Optimized Text Summarization using Customized Recurrent Neural Network Model

Dr. B Anjanadevi [a], Dr B S Nagakishore [b]

[a] Assistant Professor, Maharaj Vijayaram Gajapathi Raj College of Engineering (Autonomous),Vizianagaram, AP, INDIA.
[b]Assistant Professor, Acharya Nagarjuna University, Guntur, AP, INDIA.

In current era, the growth of big data is raising in large scale and it becomes tough to the users to load and draw conclusions from it. An efficient approach is required to generate short and precise summaries in user required form. In recent years, Sequence to Sequence models gained lot of focus on text summarization to handle various challenges like fluency, human readability and also generation of optimistic summaries. The proposed approach effectively handled these issues using Customized Recurrent Neural Networks (C-RNNs) model to generate optimistic text summary. The proposed model generated optimized text summaries where the data collected from social media/E-Commerce sites. C-RNN model are highly demanded for entrepreneur's and consumers when they need precise text summaries. Experimental results show that proposed model is outperformed the state of the art models in terms of syntactic and semantic structure and achieves qualitative results.

Keywords: Bigdata, Neural Networks, Recurrent Neural Networks, Optimistic Summaries, Customized- Recurrent Neural Networks (C-RNNs)

1. Introduction

The essential need to summarize textual data has become extremely leading in the past few years. The people have to spare a minimum time, and they are most obviously they do not wish to pay out it going through extensively lengthy, documents only to find a small segment of the text to be applicable. Hence, an interesting and likely possible solution to the above problem is to develop a program that summarizes lengthy text documents into smaller versions. Summarunner A recurrent neural network-based sequence model for extractive summarization of documents, Ramesh Nallapati, Feifei Zhai and Bowen Zhou. The approaches are usually either extractive or abstractive. SummaRuNNer, a Recurrent Neural Network (RNN) deploy a sequence model for extractive summarization of text and display that it complete performance is superior. That focus on many features of summarization like controlling out of vocabulary words and modeling syntactic features of words in the text or sentence. In extractive summarization, the model selects passages from the input document and combines them to form a shorter summary, sometimes with a post-processingstep to ensure final coherence of the output.

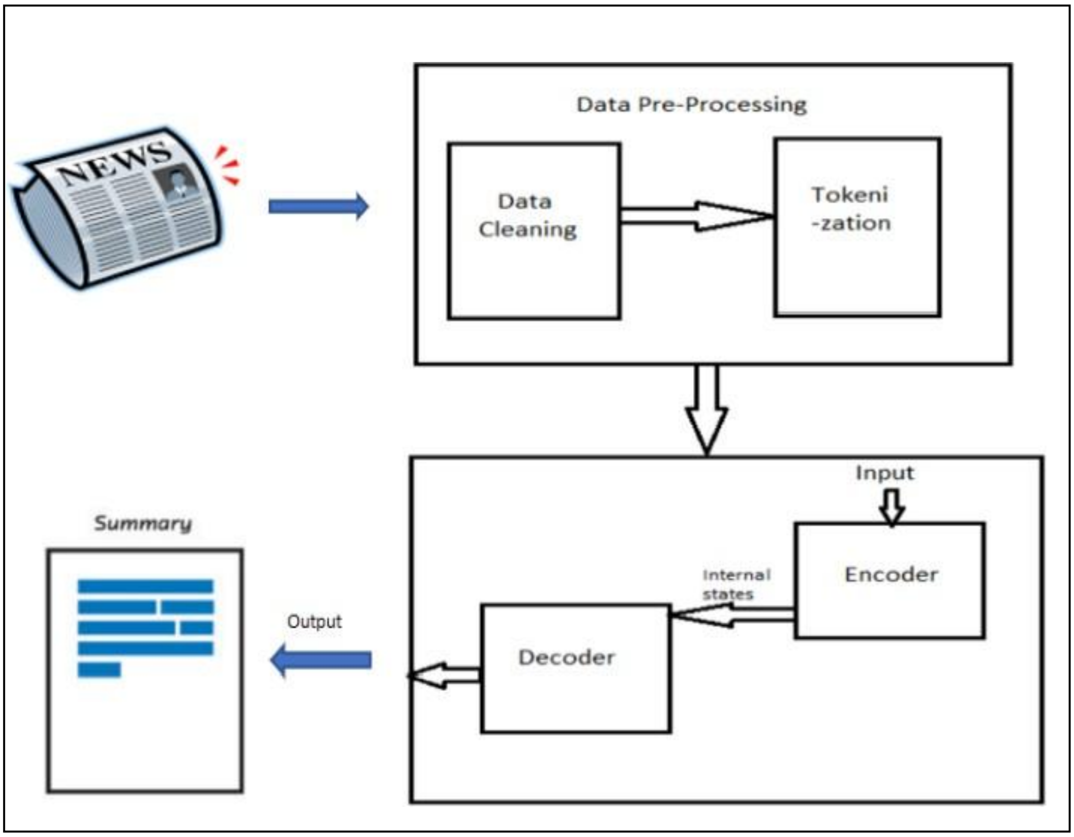
While extractive models are usually robust and produce coherent summaries, they cannot create concisesummaries that paraphrase the source document using new phrases.Text Summarization Techniques A Brief Survey, Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan, B. Gutierrez, Krys Kochut Publishing Details (IJACSA), surveys the different operations for summarization and describes the acceptability of an deficiencies of the different strategies. Concept outline helps in shorteningthe length the of document or text in packed form keeping all the important data. Programmed text summary is handled to provide a short analysis while preserving key data content and a large meaning. There are many models like web crawlers create pieces as the task of the library, various models helps in organize the new sites which helps in giving stuffed representation of news points and more often they are not review to encourage and analyze. for instance, text and expression are recurrence. The density of the sentence was suggested. The two techniques for programmed rundown are extraction and reflection. Extractive summary planning is to discriminate immediate segments of the text and making them accurately. Abstractive summary approach is to carry all important component recently. Extractive analysis gives preferred result over abstractive summary.

Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, Nallapati et al. also appeal their abstractive summarization model on the CNN/Daily Mail dataset, which carry input segment up to 800 tokens and multi sentence summaries up to 100 tokens. But their testing display a key problem with immersion of encoder-decoder models they may also invoke unnatural summaries consisting of repeated phrases.Automatic Text Summarization Approaches, Ahmad T. Al-Taani. Synopsis Text summarization systems are important in many features in a language like natural language processing. ATS generates the summary of given text which helps in saving time and resources. There are one or more text documents in text summary. Only one document is draw out in case of one document summarization whereas set of documents is selected in groupof document summarization. There are two content-based summarization generic and query-based summaries. In generic summarization If suppose a user don't know the content of the text then the information measure equal level information. But in query-based summarization, the text is verified of the original text. There are three main important techniques to text summarization i.e statistical, graph-based, and deep learning approaches. Another technique is a clustering. In statistical technique, analysts are established upon text ranking and most important text are elected from the given data, view as the important synopsis squeezing ratio. Graph-based technique focus on the semiotics analysis and relationship among text. Machine learning access the help in generating summary by applying machine learning algorithms. This approach deals with the textsummarization process as a classification problem.

Deep Learning in Text Summarization - A Survey, Athira S, Sruthy Manmadhan. This paper is survey on deep learning based approaches for text summarization over the years. In this new era world trading with huge amount of information, text summarization deals a critical role in expressinga meaningful information and exhibit precise, comprehensible information from large. There are Many ways to summarize text have been imported over the years. traditional methods create synopsis from data directly by separating words that leads to redundancy and ignore document synopsis relationship. Improving Abstraction in Text Summarization: Authors Wojciech Kryściński, Romain Paulus, Caiming Xiong, Richard Socher abstractive application are introduced based on neural sequence-to-sequence structure. Based on the sequence-to-sequence model with copy mechanism, introduce intra-temporal attention processes in the encoder and decoder to address the repetition and incoherent problem. There are two issues in previous abstractive methods. One is in this system use left-context-only decoder, thus do not have terminate context when forecast each word. Second one is that they do not handle the pre-trained explore language models on the decoder side, so it is more difficult for the decoder to discover synopsis representations, context communication and language modeling together.

2. System architecture and description

The Proposed architecture will take news articles as input in which they will be sent for pre-processing. Performing basic pre-processing steps is very important part. Because using complex and uncleaned text data is potentially an unfavorable move. So, in this



phase, we will remove all the unwanted symbols, irrelevant characters etc. from the text that do not affect the objective of the problem statement.

Figure 2.1 System Architecture

Cleaning the text by removing HTML tags's, punctuations, special characters, stop words. The tokenizer builds the vocabulary, and it converts a sequence of word to an integer. The tokenized words are fed into encoder, where it captures the information which is needed to get the summary and decoder will predicts the words and generates the summary.

3. Proposed approach

3.1 Data Preprocessing Steps

I. Read a comma-separated values (csv) file into dataframe in which it stores tabular data and it can be imported and exported from programs that store in the tables, such as Microsoft excel.

II. Drops the duplicates and missing values from the text.

III. Converting the data into lowercase and replace a string that matches a regular expression instead of perfect match.

IV. Adding the start and the end tokens at the beginning and end of the summary text.

V. Splitting the data into training and validation set by train_test_split() by specifying the test size as 0.2, to place 80% of the data into training and remaining into validation.

Tokenization Steps

I. Tokenizer ()- preparing a tokenizer.
II. Create the vocabulary index based on the frequency of the words. Fit_on_texts updates the vocabulary internally basing on the list of texts.
III. Converting text sequences into integer sequence using texts_to_sequences, it takes each word from the text and replaces with its respective integer values.
IV. Padding zeroes upto maximum length to make it equal to the total length of text. pad_sequences(), takes maxlen and padding as parameters where maxlen counts upto maximum length of all the sequences and padding is a string(post), pad after each sequence.

4. Model Building

In this phase, we will build a neural network layer in which

- We define a input layer of maximum text length for encoder
- We define one embedding layer for encoder where input layer output will be given as a input here
- We define two bidirectional lstm layers where embedding layer's output will be given as input to the 1st layer which results in encoder output and also forward, backward hidden and cell state then this output and respective states are given as input to the next bilstm layer which results in the respective output and state.
- We define a input layer of max text length for decoder
- We define one embedding layer for decoder where decoder input layer output will be given as a input here
- We define one lstm layer where embedding layer's output will be given as input to this layer and also setting its initial state to encoder'sfinal state which results in encoder output and also hidden and cell state
- We define a attention layer where encoder's and decoder's outputswill be given as input such that this layer is used to give attention to important information in the given input
- We define one concat layer which concatenates the output of attention and decoder layer
- We define one dense layer which applies activation function in our neural network layer

In final, in our model architecture we have used two input layers, two embedding layers for encoder and decoder respectively and also two bilstm layers for encoder, one lstm layer for decoder also one attention layer and dense layer.

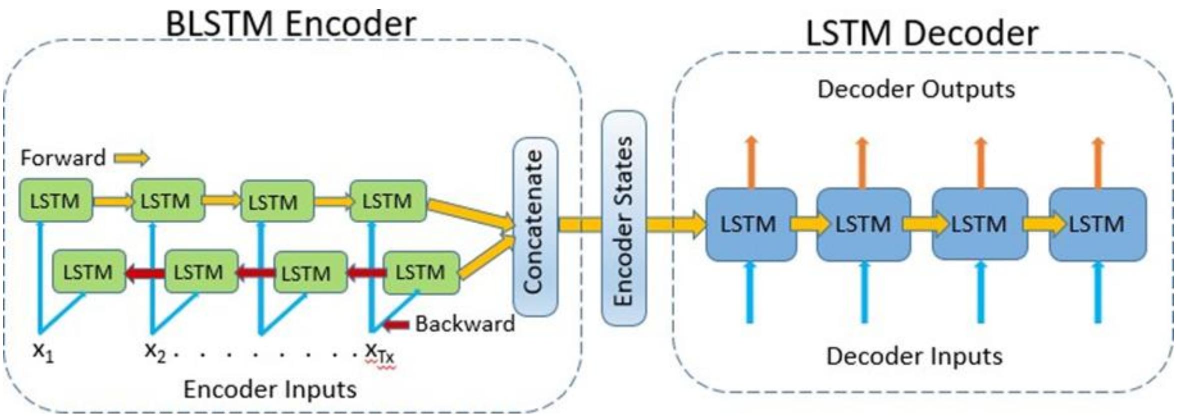


Figure 4.1 Encoder-decoder model

5. Results
5.1 Validation loss graph

A plot of loss on the training and validation datasets over training epochs. A loss is a number indicating how bad the model's prediction was on a single example. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater. The loss is calculated on training and validation and its interpretation is based on how well the model is doing in these two sets. It is the sum of errors made for each example in training or validation sets.

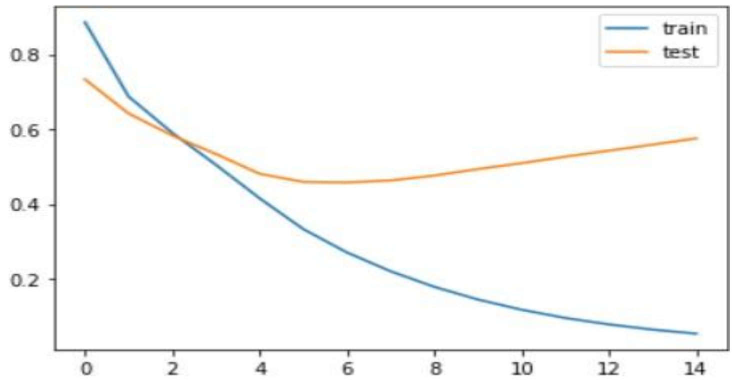


Figure 5.1 Validation Loss Graph

5.2 Text Summarizer Output

```
Text :  an of and program in machine learning and artificial intelligence was a systems engineer at infosys with almost 5 years of work experience pr
ogram and career support helped him transition to a data scientist at tech mahindra with 90 salary hike online power learning has powered 3 lakh care
ers
Original summary : to career in al with 90 salary hike
Predicted summary :  to career in al with 90 salary hike eos


Text :  kunal shahs credit card bill payment platform gave users a chance to win free food from swiggy for one year a delhi bagged this reward after
spending 2000 coins users get one coin per rupee of bill paid which can be used to avail from brands like and more
Original summary : delhi wins free food from swiggy for one year on
Predicted summary :  mumbai player gets with food food from swiggy for 1 year eos


Text :  new zealand defeated india by 8 wickets in fourth odi at hamilton on thursday to win their first match of odi series india lost an internatio
nal match under rohit sharmas captaincy after 12 consecutive dating back to march 2018 match witnessed india getting all out for 92 their seventh low
est total in odi cricket history
Original summary : new zealand end rohit indias winning streak
Predicted summary :  new zealand end rohit to win their first match in odi cricket eos


Text :  with life insurance plan customers can enjoy tax benefits on your paid and save up to on taxes plan provides life cover up to age of 100 year
s also customers have options to against critical and accidental death benefit with a life cover up to age of 80 years
Original summary : life insurance plan helps customers save tax
Predicted summary :  life insurance plan helps customers save tax eos


Text :  speaking about sexual harassment allegations against rajkumar hirani sonam kapoor said i have known hirani for many if it is not true metoo m
ovement will get in metoo movement i always believe a woman but in this case we need to reserve our she added hirani has been accused by an assistant
who worked in sanju
Original summary : have known hirani for yrs what if metoo claims are not true sonam
Predicted summary :  have known hirani for yrs sexual harassment are sonam eos
```

5.3 Evaluation Score

Used Rouge score to get the precision, recall and f-score for respectivepredicted summary based on the original summary.

	text	headlines	summary	f-score	precision	recall
0	an of and program in machine learning and arti...	to career in al with 90 salary hike	to career in al with 90 salary hike eos	0.941176	0.888889	1.000000
1	kunal shahs credit card bill payment platform ...	delhi wins free food from swiggy for one year on	mumbai player gets with food food from swiggy...	0.476190	0.454545	0.500000
2	new zealand defeated india by 8 wickets in fou...	new zealand end rohit indias winning streak	new zealand end rohit to win their first matc...	0.400000	0.307692	0.571429
3	with life insurance plan customers can enjoy t...	life insurance plan helps customers save tax	life insurance plan helps customers save tax eos	0.933333	0.875000	1.000000
4	speaking about sexual harassment allegations a...	have known hirani for yrs what if metoo claims...	have known hirani for yrs sexual harassment a...	0.608696	0.700000	0.538462

6. Conclusion

The Several research conducted for text summary generation in recent days. In proposed model, we have developed automatic text summarization method on new summary dataset. we apply the attentional encoder decoder for the task of automatic text summarization with very promising results. We have used BI- LSTM encoder and LSTM decoder for generation of optimized textsummary. Obtained result is tested as per evaluation metrics. The evaluation metrics considered in the proposed text summarization model are precision, recall and f1-score. Experimentation results are shown clearly that, proposed model is efficient than state of the art model. Further enhancement to the proposed approach can be done by considering bidirectional LSTM processing in decoder side may compute the processing in less computation time.

References

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

AthiraS,SruthyManmadhan,, Deep Learning in Text Summarization -A Survey (AICAAM), April 2019.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan, B. Gutierrez, Krys Kochut, Text Summarization Techniques: A Brief Survey.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network-based sequence model for extractive summarization of documents. In AAAI.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, Bing Xiang, Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond.

RomainPaulus, Caiming Xiong, Richard Socher, A Deep Reinforced Model for Abstractive Summarization, 2017.

Wojciech Kryściński,Romain Paulus, Caiming Xiong, Richard Socher, Improving Abstraction in Text Summarization 2018.

https://towardsdatascience.com/text-summarization-from-scratch-using-encoder-decoder-network-with-attention-in-keras-5fa80d12710e.

https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/.

https://medium.com/@raghavaggarwal0089/bi-lstm-bc3d68da8bd0.

https://towardsdatascience.com/text-summarization-using-deep-learning- 6e379ed2e89c.

https://snyk.io/advisor/python/rouge.