End-End Summarization Tool using NLP

Prayag Gavshinde[1], Raghav Agrawal[2], Rishi Somani [3] Rishabh Rathore[4] Sambhav Jain [5] Prof. Kapil Sahu[6]

* Information Technology, Acropolis Institute of Technology and Research, RGPV, Indore, India*5Associate Professor, Department of Information Technology, Acropolis Institute of Technology and Research Indore, India

e-mail: prayaggavshinde@gmail.com[1], raghavagarwal019@gmail.com[2], somanirishi81@gmail.com[3]
rishabhrathore474@gmail.com[4],sambhavjain896@gmail.com[5],kapilsahu@acropolis.in[6]

Abstract

The size of data on the Internet has risen in an exponential manner over the past decade. Thus, the need for a solution emerges, that transforms this vast raw information into useful information which a human can easily grasp. One such common technique in research that helps in dealing with enormous data is text summarization. Automatic summarization is a renowned approach that is used to reduce a document to its main ideas. It operates by preserving substantial information by creating a shortened version of the text. Usually there are two methods for text summarization as Abstractive and Extractive. Extractive methods are simple which extract the top meaningful sentence from the document as summary while Abstractive Methods explain the document by modifying some sentences in meaningful grammar. Although there are a ton of methods, researchers specializing in Natural Language Processing (NLP) are particularly drawn to extractive methods. A deep study of Abstractive Approach for text summarization, Pretrained models for Image summarization has been made in this paper. This paper also analyses the above-mentioned methods which yield a less repetitive and more concentrated summary.

Keywords— Text Generation, NLP, Machine Learning, Deep Learning

I. INTRODUCTION

A summary is a short paragraph that explains the complete input document. The main motive behind automatic text summarization is to find a short subset of the most essential information from the entire set of data and present it in a human-readable format. As there is a huge amount of data present on the web and we need to learn many things from it, the problem arises in content where automatic text summarization methods prove to be very helpful in extracting meaningful information that can be read in a short period of time. The method of extracting these summaries from the huge amount of text without losing vital information is called **Text Summarization**. The most important advantage of using a summary is, it reduces the reading time. Abstraction involves paraphrasing sections of the source document. In general, abstraction can produce summaries that are more condensed than extraction. Both techniques exploit the use of natural language processing and/or statistical methods for generating summaries. And, the classical approaches to text summarization proposed by Luhn et al have established the basis for the discipline of text summarization techniques

Our goal is to extend this applicability to the meeting domains to produce high-quality meeting summaries. The Goal of the tool is to provide short and sweet summary that the user can read, listen, and download at his own pace. To accomplish our task at hand requires a text summarization tool. The goal of this report is to capture the product evaluation process in 4 distinct phases:

1. Preparation
2. Criteria establishment
3. Characterization, and
4. Testing

II. EXISTING SYSTEM

The below table contains some websites that provide the service for text summarization using different methods to meet specific requirements.

No.	Existing system/website/software	Features	Disadvantages	Limitations/ Gaps
1.	TLDRthis (website)	Summarizing any article through URL or pasting the text	Text is allowed only in the form of copy/paste. No uploads can be done	It cannot summarize pdfs. only applicable for articles or blogs through links.
2.	Resoomer(Website)	Summarize text for everyone By copy-paste content. Options for downloading generated summary in pdf of any doctype.	summarizing documents by uploading them from the local system is not possible.	cannot summarize pdfs. only applicable for text content.
3	IBM Image Caption Generator	It is a web app to provides short and quick explanations of images known as Image captioning.	They do not make intuitive feature observations on objects or actions in the image.	The only work is to describe the Image in one sentence. We cannot download image with caption.

III. PROPOSED SYSTEM

The solution we proposed is well suited for extracting summaries of each kind on a single interface where the user enjoys all the benefits along with a voice clone feature to listen as well as read.

The proposed system as shown in figure 1 uses a T5 transformer to summarize documents from the user. The user inputs a document to the user interface to get a summary(denoted by the dashed box) which has classes derived from the NLP libraries. These classes are a collection of semantic rules (which allows the system to group the content using world knowledge) and dictionaries, which aid in the semantic analysis and SVD phases in the summarizer. The input document is first parsed or pre-processed, wherein there is a removal of unneeded words such as ‘stop words’ which are simply small function words, like “the”, “and”, “a”, which do not contribute meaning to the text summary. The next stage is the generation of a Singular Value Decomposition(SVD) matrix or document term matrix which is a m x n matrix, where m is the total number of terms in the original text and n is the number of sentences in the original text. The SVD Analysis stage derives the latent semantic structure from the document represented by matrix A. Finally, in the summarization process, the system arranges the sentences generated from the SVD Analysis stage by semantically placing them in a way that the summary encompasses all the concepts of the original text. The final summary is then given back to the user

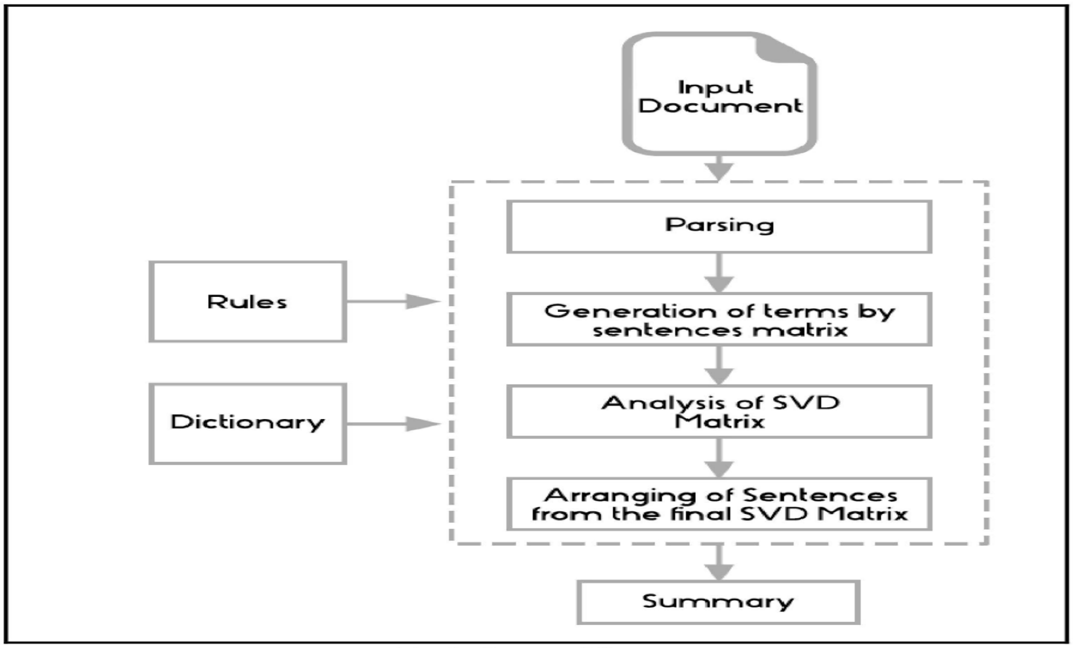

Input Document
Parsing
Rules
Generation of terms by sentences matrix
Dictionary
Analysis of SVD Matrix
Arranging of Sentences from the final SVD Matrix
Summary

Fig.1: Proposed System

The major functionality that the project enjoys are

- **Document Summarization**: Any pdf, doc, or text files like research papers, chapter notes, or anything if you upload it will find the best sentences with the highest dominating score and generate a short and sweet summary. [7]
- **The article, Blog summary through URL**: If you want to read a summary of any online article available on any website, just provide a link. You will get important details like author, Published date, and article summary.
- **Topic Modelling**: Topic modelling is something which enables you to take your reading of a particular article forward because it lets you know the topics which particular article is dependent on. Think when you have to analyze or go through thousands of documents and categorize under 10 – 15 buckets. How tedious and boring will it be? Thanks to Topic Modeling where instead of manually going through numerous documents, with the help of Natural Language Processing and Text Mining, each document can be categorized under a certain topic.
- **Image Captioning**: Sometimes It feels tedious to describe any Image. We will integrate a model which will describe and provide a perfect description of an Image.

IV. Implementation And Results

Data Collection

Flickr 8K dataset is the most popular Image captioning dataset which we collected from Kaggle. Kaggle is a popular data science learning platform where you can participate in various data science competitions hosted by organizations. You can learn from others' code as well as contribute your code which helps the open-source community to grow and learn more. The Flickr 8K dataset is used to train the Image captioning model which contains 8000 images and a separate file describing captions of each image.

Topic Modeling and Text summarization are achieved using Transformer Models so various documents, articles, text paragraphs were picked from different websites on google to validate and test for correct parameters to pass to Transformer. [10]

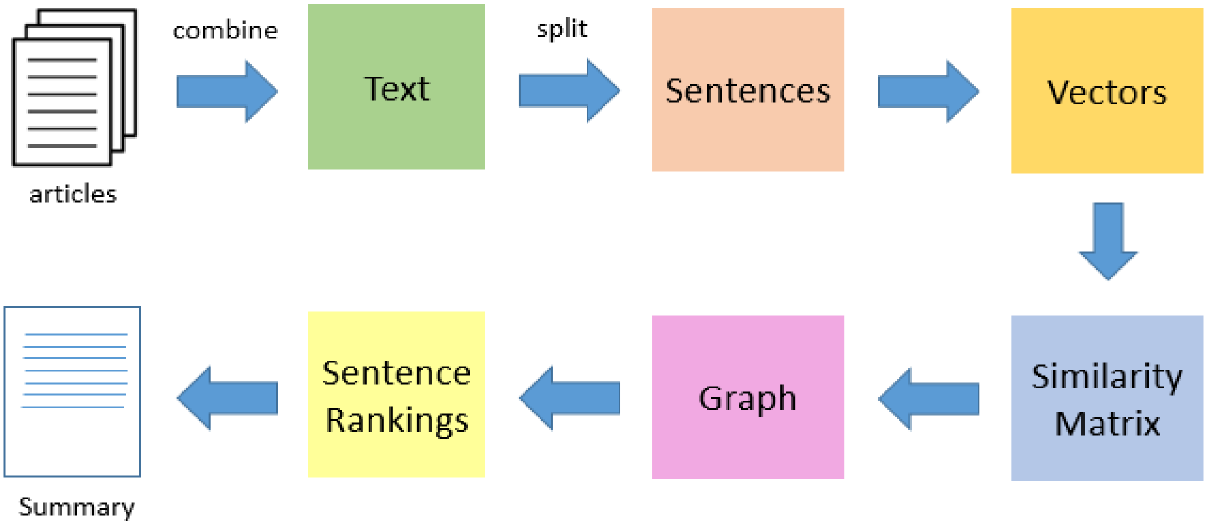


Figure 2:

Method Used

Abstraction Based summarization using T5 Transformer

T5 stands for Text-to-Text Transformer, which is **a Transformer based architecture that uses a text-to-text approach**. It helps in performing many NLP applications including language translation, question-answering, text classification, and summarization in none other than these applications. **[4]** In abstraction-based summarization, advanced deep learning techniques are applied to paraphrase and shorten the original document, just like humans do. ***T5*** gives state-of-the-art results in many NLP tasks but also has a very radical approach to NLP tasks. T5 uses common crawl web extracted text. When we feed any text to the T5 transformer we have to add a prefix as a summary to let it know the function to perform. T5 transformer is trained in such a way that it preprocesses the data before summarization like it removes punctuation, and also removes any HTMl tags or javascript typing (since it often appears in code). It deduplicates the dataset by taking a sliding window of 3 sentence chunks and deduplicates it so that only one of them appears in the dataset.T5 is a new transformer model from Google that is trained in an end-to-end manner with text as input and modified text as output. [2]

LDA(Latent Dirichlet allocation) For Topic Modelling

LDA is a form of unsupervised learning that views documents as bags of words. LDA is the same as PCA and PCA works for numerical values. PCA decomposes the input in a smaller dimension. LDA works on text data. LDA works by first making a key assumption: the way a document was generated was by picking a set of topics and then for each topic picking a set of words.[1]

1. Assume there are *k* topics across all of the documents
2. Distribute these *k* topics across document *m* (this distribution is known as α and can be symmetric or asymmetric, more on this later) by assigning each word a topic.

3. For each word *w* in document *m*, assume its topic is wrong but every other word is assigned the correct topic.
4. Probabilistically assign word *w* a topic based on two things:
   – what topics are in document *m*
   – How many times *has* word *w been* assigned a particular topic across all of the documents (this distribution is called *β*, more on this later) and repeat the Process till you are done with the document.

The Model

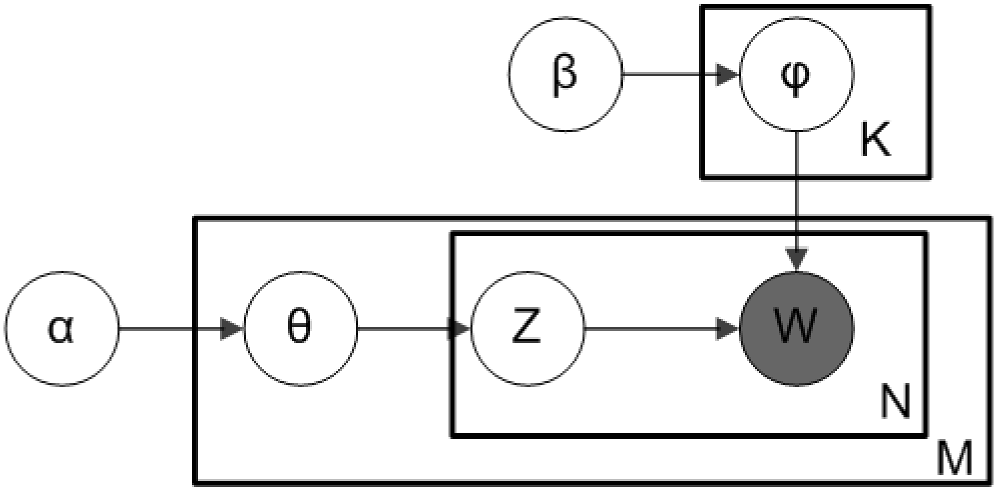


Figure 3:

α is the per-document topic distribution,
β is the per-topic word distribution,
θ is the topic distribution for document *m,*
φ is the word distribution for topic *k,*
z is the topic for the *n*-th word in document *m*, and
w is the specific wor

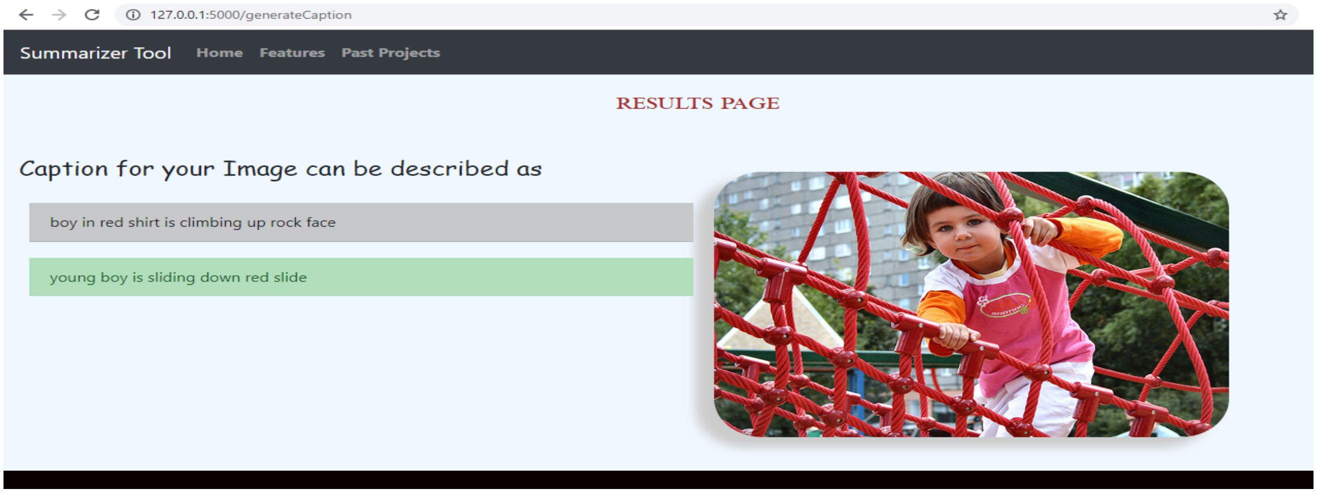
Images Captioning Using Autoencoders(CNN-RNN)

VGG- Network is a convolutional neural network model proposed by K. Simonyan and A. Zisserman in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition". This architecture achieved top-5 test accuracy of 92.7% in ImageNet, which has over 14 million images belonging to 1000 classes.

VGG is used as a Feature Extractor vector in image captioning and feature encoder part which is further used to decode the caption generation part. LSTM based RNN is very efficient for a sequence-based generation.

Testing

Input Article	**Original Summary (Title)**	**Generated Summary (Title)**
File names are appearing on the attachment details page twice and blank file names are observed. PFA screenshot	File names are appearing on the attachment details page twice also blank file names observed	blank file names are appearing on the attachment details page twice. Also, blank file names are observed on the attachment details page.
M160 FRUs Hot-swappable Routing Engine Switching and Forward… Requires Power Down Circuit Breaker Box Connector Interface Panel CIP M160 Chassis includes backplane Please refer to the M160 hardware guide for more detail.	[Archive] What are the specific Field Replaceable Units (FRU) on the M160?	M160 Chassis includes a backplane. requires a power-down circuit breaker box.

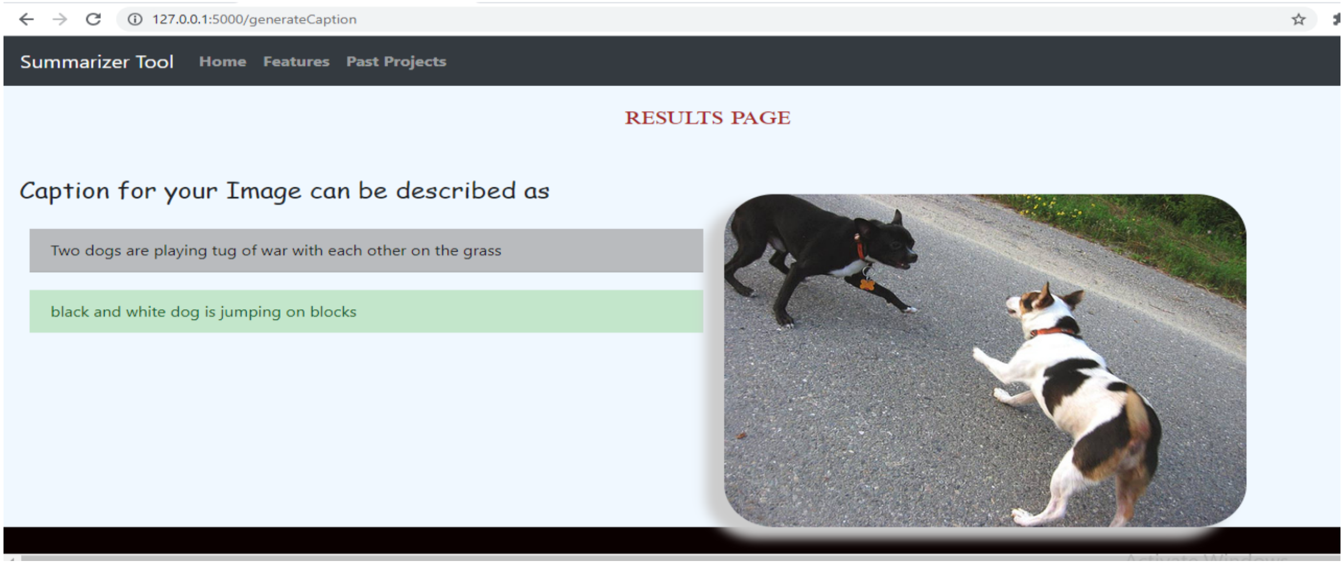


RESULTS PAGE

Caption for your Image can be described as

boy in red shirt is climbing up rock face

young boy is sliding down red slide



RESULTS PAGE

Caption for your Image can be described as

Two dogs are playing tug of war with each other on the grass

black and white dog is jumping on blocks

Document	Extracted Topics
Health experts say that Sugar is not good for your lifestyle.	Health, Sugar, Bad
My father spends a lot of time driving my sister around to dance practice	Drive, Sister, Dance, Practice

V. Conclusion and Future Work:

The automatic Text Summarization method depends on the semantic data of the concentration in a document. So this way, gathered parameters like approaches, spots of different substances are not considered. In this recommendation, Lesk means word sense disambiguation by utilizing the vocabulary definitions to the electronic dictionary information based on utilizing wordnet. This goal is clear from covering sentence, a couple of fusing words that give the setting of the word, in this not utilizing the late using the definitional shines of those words, other than those of words related to them through with the unmistakable relations portrayed in wordnet. So furthermore we are endeavoring to use another enlightening record away by wordnet for each word. For example, design sentences and identical words et cetera.

Among future work is the use of all the more balanced gathering to upgrade. Attempting diverse things with more tongue-specific segments, for instance, morphological parsers, printed entailment, and anaphoric assurance is open research for more updates later on. Programmed content summarization should be possible for various archives. A client can be given an office to print the record from the interface specifically. A point of confinement to re-synopsis alternative perhaps included for record Shorter long. Additional line holes acquired in the outline can be evacuated. Spare as a choice can be added to a tedious ever application for the client to spare the synopsis in various arrangements.

References

[1] Jagadeesh J, Prasad Pingali, Vasudeva Varma, “Sentence Extraction based single Document Summarization”, Workshop on Document Summarization, 19th and 20th March 2005, IIIT Allahabad.
[2] Arman Kiani B, M. R. Akbarzadeh ―Automatic Text Summarization Using: Hybrid Fuzzy GA-GP‖, IEEE International Conference on Fuzzy Systems. July 16-21, 2006.
[3] M. Abramowitz and I. Stegun, editors. Handbook of Mathematical Functions. Dover, New York, 1970.
[4] Ronning. Maximum likelihood estimation of Dirichlet distributions. Journal of Statistical Computation and Simulation, 34(4):215–221, 1989.
[5] H.P. Edmundson, “New Methods in Automatic Extracting”, Journal of the Association for Computing Machinery, April 1969.
[6] A.Das, M.Marko, A.Probst, M.A.Portal, C.Gershenson ―Neural Net Model For Featured Word Extraction‖, 2002.
[7] Goncalves Luís, *Automatic Text Summarization with Machine Learning - An overview.*,
[8] S. Doha, V. Gupta and H. Kamick, "Unsupervised semantic abstractive summarization", *Proceedings of ACL 2018 Student Research Workshop*, pp. 74-83, July 2018.
[9] Martin Ponweiser (2012) Latent Dirichlet Allocation in R, Vienna University of Business and
Economics.
[10] Bettina Grun, Kurt Hornik (2011) “topic models: An R Package for Fitting Topic Model”, Journal of
Statistical Software Vol. 40, No. 13.