Text Summarizer Using NLP (Natural Language Processing)

AAKASH SRIVASTAVA[1], KAMAL CHAUHAN[2], HIMANSHU DAHARWAL[3], NIKHIL MUKATI[4], PRANOTI SHRIKANT KAVIMANDAN[5]

[1, 2, 3, 4, 5] *Department of Computer Science and Business System Bharati Vidyapeeth Deemed University College of Engineering, Pune*

Abstract- ***Enormous amounts of information are available online on the World Wide Web. To access information from databases, search engines like Google and Yahoo were created. Because the amount of electronic information is growing every day, the real outcomes have not been reached. As a result, automated summarization is in high demand. Automatic summary takes several papers as input and outputs a condensed version, saving both information and time. The study was conducted in a single document and resulted in numerous publications. This report focuses on the frequency-based approach for text summarization.***

Indexed Terms- ***Automatic summarization, Extractive, frequency-based, Natural Language Processing.***

I. INTRODUCTION

Text summary is the way of selecting important points from the provided article or a document that can be reduced by a program. As the data overload problem increased, so did the interest in capturing the text as the amount of data increased. Summarizing a large document manually is challenging since it requires a lot of human effort and is time-consuming.

There are mainly two methods for summarizing the text document that can be done by using extractive and abstractive techniques.

Extractive summaries concentrate on selecting important passages, sentences, words, etc. from the primary text and connecting them into a concise form. The importance of critical sentences is concluded on the basis of analytical and semantic features of the sentences.

Summary systems are usually based on sentence delivery methods and for understanding the whole document properly as well as for extracting the important sentences from the document.

The technique of generating a brief description that comprises a few phrases that describe the key concepts of an article or section is known as abstractive summarization.

This function is also included to naturally map the input order of words in a source document to the target sequence of words called the summary.

II. LITERATURE SURVEY

The Internet is a vast source of electronic information. But the result of information acquisition becomes a tedious task for people. Therefore, automated summaries began the search for automatic retrieval of data from documents using our precious time. H.P. Luhn was the first to invent an automatic summary of the text in 1958.

There are helpful ways to produce a. summary - extraction and abstraction. Extraction is independent of the domain and takes key sentences and provides a summary on the other hand, abstracting depends on the domain and taking personal information by understanding the entire text and adjusting the policy to produce a summary. There are several methods that use different methods to obtain a summary of a text.

A. Abstractive Summarization Approach

Summarizations using abstractive techniques are broadly classified into two categories: Structured based approach and Semantic based approach.

1) Structured Based Approach:

Structured based approach encodes most important information from the document through cognitive schemes such as templates, extraction rules and other structures such as tree, ontology, lead and body phrase structure.

Tree Based Method

- It uses a dependency tree to represent the text of a document. -It uses either a language generator or an algorithm for generation of summary.
- It walks on units of the given document read and easy to summary.

Template Based Method

- It uses a template to represent a whole document. -Linguistic patterns or extraction rules are matched to identify text snippets that will be mapped into template slots.
- It generates summary is highly coherent because it relies on relevant information identified by IE system.

Ontology Based Method

- Use ontology (knowledge base) to improve the process of summarization. -It exploits fuzzy ontology to handle uncertain data that simple domain ontology cannot.
- Drawing relation or context is easy due to ontology
- Handles uncertainty at reasonable amount.

2) Semantic Based Approach:

In Semantic based approach, semantic representation of document is used to feed into natural language generation (NLG) system. This method focuses on identifying noun phrase and verb phrase by processing linguistic data. Brief abstract of all the techniques under semantic based approach is provided.

Multimodal semantic model

A semantic model, which captures concepts and relationship among concepts, is built to represent the contents of multimodal documents.

An important advantage of this framework is that it produces abstract summary, whose coverage is excellent because it includes salient textual and graphical content from the entire documents.

Information Item Based Method

- The contents of summary are generated from abstract representation of source documents, rather than from sentences of source documents. -The abstract Representation is Information Item, which is the smallest element of coherent information in a text.
- The major strength of this approach is that it produces short, coherent, information rich and less redundant summary.

B. Extractive Summarization Techniques

An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences.

Term Frequency Inverse Document Frequency Method

- Sentence frequency is defined as the number of sentences in the document that contain that term.
- Then this sentence vectors are scored by similarity to the query and the highest scoring sentences are picked to be part of the summary.

Cluster Based Method

- It is intuitive to think that summaries should address different “themes” appearing in the documents.
- If the document collection for which summary is being produced is of totally different topics, document clustering becomes almost essential to generate a meaningful summary.
- Sentence selection is based on similarity of the sentences to the theme of the cluster (Ci). The next factor that is location of the sentence in the document (Li). The last factor is its similarity to the first sentence in the document to which it belongs (Fi).

Si =W1 * Ci + W2 * Fi+ W3 *Li

Where, W1, W2, W3 are weight age for inclusion in summary.

- The clustering k-means algorithm is applied.

Graph Theoretic Approach

- Graph theoretic representation of passages provides a method of identification of themes.
- After the common pre-processing steps, namely, stemming and stop word removal; sentences in the documents are represented as nodes in an undirected graph.

2.1 Frequency based approach:

- Term frequency (TF):

TF mainly determines that how often a word appears in a text document and it is considered to be an important factor. The paragraphs in the document are divided into sentences based on the punctuation marks that appears at the end of every sentence.

- Keyword frequency:

The high frequency words in the sentence are known as keyword. It measures the frequency for every word once you've refined the content. Keywords are the terms that have the most important frequency. The word score is organized as a keyword, and the phrase is given some fixed points for each keyword found in the text based on this feature.

- Stop words filtering:

Any document will have a lot of words that appear regularly but do not give the document less or more meaning. Words like 'on', 'the', 'is' and 'and' appear frequently in the English language and there are many examples of many texts. While searching, these words do not add up value to the information when users submit a query.

2.2 Clustering approach

- K-means clustering:

This approach aims to classify n observed in k groups where each recognition belongs to a category with a descriptive meaning, acting as a collective example.
k-means can be applied to data with small size, is numerical, and continuous. The applications that can be benefited by the k-means algorithm are public transport data analysis, targeting crime hotspots, insurance fraud detection, customer segregation, document collection, etc.

In our project we have used extractive approach for text summarization. To be specific we have used TF-IDF method for summarization. From these discussions, we have observed that many techniques suffer from various challenges, for example, the graph-based methods have imitation in data size, the clustering-based methods require prior knowledge of the number of clusters, the MMR approaches have uncertainty for the coverage and non-redundancy aspects in the summary, etc. Tree Based Method lacks a complete model which would include an abstract representation for content selection. Template Based Method Requires designing of templates and generalization of template is to difficult. Ontology Based Method This approach is limited to Chinese news only. And Creating Rule based system for handling uncertainty is a complex task.

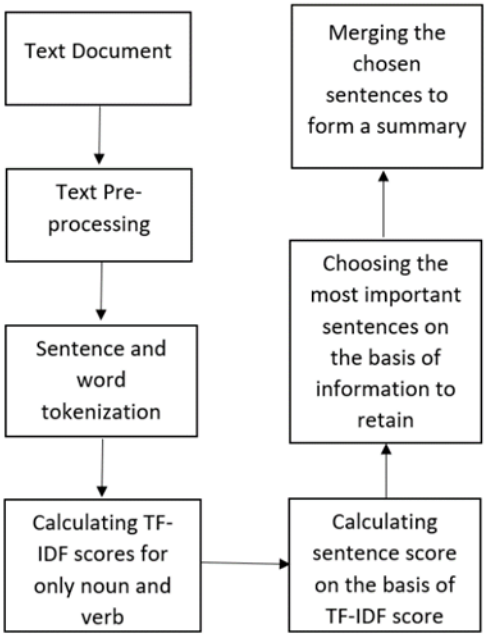
III. PROBLEM STATEMENT

We have used NLP, which seeks to summarize articles by picking a collection of words that hold the most essential information, can address this problem with the help of extractive summarizer. This approach takes a significant portion of a phrase and utilizes it to create a summary. To define sentence verbs and subsequently rank them in terms of significance and similarity, a variety of algorithms and approaches are utilized.

There is a great need for text summary techniques to address the amount of text data available online to help people find the right information and use the right information quickly. In addition, the implementation of text summaries reduces reading time, speeds up the process of researching information, and increases the information that may not be in one field.

This research paper focuses on the frequency-based approach for text summarization.

The steps involved in text summarizer are Sentence and word tokenization and then calculating sentence score on the basis of TF-IDF score which is being used to select the most important sentences to retain the information and merge it to form a summary.



STEP-1: Import all necessary libraries

NLTK (Natural Language toolkit) is a widely used library while we are working with text in python. Stop words contain a list of English stop words, which need to be removed during the pre-processing step.

STEP-2: Generate clean sentences

Text processing is the most important step in achieving a constant and positive approach result. The processing steps removes special digits, word, and characters.

STEP-3: Calculate TF-IDF and generate a matrix

We'll find the TF and IDF for each word in a paragraph.

TF (t) = (Frequency of t from document) / (total_no. Of t in the document)

IDF (t) = log_e (total_no. Of documents / No. of documents with t it) [4]

Now, we will be generating a new matrix after multiplying the calculated TF and IDF values.

STEP-4: Score the sentences

Here, we use TF-IDF word points in a sentence to give weight to a paragraph. However, Sentence scoring varies with different algorithms.

STEP-5: Generate the summary

This is the last stage of text summarization. Top sentences are calculated based on the score and retention rate given to the user are included in the summary and finally, a summary is created.

IV. TEST RESULTS

- Short Input - While performing the testing for smaller inputs we get an error of minimum value where it denotes about the word's frequency is not greater than required frequency to calculate the summary.

- Foreign Language - While giving input in any language, it successfully performs the summarization process and a meaningful summary is obtained.

- Improper URL - If the given URL hasn't a defined and a sequential data which can be summarize then it displays the error as mentioned below since the web scrapper can't get the exact data from the URL from which our summary could be generated.

- Illogical Text - If any illogical or meaningful text is given as an input, then the summary won't come as it will not make sense to generate a summary of punctuation marks or any stop words. As given below the output is generated where it shows that the given text could be stop words which gets eliminated in the pre-processing phase of summarization.

- Repeated Text - If the repeated text is given as input to generate the summary, then the summary will be obtained but it will also be in repeated manner since the text are repeating due to which the program can't differentiate between the meaning of the generated summary. So based on the repeated input, summary is generated.

Test ID	Test case	Description	Expected Result	Status
T01	Short Input	Input is too short	Should not generate summary	Pass
T02	Foreign Language	Input is in different language other than English	Summary should be generated	Pass
T03	Improper URL	If the data cannot be scrapped from the given URL	Cannot extract data	Pass
T04	Repeated Text	When the same text is given number of times	Summary for the text repeating multiple times should only be shown once on the output screen	Fail
T05	Illogical input	Meaningless inputs like symbols, punctuation marks etc.	It should not generate a summary and should show an error message	Pass

CONCLUSION

Text summaries have been shown to be useful for natural language processing tasks such as question and answer or other related fields of computer science such as text classification and data retrieval. And access time for information search will be improved. At the same time, sequencing enhances the effect and its algorithms are less biased than human creams. Using a text summary system, commercial capture services allow users to increase the number of texts they can process.

FUTURE SCOPE

In this section, we will list some of the future extensions for this study. In this article, we focused on summarizing news articles under the auspices of sports and technology. The strategies proposed here are flexible in some domains. One of the future plans would be to use an overview framework that focuses on the topic in news articles or blogs and to increase work on machine-dependent methods. Summaries focused on the headline article can be very accurate and very important for users. It would be even more interesting to work on topic modeling and summarizing in the future media domain.

REFERENCES

[1] Adhika Widyassari, S. R. (2020). Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, 18.

[2] Amigó E, G. J. (2005). a framework for the evaluation of text summarization systems. *proceedings of the 43rd annual meeting on association for computational linguistics.* ACL ’05.

[3] Antiqueira L, O. O. (2007). A complex network approach to text summarization. *Information Sciences*.

[4] B. Cretu, Z. C. (2002). Automatic summarization based on sentence extraction. *International Journal of Applied Electromagnetic and mechanics*.

[5] Brownlee, J. (2019, August 7). *A Gentle Introduction to Text Summarization*. Retrieved from https://machinelearningmastery.com/: https://machinelearningmastery.com/gentle-introduction-text-summarization/

[6] Changjian Fanga, D. M. (2016, March 5). *Word-sentence co-ranking for automatic extractive text summarization*. Retrieved from https://www.sciencedirect.com/: https://www.sciencedirect.com/science/article/abs/pii/S0957417416306959?via%3Dihub

[7] Conroy, J. M. (2001). Text summarization via hidden markov models. *Proceedings of SIGIR '01*.

[8] D. Gillick, K. R. (2009). A global optimization network for meeting summarization. *Proc. IEEE Int. Conf. Acoust*, 1-4.

[9] Darji, H. (2020, January 8). *Text Summarization-Key Concepts*. Retrieved from https://medium.com/: https://medium.com/@harshdarji_15896/text-summarization-key-concepts-23df617bfb3e

[10] Evans, D. K. (2005). *Similarity-based multilingual multidocument summarization*. Technical Report CUCS-014-05.

[11] Gupta V, L. G. (2010). A survey of text summarization extractive techniques. *J Emerg Technol Web Intell*, 258-268.

[12] J. Patel, P. (2015). https://machinelearningmastery.com/gentle-introduction-text-summarization/. *International Journal of Engineering and Computer Science*, 5.

[13] Jain, A. (2019, April 1). *Automatic Extractive Text Summarization using TF-IDF*. Retrieved from Medium.com: https://medium.com/voice-tech-podcast/automatic-extractive-text-summarization-using-tfidf-3fc9a7b26f5

[14] KS, J. (2007). Automatic summarising: the state of the art. *Inf Process Manag 43*, 1449-1487.

[15] Kumar, T. (2014). *Automatic Text Summarization.* Rourkela.

[16] Mayo, M. (2019, November). *Getting Started with Automated Text Summarization.* Retrieved from https://www.kdnuggets.com/: https://www.kdnuggets.com/2019/11/getting-started-automated-text-summarization.html

[17] Mr. Vikrant Gupta, M. P. (2012). An Statiscal Tool for Multi-Document Summarization. *International Journal of Scientific and Research (ISSN 2250-3153)*.

[18] Neelima Bhatia, A. J. (2015). Literature Review on Automatic Text Summarization: Single and Multiple Summarizations. *International Journal of Computer Applications*, 1-5.

[19] Okumura, H. T. (2009). Text Summarization Model based on the budgeted median problem. *Proc. 18th ACM Conf. Inf. Knowledge*, 1-4.

[20] Opidi, A. (2019, April 15). *A Gentle Introduction to Text Summarization in Machine Learning*. Retrieved from https://blog.floydhub.com/: https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/

[21] Panchal, A. (2019, June 10). *NLP — Text Summarization using NLTK: TF-IDF Algorithm*. Retrieved from https://towardsdatascience.com/: https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3

[22] *Recent automatic text summarization techniques: a survey*. (2019, March 29). Retrieved from https://link.springer.com/: https://link.springer.com/article/10.1007/s10462-016-9475-9

[23] S, S. (2011). Automatic Text Summarization: The current state of the art. *International Journal of*.

[24] YLLIAS CHALI, S. A. (2011). Query-focused multi-document summarization: automatic data annotations and supervised learning approaches. *Cambirdge University Press*.