Text Summarization System

Parth Shrivastava
Department of CSE
Lovely professional University

Ramakant
Department of CSE
Lovely professional University

Abstract :

Text summarization is used to summarize the main points of the document provided. We've gone through many techniques of text summarization in this paper. Techniques covered under are from many research papers launched over duration of 11 years from 2008-2019. The study includes various journals and conference publications that were carefully selected for extraction, identification, and analysis to capture and describe the primary research topics and trends in text summarization. This analysis offers a comprehensive overview of the focus areas in this field, including commonly used datasets, preprocessing steps, feature selection, techniques, methods, evaluation metrics, and challenges. Through this examination, the paper identifies and explains trends within text summarization research, references available public datasets, discusses preprocessing methods and key features, and highlights frequently employed techniques and methods as benchmarks for further development. Finally, We've made a system for text summarization and have got my own unique summary for the given dataset. The process involves Text Cleaning then word Tokenization, word frequency table succeeded by sentence tokenization finally giving out the required Summary.

Keywords: Natural language processing , Text summarization, tokenization, summary, frequency based

Introduction:

With the proliferation of internet usage and the vast growth in data availability, people are increasingly overwhelmed by the massive volumes of information and documents accessible online. This influx has motivated researchers to develop automated approaches for summarizing texts. Automatic text summarization aims to produce concise summaries that retain the essential points of the original text, enabling users to quickly grasp key information without losing the intent of the source material (Allahyari et al., 2017; Gambhir & Gupta, 2017).

Research in this area began in the mid-20th century, initially explored by Lun (1958) through statistical methods like word frequency diagrams, and has since evolved to include various techniques. These methods can be broadly classified based on the number of documents and the type of summary generated, distinguishing between single-document and multi-document summarization and between extractive and abstractive approaches. Single-document summarization derives summaries from one source (Radev et al., 2001), while multi-document summarization consolidates information from multiple sources on a common topic (Qiang et al., 2016; Ansamma et al., 2017).

Extractive summarization selects and reorders sentences or phrases directly from the text (Khan & Salim, 2014), while abstractive summarization generates new sentences or paraphrases based on the original content, often requiring advanced natural language processing techniques (Gambhir & Gupta, 2017). The latter is generally more complex but can yield more natural, coherent summaries. The ongoing development in text summarization has led to both conventional and cutting-edge techniques, including neural networks and real-time summarization, that aim to provide accurate, informative, and adaptable summaries.

Literature Review:

The development of automated text summarization techniques can be traced back to H.P. Luhn's pioneering work in 1958, which aimed to simplify data retrieval from documents by generating automated summaries. Since then, two primary techniques for summarization have emerged: extractive and abstractive methods. Extractive summarization, which is domain-independent, involves selecting key sentences or phrases from the original text to form a coherent summary. Abstractive summarization, on the other hand, depends on a comprehensive understanding of the entire text and generates new sentences based on the extracted information, often requiring advanced NLP capabilities to maintain contextual relevance.

One of the earliest methods within extractive summarization is the frequency-based approach, where Term Frequency (TF) is used to quantify the importance of words by their frequency of appearance in the document. High-frequency terms, or keywords, help identify essential content, and stop words—common words like “the,” “is,” or “and”—are filtered out to improve summary relevance. Another widely adopted technique is the clustering approach, specifically the K-means clustering algorithm. This method segments text into clusters based on descriptive patterns, with applications across fields such as customer segmentation, insurance fraud detection, and document collection.

Between 2008 and 2019, the field of text summarization saw growing research interest, evidenced by an increasing number of published papers. Early years, from 2008 to 2012, saw minimal engagement, with only a few publications annually. However, interest expanded significantly from 2013 onward, peaking in 2018 with 18 studies published in that year alone. This pattern underscores the evolving importance of summarization research, with a considerable body of work developed in recent years. This growth trend reflects not only advancements in natural language processing (NLP) techniques but

also the persistent demand for efficient information retrieval systems capable of handling the modern data deluge.

Process And Approach:

With the rapid expansion of online content, there's an increasing demand for automated methods to quickly extract relevant information from large amounts of text. NLP, specifically through extractive summarization techniques, addresses this need by selecting and highlighting the most crucial phrases or sentences from a document. This method saves time, assists in research, and allows for efficient access to information without combing through entire texts. The extractive summarization process generally follows a series of structured steps, including text cleaning, tokenization, constructing word frequency tables, scoring sentences, and assembling the final summary.

Text Cleaning

Text cleaning is the foundational step that prepares the document for further processing by removing extraneous elements like punctuation, digits, and special symbols. This step also eliminates "stop words" (e.g., "is," "and," "the")—common words that generally do not contribute meaningful content to the text. By focusing on essential terms and discarding irrelevant ones, this process simplifies the dataset, setting the stage for effective summarization.

Word Tokenization

Following text cleaning, word tokenization splits the text into individual words, known as tokens, allowing each word to be examined in isolation. Tokenization supports frequency calculations and helps to build a dataset where each token can be independently analyzed for significance. This step is pivotal for identifying which

terms recur frequently, as these often indicate key themes in the document.

Word Frequency Table

The word frequency table plays a central role in identifying significant words within the text. This table contains each word and its frequency of occurrence after stop words are filtered out. High-frequency words are generally more central to the main ideas of the document. Combining frequency counts with metrics like Term Frequency-Inverse Document Frequency (TF-IDF) helps distinguish the most relevant words, enhancing the summarization by pinpointing key phrases accurately.

Sentence Tokenization

Next, sentence tokenization divides the document into sentences, facilitating the scoring process. During this step, each sentence is assessed for its relevance by considering the cumulative frequency scores of the words it contains. Sentences containing more high-frequency or key terms are given higher scores, making them more likely candidates for inclusion in the summary.

Summary Generation

Finally, the highest-scoring sentences are compiled to create a summary, with each sentence contributing significant content from the original text. This method ensures that only the most informative and relevant sentences are included, yielding a concise summary that conveys the core message of the document effectively. By following these structured steps, extractive summarization provides a powerful tool to quickly capture essential information and support users in managing large volumes of textual data.

Fig1 *Importing libraries*

```
import spacy
from spacy.lang.en.stop_words  import STOP_WORDS
import string
```

```
# List of stop words
stopwords=list(STOP_WORDS)
stopwords[:100]
```

Fig2 Word frequency calculation

```
def word_frequency(doc):
    mytokens=[token.text.lower().strip() for token in doc if token.text not in punctuations]
    mytokens=[token for token in mytokens if token not in stopwords]

    return mytokens

def calc_word_frequency(temp):
    for word in temp:
        if word not in word_frequencies.keys():
            word_frequencies[word]=1
        else:
            word_frequencies[word]+=1
```

Fig3 Sentence Tokenization

```
mysentences=[sents for sents in doc.sents]
mysentences
```

```
[
 There are broadly two types of extractive summarization tasks depending on what the summa
rization program focuses on.,
 The first is generic summarization, which focuses on obtaining a generic summary or abst
act of the collection (whether documents, or sets of images, or videos, news stories et
c.).,
 The second is query relevant summarization, sometimes called query-based summarization,
```

Conclusion:

Text summarization is a valuable area within natural language processing that supports tasks such as question answering, text classification, and data retrieval. Automated text summarization systems enhance information access by reducing search time and allowing users to process larger amounts of content. Through researchers have identified key trends, datasets, techniques, and challenges within this field, providing a well-organized overview that guides future research.

Extractive summarization, which focuses on selecting high-frequency, relevant sentences, has proven more straightforward than abstractive summarization, which involves rephrasing and synthesis and remains more complex and less explored. Key features influencing effective summarization include keywords, frequency, similarity, and sentence structure, with machine learning approaches standing out due to their adaptive capabilities. Statistical methods also play a critical role and are often combined with machine learning and fuzzy logic to enhance performance.

Future work in text summarization may focus on refining feature selection, improving preprocessing techniques, and integrating various methodologies, such as combining statistical and fuzzy-based methods with machine learning. Challenges remain in optimizing coherence and grammar, especially in abstractive summarization. Exploring less commonly used datasets, like legal documents and tourism information, may further validate and expand summarization techniques across diverse applications.

Citations:

1. T. Kumar, "Automatic Text Summarization," Rourkela, 2014.

2. P.J. Patel, "https://machinelearningmastery.com/gentle-introduction-text-summarization/," International Journal Of Engineering And Computer Science, p. 5, 2015.

3. A. Jain, "Automatic Extractive Text Summarization using TF-IDF," 1 April 2019. [Online].Available: https://medium.com/voice-tech-podcast/automatic-extractive-text-summarization-using-tfidf-3fc9a7b26f5.

4. A. Panchal, "NLP—Text Summarization using NLTK: TF-IDF Algorithm," 10 June 2019. [Online].Available: https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3

5 Abbasi-ghalehtaki, Razieh, Hassan Khotanlou, and Mansour Esmaeilpour. "Fuzzy evolutionary cellular learning automata model for text summarization." *Swarm and Evolutionary Computation* 30 (2016): 11-26.

6. J. Brownlee, "A Gentle Introduction to Text Summarization," 7 August 2019. [Online]. Available:https://machinelearningmastery.com/gentle-introduction-text-summarization/.

7. A. Opidi, "A Gentle Introduction to Text Summarization in Machine Learning," 15 April 2019.[Online]. Available: https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/.

8. H. Darji, "Text Summarization-Key Concepts," 8 January 2020. [Online]. Available:https://medium.com/@harshdarji_15896/text-summarization-key-concepts-23df617bfb3e.

9. J.M.a.O.D.P. Conroy, "Text summarization via hidden markov models," Proceedings of SIGIR'01, 2001.

10. D.M.D.W. Changjian Fanga, "Word-sentence co-ranking for automatic extractive textsummarization," 5 March 2016. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417416306959?via%3Dihub

11. (PDF) Text Summarizer using NLP (Natural Language Processing). Available from: https://www.researchgate.net/publication/365790121_Text_Summarizer_using_NLP_Natural_Language_Processing [accessed Nov 11 2024].