AUTOMATIC ABSTRACTION BASED TEXT SUMMARIZATION

by Pradnya Sanjeev

Submission date: 08-Dec-2021 02:18PM (UTC+0530)
Submission ID: 1724269346
File name: 57_The_Insightees-FinalReport.pdf (157.81K)
Word count: 2969
Character count: 16626

AUTOMATIC ABSTRACTION BASED TEXT SUMMARIZATION

1st Ritu V Malage
Computer Science
PES University
Bangalore, India
ritu172000@gmail.com

2nd Rajeshwari Raghu
Computer Science
PES University
Bangalore, India
raji160201@gmail.com

3rd Pradnya Sanjeev
Computer Science
PES University
Bangalore, India
npradnya8@gmail.com

4th Sravya Yepuri
Computer Science
PES University
Bangalore, India
sravya.bg@gmail.com

***Abstract*—Globally, massive amounts of data are being generated in all possible areas. Attempting to comprehend such a wide range of content in such a short period of time is difficult. The main motivation behind is how to automate the process of reducing the size of the content without compromising the meaning of the context. In this paper, we will present an automatic text summarization method that reduces the compression rate while optimizing the precision.**

***Index Terms*—Text summarization, Summary, Tokenization, Abstrative Text Summarization**

I. Introduction

Data is to the current century what oil was to the previous one in today's era of contemporary technical breakthroughs. The collecting and dissemination of massive volumes of data has elevated and carried our world today. With such a large volume of data circulating in the digital domain, machine learning algorithms must be developed to mechanically compress lengthy texts and give accurate summaries that can seamlessly pass the intended information. Moreover, the use of text summarisation lessens the time to read, will increase and improves the method of researching for information, and accelerates the quantity of data that can slot in a location. The strategy of extracting these summaries from the first immense text while not losing important data is named as Text Summarization. it's needed for the outline to be fluent, accurate, continuous associated depict the many information. the present search engines are incorporating text summarisation into displaying search results. the present search engines like, Google makes a brief text summarization of the foremost important item and places the summary at the pinnacle of the list of search leads to response to the queries.

In NLP, there are two sorts of ways to summarise text: 1)Extraction-based summarization: The text summarization technique used in this case entails collecting important words from the main source document and combining them to create a summary. The extraction is carried out in accordance with the prescribed metric, with no alterations to the texts or their meaning.

2)Abstraction-based summarization: The technique in this case is paraphrasing and condensing sections of the main source document. When used to text summary in deep learning and NLP problems, the abstraction can overcome any grammar errors found in the extractive method. Text summarising algorithms employ this technique to generate new phrases and sentences that project the most useful information from the original text, similar to how people do so.

The objective is to identify the significant sentences of the text and include them to the summary. We need to note that the summary obtained contains exact sentences from the original text. Our preferred method is to capture all of the different data that can be found in the primary source materials. Despite the difficulty of the dataset, we determined that our model produces vitalizing ROUGE results and summaries when compared to other existing extractive and abstractive text summary methods after conducting rigorous tests inside our experimental setting.

The main steps are as follows, the first step is to extract the keywords from the main source document by applying NLP techniques. Second step is to assemble all the sentences which are having the keywords which are positively labelled. To produce the text summary, the next stage is to create a binary machine learning classifier. Finally, we identify all of the key words and sentences during the testing phase and then classify them. The benefits of automatic text summarization go above and beyond solving apparent problems. It saves time, gives instant response, increases productivity level and ensures all important facts are covered. Best n, maximal marginal relevance, and global selection are all methods for picking phrases for the summary. The top n ranked sentences with the appropriate summary length are chosen in the first approach.

Because abstractive summary necessitates a great deal of real-world knowledge and semantic class analysis, it is more difficult and sophisticated than extractive summarization.Abstractive summarisation is also beneficial to extractive summarisation since the summary is a more exact and precise portrayal of a human-generated summary, making it more understandable. For both of them, accurate summarisation must consist the following:

1)Sentences that maintain the order of the original text's primary ideas and notions.

2) There should be little or no repetition.

3) Sentences with a lot of consistency and coherence.

4)The ability to retain the sense of a document, especially

Identify applicable funding agency here. If none, delete this.

when it is written in long sentences.
The generated summary must be brief and to-the-point while delivering key details from the original content. Text summarising is a unique form of general summarization. Topic identification, interpretation, summary production, and evaluation of the resulting summary are among the main challenges of text summarization.

II. THEORY

A. Problem Area

In today's world, where so much information is available on the internet, it is critical to create a better system for extracting knowledge swiftly and efficiently. Manually extracting the outline of out-sized text pages is extremely challenging for humans. On the Internet, there is a wealth of text material. As a result, sorting out relevant papers from the large number of documents available and aggregating useful information from them is extremely difficult. Automated text summarising is critical for solving the two challenges mentioned above.

B. Problem we seek to solve

The technique of selecting the most significant relevant information in a document or group of linked papers and merging it into a shorter version while keeping the general meanings is known as text summarising. Extractive and abstractive summarization are two types of text summarization.The approach of extractive summarising is selecting key sentences, paragraphs, and other elements from the original document and concatenating them into a shorter form. An abstractive summarization is when you grasp the most significant concepts in a document and then convey them in plain natural language. We first solve extractive summarization using NLP and then do abstractive summarization using transformers on the original text. (Text-To-Text Transfer Transformer (T5) and Bart)

C. Why is it important

Before we move on to the text summarization, we must first define what a summary is. A summary is a text created from one or more texts that delivers key information from the source material in a condensed style. The purpose of automatic text summarization is to reduce the length of the input text while maintaining meaning. The most important benefit of adopting a summary is that it cuts down on reading time.

D. Applications of text summarization

There are many important applications of text summarization in the real world. Some of them are as follows:
1)Media:This summarization can be used to tackle the problem of information overload and "content shock" (too much content created).

2)Newsletters financial research: Many weekly emails begin with an introduction and then feature a carefully curated selection of relevant content. Summarization helps businesses to enrich their newsletters by adding a stream of summaries (rather than a list of links), which is a particularly useful format on mobile.

3)Social media marketing:Multi-document summary is a useful method for swiftly analysing a large number of search results, identifying common themes, and skimming the most relevant aspects.

4)Video scripting:When it comes to writing a script that integrates research from a variety of sources, summarization can be a valuable ally.

5)Medical cases:When it comes to assessing medical situations and directing them to the proper health expert, summarization can be a critical component in the tele-health supply chain.

E. Preprocessing

The dataset is based on text summarization and contains 417 political news articles from the BBC between 2004 and 2005 in the News Articles folder. The Summaries folder has 5 summaries for each article. The title of the article is the first clause of the text. This dataset was constructed using a dataset for data classification that consists of 2225 documents from the BBC news web site matching to stories in five thematic categories from 2004 to 2005, as described in D. Henry Graham Greene and P. Cunningham's study. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," Proc. ICML 2006; all rights to the content of the original articles, including copyright, are owned by the BBC. This dataset contains news items and summaries for each of the five directories: business, entertainment, politics, and sport, each of which has around 385 - 450.txt files.

F. Problem statement and Proposed solution

The problem is automatic abstractive based text summarization. Text summarization will be accomplished in 2 strategies one, the extraction based text summarization and therefore the alternative is that the abstraction based text summarization. Extraction based text summarization is in order to obtain solely the most important phrases from the document and merge these sentences methodically the same method as they seem in the document that will not guarantee continuity within the new summarized document and it's going to not be swish and fluent. Generally there is virtually no affiliation in between two phrases in the outline ensuing within the text is deficient in readability. Whereas because of the abstraction, the predominantly based text summarisation can draw out the key sentences and leads to merge these sentences to form a persistent document.The resolution we have a tendency to propose will believe on the methodology of victimization abstraction based text summarization. Text document or a text file is given as an input to our model. every file can be composed of n variety of sentences and paragraphs.Conversion of each paragraphs into sentences is performed.This is followed

by the text preprocessing. On these sentences tokenization is performed that contributes to tokens. every token will currently be allotted to weights supported the frequency of their look within the sentences. Summation of all the to-ken weights that are relating a sentence are going to be calculated.Filtering out those sentences that have the highest ranks supported the brink worth are wrapped. Finally an attention matrix has to be created which provides North American nation an data of however every sentence is contingent on the opposite sentence.

G. Model Building

For extractive textual summarization we use natural language processing.The text is then preprocessed. It is in these sentences that the tokenization that contributes to the tokens is carried out. Each token is now assigned a weight based on how often it appears in sentences. the rate is calculated. Filter the sentences that have the highest reach based on the threshold. Finally, an attention matrix must be created that tells us how each sentence depends on the other, and an abstract summary is generated as output.

For abstracts based on abstraction, the focus is on identifying the important sections, interpreting the context and reproducing them anew. This ensures that the core information is conveyed in the shortest possible text. The summary sentences are produced , not just extracted from the original text.

Libraries used to implement abstract summaries

Transformers: It offers hundreds of pre-trained models in over 100 languages to perform tasks in texts such as categorization, information extraction, question and answer, summary, translation, text generation, and many others.

Modules for T5

- T5ForConditionalGeneration
- T5Tokenizer
- T5Config

Modules for BART

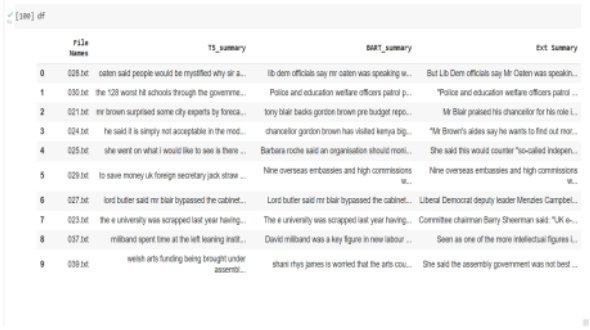
- BartForConditionalGeneration
- BartTokenizer
- BartConfig language

T5 is an abstract aggregation algorithm. It is an encoder-decoder model that is very accurate. Convert any language problems to text-to-text format. This means that if necessary, it will rewrite sentences instead of just taking sentences straight from the original. Text. First we import the modules and then we get the text that is to be summarized.We begin by importing the modules and then obtaining the text that needs to be summarized.then epitomizing the model and the tokeniser. Then we stripit according to the needs and then concatenating the word ”summarize” to raw text.Now encode the input text and generate summary IDs and now we must decode to get the summary of the text.

BART is an abstractive summarization algorithm. BART is a pre-training denoising auto-encoder for sequence-to-sequence models. It is trained by obfuscating text with a haphazard noise function and developing a model to reconstruct the original content. The neural machine translation architecture is based on a typical Transformer-based neural machine translation architecture. Import the modules required and load the module and tokenizer for bart-large-cnn. Now give the original text .encode the inputs and then pass them to model.generate() and then decodes to provide us the summary.

H. Experimental results

On applying the abstractive method for text summarization, the model was imputed with Politics related news reports which had 20 reports. Our model accepted each news report and then was able to reduce the size of the contents to a maximum extent possible even without hindering the actual meaning of the entire content. The minimized version of each news report is termed as a summary which is separately stored in summary files. Based on the sentence score, extraction method resulted in the summary which was bigger than the summary generated by the abstractive method. The size of the extractive method can be reduced by reducing the sentence score. So this method is not so efficient and which may not give the exact summary of that particular text(original text).



[100] df

	File Names	T5_summary	BART_summary	Ext Summary
0	028.txt	oaten said people would be mystified why sir a...	lib dem officials say mr oaten was speaking w...	But Lib Dem officials say Mr Oaten was speakin...
1	030.txt	the 128 worst hit schools through the governme...	Police and education welfare officers patrol p...	"Police and education welfare officers patrol ...
2	021.txt	mr brown surprised some city experts by foreca...	tony blair backs gordon brown pre budget repo...	Mr Blair praised his chancellor for his role i...
3	024.txt	he said it is simply not acceptable in the mod...	chancellor gordon brown has visited kenya big...	"Mr Brown's aides say he wants to find out mor...
4	025.txt	she went on what i would like to see is there ...	Barbara roche said an organisation should moni...	She said this would counter "so-called indepen...
5	029.txt	to save money uk foreign secretary jack straw ...	Nine overseas embassies and high commissions w...	Nine overseas embassies and high commissions w...
6	027.txt	lord butler said mr blair bypassed the cabinet...	Lord butler said mr blair bypassed the cabinet...	Liberal Democrat deputy leader Menzies Campbel...
7	023.txt	the e university was scrapped last year having...	The e university was scrapped last year having...	Committee chairman Barry Sheerman said: "UK e-...
8	037.txt	miliband spent time at the left leaning instit...	David miliband was a key figure in new labour ...	Seen as one of the more intellectual figures i...
9	039.txt	welsh arts funding being brought under assembl...	shani rhys james is worried that the arts cou...	She said the assembly government was not best ...

Fig. 1. Summarized output texts of the models

III. Conclusion

Abstractive summarization is the foundation of the most practical text summarising systems. Abstraction-based text summarization is more structurally accurate than extractive text summarization, making it a hot topic in research. Automatic text summarising is a long-standing problem, but the research focus is shifting from extractive to abstractive summarization. The benefits of having an automatic text summarization system increase the need for such systems; one of the most important advantages of employing a summary is that it reduces reading time and provides a rapid guide to the exciting content. We first apply the extractive text summarization technique for summarising the input text. It simply extracts the sentences from the original one and not very clearly, henceforth we move to the abstractive technique of text summarization where it identifies the vital sections, interpret the context and reproduce in a new approach which is evident and concise. It’s a well-known truth that present abstractive text summarization models frequently produce erroneous results. This can happen at the entity level (additional

entities are created) or at the entity relation level (extra entities are generated) (context during which entities occurring is incorrectly generated).

The extractive and abstractive summarising approaches are the subject of this survey work. An extractive summary is a technique for selecting key sentences from a large body of literature. The significance of sentences is determined using applied mathematics and semantic features. Within the last decade, several variations of the extractive technique have been tried. Nonetheless, it's difficult to quantify how much more interpretative expertise helps to performance at the phrase or text level. The resulting summary could suffer from a lack of cohesiveness and semantics if Natural Language Processing (NLP) is not used. The resulting summary may not be balanced if the contents contain many subjects. It is critical to determine the true weights of different features because the quality of the final summary is dependent on it. We should urge for more time when deciding on feature weights.

The most difficult aspect of text summarising is ensuring that content from a wide range of textual and semi-structured sources, as well as databases and web pages, is summarised in the appropriate manner (language, format, size, and time) for each user.The text summarising software should generate an effective summary in the shortest amount of time and with the least amount of repetition possible. Extrinsic and intrinsic measurements are frequently used to evaluate summaries. While intrinsic approaches use human analysis to determine summary quality, extrinsic methods use a task that is primarily based on performance measures, such as an information retrieval oriented task, to do so.

ACKNOWLEDGMENT

We owe a huge debt of gratitude to our teacher, Dr. Gowri Srinivasa, and the Teaching Assistants, as well as the entire Data Analytics team, for providing us with valuable knowledge and resources to help us gain a deeper understanding of the concepts we've learned, as well as for providing us with constant supervision and guidance, as well as for providing us with necessary project information and support.

We'd like to express our appreciation to Prof. Shylaja Sharath, the Head of the Computer Science Department, for allowing us to participate in such projects. Our gratitude and appreciation extend to our project development colleagues as well as those who have volunteered their time and skills to assist us.

CONTRIBUTIONS

Ritu V Malage: Literature Survey, Preprocessing of the dataset, Exploratory Data Analysis, Trying out different models for abstractive text summarization, IEEE Final Report.

Pradnya Sanjeev: Literature Survey, Exploratory Data Analysis, Testing the models, Preprocessing of the dataset, IEEE Final Report.

Rajeshwari Raghu: Literature Survey, Trying out different models for extractive and abstractive text summarization, IEEE Final Report.

Sravya Yepuri: Literature Survey, Trying out different models for extractive text summarization, IEEE Final Report.

The project collaboration was an associated positive effort and a constructive process. It gave us a chance to learn. This project was a great experience in learning how applied mathematical analytics is employed in the real-world scenarios.

REFERENCES

[1] https://medium.com/analytics-vidhya/seq2seq-abstractive-summarization-using-lstm-and-attention-mechanism-code-da2e9c439711

[2] https://www.machinelearningplus.com/nlp/text-summarization-approaches-nlp-example/

[3] https://towardsdatascience.com/simple-abstractive-text-summarization-with-pretrained-t5-text-to-text-transfer-transformer-10f6d602c426

[4] https://www.sciencedirect.com/topics/computer-science/text-summarization

[5] https://analyticsindiamag.com/hands-on-guide-to-extractive-text-summarization-with-bertsum/

AUTOMATIC ABSTRACTION BASED TEXT SUMMARIZATION

ORIGINALITY REPORT

17%	2%	7%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Essex Student Paper	7%
2	www.kaggle.com Internet Source	2%
3	Submitted to The NorthCap University, Gurugram Student Paper	2%
4	Submitted to Liverpool John Moores University Student Paper	2%
5	Gupta, Vishal, and Gurpreet Singh Lehal. "A Survey of Text Summarization Extractive Techniques", Journal of Emerging Technologies in Web Intelligence, 2010. Publication	1%
6	Venkat N. Gudivada. "Natural Language Core Tasks and Applications", Elsevier BV, 2018 Publication	1%
7	"Advances in Artificial Intelligence and Data Engineering", Springer Science and Business	1%

Media LLC, 2021
Publication

8	Babar, S.A., and Pallavi D. Patil. "Improving Performance of Text Summarization", Procedia Computer Science, 2015. Publication	1 %
9	P. B. Tumpa, S. Yeasmin, A. M. Nitu, M.P. Uddin, M. I. Afjal, M. A. A. Mamun. "An Improved Extractive Summarization Technique for Bengali Text(s)", 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018 Publication	<1 %
10	Lorick Jain, H V Srinivasa Murthy, Chirayush Patel, Devansh Bansal. "Retinal Eye Disease Detection Using Deep Learning", 2018 Fourteenth International Conference on Information Processing (ICINPRO), 2018 Publication	<1 %

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		