Automatic Text Summarization Using Deep Learning (compare the neural network model RNN, LSTM)

Divya Amade[1]Mahadev Bag[2] Rashmi Chandra[3]

[1,2,3] Department of Computer Science and Engineering, Raipur Institute of Technology
[Affilated by Chhattisgarh Swami Vivekanand Technical University, Bhilai]

***Abstract*—Automatic text summarization using Deep learning is very essential way for summarization large content into summarize form. This paper presents a method of achieving text summaries accurately using deep learning methods, we propose a method of text summarization which focuses on the problem of identifying the most important portions of the text and producing coherent summaries with the help of deep learning. Deep learning techniques are proved to be effective in generating summaries form of volume text. The study explores both extractive and abstractive summarization methods using sequence to sequence model and LSTM. The research findings reveal the strengths and limitations of LSTM in text summarization and demonstrate its potential for facilitating efficient information extraction from textual data.**

***Index Terms*—LSTM, RNN, ATS, abstractive, extractive**

I. Introduction

In the modern world, where tremendous amount of data is accessible on digital platforms, it is important to make an enhanced tool to get the desired data rapidly. Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. It is a toughtask for individuals to manually select the gist of elaborated text. There is an issue of scanning such large reports from the accessible archives/text. Also, the main concern is to recognize the most important data in the document, large text records or set of related text. With the revolutionary and rapidly growing amount of data, discovering the crisp amount of information is challenging. There should be some tool which compresses them into a shorter interpretation looking after its implications. Hence, it is essential to make a model that could condense data like us. Designing such a model is the real task. The purpose of this project is to produce such a model as the solution which is based on Extractive Approach for summarizing text, starting with the sequence to sequence and LSTM. The extractive approach is actually successful in delivering the summary using the same set of words which are actually most important words present in the actual text/archive; hence, it delivers the relevant information.

From here, we come across with the effectiveness of different methods for distinguishing them on the basis of size accuracy of summary. Here, these methods try to first understand the text and then mark the words according to their importance and then selecting the sentences containing the most important words in it and using them or the words used in their place to form the actual summary shortening the length of actual text. The procedure in all the methods is same i.e., Text->Text-Processing->Summary, where text is the input,text- processing is the intermediary step summary is the final output. One of the approaches referred as the abstractive approach which is one of the two important methodologies involved in Automatic Text Summarization works by giving the synopsis that includes new set of words.

It is used to deliver the required summary with the same meaning as the original text. As it is clear, the extractive approach basically first selects various and unique sentences/sections of the text/document then combines them to form a summary. These sentences are selected on the basis of accurate highlights/scores described as the importance of the sentences. The importance plus meaningof the first record is maintained preserved in both the cases. Here, we have opted for the Extractive-Approach. Using this approach, we could produce the relevant summary in efficient form as well as more accurate summarize result produced.[2]

A. Content

A text is a written document or any object that can be ‘read', ‘written', ‘displayed', ‘visualized', ‘typed', ‘interpreted', ‘scanned' or ‘printed' whether this object is a work of literature, articles published in newspapers, magazines, a type of document. It is a coherent set of signs, symbols, semantics and syntax that transmits some kind of information. Text represents textual document which is a written or printed work and regarded in terms of its content [1].

B. Summarization

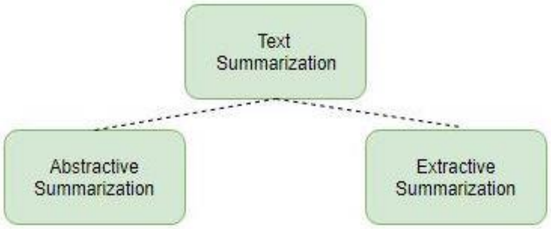
Summarization is the process of making a summary of any text. A summary is a crisp statement or restatement of major points, especially as a conclusion to a work, it is actually a comprehension and usually brief extract, abstract or recapitulation of previously stated facts or statements. To summarize means to sum up the main points of something —a summarization is the kind of summation of a large document or huge amount of text [1].

C. Text-Summarization

Text summarization is the process in which long piece of texts gets a crisp format with lesser number of words than the actual text still reflecting the same meaning as the original doc/text.

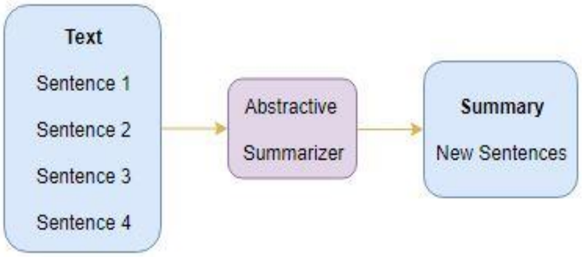
I. TEXT SUMMARIZATION METHODS

The various dimensions of automatic text summarizationcan be generally categorized as different approaches based oncertain characteristics like single or multiple document(s) specific or general purpose, learning Algorithm output-based(extractive or abstractive) [12].



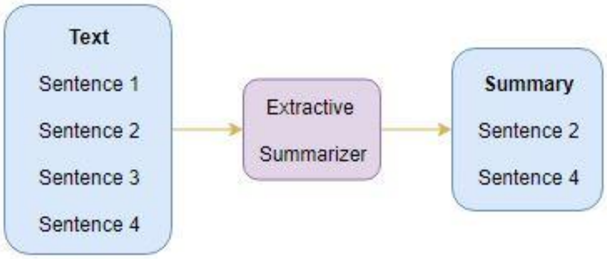
A. Abstractive Approach

Abstractive summarization, is all about trying to comprehend the content of the text, generating synonyms or completely new words and utilizing them to make the Summary. It perhaps not contains the same sentences as present in the original text. This approach incorporates learning methods to make its own sentences but reflect the same meaning as the original text was providing [12].



B. Extractive Approach

Extractive summarization is one of the methods which incorporates making a summary on the basis of scoring technique. It marks the sentences containing important words with higher value as compared to the sentences containing least valued words. A subset of these high-valued sentences isselected within the boundaries of the text. There are two important parts for accomplishing this approach: extraction and expectation both required for extracting & grouping words & sentences according to their score to display them as the appropriate summary [12].



II. Model used for Text summarization

For text summarization various types of model used such as sequence to sequence model, Recurrent neural network and long short term memory model. Conventional technique used for text summarization is seq2seq,it can summarize simple and smaller content but cant not summarize sparse data. Overcome the constraint arises from the previous model by RNN,it has hidden state i.e store previous data in memory for further output. Complex and voluminous content can summarize by lstm model, it can resolve the issue of recurrent neural network model.

A. Sequence to sequence model

Seq2seq revolutionized the process of translation by making use of deep learning.Seq2Seq (Sequence-to-Sequence) is a type of model in machine learning that is used for tasks such as text summarization. The model consists of two main components:

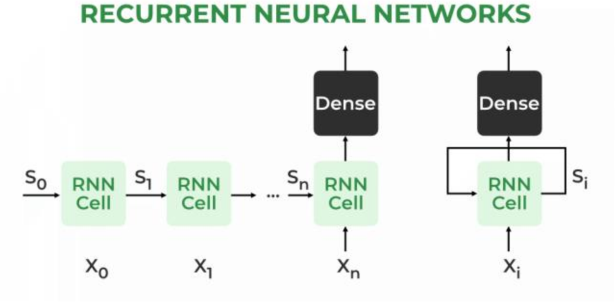
Encoder
Decoder

Sequence modelling

It can be categorize in based on the type of input and output sequences. Inputs and outputs can be one of the following: Scalar, Trend, Text, Image, Audio or Video.This model generally used for text summarization process to trained machine learning with the help of using these technique short and not complex sparse matrix can summarise. seq2seq takes as input a sequence of words(sentence or sentences) and generates an output sequence of words It does so by use of the recurrent neural network (RNN).[14]

B. RNN(Recurrent Neural Network)

Recurrent Neural Network also known as **(RNN)** that works better than a simple neural network when data is sequential like Time-Series data and text data. RNN is type of Neural network, ouput is the previous layer is input to the current layer,if we want to predict next word or sentence required to remember previous word. RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is its Hidden state, which remembers some information about a sequence. The state is also referred to as Memory State since it remembers the previous input to the network.



The formula for calculating the current state:

$$h_t = f(h_{t-1}, x_t)$$

where
h_t -> current state
h_{t-1} -> previous state
x_t -> input state

Formula for applying Activation function(tanh):

$$h_t = \tanh (W_{hh}h_{t-1} + W_{xh}x_t$$

where
W_{hh} -> weight at recurrent neuron
W_{xh} -> weight at input neuron

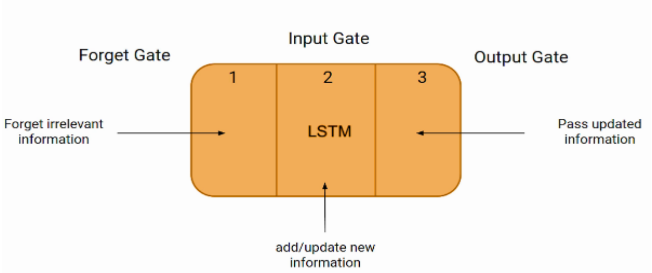
Formula for calculating the output:

$$y_t = W_{hy}h_t$$

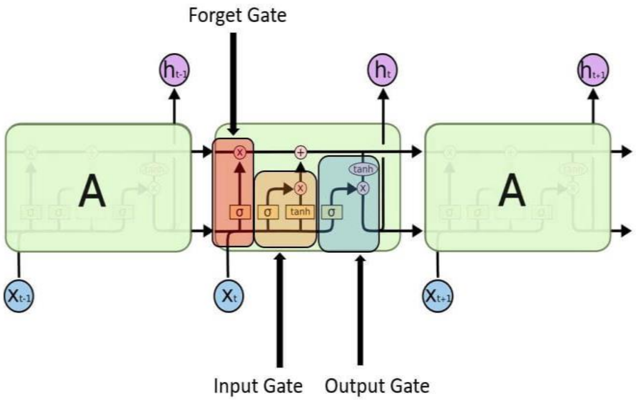
where
Y_t -> output
W_{hy} -> weight at output layer

C. LSTM (Long Short Term Memory)

Overcome the problem of RNN proposed new version of predictive model is LSTM. Long short term memory is a special form of RNN that are capable of learning long term dependencies. In some scenarios, only recent information are required to perform a given task such as language models trying to predict the last word in a sentence .In situations where the gap between relevant information and the place where it is needed is small RNN learns to use the past information without the occurrence of problems discussed earlier But there are cases where more context are needed where the gap between relevant information and where it is needed are large. In such cases LSTM network are proven to be efficient. LSTM has a chain like structure where the repeating module has four interacting modules that are the cell state, output gate, update gate and the forget gate.



Architecture of LSTM



A. Forget Gate

Forget Gate:

- $f_t = \sigma (x_t * U_f + H_{t-1} * W_f)$

where
Xt: input to the current timestamp.
Uf: weight associated with the input
Ht-1: The hidden state of the previous timestamp
Wf: It is the weight matrix associated with the hidden state 0 Sigmoid function applied to it for timestamp.

B. Input Gate

The input gate is used to quantify the importance of the new information carried by the input. Here is the equation of the input gate

Input Gate:

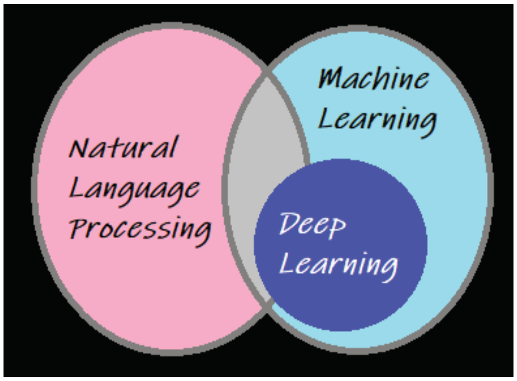
- $i_t = \sigma (x_t * U_i + H_{t-1} * W_i)$

where
Xt: Input at the current timestamp t
Ui: weight matrix of input
Ht-1: A hidden state at the previous timestamp
Wi: Weight matrix of input associated with hidden state

New Information

- $N_t = \tanh(x_t * U_c + H_{t-1} * W_c)$ (new information)

New information that needed to be passed to the cell state is a function of a hidden state at the previous timestamp t-1 and input x at timestamp t. The activation function here is tanh.



Due to the tanh function, the value of new information will be between -1 and 1. If the value of Nt is negative, the information is subtracted from the cell state, and if the value is positive, the information is added to the cell state at the current timestamp.

Update

$$C_t = f_t * C_{t-1} + i_t * N_t \text{ (updating cell state)}$$

C. Output Gate

Output Gate:

- $o_t = \sigma(x_t * U_o + H_{t-1} * W_o)$

Its value will also lie between 0 and 1 because of this sigmoid function. Now to calculate the current hidden state, we will use Ot and tanh of the updated cell state.

$$H_t = o_t * \tanh(C_t)$$

Hidden state is a function of Long term memory (Ct) and the current output. If you need to take the output of the current timestamp, just apply the SoftMax activation on hidden state Ht.

$$\text{Output} = \text{Softmax}(H_t)$$

III. Text-Processing

The automated process of analysis & manipulation of the text is known as text processing [1]. It takes the text as input, processes it & finally provides the required outcome; it could be widely used within different areas of an organization, suchas product teams could get insights from customer feedbacks to automate customer services. Here, words/tokens of the text represent discrete, categorical features.

A. Tokenization

Splitting into tokens**.** Tokens refers to any individual unitin the program which is meaningful to either the machine or the human.

Word-Tokenization: When the entire text is divided into individual words and word-score is generated for every word according to it's count.

Sentence-Tokenization: When the entire text is divided into individual sentences and each sentence is provided it's sentence-score according to the occurrence of the high-scored words.

A. Relationship Between ML, DL & NLP

Venn Diagram

Table 1: *Ratio Table*

INPUT	OUTPUT
Text	Summary
2(Two)	1(One)
Eg. Extract fromIBM reports	Coherent Summary

A. Comparision Table [1, 13,15]

Table 2: *LSTM, CoreNLP & NLTK (statistically)*

Package	Precision	Recall	F-Score
LSTM	0.9969	0.9960	0.9965
CoreNLP	0.7856	0.7255	0.7589
NLTK	0.5123	0.6523	0.5804

B. Schematic Diagram

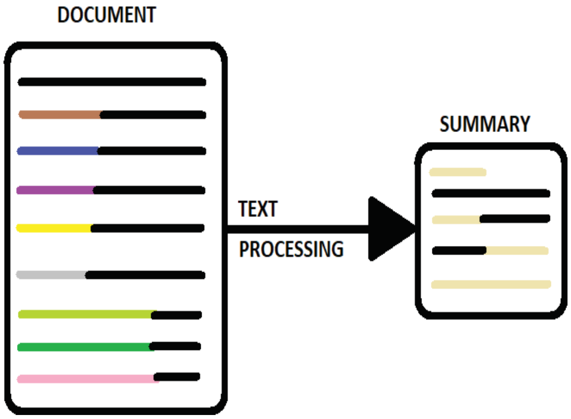


Fig.1 Text Processing

I. Conclusion

The foundation for existing models is LSTM, a package for processing text string by string. NLTK accepts strings (a series of characters) as input and output. This NLP process overcome the vanishing gradient and explode gradient problem. The vanishing gradients problem is one example of unstable behavior that you may encounter when training a deep neural network. Long short term model has used hidden layer(memory) and store input and previous information for getting the appropriate output. Explode gradients are used to update the weights, if the gradients are large, the multiplication of these gradients will become huge over time. This results in the model being unable to learn and its behavior becomes unstable. Due to the above problem resist in deep learning technique difficult to trained the model and has not find the accurate result,lstm overcome the gradient problem and trained the more accurately than other model.According to this it converts the text into one object as a whole. It includes word-vectors and this is lagging in previous tool, creating word vectors helps in proper assignment of real numbers to represent the meaning/efficacy of a word & clustering them accordingly, this makes Mathematical operation easy to use on these vectors.

II. Future Scope

In this project the extractive approach is explained, utilized and implemented; the next approach named as abstractive approach of Automatic Text Summarization could be the upcoming challenge, it is a technique wherein task of summarization becomes very complex as the whole text is required to be understood by the machine to generate a summary with entirely new words delivering the same meaning as the original text. Here, the model has to be trained with a lot of words & their synonyms, one word replacing many words & the correct usage of each word; LSTMs became popular because they could solve the problem of vanishing, explode gradients. But it turns out, they fail to remove it completely. The problem lies in the fact that the data still has to move from cell to cell.Do not completely overcome the gradient problem by ,lstm to trained the model for getting more accurate summarization of large content(Text summarization). We would be extending the project in future to create the automatic text summarizer having the combination of both Extractive & Abstractive approach for getting more accurate result.

References

[1] https://www.google.com for certain terms & their proper reference.

[2] Athira S*,SruthyManmadhan "Deep learning in Text Summarization-A Survey" Alliance International Conference on Artificial Intelligence and Machine Learning (AICAAM), April 2019

[3] Kasimahanthi Divya, Kambala Sneha "Text Summarization using Deep learning in -A Survey" Alliance International Conference on Artificial Intelligence and Machine Learning (AICAAM), April 2019

[4] Ahmad T. Al-Taani. "Automatic Text Summarization Approaches" International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017).

[5] Neelima Bhatia, Arunima Jaiswal, "Automatic Text Summarization: Single and Multiple Summarizations", International Journal of Computer Applications.

[6] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, "Text Summarization Techniques: A Brief Survey", (IJACSA) International Journal of Advanced Computer Science and Applications.

[7] Pankaj Gupta, Ritu Tiwari and Nirmal Robert, "Sentiment Analysis and Text Summarization of Online Reviews: A Survey" International Conzatiference on Communication and Signal Processing, August 2013.

[8] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques." JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010.

[9] Jiwei Tan, Xiaojun Wan, Jianguo Xiao Institute of Computer Science and Technology, Peking University "Abstractive document summarization with a Graph- Based attentional neural model.

[10] "Seonggi Ryang, Graduate school of Information science and technology, University of Tokyo "Framework of automatic text summarization using Reinforcement learning".

[11] Luhn, Hans Peter. "The automatic creation of literature abstracts." IBM Journal of research and development 2.2 (1958): 159–165.

BIOGRAPHY

Divya Amade is currently pursuing her M.Tech degree in Computer Science and Engineering from Raipur Institute Of Technology affiliated to Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India. She has completed BE in Information Technology from Rungta College of Engineering & Technology, Raipur Chhattisgarh, India in 2006. Her research interest fields are Data Science, Neural Network, and Artificial Intelligence.

Mahadev Bag is currently working as a HOD in the Computer Science and Engineering from Raipur Institute Of Technology affiliated to Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India .He has 18 Years of experience in teaching.He has published 12 research papers SCI and Scopus.

Rashmi Chandra is currently working as an Assistant Professor in the Computer Science and Engineering from Raipur Institute Of Technology affiliated to Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India .She has 11.8 Years of experience in teaching.She has published 2 research papers SCI and Scopus.