© November 2024 | IJIRT | Volume 11 Issue 06 | ISSN: 2349-6002

Survey Paper on Automated Video Transcript Summarizer

Ms. Seema R. Mane, Ms. Khushbu Agrawal, Mr. Krushna Gajare, Mr. Pranav Pisal, Ms. Manasi Wagh

[1] *Assistant Professor, Dept. of Information Technology, Sinhgad Institute of Technology and Science, Narhe, Pune*

[2,3,4,5] *Student, Dept. of Information Technology, Sinhgad Institute of Technology and Science, Narhe, Pune*

***Abstract*—Nowadays, a lot of videos that provide information about various topics are posted every day. Finding the right video and comprehending its information is the main issue since, although there are many videos available, some of them contain useless content, even though we should be able to get the best content. It is a waste of time and effort to extract the correct usage full information if we are unable to find the correct one. We put out a novel concept that employs BERT Summarization for text summarization and NLP processing for text extraction. Users can distinguish between pertinent and irrelevant information based on their needs thanks to this abstractive summary and text description of the video's key content. Additionally, our trials demonstrate that the joint model may get good results in a human review using a multi-line video description and summary that is informative, succinct, and legible.**

***Index Terms*— Plant diseases, Crop security, infection, etc**

I. INTRODUCTION

NLP is a branch of artificial intelligence that focuses on how humans and machines communicate through language. Its main goal is to provide machines the ability to comprehend, interpret, and produce natural language speech or writing. An important use of natural language processing is video summarizing, which is the process of creating accurate and succinct summaries of lengthy recordings. Creating concise, well-organized summaries that highlight the most important details from the original video is the aim. When time is of the essence or a brief synopsis of the video content is required, this technology may prove useful. Typically, video summarization combines a number of methods, including image processing, audio analysis, and text extraction. Finding a brief synopsis of a YouTube video and presenting it in text format is our major goal. Users will benefit greatly from short text summarization techniques since they allow us to quickly read the most important information. In the field of NLP we build summaries of transcripts and produce human readable outputs. These days, YouTube videos may be easily seen by people of all ages for a variety of reasons, including education, entertainment, and many other genres. It is crucial to determine the precise information we need. There are a lot of lengthy videos on the Internet that don't even offer helpful information, wasting our time. By eliminating the videos' superfluous information, we may go straight to the primary content of YouTube videos. Our paper's primary objective is to improve productivity and save users' time. Many people waste their time watching pointless content after only viewing the YouTube video's thumbnail and attention-grabbing title. Before tests, students frequently look for YouTube videos, but because of time limits, they might view them twice as quickly, which could cause them to become confused about the material. Time and effort can be saved by having access to meeting transcripts and recorded sessions, which can be useful in getting a synopsis of the video content. Our paper's primary goal is to distill the most significant information from the transcript into a single, succinct paragraph. Our goal is to save users' time by giving them pertinent and helpful information about the subject they have chosen. In recent years, methods for automatically summarizing texts have been created to make this procedure easier. The process of turning text into a reduced version that gives users the essential content is known as text summarizing. Text summarizing has

many advantages, including content organization, summarization, data retrieval, and question answering, while being a difficult endeavour because machines cannot fully comprehend human language and knowledge. Advances in technology have opened the door for more effective and quick text summary techniques, whereas previous research mostly concentrated on condensing and summarizing texts from a single document.

II. LITERATURE REVIEW

Alrumiah, S. S., Al-Shargabi, A. A. "Educational Videos Subtitles' Summarization Using Latent Dirichlet Allocation and Length Enhancement" [1], Online resources like instructional videos and courses are used by people these days. These videos and courses are typically lengthy, though, so describing them would be helpful. It is possible to create written summaries, or notes, by analyzing the video's visual, aural, and subtitle components. Subtitles for videos contain important information. As a result, summarizing subtitles helps you focus on the important information. Latent Semantic Analysis (LSA) and Term Frequency–Inverse Document Frequency (TF-IDF) models were utilized in the majority of previous studies to generate lecture summaries. Latent Dirichlet Allocation (LDA), which has been shown to be useful in document summarization, is used in this study in a different manner. In particular, there are three stages in the suggested LDA summarization model. By carrying out certain preparation operations, like eliminating stop words, the first phase seeks to get the subtitle file ready for modeling. The LDA model is trained on subtitles in the second phase to produce the list of keywords that are utilized to extract significant sentences. In contrast, the keywords list is used to construct a summary in the third phase. Since the summaries produced by LDA were lengthy, a length enhancement technique has been suggested. The authors created manual summaries of the "EDUVSUM" dataset of instructional videos for the evaluation. The authors compared the produced summaries with the manual-made outlines using two methodologies, (i) Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and (ii) human evaluation. The summaries produced by TF-IDF and LSA perform worse than the summaries produced by LDA. The suggested length enhancement technique not only decreased the length of the summaries but also increased their accuracy rates. Other areas, such as news videos, can utilize the proposed method for video summarizing.

Sangwoo Cho, Franck Dernoncourt , Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, Fei Liu, "StreamHover: Livestream Transcript Summarization and Annotation", The need for new summarizing technology that allows us to generate a preview of streamed information and access this richness of knowledge is critical given the increasing expansion of livestream broadcasting. However, because spoken language is informal, the issue is not trivial. Furthermore, the annotated datasets required for transcript summarization have been hard to come by. In this work, we introduce StreamHover, a livestream transcript annotation and summarization platform. Our benchmark dataset, which consists of more than 500 hours of films with extractive and abstractive summaries, is far bigger than the annotated corpora that are currently available. Utilizing a vector-quantized variational autoencoder, we investigate a neural extractive summarization model that learns latent vector representations of spoken utterances and extracts prominent utterances from the transcripts to create summaries. We demonstrate that our model performs better over robust baselines and has better generalization. The study's findings provide up a possibility for further investigation into better summarizing techniques for effective livestream browsing.

S. Chopra, M. Auli, and A. M. Rush, ''Abstractive sentence summarization with attentive recurrent neural networks'' [3], Sentence with Abstraction Summarization attempts to maintain the meaning of a sentence while producing a condensed version of it. We present a conditional recurrent neural network (RNN) that produces a sentence summary from input. A new convolutional attention-based encoder provides the conditioning, making ensuring the decoder concentrates on the right input words at every stage of generation. Our model is simple to train end-to-end on big data sets and only uses learnt features. Our tests demonstrate that the model performs competitively on the DUC-2004 shared task and greatly outperforms the previously proposed state-of-the-art technique on the Gigaword corpus.

Ghadage, Yogita H. and Sushama Shelke. “Speech to text conversion for multilingual languages” [4], A multilingual speech-to-text conversion system is presented in this study. The information in the speech signal serves as the basis for conversion. The most essential and natural way for humans to communicate is through speech. A human speaking utterance is fed into a speaking-To-Text (STT) system, which outputs a string of words. This system's goal is to extract, describe, and identify speech-related information. Mel-Frequency Cepstral Coefficient (MFCC) feature extraction, Minimum Distance Classifier, and Support Vector Machine (SVM) techniques are used in the implementation of the suggested system for speech classification. Prerecorded speech utterances are kept in a database. The database is primarily separated into two sections: training and testing. After completing the training phase, features are retrieved from samples taken from the training database.

Pravin Khandare, Sanket Gaikwad, Aditya Kukade, Rohit Panicker, Swaraj Thamke, “Audio Data Summarization system using Natural Language Processing” [5], The methods for converting speech audio files to text files and text summarizing on the text files are presented in this study. In the earlier instance, we converted the audio files to text format using Python modules. In the latter instance, text summarizing is done using the modules of Natural Language Processing. The English data functions are implemented using a Python library called SpaCy. Important sentences gleaned from the extraction process are used in the summarization procedure. Words are given weights based on how frequently they appear in the text document. Using this method, summaries of the primary audio file are created.

S.M.Mahedy Hasan, Md. Fazle Rabbi, Arifa Islam Champa, Md. Asif Zaman, “An Effective Diabetes Prediction System Using Machine Learning Techniques”[6], Diabetes, a chronic illness that affects everyone, is brought on by a loss of insulin sensitivity. The only method to lessen the threat of other deadly diseases is to discover diabetes early and take the appropriate treatment, as there is no permanent cure for it. Many studies have been carried out to detect diabetic individuals early utilizing various machine learning algorithms. But increasing the prediction accuracy is never easy due to the dataset's skewed class distribution, missing values, and irrelevant characteristics. In this study, we propose a tree-based machine learning approach for the Pima Indians Diabetes Dataset (PIDD) classification. To improve prediction, a feature selection strategy based on Mutual Information (MI) is used to exclude less significant features. Lastly, the Adaptive Boosting (AB) method is used to enhance the performance of Tree-Based algorithms. The AdaBoost classifier's basis estimator, the Extra Tree (ET) technique, produces the maximum accuracy of 90.5% when compared to experimental data. As a result, our suggested tree-based machine learning model might help medical professionals diagnose diabetes.

Aiswarya K R, “Automatic Multiple Language Subtitle Generation for Videos” [7], In movies, TV shows, video games, and the like, subtitles are textual content material that is derived from a transcript or screenplay of the dialogue or commentary. They are typically shown at the bottom of the screen, but they can also be at the top if there is already textual content at the back of the screen. They can be either a written translation of a dialogue in a foreign language or a written rendition of the dialogue in the same language, with or without statistics to help viewers who are hard of hearing or deaf understand the dialogue, or people who have trouble understanding spoken language or who have trouble with accent recognition. Three components are often introduced in this paper for making subtitles. Speech recognition, audio extraction, and the creation of subtitles in multiple languages.

Savelieva, Alexandra & Au-Yeung, Bryan & Ramani, Vasanth “Abstractive Summarization of Spoken and Written Instructions with BERT” [8], Speech summarization is a challenging task because of the unplanned flow, disfluencies, and other problems that are uncommon in written texts. The BERTSum model is used to conversational language for the first time in our work. We produce abstractive summaries of narrated educational movies on a broad range of subjects, from software configuration and sports to gardening and cookery. We employ transfer learning and pretrain the model on a few sizable cross-domain datasets in both spoken and written English to enhance the vocabulary. In order to restore sentence segmentation and punctuation in the output

of an ASR system, we also preprocess transcripts. ROUGE and Content-F1 scoring are used to assess the outcomes for the How2 and WikiHow datasets. A collection of summaries chosen at random from a dataset gathered from YouTube and HowTo100M are scored by human assessors. We attain a degree of textual fluency and utility that is comparable to summaries produced by human content authors, according to blind review. When applied to WikiHow articles with a wide range of topics and styles, the model outperforms the existing SOTA without exhibiting any performance regression on the standard CNN/DailyMail dataset. The approach has a lot of promise to increase internet content's discoverability and accessibility because of its strong generalizability across many styles and areas. When asked, intelligent virtual assistants will be able to summarize spoken and written instructive content thanks to this functionality.

Patil, S. et al. "Multilingual Speech and Text Recognition and Translation using Image" [9], The aforesaid application will carry out the different elements in the application. Our project's goal is to automate the application to overcome the language barrier between nations and states within the nation. In order to facilitate expressive communication, the application can recognize human speech in one language and translate it into another user-defined language. It provides audio in the translated language and has four modules: speech synthesis, image translation, voice recognition, and translation. Additionally, the application takes in written material and changes it to the appropriate language. The application can identify the text in a picture that has been saved to the system or taken with a camera, translate it into the appropriate language, and then show the translation back on the system screen.

S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'Hom meaux and R. Ptucha, "Semantic Text Summarization of Long Videos" [10], Despite being associated with some of the most significant events in their life, long videos that customers record are, strangely, the ones that are seen the least. It can take an intimidating amount of time to first obtain and view sections. We present new methods for summarizing and annotating lengthy videos in this work. Currently available methods for video summarizing only identify keyframes and subshots, but analyzing these condensed films is difficult. Using recurrent networks, our study suggests ways to annotate and produce textual summaries of lengthy movies in addition to methods for creating visual summaries of them. Interesting parts of lengthy videos are taken out based on cinematography, user desire, and image quality. Sequential encoding and decoding deep learning models are used to transform key frames from the most influential portions into textual annotations. The VideoSet dataset serves as a baseline for our summarization method, which is assessed by humans for linguistic and informative content. We think this is the first completely automated technique that can summarize lengthy consumer movies both visually and textually at the same time.

III. CONCLUSION

We are studied different methods used to video Transcript summarizer. The Automated Video Transcript Summarizer project aims to develop a system that efficiently generates concise summaries of video content. By leveraging a diverse dataset of videos and corresponding transcripts, the project employs advanced Natural Language Processing (NLP) techniques to preprocess and analyze the text. Using both extractive and abstractive summarization methods, and harnessing the power of transformer-based machine learning models like BERT and GPT, the system identifies and conveys the essential information from video transcripts. Performance evaluation and iterative refinements ensure the summaries are coherent and relevant. The final product seamlessly integrates with video platforms, providing users with quick and accurate summaries, thereby enhancing their content consumption experience.

REFERENCES

[1]. Alrumiah, S. S., Al-Shargabi, A. A. Educational Videos Subtitles' Summarization Using Latent Dirichlet Allocation and Length Enhancement. CMC-Computers, Materials & Continua, **70**, 3 (2022).

[2]. Sangwoo Cho, Franck Dernoncourt , Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, Fei Liu, "StreamHover: Livestream Transcript Summarization and Annotation", (2022).

[3] S. Chopra, M. Auli, and A. M. Rush, ‘‘Abstractive sentence summarization with attentive recurrent neural networks,’’ in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol., (2016).
[4]. Ghadage, Yogita H. and Sushama Shelke. “Speech to text conversion for multilingual languages.” 2016 International Conference on Communication and Signal Processing (ICCSP) (2016).
[5]. Pravin Khandare, Sanket Gaikwad, Aditya Kukade, Rohit Panicker, Swaraj Thamke, “Audio Data Summarization system using Natural Language Processing,” International Research Journal of Engineering and Technology (IRJET), **6**, 9 (2019).
[6]. S.M.Mahedy Hasan, Md. Fazle Rabbi, Arifa Islam Champa, Md. Asif Zaman, “An Effective Diabetes Prediction System Using Machine Learning Techniques”, 2nd International Conference on Advanced Information and Communication Technology (ICAICT), (2020).
[7]. Aiswarya K R, “Automatic Multiple Language Subtitle Generation for Videos” International Research Journal of Engineering and Technology (IRJET), **7**, 5 (2020).
[8]. Savelieva, Alexandra & Au-Yeung, Bryan & Ramani, Vasanth. Abstractive Summarization of Spoken and Written Instructions with BERT (2020).
[9]. Patil, S. et al. “Multilingual Speech and Text Recognition and Translation using Image.” International journal of engineering research and technology **5** (2016).
[10]. S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'Hom meaux and R. Ptucha, "Semantic Text Summarization of Long Videos," IEEE Winter Conference on Applications of Computer Vision (WACV), (2017).