Journal of Information Systems Engineering and Management
2025, 10(25s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Text Summarization Framework Using Machine Learning

Sayali Gaikwad[a],, Gayatri Bhandari[b],, Shrishail Patil[c], Venkat Ghodke[d]
[a,b,c]*Department of Computer Engineering JSPM, Bhivarabai Sawant Institute of Technology and Research, Wagholi, Pune,*
[d]*Assistant Professor, E&TC Engineering Department, AISSMS Institute of Information Technology Pune -411001*
Corresponding Author: Venkat Ghodke (venkatghodke@aissmsioit.org)

ARTICLE INFO

Received: 30 Dec 2024

Revised: 18 Feb 2025

Accepted: 01 Mar 2025

ABSTRACT

A crucial problem in natural language processing is text summarization, which calls for models to provide succinct, logical summaries while maintaining the accuracy and consistency of the original data. However, the practical applications of modern abstractive summarizing techniques are limited by the ongoing compromise between diversity and consistency of facts.By adding a new factuality-guided module to the diffusion process, our work suggests a Factuality-Guided Diffusion-Based Abstractive Summarization Model. Intermediate representations are ensured to closely reflect the original text by the factuality module. and the model repeatedly denoizes random noise to provide summaries. Without requiring the fundamental summarization model to be retrained, this approach guarantees variety and factual consistency. The suggested approach outperforms current methods in terms of factual correctness, according to experimental results on two benchmark datasets. The results demonstrate that this approach is a viable way to get over the limitations of the existing abstractive summarizing techniques.

Keywords: Word vectors, word analogies, quick text, integer linear programming, text summarisation, and natural language processing.

INTRODUCTION:

The creation and consumption of textual data have changed dramatically with the advent of language models and automated text generation techniques. From writing product reviews and news articles to fabricating stories. Effective summary methods it used this means of data into brief and understandable summari are becoming more and more necessary as text output increases.

Summarization is an important job that helps Making decisions by extracting crucial data from a variety of sometimes overwhelming sources. The majority of summary models, however, struggle to strike a balance between diversity and consistency of facts.

Factual consistency is usually given priority by conventional abstractive summarizing models like BART, although they are unable to provide distinct summaries. In contrast, diffusion-based models, which are intended to enhance diversity, frequently have factual contradictions. This paper suggests a Factuality-Guided Diffusion-Based Abstractive Summarization Model with a unique factuality module in order to get around this trade-off. This module dynamically activates the summarizer to provide factually correct and varied summaries during the intermediate phases of the diffusion process.

LITERATURE SURVEY

Text summarization involves producing a concise and clear synopsis of a specific document. While maintaining the original meaning, abstractive summarization may include additional terms that were absent from the original text. Dialogue summarizing is a variant of text summary where the original material is a dialogue between two or more people. It's critical to summarize talks, especially when handling long and intricate exchanges like those that take place at contact centers. Large language models (LLMs) are a good option for abstractive text summarization because of their superiority in natural language creation. There is still little research on underrepresented languages like Turkish, despite the fact that LLMs have been extensively investigated for major languages. A thorough examination of LLMs for Turkish abstractive discourse summary is provided by this paper [1].

In order The proposed method seeks to providea Semantic Oriented Abstractive Summarization to generate abstractive summaries with better qualitative content and readability.Better capture the semantic representation of text, we propose Semantic Role Labeling and Predicate Sense Disambiguation in a Joint Model (PSD+SRL) in our work. Semantically based content selection is used, and the Genetic Algorithm is used to extract features. DUC, a common corpus for text summarization, is used in our experimental investigation [2].

The proposed conceptual framework comprises five essential components: mechanisms, dataset, training strategies, optimization algorithms, encoder-decoder architecture, and evaluation metrics.

This article provides a explained of each aspect. This paper aims to enhance comprehension of The components of modern notional data summarization methods based on neural networks

through a current review, while also highlighting the challenges and issues associated with these systems.

To identify common patterns Modern neural abstractive summarization algorithms are designed with a qualitative analysis was conducted using an idea matrix. The new standard of excellence is models with an encoder-decoder architecture based on transformers [3].

A novel GBAS model based on SciBERT and GTN is called in this study.SciBERT encodes scientific data, the Scientific Information Extractor (SciIE) system extracts terminology-related words from articles, and GTN encodes and summarizes large documents. The proposed model is compared to the baseline models [4].

This work introduces a substantial bookmark corpus of 2,069,784 articles in Urdu for the abstractive data. summarizing problem [5].

The authors provide an innovative joint end-to-end solution, Abstractive Summarization of Video Sequences, which use a deep neural network to produce both a natural language description and an abstractive text summary of an input video. This provides a text-based video description and an abstractive summary, enabling viewers to differentiate between pertinent and extraneous information according to their needs. Furthermore, our evaluations indicate that the combined model may achieve superior results compared to baseline methods in individual tasks, utilizing informative, concise, and comprehensible multi-line video descriptions and summaries in a human assessment [6] [13].

The suggested method continuously steers factuality into the intermediate noise at every stage of the denoising process, producing summaries that are both varied and consistent with the original text. In this study, In order to guide factuality throughout the denoising process, Using token-level contextual matching between the source text and the intermediate noise, a method for assessing factuality is presented. Three benchmark datasets are used to verify the effectiveness of the suggested factuality-guided summarizing model. According to experimental results [7].

This paper offers a comprehensive review of contemporary text summarization methodologies. The summarization methods according to many criteria, including technique, quantity of documents, document language, output summary type, summarization domain, and other factors. The study's conclusion looks at a number of issues and future directions for research on research text summaries that may be pertinent for future scholars in this field [8].

This research presents a unified approach to abstractive summarization of medical scientific literature by combining a Transformer model on a neural network graph. The methodology preserves document-level characteristics necessary for generating high-quality summaries while utilizing Latent Dirichlet Allocation (LDA) for topic modeling to uncover hidden subjects and global insights.

The system included a (HGNN) in addition to topic modeling, enabling the simultaneous updating of local and global data and captures sentence relationships through graph-based document representation. Ultimately system integrates a Transformer decoder, significantly enhancing the model's ability to provide precise and useful abstractive summaries [9] [14].

This research investigates performance inconsistencies and explores the impact of pre-training pseudo-summarization in low-resource environments. Using five datasets from different domains, we assessed five pre-trained abstractive summarization models, looking at how well they performed in terms of attention processes and extractive tendencies. Although extractive features are present in all models, some do not have enough predictive confidence to replicate longer text segments effectively, and their attention distributions deviate from the organic real data. The latter is the principal factor contributing to underperformance in the categories of fiction, news, and

scientific articles, while BART's superior initial attention alignment produces the highest benchmark scores across all few-shot scenarios.
Subsequent analysis reveals that BART's summarization capabilities stem from the interplay between the phrase permutation task and the pre-training dataset's characteristics. In light of these results, we present Pegasus-SP, an enhanced abstractive summarization model that has already been trained. that combines sentence permutation with pseudo-summarization. In resource-constrained settings, the new model surpasses its predecessors and exhibits greater adaptability. Furthermore, we illustrate, which can increase relative ROUGE on model-data pairs by up to 10% exhibiting the most significant distributed disparities [10].

This study presents a multi-encoder transformer-based method for Korean abstractive text summarization. In numerous (NLP) applications, the application of recently, for transfer learning, pre-trained language models (PLMs) have yielded impressive outcomes. For pre-training and subsequent tasks, transformer-based models, including Bidirectional Encoder Representations from Transformers (BERT), provide state-of-the-art performance, including abstractive text summarization. The selection of only one PLM at a time is necessary since text summarization models currently frequently use a single pre-trained model for each architecture [11].

Summarization generates a concise and clear overview that encapsulates the main concept of the original material. Summarization can be categorized into two types: abstractive and extractive. Extractive summarization identifies and selects pivotal statement from data to create a summary, while abstractive summarization rephrases the content using more sophisticated and human-like expressions by using story sentence or phrases. Creating a summary of a text requires significant time and financial resources for a human annotator, as it involves thoroughly reading the complete content and composing a concise summary. A sophisticated, automated methodology for text summarization is proposed, capable of minimizing effort and generating swift summaries by integrating extractive and abstractive methodologies [12].

This research aims to improve the quality of summary statements by implementing a reinforcement learning-based reward function for text summarizing. As improved ROUGE functionalities, we suggest ROUGE-SIM and ROUGE-WMD. ROUGE-SIM supports semantically linked concepts, unlike ROUGE-L. ROUGE-WMD is a metric that enhances semantic similarity as opposed to ROUGE-L. The semantic similarity of articles and summary text was assessed using the Word Mover's Distance (WMD) approach. Our model with the two proposed reward functions outperformed ROUGE-L on ROUGE-1, ROUGE-2, and ROUGE-L. Our two models, ROUGE-SIM and ROUGE-WMD, achieved ROUGE-L scores of 0.418 and 0.406 on the Gigaword dataset, respectively [13].

VATMAN (Video-Audio-Text Multimodal Abstractive Summarization) is an innovative method for producing Trimodal Hierarchical Multihead Attention is used in hierarchical multimodal summaries. VATMAN emphasizes visual, aural, and textual modalities through a hierarchical attention mechanism, in contrast to previous generative pre-trained language models [15].

PROPOSED SYSTEM

The procedure consists of two phases: document representation and sample phrase selection. A document is represented as a continuous vector utilizing pre-trained distributed vectors from phrase embedding models [15]. The significance score of the sentence is subsequently computed using ILP and PCA, and representative phrases for the summary are chosen. A collection of diverse pre-trained sentence representations was utilized to enhance model performance further. The initial stage is to amalgamate all the text from the articles. Subsequently, partition the content into discrete sentences. In the subsequent phase, we will establish a vector representation for each sentence. The similarities among sentence vectors are subsequently calculated and recorded in a matrix. The similarity matrix is further transformed into a graph, with sentences represented as vertices and similarity scores as edges, to compute sentence rank. The concluding summary consists of a predetermined quantity of highest-ranked sentences [16].

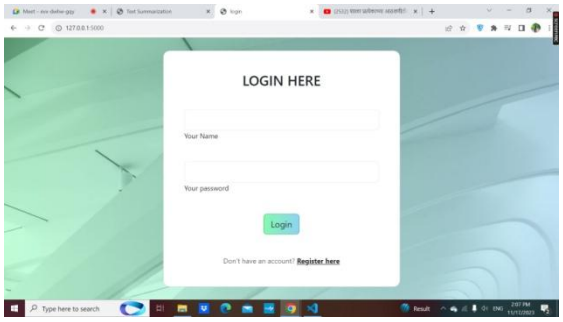
Algorithm: ERT - Bidirectional Encoder Representations from Transformers

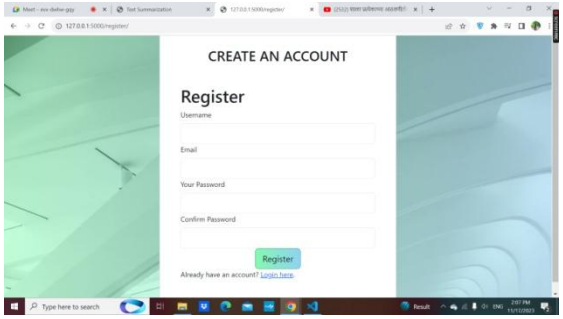
ERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that can understand the context of words in a sentence by looking both before and after a given word. This bidirectional approach allows BERT to grasp the nuanced meaning of words depending on their context.

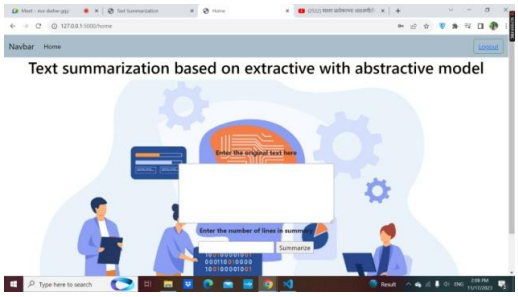
SciBERT is essentially a version of BERT tailored for the scientific domain, equipped with a deep understanding of scientific language and context. By using a custom vocabulary and training on scientific literature, SciBERT serves as

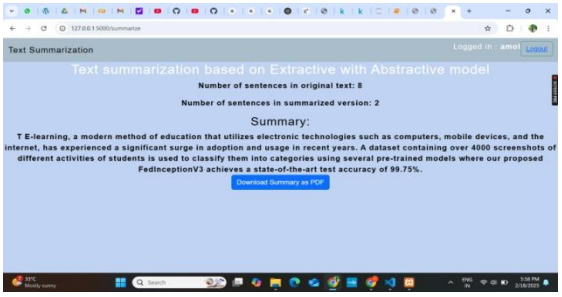
a foundation for tasks that involve understanding, analyzing, and summarizing research articles. Its integration with transformer-based models like Graph Transformer Networks (GTN) allows it to contribute significantly to complex tasks like generating abstractive summaries that capture the core ideas and relationships within scientific documents.

RESULTS AND DISCUSSION









CONCLUSION

A new method for abstractive summarization that tackles the trade-off between diversity and factual consistency is presented in this work:

- Iterative instruction during denoising ensures high factual consistency.
- The diffusion model creates diverse summaries, resulting in increased diversity.

REFERENCES

[1] Osman Büyük ," A Comprehensive Evaluation of Large Language Models for Turkish Abstractive Dialogue Summarization" , 2024

[2] N.Moratancha1 and S. Chitrakala3 , “A Novel Framework for Semantic Oriented Abstractive Text Summarization” , 2019
[3] Ayesha Ayuab Syed2 , Ford Lumban Gaol 1 , And Tokurao Matsuo2,3, (Member, IEEE) , “A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization” , 2021
[4] Mehatap Ulker , A. Bedri Ozer “Abstractive Summarization Model for Summarizing Scientific Article” , 2024
[5] Muahammad Awis, Rao Muhammd Nawab , “Abstractive Text Summarization for the Urdu Language: Data and Methods” , 2024
[6] Dilawari 2, Muhammad Usmana Ghani Khan2, “Asovs: Abstractive Summarizationof Video Sequences” , 2019
[7] Jeongwan Shin 1, Hyeyoung Park 1, (Member, IEEE),AND HYUN-JE SONG 2,(Member, IEEE) , 2024
[8] Divakara Yadav 2 , Rishabh Katna2 , Arun Kumar Yadav1 , And Jorge Morato 3 , “Feature Based Automatic Text Summarization Methods: A Comprehensive State-of-the-Art Survey” 2022
[9] Daniila Chernyshev 2 And Boris Dobrov1, , “Investigating the Pre-Training Bias in Low-Resource Abstractive Summarization” , 2024
[10] Youhyun Shin , “Multi-Encoder Transformer for KoreanAbstractive Text summarization” , 2023
[11] Aniqa Dilawari1 , Muhammad Usman Ghani Khan 3 “ Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space”, 2023
[12] Shrishail Patil, Diksha Shelke, Nilam Gavhane, Simran Sonawane , Vaishnavi Patankar , “Detection and Classification of Brain Tumor Using Convolutional Neural Network”, 6th Edition of International Conference on Communications and Cyber- Physical Engineering (ICCCE–2023).
[13] Sonawane, V.D., Mahajan, R.A., Patil, S.S., Bhandari, G.M., Shivale, N.M., Kulkarni, M.M., “Predicting Software Vulnerabilities with Advanced Computational Models”, Advances in Nonlinear Variational Inequalities, 2024.
[14] Shivale, N.M., Mahajan, R.A., Bhandari, G.M., Sonawane, V.D., Kulkarni, M.M., Patil, S.S., “Optimizing Blockchain Protocols with Algorithmic Game Theory”, Advances in Nonlinear Variational Inequalities, 2024.
[15] Patil, S.S., Mahajan, R.A., Sonawane, V.D., Shivale, N.M., Kulkarni, M.M., Bhandari, G.M., “Deep Learning for Automated Code Generation: Challenges and Opportunities”, Advances in Nonlinear Variational Inequalities, 2024.
[16] Kulkarni, M.M., Mahajan, R.A., Shivale, N.M., Patil, S.S., Bhandari, G.M., Sonawane, V.D., “Enhancing Social Network Analysis using Graph Neural Networks”, Advances in Nonlinear Variational Inequalities, 2024.