CHAPTER ONE

GENERAL INTRODUCTION

Author: AbdulAfeez Babalola

Co-Author / Corresponding Author: Jimoh Abdulganiyu

1.1 INTRODUCTION

Text summarization is basically the task for a software to reduce a large amount of text into a meaningful short summary which helps the reader to understand what information the document contains in a short descriptive form so as to save the efforts and time of the user. There are mainly two fundamental ways to automatically summarize text. Those are extraction and abstraction. In text summarization, Extractive methods work on choosing between a subset of words, phrases, or sentences present in the document in its original text to produce an extracted summary. While in, Abstractive summaries attempt to improve the coherence among sentences by eliminating redundancies and clarifying the contest of sentences. abstractive methods the algorithm builds an internal semantic representation and then they use natural language generation techniques that is in the technique the machine acts as if it has a human brain and has the ability to generate correct and meaningful summary by understanding the text present in the document. This process creates a summary that is far closer to what a person might actually extract and present as a summary of text.

According to Lee et.al, 2009. The amount of information on the Internet continues growing, but much of this information is redundant. Therefore, we need new technologies to efficiently process information. The automatic generation of document summaries is a key technology to overcome this obstacle. Given this, it is essential to develop automated methods that extract the most relevant information from a text, researched by Automatic Text Summarization (ATS). ATS is an active research area that deals with single- and multi-document summarization tasks.

Lloret et.al, 2012. Text summarization (TS) is the process of automatically creating a compressed Version of one or more documents. It attempts to get the "meaning" of documents. Essentially, TS techniques are classified as Extractive and Abstractive. Extractive summaries produce a set of the most significant sentences from a document, exactly as they appear. Abstractive summaries attempt to improve the coherence among sentences by eliminating redundancies and clarifying the contest of sentences. It may even produce new sentences to the summary.

Lin et.al, 2004. Were used to perform the quantitative assessment of the studied methods. The qualitative assessment was performed by four people who analyzed each original text and selected the sentences that they feel ought to be in the summary. The qualitative evaluation is done by counting the numbers of sentences selected by the system that match the human gold standard. Processing-time performance of each of the algorithms implemented is also taken into account. It is important to notice that Lloret and Palomar (2012) and Nenkova and McKeown (2012) present two recent and comprehensive surveys on text summarization. They do not present any assessment of any sort of the techniques and this research work targets at filling in such an important gap.

Therefore, recurrent neural network (RNN) might also be a promising alternative for extracting text summarization. Extractive RNN are able to generate near similar performance with large-scale approximate summaries as with small scale human extracted summaries. This generated summary includes verbal innovations. Research by this date has focused primarily on extractive methods, which are pertinent for image collection summarization, text summarization and video summarization. Currently, the extractive summaries are commonly used because they are easier to create. Due to this in this work we focus on them.

1.2 BACKGROUND OF THE STUDY

Nallapati et al. (2017), use a hierarchical encoder comprising of word and sentence level RNNs as well as incorporate an abstractive co-training mechanism to determine which sentences from a document should be included in the extractive summary while also taking into consideration features like sentence position, content, and saliency.

According to Verma and Lee (2017), the gold-standard summaries of DUC01 and DUC02 employ approximately 9% of words not found in the original documents. Consequently, the level of maximum similarity will be less than 100%, and even more, if compared from several gold-standard summaries, the upper bounds will be lower for any AETS method.

The growth of the internet yielded a massive increase of large amounts of information available, especially regarding text documents such as news articles, electronic books, scientific papers, blogs among others. Due to the huge volume of information in the internet, it has become unachievable to efficiently extract useful information from the huge mass of documents. Therefore, it is necessary to extract information in a clear and concise way, due to the fact that extractive summaries produce a set of the most significant sentences from a document, exactly as they appear, which helps and allow users to manage, save time and resources and also gives human a high level idea of what the document is about.

1.3 STATEMENT OF THE PROBLEM

According to Ahmed E. et.al (2018) View of a significant increase in the burden of information over and over the limit by the amount of information available on the internet, there is a huge increase in the amount of information overloading and redundancy contained in each document. Extracting important information in a summarized form would help a number of users. It is therefore necessary to have proper summaries. Subsequently, many research papers are proposed continuously to develop new approaches to automatically summarize the text. “Automatic Text Summarization” is a process to create a shorter version of the original text (one or more documents) which conveys information present in the documents.

Therefore, this study will focus on Extraction of Text Summarization using Natural Language Toolkit, attempting to achieve more consistent, non-recurring and meaningful summaries.

1.4 MOTIVATION FOR THE STUDY

Extractive text summarization is one of the applications of natural language processing and has become increasingly popular in the last few years for information condensation. Text summarization is a process of reducing the size of an original document and producing a summary by retaining important information of the original document.

In order to achieve the advantage of extracting text summaries, Natural Language Processing Toolkit will be implemented to extract and summarize intermediate representation of original text document**.**

1.5 AIM AND OBJECTIVE OF STUDY

1.5.1 AIM

The aim of this research work is to Extract Text Summarization Using Natural Language Toolkit.

1.4.2 OBJECTIVE

To achieve the set aim, the research work will cater for these objectives:

- To set features containing valuable information of original text will be collected
- To Select Machine Learning approach using Natural Language Toolkit which will be used to extract text summarization in large document
- To achieve more consistent, non-recurring and meaningful summaries.

1.5 Methodology of the study

This study adopts the use of the Natural Language Toolkit. For successful execution of this study, the following steps shall be followed:

Data Collection: input of Most publicly available dataset for long documents and articles shall be considered.

Date Preprocessing: Pre-processing is the most primary step in any summarization method. Pre-processing is carried out to clean text, normalization and grammatical errors from documents. Pre-processing methods applied are tokenization, stop word removal and stemming.

Feature Extraction: After preprocessing steps, each sentence of the document is represented as an attribute vector of features.

NLTK (Natural Language Toolkit): is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries

Extractive Summary: Final extractive summary of the document will be displayed.

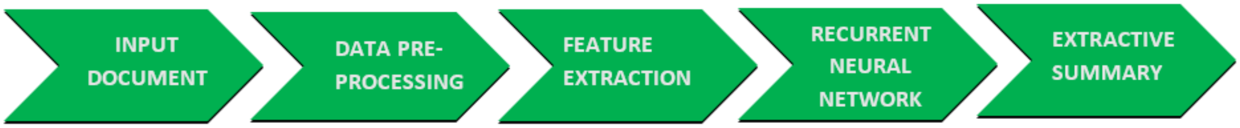

Fig 1: Diagram Showing The Research Methodology

1.6 Conceptual Model

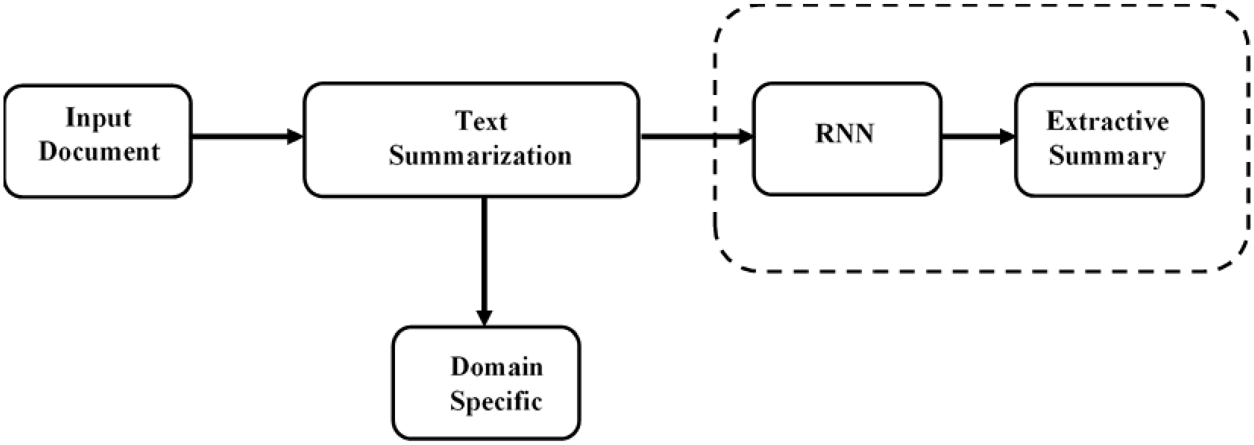

Fig 2: Diagram Showing The Conceptual Model

1.8 Scope of the Study

This research work focuses on Extractive Text Summarization Using Recurrent Neural Network. However, the scope of this research work is Extractive Text Summarization, The data used in this research paper will be provided by www.kaggle.com

1.9 Mathematical Model

Term Frequency

Term frequency (TF) is how often a word appears in a document, divided by how many words there are

TF(t) = (Number of times term t appears in a document)/ (Total number of terms in the document)

Inverse Document Frequency

Term frequency is how common a word is inverse document frequency (IDF) is how unique or rare a word is.

IDF(t) = log_e(Total number of documents/ Number of documents with term t in it)

1.10 Significance of the Study/Expected Contribution

This study will be of immense benefit in that;

1. It Create an intermediate representation of the original text in a large document.
2. It helps in shortening the documents by preserving the important contents of the text.
3. It also helps to preserve and show the main purpose of textual information.
4. It would be of great benefit for newscasters and other organizations to easily extract contents in large documents.

1.11 Definition of Terms

- **NLTK (Natural Language Toolkit):** is a powerful Python package that provides a set of diverse natural languages algorithms. It is free, opensource, easy to use, large community, and well documented.
- **Recurrent Neural Network:** also known as RNNs, are a class of neural networks that allow previous outputs to be used as inputs while having hidden states.
- **Text Summarization:** is a process of reducing the size of original document and producing a summary by retaining important information of original document.
- **Natural Language Processing:** is a branch of artificial intelligence that helps computers to understand, interpret and manipulate human language.
- **Machine Learning:** is the study of computer programs that leverage algorithms and statistical models to learn through inference and patterns without being explicitly programmed.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

2.0 Literature Review

2.1 Related Works

Julien R. et al (2017), text summarization in NLP has largely been limited to extractive methods that crop out and stick together portions of the original text in order to capture the bulk of the meaning concisely. What has been less attempted and is poorly understood are abstractive methods that might produce words not in the original text, giving more natural summaries and dynamic paraphrases. We employ a fully data-driven approach using a recurrent neural network.

Manish Shrivastava et al (2018), proposed a paper BoWLer: A neural approach to extractive text summarization, in this work, we present a simple, yet effective approach for extractive summarization of news articles. In line with many recent works in this area we propose an encoder-decoder architecture with a simple bag of word encoder for sentences followed by an attention based decoder for relevant sentence selection. Our model is trained end-to-end and its performance is comparable to the state-of-the-art models while being simpler both in terms of the number of parameters (significantly lesser) as well as the representational complexity.

Jonathan Rojas Sim´on et al (2018). In this paper, we present a new method based on a Genetic Algorithm to determine the best sentence combination of DUC01 and DUC02 datasets to rank the newest methods of AETS. Using three heuristics presented in the state-of-the-art, we rank the most recent AETS methods, obtaining upper bounds and recovering lower bounds of the state-of-the-art.

Siya Sadashiv Naik et al (2017). This paper highlights an extractive approach. Main aim is to select the best sentences by weighting them. We carried out our experiment on 15 documents from the DUC 2002 dataset. Each test document was first pre-processed. Then, all the sentences were represented as attribute vectors of features by calculating their scores. Rule-based method was proposed to select the best sentences. Results were compared with GSM summarizer and a conclusion was drawn that the best average recall, precision and f-measure values was obtained for Rule-Based Summarizer.

Tian Shi et.al (2019). Neural abstractive text summarization (NATS) has received a lot of attention in the past few years from both industry and academia. In this paper, we introduce an open-source toolkit, namely LeafNATS, for training and evaluation of different sequence-to-sequence based models for the NATS task, and for deploying the pre-trained models to real-world applications. The toolkit is modularized and extensible in addition to maintaining competitive performance in the NATS task. A live news blogging system has also been implemented to demonstrate how these models can aid blog/news editors by providing them suggestions of headlines and summaries of their articles.

Table 1: The table below summarizes the related works surveyed:

S/N	Author	Article	Aims	Approach	Achievement	Attention
1	Julien R. et al (2017).	Abstractive text summarization with Neural network	The aim was to explore text summarization	The research work was carried out using Recurrent Neural Network Approach	The paper was able to implement a network to perform character level summarization.	The work can be continue to be improve on "Extractive Summarization"
2	Manish Shrivastava et al (2018)	BoWLer: A neural approach to extractive text summarization	The aims was to summarize a single text document and present the user with the most important aspects of the same.	This study makes used of Bag Of Word Embeddings LearnER (BOWLER), a simple neural network based approach	The paper was able to implement a simple and effective neural sequence labeling model for extractive summarization for news domain and show that its performance is close to current state-of-the-art extractive summarization system.	Further work could be done to explore this aspect as well as study how different types of word embeddings affect the system performance.

3.	Jonathan Rojas Sim´on et al (2018).	Calculating the significance of automatic extractive text summarization using a genetic algorithm	the methods based on evolutionary approaches show a good tendency to obtain the best levels of performance compared to the machine learning methods.	The Research Work was carried out using Genetic Algorithm	the methods based on evolutionary approaches show a good tendency to obtain the best levels of performance compared to the machine learning methods.	Genetic Algorithms have not been used to obtain extractive summaries from Topline heuristic to reweigh the performance of the AETS methods.
4.	Siya Sadashiv Naik et al (2017)	Extractive Text Summarization By Feature-Based Sentence Extraction Using Rule-Based Concept	Main aim is to select the best sentences by weighting them.	The Research Work was implemented using Rule-Based Concept	Result produced by this summarizer was compared with existing GSM summarizer. It was seen that proposed summarizer gives better average recall, precision and f-measures vales than existing summarizer.	future work can be extended to summarize multiple documents. The proposed method could also be combined with existing learning methods for large dataset.
5.	Tian Shi et.al (2017).	LeafNATS: An Open-Source Toolkit and Live Demo System for Neural Abstractive Text Summarization	The aim of the study is to provide them suggestions of headlines and summaries of their articles.	recurrent neural network (RNN)-based sequence to-sequence (Seq2Seq) models	An extensive set of experiments on different benchmark datasets has demonstrated the effectiveness of our implementations.	Future research work and be develop on Extractive Summarization

CHAPTER THREE

METHODOLOGY

3.1 Methodology

Irny et al. (2005), defined methodology as the systemic, theoretical analysis of the methods applied to a field of study. It comprises the theoretical analysis of the body of methods and principles associated with a branch of knowledge. Typically, it encompasses concepts such as paradigm, theoretical model, phases and quantitative or qualitative techniques. It comprises the theoretical analysis of the body of methods and principles associated with a branch of knowledge, outlines the way in which research is to be undertaken and identifies the methods to be used in it. In this study, extractive text summarization methodology is embraced.

3.2 SYSTEM ARCHITECTURE

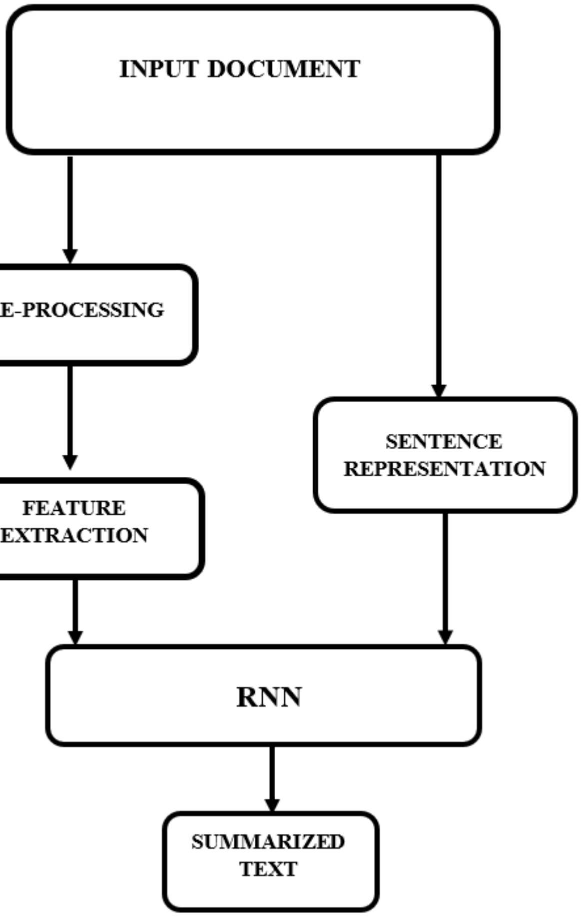

Figure 3: System Architecture Of The Model

REFERENCE

Ahmed Elrefaiy, Ahmed Rafat Abas, & Ibrahim Elhenawy. Review Of Recent Techniques For Extractive Text Summarization. Journal of Theoretical and Applied Information Technology,15th December 2018. Vol.96. No 23.

C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, volume 8. Barcelona, Spain, 2004.

Chandan K. Reddy, Ping Wang and Tian Shi: LeafNATS: An Open-Source Toolkit and Live Demo System for Neural Abstractive Text Summarization. arXiv:1906.01512v1 [cs.CL] 28 May 2019.

Irny, S.I. and Rose, A.A. (2005) "Designing a Strategic Information Systems Planning Methodology for Malaysian Institutes of Higher Learning (isp- ipta), Issues in Information System, Volume VI, No. 1, 2005

Jonathan Rojas Sim´on*, Yulia Ledeneva* and Ren´e Arnulfo Garc´ıa-Hern´andez* (2018). Universidad Aut´onoma del Estado de M´exico, Unidad Acad´emica Profesional Tianguistenco, Instituto Literario, CP, Toluca, Edo. Mex, M´exico. Journal of Intelligent & Fuzzy Systems xx (20xx) x–xx DOI:10.3233/JIFS-169588 IOS Press

Julien Romero et al (2017). Abstractive Text Summarisation with Neural Networks. (2017). Master Thesis in the Data Analytics Lab. Eidgenossische Technishche Hochschule Zurich, Swiss Federal Institute of Technology Zurich.

Lee, J.-H., Park, S., Ahn, C.-M., Kim, D. (2009). Automatic Generic Document Summarization Based on Non-negative Matrix Factorization. Information Processing and Management 45, 20–34.

Lloret, Elena, & Palomar, Manuel (2012). Text summarisation in progress: A literature review. Artificial Intelligence Review, 37(1), 1–41.

Nenkova, Ani, & McKeown, Kathleen (2012). A survey of text summarization techniques. In Mining text data (pp. 43–76). Springer.

Pranav Dhakras, Manish Shrivastava (2018). Language Technologies Research Center, Kohli Center On Intelligent Systems, International Institute of Information Technology, Hyderabad, India pranav.dhakras@research.iiit.ac.in, m.shrivastava@iiit.ac.in

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summa runner: A recurrent neural network based sequence model for extractive summarization of documents. In AAAI, pages 3075–3081.

R. Verma and D. Lee, Extractive Summarization: Lim838 its, Compression, Generalized Model and Heuristics, 2017, p. 19.

Siya Sadashiv Naik, Manisha Naik Gaonkar (2017). Extractive Text Summarization By Feature-Based Sentence Extraction Using Rule-Based Concept. 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India