ISSN 2085-4552

Improving Multi-Document Summarization Performance by Utilizing Comprehensive Document Features

Rosalina
Faculty of Computing, President University, Cikarang - Bekasi, Indonesia
rosalina@president.ac.id

Diterima 12 April 20165
Disetujui 7 Juni 2016

Abstract— **The rapid growth of information technology and communication technology makes the volume of information available on the web increase rapidly. This development is leading to information overload. Multi-document summarization appears as a way to resolve the information overload problem in an effective way. In order to improve the performance of the multi-document summary this research combined the sentence features: *sentence centroid*, *sentence position*, *sentence length* and *IsTheLongestSentence* value to weight the sentences in order to find the most informative information of a text. In addition, this research uses a new method to calculate the weight of sentence position feature. The performance of the research result was evaluated using ROUGE metrics: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU. The research result outperform MEAD system if it was evaluated using the dataset of cluster D133C and D134H and if it was evaluated using ROUGE-1, ROUGE-S and ROUGE SU for cluster D133C and ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L and ROUGE-W for cluster D134H. This shows that the research result captures the important words in the extracted summary and it generates longer sentences as longer sentence contains more material that would match the one in the reference summaries.**

Index Terms— ***multi-document summarization, document features, centroid based summarization***

I. INTRODUCTION

The rapid growth of information technology and communication technology makes the volume of information available on the web increase rapidly. This development is leading to information overload. Automatic multi-document summarization can solve this problem by providing shortened versions of texts. In summarizing the document, no important information must be omitted and no information should be repeated. The summary should contain the most important information from the original text.

In finding the most important information, four features were used to weighted the sentences such as the sentence position, the frequency of words that appear in the document, the word occurrence in the document that also appear in the heading or title, and sentence words matching occurrence a pre-compiled list of cue words [3]. Another approach in finding the most important sentence is by studying the association of word frequency and human generated summary. From the observation, it was found out that the words that appeared in some human model were the words that appeared very frequently in the source document [6].

In order to identify the most important sentence, instead of scoring word frequency the scoring also can be done using a combination of centroid value, positional value, and first sentence overlap [1]. In addition, to enhance the performance of summarization, the cluster-based ranking approaches were explored. These approaches applied clustering algorithm to obtain the theme clusters first and then ranked the sentence within each cluster or by exploring the interaction between sentences and obtained clusters.

Therefore, in extraction methods it is important to find the most important sentence in the source documents. This research aimed to combine the clustering methods with sentence features to improve the performance of the multi-document summarizer.

II. LITERATURE STUDY

Multi-document summarization is a short version of a set of documents containing informative sentences from different sources. Since the source document came from different sources, it might contain dissimilar information therefore it must be clustered first. One of the issues in multi-document summarization was that the most important information from each source could be captured and then extracted into the summary.

Different approaches were used to decide which

sentences from the source documents to be included in the summary. One of the approaches was done by scoring the sentences. There were various combinations of feature used to score the sentences: position, sentence length and modified version of document frequency [4]. The frequency of words was another feature that could be used in scoring the sentences as the high *wordfrequency* from the input that was very likely to appear in the human models [6]. In addition, the term frequency could be combined with position features in scoring candidate sentences where the position of feature computed based on its occurrence either it appeared at the beginning or at the last part of the document and based on the average position of the word [7].

Another approach was done by using the centroid methods to decide which sentences to be included in the summary. With the centroid methods, the sentences were compared to a centroid score of a document set. For each sentence in the source documents, a score was calculated using the combination of sentence features: centroid, sentence position in the document, and the word overlap with the first sentence in the [1].

Term relevancy is one of a possible basis for a sentence scoring. Terms that are important for a document cluster have a high relevancy in that cluster. For example, the term "Jakarta" might be very relevant for a cluster of books about Indonesia, but not so relevant for cluster of books about ancient cluster of China.

III. Methodology

The multi-document summarization system's pipeline which is used in this research is shown in Figure 1.

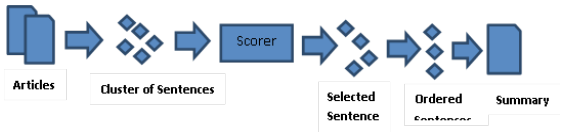

Figure 1. Multi-document summarization pipelines

1. Sentence Clustering

To generate a summary, at first all the relevant sentences need to be identified and then those documents are clustered based on its similarity. This research uses the output from clustering engine CIDR to group the relevant documents into clusters. CIDR stands for Columbia Intelligent Document Relater [2].

CIDR generates document centroid by only the first document in the cluster. Centroid contains the most highly relevant words to the entire cluster. Each document in the cluster is represented as a weighted vector of TF*IDF. Term Frequency (TF) of term *t* with N_t number of times term appears in a documents and D is the number of documents with term *t* in it is measured using Formula 1 [2].

$$TF(t) = \frac{N_t}{D} \quad (1)$$

The Inverse Document Frequency (IDF) is used to measure how important a term is, the IDF values in this research were taken from default IDF databases in MEAD distribution. The IDF database consists of 65300 pairs of words and its IDF values.

As new documents are processed, their TF*IDF values are compared with the centroid. The similarity value between a document and a centroid is measured by the cosine (normalized inner product) of the corresponding TF*IDF vectors as shown in Formula 2 [1].

$$sim(d,c) = \frac{\sum_k d_k * c_k * idf(k)}{\sqrt{\sum_k (d_k)^2}\sqrt{\sum_k (c_k)^2}} \quad (2)$$

Where:

d_k : TF value of term *k* in document *d*

c_k : TF value of term *k* in document *c*

idf : IDF value of term *k*

If the value of sim(d,c) within threshold, the new document is grouped in the cluster otherwise the new cluster should be created. The threshold value is 0.1.

2. Sentence Scoring

There are four different features that are used to score a sentence:

2.1 Centroid value

The centroid value C_i for sentence S_i is the sum of centroid value $C_{w,i}$ of all words in the sentence, the formula is shown in formula 3 [1].

$$C_i = \sum_w C_{w,i} \quad (3)$$

The centroid is the multiplication of term frequency with inverse document frequency. For example, the sentence *"U.S. Treasury Secretary Rubin arrives in Malaysia"* would get a score of 120.37 which is the sum of the individual centroid values of words as shown in Table 1.

Table 1. Centroid value

Term	TF	IDF	Centroid
Rubin	0.7	8.014115254	5.609880678
U.S.	0.5	6.404677342	3.202338671
Secretary	0.2	1.517340264	0.303468053
Treasury	0.2	4.229926521	0.845985304
Arrives	0.1	6.915502966	0.691550297
Malaysia	0.3	4.612917873	1.383875362
			12.03709836

2.2 Position Value

The positional value of sentence *s* is computed according to formula 4.

$$Position\ Value\ (s) = \frac{Max\_S - P + 1}{Max\_S} * \frac{P}{Max\_N} \quad (4)$$

Where

Max_S: Total number of sentences in the document

P : Sentence position in the paragraph

Max_N: Total number of sentences in the paragraph

2.3 Sentence Length

Sentence length is the length of each sentence in a document. The length of sentence is counted using formula 5.

$$L(s) = \frac{\sum w_i}{MaxWords} \quad (5)$$

Where w_i is the number of words in the sentence, and *MaxWords* is the maximum number of words in the sentence in the document.

For example the sentence *"Rubin will leave on Monday for Thailand and South Korea"* and suppose the *MaxWords* in the document is 29, then this sentence has the length value of 10/29 = 0.34.

2.4 IstheLongestSentence Value

IsTheLongestSentence is computed according to formula 6.

$$IsTheLongestSentence = \begin{cases} 0 & : \quad Length < Longest \\ 1 & : \quad Longest \end{cases} \quad (6)$$

3. Sentence Extraction

Sentence extraction is a method for generating summaries of a document, or document set by extracting sentences from the original document(s) and using those to generate a summary.

The sentences extracted as a sequence of n*r sentences from the original document presented in the same order as the input documents. In sentence extraction phase, the sentence scores are modified based on the relationship between pairs of sentences. The sentence score is computed based on formula 7.

$$SCORE(S_i) = C_i + P_i + L_i + ILS_i \quad (7)$$

Where:

i = Sentence number within the cluster

C = Centroid score

P = Position Score

L = Length Score

ILS = IsTheLongest Score

The next step in sentence extraction phase is ordering the sentences by score from highest to lowest.The length of the generated summary is based on compression rates given.

IV. DATASETS

The dataset was obtained from the Document Understanding Conference (DUC) of year 2004 task 5 where newswire data is grouped into clusters of documents based on topics. The news is collected from:

- AP Newswire, 1998 – 2000
- New York Times newswire, 1998 – 2000
- Xinhua News Agency (English Version), 1996 – 2000

The experiment of this research uses a total of 5 clusters which consists of 1274 sentences. In addition, the experiment uses MEAD summarization tool as a baseline to provide baseline summaries to be compared with the result of this research.

V. RESULTS AND DISCUSSIONS

The research results were evaluated using various ROUGE metrics. These metrics contain ROUGE-N, ROUGE-L, ROUGE-S and ROUGE-SU. The performance of research result was evaluated by comparing research results with MEAD summarizer.

Table 2 shows the result of ROUGE-1 evaluation of each cluster, the values for average-F lies in an interval of 0.24 to 0.37, the highest average-F value is achieved when average-R = 0.54 and average-P=0.29 using dataset of cluster D134H. This indicates that research result capture the important of individual words for ROUGE-1 average-F.

Table 2. The ROUGE-1 evaluation result of each cluster

ROUGE-1	MEAD			Research Result		
	Average-R	Average-P	Average-F	Average-R	Average-P	Average-F
D132D	0,15152	0,34884	0,21127	0,43434	0,17623	0,25073
D133C	0,10185	0,31429	0,15384	0,51852	0,28718	0,36964
D134H	0,10092	0,32353	0,15385	0,54128	0,29208	0,37946
D135G	0,05645	0,26923	0,09333	0,24194	0,28037	0,25974
D136C	0,11818	0,24528	0,15951	0,59091	0,15550	0,24621

It can be seen in Table 3 that the cluster D133C achieves the highest score of average-F if it is evaluated using ROUGE-2. This indicates that the research result captures the importance of bigram words co-occurrences.

Table 3. The ROUGE-2 evaluation result of each cluster

ROUGE-2	MEAD			Research Result		
	Average-R	Average-P	Average-F	Average-R	Average-P	Average-F
D132D	0,03061	0,07143	0,04286	0,08163	0,03292	0,04692
D133C	0,01869	0,05882	0,02837	0,18692	0,10309	0,13289
D134H	0,00926	0,0303	0,01418	0,29208	0,06468	0,08415
D135G	0,00000	0,00000	0,00000	0,01626	0,01887	0,01747
D136C	0,02752	0,05769	0,03726	0,13761	0,03597	0,05703

Moreover the research result using cluster D133C outperforms MEAD system on ROUGE-3 as seen on Table 4. This indicates that the research result not only captures the importance of bigram word co-occurances but as well as three-gram word co-occurances.

As shown in Table 5, the research result using cluster D133C also outperforms MEAD system if it is evaluated using ROUGE-4, it means that the research result also capture the importance of word four-gram co-occurrences.

Table 4. The ROUGE-3 evaluation result of each cluster

ROUGE-3	MEAD			Research Result		
	Average-R	Average-P	Average-F	Average-R	Average-P	Average-F
D132D	0,01031	0,02439	0,01449	0,01031	0,00413	0,00590
D133C	0,00000	0,00000	0,00000	0,09434	0,05181	0,06689
D134H	0,00000	0,00000	0,00000	0,03738	0,02000	0,02606
D135G	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
D136C	0,00000	0,00000	0,00000	0,04630	0,01202	0,01909

Table 5. The ROUGE-4evaluation result of each cluster

ROUGE-4	MEAD			Research Result		
	Average-R	Average-P	Average-F	Average-R	Average-P	Average-F
D132D	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
D133C	0,00000	0,00000	0,00000	0,04762	0,02604	0,03367
D134H	0,00000	0,00000	0,00000	0,00943	0,00503	0,00656
D135G	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
D136C	0,00000	0,00000	0,00000	0,01869	0,00482	0,00766

While using cluster D133C the research result also outperforms MEAD system if it is evaluated using ROUGE-L. As shown in Table 6, the research result not only captures the importance of word bigram, three-gram, and four-gram co-occurrences but as well as the longest common subsequence of words.

Table 6. The ROUGE-L evaluation result of each cluster

ROUGE-L	MEAD			Thesis Result		
	Average-R	Average-P	Average-F	Average-R	Average-P	Average-F
D132D	0,14141	0,32558	0,19718	0,39394	0,15984	0,22741
D133C	0,09259	0,28571	0,13986	0,47222	0,26154	0,33663
D134H	0,09174	0,29412	0,13986	0,44954	0,24257	0,31511
D135G	0,05645	0,26923	0,09333	0,20968	0,24299	0,22511
D136C	0,11818	0,24528	0,15951	0,54545	0,14354	0,22727

As shown in Table 7, the research result average-f scores outperforms MEAD system if it uses cluster of D133C on ROUGE-W. This indicates that the research result not only captures the importance of word bigram, three-gram, four-gram co-occurrences, and the longest common subsequence of words but as well as weighted longest common subsequence of words.

Table 7. The ROUGE-W evaluation result of each cluster

ROUGE-W	MEAD			Research Result		
	Average-R	Average-P	Average-F	Average-R	Average-P	Average-F
D132D	0,05291	0,22256	0,08549	0,134435	0,09960	0,11439
D133C	0,03622	0,19947	0,06131	0,15983	0,15798	0,15890
D134H	0,03861	0,21751	0,06558	0,15349	0,14555	0,14941
D135G	0,02336	0,20153	0,04187	0,07085	0,14851	0,09593
D136C	0,04972	0,18037	0,07795	0,18500	0,08509	0,11657

Meanwhile the research result outperforms MEAD system on ROUGE-S and SU if is evaluated using cluster D134H as shown in Table 8 and 9.

Table 8. The ROUGE-S evaluation result of each cluster

ROUGE-S	MEAD			Research Result		
	Average-R	Average-P	Average-F	Average-R	Average-P	Average-F
D132D	0,01855	0,09967	0,03128	0,15689	0,02567	0,04412
D133C	0,00952	0,09244	0,01726	0,21478	0,06561	0,10052
D134H	0,07565	0,08021	0,01397	0,27506	0,07975	0,12365
D135G	0,00341	0,08000	0,00654	0,04839	0,06507	0,05550
D136C	0,01540	0,06573	0,02495	0,31111	0,02169	0,04055

Table 9. The ROUGE-SU evaluation result of each cluster

ROUGE-SU	MEAD			Research Result		
	Average-R	Average-P	Average-F	Average-R	Average-P	Average-F
D132D	0,02101	0,11005	0,03528	0,16246	0,0269	0,04616
D133C	0,01105	0,10334	0,01997	0,22022	0,06782	0,10370
D134H	0,00934	0,09428	0,01700	0,27995	0,08185	0,12667
D135G	0,00426	0,09429	0,00815	0,05149	0,06907	0,05900
D136C	0,01540	0,06573	0,02495	0,31111	0,02169	0,04055

VI. CONCLUSIONS

From the experiment, it is found out that if the proposed system was evaluated using ROUGE-1, the values of the average-F score value lie in an interval of 0.24 to 0.37, the highest value was achieved when the average-R = 0.54 and average-P = 0.29 if it uses the dataset of cluster D134H. Moreover, if the proposed system uses the dataset of D134H, it also achieved the highest average-F score value if it was evaluated using ROUGE-S and ROUGE-SU. However, if the proposed system evaluated using ROUGE-2, the average-F score value lie in an interval of 0.01 to 0.13, and the highest value of the average-F score is achieved when the average-R = 0.18 and average-P =0.10 if it uses the dataset of cluster D133C. Using the cluster of D133C the proposed system achieved the highest average –F score value if it was evaluated using ROUGE-3, ROUGE-4, ROUGE-L and ROUGE, W.

Thus, it can be concluded that the proposed system outperformed MEAD system if it was evaluated using the dataset of cluster D133C and D134H and evaluated using ROUGE-1, ROUGE-S and ROUGE SU for cluster D133C and ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L and ROUGE-W for cluster D134H. This shows that the proposed system captures the important words in the extracted summary and it generates longer sentences as longer sentence contains more material that would match the one in the reference summaries.

REFERENCES

[1] D. R. Radev, Hongyan Jing, Malgorzata Stys, Daniel Tam (2004): Centroid-based summarization of multiple documents, Information Processing and Management 40, Elsevier, 919 – 938.

[2] D. R. Radev., Hatzivassiloglou, V., & McKeown, K. R. (1999): A Description of the CIDR system used for TDT-2, DARPA broadcast news workshop, Virginia.

[3] Edmunson (1969), New Methods in Automatic Extracting, Journal of the ACM, 16 No. 2, 264-285.

[4] Jun-Ping Ng, Praveen Bysani, Ziheng Lin, Min-Yen Kan, Chew-Lim Tan (2012): Exploiting Category-Specific Information for Multi-Document Summarization,Proceedings of the International Conference on Computational Linguistics.

[5] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, Ani Nenkova (2007): Beyond SumBasic: Task-Focused Summarization with Sentence Simplification and Lexical Expansion", Information Processing and Management, Special issue on summarization, Vol 43 (6).

[6] Vanderwende, A. Nenkova and L. (2005): The Impact of Frequency on Summarization, Microsoft Research, MSR-TR-2005-101.

[7] Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, Hisami Suzuki (2007):Multi-Document Summarization by Maximizing Informative Content-Words, International Joint Conference on Artificial Intelligence,1776-178.