AIN SHAMS UNIVERSITY

FACULTY OF ENGINEERING

Computer and Systems Engineering

Automatic Text Summarization using Natural Language Processing and Artificial Intelligence Techniques

A Thesis submitted in partial fulfilment of the requirements of the degree of

Doctor of Philosophy in Electrical Engineering

(Computer and Systems Engineering)

by

Wafaa Samy Abdul-Hamed El-Kassas

Master of Science in Electrical Engineering

(Computer and Systems Engineering)

Faculty of Engineering, Ain Shams University, 2015

Supervised By

Prof. Dr. Hoda Korashy Mohamed

Prof. Dr. Ahmed Abdelwahed Rafea

Dr. Cherif Ramzi Salama

Cairo - (2020)

AIN SHAMS UNIVERSITY

FACULTY OF ENGINEERING

Computer and Systems Engineering

Automatic Text Summarization using Natural Language Processing and Artificial Intelligence Techniques

By

Researcher Name: Wafaa Samy Abdul-Hamed El-Kassas

Degree: Doctor of Philosophy in Electrical Engineering (Computer and Systems Engineering)

Examiners' Committee

Name and Affiliation	**Signature**
Prof. Dr. Mohsen Abdel Razek Rashwan Electronics and Communication Department, Faculty of Engineering, Cairo University	.....
Prof. Dr. Mahmoud Ibrahim Khalil Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University	.....
Prof. Dr. Hoda Korashy Mohamed Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University	.....
Prof. Dr. Ahmed Abdelwahed Rafea Computer Science and Engineering Department, American University in Cairo	.....

Date: 23 September 2020

Statement

This thesis is submitted as a partial fulfilment of Doctor of Philosophy in Electrical Engineering, Faculty of Engineering, Ain Shams University.

The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Wafaa Samy Abdul-Hamed El-Kassas

Signature

……...…….

Date: 23 September 2020

Researcher Data

Name	: Wafaa Samy Abdul-Hamed El-Kassas
Last academic degree	: Master of Science
Field of specialization	: Computer and Systems Engineering
University issued the degree	: Ain Shams University
Date of issued degree	: 2015

Abstract

The Internet has an exponentially increasing amount of textual data. Searching for a certain topic can become a daunting task because users cannot read and comprehend all potentially long documents in the search results. As a result, it becomes urgent to help users by summarizing textual content. Manual text summarization consumes a lot of time, effort, cost, and even becomes impractical with the gigantic amount of textual content. Therefore, Automatic Text Summarization (ATS) in this case is clearly beneficial. Researchers have been trying to improve ATS techniques since the 1950s. ATS approaches are either extractive, abstractive, or hybrid. The extractive approach selects the most important sentences in the input document(s) then concatenates them to form the summary. The abstractive approach represents the input document(s) in an intermediate representation then generates the summary with sentences that are different than the original sentences. The hybrid approach merges between both the extractive and abstractive approaches. This thesis provides a comprehensive survey for the researchers by presenting the different aspects of ATS: approaches, building blocks, techniques, evaluation methods, and future research directions. Despite all the proposed methods in the literature, the generated summaries are still far away from the human-generated summaries. To enhance ATS for single documents, this thesis also proposes a novel extractive graph-based framework "EdgeSumm" that relies on four proposed algorithms. The first algorithm constructs a new text graph representation model from the input document. The second and third algorithms search the constructed text graph for sentences to be included in the candidate summary. When the number of words of the resulting candidate summary still exceeds a user-required length

limit, the fourth algorithm is used to select the most important sentences then add them to the final summary. EdgeSumm combines a set of extractive ATS methods (namely graph-based, statistical-based, semantic-based, and centrality-based methods) to benefit from their advantages and overcome their individual drawbacks. EdgeSumm is general for any document genre (not limited to a specific domain) and unsupervised so it does not require any training data. The standard datasets DUC2001 and DUC2002 are used to evaluate EdgeSumm using the widely used automatic evaluation tool: Recall-Oriented Understudy for Gisting Evaluation (ROUGE). EdgeSumm gets the highest ROUGE scores on DUC2001. For DUC2002, the evaluation results show that the proposed framework outperforms the state-of-the-art ATS systems by achieving improvements of 1.2% and 4.7% over the highest scores in the literature for the metrics of ROUGE-1 and ROUGE-L respectively. In addition, EdgeSumm achieves very competitive results for the metrics of ROUGE-2 and ROUGE-SU4.

Keywords: Automatic Text Summarization; Text Summarization Approaches; Text Summarization Techniques; Text Summarization Evaluation; Extractive Text Summarization; Single-Document Summarization; Text Graph Representation Model; EdgeSumm

Summary

“Automatic Text Summarization using Natural Language Processing and Artificial Intelligence Techniques”

Researcher Name: Wafaa Samy Abdul-Hamed El-Kassas

This thesis provides a comprehensive survey for the researchers by presenting the different aspects of Automatic Text Summarization (ATS): approaches, building blocks, techniques, evaluation methods, and future research directions. To enhance ATS for single documents, this thesis also proposes a novel extractive graph-based framework “EdgeSumm” that relies on four proposed algorithms. EdgeSumm combines a set of extractive ATS methods (namely graph-based, statistical-based, semantic-based, and centrality-based methods) to benefit from their advantages and overcome their individual drawbacks. EdgeSumm is general for any document genre (not limited to a specific domain) and unsupervised so it does not require any training data. The standard datasets DUC2001 and DUC2002 are used to evaluate EdgeSumm using the widely used automatic evaluation tool: Recall-Oriented Understudy for Gisting Evaluation (ROUGE). EdgeSumm gets the highest ROUGE scores on DUC2001. For DUC2002, the evaluation results show that the proposed framework outperforms the state-of-the-art ATS systems by achieving improvements of 1.2% and 4.7% over the highest scores in the literature for the metrics of ROUGE-1 and ROUGE-L respectively. In addition, EdgeSumm achieves very competitive results for the metrics of ROUGE-2 and ROUGE-SU4.

Chapter 1 introduces the thesis research including the problem definition, the research methodology, the research contributions, and the organization of the thesis.

Chapter 2 presents an introduction and background information about the different classifications, applications, standard datasets, and evaluation methods of the ATS systems.

Chapter 3 introduces the ATS approaches in more details and explores the different methods applied for each approach in the literature.

Chapter 4 highlights the techniques and building blocks that are used to implement the ATS systems. This chapter includes the text summarization operations, the statistical and linguistic features, the text representation models, the linguistic analysis and processing techniques, and the soft computing techniques.

Chapter 5 illustrates the research methodology of the proposed graph-based framework "EdgeSumm". This chapter starts with an introduction about the related work of extractive ATS methods and systems in order to introduce the motivation for proposing EdgeSumm. Then, the research methodology of EdgeSumm is illustrated along with an introduction about the EdgeSumm proposed architecture.

Chapter 6 presents a detailed explanation of the EdgeSumm architecture phases including: pre-processing, processing, and post-processing. The proposed algorithms are explained in details in this chapter.

Chapter 7 presents the evaluation results and discussion. This chapter includes the evaluation baseline systems, the used datasets, the evaluation metrics, and the evaluation results along with the experiments details.

Chapter 8 gives the work conclusions, thesis contributions, and future work. In addition, this chapter ends with a detailed explanation of the future directions for ATS research in general.

References exist at the end of this thesis.

Acknowledgment

First praise is to Almighty Allah who gives me the power to complete this work. I would like to thank my parents. I pray to God for them who taught me invaluable lessons in life. Their care, patience, and love are what guided me through my whole life.

I am grateful to my father and I revere his patronage and support. He was encouraging me continuously and dreaming with the day when I will get my PhD and he was not expecting less than excellence from me. I was always remembering his patience, support, and encouragement to me for overcoming numerous obstacles I have been facing through my research and throughout writing this thesis.

I would like to thank my supervisors Prof. Dr. Ahmed Rafea, Prof. Dr. Hoda Korashy, and Dr. Cherif Salama for their efforts, help, guidance and patience. I learned so many valuable things from them, but above all, they taught me how to be devoted to research.

I would like to thank my family, friends, and colleagues who always encourage me to complete my research.

Wafaa Samy Abdul-Hamed El-Kassas
Computer and Systems Engineering
Faculty of Engineering
Ain Shams University
Cairo, Egypt

October 2020

Table of Contents