Efficient Multi-Document Summary Generation Using Neural Network

Ms. Sonali Igave, Prof. C.M. Gaikwad

Department of Computer Science Engineering, Government Engineering College, Aurangabad , India

Abstract*—From last few years online information is growing tremendously on World Wide Web or on user's desktops and thus online information gains much more attention in the field of automatic text summarization. Text mining has become a significant research field as it produces valuable data from unstructured and large amount of texts. Summarization systems provide the possibility of searching the important keywords of the texts and so the consumer will expend less time on reading the whole document. Main objective of summarization system is to generate a new form which expresses the key meaning of the contained text. This paper study on various existing techniques with needs of novel Multi-Document summarization schemes. This paper is motivated by arising need to provide high quality summary in very short period of time. In proposed system, user can quickly and easily access correctly-developed summaries which expresses the key meaning of the contained text. The primary focus of this paper lies with the f_β-optimal merge function, a function recently presented here, that uses the weighted harmonic mean to discover a harmony in the middle of precision and recall. Proposed system utilizes Bisect K-means clustering to improve the time and Neural Networks to improve the accuracy of summary generated by NEWSUM algorithm.*

***Keywords*—Multi-document summarization, Clustering, f_β-optimal merge function, Neural Network.**

I. INTRODUCTION

In recent years use of the internet is increased rapidly thus online information is growing tremendously on web or on user's desktops. Online information generated which may be in the form of structured or unstructured and it is very difficult to read all data or information of that form. So problem of overloading information increases as use of World Wide Web and many sources like Google, Yahoo! surfing also increases.

Text mining has become a significant research field as it produces valuable data from unstructured and large amount of texts.

Main aim of summarization system is to generate a new form which expresses the key meaning of the contained text. Summarization systems provide the possibility of searching the important keywords of the texts and so the consumer will expend less time on reading the whole document. Clustering is process of grouping similar types of objects into one cluster. Data clustering is useful for data analysis. Finally main objective of summarization is to create summary which generates minimum redundancy, maximum relevancy.

This paper uses the concept of neural networks for efficient summary generation of multiple documents. For this, it uses number of attributes such as, sentences to word count, sentence position, and number of stop-words in sentence etc. Neural network verify every sentence against each of these attributes and generate output and calculate the average of all output. Then this average is used to decide the class of each sentence. Sentence is classified as either positive or negative.

The common definition captures three important features that characterize research on automatic summarization:

- Summaries may be generated from a single document or multiple documents.
- Summaries should protect important information.
- Summaries should be very short like one paragraph.

This paper study the related work done by a different publishers and researchers, in section II, the implementation details in section III where we see the system architecture, modules description, algorithms, mathematical models, and experimental setup. In section IV we discuss about the expected results and at last conclusion is provided in section V.

II. RELATED WORK

In paper [2], author proposed novel system called CATS as a multi-document summarizing system. Proposed automatic summarization system mines sentences to create 50 summaries of 250 words each, to resolve 50 complex questions on different topics. Author utilizes various

statistical techniques to generate a score for particular sentence in the documents. Furthermore summaries are shortening using sentence compression and a cleaning algorithm. Further to improve the performance author need to work on two features such as sentence compression and the distinction between the two granularities. . Proposed system achieves advantage, it retrieves the right sentences from the documents to answer a given question.

In paper [3], author proposed novel approach to automatic document summarization on the basis of clustering and extraction of sentences. Author proposed twofold approaches: in first step, sentences are clustered, and then in second step, sentences are generated based on each cluster. Proposed approach improves the summarization results significantly and this method evaluated using ROUGE-1, ROUGE-2 and F1 metrics. Finally author concludes that summarization result depends on the similarity measure.

In paper [4], author proposed a novel technique called ROUGE (Recall-Oriented Understudy forgetting Evaluation) an automatic evaluation package for summarization. Proposed scheme has some measures to automatically establish the quality of a summary by comparing it to other summaries which is generated by humans. Proposed scheme illustrates four different ROUGE measures such as ROUGE-N, ROUGE-L, and ROUGE-Wand ROUGE-S.

In paper [5], author proposed a multi-document summarization novel system, called NeATS. Author is motivated by content and readability of the results. Proposed scheme attempts to mine relative or required portions from multiple documents about some topic and finally arrange them in coherent order. Proposed scheme is outperforms in the large scale summarization evaluation.

Proposed method utilizes the common methods guided with some principle such as extracting significant concepts based on reliable statistics, filtering sentences by their positions and stigma words, falling redundancy using MMR and finally present summary sentences in their chronological order with time remarks.

In paper [6], author proposed a novel query expansion method to solve the problem of information limit in the original query. Proposed query expansion method is added in graph-based algorithm to resolve the problem. To select the query biased informative words from the document set and utilized it as query expansions to enhance the sentence ranking result author utilized the sentence-to-sentence relations and the sentence-to-word relations. Proposed method gains more related information with less noise is main benefit. System performance is enhanced by utilizing the proposed query expansion method.

In paper [7], author exhibits brief overview multi-document summarization system which was designed by Webclopedia team from ISI for DUC and designed based on the fundamentals of Basic Elements. Compare to existing DUC, proposed version of summarizer includes a query-interpretation component that make analysis of the provided user profile and topic narrative for each document cluster before generating an equivalent summary. From evolution perspective a query-interpretation component is dangerous to deal with summarization need for topic based tasks. Proposed system awarded with 4th position on ROUGE-1, 7th position on ROUGE-2and ROUGE-SU4.Assessmentcarried out by utilizing basics elements, among 32 automatic systems proposed system achieved 6^{th}position.

In paper [8], author proposed a Merge split distance for resolving the segmentation problems by integrating various a multi-purposes merge cost function. Proposed approach is basically designed for word spotting on basis of the matching of character features by making use of both of DTW and Merge-Split Edit distance. Functioning provided by proposed system is catering of improper segmented characters underlying the matching process. System depends upon the extraction of words and characters in the text and then attributing each character with a set of features. The characters and words are matched by utilizing the proposed Merge-Split Edit distance algorithm and Dynamic Time Warping (DTW). As compared to the existing work, proposed scheme achieve better performance as query words missed is very less.

In paper [9], author proposed novel approach for multi-document summarization on the basis of graph based approach. A greedy algorithm is used to enforce variety penalty on sentences and the sentences with both high information richness. Finally vital information's are selected to generate summary. Author integrates the diffusion process to achieve semantic relationships between sentences, and then information richness of sentences is calculated by a graph rank algorithm.

In paper [10], author explored overview on how to apply machine learning techniques to design a regression-style sentence ranking scheme for query-focused multi-document summarization. Support Vector Regression (SVR) is used to compute the significance of a sentence in a document set to be summarized by using a set of pre-defined features. From assessment it is conclude that regression models are to be

preferred over classification models to compute the importance of the sentences.

III. IMPLEMENTATION DETAILS

A.System Architecture

In proposed system, multiple documents are taken as input and perform preprocessing of documents with stemming and stopword process. This preprocessing step produces the dictionary words. Next, the bisect k-means clustering is applied on preprocessed data. In clustering step, number of clusters is generated according to field. Clustered documents are merged using f_β optimal merge function. This step finds out the important keywords from each cluster. Then system applies the NEWSUM algorithm to generate the primary summary related to each keyword, till keyword set is empty. At the beginning, system generates the training set with sentence classes by using neural network. The generated Primary summary is tested with training data using neural network. If the sentence belongs to positive class then and only then it is consider as final summary which is more accurate.

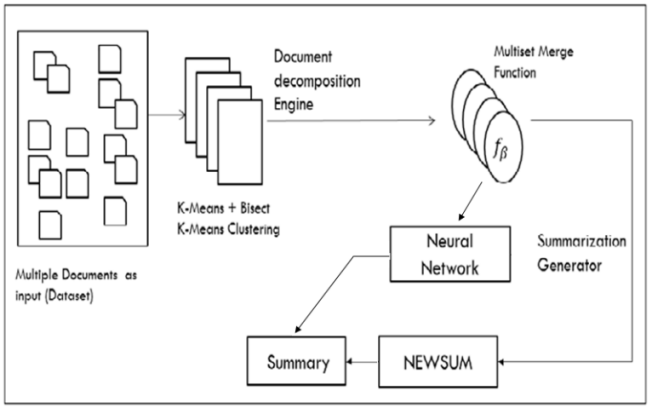


Fig.1: System architecture

B. Algorithm

Algorithm 1: Bisecting K-means Clustering

Input: Document Vectors DV
Number of Clusters k
Number of iterations of k-means ITER

Output: K-Clusters

1. Select a cluster to split (split the largest)
2. Find two sub-clusters by using the basic K-means algorithm
3. Repeat step 2
4. The bisecting step is doing for ITER times and take the split process that generate clustering with the highest overall similarity
5. Repeat steps 1, 2 and 3 till the desired number of clusters k are generated.

Algorithm 2: NEWSUM Algorithm

Assume the key concepts K for a cluster C are known:

1. Procedure SUMMARIZER(C ,K)
2. While K:size != 0 do
3. Rate all sentences in C by key concepts K (1)
4. Select sentence s with highest score and add to S (2)
5. Remove all concepts in s from K (3)
6. End while
7. Return S
8. End procedure

Algorithm 3: Neural Network

Backpropagation Method

Given are the Inputs

$$\{x_1, x_2, \ldots, x_n\},$$

Where x_i is the input for Input layer I, and i=1,2,….,n. J is the hidden layer where Sigmoid Transfer function is used to calculate output of each neuron in hidden layer. K is the output layer. W_{ij} and W_{jk} are weights for the hidden and the output layer.

The sigmoid transfer function is given by :

$$\frac{1}{(1+e^{-input}})$$

Step 1: Run network forward with the input data to get network output.

Step 2: Error value is computed:

$$E \leftarrow \frac{1}{2}(d_k - O_k)^2 + E, for\ k = 1,2, \ldots . K$$

Step 3: Error signal vectors δ_k and δ_j of both layers are computed. Vector δ_k is for output layer, δ_j is for hidden layer. The error signal terms of the output layer in this step are,

$$\delta_k = \frac{1}{2}(d_k - o_k)(1 - o_k 2), for\ k = 1,2, \ldots K$$

The error signal terms of hidden layer in this step are

$$\delta_j = \frac{1}{2}(1 - y_j 2) \sum_{k=1}^{K} \delta_k W_{jk}\ , for\ j = 1,2, \ldots, J$$

Step 4: Output layer weights are adjusted:

$$W_{jk} \leftarrow W_{jk} + \eta \delta_k y_j, for\ k = 1,2, \ldots, K\ and\ j = 1,2, \ldots, J$$

Step 5: Hidden layer weights are adjusted:

$$W_{ij} \leftarrow W_{ij} + \eta \delta_j x_i\ , for\ j = 1,2, \ldots . J\ and\ i = 1,2, \ldots . . I$$

Step 6: Go to step 1

Step 7: The training cycle is completed.

Algorithm 4: Enhanced Summary Generation algorithm

In existing algorithm of summarization i.e. NEWSUM summary contains the number of sentences are equal or less than size of the keyword set.

As per design of NEWSUM algorithm, in each iteration only one sentence is selected and keyword covered in that sentence are removed from the keyword set to reduce redundancy, but from next iteration removed keyword are not considered for the scoring of the sentence therefore there is some possibility to miss sentences which are important than selected sentences in previous iteration. In this paper new algorithm is proposed for summarization which will overcome this issue. Algorithm works in following steps;

Input: Trained dataset

Output: Enhanced Summary

Process:

1. Generate trained dataset file as input for neural network testing phase; for this use all Equations from section C (Equation Used).
2. Use the test dataset as input for Testing and pass the to (Algorithm 3)
3. Get all sentences from test file with relevant and non-relevant class
4. Initialize Enhanced summary = null
5. If(sentence class = relevant class) then add the sentence in Enhanced summary.
6. Else Skip that sentence
7. Return Enhanced Summary.

C. Equations Used

a) Term Feature(f1):

Term Feature (TF) is defined as number of times a term occurs in a sentence

$$TF(S_{i,k}) = \frac{f(t, S_{i,k})}{T(S_{i,k})}$$

Where,

f (t , $S_{i,k}$) is the frequency of each term t in sentence $S_{i,k}$.

$T = Total\ terms$

b) *Sentence Position(f2*):

Sentence position is a sentence location in a paragraph. We assumed that the first sentence of each paragraph is the most important sentence. Therefore, we sort the sentences based on its position.

Sentence position is defined as-

$$SP(S_{i,k}) = \frac{X}{N}$$

Where,

X is the position of the sentence in paragraph,

N is the number of sentences in paragraph

c) Sentence inclusion of name entity (f3):

Usually the sentence that contains more proper nouns is important and it is most probably included in the document summary.

Proper nouns (PN) in the sentence is

$$SPn(S_{i,k}) = \frac{Pn\_Count(S_{i,k})}{Length(S_{i,k})}$$

Where,

Pn_Count is no of nouns contained in sentence,

Length is Total no. of words in sentence $S_{i,k}$,

i is sentence number,

k is no of documents

d) Sentence Length (f4):

This feature is employed to penalize sentences that are too short, since these sentences are not expected to belong to the summary.

Sentence length is defined by

$$SL(S_{i,k}) = \frac{No\ of words\ in\ Sentence}{Unique\ words\ in\ doc(d_k)}$$

Where,d_k

d_k is document no.

e) Final score of each sentence:

$$final_{score}(S_{i,k}) = f_1 + f_2 + f_3 + f_4$$

D. Mathematical Expressions

Merge Function:

Functions that maps multisets of object into single object are called as merge functions. A merge function over a universe U is defined by a function:

1^{st} Order Merge Function: ϖ: μ(U)→U

2^{nd} Order Merge Function: ϖ*: μ(μ(U))→ μ(U)

Local precision and Recall:

Consider a Multiset of sources M=S1, S2… Sn Local precision and recall are defined by functions P*and r* such that:

$$\forall_u \in U : \forall_j \in N : p^*(u, j|M) = \frac{1}{|M|} \sum_{S \in M \wedge S(u) \geq j} M(S)$$

$$\forall_u \in U : \forall_j \in N : r^*(u, j|M) = \frac{1}{|M|} \sum_{S \in M \wedge S(u) \leq j} M(S)$$

f_β-Optimal Merge Function:

Consider a Multiset of sources M=S1, S2, Sn.

$$\varpi^*(M)=\frac{arg\ max}{\zeta \varepsilon M(U)}\ f_\beta(\zeta|M)$$

$$\varpi^*(M)=\frac{\arg\ max}{\zeta \varepsilon M(U)}\ \left(\frac{(1+\beta^2).p(\zeta|M).r(\zeta|M)}{\beta^2.p(\zeta|M)+r(\zeta|M)}\right)$$

β<1, Preference is given to precision.

β>1, Preference is given to recall.

E. EXPERIMENTAL SETUP

The system is built using Java (Version JDK 8) to evaluate the efficiency, effectiveness. The development tool used is NetBeans (Version 8). The experiments performed on 16GB RAM under Windows 8, Intel Core2Duo 2.93GHz. The system requires no any specific hardware to run; any standard machine is capable of running the application. This system takes DUC 2005and News dataset as an input

.

IV. RESULTS AND DATASET

A. Dataset

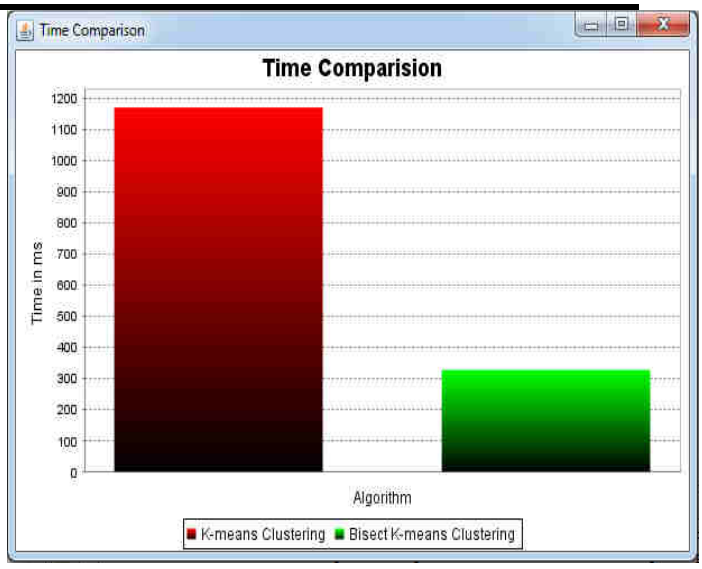
System conduct a large experiment on the Document Understanding Conference (DUC) 2005 dataset, to evaluate the performance of proposed system. In DUC 2005, participants were asked if they would be willing to use. The summary of each topic is included in the sets, use for further evaluation. Two summaries were included in each set as controls manually, and their authors have also rated a set of summaries. There are total 50 topics in DUC 2005.

B. Expert Summary Generation

We generate the Expert Summary using online tool. http://autosummarizer.com/ is used to generate expert summary. This expert summary is compared with summary generated by our proposed approaches. That is expert summary is compared with the summary generated by NEWSUM algorithm and by enhance summary generator. Proposed system is better in terms of efficient and accurate summary generation.

C. *Results and Discussion*

The fig. 2 shows the time graph between k-means clustering and bisect k-means clustering algorithm. The bisect k-means clustering algorithm take less time than the k-means clustering algorithm. The k-means algorithm works on k number of clusters which is time consuming process. But in bisect k-means cluster algorithm the clusters bisect in clusters upto equal result occurred in leaf node. This method saves the time.



Fig,2: Time Graph

The fig. 3 shows the accuracy graph between existing system summary and proposed system summary. The proposed system has more accuracy than the existing system. The neural network find out the –ve and +ve generated summary, gives final enhanced summary.

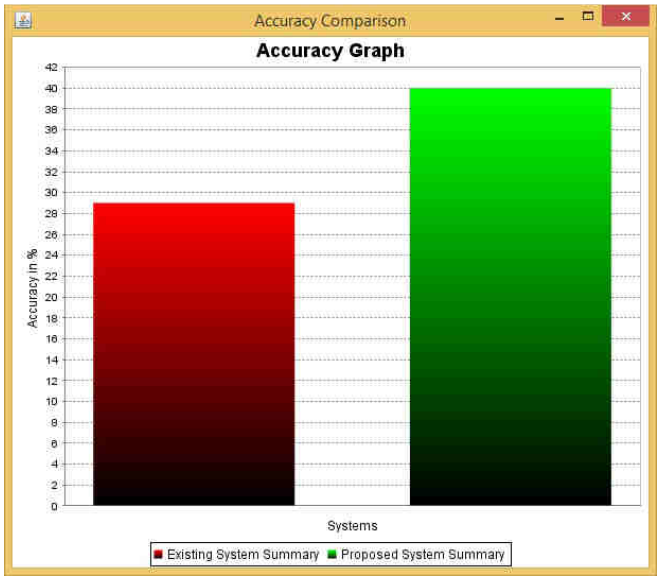


Fig.3: Accuracy Graph

V. CONCLUSION

Multi-document Summarizations schemes are mainly focused in this paper. The main features, the advantages and disadvantages of each system are described. Summarization systems provide the possibility of searching the important keywords of the texts and so consumer will expend less

time on reading the whole document. Thus there is a need to have such System which reduces the large information and generates the summarized result without changing the overall objective of user's search.

In proposed system, user can quickly access correctly-developed summaries. The primary aim of the paper lies in the f_β-optimal merge function, a function recently presented here, that uses the weighted harmonic mean to discover a harmony in the middle of precision and recall. Purpose of Bisect K-means clustering and neural network utilization is to improve the time and accuracy of system.

VI. ACKNOWLEDGMENT

It is our great pleasure to express a deep sense of gratitude to the staff members of Government College of Engineering, Aurangabad for their valuable guidance, inspirations and wholehearted involvement during this research. Their experience, perception and thorough professional knowledge, being available beyond the stipulated period of time for all kind of guidance and supervision and ever-wiling attitude to help, have greatly influenced the timely and successful completion of this implementation work.

REFERENCES

[1] Daan Van Britsom, AntoonBronselaer, Guy De Tr´e, "Using data merging techniques for generating multi-document summarizations", IEEE TRANSACTIONS ON FUZZY SYSTEMS.

[2] A. Farzindar, F. Rozon, and G. Lapalme, "Cats a topic-oriented Multidocumentsummarization system," in DUC2005 Workshop, NIST.Vancouver:NIST, oct 2005, p. 8 pages.

[3] R. M. Aliguliyev, "A new sentence similarity measure and sentencebased extractive technique for automatic text summarization," ExpertSyst. Appl., vol. 36, no. 4, pp. 7764–7772, May 2009.

[4] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries,"in Text Summarization Branches Out: Proceedings of the ACL-04 Workshop,S. S. Marie-Francine Moens, Ed. Barcelona, Spain: Associationfor Computational Linguistics, July 2004, pp. 74–81.

[5] C.-Y. Lin and E. Hovy, "From single to multi-document summarization:a prototype system and its evaluation," in Proceedings of the 40th AnnualMeeting on Association for Computational Linguistics, ser. ACL '02.Stroudsburg, PA, USA: Association for Computational Linguistics, 2002,pp. 457–464.

[6] L. Zhao, L. Wu, and X. Huang, "Using query expansion in graph-basedapproach for query-focused multi-document summarization," InformationProcessing& Management, vol. 45, no. 1, pp. 35 – 41, 2009.

[7] Zhou, Liang, Chin-Yew Lin, and Eduard Hovy. "A BE-based Multi-dccument Summarizer with Query Interpretation." Proceedings of Document Understanding Conference, Vancouver, BC, Canada. 2005.

[8] Khurshid, Khurram, Claudie Faure, and Nicole Vincent. "A novel approach for word spotting using merge-split edit distance." Computer Analysis of Images and Patterns. Springer Berlin Heidelberg, 2009.

[9] X. Wan and J. Yang, "Improved affinity graph based multi-document summarization," in Proceedings of the Human Language TechnologyConference of the NAACL, Companion Volume: Short Papers, ser.NAACL-Short '06. Stroudsburg, PA, USA: Association for ComputationalLinguistics, 2006, pp. 181–184.

[10] Y. Ouyang, W. Li, S. Li, and Q. Lu, "Applying regression models toquery-focused multi-document summarization," Information Processing& Management, vol. 47, no. 2, pp. 227 – 237, 2011.