Automatic Textual Summary Generation Using Multi-Modal Summarization

Rutuja Deshmukh

Department of Information Technology

Smt. Kashibai Navale College of Engineering, Pune, India

Dr. Lalit.V.Patil

Department of Information Technology

Smt. Kashibai Navale College of Engineering, Pune, India

Abstract **– Multi-Modal summarization is a valuable research tool for both professional and academicians. In this paper, we propose a Multi-Modal summarization method that can automatically generate a textual summary from asynchronous combination of Multi-Modal data such as text, image, audio and video. Our basic goal is to bridge the semantic gap between Multi-Modal content. The summary generation process for textual data includes techniques such as Stemming, Spell correction and N-gram generation. Similarly, for audio data techniques used are Speech transcriptions, which are then converted to text. In the image processing mechanism, the text contained in the images is extracted using OCR (Optical Character Recognition) technique, followed by techniques such as text image matching, topic modeling and theme identification. Similarly, for video processing the video is converted into frames, after which image processing and audio processing techniques are applied to it. Finally, all the Multi-Modal data is collected to generate a textual summary using the LDA (Latent Dirichlet Algorithm). The contribution work is theme identification amongst the visual data. The experimental results show that the Multi-Modal Summarization method outperforms other competitive techniques.**

Keywords **– Summarization, Multimedia, Multi-Modal, Cross-Modal, Natural Language Processing, Computer Vision, OCR Technique.**

1. INTRODUCTION

Text summarization refers to the technique of reducing the length of text in order to create a coherent and fluent textual summary with only the main points outlined in the document. In the recent years, much work has been performed to summarize meeting recordings, sport videos, movies, pictorial storylines and social multimedia. All of the above videos contain data in a synchronous form. Videos such as meeting recordings or sport videos contain synchronous collection of voice and images along with captions. Input given for the summarization of pictorial storylines consists of a set of images with text descriptions. In all of the summarization done above, all the multimedia data that needed to be summarized was in a synchronous form. Our aim is to generate a textual summary from the data which consists of asynchronous information. Multi-Modal summarization can provide users with textual summaries of the data, so that they can acquire a gist of information within short time, instead of reading the whole document or watching the lengthy videos. Almost, many of the summarization methods currently available focus mainly on NLP (Natural Language Processing). Here we try to elevate the quality of the summary generated using techniques such as ASR (Automatic Speech Recognition), OCR (Optical Character Recognition) and Computer Vision. The summary generation process for textual data includes techniques such as Stemming, Spell correction and N-gram generation. Similarly, for audio data techniques used are Speech transcriptions, which are then converted to text. In

the image processing mechanism, the text contained in the images is extracted using OCR (Optical Character Recognition) technique, followed by techniques such as text image matching, topic modeling and theme identification. Similarly, for video processing the video is converted into frames, after which image processing and audio processing techniques are applied to it. Finally, all the Multi-Modal data is collected to generate a textual summary using the LDA (Latent Dirichlet Algorithm). To keep the Multi-Modal summaries generated, simple and easy to understand, we simplify text so that it contains only the most important information, instead of the whole lengthy text. We define the most important information such as the subject (who did it), the object (to whom or what it was done), the event (the action that was performed) and the other main events related to it and thus leaving out the remaining unnecessary textual stuff. In this way, we convert the lengthy and complex text into simpler sentences. For this purpose, we make use of the LDA (Latent Dirichlet Algorithm). Multi-Modal summarization can potentially be helpful to a large number of readers. For example, professional readers may use Multi-Modal summarization to skim over the content, take the important information easily and thus save their time. A more simplified application can be to the children who are learning to read or write a new foreign language. By getting a shortened and simplified version of the text, they can grasp the meaning behind the foreign words quickly.

Our main contributions are as follows –

- We design a Multi-Modal summarization method that can automatically generate a textual summary from a set of asynchronous data such as text, audio and videos related to a specific topic.

- To make the summary generated more accurate we cover most of the important visual information mentioned and apply theme identification for the visual content available.

The subsequent sections explain the Related Work, Technical details, Existing system architecture, Proposed system architecture, Conclusion and References.

2. RELATED WORK

Here, the benefits and challenges of the Multi-Modal summarization have been discussed.

Learning to summarize images by text and visualize text utilizing images is called Mutual – Summarization. This separates the web image-text data space into three subspaces, namely pure image space (PIS), pure text space (PTS) and image-text joint space (ITJS) [1]. Advantages are – In image summarization procedure; map images from PIS to ITJS via image classification model and describe these images utilizing several high level semantic sentences. In text visualization procedure, map text from PTS to ITJS via text categorization model and then give a visual display utilizing images with high confidence in ITJS. Disadvantages are – Need to improve the Mutual – Summarization performance.

The multimodal saliency representations of audiovisual streams, in which signal (audio and visual) and semantic (linguistic/textual) cues are integrated hierarchically are studied by the author in [2] . Each modality is independently analyzed in individual saliency representations: spectro-temporal for the audio channel, spatio-temporal for the visual channel, and syntactic for the transcribed subtitle text. Advantages are – The produced summaries, based on low-level features and content-independent fusion and selection, are of subjectively high aesthetic and informative quality. Disadvantages are – Need to work on non-linear feature correlation algorithms.

Here the author J. Bian, Y. Yang, and T.-S. Chua proposes a multimedia microblog summarization framework to automatically generate visualized summaries for trending topics [3]. Specifically, a novel generative probabilistic model, termed Multi-Modal-LDA (MMLDA) is proposed to discover subtopics from microblogs by exploring the correlations among different media types. Advantages are – Well organized the messy microblogs into structured subtopics. Generates high quality textual summary at subtopic level. Selects images relevant to subtopic that can best represent the textual contents. Disadvantages are – Only focus on summarizing synchronous Multi-Modal content.

A generic video highlights generation scheme based on information theoretic measure of user excitability was presented by, T. Hasan, H. Boˇril, A. Sangwan, and J. H. Hansen [4]. Excitability is computed based

totally on the likelihood of the segmental capabilities residing in positive regions of their joint chance density function space which are taken into consideration each interesting and rare. The proposed measure is used to rank order of the partitioned segments to compress the general video collection and convey a contiguous set of highlights. Advantages are – This scheme utilizes audio excitement and low-level video features. Effectively combine the Multi-Modal features in video segments and found to be highly correlated with a perceptual assessment of excitability. Disadvantages are – Only work on summarizing synchronous Multi-Modal content.

P. Goyal, L. Behera, and T. M. Mcginnity, in their "A context-based word indexing model for document summarization"[5], propose the novel idea of using the context sensitive document indexing to improve the sentence extraction-based document summarization task. This includes, a context sensitive document indexing model based on the Bernoulli model of randomness. Advantages are – The new context- based word indexing gives better performance than the baseline models. Disadvantages are – Need to calculate the lexical association over a large corpus.

Here the author J. Bian, Y. Yang, H. Zhang, and T.-S. Chua in their "Multimedia summarization for social event",[6] study a summarization framework to automatically generate visualized summaries from the microblog stream of multiple media types. Specifically, the proposed framework comprises three stages: 1) A noise removal approach is first devised to eliminate potentially noisy images. 2) A novel cross-media probabilistic model, termed Cross-Media-LDA (CMLDA), is proposed to jointly discover subevents from microblogs of multiple media types. 3) Finally, based on the cross-media knowledge of all the discovered subevents. Advantages are: Eliminates the potentially noisy images from raw microblog image collection. Disadvantages are: Need to extend the cross-media framework for automatically detecting social events and retrieving related candidate microblogs. Need to personalized microblog summarization based on user profile.

In, "Robust structured subspace learning for data representation," [7] paper, Z. Li, J. Liu, J. Tang, and H. Lu propose a novel Robust Structured Subspace Learning (RSSL) algorithm by integrating image understanding and feature learning into a joint learning framework. The learned subspace is followed as an intermediate space to reduce the semantic hole between the low-level visible functions and the high-
level semantics. Advantages are: The proposed RSSL enables to effectively learn a robust structured subspace from data. The proposed framework can reduce the noise-induced uncertainty.

In "A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization," W. Y. Wang, Y. Mehdad, D. R. Radev, and A. Stent, propose a novel matrix factorization approach for extractive summarization [8], leveraging the success of collaborative filtering. First to consider representation mastering of a joint embedding for text and pictures in timeline summarization. Advantages are: It is easy for developers to deploy the system in real- world applications. Scalable approach for learning low-dimensional embedding's of news stories and images. Disadvantages are: Only work on summarizing synchronous multi-modal content.

In,"Multimedia news summarization in search," Z. Li, J. Tang, X. Wang, J. Liu, and H. Lu [9], broaden a singular technique of multimedia news summarization for looking effects on the Internet, which uncovers the underlying subjects among question-associated news information and threads the news events inside each topic to generate a question- related quick review. hLDA is adopted to discover the hierarchical topic structure from the query-related news articles, and an approach based on the weighted aggregation and max pooling to identify the typical news article for each topic is proposed. Advantages are: Proposed system can present vivid and comprehensive information conveniently. Readers can quickly understand the information that they require via the multimedia summarization in this system. Disadvantages are: Need to apply on news video.

In, "Multi-modal summarization for asynchronous collection of text, image, audio and video." H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, [10] proposes an extractive multi-modal summarization method that can automatically generate a textual summary given a set of documents, images, audios and videos related to a specific topic. The key idea is to bridge the semantic gaps among multi-modal content material. Advantages are: It avoids redundant

information. It provides good readability. Disadvantages are: This works only limited dataset.

3. EXISTING SYSTEM ARCHITECTURE

The traditional applications of MMS include meeting summarization, sports video summarization, pictorial storyline summarization etc. All of these videos include synchronized voice and visible captions. None of them focus on summarization of data that contains asynchronous information. Figure 1 overcomes the limitation of synchronous data by generating textual summary from asynchronous combination of Multi-Modal data. It functions as follows – the text and the image data are collected from the documents, which is then pre-processed. The visual data is divided into key frames by shot detection technique. The audio data is processed using the ASR (Automatic Speech Recognition) technique which is then converted to text using the speech transcriptions. Thereafter two techniques are applied on the data which are – Salience calculating and Text – image matching. Saliency of data is the quality of data by which it stands important from amongst the rest of the huge heap of data. The text-image matching model is trained and the sentences are simplified, for each text-image pair. We can identify the matched pairs if the score s (T,I) is greater than a set threshold value. Finally all the data is jointly optimized for checking saliency, readability, coverage for image and non-redundancy. They make sure that summary included the important content of the input documents and the summary contained very little of the unnecessary lengthy text. Readability makes sure that the summary which will be generated at the last should be in a textual form thus overcoming the other drawbacks.

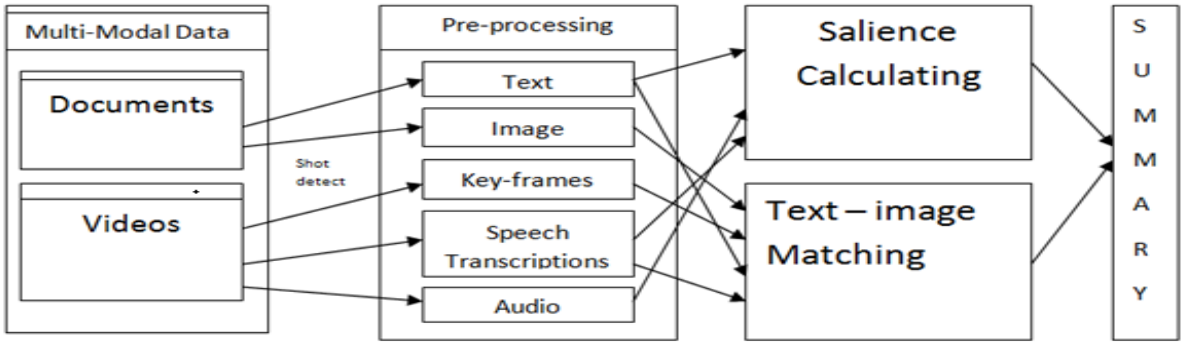

Figure 1 – Existing system architecture

Disadvantages –

1. Especially for the visual data, better models are required since text-image matching does not accurately sum up the data for summarization.

2. This existing system cannot perform well with large datasets

4. PROPOSED SYSTEM ARCHITECTURE

To overcome the drawbacks of the existing system [10], we propose a new system architecture which will not only generate a textual summary from the asynchronous multi-media data such as text, image, audio and video, but also will overcome the limitations regarding the visual data, with help of theme identification. Theme identification will allow us to understand the theme behind the visual data which makes the summary much more reliable and accurate. The system consists of four pre-processing modules such as text pre-processing, image pre-processing, audio pre-processing and video pre-processing; which are then combined together and the textual summary of the data is generated using the LDA (Latent Diritchlet Allocation) algorithm. The first module is the Text pre-processing module as shown in Figure 2. Once we get the textual data as the input, the Stemming process reduces the complex words to their simpler base words. Part-of-speech tagging identifies the correct part of speech

of the word and the N-gram generator are basically set of co-occurring words within a given window.

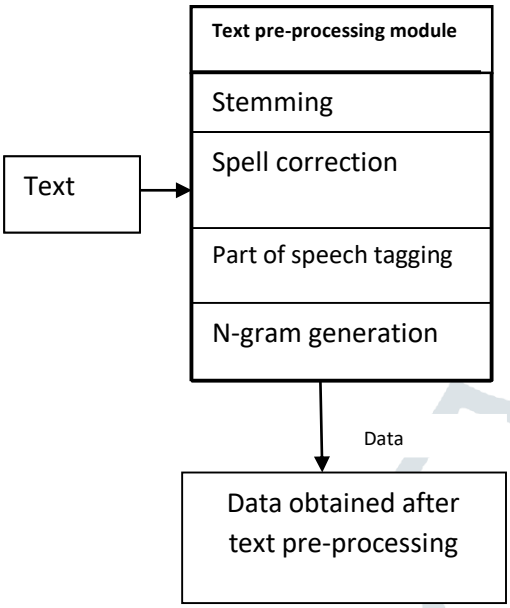

Figure 2 – Text pre-processing

For example, if in a Machine learning problem you want to classify the documents to their respective categories, a N-gram generator is used. The second module is the Image pre-processing module as shown in figure 3.

Once we get the visual data as the input, the text-image matching model is trained and the sentences are simplified, for each text-image pair. We can identify the matched pairs if the score s (T,I) is greater than a set threshold value. The Theme identification is the most important part of the visual data, as it helps us to overcome the drawbacks of the existing system [10]. Theme identification allows us to understand the theme behind the visual data which makes the summary much more reliable and accurate. A topic model allows us to examine a set of documents and discover, based on the statistics of the word, exactly what the topic might be and what each documents balance of the topic is.

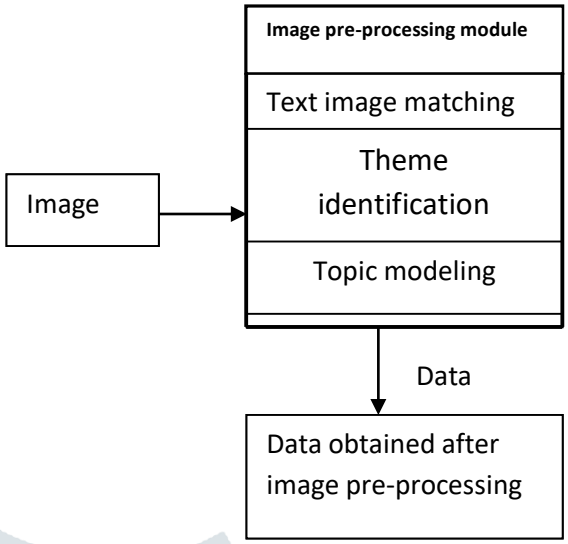

Figure 3 – Image pre-processing

The third module is the Audio pre-processing module as shown in figure 4.

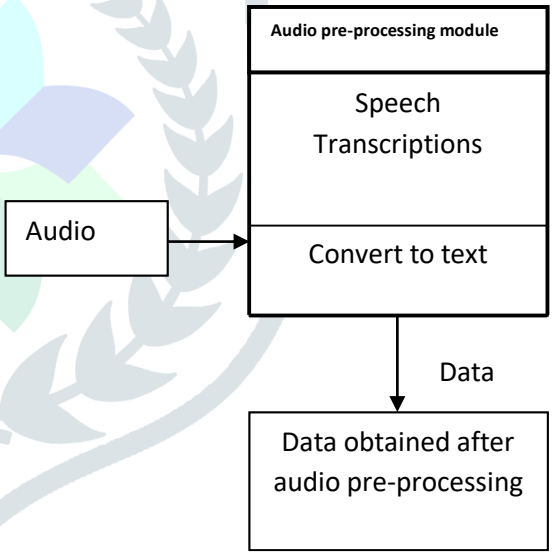

Figure 4 – Audio pre-processing

Once we get the audio data as the input, speech transcriptions are used to convert the audio data to text using ASR (Automatic Speech Recognition). Speech transcriptions are systematic representations of language in written form. Transcription was originally a process carried out manually, however now it is done by software's in computer. The fourth module is the Video pre-processing module as shown in figure 5.

Once we get the video data input, we convert the video into frames using frame conversion technique and then the image pre-processing and audio pre-

processing techniques are applied to it as mentioned in Figure 3 and 4. Once all the four modules are pre-processed, the data is collected and the LDA (Latent Dirichlet Allocation) algorithm is applied to it. It is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

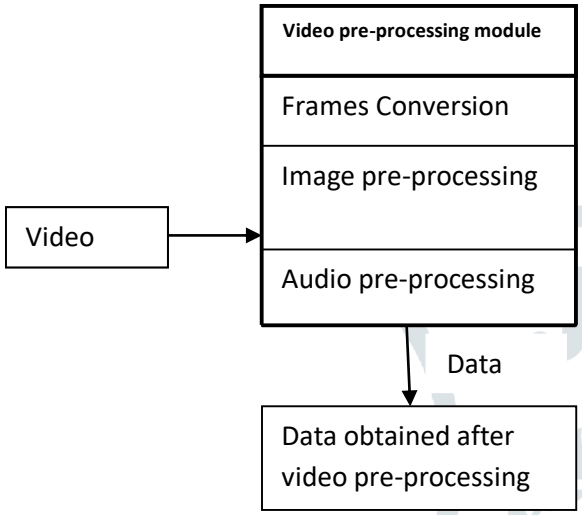

Figure 5 – Video preprocessing

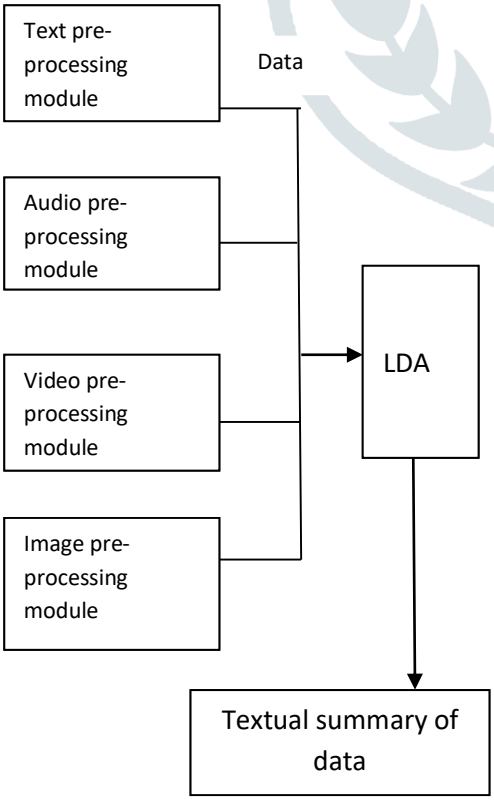

Figure 6- Summary generation using LDA

The working of LDA is given as follows –

LDA affords a generative version that describes how the files in a dataset were created. In this context, a dataset is a group of D files. Document is a group of phrases. So our generative model describes how every report obtains its phrases. Initially, permit's anticipate that recognize K topic distributions for the given dataset, meaning K multinomials containing V elements every, where V is the range of terms in our corpus. Let βi represent the multinomial for the ith topic, where the size of βi is V:|βi|=V. Given these distributions, the LDA generative process is as follows:

Steps:

1. For each document:

(a) Randomly choose a distribution over topics (a multinomial of length K)

(b) For each word in the document:
(i) Probabilistically draw one of the K topics from the distribution over topics obtained in (a), say topic βj
(ii) Probabilistically draw one of the V words from βj.

The collective data is processed with the help of the LDA algorithm as given in the figure 6.

This is how the summary is generated from the asynchronous Multi-Modal data in our proposed system. Theme identification of the visual data is a major contribution in the proposed system.

The advantages of the proposed system are given as follows –

- The proposed system generates a textual summary from the asynchronous Multi-Modal data such text, images, audio and video and proves as useful research tool for academicians and professionals.
- Theme identification of visual data makes summary more accurate than the existing system.

5. CONCLUSION

Thus we have studied the existing system of Multi-Modal summarization and its drawbacks and proposed a new system which overcomes the drawbacks of the existing system. The experimental results show that for any input combinations such as .txt, .docx, .mp3, .jpg, .mp4, .avi files etc. the output is a automatic textual summary which is generated within less time.

6. REFERENCES

1. .Li, J. Ma, and S. Gao, "Learning to summarize web image and text mutually," in Proceedings of the 2nd ACM International Conference on Multimedia Retrieval. ACM, 2012, p. 28.
2. G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," IEEE Transactions on Multimedia, vol. 15, no. 7, pp. 1553–1568, 2013.
3. J. Bian, Y. Yang, and T.-S. Chua, "Multimedia summarization for trending topics in microblogs," in CIKM. ACM, 2013, pp. 1807–1812.
4. T. Hasan, H. Boˇril, A. Sangwan, and J. H. Hansen, "Multi-modal highlight generation for sports videos using an information theoretic excitability measure," EURASIP Journal on Advances in Signal Processing, vol. 2013, no. 1, p. 173, 2013.
5. P. Goyal, L. Behera, and T. M. Mcginnity, "A context-based word indexing model for document summarization," IEEE Transactions on Knowledge & Data Engineering, vol. 25, no. 8, pp. 1693–1705, 2013.
6. J. Bian, Y. Yang, H. Zhang, and T.-S. Chua, "Multimedia summarization for social events in microblog stream," IEEE Transactions on Multimedia, vol. 17, no. 2, pp. 216–228, 2015.
7. Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 10, pp. 2085–2098, 2015.
8. W. Y. Wang, Y. Mehdad, D. R. Radev, and A. Stent, "A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization," in NAACL-HLT, 2016, pp. 58–68.
9. Z. Li, J. Tang, X. Wang, J. Liu, and H. Lu, "Multimedia news summarization in search," ACM Transactions on Intelligent Systems and Technology, vol. 7, no. 3, p. 33, 2016.
10. H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Multi-modal summarization for asynchronous collection of text, image, audio and video." in EMNLP, 2017, pp. 1092–1102.