International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Review Paper on Automatic Text Summarization

Ujjwal Rani[1], Karambir Bidhan[2],

[1]*M.Tech, CSE Department, UIET, Kurukshetra University , Kurukshetra , India*
[2]*Assistant Professor, CSE Department, UIET, Kurukshetra University , Kurukshetra , India*

Abstract

Content summarization is the way toward shortening the source archive into dense structure keeps generally thought regarding the record. The systems of content summarization are abstractive and extractive. The abstractive summarization requires characteristic language preparing devices for outlining the records. The extractive summarization requires factual, etymology and heuristics strategies for positioningthe sentences. Numerous strategies have been produced for the summarization of content in different dialects. This paper examine about the strategies for abstractive & extractive text summarization.

Key Words: Natural Language Processing (NLP), Text Summarization Techniques, Extractive summarization, Abstractive summarization.

1. INTRODUCTION

At present, fast development of cumbersome measure of data is required to process, store, and oversee proficiently. Some of the time, it is so hard to locate the right data from tremendous measure of put away information or large information archives. In the time of technologies like big data, fast growing of data in various formats whether textual or graphical have the capability to mine the information and hence data mining is important in this scenario. Text summarization is an emerging research field. H.P. Luhn in 1958(Luhn, 1958) introduced this research methodology. He proposed a method to extract the important sentences from the text using features such as phrase and word frequency (Allahyari *et al,* 2017).]. Process of Automatic Text Summarization includes extricating or gathering significant data from original content and exhibits that data as summary (Nitu et al, 2017). It reduces the effective time to get the crux of the data. Need for summarization can be seen in different reason and in numerous space, for example summarization of news articles , emails, market research, information related to government authorities, medical history of patients and diseases etc. Summarization has variants as it can be done on a single document and also on multiple documents on similar topic.

Summarization tools are also available online based on the type of data to be processed for different fields like news article summarizers like Columbia News blaster (Saranyamol & Sindhu, 2014) and medical field related summarization tools like Sum Basic (Gaikwad & Mahender, 2016).Classification of text summarization is based upon various standards (Rani & Tandon, 2018). Based on these standards three criteria are being set : input type of document, purpose critera , document output criteria (Aries *et al,* 2019). Early summarization was done on the single document which produces summary if a single document (Khan & Salim, 2014). But as the data increases multi document summarization emerged.

2. TEXT SUMMARIZATION TECHNIQUES

Text summarization approaches are comprehensively isolated into two categories: *Extractive summarization* & *Abstractive summarization* (Khan & Salim, 2014).

2.1 Abstractive Summarization

Contrasted with extractive summarization, abstractive summarization is nearer to what people typically anticipate from content analysis (Nayak &Sahoo, 2018) . The procedure is to comprehend the first record and reword the report to a shorter format while catching the key focuses (Mozhedehi *et al*, 2017). Text abstraction is primarily done using the concept of artificial neural networks.

Abstractive summarization approaches are classified into two categories: Structured based approach and Semantic based approach (Saranyamol & Sindhu, 2014).

2.1.1. Structure Based Approach

The centre of structure-based methods is utilizing earlier information and mental element compositions, for example, layouts, extraction administers just as flexible elective structures like trees, lead and body, graphs, to encode the most essential information.

2.1.1.1 Tree based

Tree based approach utilizes a dependency tree to process the content of a report. To get the precise summary this method uses the language generators (Gaikwad & Mahender, 2016). This approach includes the formation of dependency tree in it's first step, which is formed by the division of sentences in chunks (Takamura *et al* ,2014). Next step include the determination of the centrality of dependency tree. Sub trees of different are recognised and added in the dependency tree to increase its level. At last it prunes the predefined constituents (Khan & Salim, 2014).

2.1.1.2 Template based

This approach includes the processing of the document based upon some template as a reference. To point out the

text phrases that will be ripped into slots which will form a database (Khan & Salim, 2014). This is all done by using combination of linguistic patterns or extraction rule matching (Gaikwad & Mahender, 2016). The resulted phrases generate the outline of summary to be produced. Information Extraction Systems used in this approach extracts much precise information as it values only relevant information (Saranyamol & Sindhu, 2014).

2.1.1.3 Rule based

In this approach categorisation of the document is done and a feature based list of documents is prepared to summarize them. Information extraction rules generates output, from which the best option is selected by the content selection module which can respond suitably according to categorisation (Saranyamol & Sindhu, 2014).The main target of this method is to generate a summary which is more informative as compare to the existing one. Manually writing the principles and example is the limitation of this method, which is a monotonous and tedious task (Khan &Salim, 2014).

2.1.1.4 Ontology Based

An Ontology is characterized as an unequivocal conceptualization of terms and their relationship to a space. These are defined for particular domains. It is generally perceived that building an area model or ontology is a significant advance in the betterment of information based framework (Khan & Salim, 2014). The advantage of this methodology is that it abuses fuzzy ontology to deal with questionable information that simple domain ontology cannot.

2.1.1.5 Lead and body phrase based

This technique depends on the processing of the expressions (inclusion and substitution). These expressions possess similar syntactic head lump ahead of pack and sentences of body, which are used to work on the lead sentences and rewrite them (Gaikwad & Mahender, 2016). Enlivened by sentence combination system, common phrases are recognized in the lead and body pieces by this strategy. After sentence updations, summary is produced through the process of inclusion and substitution .Parsing mistakes which is the limitation of this method leads to degrade the sentential formation of summary sentences for example, grammatical errors and redundancy (Saranyamol & Sindhu, 2014).

2.1.2 Semantic Based approaches

Semantic based approach includes three stages:

- Document as input
- Representation of document semantically
- Feeding it to Natural language Generation Phase (NLG) to get the required results with the main focus lying in identifying noun and verb phrase.

2.1.2.1 Semantic Graph based

The strategy speaks to the report in the formation of semantic chart utilizing UMLS ideas and relations. It creates a more extravagant portrayal than the one gave by customary models dependent on terms (Khan & Salim, 2014). Rich Semantic Graph (RSG) which is a semantic diagram is created to condense the documents. Semantic and syntactic connections which are created in the pre-preparing module are used to link the concepts of the sentences (Gaikwad & Mahender, 2016). It produces compact, sound and less repetitive and syntactically right sentences. This approach is applicable only on single document for the summarization.

2.1.2.2 Multimodal semantic method

This method utilizes a semantic model, to signify the data of the multimodal document. Ideas and relationship among concepts is captured using this semantic model (Khan & Salim, 2014) (Gaikwad & Mahender, 2016).The linkage and connection between the concepts are represented by nodes. Significant thoughts are appraised utilizing data thickness metric which qualifies the fulfilment and connections with others concepts. The selected concepts lastly changed into sentences to make a summary (Khan & Salim, 2014). The physical assessment by the people is emerged as restriction of this system.

2.1.2.3 Information item (INIT) based

In this technique the abstract representation is utilized to produce the precise summary of the document but not the sentences from source document (Gaikwad & Mahender, 2016). The smallest component as abstract representation of reasonable data is the information item (INIT). Summary produced by this method is short, intelligent, informative & less repetitive rundown. A parser helps in the syntactical analysis of the document by INIT retrieval and SVO (subject-verb-object) formation at initial stage. A language generator is used to generate sentences. Next stage includes the ranking of sentences based on frequency scores. Summary generated includes the sentences which have high rank. Grammatical errors are common in this approach also limited in the generation of meaningful sentences. Secondly the quality of summary produced linguistically is low to due to incorrect parses (Saranyamol & Sindhu, 2014).

2.2 Extractive summarization

This technique is utilized to feature the words which are applicable, from input source document (Nayak & Sahoo, 2018). Summaries help in creating linked sentences taken according to the appearance. Choice is made dependent on each sentence if that specific sentence will be used for the summary or not (Mozhedehi *et al*, 2017). For instance, Search engines ordinarily utilize Extractive summary age techniques to create summary from website page. Numerous kinds of legitimate and numerical details have been used to make summary. The principal thing is positioning issue which incorporates positioning of the word. The second one choice issue that incorporates the choice of subset of specific

units of positions and the third one is soundness that is to know to choose different units from reasonable summary (Nitu *et al*, 2017).

2.2.1 Text Summarization using Fuzzy Logic:

For selecting sentences this approach uses fuzzy sets & fuzzy dependent on the features like sentences length, title feature, term weight, sentence position, similarity to other sentences etc (Aries *et al,* 2019). Input to the logic system is given in the form of these features. After that the mandatory rules which are required for summarization are entered. For each of the sentence, a numeric value which lies between 0 and 1 is received. That value determines the importance level of sentences in summary which is finalized (Saranyamol & Sindhu, 2014). The fuzzy system comprises of four parts: the fuzzifier part, inference engine, defuzzifier , the fuzzy knowledge base (Moratanch & Chitrakala, 2017).

2.2.2 TF- IDF method:

TF-IDF is defines as Tem frequency- inverse document frequency. It is a numeric method to determine the relevant words in the document. Frequency of sentence is characterized as the number of times any term occurs in the sentences in the source document. If the words appear frequently that means it is vital and should be given a high score than the word appearing less frequently (Bhatia and Jaiswal, 2016). For calculating TF – IDF formula is:

$$tf(w)= \left[\frac{\text{Total appearance of a term w in document (D)}}{\text{Total terms in D}} \right] \quad ....(1)$$

$$idf(w) = \log_e \left[\frac{\text{Total number of documents}}{\text{Number of documents with term w in it.}} \right] \quad .....(2)$$

Hence TF-IDF is calculated in the following way for a word w in the document :

TF- IDF(w) =(1) * (2).

2.2.3 Cluster Based Method :

Typically summaries are scripted in a way so that each segment of the record has a place with various subjects in a sorted out way and the documents are arranged either verifiably or unequivocally to acquire a summary (Aries *et al,* 2019) . This approach is called clustering technique. Sentences can be gathered dependent on their sentence score. Selection of sentences is based on cluster (Ci). Location of sentence Li is another factor for selection. Similarity of the sentence which is produced with the original sentence increases the sentence score (Fi). Sentence location in the document is also considered as one of factor to determine scores (Xi). (Ci), (Fi),(Xi) as a weighted sum represent the overall score of the sentence (Si)(Saranyamol & Sindhu , 2014)(Karamakar *et al*, 2015).

Si =W1 *Ci + W2 *Fi+ W3 *Xi (Saranyamol & Sindhu , 2014)(Karamakar *et al*, 2015).

2.2.4 Query Based Method

In this method sentences scores are determined in light of the frequency count of terms and those sentences are considers as highly valued sentences which contains a query word (Karmakar *et al*, 2015). Query is defined as per the user requirements. Extraction of sentences is done based on the high valued scores. For the extraction of summary along with their structural context (Rani & Tandon, 2018) . Extraction of information can be done from various parts or subsection of sentences of the content .The summary thus produces is the combination of such sub- sequences.

2.2.5 Machine Learning Method:

A training documents set along with its summaries in extractive form are given as input to the training stage (Saranyamol & Sindhu, 2014)(Karmakar *et al,* 2015). This approach views classification problem in text summarization (Moratanch & Chitrakala, 2017). Machine learning approach classified as supervised, unsupervised or semi-supervised. Sentences which posses the features are classified as summary sentences and the sentence which do not as non – summary sentences. Bayesian rule is used to statistically determine the classification probabilities (Allahyari *et al,* 2017):

$$P(s \in <S \mid F1, F2, ..., FN) = P(F1, F2, ..., FN \mid s \in S) * P(s \in S) / P(F1, F2, ..., FN) \text{ where}$$

S is a sentence from the document collection,F1, F2…FN are features which are used for classification (Nayak & Sahoo, 2018). S is the summary to be produced, P (s∈< S | F1, F2,...., FN) is the probability that sentence s are either summary sentences or non summary sentences on basis of features (F1,F2…FN)(Nayak & Sahoo, 2018).

2.2.6 Graph Theoretic Method:

The basic idea behind this approach is like voting. Graph Theoretic portrayal of entries gives a strategy for identifying the themes (Karmakar *et al*, 2015). After the pre-processing stage like stop word removal & stemming, the main parts of sentences is depicted as nodes of the tree after that they are linked to each other by edges if they have common words (Rani & Tandon, 2018). The nodes with the high cardinality are vital More the cardinality, more the chances of sentences are included in the produced summary (Allahyari *et al,* 2017).

2.2.7 Latent Semantic Analysis (LSA) Method:

It is an algebraic–statistical procedure to get the crux of the hidden capabilities of the word and sentences. Singular Value Decomposition method is applied to get the relationship between sentences and words (Allahyari *et al*, 2017). It is an unsupervised learning approach which doesn't require any kind of external knowledge (Bhatia & Jaiswal, 2016). LSA algorithm proceeds in three steps firstly an input metric is formed then applying the LSD method on the generated metric and lastly extraction of sentences (Chettri & Chakraborty, 2017).

2.2.8 Text Summarization using Neural Network:

It involves certain steps to be implemented in which first phase includes the training of neural network (Chettri & Chakraborty, 2017). A neural network (NN) after training is used which depicst the sentences type which is must to be involved within the summary to be generated (Saranyamol & Sindhu, 2014). The second phase includes feature fusion which means combining those features which are important for inclusion in the summary. The third phase is selection of sentences which uses the modified neural network. Only the highly ranked sentences are selected by the modified neural network. This step controls the summary selection in terms of their importance (Moratanch & Chitrakala, 2017).

2 RELATED STUDY

Luhn(1958) presented the first work done on generating the auto abstracts from the text documents. Generation of summary needs general familiarity with the subject and the objective is to save the reader time. The system IBM was used which started with the document in machine readable form and go further by the method of programmed sampling process which commensurate to the scanning a human reader. The mechanical sampling process selects the most appropriate sentences, which served as the clues to determine the type of the content of the document. In this procedure, significance factor also studied based on frequency of words appearing in the text. Based on significance factor clusters are formed and the frequency is counted.

Khan and Salim (2014)presented a review on the methods of abstractive text summarization. Abstractive summarization methods are grouped into two categories which are structured based & semantic based. Both of these approaches are discussed along with advantages and disadvantages. On the basis of three parameters different methods are compared which are representation of text, selection the information and production of summary. It has been presumed that the vast majority of the abstractive summarization strategies deliver exceptionally intelligible, strong, data-rich and less redundant summary.

Takamura *et al* (2014) utilized both reliance among words and reliance between sentences by building a nested tree. An approach of summarizing a single record that included relations among sentences and relations between words is formulated. Authors made a nested tree and planned the issue of summarization as that of whole integer linear programming. In this creation of summarization component as a combinatorial optimization problem, in which the nested tree was cut without losing significant portion in the source document

Carenini *et al* (2014) exhibited a hearty abstractive meeting summarization system of the conversation in meetings. It expands a novel multi-sentence combination algorithm soas to create theoretical layouts. It likewise used the relationship among transcripts of source meetings and summaries to choose the best layouts to create abstractive summarization of meetings. Algorithm adopted here is word graph algorithm and evaluation is done using ROUGE metric.

Saranyamol and Sindhu (2014) described approaches of automatic text summarization .Authors also provided an assessment of various methods .Techniques are categorically explained in extractive and abstractive approaches. Extractive approaches are sub-categorised as TF-IDF method, Graph Theoretic Approach, Cluster based, Machine Learning Approach etc while Abstractive approach as Structure based & Semantic based.

Karmakar *et al* (2015) provided an overview on Automatic Text Summarization and its approaches and the problems associated with them. This survey paper examined some of the extractive text summarization strategies. An extractive summary approach determines relevant sentences from source content. Authors provided solution methodologies as TF-IDF, Cluster based, Graph Theoretic Approach, Machine Learning and Query Based approach.

Bhatia and Jaiswal (2016) reviewed and investigated the important and popular work done in the area of Text Summarization. This paper provides an overview of work done by the researchers in the various sub fields into which Text Summarization is classified Single &. Multi Document into Extractive & Abstractive summarization, and Generic & Query–based. Authors provided the scenario in which each of the approaches has improved over time and the advances which are being achieved.

Gaikwad and Mahender (2016) introduced the basic outline of text summarization process and the approaches involved in it .This paper also provide a review on the different approaches of text summarization. Authors explained that Text Summarization is a procedure of removing the irrelevant data or combine significant data which is represented as summary from original content. Understanding the precise summaries is less time consuming. Two of its approaches extractive & abstractive are explained in detail along with advantages and disadvantages. .

Allahyari *et al* (2017) described the main approaches of text summarization. Authors have also provided an analysis of various procedures for the summarization for outline and depict the viability and deficiencies of the various techniques. They stressed on various for single & multi-document summarization based on extractive approach like Graph methods, Machine Learning.

Moratanch and Chitrakala (2017) survey conducted to exhibit a mechanism for extractive text summarization and correlation of particular methodologies and systems of extractive text summarization process. This paper deciphers

the strategies for text summarization in extractive way alongwith a less monotonous depiction, exceptionally adhesive, sound & significant details.

Chettri and Chakraborty (2017) presented a basic outline for Text Summarization and why it is essential. Authors provide an idea that the best way to summarize the text is to augment the various techniques like Machine Learning, Neural Networks etc. and implement them. This paper provided a brief over different techniques and methodologies used by various researchers for automatic text summarization.

Nitu *et al* (2017) described Text Summarization, one of the extraordinary information mining applications. Two classifications: extractive and abstractive are describes in the paper. Authors proposed that numerous systems on abstractive content outline have been produced for the dialects like English, Arabic, Hindi and so on. But, there is no abstractive technique for Bengali content because it must need the total information about every Bengali word, which is lengthy procedure for summarization.

Mozhedehi *et al* (2017) introduced the topic of text mining and its linkage with text summarization. A review has been presented on a summarization approaches and the most noteworthy extraction criteria are exhibited. The Summarization Criteria based on output summary, details, contents, limitations, number of input texts, language acceptance. Authors also pointed out the problems associated with each of the approaches and also distinguished between Text Mining and Data Mining. At last important evaluation method is mentioned which is termed as ROUGE-N, which compares the summaries generated by human with automate generated summary.

Rani and Tandon (2018) incorporated the abstractive and extractive synopses of the content. There are various methodologies of content summarization. This present reality utilizations of content summarization can be: reports summarization, news and articles summarization, audit frameworks, suggestion frameworks, online life checking, review reactions frameworks. The paper provided a study of various research works in the area of automatic content summarization.

Nayak and Sahoo (2018) provided a review on basic understanding as to why text summarization is done. Authors provided a view that there is a need of that approach which will give summarization without airing any repetition or any sort of uncertainty regardless of whether the summary don't contain any piece of the actual document. This paper is furnished with few of these methodologies which are desirable over get increasingly effective and exact summary of the actual document.

Aries *et al* (2019) provided an insight over various works in automatic text summarization specifically the latest ones. Authors provided some problems and cut off points which forestall attempts to push ahead and most of which are related to nature of processed language. has given classification of summarization systems which is based on different criteria. In this paper different approaches are discussed like Statistical, Graphs, Machine Learning, Linguistic etc . This paper also provided the comparison among each of these approaches, pros and cons, problems associated with them and the solutions. In this paper authors has worked on some of the challenges like summary informativeness.

3 CHALLENGES AND FUTURE SCOPE

Assessing rundowns (either consequently or physically) is a troublesome undertaking. The fundamental issue in assessment originates from the difficulty of building a standard against which the consequences of the frameworks that must be thought about. Further, it is exceptionally elusive out what a right summary is on the grounds that there is an opportunity of the framework to produce a superior summary that is not the same as any human summary which is utilized as an estimation to precise result. There are certain challenges both in extractive as well as abstractive method of text summarization. Extractive summarization lacks the readability of the text produced while abstractive method is difficult and complex to implement. The future of this research zone hugely relies upon the limit to discover proficient methods for consequently assessing the frameworks.

4 CONCLUSION

Content summarization is developing as a subpart of NLP as the interest for compressive, important, theoretical of the theme because of the enormous measure of data accessible on the net. Exact data assists with looking through more successfully and effectively. In this way message summarization is a need and utilized by the business investigators, advertising officials, advancement, analysts, government associations, understudies and instructors moreover. In the present scenario it is the requirement of everyone to get the desired information in less time. This paper explains both the extractive and abstractive approaches alongside the procedures utilized, its exhibition accomplished. Content summarization has its significance in both businesses just as research & enhancing network. Abstractive summarization needs to be focused more, as it more advances and produce the summaries by learning of data. Hence, it is a bit complex then extractive methodology, but it provides progressively important and proper summary as compare to extractive methodology. Text summarization as a part of NLP can be helpful in day to day basis work progress. Need of the hour is just to seek the

potential of the methodologies which can be suitably applied to get the desired information.

REFERENCES

Luhn,H.P. The automatic creation of literature abstracts. *IBM Corp*. (1958)

Khan, A.,& Salim, N. (2014). A review on abstractive summarization methods. *Journal of Theoretical and Applied Information Technology*, 59(1), 64-72.

Takamura, H., Okumura, M., Nagata, M. , Hirao, T., & Kikuchi, Y., (2014, June). Single document summarization based on nested tree structure. In Proceedings of *the 52nd Annual Meeting of the Association for Computational Linguistics (pp. 315-320).*

Carenini, G., Ng, R. Oya, T., & Mehdad, Y., (2014,June). A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In Proceedings of the *8th International Natural Language Generation Conference (INLG) (pp. 45-53).*

Saranyamol, C. S., & Sindhu, L. (2014). A survey on automatic text summarization. *International Journal of Computer Science and Information Technologies,* 5(6), 7889-7893.

Karmakar, S., Lad, T., & Chothani, H. (2015). A Review Paper on Extractive Techniques of Text Summarization. *International Research Journal of Computer Science (IRJCS)* Issue, 1.

Jaiswal, A.& Bhatia, N., (2016, January). Automatic text summarization and it's methods-a review. In *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 65-72). IEEE.*

Mahender, C. N., & Gaikwad, D. K., (2016). A review paper on text summarization. *International Journal of Advanced Research in Computer and Communication Engineering, 5(3), 154-160.*

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). Text summarization techniques: a brief survey. *International Journal of Advanced Computer Science and Applications, 2017.*

Moratanch, N., & Chitrakala, S. (2017, January). A survey on extractive text summarization. *In 2017 international conference on computer, communication and signal processing (ICCCSP) (pp. 1-6*). IEEE.

Chettri, R., and Chakraborty, U. K. (2017). Automatic text summarization. *International Journal of Computer Applications,* 161(1), 5-7.

Nitu, A.M. , Emran, A. , Afjal ,M.I. , Uddin, M. P. , Tumpa ,P. B. , & Yeasmin, S.(2017).Study of Abstractive Text Summarization Techniques. *American Journal of Engineering Research(AJER)(pp. 253-260).*

Mozhedehi, A. T., Rahimi,S. R.(2017,December). An Overview on Extractive Summarization. *IEEE International Conference on Knowledge based Engineering and Innovation (KBEI).*

Rani, R. ,& Tandon, S.(2018). Literature Review on Automatic Text Summarization. *International Journal of Current Advanced Research* (pp. 9779-9783).

Nayak ,A.K., & Sahoo, A.(2018).Review paper on Extractive Text Summarization. *International Journal of Engneering Research in Computer Science and Engineering(IJERCSE).*

Aries, A., & Hidouci, W. K. (2019). Automatic text summarization: What has been done and what has to be done. arXiv preprint arXiv:1904.00688.