World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A survey on audio analysis: Text characterization and summarization

Maheshwar Reddy V [1], Deepika K [2], Adithya Surya Prakash K [2] and Sanathan M [2]

[1] *Associate Professor, Department of Computer Science (Artificial Intelligence and Machine Learning), ACE Engineering College, Hyderabad, Telangana, India.*
[2] *IV B. Tech students Department of Computer Science (Artificial Intelligence and Machine Learning), ACE Engineering College, Hyderabad, Telangana, India.*

World Journal of Advanced Research and Reviews, 2024, 21(03), 1596–1601

Publication history: Received on 30 January 2024; revised on 14 March 2024; accepted on 16 March 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.21.3.0789

Abstract

The integration of cutting-edge natural language processing (NLP) technology for smooth audio-to-text conversion and summarization is examined in this survey. Utilizing Facebook's BART model for succinct summaries and Google's Speech-to-Text API for precise transcription. The report highlights the value of sophisticated summarization models and precise transcription. It talks about how the system can be used in a variety of fields, such as podcast and video transcript generation, automated meeting transcription and summarization, content indexing and search, and more. In addition to addressing issues like context preservation and bias reduction, the survey assesses relevant research on text generation, LSTM networks, and summarization techniques. Overall, by incorporating state-of-the-art technology, this study advances the processing of audio content and eventually makes it easier to extract valuable information.

Keywords: Natural Language Processing (NLP); Generative Adversarial Networks (GANs); Text generation; Deep learning; Word embeddings; summarization methods; Automatic text summarization.

1. Introduction

In a time when a lot of audio content is available, it might be difficult to effectively extract insightful information. In order to overcome this difficulty, this project skillfully combines two potent natural language processing technologies: Facebook's BART (Bidirectional and Auto-Regressive Transformers) model and Google's Speech-to-Text API. The goal is to enable accurate conversion from audio to text and then automatically generate written summaries that are clear and short. Advanced technologies are required for efficient content processing due to the increasing amount of audio information, which includes podcasts, customer care calls, business meetings, and educational lectures. The project recognizes that extracting useful information from spoken speech requires precise transcription and efficient summarizing.

Advanced technologies are required for efficient content processing due to the increasing amount of audio information, which includes podcasts, customer care calls, business meetings, and educational lectures. The survey recognizes that extracting useful information from spoken speech requires precise transcription and efficient summarizing.

The incorporation of cutting-edge technologies from Facebook and Google highlights the dedication to utilizing the most recent development in natural language processing. The applications of this initiative are diverse and include everything from bettering corporate analytics and instructional materials to making content more searchable and accessible. The methodology, technical specifications, and possible uses of this integrated system will be covered in detail in the upcoming parts, with an emphasis on how revolutionary it may be in terms of how we analyze and draw conclusions from audio material.

* Corresponding author: K Deepika

2. Literature Review

In the era of burgeoning digital content, the demand for efficient audio processing solutions has intensified, driven by the necessity to extract meaningful insights from spoken discourse. This paper presents an integrated solution poised at the intersection of audio-to-text conversion and advanced natural language processing (NLP) techniques, aimed at seamless transcription and subsequent summarization of audio content. Leveraging the capabilities of Google's Speech-to-Text API and Facebook's Bidirectional and Auto-Regressive Transformers (BART) model, this system embodies a novel approach towards transforming spoken discourse into concise and coherent textual summaries.

The survey [1] covers various models, including recurrent neural networks (RNNs) and transformer-based architectures, discussing their applications in machine translation, dialogue systems, summarization, and creative writing. It explores evaluation methodologies such as perplexity and BLEU score, while also addressing challenges like context preservation and bias mitigation. Through case studies and experimental insights, the survey offers valuable perspectives on the current state and future directions of text generation in deep learning." The survey evaluates these models using metrics such as perplexity and BLEU score and addresses challenges like context preservation and bias mitigation. Through insightful case studies and experimental insights, the survey provides valuable perspectives on the current state and future directions of text generation in the realm of deep learning."

2.1. Working Principle

The working principle incorporates multiple deep learning text generation models, such as transformer-based models like BERT and GPT, generative adversarial networks (GANs), and recurrent neural networks (RNNs). These models create text sequences from input data by utilizing deep learning techniques. Evaluation criteria that can be used to evaluate these models' performance include BLEU score and perplexity. Furthermore, it's possible that these technologies' practical uses—like summarization, dialogue systems, and machine translation—are investigated.

The survey [2] comprehensively investigates the application of GANs in generating textual content, highlighting their potential across various domains such as natural language generation, dialogue systems, and creative writing. By analyzing the underlying principles and architectures of GANs in text generation, the survey elucidates the advantages and challenges associated with this approach. Evaluation methodologies specific to GAN-based text generation, along with pertinent performance metrics, are discussed to assess the quality and coherence of generated text. Furthermore, the survey addresses notable advancements, emerging trends, and future research directions in leveraging GANs for text generation, underscoring their significance in advancing the field of natural language processing." The survey meticulously examines the utilization of GANs across a spectrum of text generation tasks, including language modeling, dialogue generation, and story generation. By dissecting the architecture and training procedures of GANs in text generation, the survey elucidates the challenges and opportunities inherent in this approach.

2.2. Working Principle

The generative adversarial networks (GANs) are used in text generation as part of the working concept. Deep learning models known as GANs are made up of a discriminator and a generator that are trained against each other to produce realistic text sequences. The survey may cover a range of architectures and training methods for GAN-based text creation in addition to assessment measures to gauge the caliber of the text that is produced. GANs' practical uses in text generation tasks including dialogue systems, language modeling, and creative writing are probably going to be investigated.

The paper [3] delves into the architecture and functionality of LSTM networks, emphasizing their ability to capture sequential dependencies in text data. Through an examination of context-based text generation techniques, the author showcases the efficacy of LSTM networks in generating coherent and contextually relevant textual content. The paper discusses various approaches for incorporating contextual information into LSTM-based text generation models, highlighting their significance in tasks such as machine translation, dialogue systems, and summarization. Furthermore, the author presents experimental results and performance evaluations to demonstrate the effectiveness of LSTM networks in context-based text generation, paving the way for advancements in natural language processing and artificial intelligence research." Through a detailed examination of context-based text generation techniques, the author demonstrates the effectiveness of LSTM networks in generating coherent and contextually relevant textual output. The article discusses strategies for integrating contextual cues into LSTM-based models, showcasing their applicability in diverse domains such as machine translation, dialogue generation, and summarization.

2.3. Working Principle

The journal delves into the use of Long Short-Term Memory (LSTM) networks for context-based text production. Recurrent neural networks (RNNs) with long-range dependency capture capabilities, such as LSTM networks, are well-suited for tasks like text production and language modeling. Along with training procedures and evaluation measures to gauge the caliber of generated text, the article may provide methods for integrating contextual data into LSTM-based text generation models. It may also be investigated how LSTM networks might be used in the real world for tasks like summarization, dialogue systems, and machine translation.

The survey [4] provides an overview of the progress made in automatic text summarization, detailing the underlying processes and methodologies employed in generating concise summaries from large volumes of text. Mridha and Lima explore various approaches to automatic text summarization, including extraction-based, abstraction-based, and hybrid methods, elucidating their strengths, limitations, and applications. Additionally, the survey discusses the challenges inherent in automatic text summarization, such as maintaining coherence, preserving key information, and handling domain-specific content. By synthesizing research findings and highlighting emerging trends, the authors provide valuable insights into the current state and future directions of automatic text summarization, offering guidance for researchers and practitioners in the field of natural language processing." Moreover, the survey delves into the inherent challenges of automatic text summarization, such as content selection, coherence maintenance, and scalability issues. By synthesizing existing research and identifying future research directions, the authors offer valuable insights into the advancements and challenges in automatic text summarization, serving as a comprehensive resource for researchers and practitioners in natural language processing."

2.4. Working Principle

The article gives a general summary of artificial text summarizing methods, including developments, underlying mechanisms, and difficulties encountered. It could go over different approaches to automatic text summarization, like hybrid, abstraction-based, and extraction-based techniques, and how to use technologies like deep learning models, machine learning algorithms, and natural language processing (NLP) to accomplish them. The evaluation methods and criteria used to gauge the caliber of summaries produced by various approaches may also be covered in the study. It might also go over practical uses and developments in automatic text summarization, providing information on potential future paths for the field's study.

The paper [5] focuses on bridging the gap between audio and visual modalities in text generation tasks, proposing the TAVT framework as a mean to facilitate transfer learning. Through meticulous experimentation and analysis, the authors demonstrate the effectiveness of the TAVT framework in generating coherent textual output from audio-visual inputs. By leveraging transfer learning techniques, the framework exhibits promising results in tasks such as speech recognition, image captioning, and multi-modal translation. The paper also discusses potential applications and future research directions in the realm of transferable audio-visual text generation, highlighting its significance in advancing multi-modal understanding and communication."

2.5. Working Principle

The creation of the TAVT (Transferable Audio-Visual Text Generation) framework, which aims to produce textual output from audio-visual inputs, is probably explored in this publication. In order to support knowledge transfer across domains, it might cover the integration of auditory and visual modalities in text creation tasks through the use of transfer learning techniques. The TAVT framework's architecture, training protocols, and applications in multi-modal translation, picture captioning, and speech recognition might all be covered in the article. It might also go over performance indicators and possible uses for transferable audio-visual text generation, emphasizing its importance in enhancing multimodal comprehension and communication.

The paper [6] introduces an innovative method for summarizing text content using voice input, aiming to enhance accessibility and user experience. Through meticulous experimentation and analysis, the authors demonstrate the effectiveness of their approach in generating concise summaries from spoken input. The proposed method holds promise for applications such as automated transcription and summarization of audio content, facilitating efficient information extraction from spoken discourse. By bridging the gap between voice-based input and text summarization, the paper contributes to the advancement of natural language processing techniques, offering valuable insights into the potential of voice-based text summarization systems." Through rigorous experimentation and analysis, the effectiveness of the proposed method in generating concise and coherent summaries from spoken input is demonstrated. The paper highlights the potential applications of voice-based text summarization, including automated transcription, content summarization in audio recordings, and accessibility services for individuals with visual

impairments. By integrating voice-based interaction with text summarization techniques, the paper contributes to advancing natural language processing technologies, offering new avenues for efficient information retrieval from spoken content."

2.6. Working Principle

With a focus on obtaining succinct summaries from spoken input, the journal probably offers a novel method for voice-based text summary. It might go through methods for turning spoken input into text, such voice-to-text APIs or speech recognition algorithms. The report might also discuss how to summarize the text that has been transcribed, either using extractive or abstractive summarization, which are NLP techniques. The method's ability to produce meaningful and cogent summaries from spoken input may be assessed, with possible uses in automated transcription and audio content summarization.

3. Comparative Study

Table 1 Comparison of various Algorithms

Year	Algorithm	Key Developments	Pros	Cons
2016	Rule-based Systems [4]	Integration with linguistic rules and heuristics	- Transparent and interpretable text generation	-Limited Scalability, requires manual rule crafting
2017	Hidden Markov Models	Application in speech recognition and text generation	- Widely used for text modelling and generation	-Limited capability to captu re complex dependencies
2018	Recurrent Neural Networks [1]	Mechanisms for improved context modelling	-Handles sequential data effectively	- Gradient vanishing, Slow training convergence
2019	Multi-Model Integration [5]	Integration of audio-visual data for text generation	-Enhances text generation with audio-visual cues	-Increased complexity, potential data synchronization issues
2020	Generative Adversarial Networks [2]	Enhanced training techniques for stability	-Generates diverse and realistic text samples	-Training instability, Mode collapse
2021	LSTM Networks [3]	Improved architectures for sequential modelling	-captures sequential dependencies in text data	- Limited holding of long term dependencies
2022	Facebook's BART Model [6]	Transformers-based architecture for summarization	- Coherent and concise text summarization	-Training complexity,Resource-intensive computation
2023	Google's Speech-to-Text API	Advanced Deep learning models for transcription	- High accuracy in audio transcription	-Limited customization options,requires internet connection

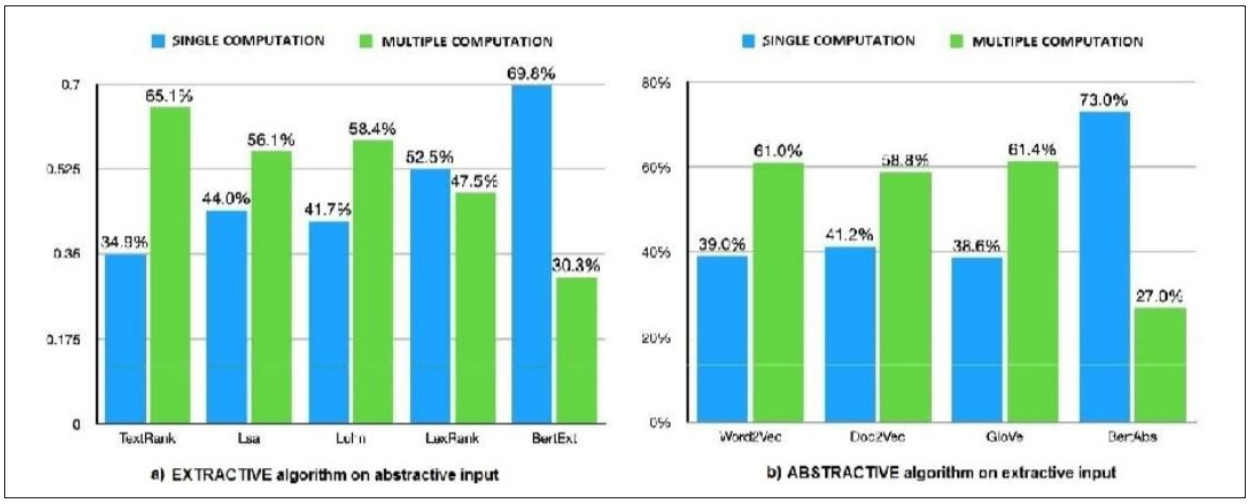

Figure 1 Extractive algorithm on abstractive input and Abstractive algorithm on extractive input

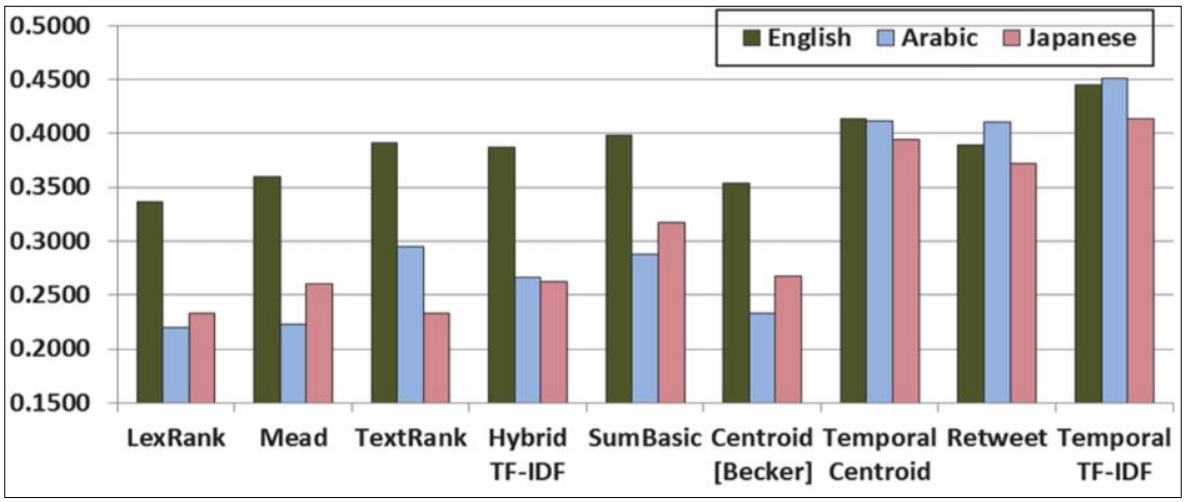

Figure 2 Various methods on different languages

4. Conclusion

We looked at a number of journal articles about text summarizing and audio to text generation. We were aware of the benefits and drawbacks of employing different approaches. Recognizing the shortcomings, we will use transformers in our suggested techniques to attempt to address them. Our method provides a smooth way to handle massive amounts of audio data processing by combining cutting-edge technology such as Facebook's BART Model and Google's Speech-to-Text API. This journal article offers a thorough examination of approaches, stressing both their advantages and disadvantages, making it an invaluable tool for natural language processing scholars and practitioners. This project will be utilized in business meetings, podcasts, education, and the media.

Compliance with ethical standards

Acknowledgments

We would like to thank our guide Dr.V.Maheshwar Reddy for his support and guidance, Associate Professor(Artificial Intelligence and Machine Learning), and Mr,Shashank Tiwari, Assistant Professor, Project Coordinator and we profoundly thank Dr.Kavitha Soppari, Head of the Department of CSE(Artificial Intelligence and Machine Learning)for her guidance and continuous support.

Disclosure of conflict of interest

The authors have no conflicts of interest to declare. All co-authors have seen and agreed with the contents of the manuscript and there is no financial interest to report.

References

[1] Touseef Iqbal , Shaima Qureshi.The Survey: Text generation models in deep learning. Journal of King Saud University – Computer and Information Sciences 34 (2022) 2515–2528.

[2] Gustavo H. de Rosa , Joao P. Papa: A survey on text generation using generative adversarial networks. Pattern Recognition 119 (2021) 108098.

[3] Sivasurya Santhanam: Context based Text-Generation Using LSTM Networks, Institute for Software Technology ,German Aerospace Center(DLR),2020.

[4] M.F. Mridha: A Survey of Automatic Text Summarization: Progress, Process and Challenges,2021.

[5] Wang Lin,Tao Jin,Ye Wang: TAVT:Towards Transferable Audio-Visual Text Generation, Zhejiang University,2022.

[6] Pratima Mohan Thorat , Prof. Dr. M. S. Bewoor: A Novel Approach for Voice Based Text Summarizer,2022