EasyChair Preprint

№ 2881

Text Summarization Framework Using Advanced Machine Learning Algorithms

Pallavi Kohakade and Sujata Jadhav

March 6, 2020

Text Summarization Framework Using Advanced Machine Learning Algorithms

Miss.Kohakade Pallavi Suresh
Department of Computer Engineering
Vishwabharti Academy College Of Engineering
Savitribai Phule Pune University
Ahmednagar, India
Kohakadepallavi1893@gmail.com

Prof.Jadhav Sujata A.
Department of Computer Engineering
Vishwabharti Academy College of Engineering
Savitribai Phule Pune University
Ahmednagar, India
jadhav.suj14@gmail.com

***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 quick development in media data transmission over the Internet requests content outline utilizing neural system from nonconcurrent blend of content. This paper speaks to a structure that uses the methods of NLP strategy to analyze the elaborative data contained in multi-modular insights and to improve the parts of content rundown. The essential idea is to connect the semantic holes among content substance. After, the created outline for significant data through multi-modular subject demonstrating. At long last, all the multi-modular components are considered to create a literary outline by expanding the significance, non-excess, believability and degree through the assigned collection of submodular highlights. The exploratory outcome shows that Text Summarization system outflanks other serious strategies.**

***Index Terms*—Summarization,Feature selection,Machine Learning,Sentence Embedding**

I. Introduction

Now a days, there are large numbers of documents or information that is present related to any particular field[1][3]. There are many sources out of which we can gather a lot of information that will be pertinent to our field of search. Much information is available at various sources like the internet. But, as we know that a huge amount of information cannot be always considered or taken into use. So, a precise amount of information is always considered and that information is drawn out from the original document that is huge in size. In other words, we can say that we pluck out the summary of the main document. A summary of any document is defined as a collection of essential data by collecting the brief statements accounting the main points of the original document. Along these lines, Summarization of a book is a methodology of isolating or getting the pertinent information out of an exceptionally enormous document[5]. It is the way toward shortening the content archive by utilizing different innovations and systems to make an intelligible rundown including the significant purposes of the first record. There are different strategies by which the synopsis procedure can be completed.

While summarization frameworks center around just common language preparing (NLP), the chance to mutually improve the nature of the synopsis with the guide of programmed discourse acknowledgment (ASR) and PC vision (CV) handling frameworks is broadly overlooked. Then again, given a news occasion (i.e., news subject), Text information are commonly offbeat in genuine life[7][8]. In this way, Text outline faces a significant test in understanding the semantics of data. Right now, present a framework that can furnish clients with literary synopses to assist with procuring the substance of nonconcurrent information in a brief timeframe without perusing records from start to finish. The motivation behind this work is to join the NLP with AI strategies to investigate another system for mining the rich data contained in multi-modular information to improve the nature of Text outline[9].

II. REVIEW OF LITERATURE

P. Sinha, S. Mehrotra, and R. Jain[1]:Proposed techniques to process quality, decent variety and inclusion properties utilizing multidimensional substance and setting information. The proposed measurements which will assess the photograph rundowns dependent on their portrayal of the bigger corpus and the capacity to fulfill client's data needs. Favorable circumstances are: The ravenous calculation for outline performs superior to the baselines. Outlines help in compelling sharing and perusing of the individual photographs. Hindrances are: Computation is costly.

H. Lin and J. Bilmes[2]:In multi-document summarization, excess is an especially significant issue since printed units from various records may pass on a similar data. A top notch (little and important) rundown ought not exclusively be educational about the rest of likewise be smaller (non-repetitive). Favorable circumstances are: The best execution is accomplished. Submodular synopsis accomplishes better ROUGE-1 scores. Impediments are: The proposed framework pricey to tackle.

M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi[3]: Eddi is a novel interface for perusing Twitter streams that bunches tweets by points drifting inside the client's own channel. A calculation for theme discovery and a subject situated UI for social data streams, for example, Twitter channels. (1) benchmark TweeTopic against other subject location approaches, and (2) contrast Eddi with a run of the mill ordered interface for devouring Twitter channels. Points of interest are: A straightforward, novel subject recognition calculation that utilizes thing phrase identification and a web index as an outside information base. Eddi is more agreeable and more effective to peruse than the conventional sequential Twitter interface. Burdens are: Users approached our customers temporarily, making it hard to extrapolate ends on how the apparatus may be utilized longitudinally. Clients were seeing the historical backdrop of their feed as opposed to tweets they had never observed, making our undertaking somewhat less practical.

P. Goyal, L. Behera, and T. M. Mcginnity[4]:Proposes theovel interface for perusing Twitter streams that groups tweets by points Proposes the clever thought of utilizing the setting touchy record ordering to improve the sentence extraction-based report outline task. Right now, a setting delicate record ordering model dependent on the Bernoulli model of haphazardness. Focal points are: The new setting based word ordering gives preferred execution over the pattern models. Weaknesses are: Need to figure the lexical relationship over a huge corpus.

D. Chakrabarti and K. Punera, "Event summarization using tweets[5]:
Right now contend that for some exceptionally organized and repeating occasions, for example, sports, it is smarter to utilize progressively refined procedures to outline the pertinent tweets. The issue of outlining occasion tweets and give an answer dependent on learning the fundamental shrouded state portrayal of the occasion through Hidden Markov Models. Focal points are: The benefit of utilizing existing inquiry coordinating innovations and for basic one-shot occasions, for example, quakes it functions admirably. The HMM can learn contrasts in language models of sub-occasions totally naturally. Drawbacks are: The hindrance that SUMMHMM needs to represent tweet words that just happen in a portion of the occasions, however not in others.

Z. Li, J. Liu, J. Tang, and H. Lu[6]:In paper, proposes a particular Robust Structured Subspace Learning (RSSL) calculation with the guide of coordinating picture information and capacity picking up information on into a joint contemplating system. The scholarly subspace is went with as a middle of the road zone to diminish the semantic empty between the low-degree seen capacities and the high-arrange semantics. Points of interest are: The proposed RSSL empowers to successfully examine a vigorous based subspace from records. The proposed structure can decrease the clamor provoked vulnerability.

W. Y. Wang, Y. Mehdad, D. R. Radev, and A. Stent[7]:The paper proposes a particular grid factorization method for extractive rundown, utilizing the accomplishment of community oriented separating. First to consider outline learning of a joint installing for printed substance and depictions in timetable synopsis. Focal points are: It is direct for developers to set up the gadget in true bundles. Versatile technique for contemplating low-dimensional inserting's of data stories and previews. Drawbacks are: Only work on condensing synchronous multi-modular substance.

III. PROPOSED METHODOLOGY

Firstly, the file which is given as info is tokenized so as to get tokens of the terms.The prevent words are expelled from the content after tokenization. The words which are remained are considered as a key word.The catchphrases are taken as a contribution for that we are joining a piece of tag to each key word.After finishing this pre-preparing step we are ascertaining recurrence of every watchword like how habitually that catchphrase has happened from this most extreme recurrence of the catchphrase is taken.Now weighted recurrence of the word is determined by partitioning recurrence of the watchwords by greatest recurrence of the key words.In this progression we are computing the total of weighted frequencies utilizing cosine similarity.then we use LDA and Generate summary.

A. Architecture

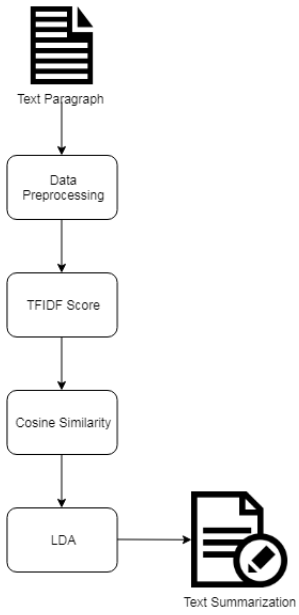

Fig. 1. Proposed System Architecture

B. Advantages

1) It provides to automatically mine and summarize subtopics (i.e., divisions of a main topic) from large paragraph related to a given topic.
2) Document contents can facilitate subtopic discovery.
3) Well organizing the messy documents into structured subtopics.
4) Generating high quality textual summary at subtopic level.

C. Algorithms

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).
Step 8 - The output list is then returned summary.

D. Mathematical Model

The mathematical model for Text Summarization System is as-

$$S = \{I, F, O\}$$

Where,
I = Set of inputs
The input consists of set of Text.
F = Set of functions

$$F = \{F1, F2, F3 \ldots .FN\}$$

F1:Data Extraction
we use text document dataset. **F2:Preprocessing**
Tokenization-This technique removes Special character and images.

Initialize feature vector bg feature =[0,0..0] for token in text.tokenize() do
if token in dict then
token idx=getindex(dict,token)
bg feature[token idx]++
else
continue
end if
end for

F3: Feature Extraction
we use word embedding sentence generator for feature extraction. **F4: Generate Summary**
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).
O:Short Summary Generation

IV. Results and Discussion

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.
Some of the parameters are considered for OCR as well as ASR for text conversion methods.
Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.
Precision=TP/TP+FP
Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class.
Recall=TP/TP+FN
F-measure - F-measure is the weighted average of Precision and Recall.
F-measure = 2*(Precision*Recall) / (Precision+Recall)
Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.
Accuracy = TP+TN/TP+FP+FN+TN

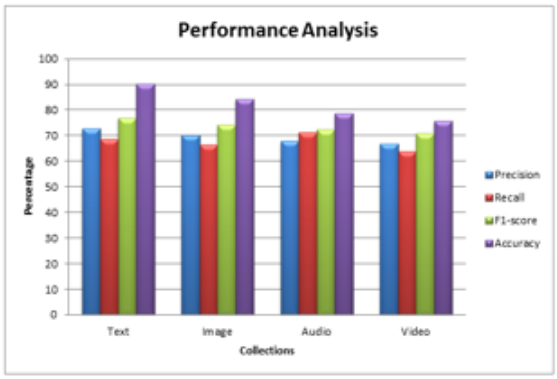

Fig. 2. Performance graph

V. Conclusion

Automatic text summarization is a complex task which contains many sub-tasks in it. Every subtask has an ability to get good quality summaries. The important part in extractive text summarization is identifying necessary paragraphs from the given document. In this work we proposed extractive based text summarization by using statistical novel approach based on the sentences ranking the sentences are selected by the summarizer. The sentences which are extracted are produced as a summarized text and it is converted into audio form. The proposed model improves the accuracy when compared traditional approach.

Future Scope: Furthermore, we intend to continue to explore new problems from the point of view of a summarization system, such as image,audio and video.

References

[1] Cheng, Jianpeng, and Mirella Lapata. "Neural summarization by extracting sentences and words."arXiv preprint arXiv:1603.07252 (2016).

[2] Nallapati, Ramesh, et al. "Abstractive text summarization using sequenceto- sequence rnns and beyond." arXiv preprint arXiv:1602.06023 (2016).

[3] Freitas, N., and A. Kaestner. "Automatic text summarization using a machine learning approach." Brazilian Symposium on Artificial Intelligence (SBIA), Brazil. 2005.

[4] Ferreira, Rafael, et al. "Assessing sentence scoring techniques for extractive text summarization." Expert systems with applications 40.14 (2013): 5755-5764.

[5] Gaikwad, Deepali K., and C. Namrata Mahender. "A Review Paper on Text Summarization." International Journal of Advanced Research in Computer and Communication Engineering 5.3 (2016).

[6] Fachrurrozi, M., Novi Yusliani, and Rizky Utami Yoanita. "Frequent Term based Text Summarization for Bahasa Indonesia." (2013): 30-32.

[7] Radev, Dragomir R., et al. "Centroid-based summarization of multiple documents." Information Processing Management40.6 (2004): 919-938.

[8] P. Sinha, S. Mehrotra, and R. Jain, "Summarization of personal photologs using multidimensional content and context," in Proc. 1st ACM Int. Conf. Multimedia Retrieval, 2011, p. 4.

[9] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in Proc. Human Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2010, pp. 912–920.

[10] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi, "Eddi: Interactive topic-based browsing of social status streams," in Proc. 23nd Annu. ACM Symp. User Interface Softw. Technol., 2010, pp. 303–312.

[11] 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.

[12] D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 66–73.