2018 2nd International Conference on Data Science and Business Analytics

Text Summarization: A Review

Sreyasi Biswas
Department of Computer Science &Engineering
Siksha 'O' Anusandhan Deemed to be University, Odisha, India
shreyasibiswas1241996@gmail.com

Rasmita Dash
Department of Computer Science &Engineering
Siksha 'O' Anusandhan Deemed to be University, Odisha, India
rasmitadash@soa.ac.in

Rasmita Rautray
Department of Computer Science &Engineering
Siksha 'O' Anusandhan Deemed to be University, Odisha, India
rashmitaroutray@soa.ac.in

Rajashree Dash
Department of Computer Science &Engineering
Siksha 'O' Anusandhan Deemed to be University, Odisha, India
rajashreedash@soa.ac.in

***Abstract*— In today's era, there are large numbers of documents or information that is present related to any particular field. There are many sources out of which we can gather a lot much information that will be pertinent to our field of search. Much information is available at various sources like the internet or various books. But, as we know that a huge amount of information cannot be always considered or taken into use. So, a precise amount of information is always considered and that information is drawn out from the original document that is huge in size. In other words, we can say that we pluck out the summary of the main document. A summary of any document is defined as a collection of essential data by collecting the brief statements accounting the main points of the original document. Therefore, Summarization of a text is a procedure of separating or getting the relevant data out of a very large document. It is the process of shortening the text document by using various technologies and methodologies to create a coherent summary including the major points of the original document. There are various methods by which the summarization process can be carried out.**

Keywords—Text summarization, summary types, summarization methods, summary evaluation.

I. INTRODUCTION

Before moving on to the text summarization, we first have to know that what a summary is. A summary is a short but brief description of a document that gives the main facts or ideas about the original document. Preparing a summary of any document is very essential. This is because it can gain access to much important information and can control the flood of information that is available in various sources [1]. It is of no use to read or to go through what is useless or less important. Summaries also save the reader's time as summaries only provide the main content and the overview of the document. Many digital libraries provide many a huge amount of information and for a researcher it is very tiresome to go through all the pages or all the information. So, a brief and a concise summary of the entire thesis are presented which helps in conquering the vast information. Text Summarization is a process of drawing out essential information out of a large document or record by undergoing various processes and representing those extracted materials in the form of a summary. The aim of text summarization is to present the original text into a brief document by using various semantics[2]. Text summarization process can be classified into two main categories such as extractive and abstractive text summarization.

There are two different groups of text summarization. It can be grouped into: indicative and informative text summarization

A. Extractive Text Summarization

In this method, only the main information from the document is drawn out and then presented in the form of a summary. Here, the main contents, statements, paragraphs, etc are selected and after that those selected sentences and paragraphs are concatenated to form a summary. Major points of the main document are found out and then those points are joined to form the concise summary out of the large amount of the information present in a particular document.

The extractive text summarization approach can also be divided into two parts:

Pre-Processing step – In this process, usually the original document is represented in the structured form.
It mainly includes:

- Sentences Boundary Identification – Whenever a dot is present at the end of any particular sentence, then the sentence boundary is identified.
- Stop-Word Elimination – These are the common words in a text with no semantics.
- Stemming – The goal of stemming is obtaining the stem or the radix of each word, emphasizing its semantics.

Processing step – In this process, based on the relevance and the importance, the statements are selected from the original text, and then the selected sentences are added in the summary. Weights are also assigned to the sentences by using the weight learning method and then the final score of the sentences are calculated by using Feature-Weight equation and then the top ranked sentences are included in the summary.

B. Abstractive Text Summarization

In this method, the main concept of the original document is first understood. After going through the complete document, the major purpose and the process should be understood and then the summary is to be formed completely in a natural language.

978-1-5386-8431-3/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDSBA.2018.00048

C. *Indicative Text Summarization*

In this method, the main concept of the original document is first understood. After going through the complete document, the major purpose and the process should be understood and then the summary is to be formed completely in a natural language.

D. *Informative Text Summarization*

In this method, the main concept of the original document is first understood. After going through the complete document, the major purpose and the process should be understood and then the summary is to be formed completely in a natural language.

II FEATURES OF TEXTS IN SUMMARIZATION

Text Summarizers are utilized to distinguish the key sentences from the first content and afterwards those sentences are extracted and concatenated to form a summary. Lists of text summarization features are described below that are required to identify the key sentences from the original document.

TABLE 1 DESCRIPTION OF TEXT FEATURES

FEATURES	DESCRIPTION
Term-Frequency	Imperative terms given by various insights depend on the term recurrence. Notable terms are those terms that happen as often as possible or many numbers of times in a particular document. The score of the sentences increases as the frequency of any particular word increases. The widely used measure that is used to calculate the word frequency is TF IDF.
Title/Headline Word	In a summary, it becomes necessary to provide a suitable title or a heading for the particular summary that is produced from the original text or any document. So, the heading word should always be positively related to the summary. The words that are included as the title or the heading word should be related or should indicate the topic or the subject of the document.
Similarity	Similarity can be calculated by various expressive knowledge. Likeness between the sentence of the document as well as the heading of the document and the likeness between two sentences in a document can be indicated by this text summarization feature.
Proper Noun	In summarization of a document, the sentences having proper nouns are very important and should be considered as an essential statement to be included in the summary. For example- name of a person, place or any organization.
Location	While forming a summary, it should always be kept in mind that the important sentences should always be included in a summary and to do this one should also know that in the original document, where these important statements are located such as at the beginning of the paragraph or at the end. In maximum cases, the first and the last sentences of a paragraph of the original document are included in the summary.
Cue Method	There are many positive and negative words that are present in the document from which the summary is to be extracted. By this method, the consequence of the constructive or the negative term used in the sentence is measured and the importance of the sentence is also measured. Many key words or cues are also there which can be used such as "in summary", "in conclusion", "the paper describes", etc.
Sentence Length	This feature helps in maintaining the size of the summary. It should always be kept in view that excessive long or short sentences are not reasonable for the formation of a summary.
Proximity	This feature is responsible for viewing the distance between the text units in the summary. It determines that sufficient distance should be there between two units so that there difference can be measured.

III STEPS OF TEXT SUMMARIZATION

Text Summarization includes various methods as well as there are various steps that must be followed for the summarization. Summarizing of any document consists of three main steps. They are as follows:

A. *Topic Identification*

The most useful and prominent information in the text is identified over here. Different techniques are used for the identification of a topic. Some of the techniques include Word Frequency, Cue Phrases, and Positions. Methods that are based on the position of phrases are usually considered to be the most useful methods for topic identification.

B. *Interpretation*

The summaries that are abstract need to go through this interpretation step. In this step, different contents are concatenated to in order to form a generalised summary.

C. Summary Generation

The text generation method is used by the system in this step.

IV METHODOLOGIES USED IN TEXT SUMMARIZATION

There are two basic approaches by which text summarization method can be carried out. They are – Abstractive and Extractive. Each kind of approach or method is discussed.

A. Abstractive Summarization approach

Summarization that is carried out using abstractive summarization approach is classified into two categories.

TABLE II DESCRIPTION OF ABSTRACTIVE SUMMARIZATION APPROACH

METHODS	DESCRIPTION	ADVANTAGES	DISADVANTAGES	REFERENCES
Tree Based Method	A dependency tree is used to represent the text of any document. It uses some algorithm for the complete generation of any summary.	It usually works on some of the units of the document by which the summary can be read and understood easily.	Some of the modals are lacked in this method which is helpful in representing the summary abstractly.	[1]
Template Based Method	Some kinds of templates are used in the representation of an entire document. Various patterns and the extraction rules are used in mapping to the template.	The kind of summary that is generated that is highly coherent because the summary that is produced is extracted using relevant items and information.	Designing of templates is required in this method and on the other hand generalising those templates is also very difficult.	[2]
Ontology Based Method	This method uses ontology method also known as knowledge based method so that the process of summarization can be improved. In many cases fuzzy ontologies are also used in handling the uncertain data which cannot be handled by simple domain ontology.	Due to this knowledge based model, drawing relations among the context becomes very easy.	In many cases, creating some of the rule based methods can be very complex sometimes.	[3].
Lead and Body Phrase Method	In this method, various operations are performed on the phrases. Operations like insertion and substitution are performed. These are performed in order to rewrite the lead sentences in the summary.	Revision of the lead sentences can be done when this method is used in text summarization process.	Lacks some of the modals that is useful in representing the summary in the abstract form.	[2].
Multimodal Semantic Model	A model called the semantic model which keeps in view the concepts and the relationship among the concepts so that the contents of the semantic model can be represented.	This model produces an abstract summary because it has got an excellent coverage and the summary contains much graphical content.	This model has a disadvantage. It is that the evaluation in this method done manually.	[2]
Information Item Based	The content of the summary that is produced from this method is the generalization	The main advantage of this summary is that it produces a concise	Due to the difficulty faced in creating meaning and	[5].

System	of the abstract representation of the original document. They are not generated from the sentences of the original documents.	summary that is richer in information and lesser in redundancy.	grammatically correct sentences, it is often rejected.	
Semantic Graph Based Method	Here, in this method the summary is formed by creating some rich semantic graph.	It produces summary that is concise and coherent and that are error free and redundant less.	It is limited to single document abstractive summarization.	[2][3]

TABLE III DESCRIPTION OF EXTRACTIVE SUMMARIZATION APPROACH

METHODS	DESCRIPTION	REFERENCES
Term Frequency-(Inverse Document Frequency Method)	Recurrence of sentences is characterized by the quantity of the sentences in the record which contains that term. At that point vectors of the sentences are identified by similitude of the question and also the most scoring sentences are picked to be a piece of the rundown.	[2]
Cluster Based Method	It is obvious to imagine that synopses should always be addressed as distinctive "topics" appearing in the reports. If the accumulation of the records for which outline is being given is of very surprising themes, archive bunching turns out to be relatively fundamental to create an important rundown. Sentence choice depends on comparability of the sentences to the topic of the group (Ci). The following component that is area of the sentence in the report (Li). The last factor is its similitude to the main sentence in the report to which it has a place (Fi). Si =W1 * Ci + W2 * Fi+ W3 *Li Where, W1, W2, W3 are weight age for consideration in rundown. The bunching k-implies calculation is connected.	[6]
Text Summarization using Neural Networks	This method includes preparation of the neural systems to include in the sorts of sentences that should definitely be incorporated into a rundown. It uses a three-layered Feed Forward neural system	[5]
Graph Theoretic Approach	Chart theoretic portrayal of entries is given by a strategy for distinguishing the actual proof of topics. After pre-handling steps, specifically evacuation of stop words; sentences contained in the reports are taken as main parts in an undirected diagram.	[4],[5].

The methods that are used in both the categories are discussed in TABLE I.

B. Extractive Summarization Approach

This approach on selecting important sentences, paragraphs, phrases from the main document and then those selected sentences are joined together to form a concise and a coherent summary. And on the other hand the sentence's importance is measured with the help of some knowledge based features. A brief description of all the methods that are included in the extractive summarization approach is discussed in TABLE II.

V EVALUATING THE SUMMARIZATION SYSTEMS

Rundown assessment is an imperative angle for content outline. By and large, outlines could be identified using the characteristic measures. Though inborn strategies endeavour to quantify outline quality utilizing human evaluation and incidental methods measure the equivalent by an assignment which is based upon the execution measure like the recovery of data. Assessment strategies are helpful in assessing the

handiness and trustfulness of the outline. Assessing the characteristics like fathom ability, rationality, and meaningfulness is extremely troublesome. Framework assessment may be performed manually (gold standard) by specialists .To measure the nature of summary, the physically master framework is utilized. The subjective assessment is finished by checking the quantities of sentences chosen by the framework that match with the human highest quality level. To measure the quantitative appraisal of the rundown the ROUGE evaluator apparatus is utilized which comprise of exactness, review and F-measure.

VI CONCLUSION

Content outline is developing as sub – part of NLP as the interest for compressive, important, and unique of theme because of substantial measure of data accessible on net. Exact data looks all the more successfully and effectively. In this way message rundown is require and utilized by business expert, showcasing official, advancement, analysts, government associations, understudies and instructors too. It is observed that official needs rundown such that at a restricted time consumed data could be prepared. This paper takes into account the points of interest of both the extractive as well as abstractive methodologies alongside those systems utilized, their execution accomplished, alongside focal points and burdens of each methodology. Content rundown has its significance in both business and in addition inquire about network. As abstractive synopsis requires additionally learning and thinking, it is bit complex then extractive methodology at the same time, abstractive rundown gives more important and fitting outline contrast with the extractive approach. Undergoing the investigation, it's likewise seen that lesser amount of work is finished utilizing abstractive strategies, there's a great deal of extension for investigating such techniques for more fitting rundown.

REFERENCES

[1] Saranyamol C S, Sindhu L, "A Survey on Automatic Text Summarization", International Journal of Computer Science and Information Technologies, 2014, Vol. 5 Issue 6.

[2] Fachrurrozi M., Yusliani Novi, and Yoanita Rizky Utami, "Frequent Term based Text Summarization for Bahasa Indonesia", International Conference on Innovations in Engineering and Technology Bangkok (Thailand), 2013.

[3] Ragunath R. And Sivaranjani N., "Ontology Based Text Document Summarization System Using Concept Terms", ARPN Journal Of Engineering And Applied Sciences, 2015, Vol. 10, No. 6.

[4] Rada Mihalcea, Niraj Kumar, Kannan Srinathan and Vasudeva Varma, (2013)

[5] Sarda A.T. and Kulkarni A.R., "Text Summarization using Neural Networks and Rhetorical Structure Theory", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, 2015.

[6] Deshpande Anjali R., Lobo L. M. R. J., "Text Summarization using Clustering Technique", International Journal of Engineering Trends and Technology (IJETT), 2013, Vol. 4 Issue8.

[7] Patil Pallavi D., Kulkarni N.J., "Text Summarization Using Fuzzy Logic", International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2014, Vol. 1 Issue 3.