EasyChair Preprint

№ 8143

Extractive Text Summarization by Deep learning

R Akhil

May 31, 2022

Extractive Text Summarization by Deep Learning

AKHIL R

Department of Computer Science & Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India..

Email : akhilradhakrishnan15@gmail.com

***Abstract*— An approach for generating short and precise summaries for long text documents is proposed. Text summarization solves this problem by generating a summary, selecting sentences which are most important from the document without losing the information. In this work, an approach for Extractive text summarization is designed and implemented for single document summarization. It uses a combination of Restricted Boltzmann Machine and Fuzzy Logic to select important sentences from the text still keeping the summary meaningful and lossless. The text documents used for summarization are in English language. Various sentence and word level features are used to provide meaningful sentences. Two summaries for each document are generated using Restricted Boltzmann Machine and Fuzzy logic. Both summaries are then combined and processed using a set of operations to get the final summary of the document. The results show that the designed approach overcomes the problem of text over loading by generating an effective summary.**

***Keywords*— Extractive Text Summarization, *Restricted Boltzmann machine, fuzzy logic, Django framework*, deep learning,**

I. INTRODUCTION

A summary can be defined as a text produced from one or more texts, containing a significant portion of the information from the original text(s), and that is no longer than half of the original text(s) . Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user and task(s). When this is done by means of a computer, i.e. automatically, we call it Automatic Text Summarization. This process can be seen as a form of compression and it necessarily suffers from information loss but it is essential to tackle the information overload due to abundance of textual material available on the Internet.

Text Summarization can be classified into extractive summarization and abstractive summarization based on the summary generated. Extractive summarization is creating a summary based on strictly what you get in the original text. Abstractive summarization mimics the process of paraphrasing a text. Text(s) summarized using this technique looks more human-like and produces condensed summaries. These techniques are much harder to implement than the extractive summarization techniques.

In this paper, we follow the extractive methodology to develop techniques for summarization of factual reports or descriptions. We have developed an approach for single-document summarization using deep learning. So this paper intends to propose an approach by referencing the architecture of the human brain. It is broken down into three phases: feature extraction , feature enhancement, and summary generation based on values of those features. Since it can be very difficult to construct high-level, abstract features from raw data, we use deep learning in the second phase to build complex features out of simpler features extracted in the first phase. These extracted features depend highly on how factual the given document is.

In 2015, S. A. Babar and P. D. Patil proposed an approach to improve the performance of text summarization using Singular value Decomposition and Fuzzy inference system. In 2016, S.P Singh, A. Kumar, A. Mangal, S. Singhal proposed a method for bilingual text summarization using Restricted Boltzmann Machine (RBM). Eleven sentence scoring features were used and given to RBM for feature enhancement. In 2017, an approach for auto text summarization was developed by H. A. Chopade and M. Narvekar using Deep network and Fuzzy logic which provided significant increase in the accuracy of the summary. This work is focused on extractive text summarization. Combining some of the significant works an approach is designed which uses RBM as an unsupervised deep learning algorithm and Fuzzy logic, along with some sentence features to improve the connectivity and the relevance in the sentences. Both pdf and txt files can be summarized.

II. *Related Work*

[1] ***Extractive Text Summarization Using Deep Learning*** **by Nikhil S. Shirwandkar, Student, Mtech Electronics Engineering and Dr. Samidha Kulkarni Associate Professor, Department of Electronics Engineering**.:
We get our proposed methodology from this paper. It uses a combination of Restricted Boltzmann Machine and Fuzzy Logic to select important sentences from the text still keeping the summary meaningful and lossless. Two summaries for each document are generated using Restricted Boltzmann Machine and Fuzzy logic. Both summaries are then combined and processed using a set of operations to get the final summary of the document.

[2] ***Extractive Text Summarization using Sentence Ranking*** **by J N Madhuri and Ganesh Kumar R, Associate Professor, Department of Computer Science and Engineering. :**
We get our dataset from this paper. J.N.Madhuri et.al (2019) has proposed a method to generate an extractive summary using sentence ranking technique using the term frequency after the removal of stop words. It works for any kind of text but cannot semantically distinguish sentences. The evaluation is done using the MSWord summarizer and human summarized summary. The summaries are converted to mp3 format so that it is easy to evaluate or know the summary .

III. PROPOSED METHODOLOGY

Fig below shows the proposed method for Extractive text summarization using Deep Learning

A.

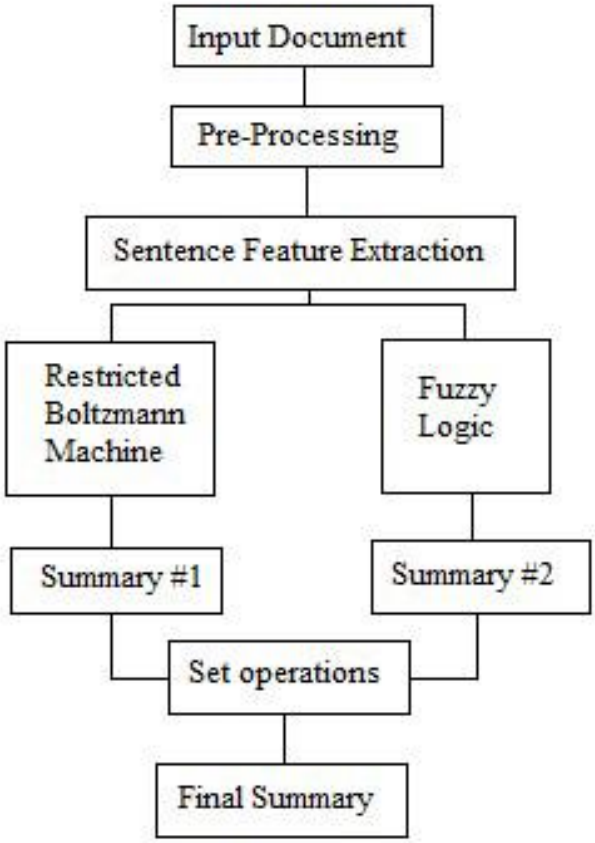

Fig. 1 Proposed Method

1. **Input Document**: The first step of the process for generating the summary is to input a text document with format as .txt. Text Documents used in this work are in English language. The text files are imported using tkinter python library.

2. **Pre-Processing**: The imported text is processed as shown in fig 2

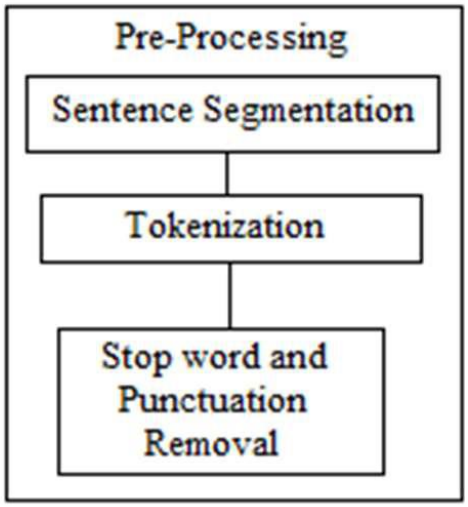

Fig. 2 Pre-Processing

In Sentence Segmentation, the whole text is broken down into sentences and is stored in an array with their respective sentence positions. In Tokenization, the sentences obtained are further broken down into words for some feature calculations. In Stop word and punctuation removal, the commonly occurring words such as the, an, a, but, and, or etc. are removed along with all the punctuation. All the pre-processing steps are carried out using a library in python called as Natural Language Toolkit (NLTK) for Natural Language Processing.

3. **Sentence Feature Extraction:** After Pre-processing the text, sentence features are calculated to find the sentence score.
   After calculating all the sentence features, a sentence feature matrix is formed. In this work, each sentence has nine feature values. The number of feature values is variable based on the number of features used.

;

4. **Restricted Boltzmann Machine (RBM):**

   Restricted Boltzmann Machine is a stochastic neural network that is a network of neurons where each neuron has some random behaviour when activated. RBM has a single layer of visible units and one layer of hidden units. There is no intra_connection between the same layers. Connections between neurons are bidirectional and symmetric (fig-3). That shows the information flows in both directions during testing and training and while the usage of the network they hold the same weights in both directions. The flowchart represents the flow of execution for summarization (fig-4). It starts with the text document and pre-processing steps, which is later converted into the matrix form to create a summary.

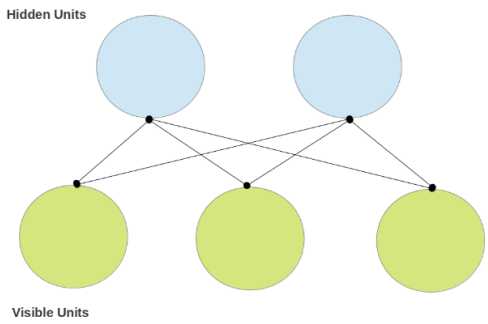

Fig 3: RBM network

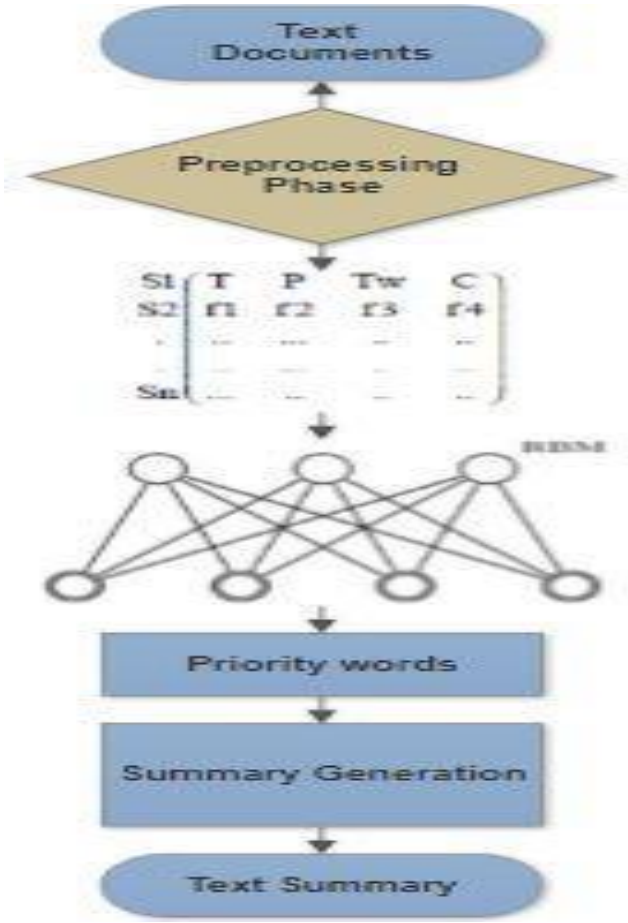

Fig 4: Block diagram for summarization

5. **Summary #1**: Fig. 5 shows the process of generating the first summary. The sum of all enhanced feature values for each sentence in the document is calculated and stored in a list.

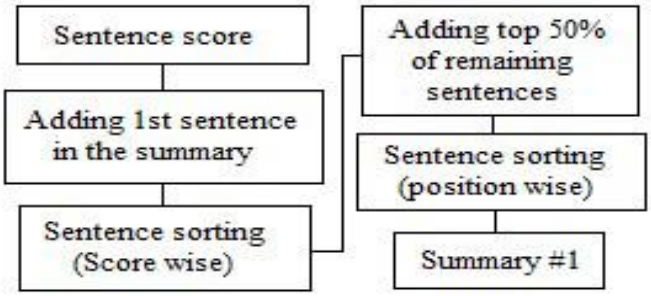

Fig 5. Generating Summary #1

Thus, for each sentence one value is generated which is the score of that sentence. On the basis of scores, sentences are arranged in descending manner. Summary always includes the first sentence as it is the most important sentence and then top 50% of the remaining sentence based on their descending scores are added in the first summary and sorted according to their original position in the document.

6. **Fuzzy Logic**: The feature scores calculated earlier are converted into percentage. Triangular membership functions are used to fuzzify each score into three levels HIGH, MEDIUM and LOW. IF-THEN fuzzy rules are then applied for de-fuzzification to determine whether the sentence is Important, Average or Unimportant e.g. IF (Feature1 is HIGH, Feature2 is HIGH, Feature3 is MEDIUM, Feature4 is MEDIUM), THEN (Sentence is Important). Python library used for fuzzy logic system is skfuzz.

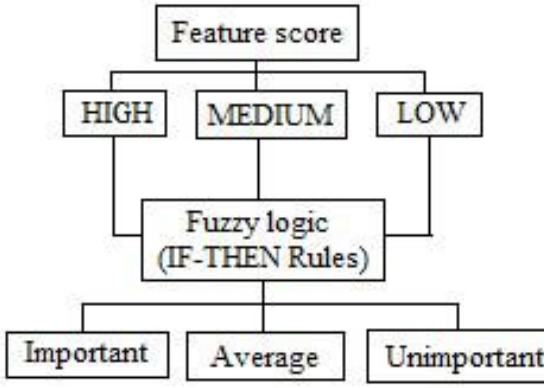

Fig. 6 Fuzzy logic system

7. **Summary #2:** After getting the computation results from the fuzzy logic system, the sentences which fit in "Important" category are added in the second summary according to their original position in the document.

8. **Set operations for Final Summary:** From the first and second summary, a common and uncommon set of sentences are found. Inclusion of common set is made in the Final Sentence.

   The uncommon sentences are sorted position wise and half part of uncommon set of sentences are added in final summary. After adding, the sentences are arranged according to their original position in the text document.

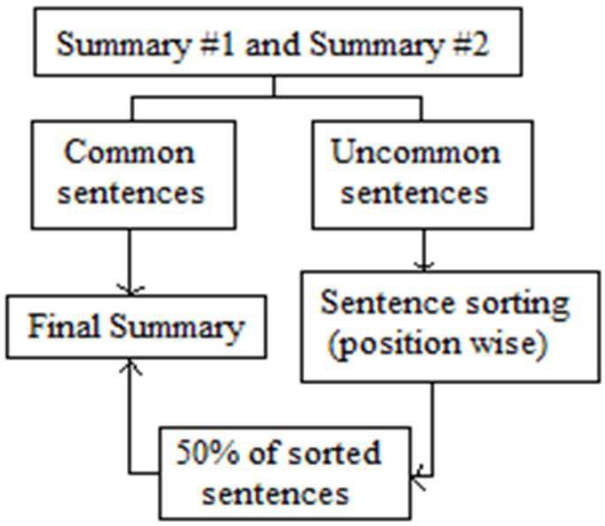

Fig 7 Generating final summary

Algorithm
Step1: Take text file as an input
Step 2: Split the entire document into sentences.
Step 3: Split the sentences into words and remove the stop words from the obtained words.
Step 4: Stemming for all the words after stop word removal is done using nltk toolkit.
Step 5: Calculate the TF-ISF for each word after stemming and take the mean of all the values of a sentence and one value is obtained which is the TF-ISF for one sentence.
Step 6: Calculate the sentence length for each sentence.
Step 7: Generate term-document matrix (TD matrix) of the data
Step 8: Generate a graph for the document to apply PageRank algorithm
step 9: Getting the rank of every sentence using page rank
step 10: Finding important sentences and generating summary using Restricted Boltzmann Machine for summary 1 and using Fuzzy Logic for summary 2
Step 11: Use set operations in summary1 and summary2 to get final summary
Step 12: Print the final summary.

IV. EXPERIMENTAL ANALYSIS AND RESULTS

In the proposed method, nine feature values are calculated for a sentence to get better relevance. RBM is trained for fifteen epochs for generating the first summary of the document. Fuzzy logic IF-THEN rules are used for generating the second summary of the same document and then using set operations a final summary is obtained. The proposed method for Extractive text summarization is implemented for single document along with the original method which uses RBM only. The generated

summaries from both the methods are compared. When doing sample test using older methods, we get below 50% accuracy. When using we use together the 3 prediction algorithms, we get 87% above accuracy in prediction.

V. CONCLUSIONS AND FUTURE SCOPE

In this work, RBM is used as an unsupervised learning algorithm along with fuzzy logic for improving the accuracy of the summary. It is observed that the proposed approach generates short and precise summaries without any irrelevant text.

Overall, the application satisfies all the requirements as mentioned in the requirement specifications and have validated all our functionalities. The application is highly flexible one and is well efficient to make easy interaction with the users. We have designed our application so that users can summarize .txt and .pdf files with minimum effort. We have tried our maximum to make this application user friendly and as simple as possible with very basic looks, but which is flexible enough to be stylized. The software securely stores the data entered for future reference. To make it user friendly we have made it easy to use and so simple that users can see it and understand what it means at the first glance. Extractive Text Summarization is an interactive web application that helps the user to summarize large files and the summary could be download to their respective system. It helps to reduce the time spend on reading and understanding the large file by providing a summary.

The proposed method can be extended for multi document summarization. Documents in different languages can be summarized. Various other features can be used and the method can be combined with other methods for improving the nature of the summary. Also it can be used in Abstractive text summarization.

VI. REFERENCES

[1] *Extractive Text Summarization Using Deep Learning* by Nikhil S. Shirwandkar, Student, Mtech Electronics Engineering and Dr. Samidha Kulkarni Associate Professor, Department of Electronics Engineering.

[2] *Extractive Text Summarization using Sentence Ranking* by J N Madhuri and Ganesh Kumar R, Associate Professor, Department of Computer Science and Engineering.

[3] G. V. Madhuri Chandu, Premkumar, A., K, S. S., and Sampath, N., "Extractive Approach For Query Based Text Summarization", in 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2019.

[4] R. Prasanna Kumar, "A Comprehensive Survey on Topic Modeling in Text Summarization", 5th International Conference on Micro-Electronics and Telecommunication Engineering, Springer book series on "Lecture Notes in Networks and Systems". 2021.

[5] N. Lalithamani, "Text Summarization", Journal of Advanced Research in Dynamical and Control Systems, vol. 10, no. 3, pp. 1368-1372, 2018.

[6] D. Raj and M. Geetha, "A Trigraph Based Centrality Approach Towards Text Summarization", in 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2018

[7] Automatic Document Summarization Using Deep Learning Mechanism with Competent Analysis .Dr. N. Yuvaraj ,Computer Science, Economics 2019

[8] Text Summarization Techniques and Applications Virender Dehru1, Manipal University Jaipur

[9] An Approach for Summarizing Hindi Text using Restricted Boltzmann Machine in Deep Learning

[10] J. Anitha ,1Department of Information Technology

[11] An Approach for Text Summarization using Deep Learning Algorithm. G. PadmaPriya, K. Duraiswamy

[12] D.K. Gaikwad, C.N. Mahender, "A Review Paper on Text Summarization", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue. 3, 2016.

[13] V. Gupta and G.S. Lehal, "A survey of text summarization extractive techniques", Journal of emerging technologies in web intelligence

[14] A Survey of Automatic Text Summarization: Progress, Process and Challenges Progress, Process and Challenges
M. F. MRIDHA ,AKLIMA AKTER LIMA

[15] Text Summarization Using Restricted Boltzmann Machine: Unsupervised Deep Learning Approach
Ashwini Ambekar1. Dept of CSE.