International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Review of Applications and Issues of Text Summarization

Author

[1] Avula Jyotsna Reddy, Department of Computer and Science and Engineering, Presidency University

[2]Keerthi Reddy. P, Department of Computer and Science and Engineering, Presidency University

[3]Abhayram, Department of Computer and Science and Engineering, Presidency University

ABSTRACT:

Text summarization deals with the creation of human-readable summaries from text documents and involves syntactic, semantic and discourse-level processing of text. The summarization system reduces a text document and conveys the key meaning of the text. Reducing data helps users to spot the specified information quickly without wasting time and energy in reading the whole document. This is often the foremost challenging task in information retrieval systems. Text summarization approaches are extractive and abstractive. Extractive summarization approaches are intended to come up with summaries by selecting key units (a subset of existing words, phrases, or sentences) of the original document. Abstractive summarization approaches require advanced NLP tools like ideal semantic parsers and language generation systems. supported input type, summarization will be classified into single and multi-document summaries.

Keywords: summarization, text, NLP, documents, Summary.

A.INTRODUCTION

Automated records retrieval structures are delivered to reduce "Information Overload". Information overload is the trouble of information trouble and making selections through someone whilst there may be an excessive number of records. Web Search Engines are the maximum considerable IR applications. Information retrieval (IR) is a challenge of retrieving files from a database in reaction to a consumer's query, and rating them primarily based totally on relevance.

This has been generally achieved the use of statistical strategies that

(a) pick terms (phrases, terms, and different units) from files which are deemed to great constitute their records, and

(b) create an inverted index report so as to offer a clean get entry to files containing those terms.

Now a days, it's miles very not unusual place that a keyword-primarily based totally search at the net through a consumer returns hundreds, or maybe heaps of hits, through which the consumer is frequently confused. The hassle is due to the shortage of a green and effective technique to locate the specified records. It could be very tough challenge for human beings to manually summarize big files of textual content. Research in computerized textual content summarization has acquired full-size interest withinside the beyond few years because of the exponential increase withinside the quantity & complexity of records reasserts at the net. Text summarization is the advent of brief model of textual content through laptop program. Generally, whilst human beings summarize textual content, we examine the complete choice to broaden a complete information, and then write a precise highlighting its most important points. Since computer systems do now no longer but have the language abilities of human beings, opportunity strategies have to be considered. In the exercise of

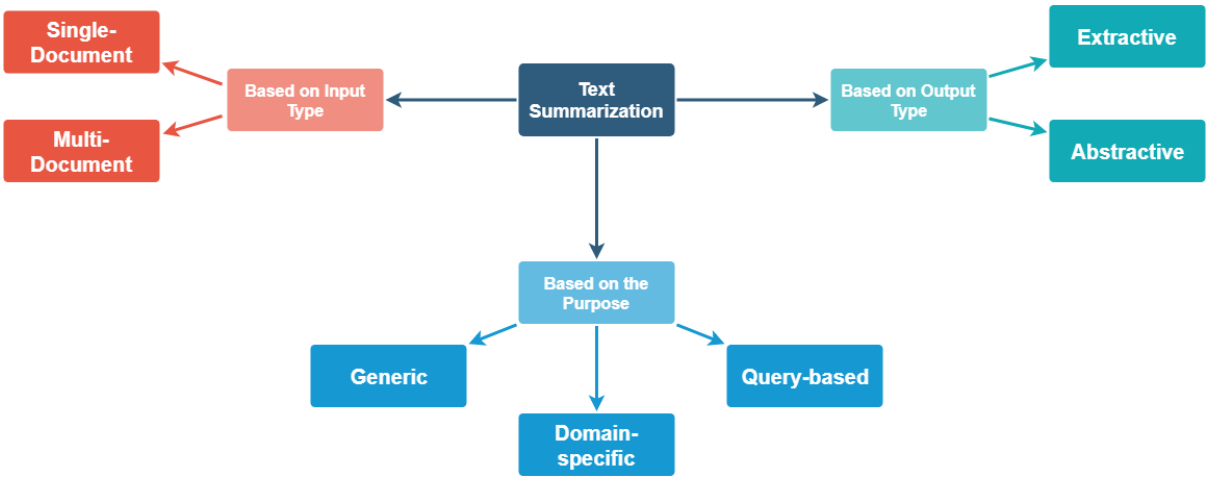
computerized textual content summarization, choice-primarily based totally technique has so far been the dominant strategy. The "maximum crucial" content material is dealt with as the "maximum frequent" or the "maximum favourably positioned" content material. Text summarization has a tendency to be a crucial challenge in content material extraction for the duration of net mining. Text summarization is a crucial NLP challenge, which has numerous applications. The large classes of tactics to textual content summarization are extraction and abstraction. Extractive strategies pick a subset of present phrases, terms, or sentences withinside the authentic textual content to shape a precis. abstractive strategies first construct an inner semantic illustration after which use herbal language era strategies to create a precise. Such a precise may comprise phrases that aren't explicitly gift withinside the authentic file. Most textual content summarization structures are primarily based totally on a few shapes of extractive summarization. subject matter identification, interpretation, precise era, and assessment of the generated precise are the important thing demanding situations in textual content summarization The crucial duties in extraction-primarily based totally summarization are figuring out key terms withinside the file and the use of them to pick sentences withinside the file for inclusion withinside the precis. abstraction-primarily based totally strategies paraphrase sections of the supply file. All extraction-primarily based totally summarizers carry out the subsequent 3 highly impartial duties

(a) taking pictures key factors of textual content and storing as an intermediate illustration,
(b) scoring sentences withinside the textual content primarily based totally on that illustration, and
(c) composing a precise through deciding on numerous sentences.

Summary era through an automated method has benefits as:

a. decreased analysing time
b. the dimensions of the precis may be controlled
c. its content material is deterministic and
d. the hyperlink among a textual content detail withinside the precis and its role withinside the authentic textual content may be effortlessly established.

B. TYPES OF SUMMARIZATIONS:



Based on input type:

Summarization methods are classified according to the source of input, which can be single or multi-document summarization. A Single Document Summarization, means only a single document is provided to generate a summary. It is earliest and simple approach for summarization, we can use both abstractive and extractive methods on single document summarization. Meanwhile in multi document summarization, where more than one information sources are provided for generating the summary.

Based on output type:

Summarization methods are classified into extractive summarization and abstractive summarization. Extractive summarization, which selects the existing words, phrases and sentences in the data to produce a summary and abstractive summarization, which generates new phrases and sentences from data to create a summary.

Based on the purpose:

Summarization methods based on the purpose are classified as generic, domain-specific and query-based summarizations. Generic summarization condenses all the information content available in source text. Domain-specific summarization generates summary of each document that contain key information and makes it easy to the users in finding the desired document easily. Query based summarization aims to extract a summary of a document which answers directly or relevant to the search query.

Automatic Summarization And Knowledge Bases

The main goal of the automatic text summarization is to create summaries that are identical to human-created summaries. But, in many cases, the understandability and robustness of the generated summaries are not good enough, because the generated summaries do not contain and cover all the semantically pertinent features of data effectively. This is because the existing summarization techniques do not consider the message conveyed by the context data.

A step towards building more precise, error-free summarization systems is to combine with knowledge bases, that are semantically based and establish different categories to divide the existing data into in an order to better understand of data.

The approach of human-generated knowledge bases and various ontologies in many various realms have opened further possibilities in text summarization, and outreaches increasing attention lately. For example, Henning et al presented advanced toward the sentence extraction that maps sentences to concepts of ontology features, they can improve the semantic representation of sentences which is useful in the selection of sentences for summaries. They experimentally showed that ontology-based extraction of sentences performs better than baseline summarizers. Chen et al introduced a user query-based text summarizer that uses the Unified Medical Language System (UMLS) medical ontology to make a summary of medical data/text. Baralis et al propose a Yet Another Great Ontology (YAGO) based summarizer that leverages YAGO ontology to identify key concepts in the documents. The concepts are evaluated and then used to select the most relevant document sentences. Sankarasubramaniam et al introduced an approach that makes use of Wikipedia in conjunction with a graph-based ranking technique. First, they create a bipartite sentences-concept graph and then use an iterative ranking algorithm for choosing summary sentences.

C. THE IMPACT OF CONTEXT IN SUMMARIZATION

Summarization systems often have additional affirmation they can utilize to specify the most important topics of documents. For example, summarizing a news article is different from a blog, cause certain text features like the length of the document and the genre of the topic like tech, movie, travel, sports, etc and while summarizing the blogs, there are conversations or comments coming after the blog posts that indicate the good source of information to determine which parts of the blog are important and interesting, make the task of summarization a serious data science problem.

In Google:

Entity timelines-

Given an entity and a time period, to provide a summary of the most important events involved in this entity. It means that it needs to instantly provide the information to the point about what happened to/about/around an entity within a particular time. In this, for a given entity it identifies and rank news collection, the importance of the entity in the collection and concisely summarizes each news cluster.

Storylines of events-

Identify and summarize events that lead to the event of interest and provide background information for an event and structured timelines for a given entity. In this, it defines a graph of events using similarity and identifies the heaviest path ending in a given event, and summarizes the whole events in this path regarding important entities in the target event.

Sentence Compression-

Given a sentence, generate a shorter one while preserving the key concepts. In this, it moves beyond Extractive summarization, which selects the main information from a source text onward to abstractive summarization which generates new phrases and sentences that represent the most important, challenging research problem.it uses syntactic information to ensure grammatically and the use of semantic or lexical information to preserve the essential content, and is limited to publicly available training data.

Event understanding-

It learns how events are referred to in the text and act for event mentioning text in predicate-argument structure. It is Unsupervised learning with limited training data with no restriction in the domain of events. It uses abstractive summarization for generating the headlines for news. For a given set of news collections, it extracts compressed patterns between entities from news collections and clustering patterns that refer to the same event.

Summarization of user-generated content-

It identifies the main aspects and summarizes for machine or user consumption like using a centrality summarization model on the comments for a video and using video categorization to provide the users a better search experience.

Summarize the web:

A search engine is designed to perform web searches. They search the world wide web in a systematic way for particular information specified in a textual web search query. It organizes information for accessibility

and usefulness. Web pages contains lots of data which cannot be summarized such as pictures and lots of textual information they have is often sparse, which makes applying summarization techniques narrow. In spite of that, they can consider the context of a web page that is pieces of information extracted from the data of all the pages linking to it, as additional information to improve summarization. The initial research in this topic is where the query web search engines and fetch the pages having links to the particular web page. They will analyse the pages and pick the best sentences containing the links web pages heuristically.

Books and literature:

Google has reportedly worked on projects that aims to produce the summary of a novel. So, it helps consumers while purchasing the book.

D. CONCLUSION

Owing to the fast growth in technology and usage of mobiles and internet, we have a large amount of data. This can be solved if we have good and precise text summarizers which helps us in producing summary of documents. Automatic text summarization guarantees us with powerful technology to make use in our struggle with information overload. Summarization tools can help us by summarizing the large quantities of information will makes us to have wider and potentially richer set of summarized data which helps in better decision making. In this paper we have discussed type of summarization which helps us to produce summary. Text summarization methods are classified into extractive summarization, which selects the existing words, phrases and sentences in the data to produce a summary and abstractive summarization, which generates new phrases and sentences from data to create a summary.

E. REFERENCES:

1. Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 500–509.
2. Rasim M Alguliev, Ramiz M Aliguliyev, Makrufa S Hajirahimova, and Chingiz A Mehdiyev. 2011. MCMR: Maximum coverage and minimum redundant text summarization model. Expert Systems with Applications 38, 12 (2011), 14514–14522.
3. Rasim M Alguliev, Ramiz M Aliguliyev, and Nijat R Isazade. 2013. Multiple documents summarization based on evolutionary optimization algorithm. Expert Systems with Applications 40, 5 (2013), 1675–1689.
4. Elena Baralis, Luca Cagliero, Saima Jabeen, Alessandro Fiori, and Sajid Shah. 2013. Multi-document summarization based on the Yago ontology. Expert Systems with Applications 40, 17 (2013), 6976–6984.
5. Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. A hybrid hierarchical model for multi-document summarization. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 815–824.
6. Ping Chen and Rakesh Verma. 2006. A query-based medical information summarization system using ontology knowledge. In Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on. IEEE, 37–42.
7. Freddy Chong Tat Chua and Sitaram Asur. 2013. Automatic Summarization of Events from Social Media.. In ICWSM.
8. John M Conroy and Dianne P O'leary. 2001. Text summarization via hidden markov models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 406–407.

9. Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 305–312.
10. J-Y Delort, Bernadette Bouchon-Meunier, and Maria Rifqi. 2003. Enhanced web document summarization using hyperlinks. In Proceedings of the fourteenth ACM conference on Hypertext and hypermedia. ACM, 208–215.
11. Günes Erkan and Dragomir R Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res.(JAIR) 22, 1 (2004), 457–479.
12. Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 19–25.
13. Vishal Gupta and Gurpreet Singh Lehal. 2010. A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence 2, 3 (2010), 258–268.
14. Ben Hachey, Gabriel Murray, and David Reitter. 2006. Dimensionality reduction aids term co-occurrence based multi-document summarization.In Proceedings of the workshop on task-focused summarization and question answering. Association for Computational Linguistics, 1–7.
15. John Hannon, Kevin McCarthy, James Lynch, and Barry Smyth. 2011. Personalized and automatic social summarization of events in video. In Proceedings of the 16th international conference on Intelligent user interfaces. ACM, 335–338.
16. Sanda Harabagiu and Finley Lacatusu. 2005. Topic themes for multi-document summarization. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 202–209.
17. Leonhard Hennig, Winfried Umbrath, and Robert Wetzker. 2008. An ontologybased approach to text summarization. In Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on, Vol. 3. IEEE, 291–294.
18. Meishan Hu, Aixin Sun, and Ee-Peng Lim. 2007. Comments-oriented blog summarization by sentence extraction. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, 901–904.
19. Meishan Hu, Aixin Sun, and Ee-Peng Lim. 2008. Comments-oriented document summarization: understanding documents with readers' feedback. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 291–298.
20. Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization-step one: Sentence compression. In AAAI/IAAI. 703–710.
21. Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. 74– 81.
22. Inderjeet Mani and Eric Bloedorn. 1999. Summarizing similarities and differences among related documents. Information Retrieval 1, 1-2 (1999), 35–67.
23. Inderjeet Mani, Gary Klein, David House, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 2002. SUMMAC: a text summarization evaluation. Natural Language Engineering 8, 01 (2002), 43–68.
24. Rada Mihalcea and Paul Tarau. 2005. A language independent algorithm for single and multiple document summarization. (2005).
25. Liu Na, Li Ming-xia, Lu Ying, Tang Xiao-jun, Wang Hai-wen, and Xiao Peng. 2014. Mixture of topic model for multi-document summarization. In Control and Decision Conference (2014 CCDC), The 26th Chinese. IEEE, 5168–5172.

26. Ani Nenkova and Amit Bagga. 2004. Facilitating email thread access by extractive summary generation. Recent advances in natural language processing III: selected papers from RANLP 2003 (2004), 287.
27. Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In Mining Text Data. Springer, 43–76.
28. Paula S Newman and John C Blitzer. 2003. Summarizing archived discussions: a beginning. In Proceedings of the 8th international conference on Intelligent user interfaces. ACM, 273–276.
29. You Ouyang, Wenjie Li, Sujian Li, and Qin Lu. 2011. Applying regression models to query-focused multi-document summarization. Information Processing & Management 47, 2 (2011), 227–237.
30. Makbule Gulcin Ozsoy, Ilyas Cicekli, and Ferda Nur Alpaslan. 2010. Text summarization of turkish texts using latent semantic analysis. In Proceedings of the 23rd international conference on computational linguistics. Association for Computational Linguistics, 869–876.