SUDAN UNIVERSITY OF SCIENCE AND TECHNOLOGY
COLLEGE OF GRADUATE STUDIES

COLLEGE OF COMPUTER SCIENSE AND INFORMATION TECHNOLOGY

TEXT SUMMARIZATION OF HOLY QURAN INTERPRETATION IN ENGLISH USING DEEP LEARNING ALGORITHM

(CASE STUDY: SURAT ALFATIHA)

تلخيص تفسيرالقرآن الكريم باللغة الإنجليزية بإستخدام خوارزميات التعلم العميق
دراسة حالة : (سورة الفاتحة)

THEESIS SUBMITTED AS A PARTIAL REQUIREMENTS OF MASTER DEGREE IN INFORMATION TECHNOLOGY

PREPARED BY

Nisreen Salih Hummeida

SUPERVISED BY

Dr.Hwaida Ali Abdalgadir

DEC 2018

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor **Dr. Hwaida Ali Abdalgader** for the continue support, for her patience, motivation, and immense knowledge. Her guidance helped me in all time of research writing of thesis. I could not have imagined a better advisor and mentor for my MSC research.

Also I would like to thanks my colleague Omer Abdel Majed for his support and guide during my research.

Finally, I must express my very profound gratitude to my parents and to my partner for providing me without unfailing support and continuous encouragement throughout the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you

Nisreen Salih

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

For long time summarization is done by human, but sometimes take long time to be done. Nowadays many researchers are going for text summarization automatically, which can be done by using some techniques. Deep learning algorithm is one of the most techniques used in text summarization. The difficulties to understand the indented meaning of the interpretation of Quran for Muslim and new comers to Islam which give us motivation to build an automatic extraction text summarization using Restricted Boltzmann Machine (RBM) to produce proper summary by applying different preprocessing techniques.

This approach consists of three phases, which are, feature extraction, feature enhancement, and summary generation, which work together to generate understandable summary. Once the features are enhanced using RBM summary of each interpretation (single document summarization) is generated by scoring the sentences based on those enhanced features and an extractive summary is constructed. The Precision and Recall used for measuring the performance of the proposed approach.

The summary that generated by RBM algorithm was compared with other existing method using same algorithm. Experimental results showed that the summary produced by proposed approach responds better than existing method.

Keywords:

Extractive Summarization, Deep learning, Single Document, RBM

المستخلص

منذ وقت طويل يتم التلخيص بواسطه الإنسان ولكن فى بعض الأحيان قد يستغرق ذلك زمناَ طويلاَ. فى الآونه الأخيره ذهب العديد من الباحثين فى مجال التلخيص لإيجاد طريقه تلقائيه لتلخيص النص بحيث يتم ذلك بإستخدام تقنيات حديثه. معماريه التعلم العميق تعتبر واحده من اكثر التقنيات المستخدمه فى هذا المجال.

على صعيد آخر يعتبر فهم تفسير آيات القرآن الكريم للمسلمين ولمعتنقى الأسلام الجدد من المشاكل التى تواجهه العديد منهم؛ وذلك لطول نصها وتعقيد كلماتها؛ وذلك اعطى الباحثين دافعاً قوياً لإيجاد آليه لتلخيص تفسير القرآن بإستخدام معماريه بولتزمان المقيدة (RBM) لإنتاج ملخص مناسب من خلال تطبيق تقنيات معالجه مختلفه للنص.

يتكون المنهج المقترج من ثلاث مراحل: استخراج الخصائص، تحسين الخصائص، توليد مخلص منها. هذه المراحل تعمل بشكل متزامن لتوليد ملخص مفهوم. حيث أن إنشاء تلخيص للمستند يتم بنائه من خلال إدخال الجمل بناءاً على تلك الخصائص المحسنه ومن ثم يتم إنشاء تلخيص بتلك الخصائص المحسنه. تم إستخدام أداة الدقه والإستدعاء لقياس أداء المنهج المقترح.

قُرن الملخص الذى تم إنشائه بواسطة معماريه بولتزمان المقيدة (RBM) مع ملخص آخر إستخدم نفس المعماريه، أظهرت النتائج التجريبيه أن الملخص الناتج عن المنهج المقترح يعطى نتيجه أفضل من الطريقه الحاليه.

CHAPTER 1

INTRODUCTION

1. 1 Introduction

For long time summarization is done by human, but sometimes take long time to be done. Nowadays many researchers are going for text summarization automatically, which can be done by using some techniques.

Text summarization relates to the process of obtaining a textual document, obtaining content from it and providing the necessary content to the user in a shortened form and in a receptive way to the requirement of user or application. [1]. Text summarization can be done on a single document which generates a shorter version of it or multiple document summarizations, which generate summary from multiple related documents.

There are different interpretations of Holy Quran, and this interpretation is difficult to be understood for most of Muslims and has complicated meaning.

Quran is the word of Allah, therefore, in order to understand it in the true sense, a person has to sharpen his or her intellectual ability as well as increase the knowledge. When one is at a particular intellectual level, only then can he or she start understanding the true message, which Allah Almighty conveyed through words of Quran.

1.1.1 The Importance of Interpretation of the Holy Quran:

There is a variety of reasons why interpretation is important, and people consult interpretation when it comes to understanding of Quran. The few important reasons that make interpretation important are as follows: [2]

1. Firstly, it tries to explain Quran so that the understanding of the reader about Quran and its message increases. Reading only the translation can give the literal meanings of the words of Quran, however, interpretation comes with context in which a particular Ayah was revealed, hence giving a context to each and every word, which consequently means a better understanding of Quran and itsverses.
2. Secondly, interpretation is also important when it comes to driving out the laws of Islam from Quran. Verses of Quran come with instructions, and then there are the hadiths of Prophet (PBUH) that also gives instructions. Thus, interpretation combines both and gives a complete and comprehendible set of instructions to thereader.
3. Thirdly, when there is no science involved in the interpretation of Quran, then ambiguity and contradiction is surely rise. Hence, instead of ambiguous interpretations there are clear explanations.

For all these reasons and other reasons which prove the necessity of interpretation of Holy Quran in our life, and our need to understand the interpretation in brief, the need for text summarization is rise.

1.1.2 Deep learning algorithm

Deep learning comprises of a set of algorithms that focuses on learning the abstract representations of the data at multiple levels through several non-linear transformations. It is one of the domains that has gained popularity in the recent times following neural networks. [3]

Deep learning algorithm is one of most techniques that used in text summarization. For summarizing the text there is a need of structuring the text into certain model which can be given to Restricted Boltzmann Machine (RBM). [3]

Restricted Boltzmann Machine (RBM) is stochastic neural network (that is a network of neurons where each neuron has some random behavior when activated). It consists of one visible unit (neuron) and one layer of hidden units. Units in each layer have no connections between them and are connected to all other units in other layer. [4]

1.2 Problem Statement

There are different interpretations of Holy Quran, and these lead to some difficulties, and they are as follow:

1. It is difficult for Muslims to read and understand all of them.
2. It's difficult for Non Muslims to understand the language of interpretation.
3. Difficulty that face people to understand the intended meaning of interpretation.

1.3 Significance of the Research

This research is significant for both learners and the interpreters, and it is going to shade a light on the text summarization that is used in the holly Quran's interpretation that as follow:

- The research helps in understanding the interpretation in a simple word.
- Easy for reading and use.
- It helps in using technology in Islamic Studies.

- It provides interpretations for non-specialized interpreters.
- Provides summary of interpretation for people who are willing to readit.

1.4 Objectives of the Research

The main objective of this research:

- ✓ To summarize Surat Alfatiha using Restricted Boltzmann Machine (RBM).
- ✓ To explore how applying preprocessing techniques (Part of speech tagging, stop word removing, stemming) can improve text summarization.
- ✓ To Enhancement the text using RBM.

1.5 Hypothesis

This research is going to test the following hypothesis:

- Text summarization give people the main idea of interpretation.
- Using RBM (Restricted Boltzmann Machine) can give proper summary.
- Applying preprocessing techniques like (Part of speech tagging, stop word removing, stemming) can improve text summarization.

1.6 Methodology

Automatic summarization is done by extractive method. Proposed approach can be classified into following phases:

1. Preprocessing phase.
2. Feature extraction phase.
3. Deep learning phase,and
4. Post processingphase

Flow Chart of proposed approach as follow:

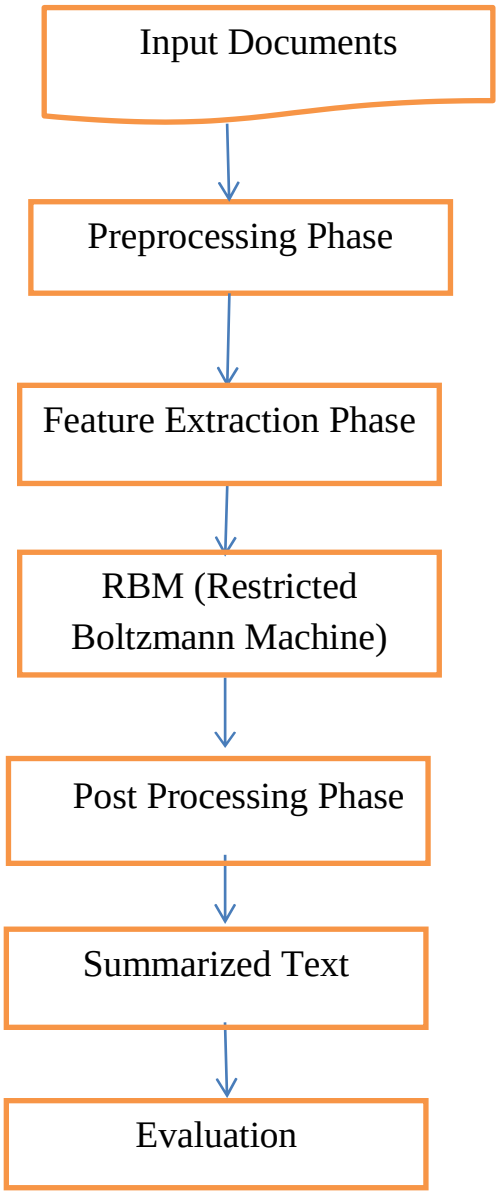

Fig 1.1 Flow Chart of Text Summarization

1.7 The Scope

The research applied on summarizing Surat Alfatiha as case study by using Deep Learning algorithm.

1.8 Layout of the Thesis:

The other part of the thesis is organized as follows:

Throw light on the domain text summarization, and deep learning. It presents the principal approaches used in the summarizing one or more documents, its implementation, performance and evaluation methods. In addition, presents overview on the background on deep learning, and related research work are reviewed in chapter two. Chapter three deals with the implementation of all modules of the text summarizer described in text summarizer in chapter 1, and the prerequisite for running the implementation of text summarizer will also describe in this chapter.

Chapter four present the results and performance evaluation of the system generated summary, and finally the conclusion and future work.

CHAPTER 2

LITERATURE REVIEW

This chapter will throw light on the domain text summarization, and deep learning. It presents the principal approaches used in the summarizing one or more documents, its implementation, performance and evaluation methods. In addition, presents overview on the background on deep learning.

2.1 Introduction

Text summarization now a day tends to be challenging problem, these refer to a large amount of information provided. The need to provide summaries has become wider spreading. It has become an important and timely tool for assisting and interpreting in growing of information. It difficult for human begins manually summarizes large documents of text. [5]

A summary can be used as a pointer some parts of original document or to cover all relevant information of the text. The important advantage of using a summary is its reduced reading time. A good summary system should reflect the desire topics of the document while keeping redundancy to a minimum. [5]

2.2 Definitions of Text Summarization

The literature provides various definitions of text summarization:

❖ Text Summarization

A text summarization is the process of distilling the most important information from the source (or sources) to produce an abridged version for a user (or users) and task (or tasks) [7].

❖ Text

'Text' can be anything ranging from on-line documents, multimedia documents, hypertexts, etc. As the amount of data available through various means has increased, so is the interest in automatic text summarization.

❖ Summary

A summary can be defined as a text that produced from or more texts that contain a significant portion of the information in the original texts, and that is no longer than half of the original texts. [6]

Summaries generated by programs may not be equal to the ones obtained by humans and this is the precise reason why automatic summarization is a thriving research area. Summary can be classified into two categories based on what is being focused on. They are as follows:

1. Generic Summary:

Summary is generated without any additional information; it is summarized by extracting sentences or by giving out an abstract without the context of query.

2. Query-oriented Summary:

With the development in the field of information retrieval query-oriented document summarization has start gaining popularity. A query-oriented summary presents the information that is most relevant to the given queries.

2.3 Summarization Techniques

There are two divergent approaches to text summarization [7] based on how the summary is generated:

2.3.1 Extractive Summarization:

Extractive summarization extracts relevant sentence from the source document by reducing redundancy of information in summary. This method is relatively easier in terms of time consumed and flexibility than the abstractive summarization.

The extraction text summarization is divided into two steps [7]:

1. Preprocessing step: is structured representation of the text. It includes: sentence boundary, Stop word elimination, tokenizing, stemming, POS tagging, etc.

2. Processing step: features affecting the importance of the sentence are decided and calculated and then weights are assigned to this feature.

2.3.2 Abstractive Summarization:

An abstractive summarization [8] tries to improve an understanding of the main notion in a document and then express those notions in a clear natural language. It uses linguistic method to examine and interpret the text and then to find the new notions and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. [6] This method aims at giving summaries that are like human summary and it is still an open research area.

2.4 Summarization methods:

Summarization method can also be classified into the following categories based on the set of documents taken into consideration for summarization:

2.4.1 Single-document Summarization:

The work on summarization began as early as fifties when the first summarization research was presented. It is considered the cornerstone for all works which followed it. Single-document summarization involves generating summary for every single source document. All the documents will belong to a single domain.

2.4.2 Multi-document summarization:

Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. Resulting summary allows users to quickly familiarize themselves with information contained in a large cluster of documents. Every topic is described from multiple perspectives within a single document.

Summarizing a single document is not an easy task but summarizing a multi document poses additional challenges. To avoid the repetitions, one must identify and locate thematic overlaps, and to deal with potential inconsistencies between documents to arrange event along a single timeline. For these reasons, multi-document summarization is much less developed than its single-document.

Various methods have been proposed to identify cross-document overlaps. SUMMONS a system that covers multi document summarization takes an information extraction approach. All input documents are parsed into templates; SUMMONS clusters the templates according to their contents, and then applies rules to extract items of major input [7]. In contrast, parse each sentence into a syntactic dependency structure using a robust parser and then match trees across documents, using paraphrase rules that alter the trees as needed.

An ideal multi-document summarization system not only shortens the source texts, but also presents information organized around the key aspects to represent diverse views. Success produces an overview of a given topic. Such text compilations should also basic requirements for an overview text compiled by a human. The multi-document summary quality criteria are as follows [9]:

- clear structure, including an outline of the main content, from which it is easy to navigate to the full text sections
- text within sections is divided into meaningful paragraphs
- gradual transition from more general to more specific thematic aspects
- good readability.

2. 5 Kinds of summarization systems

To create a summary, you must evaluate each portion (paragraph, sentence, word) of the text and decide whether to keep it or not. You may then reformulate what you have selected in coherent form and must then out it. [9]

Researches in an automated text summarization have identified three distinct stages: [9]

2.5.1 Topics Identification:

Topic identification produces the simplest type of summary. Whatever criterion of importance is used, once the system has identified the important sentence, paragraphs, etc can either simply list them or display them diagrammatically.

To perform this stage, systems employ several independent modules. Each module assigns a score to each unit of input (word, sentence or longer), then a combination module combines the score of each unit to assign a single integrated score to it. Finally, the system returns the n highest-scoring units, according to summary length requested by user.

The system generates the performance of the topic identification by using recall and precision score. [9]

2.5.2 Interpretation or topic fusion:

Interpretation is distinguishing extract type from abstract type systems. When the process starts the topics identifies as substantial are combined, represented in new phrase, and expressed in a new formulation, using word not in the original text.

The system cannot perform interpretation without previous knowledge about the domain; it must interpret the input of something strange to the text. [8]

The template representation used in information extraction to represent stories for summarization. Interpretation remains blocks because of domain knowledge acquisition problem. The system will have to solve this problem before the abstract summarization start.

2.5.2 Summary generation:

Is the major stage of summarization, when summary content has been created through abstracting and/or information extraction, it exists within the computer in internal notation, and thus requires the techniques of natural language generation, namely text planning, sentence (micro-) planning, and sentence realization. [8] Extract summaries require no generation, they are simply extracted and printed whether they are printed in order of importance score or in text order.

In the context of summarization, describe a summary revision program that takes as input simple extracts and produces shorter and more readable summaries.

2.6 Evaluation of Summary

Evaluation of summary is a very important task in the field of automatic summarization of text. Evaluating the summary besides enhancing development of reusable resources and infrastructure helps in comparing and replicating results and thus, adds competitions to improve the results. Evaluation of summary is a

challenging work too as it is not easy for humans to know what kind of information should be present in the summary. [12]

How can you evaluate the quality of the summary? The literature on this question suggests are so task and so user oriented that no single measured cover all cases. There are two types of systems [13] to evaluate summaries. They are:

2.6.1 Intrinsic

Intrinsic systems evaluate only the quality of the summary generated. Most evaluations of summarizations systems are intrinsic. The evaluators create a set of ideal summaries, and then compare the summarizer's output to it, measuring content overlap. [13]

Apart from comparing the summarizer's output to the ideals, some evaluators rate system's summaries according to some scale, for example, coverage; fluency; readability and in formativeness. [13]

2.6.2 Extrinsic:

Extrinsic system considers the user assistance in the task performance while evaluating summaries. Extrinsic evaluation is easy to motivate but the problem is to ensure that the metric applied correlates well with task performance efficiency. The largest extrinsic evaluation is TIPSTER-SUMMAC study [13]. In Classification, task tester classified a set of TREC texts and their summaries created by various systems. After the classification, the agreement between the classification of the text and their corresponding summaries is measured; the greater the agreement, the better the summary capture the full text to be classified as it is. In the Ad hoc task, tester classified query-based summaries are Relevant and Not Relevant to the query. All extraction systems performed equally well for generic summarization and that IR methods produced the best query-based extracts. [13]

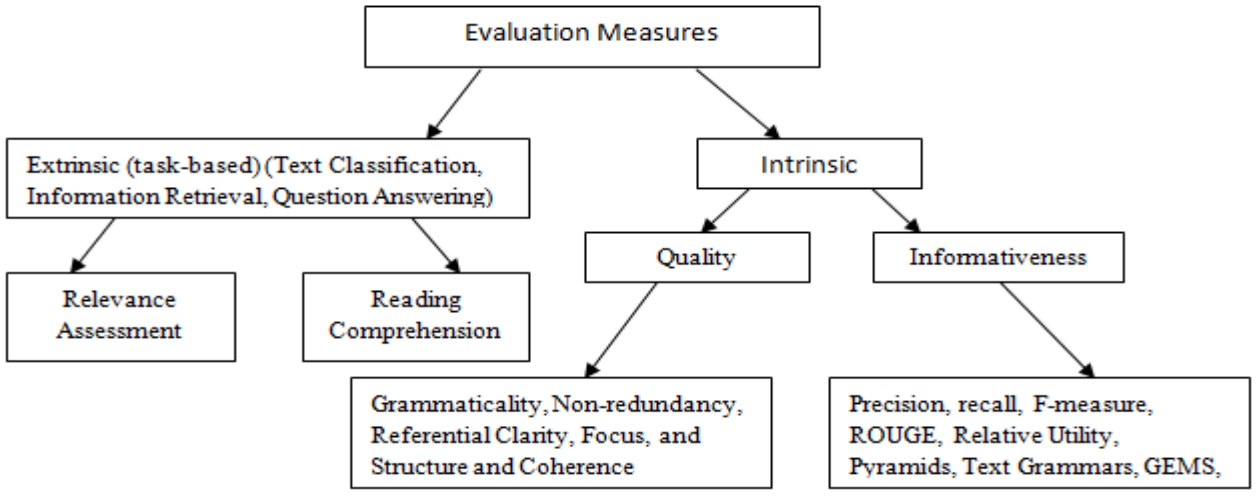

Figure 2.1: Taxonomy of summary evaluation measure [14]

2.6.3 Manual Evaluation:

Manual evaluation requires a group of human judies to read a whole original text, then read the whole summary and more or less subjectively evaluate the quality of the summary. Some methodologies to help with the process exist; with define several categories for the judies to evaluate the summary. While a working approach, if a very time consuming and requires averaging of scores from several people as there exists no ideal summary and the evaluation it always at least bit subjective. [34]

2.6.4 Precision

It determines what fraction of the sentences chosen by the humans and selected by the system are correct precision is the number of sentences found in both system and ideal summaries divided by the system summary. [33]

$$Precision = \frac{|\text{system-human choice overlap}|}{|\text{sentences chosen by system}|}$$

2.6.5 Recall

It determines what proportion of the sentences chosen by the human is even recognized by the machine. Recall is the number of sentences found in both system and ideal summaries divided by the number of sentences in the ideal summary. [33]

$$Recall = \frac{|\text{system-human choice overlap}|}{|\text{sentences chosen by human}|}$$

2.7 Deep learning

Deep learning is a set of algorithms in machine learning that attempt to model high-level abstractions in data by using architectures composed of multiple nonlinear transformations. It aims at learning distributed representations of data effectively through several levels with the highest level representing the abstract form of data. Research in this area attempts to define what makes better representations and how to create models to learn these representations. The field of deep learning has attracted a lot of attention in the recent past and the techniques have been applied in many fields including natural language processing, machine learning, information retrieval, computer vision and artificial intelligence, etc. And it is still gaining popularity because of the interesting challenges it poses. [18]

2.7.1 Background on Deep learning

Since 2006, deep learning has appeared as a new area in machine learning. Deep learning is an intersection among the researcher area of neural networks, artificial intelligence, graphical modeling, optimization, pattern recognition and signal processing. [18]

Definition: A sub field within machine learning that is based on algorithms for learning multiple levels of representation in order to model complex relationships among data. Higher level features and concepts are thus defined in terms of lower level ones, and such a hierarchy of features is called a deep architecture. Most of the models are based on unsupervised learning of representation. [18]

Deep learning is the deep architecture and good learning algorithm, which perform intellectual learning. It's composition of many layers of adaptive non-linear components

2.7.2 Different deep architecture

There are several deep architectures, but CNNs and DBNs are the two milestones in field of deep learning. Researchers tried different way to train the multi-layer neural network, but more than one of the successful even in term of efficiency and accuracy except CNNs. DBNs is a hybrid model consisting of an undirected graph model and a directed graph model. [19]

2.7.2.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a feed-forward network, but it combines three architecture ideas to ensure some degree of shift and distortion invariance. [19]

Convolutional neural networks aim to use spatial information between the pixels of an image. Therefore, they are based on discrete convolution. The basic components of convolutional neural networks are: [19]

A. Convolution Layer

In each convolutional layer, the convolutional filter exploits the local correlation by enforcing a local connectivity pattern among adjacent layers. The upper layer m is already getting from a subset of units from lower layer m-1. Comparing with MLP, another advantage of convolutional layer is the numbers of parameters are significantly reduced due to the sharing parameters. The connectivity is illustrated in Figure 2.2. [20]

Layer m-1 is feature input from retina. In Figure 2.2, we take the example of receptive fields with the width of 3 and thus 3 adjacent neurons are connected with the same parameter in the retina layer (the button layer). For the above layer, the convolutional operation follows the same process as previous layers. There is one thing we need to note here is, in layer m-1; only three adjacent neurons are connected. But for upper layer, their receptive field with respect to the input is larger (5 instead of 3). During the training processing, the filters are learnt to

produce strongest response to a spatially local input pattern. The example in figure 2.2 only represents three layers of convolution. If we stack more and more layers, the filter becomes increasingly global. [21]

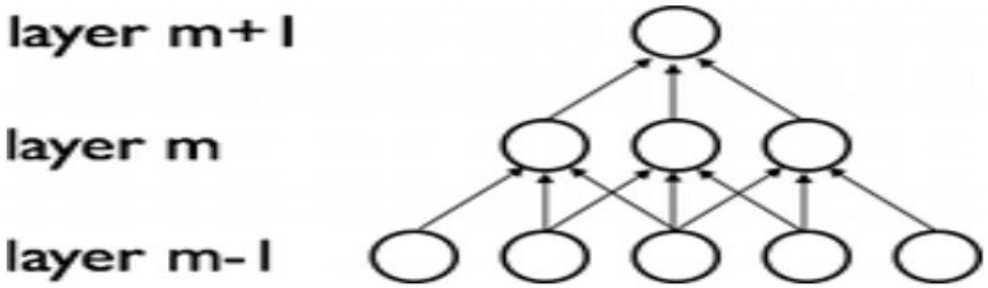

Figure 2.2: Network with three-convolution layer [22]

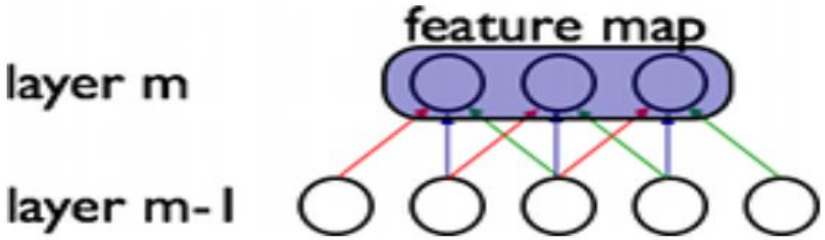

Figure 2.3: Weights are shared in convolution layer [22]

In Figure 2.3, they show 3 hidden units belonging to the same feature map. The lines with the same color represent the same weight are shared among different input part of signal. These shared weights can still be learnt from gradient descent algorithm. Applying convolutional filter repeatedly across the image allow the model to detect the interested feature regardless of the position in visual field. Furthermore, the shared weights could significantly reduce the number of parameters need to be learnt in the framework. The advantage will let CNN apply to larger training date with fewer computation times, which could be better for generalizing the problem. [21]

B. Pooling layer

The second important component of CNN is pooling. The pooling layer functions as non-linear down-sampling. The pooling algorithm only focuses on particular sub-region of feature map. The goal for pooling is trying to the most meaningful information from a particular region.

Pooling algorithm is important for three reasons:

• By eliminating non-maximal values, it reduces computation for upper layers.

• It provides a form of translation invariance. Imagine cascading a max-pooling layer with a convolutional layer.

• In some case, pooling is helpful for summarizing different feature length into same dimension.

2.7.2.2 Deep Belief Networks (DBNs)

DBNs are hybrid model, which has two parts, undirected graph model, and graph model, which is stochastically Restricted Boltzmann Machine (RBM). DBNs are a stochastically learning architecture whose object function depends on learning purpose. [18]

The greatest advantage of DBNs are its capability of learning feature, which is achieved by “layer by layer” learning strategies where the higher-level features are learned from the previous layers and the higher-level features are believed to be more complicated and better reflects the information contained in the input data's structures. [19]

DBNs extract a deep hierarchical representation of the training data. They model the joint distribution between observed vector $x$and the ℓhidden layers h^kas follows:

$$P(x, h^1, \ldots, h^\ell) = \left(\prod_{k=0}^{\ell-2} P(h^k | h^{k+1}) \right) P(h^{\ell-1}, h^\ell) \quad (1)$$

Where $x = h^0$, $P(h^{k-1}|h^k)$is a conditional distribution for the visible units conditioned on the hidden units of the RBM at level k, and $P(h^{\ell-1}, h^\ell)$is the visible-hidden joint distribution in the top-level RBM. This is illustrated in the figure below. [22]

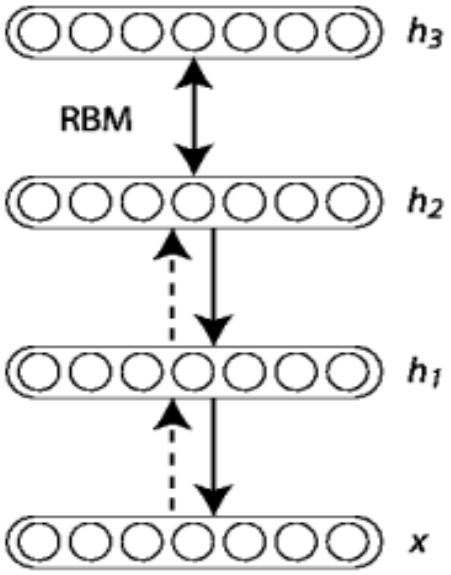

Figure 2.4: Deep Belief Networks (DBNs) [18]

The principle of greedy layer-wise unsupervised training can be applied to DBNs with RBMs as the building blocks for each layer [20]. The process is as follows:

1. Train the first layer as an RBM that models the raw input $x = h^{(0)}$as its visible layer.
2. Use that first layer to obtain a representation of the input that will be used as data for the second layer. Two common solutions exist. This representation can be chosen as being the mean activations $p(h^{(1)} = 1|h^{(0)})$or samples of$p(h^{(1)}|h^{(0)})$.
3. Train the second layer as an RBM, taking the transformed data (samples or mean activations) as training examples (for the visible layer of that RBM).
4. Iterate (2 and 3) for the desired number of layers, each time propagating upward either samples or mean values.
5. Fine-tune all the parameters of this deep architecture with respect to a proxy for the DBN log- likelihood, or with respect to a supervised training criterion (after adding extra learning machinery to convert the learned representation into supervised predictions, e.g. a linear classifier).

2.7.2.3 Restricted Boltzmann Machine

Boltzmann Machines (BMs) are a form of log-linear Markov Random Field (MRF), i.e., for which the energy function is linear in its free parameters. To make them powerful enough to represent complicated distributions (i.e., go from the limited parametric setting to a non-parametric one), it is considered that some of the variables are never observed. By having more hidden variables (also called hidden units), we can increase the modeling capacity of the BM. Restricted. [18]

Boltzmann Machines further restrict BMs to those without visible-visible and hidden-hidden connections. A graphical depiction of an RBM is shown in Figure 2.5.

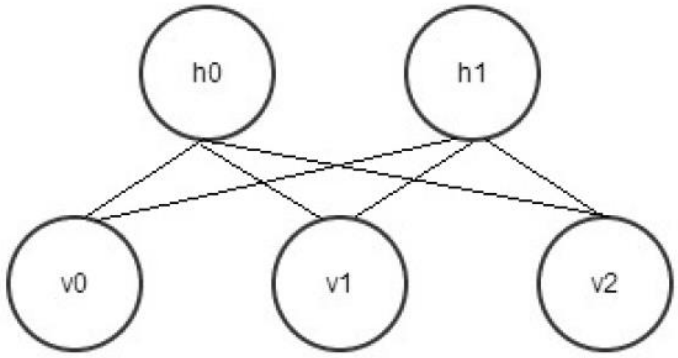

Figure 2.5: Restricted Boltzmann Machine (RBM) [18]

The energy function E (v; h) of an RBM is defined as:

$$E(v,h) = -b'v - c'h - h'Wv \tag{2}$$

where W represents the weights connecting hidden and visible units and b', c' are the offsets of the visible and hidden layers respectively. v stands for the visible units and h stands for the hidden units.

This translates directly to the following free energy formula:

$$\mathcal{F}(v) = -b'v - \sum_i \log \sum_{h_i} e^{h_i(c_i + W_i v)} \tag{3}$$

Because of the specific structure of RBMs, visible and hidden units are conditionally independent given one-another. Using this property, the following equations can be written:

$$p(h|v) = \prod_i p(h_i|v) \tag{4}$$

$$p(v|h) = \prod_j p(v_j|h) \tag{5}$$

2.8 Related Work

In the previous research, different techniques were presented for producing summary of any text.

For extractive summarization, Zhong et al. (2015) [23] used a deep architecture that is like an AE. They used the learned representations for filtering out unimportant words of a document in the early layer and discovering key words in later layer. Their training method requires sample queries during the first stage. The concept space is to extract the candidate sentences for the summary. However, in our model both the term and concept space are developed based on a sentence of a document. They use the learned representations directly to represent the semantic similar of sentences and in the ranking function.

In supervised models, Denil et al. (2014) [24] proposed a model based on a CNN to extract candidate sentences to be included into the summary. A key contribution of the paper is that CNN is trained to classify sentiment labels that are either positive or negative (i.e. a binary label) and sentence extraction can affect the results of predicted a sentiment.

Cheng and Lapata (2016) [25] treat single document summarization as a sequence labeling task and model it with recurrent neural networks. Their model is composed of a hierarchical document encoder and an

attention-based extractor; the encoder derives the meaning representation of a document based on its sentences and their constituent words while the extractor adopts a variant of neural attention to extract sentences or words

Cao et al. (2015) [26] used a Recursive Neural Network for text summarization using hand-crafted word features as inputs. With the capability of dealing with a variable length input in an RNN, the proposed system formulates the sentence ranking task in a hierarchical regression fashion. Not using any hand-crafted word representations and labeling data are significant in our proposed model.

Mahmoud El-Haj, Udo Kruschwitz, Chris Fox (2014) builds two Arabic summarization system (AQBTSS and ACBTSS) (Arabic Query Based Text Summarization and Arabic Concept Based Text Summarization System) and they found that Query-based summarizer performs much better than concept-based. This results and assumptions need more investigation and theoretical analysis because their system depends on query, if the number of queries and documents increase, the performance will be lower.

Sukriti Verma and Vagisha Nidhi (2017) proposed an extraction summarization using deep learning of factual reports. They use Restricted Boltzmann Machine (RBM) to enhance and abstract those features to improve resultant accuracy without losing any important information. The sentences are scored based on those enhanced features and an extractive summary is constructed. The proposed approach has an average precision value of 0.7 and average recall value of 0.63 which are both higher than those of the existing approach.

Shashi pal, ajaikumer, abhiasamangal, shikhaSinghal (2016) proposed Bilingul (Hindi and English) unsupervised automatic text summarization using deep learning to improve result accuracy and they used Restricted Boltzmann Machine to generate a shorter version of original document and exploring the feature to improve the relevance of sentences in the dataset. The output result of the proposed algorithms is almost 85% accurate and preserves the meaning of summarized document.

Geehansabah, sitiKhaotijah, Faris Mahdi (2015) generate an extractive text summarization from proper set of sentences from documents by deep learning. The procedure is manipulated by Restricted Botlzmann Machine (RBM) algorithms for better efficiency by removing redundant sentences. They used four different features for feature extraction phase, the feature score of the sentences are applied to the RBM. The evolution matrices considered in the proposed text summarization are recall, precision, and f-measure. They have three different documents sets; the response is 0.49, 0.46 and 052 respectively for three document sets.

Table 2.1: Summary of Related Work

Paper's Title	Date	Author	Techniques	Results	Open Issue
Bilingual Automatics Text Summarization Using Unsupervised Deep Learning	2016	ShashiPal,Ajai Kumar, AbdilashaMangal, ShikhaSinghal	RBM (Restricted Boltzmann Machine)	They compare human generated and the system generated summary. They are using ROUGE-1 as it has high recall significance test. According to F-measure their system is giving 85% accuracy.	More enhancement by adding more features to get more relevant sentences and meaningful summary
An Approach for Text Summarization using Deep Learning Algorithm	2014	PadmaPriya, G. and K.Duraiswamy	RBM (Restricted Boltzmann Machine)	The comparative analysis of the performance of the proposed approach and an existing method shows that the proposed approach responds better as compared to existing one.	More enhance in text summarization by considering more feature extraction and adding more hidden layers
Experiments in Automatic Text Summarization Using Deep Neural Networks	2011	Ben King, Rahul JHA, Tyler Johson, Vaishnavi	Backpropagation algorithm	Implementing the autoencoder for unsupervised learning led to no overall increase in Performance	It's better to embed sentence in Statistical feature to improve overall performance
Finding Summary of Text Using Neural Networks and RhetoricalStructure	2016	Sarda A. T., Kulkarni A. R.	Neural Network (Backpropagatin), Rhetorical Relations	They use human generated summary to compare with their summaries. This comparison by using ROUGE Package. They	Combination of NN and RST with adding more feature selection for better summary result
Extractive Summarization Using Deep Learning	2017	Sukriti Verma and Vagisha Nidhi	Restricted Boltzmann Machine (RBM)	The proposed approach has an average precision value of 0.7 and average recall value of 0.63	Each algorithm run separately for each document instead of learning from corpus

Experimenting with Automatic Text Summarizatio n for Arabic	2014	Mahmoud El-Haj, Udo Kruschwitz, Chris Fox	AQBTSS and ACBTSS	They found that Query-based summarizer performs much better than concept-based.	The performance will be lower if the query on the document increase.
An Extractive Text Summarization Using Deep Learning	2015	Geehansabah, sitiKhaotijah, Faris Mahdi	Restricted Boltzmann Machine	The evolution matrices considered in the proposed text summarization are recall, precision, and f-measure response is 0.49, 0.46 and 052 respectively for three document sets.	Number of feature extraction is 4 which is very little.

CHAPTER 3

IMPLEMENTATION

This chapter deals with the implementation of all modules of the text summarizer described in text summarizer in chapter 1. The prerequisite for running the implementation of text summarizer will also describe here.

3.1 Prerequisite installation

The software and packages that we need to install for running the implementation of text summarizer, which are:

3.1.1 Anaconda

Anaconda is a free an open source distribution of the Python and R programming languages for data science and machine learning applications (large-scale data processing, predictive analytics, scientific computing) that aims to simplify package management and deployment [28].

We use the Anaconda navigator, which is a desktop Graphical User Interface (GUI) included in anaconda distribution that allow us to launch applications and manage conda package, environment and channels without using command-line commands. Navigator can search for package on Anaconda Cloud or in a local Anaconda Repository, install the packages and run it. It is available for windows, macOS, and linux[29].

In addition, we need to install libraries for data preprocessing and deep learning algorithm to be running. We download these libraries through conda prompt command when we run it as administrator. Those libraries are:

3.1.2 Natural Language Tool Kit (NLTK) libraries:

The Natural Language ToolKit (NLTK) is a platform used for building python programs that work with the human language data for applying in statistical natural language processing (NLP) [30].

It contains text-processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. It also includes graphical demonstrations and sample dataset and a book, which explains the principle behind the underlying language processing tasks the NLP support [31]. NLP techniques is used for parsing, reduction of words and to generate text summary. NLTK comes with many corpora, toy grammars, trained models, etc.

3.1.3 NumPy:

NumPy is the fundamental package for scientific computing in python library that provide multidimensional array object, various derived object, an assortment for fast operations on arrays, and including mathematical, logical shape, and basic statistical operations and much more [32].

3.1.4 Theano:

Theano is python library that allows user to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays efficiently. It is an open source project primarily develops by a machine learning group at the University of Montreal. The implementation can run on both Central Processing Unit (CPU) and Graphics Processing Unit (GPU).

3.2 Design

The detailed description of the system in figure 3.1 as follow: The text summarizer has two parts based on the function to be performed. The first one performs data preprocessing, and vectors the data so, the feature can be enhanced by using Restricted Boltzmann Machine (RBM). Here we implement all modules to get the summary from given document.

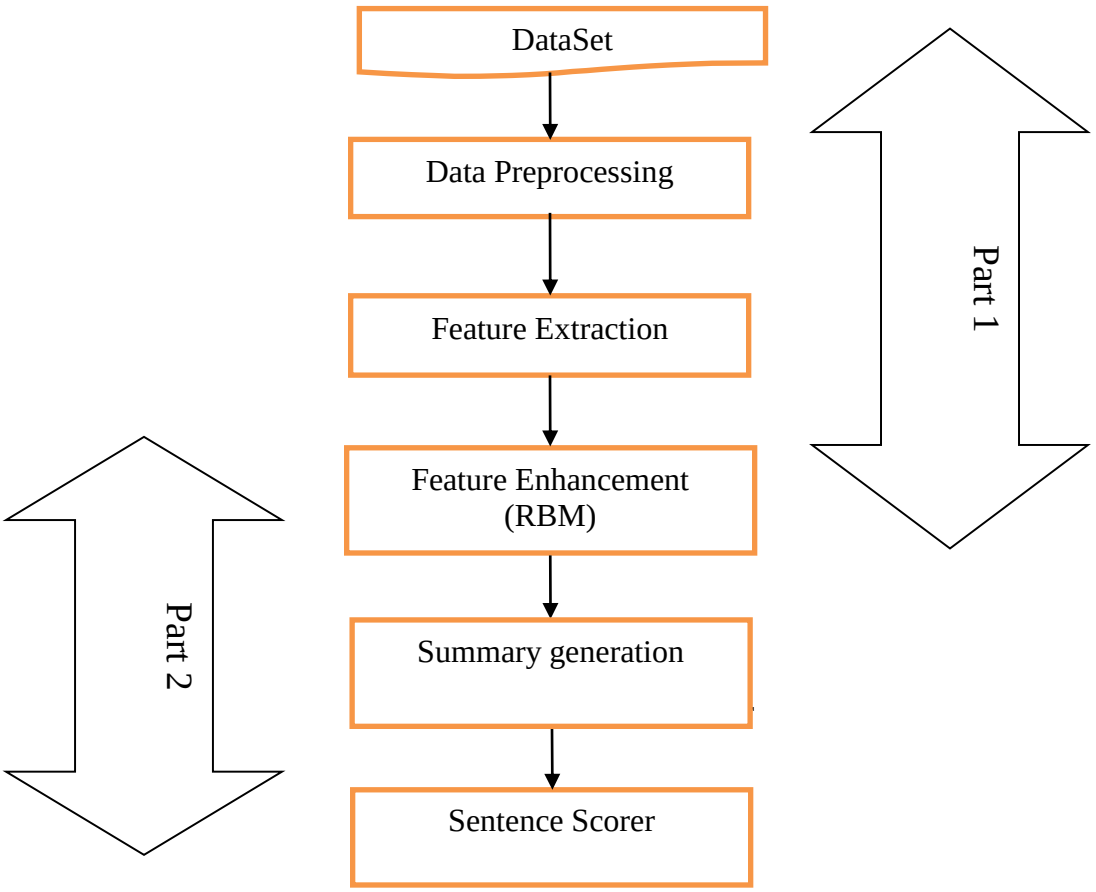

Figure 3.1: Text Summarizer

3.2.1 DataSet

The dataset used in this thesis are the interpretation of suratalfatih from different interpreter, we get the source of the data from the website www.tafseer.com in English language, and one reference [27]. The dataset in text format written in English, the text document contains punctuations, fulstop, comas.

We prepare the dataset in txt format to be ready for preprocessing phase.

3.3 Data Preprocessor

In the data preprocessor, the dataset was given i.e interpretation of suratAlfatiha in txt format, and there different preprocessing techniques are applied to the data as follows:

Firstly, we tokenize the text document into sentence by using the nltk.tokenizesent_tokenize. After that, we tokenize the sentences into words by using nltk.tokenizeword_tokenize. Once the sentences are broken into words, they are subjecting to further processing to obtain feature vector.

Sentences tokenize:

['In the Name of Allah, the All-Merciful, the Most Merciful, "In the Name of Allah," meaning: I begin with every Name belonging to Allah, Most High.', "This is because the word Name is singular and in the genitive form; therefore it subsumes all of Allah's Beautiful and Perfect Names.", '"Allah he is the God who is worshipped: the one deserving worship to the exclusion of everything else because of His qualities of divinity, attributes, all of which are perfect.',

Sentence to words:

['In', 'the', 'Name', 'of', 'Allah', 'the', 'All', 'Merciful', 'the', 'Most', 'Merciful', 'In', 'the', 'Name', 'of', 'Allah', 'meaning', 'I', 'begin', 'with', 'every', 'Name', 'belonging', 'to', 'Allah', 'Most', 'High'], ['This', 'is', 'because', 'the', 'word', 'Name', 'is', 'singular', 'and', 'in', 'the', 'genitive', 'form', 'therefore', 'it', 'subsumes', 'all', 'of', 'Allah', 's', 'Beautiful', 'and', 'Perfect', 'Names'], ['Allah', 'He', 'is', 'the', 'God', 'who', 'is', 'worshipped', 'the', 'one', 'deserving', 'worship', 'to', 'the', 'exclusion', 'of', 'everything', 'else', 'because', 'of', 'His', 'qualities', 'of', 'divinity', 'attributes', 'all', 'of', 'which', 'are', 'perfect'],

Secondly, we used the nltk.stopword to filter the stopword, and then stemming the words by using the nltk.PorterStemmer ().stem(word).

Stop words filtered:

Hereafter, yourself, former, here, thereafter, nine, bill, fifteen, most, now, otherwise, con, ever, whereby, four, something, your, between, bottom, then, nevertheless, cry, cannot

Sentences after stemming:

['In name allah merci compassion one manipractic taught islam follow begin activ name god', 'thiprincipl conscious earnestli follow necessarili yield three benefici result', 'first one abl restrain oneself manimisdesinc habit pronounc name god bound make one wonder commit offenc act reconcil say god holi name', 'second man pronounc name god start good legitim task act ensu start point mental orient sound', 'third import benefit man begin somethpronounc god name enjoy god support succour god bless effort protect machintemptatsatan', 'for whenev man turn god god turn well']

3.3.1 POS Tagger:

Before tagging the words in the documents with its corresponding POS, it must be parsed. The sentences in each file are read separately and broken down into tokens i.e words. Once the parsing is done, the POS tagger module from NLTK package is used to perform tagging. Example:

['God', 'Herein', 'O', 'Allah', 'Compassionate', 'Prophet', 'Essence', 'Adam', 'Gabriel', 'Unity', 'Commander', 'Believers', 'Greatest', 'Merciful', 'Jesus', 'My', 'First']

3.3.2 Feature Extractor:

The data preprocessor module reprocesses each document and a list containing the sentences in each document is obtained as a result. After that the document is structured into a sentence feature-matrix. A feature vector is extracted for each sentence. The combination of 9 sentences feature is most suitable to summarize the interpretation. These computations are done on its text document after reprocessing phase:

1. **Number of thematic words**: The thematic words are the most frequently occurring words of the text. For each sentence, the ratio of the number of thematic words to total words are computed as follow:

$$Sentence\_Thematic = \frac{No.\ of\ thematic\ words}{Total\ words} \quad (1)$$

2. **Sentence Position:** This feature is calculated as follows.

$$Sentence\_Position = \begin{cases} 1, \text{if its the first or last sentence of the text} \\ cos((SenPos - min)((1/max) - min)), \text{otherwise} \end{cases} \quad (2)$$

Where, SenPos = position of the sentence in the text

min = th * N

max = th * 2 * N

N is the total number of sentences in document.

th is threshold calculated as 0.2 * N.

3. **Sentence Length:** This feature is used to exclude sentences that are too short as those sentences will not be able to convey much information.

$$Sentence\_Length = \begin{cases} 0, \text{if number of words is less than 3} \\ No.\ of\ words\ in\ the\ sentence, \text{otherwise} \end{cases} \quad (3)$$

4. **Sentence position relative to paragraph:** This comes directly from the observation that at the start of each paragraph, a new discussion is begun and at the end of each paragraph, we have a conclusive closing.

$$Position\_In\_Para = \begin{cases} 1, \text{if it is the first or last sentence of a paragraph} \\ 0, \text{otherwise} \end{cases} \quad (4)$$

5. **Number of proper nouns:** This feature is used to give importance to sentences having a substantial number of proper nouns. Here, we count the total number of words that have been PoS tagged as proper nouns for each sentence.
6. **Number of numerals:** Since figures are always crucial to presenting facts, this feature gives importance to sentences having certain figures. For each sentence we calculate the ratio of numerals to total number of words in the sentence.

$$Sentence\_Numerals = \frac{No.\ of\ numerals}{Total\ words} \quad (5)$$

7. **Number of named entities:** Here, we count the total number of named entities in each sentence. Sentences having references to named entities are often quite important.
8. **Term Frequency-Inverse Sentence Frequency (TF ISF):** Frequency of each word in a sentence is multiplied by the total number of occurrences of that word in all the other sentences. We calculate this product and add it over all words.

$$TF - ISF = \frac{log(\sum_{all\ words} TF * ISF)}{Total\ words} \quad (6)$$

The importance of words increases proportionally due to the number of times a word appears in the document. Some of feature extracted: tf-isf

[0.6306972211576749, 0.2201902220252896, 0.10824670750564486, 0.14945302560159432, 0.8502972196724897, 0.24423728025045627, 0.30684469062441294, 0.23678525428721217, 0.3009246150909227, 0.28492722837799933, 0.27376221250208227, 0.17231760519269182, 0.2706960202390811, 0.5158446791710618, 0.27150781857860806, 0.08666007377053049, 0.26725247704383726, 0.5037650695664951, 0.4204688414702551, 0.13355717051742425, 0.7324057510929028, 0.39816832853031386, 0.34913760422086526, 0.41743830955125416, 0.43686902933711474, 0.23456727682781875, 0.31991383724779227, 0.16250652327119727]

9. **Sentence to Centroid similarity:** Sentence having the highest TF-ISF score is considered as the centroid sentence. Then, we calculate cosine similarity of each sentence with that centroid sentence.

$$Sentence\_Similarity = cosine\_sim(sentence, centroid) \qquad (7)$$

At the end of this phase, we have a sentence-feature matrix.

3.3.3 Feature Enhancement:

The sentence-feature matrix has been generated with each sentence having 9 feature vector values.

To enhance and abstract, the sentence-feature matrix is given as input to a Restricted Boltzmann Machine (RBM) which has one hidden layer and one visible layer. Single hidden layers will suffice for the learning process based on the size of our training data. The RBM that we are using has 9 perceptrons in each layer with a learning rate of 0.1. We have trained the RBM for 5 epochs with a batch size of 4 and 4 parallel Gibbs Chains.

Each sentence feature vector is passed through the hidden layer in which feature vector values for each sentence are multiplied by learned weights and a bias value is added to all the feature vector values which is also learned by the RBM. At the end, we have a refined and enhanced matrix. Note that the RBM will have to be trained for each new document that has to be summarized.

The structure and the working process of RBM to perform feature selection as follows:

i. First, the network is trained by using some dataset and setting the neurons on the visible layer to match data points in this data set.

ii. After the network is trained we can use it on new unknown data to make classification of data (unsupervised learning).

The given tf-idf vector divide into two sets for training, and testing, purpose of RBM respectively. In the training phase the RBM works in two step. The input of the first step, is feature vector (sentence matrix) S= [s_1, s_2,s_3, ….,s_n]. During the first cycle of RBM a new refined of sentence matrix set S'= [s'_1, s'_2,s'_3, ….,s'_n]. In step two, the same procedure will be applied to refined sentence matrix to get the more refined sentence matrix.

After obtaining the refined sentence matrix from the RBM, these refined sets are tested on a particular randomly generated threshold for each feature is calculated.

In this part, we have obtained the good set feature vector by RBM. Next, we will fine-tune the obtained feature vector set by adjusting the weight of the units of RBM. To fine-tune the feature vector set optimally we use back propagation algorithm. Back propagation algorithm is well known method to adjust the deep architecture to find good optimum feature vector set for the precise contextual summary of text.

3.3.4 Sentence Scorer:

The sentence scorer plays an important role by scoring each sentence in the document based on the feature vector. The enhanced feature vector values are summed to generate a score against each sentence. The sentences are then sorted according to decreasing score value. The most relevant sentence is the first sentence in this sorted list and is chosen as part of the subset of sentences which will form the summary

3.3.5 Summary Generator:

The summary generator generates the summary for each document by gather up the top ten high scoring sentences. The sentence score list is given by the sentence scorer is taken as an input and the sentences corresponding to the index of the top ten scores are extracted from each document (extractive summarization). This is the final stage of text summarization system.

CHAPTER 4

RESULTS AND PERFORMANCE EVALUATION

4.1 Results

The result obtained from the text summarizer is presented in this chapter. The end of text summarizer the summarize interpretation of surat alfatiha using RBM. The following is sample summary of random file chosen from the entire dataset.

Sample of the Summary

In the Name of Allah, the All-Merciful, the Most Merciful.

"Allah," He is the God who is worshipped: the one deserving worship to the exclusion of everything else because of His qualities of divinity, attributes, all of which are perfect. He has decreed it for those who fear Allah, those who follow His Prophets and Messengers: these have unrestricted mercy; anyone else only has a portion of this mercy.Therefore, all blessings are one of the resultant effects of this mercy.

This principle holds true for all of His Names. It is said concerning the Name, All-Knowing: He is All-Knowing, possessing the quality of knowledge by which He knows everything. He is All-Powerful, possessing the quality of power which makes Him omnipotent etc.

4.2 Performance Evaluation

Several interpretations from different interpreter, with varying number of sentences were used for experimentation and evaluation. The proposed algorithm was run on each of those and system-generated summary. We measure the performance by using manual evaluation, precision, and recall measures.

1. Manual Evaluation:

We ask Dr.Huda, Dr.Ahmed, and Dr. Awad (Quran AlKraim University) to give their evaluation about the summary that we get it from the system-by using the proposed approach, they said that the summary of system is give an understandable meaning to human and its approximately 70% gives the meaning of interpretation.

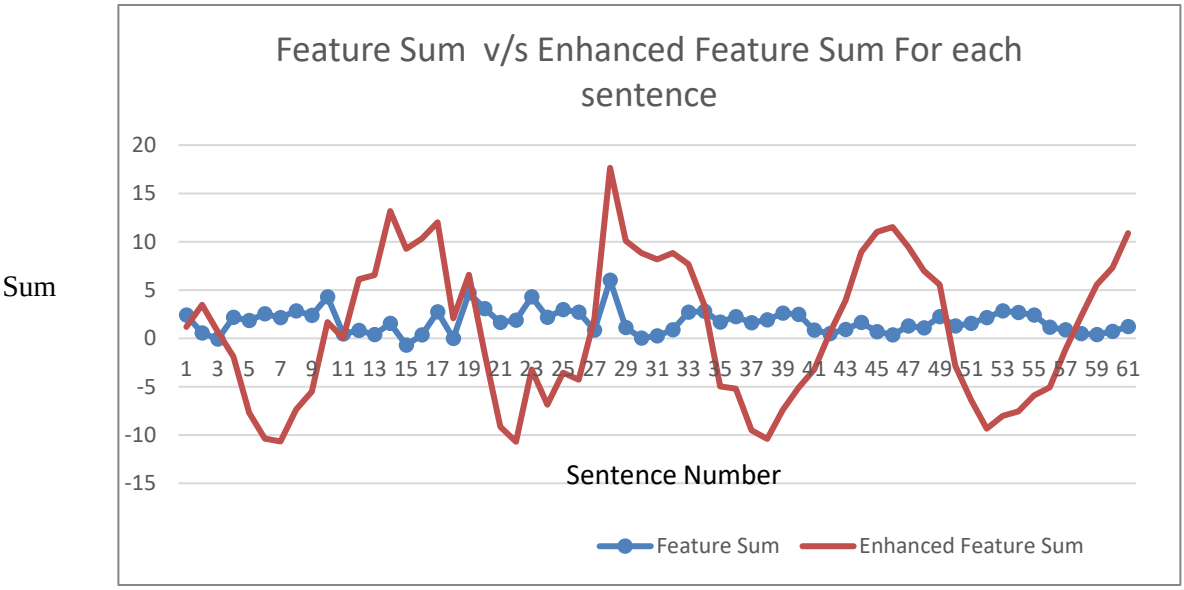

Fig4.1 Comparison between feature vector sum and enhanced feature vector

Feature extraction and enhancement is carried out for all documents. The values of feature vector sum and enhanced feature vector sum for each sentence of one document have been plotted in fig 4.1. The Restricted Boltzmann Machine extracted hierarchical representation that did not have much variation, hence discovering the interpretation. The sentence has been ranked based on final feature vector sum and summaries are generated.

Evaluation of system-generated summaries is done based on the basic measure: Precision, Recall.

2. Precision:

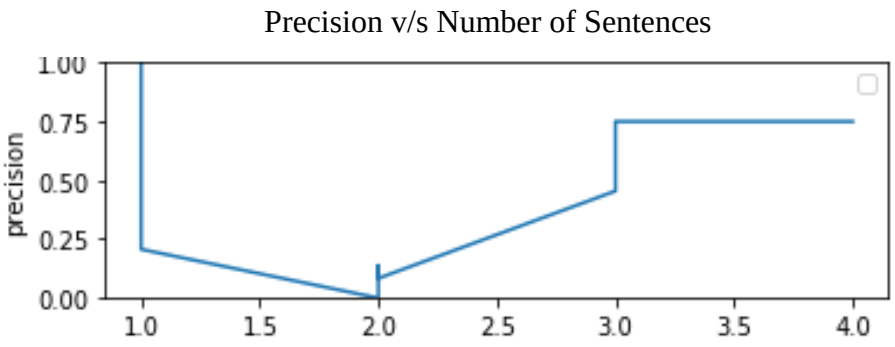

Fig 4.2: Precision values corresponding to summaries of various documents

3. Recall:

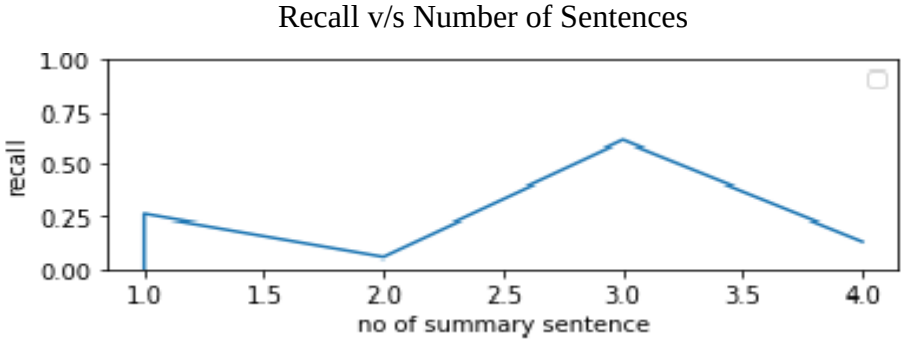

Fig4.3: Recall values corresponding to summaries of various documents

The number of sentences in the original documents with corresponding with certain of threshold, the Restricted Boltzmann Machine has sizable data to be trained successfully with precision 0.70 and recall 0.60 values are generated.

The proposed approach responds better for summarization of interpretation of suratAlfatiha.

4.3 Comparative Analysis

We make a comparative analysis of the performance of the proposed approach and existing method. Both methods are triggered based on the deep learning algorithm. The algorithm concentrates on the precision and recall values of the proposed approach and existing method and we found that the proposed gives the precision and recall respectively 0.70 and 0.6 and existing method gives 0.49, 0.46.

The result show that the proposed approach responds better as compared to the existing method.

4.4 Conclusion

We have developed a model that summarizes single document interpretation of surat Alfatiha. The algorithm runs separately for each document. We extracted 9 features from the given document and enhance them to score each sentence and comparing the system-generated summary to those written by human. The features are processed through different levels of the RBM algorithm and the text summary is generated accordingly. The experimentation of the proposed text summarization algorithm is carried out by considering six different document sets for each verse of surat alfatiha. The responses of six documents sets to the proposed text summarization are satisfactory. The performance judging parameters has got values .70 and .60 respectively for six documents sets.

4.5 Future work

The futuristic enhancement to the proposed approach can done by considering different features and by adding more hidden layers to the RBM algorithm. More enhancements by adding more features to get more relevant sentence may give better respond than proposed approach. This approach can be further develops extracting a summary from multi-document rather than single document.

References:

[1] Siya Sadashiv Naik, and Manisha Naik Gaonkar, Survey of Extractive Based Automatic Text Summarization Techniques, 2011.

[2] PadmaPriya, G. and K. Duraiswamy, An Approach for Text Summarization Using DeepLearning Algorithm, Jounral of computer science 10 (1);1-9.2014.

[3] www.quranreading.com/blog/tafsir-e-quran-its-important/.Time 04:20pm

[4] Ani Nenkova, Kathleen McKeown "Mining Text Data" 2012.Dol 10. 1007/978-1-4614-3223-4-3.

[5] Hovy.E.H., Automated text summarization in R.Mitkov, The Oxford Handbook of computational linguistics, Chapter 32 pages 583 -598,2005.

[6] Mani, I., House, D., Klein, G., et al., The Tipster SUMMAC text summarization Evaluation, In Proceedings of GACL, 1999.

[7] Edmundson, H.P., New Methods in Automatic Extracting, MIT Press, 1999.

[8] Sparck Jones, K., Automatic summarizing: factors and directions, In Mani and Maybury

[9] Strzalkowski, T., G. Stein, J. Wang, and B. Wise, A robust practical text summarizer, In Mani and Maybury.

[10] Li Deng and Dong Yu, Deep learning method and Application.

[11] Y LeCun and Y Bengio, Convolutional networks for images, speech, and timeseries, Handbook of brain theory and neural networks, 1995.

[12] Madhavi K. Ganapathiraju, Overview of summarization methods, 11-742: Self-paced lab in Information Retrieval,November 26, 2002.

[13] Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner, Automatic Text Summarization using a Machine Learning Approach, Book: Advances in Artificial Intelligence: Lecture Notes in computer science,Springer Berlin / Heidelberg, Vol 2507/2002,205-215, 2002.

[14] Khosrow Kaikhah , Text Summarization using Neural Networks, Department of Faculty Publications- Computer Science, Texas State University, eCommons,2004.

[15] Mahak Gambhir,Vishal Gupta, Recent automatic text summarization techniques: a survey , 2016.

[16] Y.leCun. Generalization and network desgin strategy. In connections in perspective, 1989.

[17] Y.leCun, K.Kavukvuoglu, and C.Farabel, Convolutional networks and applications in vision in circuits and systems, international symposium on, pages 253 – 256, 2010.

[18] D.Forsyth and J.Ponce. computer vision: A Modern Approach. Prentice Hall Professional Technical Reference, New Jersey, 2002.

[19] G.E Hinton and R.R. Salakhutdinov, Reducing the dimensionality of Data with Neural Networks, Science, 28 July 2006, Vo1.313.no.5786,pp. 504 – 507.

[20] G.E Hinton, S.Osindero, and Y.Teh, A fast learning algorithm for deep belief nets, Neural Compulation, Vol 18, 2006

[21] Hubel, D.H. and Wiesel, T.N. (1968), Receptive fields and functional artitecture of monkey striate cortex. Journal of Physiology (London), 195: 215 – 243.

[22] LISA Lab, U.o.M. (2015). Deep Learning Tutorial

[23] Sheng-hua Zhong, Yan Liu, Bin Li, and Jing Long. Query-oriented unsupervised multi-document summarization via deep learning model. Expert Systems with Applications, 42 (21):8146–8155, 2015.

[24] Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca J. Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. In ACL. 1587–1597.

[25] Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In ACL. 484–494. Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In NAACL-HLT. 93–98.

[26] Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015a. Ranking with recursive neural networks and its application to multi-document summarization. In AAAI. 2153–2159

[27] The Holy Quran English translation of the meanings and Commentary. King Fahd Holy Quran Printing Complex.

[28] Gorelick (Author), Micha; Ozsvald, Ian (September 2014). High Performance Python: Practical Performant Programming for Humans (1st ed.). O'Reilly Media. p. 370. ISBN 1449361595.

[29] Doig, Christine (21 May 2015)."Conda for Data Science". Archived from the original on 16Jun 2015. Retrieved 16 Jun 2015. Conda works with Linux, OSX, and Windows, and is language agnostic, which allows us to use it with any programming language or even multi-language projects.

[30] Edward Loper. 2004. NLTK: Building a pedagogical toolkit in Python. In PyConDC 2004. Python Software Foundation.

[31] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, pages 62–69. Somerset, NJ: Association for Computational Linguistics.

[32] Ivan Idris. NumPy: Beginner's Guide - Third Edition: June 2015.

[33] Steinberger,J.,&Jezek,K.(2012). Evaluation measures for text summarization. Computing and informatics, 28(2),251-75

[34] TORRES-MORENO, Juan-Manual; EBSCOHOST. Automatic text summarization Torres-Moreno.Inc London:Hoboken, NJ, 2014.ISBN 9781119044147 1119044146 1322166625 9781322166629 9781119004756.

Appendix A

Appendix: Case Study (Interpretation of Surat Alfatiha)

The following is a sample file from the fulltext that is being used for text summarization. Verse one from different interpreter

Verse 1

In the Name of Allah, the All-Merciful, the Most Merciful Know that one of the principles agreed upon by the Salaf and their Imams is to have faith in Allah and His Attributes, and in the rules determining how they are to be received. So, for example, they believe that he is Rahman and Rahim, possessing the quality of mercy which is extended to its recipients. Therefore, all blessings are one of the resultant effects of this mercy. This principle holds true for all of His Names. It is said concerning the Name, All-Knowing: He is All-Knowing, possessing the quality of knowledge by which He knows everything. He is All-Powerful, possessing the quality of power which makes Him omnipotent etc. The Arabic words "Rahman" and "Rahim" translated "Most Gracious" and "Most Merciful" are both intensive forms referring to different aspects of Allah's attribute of Mercy. The Arabic intensive is more suited to express Allah's attribute than the superlative degree in English. The letter implies a comparison with other being like unto Allah. Mercy may imply pity, long-suffering, patience, and forgiveness, all of which the sinner needs and Allah Most Merciful bestows in abundant measure. But there is a Mercy that goes before even the need arises, the Grace which is ever watchful, and flows from Allah Most Gracious to all His creatures, protecting them, preserving them, guiding them, and leading them to clear light and higher life.

Verse 1.2

In the Name of Allah, the All-Merciful, the Most Merciful

Opinion is divided whether Bismillah should be numbered as a separate verse or not. It unanimously agreed that it is a part of the Qur-an in Sura An-Naml. Therefore it is better to give it an independent number in first Sura. For subsequent Suras it is treated as an introduction or head-line, and therefore not numbered.

One of the many practices taught by Islam is that its followers should begin their activities in the name of God. This principle, if consciously and earnestly followed, will necessarily yield three beneficial results. First, one will be able to restrain oneself from many misdeed, since the habit of pronouncing the name of God is bound to make one wonder when about to commit some offence how such an act can be reconciled with the saying of God's holy name. Second, if a man pronounces the name of God before starting good and legitimate tasks, this act will ensue that both his starting point and his mental orientation are sound. Third - and this is the most important benefit - when a man begins something by pronouncing God's name, he will enjoy God's support and succour; God will bless his efforts and protect him from the machinations and temptation of Satan. For whenever man turns to God, God turns to him as well.

Verse 1.3

In the name of Allah, the Merciful, the Compassionate

One of the many practices taught by Islam is that its followers should begin their activities in the name of God. This principle, if consciously and earnestly followed, will necessarily yield three beneficial results. First, one will be able to restrain oneself from many misdeed, since the habit of pronouncing the name of God is bound to make one wonder when about to commit some offence how such an act can be reconciled with the saying of God's holy name. Second, if a man pronounces the name of God before starting good and legitimate tasks, this act will ensue that both his starting point and his mental orientation are sound. Third - and this is the most important benefit - when a man begins something by pronouncing God's name, he will enjoy God's support and succour; God will bless his efforts and protect him from the machinations and temptation of Satan. For whenever man turns to God, God turns to him as well.

Verse 1.4

In the Name of Allah, the All-Merciful, the Most Merciful.

"In the Name of Allah," meaning: I begin with every Name belonging to Allah, Most High. This is because the word Name is singular and in the genitive form; therefore it subsumes all of Allah's Beautiful and Perfect Names. "Allah," He is the God who is worshipped: the one deserving worship to the exclusion of everything else because of His qualities of divinity, attributes, all of which are perfect. "The All-Merciful, the Most Merciful," these are two Names proving that He, Most High, is one who possesses a great and all-encompassing mercy that includes everything and embraces every living being. He has decreed it for those who fear Allah, those who follow His Prophets and Messengers: these have unrestricted mercy; anyone else only has a portion of this mercy.

Know that one of the principles agreed upon by the Salaf and their Imams is to have faith in Allah and His Attributes, and in the rules determining how they are to be received. So, for example, they believe that he is Rahman and Rahim, possessing the quality of mercy which is extended to its recipients. Therefore, all blessings are one of the resultant effects of this mercy. This principle holds true for all of His Names. It is said concerning the Name, All-Knowing: He is All-Knowing, possessing the quality of

knowledge by which He knows everything. He is All-Powerful, possessing the quality of power which makes Him omnipotent etc.

Verse 1.5

In the Name of Allah, the All-Merciful, the Most Merciful.

The Arabic words "Rahman" and "Rahim" translated "Most Gracious" and "Most Merciful" are both intensive forms referring to different aspects of Allah's attribute of Mercy. The Arabic intensive is more suited to express Allah's attribute than the superlative degree in English. The letter implies a comparison with other being like unto Allah. Mercy may imply pity, long-suffering, patience, and forgiveness, all of which the sinner needs and Allah Most Merciful bestows in abundant measure. But there is a Mercy that goes before even the need arises, the Grace which is ever watchful, and flows from Allah Most Gracious to all His creatures, protecting them, preserving them, guiding them, and leading them to clear light and higher life.

Opinion is divided whether Bismillah should be numbered as a separate verse or not. It unanimously agreed that it is a part of the Qur-an in Sura An-Naml. Therefore it is better to give it an independent number in first Sura. For subsequent Suras it is treated as an introduction or head-line, and therefore not numbered.

Verse 1.6

In the Name of God, the Compassionate, the Merciful.

The bismillah ['in the Name of God'] is a grammatical particle of implication. That is to say 'by means of God', new things become manifest and by means of Him created things exist. There is nothing from any newly created thing or sequence of events; or from any perceived thing or trace of a thing, etc.; or anything else from rocks or clay, grass or trees, any impression left on the ground or standing remains, or any judgment or causes, that have existence except by means of the Real. The Real is its sovereign. Its beginning is from the Real and its return is to the Real. Through Him the one who declares the unity [of God] finds and through Him the rejecter abandons faith. Through Him the one who acknowledges knows, and through Him the one who perpetrates lags behind.

He said, 'In the Name of God' rather than 'In God'. According to some people, this is a way of seeking blessing through mentioning His name. According to others, it is because of the difference between this [wording] and oaths. According to scholars, it is because the name is the thing that is named. In the view of the people of mystical knowledge, [the wording is such] in order to seek the purification of hearts from attachments and the liberation of the innermost selves from obstacles so that the word 'God' may enter into a clean heart and purified innermost self.

Upon the mention of this verse, some people are reminded from the [letter] of His beneficence with His friends, and from the [letter] of His secret with his chosen ones, and from the [letter] of His grace to the people of His friendship. They know that by His beneficence, they come to know His secret, and by His grace to them, they preserve His command, and by Him (glory be to Him Most High) they recognize His measure.

Other people, upon hearing 'In the Name of God', are reminded by the [letter] of the immunity of God from every evil, and by the [letter] of His soundness from any defect, and by the [letter] of His magnificence in the exaltedness of His description.

Others are reminded at the [letter] of His splendor, and at the [letter] of His radiance, and at the [letter] of His dominion.

Because God has repeated the verse In the Name of God, the Compassionate, the Merciful in every sura and it has been established that it is part of them, we want to mention in every sura non-repetitive and

non-reiterative utterances taken from the allusions of this verse. Because of this we will not examine the words exhaustively here. Through Him there is confidence.

Appendix B

Appendix: Summary

The following is a sample summary of interpretation of verse 1

Summary 1

In the Name of Allah, the All-Merciful, the Most Merciful Know that one of the principles agreed upon by the Salaf and their Imams is to have faith in Allah and His Attributes, and in the rules determining how they are to be received.Therefore, all blessings are one of the resultant effects of this mercy.This principle holds true for all of His Names.It is said concerning the Name, All-Knowing: He is All-Knowing, possessing the quality of knowledge by which He knows everything.He is All-Powerful, possessing the quality of power which makes Him omnipotent etc.The letter implies a comparison with other being like unto Allah

Summary 1.2

In the Name of Allah, the All-Merciful, the Most Merciful Opinion is divided whether Bismillah should be numbered as a separate verse or not. Therefore, it is better to give it an independent number in first Sura. For subsequent Suras it is treated as an introduction or head-line, and therefore not numbered. This principle, if consciously and earnestly followed, will necessarily yield three beneficial results. First, one will be able to restrain oneself from many ;.misdeeds, since the habit of pronouncing the name of God is bound to make one wonder when about to commit some offence how such an act can be reconciled with the saying of God's holy name. For whenever man turns to God, God turns to him as well.

Summary 1.3

In the name of Allah, the Merciful, the Compassionate One of the many practices taught by Islam is that its followers should begin their activities in the name of God.This principle, if consciously and earnestly followed, will necessarily yield three beneficial results.First, one will be able to restrain oneself from many misdeed, since the habit of pronouncing the name of God is bound to make one wonder when about to commit some offence how such an act can be reconciled with the saying of God's holy name. Second, if a man pronounces the name of

God before starting good and legitimate tasks, this act will ensue that both his starting point and his mental orientation are sound.

Summary 1.4

In the Name of Allah, the All-Merciful, the Most Merciful."Allah," He is the God who is worshipped: the one deserving worship to the exclusion of everything else because of His qualities of divinity, attributes, all of which are perfect.He has decreed it for those who fear Allah, those who follow His Prophets and Messengers: these have unrestricted mercy; anyone else only has a portion of this mercy.Therefore, all blessings are one of the resultant effects of this mercy.This principle holds true for all of His Names.It is said concerning the Name, All-Knowing: He is All-Knowing, possessing the quality of knowledge by which He knows everything.He is All-Powerful, possessing the quality of power which makes Him omnipotent etc.

Summary 1.5

In the Name of Allah, the All-Merciful, the Most Merciful.The letter implies a comparison with other being like unto Allah.But there is a Mercy that goes before even the need arises, the Grace which is ever watchful, and flows from Allah Most Gracious to all His creatures, protecting them, preserving them, guiding them, and leading them to clear light and higher life.Therefore it is better to give it an independent number in first Sura.For subsequent Suras it is treated as an introduction or head-line, and therefore not numbered.

Summary 1.6

In the Name of God, the Compassionate, the Merciful.The bismillah ['in the Name of God'] is a grammatical particle of implication.; or anything else from rocks or clay, grass or trees, any impression left on the ground or standing remains, or any judgment or causes, that has existence except by means of the Real. Its beginning is from the Real and its return is to the Real. Through Him the one who declares the unity [of God] finds and through Him the rejecter abandons faith. He said, 'In the Name of God' rather than 'In God'. According to scholars, it is because the name is the thing that is named. They know that by His beneficence, they come to know His secret, and by His grace to them, they preserve His command, and by Him (glory be to Him Most High) they recognize His measure. Others are reminded at the [letter] of His splendor, and at the [letter] of His radiance, and at the [letter] of His dominion. Because God has repeated the verse In the Name of God, the

Compassionate, the Merciful in every sura and it has been established that it is part of them, we want to mention in every sura non-repetitive and non-reiterative utterances taken from the allusions of this verse.