Improvement of Cluster Importance Algorithm with Sentence Position for News Summarization

Nur Hayatin[1]
Department of Informatics Engineering
University of Muhammadiyah Malang
Malang, Indonesia
noorhayatin@umm.ac.id

Gita Indah Marthasari[2]
Department of Informatics Engineering
University of Muhammadiyah Malang
Malang, Indonesia
gita@umm.ac.id

Syadza Anggraini[3]
Department of Informatics Engineering
University Muhammadiyah Malang
Malang, Indonesia
sasaanggraini.sa@gmail.com

***Abstract*—Text summarization is one of the ways to reduce large document dimension to obtain important information from the document. News is one of information which usually has several sub-topics from a topic. In order to get the main information from a topic as fast as possible, multi-document summarization is the solution, but sometimes it can create redundancy. In this study, we used cluster importance algorithm by considering sentence position to overcome the redundancy. Stages of cluster importance algorithm are sentence clustering, cluster ordering, and selection of sentence representative which will be explained in the subsections below. The contribution of this research was to add the position of sentence in the selection phase of representative sentence. For evaluation, we used 30 topics of Indonesian news tested by using ROUGE-1, there were 2 news topics that had different ROUGE-1 score between using cluster importance algorithm by considering sentence position and using cluster importance. However, those 2 news topics which used cluster importance by considering sentence position have a greater score of Rouge-1 than the one which only used cluster importance. The use of sentence position had an effect on the order of sentence on each topic, but there were only 2 news topics that affected the outcome of the summary.**

Keywords— News Summarization; Redundancy; Cluster Importance; Sentence Position

I. INTRODUCTION

Information is a notice regarding news that is usually contained in the form of articles, news, scientific papers, and books. However, the information presented is usually quite difficult for some people to understand because much information contained or called information overload [1], and this also occurs to news. In Indonesia, there are more than 43,000 news sites (https://nasional.kompas.com), where at a certain moment, some news discuss the same topic. Sometimes, people read news from one site to another to compare the content from the same topic. This will be very time consuming and sometimes there are several sentences that contain the same intent, and this means the reader will repeat the same occupation. This is why the summarization system is the proper solution to the problem. By using summarization system, some news with the same topic will be summarized into one summary. This will facilitate the reader to get the outline of some news practically.

Summarization is a process to reduce the size of original document to a size which is not more than half of the original document [2]. In document summarization, there are two types of summarization; both of them are extraction and abstraction. Abstraction is a summary produced by changing the sentence but it has same meaning with the original one [3]. Meanwhile, extraction is a summary produced by taking original sentence from the document. Research in this area mostly generates summary with extraction method. A multi-document summary is a summary that involves more than one document. Fabianus [4] conducted a research to create a system of multi-document summarization for news in Bahasa Indonesia by using extraction with TF-IDF method as sentence scoring. But this study does not consider the problem of sentence redundancy. However, the important thing from extraction summary is extracting important textual units from multiple related documents, removing redundancies, and reordering the units to produce the summary [8].

Redundancy is the appearance of similar sentence in a summary. It is one of the most important factors in multi document summarization task [8]. If there is a redundancy in result of a summary, it will make the summary contains a lot of important information with the same meaning. So, it requires a method or algorithm to find similarity between sentences to overcome redundancy and optimize the sentence selection as the material for summary compilation [4]. Since the documents may contain redundant information, the performance of a multi document summarization system heavily depends on the sentence similarity measure used for removing redundant sentences from the summary [5].

Many previous research on extractive summaries used sentence features such as position in the text, words frequency,

978-1-5386-8402-3/18/$31.00 ©2018 IEEE

or key phrases that indicated the importance words of the sentences. As thought Maximal Marginal Relevance (MMR), the earlier research on extractive summaries is considered redundancy issues by using the sentence level features. Clustering is an alternative approach to ensure good coverage and avoid redundancy that groups the similar textual units (paragraphs, sentences) into multiple clusters to identify themes of common information and selects text units one by one from clusters to the final summary [6][7]. Cluster importance is one of the algorithms on multi-document text summarization system using a similarity histogram based sentence-clustering algorithm to identify multiple sub-topics (themes) from the input set of related documents and selects the representative sentences from the appropriate clusters to form the summary [8]. This algorithm is adapting of a suitable sentence-clustering algorithm, which automatically determines the number of clusters and which is unsupervised in nature. So, this algorithm plays a vital role when the number of clusters is not known in advance.

This study proposes a solution to overcome redundancy by performing multi-document summarization of news in Bahasa Indonesia by using improvement of cluster importance algorithm with sentence position feature. Based on the research [9] explained that in a document, especially news, sentence position is an important feature where the sentence located in the beginning of paragraph has the biggest score than sentence which is located in the end of paragraph. The first phase to summarize is sentence clustering to news sub-topics. Second, the result of clusters that have been formed by weighting each cluster will be sorted. Third, summarization is done by choosing a representative sentence from each cluster by weighting every sentence and choosing a sentence with the highest weight from each cluster as material for summary aggregation.

II. METHOD

This study was described into figure 1. The picture describes summarization system that covers some processes to build summary. These processes will be explained in the subsections below.

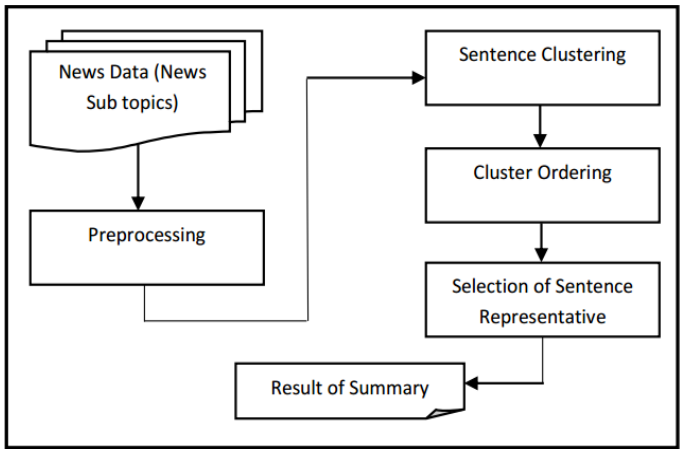

Fig. 1. Schematic of News Multi-Document Summarization System

News Dataset

The data of this study retrieved online news that as many as 30 topics. 11 topics of news were retrieved from research [9] and added another 19 topics of news. In one topic of news, there are 5 news sub-topics which every sub-topics were different to each other. The news data used was only the content without using the title of the news.

Preprocessing

In the news, data preprocessing will be several processes such as splitting the news into individual sentences, case folding, tokenizing, and stop words removal. Splitting the news into individual sentences is a process to split news in paragraph form into individual sentences. After that, every sentences is done case folding to harmonize all alphabets into lowercase and also to delete delimiter such as (.), (,), (!), (?), (:), and etc. Then, proceeding every sentences to perform tokenizing that breaks down each sentence into words carried out before stop words removal process in order to make it easier in removing important words. The list of stop words was taken from research [10].

Cluster Importance Algorithm

The concept of Cluster importance is that the more sentences that exist in a cluster, the cluster will be considered an important cluster that will be prioritized to appear at the beginning of the summary. Stages of cluster importance algorithm were sentence clustering, cluster ordering, and selection of sentence representative which will be explained in the subsections below. Furthermore, the contribution of this research is to add the position of sentence in the selection phase of representative sentence.

Sentence Clustering

Sentence clustering is the first stage of cluster importance algorithm. More clearly about the sentence clustering is illustrated in the flowchart of figure 2.

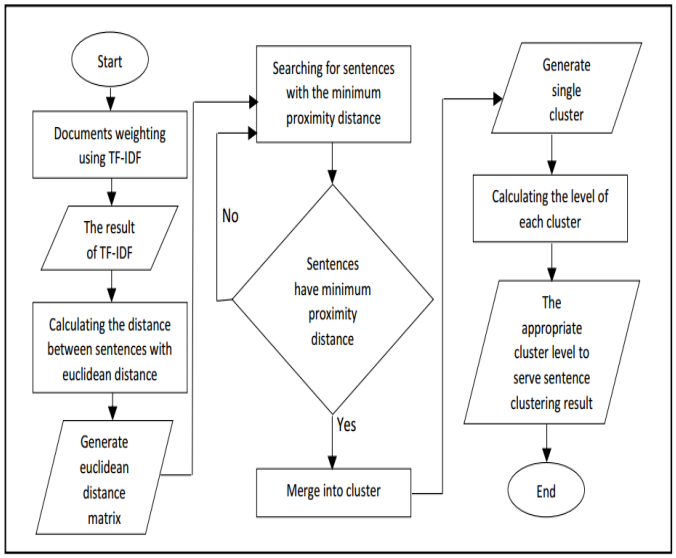

Fig. 2. Schematic of News Multi-Document Summarization System

Based on figure 2 above, the sentence clustering used a single linkage method. However, before using single linkage method, first performed documents weighting using TF-IDF. Term Frequency Inverse Document Frequency (TF-IDF) was weighting by calculating Term Frequency (TF) which was the

frequency or the number of words that appeared in a document. Besides, it also calculated Inverse Document Frequency (IDF) to calculate the importance of a word in a document seen from a number of documents as a whole [11]. The following equations 1 and 2 showed the Term Frequency Inverse Document Frequency (TF-IDF) weighting.

$$IDF\ (t) = \log \frac{N}{df(t)} \quad (1)$$

$$TF.IDF = TF(dt) * IDF(t) \quad (2)$$

After using TF-IDF as documents weighting, then measure the closeness between sentences by using euclidean distance. The distance between sentences was measured to determine the closeness between them. The following equation 3 shows euclidean distance.

$$Dis(x,y) = \sqrt{\Sigma(x^i - y^i)^2} \quad (3)$$

The next process after TF-IDF weighting and calculating minimum proximity distance using euclidean distance was sentence clustering by using single linkage. Single linkage is a clustering technique performed by consecutive way of merging begun by searching for two objects (sentences) with a minimum proximity distance. If it had minimum proximity distance, it will merge into a cluster. Next, look for other objects (sentences) that had minimum proximity to the cluster formed. If it had minimum proximity distance, it will merge into cluster formed or form a new cluster with other objects. It was done up to form single cluster. After single cluster was formed, and then choose the right level of cluster to determine the cluster as result of sentence clustering by measuring dissimilarity between them [12]. The following equation 4 calculates dissimilarity.

$$dissimilarity(cluster1, cluster2) = \frac{\Sigma Euclidian(d1,d2)}{size\ cluster1\ x\ size\ cluster2} \quad (4)$$

As for the mean of cluster [12] shown by equation 5 below.

$$Sim(X) = \Sigma_{d \epsilon X} Euclidian(d,c) \quad (5)$$

So, determining the level of cluster or appropriate cluster level as the result of sentence clustering needed to use equation 4 selected by the biggest dissimilarity value.

Cluster Ordering

Cluster ordering is the second stage of cluster importance algorithm. After completion of sentence clustering, a number of clusters were formed. The purpose of cluster ordering was to show information richness each cluster formed. A cluster consisted of sentences that are less important. Therefore, it needs cluster ordering or cluster sorting [8]. More clearly, cluster ordering can be illustrated by flowchart of figure 3.

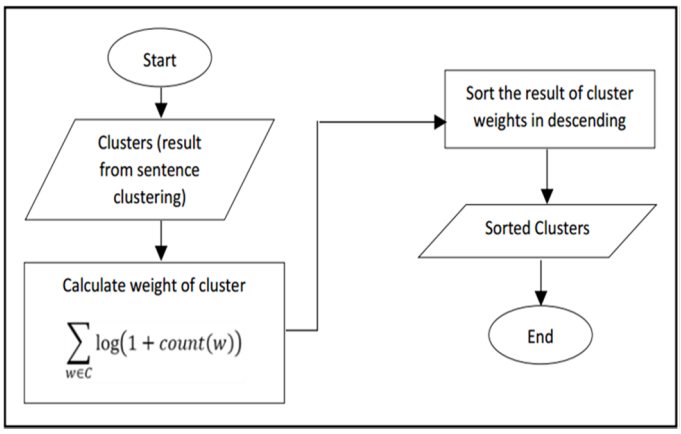

Fig. 3. Flowchart of Cluster Ordering

Based on figure 3, after a number of clusters have been formed from sentence clustering, each cluster was calculated its weight of words. After the weight of each cluster was found, do the sorting from largest to smallest. Cluster ordering is to know which cluster should be in the first as the material of summary aggregation.

Selection of Sentence Representative

The third stage was representative sentence selection. This stage was a stage to choose a sentence from each cluster which formed by weighting each sentence. More clearly, selection of sentence representative can be illustrated by flowchart of figure 4 above.

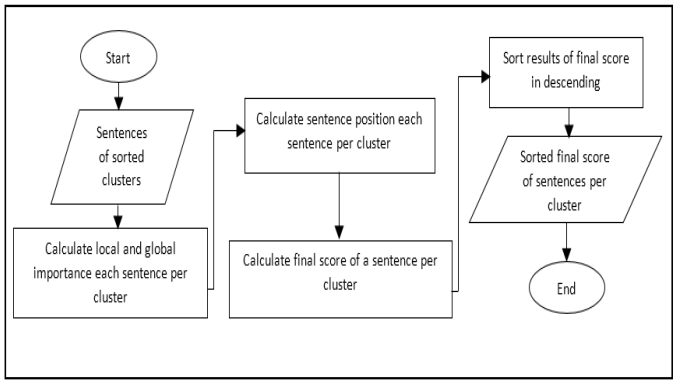

Fig. 4. Flowchart of Representative Sentence Selection

Based on figure 4, selection of sentence representative has two weighting calculation, namely W1 and W2. W1 is the first sentence weighting calculation by calculating local and global importance. Local importance is the word that indicates how many words in the formation of cluster or a number of a word in a cluster. The calculation of local importance was log(1+CTF), Cluster Term Frequency(CTF). While, global importance was the number of clusters that contained a word, the calculation using log(1+CF), Cluster Frequency(CF) [8]. The following equation 6 and 7 show W1.

$$W_1 = Score(S) = \sum Weight(w) \quad (6)$$

$$Weight(w) = \alpha_1 \log(1 + CTF) + \alpha_2 \log(1 + CF) \quad (7)$$

$$W_2(S) = \frac{1}{\sqrt{POS(S)}} \quad (8)$$

Beside calculating first weight, W1 then calculate second weight, W2. W2 is the calculation of sentence position. From equation 8, POS(S) indicated sentence index appeared in a document. Consideration of sentence position within summarization of news multi-document according to study [13] where sentences located at the beginning of document had greater score than at the end of document. Besides, based on science journalism, technical writing in online news used "inverted pyramid", important sentences were located at the beginning of document and less important sentences of news were located at the end of document [14]

Obtaining the weight of a sentence was done to add up the first and second weight (W1+W2), in other words, to calculate final score of a sentence shown by equation 9.

$$\textit{The last score(s)} = W1(s) + W2(s) \quad (9)$$

After the final score had been obtained, scores were sorted from largest to smallest. The greatest final score of a sentence from each cluster was elected as sentence representative of each cluster and also as the material for summary compilation.

TESTING DESIGN

This study will be tested against the summary result. The following picture explains the testing design.

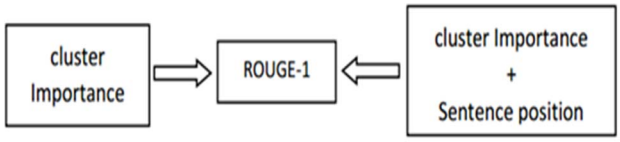

Fig. 5. Schematic of Testing Design

Based on figure 5 above, it showed testing scenario to summary results in which the testing was conducted between summary results of cluster importance and summary results of cluster importance with sentence position. This testing used Recall Oriented Understudy for Gisting Evaluation (ROUGE). ROUGE-N used to calculate the N-gram recall between system summary and reference summary [15]. N value used was 1. The following equation 10 shows ROUGE-N.

$$ROUGE - N = \frac{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count_{match}(N-gram)}{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count(N-gram)} \quad (10)$$

As this study used multi-document, the final calculation ROUGE-N was used as follows [15].

$$ROUGE - N_{multi} = argmax_i ROUGE - N(r_i, S) \quad (11)$$

Equation 11 above was to determine final value of ROUGE-N taken by maximum value of ROUGE-N.

III. RESULT AND DISCUSSION

The research evaluation was done by testing the summary results between the summary generated by the system and manually by using ROUGE-N based on equation 6 and 7. This evaluation used two ground truth sources. Testing scenarios on these news topics was to compare test results from 30 news topics using cluster importance algorithms plus the position of sentences (CI+POS) with the results of testing 30 news topics only using cluster algorithm of importance (CI). This section explains the results of this study based on the testing results from testing design as described earlier. The following will explain more about the testing results. Examples of summary results are presented below:

Kepala Bagian Kepesertaan, Badan Penyelenggara Jaminan Sosial (BPJS) Sumatera Bagian Utara, Manna Lubis mengatakan pada 2015 BPJS hadir di Gunungsitoli dan Tapaktuan Aceh. Bahkan, sesuai dengan roadmap cakupan kepesertaan yang menyebutkan di tahun 2019 seluruh rakyat Indonesia sudah menjadi anggota BPJS Kesehatan dan apabila ditahun 2019 tersebut sudah tercapai Universal Health Coverage (UHC), BPJS Kesehatan tetap membuka pendaftaran bagi peserta baru khususnya bagi bayi yang baru lahir, warga Indonesia yang baru kembali dari luar negeri, penduduk asing yang baru masuk ke Indonesia, dsb.
"Kami melakukan aksi demo di Kantor BPJS karena pelayanan BPJS yang masih buruk bahkan adanya diskriminasi," kata ketua Federasi Serikat Metal Indonesia Kota Depok Wido Praktikno, Senin (1/12/2014). Bahkan, sesuai dengan roadmap cakupan kepesertaan yang menyebutkan di tahun 2019 seluruh rakyat Indonesia sudah menjadi anggota BPJS Kesehatan dan apabila ditahun 2019 tersebut sudah tercapai Universal Health Coverage (UHC), BPJS Kesehatan tetap membuka pendaftaran bagi peserta baru khususnya bagi bayi yang baru lahir, warga Indonesia yang baru kembali dari luar negeri, penduduk asing yang baru masuk ke Indonesia, dsb.

Summary testing describes the results of testing which will be explained in the table I below:

TABLE I. ROUGE-1 VALUE OF SUMMARY RESULTS

No.	Topic	CI+Pos	CI
1	Air Asia	0.51705	0.51705
2	Angkot Tabrak Grab	0.24103	0.24103
3	Banjarnegara	0.56796	0.56796
4	BBM	0.37888	0.37888
5	BPJS	0.47429	0.42286
6	Countdown Asian Games	0.53498	0.53498
7	Demo Angkot Malang	0.46018	0.46018
8	Dokter Letty	0.43182	0.43182
9	Dolly	0.48831	0.48831
10	Ebola	0.76812	0.76812
11	Gempa Korea Selatan	0.66447	0.66447
12	Gunung Agung	0.49032	0.49032
13	Habib Rizieq	0.38462	0.38462
14	Hari Raya Nyepi	0.41341	0.41341

No.	Topic	CI+Pos	CI
15	Konser Boyband SHINee	0.44872	0.44872
16	Kunjungan Obama	0.50691	0.50691
17	Kurikulum 2013	0.47511	0.47511
18	Ledakan Gudang Mercon	0.54040	0.54040
19	Mahasiswi UI	0.34574	0.34574
20	Palestina	0.41905	0.41905
21	Penasehat KPK	0.35673	0.35673
22	Penutupan Hotel Alexis	0.40462	0.40462
23	Penyanderaan Angkot	0.24725	0.24725
24	Penyiraman Novel Baswedan	0.50459	0.50459
25	Pemilihan Presiden	0.36290	0.36290
26	Pria Pencuri Amplifier	0.64737	0.64737
27	Saksi Kunci E-KTP	0.61290	0.61290
28	Sinabung	0.24171	0.15640
29	Tora Sudiro	0.53333	0.53333
30	U19	0.54113	0.54113

The testing of summary included 30 topics using two ground truth. Based on table 1 above, from 30 topics, there were 2 topics which had different value of ROUGE-1 between cluster importance and cluster importance with sentence position. Those topics are BPJS and Sinabung. However, ROUGE-1 value of cluster importance with sentence position had the greatest value than ROUGE-1 value of cluster importance. While 28 topics had same value of ROUGE-1 between cluster importance and cluster importance with sentence position.

Figure 6 shows the illustration of ROUGE-1 result that representation with graph.

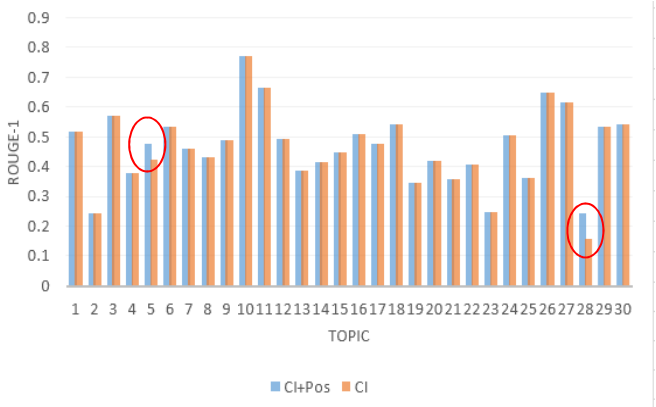

Fig. 6. Comparison Graph for ROUGE-1 Result between CI and CI+POS

The difference of ROUGE-1 value was caused by influence of sentence weight at the selection of sentence representative which affected order of sentence with highest weight as the material for summary compilation. Here is an example from a topic BPJS due to influence of sentence position weighting. From figure 7 and 8 due to changes the order of sentences, it affects the summary result that also has impact on the value of ROUGE-1 produced.

NO.	CLUSTER	SENTENCES	W1	W2	SENTENCE WEIGHT
		4.1 kepala bagian kepesertaan badan penyelenggara jaminan sosial bpjs sumatera bagian utara manna lubis mengatakan 2015 bpjs hadir gunungsitoli tapaktuan aceh	9.261	1	10.261
		1.5 melakukan aksi demo kantor bpjs pelayanan bpjs buruk diskriminasi kata ketua federasi serikat metal indonesia kota depok wido praktikno senin 1122014	9.28	0.447	9.727
		2.1 ketua asosiasi pengusaha indonesia apindo kota medan rusmin lawin mengatakan program asuransi badan penyelenggara jaminan kesehatan bpjs bagus	8.437	1	9.437
		5.8 sesuai peraturan perundangan pekerja penerima upah ppu bumn bumd badan usaha skala besar mapun wajib mendaftarkan pegawainya lambat 1 januari 2015	8.578	0.354	8.932
		3.1 pelaksanaan program bpjs badan penyelenggara jaminan sosial kesehatan diajukan kamar dagang industri kadin tahun 2019 mendatang ditunda	7.373	1	8.373
		4.4 kata bpjs cabang tapaktuan kedepannya melayani masyarakat kawasan aceh bagian barat kawasan berdekatan unit usaha disana	7.201	0.5	7.701
		2.6 rusmin menjelaskan hari sopirnya mengurus bpjs datang pukul 0600 wib antrean kantor bpjs mencapai 100 orang	6.904	0.408	7.312

Fig. 7. An example of Representative Sentence Selection with Sentence Position

NO.	CLUSTER	SENTENCES	W1	SENTENCE WEIGHT
		1.5 melakukan aksi demo kantor bpjs pelayanan bpjs buruk diskriminasi kata ketua federasi serikat metal indonesia kota depok wido praktikno senin 1122014	9.28	9.28
		4.1 kepala bagian kepesertaan badan penyelenggara jaminan sosial bpjs sumatera bagian utara manna lubis mengatakan 2015 bpjs hadir gunungsitoli tapaktuan aceh	9.261	9.261
		5.8 sesuai peraturan perundangan pekerja penerima upah ppu bumn bumd badan usaha skala besar mapun wajib mendaftarkan pegawainya lambat 1 januari 2015	8.578	8.578
		2.1 ketua asosiasi pengusaha indonesia apindo kota medan rusmin lawin mengatakan program asuransi badan penyelenggara jaminan kesehatan bpjs bagus	8.437	8.437
		3.1 pelaksanaan program bpjs badan penyelenggara jaminan sosial kesehatan diajukan kamar dagang industri kadin tahun 2019 mendatang ditunda	7.373	7.373
		4.4 kata bpjs cabang tapaktuan kedepannya melayani masyarakat kawasan aceh bagian barat kawasan berdekatan unit usaha disana	7.201	7.201
		2.6 rusmin menjelaskan hari sopirnya mengurus bpjs datang pukul 0600 wib antrean kantor bpjs mencapai 100 orang	6.904	6.904

Fig. 8. An example of Representative Sentence Selection without Sentence Position

IV. CONCLUSIONS

From the research that has been done, it can be concluded that there are two topics (BPJS and Sinabung) which have different value of ROUGE-1 between cluster importance and cluster importance with sentence position. ROUGE-1 value from cluster importance with sentence position of those topics has higher value than ROUGE-1 value from cluster importance.

Application of sentence position to cluster importance algorithm as consideration to summarize news document does not give significant result. This is shown on 30 topics that were tested, 28 topics have equal ROUGE-1 value between cluster importance and cluster importance with sentence position. However, application of sentence position shows difference of sentence order at the stage of representative sentence selection, but it does not give big impact to final summary result.

Based on explanation above, summary of application of sentence position is influenced by the news data itself where there is no exact sentence between one and another. So, it causes an influence on the weight of each sentence.

There is also a suggestion for further research that is to use another data research other than news. Therefore, it can be shown to be necessary or not regarding consideration of sentence position to summarize document.

REFERENCES

[1] W. E. Waliprana and M. L. Khodra, "Update summarization untuk kumpulan dokumen berbahasa Indonesia," *Cybermatika*, vol. 1, no. 2, 2013.

[2] N. Munot and S. S. Govilkar, "Comparative study of text summarization methods," *Int. J. Comput. Appl.*, vol. 102, no. 12, pp. 33–37, 2014.

[3] A. Agrawal and U. Gupta, "Extraction based approach for text summarization using K-Means clustering," *Int. J. Sci. Res. Publ.*, vol. 4, no. 11, 2014.

[4] Y. S. Fabianus H. Evan, P. W.P, and Pranowo, "Pembangunan perangkat lunak peringkas dokumen dari banyak sumber menggunakan sentence scoring dengan metode TF-IDF," in *Seminar Nasional Aplikasi Teknologi Informasi*, 2014.

[5] K. Sarkar, K. Saraf, and A. Ghosh, "Improving graph based multidocument text summarization using an enhanced sentence similarity measure," in *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, 2015.

[6] A. S. Asa, S. Akter, M. P. Uddin, M. D. Hossain, S. K. Roy, and M. I. Afjal, "A Comprehensive Survey on Extractive Text Summarization Techniques," *Am. J. Eng. Res.*, vol. 6, no. 1, pp. 226–239, 2017.

[7] S. KM and S. R, "Text Summarization using Clustering Technique and SVM Technique," *Int. J. Appl. Eng. Res.*, vol. 10, no. 12, pp. 28873–28881, 2015.

[8] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents," *Tech. – Int. J. Comput. Sci. Commun. Technol.*, vol. 2, no. 1, pp. 974–3375, 2009.

[9] N. Hayatin, C. Fatichah, and D. Purwitasari, "PEMBOBOTAN KALIMAT BERDASARKAN FITUR BERITA DAN TRENDING ISSUE UNTUK PERINGKASAN MULTI DOKUMEN BERITA," vol. 13, no. 1, pp. 38–44, 2015.

[10] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," Institute for Logic Languange and Computation Universiteti van Amsterdam, 2003.

[11] E. Purwanti, "Klasifikasi dokumen temu kembali informasi dengan K-Nearest Neighbour," *Rec. Libr. J.*, vol. 1, no. 2, 2015.

[12] Annisa, Y. Munarko, and Y. Azhar, "Peringkasan tweet berdasarkan trending topik twitter dengan pembobotan TF-IDF dan single linkage agglomerative hirarchical clustering," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control.*, vol. 1, no. 1, pp. 9–16, 2016.

[13] J. P. Mei and L. Chen, "SumCR: A new subtopic-based extractive approach for text summarization," *Knowl. Inf. Syst.*, vol. 31, no. 3, pp. 527–545, 2012.

[14] S. Verdianto, A. Z. Arifin, and D. Purwitasari, "Strategi pemilihan kalimat pada peringkasan multi dokumen," *J. Tek. ITS*, vol. 5, no. 2, 2016.

[15] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proc. Work. text Summ. branches out (WAS 2004)*, no. 1, pp. 25–26, 2004.