Volume-11,Issue-5(May-19)
ISSN (O) :- 2349-3585

AUTOMATIC TEXT SUMMARIZATION AND TITLE GENERATION USING LSTM

Mayur Deshmukh[1], Harsh Sharma[2], Anuragini Sunar[3], Shweta Somavanshi[4], Prof. D. T. Salunke[5]
Department of Information Technology, RajarshiShahu College of Engineering, Tathawade,Pune, India

Abstract - In this digital world, data growth is steadily increasing with the use of high-performance computing. A huge amountof data is available on the internet and social media which needs to be extracted or summarized in user required form. Text Summarization provides the conversion of large text data in a shorter version without altering the content and meaning of the information. To understand and manually summarize the document, it is very difficult for human beings to read the entire document. Several techniques have been developed for automatic text summarization but, efficiency is always a concern. As there is increase in size and number of documents that are available online, an efficient automatic summarizer is need of the hour. This paper presents an implementation of text summarization system architecture with deep learning model. Article abstracts are used to train the deep learning model to generate the candidate's titles. The purpose of this study is of this study is to implement a deep learning approach forautomatic text summarization. The main objective of this paper is that we propose an automatic text summarization system by applying LSTM to generate short summaries from the abstracts and suitable title for generated summary.

Keywords: Text summarization, LSTM, Topic progression, lexical chains, Seq2Seq.

I. Introduction

Nowadays the size of the document is extremely large and it is troublesome to browse or perceive the whole document therefore, the outline of the document is required. It's time overwhelming and really troublesome task to extract data manually from a great amount of information out there. Therefore, there's a necessity to extract relevant, non-redundant and necessary knowledge. By Text summarization, this task can be achieved in an easier way. Text summarization mechanically converts massive computer file into summarized type,that is well intelligible, clear and complete. In the absence of an Automatic Text Summarizer, it is a terribly grueling job for a human to browse out the whole document to know. The main advantage of summarization lies in the fact that it reduces the user's time in searching the important details in the document. When humans summarize an article, they first read and understand the article or document and then capture the important points. They then use these important points to generate their own sentences to communicate the gist of the article. Even though the quality of the summary generated might be excellent, manual summarization is a time-consuming process. Hence, the need for automatic summarizer is quite apparent. The most important task in extractive text summarization is choosing the important sentences that would appear in the summary. Identifying such sentences is a truly challenging task. Currently, automatic text summarization has applications in several areas such as news articles, emails, research papers, and online search engines to receive a summary of results found. In this paper, we have a tendency to propose a method of text summarization that focuses on the matter of distinguishing the foremost vital parts of the text and manufacturing coherent summaries as well as to generate a title for the article.

II.Literature survey

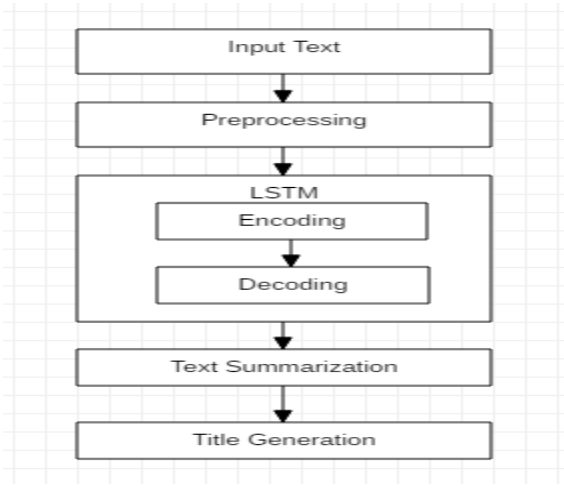
As the amount of information on the web is increasing rapidly day by day in different formats such as text, video, images. It has become difficult for individual to find relevant information of his interest. Suppose user queries for information on the internet he may get thousands of result documents which may not necessarily relevant to his concern. To find appropriate information, a user needs to search through the entire documents this causes information overload problem which leads to wastage of time and efforts [1].To deal

with this dilemma, automatic text summarization plays a vital role. Automatic summarization condenses a source document into meaningful content which reflects main thought in the document without altering information. Thus, it helps user to grab the main notion within short time span. If the user gets effective summary it helps to understand document at a glance without checking it entirely, so time and efforts could be saved [2]. Text summarization process works in three steps: analysis, transformation and synthesis. Analysis step analyses source text and select attributes. Transformation step transforms the result of analysis and finally representation of summary is done in synthesis step. Text summarization approaches generally categorized into: extractive summarization and abstractive summarization. Extractive summarization extracts important sentences or phrases from the source documents and groups them to generate summary without changing the source text. However, abstractive summarization consists of understanding the source text by using the linguistic method to interpret and examine the text. The abstractive summarization aims to produce a generalized summary, conveying information in a concise way. This paper presents extractive and abstractive text summarization techniques. The Extractive based summarization method selects informative sentences from the document as they exactly appear in source based on specific criteria to form summary. The main challenge before extractive summarization is to decide which sentences from the input document is significant and likely to be included in the summary. For this task, sentence scoring is employed based on features of sentences. It first, assigns a score to each sentence based on feature then rank sentences according to their score. Sentences with the highest score are likely to be included in final summary. Following methods are the technique of extractive text summarization [3]. A Term Frequency-Inverse Document Frequency Method Term frequency (TF) and the inverse document frequency (IDF) are numerical statistics presents how important a word in a given document. TF is number of times a term occurs in the document and IDF is a measure that diminishes the weight of terms that occur very frequently in the collection and increases the weight of terms that occur rarely. Then sentences are scored according to product and sentence having high score are included in summary. One problem with this method is sometimes longer sentences gets high score due to fact that they contain a greater number of words. Arunlfo and Ledeneva proposed an approach of term selection and weighting with the help of tf-idf. They used unsupervised learning algorithm to generate non-redundant summary. Sarkar improved news summarization result by using sentence feature along with tf-idf. Issue regarding tf-idf is discussed in her study. Baralis et al. uses weighted item set based model to accumulate information in document. This model connects various significant term then weight is given with if-idf to extract related item set to generate summary. Kamal and Sultana proposed strategy depends on co-event of biological terms in sentences. Three feature terms are used to calculate the frequency of occurrence to generate summary. Jayshree and Murthy calculate term frequency for extracting keywords. GSS is probabilistic feature selection when multiplied with tf-idf gives importance of word to be included in summary. Cluster Based Method: Documents are composed in such a manner that they address different ideas in separate sections. It is natural tothink that summaries should address different themes separated into sections of the document. In case that the document for which summary is being delivered is of entirely different subjects then summarizer assimilates this aspect through clustering. The document is represented using TF-IDF of scores of words [4]. High frequency term represents the theme of a cluster. Summary sentence is selected based on relationship of sentence to the theme of cluster. Cluster based method generate summary of high relevance, to the given query or document topic. Zhang and Li formed a cluster of sentences using K means clustering algorithm. Based on sentence features central sentence of cluster is considered as the summary. Patil and Mahajan extract and group representative sentence from a research article. Summary sentences are generated using local and global search strategy [5]. Author's implemented approach to text summarization by indexing by latent semantic analysis, which is attempted by using word queries and word of documents to overcome the problem of extraction techniques. But there may be chances of selecting unimportant or irrelevant concepts from the document in latent Semantic

ISSN:-2349-3585 |www.ijrdt.org

analysis. Because one word has a lot of meaning and if we can't provide evidence to extract text using latent Semantic techniques then user query may not find expected output. To overcome this unreliable output, the author used Latent Semantic Indexing (LSI). It uses a technique of Matrix based on the method of decomposition of Singular Value [11]. Automatic text summarizers can be described in this classification as approaching the problem at the level of the entity, surface, or discourse. Because it has observed that the current system summarization has many limitations, constraints. And the text summary generated contains poorly linked phrases and is not relevant to the topic [12] This method identifies most important sentences from the given input text using shallow linguistic features in "Summarizing text by ranking text units according to shallow linguistic features". They focused on the degree of connectivity among sentences. It results in a coherent and expected output that reduces non-coherent phrases from the summary result [13]

III.System Architecture



Preprocessing:

i) Segmentation: It is a process of dividing a given document into sentences.

ii) Removal of Stop words:Stop words are frequently occurring words such as ‘a’ an’, the’ that provides less meaning and contains noise. The Stop words are predefined and stored in an array.

iii) Tokenization: The words are assigned tokens or weights according to the usage and importance.

Iv) Word Stemming: converts every word into its root form by removing its prefix and suffix so that it can be used for comparison with other words.

Encoder-Decoder Architecture for Text Summarization:

Develop a basic seq2seq model for text summarization character - level. Use a word - level model that is quite common in the processing of text. Use a model character level for this text. As mentioned above, the architecture of encoders and decoders is a way to create LSTMs for predicting sequences. Encoders read and encode the whole input sequence in an internal representation, usually called the context vector by a fixed - length vector. On the other hand, the decoder reads from the encoder the encoded input sequence and produces the output sequence.The architecture of the encoder-decoder consists of two primary models: one reads and encodes the input sequence to a fixed - length vector, and the other decodes the fixed - length vector and produces the predicted sequence. This architecture is designed to address seq2seq issues.

Title Generation:

After Summary Generation the sentences using one of the three configurations mentioned in the previous subsection several sentences are left over. Terms appearing in the computer-generated title are taken from these sentences.

IV.Algorithm

A.LSTM:

LSTM Neural Networks, stand for Long - Term Memory, are a special type of recurrent neural networks that have recently received much attention within the machine learning community. LSTM networks simply have some internal contextual state cells that behave as long - term or short - term memory cells.

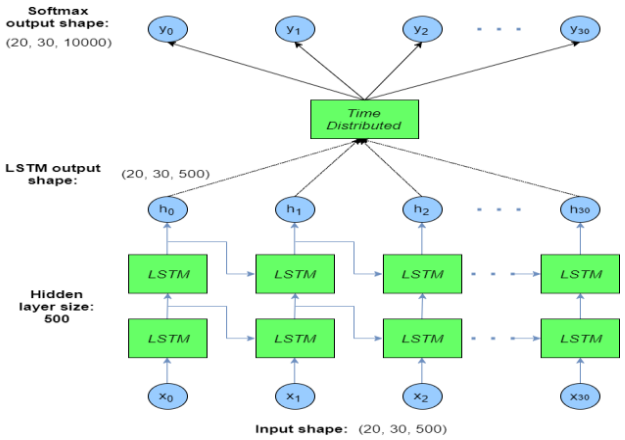


Fig: LSTM Architecture

These cells state modulates the output of the LSTM

network.This is a very important example because we need the neural network prediction to depend not only on very last input, but on the historic context of inputs. Denote ** as element wise multiplication and ignore bias term.

$$i_t=\sigma(x_tU^i+h_{t-1}W^i)$$
$$f_t=\sigma(x_tU^f+h_{t-1}W^f)$$
$$o_t=\sigma(x_tU^o+h_{t-1}W^o)$$
$$\hat{C}_t=\tanh(x_tU^g+h_{t-1}W^g)$$
$$C_t=\sigma(f_t*C_{t-1}+i_t*\hat{C}_t)$$
$$h_t=\tanh(C_t)*o_t$$

The notations are as follows:

i= input gate

f=forget gate

o=output gate

W= The recurring connection in the previous hidden layer and the current hidden layer.

U=The matrix of weight that connects the inputs to the hidden current layer.

C=The candidate stage of the respective LSTM.

h=hidden state derived from the three different gates mentioned above.

Note that they have exactly the same equations with different matrices of parameters. They closely call gates because the sigmoid function squashes such vector values between 0 and 1, and you describe how much of that other vector you want to let through by multiplying them with another vector. The gate input describes how much of the newly calculated state you want to let through for the current input. The forgotten gate defines how much of former state you want to let through. Finally, the output gate defines how much of the internal state (higher layers and next step) you want the external network to be exposed to. All gates are of the same size dh, the size of your hidden state.

V.Experimental Result

Result:

Libraries:In our project, We used Deep learning libraries like Keras and Tensorflow, Numpy for mathematical operation, NLTK for Natural Language Processing.

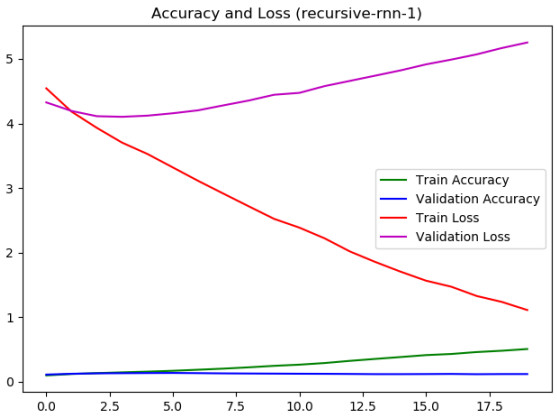


Fig: Accuracy Score

Table: Comparing different model Accuracies with respect to LSTM.

Sr. No.	Model	Accuracy
1.	NB	94.11
2.	SVM	93.31
3.	E-RNN	94.4
4.	CNN-RNN	94.01
5.	LSTM	95.2

VI. Conclusion

Thus we have implemented a text summarization system with deep learning model. Article abstracts are used to train the deep learning model to generate the summary. The contribution of this paper is that we proposed an automatic text summarization system by applying LSTM to generate short summaries from the abstracts and suitable title for the generated summary. Summarization system produces an effective summary in a short time with less redundancy having grammatically correct sentences.

References

[1] Archana AB.1, Sunitha. C.2," An Overview on Document Summarization Techniques", International Journal on Advanced Computer Theory and Engineering (IJACTE), Volume-1, Issue-2, 2013, ISSN (Print): 2319 U 2526.

[2] Josef Steinberger., KarelJezek.," Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", Department of Computer Science and Engineering, University I 22, CZ-306 14 Plzen.

[3] Rafael Ferreira.1, Luciano de Souza Cabral.2, RafaelDueire Lins.3, Gabriel Pereira Silva.4, Fred Freitas.5, George D.C. Cavalcanti.6, Luciano Favaro.7,"Assessing sentence scoring techniques for extractive text summarization",Ex-pert Systems with Applications 40 2013

Elsevier, 5755-5764.

[4] Robert M oro.1, M aria Bielikova.2" Personalized Text Summarization Based on Important Terms Identification", 23rd International Workshop on Database and Expert Systems Applications, 2012 IEEE, 1529-4188.

[5] Landauer T.K., Foltz P.W. and LahamD, "Introduction to latent semantic analysis"Discourse Processes, Vol. 25, pp. 259–284, 1998.

[6]H.Saggion and T. Poibeau, "Automatic text summarization: Past,present and future" Multi-source, Multilingual Information Extraction and Summarization,ed: Springer, pp. 3-21., 2013.

[7] M. Haque, et al.,"Literature Review of Automatic Multiple Documents Text Summarization", International Journal of Innovation and Applied Studies, vol. 3, pp. 121- 129, 2013.

[8] D. R. Radev, et al.,"Introduction to the special issue on summarization", Computational Linguistics, vol. 28, pp. 399-408, 2002.

[9]C.Fellbaum,"WordNet:AnElectronicLexicalDatabase",Cambridge, MA: MIT Press., 1998.

[10]G.A. Miller,"WordNet: A Lexical Database for English", Communications of the ACM, Vol. 38, No. 11: 39-41., 1995.

[11] Landauer T.K., Foltz P.W. and LahamD. ,"Introduction to latent semantic analysis", Discourse Processes, Vol. 25,pp. 259–284,1998.

[12]Rajesh S. Prasad, U. V. Kulkarni, Jayashree R. Prasad, "Connectionist Approach to Generic Text Summarization,", World Academy of Science,Engineering and Technology 55, 2009.

[13] Pankaj Gupta, Vijay Shankar Pendluri, Ishant Vats, "Summarizing text by ranking text units according to shallow linguistic features", Feb. 13~16, 2011 ICACT, 2011.