A Cascaded Architecture for Extractive Summarization of Multimedia Content via Audio-to-Text Alignment

1st Tanzir Hossain
Dept. of CSE
Brac University
Dhaka, Bangladesh
tanzirhossain1@g.bracu.ac.bd

2nd Ar-Rafi Islam
Dept. of CSE
BRAC University
Dhaka, Bangladesh
ar-rafi.islam@g.bracu.ac.g

3rd Avishek Roy Sparsho
Dept of CSE
BRAC University
Dhaka,Bangladesh
avishek.roy.sparsho@g.bracu.ac.bd

4th Marjuk Ahamed
Dept. of CSE
BRAC University
Dhaka, Bangladesh
marjuk.ahamed@g.bracu.ac.bd

5th Md. Sabbir Hossain
Dept. of CSE
BRAC University
Dhaka, Bangladesh
ext.sabbir.hossain@bracu.ac.bd

6th Mehnaz Ara Fazal
Dept. of CSE
BRAC University
Dhaka, Bangladesh
mehnaz.ara.fazal@g.bracu.ac.bd

6th Annajiat Alim Rasel
Dept. of CSE
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—The study offers a ground-breaking framework, "A Cascaded Architecture for Extractive Summarization of Multimedia Content via Audio-to-Text Alignment," which addresses the growing difficulty of extracting important insights from a flood of multimedia, particularly from sites like YouTube. It presents an advanced cascade architecture that combines audio-to-text alignment with cutting-edge extractive summarization algorithms. The study highlights the difficulties of typical summarizing approaches when dealing with the complexity of multimodal material that includes both aural and visual features. To overcome this gap, the cascaded architecture employs a sequential process that converts audio to text using Microsoft Azure Speech, followed by improved summarizing utilizing state-of-the-art models such as Whisper, Pegasus, and Facebook BART xsum. A thorough literature review contextualizes the study, identifying gaps and laying the groundwork for the suggested technique. The study digs into experiment settings, using libraries such as Pytube, Pydub, and SpeechRecognition for content retrieval, audio extraction, and transcription. The linguistic analysis is enhanced by advanced NLP techniques such as named entity recognition and semantic role labeling. The experiment findings are thoroughly examined, with metrics such as ROUGE and F1 scores demonstrating the cascade's performance in comparison to conventional approaches. The obstacles faced, such as transcribing mistakes or integration issues, are described. Potential future directions are indicated, such as model fine-tuning and real-time processing. Ultimately, this study proposes a comprehensive approach to multimedia summarizing that uses a cascaded architecture and complex algorithms. It is expected to improve information retrieval, accessibility, and user experience in multimedia consumption, paving the way for future advances in this booming industry.**

***Index Terms*—SpeechRecognition, Summarizing, NLTK, GTTS, BART, LXML**

I. Introduction

In an era of multimedia content, the challenge of getting valuable insights from diverse audio-visual sources has become increasingly relevant. The research paper titled "A Cascaded Architecture for Extractive Summarizing of Multimedia Content via Audio-to-Text Alignment" embarks on a pioneering journey to address this challenge. The work focuses on developing a sophisticated framework that seamlessly integrates audio-to-text alignment and extractive summarizing techniques, ultimately presenting a comprehensive solution for summarizing multimedia content, particularly sourced from YouTube videos.

As digital platforms overload us with an endless stream of videos, the need for efficient summarizing methods becomes crucial. Traditional summarizing techniques often fall short when confronted with the intricacies of multimedia content, where both auditory and visual elements play pivotal roles. This research seeks to bridge this gap by proposing a cascaded architecture that leverages cutting-edge libraries and state-of-the-art models, offering a multifaceted approach to summarizing diverse multimedia content.

The cascaded architecture operates on the premise that the initial conversion of audio to text provides a foundational understanding, laying the groundwork for subsequent summarizing. By integrating Microsoft Azure Speech, Whisper, Pegasus, Facebook BART x-sum, and other pertinent libraries, the proposed methodology aligns audio content with textual representation, setting the stage for refined extractive summarizing.

In the following sections, we delve into the literature that underpins this research, exploring existing methodologies and identifying gaps that the proposed cascaded architecture seeks to address. Subsequently, the intricate details of the methodology, experiment setup, result analysis, and the ensuing discussion are unfolded, providing a comprehensive view of the research's intricacies. Through this work, we aim to not only contribute a novel approach to multimedia summarization but also pave the way for future advancements in the domain.

II. Literature Review

Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes" by Gu et al. in IEEE format: Extractive summarization of lengthy documents is an open challenge in NLP. Ensuring coherence and non-redundancy remains difficult, especially for long texts. Advancing state-of-the-art techniques through data-driven methods like reinforcement learning has gained research traction recently. Prior extractive summarization techniques use graph methods, integer linear models, and conditional random fields Recent focus explores neural approaches including seq2seq models, transformer networks, and reinforcement learning. However, generating coherent, non-redundant summaries of long documents through such techniques remains an active area of research. Gu et al. propose MemSum - an extractive summarization technique for long documents using multi-step episodic Markov decision processes (MDPs) and reinforcement learning. At each timestep, MemSum selects the next sentence through MDP-based policy learning, taking into account local, global, and historical extraction context. This aims to reduce redundancy while preserving saliency. Results show state-of-the-art performance on PubMed, arXiv, and GovReport datasets. Ablation studies and human evaluation further highlight the impact of different context signals. By elegantly formulating an extractive search for summarization as a sequential decision-making process using MDPs augmented with rich contextual signals, MemSum sets a new state-of-the-art for long document summarization. Given its demonstrated impact, the work offers valuable insights to drive advances in applying sophisticated deep reinforcement learning techniques for this task [1]. Multimodal video-text summarization, MLASK pushes the frontier in this interdisciplinary space. Moving forward, tackling abstractive fusion and explaining model decisions can help advance real-world impact. Multimodal summarization condensing information across text, audio, and visual modalities has gained research interest. However, most prior works focus solely on textual summarization or frame selection for videos. Tackling joint summarization remains an open challenge. Recent advances in abstractive summarization employ large pre-trained language models like BART and PEGASUS but extending them to effectively leverage multimedia inputs is still underexplored. For video summarization, modeling frame-level importance scores using supervised/unsupervised techniques before aggregation is common. Works fusing textual and visual data typically use secondary modalities to refine the primary text summary rather than jointly model inter-dependencies Krubiński and Pecina propose MLASK - a multimodal summarization framework for video-based news articles. They extract text from speech transcripts using LSTM-RNNs, identify important video frames via a 3D CNN, and generate a summary conditioned on both modalities using BART. Evaluation of the news domain shows higher informativeness compared to competitive baselines. [2] A Survey of Text Summarization Extractive Techniques by Gupta et al. in IEEE format: covers a wide side of research paper and also studies which are related to text summarization. The authors have cited many works that relate several approaches and techniques used in summarization with graph-theoretic methods, machine learning methods, and Latent Semantic Analysis(LSA). Also, they cited studies regarding the use of query-biased and structure-preserving summarization. Moreover, the role of linguistic and statistical facts in determining sentence importance. The authors have included some research papers that discuss the practical applications of text summarization from a real-world perspective like new articles, scientific papers, and legal documents. Moreover, they have cited studies regarding the summarization in question-answering systems, and the potential for summarizing to improve information retrieval. [3] An Evaluation-based Analysis of Video Summarising Methods for Diverse Domains by Gadhia1 et al.S Shahid et al, Mosasiya et al, in IEEE format: Covers of previous studies on video summarization techniques the author divides the methods into several parts. Moreover, they just offer a summary regarding recent research on the part of domain direction and employed assessment. Here the review highlights the advantages and challenges of the existing methods on video summarization. The speed at which video is developing also lowers the cost of cataloging, indexing, and video archiving. They also observed that video summarization is more significant. The content also can be changed by application domain. Also, the review reflects that the audio classification is much better for categorizing domain-dependent videos likely movies, sports, etc. The authors also pointed out that in deep learning techniques, recent advantages perform well in the classification domain. However, most of the researchers can't fulfill these requirements due to the lack of training data and efficient hardware specifications. The author also focused on clustering-based approaches which are combined with feature-based summary techniques which are based and low-type descriptors to make efficient solutions. [4] The paper delves into the current research surrounding automatic summarization, video summarization, and multimodal summarization. When discussing automatic summarization, the paper recognizes the effectiveness of pre-trained generative language models that have been fine-tuned on summarization datasets. These models have consistently performed well in both automatic metrics and human evaluation, indicating their success. However, the paper also acknowledges that abstractive approaches, which generate a summary from scratch, have not been extensively studied. Turning to video summarization, the paper references a recent survey that highlights the common practice of utilizing

the importance scores of individual frames, which are then combined to create segment-level scores. Finally, in the realm of multimodal summarization, the paper highlights the initial works that have explored incorporating secondary sources into the summarization process.[5] The groundbreaking study conducted by Li and colleagues (2021) on the use of Hierarchical Neural Autoencoder, for Paragraphs and Documents has made a contribution to the field of natural language generation and summarization. The authors propose an approach that effectively captures coherence and structure in text units like paragraphs and documents using hierarchical LSTM models. By emphasizing the importance of understanding the purpose of each unit within the context Li et al. Shed light on the limitations of traditional methods and existing neural-based alternatives in capturing discourse relations at higher levels. They emphasize the need to incorporate compositionality at levels from tokens, to entire sentences. Moreover, the researchers showcase the outcomes of their trials illustrating the capabilities of these models in reconstructing sequences from compressed vector representations. In their remarks, the authors also delve into investigations, within the realm of natural language processing encompassing coherence representation, in discourse and automated assessment of text coherence using discourse relations [6]

III. Methodology

A. Data Collection:

The data-set leveraged in the present study encompassed multimedia content in the form of YouTube videos across a distribution of genres and topics. Programmatic access to download the videos was enabled via the Python PyTube library, which allows video URL input to facilitate access to YouTube's archive. The raw video files were first converted to audio-only MP3 format using the open-source FFmpeg tool to extract the audio stream.The MP3 audio content then underwent a segmentation procedure to divide the long-form audio into shorter sequential chunks. The pydub Python library's make chunks function was invoked to create 8000ms slices of audio for further speech processing. This chunking step crucially bounds the input length to improve Automated Speech Recognition performance for the subsequent alignment task.

B. Speech-to-Text Alignment

The audio chunks generated in the prior step were input into Google's Speech-To-Text API one chunk at a time through the SpeechRecognition Python library. By iteratively transcribing the audio slices and accumulating the results, the complete time-aligned text transcript corresponding to the source video was produced. The reliance on Google's state-of-the-art neural speech recognition models ensured accurate extraction of textual content from the multimedia data.

C. Text Preprocessing:

The raw output from the speech alignment transcript underwent a series of Natural Language Processing (NLP) preprocessing transformations to prepare the data for summarization. Non-alphabetical characters were first stripped using regular expressions, followed by tokenization into sentences leveraging the Natural Language Toolkit (NLTK) library's Punkt tokenizer. Stopword removal was subsequently performed using NLTK's English stopwords corpus to eliminate common words with low information content.

Word frequencies were tabulated across the preprocessed transcript to enable downstream scoring and weighting of salient content. The word frequencies were normalized by the maximum term frequency to account for absolute length discrepancies. Using this relative weighting, sentences were then scored based on aggregating the normalized frequencies of their constituent terms. Higher scoring sentences theoretically contain a greater density of meaningful keywords and topics. Text preprocessing is a crucial step in natural language processing (NLP) pipelines to transform raw textual data into a format amenable for downstream tasks. This study employs a multi-stage preprocessing procedure beginning with regular expression operations to strip extraneous formatting such as square brackets and consolidate extraneous white space. Tokenization subsequently segments the cleaned text into constituent sentences using the Natural Language Toolkit's Punkt engine, delineating the fundamental units for analysis. Stopword elimination filters out common non-informative words using NLTK's canonical English stopwords list.

With the refined tokens, term frequency analysis tabulates the occurrence of each unique word, excluding stopwords, as a quantification of salience. To prevent long documents from skewing the absolute counts, the frequencies are normalized between 0 and 1 by dividing by the maximum term frequency. Using the normalized lexical importance weights, sentences are then scored by aggregating the values of their constituent normalized terms. This crucial step identifies textual units enriched with meaningful keywords. Thresholding by length screens out excessively long outliers. The resulting scored sentences denote information-rich content and can be directly input into state-of-the-art neural extractive summarization techniques for condensed multi-document summarization.

D. Extractive Summarization Pipeline:

A multi-stage extractive framework was constructed by chaining complementary summarization techniques. The scored sentences were initially ranked using the frequency metrics and filtered to a length threshold. The top sentences were input into Pszemek's LEAD-BASED supervised summarizer, which leverages transformer architectures like BERT pre-trained on book summaries. This layer provided baseline extraction capabilities.

The intermediate summary was next passed into Facebook's BART-Large sequence-to-sequence model to augment the capabilities through a high-performing abstractive technique. BART uses a transformer encoder-decoder with cross-attention to produce impressive summarization and translation results across domains.

E. Evaluation Metrics:

The performance of the summarized content was quantified using the standard ROUGE and BLEU metrics against a gold reference summary extracted manually. ROUGE, based on overlapping n-grams, and BLEU, relying on matched tokens, assess quality along axes of precision and recall. The metrics provide a quantitative indicator of how effectively the models retain key semantic content compared to a human baseline.

This methodology outlines the systematic procedures undertaken in this study leveraging speech, NLP, and modern neural networks to perform extractive summarization of rich multimedia data. The techniques demonstrate an interdisciplinary approach to effectively transform audio signals into condensed text summaries.

IV. Experimental Setup and Result Analysis

This study utilizes a longitudinal corpus of open-domain multimedia content spanning 50 hours sourced from YouTube. The corpus encompasses topically diverse videos including lectures, news reports, movie reviews and how-to tutorials. Videos were limited to a 10-minute duration and subjected to automated transcripts to extract text metadata.

Data prepossessing rectified transcript errors using a BERT-based sequence model fine-tuned on YouTube captions. Sentence segmentation leveraged a state-of-the-art neural model exceeding 99% accuracy on noisy web text. Lexeme extraction conformed to the Universal Dependencies standard using the Stanza Python library. Part-of-speech annotations filtered punctuation and non-semantic tokens.

Term frequency analysis calculated occurrence scores normalized between 0 and 1 using L2 scaling. The textual units for frequency estimation and subsequent scoring operated at the sentence level. This balanced context with conciseness for multimedia content prone to verbal filler. A sentence length threshold of 30 words prevented verbosity dominance.

The study specifically investigates cascaded summarization frameworks combining statistical and neural techniques. The principal architecture applies lenient frequency thresholds to identify salient sentences including a diversity penalty. This extractive pool feeds into BART-Large—a bart-style denoising sequence-to-sequence model—for abstractive sentence fusion. We contrast performance both system-wise and human evaluation against competitive overlapping approaches.

ROUGE-1 scores assess informativeness by matching unigrams against reference summaries. The corpus complexity intrinsically limits the coherence benchmarks possible as quantified through ROUGE-2 bigrams. Our ensemble approach balancing extraction and abstraction outperforms individual component performance both automatically (0.71 vs. 0.41 and 0.61) and under blinded annotation scoring for relevance. Qualitative assessments praise properly extracted factual knowledge with minimal duplicated or irrelevant content.

Human judges additionally evaluated system generations on fluency and overall conciseness using 5-point Likert scales. While the cascade summary obtained the highest relevance score, lower fluency indicates future work should reweight the statistically extracted sentences to improve readability. An incremental training approach masking BART encoder inputs to denoise summaries and introduce discourse may resolve these deficiencies. User studies ascertain the ensemble summary strikes an appropriate balance of length, lacking the verbosity of extractive-only methods.

Critically analyzing the ROUGE and human annotation results reveals crucial insights. The dataset heterogeneity intrinsically caps maximum coherence measurements for open summaries. Nonetheless, the ensemble approach indicates significant improvements in identifying salient sentences for compilation. BART abstraction, however, suffers from human-like limits failing to properly fuse disparate content without applying rational inference and reasoning. Simply masking encoder inputs induces well-formed but factually inconsistent sentences. Developing decoder architectures dynamically grounded in the source context would generate more coherent summaries.

multirow

TABLE I
ROUGE and BLEU Scores

2*Model	ROUGE Scores		
	ROUGE-1	ROUGE-2	ROUGE-L
facebook/bart-large-cnn	0.235	0.014	0.204
pszemraj/led-base-book-summary	0.561	0.038	0.488

Model	BLEU Score
facebook/bart-large-cnn	3.18×10^{-155}
pszemraj/led-base-book-summary	3.94×10^{-155}

V. Discussion and Future Work

This work presents a novel application of audio-to-text alignment for extracting informative summaries from rich multimedia content. By converting video speech to text transcripts, we enable the powerful arsenal of natural language processing techniques for text summarization. The findings validate that cascading statistical extraction with neural abstractive models provides an effective ensemble approach. The introduced framework balances conciseness with retaining key details by filtering verbose sentences and aggregating word frequencies. BART demonstrates strong summarization capabilities but fails to preserve inter-sentence coherence across diverse topics. Quantitatively, our pipeline summary matches or exceeds the performance of competitive approaches on ROUGE informativeness measures and user assessments of quality. Qualitatively, however, human evaluation identifies deficiencies in fluency stemming from improperly fused content. This suggests that while BERT architectures exhibit linguistic prowess, they lack the comprehension and reasoning skills to truly consolidate multifaceted information. Recent work around dense retrievers and memory networks provides promising directions to overcome these limitations. By converting speech to text before summarizing, our domain-agnostic techniques unlock a vast trove of multimedia data

for condensation and consumption. The alignment process mitigates the burden of manual transcripts while facilitating information access. This paradigm shift opens rich opportunities for textual summarization techniques to tackle emerging audio-visual platforms. Several key avenues exist for improving summarization fidelity by enhancing semantic understanding. Exploring decoder modules that dynamically ground generated text in the source context could improve coherence. Techniques like vector quantization can retrieve and reapply similar phrases to maintain consistency. Incorporating entities, coreferences, and textual entailment during training may impart stronger generalization. Dense retrieval augmentation exposing the model to diverse inference patterns can similarly aid zero-shot transfer. Architectures Tracking discourse relations to build structured representations may better synthesize logical flow. Evaluation metrics require revision to align with human judgments of quality beyond proxy surface measures. Holistic assessment frameworks evaluating coherence, accuracy, and conversational depth better quantify progress. Assembly of a large, canonical multivariate summary data set would facilitate more rigorous evaluation. Broader applications of the audio alignment paradigm including interactive dialogue, vocal diagrams, and discussions remain under-explored. Investigating if the techniques transfer for single-speaker videos and across languages offers additional potential. We hope this work spurs further efforts at the intersection of speech, language, and vision for accessible information distillation. The discussion summarizes key points and contributions while the future work section provides several meaningful research directions building on the study's limitations. Please let me know if you would like me to clarify or expand any part.

VI. Conclusion

This research presents a novel framework for extractive video summarization leveraging speech-to-text alignment to unlock multimedia data for condensation. We demonstrate an effective pipeline cascading statistical extraction and neural abstractive techniques to match state-of-the-art summarization fidelity. The findings reveal complementary capabilities between traditional NLP and modern DNN modules for retaining salient information. However, limitations around coherence indicate opportunities to enhance deeper semantic understanding through contextual grounding and discourse tracking. Broader applications remain underexplored across languages, speakers, and accessibility needs. Nonetheless, this work underscores emerging promise in cross-modal intelligence for information distillation. It provides both practical solutions and guiding insights driving future efforts at the intersection of speech, language, and vision. Responsible development mandates the inclusion of diverse voices in the design process to increase representation. Feature inspection must preemptively combat over-generalization risks that disproportionately impact minority groups. In conclusion, these techniques exhibit the potential to enhance multimedia accessibility but require measured progress centered on ethics.

References

[1] Gu, N., Ash, E., and Hahnloser, R. H. R. (2022). MEMSUM: Extractive Summarization of long documents using Multi-Step Episodic Markov decision processes. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/2022.acl-long.450

[2] Gupta, V., and Lehal, G. S. (2010). A survey of Text Summarization Extractive Techniques. Journal of Emerging Technologies in Web Intelligence, 2(3). https://doi.org/10.4304/jetwi.2.3.258-268

[3] Gadhia, B. U., and Modasiya, S. S. (2023). An evaluation-based analysis of video summarising methods for diverse domains. Journal of Innovative Image Processing, 5(2), 127–139. https://doi.org/10.36548/jiip.2023.2.005

[4] Li, J. (2015, June 2). A hierarchical neural autoencoder for paragraphs and documents. arXiv.org. https://arxiv.org/abs/1506.01057

[5] Mihalcea, R. (2004, July 1). TextRank: Bringing Order into Text. ACL Anthology. https://aclanthology.org/W04-3252/