DOI: http://dx.doi.org/10.26483/ijarcs.v9i3.6037

ISSN No. 0976-5697

Volume 9, No. 3, May-June 2018

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

APPLICATIONS OF TEXT SUMMARIZATION

Debabrata Khargharia
Department Of Computer Science & Engineering
DUIET, Dibrugarh University
Dibrugarh, 786004, India

Nomi Baruah
Department Of Computer Science & Engineering
DUIET, Dibrugarh University
Dibrugarh, 786004, India

Nabajit Newar
Department Of Computer Science & Engineering
DUIET, Dibrugarh University,
Dibrugarh 786004, India

Abstract: As data is accessible in abundance for each point on web, gathering the critical data as rundown would profit various clients. Thus, there is developing enthusiasm among the examination group for growing new ways to deal with consequently summarise the content. Automatic text summarisation framework creates a rundown, i.e. short length message that incorporates all the essential data of the archive. Outline can be created through extractive and additionally abstractive strategies. Abstractive strategies are profoundly unpredictable as they require broad regular dialect preparing. Accordingly, look into group is concentrating more on extractive rundowns, attempting to accomplish more lucid and significant outlines. Amid 10 years, a few extractive methodologies have been created for programmed outline age that actualizes various machine learning and enhancement procedures. This paper introduces some of the different applications where text summarisation are used.

I. INTRODUCTION

Text Summarization is the way toward distinguishing the most essential important data in a record or set of related archives and compacting them into a shorter rendition saving its general implications.There are three main steps for summarizing documents. These are topic identification, interpretation and summary generation. First of all, Topic Identification; here the most important information in the text is identified. There are different techniques for topic identification are used which are Position, Cue Phrases, word frequency. Secondly, Interpretation; here step, different subjects are fused in order to form a general content. Thirdly, Summary Generation; here the system uses text generation method. Automatic text summarization framework produces rundown, i.e., condensed type of the record that contains a couple of critical sentences chose from the document. In late fifties, text summarization started and till now, there is awesome change in this field. A large number of procedures and methodologies have been created in this field of research. A summary produced by a Automatic text summarizer should comprise of the most important data in a report and in the meantime, it ought to possess less space than the first record. There are many issues like excess, worldly measurement, co-reference, sentence requesting, and so forth that need specific consideration while outlining various number of archives, along these lines, making this undertaking more unpredictable

II. TEXT SUMMARIZATION: APPLICABILITY

Programmed/Automatic a piece of machine learning and information mining. The fundamental thought of synopsis is to discover a subset of information which contains the "data" of the whole set. Such systems are broadly utilized as a part of industry today. Inquiry is a case; others incorporate outline of records, picture accumulations and recordings.

Text Summarization is used in many areas. It generally gives a rough idea of what is in the document, without reading the whole document. Some of the areas on which we are doing our survey are as follows:

1. Tourism
2. Legal Document
3. Stories
4. Medical Documents
5. Biomedical Documents
6. News

1. Tourism Text Summarization

Xiao Wu et. al. proposed a paper on Personalized Multimedia Web Summarizer for Tourist [1]where theyfeature the utilization of media innovation in creating characteristic rundowns of tourism related data. The framework uses a computerized procedure to accumulate, channel and order data on different traveler spots on the Web. The final product present to the client is a customized mixed media outline created as for clients' inquiries loaded with content, picture, video and continuous news made retrievable for cell phones. Preliminary investigations exhibit the prevalence of our introduction plot over customary strategies.

Nyaung and Thein proposed [2] that due the quick increment of Internet, web conclusion sources progressively develop which is helpful for both potential clients and item makers for expectation and choice purposes. These are the client created substance written in characteristic dialects and are sans unstructured writings conspire. Hence, conclusion mining procedures wind up noticeably prevalent to naturally process client surveys for extricating item highlights and client sentiments communicated over them. Since client surveys may contain both stubborn and accurate sentences, an administered machine learning procedure applies for subjectivity characterization to enhance the mining execution. In this paper, we devote our work is the undertaking of assessment synopsis. Along these lines, item highlight and sentiment

extraction is basic to supposition outline, since its adequacy altogether influences the ID of semantic connections. The extremity and numeric score of the considerable number of highlights are controlled by Senti-WordNet Lexicon. The issue of feeling outline alludes how to relate the supposition words regarding a specific element. Probabilistic based model of managed learning will enhance the outcome that is more adaptable and viable.

2. Legal Document Summarization

Saravanan, Ravindran and Raman proposed on Legal Document Summarization as a Conditional Random Field (CRF) is applied to segment a given legal document into seven labelled components and each label represents the appropriate rhetorical roles. Feature sets with varying characteristics are employed in order to provide significant improvements in CRFs performance. The framework is then advanced by the utilization of a term conveyance demonstrate with organized area information to extricate key sentences identified with explanatory classifications. The last organized synopsis has been seen to be nearest to 80% exactness level to the perfect rundown created by specialists in the territory. Content rundown is one of the important issues in the data time, and it should be understood given that there is exponential development of information. It tends to the issue of choosing the most imperative bits of content and that of creating intelligible synopses. [3]

Selvani et. al. proposed on “An Automatic Legal Document Summarization and Search Using Hybrid System” [4] as a technique, which rewrites the original text into a shorter version by replacing the wordy concept with shorter ones. It reduces the sentences from the original text without changing their meaning and Extract summarisation is a technique, which reuses the most important sentences form the original text for text summarization by considering existing words, sentences etc. from the original text to form summary. A summary of a judgement helps in organizing a large volume of documents and finding the relevant judgements for their cases. For this reason, the information frequently summarized by legal experts. But due to manual summarization by legal experts it requires much human time and expertise to provide manual summaries for legal documents. In automatic legal document summarization, by using extraction technique it extracts the main points from the legal document and provides the summary. So that it saves the time to provide summaries of legal documents and a lawyer can spends less time on reading the whole document.

Kanapala, Pal and Pamula on “Textsummarization from legal documents: a survey” [5] proposed enormous amount of online information, available in legal domain, has made legal text processing an important area of research. In this paper, they attempt to survey different text summarization techniques that have taken place in the recent past and put special emphasis on the issue of legal text summarization. They began with general prologue to content outline, quickly touch the current advances in single and multi-report rundown, and afterward dig into extraction based lawful content synopsis and furthermore talk about various datasets and measurements utilized as a part of outline and look at exhibitions of changed methodologies, first by and large and after that centered to legitimate content.They additionally said features of various rundown systems and quickly secured a couple of programming devices utilized as a part of legitimate content synopsis lastly finish up with some future research headings.

3. Story Summarization

A large portion of the story summarization investigate conveyed out to date has been worried about the summarization of short archives (e.g., news stories, specialized reports), and practically nothing work if any has been done on the summarization of long archives. Mihalcea and Ceylan proposed a paper on “Explorations in Automatic Book Summarization” [6] to address this hole and investigate the issue of book summarization. The informational index particularly composed for the assessment of frameworks for book synopsis, and portrays rundown strategies that expressly represent the length of the archives. A major difference between short and long documents stands in the frequent topic shifts typically observed in the later. While short stories are usually concerned with one topic at a time, long documents such as books often cover more than one topic. Thus, the intuition is that a summary should include content covering the important aspects of allthe topics in the document, as opposed to only generic aspects relevant to the document as a whole. A system for the summarization of longdocuments should therefore extract key concepts from all the topics in the document, and this task is better performed when the topic boundaries are known prior to the summarization step.

Wong on “Automatic News Summarization and Extraction System” [7] gave the client a short outline of the story. This enables the client to choose whether a video cut returned by the internet searcher is significant to the point he/she is hunting down. A content outline procedure called 'Lexical Chain' is connected to abridge the content. Aside from an outline of the first story, it could likewise endeavor to recognize vital catchphrases. Such catchphrases, if distinguished effectively, would contribute gigantic add up to demonstrate what the theme of the story is about. These watchwords were alluded as 'Key Entities'.For both story division and synopsis, it utilized a content preparing framework called GATE [8], created by Sheffield University to remove these catchphrases. Exactness of the framework has turned out to be high, and it is one of the fundamental parts which give the strength of this division procedure.

The work of Kazantseva on “Automatic Summarisation of Short Fiction” [9] is an investigation into programmed outline of shy of fiction. It shows a framework that makes rundowns out of artistic short stories utilizing two kinds of in-arrangement: data about elements integral to a story and data about the linguistic part of provisions. The rundowns are custom fitted to a particular reason: helping a peruser choose whether she would be occupied with perusing a specific story. They contain simply enough information to empower a peruser to shape sufficient assumptions about the story, however they don't uncover the plot. As indicated by these criteria, an objective rundown furnishes a peruser with a thought of whom the story is about, where and when it happens (in a way that goes past basically posting names and places) however does not re-recount the occasions of the story. Keeping in mind the end goal to fabricate such summaries, the framework endeavors to distinguish sentences that meet two criteria: they concentrate on fundamental elements in the story

and they relate the foundation of the story instead of occasions. Talking about the criteria for the sentence choice process involves an extensive piece of this paper. These criteria can be generally partitioned into two classifications: 1) data about principle elements (e.g., fundamental characters and areas) and 2) data identified with the syntactic part of conditions. By depending on this data the framework chooses sentences that contain important data to the setting of the story.

4. Medical Document Summarization

Afantenos, Karkaletsis and Stamatopoulos recommended that during the most recent decade, records outline got expanding consideration by the AI look into group. All the more as of late it likewise pulled in light of a legitimate concern for the therapeutic research group too, because of the gigantic development of data that is accessible to the doctors and scientists in medication, through the extensive and developing number of distributed diaries, gathering procedures, medicinal locales and entries on the World Wide Web, electronic restorative records, and so forth. This overview gives initial a general foundation on records outline, showing the components that synopsis relies on, talking about assessment issues and portraying quickly the different kinds of rundown procedures. It at that point analyzes the qualities of the restorative area through the distinctive kinds of therapeutic reports. At long last, it introduces and talks about the synopsis methods utilized so far in the medicinal space, alluding to the relating frameworks and their attributes. The paper talks about completely the promising ways for future research in therapeutic archives rundown. It for the most part centers around the issue of scaling to vast accumulations of reports in different dialects and from various media, on personalization issues, on versatility to new sub-spaces, and on the coordination of synopsis innovation in commonsense applications. [10]

Gayathri and Jaisankar proposed [11] on the huge amount of information available and dealt with providing condensed version of the original document and presented an extractive informative single medical document summarization approach. Here the area particular vocabulary words are utilized as prompt words. Pre-processing is done to remove the unuseful information from the info archive. The pre-handled archive is utilized for positioning. Keeping in mind the end goal to rank each sentence, the score is figured for each sentence in view of the quantity of events of the signal words. The signal word score is figured for each prompt word in the learning base. At long last, the sentences are positioned in light of their sign word recurrence. The proposed framework creates a last rundown in light of the pressure proportion with a specific end goal to limit the length of the outline delivered. Since the comparability measure is utilized as a part of the last outline creation, the framework incorporates profoundly different sentences and delivers a more useful last synopsis. The framework is tried with the contribution of 100 medicinal articles taken from different therapeutic sources. The articles with writer composed edited compositions are considered for testing. The outcome demonstrates that the proposed approach performs better contrasted with the current summarizers. The execution of the framework is additionally estimated as for the nature of the outline created by utilizing ROUGE. [12]

Elhadad proposed [13] a user-sensitive approach to text summarization. One domain which would highly benefit from creating summaries to both individual and class-based user characteristics is the medical domain, where physicians and patients access similar information, each with their own needs and abilities. Their framework is a medical digital library for physicians and patients where they describe a summarizer, which generates summaries of findings in an input set of clinical studies. When a physician is treating a specific patient, he's looking for information relevant to the patient's history and problems. The summarizer takes the user's interests into account and presents only the findings pertaining to a user model, as approximated by an existing patient record. The same synthesis of information can also be of interest to the patient. The summarizer predicts which medical terms used in a text will be too technical for patients, and augments it with appropriate definitions when necessary.

5.Biomedical Text Summarization

Mishra et. al. proposed the measure of data for clinicians and clinical scientists is developing exponentially. Content outline lessens data as an endeavor to empower clients comprehends pertinent source messages all the more rapidly and easily. Lately, considerable research has been led to create and assess different outline procedures in the biomedical area. The objective of this investigation was to methodically survey late distributed research on rundown of printed archives in the biomedical space. Of 10,786 investigations recovered, 34 (0.3%) met the consideration criteria. Characteristic dialect preparing (17; half) and a mixture procedure involving factual, Natural dialect handling and machine learning (15; 44%) were the most widely recognized outline approaches. Most investigations (28; 82%) led an inborn assessment. Late research has concentrated on a cross breed method including measurable, dialect preparing and machine learning procedures. Additional research is required on the application and assessment of content rundown in genuine research or patient care settings. [14]

Schulze and Neves proposed [15] expanding measure of biomedical data that is accessible for specialists and clinicians makes it harder to rapidly locate the correct data. Automatic summarization of different writings can give rundowns particular to the client's data needs. It investigates the utilization named-entity recognition for graph based summarization. It extended the LexRank check with data about named substances and present EntityRank, a multi-record chart based rundown assuming that is just in context of named parts.

6.News Summarization

McKeon did a research [16] on News Summarization. He proposed redundant text collections, such as the web, creates both problems and opportunities for natural language systems. On the one hand, the presence of numerous sources conveying the same information causes difficulties for end users of search engines and news providers; they must read the same information over and over again. Their examination on outline of multilingual news expects us to manage loud info; we depend on cutting edge machine interpretation frameworks

and utilize data that is accessible at the season of rundown to enhance the familiarity of the synopsis and furthermore moved to outline of other media, including email and gatherings. The first-level summary generates the summary of each article on all these topics. Sentiment Analysis is performed on the first-level summary to understand the variation in related news articles from different news agencies. The second-level summary generates the summary of the combined first-level summaries of two/three related articles on a topic. The ROUGE metric is used to evaluate the performance of summarization [17].

Wong proposed [18] to build a system which address summarization at a multimedia level. She used lexical chain analysis to do the work on news summarization. Their aim is to build an automatic news summarization system. It records broadcast television news, analysis the content to identify news stories. The text of each story is summarized and important keywords are extracted. This information is to be stored in a central database, and an Information Retrieval system is to be implemented which let users to search for any piece of news in the database. Such a framework have advantage over other web crawlers, as we utilize synopsis strategies to feature the most essential data to the client, empowering him/her to find the story he/she is searching for in a shorter time, when contrasted with a normal content based web index.

III. CONCLUSION

This study paper is focusing on various uses of text summarization. The significance of sentences is chosen in light of measurable and semantic highlights of sentences. Numerous uses of the content outline have been attempted in the past couple of decades. Be that as it may, it is difficult to state how much more prominent interpretive modernity, at sentence or content level, adds to execution. Without the utilization of content synopsis, the produced rundown may experience the ill effects of absence of union and semantics. In the event that writings containing numerous points, the produced rundown won't not be adjusted. Choosing appropriate weights of individual highlights is imperative as nature of definite rundown is relying upon it. The greatest test for content rundown is to outline content from various printed and semi organized sources, including databases and website pages, in the correct way (dialect, arrange, estimate, time) for a particular client. The content rundown programming should deliver the successful synopsis in less time and with slightest repetition.

IV. REFERENCES

[1]. Personalized Multimedia Web Summarizer for Tourist , Xiao Wu et. al. , Key Laboratory of Intelligent Information Processing Institute of Computing Technology, CAS, Beijing, China , April 21-25,2008

[2]. Dim En Nyaung, Thin Lai Lai Thein . Feature-Based Summarizing and Ranking from Customer Reviews . World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:9, No:3, 2015

[3]. Improving Legal Document Summarization Using Graphical Models. (PDF Download Available). Available from:https://www.researchgate.net/publication/220809892_Improving_Legal_Document_Summarization_Using_Graphical_Models [accessed Dec 16 2017].

[4]. Selvani Deepthi Kavila, Vijayasanthi Puli, G.S.V. Prasada Raju, and Rajesh Bandaru. An Automatic Legal Document Summarization and Search Using Hybrid System. Department of Computer Science and Engineering, Anil Neerukonda Institute of Technology and Sciences, Sangivalasa, Visakhapatnam, AP, India . 2013

[5]. Ambedkar Kanapala, Sokumal Pal, Rajendra Pamula. Text summarisation from legal documents:a survey. Artificial Intelligence Review. 2017

[6]. Rada Mihalcea and Hakan Ceylan. Explorations in Automatic Book Summarization. Department of Computer Science. University of North Texas. January 2014

[7]. Lawrence Wong. Automatic News Summarization and Extraction System. MEng Computing. Imperial College Dept. of computing. Accessed on 02.01.2018

[8]. GATE, Generic Architecture of Text Engineering - http://gate.ac.uk/

[9]. Anna Kazantseva. Automatic Summarisation of Short Fiction. University of Ottawa. December 2006.

[10]. Summarization from Medical Documents: A Survey. Available from: https://www.researchgate.net/publication/220103096_Summarization_from_Medical_Documents_A_Survey [accessed Dec 16 2017].

[11]. Gayathri, P., N. Jaisankar. Towards an Efficient Approach for Automatic Medical Document Summarization.School of Computing science and engineering . VIT university. 2015

[12]. ROGUE, Recall- Oriented Understudy for Gisting Evaluation. http://www.berouge.com

[13]. Noemie Elhadad.User-Sensitive Text Summarization: Application to the Medical Domain.Columbia University.2006

[14]. https://doi.org/10.1016/j.jbi.2014.06.009 [accessed Dec 16 2017].

[15]. Frederik Schulze and Mariana Neves. Entity-Supported Summarization of Biomedical Abstracts.Osaka, Japan, December 12th 2016.

[16]. Columbia University, Text summarization: News and Beyond, 2014

[17]. Networks & Advances in Computational Technologies (NetACT), 2017

[18]. Lawrence Wong. Automatic News Summarization and Extraction System. MEng Computing. Imperial College Dept. of computing. Accessed on 02.01.2018