International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 32285-32295
© Research India Publications
http://www.ripublication.com

Multi - Document Summarization Using Cluster Model Incorporating Building Dictionary Method

[1]Gayathri. P, [2]Dr. N. Jaisankar, [3]Ankur Jain
[1]Assistant Professor (Senior), [2]Professor, [3]B.Tech Student
[1,2,3]School of Computing Science and Engineering, VIT University, Vellore – 632014, India
[1]pgayathri@vit.ac.in, [2]njaisankar@vit.ac.in,[3]ankur.jain2013@vit.ac.in

Abstract

This paper presents the study of Multi Document Summarization (MDS) of medical documents with the help of clustering model and building dictionaries. This paper also tells about efficiency of the method in practical abstract. This paper aims to reduce the complexity of the process by using the sequence of different methods. A summary generated from this method is evaluated against human written summaries. The objective of the multi document summarization is to generate an efficient summary with the help of sequencing a series of methods and algorithms. It also takes the keyword from the document and makes a local dictionary which is helpful in extracting summary from the document.It generates dramatically better summaries than an extractive summaries based on sentence extraction, reordering components and omitting useless words or lines. This presents an inter clustering method that takes events, topics or points from multiple clusters as inputs and then selects the most salient and relevant points to make the output. This also provides users a control over the summarization process using different methods. Besides the general idea and concept, we discuss the benefits and limitations concerning these methods. With the aim of enhancing MDS with this approach we can generate a better summary.

Keywords: Multi-document summarization, Clustering, Dictionary, Document Summarization

Introduction

In today’s world everything is in large quantity whether it is knowledge or anything else. Here we are talking about knowledge about anything. If we want to read something, there is large amount of data, uncountable. So you have to choose better because of lesser time. We don’t have time but the data or material have no limits.

Tolearn something, we go to the internet and tryto understand a particular thing about the topic in lesser time and tryto avoid reading all the details. We are trying in this paper to build something that can do your half work with the help of this method of an automatic text summarization or multi document summarization [1]. It is not easy for us to summarize documents manually from this large amount of data files on internet. The need of MDS is recently increased due to the proliferation of data on the web. For example, in medical field there is lots of data on the web. If a user wants to read the data it is not possible to read all, so here is the importance, MDS summarize the data files and build a new file which can contain data which is useful and short. It takes only the important data in the file and rest useless data is deleted (or we can say ignored), which saves the time of user.

The objective of MDS is to summarize the multiple documents into a single document and build a summary [2]. It reduces all the useless sentences which is not important or might be ignored. A summary could be indicative of what type of particular document is about and could be informative about the whole document. Summary can either be abstractive or extractive [3]. In abstractive summary, the sentences which are selected from the document are further processed and restructured before putting them into final summary. While in extractive summary the important sentences are identified and directly put into the summary. In this case, summary consist of original sentences same as in the document.

This paper focuses mainly on the informative and extractive type of summary. The problem of summarization was first identified by Sparck Jones and Endres-niggemeyer in 1995. The implementation of this method is very keen in the future world. Our extracted summary is accurate and more effective than any other previous methods. Our method is based on clustering model. The concept of machine learning is very effective in this manner as it can increase the reliability of summary. Mainly it can be used for improving efficiency of the summary. And the most important thing of this concept is after every task it improves itself to give better summary next time. So it is going better after every time it works.

This paper is organized as follows: section 2 describes about the previous work done in MDS. Section 3 illustrates about our proposed approach for MDS with architecture. Performance evaluation is done in section 4 and section 5 concludes the proposed work.

Related Work

Many researchers are working on how to get summary through different method. They are very good in some techniques but their basic approach is mainly based on text feature representation. In multi document summarization various methods are used to reach the goal [4]. The very first technique if we talk about is graph method and the results are very good. They find the length of the sentences and according to shortest path algorithm then embellish the summary. But through the time, clustering method [5] came to group similar topics in induction way. Many people find out that there are two types of summary that is abstractive and extractive [6]. Both have their own use. Summary made through data mining be more accurate and indicative which is very useful for user. The resultant summary is very accurate.

In recent years, MDS has become an important research topic. Many researchers had given their best thoughts to create and increase accuracy. Text mining [7] or data mining is also used in the research of multi document summarization. Machine learning [8] is a new concept for everyone. Many of us use this concept in multi document summarization to increase the accuracy which was very important [9]. System which can work on the principle of machine learning can learn from itself and upgrades to the new system. This means that this era's system can work on their own without human interference [10]. Some researchers have made this type of machines and testing results are very good. Many researchers used this technology in MDS to increase the value of summary.Some used K-NN based method [11] to summarize the multi document which is also good. For checking the accuracy there is a tool called ROUGE [12] which helps in compare the human written summary to system made summary and give points to the summary.

Many researchers used both extractive and abstractive simultaneously (half of the summary is extractive and half is extractive) that depends on the length of the sentences [13]. Like a big sentence which is considered to be important used abstractive and short length sentences used extractive [14]. This is good for very long sentences and considered to be important. Some used different method for presenting a summary. This is very important that how you present your result? So, many have used many different techniques to present the summary. Use of naïve algorithm is very common in ordering of sentences. Other is Chronological Ordering (CO). This is very easy to use and give the best result. Many researchers have written thesis on the use of CO [15]. This can help in presenting a summary in front of user that a user can easily understand that of what is written about the topic. We used some of the common techniques which can be used before but we used in a different manner with the help of different technique. This method is composed of many techniques which can be very helpful to increase the accuracy. The methods using this are more appropriate and useful. In this paper, the method we used is more effective than previous method. The resultant summary represents in such a way that it is easy to learn and understand with the help of embellishing. So, the output of this method would be powerful, more accurate and effective than previous methods output.

Proposed Method

As we discussed above, our proposed method is based on the sequence of different steps using different algorithms and different methods. From multiple documents to an output summary passing through a multiple steps to get our result best, we follow a flow shown in fig.1 below:

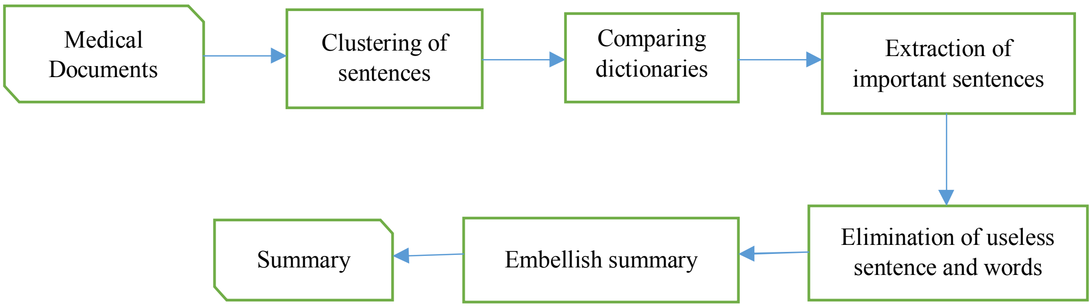

Figure 1: Proposed MDS Architecture

Clustering of Sentences

The meaning of cluster is to group similar objects into their classes. This can combine all the similar topics under same base. Mainly when user input the document many thing are repetitive so for avoiding repetition this method is used. Clustering is therefore a collection of objects which are similar between them and are dis-similar to the objects belonging to the other clusters. In this method we use many data variable, processing every word under the document. It takes up all the analysis and group all the similar topics, paragraphs and sentences under same page. Clustering method eventually ranked the sentences according to the similarity with cluster which simply represents frequent occurring time.

Some methods are quite popular like graph method and it is the best and simplest method for doing this step. Clustering of sentences or documents is best solved by graph method only. So, for using graph we use sentences as a node and frequency of the words in the sentences considered as edges. Let'stake graph G(N, E), where N denotes nodes and E denote edges. For clustering we have to use subgraph method and find the clique K in processing to build clusters in the document. Clique is a subset of vertices in a graph that is, every two vertices in a clique is adjacent in a graph. The Clique number of a graph G is the number of vertices in a maximum clique in G.We see the value of cliques which are completely connected to sub topics or we can say sub graphs from equations (1) and (2).

$$\mathrm{K} = \mathrm{N} - 1 \qquad (1)$$

$$\mathrm{n} = \frac{\mathrm{N(N-1)}}{2} \qquad (2)$$

Equation (2) checks that how close the neighbourhood of node to a clique.
Clustering coefficient(Ci),

$$Ci \; = ni/(Ki(Ki-1)/2)\{\mathrm{K} \neq 0,\ 1\ ;\} \qquad (3)$$

$$\text{Average clustering coefficient} = \frac{1}{N\sum_{i=1}^{N} Ci} \qquad (4)$$

Where,
N = number of node pairs
Ci = Clustering coefficient
K = Clique

n = Value that checks the neighbourhood of the node.

By finding clustering coefficient using equation (3) and average clustering coefficient using equation (4), we easily determine the clusters accurateness.After finding the clustering coefficient, we build the clusters with the help of cliques. So after building the clusters of sentences we proceed to the next step.

Comparing the Dictionaries

Building a Local Dictionary

Group of words are combined to form sentences and many sentences are combined to form paragraph and paragraphs leads to document. At last words are the root in every document and if we want to make our summary best, we have to work on the words. In this step our approach is to make a local dictionary by the help of tf-idf algorithm. The tf-idf algorithm works on the key basis. It counts the word frequencies and save these words to build the dictionary as shown in equation (5). This algorithm analyse all the keywords in the sentences and build up a dictionary which we called a local dictionary. The question is why we build this dictionary? The answer is bit simple; by this we find the importance of the words in the data. We already have a universal dictionary in which all the important terms and words are present. For examplein case of medical documents all the name of disease, branches, drugs etc. are mentioned in the universal dictionary.

The tf-idf is a numerical statistic that is intended to reflect how important a word t is to a document d in a collection or corpus D; mathematically it can be expressed as shown in equation (5).

$$tfidf = \frac{n_t}{n_d} x log \frac{D}{D_t} \tag{5}$$

Where,
n_t = Number of times that term t occurs in document d
n_d = Number of terms in document d
D = Total number of documents
D_t = Number of documents containing term t

Comparing Local dictionary with Universal Dictionary

By the help of tf-idf algorithm we build a local dictionary. This dictionary consists of all the words in the document except articles and pronouns, with the data of the frequency of each word present in the document. In the document many words occurred only once or twice but the word is important. So, for giving the importance to that word we use universal dictionary. Universal dictionary includes all the words which have importance of their own. We compare these both dictionaries and rank the sentences. Like the words matches with universal dictionary have higher rank with the higher number in word frequency than the words which are not in universal dictionary. We check the frequency and give them a rank. The words in the sentences have a rank now so as the sentences consist of these words have ranks. So, at the end of this, we have an order to give the priority of each sentence by assigning ranks. After finding the ranks we proceed to the extraction process.

Extraction of Important Sentences

The approach of extraction of sentences from the document is based on graph method. Sentences are the most important thing in the data. And we have to choose an important sentence which gives us information about the topic and headline, besides it, its worthless. So this is the keen step for building summary. As in the last step we rank the sentences according to the words in the sentences but according to that we can't tell that the summary is best. So, there are many other small steps for applying extraction, we choose our best which gives us more accurate data [16]. Following features are used to extract important sentences.

Sentences of short length - Sentence which constitute fewer words are more resourceful than long sentences.

Sentences having important words - Sentences which constitute the words which are important to the relevant topic. For example, we take a topic of medical analysis and the words like dentistry, different types of medicines, fields like surgeon, physician etc., are important words. We can use inbuilt dictionary for this purpose same as comparing dictionary concept.

Dates, years, time and place - Sentence which shows some data about dates and years like discovery, inventions or the places is considered to be in summary.

Cue words - There are certain words which shows their importance in their own like we take an example of results, define, introduction, significantly, advantages or disadvantages, pros or cons etc. These words wherever came we have to include the sentence into the summary.

We can find the importance of the sentences or words by finding graph efficiency. The results are good and can be considered. To avoid infinites, one can define graph efficiency (average inverse distance) from equation (6).

$$\text{Ŋ} =\frac{1}{N}\sum_{i,j\neq 1}^{l}\frac{1}{lij} \qquad (6)$$

Where,
N = Number of node pairs
L = Total number of nodes

Except graph efficiency we use many things but we choose best. The higher the graph efficiency higher is the importance. These points consist of approximately every case that could be helpful in building summary. The resultant of this method gives us the accurate ranking of the sentences in the data. After consideration of every point, the ranking [17] and priority of the sentences would be best. The lines which are compulsory, present in the summary without any error to give the result in form of best summary which is understandable, helpful and small.

Elimination of Useless Sentences and Words

The useless and extra sentences are deleted in this step. It works on the principle that only a particular amount of data can be conceived and rest of the data are ignored. The algorithm can also be set that user will put how many words or how much percentage he wants in his summary and up to that line summary can be processed and rest of the data can be deleted or ignored. The system shows that amount of data only and this much amount of data proceed to the next step. This step takes very less time to execute as it works only to conceive the required data. After this the data is preceded to the embellishing.

Embellish Summary

After getting required summary, this step is used to finalize the summary for presenting. Until this step all the data is randomly arranged in cluster form i.e., the dates are not sorted; only similar topic are bind due to clustering. In this step we used the concept of chronological ordering (CO). CO works on data that are not arranged in proper manner. Like dates, years and the sentences of past present and future. It works on all that theories to arrange the data that if someone will read all the things are arranged in proper manner. Using CO in the summary to describe the main events helps the user to understand what has happened. In this method we need to assign a date or order to a particular theme. To find the date of the theme we need to search the data and references for this manner. As mentioned earlier we arrange these themes in particular manner. In our case for medical documents we analyse the discovery, invention and details of the different theme to finalise the order in preparing summary. CO is very helpful in arrangement of this type of data. And we are using CO in this method to present the summary. Simultaneously this step can also work on how we present the final result in front of the user. Like the final summary should be in points or passages or paragraphs. So, we go with the point system as it is simple to understand the points rather than reading a whole paragraphs. Words which are not in used but comes with a lines (like “is”) which cannot affect the sentence if we replace it by “-”. It is more understandable in presenting summary like this way. The final summary is ready to display in front of the user. The resultant summary is more understandable; having points and conceives only the important data. The user can easily give time to read the summary as it is very short comparative to lots of documents.

Performance Analysis

To evaluate the proposed approach and the objective of the method to compare its comprehensibility, readability and usefulness against human written results reviews and machine generated results extracts. The extraction of sentences and the information given by the document is presenting a topical overview resembling a three level literature review. The resultant summary is similar to many extracted summaries. The resultant summaries resemble to the human written literature as they are laid out as presenting a comparative overview of similarities and unique features. For the analysis take an input of hundreds of documents and for each input document, three types of summaries were generated, each with a different kind of method i.e., our proposed method, sentence extraction method and a human written. It would be interesting to see whether these findings and differences would be replicated in a larger study. Content evaluation of the 30 sets of summaries from the documents of medical history and text is made [18]. As mentioned earlier each set of summary input has three types of summaries that are human written, proposed system based and baseline system based. The evaluation of these sets of summaries are done by a tool named ROUGE (Recall Oriented Understudy for Gisting Evaluation). To estimate the coverage of recall, precision and F-measure, we compared the system on ROGUE. ROUGE uses human written summary as a reference summary.

Content evaluation of the 30 sets of summaries by the ROUGE-1 metric revealed that proposed system summaries had a higher but not significantly different effectiveness or f-measure as compared to the baseline summaries, MEAD [19]. The MEAD summarization system was the baseline; it followed a sentence extraction approach to generate multi document extracts of information. When ROUGE gave the score for each summary we find out the sum of all the 30 set of summaries score and then divided it by 30 to find the average of all the summaries. The average score is shown in table 1 and table 2.

Table1: Results from the Content Evaluation using ROUGE

Measures	Proposed Approach	MEAD
Recall	0.70	0.64
Precision	0.5	0.4
F-measure	0.55	0.49

Table 2: Results from the Quality Evaluation using ROUGE

Measures	Proposed Approach	MEAD	Human Written
Comprehensibility	0.64	0.64	0.62
Readability	0.80	0.76	0.84
Usefulness	0.68	0.65	0.68

According to university of Washington [20], Seattle the method for analyse our system for summarizing document is based on coherence. G-flow, Lin-li and Nobata-li are estimation used for scoring the document summarization. As sentence ordering does not matter for ROUGE, we do not include Lin-li or Nobata-li in this evaluation. Because, our proposed method does not explicitly maximize the coverage while Lin does. We expected G-flow to perform slightly worse than Lin.

G-flow estimates the coherence of the summary via equation (7)

$$Coh(X) = \sum_{i=1\ldots|X|-1} w_G + (x_i\,, x_{i+1}) + \lambda w_G - (x_i, x_{i+1}) \quad (7)$$

While w_G represents edges (positive (+) and negative (-)) and lambda is a trade of coefficient while X to be a sequence of sentences $(x_1, x_2, \ldots., x_{|X|})$ used to calculate the coherence of the summary. The ROUGE-1 scores for G-flow and other recent multi document summarization system is shown in fig.2. We can conclude that good summaries have both the characteristics listed in the quality dimensions and good coverage.The summaries which are very good called gold summaries. G-flow only scores significantly lower than Lin and the gold standard summaries. An improvement to G-flow may focus on increasing coverage while retaining strengths such as coherence. The analysis of G-flow, gold and Lin-li is given below in fig. 2.

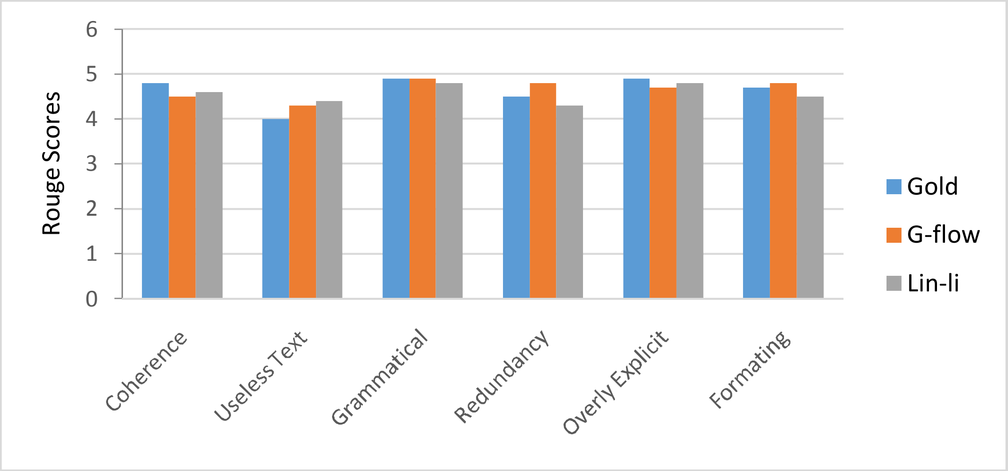

Figure 2: Ratings For The System. 0 Is The Lowest Possible Score And 5 Is The Highest Possible Score

Conclusion

The study tell us about the method for multi document summarization is exceptionally good despite that the method takes many steps to reach the summary but the combination of these steps produced very good result. In this paper, we discussed things in a sequential manner which gives us a very good and efficient summary. All the methods were best according to researchers, the things present in this paper and the things which we discussed or not discussed. This leads to a new era. We can combine a multiple document and produce a summary which is short, efficient, accurate and understandable with the help of these steps. The summary and combination of multiple documents can be used in many ways but the most important thing is that the time of user would save. The user can grasp many things in less time as before. The user can search things or read full books/documents but because of the method we proposed the frequency of grasping things will increase. The resultant output of this method is exceptionally good.

References

[1] Chin-Yew Lin and Eduard Hovy, 2002, “Manual and Automatic Evaluation of Summaries”, Proceedings of the Workshop on Automatic Summarization (including DUC 2002) Philadelphia, July, pp. 45-51.

[2] Li Wenjie, Wei Furu, Lu Qin and He Yanxiang, 2008, “PNR2: Ranking Sentences with Positive and Negative Reinforcement for Query-Oriented Update Summarization”, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 489–496 Manchester.

[3] Yogan Jaya Kumar and Naomie Salim, 2012, “Automatic Multi Document Summarization Approaches”, Journal of Computer Science 8 (1): 133-140, ISSN 1549-3636.

[4] Lin, C.Yand Hovy, E, 2003, “Automatic evaluation of summaries using n-gram co-occurrence statistics”, Proceedings of the Conference of the North America Chapter of the Association for Computational Linguistics on Human Language Technology- volume 1(pp. 71-78) Association for Computational Linguistics.

[5] Dingding Wang,Shenghuo Zhu, Tao Li,Yun Chi and Yihong Gong, 2008, “Integrating Clustering and Multi-Document Summarization to Improve Document Understanding”, CIKM’08, October 26–30, Napa Valley, California, USA. ACM 978-1-59593-991-3/08/10.

[6] Kokil Jaidka, Christopher S.G. Khoo and Jin-Cheon, 2013, “Deconstructing Human Literature Reviews – A Framework for Multi-Document Summarization”, Proceedings of the 14th European Workshop on Natural Language Generation, pages 125–135, Sofia, Bulgaria.

[7] Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari and Jordan Pascual, 2014, “KNN based Machine Learning Approach for Text and Document Mining”, International Journal of Database Theory and Application Vol.7, No.1, pp.61-70

[8] Inderjeet Mani and Eric Bloedorn, 1998, “Machine Learning of Generic and User-Focused Summarization”, The MITRE Corporation, 11493 Sunset Hills Road, W640, P~eston, VA 22090, USA From: AAAI-98 Proceedings.

[9] Judith D. Schlesinger, Dianne P. O’Leary and John M. Conroy, 2008, “Arabic/English Multi-document Summarization with CLASSY—the Past and the Future”, A. Gelbukh (Ed.): CICLing 2008, LNCS 4919, pp. 568–581, Springer-Verlag Berlin Heidelberg.

[10] Yancui Li, Yuesheng Gu, Hongyu Feng and Yanpei Liu, 2011, “Research of Multi-Document Summarization Using Affinity Propagation Clustering”, Advances in information Sciences and Service Sciences(AISS) Volume3, Number10, doi : 10.4156/AISS.vol3.issue10.51.

[11] Anton Leuski, Chin-Yew Lin and Eduard Hovy, “iNeATS: Interactive Multi-Document Summarization”, University of Southern California Information Sciences Institute 4676 Admiralty Way, Suite 1001 Marina Del Rey, CA 90292-6695.

[12] Kai Hong, John M.Conroy, Benoit Favre, Alex Kulesza, Hui Lin and Ani Nenkova, 2004, “A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization”, DUC.

[13] Hongyan Jing and Kathleen R. McKeown, “Cut and Paste Based Text Summarization”, Department of Computer Science Columbia University New York, NY 10027, USA.

[14] Jade Goldstein, Vibhu Mittal, Jaime Carbonell and Mark Kantrowitzt, 2000, “Multi-Document Summarization By Sentence Extraction”, School of Computer Science at Research Showcase, CMU 4.

[15] Regina Barzilay, Noemie Elhadad and Kathleen R. McKeown, 2001, “Sentence Ordering in Multi document Summarization” T ’01 San Diego, California USA Copyright ACM 0-89791-88-6/97/05.

[16] Kai Hongand Ani Nenkova, 2014, “Improving the Estimation of Word Importance for News Multi-Document Summarization”, Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 712–721, Gothenburg, Sweden, April 26-30.

[17] Kathleen R. McKeown, Judith L. Klavans, Vasileios Hatzivassiloglou, Regina Barzilay and Eleazar Eskin, 1999, “Towards Multi document Summarization by Reformulation: Progress and Prospects”, Department of Computer Science Columbia University 1214 Amsterdam Avenue New York, NY 10027, From: AAAI-99 Proceedings.

[18] Medical Library Association312.419.9094info@mlahq.org

[19] Dragomir R. Radev, Sasha Blair-Goldensohn and Zhu Zhang, “Experiments in Single and Multi-Document Summarization Using MEAD”, School of Information and Department of EECS University of Michigan Ann Arbor, MI 48109.

[20] Janara Christensen, Mausam, Stephen Soderland and Oren Etzioni, “Towards Coherent Multi-Document Summarization”, Computer Science & Engineering University of Washington Seattle, WA 98195.