IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906041 |

A Novel Approach for Text Summarization Using Hybrid Algorithm

Nidhi Mishra, Ms.Neelam Sharma, Mr. Lalitkumar P Bhaiya

M.Tech Student (Sem-4th), Department of Computer Science, Bharti College of Engineering & Technology, Durg, Chhattisgarh, India

Assistant Professor & M.Tech Coordinator, Department of Computer Science, Bharti College of Engineering & Technology, Durg, Chhattisgarh, India

Associate Professor, Department of Computer Science, Bharti College of Engineering & Technology, Durg, Chhattisgarh, India

ABSTRACT: Now days many research is going on for text summarization. Because of increasing information in the internet, these kind of research are gaining more and more attention among the researchers. Extractive text summarization generates a brief summary by extracting proper set of sentences from a document or multiple documents by deep learning. The whole concept is to reduce or minimize the important information present in the documents. The procedure is manipulated by Restricted Boltzmann Machine (RBM) algorithm for better efficiency by removing redundant sentences. The restricted Boltzmann machine is a graphical model for binary random variables. It consist of three layers input, hidden and output layer. The input data uniformly distributed in the hidden layer for operation. The experimentation is carried out and the summary is generated for three different document set from different knowledge domain. The f-measure value is the identifier to the performance of the proposed text summarization method. The top responses of the three different knowledge domain in accordance with the f-measure are 0.85, 1.42 and 1.97 respectively for the three document set.

KEYWORDS: Multi- Document, Summary, Redundancy

I. INTRODUCTION

From many years, summarization is done by humans manually. In the present time, the amount of information is increasing gradually by the mean of internet and by other sources. To overcome this problem, text summarization is essential to tackle the overloading of information. Text summarization helps to maintain the text data by following some rules and regulations for efficient usage of text data. For example, the extraction of summary from a given document for the extraction of a definite content from the whole document or multi-documents. Text summarization relates to the process of obtaining a textual document, obtaining content from it and providing the necessary content to the user in a shortened form and in a receptive way to the requirement of user or application. Automatic summarization is linked closely with text understanding which imposes several challenges comprising of variations in text formats, expressions and editions which adds up to the ambiguities (Sharef *et al*., 2013). Researchers in text summarization have approached this problem from many aspects such as natural language processing (Zhang *et al*., 2011), statistical (Darling and Song, 2011) and machine learning and text analysis is the fundamental issue to identify the focus of the texts.

Text summarization can be classified in two ways, as abstractive summarization and extractive summarization. Natural Language Processing (NLP) technique is used for parsing, reduction of words and to generate text summery inabstractive summarization. Now at present NLP is a low cost technique and lacks in precision. Extractive summarization is flexible and consumes less time as compared to abstractive summarization (Patil and Brazdil, 2007). In extractive summarization it consider all the sentence in a matrix form and on the basis of some feature vectors all the necessary or important sentences are extracted. Afeature vector is an n-dimensional vector of numerical features that represent some object. The main objective of text summarization based on extraction approach is the choosing of appropriate sentence as per the requirement of a user.

Generally, text summarization is the process of reducing a given text content into a shorter version by keeping its

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906041 |

main content intact and thus conveying the actual desired meaning (Mani, 2001a; 2001b). Single document summarization is a process, which deals with a single document only. Multi-document summarization is the method of shortening, not just a single document, but a collection of related documents, into a single summary (Ou *et al*., 2008). The concept looks easy, but while implementation it is a tough task to compile. Sometimes it may not be able to fulfill our desired goal. Most of the similar techniques employed in single-document summarization are also employed in multi-document summarization. There exist some notable disparities (Goldstein *et al*., 2000): (1) The degree of redundancy contained in a group of topically-related articles is considerably greater than the redundancy degree within an article, since each article is appropriate to illustrate the most important point and also the required shared background. So, anti-redundancy methods play a vital role. (2) The compression ratio (that is the summary size with regard to the size of the document set) will considerably be lesser for a vast collection topically related documents than for single document summaries. In order to provide a lot of semantic information, guided summarization task is introduced by the Text Analysis Conference (TAC). It aims to produce semantic summary by using a list of important aspects. The list of aspects defines what counts as important information but the summary also includes other facts which are considered as especially important. Furthermore, an update summary is additionally created from a collection of later Newswire articles for the topic under the hypothesis that the user has already read the previous articles. The summary generated is guided by pre- defined aspects that is employed to enhance the quality and readability of the resulting summary (Kogilavani and Balasubramanie, 2012).

In this study, we have developed a multi-document summarization system using deep learning algorithm Restricted Boltzmann Machine (RBM). Restricted Boltzmann Machine is an advance algorithm based on neural network, it performs the entire necessary task for text summarization. Initially, the preprocessing steps are applied, those steps include (1) Part of speech tagging,

(2) Stop word filtering, (3) steaming. Then comes the feature extraction part. In this part of the text summarization certain features of sentences are extracted. The features we are extracting are: Title Similarity, Positional Feature, Term Weight and Concept Feature. All most all the text summarization models face two major problems, first the ranking problem and the second one is how to create the subset of those ranking or top ranked sentences. There are varieties of approaches for the ranking problem. In this study we are solving the ranking problem by finding out the intersection between the user query and a particular sentence. On the basis of this, a sentence score is generated for every sentence and they are arranged in descending order. Out of this ranked sentences some of sentences are selected on the basis of compression rate entered by the user. In this way we solve the ranking problem. In the end we have used DUC 2002 dataset to evaluate the summarized results based on the measures such as Precision, recall and f-measure.

II. LITERATURE REVIEW

Wafaa et al (2020) Searching the Internet for a certain topic can become a daunting task because users cannot read and comprehend all the resulting texts. Automatic Text summarization (ATS) in this case is clearly beneficial because manual summarization is expensive and time-consuming. To enhance ATS for single documents, this paper proposes a novel extractive graph-based framework “EdgeSumm” that relies on four proposed algorithms.

Alguliyev et al (2019) Text summarization is a process of extracting salient information from a source text and presenting that information to the user in a condensed form while preserving its main content. In the text summarization, most of the difficult problems are providing wide topic coverage and diversity in a summary. Research based on clustering, optimization, and evolutionary algorithm for text summarization has recently shown good results, making this a promising area. In this paper, for a text summarization, a two-stage sentences selection model based on clustering and optimization techniques, called COSUM, is proposed. At the first stage, to discover all topics in a text, the sentences set is clustered by using k-means method.

Shrabanti et al (2018) In the present scenario we are living in a digital media and virtual world. To conveniently communicate in digital world electronic data have to gradually increase. So it is a serious challenge to manage the huge digital and electronic resources efficiently and accurately. One of the important solutions of the above problem is text summarization i.e. an application of text mining. Representing the gist of a text document is called summary.

K. Kurniawan et al (2018) Automatic text summarization is generally considered as a challenging task in the NLP community. One of the challenges is the publicly available and large dataset that is relatively rare and difficult to

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906041 |

construct. The problem is even worse for low-resource languages such as Indonesian. In this paper, we present INDOSUM, a new benchmark dataset for Indonesian text summarization. The dataset consists of news articles and manually constructed summaries.

Mirani et al (2017) People tend to read multiple news articles on a topic since a single article may not contain all important information. A summary of all the articles related to topic will save the time and energy. Text Summarization is a way of minimizing a textual document to a meaningful summary. In this research, an extractive-based approach is used to generate a two-level summary from online news articles.

III. PROBLEM STATEMENT

At present, there is a plethora of data presented by way of Internet and parallel sources. To efficiently tackle the relevant data, there is an indispensable necessity for a device for extracting appropriate group of sentences from the specified documents. Summarization of content is a must. One has to attain ground-breaking data while addressing mammoth group of documents. The onset of World Wide Web with a big bang has affected a sea of change in peoples' lives that it is impossible to spend even a second, devoid of data. It is humanly impossible to commit to memory the ins and outs of each piece of information. With the result, summarization of text documents has begun to play a significant part in information gathering.

A more effective approach is needed for multidocument summarization. The problem is to find feasible methods to determine the most significant sentences in the given documents and to retrieve it without any redundancy so as to improve the quality and readability of the summary. This research work proposes an efficient hybrid approach for multidocument summarization based on deep learning algorithm integrated with fuzzy model and hybrid GA-PSO.

IV. METHODOLOGY

In the proposed method, initially the input documents perform preprocessing and feature vector extraction. The extracted feature vector is given as input to fuzzy model and obtains the optimized feature vector. Based on optimized feature vector, the summary is generated by the proposed system. Block diagram of the proposed approach is given in Figure 5.2. The proposed approach consist of the following process preprocessing, feature vector extraction, fuzzy model optimized by hybrid genetic PSO, deep learning for summarization, summary generation.

- Preprocessing and Feature Vector Extraction

Initially the input documents are subjected to preprocessing phase. These processes are already discussed in the Section 4.6.1. After preprocessing the essential features of the input documents are extracted. This is called as feature vector extraction phase.

- Fuzzy Model based Optimized by Hybrid GA-PSO

The fuzzy model based on hybrid genetic PSO used in the proposed approach performs the following process namely, fuzzfier, fuzzy rule base, inference engine integrated by hybrid genetic PSO, defuzzifier.

- Fuzzifier

The fuzzifier is used to translate the feature vector into linguistic values based on the given membership function which in turn used as linguistic variables for the input. The fuzzier is explained detailed in the section 4.2.

- Fuzzy rule base

It is the most important part in the fuzzy model where the IF-THEN rules are defined. It is explained detailed in the Section 4.6.

- Inference engine base integrated by hybrid GA-PSO

Basically, Inference engine is used to obtain the input values from the fuzzifier which then checks them with the knowledge base to decide the significance of the sentence and to optimize for the getting expected results. So, to reduce the redundancy in the set of rules, an optimization algorithm is used. The optimization algorithm used by the proposed approach is a hybrid of GA and PSO.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906041 |

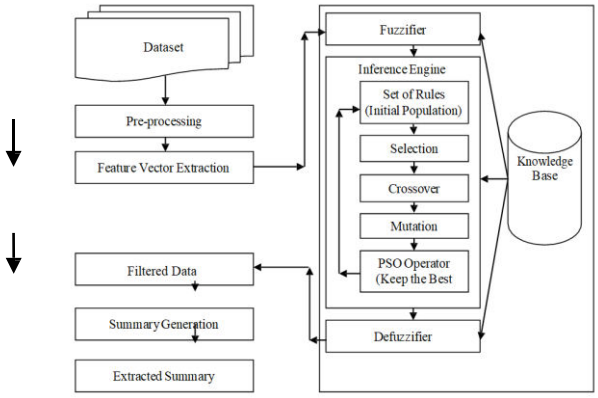

Figure 4.2 Block Diagram of the proposed system

V. RESULT AND DISCUSSION

The experimentation carried out throughout the testing phase is based on a number of different Datasets. On the training data, preprocessing and feature vector extraction are carried out, and the matching feature vector and fuzzy score are produced.

Table 5.1 Average precision, recall and F-Measure using proposed system

Dataset	Precision	Recall	F-Measure
Dataset-1	49.7	48.6	52.3
Dataset-2	50.1	52.6	51.9
Dataset-3	52.3	49.5	48.2
Dataset-4	49.9	50.7	49.3
Dataset-5	51.3	54.1	54.9
Dataset-6	54.9	53.6	55.8
Dataset-7	55.8	54.2	54.8
Dataset-8	53.1	48.1	52.4
Dataset-9	52.1	49.4	53.2
Dataset-10	55.3	53.9	51.4
Average	53.6	52.5	51.6

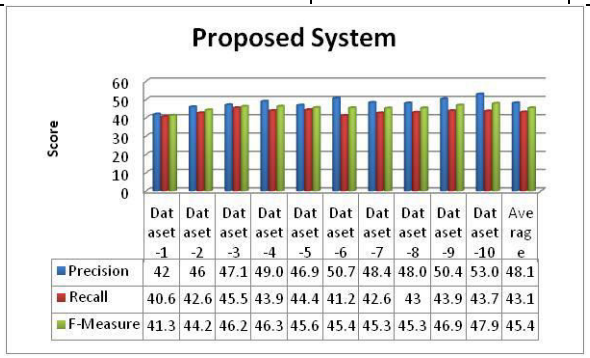

Figure 5.1 Graphical representation of evaluation measures for proposed system

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906041 |

VI. CONCLUSION

Several researches were conducted for summery generation from the multiple documents in recent days. We have developed automatic multi-document summarization system which incorporates the RBM. We have used four different features for feature extraction phase. The feature score of the sentences is applied to the RMB in which the RBM rules are optimized with the help of Deep Learning Algorithm. The features are processed through different levels of the RBM algorithm and the text summary is generated accordingly. The generated result is tested as per the evaluation matrices. The evolution matrices considered in the proposed text summarization algorithm are recall, precision and f- measure. The experimentation of the proposed text summarization algorithm is carried out by considering three different document sets. The responses of three documents sets to the proposed text summarization algorithm are satisfactory. The performance judging parameter f-measure has got values, 0.49, 0.469 and
0.520 respectively for the three document sets. The futuristic enhancement to the proposed approach can done by considering different features and by adding more hidden layers to the RBM algorithm.

REFERENCES

1. Darling, W.M. and F. Song, 2011. Probabilistic document modeling for syntax removal in text summarization. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (CL' 11), ACM Press, Stroudsburg, PA., pp: 642-647.
2. Goldstein, J., V. Mittal, J. Carbonell and M. Kantrowitzt, 2000. Multi-document summarization by sentence extraction. Proceedings of the NAACL-ANLP Workshop on Automatic Summarization, (WAS' 00), ACM Pres, Stroudsburg, PA, USA., pp: 40-48. DOI: 10.3115/1117575.1117580
3. Kogilavani, A. and P. Balasubramanie, 2012. Sentence annotation based enhanced semantic summary generation from multiple documents. Am. J. Applied Sci., 10.3844/ajassp.2012.1063.1070
4. Mani, I., 2001a. Automatic Summarization. 1st Edn., John Benjamins Publishing, Amsterdam, ISBN-10: 9027249865, pp: 285.
5. Mani, I., 2001b. Recent developments in text summarization. Proceedings of the 10th International Conference on Information and Knowledge Management, Nov. 06-11, ACM Press, McLean, VA, USA., pp: 529-531. DOI: 10.1145/502585.502677
6. Ou, S., C.S.G. Khoo and D.H. Goh, 2008. Design and development of a concept-based multi-document summarization system for research abstracts. J. Inform. Sci., 34: 308-326.
7. Patil, K. and P. Brazdil, 2007. Text summarization: Using centrality in the pathfinder network. Int. J. Comput. Sci. Inform. Syst., 2: 18-32
8. Sharef, N.M., A.A. Halin and N. Mustapha, 2013. Modelling knowledge summarization by evolving fuzzy grammar. Am. J. Applied Sci., 10: 606-614. DOI: 10.3844/ajassp.2013.606.614
9. Zhang, Y., D. Wang and T. Li, 2011. iDVS: An interactive multi-document visual summarization system. Mach. Learn. Know. Disco. Databases, 6913: 569-584. DOI: 10.1007/978-3-642-23808-6_37

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details