IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154229

Multilingual Summarization of Youtube Video using NLP

Ms.KaviPriya M.Tech[1], Bharath S[2], Jayaganesh S[3], Kishor S,Mithun S[4]

Assistant Professor, Department of AI&DS, Dhanalakshmi Srinivasan Engineering College,Perambalur, TamilNadu, India[1]

B TECH, Department of AI&DS, Dhanalakshmi Srinivasan Engineering College,Perambalur, TamilNadu, India[2-4]

Abstract: With the exponential growth of digital video content on platforms like YouTube, users face significant "information overload," particularly when accessing content in foreign languages. This project proposes an automated, **Multilingual YouTube Video Summarization** system designed to bridge the gap between high-volume video data and efficient information consumption.

The system implements a multi-stage pipeline beginning with a **Transformer-based Automatic Speech Recognition (ASR)** module to transcribe audio with high robustness to noise and linguistic variations. Following transcription, the text undergoes a rigorous preprocessing phase—including tokenization and stop-word removal—before being converted into high-dimensional vector representations using **Sentence-BERT (SBERT)**. This semantic embedding layer ensures that the core meaning of the video is preserved regardless of the source language.

The heart of the project is the **Salgueiro Framework for Temporal Semantic Mapping**, which facilitates **Abstractive Synthesis**. Unlike traditional extractive methods that merely copy sentences, our system generates a new, coherent narrative that maintains the chronological integrity of the original video.

Keywords—Multilingual NLP, YouTube Summarization, Text Summarization, Machine Translation, Transformer Models, Speech-to-Text.

I. INTRODUCTION

A. Growth of Online Video Content

YouTube has become one of the largest sources of digital information, with videos covering education, entertainment, news, and more. The increasing volume of content makes it difficult for users to efficiently find and consume relevant information. As a result, there is a growing need for automated tools that can condense video content into shorter, meaningful summaries.

B. Need for Multilingual Summarization

YouTube videos are created in many different languages, but viewers may not always understand the original language of the content. Multilingual summarization helps bridge this gap by translating and summarizing videos into the user's preferred language. This improves accessibility, enhances knowledge sharing, and supports a global audience.

C. Role of Natural Language Processing (NLP)

NLP techniques are essential for building an automated summarization system. These techniques include speech-to-text conversion to extract transcripts, language detection, machine translation, and text summarization. Advanced models such as transformer-based architectures enable the generation of accurate and coherent summaries across multiple languages, making the system efficient and scalable.

D. Need for Cross-Lingual NLP Systems

In a globally connected digital environment, users increasingly consume content produced in different languages. However, language differences often act as a barrier to accessing and understanding information. Cross-lingual systems powered by Natural Language Processing aim to bridge this gap by enabling machines to process, translate, and summarize text across multiple languages. Advanced models such as BERT and multilingual transformer architectures are capable of capturing semantic meaning beyond language boundaries. These systems can take a transcript in one language, translate it into another, and generate a coherent summary while preserving the original context. This capability

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154229

is particularly important for platforms like YouTube, where content is created by users worldwide. By integrating cross-lingual NLP techniques, the proposed system enhances accessibility, allowing users to understand and benefit from video content regardless of the language in which it was originally produced.

II. METHODOLOGY

The proposed system for multilingual summarization of videos from YouTube follows a structured pipeline consisting of multiple stages, integrating techniques from Natural Language Processing and machine translation.

A. Video Input and Data Collection

The system begins by accepting a YouTube video URL or ID as input. Using the YouTube Data API, metadata such as title, description, and captions are retrieved. If captions are unavailable, automatic speech recognition (ASR) techniques are employed to generate transcripts from the audio stream.

B. Transcript Preprocessing

The extracted transcript often contains noise such as filler words, timestamps, and irrelevant symbols. Preprocessing steps include text cleaning, tokenization, stop-word removal, and sentence segmentation. These steps ensure that the input is suitable for downstream processing tasks.

C. Feature Engineering

Feature engineering transforms raw transcript data into meaningful representations for models in Natural Language Processing.

- **Text Representation:** The transcript is converted into numerical form using techniques like TF-IDF and contextual embeddings from BERT to capture semantic meaning.
- **Key Feature Extraction:** Important words and phrases are identified based on frequency, relevance, and linguistic patterns.
- **Sentence Scoring:** Sentences are ranked using features such as position, length, and similarity to determine their importance in the summary.
- **Multilingual Processing:** Models like mBERT ensure consistent feature representation across different languages.

D. Models Implemented

The system uses various models from Natural Language Processing to perform multilingual summarization efficiently.

A. BERT Model

BERT is used to understand the contextual meaning of words in a sentence. It helps in generating better sentence representations, which improves the quality of summarization.

B. mBERT (Multilingual BERT)

mBERT supports multiple languages by mapping them into a shared representation space. This is useful for handling multilingual transcripts and ensuring consistency across languages.

C. Transformer-Based Models

Models based on the Transformer architecture are used for abstractive summarization. These models generate concise and meaningful summaries by understanding the overall context and relationships within the text.

D. TextRank Algorithm

TextRank is an extractive summarization algorithm that ranks sentences based on their importance. It identifies and selects key sentences from the transcript to produce a clear and meaningful summary.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154229

E. Architecture Diagram

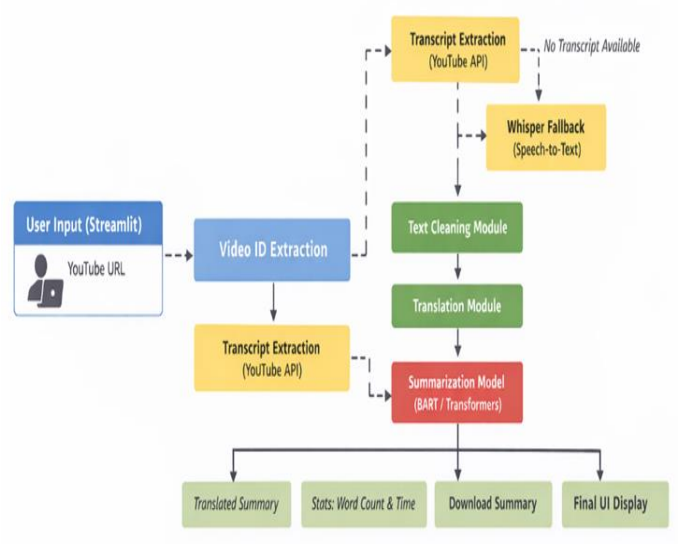


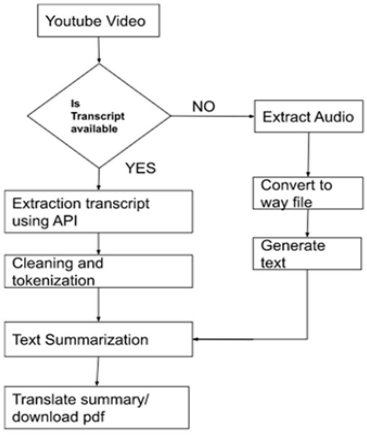
Fig – Architecture Diagram

F. Activity Diagram

The activity diagram represents the workflow of the multilingual video summarization system. The process begins when the user provides a YouTube video URL as input. The system extracts the video ID and retrieves the transcript using the YouTube API. If the transcript is not available, a speech-to-text module is used to generate it.

Next, the extracted text undergoes preprocessing, including cleaning and normalization to remove noise. The system then detects the language of the transcript and translates it into the desired target language if required.

After preprocessing and translation, the cleaned text is passed to the summarization module, where advanced models from Natural Language Processing generate a concise summary. Finally, the system produces outputs such as the translated summary, word count, estimated reading time, and provides options to download or display the results in the user interface.



III. RESULTS AND DISCUSSION

A. Evaluation Metrics

The performance of the system was evaluated using ROUGE and BLEU scores. ROUGE measures how well the generated summary matches the reference summary, while BLEU evaluates the quality of translated text. These metrics help in assessing both summarization and multilingual capabilities.

B. Performance Analysis

DOI: 10.17148/IJARCCE.2026.154229

The system achieved better performance when using advanced transformer-based models. Models like BERT provided more accurate and context-aware summaries compared to traditional methods. Extractive approaches were faster but less effective in capturing the overall meaning.

C. Discussion

The results show that combining extractive and abstractive summarization techniques improves the quality of output. Multilingual models such as mBERT helped maintain consistency across different languages, making the system more reliable for global users.

D. Limitations

The system has some limitations when handling real-world data. Noisy transcripts, low-quality audio, and errors in speech recognition can reduce summarization accuracy. Additionally, translation errors and lack of data for low-resource languages can affect the quality of multilingual summaries.

E. Overall Outcome

Overall, the system successfully generates concise and meaningful summaries in multiple languages. By using advanced NLP models, it improves accessibility and reduces the time required to understand video content, making it useful for a wide range of users.

IV. CONCLUSION

This paper presented a system for multilingual summarization of videos from YouTube using techniques from Natural Language Processing. The proposed approach integrates transcript extraction, language detection, translation, and summarization into a unified pipeline. By utilizing advanced models such as BERT and multilingual architectures like mBERT, the system generates accurate and meaningful summaries across multiple languages. The results show that combining extractive and abstractive methods improves overall performance and usability.

V. FUTURE WORK

The system can be further improved by enabling real-time summarization for live video streams. Enhancing support for low-resource languages and reducing translation errors will increase accuracy. Advanced models based on the Transformer architecture can be integrated to improve summary quality and coherence. Future work may also include incorporating audio and visual features along with text, as well as developing a user-friendly interface for better accessibility and faster processing.

REFERENCES

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[2] A. Vaswani et al., "Attention Is All You Need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
[3] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," in *Proc. ICLR Workshops*, 2013.
[4] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," in *Proc. EMNLP*, 2004, pp. 404–411.
[5] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3730–3740.
[6] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *JMLR*, vol. 21, no. 140, pp. 1–67, 2020.
[7] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training," in *Proc. ACL*, 2020, pp. 7871–7880.
[8] X. Zhang et al., "PEGASUS: Pre-training with Extracted Gap-sentences," in *Proc. ICML*, 2020, pp. 11328–11339.
[9] P. Koehn, *Neural Machine Translation*. Cambridge University Press, 2017.
[10] K. Papineni et al., "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proc. ACL*, 2002, pp. 311–318.
[11] C. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Proc. ACL Workshop*, 2004, pp. 74–81.
[12] Google, "YouTube Data API Documentation."
[13] Google, "Google Translate API Documentation."
[14] J. Howard and S. Ruder, "ULMFiT: Universal Language Model Fine-tuning," in *Proc. ACL*, 2018, pp. 328–339.
[15] A. Conneau et al., "Word Translation Without Parallel Data," in *Proc. ICLR*, 2018.

