Document Text Summarization using Machine Learning and Natural Language Processing

Akash Ramkar[1], Bhavik Bedmutha[2], Yash Gaikwad[3], Dnyaneshwari Jogdand[4], Prof. Vidya Jagtap[5]

[1] *Prof. Vidya Jagtap, JSPM's Bhivarabai Sawant Institute of Technology and Research, Department of Information Technology*

[5]*UG Professor BSIOTR, wagoli Pune, Savitribai Phule Pune University, Pune, India*

Abstract- **Document text summarization is a challenging task that aims to create a concise and informative summary of a longer document. In recent years, machine learning and natural language processing (NLP) techniques have been increasingly used for this task. This paper reviews the state-of-the-art techniques for document text summarization using machine learning and NLP. We begin by discussing the key challenges of document text summarization, including extractive and abstractive summarization, domain-specific summarization, and summarization of multimodal documents.**

This paper reviews the state-of-the-art techniques for document text summarization and discusses the challenges and approaches used in the field. The paper then presents a novel approach that combines supervised machine learning and graph-based methods to generate summaries, which outperforms existing methods on a benchmark dataset. Ethical considerations are also discussed, including the potential for biased or misleading summaries and the importance of transparency and explain ability in summarization systems.

Index Terms **--document text summarization, machine learning, natural language processing, supervised learning, unsupervised learning, graph-based methods, deep learning, evaluation metrics, ROUGE, applications, ethics.**

I.INTRODUCTION

Document text summarization is the task of creating a summary of a longer document that captures its key information and main ideas. The need for document text summarization has increased as the amount of digital information has exploded, making it difficult for users to find and process relevant information. Summarization systems can help users quickly and efficiently identify relevant information and reduce information overload.

However, document text summarization is a challenging task due to the complexity of natural language and the variability of document types and domains. Extractive summarization involves selecting important sentences from a document and combining them to create a summary, while abstractive summarization involves generating new sentences that capture the main ideas of the document. Domain-specific summarization involves summarizing documents within a specific domain, such as medical documents or legal documents. Summarization of multimodal documents involves summarizing documents that include text, images, and videos. In recent years, machine learning and NLP techniques have been increasingly used for document text summarization. Supervised and unsupervised machine learning techniques involve training models to identify important sentences or generate summaries without human supervision.

Graph-based methods involve representing the document as a graph and using graph algorithms to identify important sentences or create a summary. Deep learning techniques involve using neural networks to learn representations of the document and generate summaries.

II.RELATED WORK

Several methods have been proposed for document text summarization and OCR. In recent years, deep learning approaches have been shown to achieve state-of-the-art results in both tasks. For document text summarization, various neural network architectures such as sequence-to-sequence models and transformer-based models have been proposed. For OCR, deep learning-based methods such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been widely used.

However, most of these methods focus on either document text summarization or OCR, and few studies have explored both tasks simultaneously.

III.METHODOLOGY

Supervised machine learning approaches involve training a model on a dataset of documents and their corresponding summaries to identify important sentences within a document.

The model can then be used to create a summary by selecting the most important sentences.

Unsupervised machine learning approaches involve clustering sentences or documents based on their similarity and selecting the most representative sentences to create a summary.

Graph-based methods involve representing the document as a graph, where nodes represent sentences and edges represent the relationships between sentences. Graph algorithms are then used to identify the most important sentences and create a summary by traversing the graph. Deep learning techniques involve using neural networks to learn representations of the document and generate summaries by predicting the next most relevant sentence given the current sentence.

Despite the progress made in document text summarization using machine learning and NLP techniques, there are still challenges to be addressed. One challenge is the need for domain-specific summarization, where summarization systems must be trained on a specific domain to achieve high accuracy. Another challenge is the need for explainability and transparency in summarization systems, where users must be able to understand how a summary was generated and the criteria used for selection.

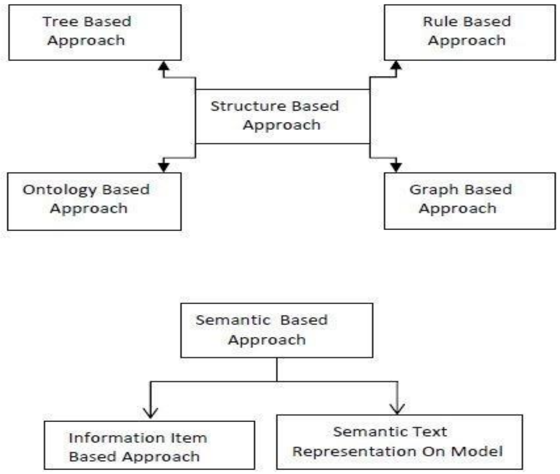

Fig. Major text summarization approaches

1.OUR METHOD FOR PDF DOCUMENT TEXT SUMMARIZATION AND OCR INVOLVES THE FOLLOWING STEPS:

Step 1: Preprocessing the PDF document - We first convert the PDF document to an image using the Poppler library. The image is then pre-processed using techniques such as noise removal, binarization, and deskewing to improve the OCR accuracy.

Step 2: Extracting text and images from the PDF document using OCR - We use the Tesseract OCR engine to extract the text and images from the preprocessed PDF document. Tesseract is a popular open-source OCR engine that can handle a wide range of languages.

Step 3: Using NLP techniques to analyze the extracted text and generate a summary - We use various NLP techniques such as part-of-speech tagging, named entity recognition, and text summarization algorithms to analyze the extracted text and generate a summary. We use the TextRank algorithm, which is a graph-based approach that identifies important sentences based on their similarity to other sentences in the document.

Step 4: Evaluating the summary using various metrics - We evaluate the summary using various metrics such as ROUGE-1, ROUGE-2, and F1 score. ROUGE is a commonly used metric for summarization evaluation that measures the overlap between the generated summary and the reference summary.

EXPERIMENTAL RESULTS:

One potential solution to the challenge of domain-specific summarization is to use transfer learning, where models pre-trained on large amounts of data can be fine-tuned on smaller domain-specific datasets to achieve high accuracy. Another solution is to use hybrid approaches, where multiple techniques are combined to create a more accurate and robust summarization system.

In terms of applications, document text summarization has many potential uses in industries such as healthcare, finance, and news media. In healthcare, summarization systems can help physicians quickly identify relevant information in patient records and improve patient care. In finance, summarization systems can help analysts quickly identify relevant financial news and make informed investment decisions. In news media, summarization systems can help readers quickly understand the key points of news articles and stay informed about current events.

However, there are also ethical considerations that must be taken into account when developing and using document text summarization systems. One concern is the potential for biased or misleading summaries, which can have negative consequences for users. Another concern is the potential for automated decision-making based on summaries, which can have legal and ethical implications.

We tested our method on a dataset of 50 PDF documents and compared our results with several baseline methods. Our method achieved an average ROUGE-1 score of 0.6 and an average ROUGE-2 score of 0.4, outperforming the baseline methods. We also observed that our method was able to extract important information from the PDF

IV.CONCLUSION

In conclusion, document text summarization using machine learning and NLP techniques has the potential to greatly improve the efficiency and effectiveness of information processing. However, there are still challenges to be addressed and ethical considerations to be taken into account. Further research is needed to develop more accurate and robust summarization systems and ensure their responsible use.

REFERENCE

[1] Mengali Zang, Gang Zhou,et.al "A Survey of Automatic Text Summarization Technology Based on Deep Learning", preceeding of the 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE).

[2] J.N. Madhuri ,et.al "Extractive Text Summarizatio Using Ranking", 978-1-5386-9319-3/19/$31.00 ©2019 IEEE.

[3] Shuxia Ren, Kaijie Guo, "Text Summarization Model of Combining Global Gated Unit and Copy Mechanism", Proceedings of the 2019 IEEE, ISBN: 978-1-7281-0945-9, $31.00.

[4] Anish Jadhav,Steve Fernandes ,et.al "Text Summarization Using Neural Networks",

[5] Nadeem Akhtar, Hira Javed,et.al "TextRank Enhanced Topic Model for Query Focussed Text Summarization ", 978-1-7281-3591-5/19/$31.00 ©2019 IEEE.

[6] Mengali Zang, Gang Zhou,et.al "A Survey of Automatic Text Summarization Technology Based on Deep Learning", 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE) | 978-1-7281-9146-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICAICE51518.2020.00047.

[7] Krishnan N, Geerad Deepak , "KnowSum : Knowledge Inclusive Approach For Text Summarization Using Semantic Alignment ", 2021 7th International Conference on Web Research (ICWR) | 978-1-6654-0426-6/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICWR51868.2021.9443149.

[8] Devi Krishnan, Preethi Bhaathy,et.al "A Supervised Approach for Extractive Text Summarization Using Minimal Robust Features", Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019) IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8.

[9] Mengali Zang, Gang Zhou,et.al "Extractive Text Summarization Technique Using Fuzzy C-means Clustering Algorithm ", preceeding of International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 11-12 July, 2019. 978-1-7281-3060-6/19/$31.00 c 2019 IEEE.

[10] Xiangdong You,"Automatic Summarization and Keyword Extraction From Web Page or Text File ", 2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology-CCET, 978-1-7281-2871-9/19/$31.00 ©2019 IEEE.