IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303084 |

# Text Summarization in Hindi

**Hutesh Vidhate, Dr. Diptee Chickmurge, Dwarkesh More**

School of Computer Engineering & Technology MIT Academy of Engineering, Pune, India

School of Computer Engineering & Technology MIT Academy of Engineering, Pune, India

School of Computer Engineering & Technology MIT Academy of Engineering, Pune, India

**ABSTRACT**: A Deep Learning approach for text summarization in hindi, has been explored in this paper. Automatic text summarization is a technique which compresses large amounts of text to a shorter summarized format including the important information.

Here we present an approach to design an automatic text summarizer for hindi text that generates a summary by extracting sentences.

The Deep learning approach deals with the abstractive text summarization for a single document based on the Sequence-to-Sequence model with LSTM. The system has the backend model for the processing of the document , a frontend for the user to input the document which is forwarded to the deep learning model through which the summarized document is given to the user. The implementation of the backend is done in python for the creation and the training of the deep learning model. Two evaluation techniques ROUGE and BLUE have been used for the evaluation of the model accuracy.

**KEYWORDS:** Text Summarizer, NLP, LSTM-RNN, Abstractive text summarization, Encoder-Decoder Model, Sequence to Sequence Model.

## I. INTRODUCTION

The amount of data produced in the world is increasing day by day. Data amounting to more than 2.5 quintillion bytes has been produced in the last two years which is almost equal to 90 percent of the total data the world has ever produced. This rise in the amount of data is due to the widespread reach of the internet in the world and many social media apps , blogs , posts related apps etc. There is a large amount of textual data produced due to the extensive use of the internet by various means such as social media, digital marketing, newspapers, reviews, blogs etc.

Data is getting incremented every second and the rate of addition of data to the internet also keeps increasing. With the production and availability of such huge amounts of data there arises a need for processing the data and deriving insights from the textual data. One way to do it is using summarization. A precise summary of data helps readers to understand the gist of the data and derive usable and necessary information from it. With this in mind , a need for making a system which produces precise and accurate summaries arises. The system can save a lot of time for the users and also in cases , where a user needs to store data , it can alsosave a lot of space.

Text summarization is divided into two main types: abstractive and extractive text summarization. In these two branches , there are a lot of data summarization techniques available in the market. Many algorithmic techniques pertaining to machine learning and deep learning have also been employed to procure good results for text summarization and a lot of text summarization methods in many different languages has also been developed through various researchers , these research techniques are published in various the research papers of various professional organizations such as IEEE, Springer, Elsevier to name a few.

Since as said earlier there has been a lot of research and development done for summarization in the English language , we use this as a base for making a system capable of summarizing and producing summarized results in Hindi. In the upcoming documentation we present a deep learning technique for abstractive text summarization for Hindi language using

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303084 |

sequence - to - sequence model with LSTM and attention mechanism.

## II. RELATED WORK

**Summcoder: An Unsupervised Framework For Extractive Text Summarization Based On Deep Auto-Encoders.**
In this paper the author used a novel methodology for generic extractive text summarization of single documents. Their approach generates a summary according to three sentence selection metrics: sentence content relevance, sentence novelty, and sentence position relevance. A sentence ranking and a selection technique are developed to generate the document summary by ranking the sentences according to the final score obtained through the fusion of the three sentenceselection metrics. The comparative study is conducted using the well known Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric for text summarization. Their approach obtained highly competitive performance on the DUC 2002 dataset, and found that it outperforms most of the recent state-of-the-art methods, even those based on supervised learning.

**Automatic Text Summarization Using Supervised Machine Learning Technique For Hindi Language**
This paper discusses single document automatic text summarization for Hindi text using Supervised Machine Learning Technique (SVM).
It was performed on news articles of different categories such as Bollywood, politics and sports.
The performance of the technique was compared with the human generated summaries.
In this experiment, each sentence in the document was represented by a set of various features namely- sentence paragraph position, sentence overall position, numeric data, presence of inverted commas, sentence length and keywords in sentences. Then they were classified into one of four classes namely- most important, important, less important and not important.
From the experimental results, it shows that the summarization is more difficult if they need more compression.
In future, more features like named entity recognition, cue words, context information, world knowledge etc., can be added in their work to improvise the technique and can also be extended to work on multiple documents.

**Dual Encoding For Abstractive Text Summarization**
In this paper the author proposed method employs a dual encoder including the primary and the secondary encoders and 1 decoder. The primary encoder calculates the semantic vectors for each word in the input sequence. The secondary encoder first calculates the importance weight for each word in the input sequence and then recalculates the corresponding semantic vectors. The decoder with attention mechanism decodes by stages and generates a partial fixed-length output sequence at each stage. This paper also discusses dual encoding abstractive text summarization using deep learning.

**News Article Summarization with Attention-based Deep Recurrent Neural Networks**
In this paper the author focuses on an approach for building a text automatic summarization model for news articles, generating a one-sentence summarization that mimics the style of a news title given some paragraphs using Encoder-decoder using LSTM and GRU cells, and with/without attention. In this paper the results were measured by calculating their respective ROUGE scores The authors managed to build and train two relatively complex deep learning models and outperformed our baseline model, which is a simple feed forward neural network.

**An Improvised Extractive Approach to Hindi Text Summarization**
This paper emphasizes an extractive approach for text summarization and its implementation on Java**.** In this paper a survey of the extractive and abstractive text summarization for English language has been carried out to form the basis and understanding of the techniques that can be used for Hindi text summarization.
The system proposed in this paper followed 4 major steps : pre-processing, thematic word generation, sentence scoring and summary generation.
Preprocessing involved removal of stopwords in Hindi(a collection of 170 stopwords were present in the Hindi Wordnet used) . In the thematic word generation steps the words pertaining to the theme of the document to be summarized are extracted and ranked. From the sentences the hindi words are reduced to their radix term and then thematic terms are determined. The sentences are scored on their relevance and ranking based on the previous steps and the summary is generated.
The average accuracy of the system with the expert's manual summary is found to be 85 % and also the average retention ratio is 81.1 %. This is found to be a considerably good percentage when compared with the system without the post processing stage.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303084 |

**A novel technique for multi document hindi text summarization**

Extractive text summarisation using machine learning techniques. The extraction technique of text summarization consists of selecting important sentences from source documents and arranging them in the destination documents.
Fuzzy logic is used for sentence ranking.
The summary generated by the system is found very close to the summary generated byhumans. The Precision, Recall & F-score values show very good accuracy of summary generated by the system. The system achieves an average precision of 73% over multiple Hindi documents.

**Neural Abstractive Text Summarization with Sequence-to-Sequence Models**

In this paper the author provided a comprehensive survey on the recent advances of seq2seq models for the task of abstractive text summarization. They discussed two categories of training methodologies, i.e., word-level and sequence-level training. The first model is an Encoder-decoder using RNN Seq2Seq Model which uses attention and pointing/coping mechanisms. The second model is Encoder-decoder using CNN Seq2Seq Model which uses position embedding mechanism. They measured results by calculating models respective ROUGE and BERT Scores.

**DATASET AND EVALUATION**

The dataset consists of 4515 examples of and contains Author_name, Headlines, Url of Article, Short text, Complete Article.
**Shape of Dataset:** (4515, 6)

**Evaluation metrics** used are ROUGE (Recall- Oriented Understanding for Gisting Evaluation) and BLEU(Bilingual Evaluation Understudy)

**ROUGE (Recall-Oriented Understanding for Gisting Evaluation)**

ROUGE is a recall-oriented measure that works by comparing the number of machine-generated words that are a part of the reference sentence with respect to the total number of words in the reference sentence.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

ROUGE-N measures the number of matching 'n-grams' between our model-generated text and a 'reference'.
The N represents the n-gram that we are using.

$$\textbf{ROUGE-N Recall} \Rightarrow \frac{\text{number of n-grams found in model and reference}}{\text{number of n-grams in reference}}$$

$$\textbf{ROUGE-N Precision} \Rightarrow \frac{\text{number of n-grams found in model and reference}}{\text{number of n-grams in } \textbf{model}}$$

$$\textbf{ROUGE-N F1 Score} \Rightarrow 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303084 |

**BLEU(Bilingual Evaluation Understudy)**

BLEU metric is basically calculated by comparing the number of machine-generated words that are a part of the reference sentence with respect to the total number of words in the machine-generated output.

This metric has precision values between [0,1], where 1 represents a perfect match and 0 represents a complete mismatch.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

## III. IMPLEMENTATION DETAILS

Text Summarization Using an Encoder-Decoder Sequence-to-Sequence Model

- Step 1 - Importing the Dataset
- Step 2 - Cleaning the Data
- Step 3 - Determining the Maximum Permissible Sequence Lengths
- Step 4 - Selecting Plausible Texts and Summaries
- Step 5 - Tokenizing the Text
- Step 6 - Removing Empty Text and Summaries
- Step 7 - Creating the Model
- Step 8 - Training the Model
- Step 9 - Generating Predictions

- We feed in our input (text from news articles) to the Encoder unit. Encoder reads the input sequence and summarizes the information in the form of internal state vectors (in case of LSTM these are called the hidden state and cell state vectors).
- The encoder generates the context vector, which gets passed to thedecoder unit as input. The outputs generated by the encoder are discarded and only the context vector is passed over to the decoder.
- The decoder unit generates an output sequence based on the context vector.
- We can set up the Encoder-Decoder Layer with LSTM in 2 phases:

  - Training phase
  - Inference phase

**Training phase**

In the training phase, we will first set up the encoder and decoder. We will then train the model to predict the target sequence offset by one timestep.

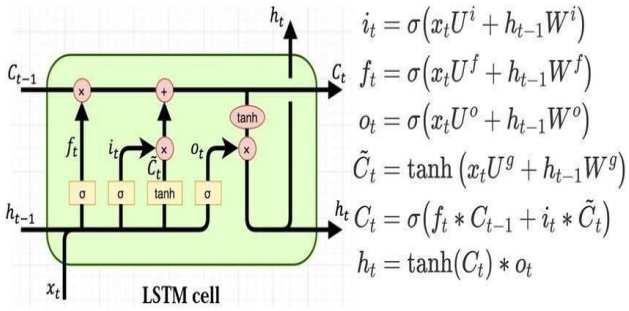
**Inference Phase**

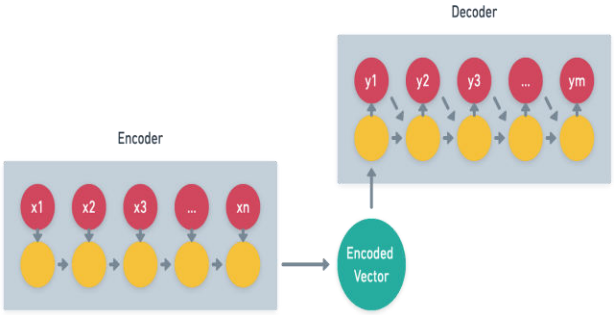
After training, the model is tested on new source sequences for which the target sequence is unknown. So, we need to set up the inference architecture to decode a test sequence. In the end, the inference architecture drives us to decode a test sequence.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303084 |




**Encoder**

The input length that the encoder accepts is equal to the maximum text length which is estimated. This is then given to an Embedding Layer of dimension {(total number of words captured in the text vocabulary) * (number of nodes in an embedding layer)}. This is followed by 3 LSTM networks wherein each layer returns the LSTM output, as well as the hidden and cell states recorded at the previous time steps.

**Decoder**

In the decoder, an embedding layer is defined followed by an LSTM network. The initial state of the LSTM network is the last hidden cell state taken from the encoder. The output of the LSTM is given to a Dense layer wrapped in a TimeDistributed layer with an attached softmax activation function.

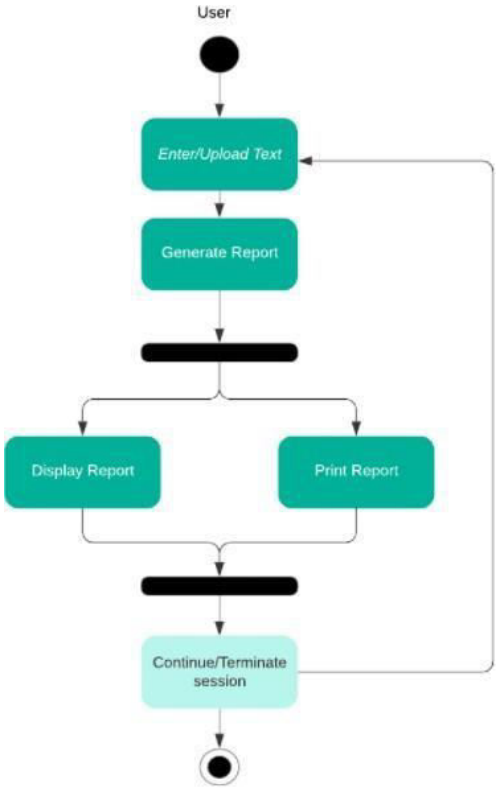
Altogether, the model accepts encoder (text) and decoder (summary) as input and it outputs the summary. The prediction happens through predicting the upcoming word of the summary from the previous word of the summary.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303084 |

**FLOWCHART**




The flowchart describes the flow in which we sample the input and train the model to get the final summarized text

## IV. RESULT

On testing our model against rouge matrix the result which we got was as follows:

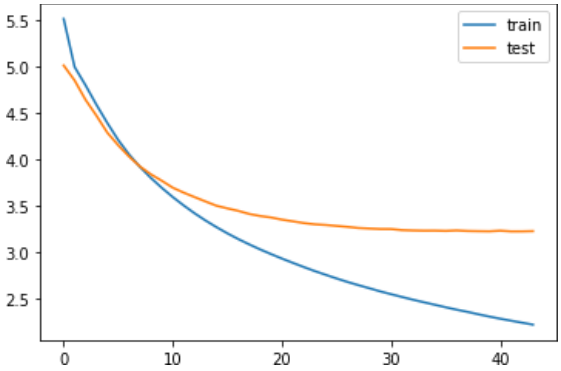



Fig:Train and Validation Loss (Loss v/s Epoch)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303084 |

## V. CONCLUSION & FUTURE SCOPE

In this paper, we present an approach to design an automatic text summarizer for hindi textthat generates a summary by extracting sentences. The Deep learning approach deals with the abstractive text summarization for a single document based on the Sequence-to-Sequence model with LSTM . We obtained results that were measured by calculating ROUGE scores with respect to the actual references. We think the deficiencies currently embedded in our language model can be improved by better
fine-tuning the model, more deep learning method exploration, as well as larger training dataset.The approach gives reasonably good performance. Furthermore the possibilities of the further expansion of the model to text summarization for multiple documents and for other languages can be explored in the future.

## REFERENCES

[1] Joshi, Akanksha & Fidalgo, Eduardo & Alegre, Enrique & Fernández-Robles, Laura. (2019). SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. Expert Systems with Applications. 129. 200-215. 10.1016/j.eswa.2019.03.045.

[2] S. Vijay, V. Rai, S. Gupta, A. Vijayvargia and D. M. Sharma, "Extractive text summarisation in Hindi," 2017 International Conference on Asian Language Processing (IALP), 2017, pp.318-321, doi: 10.1109/IALP.2017.8300607.

[3] Mr. Sarda A.T., Mrs. Kulkarni A.R. "Text Summarization using Neural Networks and Rhetorical Structure Theory" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, June 2015.

[4] Arti S. Bhoir, Archana Gulati
"Multi-document Hindi Text Summarization using Fuzzy Logic Method" International Journal of Advance Foundation And Research In Science & Engineering (IJAFRSE) Volume 2, Special Issue , Vivruti 2016.

[5] Nitika Jhatta, Ashok Kumar Bathla "A Review paper on Text Summarization of Hindi Documents" IJRCAR VOL.3 ISSUE.5 may 2015.

[6] K. Yao, L. Zhang, D. Du, T. Luo, L. Tao and Y. Wu, "Dual Encoding for Abstractive Text Summarization," in *IEEE Transactions on Cybernetics*, vol. 50, no. 3, pp. 985-996, March 2020, doi: 10.1109/TCYB.2018.2876317.

[7] Gaikwad, Deepali K and Mahender, C Namrata. A Review Paper on Text Summarization International Journal of Advanced Research in Computer and Communication Engineering, 5(3). 2016. ACM.

[8] Chopra, S., Auli M. & Rush, A.M. (2016) Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. 2016 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology

[9] Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2020. Neural Abstractive Text Summarization with Sequence-to-Sequence Models. ACM Trans. Data Sci. 1, 1, Article 1 (January 2020),35 pages. https://doi.org/10.1145/3419106

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details