IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

Survey on Abstractive Transcript Summarization of YouTube Videos

S. Tharun[1], R. Kranthi Kumar[2], P. Sai Sravanth[3], G. Srujan Reddy[4], B. Akshay[5]
Assistant Professor, Department of Computer Science and Engineering[2]
Students, Department of Computer Science and Engineering[1,3,4,5]
VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana, India

Abstract: *Thousands of video recordings are created and shared on the internet every day. It is becoming increasingly difficult to spend time to watch such videos, which may take longer than anticipated, and our efforts may go in vain if we are unable to extract meaningful information from them. Summarizing transcripts of such videos helps us to quickly search for relevant patterns in the video without having to go through the entire content. Abstractive transcript summarization model is very useful in extracting YouTube video transcripts and generates a summarized version. An automatic summarizer's purpose is to shorten the time of reading, enable easier selection, be less prejudiced compared to humans, and portray content that is compressed while preserving the important material of the actual document. Extractive and abstractive approaches are the two most common ways to summarise text. Extractive approaches choose phrases or sentences from input text, whereas Abstractive methods generate new words from input text, making the task much more difficult.*

Keywords: Transcripts, Text Summarization, Natural Language Processing, REST API, Chrome Extension.

I. INTRODUCTION

Since the development of Digital multimedia, massive amounts of digital content have been posted every day via social platforms. For example: news, documentaries, movies, talk programmes, and sports. This vast amount of disparate content necessitates both processing time and memory. Extraction of condensed form of videos is necessary so that consumers may get the most information in the least amount of time. Reading is a hands-on activity, whereas watching is a passive one. When you're watching educational videos, the last thing you want to do is take notes and miss half of the information. Transcripts of videos provide complete content while allowing viewers to completely participate in the video's vision without the worry of missing anything. The advantages of video transcripts are incalculable, whether you're a YouTube watcher or a video producer. The transcript is useful for doing modifications and capturing audio for legal documentation as the owner of a shared video, not to mention the time saved by talking it rather than typing it. There may be occasions when a video is not permitted or accessible, but a YouTube video transcript might readily serve as a trade saver.

We can construct a brief and articulate summary of a long text content using artificial text summarizing technologies. These approaches can extract applicable details from the input text and use that information more quickly. Not only does this textual conversion lower the quantity of video data, but it also allows users to traverse through it. The users can have a quick glance at the summary of the YouTube video and know whether watching the video is worth their time and if the content in the video matches with the topics that they are looking for.

Automatic Text Summarization is one of the most exciting and challenging problems in the fields of Natural Language Processing and Machine Learning. Extractive and abstractive approaches are the two most common ways to summarise text. Extractive approaches depend only on extracting significant phrases from a document and later combining these significant phrases into a coherent summary. Abstractive approaches analyse the document of text and generate a coherent summary with new words and phrases just like humans do.

In this paper, we have presented the previous work related to the problem statement and also described various text summarization techniques by emphasising more on abstractive text summarization methods.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

II. LITERATURE REVIEW

They used Latent Dirichlet Allocation (LDA) in Paper [1], which has been shown to be effective in summarization of documents. The suggested LDA summarizing model is divided into three stages. The first phase prepares the subtitle file for modelling by deleting stop words and doing other pre-processing tasks. The subtitles are used to train the LDA model in the second step in order to generate the list of keywords that will be employed to extract relevant sentences. The summary is prepared based on the list of keywords generated in the third phase. The quality of LDA-based generated summaries beats that of TF-IDF and LSA summaries.

Stream Hover which is a platform for explaining and summarizing transcripts of live streamed videos was presented in the paper [2]. They looked at a neural extractive summarization model that learns vector representations of audio file and extracts significant observations from subtitles to construct summaries using a vector quantized autoencoder.

The paper [8] offers a system that generates subtitles for movies in either of three languages: English, Hindi, or Malayalam, depending on the user's preference. Audio extraction, voice recognition, and subtitle generation are the three components in the model. Using the FFMPEG platform, audio extraction converts an entire file of any structure to.wav (Waveform Audio) format. The extracted.wav file is then used to create subtitles in the form of a .srt file, which contains the phrases (lyrics) spoken in the audio file. The Google Translate API is used to achieve speech attention. The srt file is then included in the video in the Subtitle creation module, where the lyrics are synchronized with the time and presented alongside the video. The Moviepy video enhancement library in Python is used to accomplish this.

The paper [9] presents a system for generating abstractive summaries of videos on various topics including cooking, cuisine, software configuration, sports, etc. They used Transfer learning & pre-trained the model on large datasets in English to extend the vocabulary. They also did transcript pre-processing to get better sentence formation and punctuation in ASR systems results. For the How2 and WikiHow datasets, ROUGE and Content-F1 scoring are used to assess the results.

Different Text Summarization Methods are classified in the paper [6]. Abstractive Text Summarization is given more weight in the paper. The authors feel that, although being more difficult and computationally intensive than extractive summarizing, abstractive summarization holds more promise in terms of producing more natural and human-like summaries. As a result, we might anticipate further approaches in this subject that provide new viewpoints from computational, cognitive, and linguistic perspectives.

A hybrid end-to-end model, ASoVS, was recommended in the study [12], which uses a deep neural network for generating description and text summary in abstractive method for any video which is given as input. They created a framework which draws people & their traits like gender, age, emotion, etc. as well as scenes, objects, and behaviors for providing a multiple-line video description. They utilized an attention model for abstractive summarization which outperformed baseline techniques and got better results.

The paper [5] discusses methods for converting speech audio files to text files as well as text summarization on the text files. They used Python libraries to convert the audio files to text format in the first scenario. Natural Language Processing modules are utilized for text summarization in the latter situation. The English data functions are implemented using the spaCy Python library. The important sentence obtained when the extraction is studied is used in the summarization approach. Words are allocated weights based on the number of times they appear in the text file. This method is used to create summaries from the original audio recording.

The paper [11] provides a system in which textual explanations are summarized using extractive methods, with the best results coming from the LSA, LexRank, and SumBasic approaches. The results of human reviews of video summaries were favourable.

A conditional recurrent neural network (RNN) was presented in the study [3], which provides the summary of any text which is given as input. The governing is delivered by a unique convolutional encoder, which ensures that decoder targets on correct words during every phase of the generating process. The model is straight forward to train end-to-end on big data sets and depends solely on learned features. The model exceeds the state-of-the-art technique on Gigaword Dataset while competing on the DUC-2004 shared task, according to their experiments.

The paper [10] presents a method in which the application recognizes voice in one language and converts it to a user-defined language for expressive communication. It has four modules: voice recognition, translation, speech synthesis, and visual translation, with audio output of the translated language. In addition, the application receives typed content and translates it into the required language. They employed OCR technology to convert images into text. The abbreviation of

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

OCR is Optical character recognition. It is a technique for extracting text from photographs such as handwritten signs and billboards.

The strategies for generating subtitles are described in the work [7]. Automatically with extensive descriptions of three modules: Audio Extraction, which transforms an MPEG-compliant input file to.wav format, and Speech Recognition, which recognizes extracted speech. The HMM (Hidden Markov Model) determines the likelihood of occurrence of words using the Language model and the Acoustic model. Subtitle generation, which produces a.txt/.srt file that is synchronized with the input file. It is extremely beneficial to persons who are deaf, have reading and literacy difficulties, and are learning to read. A multilingual speech-to-text conversion method is presented in the work [4]. The information in the voice signal is used to convert. For humans, the most natural and important mode of communication is speech. A human voice utterance is input into a Speech-To-Text (STT) system, which needs a string of words as output. The goal of the system is extracting, characterizing, and recognising speech-related content. For voice classification, the suggested system uses Mel-Frequency Cepstral Coefficient feature extraction methodology, as well as the Minimum Distance Classifier and Support Vector Machine techniques.

ROUGE Metrics for evaluating Text Summaries are provided in article [13]. The abbreviation of ROUGE is Recall-Oriented Understudy for Gisting Evaluation. It contains measures for deciding the quality of summary by comparing it to other summaries generated by humans. Between the summary generated by the computer to be evaluated and the ideal summaries written by humans, the measurements count the number of overlapping units like n-grams, word pairs, and word sequences.

The systems proposed in the papers [1], [11] don't work for videos which do not have readily available subtitles. The work in articles [2], [9] is restricted only to English videos. There is absence of Media player in the work presented in paper [8], the entire video must be uploaded in their system in order to generate the subtitles. The system proposed in the work [12] generates video description based on the content present in the video and doesn't include its audio, this system best suits for generating descriptions for CCTV footages. The work presented in paper [5] is applicable only to audio files and is not compatible with video files. The system proposed in article [3] only works for text input, it cannot neither extract subtitles from videos nor generate subtitles for videos. There is only translation and no summarization of text in the work presented in the paper [10], the text which is generated after speech recognition is translated into the target language as output. The work presented in the paper [7] is limited only to extraction of subtitles from the videos. The summarized version of the text is not provided in the work [4].

III. TEXT SUMMARIZATION METHODS

The job of constructing a brief and fluent summary while keeping vital information and overall meaning is known as automatic text summarizing.

For text summarizing, there are essentially two methods:

- Extractive
- Abstractive

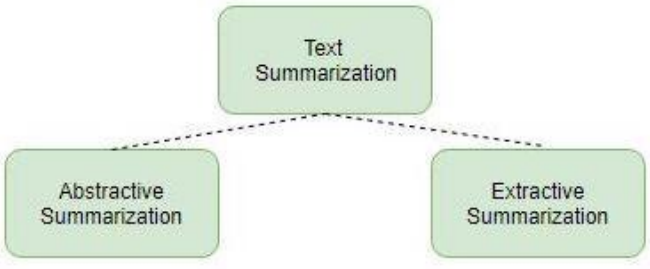

Figure 1: Text Summarization methods

3.1 Extractive Summarization

This approach's name is self-explanatory. We select only the most important sentences or words from native text and extract them. Our summary would be those extracted sentences. Extractive summarization is depicted in the diagram below:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

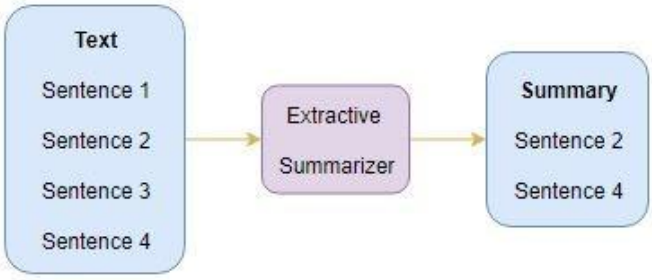

Figure 2: Extractive Summarization model

3.2 Abstractive Summarization

This is an intriguing strategy. We create new sentences from the original content in this step. This is in contrast to our previous extractive technique, in which we only utilised the sentences that were present. It's possible that the sentences formed by abstractive summarization aren't found in the original material:

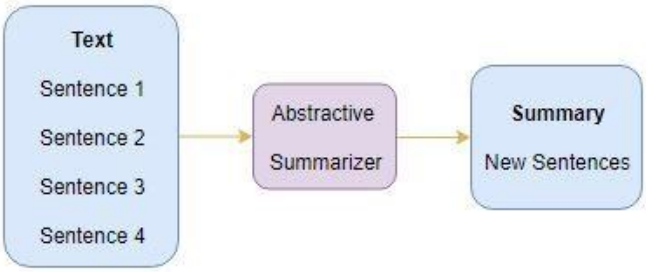

Figure 3: Abstractive Summarization model

3.3 Structure-Based Approaches to Abstractive Summarization

Previous knowledge and psychological feature schemas are used in structure-based approaches. To encapsulate the most significant data, it comprises templates, extraction procedures, and a variety of alternative structures.

A. Tree-Based Methods

The primary notion behind tree-based approaches is that the text or document content is represented by a dependency tree. At the same time, the content selection algorithms varied dramatically from one theme intersection to the next. These algorithms, such as algorithmic programme try crosswise of analysed sentences, are utilised to choose content for outline. A language generator is used to create the outline. Sentence fusion is an example of this type of method. This method parses various documents, discovers similar information by matching input sentences' syntactic trees, and adds paraphrased data. Subsets of the sub-trees are then matched using bottom-up local multi-sequence alignment, fragments are combined using a fusion grid enclosing the alignment, and the grid is transformed into a sentence with the help of a language model. As a result, this strategy integrates statistical approaches like local, multi-sequence alignment and language modelling with grammatical representations generated from input materials.

B. Template-Based Methods

Template-based methods use a predefined guide to represent a complete document. To create a database, grammatical patterns are paired to find word fragments which can be drawn into the guide slots. The data of the outline is specified by the text samples. GISTEXTER is an example of such a method. It is a summarising system that aims to identify subject-related content from input data. This method converts data into database entries and then inserts sentences into ad hoc summaries from the database.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

C. Lead and Body Phrase Method

Tanaka's lead and body phrase technique summarises news and necessitates syntactic analysis of the sentence's head and body pieces. This method, which is based on the sentence fusion methodology, discovers familiar sentences in head and body parts, then inserts and replaces them to construct a summary by modifying sentences. Syntactical parsing of the head and body parts by detecting trigger search pairs, as well as sentence alignment by using various metrics, are among the procedures. Finally, to create a new sentence, both insertion and substitution are used. Substitution happens if body sentence has more content & contains same comparable sentence; if body sentence does not have a counterpart, insertion occurs.

D. Rule-Based Methods

Documents given as input are represented in rule-based techniques as classes and lists of aspects. This technique employs rule-based content extraction unit, text selection heuristics, and one or multiple designs to construct a phrase. Related nouns & verbs are found to build extraction rules, and several candidate rules are chosen to be handed on to the summary creation module. At the end, outline sentences are generated using generation patterns. This method produces the finest summary, but it is time consuming because rules and patterns must be manually written.

E. Graph-Based Methods

In extractive and abstractive summarization algorithms, a graph data structure is commonly used. The system's unique feature is each block describes a word unit, which provides the structure of phrases. Opinosis, a framework which provides condense abstractive summaries of excessively unnecessary opinions, is one of the most well-known projects that uses this technique. This model creates an abstractive summary by continually looking for the Opinosis graph encoding a rational phrase with heavy repetition scores in order to uncover relevant pathways, that become candidate summary phrases. To provide a short summary, all of the paths are ordered in decreasing sequence of their scores, & repeated sequences are deleted using Jaccard measure. Because of the way paths are examined, the Opinosis summarizer is regarded a "shallow" abstractive summarizer because it utilizes the native data to produce summaries. However, because of the way paths are explored, it yields sentences that are invisible in the actual data. As a result, rather from being purely extractive, it is abstract.

F. Ontology-Based Methods

Ontology is widely used in NLP and are used for both extractive and abstractive summarization. Because summarization is usually limited to the same topic or domain, this is useful. Each domain has its unique knowledge structure, which ontology can better represent. All ontology-based summarising methods include compressing and reformulating phrases using linguistic and natural language processing techniques. The "fuzzy ontology" method, that is utilised for summarizing news in China to model ambiguous content, is one of the most influential ways. This method includes a lengthy pre-processing phase that includes domain specialists sketching up a domain ontology for the news occurrences and the extraction of similar words from news dataset. The word classifier then categorises similar words based upon news events. The membership degrees are generated by the inference step for each fuzzy ontology topic. Each concept has a set of membership degrees that are linked to specific domain ontology events. Finally, a news agent uses the ontology to summarise the news.

G. Semantic-Based Approaches

Semantic techniques use linguistic illustrations of data to feed into a natural language generation system, the primary goal of noun and verb phrases identification. Semantic model with multiple modes the multi-modal semantic model represents text and images in multi-modal documents by capturing concepts and forming relationships between them. The under structure of a semantic model is knowledge representation on the basis of objects. Nodes represent concepts, while connections reflect the relationships that exist between them. The entireness, connection with other words, & the total appearances of a statement are all considered when rating important concepts using an information density metric. At the end, the concepts selected are translated into sentences to make a summary.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

H. Information Item-Based Methods

A notional representation of the input document is used to generate this methodology. The abstract representation is tiny block of native content. Content about summary is created from the notional representation of input files rather than words from input files. The main purpose is identification of all text units, their qualities, the predicates that connect them & features of predicates. This method's architecture was introduced in conditions of the Text Analysis Conference for multiple files news summarization. Subject-verb-object triples are produced by syntactical examination of words using a parser at the beginning of the Information Item (INIT) retrieval process. Most INIT do not produce complete sentences, so they must be combined into a sentence structure before being used to generate text. The generation of sentences is done using a language generator during the sentence generating step. The ranking of each sentence is done in the next phase, the sentence selection phase, based on the average document frequency score. Finally, high-ranking sentences are grouped, and the abstract is created with care. This strategy produces a concise, well-organized, information-dense, and less redundant summary. However, because it excludes a lot of useful information, the grammatical quality of the resulting Summary suffers.

I. Semantic Text Representation Model

This method uses semantics rather than syntax or structure to examine input text. Semantic Role Labelling, according to Atif et al., extracts predicate argument structure from every sequence & divides the files set into sequences with file and position numbers. The SENNA semantic role labeller API is used to assign the position number. For semantic similarity scores, the similarity matrix is built using a semantic graph. The predicate structure, semantic similarity, & files set association are then determined using a modified graph-based ranking method. Finally, for summarising, MMR is employed to eliminate redundancy.

J. Semantic Graph Model

The Semantic Graph Model technique creates a summary by generating a rich semantic graph (RSG). There are three stages to the strategy. The rich semantic graph is used to represent input documents in first stage (RSG). The RSG represents the input document's verbs and nouns as graph nodes, with edges corresponding to semantic and topological relationships between them. The syntactic and semantic links produced in pre-processing module connect the sentence concepts. Using heuristic criteria, the actual graph is then reduced to compact graph. At the end, the reduced linguistics graph is used to produce the abstractive outline. This method generates fewer unnecessary and semantically correct sentences, but it is only applicable to a single file.

3.4 Abstractive Summarization Models

A. PEGASUS

Pre-Training with Extracted Gap-Sentences for Abstractive Summarization: During fine-tuning of downstream NLP tasks like text summarization, recent work pre-training Transformers with self-inspected objectives on huge text datasets has demonstrated remarkable results. Pre-training objectives for abstractive text summarization, on the other hand, have not been investigated. Furthermore, systematic evaluation across several domains is lacking. They suggested a new self-supervised aim for pre-training big Transformer-based encoder-decoder models on vast text Datasets. Key sequences from an input document are removed/masked in PEGASUS, and the remaining sentences are formed as single output sentence from the remaining sequences, similar to an extractive summary.

B. Seq2Seq

Seq2seq converts a sequence into another (sequence transformation). In order to avoid problem of vanishing gradient, it employs a recurrent neural network or most commonly LSTM or GRU. The previous stage's output serves as the context for each item. The basic components are an encoder and a decoder network. Each item is encoded into a hidden vector that contains both the object and its context. While using the previous output as input context, the decoder turns the vector into an output item. Optimizations include:

- Attention: The decoder's input is a single vector that contains the complete context. The attention feature allows the decoder to selectively examine the input sequence.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

- Beam Search: Rather than selecting a single output (word), numerous highly probable possibilities are preserved and organised as a tree (using a SoftMax on the set of attention scores). Weighted by the attention distribution, average the encoder states.
- Bucketing: Padding with 0s, which can be done to both input and output, allows for variable-length sequences. When the length of the sequence is 100 and the input is just three items long, however, valuable space is wasted. Buckets come in a variety of sizes, with the lengths of input and output specified.

A cross-entropy loss function is commonly used in training, in which one output is penalised to the extent that the probability of the next output is less than 1.

IV. CONCLUSION

The number of YouTube users in 2020 was approximately 2.3 billion and has been increasing every year. Every minute, 300 hours of YouTube videos are posted. It is frustrating and time consuming to search for the videos that contains the information we are actually looking for. For instance, there are many Ted Talk videos available online in which the speaker talks for a long time on a given topic, but it is hard to find the content the speaker is mainly focusing on unless we watch the entire video. Existing video summarization systems require strong prior technical knowledge. Machine learning based algorithms require high processing power. Summarizing video based on its subtitle is the fastest way of generating video summary, because dealing with text is easier and faster compared to training various videos using machine learning models. This paper presents the users a predominant advantage of producing summaries of YouTube videos. Various researches have been carried out to accomplish different modules such as: Text summarization, Transcript generation and translation. A brief summary of different text summarizations, Extractive Summarization and Abstractive Summarization has been described. Subsequently different models in abstractive summarization have also been described. The fundamental goal of this challenge is to develop a software that produces summaries of YouTube videos automatically to get the gist of entire video before watching it. To develop the model two methods have been described: PEGASUS and Seq2Seq.

V. FUTURE SCOPE

The hearing-impaired individuals and the students are the groups of people in the society that would be benefitted the most with the Transcript Summarization of YouTube videos. The hearing-impaired people find it difficult to understand the videos without the transcripts or subtitles. It would be helpful for them if summary is generated even for the videos which do not have readily available transcripts. Summarising the transcripts of the YouTube videos would also help the students to pick lecture/tutorial videos based on their preferences. The concept of Transcript Summarisation can also be extended to other streaming services as well.

REFERENCES

[1]. Alrumiah, S. S., Al-Shargabi, A. A. (2022). Educational Videos Subtitles' Summarization Using Latent Dirichlet Allocation and Length Enhancement. CMC-Computers, Materials & Continua, 70(3), 6205–6221.

[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" , arXiv : 2109.05160v1 [cs.CL] 11 Sep 2021

[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., June 2016, pp. 93–98.

[4]. Ghadage, Yogita H. and Sushama Shelke. "Speech to text conversion for multilingual languages." 2016 International Conference on Communication and Signal Processing (ICCSP) (2016): 0236-0240.

[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) Volume 06, Issue 09, [September - 2019], e-ISSN: 2395-0056; p-ISSN: 2395-0072.

[6]. Hugo Trinidad and Elisha Votruba, " Abstractive Text Summarization Methods "

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

[7]. Prof. S. A. Aher , Hajari Ashwini M , Hase Megha S, Jadhav Snehal B, Pawar Snehal S, " Generating Subtitles Automatically For Sound in Videos, " International Journal of Modern Trends in Engineering and Research (IJMTER) Volume 03, Issue 03, [March – 2016] ISSN (Online):2349–9745; ISSN (Print):2393-8161

[8]. Aiswarya K R, " Automatic Multiple Language Subtitle Generation for Videos, " International Research Journal of Engineering and Technology (IRJET) Volume 07, Issue 05, [May - 2020], e-ISSN. 2395-0056, p-ISSN: 2395-0072.

[9]. Savelieva, Alexandra & Au-Yeung, Bryan & Ramani, Vasanth. (2020). Abstractive Summarization of Spoken and Written Instructions with BERT.

[10]. Patil, S. et al. "Multilingual Speech and Text Recognition and Translation using Image." International journal of engineering research and technology 5 (2016): n. Pag.

[11]. S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'Hommeaux and R. Ptucha, "Semantic Text Summarization of Long Videos," 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 989-997, doi: 10.1109/WACV.2017.115.

[12]. A. Dilawari and M. U. G. Khan, "ASoVS: Abstractive Summarization of Video Sequences," in IEEE Access, vol. 7, pp. 29253-29263, 2019, doi: 10.1109/ACCESS.2019.2902507.

[13]. Lin, Chin-Yew, "ROUGE: A Package for Automatic Evaluation of Summaries," In Proceedings of 2004, Association for Computational Linguistics, Barcelona, Spain.