Automatic Text Summarization: The Future Need

Kavita Jain[1] Mr. Manish Dubey[2]
[1]M. Tech scholar [2]Assistant Professor
[1,2]Department of Computer science and Engineering
[1,2]Arya Institute of Technology & Engineering, Jaipur Rajasthan, INDIA

Abstract— Numerous documents are available on internet and the number of documents is also rising day by day. In such a scenario reading the complete documents related to a topic is quite impossible , if we get the summary of the related one it will be quite useful and speed up the total process. In our paper we review the concept of automatic text summarization, its approaches etc.

Key words: Abstractive, Extractive, Information Retrieval, Summary Generation, Text Summarization

I. INTRODUCTION

Looking at feeling of content is a trying issue in Text summarization has turned into an essential and auspicious tool for helping and deciphering content information in today's quickly developing information age. Colossal expanding and simple accessibility of information on the World Wide Web have as of late brought about reviewing the classical linguistics problem [1] – the condensation of information from content documents. This undertaking is basically a data decrease prepare. The objective of automatic content summarization is gathering the source content into a shorter form safeguarding its information substance and general significance. Content summarization is the procedure of automatically making a packed form of a given document pre-serving its information content. Automatic document summarization is an essential research region in characteristic dialect preparing (NLP). The innovation of automatic document summarization is creating [2] and may give an answer for the information over-burden problem.

Summary may allude to abstract summary, edited version or executive summary. An abstract is a brief summary of an exploration article, theory, audit, gathering continuing or any top to bottom investigation of a specific subject or teach, and is frequently used to help the peruser rapidly learn the paper's motivation. [3] When utilized, an abstract dependably shows up toward the start of an original copy or typescript, going about as the purpose of-passage for any given scholastic paper or patent application.

A compressed version (or encapsulation) is a consolidating or diminishment of a book or other inventive work into a shorter shape while keeping up the solidarity of the source.[4] The abstract can be consistent with the first work regarding disposition and tone, catching the parts the condensing writer sees to be most imperative; it could be an entire satire of the first; or it could fall anyplace in the middle of, either by and large catching the tone and message of the first writer however missing the mark in some way, or quietly curving his words and message to support an alternate elucidation or plan.

An executive summary, once in a while known as an administration summary, is a short document or segment of a document, created for business purposes, that condenses a more drawn out report or proposition or a gathering of related reports such that perusers can quickly wind up noticeably familiar with a huge group of material without reading it all. It more often than not contains a brief proclamation of the problem or proposition shrouded in the major document(s), foundation information, compact investigation and fundamental conclusions. It is expected as a guide to basic leadership by managers [5] and has been portrayed as perhaps the most essential piece of a business plan.

Automatic content summarization systems can be sorted into a few unique types [6] The distinctive measurements of content summarization can be for the most part arranged based on its input type (single or multi document), reason (generic, space particular, or question based) and yield type (extractive or abstractive).

Single document summarization produces summary of single input document. Then again, multi document summarization produces summary of multiple input document. These multiple inputs are frequently documents talking about a similar theme. A considerable lot of the early summarization systems managed single document summarization.

Generic summarization intention is to condense all texts paying little mind to its point or space; i.e., generic summaries make no suppositions about the area of its source information and view all documents as homogenous texts. Most of the work that has been done rotates around generic summarization [7]. There have additionally been developments of summarization systems which are focused upon different area of intrigue. For instance, compressing money articles, biomedical documents, climate news, psychological militant occasions and numerous [8]. Regularly, this type of summarization requires area particular learning bases to help its sentence determination handle. Inquiry based summary contains just information which are questioned by the client. The inquiries are normally characteristic dialect inquiries or keywords that are related to a specific subject. For example, pieces delivered via web crawlers is a case of inquiry based application [9].

Extractive summaries or concentrates are delivered by recognizing essential sentences which are straightforwardly chosen from the document. A large portion of the summarization systems that have been created are for extractive type summaries [10]. In abstractive summarization, the chose document sentences are consolidated soundly and packed to reject insignificant areas of the sentences [11].

II. APPROACHES TO SENTENCE EXTRACTION OF TEXT SUMMARIZATION

Analyzing conclusion of text is a trying issue in Text summarization has turned into an imperative and opportune tool for assisting and deciphering text information in today's fast-developing information age. Colossal increasing and easy accessibility of information on the World Wide Web

have as of late brought about looking over the classical linguistics problem [1] – the condensation of information from text documents. This task is basically a data lessening process. The objective of automatic text summarization is gathering the source text into a shorter form protecting its information substance and general significance. Text summarization is the procedure of automatically making a packed form of a given document pre-serving its information content. Automatic document summarization is an imperative research territory in normal dialect preparing (NLP). The innovation of automatic document summarization is creating [2] and may give an answer for the information over-burden problem.

Summary may allude to abstract summary, edited version or executive summary. An abstract is a brief summary of a research article, thesis, review, conference proceeding or any top to bottom analysis of a specific subject or train, and is regularly used to help the per user rapidly as The key idea of extractive summarization is to distinguish and separate imperative document sentences and set up them together as a summary; i.e., the produced summary is an accumulation of unique sentences. There are a few ways to deal with sentence extraction. The accompanying subsections will depict three methodologies, specifically, recurrence based approach, feature based approach and machine learning based approach.

A. Fequency Based Approach

As we talked about in the presentation area; in the early work on text summarization, which was spearheaded by Luhn, it was assumed that imperative words in document will be rehashed commonly contrasted with alternate words in the document [12] Thus Luhn proposed to demonstrate the significance of sentences in document by utilizing word recurrence. From that point forward, a number of the summarization systems utilize recurrence based methodologies in their sentence extraction handle [13]. Two methods that utilization recurrence as a basic type of measure in text summarization are: word likelihood and term recurrence backwards document recurrence.

B. Feature Based Approach

One of the normal approach to decide the significance of a sentence is to recognize the features that mirrors the pertinence of that sentence. [14] Defined three features esteemed characteristic to sentence pertinence i.e., sentence position, nearness of title word and signal words. For instance, the starting sentences in a document generally portrays the principle information concerning the document. In this way, selecting sentences based on its position could be a reasonable system. The accompanying features are usually used to decide sentence importance [15].

1) Title/Headline Word

Title words showing up in a sentence could propose that the sentence contains vital information.

2) Sentence Position

The starting sentences in a document more often than not depicts the principle information concerning the document.

3) Sentence Length

Sentences which are too short may contain less information and long sentences are not appropriate to speak to summary.

4) Term Weight

Words or terms which have high event inside a document is utilized to decide the significance of a sentence.

5) Proper Noun

Formal person, place or thing and named entities such as person, organization and location specified in a sentence are thought to convey vital information.

Figure 1 portrays the general model of a feature based summarizer. The scores for each feature are processed and consolidated for sentence scoring. Preceding sentence scoring, these features are offered weights to decide its level of significance. In this case, feature weighting will be connected to decide the weights associated to each feature and the sentence score is then processed utilizing the linear combination of each feature score multiplied by its relating weight:

$$\text{Score} = \sum_{I=1}^{N} wI \times fI$$

Where:

wI = The weight of feature i

fI = The score of feature i

[16] Proposed a text summarization display based on Particle Swarm Optimization (PSO) to decide the feature weights. [17] Used hereditary calculation to inexact the best weight combination for their multi document summarizer. Differential advancement calculation has likewise been utilized to scale the pertinence of feature weights [18]. Examination on the impact of various feature combination was conveyed by [19], where it was found that better outcomes were gotten by consolidating term recurrence weight with position and hub weight.

In later works, the joining of fluffy principles was examined by [20] for scoring sentences. For example, one of their built tenets states "if (No WordIn Title is VH) and (Sentence Length is H) and (Term Weight is VH) and (Sentence Position is H) and (Sentence Similarity is VH) and (Proper Noun is H) and (Thematic Word is VH) and (Numberical Data is H) then (Sentence is essential)". Their trial finding (tried on the DUC 2002 data set) demonstrated that the fluffy rationale based technique could beat a general factual strategy. A current review additionally bolsters the benefits of utilizing fluffy reasoning to decide the significance of a sentence [21].

Certain the paper's motivation. [3] When utilized, an abstract dependably shows up toward the start of a composition or typescript, going about as the purpose of-passage for any given scholarly paper or patent application.

An abbreviated version (or concise edition) is a gathering or decrease of a book or other inventive work into a shorter shape while keeping up the solidarity of the source.[4] The compressed version can be consistent with the first work regarding state of mind and tone, catching the parts the condensing writer seems to be most essential; it could be an entire spoof of the first; or it could fall anyplace in the middle of, either for the most part catching the tone and message of the first writer however missing the mark in some way, or unobtrusively turning his words and message to support an alternate translation or plan.

An executive summary, at times known as an administration summary, is a short document or segment of a document, created for business purposes, that abridges a

longer report or proposition or a gathering of related reports in such a way that per users can quickly end up plainly familiar with a substantial assortment of material without reading it all. It normally contains a brief articulation of the problem or proposition shrouded in the major document(s), foundation information, compact analysis and fundamental conclusions. It is proposed as a guide to basic leadership by managers [5] and has been portrayed as potentially the most vital piece of a business plan.

Automatic text summarization systems can be classified into a few distinct types [6] The diverse measurements of text summarization can be for the most part ordered based on its input type (single or multi document), reason (generic, area particular, or question based) and yield type (extractive or abstractive).

Single document summarization produces summary of single input document. Then again, multi document summarization produces summary of multiple input document. These multiple inputs are regularly documents talking about a similar subject. A number of the early summarization systems managed single document summarization.

Generic summarization reason for existing is to outline all texts regardless of its subject or area; i.e., generic summaries make no assumptions about the space of its source information and view all documents as homogenous texts. Most of the work that has been done spins around generic summarization [7]. There have additionally been developments of summarization systems which are focused upon different space of intrigue. For instance, outlining account articles, biomedical documents, climate news, psychological militant occasions and numerous [8]. Frequently, this type of summarization requires area particular learning bases to assist its sentence choice process. Inquiry based summary contains just information which are questioned by the client. The inquiries are ordinarily normal dialect inquiries or keywords that are related to a specific subject. For example, scraps delivered via web search tools is a case of inquiry based application [9].

Extractive summaries or concentrates are created by distinguishing essential sentences which are straightforwardly chosen from the document. The vast majority of the summarization systems that have been created are for extractive type summaries [10]. In abstractive summarization, the chose document sentences are joined rationally and packed to bar irrelevant areas of the sentences.

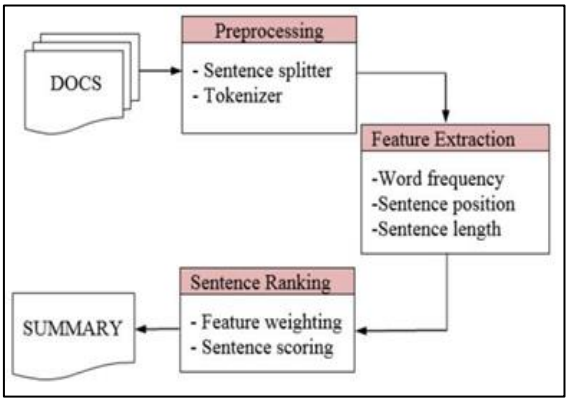

Fig. 1: A Feature Based Summarization Model

C. Machine Learning Approach

Machine Learning (ML) approach can be connected on the off chance that we have an arrangement of training document and their comparing summary extracts [22]. The objective of machine learning can be firmly related to a classification problem, i.e., to gain from a training model so as to decide the appropriate class where a component belongs to. On account of text summarization, the training model comprises of sentences marked as "summary sentence" on the off chance that they belong to the reference summary, or as "non-summary sentence" generally. Sentences are generally spoken to as feature vectors.

III. Techniques of Text Summarization

A. Domain Specific Summarization

A significant part of the work we reviewed in the previous sections included generic summarization whereby the importance of a summary is chosen quite recently based on the input document without identifying with its domain or the client needs [23]. For instance, inputs such as restorative documents, news documents or emails; have uncommon structures or remarkable qualities which ought to be considered by the summarizer to deliver more precise information. Next, we will review a portion of the works concerning domain specific text summarization.

B. Restorative Summarization

The review on automatic summarization was observed to be exceptionally helpful to the medicinal field. Summarization can help specialists to acquire important information about a specific disease or information from the patient records [24]. It will likewise be advantageous to patients or clients whom swing on the web to discover information related to their medical issues [25]. Moreover, there are broad assets that give access to medicinal information and restorative related databases. For example, there are more than 20 million articles in MEDLINE; a biomedical database. Summarization is consequently basic in such condition to treat the problem of information over-burden.

C. News Summarization

Early work on news summarization can be gone back to 1990s when SUMMONS summarizer was made [26]. SUMMONS was intended for outlining single occasions (news articles related to psychological oppressor occasions). It was fabricated utilizing a template-driven message understanding framework, MUC-4 [27]. The framework initially forms the full text and fills the template slots before incorporating the summary from the extracted information.

Like the SUMMONS framework is a framework called RIPTIDES [28]. It consolidates information extraction to bolster summarization. They utilize catastrophic event situation templates for every text and give them as input to the summarization framework. The summarizer first merges the templates into occasion situated structure and then the significance scores are assigned to each slot/sentence to choose the summary sentences.

IV. Comparison and Evaluation

Summary evaluation is a vital aspect of text summarization [33]. Evaluation strategies assess the convenience and

honesty of the summary [30]. Assessing the characteristics of summary like comprehensibility, coherence and readability are difficult tasks. By and large for summary evaluation inborn and extrinsic measures are utilized [33,31,32].

In characteristic strategies, people assess the nature of summary. While extrinsic techniques measure the quality by a task-based execution measure [33]. Characteristic measures are known as glass-box testing while extrinsic measures are known as discovery testing [31]. Characteristic evaluations have assessed mostly the coherence and education of summaries. Extrinsic evaluations, then again, have tried the effect of summarization on tasks like relevance assessment, reading comprehension, and so on.

Two normally utilized criteria for summary evaluation are:

1) Precision and Recall
2) Compression Ratio and Retention Ratio

A. Precision and Recall

For assessing the comparability between human created and framework produced summaries accuracy and recall are utilized [29].

Where, Correct shows sentences that are same in mechanized summary as well as manual summary. Wrong speaks to sentences introduce in computerized summary however not in manual summary.

Missed gives sentences that are found in manual summary yet not present in robotized summary.

In this way, Precision demonstrates the quantity of appropriate sentences extracted by the framework and Recall gives the quantity of reasonable sentences missed by the summarization framework.

B. Compression Ratio and Retention Ratio

As a rule, a text is said to be a summary, when it must obey two necessities:

- It must be shorter than the original input text;
- It must contain the vital information of the original text [29].

Pressure Ratio [31,32] measures how much shorter the summary is as of the original text.

Compression Ratio = Length of Summary/Length of FullText

Retention Ratio determines how much information is retained [31]. A good summary is one that has high retention ratio and low compression ratio. [29]

V. CONCLUSION

The importance of text summarization is growing day by day and our review paper have given an idea about what the automatic text summarization is all about, its methodologies and its evaluation process.

REFERENCES

[1] Jezek K., Steinberger J., Automatic Text Summarization, in Snasel, V. (ed.) Znalosti 2008, pp 1-12. FIIT STU Brarislava, UstavInformatiky a softveroveho inzinierstva (2008) ISBN 978-80-227-2827-0
[2] Rasim ALGULIEV, Ramiz ALIGULIYEV, Evolutionary Algorithm for Extractive Text Summarization, in Intelligent Information Management, 2009, pp 128-138. doi:10.4236/iim.2009.12019
[3] Gary Blake and Robert W. Bly, The Elements of Technical Writing, pg. 117. New York: Macmillan Publishers, 1993. ISBN 0020130856
[4] "Abridgment". m-w.com. Merriam-Webster.
[5] Definition of Executive Summary from Colorado State University.
[6] Nenkova, A. and K. McKeown, 2012. A Survey of Text Summarization Techniques. In: Mining Text Data, Aggarwal, C.C. and C. Zhai (Eds.), Springer Science and Business Media, New York,
[7] Nenkova, A. and K. McKeown, 2011. Automatic summarization. Foundat. Trends Inform. Retrieval, 5: 103-233.
[8] Radev, D.R. and K.R. McKeown, 1998. Generating natural language summaries from multiple on-line sources. Comput. Linguist., 24: 470-500.
[9] Nenkova, A. and K. McKeown, 2011. Automatic summarization. Foundat. Trends Inform. Retrieval, 5: 103-233.
[10] Aliguliyev, R.M., 2009. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Syst. Applic., 36: 7764-7772. DOI: 10.1016/j.eswa.2008.11.022.
[11] Ganesan, K., C. Zhai and J. Han, 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. Proceedings of the 23rd International Conference on Computational Linguistics, (CCL' 10), ACM, pp: 340-348.
[12] Luhn, H.P., 1958. The automatic creation of literature abstracts. IBM J. Res. Dev., 2: 159-165
[13] Klassen, P.P., 2012. Calculating LLR topic signatures with dependency relations for automatic text summarization. University of Washington.
[14] Edmundson, H.P., 1969. New methods in automatic extracting. J. ACM, 16: 264-285.
[15] Gupta, V. and G.S. Lehal, 2010. A survey of text summarization extractive techniques. J. Emerg. Technol. Web Intell., 2: 258-268.
[16] Binwahlan, M.S., N. Salim and L. Suanmali, 2009. Swarm based text summarization. Proceedings of the International Association of Computer Science and Information Technology-Spring Conference, Apr. 17-20, IEEE Xplore Press, Singapore, pp: 145-150.
[17] Bossard, A. and C. Rodrigues, 2011. Combining a multi-document update summarization system–CBSEAS– with a genetic algorithm. Combinat. Intell. Methods Applic., 8: 71-87. DOI: 10.1007/978-3-642-19618-8_5 Brin, S. and L. Page, 2012. Reprint of: The anatomy of a large-scale
[18] Abuobieda, A., N. Salim, M.S. Binwahlan and A.H. Osman, 2013a. Differential evolution cluster-based text summarization methods. Proceedings of the International Conference on Computing, Electrical and Electronics Engineering, Aug. 26-28, IEEE Xplore Press, Khartoum, pp: 244-248.
[19] Hariharan, S., 2010. Multi document summarization by combinational approach. Int. J. Comput. Cognit., 8: 68-74.
[20] Suanmali, L., N. Salim and M.S. Binwahlan, 2011. Fuzzy genetic semantic based text summarization.

Proceedings of the IEEE 9th International Conference on Dependable, Autonomic and Secure Computing, Dec. 12-14, IEEE Xplore Press, Sydney, NSW, pp: 1184-1191.

[21] Babar, S.A. and P.D. Patil, 2015. Improving performance of text summarization. Procedia Comput. Sci., 46: 354-363. DOI: 10.1016/j.procs.2015.02.031.

[22] Neto, J.L., A.A. Freitas and C.A.A. Kaestner, 2002. Automatic Text Summarization using a Machine Learning Approach. In: Advances in Artificial Intelligence, Bittencourt, G. and G.L. Ramalho, Springer, pp: 205-215.

[23] Nenkova, A. and K. McKeown, 2011. Automatic summarization. Foundat. Trends Inform. Retrieval, 5: 103-233.

[24] Becher, M., B. Endres-Niggemeyer and G. Fichtner, 2002. Scenario forms for web information seeking and summarizing in bone marrow transplantation. Proceedings of the Conference on Multilingual Summarization and Question Answering, (SQA' 02), Stroudsburg, PA, USA, pp: 1-8.

[25] Kaicker, J., V.B. Debono, W. Dang, N. Buckley and L. Thabane, 2010. Assessment of the quality and variability of health information on chronic pain websites using the DISCERN instrument. BMC Med., 8: 59-59.

[26] McKeown, K. and D.R. Radev, 1995. Generating summaries of multiple news articles. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 09-13, Seattle, WA, USA, pp: 74-82.

[27] Sundheim, B.M., 1992. Overview of the fourth message understanding evaluation and conference. Proceedings of the 4th Conference on Message Understanding, (CMU' 92), ACM, pp: 3-21.

[28] White, M., T. Korelsky, C. Cardie, V. Ng and D. Pierce et al., 2001. Multidocument summarization via information extraction. Proceedings of the 1st International Conference on Human Language Technology Research, (LTR' 01), pp: 1-7.

[29] Eduard Hovy, The Oxford Handbook of Computational Linguistics, Oxford University Press, Oxford, chapter 32, (2003).

[30] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi ,Bahareh Gholamzadeh "A Comprehensive Survey on Text Summarization Systems ", Computer Science and its Applications, CSA '09. 2nd International Conference on 10-12 Dec. 2009.

[31] http://www.isi.edu/natural-language/people/{hovy,cyl,marcu}.html.

[32] Martin Hassel, "Evaluation of Automatic Text Summarization", A practical implementation Licentiate,Thesis Stockholm,Sweden,(2004).

[33] Vishal Gupta, Gurpreet Singh Lehal," A Survey of Text Summarization Extractive Techniques", journal of emerging technologies in web intelligence, vol. 2, no. 3, august 2010.

