Automatic Text Summarization: A Critical Review

Harsh Hardel,[1] Jay Sharma,[1] Chirag Sahuji,[1] Dr. Rajesh R Prasad[1]

[1]Department of Computer Science and Engineering, School of Engineering, MIT Art, Design and Technology University, Pune, Maharashtra, India- 412201

---***---

Abstract - *Amount of data on internet is increasing day by day. So it becomes more difficult to retrieve important points from large documents without reading the whole document. Manually extracting key components from large documents requires lot of time and energy .Therefore a automatic summarizer is must needed which will produce the summary of document on its own. In this a text is given to the computer and the computer using algorithms produces a summary of text as output. In this paper we have discussed our attempt to prepare an extraction based automatic text summarizer in which paragraphs of documents are split in sentences and then these sentences are ranked on the basis of some features of summarization where higher rank sentences are found to be important one which are used to generate the summary.*

1. INTRODUCTION

World of text is vast and evolving constantly. Most information available on the internet is still in a text format. We humans as the frequent user of this vast information have to go through large and many different documents to gain that information which takes lots of our time as well as energy .Due to this massive increase in text information a Text summarization is must needed to extract the Important points from large set of information or documents. Text Summarization is process of producing a concise summary while maintaining important information and overall meaning of text. We can divide Text summarization into two categories:-

1.1 Abstractive Summarization:-

In this method summary involves the words which are generally not present in actual text it means that it produces summary in a new way by selecting words on semantic analysis just like how humans read articles and then writes the summary of it in their own words.

1.2 Extractive Summarization:-

In this the most important parts of sentences are selected from text and summary is generated from it.

Text Summarization is a one of use cases of Natural Language Processing (NLP).So what is NLP? Natural Language Processing is the branch of Artificial Intelligence concerned with making computers understand text and words in same way as human beings can understand it .Some of Applications of NLP are:-

- Spam Detection
- Machine translation
- Chatbots
- Sentiment Analysis
- Text Summarization

During our process of generating text summarization there comes an important step know as feature extraction which is basically extracting the features of summarization to score the sentences. Following are some features of text summarization:-

Sentence Length - This feature is used to prevent too short or too long sentences from appearing in the summary. basically it is ratio of words in sentence divided by amount of words coming in largest sentences.

Sentence Position- It tells the position of each sentence in the document and also signifies that each sentence has different importance.

Term Frequency- It refers to how many times a term appears in a document..

Title Similarity- Title always gives the clues of text content. If sentence has high degree of intersection with title words then this feature has higher score and more probability to be part of summary.

Sentence Similarity- It tells how much one sentence is similar to another one.

Upper-Case Feature- It assigns higher score to words containing more than one uppercase letters, mainly proper nouns.

Cue phrases- Phrases such as "the most important", "according to survey/study", "in conclusion/summary" indicate significant content of text.

It is very important to understand the intuition of words at different positions as well as to know the similarity between those words, so to solve these linguistic problems in NLP a library is called WordNet. WordNet is a basically a lexical database which tells the semantic relations between words in more than 200 languages. It contains adjectives, nouns, adverbs, verbs grouped together into coherent synonyms where each word has different concept. It is easily available online to download. WordNet differs from thesaurus as thesaurus link words by their meaning but wordnet interlinks sense of words. Wordnet library is also present in NLTK library of python.

2. LITERATURE SURVEY

There has been a lot of development in the field of Text Summarizing in the past 9 years. With every advancement comes a new techniques with more optimized approach. Some of these techniques are mentioned below -

A. Abstractive Text Summarization Approaches with Analysis of Evaluation Techniques [1]

Aim is to summarize large documents to extract the key components from the text which makes it very convenient to understand the context and main points of the text. This paper was published in the year 2021 and the author was Abdullah Khilji.

The approach described in the paper uses a baseline machine learning model for summarizing long text. This model takes raw text as input and gives predicted summary as output. They experiment with both abstractive and extractive based text summarization techniques. For verifying the model they use BELU, ROGUE and a textual entailment method.

Dataset being used is Amazon fine foods reviews which consist of half a million reviews collected over a period of 10 years. The methodology being used is Sequence to Sequence and Generative Adversial. However the paper is missing details of extractive summarization techniques.

B. A Comprehensive Survey on Text Summarization Systems [2]

This paper describes about the classification of summarization techniques and also about the criteria that are important for the system to generate the summary. This paper was published in 2009 and the author was Saeedeh Gholamrezazadeh.

Extractive approach is described as reusing the portions of the main text which is italics, bold phrases, first line of each paragraph, special names, etc. For Abstractive approach involves rewriting the original text in a short version by replacing large words with their short alternatives.

Drawbacks of Extractive summarization includes inconsistencies, lack of balance and lack of cohesion. Examples of extractive summarization are Summ-It applet and examples of abstractive summarization are SUMMARIST.

C. Extractive Automatic Text Summarization Based on Lexical-Semantic Keywords [3]

This paper focuses on Automatic Text Summarization to get the condensed version of the text. The method mentioned also takes into account the title of the paper, the words mentioned in it and how they are relevant to the document. However the summary might not include all the topics mentioned in the input text. This paper was published in the year 2016 by Angel´ Hern´andezCastan˜eda.

Proposed approach firstly passes text to feature generation methods like D2V, LDA, etc. to generate vectors for each sentence. Then it undergoes clustering which measures the proximity among different vectors generated in the previous step. Then LDA is used to get main sentences above the rest all sentences which are then used to generate the summary.

This paper explains the whole process in a very precise manner however there is less information on the methods being used in some steps of the proposed flow. No info on term frequency and inverse term frequency is present in the paper.

D. Abstractive Text Summarization based on Improved Semantic Graph Approach [4]

This paper talks about graph based method for summarizing text. It also tells about how graph based method can be used to implement abstractive as well as extractive text summarization. This paper was published in the year 2018 and is authored by Atif Khan.

According to almost all graph based method text is considered as a bag of words and for summarization it uses content similarity measure but it might fail to detect semantically equivalent redundant sentences. The proposed system has two parts one for making the semantic graph and other part for improving the ranking algorithm based on the weighted graph. Finally after both parts are executed successfully, abstractive summary of the text is generated.

The paper has good information on graph based ranking algorithm and improved versions of it. However there is no mention of textrank algorithm. Also a comparison between graph based ranking algorithm and textrank ranking algorithm is absent.

E. Text Summarizer Using Abstractive and

Extractive Method [5]

In this paper the main motivation is to make computers understand the documents with any extension and how to make it generate the summary. In this method the system uses a combination of both statistical and linguistic analysis. This paper was published in the year 2014 and is authored by Ms. Anusha Pai.

The system introduced in the paper takes text input from user. For summarizing this input it firstly separates the phases, then removes the stop words from the input. After this it performs Statistical and Linguistic analysis to generate the summary. This output is then sent and stored in the database.

The system proposed generates summary according to the input given by the user. This can further be improved by adding synonyms resolution to the model which will treat synonyms words as same. Also multiple documents summarization support can be added.

F. Evolutionary Algorithm for Extractive Text Summarization [6]

In this paper a new possibility is introduced, abstractive text summarization might compose of novel sentences, which are not present in the original document. The method introduced uses unsupervised document summarization method which uses clustering and extracting to generate summary of the text in the main document. This paper was published in 2009 by Rasim Alguliev.

This method uses Sentence clustering to classify sentences based on similarity measures which classifies the sentences in clusters. After this objective function is used to calculate their importance. Along with this Modified Discrete Differential Evolution Algorithm is also mentioned in the paper.

No information is given about Graph based approach and also no comparison between graph based approach and objective function is given. Also the result is less optimized than approaches we saw in other papers.

G. Automatic Keyword Extraction for Text Summarization: A Survey [7]

This paper talks about currently present approaches for summarization. It also talks about Extractive and Abstractive text summarization. This paper was published in the year 2017 and is authored by Santosh Kumar Bharti.

This paper has divided Automatic Text Summarization in 4 types Simple Statistics, Linguistic, Machine Learning and Hybrid. These are techniques in which we can implement Automatic Text Summarization. In Simple Statistics we have Inverse Document Frequency, Relative Frequency Ratio, Term Frequency, etc. In Linguistic Approach we have Electronic Dictionary, Tree Tagger, n-Grams, WordNet, etc. In Machine Learning we have SVM, Bagging, HMM, etc. Hybrid is the combination of the previous three mentioned.

This paper contains of all the techniques present till the date however since 2017 there has been a lot of improvement in this field. This paper needs to be updated with the latest technologies.

H. Automatic Keyword Extraction for Text

Summarization in Multi-document e-Newspapers Articles [8]

This paper was published in the year 2017 and is authored by Santosh Kumar Bharti. It takes about implementing Extractive Summarization in everyday life for summarizing e-Newspaper articles.

This paper mainly focuses on Extractive Summarization. It parses the article and scans for keywords relevant to the title of the Newspaper. Then we use NLP to look for the keywords which have the most weightage, along with this we avoid the stop words which would be present a lot in the article. Summary produced is usually around 17 percent of the whole article.

This paper shows the comparison between TF-IDF, TF-AIDF, NFA and Proposed methods. It uses Fmeasure as a parameter to measure efficiencies of the techniques. It can be clearly seen that the proposed method is more efficient than the other three methods.

The paper only mentioned model implementation on Newspaper articles and not on any other type of articles. Also only Extractive Summarization is used for this model. This model should be tested on different articles and also using different methodologies.

I. Graph-based keyword extraction for single-document summarization [9]

This paper focuses on approach used for selecting keywords from the text input given by the user. The comparison is done between supervised and unsupervised approach to identify keywords. This paper was published in the year 2008 by the writer Marina Litvak.

This approach takes into account some structured document features using graph-based syntactic representation of the text and web documents which improves the traditional vector space model. For the supervised approach, a summarized collection of documents is used to train the classification algorithm which induces a keyword identification model. Similarly for unsupervised approach, HITS algorithm is used on the document graphs. This is done under the assumption that the top-ranked nodes are representing the document keywords.

Supervised Classification model provides the best keyword identification accuracy. But a simple degree-based ranking reaches the highest F-measure. Also only first iteration of HITS is enough instead of running it till we get convergence,

J. Extractive approach for text summarization using graphs [10]

This paper implements Extractive approach but using a different approach. It uses two matrices sentence overlap and edit distance to measure sentence similarity.

This paper was published in the year 2021 by the author Kastriot Kadriu.

The proposed model takes document as input. The document is then tokenized after which lemmatization is performed and is checked for dependency parsing. Then the model checks for sentence overlap and edits distance if necessary. After this graph representation is done. Now we can apply different algorithms to generate summary of the text.

In the paper author has taken into account different methods for generating the summary. This includes pagerank, hits, closeness, betweenness, degree and clusters. Finally the summary is generated. The paper also shows a comparison between different methods and how accurate their summary is. We use F-score to measure their effectiveness.

K. Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization [11]

This paper implemented and compared the performance of three text summarization developed using existing summarization systems to achieve the connectionism. This paper was published in 2010 by the authors Rajesh Shardanand Prasad and Uday Kulkarni.

Three approaches used in paper are based on semantic nets, fuzzy logic and evolutionary programming respectively. Results they got where that the first approach performs better than MS word, second approach was resulting in an efficient system and third approach showed most promising results in case of precision and F-measure. The paper has used the DUC 2002 dataset to evaluate summarized results based on precision and F-measure.

Approaches used in this paper focus only on small details related to general summarization rather than developing an entire summarization system thus are only helpful in research purposes.

L. Feature Based Text Summarization [12]

This paper aims at creating a feature based text summarizer which are applied for different size of documents. This paper was published in the year 2012 by authors Dr. Rajesh Prasad.

This paper follows the extractive way of summarizing the text and utilizes the combination of nine features to calculate the feature scores of each sentences and rank them according to scores they get. The higher rank sentences are part of final summary of text. They have used different types of documents which requires different features to get summary as a dataset for their model. This approach gave better results when compared to MS word in terms of precision, Recall and F-measure in most type of documents.

This paper shows great results for different types of documents but some documents may require features more than this paper has used so further research is needed in this approach.

M. Review of Proposed Architectures for Automated Text Summarization [13]

This paper aims to review various architectures which has been proposed for automatic text summarization. This paper was published in the year 2013 by its authors Tejas Yedke,Vishal Jain and Dr.Rajesh Prasad.

In this paper different techniques for text summarization are discussed and their advantages and drawbacks are also reviewed. DUC 2002 dataset is used to calculate results that which approach performs better for which type of document in terms of precision, Recall and Fmeasure. The limitation of this paper is that it doesn't provide the most effective technique for summarization.

N. Automatic Extractive Text Summarizer(AETS): Using Genetic Algorithm [14]

This paper aims at developing a extractive text summmarizer using Genetic algorithm. This paper was published in the year 2017 by its authors Alok Rai,Yashashree Patil,Pooja Sulakhe,Gaurav Lal,and Dr.Rajesh Prasad.

The approach this paper follows involves the feature extraction, fuzzy logic and genetic algorithm to train machine to produce better results in automatic summarization. This paper defines Genetic algorithms as the search strategies which cop with the population of simultaneous seek positions. According to input file and compression rate given by user the authors resulted in forming a meaningful summary as output text. According to paper genetic algorithm is sentence-choice based techniques and gives best results when text summarization is done.

O. A Novel Evolutionary Connectionist Text Summarizer (ECTS) [15]

This paper aims to create a efficient tool that can summarize large documents easily using evolving connectionist approach.This paper was published in the year 2009 by Rajesh S.Prasad,Dr. U.V Kulkarni,and Jayashree R.Prasad.

In this paper a novel approach is proposed for part of speech disambiguation using a recurrent neural network, which deals with sequential data. Fifty random different articles where used as dataset for in this paper. Authors found the accuracy of ECTS was ranging from 95 to 100 percent which average accuracy of 94 percent when compared to other summarizers.

Though this paper has explained connectionist approach very well but their lies some issues in POS disambiguation and deviations is found in ECTS for couple of sentences.

P. Abstractive method of text summarization with sequence to sequence RNNs [16]

This paper was published in the year 2019 and is authored by Abu Kaisar Mohammad Masum. It takes about how bi-directional RNN and LSTM can be used in encoding layer along with attention model in decoding layer for performing abstractive text summarization on amazon fine food review dataset.

This focus of this paper is Abstractive Summarization by the use of sequence to sequence RNNs. It performs Data Processing in which we Split Text, Add Contractions, Remove stop words and perform Lemmarization. After verifying if the text is purified, we count the size of vocabulary and add word embedding file.

Next step is the addition of special tokens which mark important waypoints in the dataset. After this an Encoder layer and Decoder Layer is present in LSTM and then the Sequence to Sequence model is built.

The above model is then trained with data for generating the response summary. Even though above model performs well for short text it suffers when long text input is given. Other drawback is that it is currently trained for English language but no such summarizer is available for other languages.

Q. Automatic Text Summarization Using Local Scoring and Ranking [17]

This paper was published in the year 2017 and is authored by Diksha Kumar. Instead of building a text summarizer this paper focuses on improving the currently available Automatic Text Summarizer to achieve more coherent and meaningful summaries.

The model introduced in the paper uses automatic feature based extractive text summarizer to better understand the document and improve the coherence. The summary of the given input is generated on the basis of local scoring and local ranking. We can select the top n sentences in the ranking for the summary. Here n depends on the compression ratio of the summarizer.

Feature Extraction can be done on the basis of different criteria. It can be done the basis of frequency of a word appearing in a sentence by selecting the words occurring the most, on the basis of the length of the sentence by avoiding too short or too long sentences, on the basis of the position of the sentence by giving high score to first sentence and less to second sentence and so on, on the basis of sentences overlapping the title or heading which can be considered important and on the basis of similarity of a sentence with respect to all other sentences in the document.

R. A Genetic Fuzzy Automatic Text Summarizer [18]

This paper was published in the year 2009 and is authored by Daniel Leite. This paper focuses on fuzzy based ranking system to select the sentences for performing extractive summarization on the input dataset.

The fuzzy knowledge base used in this model was generated by a genetic algorithm. For fitness function ROGUE informativeness measures were adopted and a corpus of newswire text is being used along with their human generated summaries for defining the fuzzy classification rules.

The paper also talks about SuPor-2 features which uses Na¨ıve-Bayes probabilistic classifier to find the relevance of a sentence to be extracted from the dataset to be in the generated summary. It has 11 features that address either the surface or the linguistic factors that interact with one another to find the relevance of a sentence. For future scope the paper talks about using ideal mutation and crossing rates in the evolution phase for the genetic algorithm. Also membership functions can be explored for modelling fuzzy sets.

S. Enhancing Performance of Deep Learning Based Text Summarizer [19]

This paper was published in the year 2017 and is authored by Maya John. This paper aims to enhance the performance of currently present deep learning model used for text summarization.

In current deep learning models the summary sentences form the minority class which is very small when compared to the majority class which leads to inaccuracy in summary generation. To enhance the performance, data can be resampled before giving it to the deep learning model.

The proposed system is divided in steps for simplicity. These are text preprocessing, feature extraction, resampling and classification. Inside text preprocessing we have tokenization, stop words removal, stemming and lemmatization. Along with this several resampling mechanisms are discussed in the paper which can be used on the dataset for improving the classifier performance.

T. Cut and Paste Based Text Summarization [20]

This paper was published in the year 2000 and is authored by Hongyan Jing. This paper's text summarization model is based on the examination of human written abstract for a specific text.

The model extracts text from the input for generating summary and removes the inessential phrases from the text. The phrases given as output then join together to form coherency. This is done on the basis of a statistically based sentence decomposition program which finds where the phrases of a summary begins in the original text input. By this it produces an aligned corpus of the summary along with the articles used to make the summary.

The model uses Corpus of human-written abstract for analyzing the input. It also uses WordNet for Sentence reduction and Sentence combination. Even though it is very accurate, when we test this model with current models it is very simple and old compared to the ones that are being used currently. A lot of advancement has been done in this field after the time this paper was published.

3. CONCLUSIONS

Automatic text summarization is an attractive topic of research with a variety of business applications. Summarizing can help with a variety of downstream applications, including news digests, report generation, news summarization, and headline development, by condensing enormous amounts of information into short bursts. Summarization algorithms can be divided into two types.

Summaries are created by copying and rearranging portions of the source text in extractive summarizing systems. Second, abstractive summarization methods create new phrases by rephrasing or substituting words not found in the original text. The vast bulk of previous work has been done in the field of extractive summarization since abstractive summarization is more complex than extractive.

The extractive approach is more straightforward because copying big sections of text from the original document ensures grammar and accuracy. Paraphrasing, Generalization and Assimilation are a part of abstractive summarization. Even though abstractive summarization is a more difficult process, thanks to recent improvements in the deep learning field, there have been some progress.

REFERENCES

[1] Abdullah Faiz Ur Rahman Khilji, Utkarsh Sinha, Pintu Singh, Adnan Ali, and Partha Pakray. Abstractive text summarization approaches with analysis of evaluation techniques. In International Conference on Computational Intelligence in Communications and Business Analytics, pages 243–258. Springer, 2021.

[2] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, and Bahareh Gholamzadeh. A comprehensive survey on text

summarization systems. In 2009 2nd International Conference on Computer Science and its Applications, pages 1–6. IEEE, 2009.

[3] Angel´ Herna´ndez-Castan˜eda, Ren´e Arnulfo Garc´ıaHerna´ndez, Yulia Ledeneva, and Christian Eduardo Milla´n-Herna´ndez. Extractive automatic text summarization based on lexical-semantic keywords. IEEE Access, 8:49896–49907, 2020.

[4] Atif Khan, Naomie Salim, Haleem Farman, Murad Khan, Bilal Jan, Awais Ahmad, Imran Ahmed, and Anand Paul. Abstractive text summarization based on improved semantic graph approach. International Journal of Parallel Programming, 46(5):992–1016, 2018.

[5] A Pai. Text summarizer using abstractive and extractive method. International Journal of Engineering Research & Technology, 3(5):2278–0181, 2014.

[6] Rasim Alguliev, Ramiz Aliguliyev, et al. Evolutionary algorithm for extractive text summarization. Intelligent Information Management, 1(02):128, 2009.

[7] Santosh Kumar Bharti and Korra Sathya Babu. Automatic keyword extraction for text summarization: A survey. arXiv preprint arXiv:1704.03242, 2017.

[8] Santosh Kumar Bharti, Korra Sathya Babu, Anima Pradhan, S Devi, TE Priya, E Orhorhoro, O Orhorhoro, V Atumah, E Baruah, P Konwar, et al. Automatic keyword extraction for text summarization in multidocument e-newspapers articles. European Journal of Advances in Engineering and Technology, 4(6):410–427 , 2017.

[9] Marina Litvak and Mark Last. Graph-based keyword extraction for single-document summarization. In Coling 2008: Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization, pages 17–24, 2008.

[10] Kastriot Kadriu and Milenko Obradovic. Extractive approach for text summarisation using graphs. arXiv preprint arXiv:2106.10955, 2021.

[11] Rajesh Shardan and Uday Kulkarni. Implementation and evaluation of evolutionary connectionist approaches to automated text summarization. 2010.

[12] Rajesh Shardanand Prasad, Nitish Milind Uplavikar, Sanket Shantilalsa Wakhare, VY Jain, Tejas Avinash, et al. Feature based text summarization. International journal of advances in computing and information researches, 1, 2012.

[13] Tejas Yedke, Vishal Jain, and RS Prasad. Review of proposed architectures for automated text summarization. In Proceedings of International Conference on Advances in Computing, pages 155–161. Springer, 2013.

[14] Alok Rai, Yashashree Patil, Pooja Sulakhe, Gaurav Lal, and Rajesh S Prasad. Automatic extractive text summarizer (aets): Using genetic algorithm. no, 3:2824–2833 , 2017.

[15] Rajesh S Prasad, UV Kulkarni, and Jayashree R Prasad. A novel evolutionary connectionist text summarizer (ects). In 2009 3rd International Conference on Anticounterfeiting, Security, and Identification in Communication, pages 606–610. IEEE, 2009.

[16] Abu Kaisar Mohammad Masum, Sheikh Abujar, Md Ashraful Islam Talukder, AKM Shahariar Azad Rabby, and Syed Akhter Hossain. Abstractive method of text summarization with sequence to sequence rnns. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pages 1–5. IEEE, 2019.

[17] Hari Disle Sameer Gorule Ketan Gotaranr Diksha Kumar, Sumeet Bhalekar. Automatic text summarization using local scoring and ranking. In 2017 international conference on computing methodologies and communica-tion (ICCMC), pages 59–64. IJRTI, 2019.

[18] Daniel Leite and L Rino. A genetic fuzzy automatic text summarizer. Csbc2009. Inf. Ufrgs. Br, 2007:779– 788, 2009.

[19] Maya John and JS Jayasudha. Enhancing performance of deep learning based text summarizer. Int. J. Appl. Eng. Res, 12(24):15986–15993, 2017.

[20] Hongyan Jing and Kathleen McKeown. Cut and paste based text summarization. In 1st Meeting of the North American Chapter of the Association for Computational Linguistics, 2000.