Comparative Study on Text Summarization using NLP and RNN Methods

N G Gopikakrishna*, Parvathy Sreenivasan, Vinayak Chandran, Yadhu Krishna K P, Sanuj S Dev, Krishnaveni V V#
Department of Information Technology,
College of Engineering Kidangoor, Kottayam, India

Abstract **- Text summarization is the process of generating short, understandable, and accurate summary of a longer text document. Text summarization is having an important role to gain right amount of information within less amount of time. Text data is more difficult to understand due to larger number of characters. So, text summarization is an important tool for today. Text summarization is divided into two subparts that are Extractive Text Summarization (ETS) and Abstractive Text Summarization (ATS). ETS is simpler than ATS. ETS is based on algorithms and it extracts the important words or sentences from the input text document. Where ATS generates summary by itself. This paper represents a comparative study text summarization based RNN and NLP based techniques for text summarization.**

Keywords: - Text Summarization; Extractive Text Summarization; Abstractive Text Summarization; NLP; RNN

I. INTRODUCTION

The Technology has drastically reformed the aspect of human civilization within the last decades. In our lifestyle we are addressing lot of textual data consider the online, comprised of internet sites, news articles, status updates, blogs and so far. It will take most of our time during every day, time could be a vital consider our life, so text summarization may be an important aspect to chop back time for reading this text. Automatic summarizers are designed to reduce the document text size by building a summary that has the foremost important ideas therein document and will provides a higher understanding of lots of knowledge in an exceedingly very short time. There are two main approaches to summarizing text documents; They are: Extractive Methods and Abstractive Methods [1]. Extractive summarizations extract important sentences from the initial documents and group them to produce a summary without changing the initial text. Abstractive summarization has the ability to produce new words that are no longer present on the original file. Abstractive methods can be classified into two categories namely, structured based approach and semantic based approach.

In our proposed system, summarization relies on both NLP and RNN method. Here we are using Natural Language processing (NLP) algorithm for summarization. NLP can be a subfield of computing, information engineering, and applied science concerned with the interactions between computers and human (natural) languages, particularly the way to program computers to process and analyse large amounts of language data Recurrent Neural Network (RNN) is also a category of artificial neural networks where connections between nodes form a directed graph along a temporal sequence this allows it to exhibit temporal dynamic behaviour, Majority of the work has traditionally focused on extractive approaches because of the easy of defining hardcoded rules to select important sentences than generate new ones. Also, it promises grammatically correct and coherent summary but they often summarize long and complex texts well as they are very restrictive [2].

Summary construction is, in general, a flowery task which ideally would involve deep language processing capacities. A summary could also be employed in an indicative way as a pointer to some parts of the initial document, or in an informative way to hide all relevant information of the text [3]. In both cases the foremost important advantage of employing a summary is its reduced reading time. Summary generation by an automatic procedure has also other advantages: the dimensions of the summary are controlled, its content is determinist, and the link between a text element within the summary and its position within the original text are often easily established. Summarization involves compression, so it is vital to be ready to evaluate summaries at different compression rates. Methods for evaluating text summarization (and, indeed, communication processing systems) are also broadly classified into two categories [4]. The first, an intrinsic evaluation, tests the summarization of the system itself. The second, an extrinsic evaluation, tests the summarization supported how it affects the completion of another task.

Text Summarization supported Input Type: during this, had two types, Single Document and Multi Document techniques. Single Document Text Summarization (SDTS): during this sort of Summarization, the length of the input is brief. There will be just one document given because the input for Summarization. Multi Document Text Summarization: this is often a process during which the length of the input on a selected topic is simply too long and then multiple documents are provided as an input for a summarization technique [5].

The evaluation of a summary quality could also be a really ambitious task. Serious questions remain concerning the acceptable methods and form of evaluation. There is a spread of possible bases for the comparison of

summarization systems performance. We are ready to compare a system summary [6] to a personality's generated summary or to a distinct system summary during this paper we are mainly target the text summarization supported the NLP and RNN. After the evaluation of output, we compare the both output and produce the comparative results. Through this study we are able to learn which method is healthier. Through this text summarization method this might help us for save our time in way of life.

II. RELATED WORK

Abu Kaisar Mohammad Masum,Sheikh ,Abujar,Md Ashraful Islam Talukder,AKM Shahariar Azad Rabby and Syed Akhter Hossain in [7] proposed a text summarization method with sequence to sequence RNN. Which is an abstractive summarization method and it use input dataset as amazon fine food reviews dataset, which is available on Kaggle. This consists of bi-directional RNN with LSTMs in encoding layer and attention model in decoding layer, also sequence to sequence model is used to get the summary. Neural Machine Translation, RNN Encoder-Decoder and Sequence to Sequence model are the three kinds of deep learning models used. They have also identified the limitations of their experiment; machine provide a correct summary only short text. The maximum output of long text provides incorrectly. Another important limitation is needed to fixed the text and summary length. Long time and strong hardware configuration need to train the dataset.

Ravali Boorugu and Dr. G. Ramesh in [8] explained in detail some of the remarkable works in arena of text summarization. They have also discussed about the types of text summarization. Text summarization is mainly categorized into three. And they are, based on input type, based on purpose and based on output type. The text summarization techniques can be used to summarize the product reviews in an online market. This can help the customers with reading the long reviews. According to the survey in this paper they conclude that Summarization of Online product review can be achieved with higher accuracy by using Seq2Seq model along with the LSTM and attention mechanism.

Saiyed Saziyabegum and Dr. Priti S. Sajja in [9] provided a review on evaluation methods used for text summarization. This paper discusses about two types of evaluation methods, intrinsic and extrinsic. Extrinsic evaluation evaluates the summarization based on how it influences the completion of some other task such as text classification, information retrieval, answering of question etc. Intrinsic evaluation evaluates the summarization system in of itself. Comparison between the human summary automatically generated summary can be used to find the efficiency of summary. In this paper it also describes about two aspects used in intrinsic evaluation and they are Informativeness and Quality.

YAN DU and HUA HUO in [10] proposed a text summarization method which can be used for summarizing news based on Multi-Feature and Fuzzy Logic. There are mainly four aspects of this work and they are pre-processing the news using NLTK, extraction of features from text like word features, word frequency, word property and so on. And the next aspect is assigning weights of the extracted news feature using Genetic algorithm which is a heuristic search algorithm and at last assigning scores to the sentences using fuzzy logic. Fuzzy logic is imitating human brain's thinking mode in concept judgment and reasoning. This paper also shows a comparison of the proposed method with other methods like MS word, GCD, SOM, System19, System31, SDS-NNGA, System21, and Ranking SVM. And the proposed method performs better than other methods.

Md Ashraful Islam Talukder, Sheikh Abujar, Abu Kaisar Mohammad Masum, Sharmin Akter and Syed Akhter Hossain in [11] discusses about Abstractive summarization. Text summarization methods are divided into two categories and those are Abstractive and Extractive. Abstractive is better than Extractive because it produces a summary like human. In this paper they compared different abstractive methods which are Word Graph Methodology, Semantic Graph Reduction Algorithm and Markov Clustering Principle. They compared these based on Utilized method for eliminate redundancy, semantically correctness, syntactically correctness, Expected correctness percentage and reduced text percentage.

III. PROPOSED APPROACH

Text summarization [12] has been extensively used in various fields like science, medicine, law, engineering, etc. In the proposed system we are doing a comparative study on RNN [7] (Recurrent Neural Networks) based text summarization and NLP [12] (Natural Language Processing) based techniques for text summarization. We will use some evaluation measures to find which model is best for summarization process.

A. System Architecture

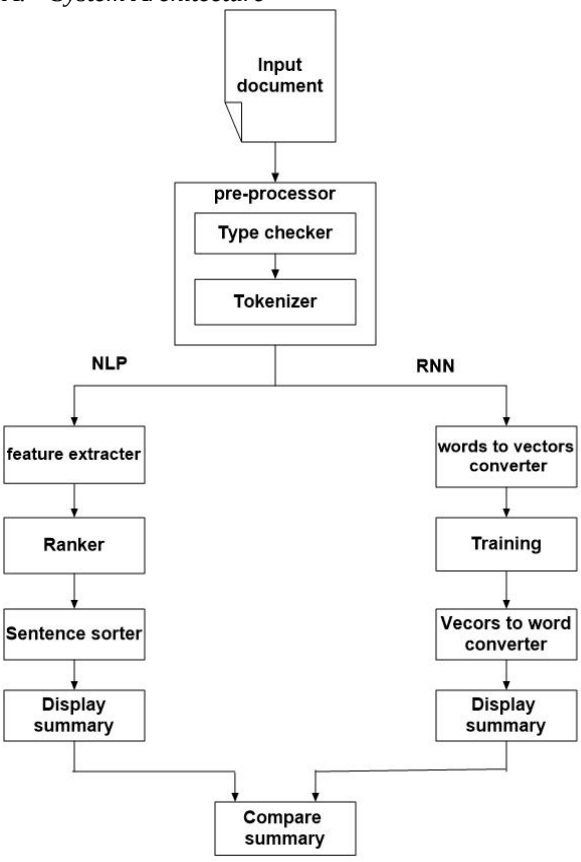

Fig. 1. System Architecture

B. RNN Based System

Traditional neural networks don't have the memory when moving on to next steps. However, for process like text summarization, the sequence of words in input documents is crucial. In this case, we want a model that remembers the previous state. To achieve that, we have to use recurrent neural networks (RNN). RNN uses loops for memorizing the past data. The output from the previous step is passed to the next step as a part of the calculation. Therefore, the information gets stored from the previous timestamp. Basically, traditional RNNs often do not remember information efficiently due to the increasing distance between the connected information. Here the activation functions are nonlinear, due to this reason it is hard to trace back to hundreds or thousands of operations and it leads to a difficulty in getting the information. LSTM [2] (Long Short-Term Memory) networks can convey information in the long term and for a long time. Inside each LSTM cell, there are several linear operations inside for making ease of computational purpose. The previous cell state containing all the information so far smoothly goes through an LSTM cell by doing some linear operations. Inside each LTSM cell makes decisions about what information to keep, and when to allow reads, writes and erasures of information via three gates (input, forget, output) that open and close.

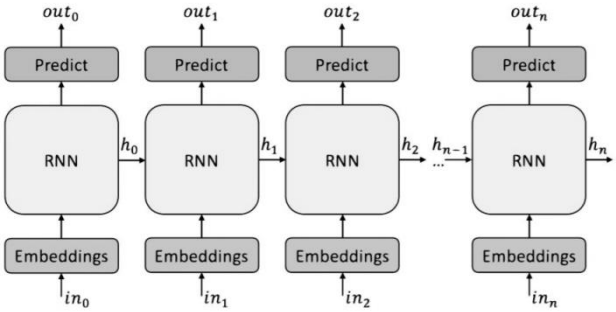

Fig. 2. Recurrent Neural Networks

- Formula for calculating 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 output:

$y_t = w_{hy}h_t$

where:

y_t -> output
w_{hy} -> weight at output layer

1. Cleaning of Source Input
   This is the initial step to clean and pre-process the data.

2. Conversion of Words to Vectors
   The Document after pre-processing is converted to word vectors using simple numbering to get index-based embeddings. So, each word is converted to vectors, which helps in understanding the similarity between the words.

3. Training of RNN
   The obtained word embeddings are trained by using neural network which contains 2 layers RNN's. They are encoder layer and decoder layer. Encoder contains a fixed length of a sentence and decoder contains the sequence of output. These layers are trained and provide a sequence-to-sequence model. The hidden unit used to improve memory capacity and training.

4. Conversion of Vectors to Words
   The obtained summary vectors are converted back to

words from the captured vectors. Which are then combined to form the summary.

5. Display of Summary

The obtained summary is displayed on the screen.

C. NLP Based System

We have seen that because of abundant availability of information's, text summarization features a very vital role in saving user's time and resources. This is an approach to generate summary of lengthy text based on Abstractive text summarization. Natural Language Processing technique using the NLTK [14] which is going to use an enormous tool compartment, and is going for make a favour for people with the entire common dialect handling procedure. NLP uses structure and semantic based ways of summarization. This method has the ability to generate new sentence which improves the focus of summary. Here we are using Natural language processing (NLP) algorithm for summarization. This deals with the interactions of computers and human (natural) languages. That is how to program computers to process and analyse large amounts of natural language data. And also, here we use TF-IDF [12] to find the similarity and relevance in words.TF-IDF is a statistical measure that evaluates how relevant a word is to a document by finding how many times a word appears in a document. The methodology involves Pre-processing, Document set as a graph, Scoring sentences and summary generation.

1. Pre-processing

This is the process of removing irrelevant words, symbols, unnecessary tags etc. from the input document. There can be various steps inside this process. This includes Tokenization, Stopword detection, Stemming.

a) Tokenisation: The whole text is changed to stream of words.
b) Stemming: Categorising same patterns of words.
c) Stopword detection**:** Finding stopwords and remove words with no informative meaning.

2. Ranking

This includes various steps which are:

a) Generate term-document matrix (TD matrix) of the data.
b) Generate a graph for the document to apply PageRank algorithm.
c) Getting the rank of every sentence using PageRank.

3. Sort The Sentences

Finding important sentences and generating summary. This will separate out the sentences that satisfy the criteria of having a score above the threshold.

Calculation of threshold**:**

We take the mean value of normalized scores, any sentence with the normalized score 0.2 more than the mean value is considered.

4. Display Summary:

Sentences that satisfy score above threshold are combined to form the summary.

D. COMPARATIVE STUDY

Here we mainly use three factors to perform the comparative study.

Time: - The amount of time (in seconds) taken to generate the resultant summary.

Length: - The length of the displayed summary, that's total number of words within in the summary.

Quality evaluation [9]: The meaningfulness of the summary by evaluating Redundancy, Grammaticality and Basic Elements etc. This factor is just possible to small scale summary evaluation.

IV. EXPERIMENT AND RESULTS

A. Pre-processing

Pre-processing is the process of cleaning up and preparing text before generating summaries. Here we are using 3 options for taking the input from user. That are input by uploading file (the file can be either in pdf format or in document format), input by pasting the URL of the websites for getting a brief description about that particular site, the next is by just pasting the contents in a text area. The text pre-processing process is mainly had two steps: checking the type of input and tokenizing the sentence. In this paper, we use Natural Language Toolkit (NLTK) to pre-process text.

B. Comparative Study Between Models

We are using as said in the previous section a comparative study on text summarization using NLP and RNN methods. The comparative study between the models is represented in the table-1. Here, we are taking mean of results from five inputs, this consists of how long the summary is, how much time (in seconds) each model took for generating the summary, how much syntactically correct and also percentage level of matching with the human summary of comparing both the models.

NLP system is always taking short time foe generating the summary when comparing with RNN system. In the case of length of the summary RNN only generate most important sentences and hence the summary length will be shorter than NLP. But there can be a situation where missing of some necessary sentences in case of RNN. Both the methods a good semantically correct summary.

Comparing with a set of human summaries NLP based system is having higher percentage level matching with it. Hence, we can summarise that NLP based system is better than RNN based system in providing more understanding and better summary within a short period of time.

V. CONCLUSION AND FUTURE WORKS

Automatic text summarization is the process of

TABLE I. COMPARING THE RESULTS

SYSTEM	AVERAGE TIME(s)	AVERAGE LENGTH (words)	SYNTACTICALLY CORRECT	HUMAN SUMMARY MATCHING (%)
NLP	0.088	112	✓	92%
RNN	17.336	69	✓	75%

reducing the text content and retaining the important point of the document. We have seen that due to abundant availability of data, text summarization has a very vital role in saving user's time, as well as resources and it is an important tool in our day-to-day life. It is able to find a short subset of the most essential information from the entire set and present it in human-readable format quickly.

And the text summarization refers to the technique of shortening long pieces of text. The intention is to create a coherent and fluent summary having only the main point outlined in the document. Then we have seen the use of various algorithms and methods for this purpose and mainly used two types are the NLP and RNN methods. This is two methods to give their individual summary in simplest ways and together gives different types of summaries. User can give any kind of data (URL, paste data and also the text data) for generating summary.

In the comparative study on Text summarization using NLP and RNN. The best one of comparative is the NLP model of Text summarization. Because of the, meaningful sentence and the clear image of the whole topic, it is also similar to the human summarization. In the case of the RNN, it is very short to compare to NLP, but meaning of sentence is very poor, important sentences are missing for the summary. Time consumption RNN is too late than NLP for providing the summary.

With this increasing growth of the Internet, it has made a huge amount of information available. For humans it is difficult to interpret large volume of data. Thus, a tool that can reduce the workload of human are more important to be build. Therefore, this help people in reducing their time consumption and also their work reduction in internet. When researching document, summaries, make the selection process easier. Summary can contain words that are not explicitly present in the original document.

In the future we plan to use multi document summarization. Also, we plan to find some more advanced methods for pre-processing. We would like to add more measurement metrics for comparison.

REFERENCES

[1] Suneetha Manne, Zaheer Parvez Shaik Mohd., Dr. S. Sameen Fatima, "Extraction Based Automatic Text Summarization System with HMM Tagger", Proceedings of the International Conference on Information Systems Design and Intelligent Applications, 2012, Vol. 132, P.P 421-428

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[3] https://www.researchgate.net/publication/220974615 Automatic Text Summarization Using a Machine Learning Approach

[4] Sparck-Jones, K., and Galliers, J. 1996. Evaluating Natural Language Processing Systems: An Analysis and Review. Lecture Notes in Artificial Intelligence 1083. Springer-Verlag

[5] A Survey on NLP based Text Summarization for Summarizing Product Reviews Ravali Boorugu Dr. G. Ramesh

[6] Evaluation Measures for Text Summarization Josef Steinberger, Karel Je˘zek

[7] Abstractive Method of Text Summarization With Sequence To Sequence RNNs - Abu Kaisar Mohammad Masum Sheikh Abujar, Md Ashraful Talukder [2019].

[8] A Survey on NLP based Text Summarization for Summarizing Product Reviews-Ravali Boorugu,Dr. G. Ramesh(2020).

[9] Review on Text Summarization Evaluation Methods- Saiyed Saziyabegum, Dr. Priti S. Sajja [2017].

[10] News Text Summarization Based on Multi-Feature and Fuzzy Logic-YAN DU and HUA HUO (2020)

[11] Comparative Study on Abstractive Text Summarization-Md Ashraful Islam Talukder,Sheikh Abujar,Abu Kaisar Mohammad Masum,Sharmin Akter and Syed Akhter Hossain(2020)

[12] NLP based Machine Learning Approaches for Text Summarization-Rahul, Surabhi Adhikari, Monika [2020].

[13] Automatic Text Summarization Using Natural Language Processing-Pratibha Devihosur, Naseer R [2017].