ENHANCED SENTENCE EXTRACTION THROUGH NEURO-FUZZY TECHNIQUE FOR TEXT DOCUMENT SUMMARIZATION

MUHAMMAD AZHARI BIN AHMAD KAMIL

MASTER OF SCIENCE
IN INFORMATION AND COMMUNICATION TECHNOLOGY

2021

Faculty of Information and Communication Technology

ENHANCED SENTENCE EXTRACTION THROUGH NEURO-FUZZY TECHNIQUE FOR TEXT DOCUMENT SUMMARIZATION

Muhammad Azhari bin Ahmad Kamil

Master of Science in Information and Communication Technology

2021

ENHANCED SENTENCE EXTRACTION THROUGH NEURO-FUZZY TECHNIQUE FOR TEXT DOCUMENT SUMMARIZATION

MUHAMMAD AZHARI BIN AHMAD KAMIL

A thesis submitted
in fulfilment of the requirements for the degree of Master of Science
in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2021

DECLARATION

I declare that this thesis entitled “Enhanced Sentence Extraction Through Neuro-Fuzzy Technique for Text Document Summarization” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in the candidature of any other degree.

Signature : ……

Name : Muhammad Azhari Bin Ahmad Kamil

Date : ……1/11/2020……

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Master of Science in Information and Communication Technology.

Signature : ……...

Supervisor Name : Dr. Yogan A/L Jaya Kumar

Date : …….17/08/2021……

DEDICATION

To my beloved parents, Ahmad Kamil Bin Tajuddin and Fuziah Binti Hamid, your love and support are my greatest inspiration upon accomplish this study.

To my dearest supervisors, Dr. Yogan Jaya Kumar and Dr. Norhayati for being responsible, receptive and always by my side to encourage and motivate me.

To my dear friend, especially Amirul Ramzani Radzid, Nur Shida Ahmad Sharawadi and Nur Atikah Arbain for your support and motivation throughout this study.

ABSTRACT

A summary system comprises a subtraction of text documents to generate a new form that delivers the essentials contents of the documents. Due to the hassle of documents overload, getting the right information and effectively-developed summaries are essential in retrieving information. Reduction of information allows users to find the information needed quickly without the need to read the full document collection, in particular, multi documents. In the recent past, soft computing-based approaches have gained popularity in its ability to determine important information across documents. A number of studies have modelled summarization systems based on fuzzy logic reasoning in order to select important sentences to be included in the summary. Although past studies support the benefits of employing fuzzy based reasoning for extracting important sentences from the document, there is a limitation concerning this method. Human or linguistic experts are required to determine the rules for the fuzzy system. Furthermore, the membership functions need to be manually tuned. These can be a very tedious and time-consuming process. Moreover, the performance of the fuzzy system can be affected by the choice of rules and parameters of membership function. Therefore, this study proposes a text summarization model based on classification using neuro-fuzzy approach. A classifier is first trained to identify summary sentences. Then, we use the proposed model to score and filter high-quality summary sentences. We compare the performance of our proposed model with the existing approaches, which are based on fuzzy logic and neural network techniques. In this study, we also evaluate the performance of sentence scoring and clustering in the process of generating text summaries. The proposed neuro-fuzzy model was used to score the sentences and clustering were performed using K-Means and Hierarchical Clustering (HC) approaches. The proposed approach showed improved results compared to the previous techniques in terms of precision, recall and F-measure on the Document Understanding Conference (DUC) data corpus. However, it was found that no improvements in the quality of the generated summaries obtained by simply performing clustering.

PENGEKSTRAKAN AYAT YANG DITINGKATKAN MELALUI TEKNIK NEURO-KABUR UNTUK RINGKASAN DOKUMEN TEKS

ABSTRAK

Sistem ringkasan terdiri daripada pengekstrakan ayat-ayat di dalam dokumen untuk menghasilkan rumusan yang menyampaikan kandungan penting dokumen. Disebabkan masalah dokumen yang berlebihan, adalah penting untuk mendapatkan maklumat yang betul dan rumusan dibuat dengan berkesan. Pengurangan maklumat membolehkan pengguna mencari maklumat yang diperlukan dengan cepat tanpa perlu membaca penuh kesemua dokumen, khususnya, pelbagai jenis dokumen. Pada masa lalu, pendekatan berasaskan pengkomputeran lembut telah mendapat populariti dalam keupayaannya untuk menentukan maklumat penting di dalam dokumen yang banyak. Beberapa kajian telah memodelkan sistem ringkasan berdasarkan penalaran logik kabur untuk memilih ayat-ayat penting untuk disertakan dalam ringkasan. Walaupun kajian lepas menyokong manfaat menggunakan kaedah logik kabur untuk mengekstrak ayat penting daripada dokumen, terdapat batasan mengenai kaedah ini. Manusia ataupun pakar linguistik diperlukan untuk menentukan peraturan untuk sistem kabur. Tambahan pula, fungsi keahlian perlu disetkan secara manual. Ini boleh menjadikan proses yang memenatkan dan memakan masa. Selain itu, prestasi sistem kabur boleh dipengaruhi oleh pilihan peraturan dan parameter fungsi keahlian. Oleh itu, kajian ini mencadangkan model ringkasan teks berdasarkan klasifikasi menggunakan pendekatan pengendali neuro-kabur. Pengkelasan dilatih terlebih dahulu untuk mengenal pasti ringkasan ayat. Kemudian, kami menggunakan model yang dicadangkan untuk memberi skor dan menapis ayat ringkasan berkualiti tinggi. Kami membandingkan prestasi model yang dicadangkan dengan pendekatan yang sedia ada, yang berdasarkan teknik logik kabur dan teknik rangkaian neural. Dalam kajian ini, kami juga menilai skor ayat dan prestasi pengklusteran dalam proses menghasilkan ringkasan teks. Model pengendali neuro-kabur yang dicadangkan digunakan untuk mengira skor ayat dan proses klusteran dilakukan menggunakan pendekatan K-Means dan Hierarchical Clustering (HC). Pendekatan yang dicadangkan menunjukkan hasil yang lebih baik berbanding teknik terdahulu dari segi ketepatan, recall dan F-measure pada korpus data Document Understanding Conference (DUC). Walau bagaimanapun, didapati bahawa tiada penambahbaikan dalam kualiti ringkasan yang dijana yang diperoleh dengan sekadar melakukan pengklusteran.

ACKNOWLEDGEMENTS

First and foremost, I would like to take this opportunity to express my sincere acknowledgement to my supervisor Dr. Yogan Jaya Kumar from the Faculty of Technology and Communication, Universiti Teknikal Malaysia Melaka (UTeM) for his essential supervision, support and encouragement towards the completion of this thesis. I would also like to express my greatest gratitude to Ts. Dr. Norhayati Harum from the Faculty of Technology and Communication, co-supervisor of this project for her advice and suggestions. Special thanks to UTeM short term grant funding for the financial support throughout this project. Particularly, I would also like to express my deepest gratitude to Mr. Badrul, the technicians for CIT Labs, Faculty of Technology and Communication, Ts. Dr. Halizah Basiron, Prof. Dr. Abd Samad Hassan Basari lecturer and Allahyarham Prof. Dr. Burairah Hussin for their assistance and support through my year of research. Special thanks to all my peers, my CIT Labs mate, my beloved mother, my beloved father and siblings for their moral support in completing this master. Lastly, thank you to everyone who had been to the crucial parts of realization of this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
ANFIS	-	Adaptive Neuro Fuzzy Inference System
ANN	-	Artificial Neural Network
BDD	-	Boolean Decision Diagram
BNN	-	Boolean Neural Net
CDISI	-	Crack Detection and Impact Source Identification
DUC	-	Document Understanding Conference
FL	-	Fuzzy Logic
GA	-	Genetic Algorithm
HC	-	Hieratical Clustering
LSTM	-	Long Short-Term Memory

MF	-	Mean Frequency
NMT	-	Neural Machine Translation
NN	-	Neural Network
OCR	-	Optical Character Recognition
OFL	-	Optimize Fuzzy Logic
PSO	-	Particle swarm optimization
ROUGE	-	Recall-Oriented Understudy for Gisting Evaluation
TF-IDF	-	Term Frequency–Inverse Document Frequency

LIST OF PUBLICATIONS

Azhari, M., Kumar, Y.J., Goh, O.S., Choon, N.H., and Pradana, A., 2018. Multi document summarization using neuro-fuzzy system. In: *Advances in Intelligent Systems and Computing*.

Azhari, M., Kumar, Y.J., Goh, O.S., and Ngo, H.C., 2018. Automatic Text Summarization: Soft Computing Based Approaches. *Advanced Science Letters*.

Azhari, M., Kumar, Y.J., Goh, O.S., and Raza, B., 2018. Text summarization evaluation based on sentence scoring and clustering. *Journal of Engineering Science and Technology*.

Azhari, M. and Kumar, Y.J., 2017. Improving text summarization using neuro-fuzzy approach. *Journal of Information and Telecommunication*, 0 (0), pp.1–13.

CHAPTER 1

INTRODUCTION

1.1 Overview

The world we live today witnesses a fast-moving information age. The sole contributor to this exponential growth is the World Wide Web (WWW). People are being exposed to loads of information; let it be in the form of web pages or text documents which are retrieved through the web daily. Even with simple access to online data had caused big effect on the individuals. At the same time, due to the fact of information overload, it has also caused them problems. To deal with the issue, it is essential to provide alternatives to digest different sources of data. One of the studies currently under review is the field of automatic text summarization, particularly in the case of online text resources. A summary is a concise representation of a document's content that contains a significant portion of the information in the original text(s) (Saleem et al., 2015). Text summarization is a reductive transformation of source text to summary text through content reduction by selection or generalization on what is important in the source (Jones, 2007). Automatic text summarization is therefore the process of automatically producing such summaries, i.e., through a computer. Automatic text summarization studies often discover distinct methods to generate summaries to meet the needs of various activities and users.

The purpose and type of summary explains the kind of summary produced. For example, “approach towards text summarization can be either extractive or abstractive” (Radev et al., 2000). In summarization extraction types, important verses are identified and extracted directly from the original document, i.e., the final summary contains the original

sentence. On the contrary, in the abstractive summarization (Salim, 2015), “selected verses from the original document are further processed to rearrange them before incorporating them into the final summary”. This process usually involves a deep understanding of natural language and compression sentence. Summaries could also be labelled as generic or domain specific (Svore et al., 2006). The generic summarization is to summarize all the texts regardless of their topic or domain; a generic summary does not make assumptions about its source information domain and view all documents as homogeneous texts. There is also a summary system development that is based on various domain interests. For example, to summarize biomedical documents, financial articles, , terrorist events, weather news and more (Radev and McKeown, 1998; Verma et al., 2007; Hennig et al., 2008); Apart from generic and domain specific summarization, another type of approach is called query-based summarization (Tang et al., 2013) .In query-based approach, summaries are produced based on a user query on a particular subject.

By understanding the types of summary i.e., extractive, abstractive, generic, domain specific and query based, either single document or multi document texts we can then apply them to produce summaries. Single document summarization systems produce summary of one input document. The distinct characteristic that makes multi document summarization rather different from single document summarization is that the former involves multiple sources of information as inputs. These multiple inputs are often a collection of documents on the same topic or the same event. Early work in online systems has shown multi-document summarization been applied to clusters of news articles (Radev and McKeown, 1998). Besides understanding the types of summarization, the core of an automatic summarization system is the method used to generate its summary. To date, various methods have been proposed and used to perform text summarization (Poibeau et al., 2013). Feature-based method is one of the most commonly used method in this area. In the process of identifying

important sentence, the characteristics that influence the relevance of the sentence are determined. Some of the features that are often considered for sentence selection are frequency, title words, cue words, sentence location and sentence length (Gupta et al., 2010). These features often increase the nomination of a sentence to be summarized. Feature-based methods are usually used to summarize a document. In the case of diverse document summaries, the two most commonly used mainstream methods are cluster-based methods and graph-based methods (Aliguliyev, 2010b; Galluccio et al., 2012). Further details and other methods will be discussed later in Chapter 2.

Ultimately, to test the effectiveness of any proposed summarization method, evaluation on the automatic text summarization systems need to be carried out to determine the accuracy of the summarized output. Usually, a gold-standard is used; against which the results of the systems can be compared. The gold-standard is a set of human-made summaries written by experts. Researchers use the human summaries as benchmark to show how much the performance of their proposed summarization model is acceptable compared with that performance of human (H2-H1) (Kumar et al., 2017). A widely used evaluation tool for text summarization is ROUGE (Recall Oriented Understudy for Gisting Evaluation) (Lin, 2004; Liu, 2011). Basically, ROUGE would compare two generated summaries (i.e., between the human generated summaries and the automatic generated summaries) in order to determine the precision, recall and the F-measure of the summarized text. In this study, we propose to enhance sentence extraction through neuro-fuzzy technique for multi document summarization, where the domain covers news stories. The related source documents used for the purpose of this study are obtained from the DUC (Document Understanding Conference) 2002 data set; a standard corpus used for text summarization studies (Kumar et al., 2017). This chapter proceeds as follows: Section 1.2 reviews the problem background; Section 1.3 presents the problem statements; Section 1.4 describes the

objectives of the study; Section 1.5 gives the scope of the study; Section 1.6 mentions the significance of the study; Section 1.7 lists the expected contribution and Section 1.8 describes the organization of the thesis.

1.2 Project background

As stated earlier, owing to the proliferation of data on the Internet, the need for automatic text summarization has lately risen. Finding data from online documents has been decreased to the fingertips of the user with the availability and speed of Internet. However, manually summarizing large internet information is not simple for users. For example, when a user searches for information regarding the recent tsunami in Sumatra, Indonesia (in December 2018), users may receive a large article related to the event. The user would choose for the summary of those articles without the need to go through each and every article. The goal of automatic text summarization (in this case multi-document summarization) is condensing the source texts into a shorter version whilst preserving its information content (Haque et al., 2013).

In the recent past, soft computing-based approaches have gained popularity in its ability to determine important information across documents (Aik, 2008; Aliguliyev, 2009; Berker and Güngör, 2012; Balcan and Gupta, 2014). For instance, a number of studies have modelled summarization systems based on fuzzy logic reasoning in order to select important sentences to be included in the summary (Barzilay and Lapata, 2008; Babar and Patil, 2015). First, the features influencing the importance of a sentence are determined, such as, title word, sentence position, thematic word, etc. Then selected sentence features are used as the input to the fuzzy system. The scores for each sentence are then derived using fuzzy rules scoring. The sentences with high fuzzy score will be selected to be included in the summary until the desired summary length is obtained.

Apart from sentence scoring, fuzzy logic has also been used for semantic analysis to produce text summary. For example, Kumar et. al (2017) investigated the cross-document relations that exist between sentences and used fuzzy logic to rank sentences based on the type of cross document relations. The authors in (Republic, 2009) extracted the semantic relations between concepts using fuzzy reasoning to select summary sentences. This method which is based on latent semantic analysis improves the quality of summary. Although all the above works support the benefits of employing fuzzy based reasoning for extracting important sentences from the document, there is a limitation concerning this method. Human or linguistic experts are required to determine the rules for the fuzzy system. Furthermore, the membership functions need to be manually tuned. These can be a very tedious and time-consuming process. Moreover, the performance of the fuzzy system can be affected by the choice of rules and parameters of membership function (Jang, 1991).

Besides fuzzy logic, neural network models have also been employed in text summarization studies whereby its learning capabilities are used to identify summary sentences from the input text document (Aik, 2008). Megala et. al (2014) used a three-layered feedforward network model to learn the patterns in summary sentences. The resulting trained network is then applied to new input documents to determine if a sentence should be included in the summary. In another related work, Sarda and Kulkarni (2015) used a similar neural network model with the combination of Rhetorical Structure Theory (RST). The RST relations that exist in the sentences which are picked out by their neural network model are used to form high quality summaries. Fattah and Ren (2008) in their study proposed an improved content selection approach using probabilistic neural network. They used probability function to better estimate the weights of their neural network model. Although neural network model has been useful in term of its learning capabilities, the model provides little information about the relationship between the input and output.

1.3 Problem statement

By understanding the problem background which has been discussed in the previous section, it can be concluded that issues in the field of text summarization studies still need more investigation. It can be observed that among the two soft computing techniques that have been associated with text summarization, fuzzy logic implementation is based on knowledge–driven reasoning whereas neural network is based on data–driven approximation. Taking these observations into consideration, a better summarization system can be modelled by considering the advantages of both approaches and avoiding their drawbacks. The limitations and gaps left by past research work add to some important questions to be answered in this thesis.

The primary research question for this study is "How to overcome the limitations and gaps left by past research work in soft computing approach based?" Derivative questions are as follows:

1. Will the integration of neural network over fuzzy reasoning method produce better classification of summary sentences?
2. Can the proposed neuro-fuzzy classifier improve the overall quality of the generated summary?
3. Does clustering provide significant impact on the process selecting summary sentences?