Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 10/2 (2017), 77-85
DOI: http://dx.doi.org/10.21609/jiki.v10i2.418

SENTENCE ORDERING USING CLUSTER CORRELATION AND PROBABILITY IN MULTI-DOCUMENTS SUMMARIZATION

I Gusti A. S. Adi Guna, Suci Nur Fauziah, and Wanvy Arifha Saputra

Informatics Department, Faculty of Information and Technology, Institut Teknologi Sepuluh Nopember, Jl. Raya ITS Kampus Sukolilo, Surabaya, 60111, Indonesia

E-mail: socrates.adiguna@gmail.com, wanvy15@mhs.if.its.ac.id

Abstract

Most of the document summary are arranged extractive by taking important sentences from the document. Extractive based summarization often not consider the connection sentence. A good sentence ordering should aware about rhetorical relations such as cause-effect relation, topical relevancy and chronological sequence which exist between the sentences. Based on this problem, we propose a new method for sentence ordering in multi document summarization using cluster correlation and probability for English documents. Sentences of multi-documents are grouped based on similarity into clusters. Sentence extracted from each cluster to be a summary that will be listed based on cluster correlation and probability. User evaluation showed that the summary result of proposed method easier to understanding than the previous method. The result of ROUGE method also shows increase on sentence arrangement compared to previous method.

Keywords: *Document Summarization, Cluster Ordering, Cluster Correlation, Probability*

Abstrak

Sebagian besar ringkasan dokumen dihasilkan dari metode *extractive*, yaitu mengambil kalimat-kalimat penting dari dokumen. Ringkasan dengan metode *extractive* sering tidak mempertimbangkan hubungan antar kalimat. Pengurutan kalimat yang bagus menunjukkan hubungan *rhetorical,* seperti hubungan sebab akibat, topic yang relevan, dan urutan yang kronologis diantara kalimat. Berdasarkan permasalahan ini, diusulkan sebuah metode baru untuk pengurutan kalimat pada peringkasan dari beberapa documen menggunakan *cluster correlation* dan *probability*untuk dokumen berbahasa inggris. Kalimat dari beberapa dokumen dikelompokkan berdasarkan kemiripannya ke dalam cluster-cluster. Kalimat diekstrak dari setiap cluster untuk menjadi ringkasan, ringkasan akan diurutkan berdasarkan *cluster correlation* dan *probability.* Hasil evaluasi pengguna menunjukkan hasil ringkasan dari metode usulan lebih mudah dipahami dari pada metode sebelumnya. Hasil ROUGE juga menunjukkan peningkatan susunan kalimat dari metode sebelumnya.

Kata Kunci: *Peringkasan Dokumen, Pengurutan Cluster, Cluster Correlation, Probability*

1. Introduction

In the present, huge electronic textual information is available and accessible. The information retrieval technology has make everyone can obtain large number of related documents using search engine. However, this situation also makes people need large of time to obtain the necessary information from all documents that they found. Automatic document summarization has been concern of the researchers for decade to solve this problem [1].

Most of existing automatic document summarization methods are extractive based, which mean they need to find the significant sentence or paragraph in documents, and arrange them to become a summary. However, extractive based summarization has a big hole on it. How the system arranges the sentences (or paragraphs) to become a proper summary is crucial part of this method. Sentence ordering without considering the relation among them can cause in incoherent summary [1].

A good sentence ordering should aware about rhetorical relations such as cause-effect relation, topical relevancy and chronological sequence which exist between the sentences. For example, if sentence A mention the event that caused by sen-

tence B, then we might want to order the sentence A before the sentence B in a summary that contains both sentences A and B [2].

Sentence ordering is more difficult in multiple documents, because the sentences that will draw up the summary is extracted from different writing styles documents and authors, which no one of the document can provide a standard sequence for all sentences that extracted. The way it orders the sentences should be context-aware and represent all of the source documents [3].

The previous research [4-7] proposed summarization document using Similarity Based Histogram Clustering (SHC) and cluster importance for sentence ordering method. SHC method capable to prevent the cluster contain duplicate sentence in it. SHC also prevent the situation that led the summary become redundant. Cluster importance is an ordering cluster method based on frequency information density among cluster's member. The information density is calculated by count the number of terms in cluster that has frequency above the predefined threshold. However, cluster importance ignores structure or relation among the cluster that provides the sentences that form the summary is not associated with each other.

Based on this problem, we propose a new method for sentence ordering in multi document summarization using cluster correlation and probability for English documents. This method is inspired by correlation coefficient method that used to measure the degree of relatedness of two vectors [8].

Sentence Clustering

Sentence clustering is used to find the similarity and dissimilarity across the documents. In sentence clustering, if the number of cluster has been determined, there is possibility that some sentences will be forced to become member of a cluster although it should not be. This probable error in cluster's member placement may cause some clusters to have duplicate member of sentence and led the summary become redundant. To avoid the problem in cluster member placement, we use Similarity Histogram-based Clustering (SHC) for the sentence clustering method. SHC can be used to make clusters by measure the similarity among the sentences [6].

The histogram that used in SHC is statistical representation of the similarity between the cluster members. Each value on histogram shows a certain similarity interval. A Similarity threshold is used in cluster member's selection. Every sentence will be registered into an appropriate cluster. The appropriate cluster can be found only if the sentence does not make the similarity value of these cluster members reduce. If that sentence can't be fit on any cluster, it should make a new cluster by itself [4]. The uni-gram matching similarity is used as function of similarity of two sentences, S_j and S_i. The similarity between sentences is calculated by count the corresponding words between S_j and S_i ($|s_i| \cap |s_j|$). Then it is divided by the total length of the words that form S_j and S_i ($|s_i|+|s_j|$) as shown in equation(1).

$$sim(s_i, s_j) = \frac{(2 * | s_i | \cap | s_j |)}{| s_i | + | s_j |} \quad (1)$$

$Sim = \{sim_1, sim_2, sim_3, \ldots. sim_m\}$ is collection of similarity between a couple sentences with $m= n(n=1)/2$. The equation determines the histogram function equation(2),

$$h_i = count(sim_j) \quad \text{for } sim_{li} \leq sim_j \leq sim_{ui} \quad (2)$$

Where sim_{li} shows the minimum similarity bin to-i and sim_{ui} is maximum similarity bin to-i. Histogram Ratio (*HR*) of cluster calculated by equation (3) and threshold determined by equation (4).

$$HR = \frac{\sum_{i=T}^{n_h} hi}{\sum_{j=1}^{n_b} hj} \quad (3)$$

$$T = [S_T * n_b] \quad (4)$$

S_T is similarity *threshold*, where bin number that corresponds to the similarity threshold (S_T) annotated with *T*.

Cluster Correlation

The inter-cluster correlation calculation based on frequencies a term that contained in each cluster. Some words are not always available in each cluster, that make so many comparisons with zero that does not affect to the results. For that, we simplify the calculation by use only important words to calculate the correlation of the cluster.

The important words are the words that often appear in all document. Determination of important words based on the frequency of occurrence terms that meet the threshold (θ) in all documents. The inter-cluster correlation is calculated by determine the cluster a and $T = \{t_1, t_2, t_3, \ldots t_n\}$ is set of term in cluster a. $w_{x,ti}$ is number of frequency term $t_i \in T$ in cluster a, then a has member $a = \{w_{a,t_1}, w_{a,t_2}, \ldots. w_{a,t_n}\}$. Weight of term frequency

in cluster *a* and *b* ($w_{t\ ab}$) calculated by multiply the number of distinct term *n* with the total of the result of multiplication the weight of each term t_i in cluster *a* (w_{a,t_i}) with the weight of each term t_i in cluster b (w_{b,t_i}), as the equation(5).

$$w_{t\ ab} = n \sum_{i=1}^{n} w_{a,t_i} * w_{b,t_i} \quad (5)$$

The Term frequency in cluster *a* and b calculated by multiply the cluster term frequency TF_a and TF_b in equation (6),

$$TF_a = \sum_{i=1}^{n} w_{a,t_i} \quad (6)$$

$$TF_b = \sum_{i=1}^{n} w_{b,t_i} \quad (7)$$

$$TF_{a,b} = TF_a * TF_b, \quad (8)$$

Correlation coefficient is generated by reduce the weight term frequency $w_{t\ ab}$ (equation (5)) with the term frequency of all cluster $TF_{a,b}$ (equation(8)), then it was divided by square root of weight term of every cluster. The function to calculate correlation is shown in the equation(9):

$$r_{(\vec{ta},\vec{tb})} = \frac{w_{t\ ab} - TF_{a,b}}{\sqrt{\left[n \sum_{i=1}^{n} w_{a,t_i}^2 - TF_a^2\right]\left[n \sum_{i=1}^{n} w_{b,t_i}^2 - TF_b^2\right]}} \quad (9)$$

The value of correlation has ranges from -1 (strong negative correlation) to 1 (strong positive correlation). If the value of correlation is 0, it is mean it has not any correlation with the other cluster [8]. The correlation coefficient was used as the basis for calculate cluster order on this paper.

Distributed Local Sentence

Distributed local sentence is a method of sentence arrangement. This sentences distribution method was proposed method by [4]. Distributed local sentence method is constructed through calculated the probability distribution, calculated sum of distribution, calculated expansion of distribution, calculated the weight of the component sentences and calculated the weight distributed local sentence. The weight local sentence is obtained by summing the entire component forming sentence *i* in cluster *k* which is divided by the number of component forming sentence *i* in cluster *k*. The equation (10) to calculate weight of distributed local sentence.

$$W_{ls}(s_{ik}) = \frac{1}{|S_{ik}|} \sum_{W_{t1,jk} \in S_{ik}} Wt_{l,jk} \quad (10)$$

2. Methods

There are five main steps that used in this research, i.e. preprocessing, sentence clustering, sentence ordering, sentence extraction and sentence arrangement. Figure 1 shows the process to generate a summary. The research method was adopted from the research [4].

Preprocessing

This process aims to prepare the data which is used in the sentence clustering. The preprocessing process consist of tokenizing, stopword removal, and stemming. Tokenizing is process of beheading the sentence into standalone words. Stopword removal is process of removing unused words. Stemming is process of getting the words to lower-case form. In this research, tokenizing process uses Stanford Natural Language Processing, stopword removal process uses stoplist dictionary and stemming process uses Library English Porter Stemmer [4].

Sentence Cluster

Data which is generated from preprocessing process will be grouped by using Similarity-based Histogram Clustering (SHC). SHC method was chosen because this method ensures the results of cluster remain convergent. Similarity between

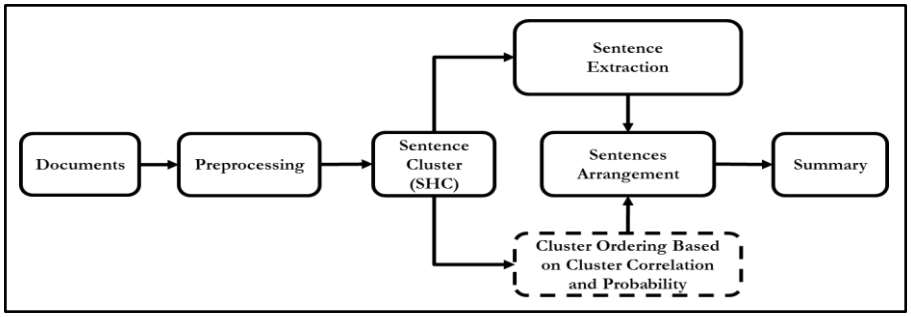

Figure 1. Sentence ordering based on cluster correlation

sentences was calculated by using uni-gram matching-based similarity as equation(1-4).

A sentence can be entered on certain cluster if it passes the criteria of the cluster. But if the sentence does not meet the criteria of all existing cluster, it will set up a new cluster. The SHC method is used in this research was adapted from research [4].

Cluster Ordering

The proposed method or contribution of this research is on the cluster ordering section. Sorting is useful to determine the order of sentences that represent in the summary. The summary will be easily understood if the sentences sorted according to the proximity of it with the topic or content.

Cluster ordering is used to calculate the order of each cluster using correlation between cluster and probability each cluster. There are three steps in this section, i.e. inter-cluster correlation calculation, probability calculation and weighted coherence.

Cluster Correlation

After sentence clustered using SHC, the first step was to determine the correlation of sentence cluster. The cluster correlation formula that adopted was equation (9) from correlation coefficient method [8] which can be used to measure the degree of relatedness for two vectors. That correlation can assume the value of correlation has ranges from -1 (strong negative correlation) to 1 (strong positive correlation). If the value of correlation is 0, it means that it uncorrelated with the other cluster.

In this research, the correlation was used without regard it has positive or negative value. So, in this research we add absolute value as the equation (11).

$$Correlation_{(\vec{ta},\vec{tb})} = |r_{(\vec{ta},\vec{tb})}| \qquad (11)$$

Correlation cluster values between the two clusters will determine the order of the cluster. In equation (11), the value of correlation has ranges from 0 to 1. If the value correlation is 0, it means has not any correlation with other cluster, but if the value reach 1, it means has any correlation regardless positive or negative correlation as equation (9).

Cluster Probability

The second step, to determine the probability. Probability calculated by use the frequency of occurrence of important sentences. That means Importance sentence is a sentence which consists of number of important words on it. Some cluster with large of member should have more weight because they have sentence information density.

It is difference with the cluster that has one or little members. Therefore, it required probability calculation. The Probability value obtained by comparing the frequency of important sentences in *cluster a* with frequency of importance sentence in all document (12).

$$\mathrm{P}(cluster_x) = \frac{freqImpSentece_{cluster\,x}}{freqImpSentece_{all\,document}} \qquad (12)$$

Weight Coherency

The third step, Weight coherency will determine the order of the sentence, the highest coherence weight will be concept of sentence ordering. WC_a is weight coherency which obtained by multiplying the sum of correlation value *a* to *x* with probability cluster *a,* where cluster *x* is correlation pair of cluster *a,* as equation(13).

$$WC_a = \sum_{x=1}^{n} Correlation_{a,x} \times P(a) \qquad (13)$$

The equation(13) can be assumed correlation cluster in equation(11) multiply by cluster probability in equation(12). The weight coherency can stabilize the value of correlation and frequency importance word in sentence within cluster. So the cluster that have most important sentences will be select as priority and the other cluster follow with sorted by descending based on weight coherency value.

It different in equation(9) that just only calculate cluster correlation and not focus the importance word in sentence within cluster. So the result of cluster ordering just sorted by high value of correlation in descending. The effect of that, negative correlation become last priority, and althought it have negative correlation, it still have many importance sentence rather than zero correlation.

Sentence Extraction

Sentence extraction is a phase to select a sentence that represents the cluster. The extracted sentences then used to arrange the summary. Sentence extraction uses sentence distribution from researcher [4]. Sentence distribution is used to determine the position of each sentence in the cluster, if a sentence that has elements spreader in a cluster will have a higher position in the cluster. The equation-

$$ROUGE - N = \frac{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count_{match}\,(N-gram)}{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count\,(N-gram)} \quad (14)$$

(10) is used for sentence distribution that used in this research. Sentence distribution that has the highest weight on each cluster will be used to arrange the summary.

Sentences Arrangement

The extracted sentence that used to arrange the summary will be listed based on the cluster ordering result. Every cluster contains one or more sentence that chosen based on sentence extraction section using sentence distribution method. Number of cluster is equal to number of sentence in summary

Evaluation method

The evaluation of automatic summarization is use ROUGE (Recall Oriented Understudy for Gisting Evaluation) method. ROUGE-N measures the ratio of n-grams between the candidate summary and the set of reference summaries. ROUGE is effective to evaluate the document summary result.

ROUGE-N is computed as equation (14) [9]. N is the length of N-gram, $Count_{match}$ (N-gram) is the maximum number of N-gram between the candidate summary and the set of reference summaries. Count(N-gram) is the number of N-gram in reference summaries. In this research, we use ROUGE-1 and ROUGE-2 where the best condition of ROUGE-1 and ROUGE-2 is 1.

3. Result and Analysis

Data Set

In this research, we use DUC (document understanding conferences) 2004 task 2 dataset from http://www-nlpir.nist.gov/projects/duc/data/2004-_data.html. DUC is one of the most popular data set for document summary. It is consisting of news documents collection from Associated Press and New York Times. 25 topics dataset is used where every topic consists of 10 documents.

Results

The experiment is performed using java programming. We compare the proposed method result with cluster importance method from [4]. The research [4] uses sentence clustering (SHC) and sentence distribution method for summary extraction and uses cluster importance for ordering sentence.

There are four parameters in this research, i.e. HR_{min}, ε, S_T, and α. HR_{min}, ε, and S_T are parameter for sentence clustering with SHC method. And then parameter α is parameter for sentence extraction. This research uses HR_{min}= 0.7 ,ε = 0.3, S_T = 0.4, θ = 10 , and α=0.4 or α=0.2. The other comparator method is also using parameter θ=10 for cluster ordering with cluster importance method.

Our focus in this research is to make a proper summary which more readable. We spread form survey to 20 volunteers and use ROUGE method. The form survey consists 25 topics of generated summary, all topics has been evaluated by 20 volunteers (post graduate students and English teacher) for evaluation. For each topic, every volunteer chooses the most coherent among the summaries which generated using cluster importance method, cluster correlation method, and cluster correlation + probability method.

Figure 2 shows the generated summary by cluster correlation method and cluster importance method for topic "Lebanese presidential election". 12 volunteers out of 20 volunteers have chosen summary using cluster correlation method with probability and another 8 volunteers have chosen cluster correlation method without probability for this topic.

The resulting summary using cluster correlation without probability have similar sentence ordering with the resulting summary using cluster correlation with probability. But, both have very different sentence ordering with the resulting summary using cluster importance.

In our analysis, there are some reasons that make the resulting summary using cluster correlation is better than the resulting summary using cluster importance, as follows: The cluster correlation can make chronological sequence of summary which exist between the sentences, The words in previous sentence are described in the next sentence.

Cluster correlation can avoid pronouns usage in the first sentence; because of the pronouns usage in the first sentence make the summary difficult to understand.

Table 1 shows the number of volunteers which choose cluster importance or cluster correlation as the satisfactory method for each topic. Figure 3 shows the volunteer's choice for cluster importance, cluster correlation with probability and cluster correlation without probability. Table 1 and figure 3 show that most of volunteers chose

Summarization's result using cluster correlation without probability

[1] Parliament on Thursday formally elected Gen. Emile Lahoud, the popular army commander who has the backing of powerful neighbor Syria, as Lebanon's next president.
[2] Lahoud's nomination complies with a tradition that the president be a Maronite Christian, the prime minister a Sunni Muslim and Parliament speaker a Shiite Muslim.
[3] Prime Minister Rafik Hariri, the business tycoon who launched Lebanon's multibillion dollar reconstruction from the devastation of civil war, said Monday he was bowing out as premier following a dispute with the new president.
[4] Such political disputes in Lebanon in the past were solved only with the intervention of Syria, the main power broker in this country.
[5] The new president will be sworn in Nov. 24, the day Hrawi leaves office.
[6] ``Congratulations, your excellency the general," Lebanese Prime Minister Rafik Hariri told army commander Emile Lahoud in a telephone conversation Monday that was headlined on the front-page of the leftist newspaper As-Safir.
[7] But many legislators, who in the past gave their overwhelming support to Hariri, did not name him and, instead, left it to the president to select a prime minister.
[8] A Cabinet minister and a close Syria ally on Wednesday criticized the Syrian-backed choice of the army commander as president, and said he will boycott a vote to elect the military man for the executive post.
[9] Lahoud pledged in a tough inauguration speech to clean up the graft-riddled administration.
[10] Lahoud, a 62-year-old naval officer, enjoys wide public and political support at home and has good relations with Syria.

Summarization's result using cluster correlation with probability

[1] Parlement on Thursday formally elected Gen Emile Lahoud, the popular army commander who has the backing of powerful neighbor Syria, as Lebanon's next president.
[2] Lahoud's nomination complies with a tradition that the president be a Maronite Christian, the prime minister a Sunni Muslim and Parliament speaker a Shiite Muslim.
[3] Prime Minister Rafik Hariri, the business tycoon who launched Lebanon's multibillion dollar reconstruction from the devastation of civil war, said Monday he was bowing out as premier following a dispute with the new president.
[4] Such political disputes in Lebanon in the past were solved only with the intervention of Syria, the main power broker in this country.
[5] The new president will be sworn in Nov. 24, the day Hrawi leaves office.
[6] But many legislators, who in the past gave their overwhelming support to Hariri, did not name him and, instead, left it to the president to select a prime minister.
[7] ``Congratulations, your excellency the general," Lebanese Prime Minister Rafik Hariri told army commander Emile Lahoud in a telephone conversation Monday that was headlined on the front-page of the leftist newspaper As-Safir.
[8] A Cabinet minister and a close Syria ally on Wednesday criticized the Syrian-backed choice of the army commander as president, and said he will boycott a vote to elect the military man for the executive post.
[9] Lahoud pledged in a tough inauguration speech to clean up the graft-riddled administration.
[10] Lahoud, a 62-year-old naval officer, enjoys wide public and political support at home and has good relations with Syria.

Summarization's result using cluster Importance

[1] His word is referred to in the Beirut media as "the password".
[2] Criticism of the nomination process also came from a meeting of Catholic bishops on Wednesday.
[3] Eleven deputies were absent.
[4] All the 118 legislators present at the session cast votes in his favor.
[5] The leading An-Nahar and other newspapers said the delay could last for days.
[6] The two leaders met Friday, but no presidential decree followed.
[7] It added that the money markets remained stable.
[8] But the dispute between the two leaders appears to be over who will have the upper hand in governing the nation of 3.2 milllion.
[9] '"I'm not a candidate, " Harari said in a live interview with CNN.
[10] Lahoud's nomination ends weeks of suspense over the identity of the next head of state.

Figure 2. Topic: "Lebanese Presidential Election"

cluster correlation with probability and cluster correlation without probability as the satisfactory system.

In this research, we used ROUGE-1 and ROUGE-2 as the metric evaluate the difference between ordering generated by automatic summarization and human. The Average ROUGE's score of the proposed method (cluster correlation and probability) and cluster important are presented in Table 2. It shows the comparison of ROUGE's score between the proposed method (cluster correlation and probability) and cluster importance where the proposed method gets better score than cluster importance.

The ROUGE-1's score comparison between cluster importance and proposed method (cluster correlation and probability) of each document are presented on Figure 4. Average documents show adjacent values between cluster importance and proposed method (cluster correlation and probabi-

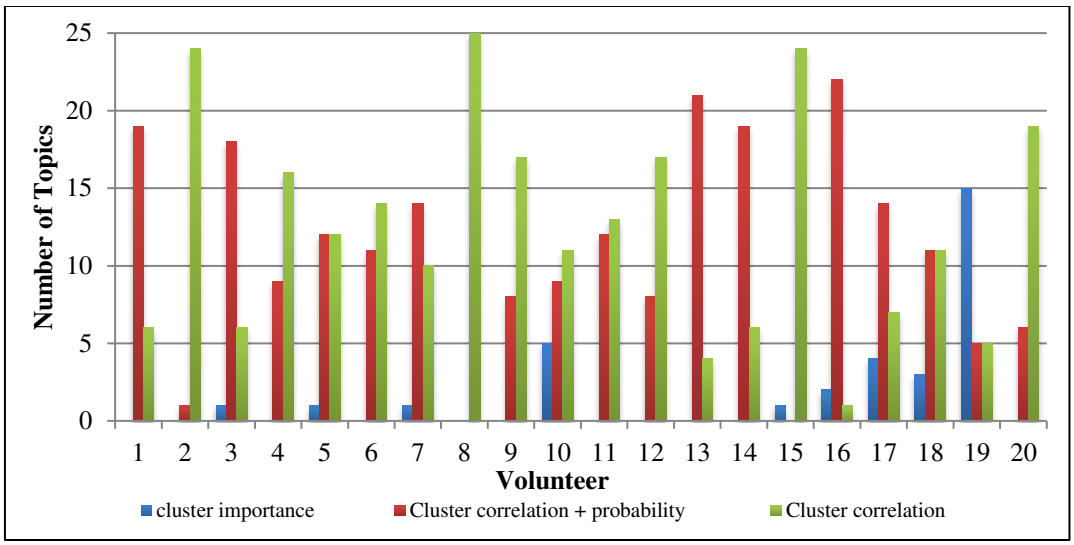

Figure 3. Graph showing the volunteers choice

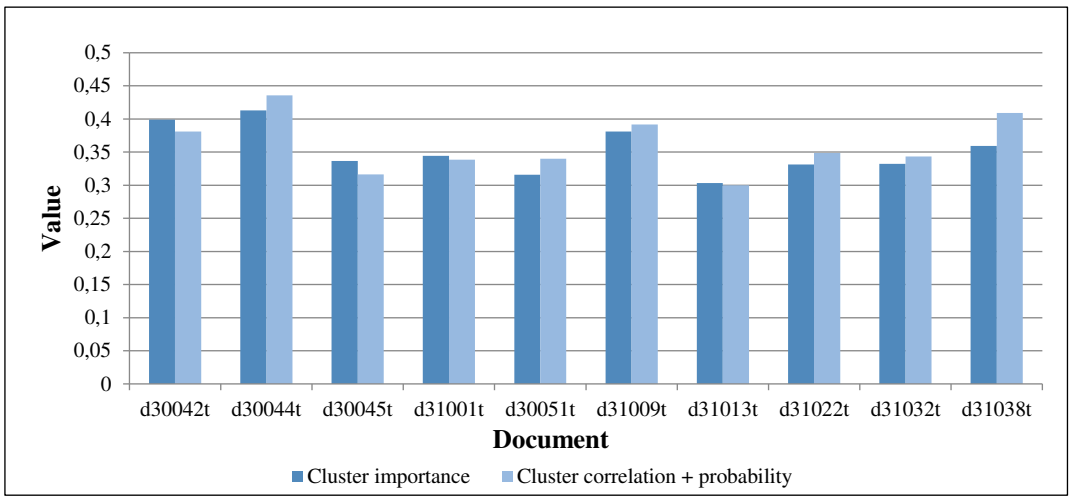

Figure 4. Graph Showing ROUGE 1 Every Document

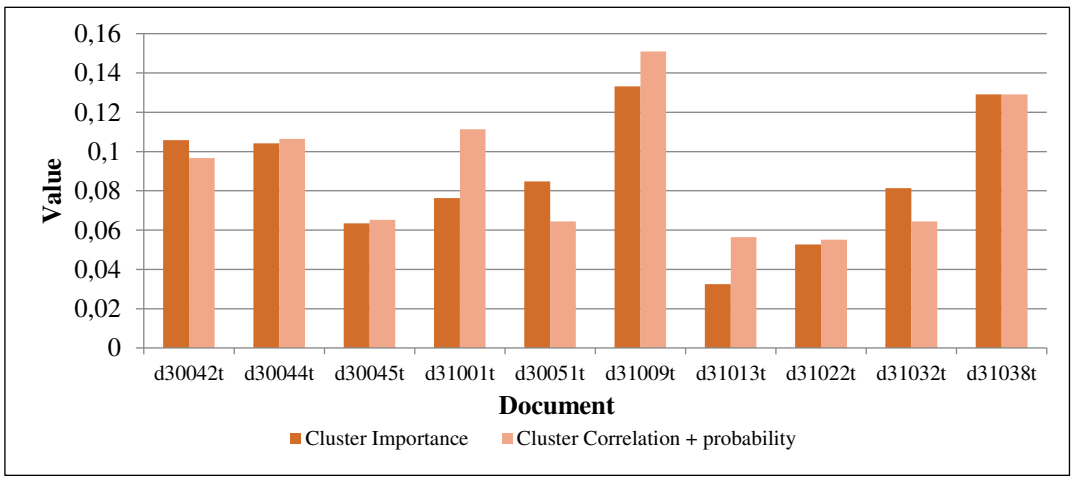

Figure 5. Graph showing ROUGE 2 every document

-lity). The comparison of ROUGE-2 score is presented in Figure 5. Figure 5 shows that proposed method gets better value than cluster importance in some documents.

TABLE 1
SURVEY RESULT FOR EACH TOPIC

Topic	Cluster importance	Cluster correlation + probability	Cluster correlation
T1	2	10	8
T2	1	8	11
T3	1	8	11
T4	3	7	10
T5	1	11	8
T6	0	11	9
T7	0	6	14
T8	2	7	11
T9	0	11	9
T10	1	7	12
T11	2	9	9
T12	1	7	12
T13	2	8	10
T14	1	11	8
T15	2	9	9
T16	3	11	6
T17	3	8	9
T18	2	8	10
T19	2	9	9
T20	1	11	8
T21	1	11	8
T22	2	11	7
T23	1	5	14
T24	0	12	8
T25	2	8	10

TABLE 2
TESTING OF SENTENCE ORDERING

Sentence Ordering Method	ROUGE 1	ROUGE 2
Cluster correlation + probability	0.360	0.090
Cluster importance	0.351	0.086

4. Conclusion

The proposed method is sentence ordering using cluster correlation and probability in multi document summarization has been successful. The result showed that the proposed method gives better summary than cluster importance. Summary using cluster correlation is preferred by most of volunteers.

Cluster correlation method has proven better than cluster importance which has average score 0.360 for ROUGE 1 and 0.090 for ROUGE 2 in Table 2. There is an increase value by 0.004 on ROUGE 1 and ROUGE 2. The increase in numbers is due to the evaluation on the ground-truth that measured from the summary of documents based on important sentences regardless of its sentence ordering, to know the sentences ordering in the document objectively and more significant, then the evaluation that used was human perceptions like surveys.

From the ordering of sentences, the cluster importance method not consider the correlation between important sentences in the cluster. Then the result of sentences ordering is different as shown in Figure 2, Figure 3 shows the proposed method is more objective on a topic than cluster importance method.

Combination cluster correlation with probability generate similar summary with cluster correlation without probability. The usage of probability in this method has no effect on the first to fifth sentence of the summary results. The resulting summary from the cluster correlation method represents not only the topic of document, but also the chronological sequence which exists among the sentences.

In the sentence ordering of multi document summarization, source documents cannot provide enough information in sentence ordering of summary as ground truth for evaluation. Figure 1 explains sentence ordering method of this research. The proposed method consists four steps:

The first step, preprocessing for prepare data which used sentence clustering. The second step, document is categorized using SHC method and produces clusters. The third step, sentence extraction uses sentence distribution method. The fourth, sentence ordering uses cluster correlation between clusters.

Because of sentence ordering is non-standard of ordering method; it is difficult to make a proper order of the sentence. The summary is extracted from different writing styles documents and authors. Our method is formed based on the information of source documents, which must be in multi document summarization.

The results show that the proposed method can improve the quality of the summary from multi documents which use SHC as sentence clustering method. It is related with some of previous researches which used SHC clustering method [4]–[7].

In future work, we will focus in how to deal with a large amount of document. The large quantity of document can make a huge amount of cluster, which makes the complexity of correlation calculation will be increased and hard to handle.

References

[1] U. Hahn and I. Mani, "Challenges of automatic summarization," *Computer (Long. Beach. Calif).*, vol. 33, no. 11, pp. 29–36, 2000.

[2] D. Bollegala, N. Okazaki, and M. Ishizuka, "A bottom-up approach to sentence ordering for multi-document summarization," *Inf. Sci. (Ny).*, vol. 217, no. 1, pp. 78–95, 2012.

[3] G. Peng, Y. He, N. Xiong, S. Lee, and S. Rho, "A context-aware study for sentence ordering," *Telecommun. Syst.*, vol. 52, no. 2, pp. 1343–1351, 2013.

[4] Wahib, A., Arifin, A.Z. and Purwitasari, D., 2016. "Improving Multi-Document Summary Method Based on Sentence Distribution". *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 14, no.1, pp.286-293.

[5] Suputra H. G.I, Arifin Z.A, and Y. A, "Stategi Pemilihan Kalimat pada Peringkasan Multi-Dokumen Berdasarkan Metode Clustering Kalimat," *J. Ilmu Komput.*, 2013.

[6] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents," *Tech. – Int. J. Comput. Sci. Commun. Technol.*, vol. 2, no. 1, pp. 325–335, 2009.

[7] R. Azhar, M. Machmud, H. A. Hartanto, and A. Z. Arifin, "Pembobotan Kata Berdasarkan Klaster pada Optimisasi Coverage , Diversity dan Coherence untuk Peringkasan Multi Dokumen."

[8] S. Al-anazi, H. Almahmoud, and I. Al-turaiki, "Finding Similar Documents Using Different Clustering Techniques," in *Procedia - Procedia Computer Science*, 2016, vol. 82, no. March, pp. 28–34.

[9] 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.