TEXT SUMMARIZATION METHODS USING NLP

Vedant Sharma,Nischay Goyal
Bachelor of Technologyin Computer Science and Engineering,
Galgotias University,Greater Noida
(harshit.20scse1010943@galgotiasuniversity.edu.in,
md.20scse1010767@galgotiasuniversity.edu.in

ABSTRACT

There is a vast amount of textual content available, and it continues to develop every day. Consider the internet, which is made up of web pages, news stories, status updates, blogs, and many other things. Because the data is unstructured, the best we can do is perform search and glance through the results. Much of this text material needs to be reduced to shorter, focused summaries that capture the important elements, both so we can explore it more easily and to ensure that the larger papers include the information we need. The proposed solution for this problem is a web application that uses abstractive text summarization to provide a paraphrase of the essential points. text, using a vocabulary set not the same as the original document Summaries cut down on reading time and make the selection process easier when investigating documents. The efficacy of indexing is improved by automatic summarization. Human summarizers are more prejudiced than automatic summarising techniques. Flask, Python 3, web development (HTML, CSS, BOOTSTRAP), and the Text Summarizing API are the tools and technologies utilised in text summarization (which uses NLP). One of the next goals could be to apply the topic-focused summarization framework to news articles or blogs, as well as to expand the machine learning research.

1.1 INTRODUCTION

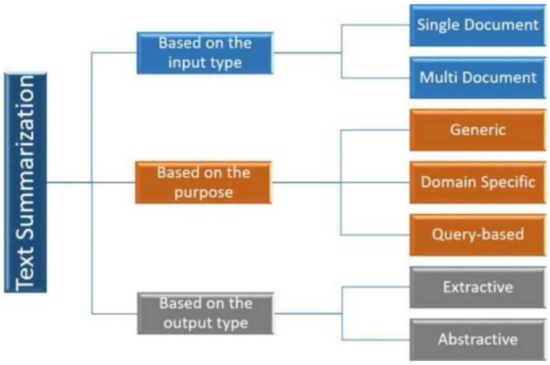
The Internet is a data warehouse. The internet provides access to news, movies, education, medicine, health, nations, weather, geology, and other topics. This could be data in the form of statistics, numerical, mathematical, or text[1]. Because of the increased number of characters, text data is more difficult to interpret. Because of the massive amount of data available, a system must be in place to ensure that only the most important aspects of the data are accessed. This can be accomplished by text summarization. Text summary has been the subject of decades of inquiry and study. To generate brief summaries, various methods have been developed and tested on various datasets. Different comparison scores are used to compare them. Extractive or astractive text summarising, single or multi-document summary are all options. Extractive text summarising is a technique for creating summaries from documents that use the same language as the original. ABS is more generic and concentrates on the document's main ideas. Single document summarising approaches produce summaries of a single document's text, while multi document summarization techniques produce summaries of numerous documents. Natural language processing (NLP) is a branch of computer science, artificial intelligence, and linguistics dealing with computer-human interaction. The process of designing a system that can process and produce language as well as a human can is known as natural language processing. As the use of the World Wide Web has grown, so has the problem of information overload. As a result, a system that automatically acquires, categorises, and summarises documents as needed by users is required. Text summarization can be used in a variety of applications; for example, academics need a tool to create summaries for selecting whether to read a text in its entirety or not, and for summarising material found on the internet. Multi document summarising can be used by news organisations to cluster and summarise material from many sources.

EXTRACTIVE SUMMARIZATION

To generate a summary, extractive text summarization selects a subset of existing words, phrases, or sentences from the original text. Extractive summarization selects essential phrases or keywords from a document using a statistical approach. Extractive summarization selects essential phrases or keywords from a document using a statistical approach.

ABSTRACTIVE SUMMARIZATION

The abstractive text summarization approach develops a phrase using a semantic representation before using natural language generation techniques to provide a summary that is more human-like. A summary like this could include words that aren't in the original. It entails comprehending the original content and retelling it in a more concise manner.



1.2 TYPES OF SUMMARIZATION TECHNIQUES

1.2.1 STATISTICAL APPROACHES

Statistical techniques [2] can summarise a document by employing statistical elements of the sentence such as title, location, and phrase frequency, giving weights to the keywords, calculating the sentence's score, and then selecting the sentence with the greatest score into the summary.

Title Method [3]

According to this strategy [3] [4], sentences in the title are deemed more important and are more likely to be included in the summary. The number of words usually used between a sentence and the title determines the phrase's score.

Location Method [3]

Text is given different weights depending on where it comes in a paragraph: first, middle, or last, or in a prominent area of the work like the conclusion or opening. The first few sentences of a document, as well as the last few phrases or the conclusion, are deemed more important and are included in the summary.

Cue Word Method [3]

Positive weights like "confirmed, significant, best, this document" and negative weights like "hardly, impossible" are given to text based on its importance. Cue phrases are typically genre-specific. In summary, a sentence including such cue phrases can be inserted.

1.2.2 LINGUISTIC APPROACHES

Linguistics is a branch of linguistic science that studies semantics and pragmatics. Semantics is the study of how meaning is inferred from words and concepts, while pragmatics is the study of how meaning is inferred from context.

Lexical Chain [5] [6]

Lexical chains take advantage of the cohesiveness between any number of linked words. Lexical chains are created by grouping (chaining) semantically similar groupings of words in a source document. The relationships between words that may cause them to be placed into the same lexical chain include identities, synonyms & hypernyms/hyponym.

Word Net [7]

It divides English words into sys-nets, which are collections of synonyms. Word Net additionally gives a short definition for each sys-net as well as a semantic relationship between them. Word-net also functions as a thesaurus and an online dictionary, and many systems use it to determine word relationships.

Graph Theory [8]

The structure of the text as well as the relationship between sentences in the document can be represented using graph theory [8]. The document's sentences are represented as nodes. Connections between sentences are regarded the edges between nodes. These links are linked by a similarity relationship.

1.3 FORMULATION OF PROBLEM

There is a vast amount of textual content available, and it continues to develop every day. Consider the internet, which is made up of web pages, news stories, status updates, blogs, and many other things. Because the data is unstructured, the best we can do is perform search and glance through the results. Much of this text material needs to be reduced to shorter, focused summaries that capture the important elements, both so we can explore it more easily and to ensure that the larger papers include the information we need. Everyone needs summaries of large textual data that are fast and easy to read and understand, whether they are researchers, students, doctors, or teachers. For example, researchers need a tool to generate summaries for deciding whether to read the entire document or not, and for summarising information searched by users on the Internet.

1.3.1 TOOLS & TECHNOLOGY USED

The tools and technology that we will use for the proposed system are:

NLP: NLP stands for natural language processing, which is the ability of a computer software to interpret human language as it is spoken and written. It's an artificial intelligence component.

TENSORFLOW: TensorFlow is an open source machine learning platform that runs from start to finish. It has a large, flexible ecosystem of tools, libraries, and community resources that allow academics to advance the state-of-the-art in machine learning and developers to quickly construct and deploy ML applications.

TEXT SUMMARIZATION API: The summarising API allows you to take the most important sentences from a document and summarise their meaning.

FLASK: Flask is a web framework and a Python module that makes it simple to create web applications. It has a simple and extensible core: it's a microframework without an ORM (Object Relational Manager) or other things like that.

PYTHON3: Python is a dynamically semantic, interpreted, object-oriented high-level programming language. Python's concise, easy-to-learn syntax prioritises readability, which lowers software maintenance costs. Modules and packages are supported by Python, which fosters programme modularity and code reuse.

HTML / CSS: HTML (Hyper Text Markup Language) is a markup language that is used to define the meaning and structure of web content. CSS (Cascading Style Sheets) is a stylesheet language for describing the presentation of an HTML document.

LITERATURE SURVEY

There have been a number of notable works in the field of text summarization in recent years. Earlier research focused primarily on single-document text summarization. When compared to previous ways, technology has advanced and computing power has improved, paving the way for a faster, more effective, and more accurate means of processing documents.

Niladri[9] Chatterjee, Amol Mittal, and Shubham Goyal presented an extractive based Text Summarization technique based on Genetic Algorithms. They modelled the single document as a Directed-Acyclic-Graph in this study. Each DAG edge is given a weight based on a schema described in the paper. They employ an Objective function to express the summary's quality in terms like readability (readability factor), cohesiveness (cohesion factor), and topic relation (topic relation factor).

The most common words, according to Luhn[10] convey the text's most essential topic. His plan was to assign a score to each sentence depending on the number of times the terms appeared, and then chose the one with the greatest score.

Methods based on location, title, and cue words were proposed by Edmunson[14]. He stated that the topic information should be presented in the first few sentences of a paper or the first paragraph, and that it should be included in the summary. One of the shortcomings of statistical approaches is that they ignore semantic relationships between texts.

Goldstein [18] introduced query-based summarization, which extracts relevant sentences from a document based on the query that was executed. The extraction criterion is expressed as a query. The likelihood of being included in a summary grows as the number of words in the query and a sentence increases. Goldstein[11] [12] also looked at news article summarization and ranked sentences in the document using statistical and linguistic criteria.

For lexical chain computation, H. Gregory Silber and McCoy [15] devised a liner time approach. By constructing an intermediate representation, the author follows Barzilay and Elhadad [20] in using lexical chains to extract significant concepts from the source text.

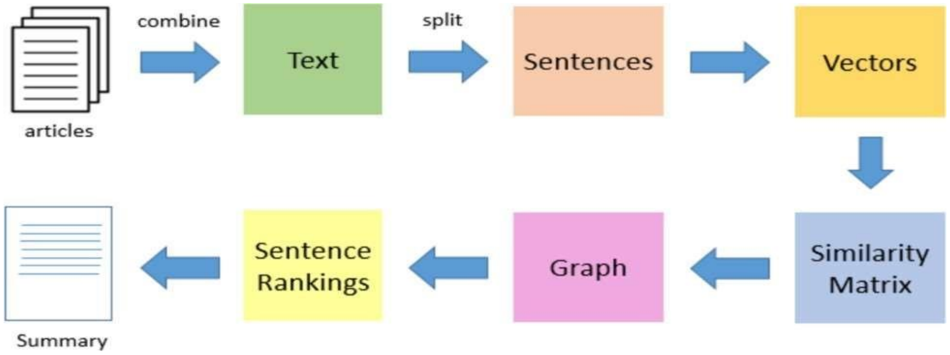
The use of graph theory [16] is another way for summarization. To generate a semantic network of the original content, the author presented an approach based on subject-object-predicate (SOP) triples from individual phrases. The relevant concepts, which contain the meaning, are dispersed among the sentences. Identifying and exploiting links among them, according to the author [16], could be useful for extracting pertinent text.

One of the researchers, IIT Bombay's Pushpak Bhattacharyya [17], proposed a Word Net-based summarising method. The document is summarised by using Word-net to generate a sub-graph. The Word Net is used to apply weights to nodes in the sub-graph in relation to the synsnet. The most widely used text summarising approaches employ either a statistical or linguistic approach, or a combination of both.

COMPARISON OF SUMMARIZATION METHODS:

Type of summarization methods	Subtype	Concept	Advantages	Disadvantages	Application/ Work done
1) Approaches	Abstractive	It is the process of reducing a text document in order to create a summary	Good compressi on ratio. More reduced text and semantica lly related summary.	Difficult to compute	[18]
	Extractive	It consists of selecting important sentences	Easy to compute	Suffers from inconsistency	[19]
2) Languages	Mono-lingual	Can accept input only with specific language	Need to work with only one language	Cannot handle different languages	[20]
	Multi-Lingual	Can accept documents in different language	Can deal with multiple language	Difficult to implement	[18]
3) No. of input document	Single-document	Can accept only one input document	Less overhead	Cannot summarize multiple related documents	[19]
	Multi-document	Can accept multiple input documents	Multiple documents of same topic can be summariz ed to single document	Difficult to implement	[19]

BLOCK DIAGRAM :



- ➢ The first step would be to concatenate all the text contained within the articles.
- ➢ Then split the text into individual sentences after that,
- ➢ We'll find vector representations (word embeddings) for each and every sentence in the next stage.
- ➢ After that, similarity between sentence vectors is determined and stored in a matrix.
- ➢ For sentence rank calculation, the similarity matrix is turned into a graph with sentences as vertices and similarity scores as edges.
- ➢ Finally, a set of high-ranking sentences from the final summary.

RESULT /CONCLUSION:

There is an information overload as a result of the rapid advancement of technology and the widespread usage of the Internet. This difficulty can be overcome if there are powerful text summarizers that provide a document summary to assist users. As a result, a system must be developed that allows a user to quickly access and obtain a summary document. A document can be summarised in one of two ways: extractive or abstractive procedures. Extractive text summarization is simpler to construct. However, abstractive text summarization is more powerful since it produces a summary that is semantically related but difficult to construct.

REFERENCES :

[1] Gonçalves, Luís. 2020. “Automatic Text Summarization with Machine Learning — An overview.”Medium.com. https://medium.com/luisfredgs/au(Gonçalves,2020)Automatic-text summarization-with-machine-learning-an-overview-68ded5717a25. text summarizer”, International Conference on Computational Linguistics, 2004.

[2] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh.A Comprehensive Survey on Text Summarization Systems.

[3] Edmundson, H.P., 1968.New methods in automatic extraction. J. ACM16 (2), 264–285. S.

[4] Luhn, H.P., 1959. The automatic creation of literature abstracts. IBM J.Res. Develop., 159–165

[5] Barzilay, R., Elhadad, M.Using Lexical Chains for Text Summarization. In Proc. ACL/EACL’97 Workshop on Intelligent Scalable Text summarization, Madrid, Spain,1997, pp. 10–17.

[6] Silber G.H., Kathleen F. McCoy. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization.

[7] William P. Doran, Nicola Stokes, John Dunnion, and Joe Carthy, “Comparing lexical chain-based summarization approaches using an extrinsic evaluation,” In Proc. Global Word net Conference (GWC 2004), 2004.

[8] Kedar Bellare, Anish Das Sharma, Atish Das Sharma, Navneet Loiwal and Pushpak Bhattacharyya.Generic Text Summarization Using Word net. Language Resources Engineering Conference (LREC 2004), Barcelona, May, 2004.

[9] Niladri Chatterjee, Amol Mittal and Shubham Goyal's "Single Document Extractive Text Summarization Using Genetic Algorithms" (2012)

[10] Luhn, H.P., 1959. The automatic creation of literature abstracts. IBM J.Res. Develop., 159–165.

[11] Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999.Summarizing text documents: Sentence selection and evaluation metrics. In: Proc. ACM-SIGIR'99, pp. 121–128.

[12] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh.A Comprehensive Survey on Text Summarization Systems. 2009 In proceeding of: Computer Science and its Applications, 2nd International Conference.

[13] Barzilay, R., Elhadad, M.Using Lexical Chains for Text Summarization. In Proc. ACL/EACL'97 Workshop on Intelligent Scalable Text summarization, Madrid, Spain,1997, pp. 10–17.

[14] Edmundson, H.P., 1968.New methods in automatic extraction. J. ACM16 (2), 264–285. S.

[15] Silber G.H., Kathleen F. McCoy. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. Computational Linguistics 28(4): 487-496, 2002.

[16] Kedar Bellare, Anish Das Sharma, Atish Das Sharma, Navneet Loiwal and Pushpak Bhattacharyya.Generic Text Summarization Using Word net. Language Resources Engineering Conference (LREC 2004), Barcelona, May, 2004.

[17] J. Leskovec, M. Grobelnik, N. Milic-Frayling.Extracting Summary Sentences Based on the Document Semantic Graph. Microsoft Research, 2005.

[18] Eduard Hovy and Chin Yew Lin, "Automated text summarization in SUMMARIST", MIT Press, 1999, pages 81–94. [15] Http://Web.science.mq.edu.au/swan/summarization/proje cts_full.html

[19] [19] Http://Web.science.mq.edu.au/swan/summarization/proje cts_full.html

[20] Martin Hassel, Nima Mazdak, "A Persian text summarizer", International Conference on Computational Linguistics, 2004.