ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

Automatic Textual Summary Generation Using Multi-Modal Summarization Framework

Prasanna Vidhate, Amar Pawar, Jai Golande, Rushikesh Dalvi, Girish Navale

AISSMS Institute of Information Technology, Pune, Maharashtra, India

ABSTRACT: In recent years, much work has been performed to summarize meeting recordings, sport videos, movies, pictorial storylines and social multimedia. Automatic text summarization is an essential natural language processing (NLP) application that goals to summarize a given textual content into a shorter model. The fast growth in multimedia information transmission over the Internet demands multi-modal summarization (MMS) from asynchronous combination of text, image, audio and video. This paper represents an MMS framework that utilizes the techniques of NLP, speech processing and OCR watermarking technique to examine the elaborative information contained in multi-modal statistics and to enhance the aspects of multimedia news summarization. The basic concept is to bridge the semantic gaps among multi-modal content. For audio scheme, convert the audio signals into textual format. For visual scheme, extracts text from images using OCR technique. After, the generated summary for important visual information through content-picture matching or multi-modal topic modeling. Finally, all the multi-modal factors are considered to generate a textual summary by maximizing the importance, non-redundancy, credibility and scope through the allocated accumulation of submodular features. The contribution work is to identify theme of visual scheme. The experimental result shows that Multi-Modal Summarization framework outperforms other competitive techniques.

KEYWORDS: Summarization, Multimedia, Multi-Modal, Cross-Modal, Natural Language Processing, Computer Vision, OCR Technique

I. INTRODUCTION

TEXT summarization plays a vital role in our daily life and has been studied for several decades. From information retrieval to text mining, we are frequently exposed to text summarization. Multimedia data (including text, image, audio and video) have increased dramatically recently, which makes it difficult for users to obtain important information efficiently. Multi-modal summarization (MMS) can provide users with textual summaries that can help acquire the gist of multimedia data in a short time, without reading documents or watching videos from beginning to end. When summarizing meeting recordings, sport videos and movies, such videos consist of synchronized voice, visual and captions. For the summarization of pictorial storylines, the input is a set of images with text descriptions. None of these applications focus on summarizing multimedia data that contain asynchronous information about general topics. Intuitively, readers can grasp the gist of the event more easily by scanning the image or the video than by only reading news document, and thus we believe that the multi-modal data will also reduce the difficulty for machine to understand a news event. While most summarization systems focus on only natural language processing (NLP), the opportunity to jointly optimize the quality of the summary with the aid of automatic speech recognition (ASR) and computer vision (CV) processing systems is widely ignored. On the other hand, given a news event (i.e., news topic), multimedia data are generally asynchronous in real life, which means there is no given explicit description for images and no subtitles for videos. Thus, multi-modal summarization (MMS) faces a major challenge in understanding the semantics of visual information. In this work, we present an MMS system that can provide users with textual

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

summaries to help to acquire the gist of asynchronous multimedia data in a short time without reading documents or watching videos from beginning to end. The purpose of this work is to unite the NLP, ASR and CV techniques to explore a new framework for mining the rich information contained in multi-modal data to improve the quality of multimedia news summarization.

Motivation

- The Multi-modal Topic Modeling is that textual descriptions of images often provide important information about semantic aspects (topics), and image features are often correlated with semantic topics.
- Automatically generate a fixed-length textual summary to represent the principle content of the multi-modal data.

II. RELATED WORK

The paper [1] 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 between multi-modal content. Advantages are: It avoids redundant information. It provides good readability. Disadvantages are: This works only limited dataset.

The paper [2] proposes a multimedia microblog summarization framework to automatically generate visualized summaries for trending topics. Specifically, a novel generative probabilistic model, termed multimodal-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.

The paper [3] Proposes a multimedia social event 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. Generates multimedia summary for social events utilizing the cross-media distribution knowledge of all the discovered subevents. 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.

The paper [4] proposes a novel matrix factorization approach for extractive summarization, leveraging the success of collaborative filtering. First to consider representation learning of a joint embedding for text and images 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.

The paper [5] proposes the novel idea of using the context sensitive document indexing to improve the sentence extraction-based document summarization task. In this paper, proposes 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.

III. OPEN ISSUES

The existing applications related to MMS include meeting record summarization, sport video summarization, movie summarization, pictorial storyline summarization, timeline summarization and social multimedia summarization etc. Meeting recordings, sport videos and movies consist of synchronized voice, visual and captions and pictorial storylines consist of a set of images with textual descriptions. Most of the existing applications focus on synchronous multimodal content, in which images are paired with textual descriptions and videos, are paired with subtitles.
Disadvantages are:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

1. The existing applications focus on only synchronous multimodal content.
2. MMS discovers the semantic gap between different modalities.
3. MMS shows less quality of generated summaries.

IV. SYSTEM OVERVIEW

This paper proposes an approach to generate a textual summary from a set of asynchronous documents, images, audios and videos on the same news event, as shown in Fig. 1. Because multimedia data are heterogeneous and contain more complex information than that contained in pure text. The MMS framework shows in Fig. 1. For the audio information contained in videos, obtain speech transcriptions through ASR and design a method to selectively use these transcriptions. For visual information, including the key frames extracted from videos and the images that appear in documents learn the joint representations of text and images with a neural network; then identify the text that is relevant to the image based on text-image matching or multi-modal topic modeling. In this way, audio and visual information can be integrated into a textual summary by joint optimization. Contribution work is, to design an MMS method that can automatically generate a textual summary from a set of asynchronous documents, images, audios and videos related to a specific event. Consider four criteria like, salience, non-redundancy, readability and coverage for visual information that are jointly optimized by the budgeted maximization of submodular functions to select the representative sentences. Bridge the semantic gap between the textual and visual data.

A. Architecture

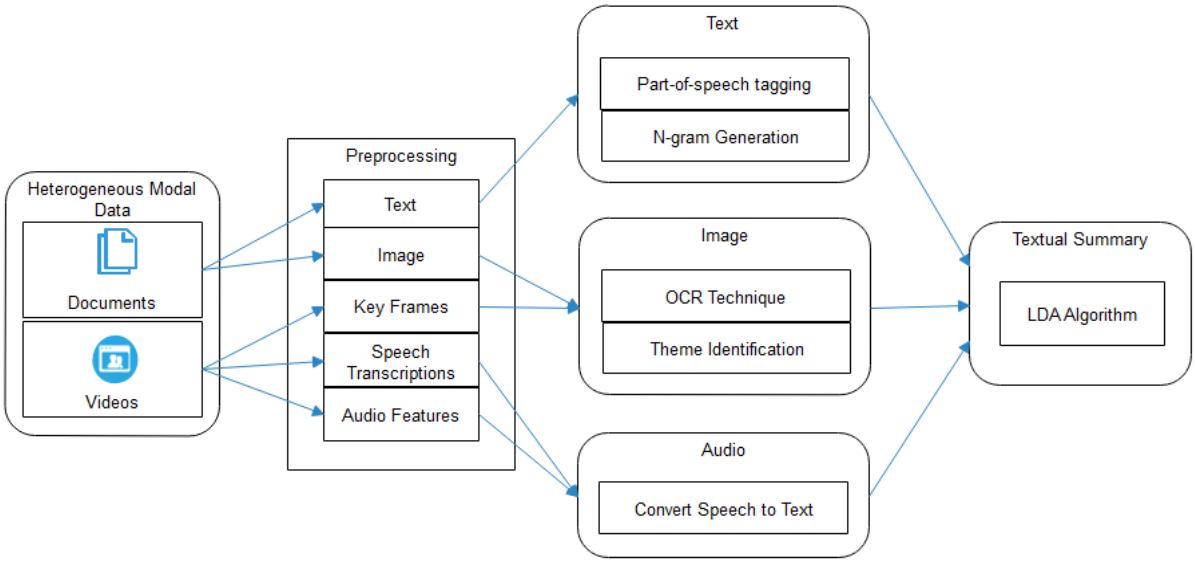

Fig. 1 System Architecture

Advantages are:

1. Provides users with textual summaries to help to acquire the gist of asynchronous multimedia data in a short time without reading documents or watching videos from beginning to end.
2. Automatically generate a fixed-length textual summary to represent the principle content of the multi-modal data.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

B. Mathematical Model

The input is a collection of multi-modal data $M = \{D_1, \dots, D_{|D|}, V_1, \dots, V_{|V|}\}$related to a news event T, where each news document $D_i = \{T_i, I_i\}$consists of text T_i and image I_i (there may be no image for some documents). $V_i$denotes a news video and $|.|$denotes the cardinality of a set. The objective of our work is to automatically generate a fixed-length textual summary to represent the principle content of the multi-modal data M.

An extractive summarization method in which all these aspects can be jointly optimized through the budgeted maximization of submodular functions defined as follows:

$$\max_{S \subseteq T} \{F(S) : \sum_{s \in S} l_s \leq L\} \quad (1)$$

where T is a set of sentences, S is a summary, l_s is the length (number of words) of sentence s, L is budget, i.e., length constraint for the summary, and submodular function F(S) is the summary score.

- Text Salience of Summarization

The text salience of summary S by a diversity-aware objective:

$$F_s(S) = \sum_{i=1}^{K} \sqrt{\sum_{s_J \in P_i \cap S} Sa(s_j)} \quad (2)$$

where$P_i, i = 1, \dots, K$ is a disjoint partition of the set of all the sentences and speech transcriptions V into separate clusters, and the salience scores $s(t_j)$are normalized to [0,1] by dividing by the maximum value among all the sentences.

- Matching-based Image Coverage of Summarization

We model the summary S coverage for the image set I as follows:

$$F_m(S) = \sum_{p_i \in I} Im(p_i).m(p_i, C_j) \quad (3)$$

where the $Im(p_i)$is the weight for image pi. For keyframep_i, $Im(p_i)$is the average salience score of the speech transcriptions within the shot to which p_i belongs. For document image p_i, $Im(p_i)$is the average salience score of the sentences in the document in which p_i is embedded. $C_j$is a sentence or a sentence segment obtained based on semantic framing, chunking or word tokenizing.

C. Algorithms

1. Text Summary Generator Algorithm

Step 1 - It takes a text as input.
Step 2 - Splits it into one or more paragraph(s).
Step 3 - Splits each paragraph into one or more sentence(s).
Step 4 - Splits each sentence into one or more words.
Step 5 - Gives each sentence weight-age (a floating point value) by comparing its words to a pre-defined dictionary called "stopWords.txt".

If some word of a sentence matches to any word with the pre-defined Dictionary, then the word is considered as Low weighted.

Step 6 - An ordered list of weighted sentences is then prepared (Relatively High weighted sentences comes first and low weighted sentences comes At last position).
Step 7 - Now, we have the ordered list of weighted sentences, it continues to Store each sentence (from ordered weighted sentences) in the output Variable (i.e. a list) until it reaches the reduction ratio (It uses a formula to determine max number of sentences to put in the output List).

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

Step 8 - The output list is then returned summary.

2. OCR Algorithm

Process 1: Generating the edge image

Image generateEdgeImage(Image grayImg)
//Create an $X \times Y$ output image edgeImg
//grayImg is the $X \times Y$ result image created in step 1
Step 1: $x \leftarrow 0; y \leftarrow 0; left \leftarrow 0; upper \leftarrow 0; rightUpper \leftarrow 0;$
Step 2: **for all** $pixel_{x,y} \in grayImg$ **do**
Step 3: **if** $(0 < x < X - 1)$ ***and*** $(0 < y < Y)$ **then**
Step 4: $left \leftarrow |pixel_{x,y} - pixel_{x-1,y}|$
Step 5: $upper \leftarrow |pixel_{x,y} - pixel_{x,y-1}|$
Step 6: $rightUpper \leftarrow |pixel_{x,y} - pixel_{x+1,y-1}|$
Step 6: $edgeImg_{x,y} \leftarrow \max(left, upper, rightUpper)$
Step 7: **else**
Step 8: $edgeImg_{x,y} \leftarrow 0$
Step 9: **end if**
Step 10: **end for**
Step 11: $edgeImg_{x,y} \leftarrow sharpen(edgeImg)$
Step 12: return($edgeImg$)

Process 2: Localizing text candidates

textRegion[] detectTextRegions(Image edgeImg)
// edgeImg is created using process 1
//textRegion is a data structure with 4 fields: x_0, y_0, x_1, y_1
//determineYCoordinates uses the process 3
// determineXCoordinates uses the process 4
Step 1:$Integer[]\ H \leftarrow calculateLineHistogram(edgeImg)$
Step 2: $textRegions[]\ TC \leftarrow determineYCoordinate(H)$
Step 3:$TC \leftarrow determineXCoordinate(edgeImg, TC)$
Step 4: return(TC)

Process 3: Determining the Y-coordinates of text regions

$textRegion[]\ determineYCoordinate$
$(Integer[]\ H)$
//H is the line histogram, see step 3
Step 1: $textRegion\ rect;$
Step 2: $textRegions[]\ TC; y \leftarrow 1, j \leftarrow 0; insideTextArea \leftarrow false;$
Step 3: **for** $el_y \in H$ **do**
Step 4: **if** $\Big((el_y > MinEdges)$ ***or*** $\big((el_y - el_{y-1}) > MinLineDiff\big)\Big)$ **then**
Step 5: **if not** $insideTextArea$ **then**
Step 6: $rect.y0 \leftarrow y$
Step 7: $insideTextArea \leftarrow true$
Step 8: **end if**
Step 9: **else if** $insideTextArea$ **then**
Step 10: $rect.y1 \leftarrow y - 1$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

Step 11: **if** $((rect.y1 - rect.y0) > MinLines)$**then**
Step 12: $TC[j] \leftarrow rect$
Step 13: $j \leftarrow j + 1$
Step 14: **end if**
Step 15: $insideTextArea \leftarrow false$
Step 16: **end if**
Step 17: **end for**
Step 18: **return**(TC)

Process 4:Determining the X-coordinates of text regions
$textRegion[]\ determineXCoordinate$
$(Image\ edgeImg, textRegion[]\ TC)$
Step 1: $left \leftarrow maxInt, right \leftarrow -1;$
Step 2: **for**$textCandidate_i \in TC$**do**
Step 3: **for all** $pixel_{x,y} \in textCandidate_i$**do**
Step 4: **if** $(edgeImg_{x,y} \neq 0)$**then**
Step 5: **if** $(left > x)$**then**
Step 6: $left \leftarrow x$
Step 7: **end if**
Step 8: **if** $(right < x)$**then**
Step 9: $right \leftarrow x$
Step 10: **end if**
Step 11: **end if**
Step 12: **end for**
Step 13: $textCandidate_i.x0 \leftarrow left$
Step 14: $textCandidate_i.x1 \leftarrow right$
Step 15: **end for**
Step 16: **return**(TC)

Process 5:Generating the text image
Image $segmentTextRegions$
$(Image\ edgeImg, textRegion[]\ TC)$
//$edgeImg$ is created with Alg. generating the edge image
//TC is the array returned from Alg. $determineXCoordinate$of text regions
Step 1: Image $reducedImg \leftarrow erase(TC, edgeImg)$
Step 2: Image $binaryImg \leftarrow binarize(reducedImg)$
Step 3: Image $gapImg \leftarrow fillGaps(binaryImg)$
Step 4: $TC \leftarrow refineCoordinates$
$(edgeImg, gapImg, TC)$
Step 5: Image $\text{text}Img \leftarrow extractImage(grayImg, TC)$
Step 6: $\text{text}Img \leftarrow enhanceContrast(textImg)$
Step 7: **return**$(\text{text}Img)$

3. Hidden Markov Model (HMM) algorithm for speech recognition:

A HMM is characterized by 3 matrices viz., A, B and PI.
A - Transition Probability matrix $(N \times N)$
B - Observation symbol Probability Distribution matrix $(N \times M)$
PI - Initial State Distribution matrix $(N \times 1)$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

Where, N =Number of states in the HMM
M = Number of Observation symbols
After can apply HMM for speech recognition by using following steps:
1. Recursive procedures like Forward and Backward Procedures exist which can compute P (O|L), probability of observation sequence.

Forward Procedure:
Initialization:

$$\alpha_1(i) = \pi_i b_i o_1, \qquad 1 \le i \le N$$

Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), \qquad 1 \le t \le T-1, 1 \le j \le N$$

Termination

$$P(O|\lambda) \sum_{i=1}^{N} \alpha_T(i)$$

Backward Procedure:
Initialization:

$$\beta_T(i) = 1, \qquad 1 \le i \le N$$

Induction

$$\beta_T(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1}) \beta_{t+1}(j), \qquad T-1 \le t \le 1, 1 \le i \le N$$

Termination

$$P(O|\lambda) \sum_{i=1}^{N} \alpha_T(i)$$

2. The state occupation probability $t(sj)$ is the probability of occupying state sj at time t given the sequence of observations

$$O_1, O_2, \dots, O_N.$$

3.Baum-welch algorithm for parameter re-estimation.

V. RESULT AND DISCUSSIONS

Experiments are done by a personal computer with a configuration: Intel (R) Core (TM) i3-2120 CPU @ 3.30GHz, 4GB memory, Windows 7, MySQL 5.1 backend database and jdk 1.8. The application is dynamic web application for design code in Eclipse tool and execute on Tomcat server. Some functions used in the algorithm are provided by list of jars like standford core NLP jar for keywords extraction using POS tagger method. TalkingJavaSDK jar uses for speech to text conversion and imageio jar uses for image read and write.

For English Multi-modal Summarization Framework, Fig. 2 shows that when summarizing textual data like, .txt, .doc, .docx etc. file format it gives better accuracy. The image file applies the OCR algorithm for textual part extraction and generates summary. The speech transcriptions HMM model performs better on audio files.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

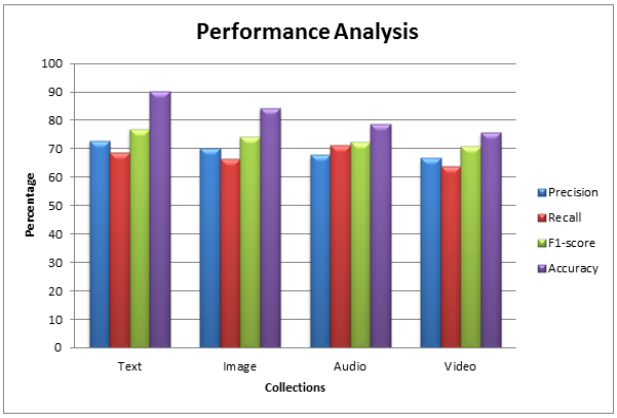

Fig. 2 Performance graph for MMS framework

VI. CONCLUSION

This paper addresses an asynchronous MMS task, namely, how to use related text, audio and video information to generate a textual summary. We formulate the MMS task as an optimization problem with a budgeted maximization of submodular functions. We address readability by selectively using the transcription of audio through guidance strategies. The experimental results shows that giving input any combination like, .txt, .docx, .jpg, .mp3, .mp4, .avi etc. file and output is automatic generated textual summary showing within time.

REFERENCES

[1] 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.
[2] J. Bian, Y. Yang, and T.-S. Chua, "Multimedia summarization for trending topics in microblogs," in CIKM. ACM, 2013, pp. 1807–1812.
[3] 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.
[4] 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.
[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.