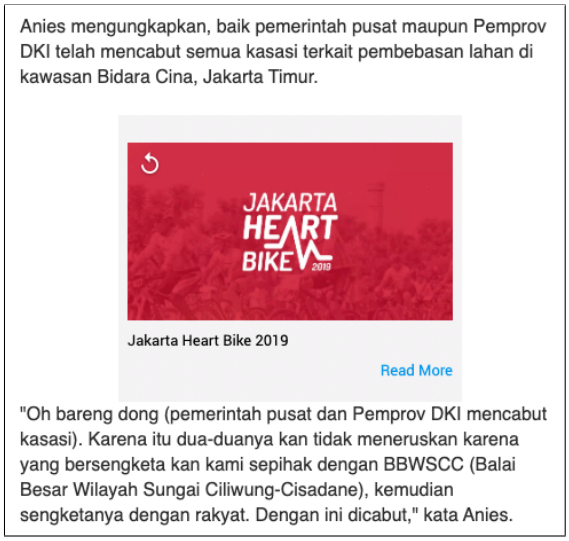

Fig. 2. Text with an advertisement

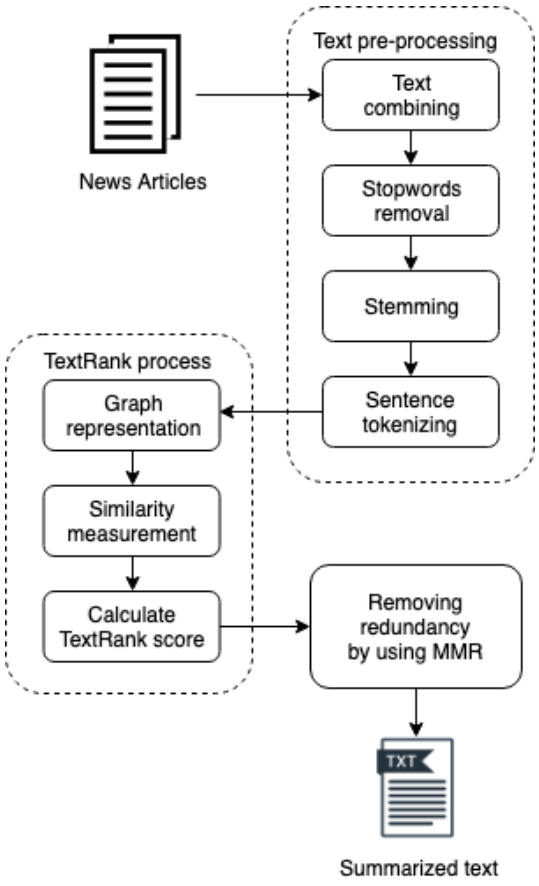

Fig. 3. General Architecture

the root words of each sentence, we perform the stemming algorithm.

The pre-process continues with the sentence tokenization process. It has similar behavior with regular tokenization, which generally yields words as the output. However, sentence tokenization extracts paragraphs into sentences instead of words. The process to yield the sentence tokens is by splitting the sentences in a paragraph based on punctuation such as period, exclamation, or question mark. Fig. 4 shows the snippet text before pre-processing and the table I.

C. Text Summarization Process

This research applies extractive method to summarize the text. It means, we extract the important sentences and then combine them into a summarized text. In order to yield the expected result, we apply several tasks, such as term weighting, sorting the sentences by the document similarities, select the most related sentences into being included in the summarization result, and the final task is eliminating redundant sentences in the summarized text. The result of these tasks is the final, text summary. We implement the TextRank algorithm to obtain multi-document summarization. This algorithm is based on graph ranking model for text processing. The text summarization process begins by calculating the weight of the sentence tokens from the previous pre-processing stage. The calculation is divided into three sections, such as establishing graph representation, sentence similarity calculation, and TextRank score calculation.

Jakarta - Pembantaian 292 ekor buaya di Sorong menjadi sorotan dunia. Media massa dari sejumlah negara ikut memberitakan pembantaian yang dipicu oleh tewasnya seorang warga yang dimakan buaya. Majalah TIME menulis judul 'Hundreds of Crocodiles Slaughtered in Retaliation for Attack on a Villager in Indonesia' pada artikelnya soal insiden ini.

Fig. 4. The example of text before pre-processing

TABLE I
THE EXAMPLE OF TEXT AFTER PRE-PROCESSING

No.	Sentence Example
1.	Jakarta bantai 292 ekor buaya Sorong menjadi sorot dunia
2.	Media massa jumlah negara ikut berita bantai picu tewas orang warga makan buaya
3.	Majalah TIME tulis judul 'Hundreds of Crocodiles Slaughtered in Retaliation for Attack on a Villager in Indonesia' artikel soal insiden

1) Graph Representation: The graph representation is the process to obtain the relation among the words in sentences. As shown in Fig. 5, each vertex represents the sentence token in the Table I. The edge that connects two vertexes represents the similarity between two sentences. The higher the edge value means the similarity between two sentences is strong. Sentence 1 and 2 share similar words, meanwhile sentence 3 has no similar word to sentence 1 and 2.

2) Similarity Measurement: According to formula (1), the similarity measurement between sentence 1 and 2 are calculated as follow.

$$Similarity(S_1, S_2) = \frac{2}{9+13} = \frac{2}{22} = 0.091$$

3) Calculating the TextRank Score: The TextRank score is obtained by calculating the similarity score between two vertexes in all sentences. The next process is calculating the

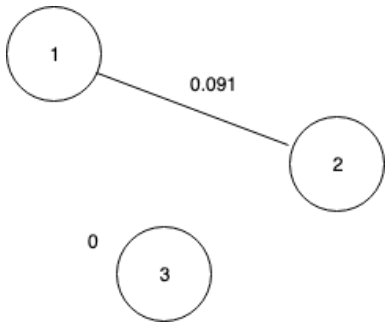

Fig. 5. Graph Representation

final score by sum up all the edges value for each vertex. We will select the vertexes which have a high score as the summary result candidates.

4) Eliminating Redundant Sentences: Maximum Marginal Relevance (MMR) is a method that uses an extraction technique to summarize a document. This method combines cosine similarity of sentences with other sentences that have been selected as summary result candidates from the previous stage. The aim of this method is minimizing the sentences that have a similar meaning in the summary results candidates. Relevant summaries can be obtained by measuring the relevance of information in sentences and queries [9]. MMR value for each sentence s_i can be obtained by using equation 3.

$$MMR_i = arg\ max[\lambda\ sim_1(D_i, Q) - (l - \lambda)\ max\ sim_2(D_i, Dj)] \quad (3)$$

As shown in equation 3, l is a parameter with the interval [0-1] to set the relative importance level between relevance and redundancy. sim_1 is a measurement of the sentence similarity with queries, while sim_2 is a measurement of the sentence similarity with other sentences that have been selected in the summary. The higher value means higher accuracy, meanwhile the lower value means higher diversity. We select the highest MMR value from the summary result candidates to be included in the final summary result.

D. Evaluation

The commonly used evaluation method for validating the result of automatic text summarization in Bahasa Indonesia is by comparing the system generated text summary and human-generated text summary. This method is considered as the gold standard because the text is generally summarized by the expert. The human-generated text summary is required to build a valid dataset for automatic text summarization evaluation benchmark [10]. Therefore, we assign a volunteer to build the human-generated text summary. The volunteer has expertise in the field of journalism.

This research applies Recall-Oriented Understudy of Gisting Evaluation (ROUGE) method to evaluate the system generated text summary [11]. The measurement used in this research is ROUGE-N (N-gram). This measurement consists of ROUGE-1 and ROUGE-2. The calculation of ROUGE-1 is done by comparing the system generated text summary and human-generated text summary word by word. Meanwhile, ROUGE-2 compares all the system generated text summary and human-generated text summary.

IV. Result and Discussion

The multi-document summarization has a problem regarding the sentence similarity among the combined texts. In this research, we are going to observe the combination of TextRank + MMR to reduce the sentence similarity in the summarized text result. As a result, the TextRank + MMR is successfully reduce the similar sentences in a summarized text. For example, the Fig. 6a show the snippet of the summarized text by using the TextRank algorithm. It shows that both sentences are similar. To reduce the similar sentences, we apply additional step, which is Maximal Marginal Relevance (MMR). The Fig. 6b shows the summarized text by using the TextRank + MMR. It shows that one of the similar sentences has been reduced.

This research compares the performance of TextRank and TextRank + MMR to summarize the provided multi-document. The Fig. 7 shows the comparison of the summarized text size among TextRank, TextRank + MMR, and the expert. The TextRank method produces an average of 28.3% text summaries from the original text. Meanwhile, the TextRank + MMR method yields 4.2% more concise text summaries. It produces an average of 24.1% text summaries from the original text. The TextRank + MMR yields more concise result because it reduces sentences that are considered redundant in summary. On the other hand, the expert produces an average of 52.1% more concise text summaries from the original text.

As mentioned in section III-D, the automatic text summarization evaluation is performed by using Recall-Oriented Understanding for Gisting Evaluation (ROUGE) method. ROUGE perform the evaluation method by comparing the text summary produced by the TextRank and TextRank + MMR with the human-generated text summary. The human-generated text summary by the expert is considered as the ideal summary (gold standard). This research utilizes ROUGE-1 and ROUGE-2. The ROUGE method produces the F-score value from the text summary. The F-score is the harmonic mean of the recall and precision. It is expected to produce a reasonable evaluation result. The F-score value lies between 0 to 1. If the F-score value closes to 0, then the system generated text summary is not similar to the human-generated text summary. On the other hand, if the F-score value closes to 1, then the system generated text summary is identical with the human-generated text summary. Hence, the higher value of the F-score, then the text summarization result is considered excellent.

Laporan HRW : 1.000 Lebih Rumah Kaum Rohingya Dimusnahkan Bukti-bukti citra satelit bahwa lebih dari 1.000 rumah di desa kelompok minoritas Rohingya telah dimusnahkan di barat laut Myanmar telah dikumpulkan oleh Human Rights Watch HRW.
Human Rights Watch menyatakan pihaknya memiliki bukti-bukti citra satelit bahwa lebih dari 1.000 rumah di desa kelompok minoritas Rohingya telah dimusnahkan di barat-laut Myanmar, yang dulu dikenal sebagai Burma.

(a) Original TextRank Summarization Result

Laporan HRW : 1.000 Lebih Rumah Kaum Rohingya Dimusnahkan Bukti-bukti citra satelit bahwa lebih dari 1.000 rumah di desa kelompok minoritas Rohingya telah dimusnahkan di barat laut Myanmar telah dikumpulkan oleh Human Rights Watch HRW.

(b) TextRank + MMR Summarization Result

Fig. 6. Summarization Result

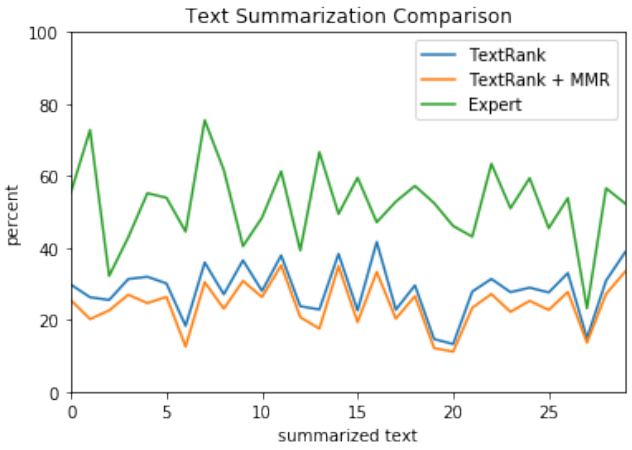

Fig. 7. The comparison between several methods

TABLE II
ROUGE-1 AND ROUGE-2 OF EACH CATEGORY

Category	ROUGE-1	ROUGE-2	Average	
			ROUGE-1	ROUGE-2
General	0.5679 0.65363 0.49649 0.53097 0.40527	0.42308 0.5814 0.42233 0.4594 0.35739	0.530852	0.44872
Economic	0.53305 0.34125 0.67368 0.53333 0.47248	0.45982 0.2381 0.59259 0.48485 0.31829	0.510758	0.41873
Entertainment	0.56 0.35333 0.4492 0.52291 0.4573	0.4964 0.23468 0.34286 0.40113 0.34783	0.468548	0.36458
Criminal	0.45476 0.37369 0.46008 0.64686 0.5	0.27451 0.3748 0.42742 0.52083 0.43459	0.487078	0.40643
Sport	0.47685 0.31481 0.54852 0.52958 0.56604	0.2871 0.17705 0.49667 0.40483 0.47264	0.48716	0.367658
Politic	0.55635 0.39196 0.45125 0.58911 0.69003	0.39196 0.48048 0.27647 0.52332 0.60606	0.53574	0.455658

Table II shows that the lowest F-score average lies in the criminal and entertainment categories. We examine the criminal category has an initial writing issue. There are many initials which consist of one letter that is not included in the system generated text summary. However, the expert reckons that the initials should be included in the text summarization. Meanwhile, the entertainment category has a different issue. The articles in the entertainment category have some irregular words such as "wkwkwkwk", "yay" and "ulala". These issues cause the average F-score for both categories are low. This research obtains that the average F-score for ROUGE-1 is 0.5103, and ROUGE-2 is 0.4257.

V. CONCLUSION

According to the previous research, the TextRank algorithm has shown a good result to summarize the text. The author claims that this algorithm does not depend on one specific language. This research utilizes the TextRank algorithm to perform multi-document summarization. We found that there are similar sentences included in the text summary result because the TextRank algorithm extracts the important sentences by using similarity measurement. Therefore, we perform one more process to reduce similar sentences by using Maximal Marginal Relevance (MMR). This process reduces similar sentences and produces a more concise text summary. The Maximal Marginal Relevance (MMR) shows a good performance. We do not find the redundant sentences in the text summary. Besides, we also found that some irregular words and initials might cause low accuracy. The expert considers that the initials are important. Meanwhile, according to the similarity measurement in the TexRank algorithm, they are not important.

REFERENCES

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*, 2nd ed. Prentice-Hall, Inc., 2009.

[2] V. Gupta and G. Singh Lehal, "A Survey of Text Summarization Extractive Techniques," 2010.

[3] D. Gunawan, A. Pasaribu, R. F. Rahmat, and R. Budiarto, "Automatic Text Summarization for Indonesian Language Using TextTeaser," *IOP Conference Series: Materials Science and Engineering*, vol. 190, no. 1, p. 012048, apr 2017.

[4] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004. [Online]. Available: https://aclanthology.coli.uni-saarland.de/papers/W04-3252/w04-3252

[5] Xu Yong-dong, Wang Xiao-long, Liu Tao, and Xu Zhi-ming, "Multi-document summarization based on rhetorical structure: Sentence extraction and evaluation," in *2007 IEEE International Conference on Systems, Man and Cybernetics*, Oct 2007, pp. 3034–3039.

[6] S. Xiong and D. Ji, "Query-focused multi-document summarization using hypergraph-based ranking," *Information Processing & Management*, vol. 52, no. 4, pp. 670 – 681, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457315001491

[7] G. Yapinus, A. Erwin, M. Galinium, and W. Muliady, "Automatic multi-document summarization for indonesian documents using hybrid abstractive-extractive summarization technique," in *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Oct 2014, pp. 1–5.

[8] R. Reztaputra and M. L. Khodra, "Sentence structure-based summarization for indonesian news articles," in *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*, Aug 2017, pp. 1–6.

[9] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 335–336. [Online]. Available: http://doi.acm.org/10.1145/290941.291025

[10] D. Gunawan and A. Amalia, "Review of the recent research on automatic text summarization in bahasa indonesia," in *2018 Third International Conference on Informatics and Computing (ICIC)*, Oct 2018, pp. 1–6.

[11] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013