JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Review Of Text Summarization Techniques using NLP For Transcripts And Articles

Bhushan Aher
Computer engineering
Modern Education Society's College of Engineering.
Pune, India.
aherbhushan306@gmail.com

Rohit Ushir
Computer engineering
Modern Education Society's College of Engineering. Pune, India
rohitushir27@gmail.com

Prof. Shobha Raskar
Computer engineering
Modern Education Society's College of Engineering. Pune, India
shobha.raskar@mescoepune.org

Onkar Chaudhari
Computer Engineering
Modern Education Society's College of Engineering. Pune, India
onkarnchaudhari@gmail.com

Naresh Barule
Computer Engineering
Modern Education Society's College of Engineering
Pune, India
nareshbarule6@gmail.com

***Abstract*—** Enormous number of video recording and articles are being created and shared across the internet throughout a day. It has become really difficult to spend time watching such videos or reading such articles which may have a longer duration or length than expected and sometimes our efforts may become futile if we couldn't find relevant information out of it. Summarizing transcripts of such videos or summarizing such articles automatically allows us to quickly lookout for the important patterns in the video and helps us save time and effort to go through the whole content of the video. The analysis is completely based on the NLP state of the art technique which is a part of artificial intelligence which helps in language recognition, summarization etc. This paper focuses on the algorithms which helps in summarizing the texts and transcripts generated. This text provides information about how we can use different algorithms to summarize the texts and also converting speech to texts for the videos which do not have transcripts and summarizing it to give an overview about the contents by extractive and abstractive text summarization techniques. This involve NLP, a type of AI that deals with analyzing, understanding and generating natural human languages so that computers can process written and spoken human languages without using computer-driven language

Keywords—: NLP, artificial intelligence, text summarization techniques, abstractive, extractive

INTRODUCTION

Before going to the Text summarization, We must first understand what a summary is. A summary is a text created from one or more texts that delivers information.In the original text, there is a lot of useful information., It's also in a shorter form. The purpose of automatic text summarization is to display the source text in a condensed, semantically rich form. The most significant benefit of adopting a summary is that it cuts down on reading time. Extractive and abstractive summarization are two types of text summarization approaches. Choosing crucial terms, An extractive summarising method involves taking paragraphs and other elements from the source content and concatenating them into a shorter version. Abstractive summarization is the process of understanding the primary concepts in a document and then articulating those notions in simple natural language.

Text summary can be classified into two types: indicative and informative. Inductive summary merely conveys to the user the text's core point. This form of summarization is usually 5 to 10% of the original text's length. Informative summary methods, on the other hand, provide brief information on the primary text. The informative summary should be 20 to 30 percent of the length of the main content.

Main steps for text summarization:There are three main steps for summarizing documents. These are topic identification, interpretation and summary generation.

- Topic identification: The text's most important information is highlighted. Position, Cue phrases, and word frequency are some of the approaches utilised for topic identification. The most useful strategies for subject identification are those that are based on the position of sentences.
- Interpretation: The understanding of abstract summaries is required. Various subjects are combined in this step to generate a general content.

- Summary generation: The understanding of abstract summaries is required. Various subjects are combined in this step to generate a general content.

I. LITERATURE REVIEW In the last few years, there have been a slew of notable works on text summarising. Earlier research focused primarily on single-document text summarization. When compared to previous approaches, technology has advanced, as has computing power, paving the door for a faster, more effective, and more precise form of document processing.

Ravali boorugu and Dr. G. Ramesh proposed an extractive based technique which makes use of various text summarization types like summarization based on input, based on purpose, based on output type for product reviews to get an idea of the reviews in summarized version. It also includes single document text summarization(SDTS) and multi document text summarization(MDTS). For the category of based on the input type they have used single document and multi document approach. For the category of based on purpose, it includes generic, domain specific, query based techniques and for output type based, it includes extractive and abstractive based summarization. They proposed various methods by which extractive summarization can be done in the initial phases, discussed the latest research in this arena.

Adhika Widyasari, Edy Noersasongko and Abdul Syukur proposed an automatic text summarization approach . It includes automatically machine generated summary. They aimed on identifying and analyzing methods, datasets, and trends in automatic text summarization research from 2015 to 2019. This includes concept based automatic text summarization. It again classifies into single document, multi document, extractive, abstractive, supervised learning, unsupervised learning in the machine learning approach. They aimed in increasing the performance and giving quick results. Comparison of mostly used methods and algorithms in the text summarization context.

For text summary, Rahul, Surabhi Adhikari, and Monika suggested NLP-based machine learning algorithms. It includes various researches presenting machine learning approaches for text summarization. Classifying spams on twitter using algorithms like Naive Bayes, Random Forest, support vector algorithm and generating EXT text summaries .This also includes various deep learning concepts like convolutional neural networks(CNN), recurrent neural network(RNN).It also includes other algorithms like k-nearest neighbor Newtonian method artificial bee colony, human learning algorithm.

Parth Dedhia, Hardik Pachgade, Meghana Naik proposed a study which completely focuses on abstractive text summarization techniques. Abstractive technique is an advanced summarization technique which also generates grammatically correct summaries unlike extractive technique. To achieve abstractive text summarization they have used RNN. One of the most used is Long Short Term Memory (LSTM). It also includes detailed explanation of LSTM like the equations input types, model architectures.

NEED FOR TEXT SUMMARIZATION

The current explosion of data circulating in the digital world, the majority of which is non-structured textual content, necessitates the development of automatic text summarising technologies that enable users to quickly draw insights from it. For example, if you want to extract specific information from an online news storey or a video, you may have to sift through its content and spend a significant amount of time weeding out the irrelevant material before finding what you need. As a result, adopting automatic text summarizers capable of collecting crucial information while excluding inessential and irrelevant data is becoming increasingly important. As a result, extraction is critical.

II. EXTRACTIVE TEXT SUMMARIZATION

This procedure can be broken down into two parts: preprocessing and processing. Preprocessing is the representation of the original text in a structured format. It usually consists of the following steps: a) Sentences boundary identification, b) Stop word deletion, and c) Stemming. The processing step determines and calculates the features that influence the relevance of sentences, and then weights are assigned to these weights using the weight learning approach. The feature-weight equation is used to get the final score of each phrase. The top-ranking sentences are chosen for the final summary.

III. ABSTRACTIVE TEXT SUMMARIZATION

Abstractive summarization, on the other hand, analyses the whole content to reproduce the original content in a new and optimized way using advanced natural language techniques. The newly generated content is shorter and more importantly, conveys the most critical information of the original content . Abstractive summaries also generate fluent sentences that are grammatically correct, unlike extractive methods, which may lead to disfluent sentences.

In our heads, we generate a semantic representation of the document. We next select words from our general vocabulary (words we frequently use) that meet the semantics to construct a brief summary that encompasses all of the document's main ideas. As you can see, creating this type of summarizer could be tough because it would require the NaturalLanguage Generation.

IV. EXTRACTIVE TEXT SUMMARIZATION WITH SUMMY

The Sumy library includes many algorithms for text summarising. Rather of needing to create your own algorithm, you can simply import it.

1. *LexRank*, In LexRank algorithms, a sentence that is similar to many other sentences in the text has a higher chance of being crucial. The Lex rank approach assumes that a sentence is endorsed by other like sentences and hence is ranked higher. The higher the rank, the more important the information in the summary text is..

2. *LSA(Latent semantic analysis),* is an unsupervised learning algorithm that can be used for extractive text Summarization. It extracts semantically significant sentences by applying singular value decomposition(SVD) to the matrix

of term-document frequency. Importing the lsa from sumy and passing the document will summarize the text using this algorithms.

3. *Luhn,* the approach of this algorithms is based on TFIDF(Term - frequency- inverse document frequency). It is useful when very low frequent words as well as highly frequnt words (stop words) are both not significant. Based on this, sentence scoring is carried out and the high ranking sentences make it to the summary.

4. *TextRank,* is an extractive summarization technique from genesim. It is based on the concept that words which occur more frequently are significant. Hence, the sentences containing highly frequent words are important. Based on this, the algorithm assigns scores to each sentence in the text. The top-scoring sentences are included in the summary.

V. ABSTRACTIVE SUMMARIZATION WITH TRANSFORMERS

Hugging face's transformers supports various common Models like GPT-2, GPT-3, BART, Open AI, GPT, T5.

T5 transformer, is an encoder-decoder model. It converts all language problems into a text-to-text format. Provide the text for decoding purpose to the transformer. It will perform the operations and in the final step encode the output to get the original text.

BART transformer, the bidirectional and auto-regressive or BART is a transformer that combines bidirectional encoder with an auto regressive decoder into seq2seq model.

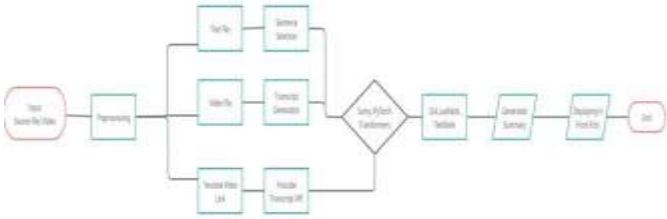
GPT-2 transformer, is another major player in the text summarization, introduced by open AI with the process similar to BART. **Algorithm:**

1. Send an input file in the video or text format.
2. Send the request to the back end.
3. Taking the input file as a video or as a text.
4. Generate the transcript for the video by speech to text.
5. Apply appropriate algorithm/method.
6. Send the summarized version to front end.
7. Display the summary.

VI. TEXT SUMMARIZATION HISTORY

Extractive summarizers have traditionally relied heavily on scoring sentences in the original document. Statistical methodologies or linguistic strategies are used in the most prevalent and current text summarising systems. The sentences are weighted using high frequency terms, standard keywords, Cue Method, Title Method, and Location Method.

DATA FLOW



VII. CONCLUSION

Although automatic text summarization is an ancient problem, current research is focusing on growing trends in biomedicine, product reviews, education domains, emails, and blogs. This is owing to an abundance of knowledge in these fields, particularly on the World Wide Web. In the field of NLP (Natural Language Processing), automated summarization is a hot topic. It entails constructing a summary of one or more texts automatically. Extractive or abstractive document summarization selects a number of indicative sentences, chapters, or paragraphs from the original content automatically. Text summarising techniques based on neural networks, graph theory, fuzzy logic, and clustering have all been successful in producing an effective summary of a document to some extent. Methods that are both extractive and abstractive have been studied. The majority of summarising approaches rely on extractive techniques. The abstractive method is akin to human summaries. Currently, abstract summarization necessitates a lot of work

ACKNOWLEDGMENT

Without the help of our guides, Prof. Shobha Raskar and Prof. Jaya Mane, this article and research would not have been possible.

REFERENCES

[1] Parth Rajesh Deshia, Hardik Pradeep Pachgade: Study on abstractive text summarization techniques, 2020. Emerging advances in information technology and engineering are the focus of this international conference. 2020.

[2] Rahul, Saurabhi Adhikari,Monika: NLP based machine learning approaches for text summarization, 2020. fourth international conference on computing methodologies and communication. 2020.

[3] Ravali Boorugu and Dr. G. Ramesh: a survey on NLP based text summarization for summarizing product reviews.Second international conference on inventive research and computing application. 2020.

[4] Dr. Gajula Ramesh, Dr.J.Somasekar, Dr. Karanam Madhavi, Dr. Gandikota Ramu, Best keyword set recommendations for building service-based systems International Journal of Scientific and Technology Research, volume 8, issue 10, October, 2019.

[5] Adhika Widyasari, Edy Noersasongko, abdul syukur.International conference on information and communication technology. 2019.

[6] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 Int. Conf.
Data Sci. Commun. IconDSC 2019, pp. 1–3, 2019, doi: 10.1109/IconDSC.2019.8817040

[7] Prabhudaas Janjanam and Pradeep Reddy: Text summarization-an essential study. Second international conference on computational intelligence in data science.2019.

[8] T. Jo, "K nearest neighbor for text summarization using feature similarity," Proc. - 2017 Int. Conf. Commun.
ICCCCEE 2017, pp. 1–5, doi: 10.1109/ICCCCEE.2017.78667059. Control. Comput. Electron. Eng. ICCCCEE 2017, pp. 1–5, doi: 10.1109/ICCCCEE.2017.78667059.

[9] B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Multidocument extractive text summarization: A comparative assessment on features," Knowledge-Based Syst., vol. 183, p. 104848, 2019, doi: 10.1016/j.knosys.2019.07.019.

[10] M. Afsharizadeh, H. Ebrahimpour-Komleh, and A. Bagheri, "Query-oriented text summarization using sentence extraction technique," 2018 4th Int. Conf. Web Res. ICWR 2018, pp. 128–132, 2018, doi: 10.1109/ICWR.2018.8387248.

[11] L. Cuiling, "Text Automatic Summarization Generation Algorithm for English Teaching," 2016 Int. Conf. Intell. Transp. Big Data Smart City, p. 2016, 2017