May-June 2020
ISSN: 0193-4120 Page No. 4351-4354

Abstract Extraction from Audios

Aditi H M[1], Akhila Kulkarni[2], Akshatha Janardhan[3], Anuhitha N[4], Shruthi G[5*]

[5]Professor, [1,2,3,4,5]School of Computing and Information Technology, REVA University, Bengaluru, India
[1]aditihm311@gmail.com, [2]akhila.kulkarni1998@gmail.com, [3]akshatha.janardhan98@gmail.com,
[4]anuhithan@gmail.com, [*5]shruthig@reva.edu.in

Article Info
Volume 83
Page Number: 4351-4354
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract
Abstract extraction from audios is the method of identifying the beneficial and essential information from a given audio. This process involves speech signals to be converted into a sequence of words, other linguistic units by making use of an algorithm which is implemented as a computer program. An extractive speech abstraction method is devised on the obtained sequence of words. Extractive speech abstraction normally works as a binary classifier determining whether a sentence is a part of the abstract or not. As a result, using rank based classifiers, we represent the importance among sentences. Based on the weights assigned, the sentences are ranked. Consequently, a high-quality abstract is extracted from the input by choosing the sentences which are highly ranked and formed extract is stored as a text.

Keywords**:** *abstract, text, document, ranking, sentence*

1. Introduction

In today's world, a large amount of information is being generated every day. It thus becomes mandatory to have a better method to extract the gist of the information most effectively. Audio abstract extraction is the method of developing small, precise, and eloquent text gist of a larger document. Abstract can be defined as "a text that can be produced from one or more texts which conveys important information present in the original file". The basic description summarizes some essential views that define automated automatic abstraction: Abstracts are created using one or more documents. They must conserve essential information. Abstracts need to be small in size. Thus we use Text Summarization using rank algorithms for this. Text summarization is a technique that generates abstracts of any given document by extracting useful information from it. It is widely used providing a brief version of large documents. Texts are summarized in two ways: Abstractive and Extractive. The former method involves human knowledge and understanding the content to generate the abstract whereas the later method picks significant sentences from the given text and generates the abstract. In this paper, summarization is done using the Extractive method. This can be extended to summarizing multiple documents. Here, the text summarization is done using sentence scoring. The text obtained from the audio is broken down into sentences and tokens during the pre-processing stage. Combination of various parameters is considered for scoring each sentence. The scores determine the rank of the sentences. The abstract is generated by choosing the specified number of sentences which have the highest rank. Abstracts reduce the time that is needed for listening and understanding the audio and also reduces the storage capacity that is needed for storing information. Automatic abstract extraction methods are mostly required to address the rapidly increasing amount of data present online to help explore related information and to absorb related information rapidly.

2. Related Work

This section contains a brief discussion about other researches carried out on Speech summarization. The first important step towards an automatic speech recognition system is to discard noise and give prominence to parts of the speech signal which can identify linguistic content. In [7] Worked with the speech command dataset to construct a system that recognizes and understands verbal instructions using python libraries. There are 3 essential steps for summarization as mentioned in [8] [9]. The first step involves analyzing the original text and selecting essential features. The next step transforms the outcome of the previous step and the Synthesis step includes the final representation of the summary. The gist of a document can either be obtained by extractive summarization or abstractive summarization. The former [1] chooses sentences from a document to give the gist while the later [2] [3]

May-June 2020
ISSN: 0193-4120 Page No. 4351-4354

understands the meaning or the content of the whole document. The main challenge faced in Extractive summarization is to be able to choose sentences from the input that are important and those that should be included in the gist [10]. LUHN's work [4] still holds good even after so many years. His work says that the frequency i.e. occurrence of a particular word is the most important factor that contributes to the summary. Jing's work suggests that removal of phrases or words that are not significant in the process of getting a gist is the effective way for extraction summarization. BAXENDALE from his work came to a conclusion that accuracy of the summary increases when the very first sentence of the paragraph is included in the gist. He found that around 85% of the paragraph contained the topic of it in the first line and around 7% in the last. Hence his work suggests that position is the key to summary. Fang Chen in their work emphasizes on the length of the sentence. They say if a sentence is too long or too short it is considered irrelevant [5]. EDMUNDSON introduced two new factors: skeleton and cue words in addition to the combination of frequency and the position to get accurate summaries. They considered 3 factors to get accurate summaries, frequency of phrases, frequency of words and the length of the sentence in a paragraph. A survey by Moratanch and Chitrakala suggested that there are two levels of feature extraction. Those are token level and sentence level. The sentence level includes sentence length, location and sentence to sentence cohesion [6]. Arunlfo and Ledeneva [11] in their work talks about using Term Frequency to give weight for each word and find whether the term belongs to the summary. Zhang and Li [12] used the K-means clustering algorithm to form a cluster of sentences. Central sentence of the cluster is considered as the summary.

3. Methodology

In this research, we are using datasets with different kinds of documents to test the model and extracting abstract from them in an efficient manner. Once the model has been tested, we use a procedure to extract abstracts from audio files

The procedure followed in doing this is as follows:

- Input an audio file
- Extract the text from audio and give it as input to the extractive Text summarization Model
- Tune the model to give an efficient output abstract
- Obtain an end to end automatic abstract extractor

Inputting an audio file:

The libraries in Python that are helpful in audio signal processing are LibROSA and SciPy. Using these libraries, visualization of the audio files can be done. Visualization enables us to understand the data and the preprocessing steps in a better way. Once this is done, the audio can be processed and given to the abstract extraction model.

Extraction of text from the audio file:

1. Sentence segmentation: This involves segmentation of extracted content into sentences. The segmentation is done by recognizing the extremity of the sentence which can be (.), (!) or (?) along with the complete count of sentences in the extracted content.

2. Tokenization: In this, sentences are broken down into words by recognizing the spaces and special characters.

3. Removal of stop words: Stop words are the words which do not have much importance and are supplied to the system which in turn compares the tokens obtained from tokenization with every stop word defined and then disposes them as they can influence the abstract that is going to be generated.

4. Stemming: For convenience, words which are in distinct forms in the text are converted to their root form.

After this, the input is segmented to form a group of words and these are ranked based on frequency, sentence position, cue words, similarity with the title, sentence length, proper noun and sentence reduction.

The ranked sentences are arranged, starting from the highest rank to the lowest. Once this is done, the user specified number of highly ranked sentences are selected. Rearrangement of sentences according to the original text is done to make the abstract more meaningful.

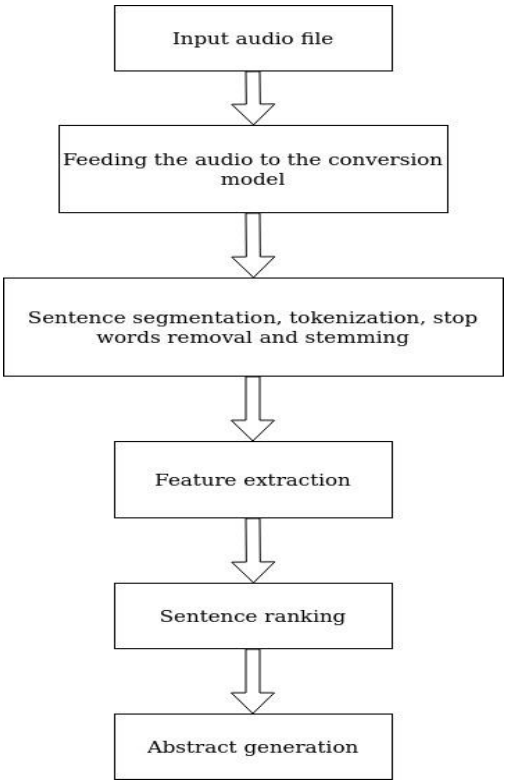

Figure 1: Workflow Diagram

May-June 2020
ISSN: 0193-4120 Page No. 4351-4354

Techniques used in ranking the sentences:

In this research, we use two of the commonly used techniques to determine the more relevant words in the topic, which are Word Probability (WP) and Term Frequency-Inverse Document Frequency (TFIDF). Word Probability is used when the system tells which word can be used in the sentence.

Here frequency is used as an indicator.

- Firstly, all the text present in the text is linked in a chain.
- Then the text is divided into separate sentences.
- Then there is a vector representation for individual sentences.
- Then a matrix is populated by calculating the closeness among sentence vectors.
- This similarity matrix obtained is transfigured to form a graph, in which the vertices represent the sentences and the edges represent the similarity scores. This will be used for calculating sentence rank.

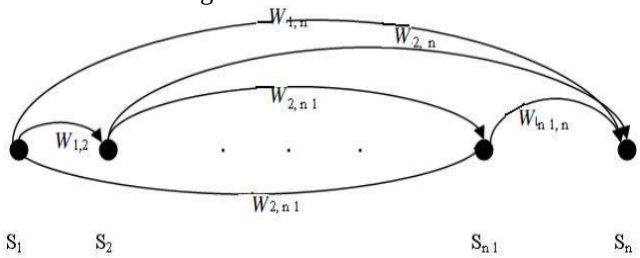

Figure 2: Vector representation to calculate similarities between sentences

- The last step will be extracting an abstract from the top ranked sentences.

The above technique is known as The Extractive text summarization [1] that includes choosing the sentences and phrases from the original text to generate the new summary. Here, using the Sentence Rank algorithm comes into picture.

4. Experimental Results

The existing systems use algorithms to generate a summary from a text document. In the proposed system, we have included an audio module which takes in an audio file as the input, goes through a set of processes where all the unnecessary data is removed from the content and only the important information is retained, the sentence rank algorithm is applied and the resultant abstract is generated. The size of the abstract is provided and can be changed when needed. The resultant abstract will contain the highly ranked sentences and are arranged in the order in which they are present in the original data.

5. Conclusion and Future Enhancement

The massive growth of the internet has made abundant information accessible. Humans are having a tough time summarizing vast quantities of data. Hence, tools for summarization are immensely needed. In this paper, we have used methods that convert a given audio to its corresponding text which is further converted to an abstract using the Sentence Ranking algorithm, an algorithm for ranking keywords and text segments using measures of semantic relatedness and text rank algorithm. Segmentation as proposed in this paper plays an important role in retaining all retrievable information. The purpose of summarization is to select a number of indicative sentences from the original text. It diminishes a content to form an abstract that holds the foremost vital points of the original content. It provides a concise and informative text. It quickly determines which articles are worth reading. It reduces information overload. Automatic summarization improves the effectiveness of indexing and provides unbiased summaries compared to human beings.

This research work can be extended to be developed and used as means of communication in different languages. It can also be used to extract summaries of speeches and news which would convey the same information in a written form. To make it convenient, the extracted abstract could be converted back to speech to help people with reading disabilities.

References

[1] J. N. Madhuri and R. Ganesh Kumar, *"Extractive Text Summarization Using Sentence Ranking*," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3.

[2] Vishal Gupta,G.s. Lehal. *"A survey of text mining techniques and applications"*, Journal of Emerging Technologies in Web intelligence,VOL 1,NO 1,6076,August 2009

[3] G.Erkan, Dragomir R.Radev. *"LexRank:graph based centrality as salience in Text summarization*",Journal of Artificial

[4] Luhn, H (1958). *"The automatic creation of literature abstracts*". IBM Journal of Research Development, 2(2):159-165.

[5] Y. Liu, L. Jin and C. Fang, *"Arbitrarily Shaped Scene Text Detection With a Mask Tightness Text Detector*," in IEEE Transactions on Image Processing, vol. 29, pp. 2918-2930, 2020.

[6] N. Moratanch and S. Chitrakala, *"A survey on extractive text summarization,"* 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, 2017, pp. 1-6.

[7] https://www.analyticsvidhya.com/blog/2019/07/learn-build-first-speech-to-text-model-python/

[8] U. Hahn and I. Mani, "*Automatic Researchers are investigating summarization tools and methods"* IEEE Computer 33.11, no. November, pp. 29–36, IEEE, 2000.

May-June 2020
ISSN: 0193-4120 Page No. 4351-4354

[9] K. Sp¨arck Jones, "*Automatic summarising: The state of the art,"*Infor-mation Processing & Management, vol. 43, pp. 1449–1481, nov 2007

[10] R. Ferreira, L. De Souza Cabral, R. D. Lins, G. Pereira E Silva,F. Freitas, G. D. C. Cavalcanti, R. Lima, S. J. Simske, and L. Favaro,"*Assessing sentence scoring techniques for extractive text summarization,"* Expert Systems with Applications, vol. 40, no. 14, pp. 5755–5764,2013.

[11] R. A. Garc´ıa-Hern´andez and Y. Ledeneva, *"Word sequence models for single text summarization*," inProceedings of the 2nd Internationa lConferences on Advances in Computer-Human Interactions, ACHI 2009,pp. 44–48, IEEE, 2009.

[12] P.-y. Zhang and C.-h. Li, "*Automatic text summarization based on sen-tences clustering and extraction*," in2nd IEEE International Conference On Computer Science and Information Technology, vol. 1, pp. 167–170,IEEE, 2009