IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11367

Automatic Text Summarizer and Translator

Amit Kumar[1], Vishal Kumar Saha[2], Bhupinder Singh Mann[3], Abhishek Kumar Yadav[4]

Student, IT, Sinhgad Institute of Technology, Lonavala, India[1-4]

Abstract: Text-summarization is one of the most challenging applications in the field of NLP where appropriate analysis is needed of given input text. Result of summarized text always may not identify by optimal functions, rather a better summarized result could be found by measuring sentence similarities. The current sentence similarity measuring methods only find out the similarity between words and sentences. There are two major problems to identify similarities between sentences. These problems were never addressed by previous strategies provided the ultimate meaning of the sentence and added the word order, approximately.

In this project, main objective is to try to measure sentence similarities, which will help to summarize text of any language, but we considered English here.

We have seen several text summarizing software, but the one we intend to develop will comprise of two factors summarization and translation. As English is one of the most popular languages around the globe, it is difficult for a lot of people to read long documents and lengthy texts hence summarization comes in to give a brief informative summary of the language. Not just that, we are also focusing on translation of the output into the simplest form of Hindi language.

Keywords: Text Summarizer, Translator, BERT, BART

I. INTRODUCTION

Automatic text summarization, or just text summarization, is the process of creating a short and coherent version of a longer document. 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 (or users) and task (or tasks). Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. Single-document text summarization is the task of automatically generating a shorter version of a document while retaining its most important information. The task has received much attention in the natural language processing community. Since it has immense potential for various information access applications. Examples include tools which digest textual content (e.g., news, social media, reviews), answer questions, or provide recommendations.

The summarization model could be of two types: -
Extractive Summarization—Is akin to using a highlighter. We select sub segments of text from the original text that would create a good summary
Abstractive Summarization—Is akin to writing with a pen. Summary is created to extract the gist and could use words not in the original text. This is harder for machines to do.

II. PROBLEM STATEMENT

Recently, there has been a surge in the amount of text data from many sources. This amount of text is a valuable source of knowledge and information that must be appropriately summarised In order to be useful. The main goal of this challenge is to automatically summarise material. The procedures are outlined below to help you learn more about them. People are overwhelmed by the vast amount of online information and papers as the Internet has grown dramatically. This expanding availability of documents has demanded exhaustive research in automatic text summarisation.

III. MOTIVATION

Text summarization (TS) is the process of identifying the most salient information in a document set of related documents and conveying it in less space (typically by a factor of five to ten) than the original text. Some of the reasons for motivation of text summarization are as follows:

- To keep up with the world affairs by listening to news.
- People base investment decisions on stock market updates.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11367

- People even go to movies largely on the basis of reviews they've seen.
- With summaries, People can make effective decisions in less time.
- The motivation here is to build such tool which is computationally efficient and creates summary automatically and translate that summary to Hindi.

IV. AIM AND OBJECTIVE

The goal of automatic text summarization is to convert the information from source document into a shorter version with proper semantics as it is so much tedious task for human being to generate the summarized text from large documents. Main advantage of this process is that it reduces the reading time and conveys the proper information to readers by rigorously analysing the text. In addition to the summarization, we are also trying to implement one more feature that will be conversion of the summarized text into Hindi Language. As Hindi is the most spoken language in India (528 million) so we have decided to convert that into Hindi.

V. LITERATURE SURVEY

Work done by Derek Miller (2019) which summarizes lectures. The result of the study is a restful API service and tool operable from the command line to summarize any given lecture [5]. There are two main components in this lecture. Firstly, users can manage the creation and storage of lectures and its corresponding summaries. Secondly, the BERT model produces sentence embeddings from the K-means clustering which are then used in the inference. The input paragraph is firstly tokenized into sentences which are passed to the BERT model to produce the embeddings which are then clustered using K-means clustering. Finally, sentences which are closest to the centroid of each cluster are chosen to be a part of the final summary.

Tacho Jo [6] In this paper the author proposed a particular version of KNN (K Nearest Neighbour) where the words are assumed as features of numerical vectors represents text. The similarity between feature vectors is computed by considering the similarity among attributes as well as among values. Text summarization viewed as the task of classification. The text is partitioned into paragraphs or sentences. Each paragraph or sentence is classified into 'summary or 'non summary' by the classifier. The sentences which are classified into 'summary' are extracted as results from summarizing the text and other text rejected. Improved results are obtained with the proposed version of KNN in text classification and clustering. The modified version of KNN leads to a more compact representation of data item and better performance.

N. Moratanch et al. [7] In this paper the author presents the comprehensive review of extraction-based text summarization techniques. In this paper the author provides survey on extractive summarization approach by categorized them in: Supervised learning approach and Unsupervised learning approach. Then different methodologies, the advantages are presented in the paper. The author also includes various evaluation methods, challenges and future research direction in the paper.

Pankaj Gupta et al. [10] In this paper author has reviewed different techniques of Sentiment analysis and different techniques of text summarization. Sentiment analysis isa machine learning approaching which machine learns and analyse the sentiments, emotions present in the text. The machine learning methods like Naive Bayes Classifier and Support Machine Vectors (SVM) are used. These methods are used to determine the emotions and sentiments in the text data like reviews about movies or products. In Text summarization, uses the natural language processing (NPL) and linguistic features of sentences are used for checking the importance of the words and sentences that can be included in the final summary. In this paper, a survey has been done of previous research work related to text summarization and Sentiment analysis, so that new research area can be explored by considering the merits and demerits of the current techniques and strategies.

The paper by Sandeep Subramanian, Raymond Li, Jonathan Pilault and Christopher Pal (2019) takes a different approach to automatic summarization [8]. Here, in contrast to the traditional encoder-decoder architecture, the decoder is replaced with a language model. Also, instead of using a recurrent neural network (RNN), this paper reports that a feedforward architecture such as convolutional neural network (CNN) or fully attentive models known as transformers would be more efficient (logarithmic) compared to the already in use RNN (Linear in time). To deal with extremely long documents sentence extraction is performed using two different hierarchical document models – pointer networks and sentence classifiers. This extracts sentences from the document that can be used to better condition the transformer language model.

VI. METHODOLOGY

BARTForConditionalGeneration: -

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022
DOI: 10.17148/IJARCCE.2022.11367

INPUT PARAMETER:-

- Input_ids-

Indices of input sequence tokens in the vocabulary. Padding will be ignored by default should you provide it.

- Attention_mask:-

Mask to avoid performing attention on padding token indices. Mask values selected in [0, 1]

- Decoder_input_ids : -

Indices of decoder input sequence tokens in the vocabulary.

- Decoder_attention_mask

Default behavior:- generate a tensor that ignores pad tokens in decoder_input_ids. Causal mask will also be used by default.

OUTPUT PARAMETER:-

- A Seq2SeqLMOutput

BartTokenizer: -

INPUT PARAM:

- vocab_file (str) – Path to the vocabulary file.
- bos_token (str, optional, defaults to "<s>") –The beginning of sequence token that was used during pretraining. Can be used a sequence classifier token.
- eos_token (str, optional, defaults to "</s>") –The end of sequence token

OUTPUT PARAM:-

- List of input IDs with the appropriate special tokens.

BART

BART stands for Bidirectional Encoder and Autoregressive Decoder.

BART is **a denoising autoencoder built with a sequence-to-sequence model** that is applicable to a very wide range of end tasks. Pretraining has two stages

(1) text is corrupted with an arbitrary noising function, and

(2) (2) a sequence-to-sequence model is learned to reconstruct the original text.

VII. APPLICATIONS

Following are the applications of automatic text summarizer and translator:-

- Summaries reduce reading time.
- When researching documents, summaries make the selection process easier.
- Automatic summarization improves the effectiveness of indexing.
- Automatic summarization algorithms are less biased than human summarizers.
- Personalized summaries are useful in question-answering systems as they provide personalized information.
- Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of texts they are able to process

VIII. FUTURE SCOPE AND CONCLUSION

- BERT is undoubtedly a breakthrough in the use of Machine Learning for Natural Language Processing. The fact that it's approachable and allows fast fine-tuning will likely allow a wide range of practical applications in the future.
- The research in the field of NLP is trying to reach human-level every day. With models like this, there is a wide range of applications where it finds use.
- The ease in which these methods can be implemented makes it accessible to not only developers but even business analysts.

ISSN (O) 2278-1021, ISSN (P) 2319-5940

REFERENCES

[1] Anirudh Srikanth, Ashwin Shankar Umasankar, "Extractive Text Summarization using Dynamic Clustering and Co-Reference on BERT" , 2020 IEEE

[2] Nikhil S. Shirwandkar, Dr. Samidha Kulkarni, "Extractive Text Summarization using Deep Learning" , 2018

[3] Reeta Rani and Sawal Tandon, "LITERATURE REVIEW ON AUTOMATIC TEXT SUMMARIZATION" , 2018 International Journal of Current Advanced Research

[4] Nikhil S. Shirwandkar, Dr. Samidha Kulkarni, "Extractive Text Summarization using Deep Learning", 2018 IEEE

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11367

[5] Miller, Derek, "Leveraging BERT for extractive text summarization on lectures." arXiv preprint arXiv:1906.04165, 2019

[6] Taeho Jo, “K Nearest Neighbor for Text Summarization using Feature Similarity.” International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE), 2017.

[7] N.Moratanch, S. Chitrakala, “A Surveyon Extractive Text Summarization.” IEEE International Conference on Computer, Communication and Signal Processing (ICCCSP), 2017.

[8] Sandeep Subramanian, Raymond Li, Jonathan Pilault and Christopher Pal, “On extractive and abstractive neural document summarization with transformer language models." arXiv preprint arXiv:1909.03186, 2019

[9] Akshi Kumar, Aditi Sharma, Sidhant Sharma, Shashwat Kashyap, “Performance Analysis of Keyword Extraction Algorithms Assessing Extractive Text Summarization.” International Conference on Computer, Communication, and Electronics (Comptelix), 2017.

[10] Pankaj Gupta, RituTiwari and Nirmal Robert, “Sentiment Analysis and Text Summarization of Online Reviews: A Survey.” International Conference on Communication and Signal Processing, 2016.