JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AUTOMATIC TEXT SUMMARIZATION USING DEEP LEARNING AND NLP MODEL

Chaluvadi Abhishek[1], Mula Anvesh Reddy[1], Madishetti Raviteja[1], Sontenam Sai Sampath Ashish[1], Smt. M. Venkata Ramana[2]

Students, Department of Computer Science Engineering, GITAM Deemed to be University, Visakhapatnam, India-530045[1]
Assistant Professor, Department of Computer Science Engineering, GITAM Deemed to be University, Visakhapatnam, India[2]

Abstract: In this modern era, with so much information on the Internet, it's vital to provide a fast and efficient method of extracting data. It is difficult for people to manually retrieving the summary of a large written document. You may find a large range of written content on the Internet. As a result, finding relevant documents among the large number of documents available, let alone extracting important information from it, is difficult.
In order to address the two concerns mentioned above, automatic text summarizing is required.
Text summarizing is the process of picking the most important and meaningful information from a document or group of texts and summarizing it into a simpler form while keeping the general meanings.

Keywords: **Text Summarization, Nlp, Extractive Summary, Abstractive Summary, and Deep Learning are some of the terms used in this paper.**

I. INTRODUCTION

As the internet and big data have increased in popularity, people have gotten overwhelmed by the massive amount of information and documents available on the internet. As a result, many academics are encouraged to provide a technology solution that can automatically summarize texts. Summaries created by automatic text summarization include all critical information from the original material as well as key sentences. As a result, the information is delivered quickly while still meeting the document's original goal. Text summarization has been studied since the mid-twentieth century, with LUN (1958) being the first to use a statistical approach called word frequency diagrams to openly discuss it. Many other approaches have been developed to date. Depending on the amount of documents, there are single and multi-document summary choices. In the meanwhile, depending on the type of process summarization follows it has 2 methods abstractive and extractive text summarization. Summarization systems usually have access to additional evidence in order to find the most important document themes. When summarizing blogs, for example, arguments or comments that follow the blog post might be helpful in evaluating which parts of the blog are crucial and intriguing. Scientific article summaries contain a substantial amount of information, such as referenced papers and conference information, that can be utilised to emphasise key sentences in the original study. The sections that follow go over some of the contexts in more detail

Abstarctive text summarization is the process of extracting new context from the entire document which doesn't actually present in the input document.

Extractive text summarization make use of input sentences or document inorder to produce the output.

II. LITERATURE SURVEY

In [1], JINGWEI CHENG implemented “Automatic text summarising technology” has shown to be a useful tool for dealing with information overload. Automatic text summarising model that uses a template encoder and a title of the article decoder to extend the classic Seq2Seq neural summarization model. In the sentence embedding, the encoder encodes both the syntactic structure and the word data of a sentence. To pay attention to syntactic units, a hierarchical attention method is developed. To produce a greater quality of generated summaries, the decoder is modified with a top attention mechanism with dual LSTM network. On the CNN/DM datasets, we performed experiments to evaluate the suggested technique to baseline models. The findings of the experiment reveal that the suggested method outperforms abstractive benchmark models in term of ROUGE evaluation criteria and achieves summarize generating comparable performance to the extraction baseline method.

In [2], Pratibha Devihosur and Naseer R recommended "The Automatic Text Summarizing "uses an unsupervised learning system to complete the summarization task. Simplified Lesk calculation is used to determine the importance of a sentence in information content. WordNet is used as an internet semantic lexicon. Word - based Annotation is a crucial and rigorous approach for dealing with distinctive dialects. In different contexts, a term may have a different meaning. As a result, the primary goal of sentiwordnet is to determine the appropriate emotion of a word used in a certain context. To begin, Automatic Text Summarization uses the Simplified Lesk method to determine the weights of a large number of sentences in a text and then organises them in decreasing order based on their weights.Next, a certain amount of words are chosen from the requested rundown, as suggested by the specific level of rundown. The proposed method produces the best results up to 50% summary of the first material and attractive results up to 25% sketch of the first material.

In [3], Mihir Vaidya and Varad Ahirwadkar implemented the practise of reducing lengthy chunks of writing is known as text summarization. For abstractive text summarization, neural network models have been supplied with a novel practicable approach. The term "abstractive" refers to the creation of new words from the original document that may or may not exist in the original document. These neural network models have two flaws: they tend to repeat themselves and are prone to reproduce factual details incorrectly. We suggest a novel architecture that enhances the classic sequence-to-sequence selective attention model with two orthogonal ways in this paper. We primarily used a pointer-generator network, which uses pointing to select words from a given text, allowing for precise data replication while retaining the capacity to introduce new words via the generator.

III. PROPOSED METHODOLOGY

3.1 Dataset

The dataset that was used is Amazon Fine Food Reviews which was available in the kaggle website.
The dataset was hosted on Kaggle.
This data is totally based food reviews which are collected from amazon. This dataset contains more than 6 columns like those 6 columns will be act as features of the model that we are building and this dataset contains approximately 5 lakh tuples (reviews)

Dataset includes:

Reviews are in the range from 2009 – 2013.
589,309 reviews from users.
235,789 members of users
73,689 different products

About the dataset:

1.id
2.product Id
3.User Id
4.Profile Name
5.Time
6.Summary
7.Text

3.2 Algorithms

3.2.1 Graph based Algorithms:

The graph-based approach to text summarization is indeed an unsupervised method in that we use a graph to rank the necessary sentences or words. The basic goal of the graphical method is to extract the most important sentences from a single document. In a summary, we determine the importance of a vertex in a graph. Text-based ranking is accomplished using unidirectional and weighted graphs. Either the documents or the statements are presented as nodes in this technique. Edges are used to link any two nodes that share the same data. The initialization of weightings to the nodes of the graph is used to score sentences
Other text summarization methods using a graph-based approach include:

1)page rank:
It is a Google algorithm that controls the connection of web sites with comparable content.

2)Lex rank:
This alogorithm make use of cosine similarity and tf-idf models.Cosine similarity is a measure of how similar two vectors are in an n-dimensional space. It evaluates the cosine similarity of two vectors and see whether they are heading in almost the same general direction. It's often used in text analysis to determine document similarity.

3)Text rank:
It is an extention to the page rank algorithm and similar to the Leax rank algorithm in order to normalize the data.

3.2.2 Neural Networks algorithms:

Humans will not rethink their thoughts every second. As you read this article, you grasp each term based on the understanding of prior statements. You don't start over from the beginning and throw everything out. Your ideas are powerful. Something that typical neural networks are incapable of, and which appears to be a great weakness. Consider the case below: You want to categorise the different types of events that happen in a movie at various moments. It's unknown how a normal neural network can also use previous movie events to inform future ones.

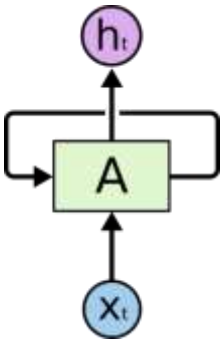

Fig 3.1 RNN looping structure

Recurrent neural networks are used to solve this problem. They're networks with built-in cycles that keep data endure
In the diagram above, a slice of a neural network, A, evaluates a set of inputs Xt and returns a value ht. A loop can be used to communicate data from one network phase to the next.
Because of the loops, Rnn looks mysterious. Yet, when you consider it, they're not really all that dissimilar from a perceptron. A recurrent neural network network is made up of several copies of the same network, each of which sends the message to another in line.
Consider what happens if you unwind the loop:

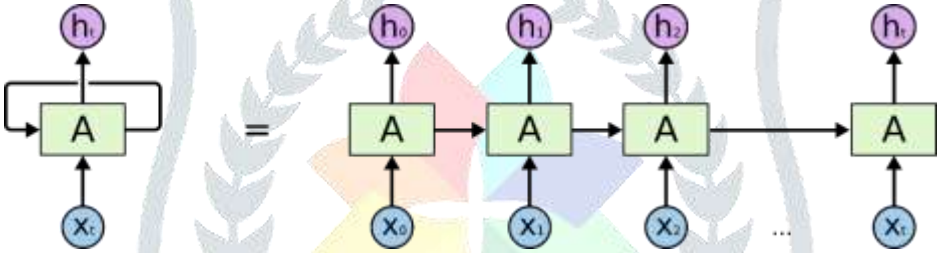

Fig 3.2 RNN

One of the advantages of RNNs is its ability to integrate earlier information to the current task, such as using a prior video sequence to help know the current screen. To complete a given task, we sometimes only need to glance at recent data. Consider a learning algorithm that tries to guess the words based on the ones that have came before it. We don't need much more information
to figure out how the last word in "the clouds are already in the sky" means - it's clear that next word will be sky. In situations where the distance between relevant information and the location where it's needed is small, RNNs can learn to leverage prior knowledge. However, they will be some occasions when we require further information. Take the text's concluding sentence: "I grew up in France... I am proficient in French." According to current findings, the next word would most likely be a language name, and we'll need more information for France from the past to narrow down which language it will be. It's entirely possible that the time separating relevant data and when it's required may increase exponentially. RNNs, on the other hand, lose their capacity to learn to connect the dots as the gap expands.

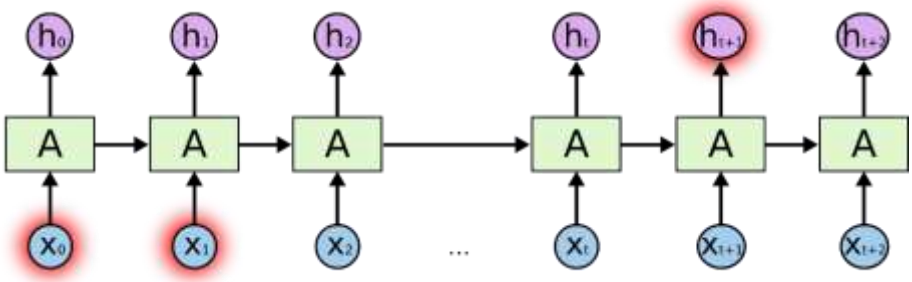

Fig 3.3 lstm architecture

LSTMs, or Long ShortTerm Memory networks, are a form of Rnn which can learn long-term dependencies. They were first introduced by Hochreiter & Schmidhuber (1997), and they have since been refined and popularised by a number of people. They are being used more and more frequently, and they are incredibly useful in a variety of situations

Lstm networks were created expressly to address the issue of long-term reliance. They shouldn't have tried extremely hard to recall knowledge for extended periods time; it should be a non-issue for them! A series of repeated neural network modules helps

to compensate every rnn. This repeating module in conventional RNNs will have a relatively simple structure, such as a single tanh layer.

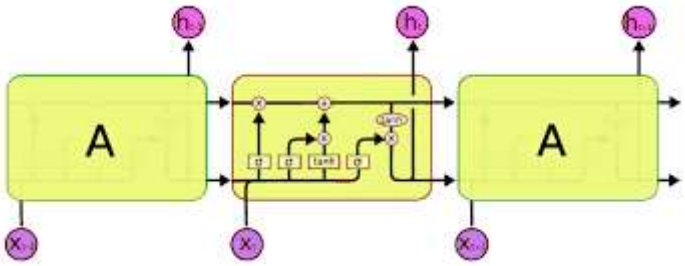

Fig 3.4 lstm architecture

LSTMs, on the other phase , utilize multiplications with additions to compute small changes in data. In LSTMs, flow of information through a mechanism called cell states. Lstm networks have the flexibility of remembering or not remembering information in this manner.

There appear to be three distinct reliances on information at a given cell state.We'll provide an example to illustrate this. Take, for example, projecting stock values for a specific stock. Today's stock price will be determined by:

1.the stock trend of a particular stock from the past flows may be uptrend or downtrend in present situation.
2.price of the stock may also vary from the previous days.
3.fators that affecting price fall or up includes company policies and decisions which are made on the management level.

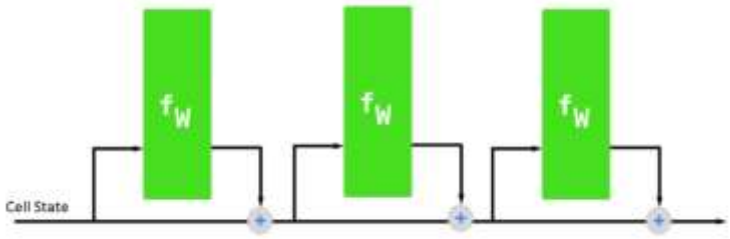

Fig 3.5 convoy belt

The most important feature of LSTM are convoy belt.cell state machine is related to a conveyor that transmits data all through the cell. While it isn't exactly a gateway, it is essential for data to flow through each cell and to other cells. According to the findings of the forget and input gates, the data passing through it is modified and updated before being passed to the next cell.

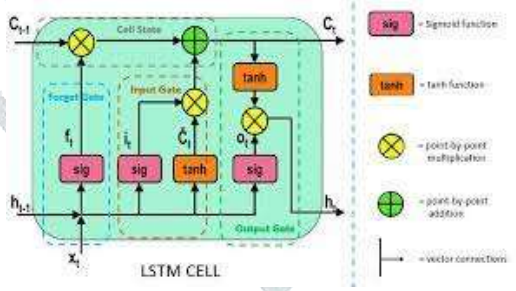

Fig 3.6 lstm cell

Forget Gate:

Just like humans who are not to consider some of the events and situations in their daily life activities forget gate also remove unwanted information from merging it with the cell states.
It takes two inputs(xt and ht-1) and send it to sig gate which removes the unwanted information from input data and send to the multiplication process.

Output Gate:

Based on cell condition, previous cell outputs, and new data, the penultimate gate selects useful information. It achieves this using a tanh function to create a vector from of the cell state after the inputs and forget gates have merged. The new input and prior cell output are then run through a sigmoid function to evaluate which values must be outputted. The output of this cell is multiplied by the results of those two procedures.

3.2.3 Bi-directional LSTM:
The method of allowing any neural network to store sequence data in both backwards and forward directions is known as bi-directional long-short term memory (bi-lstm).
A bidirectional LSTM differs from a conventional LSTM in that its input is split into two channels. With a conventional LSTM, we may make input travel in one direction, either forwards or backwards. We can, however, have flow of information in both

directions with bi-directional input, storing both of the future as well as the past.. Let's consider the example for a greater understanding.

The blank space in the sentence "boys go to..." cannot get completed. Still, we can simply forecast the prior blank space and also have our model do the same thing while we have a forthcoming sentence like "boys come out of school," and bidirectional LSTM allows the neural net to work

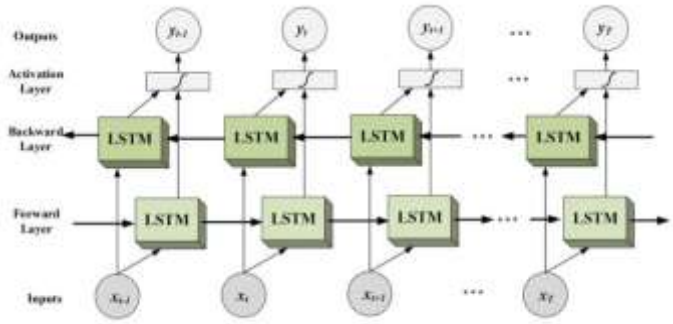

Fig 3.7 Bi-directional LSTM architecture

IV. RESULTS AND DISCUSSION

4.1 Streamlit Aplication:

Streamlit one of the most important and easiest way implement an website using python which usally doesn't require to integrate with the HTML/CSS part of the code. Most of the blocks that are required for building the website are predefined in the streamlit application

4.2. Summarizing Using Url:

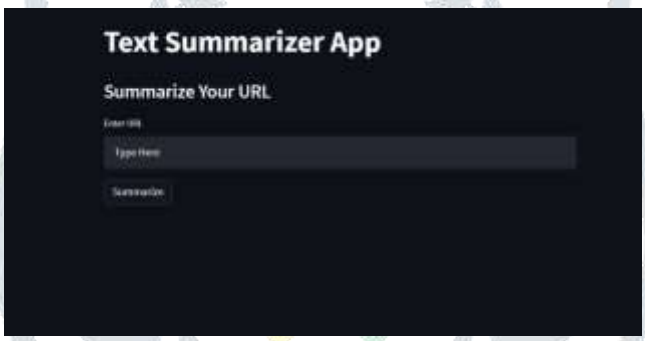

Fig 4.1 Home Page

When we enter any web page url which supports HTML parser then entire content of that page will be extracted and summarized.

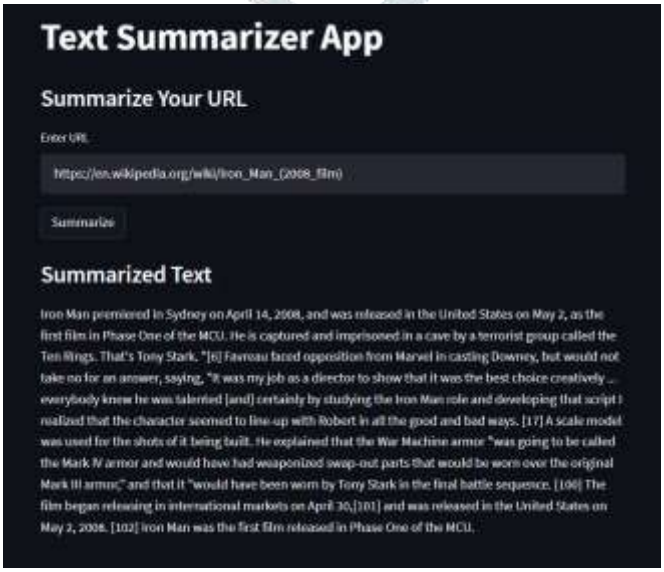

Fig 4.2 Input with URL

4.3 SUMMARIZING USING TEXT:

INPUT:

Iron Man is a 2008 American superhero film based on the Marvel Comics character of the same name. Produced by Marvel Studios and distributed by Paramount Pictures,[N 1] it is the first film in the Marvel Cinematic Universe (MCU). Directed by Jon Favreau from a screenplay by the writing teams of Mark Fergus and Hawk Ostby, and Art Marcum and Matt Holloway, the film stars Robert Downey Jr. as Tony Stark / Iron Man alongside Terrence Howard, Jeff Bridges, Shaun Toub, and Gwyneth Paltrow. In the film, following his escape from captivity by a terrorist group, world famous industrialist and master engineer Tony Stark builds a mechanized suit of armor and becomes the superhero Iron Man.

A film featuring the character was in development at Universal Pictures, 20th Century Fox, and New Line Cinema at various times since 1990, before Marvel Studios reacquired the rights in 2005. Marvel put the project in production as its first self-financed film, with Paramount Pictures distributing. Favreau signed on as director in April 2006, and faced opposition from Marvel when trying to cast Downey in the title role; the actor was signed in September. Filming took place from March to June 2007, primarily in California to differentiate the film from numerous other superhero stories that are set in New York City-esque environments. During filming, the actors were free to create their own dialogue because pre-production was focused on the story and action. Rubber and metal versions of the armor, created by Stan Winston's company, were mixed with computer-generated imagery to create the title character.

OUTPUT:

SUMY LEX RANK:

Fig 4.3 Input with Text

4.4 USING LSTM:

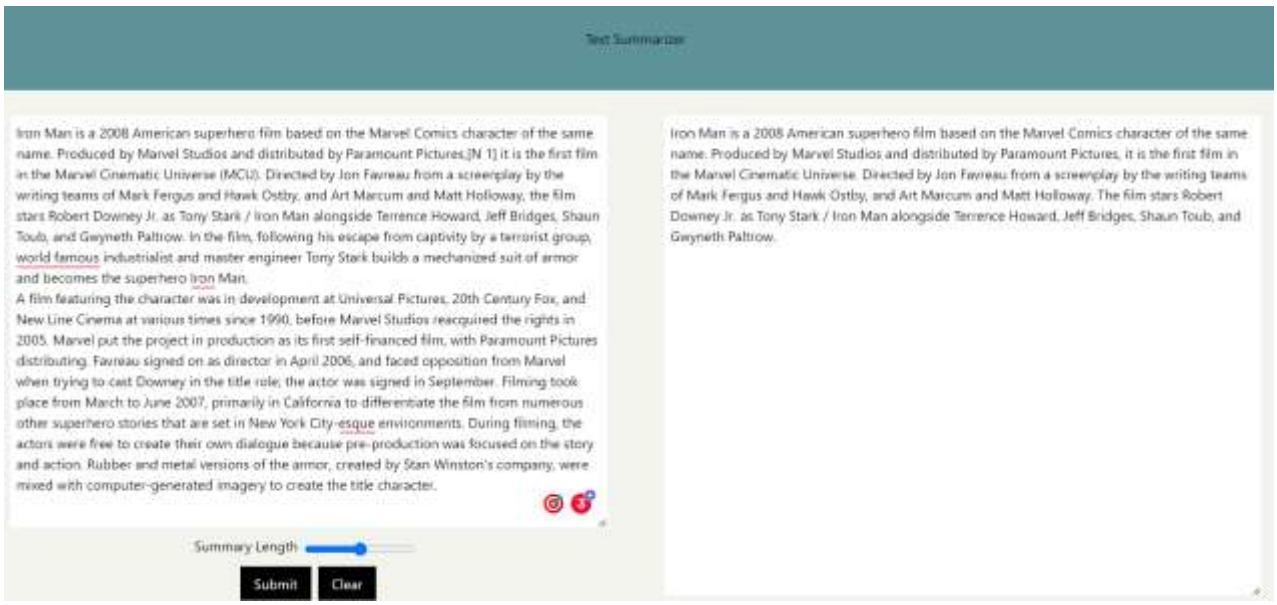

Fig 4.4 Input with LSTM

V. CONCLUSION

Automatic text summarization is one of the most common and effective techniques in natural language processing. Text summarizing can be accomplished in one of two ways: extractive summarization or abstractive summarization. The area of automatic text summarizing has a long history of research, and the focus is shifting from extractive to abstractive summarization. The abstractive summary technique creates a summary that is relevant, precise, content-rich, and less repetitive. Abstractive summarization is a difficult area since it focuses on providing a summary that is closer to human intellect. As a result, this study puts the methodologies for abstractive summarization to the test, as well as the benefits and drawbacks of various approaches. All of the techniques are contrasted and discussed. This survey is a good method to learn more about abstractive summarization.

REFERENCES

[1] Jingwei Cheng, Fu Zhang, Xuyang Guo "A Syntax-Augmented and Headline-Aware Neural Text Summarization Method", IEEE", Communication and Technology, 2021
[2] Keerthana P, Automatic Text Summarization Using Deep Learning", 2020
[3] Mihir Vaidya, Varad Ahirwadkar "Deep Learning Approach for Text Summarization", IJRET,2021
[4] Pratibha Devihosur, Naseer R, "Automatic Text Summarization Using Natural Language Processing", IJRET, 2019
[5] Xiao Gao, Amin Ali, Hassan Hassan, Eman M, "Improving the accuracy of text summarization based on ensemble method", Hindawi, 2021
[6] Kasimahanthi Divya, Kambala Sneha, Baisetti Sowmya, G Sankara Rao "Text Summarization using Deep Learning". "IEEE",2020
[7] Nikita Desai , Prachi Shah2"Automatic Text Summarization Using Supervised Machine Learning". "IEEE",2020