International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Result Evaluation of Graph Based Multi Document Summarization

Vijay Sonawane[1], Rakesh Salam[2]

Information Technology Department, Technocrats Institute of Technology, Anand nagar, Bhopal, India

Abstract- ***Summarization is the process of decreasing large source document to shorten version of summary which will be easy to read. Document summarization is an emerging technique which is used for understanding the main purpose of any kind of documents. Summarization can be either single or multi document summarization. If summary is to be generated for single document then it is called as single document summarization. If summary is to be created for multiple relevant documents then it is called as multi document summarization. An Graph based approach for Multi Document Summarization is a graph based multi document summarization technique in which, set of documents is preprocessed, undirected graph will be constructed to calculate similarity between sentences, the word class is attached to each sentence, sentences are ranked according to word class and similarity of sentences and top ranked sentences are included in the summary.***

Keywords: Single Document, Multi Document, Summarization and Sentence Ranker

1. Introduction

A summary can be defined as a text that is generated from one or more texts, that include a major part of the information in the original text(s), and that is no longer than half of the original text(s) [6].Text summarization is the process of distilling the most important information from a source (or sources) to produce a shorter version for a particular user (or users) and task (or tasks) [10]. Roughly summarization is the process of decreasing a large volume of information to a summary or abstract preserving only the most essential items.

Due to the rapid growth of the Internet and the emergence of low-cost, large-capacity storage devices, we are now exposed to a lot of online information in daily life [1]. This situation makes it difficult for us to find and gather which exact information we need. Automatic text summarization is a key technology to solve this difficulty [2], with the properly summarized information, we can quickly and easily understand what the major points of the original document are and find how relevant the original document is to our own needs. We need to get right information without having gone through the source document [12]. Therefore we need a summary of document so that we can get the main purpose of the whole document.

2. Proposed Methods of Text Summarization

Graph Based Multi Document Summarization Multi Document Summarization is graph based multi document summarization algorithm. The Algorithm consists of the steps mentioned in Fig.1.The input passed to the system is a set of text documents. Firstly, the input set of related documents is pre-processed. Classes are attached to each sentence of the document and sentence length is calculated. The undirected graph will be constructed for each text document with sentences as vertices and similarities as edges. Thereafter, the sentences are ranked according to their absolute class, summed class and salient scores. The select top-ranking sentences to form the summary for each document and semantic checking are also used to filter out redundant information. Next, the single summary of each document will be assembled into only one document. Finally, the above described process is applied to this combined document to form the desire extractive summary.

2.1 Preprocessing

Before attaching a class to a sentence, the input set of related documents will be required to preprocess. Initially, the input documents are parsed to select all sentences. Those sentences, which are too short or almost, contain no information [12], then they are eliminated. Here all stop words are removed from each document and words are converted to their respective root form. Stemming is applied to reducing inflected words to their root form. For example, “finding” is converted to “find” [23]. In GBMDS, text file of stop words is maintained. If a sentence contains stop word present in a file then it is removed.

2.2 Class Attachment to the Sentence

Before constructing the graph, class is attached to each sentence of the documents. Here the database of word class is maintained. The sentences words attach to word class using predefined word class [23]. According to the database the absolute and summed class is attached to each sentence and calculated length of each sentence [7]. Length of each sentence is calculated as a number of characters present in a sentence. If sentence contains n characters then length of that sentence is n.

2.3 Graph Construction

The graph G = (V x E) which represents each sentence presenting in the document becomes a node and the edges of the graph represent similarity between the sentences.

$$\text{Similarity } (S_i, S_{i+1}) = \frac{\text{sum } (A_i\text{-}A_{i+1}, B_i\text{-}B_{i+1} \ldots Z_i\text{-}Z_{i+1})}{\text{Number of Characters}}$$

Where,
i = i^{th} sentence of the document.

A_i = Count indicating the number of times A has occurred in i^{th} sentence.

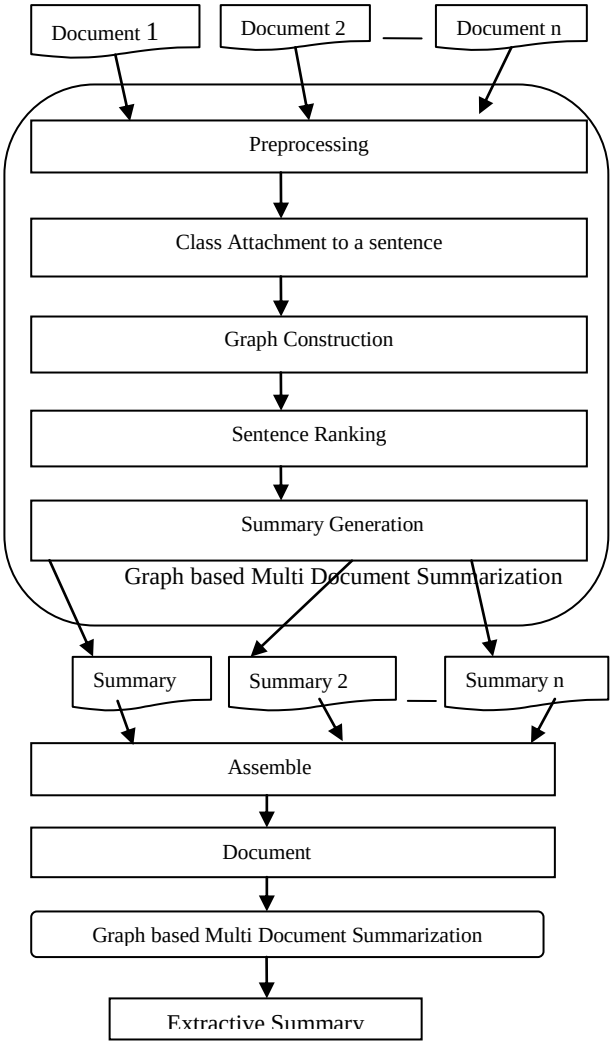


Fig.1: Main Process Graph Based Multi Document Summarization

A_{i+1} = Count indicating the number of times A has occurred in $i+1^{th}$ sentence.
B_i = Count indicating the number of times B has occurred in i^{th} sentence.
B_{i+1} = Count indicating the number of times B has occurred in $i+1^{th}$ sentence.
Up to
Z_i = Count indicating the number of times Z has occurred in i^{th} sentence.
Z_{i+1} = Count indicating the number of times Z has occurred in $i+1^{th}$ sentence.

Using this formula to calculate the similarity between the sentences, this means calculate the graph value from each sentence of source document.

2.4 Sentence Ranking

Once the document graph is constructed, the sentences in a source document will be ranked based on the absolute class, similarity between sentences and length of sentence [13]. The sentence is given high rank if its absolute class is higher than the remaining sentences of absolute class. If an absolute class between two sentences are given same value then the sentence is ranked based on the length of sentences. i.e. The sentence which has highest length will be given to next higher rank or else on the basis of similarity between sentences [12].

2.5 Summary Generation

In this step, final summary is generated by using selecting top ranking of sentence. Here, top rank of each sentence is refined according to the summed class. Summed class is used for arrangement of summary in proper sequence [10]. Simply, high ranking scores with sentences may be selected as the final ones in the summary. The sentences score is calculated based on relevant value and in-formative value.

3. Experimental Results

The three summarization techniques that we used in our comparative result have already been established. Summaries produced by Graph based Multi Document Summarization approach is compared with these established automatic generic multiple-document summarization methods: Random, LEAD, MEAD.

RANDOM [12] based technique randomly selects the sentences and put them inside summary. It uses threshold as a sentence length for selecting sentences for summary. Random based technique sets as lower bound. In LEAD [12] based technique first or first and last sentence is contained in the summary depending upon sentence length. It is best suitable for news summarization. This method involves selecting the highest score to the first sentence in each document, and then select second sentence in each document, and so on until desire summary is constructed. MEAD [8] is generates a centroid (vector) for all of the lines and then selects those lines which are closest to the centroid. MEAD [8] is also properly adjusts a sentence's score based on its length, its similarity to already selected sentences for the extract and its position in the original document. In Graph based Multi document Summarization, first set of input documents is pre-processed, class is attached to each sentence, similarity between each sentence is calculated, and sentences are given a rank according to their class and finally summary is generated. GBMDS uses absolute class, summed class, sentence length and sentence similarity to generate a summary. Summed class is applied to arrangement of summary in proper sequence.

Random based summarizer sets the lower bound i.e. it randomly selects the sentences whose length is better than a threshold. LEAD based summarizer technique selects first sentence of each text document, then select second sentence of each document, etc. until the final summary constructed. So, it is best suitable for news summarization. It sets upper bound. Random and LEAD are suitable for specific kind of documents. MEAD technique is a commonly used which may perform many different summarization tasks. It can also summarize individual documents summaries or clusters of related documents summaries MEAD [7] is the combination of lead-based and random based. It is a two baseline summarizer. A random based summary consists of enough selected sentences randomly (from the cluster) to generate a summary of the desire size. In GBMDS considers database of class, sentence length and similarity between sentences to

include them into a summary. The class is used for selection of sentences to include into summary and to arrange sentences in the appropriate order in the summary.

The two summarization methods that we used in comparative results have already been established. To evaluate Multi Document Summarization is graph based approach, it has compared with two summarizers techniques: Random and LEAD.

1. Random Summarizer: A summarization system that randomly selects lines with no overlapping till it reaches the final length of 40 words [7]. In this technique sentences are selects randomly and put them in the summary.
2. LEAD Summarizer: In LEAD based technique selects first sentence is containing in the summary depending upon sentence length size [7]. LEAD based summarizer techniques selects first sentence of each text document, then the second sentence of each document, etc. until the desired summary constructed. A LEAD [7] summarization system that chosen sentences with no overlapping till it reaches the final length of 40 words. The using the manual evaluation and automatic metric ROUGE evaluation to obtained resulting summaries. Examples of summaries are given in Table I.

Table 1: Summaries Generated from RANDOM, LEAD AND GBMDS

Random Summary
Granted most of this can be fixed by jail breaking you phone
Thing that you thought you wouldn't use.
No copy paste is not a big deal.
The internet is great but lack customization gas.
Still Camera takes nice shots.
Super cheap with the mobile and the third party apps and the zones hack.
LEAD Summary
Thing that you thought you wouldn't use.
Granted most of this can be fixed by jail breaking you phone
Still Camera takes nice shots.
The internet is great but lack customization gas.
Super cheap with the mobile and the third party apps and the zones hack.
No copy paste is not a big deal.
GBMDS Summary
Super cheap with the mobile and the third party apps and the zones hack.
Thing that you thought you wouldn't use.
The internet is great but lack customization gas.
My one big gripe is that it lacks customization.
Granted most of this can be fixed by jail breaking you phone
User friendly touch screen keyboard and a great experience.
No copy paste is not a big deal.

3.1 Manual Evaluation

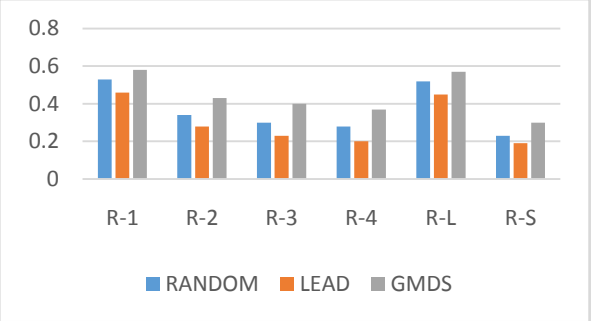
In the manual evaluation method, evaluation to obtained the readability of the created summaries. Without showing the reference summary of evaluation [1], we asked each people to rate of linguistic sentences with a scale range rate from a max of 5 (very good) to a min of 1 (very poor).

1. Grammaticality: sentences grammatically correct without artefacts.
2. Redundancy: The absence of unnecessary repetitions.
3. Clarity: Will be easy to read.
4. Coverage: cover of overall the aspects.
5. Coherence: organized and well-structured.

The each criterion included average score are shown in Table II.

Table 2: Manual Evaluations

	Random	LEAD	GBMDS
Grammatically	3.54	3.68	3.71
Redundancy	2.84	2.90	3.10
Clarity	2.80	2.97	3.05
Coverage	2.69	2.33	3.36
Coherence	2.05	2.60	2.62



Graph I: Manual Evaluations Of Three Summarization Techniques.

From Table II we have seen that the system contain scores of Grammaticality, Redundancy, Clarity, coverage and Coherence are close to each other. We observe only gap between in the Coverage metric [1]. This metric to describe how many aspects and opinions are actually covered in desire summary. The scores indicate that GBMDS graph based is able to generate summaries with a more efficient range of aspect than the other two systems.

3.2 ROUGE Evaluation

The ROUGE [19] is software package for automatically evaluate summary. It is technique of evaluation method for summarization, which is depending upon on the similar sentences between one or more model summaries [1]. Rouge is software package which is used for automatically evaluating summary and translation in natural language processing [20]. It is a set of metrics and metrics compare an automatically produced summary against with other summary created by human.

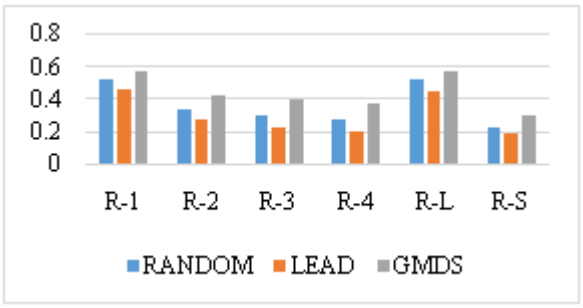
The Run Ids taken into realisation for this evaluation are ROUGE-1 (R-1), (R-2), (R-3), (R-4), ROUGE-L (R-L), and ROUGE-S (RS) [1]. The matrix id R-1 and R-2 which is used for calculate the number of bigrams and unigrams, respectively that coincides in the candidate and references summaries. R-S indicates the overlapping of skip bigrams between reference and candidate summaries [1][20]. ROUGE-L stand for Longest Common Subsequence (LCS) based statistics**.**

Table 3: Rouge Evaluations

Sr. No.	Metric (Run ID)	LEAD	RANDOM	GBMDS
		F-Score	F-Score	F-Score
1	R-1	0.46189	0.53255	0.58142

2	R-2	0.28583	0.34237	0.43768
3	R-3	023873	0.30656	0.40324
4	R-4	0.20116	0.28154	0.37954
5	R-L	0.45517	0.52814	0.57914
6	R-SU	0.19332	0.23866	0.3074

The problem of longest common subsequence takes into structure of account sentence level similarity naturally and identifies longest co-occurring in the sequence automatically.



Graph II: Comparative Results of Three Summarization Techniques Using F-Score Method.

From Table III we can see that in ROUGE metrics, GBMDS graph based compare with other two systems. This is, according to ROUGE [19], our summarizer produces summaries whose lexical sentences is closer to human created summaries and thus is more capture efficient the summaries other than the two systems.

4. Conclusion

A summary can be defined as a text that is generated from one or more texts, that include an important part of the information in the original text(s), and that is no bigger than half of the original text(s). Graph based approach for multi document summarization technique. In this technique, sentences are preprocessed, class is attached to each sentence, sentence length is calculated, undirected graph will be constructed, and each sentence is given rank based on class and then top ranked sentences has selected in summary, therefore its more efficient than other technique.

References

[1] Giuseppe Di Fabrizio, Ahmet Aker, "STARLET: Multi-Document summarization of Pro Product and Service Reviews with balanced rating Distributions", 2011 1th IEEE International Conference on Data Mining Workshops, DOI 10.1109/ ICDMW.2011.158, 2011 IEEE.
[2] Daan Van Bristsom, Antoon Bronselaer, Guy De Tr'e "Automatically Generating Multi Document Summarization", 2011 11th International Conference on Intelligent System Design and Application.
[3] J. Feng, M. Johnston, and S. Bangalore, "Speech and multi-modal interaction in mobile search," Signal Processing Mag-azine, IEEE, vol. 28, no. 4, July 2011.
[4] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target Extraction through double propagation," Comput. Linguist, vol. 37, 2011.
[5] N. Gupta, G. Di Fabbrizio, and P. Haffner, "Capturing the stars: predicting ratings for service and product reviews," in Proceedings of the NAACL HLT 2010 Workshop on Semantic Search, ser. SS '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 36–43.
[6] A. Aker, T. Cohn, and R. Gaizuaskas, "Multi document summarization using a* search and discriminative training", in proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics, 2010, pp. 482-491.
[7] Naomi Daniel, Dragomir Radev, Timothy Allison "Sub-event based multi-document summarization" Proceedings of the HLT-NAACL 03 on Text summarization workshop - Volume 5
[8] Jayabharathy, Kanmani, Buvana "An Analytical Framework for Multi-Document Summarization" International Journal of Computer Science Issues (IJCSI);May2011, Vol. 8 Issue 3, p308
[9] Antoon Bronselaer, Saskia Debergh, Dirk Van Hyfte, Guy De Tr'e, " Estimation of topic cardinality in document collections," in Proceeding of the 10th SIAM 2010 conference on data mining.
[10] Antoon Bronselaer, Guy De Tr'e, "Aspects of object merging", in Proceeding of the NAFIPS Conference, Toronto ,Canada, 2010.
[11] G. Carenini and L. Rizoli, "A multimedia interface for fa¬cilitating comparisons of opinions," in IUI '09: Proceedings of the 13th international Conference on Intelligent user inter-faces. ACM, 2009, pp. 325–334.
[12] Mohsin Ali, Monotosh Kumar Ghosh, "Multi-document Text Summarization:
[13] SimWithFirst Based Features and Sentence Co-selection Based Evaluation", 2009 International conference on Future Computer and Communication. Department of Computer Science and Engineering, Khulna University, Bangladesh.
[14] W.Duan, B.Gu, A.B.Whinston," Do Online Reviews Matter?- An Empirical Investigation of Panel Data," Journal Decision Support System, vol.45,2008.
[15] B.Pang, L.Lee , "Opinion mining and sentiment analysis," Foundation and Treands in Information Retrieval, vol.2 2008.
[16] A.Esuli, "Automatic generation of lexical resources for opinion mining: models, algorithms and application," SIGIR forum vol.42, 2008.
[17] D. Park, J. Lee, and I. Han, "The effect of on-line consumer reviews on Consumer purchasing intention: The moderating role of involvement," Int. J. Electron.Commerce, vol. 11, pp. 125–148, July 2007.
[18] G.Carenini, R.Ng, and A.Pauls, "Multi-document summarization of evaluative text," in 11th Meeting of the European Chapter of the Association for Computational Linguistics, 2006.
[19] Mokoto Hirohata, Yousuke Shinnaka "Sentence Extraction Based Presentation Summarization Techniques and Evaluation Metrics" 2005 IEEE ICASSP.
[20] Chin Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries".
[21] Chin Yew Lin. "ROUGE Working Notes" 2004 IEEE.

[22] Sparck Jones, K. Automatic summarizing: factors and directions. Advances in Automatic Text Summarization.MIT Press
[23] P,Naveen Kumar, A.P.Shiva Kumar "Concept Frequency: A Feature set Based Text Compression Model" 2012 International Journal of Advanced Research Computer Science and Software Engineering.
[24] Rafeeq Al-Hashemi, "Text Summarization Extraction System (TSES) Using Extracted Keywords " International Arab Journal of e-Technology, Vol. 1, No. 4, June 2010
[25] E Balagurusamy, "Programming with a Java", Fourth Edition, Tata McGraw Hill Publication.
[26] Herbert Schildt, "The Complete Reference Java", Seventh Edition, Tata McGraw Hill Publication.
[27] G. Booch, James Rumbaugh, "Object Oriented Modeling and Design", Second Edition, Prentice Hall
[28] R. Pressman, "Software Engineering: A practitioner's Approach", Seventh Edition, McGraw International Edition, 2010,