A Survey on Multimodal Summarization using Deep Learning

K. P. Moholkar[1], Isha Patil[2], Rutuja More[3], Snehal Deore[4], Dhanashri Bhise[5]
JSPM'S Rajarshi Shahu College of Engineering, Pune

(kavita.moholkar@gmail.com[1], ishapatil68@gmail.com[2], rutujamore1706@gmail.com[3], snehideorekrish@gmail.com[4], dhanashribhise26@gmail.com[5])

Abstract

In the present situation of fast-growing consumption of the Internet, searching for the required information becomes monotonous and sluggish task. Apart from finding relevant content from given results there is another very difficult task for a user, which is to manually summarize the large multimodal data. Now-a- days there is a rapid growth of multimodal data. Summarization is the process of truncating the enormous amount of data available in various formats into summarized documents, conserving the most important points and the overall meaning of the document. The process of summarization is divided into extractive summarization and abstractive summarization. Extraction is the process which concatenates sentences from the document depending on their importance, whereas abstraction involves generating novel sentences from information extracted from corpus. In this survey paper we investigate the popular and important work done in the field of multimodal summarization, various summarization methods that can create a compacted version of a set of documents and audios related to a specific topic without any human control.

Keywords: *Text Summarization, RNN, Encoder-Decoder, Natural Language Processing, Multimodal data.*

1. Introduction

Automatic summarization rebuild the large documents into truncated form which could be challenging and costly to undertake otherwise. Machine learning algorithms are used to encompass documents and find out important information before generating the required summarized data. Mainly, Summarization can be done with two different methods which are extraction and abstraction. The method which selects phrases and sentences from the original document as it is without changing its meaning to generate the new summarized version is Extractive summarization. This method involves grading the phrases according to their importance so that only important data is selected in the summary. The other method generates altogether new data to retain the nuance of the original document. This method is similar to what humans use while generating summary manually. RNN based sequence-to-sequence methods have confirmed very important in abstractive methods of summarization. In our paper, we are using abstractive summarization with dual encoder and a single decoder. Unlike the previous work in this field which use only a single encoder, our method uses a set of two encoders, the primary and the secondary encoder.

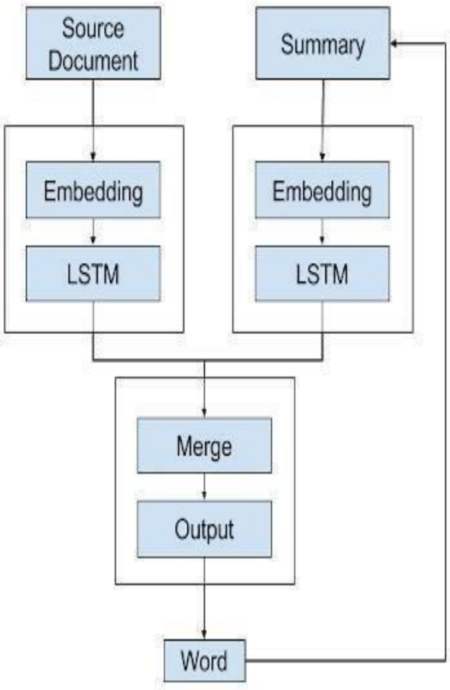

Figure 1 - Process Flow of Summarization

2. Literature Survey

Furui et al [1] evaluates results for various techniques for compaction- based automatic speech summarization. This includes two stages such as salience feature extraction and word based sentence densification. Research on speech summarization helps in improving QA systems using speech documents. By densifying salient features of long presentations and lectures can provide user with valuable expedient for absorbing valuable information in a much lesser time. This paper examines speech summarization techniques with the two presentation methods in unbounded domains, one stage model and two stage model with sentence extraction and sentence compaction. The key drawback of this mode is the errors and disfluencies generated during recognition. It has an average accuracy 45.5%.

Yogita et al in [2] presents a speech to text conversion system which can work with more than one language. Speech-To-Text (STT) is the system which takes various expressions by humans such as their conversations as an input to extract important information about the input data. Various methods such as Mel-Frequency Cepstral Coefficient (MFCC) along with dimensionality reduction technique and Minimum Distance Classifier, also methods that perform classification by finding the examples that maximizes the gap between two classes are used in this system for speech classification. Continuous pieces of speech beginning and ending with a clear pauses are recorded in advance and stored in a database. Training and testing are the two sections of the database. Training phase gets its input from training section of the database. In the training phase the features are extracted from the samples available. The features for each sample are combined to generate feature vector. System is provided with the samples for testing which are further passed to generate feature vectors. These features and reference feature vector are computed for resemblance and vectors with maximum resemblance are given as output. This SVM and MFCC techniques gives accuracy about 65%.

Haoran Li et al [3] articulated a multimodal summarization task as the problem of finding the best solution from all feasible solutions. In this paper, the author has stated the technique to handle text and audio inputs. This model addresses generation of readable contexts by using audio transcripts through guidance strategies. This model is based on graphs which effectively calculates the salience score for each input to generate more accurate summaries. The paper investigates different techniques for establishing significant connection between image and text, to add value to the performance of the model. The accuracy of this model decreases for audio and video inputs whereas the system gives better performance for textual data.

Kaichun Yao et al [4] presents an extended sequence-to- sequence framework for text summarization. This is a dual encoding model which involves encoder and decoder model along with the attention mechanism. Unlike the traditional encoder-decoder model, this model considers the whole sequence for decoding, generating a fixed length output sequence at every stage. The main drawback of traditional encoder decoder models is that they generates summaries with frequent phrases or words. This model overcomes these drawbacks by introducing Pointer generator concept, which is used to set the rear and OOV (Out Of Vocabulary) problem. It uses Daily Mail dataset and DUC 2004.

Sheetal Shimpikar et al [5] has discussed different methods of summarization of textual data for various Indian languages. Various Technologies that consider variables to make a logical and consistent summary are discussed in this paper. Summarization mainly focuses on finding a prototypical batch of the data containing the information of the complete dataset. The systematic observation of the similarities or dissimilarities between two or more text summarization techniques are considered in this paper. These techniques are mainly used for Indian regional languages. Various methods like indexing and feature selection are discussed in this paper.

Gabriel Silva et al in [6] has presented a model which employs extractive summarization technique to generate short summaries. The model uses machine learning concepts for its implementation. The extractive summarization technique includes extracting the salient features from large text documents based on qualitative and quantitative analysis. Then these sentences are concatenated to form a meaningful summary. The model is trained on CNN-corpus dataset. The main drawback of this model is that it includes redundancy of words and phrases.

Nikhil S. Shirwandkar et al [13] has introduced an extractive text summarization model. This model uses a combination of Restricted Boltzmann Machine and Fuzzy Logic. It extracts salient sentences from the document and concatenate them to form a short summary. It uses Restricted Boltzmann Machine as unsupervised algorithm. Important sentences are extracted from the document using both these approaches separately and then they are condensed and processed to generate summary. The model is designed for English language. It overcomes the problem of text overloading. The drawback of this model is that it works only for a single document. It has an average 84% accuracy.

Table 1 - Summary of Literature Survey

or	**ıods**	**set**	**ıntages**
oki, Furui, Tomonori	nce extraction od	M35, M31	nmarization ratio is smaller , lts are effective.
an Li, Junnan Zhu, CongMa,	n based LexRank , system	ws events	narization Process without speech criptions, the text and audio, guide l. Performs better than the audio l
hun Yao, Libo Zhang ,	der- Decoder method	/ DailyMail -2004	er generator is used to deal with nd OOV word
ta H.Ghadage, Sushama elke	, MFCC	-2002	ks with English Marathi mix data
ta Moholkar, Dr. S. H. Patil	l2Vec		ks on only text data

3. Conclusion

Multimodal summarization is drawing a lot of attention these days due to increasing multimedia data. Going through all this lengthy data is very time consuming, people often are more interested in key points from the original volume of data. This survey focuses on various models which employs various techniques for multimodal summarization. But multimodal data also includes real time spontaneous data. A lot of research is going on revolving around multimodal input and output data.The proposed model in this paper employs deep learning techniques in order to process spontaneous data and generate summaries with maximum accuracy.

4. References

[1] Sadaoki Furui , Tomonori Kikuchi “Speech-to-text and Speech-to-speech Summarization of spontaneous speech”.

[2] Yogita H. Ghadage, Sushama D. Shelke ,”Speech to Text Conversion for Multilingual Languages”.

[3] Haoran Li, Junnan Zhu, Cong Ma,etal ,”Read, Watch, Listen and Summarize: Multi-modal Summarization for Asynchronous Text, Image, Audio and Video.”

[4] Kaichun Yao, Libo Zhang , etal ,”Dual Encoding for Abstractive Text Summarization.”.

[5] Sheetal Shimpikar, Sharvari Govilkar, etal.”A Survey of Text Summarization Techniques for Indian Regional Languages. “.

[6] Gabriel Silva, Rafael Lins, Luciano,etal “Automatic Text Document Summarization Based on Machine Learning.”.

[7] Kavita Moholkar, Suhas Patil , “Hybrid CNN-LSTM Model for Answer Identification”International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277- 3878, Volume-8 Issue-3, September 2019.

[8] Litvak M. and Last M.: “Graph-based keyword extraction for single-document summarization. In Proceedings of the workshop on Multi-sourceMultilingual Information Extraction and Summarization”, pp. 17{24, ACL (2008).

[9] Michael J. Giarlo. A comparative analysis of keyword extraction techniques. Rutgers,The State University of New Jersey and Chengzhi Zhang, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, Bo Wang.

[10] Renjith SR and Sony P, “Automatic text summarization for Malayalam using sentence extraction”. Proceedings of 27th IRF International Conference, 14th June 2015, Chennai, India, ISBN: 978-93-85465-35-2.

[11] Harabagin and Lacatusu, “Generating single and multi-document summaries with gistexter”, in document understanding conference, 2002.

[12] Lee and Jian, “Ontology based method, text representation on Fuzzy ontology and content selection isclassifier”, Systems, Man and Cybernetics, Part B: Cybernetics, IEEE Transaction on, vol.35, pp. 859-880, 2005.

[13] Nikhil S. Shirwandkar, Dr. Samidha Kulkarni, “Extractive Text Summarization using Deep Learning”, 2018 IEEE.

AUTHORS PROFILE

Prof. K.P. Moholkar is currently working as Assistant Professor in Department of Computer Engineering, JSP hu College of Engineering, Pune. She is currently pursuing Ph.D. from Bharti Vidyapeeth Dee sity and She has published more than 27 research papers in reputed international journals. Her rk focuses on Data Mining, Artificial Intelligence, Learning and Deep Learning. She has 18 y xperience.

Ms. Isha Patil, BE Computer Engineering, JSPM’s Rajarshi Shahu College of Engineering, Pune.

ISSN: 2233-7857 IJFGCN

Ms. Rutuja More, BE Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Pune

Ms. Snehal Deore, BE Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Pune

Ms. Dhanashri Bhise , BE Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Pune