Intelligent Content Conciser using Pointer Generated Network

Kalpana Devi S, Nithya R, Oviyaa B, Sandhya G

***Abstract**: The combination of best suited architecture and successful algorithm results in the increased nature of efficient learning among the end users. To increase the number of quality learner's text summarization provides the best initiative among the readers and learners. As words and sentences comprise a document, document summarization finds diverse words with different sets of synonyms by performing training activity for the process. The S2S(Sequence to Sequence) training mechanism describes the embedding way of sentences and documents. The pointer generation enhances the new hybrid model for summary extraction. The proposed model implements attention mechanism and uses Recurrent Neural Network with LSTM cells at encoder and decoder. The working model focuses on many factors for summary extraction such as sentence/document similarity, repeatedness, indexing and sentence-context richness. It also keeps track of summarized text using coverage to avoid repetition.*

***Keywords** : abstractive summarization ,hybrid model ,LSTM cells , Neural Network, pointer generator, Recurrent Neural Network , S2S(Sequence to Sequence),*

I. INTRODUCTION

A summary is stated as the actual meaning in the modified brief document containing all the required information and not deviating from the content of original document. A non exceptional content concise will produce exact words as expected to complete the transition. The process of summarization is classified as Abstractive and Extractive. An extractive summary extracts the absolute information from the original report or an article. Extraction also reveals the rate of compression of the original document. The length of the document helps to generate scores for the tokens. To summarize the text, it is a healthy and straightforward method. The original document is scanned completely to divide it as tokens. Each token is assigned with a score. The maximum score is used for generating relevant summary. It also follows some standardization techniques based on their increased feasibility. But, an abstractive summary results in a summary that includes words and phrasesthat are different from those in the source document. Though it is found tough to generate summary than extractive process, it extends its hands in natural language processing. The dataset from CNN news channel is considered here for summary extraction. The dataset is used to avoid a summary of long sentences that is more challenging and requires progress on eliminating repetitions. The pointer generator network copies the source text document through pointing, handling the out-of-vocabulary while maintaining the words. The pointer network maintains a balance among the hybrid model. The subsequent contributions during this work are,

1. To apply to summarization the primary cognitive process encoder - decoder RNN this was initially developed for computational clustering linguistics.

2. The proposed novel method, achieves high efficient performance on machine translation.

II. RELATED WORK

A vast majority of past added account has been extractive, that consists of distinctive key sentences or passages within the supplied document and reproducing them as an outline. The similarities and variations of our planned models with connected work on theoretic account below are analyzed.

Al-Sabahi et.al [3] in his paper segregates the documents in a tree like format comprising of sentences. Maintaining the original meaning of the sentences constructed from tree is the primary task of the work than identifying similarity. Though tree structure consumes time, similarity matching is of less concern. Since the paper focuses on originality maintenance of the document its vocabulary table is restricted in size.

Cristian Felix et al [7] discusses about the mapping of each word with an assigned value. The mapping of assigned values constitutes an expected document. Moreover, the resulted document at least generates the similar summary related to user expectation. But the comparison with other techniques does not yield best results.

Kuan-Yu Chen et.al [14] proposed a model which generates unique words. The concatenation of those unique words generates a summary with set of related sentences. The unique words may be expanded in future by referring the internet or other table dictionaries. The main disadvantage of the proposed paper is that it consumes more time in referring the connected tables to fix a suitable word or sentences to the summary.

Xiaoping Sun et al [21] in his paper explains about the summary extraction. The document is read thoroughly and isolates the sentences and words. Each isolated content is given a priority number.

Manuscript published on November 30, 2019.
* Correspondence Author
Kalpana Devi S, Assistant Professor, CSE Department, Easwari Engineering College, Chennai, Tamil Nadu.
Nithya R, CSE Department, Easwari Engineering College, Chennai, Tamil Nadu.
Oviyaa B, CSE Department, Easwari Engineering College, Chennai, Tamil Nadu.
Sandhya G, CSE Department, Easwari Engineering College, Chennai, Tamil Nadu.

The high prioritized word or sentence helps in summary generation and summary extraction. The low prioritized word or sentences make some assumptions to become a part of summary extraction. The paper helps summary extraction to become more effective than the other available techniques.

Cheng-Ying Liu et al [6]in his proposed work reads the comments from the social network services and constructs a short-sized summary. Comments may be either positive or negative. Positive comments are assigned with high ranks than the negative ones. The summary is defined as the short-sized document with the mere meaning about the targeted article. The summary tries to fit the high ranked sentences to get exact content of the summary. Though there is an existence of negative comments it is also used appropriately without any negligence in the summary modeling. Identifying the exact summary is a tedious process in such processed model.

Glorian Yapinus et al [9]. talk about multiple document summary. Reading a single document and creating summary is simple process. In his work, he exploits the outlines of every document. Each outline is expanded as summary. Each summary is connected together as single integrated document with multiple targeted documents. People by nature assume to become efficient by reading the integrated summary. Since the model focuses on outlines alone for summary generation, time spent is very less. The similar models deviate by consuming more time than the proposed idea.

Kaiz Merchant et al[12] helps the professional lawyers and social workers for their effective professionalism by enabling them to learn more content than the ordinary person through summarized document relevant to their topics. The summarized content is read number of times than the original document to get clear idea and it extremely avoids exhaustive learning. Repeated document learning leads to trained summary concept than the untrained document. The Latent Semantic analysis checks the document synonym and finds the deviation percentage for the learner requirement. Trained summaries are classified for repeated study and create intuitiveness among the professionals.

Dharmendra Hingu et al [8] proposed to extract the vital information from a document. The vital information are identified by tokenizing the sentences as words of frequent nature. The frequent words are meant to embed sensitive information in a sentence. The similar meaning of such identified words is replicated in the second appearance of the sentences. The cited words play a major impact over summarizing the document or a paper. The proposed model yields straightforward results.

Hossein Ebrahimpour-Komleh et al [11]. explained about the feature of every document and its extraction. Feature extraction is the major role in this paper. Evaluating the extracted feature is the next step. The highest evaluated score of the document is meant for creating summary for any document. Every score of evaluation is maintained separately in different aspects. The score is compared every time when the document is summarized. Any change in the score leads to changes in the document. The frequent change leads to effective content of the available concept. But the continuous change consumes more time in the generation of simple summary. It facilitates the increased number of active learners from skipping the intended activity.

The above mentioned review proves and projects the importance of summarization with various techniques / methods. Every aspect is thoroughly investigated and proposed a new idea of pointer generated approach.

III. PROPOSED SYSTEM

The model uses S2S(sequence to sequence) attention mechanism using pointer generator approach to summarize the news articles by training Recurrent Neural Network(RNN). The dataset is taken from CNN/DAILY news which is trained through the RNN to compare the source text to form an efficient summarization using both extractive and abstractive approach.

The model describes to the neural machine interpretation demonstrate utilized in Bahdanau et al. (2014). The proposed model comprises of encoder and decoder for input transitions to yield output. The encoder is designed with hidden state. The two way encoder and the one way decoder play a major role in the process.

The encoder performs initial stage of cleaning and separating the words for reference section generation. The hidden state simulates epochs in the coding section to iterate for similar mapping and indexing. The decoder generates valid and precise summary from the generalized mapping words and documents with less hit over the error rate.

The decoder-vocabulary of every short group is limited to words in the source archives of that group. Furthermore, the most incessant words in the target word reference are included until the vocabulary achieves a fixed size. The point of this system is to diminish the measure of the soft max layer of the decoder which is main computational bottleneck. The model consists of three modules namely,

1. Preprocessing module
2. Training RNN
3. Summary construction

A. PREPROCESSING

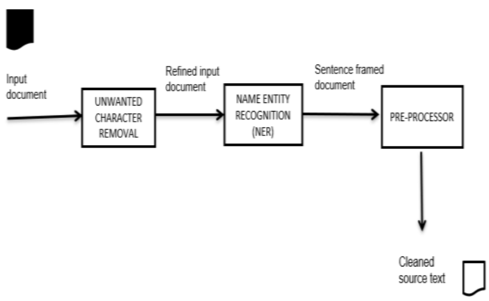

Fig. 1 Preprocessing Module

In this module, the source document is tokenized to break into small pieces called tokenization. Concurrently eliminating certain characters, such as punctuation and special symbols. The main motive of the module is to structure the given source input.

Name Entity Recognition (NER) is used which is capable of finding entity elements from a source document and can analyze the category in which the element belongs. This phase is included in every text processing and natural language processing contains word elimination, removal of nouns and proper names, elimination of word repetition. The objective of this phase is to concise the size of the text.

B. TRAINING RNN

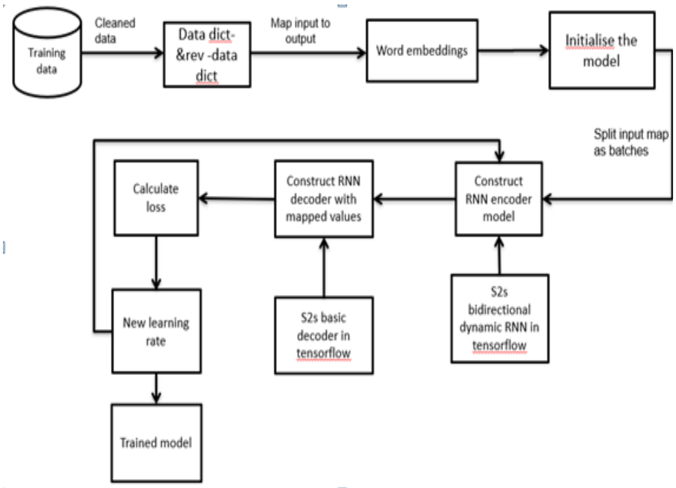

Fig. 2 Training RNN

The input to the process is made sequentially for the information.RNN is constructed with Long Short-Term Memory(LSTM) cells to carry forward content of the document from the beginning to the end. The total input dataset is split into chunks for training purposes. During each epoch the accuracy of the model is measured in the error rate. Each batch is send through the training to measure and improve the accuracy until the error rate is negligible. The input sentence is transformed into a vector value using GLoVe, which assigns similar or closer values to related words. The vectorization of the input can be done either by using Latent Semantic Analysis, Word2vector or GLoVe. In the proposed model GLoVe is used for vectorization due to the better result yielded by it. During the training of the model all new words are stored in a word dictionary to make the process of vectorization simpler during the computation of actual output.

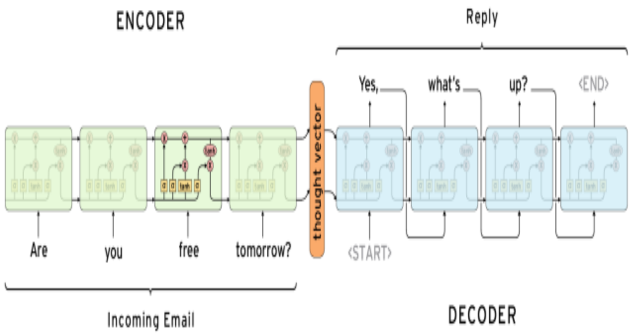

Fig. 3 Encoder-Decoder Model

The model describes to the the artwork. Please verify that the figures and tables you mention in the text actually exist. **Do not put borders around the outside of your figures.** Use the abbreviation "Fig." even at the beginning of a sentence. Do not abbreviate "Table." Tables are numbered with Roman numerals. Include a note with your final paper indicating that you request color printing.

$$W_i \ - \ input\ tokens\ of\ source\ article$$

$$h_i \ - \ Encoder\ hidden\ states$$

$P_{vocab} \ = \ softmax(Vh_i \ + \ b)$is the distribution over vocabulary from which we sample the prediction

The encoder used is bidirectional LSTM-RNN which help maintain the context of the input making the result semantically sound. The RNN has three gates namely the forget gate, input gate and the output gate. The RNN reads the input word by word and encodes it passing the context into the hidden layers in both the forward and backward direction. The major bottleneck is that the encoder is a fixed size encoder making it difficult the main the context in long sentences to overcome this the attention is added.

- $importance_{it} = V * tanh(e_i W_1 \ + \ h_t W_2 \ + \ b_{attn})$
- $AttentionDistributiona_t = softmax(importance_{it})$
- $ContextVectorh_t^* = \sum_i \quad e_i * a_i^t$

The attention mechanism helps the process to remember certain phase of input like names and numbers. It is used in the process of output decoder. It computes a weight on each input source that determines how much attention a word should be paid. The weight median of hidden layers is generated. The context vector holds the distribution generation over the input sentence and holds the whole information.

First, the encoder RNN scan the raw data as word by word producing a sequence of hidden layers. Once the encoder completes the scan, the decoder begins to produce output in a sequential manner. The decoder acquires the input of the previous word of the summary and uses to update the decoder hidden state.

- $P_{vocab}(w) = softmax(V'(V[h_t\ , h_t^*] + b) + b')$
- $For\ the\ loss\ at\ time\ step\ t, loss_t = -log\ P(w_t^*),$ $where\ w_t^* is\ the\ target\ summary\ word$
- $LOSS = \frac{1}{T}\sum_{t=0}^{T} \quad loss_t$

The total loss across all the time steps is calculated by summing the individual loss at each timestamp, which is obtained by the function performed on the predicted value and the expected value. The sequence to sequence mechanism is difficult to copy the text from source code and sometimes even with robust word embedding, the mechanism takes another word for the meaning. The decoding uses the beam search a variation of the greedy search algorithm which is more effective since a specific number of possible predictions that are taken into considerations are lesser. This reduces the number of decision variables to choose from into a fixed k value. The beam search uses a softmax layer to make the prediction easier by choosing the probability of the possible states. This results in a much faster model and also makes it space efficient. The pointer generator is used to copy the text in easier way also able to copy out of vocabulary(OV) from raw data and it is easy to train the data than sequence to sequence mechanism In this module, the dataset CNN/DAILY NEWS is taken and summary of the news articles is compared with the cleaned data from preprocessing module to produce an efficient summarization

C. SUMMARY CONSTRUCTION

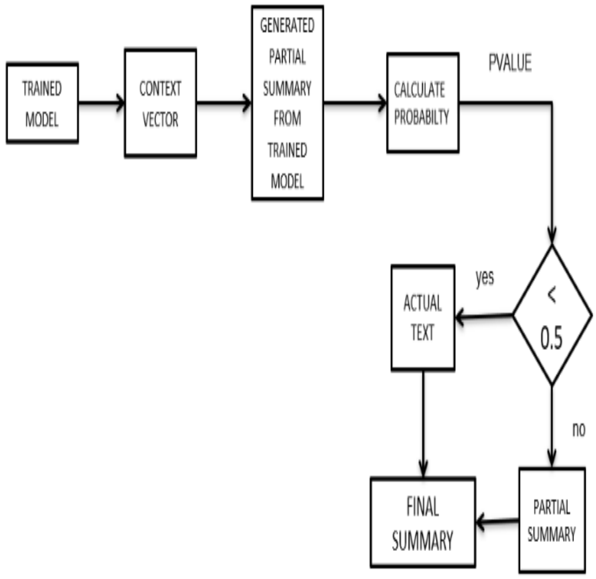

Fig. 4 Summary Construction

The trained model is given to the context vector(the attention mechanism use to produced weighted sum of hidden states). The probability range of word generation vs word copy lies between 0 and 1. P value is used to combine the vocabulary and attention mechanism. P value compares the actual summary of trained data to the partial summary production to form a sequential summarization. The model comparatively produces the output using the attention mechanism with high weighted values.

IV. DATASET

In the dataset each article has a maximum of 50 words, and the corresponding headline a limit of 15 words.The training phase consumes 75% of the dataset and remaining 25% for testing phase.The following tabulation gives the average list of facts such as tokens and sentences.

TABLE 1: Number of tokens on an average Vs Number of sentences

Number of tokens on an average article / document	Number of tokens on average sentences	Number of sentences
781	56	3.75

V. EXPERIMENTAL RESULTS AND DISCUSSION

From the results and discussions, it is proved that the pointer generated approach over summary is more reliable. The performance shows increased variation in producing the actual factual information of the produced document or dataset. The parameter PGEN denotes the generation probability.

During training phase, each token or word is mapped with the PGEN value .The highest PGEN value ensures the availability of factual information similar to the document inputted. Uncertainty at the initial point of a sentence, in connecting the words etc., also plays an important role in the existing mapping process by simply copying the sentence without being tested. But, in our mixture model, the mapping strategy allows the network to copy but with limited restrictions in the summary.In the testing phase, error rate is the main factor to calculate the performance achieved. Minimum error rate is improved by increased epochs. Repeated mapping reduces the error rate and the mapping process is updated when there is an uncertainty such as grammatical views of the factual document. The pointer generated approach clicks many ways for the future enhancements by the increased factor mortality.

TABLE 2:Training error rate

A	B
EPOCH	ERROR RATE
1000	48.8
2000	36
3000	35
4000	32
5000	35.9
6000	26.835
7000	30.391
8000	22.094
9000	21.575
10000	19.957
11000	17.086
12000	14.508
13000	19.026
14000	12.483
15000	16.992
16000	16.651
17000	16.09
18000	15.805
19000	16.098
20000	13.493
21000	10.044
22000	9.907
23000	10.243
24000	8.911

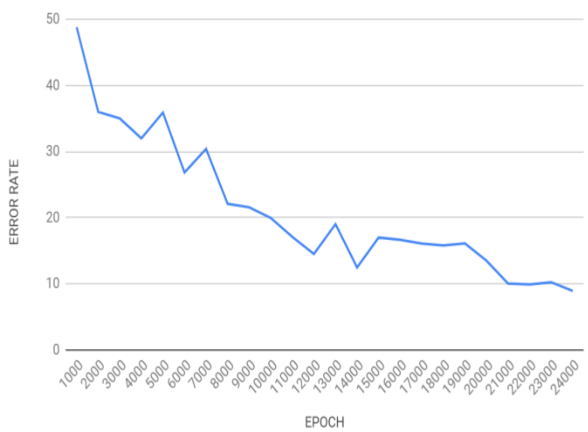

Fig. 5 Error Rate vs Epoch graph

The above figures show the performance of the model during the training phase.

Retrieval Number: D9198118419/2019©BEIESP
DOI:10.35940/ijrte.D9198.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The model once trained can be used for prediction; a sample work of the prediction by the model is as below.

```
the five-time world champion michelle kwan
withdrew from the []# us figure skating
championships on wednesday , but will
petition us skating officials for the chance
to compete at the #### turin olympics .us
business leaders lashed out wednesday at
legislation that would penalize companies for
employing illegal immigrants .
```

Fig. 6 Input to the trained mode

```
['world champion league championships
results', 'us businessmen slam legislation to
override illegal immigrants']
```

Fig. 7 Output from the trained model

The Fig.6 shows 2 news articles with a word limit of 50, and along with few unwanted characters. The Fig.7 shows the output from the trained model having the headlines of the two input articles. Both the generated articles headlines are close to the human generated headlines for the articles.

VI. CONCLUSION

In this model, the LSTM encoder-decoder yields significant results with hybrid summarization for news dataset. The epoch vs error rate illustrates the concise summary of the reported article. The working model outperforms concise summarization applied on text dataset and also provides a way of hint for further extension. Identifying specific problems and grouping on similarity plays a broad way of future enhancements thereby improving the performance incrementally.

REFERENCES

1. AdityaJain,DivijBhatia,Manish K Thakur,"Extractive Text Summarization Using Word Vector Embedding",2017 International Conference on Machine Learning and Data Science (MLDS).
2. Ailin Li , Tao Jiang , Qingshuai Wang , Hongzhi Yu(2016),"The Mixture of Textrank and Lexrank Techniques of Single Document Automatic Summarization Research in Tibetan",2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC).
3. Al-Sabahi, K., Zuping, Z., &Nadher, M. (2018)" A Hierarchical Structured Self-Attentive Model for Extractive Document Summarization (HSSAS)" IEEE Access, Vol.no 6, Page no 24205–24212.
4. Ayman El-Kilany ,ImanSaleh(2012),"Unsupervised document summarization using clusters of dependency graph nodes",2012 12th International Conference on Intelligent Systems Design and Applications (ISDA).
5. CengizHARK ,Taner UÇKAN , Ebubekir SEYYARER , Ali KARCI,"Graph-Based Suggestion For Text Summarization",2018 International Conference on Artificial Intelligence and Data Processing (IDAP).
6. Cheng-Ying Liu , Ming-Syan Chen , Chi-Yao Tseng(2015),"IncreSTS: Towards Real-Time Incremental Short Text Summarization on Comment Streams from Social Network Services",IEEE Transactions on Knowledge and Data Engineering, Vol no. 27 ,issue no 11.
7. Cristian Felix, Steven Franconeri ,EnricoBertini,(2018)," Taking Word Clouds Apart: An Empirical Investigation of the Design", IEEE Transactions on visualization and computer graphics,Vol. no24,Issue no 1 ,Page no 657-666.
8. DharmendraHingu ,Deep Shah , Sandeep S. Udmale(2015),"Automatic text summarization of Wikipedia articles",2015 International Conference on Communication, Information & Computing Technology (ICCICT).
9. GlorianYapinus , Alva Erwin , Maulhikmah Galinium , WahyuMuliady, (2014)"Automatic multi-document summarization for Indonesian documents using hybrid abstractive-extractive summarization technique",2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE).
10. Honey Gupta ,AveenaKottwani , SoniyaGogia ,SheetalChaudhari,"Text analysis and information retrieval of text data",2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET).
11. HosseinEbrahimpour-Komleh , AyoubBagheri(2018) ,"Query-oriented text summarization using sentence extraction technique", 2018 4th International Conference on Web Research (ICWR).
12. KaizMerchant ,YashPande(2018),"NLP Based Latent Semantic Analysis for Legal Text Summarization",2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI).
13. Kamal Sarkar (2018),"Automatic Text Summarization Using Intenal and Extemal Information", 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT).
14. Kuan-Yu Chen , Shih-Hung Liu ,Berlin Chen ,His-Min Wang,Ee-EeJan,Wen-LianHsu,Hsin-His Chen,2015, "Extractive Broadcast News Summarization Leveraging Recurrent Neural Network Language Modeling Techniques",IEEE, Vol. no 23 , Issue no 8, Page no1322 - 1334.
15. M Indu ; K V Kavitha(2016),"Review on text summarization evaluation methods", 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS).
16. Neelima Bhatia, ArunimaJaiswal,"Automatic text summarization and it's methods - a review",2016 6th International Conference - Cloud System and Big Data Engineering (Confluence).
17. NithinRaphal ,HemantaDuwarah ,Philemon Daniel,"Survey on Abstractive Text Summarization",2018 International Conference on Communication and Signal Processing (ICCSP).
18. P. Krishnaveni , S. R. Balasundaram(2016),"Automatic text summarization by local scoring and ranking for improving coherence", 2017 International Conference on Computing Methodologies and Communication (ICCMC).
19. Shohreh Rad Rahimi , Ali Toofanzadeh Mozhdehi, MohamadAbdolahi,"An overview on extractive text summarization",2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI).
20. SiyaSadashivNaik, ManishaNaik Gaonkar,"Extractive text summarization by feature-based sentence extraction using rule-based concept",2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT).
21. Xiaoping Sun , HaiZhuge (2018),"Summarization of Scientific Paper Through Reinforcement Ranking on Semantic Link Network",IEEEAccess,Vol. no 6, Page no 40611 – 40625.