Automatic Text Summarization for Sinhala

A thesis submitted for the Degree of Master of Philosophy

W V Welgama

University of Colombo School of Computing

December 2012

Declaration

The Thesis is my original work and has not been submitted previously for a degree at this or any other university/institute. To the best of my knowledge it does not contain any material published or written by another person, except as acknowledge in the text.

Author's name: Mr. W Viraj Welgama Date: 04 – 12 - 2012

Signature: ……..……

This is to certify that this thesis is based on the work of Mr. Welgamage Viraj Welgama under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by

Supervisor Name: Dr. A R Weerasinghe Date: ……..……

Signature: ……..……

Acknowledgement

First and foremost I offer my sincerest gratitude to my supervisor Dr. A R Weerasinghe, the head of the Language Technology Research Lab (LTRL) and a Senior Lecturer of University of Colombo School of Computing who has always supported and guided me throughout my research with his vast experience and valuable knowledge. It is impossible to carry out this research without his encouragements and patience. He has made available his support in a number of ways and I am feeling lucky to have such a supervisor for my academic carrier.

Then I owe my deepest gratitude to Dr. Inderjeet Mani, a senior principal scientist at the MITRE Corporation and the author of two very valuable books on Text Summarization, who visited the LTRL during my research period and gave his valuable advices and comments on my research. This thesis would not have been possible unless the kind donation of his book *Automatic Summarization* to the LTRL. This book is greatly helped me to gather the comprehensive background knowledge of the field.

I am grateful to Dr. Dilhari Attygalle, the Head of the Department of Statistics, University of Colombo, and two of her assistants who supported me for the analysis of results. Their advices on statistical analysis were greatly helpful for me to finalize my experimental results.

I would like to express my gratitude to my colleague Mr. Dulip Herath who always shares his vast knowledge of the field with me and gives valuable suggestions and feedbacks at the crucial stages of the research. Mr. Vincent Halahakone, Mr. Nishantha Medagoda and Mr. Asanka Wasala the former members of the LTRL who were always giving me enthusiastic encouragements which seriously pushed me to complete this research timely.

It is an honor for me to give my sincere gratitude to Mr. Namal Udalamatta, Mr. Chamila Liyanage and Ms. Thilini Nadungodage who voluntarily stood to be the human annotators. This research is not possible without their expert knowledge on human summarizing and their patience on annotating 120 editorials.

I am grateful to all my colleagues at the LTRL and the UCSC who always gave their maximum support and encouragements to complete this research. Last but not least, I wish to acknowledge the patience and endurance of my loving wife, Niluka who never complained when I had to spend many time over nights and weekends to complete this research. Appealing smile of my loving baby Vihan was also greatly helpful me to relax for a moment during this hard work.

Abstract

With the rapid development of information and communication technology, people are surrounded with vast amounts of information albeit with less and less time or ability to make sense of it. The field of automatic summarization which has been in existence since the 1950's is anticipated to find solutions to this issue. With the adaptation of Unicode technology in 2004, the Sinhala language began to appear in computers rapidly and Sinhala language users also began to experience the above issue. This research on Automatic Text Summarization in Sinhala is carried out to find the possible approaches to address the above issue with the minimum linguistic resources.

The field of automatic text summarization began with some classical approaches which attempted to indentify the most salient information of an article using some thematic features. This research was intended to indentify such features for the Sinhala language with the most suitable approach to define each of these features for achieving accurate summaries. In order to benefit from all these features, this research proposes a best possible linear combination of identified features.

The proposed method was evaluated by comparing the machine generated and human extracted summaries based on the primary assumption that the human summaries are perfect. Results show that the sentence location feature is the best individual feature for extracting most informative sentences from Sinhala articles while the linear combination of keyword feature, title words feature and the sentence location feature giving the best performance for a summarizer. Results revealed some equations to define the flow of information over a Sinhala article which can be used in many such applications. Further, this research provides a benchmark for future research on Sinhala automatic text summarization.

Table of Contents

Acknowledgement - - - - - - - - - - - - ii

Abstract - - - - - - - - - - - - - - iii

List of Figures - - - - - - - - - - - - - vi

List of Tables - - - - - - - - - - - - - vii

Chapter 01 – Introduction - - - - - - - - - - - 01

1.1 Overview - - - - - - - - - - - - 01

1.2 Motivation - - - - - - - - - - - - 03

1.3 Objectives - - - - - - - - - - - - 04

1.4 Scope - - - - - - - - - - - - 05

1.5 Outline of the Thesis - - - - - - - - - - - 06

Chapter 02 – Background - - - - - - - - - - - 07

2.1 Basic Norms of Text Summarization - - - - - - - - 07

2.1.1 Genres of a Summary - - - - - - - - - 07

2.1.2 Summarization Parameters - - - - - - - - 08

2.1.3 Aspects of Summarization - - - - - - - - 09

2.1.4 Summarization Machine - - - - - - - - 09

2.2 History of Automatic Summarization - - - - - - - - 10

2.3 Summarization Approaches - - - - - - - - - 11

2.3.1 Classical Approaches - - - - - - - - - 12

2.3.2 Corpus based Approaches - - - - - - - - 13

2.3.3 Exploiting Discourse Structure- - - - - - - - 16

2.3.4 Knowledge-Rich Approaches - - - - - - - - 17

2.4 Summarization for Indic Languages - - - - - - - - 18

Chapter 03 – Methodology - - - - - - - - - - **21**

3.1 The Edmundonian Paradigm - - - - - - - - - - 21

3.2 Adopting Edmundonian Paradigm for Sinhala - - - - - - 23

3.2.1 Identifying Keywords - - - - - - - - - - 24

3.2.2 Title Words - - - - - - - - - - 28

3.2.3 Sentence Location - - - - - - - - - - 29

3.2.4 Paragraph Location - - - - - - - - - - 32

3.3 Data-set - - - - - - - - - - - - - 33

3.3.1 Cleaning the Articles - - - - - - - - - - 35

3.3.2 Splitting Sentences - - - - - - - - - - 36

Chapter 04 – Experiments and Results - - - - - - - - **40**

4.1 Creating Manual Summaries - - - - - - - - - - 40

4.2 Defining the Evaluation Criteria - - - - - - - - - - 42

4.3 Identifying the Best Approach for each Individual Feature - - - - 43

4.3.1 Weighting for Keywords - - - - - - - - - - 43

4.3.2 Weighting for the Sentence Location - - - - - - - 45

4.3.3 Weighting for the Paragraph Location - - - - - - - 46

4.4 Tuning the Parameters - - - - - - - - - - - 48

4.5 Experiments with the Test Data - - - - - - - - - - 53

Chapter 05 – Conclusion and Future Works - - - - - - - **55**

5.1 Conclusion - - - - - - - - - - - - - 55

5.2 Future Works - - - - - - - - - - - - - 58

References - - - - - - - - - - - - - - 60

Appendix A - Samples of Source Article, Human Extracted Summaries and Machine Extracted Summaries - - - - - - - - I

Appendix B - Parameter Tuning Data - - - - - - - - - VII

List of Figures

Figure 2.1: A high-level architecture of a summarizer - - - - - - 09

Figure 3.1: The lightweight algorithm defined for stemming Sinhala words - - 27

Figure 3.2: Linear function defined to assign weight for the sentence location - - 30

Figure 3.3: Hyperbolic function defined to assign weight for the sentence location - 31

Figure 3.4: Quadratic function defined to assign weight for the sentence location - 32

Figure 3.5: Distribution of number of sentences per article - - - - - 35

Figure 3.6: ZWJ character appeared in unwanted places in a paragraph - - - 36

Figure 3.7: Algorithm defined for the sentence boundary detection - - - - 37

Figure 3.8: Distribution of number of sentences per paragraph - - - - - 38

Figure 3.9: Distribution of number of words per sentence - - - - - 39

Figure 4.1: Algorithm for selecting articles for human annotation - - - - 40

Figure 4.2: Distribution of number of paragraphs per article - - - - - 41

Figure 4.3: Algorithm defined to assign all possible values for parameters - - - 49

List of Tables

Table 3.1: Sample of results of the lightweight stemming algorithm - - - - 26

Table 3.2: Basic statistics of the Sinhala Editorials Corpus - - - - - 34

Table 4.1: F-Score values calculated based on KBA, DRA and WSW for 10 articles - 44

Table 4.2: Mean, Standard Deviation and CV values for DRA, KBA and WSW - - 44

Table 4.3: F-Score values generated by the sentence location feature, calculated based on three defined equations for 15 articles - - - - - 45

Table 4.4: Mean, Standard Deviation and CV values for three equations based on sentence location feature - - - - - - - - - 46

Table 4.5: F-Score values generated by the paragraph location feature, calculated based on three defined equations for 15 articles - - - - - 47

Table 4.6: Mean, Standard Deviation and CV values for three equations based on paragraph location feature - - - - - - - - - 47

Table 4.7: First 20 combinations generated from the algorithm defined in figure 4.3 - 50

Table 4.8: Averaged F-Score values for first 20 combinations of α, β, γ and δ - - 51

Table 4.9: Averaged F-Score values for each individual feature and for best three combinations - - - - - - - - - - - 52

Table 4.10: Averaged F-Score values for the best three combinations with and without stemming the words - - - - - - - - 53

Table 4.11: Averaged F-Score values for the best three combinations - - - - 53

Table 4.12: Averaged F-Score values for the worst three combinations - - - 54

Chapter 01 – Introduction

This thesis explains the research carried out to summarize Sinhala language text automatically and to present the results of the proposed approach. This chapter will give an overview of the topic with the motivation behind the research and its main objectives. The scope which was defined for the research will be explained and finally the flow of the rest of the thesis will be presented.

1.1 Overview

With the rapid development of information technology, the world is flooded with information. Also information has become the most valuable and important resource of this fast growing information society. However, according to the theory of information, the value of information is inversely proportional to the time taken to access such information. Therefore, the most important fact is to access the right information at the right time. On the other hand, as a result of commercialization, time has become the most valuable factor in almost all day to day activities of people. A day of an average person is packed full of activities and it is hard to find time to read or listen to all relevant information available for him. This situation leads the people to find a solution to get the maximum benefit from both information and time by balancing these two important factors. This is where "Summarization" becomes all important.

Summarization is a process of squeezing the most important information from a source and presenting it in a way that people can grasp as much information as possible in a short time. Summaries are everywhere. It is hard to imagine a day of an average person without summaries. Newspaper headlines, previews of movies, abstracts of scientific articles, score tables of games, road maps, TV program guides, minutes of a meeting, a program of a conference, weather forecasts, stock market bulletins, library catalogues, obituaries, the menu in a restaurant, table of contents of a book are all kinds of summaries we daily deal with (Mani, Automatic Summarization, 2001). These summaries can be in video, audio, image or in text form. People and companies are expending significant time and money to generate these summaries to keep their busy customers up to date with their products and services.

Since inventing the programmable computer in the middle of the 20th century, people have been trying to handover most of their routine activities to computers. Use of computers has spread rapidly to all the fields since computers were able to process data rather than being just used for typesetting documents. Much research has been carried out by computer scientists over the last six decades to enable computers to work with human languages in the field of Natural Language Processing (NLP). Automatic Text Summarization is one of the major

research fields in NLP, which researchers are trying to automatically summarize information from one or many sources.

The goal of automatic summarization is *to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs* (Mani, Automatic Summarization, 2001). To achieve the above goal completely, computers have to understand the information available in the source and have to be able to regenerate the gist of that information based on the user requirement. Many techniques including data driven approaches and machine learning techniques have been used over the last six decodes to achieve the above goal to a certain extent. It is hard to determine the quality of a good summary since it depends on many parameters such as the user's background knowledge, user requirements and compression ratio among others, but scientists have been able to reach a level such that computers are able to generate human consumable summaries.

The United States, European Community and Pacific Rim countries have identified the importance of automatic text summarization and they have begun to invest on it (Mani & Maybury, Advances in Automatic Text Summarization, 1999). Automatic summarization is increasingly being exploited in commercial sector applications such as in the telecommunication industry, data mining and databases, filters for web-based information retrieval and word processing tools such as Microsoft Word. British Telecom produced a summarization tool called *Prosum* for both offline and online texts which works by selecting key sentences and extracting them as a summary. *Inxight's summarizer* used in AltaVista Discovery search engine is using summarization techniques to filter for web-based information retrieval. IBM Japan has invested on two summarization tools called *Internet King of Translation* (Japanese) and *Lotus Word Pro* (Japanese Version) and those applications are now being used in commercial level. Text Analysis toolset owned by IBM has one component for automated text summarization while the Apple Company also uses summarization tool for their word processing applications.

The application of automated summarization is not just limited only for generating summaries, but to many kinds of other NLP applications such as Search Engines, Intelligence Gathering Systems and others. These applications use summarization techniques to capture the most relevant information and present them appropriately as an essential part of their main function. Some research has been carried out to find ways to adopt summarization techniques to hand-held electronic devices such as mobile phones (Nakao, 2000). These attempts are essential and helpful for people to ease their day to day activities.

Sinhala, the mother tongue of the majority of Sri Lankans is one the official and national language of Sri Lanka. It is spoken widely in the island except in the north and some parts of the east and around three million people worldwide speak Sinhala (Wikipedia). Sinhala belongs to the Indo-Aryan branch of the Indo-European languages and the languages such as Hindi and Gujarati are siblings of Sinhala in the language family tree. However, due to the separation of Sinhala by the Dravidian Belt from its Indo-Aryan roots, Sinhala has evolved into a unique language in the Indo-Aryan family. Sinhala language has been influenced by the Tamil language and then Portuguese, Dutch and English language terms were combined with Sinhala words due to colonialism.

With the introduction of microcomputers in the early 1980's, Sri Lanka too embarked on the use of computers with local language input and output. Basic infrastructure facilities like fonts and keyboard drivers have been developed for Sinhala and then Sinhala language scripts begun to appear on computer screen. Initially computers used mainly for printing purposes and all the technology made were focused on printing. However, the experiments on Sinhala language processing using computers have been done in micro level, especially as undergraduate research projects in local universities. This environment provided necessary background to develop NLP research on Sinhala language but the applicability on state-of-the-arts technologies was poor due to lack of required linguistic resources. With the introduction of Unicode technology in 2004, the situation was changed to the positive direction by opening opportunities to researchers to carry out more NLP research on Sinhala.

1.2 Motivation

With the introduction of Unicode technology, many languages were able to work with computers directly in their own scripts. Sinhala script appeared in computers with the Unicode standard from 2004 solving many standardizing problems caused hitherto by proprietary encodings. It also allowed people to generate e-contents, store them and publish securely without having any issues, especially with exchanging them with others. Most of the information generators including daily newspapers started publishing their contents on the Internet and therefore Sinhala language users started having to deal with the problem of information overload as of recent.

Apart from storing and representation, the Unicode standard allows us to process languages. Basic infrastructures such as input methods, fonts and rendering engines were built gradually and linguistic resources such as corpora, lexicons and tag sets among others were developed on top of these infrastructure facilities. This situation motivated local language researchers to

apply existing language processing theories and develop new theories for their mother tongues.

With the rapid increase of local language contents in electronic form and the gradual improvements of Sinhala language resources for computational models, the possibility of developing some language processing applications for Sinhala has increased. An automatic summarizer was one such major application which many people can benefit it through because it helps people to get the most important and relevant information in a shorter time. The motivation of developing an automatic text summarizer for Sinhala was empowered under these circumstances.

To be alive in this rapidly changing cyber world, languages have to adopt the technology and have to be represented on the Internet. Much research has been carried out over the last six decades to adopt technology to most common languages such as English, because the technology was born with these languages. Later, when the other languages were enabled with the technology, the first step was to apply the existing techniques and findings to these languages rather than reinventing techniques for the same issues. This helps such languages to adapt to the technology rapidly in shorter time and less cost while it also helps linguists to identify the language families based on the adaptability. This scenario also motivated the author to apply such existing summarizing techniques to Sinhala and find their applicability to languages such as Sinhala.

1.3 Objectives

The major objective of this research is to find the most suitable approach to summarize text written in Sinhala. Since Sinhala is considered as a less resourced language in the field of NLP, the challenge is to find an approach which does not need many linguistic resources in order to achieve acceptable performance. Even though the current state of the art is using rich linguistic resources such as summary corpora, annotated data, parsers, WordNet and named entity recognizers among others, the techniques developed at the early stage of automatic summarization did not use such resources. Therefore, one of the goals in this research is to apply techniques suitable for law-resourced languages and to find the adaptability of them to a language such as Sinhala, which comes from a different language family.

There were no previous attempts recorded in the literature for developing an automated summarizer for Sinhala. As such we can assume that no studies have been carried out to find how the features of Sinhala language behave in a summarizing context. Therefore, finding out how significant information is distributed over an article, how Sinhala writers use paragraphs

for the flow of information and how they focus into the title are some supplementary objectives of this research.

When Sinhala language is adopting technology, many linguistic resources for Sinhala will be developed in the future and future researchers will be able to find further advanced techniques to summarize Sinhala text. However, there would be a point of previous reference which they can use to compare their results and prove the success of their new approaches. Therefore, one of the other goals of this research is to provide a benchmark for automatic Sinhala text summarization for future researches on Sinhala language.

1.4 Scope

Sinhala sentences can have many different structures. Even though some artificially generated basic simple sentences are commonly used especially for language teaching purposes, real world sentences have more complex structures. Moreover, Sinhala has specific word separation policies, but writers often use their own policies even though there are some attempts to formalize it (NIE, 1989). Therefore, the sentence structures of an arbitrary text would be more complex to analyze using computers especially without some linguistic resources such as taggers, parsers and other tools.

To overcome the above issues, this research used editorials of three national newspapers namely *Dinamina*, *Lankadeepa* and *Divaina*. It was assumed that editorials are written by the chief editor of the newspaper, who is professional to write in refined Sinhala. Furthermore, all the editorials have an appropriate title and are around 50 sentences in length, which is more suitable for the proposed design of the research. Selected articles from this editorial corpus are manually annotated by three human annotators and this average length of an editorial is neither too long nor too short for such tedious manual task.

As explained in the introduction, Sinhala is considered a less resourced language in the field of NLP and therefore it is difficult to use the recent approaches used in languages such as English in the field of Automatic Text Summarization. Therefore, the scope of this research was limited to apply technologies applicable to low-resourced languages, to find out the most suitable factors for achieving accurate summaries automatically.

1.5 Outline of the Thesis

The next chapter of the thesis will explain the technical background of automatic text summarization with a comprehensive literature review carried out to find the current status of the field. Basic norms of summarization will be explained with different approaches and resources used over the last six decades to automatically summarize text. Methodology and the resources used to carry out the research will be described under the Methodology Chapter (Chapter 03) while the experiments carried out with the proposed methodology and then the results will be discussed under Experiments and Results Chapter (Chapter 04). Evaluation of the research will be explained in the same chapter with comparison of results. The author's view of the final results and then the research conclusion and possible future works will be discussed in the last chapter (Chapter 05). References which were used to carry out the research will be listed at the end of the thesis.

Chapter 02 – Background

This chapter explains the technical background behind the automatic summarization and the comprehensive study carried out to find the current status of the field. Different approaches that have been taken over the last six decades to automatically summarize text will be explained and the applicability of such approaches for less resourced languages such as Sinhala will be discussed.

2.1 Basic Norms of Text Summarization

2.1.1 Genres of a Summary

Summaries can be viewed in many dimensions. One angle would be the relationship between the summary and its input and the fundamental distinction between *Extracts* and *Abstracts* can be seen through it. Extracts contain the exact sentences appeared in its input while the abstracts are rewritten forms of the input. Extract need not consist of sentences but it may consist of a list of technical terms, proper nouns, noun phrases, truncated sentences among others. Abstracts contain at least some materials which are not present in its input. However, a short abstract may offer more information than a longer extract.

Another way to look at summaries is in terms of the traditional distinction between *Indicative* and *Informative* summaries (Borko & Bernier, 1975). Indicative summaries provide a reference function for selecting documents for more in-depth reading while informative summaries are aimed at helping the user to decide whether to read the information source or not. In the standard guidelines provided for abstractors by American National Standards Institute (ANSI) has specified that the indicative summaries are to be used for less-structured documents like editorials, essays, annual reports and others, whereas informative summaries are generally used for other documents. Also, it has been mentioned that, in scientific investigation reports, an indicative summary should contain information about the article's purpose, scope and approach but not the results, conclusions and recommendations while an informative summary should cover all of these aspects (ANSI, 1997).

Another dimension of viewing summaries is the type of users that the summary is intended for. Two different summary types can be seen through it namely *User-Focused* summaries and *Generic* summaries. User-Focused summaries (also called topic-focused summaries or query-focused summaries) are for specific user or user groups and some users' interest will be taken into account when making summaries. User query and user background knowledge of the subject are most important factors for user-focused summaries. Generic summaries are aimed

at a particular readership community and traditionally those are written by professional abstractors served as surrogates for full text. However, user-focus summaries have increasing importance in computing environments since it is always able to capture user's requirements and the interest.

2.1.2 Summarization Parameters

Automatic summarization is a highly interdisciplinary application, involving natural language processing, information retrieval, library science, statistics, cognitive psychology and artificial intelligence (Mani, Automatic Summarization, 2001). Therefore, many parameters from these paradigms are involved to fine-tune the summary against its input. There can be many lists for these parameters albeit most common parameters can be described as follows.

Compression Rate is the typical parameter for every summary, which is the ratio between the summary text length and the source text length. It allows user to determine how much information he needs from the source and usually it is set anywhere from 5% to 30% (Mani, Automatic Summarization, 2001). *Function* allows user to select the types of summaries he needs. That can be just an indication of topics or informative as to content or evaluation of the content. *Audience* is the parameter to set the user's type. It can be either user-focused summary or generic summary. *Relation to the source* is to select whether user needs extracted summary or abstracted summary.

Summaries can be generated using either from a single document or from multiple documents. That can be set from the parameter called *Span*. Summaries can be monolingual (processing a single language and give the output in the same language) or multilingual (processing several languages and give the output in the same language as input) or cross-lingual (processing several languages and give the output in a different language from input) and *language* parameter can be set to get one of these values.

Summarizer will use different strategies for various types of text such as scientific or technical reports, news stories, email messages, editorials, books and others. *Genre* of a summarizer is to set such different varieties of the input. Summaries can take different media types such as text, audio, tables, pictures and diagrams and movies as the input and can produce the output in one of these different forms. *Media* can be set to indicate this feature for a summarizer.

Importance of these parameters will vary according to the application. It is unlikely that any single summarizer will handle all of these parameters. However, the summarizers are built including only the relevant parameters to satisfy the purpose of the summarizer.

2.1.3 Aspects of Summarization

A summary can be described mainly using three aspects, namely *Input*, *Purpose* and *Output* (Hovy & Marcu, Automated Text Summarization Tutorial — COLING/ACL'98, 1998). The domain of the source text, genre of the source (newspaper articles, editorials, letters, technical reports, emails etc.), form of the source text (whether it is a regular text structure or a free-form) and the source text size (single document or multi documents) are the parameters for the aspect of input. These parameters can be set to define the input form and then the output will depend on it. Purpose of a summary can be described based on the situation, audience and usage of the summary. Audience can be a focus group which has some background knowledge about the source or it can be a general audience. Output of the summary depends on its completeness, format and the style. Completeness is to indicate the level of user requirements while the style is to set the output form of the summary. It can be informative, indicative, aggregative or a critical summary. The format of the output will be a paragraph or a table or a chart.

2.1.4 Summarization Machine

If the summarizer is considered as a machine, the typical architecture of it will be as in figure 2.1. It references some parameters described above and basic three phases in automatic summarization.

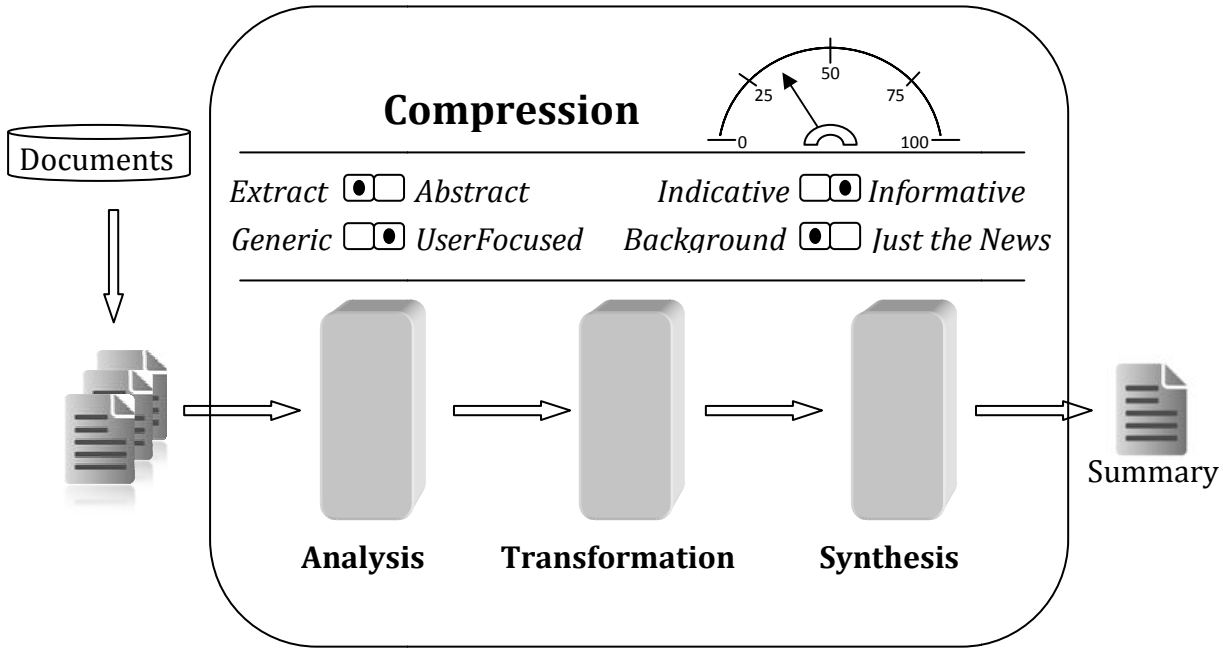

Figure 2.1: A high-level architecture of a summarizer (Source: Mani, Automatic Summarization, 2001)

Researchers have identified three basic phases in text summarization, namely *Analysis*, *Transformation* and *Synthesis*. Summarizer analyzes the input and builds an internal representation of the input in the analysis phase. Transforming the internal representation into a representation of a summary is happening in the transformation phase. Summary representation is turned back into natural language in the synthesis phase. These two phases are mostly applicable to the systems which produce abstracts or which perform compaction or multi-document summarization. Systems which produce single-document extracts without compaction will be directly going from the analysis phase to the output.

Three basic condensation operations which summarizers carry out can be identified in any of the above phases. *Selection* is the operation for filtering of elements to obtain more salient information from the input. *Aggregation* is for merging the identified elements which were identified in the previous operation. Finally, the operation called *Generalization* is the substitution of elements with more general or abstract ones to make the summary. Other more complex operations such as *paraphrasing* or *simplification* can be described in terms of these three basic operations (Mani, Automatic Summarization, 2001).

2.2 History of Automatic Summarization

Experiments on summarizing text using computers were begun in the late 1950's by characterizing surface level approaches. Luhn describes a simple, genre-specific approach that uses term frequencies for weighting sentences which are then extracted to make abstracts (Luhn, 1958). This work can be considered as the first computational paper on automated extraction (Mani & Maybury, Advances in Automatic Text Summarization, 1999). Luhn was motivated by the need of dealing with information overload and it indicates that the problem of information overloading existed even before the 1950's. Rath, Resnick, and Savage have used five different word frequency and distribution based sentence selection approaches as Luhn did in his work (Rath, Resnick, & Savage, 1961). Using these thematic features such as word frequency gave a positive start for the research in automatic summarization.

In the early 1960's, researchers started to use entry level approaches based on syntactic analysis. Climenson, Hardwick and Jacobson's work has used such syntactic analysis for machine indexing and abstracting (Climenson, Hardwick, & Jacobson, 1961). Using the sentence location as a feature was introduced to the field in 1969 by Edmundson (Edmundson, 1969). He has used additional three features in addition to word frequencies, namely cue phrases, title and heading words and the sentence location. He has found that the combination of cue phases, title words and the sentence location was the best features. He also has

mentioned that the location being the best individual feature whiles the keywords alone the worst performing features. When early 1970's, there was a renewed interest in the field which led to develop first commercial application for automatic abstracting. Pollock and Zamora have developed an automatic abstractor for the Chemical Abstracts Service (CAS) mainly using cue phrases specific to chemistry sub domain which they later used as a commercial product (Pollock & Zamora, 1975).

More extensive entry level approaches have been used in the late 1970's. First discourse-based approaches based on story grammars were experimented in this time. Correira's work on computing story trees was one of early attempts for such approaches (Correira, 1980). Entry level approaches based on artificial intelligence such as use of scripts, logic and production rules, semantic networks as well as some hybrid approaches were experimented in the 1980's (Mani & Maybury, Advances in Automatic Text Summarization, 1999).

In the late 1990's the field of automatic summarization grew aggressively with all type of approaches being explored already due to the government and commercial interest for the applications. Currently the research works have exclusively focused on extracts rather than abstracts along with a renewed interest in earlier surface-level approaches. However, more natural language generation works have been begun to focus on automatic summarization and the field is now exploring new areas such as multi-document summarization, multi lingual summarization and multimedia summarization rather than focusing on single document text summarization.

2.3 Summarization Approaches

Basic methods of automatic summarization can be identified in terms of the level in Linguistic Space. Two broad approaches can be identified as *Shallow Approaches* and *Deeper Approaches* (Mani, Automatic Summarization, 2001). Shallow approaches use techniques which do not require the linguistic analysis beyond syntactic level. These approaches typically use to produce extracts, by extracting sentences from the original source. Some smoothing techniques use to repair any incoherence occurring in such extractions such as breaking of anaphoric references. The main advantage of these approaches is the robustness because it uses some straight forward methods to select summary sentences. However, there are some limitations in terms of the quality of the summary because it is hard to understand the real meaning of a sentence using these approaches.

Deeper approaches are used at least a sentential semantics level representation of sentences. Those approaches are able to produce abstracts, which involves natural language generation from a semantic or discourse level representation. Since the output texts of such approaches are generated by the machine, it requires rich linguistic resources such as sentence parsers, morphological parsers, WordNet, domain specific corpora among others. These approaches were initially originated for specific domains which have structured data as the input source such as the results and statistics of sport events, stock market bulletins and others. They produce more informative summaries since they are capable to identify more salient information of the input.

Even though summarization approaches can be divided into two of these broad categories, there are some hybrid approaches that have also been attempted in automatic summarization. These kinds of hybrid approaches are specially used in multi-document summarization, which merge different text elements drawn from multiple sources to produce abstracts (Mani, Automatic Summarization, 2001). However, the approaches used to automatically generate summaries from text can be classified in different ways based on the techniques and linguistic resources used by them. Such classification can be described as follows.

2.3.1 Classical Approaches

Initial attempts on automated summarization began with classical approaches. These approaches were founded in the 1950s and they are still serving as a fundamental basis for both practical applications and subsequent researches in the field. In 1958, Luhn started the field with his work on automatic creation of literature abstracts (Luhn, 1958). He used term frequencies to weight sentences. He started with filtering out closed-class words such as pronouns, prepositions and articles using a stop-word list and then normalizing terms by aggregating together the autographically similar terms. He has counted the frequencies of those aggregated terms and then has removed the low frequency terms. Sentences are then weighted using the resulting set of significant terms and a term density measure. Each sentence is divided into segments in the way that one segment is contained maximum of four non significant terms. Each segment is scored by taking the square of the number of bracketed significant terms divided by the total number of bracketed terms. The score of the highest scoring segment is taken as the sentence score.

Even though Luhn describes few possible extensions to his basic algorithm such as varying the length of the abstract and giving domain-specific word list as bonus words, Edmundson explained that the thematic features such as term frequency is less significant with compare to

other features such as title words or the term location (Edmundson, 1969). Edmundson extended the Luhn's work to look at these features in addition to the term frequency. He identified cue-phrases as well to score a sentence. He used manually created extracts to evaluate the performance of his algorithm and then he found that the keywords identified using the term frequency is the worst individual factor for extraction.

Classical approaches of automatic summarization basically use thematic features of the text. Identifying the bonus and stigma words for a specific domain is one of such feature which researchers have used to score sentences for making summaries. The work carried out by Pollock and Zamora at the CAS has relied on positive and negative cue phrases specific to the chemistry sub-domains (Pollock & Zamora, 1975). They have assigned a less weight to the positive terms which occur frequently in the text to avoid being lengthy the summary. This work has used elimination operations to compact the sentences and has used part-of-speech information to identify the clause boundaries. Authors conclude their system is functionally adequate even though the quality of the system abstracts are lower than the good manual abstracts.

These simple and limited resourced classical approaches gave a solid start for the field of automatic summarization. These kinds of fairly straight forward approaches are still being used for some commercial extracting systems. Even though the field of automatic summarization is growing aggressively by using some advance algorithms and high level linguisitic resources for more resourced languages, these classical approaches are still being applied for the less resourced languages such as Sinhala to avoid the issue of lacking adequate linguistic resources.

2.3.2 Corpus based Approaches

Even though the field of automatic summarization took a solid start with the classical approaches researchers faced a problem of identifying the relative contribution of each of these thematic features. The contribution of these features is dependent on the text genre because the writing styles and formats can be varied for each domain. For example, more salient information is laying on different locations for newspaper text and TV news broadcasts while the abstract and the conclusion of a scientific article will give more critical information other than the information available in other locations. Therefore, the location feature is behaved differently in different genres. Researchers have used corpus based approaches to determine the importance of different such features in different genres. Corpora can be used to count the occurrences of any feature to determine its importance. On the other hand,

researchers have used these corpus based approaches to learn rules or techniques for automated summaries by analyzing the corpus of human generated summaries along with their full-text sources. It is also useful for building empirically-based language models and researchers can share data-sets to try and compare different techniques.

A common use of a corpus is in computing weight based on term frequency. The *tf.idf* (term frequency verse inverse document frequency), which is widely used in information retrieval as well as text summarization is used to take out terms that distinguish one document from the other documents in a corpus. The importance of a word increases proportionally to the number of times that word appears in the document but is offset by the frequency of the word in the corpus. That feature can be captured using a text corpus of a given genre.

One of major challenge in these corpus based approaches is to creating and making available a suitable text corpus. Corpus should contain a representative sample of text from a specific domain, which needs to be summarized and then it needs to take permission from authors or abstractors to exploit and distribute the text for further experiments. It also needs to be considered the quality of collected summaries because the results of each experiment rely on them. Summaries can be created by both authors who write abstract along with the main article (especially for scientific articles) and professional abstractors who are trained to follow certain prescriptive guidelines. Special attention needs to be paid to select author made summaries since they are not systematic as one made by professional abstractors. Evaluation of these approaches against unseen test data is also another challenge of these corpus based approaches since it needs considerable amount of test data.

A research carried out by the Xerox Palo Alto Research Center has used a collection of 188 full text and summary pairs, obtain from 21 different scientific collections (Kupiec, Pedersen, & Chen, 1995). Each summary was written by a professional abstractor and it was three sentences long on average. They have used Bayesian classifier to calculate the probability of having the given sentence is in summary. They have used summary corpus to annotate full text sentences as positive or negative examples for a summary. Thematic features such as sentence length, sentence location, presence of cue-phrases and high frequency words and proper names have been used as the features of its feature vector. They have achieved 42% Recall on test documents. When the compression rate is increased (when the summary is lengthened), they reached 84% sentence Recall at 25% of the full-text length. This work confirms the finding of Edmundson (1969) using classical approaches that is the best individual feature is the location. Also they have concluded that the combination of location, cue phrases and the sentence length features was the best combination for making summaries. This work is referred as KPC-approach in successive research.

Myaeng and Jang have been experimented a variant of above method with technical texts in Korean (Myaeng & Jang, 1999). They have considered only the materials in Introduction and Conclusion sections and manually tagged these sentences to indicate whether they represent the background, a main theme, an explanation of the document structure or a description of future work. They have found that more than 96% of the summary sentences were main theme sentences. They have also used the Bayesian classifier to determine whether a sentence belongs to a main theme and then have combined evidence from multiple Bayesian feature classifiers to determine whether a sentence belong to a summary. They have also concluded that the combination of cue words, sentence location and the presence of title words in a sentence gives the best results.

The work of Aone, Okurowski, Gorlinsky and Larsen has proved that the use of different ways of aggregating terms is effective for summarization performance (Aone, Okurowski, Gorlinsky, & Larsen, 1999). Counting morphologically variant forms together with its root, adding synonym occurrences to the same concept and treating name aliases as occurrences of the same entity among others were such different ways of aggregating terms. They have shown that the performance of the summarizer can be improved when place names and organization names are identified as terms and when the person names are filtered out. The reason they gave for filtering person names is the documents in the corpus they used to train and test their system are generally not person-focused.

Hovy and Lin's work on Automated Text Summarization in SUMMARIST, they have used a 13,000 articles corpus which containing texts, abstracts and keywords to identify the location-relevant information and that gave them a ranked list of sentence positions that tend to contain the most topic-related keywords (Hovy & Lin, Automated Text Summarization in SUMMARIST, 1999). This introducing new method has got some encouraging results (Mani & Maybury, Advances in Automatic Text Summarization, 1999).

However, having a trainable summarizer is not a guaranteed resource for making useful summaries. Sentences extracted from the original text can be out of context and may have some anaphoric references which do not appear in the summary. Also it may have gaps between extracted sentences. Researchers were looking for new approaches such as Exploiting Discourse Structure and Knowledge-Rich Approaches to overcome such issues.

2.3.3 Exploiting Discourse Structure

A useful summary or an abstract is not just a collection of text which is available in the source, but some salient information of the source. It has to have some internal organization which reflects the need of having the abstract be coherent and to represent some of the argumentation used in the source. Human abstractors are used some techniques to keep these structures on their abstracts. Discourse structures are used to capture this feature for automated abstracting (Mani & Maybury, Advances in Automatic Text Summarization, 1999). Linguistic distinction between *cohesion* and *coherence* is used to classify the discourse models. Text cohesion involves the relations between words or referring expressions which determine the tightness of them when connected the text. The relations among linguistic components such as anaphora, ellipsis and conjunction and the lexical relations such as reiteration, synonymy and hyponymy are involved with the cohesion. Coherence represents the overall structure of a multi-sentence text in terms of macro level relations between sentences.

Barzilay and Elhadad have grouped the related terms together by text cohesion relationships and they called it “lexical chains” (Barzilay & Elhadad, 1999). They have used the WordNet (Miller, 1995) to identify the relations between terms and then suggested that the reader might get a better identification of the topic of a text by grouping together words into lexical chains than simply taking the most frequent words in the text. They have used the number and the weight of different relations in the chain to select the best chain among many possible chains being formed when using the WordNet. Even though the authors have pointed out some limitations of their system such as the inability of controlling the length of the summary and inability of selecting constituents smaller than a sentence among others, their evaluation of the system against an ideal human constructed summaries has proved that their system gives better performance than commercial summarizers available before 1997.

Teufel and Moens extended the KPC approach (KPC approach has been described under corpus based approaches) to address the discourse structure of the abstracts (Teufel & Moens, 1999). They have used a corpus of computational linguistics articles which have author supplied abstracts and they have labeled each sentence in the ideal summary with a list of pre-defined seven possible rhetorical roles namely, *Background*, *Topic/Aboutness*, *Related Work*, *Purpose/Problem*, *Solution/Method*, *Result* and *Conclusion/Claim*. They have divided the summarization task into two stages as extraction of sentences and the identification of rhetorical roles for each extracted sentences. They have used Bayesian classifiers modeled in KPC approach for both these stages and reported that the Indicator Quality feature led to the

best performance, which is 54.4% Precision. The combination of indicator quality, location, sentence length, title, presence of section header keywords and thematic words were the best mixer, which has given 66% Precision.

Approaches that used to capture the discourse structure associated with the source text and use them to generate more effective summaries have shown considerable improvements of the field of automatic summarization. Still the problem of identifying the most appropriate primitive linguistic unit for summarization process is existed. However, this problem may be solved in the future by improving such methods associated with discourse structure. Increasing the use of corpora with discourse annotations to provide additional data for discourse modeling and further development of theoretical models will be supported for such improvements of these models.

2.3.4 Knowledge-Rich Approaches

Knowledge-Rich approaches introduced to the field to focus on structured information rather than addressing on linguistic complexities and variability of input. The major attention of these approaches is for the transformation and synthesis phrases of the summarization machine (figure 2.1). Hahn and Reimer (1999) have considered the summarization as an operator-based transformation that takes output from a natural language analyzer and creates conceptually more abstract condensed knowledge structures (Hahn & Reimer, 1999). They have proposed a formal model that is embedded in a classification-based model of terminological reasoning for their system which they have implemented in information technology reviews and legal reports domains.

McKeown, Robin and Kukich's work on Generating Concise Natural Language Summaries; they have proposed techniques for opportunistically packing information into sentences using linguistic constructions (McKeown, Robin, & Kukich, 1995). The resultant sentences refine using elimination operations such as deleting repetitions and aggregation operations such as conjoining similar contents to pack maximum information in minimum space. Authors have developed two summary systems namely, STREAK and PLANLOC to generate summaries for basketball games and network planning activities respectively. STREAK system uses a revision approach, which is editing the draft for essential facts while the PLANLOC system uses discourse planning, that is looking ahead to perform repetition, deletion and conjunction operations. Finally they have emphasized the practicality of summary generation and the advances of using information extraction methods to generate summaries.

These knowledge-rich approaches have used hybrid application of statistical analysis and domain specific techniques and the condensation operation have been used in all the phases in summarization. However, unexplored issues such as the need for more comprehensive corpus analysis of conceptual and linguistic summarization operations, the formalization of the range of condensation operations, the evaluation approaches and the moving towards more general purpose, domain independent approaches and others have not addressed with these approaches. A key challenge for the field of automatic summarization will be the effective integration and reuse of previous and future results in Information extraction, text planning and linguistic realization to maximize the progress (Mani & Maybury, Advances in Automatic Text Summarization, 1999).

2.4 Summarization for Indic Languages

Most of the approaches explained above have been experimented mainly with more resourced languages such as English and the techniques and theories have been developed based on the linguistic behavior of such languages. Highly inflected languages such as Sinhala and Tamil have different linguistic structure than English and hence the techniques explored in above researches might not fit well with such languages. Few attempts for applying such techniques have been recorded in the literature for Tamil language, but none of previous attempts recorded for Sinhala Language.

The work carried out by Banu, Karthika, Sudarmani, and Geetha for making summaries in Tamil has used a natural language sentences representation system called Language-Neutral Syntax (LNS) by considering the semantics of documents (Banu, Karthika, Sudarmani, & Geetha, 2007). They have applied a syntactic analysis of the text for each sentence and Subject-Object-Predicate (SOP) triples have been extracted from individual sentences to create a semantic graph of the original document and the corresponding human extracted summary. Then they have used semantic normalization to SOP triples for reducing the number of nodes in the semantic graph of the original document. A classifier has been trained using Support Vector Machine (SVM) learning algorithm to identify SOP triples from the document semantic graph that belong to the summary. Finally the classifier has been used to automatically extract summaries from test documents.

Jeganathan (2005) has tried to summarize Tamil text using sentence extraction approach based on GistSumm (GIST SUMMarizer) (Pardo, Rino, & Nunes, 2003) for his final year research of the degree in computer science at University of Colombo (Jeganathan, 2005). The main assumption of the GistSumm is, when a person summarizes a text, he first tries to

identify the gist and then, adds information drawn from the text to complement it. Pardo, Rino and Nunes have determined the gist sentence either through keywords or through text mining method. The gist sentence corresponds to the most significant distribution of keywords and that will be the most representative of the source text. Once the gist sentence is determined, they use Cosine Similarity Measure to find the most closed other sentences to the gist sentence by defining the feature vector once removing stop words and stemming the other words. They have achieved average 50% of quality, by measuring the results using human evaluators. Finally they have concluded that getting N-gram counts rather than calculating *tf.idf* values to select keywords gives better performance. The similar work carried out for Tamil text shows this result in the other way, i.e., *tf.idf* based keywords identify the gist sentence correctly most of the time. It has scored 47% Precision and 28% Recall on average at 30% compression rate, even though that work has suffered with data. However, Jeganathan has compared his work with the Tamil text summarizer in Microsoft Word and has reported that his system gives 2.16 average score for Precision and Recall while MS Word gives only 1.16.

Generally, the languages of Indian subcontinent have received little attention in the field of automatic text summarization, primarily because the amount of digital information available in those languages is less. However, with the rapid development of the technology with the Unicode standard, the scenario is changing now for the positive direction and few researches have been carried out for such languages. Alkesh Patel, Tanveer Siddiqui and US Tiwary's work on language independent approach to multilingual summarization has described a language independent algorithm to generate extractive summaries from a single document (Patel, Siddiqui, & Tiwary, 2007). They have used structural and statistical factors rather than semantics of the languages. They have tested their algorithm for English, Hindi, Gujarati and Urdu and claimed that their algorithm performed equally well regardless of the language. They have used two feature vectors called *Title Feature Vector* and *Theme Feature Vector* for each language to calculate sentence weights. Stop words removal and Stemming have been done for each language by using predefined language dependent techniques and then the sentences have been weighted based on two feature vectors and the sentence location feature. They have used a partitioning scheme and their own hypotheses to assign weights for the sentence location feature. The quality of the summaries is tested with respect to its degree of representativeness for the languages and the authors claimed that the results are encouraging as a robust language independent algorithm.

Vishal Gupta and Gurpreet Singh Lehal have attempted to prepare Panjabi language text for automatic summarizing on their work on Preprocessing Phase of Punjabi Language Text Summarization (Gupta & Lehal, 2011). They have only considered the preprocessing stage of summarization by identifying various sub phases of preprocessing stage, such as Punjabi words boundary identification, Punjabi language stop words elimination, Punjabi language noun stemming, finding Common English Punjabi noun words, finding Punjabi language proper nouns, Punjabi sentence boundary identification and identification of Punjabi language Cue phrase in a sentence. In depth analysis of Panjabi corpus, Punjabi dictionary and Punjabi morphs have been carried out by authors to define required linguistic resources for these sub phases. Authors claimed that the successive researchers can use their preprocessing techniques to summarize Panjabi text.

The Centre for Development of Advanced Computing, Noida (CDAC-Noida) has developed a text summarization system for Hindi and has published through their website for common use (CDAC-Noida, 2006). They have developed a comprehensive tool to summarize Hindi and Panjabi texts and the user can select the language and the compression rate of the summary. Developers have not published their approach on developing the system, but it seems that they have used some sentence ranking algorithm which uses key terms to weight sentences, because the user can specify the keywords for the summary.

Even though researchers are trying to generate good human consumable summaries using computers over the last six decades, the issue of evaluation is continued as an outstanding problem (Mani, Automatic Summarization, 2001). It is hard to determine a better summary even based on the intuition but sometimes it is easier to tell if something is a poor summary. This problem caused an issue of defining a gold standard of a unique reference summary against which system summaries can be compared. However, researchers are using some pre-defined evaluation techniques to compare their approaches with the previous attempts. Since this work is the first attempt for summarizing Sinhala text, the methodology used is based on the work carried out by Edmundson (1969) with some improvements and evaluated using standard statistical methods.

Next chapter explains the methodology used in detail while the following chapters explain the experiments, results and the evaluation.

Chapter 03 – Methodology

This chapter explains the methodology adopted to carry out the research on automatically summarizing Sinhala text. It is intended to generate the summary as extract rather than abstract due to certain limitations of prior infrastructure resources. Abstracts need some kinds of natural language generation techniques which require rich linguistic resources while extracts can be generated using classical approaches. The methodology used in this research was based on the groundbreaking work carried out by Edmundson (1969), which is used as the fundamental work for the most of other research on extraction. The adaptation of the Edmundson's work to Sinhala is explained by this chapter along with the steps carried out to prepare the data-set.

3.1 The Edmundonian Paradigm

The research work carried out by Edmundson in 1969 to propose new methods in automatic extracting is considered as the foundation for work on extraction. His work is still continued to influence extraction work today. Subsequent research has expanded the set of features he used and has developed more sophisticated methods for weighting different features based on information from a corpus (Mani, Automatic Summarization, 2001).

Edmundson (1969) has used four thematic features to assign weights on sentences in the source document. Three of them are word level features chosen after excluding all stop words while the other feature derived based on the structure of the source article. Four features he used can be described as follows.

1. **Cue Words:** Cue words are connective expressions that link spans of discourse and signals semantic relations in a text. Two types of cue words can be identified with respect to creating summaries as *bonus words* and *stigma words*. Bonus words are above an upper corpus frequency threshold, which are then used as evidence of selection. Edmundson has identified that the bonus words are consisted of comparatives, superlatives, adverbs of conclusion value terms, relative interrogatives and causality terms. Stigma words in the other hands are the words below a lower frequency cutoff which are then used as evidence of non-selection. The words like "impossible", "as an example", "hardly" and others can be stigma words in some domains. Edmundson has showed that the stigma terms can be consisted of "anaphoric expressions, belittling expressions, insignificant detail expressions and hedging

expressions. He has extracted this cue words from the training corpus based on corpus frequency.

2. **Title Words:** The words appear in the title, subtitles and headings are considered as title words. Edmundson has assigned a hand assigned weight for each title word to get the best performance. The main assumption of it is that authors will tend to use informative titles. Sentences which contain title words will be scored based on a defined equation.

3. **Key Words:** Most frequent words that appear in the document are considered as keywords. The list of keywords can be identified by sorting the list of words in the document by their frequency. Words that appear above a defined cutoff are considered keywords and their document frequencies will be the word's weights. If identified cue words appear in this list of keywords, they will be ignored since they are weighted separately as cue words.

4. **Sentence Location:** Sentence location is the feature to assign a weight for a sentence based on its position in the document. Edmundson has used two methods to assign weight for the sentence location. One is, he has manually short listed particular section headings such as "Introduction" and "Conclusions" and then assigned a positive weight for the location feature for sentences which occurred under such headings. The second method is to assign weights based on their ordinal position in the text. If the sentence occurred in the first and last paragraphs, or if they were the first or last sentences in a paragraph, they were assigned a positive weight.

The overall method of scoring sentences for extraction was based on a linear combination of these four features, as shown in the equation 01.

$$W(s) = \alpha C(s) + \beta K(s) + \gamma T(s) + \delta L(s) \quad - - - - (01)$$

Where,

W(s) is the overall weight of the sentence *s,*

C(s) is the score given to sentence *s* based on the presence of Cue Words,

K(s) is the score given to sentence *s* based on the presence of Key Words,

T(*s*) is the score given to sentence *s* based on the presence of Title Words,
L(*s*) is the score given to sentence *s* based on the Sentence Location and
α, β, γ and δ are constant values.

Edmundson has used a corpus of 200 scientific papers on chemistry in which the length of each paper is between 100 and 3,900 words. He has divided his set of articles into the training set and the test set and in the training phase, he has adjusted feature weights manually and has used the feedback from comparisons against manually created training extracts to tune α, β, γ and δ parameters. Final system was tested and evaluated on the test data. He has found in the evaluation that keywords were poorer than other three features and that the combination of cue words, title words and sentence location was the best combination while the sentence location being the best individual feature.

3.2 Adopting Edmundonian Paradigm for Sinhala

As explained in Chapter 02, much research has been carried out over the last five decades based on Edmundonian paradigm for automated summarization as classical approaches. Different varieties of this paradigm have been tested and some obvious improvements such as using *tf-idf* values rather than using most frequent words and others have been found. This research attempted to find the most applicable such thematic features for Sinhala language to automatically summarize the Sinhala text.

Sinhala is a highly inflectional language, which uses suffixes with roots to form words. Unlike English, Sinhala nouns can be inflected for 130 word forms while verbs can be inflected around 240 forms (Weerasinghe, Herath, & Welgama, 2009). It can be clearly identified that there are two major varieties of Sinhala as *Spoken Sinhala* and *Written Sinhala* and written Sinhala is more structured than spoken Sinhala. However, written Sinhala also follows many different formats especially for word separation and therefore identifying thematic features for Sinhala text is highly domain specific.

This research is carried out for the domain of editorials of three daily national newspapers (Details of the data-set will be explained later in this chapter). The main assumption behind the selection of this domain is that the editors of national newspapers are professionals who write refined Sinhala. Also, it was assumed that the editorials have a unique structure, which is more suitable to be summarized. The research was conducted based on four thematic features of these selected editorials. Three of them are overlapped with the Edmundson's features, while the other feature is newly introduced for this research.

Cue Words, which Edmundson (1969) has used as one of his four features did not consider as a feature for these editorials due to several reasons. Cue words are highly specific for a particular subject or a domain. Bonus terms for a particular subject will be stigma terms in different domains. Even though the selected corpus can be considered as belonging to a specific domain called "news editorials", a theme of an editorial can be anything. Usually editorials are used to present the publisher's opinion for a temporal incident of the society and that can be about politics, religion, sociology, humanities, science or any other subject. Therefore, identifying subject specific cue terms of such articles will not be useful to summarize them automatically.

The other three features which Edmundson (1969) used (*Key words*, *Title Words* and *Sentence Location*) are considered along with the new additional feature called *Paragraph Location*. Different approaches are used to define each of these features and the best approach to assign weights for each individual feature is selected by evaluating each feature separately against the manually created summary. The approaches taken for selecting each individual feature can be described in detail as follows.

3.2.1 Identifying Keywords

Keywords of an article are primarily identified based on the term frequency. The main assumption of this paradigm is called "Thematic Term Assumption", that is *relatively more frequent terms are more salient* (Mani, Automatic Summarization, 2001). Original motivation of using this thematic term features was the pioneering work of Luhn (1958) who suggested finding keywords in a document by filtering against a stop list of function words such as prepossessions, determiners, conjunction and others. However, both Luhn's and Edmundson's works use most frequent words of the article irrespective of their document frequency.

Spärck-Jones (1972) introduced a variant of the Thematic Term Assumption which is, the term importance is proportional to its frequency in the document, but it is inversely proportional to the total number of documents which that term occurred (Spärck-Jones, 1972). More research has been carried out after Spärck-Jones's work by using this assumption and researchers calculated *tf.idf* weight rather than considering word frequency alone. This was successfully being used for stop word removal as well since its definition proves it.

tf.idf is the multiplication of term frequency with its inverse document frequency. Although more variations exist to calculate the *tf.idf* values, only a simple form of that is used in the research is described below.

If the term frequency of the term *t* in the document *d* is *tf(t, d)*,
tf(t, d) can be simply defined as the number of occurrences of term *t* in the document *d*.

If the inverse document frequency of the term *t* is *idf(t)*, it can be defined as in equation 02.

$$idf(t) = \log \frac{|D|}{|\{d: t \in d\}|} \quad ----(02)$$

Where,
|*D*| is the total number of documents in the corpus and,
$|\{d: t \in d\}|$ is the number of documents where the term *t* appears.

Then *tf.idf* can be defined as,

$$tf.idf(t,d) = tf(t,d) \times idf(t) \quad ---(03)$$

If a term occurs more frequently in a document and if it appears in most of the documents in the corpus, it is considered as a less important word. Terms which are specific only for a particular document will be positively weighted by the equation 03.

While counting the term frequencies for each document in the corpus, function words defined in Weerasinghe, Herath and Welgama (2009) were skipped. They have identified 440 such words as Sinhala stop words (*nipatha pada*) among conjunctions, determiners, interjections, particles and post positions.

Stemming of Words

Stemming is an essential process in the field of NLP. It is the process of reducing the inflected form of a word to its stem. Many NLP applications which use words as basic elements employ stemmers to extract the stems of words. This is a very efficient and lightweight approach compared to morphological parsing. Even though there are some advanced stemmers for languages such as English, the algorithms which they employ do not work well for highly inflected languages such as Sinhala.

As Sinhala is a highly inflectional language, there are many word forms to denote a single concept. This situation is highly effected for the frequency of a term and therefore words have to be stemmed before getting their frequencies. They are no any previous attempts recorded in the literature for defining proper stemming algorithm for Sinhala. Therefore, two stemming

algorithms as explained below were defined to identify the stem of each word and then a single approach was selected based on the experimental results. The impact on stemming on the task like text summarization is evaluated while evaluating the summarization accuracy and the results have been explained in Chapter 04.

Knowledge-Based Approach: The *Gold Standard* for Sinhala stemming is defined by extracting 33,684 stem values and their 1,325,273 corresponding word forms from the lexicon described at Weerasinghe, Herath and Welgama (2009). This list of (Word, Stem) pairs were used as a lookup for identifying the stem of a given word. If the given word does not exist in the gold standard, the word itself is considered as its stem.

Data-Driven Approach: A simple lightweight algorithm is used to identify the stems of a list of words. Initially the list of words extracted from the *Sinhala News Editorials Corpus* (the composition of the corpus is described in section 3.3) is sorted alphabetically to get the similar word forms together and then the algorithm described in Figure 3.1 uses to define the stem of each word.

The *suffixes list* is a list of all possible Sinhala suffixes identified from the work carried out by Weerasinghe, Herath and Welgama (2009). Table 3.1 shows a part of the word, stem pairs list generated using the proposed algorithm.

Table 3.1: Sample of results of the lightweight stemming algorithm

Word	**Stem**	**Word**	**Stem**	**Word**	**Stem**
අ.පො.ස	අ.පො.ස	අංකල්	අංකල්	අංගනයක	අංගනය
අංක	අංක	අංකවලට	අංක	අංගප්‍රත්‍යංග	අංගප්‍රත්‍යංග
අංකද	අංක	අංකුර	අංකුර	අංගපුලාවක්	අංගපුලාවක්
අංකය	අංක	අංකුරවල	අංකුර	අංගම්පොර	අංගම්පොර
අංකයක්	අංක	අංග	අංග	අංගය	අංගය
අංකයද	අංක	අංගන	අංග	අංගයක්	අංගය
අංකයෙන්	අංක	අංගනය	අංගනය	අංගයකි	අංගය

The stem values generated from the Data-Driven approach are evaluated against the Gold Standard defined from the Knowledge-Based approach. The efficiency of the Data-Driven approach is defined as the ratio between the number of correctly identified stems and the total number of words in the list. The evaluation revealed that the simple lightweight algorithm defined in Figure 3.1 is capable to define 26,272 linguistically correct stems out of 46,874 words. So the efficiency of the Data-Driven approach for identifying linguistically correct stems is 56.04%.

However, according to the definition of stemming, a stemming algorithm does not need to identify the linguistically correct stem, but it is sufficient to map all the forms of a word to a single form. Therefore, the efficiency of the Data-Driven approach may be different in the context of text summarization. The impact of stemming for the Sinhala text summarization is evaluated separately and the results revealed that the stemming of words before calculating their frequencies is caused to increase the performance of the summarizer. More details on evaluation are described in Chapter 04.

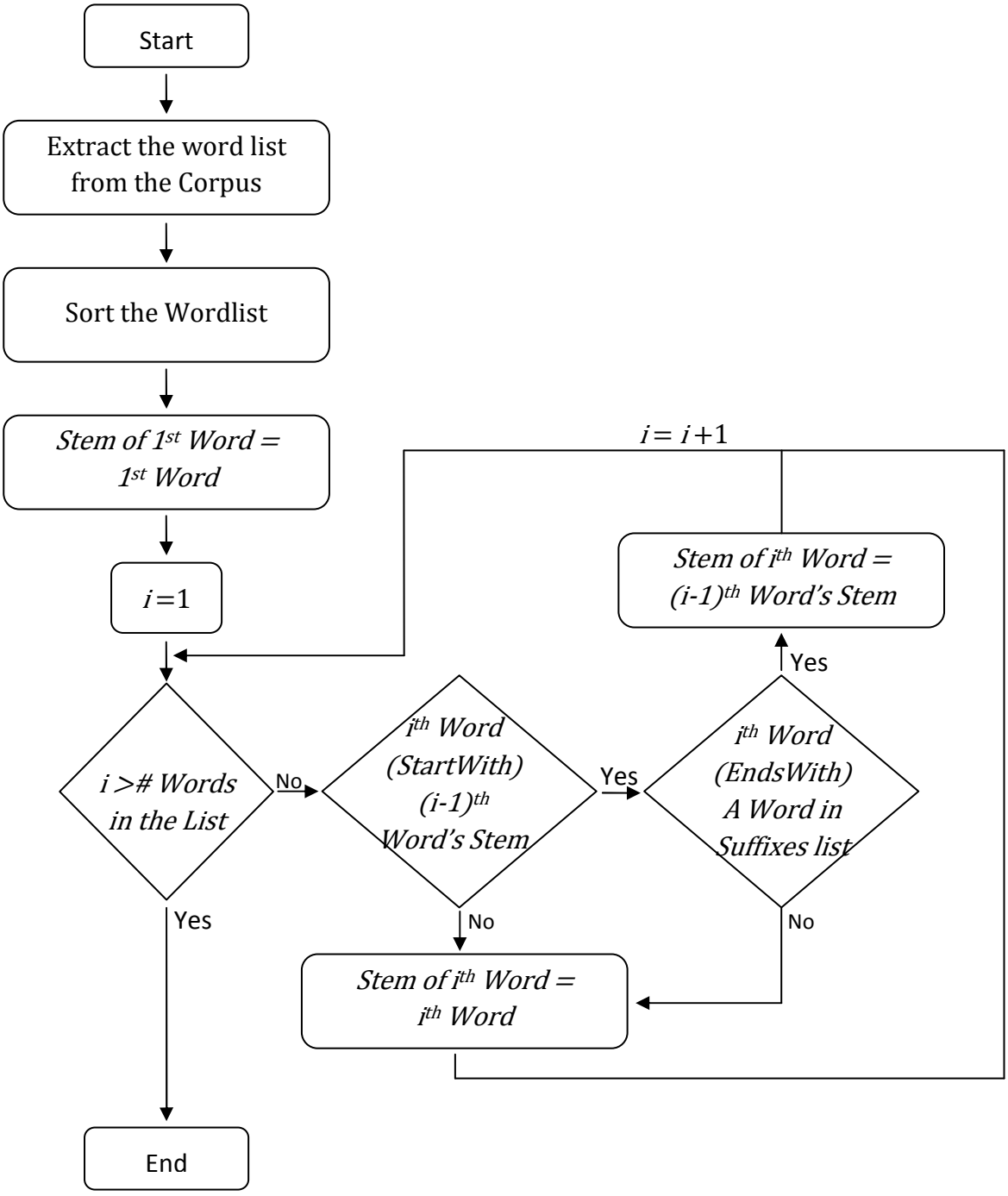

Figure 3.1: The lightweight algorithm defined for stemming Sinhala words

Two sets of *tf.idf* values were calculated based on above two approaches. *tf.idf* value itself was considered as the term weight and $W(s)$, the total weight to be assigned to the sentence *s* based on keywords is calculated as in equation 04.

$$w(s) = \frac{\sum_{i=1}^{n} tf.idf(t_{i,s}, d)}{n} \quad - - - - (04)$$

Where,

$tf.idf(t_{i,s}, d)$ is the *tf.idf* value for the i^{th} term of the sentence *s* in the document *d* and,

n is the number of words in the sentence *s* (the sentence word length).

Two approaches were evaluated separately by comparing the *F-Score*, which are obtained by comparing 100 human summarized articles against 100 computer generated summaries. Only the keyword feature is used as a feature to extract this computer generated summaries. The method of evaluation and the results will be discussed in detail in Chapter 04.

3.2.2 Title Words

Usually all the editorials have an appropriate title given by the author. It is assumed that authors always use contents related to the title for filling the article. Therefore, the title can be considered as the gist of the article, especially in news editorial domain. However, exceptions for this primary assumption were found in some articles of the corpus and such articles were avoided as much as possible from the corpus. That seems to happen when authors use more discourse level titles for their articles.

Edmundson (1969) has defined title words as a feature and that is used to assign a weight to the sentence *s* based on the terms in it that are also present in the title. Edmundson has used the title, subtitles and headings to identify the title words and has manually assigned weight as it leads to the best performance.

The selected editorials do not have any subtitles and headings, but each article consists of an appropriate main title. Words in the title were stemmed to capture its inflected forms as well in the sentence *s* and weight for title words was given as shown in the equation 05.

$$W(s) = \frac{No\ of\ title\ words\ in\ the\ sentence\ s}{Total\ number\ of\ words\ in\ the\ sentence\ s} \quad - - - - (05)$$

Where,

$W(s)$ is the weight assigned for the sentence *s* based on title words.

Equation 05, which is defined to assign weight for the sentence *s* due to title words, always gives a value between zero and one. According to this definition, if the sentence *s* contains more title words, it will be positively weighted while if it does not contain any title words, the weight due to title words is zero.

3.2.3 Sentence Location

Assigning a weight for the location of a sentence is proposed by Baxendale (1958) in his experiment on man-made indexing (Baxendale, 1958). He found that the important sentences were located at the beginning or the end of paragraphs. He claimed that salient sentences were likely to occur as either the first sentence in paragraph 85% of the time or the last sentence in the paragraph 7% of the time. Edmundson has assigned weight for the location based on article's sections such as Introduction and Conclusion and then based on its ordinal position of the text. He has followed Baxendale's (1958) findings and has assigned positive weights for sentences in first and last paragraphs and first and last sentences in a paragraph.

Articles in the selected editorial corpus do not contain sub sections and therefore Edmundson's first approach is not applicable for them. However, since no any previous attempts recorded to identify how Sinhala sentences laid in a document, three different equations to assign weights for the location feature was experimented. Sentence location is defined locally within a paragraph and its global position of the text was not considered since the paragraph location would be considered as a separate feature.

Assign Weight using a Linear Function: First approach to assign weights for the sentence location was done by using a linear function. Main assumption of defining a linear function for assigning weights is that the most salient sentence of a paragraph is located at the beginning and then the importance of the other sentences will decrease gradually through in the paragraph. The equation 06 denotes the linear function, which defines to weight the sentence *s* based on its location.

$$W(s) = \frac{(n - i) + 1}{n} \quad - - - -(06)$$

Where,

$W(s)$ is the weight assigned for the sentence *s* based on its location,

i is the sentence location within the paragraph and

n is the total number of sentences in the paragraph.

According to the equation 06, first sentence of a paragraph will be scored as one for its weight for the location and all the other sentences will be scored between one and zero. The figure 3.2 will illustrate the defined function from the equation 06 for a paragraph which has 10 sentences.

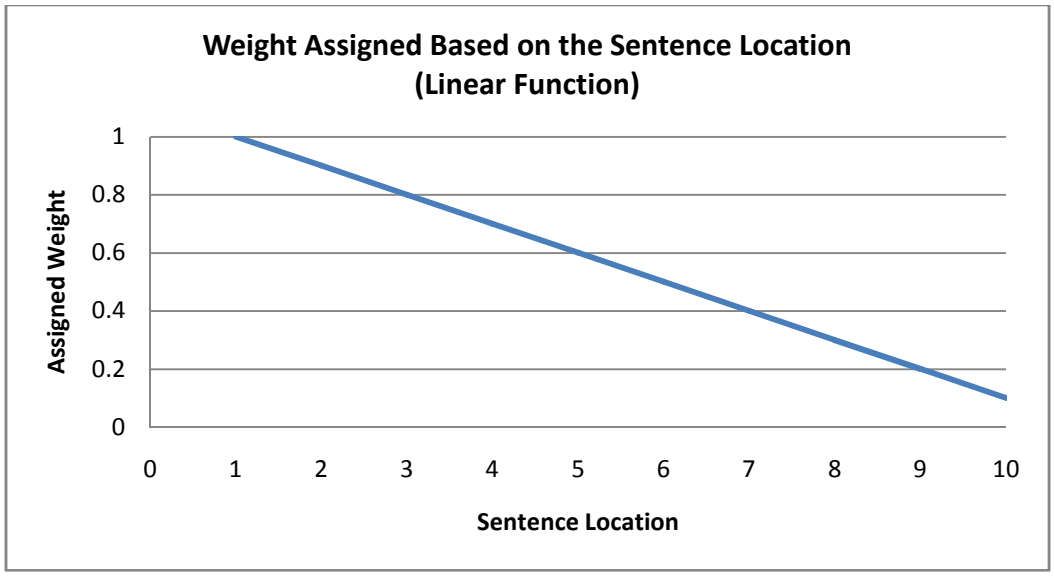

Figure 3.2: Linear function defined to assign weight for the sentence location

Assign Weight using a Hyperbolic Function: Second approach to assign weight for the sentence location is using a hyperbolic function. The main assumption of defining a hyperbolic function is same as defining a linear function, but it is assumed that the importance of the sentences decreases significantly through the paragraph. Equation 07 denotes the hyperbolic function defined to assign weight for the sentence *s* in a given paragraph.

$$W(s) = \frac{1}{i} \quad - - - - (07)$$

Where,

$W(s)$ is the weight assigned for the sentence *s* based on its location and,

i is the sentence location within the paragraph

As defined in equation 07, first sentence of a paragraph will be weighted as one and then other sentences will be weighted from a value between one and zero. Figure 3.3 graphically illustrates the weight decreased along with the sentence location.

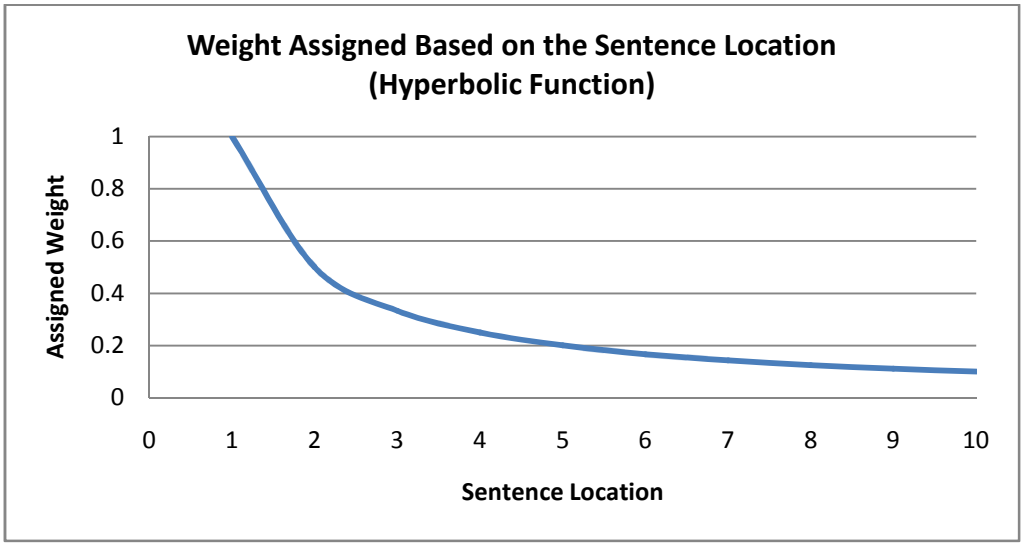

Figure 3.3: Hyperbolic function defined to assign weight for the sentence location

Assign Weight using a Quadratic Function: Third approach used in this research to assign weight for the sentence location is using a quadratic function. Main assumption behind this definition is that the most salient sentences of a paragraph have laid on the beginning and the end of the paragraph and less important sentences are in the middle of the paragraph.

Equation 08 shows the definition of the quadratic function which set to assign a weight for the sentence *s*.

$$W(s) = \begin{cases} 1 - \left\{\frac{2}{n-1} \times (i-1)\right\}, & i < \frac{n+1}{2} \\ 1 - \left\{\frac{2}{n-1} \times (n-i)\right\}, & i > \frac{n+1}{2} \\ 0.1, & i = \frac{n+1}{2} \end{cases} \quad - - - - \ (08)$$

Where,

$W(s)$ is the weight assigned for the sentence *s* based on its location,

i is the sentence location within the paragraph and

n is the total number of sentences in the paragraph

Figure 3.4 illustrates the graphical view of the quadratic equation. According to the equation 08, the sentence appeared in the first and the last location of a paragraph will be scored as one while the middle sentence will get the lowest value, 0.1. Sentences in other positions will be scored from a value between 1 and 0.1.

Finally, summaries were generated only using the location feature based on three of these equations and evaluated against the manually extract summaries by calculating *F-Score*

measure. Best approach to weight Sinhala sentences based on its location was identified by analyzing these F-Score values and that approach is selected for assigning weights for the location feature. The experiments carried out with these three approaches and their results will be discussed in detail in Chapter 04.

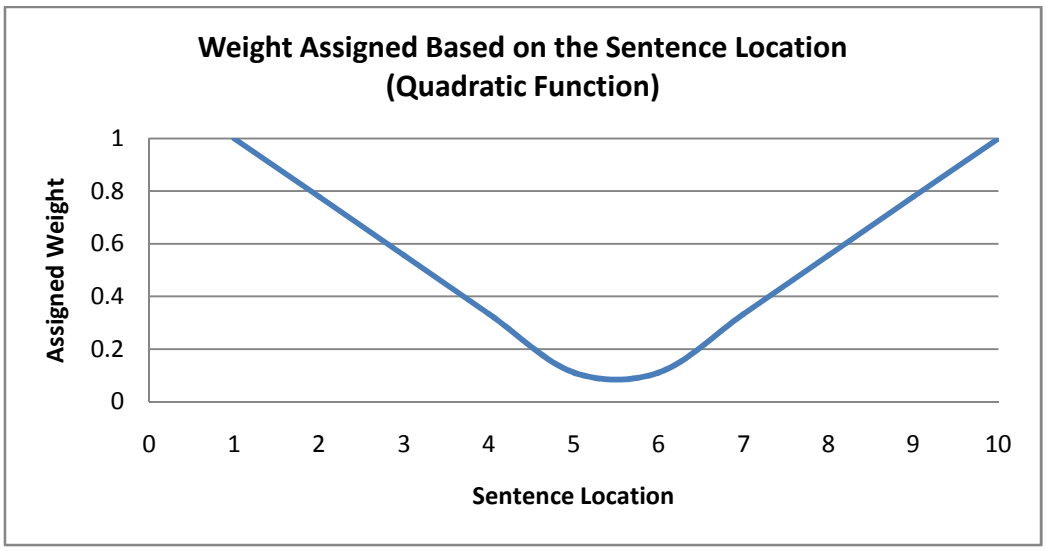

Figure 3.4: Quadratic function defined to assign weight for the sentence location

3.2.4 Paragraph Location

Paragraph location is introduced as a new feature to weight sentences which are not used directly in Edmundson (1969) approach. However, Edmundson has considered the subsections of the articles when assigning weights for the sentence location. The main purpose of adding this feature to weight sentences is to find whether the Sinhala newspaper editors follow some underlying rules to structure their editorials.

Three individual experiments were carried out to identify the most suitable function which is able to describe the flow of information among Sinhala paragraphs most appropriately. Three functions defined to weight sentences based on its location (equation 06, 07 and 08) were used to weight sentences based on its paragraph location in the article. All the sentences in a given paragraph will be scored by a unique weight which are calculated based on the paragraph location. Same evaluation techniques were used to identify the most suitable approach to assign a weight for a sentence based on its paragraph location.

After identifying the most suitable approach for each of these individual features, the final score for a given sentence will be calculated using a linear combination of these four features. Edmundson (1969) has also used a linear combination of selected four features to assign the final weight for the sentence *s*. The linear function can be defined as in equation 09.

$$W(s) = \alpha K(s) + \beta T(s) + \gamma L(s) + \delta P(s) \quad - - - - (09)$$

Where,

W(s) is the overall weight of the sentence s

K(s) is the score given to sentence s based on the presence of Keywords,

T(s) is the score given to sentence s based on the presence of Title Words,

L(s) is the score given to sentence s based on the Sentence Location,

P(s) is the score given to sentence s based on the Paragraph Location and,

α, β, γ and δ are constant values.

W(s) will be calculated for all sentences of the article and sorted according to the descending order of the weights. Finally, the n number of sentences will be extracted as the summary of the given article where n is defined as in equation 10.

$$n = \frac{N \times C}{100} \quad - - - - (10)$$

Where,

N is the total number of sentences in the article and,

C is the required compression rate for the summary

3.3 Data-set

Sinhala has its own writing system and the literary tradition of Sinhala goes as far back as two thousand years. Even though the Sinhala writing system appeared in computers in the early 1980's with proprietary encoded fonts, it boosted up developing the language processing applications for Sinhala with the introduction of the Unicode standard in 2004. Bloggers, volunteers for wikis and then newspapers started to appear on the Internet in Sinhala scripts and hence the web content in Sinhala gradually grew up. Most of daily national newspapers appeared on the Internet by opening opportunities for researchers to carry out local language research with their vastly increasing data.

To carry out research on automatically summarizing Sinhala text, it was decided to look for well structured Sinhala articles from a specific domain. After observing many regular articles on online daily newspapers, editorials of three national newspapers; namely *Dinamina*, *Divaina* and *Lankadeepa* were selected due to following reasons.

- Editorials are written by the chief editor or one of senior journalist of the newspaper who is professional to use refined Sinhala. Therefore, it is reasonable to assume that the most common language errors such as misspelled words, word separation errors and others are rare to appear in these editorials. This will help to get more accurate word frequencies, which significantly affect assigning weights based on keywords.
- Each editorial contains an author assigned title, which is most appropriate to its contents. This allows the title word feature to be more meaningful.
- All the editorials have approximately equal number of sentences which is more suitable for a research like automatic text summarizing. The average number of sentences per selected editorial is 52 and that is neither too short nor too long for the defined research. Figure 3.5 shows the frequency distribution of the number of sentences per article, which approximately follows the poison distribution.
- All the editorials have simple unique structure, which do not contain different formatting styles such as tables, graphs, images and others. It contains only text which is written fluently by separating paragraphs. This simple structure helps to calculate weights based on the sentence location and the paragraph location accurately.

1,400 editorials were collected from three daily newspapers and the *Sinhala News Editorials Corpus* was created for carrying out the research on automatic text summarization in Sinhala. It was assumed that nearly one million words text corpus is sufficient to represent the language for a research like automatic text summarization. The collected articles are stored in computer in txt file format and table 3.2 shows the basic statistics of the defined corpus.

Table 3.2: Basic statistics of the Sinhala Editorials Corpus

Feature	Amount
Number of articles	1,400
Total number of words	952,948
Total number of distinct words	52,334
Total number of sentences	72,143
Average number of sentences per article	52
Total number of paragraphs	17,583
Average number of sentences per paragraph	4
Average number of paragraphs per article	13

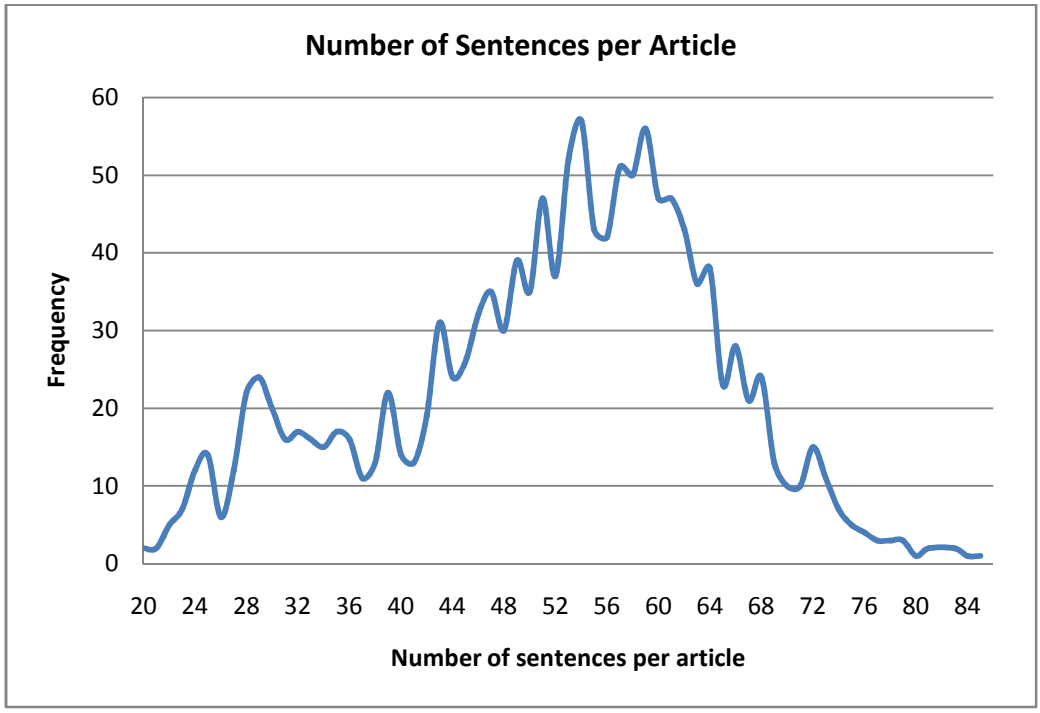

Figure 3.5: Distribution of number of sentences per article

3.3.1 Cleaning the Articles

Sinhala has three special characters called *Yansaya, Rakaranshaya* and *Repaya* which do not appear in the Sinhala Unicode Table (defined at http://unicode.org/charts/PDF/U0D80.pdf). These are used to represent some combinations of consonant and vowel modifiers, especially to shorten the length of the word. Yansaya and Rakaranshaya are widely used even in modern Sinhala while Repaya is less used in current writing system.

To represent these three characters in the Unicode representation, a special character called *Zero-Width Joiner* (ZWJ) is used. Sinhala keyboard drivers allow typing these combined characters by just pressing a single key and then all required key values including this ZWJ character will be passed to the memory. Most of Sinhala typists are not aware with this scenario since the ZWJ character is not visible in the text, but they can see the composite character which they intend to type. However, when deleting such a composite character, the ZWJ character will remain in the text if the typist does not delete them properly. This issue has been addressed in newer version of keyboard drivers, but older keyboard drivers allow user to make this a significant issue. This has become an issue especially due to typist's unawareness about this scenario. Figure 3.6 shows a paragraph of an article which contains ZWJ character (marked in circles) in unwanted places.

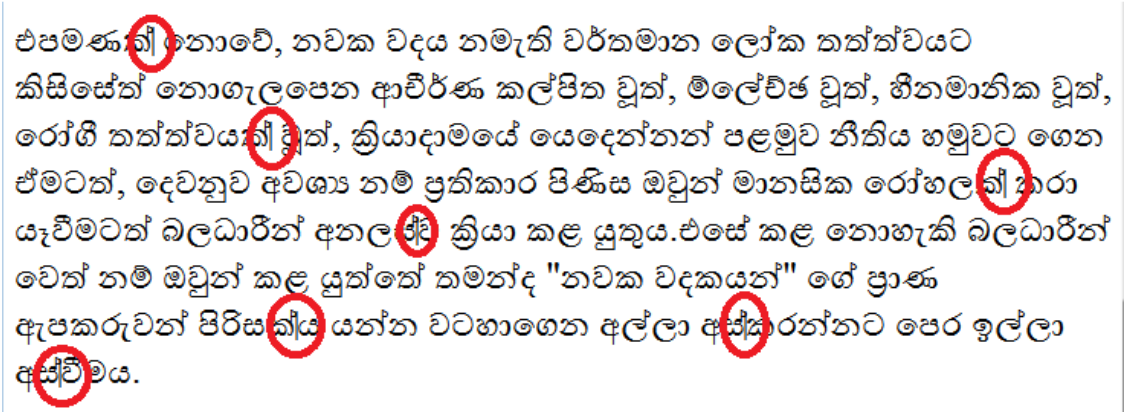

Figure 3.6: ZWJ character appeared in unwanted places in a paragraph

Most of the articles in the collected corpus effected with the above issue and it would cause for miscalculation of the word frequencies of the corpus. The word with an unwanted ZWJ character will be treated as a separate word when counting the word frequencies because its internal representation differs from the accepted form. Therefore, all the articles in the corpus had to be cleaned by removing unwanted ZWJ characters from them.

3.3.2 Splitting Sentences

Editorials contain author defined paragraphs which may contain two or more sentences per paragraph. However, a paragraph is defined for this research as “the text between two new lines” and therefore a single sentence also had to be considered as a paragraph, if it appears alone in between two lines. Sentences within a paragraph were detected automatically using an algorithm developed based on heuristics. Figure 3.7 shows the algorithm defined for the sentence boundary detection.

The exception list is created by adding misleading words which can be followed by a dot, but not as the full stop. The most common such words are the letters in English alphabet which are written in Sinhala scripts. Those are used to write the initials of personal names with followed by a dot. Abbreviations are another common such words, which are written by Sinhala letters followed by a dot. More words were added to the exception list while running and testing the above algorithm in trial and error basis. The minimum character length for a sentence is defined as five, based on the experiments and it was set as the *defined threshold*.

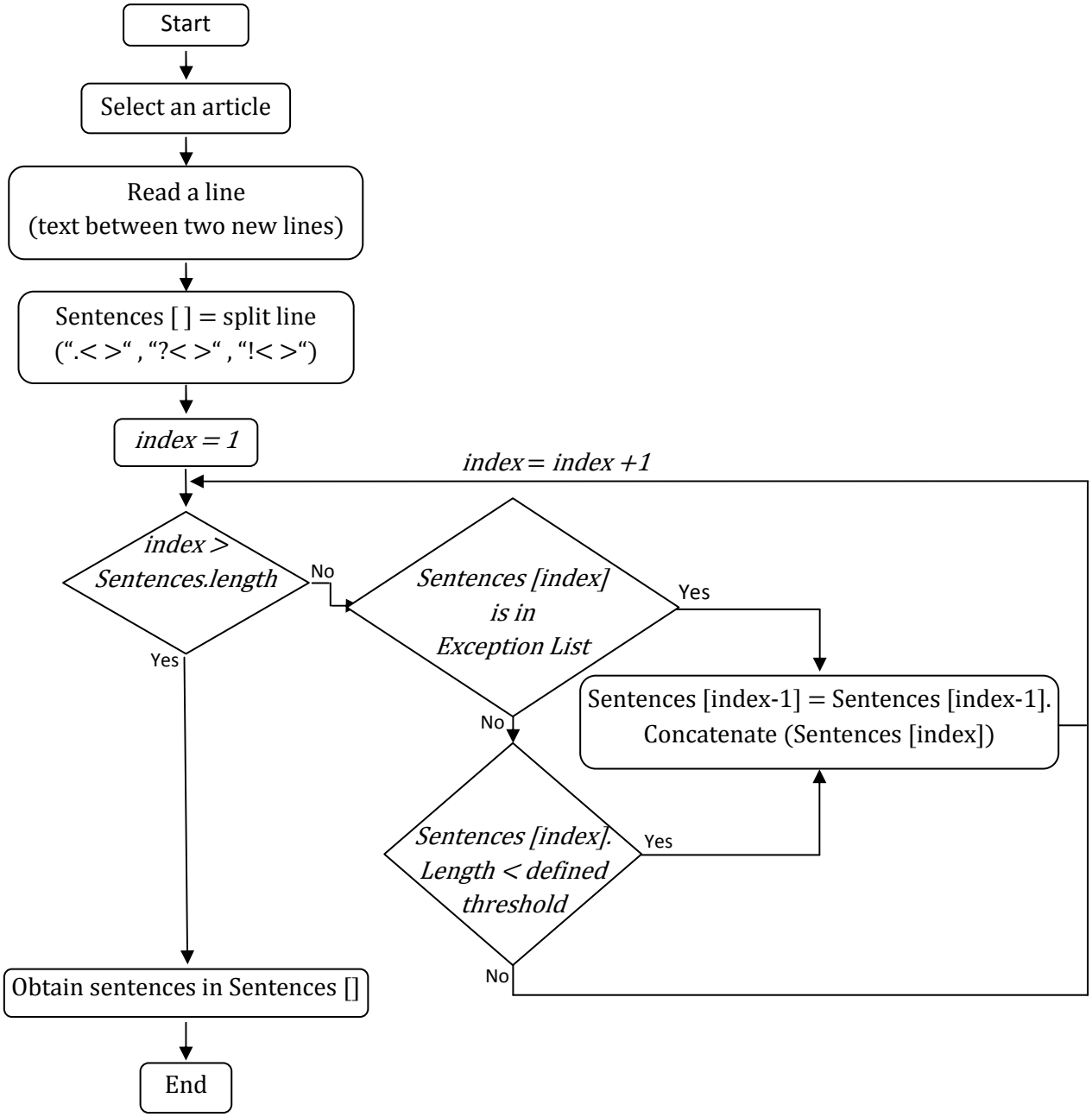

Figure 3.7: Algorithm defined for the sentence boundary detection

Even though the average sentences per paragraph is four (Table 3.2), the most frequent number of sentences per paragraph is two. Figure 3.8 shows the frequency distribution of the number of sentences per paragraph for all articles.

According to the graph in figure 3.8, number of sentences per paragraph has skewed towards the left, but most of the paragraphs of the collected articles contain two to seven sentences, which is fare enough for the defined research.

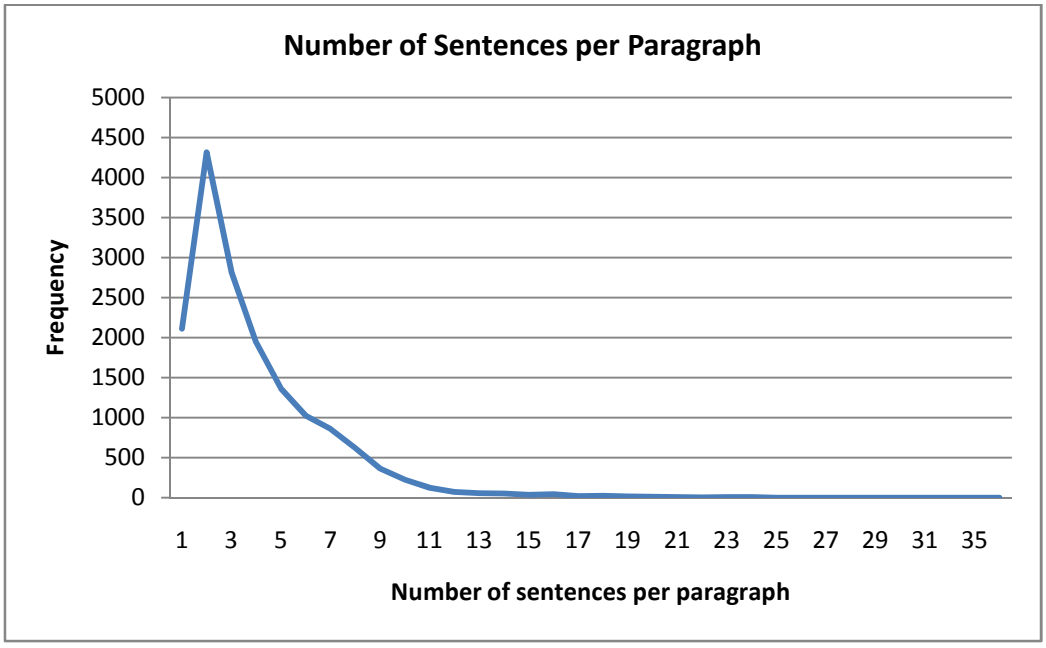

Figure 3.8: Distribution of number of sentences per paragraph

The unpublished statistics of *UCSC 10M Words Sinhala Corpus* has proved that the average word length of a Sinhala sentence is 12. The average word length of a sentence of this selected editorial corpus is 13, and it is a reasonable figure especially for the Sinhala text in newspaper articles. Figure 3.9 shows the frequency distribution of the sentence length (in words) of the editorials.

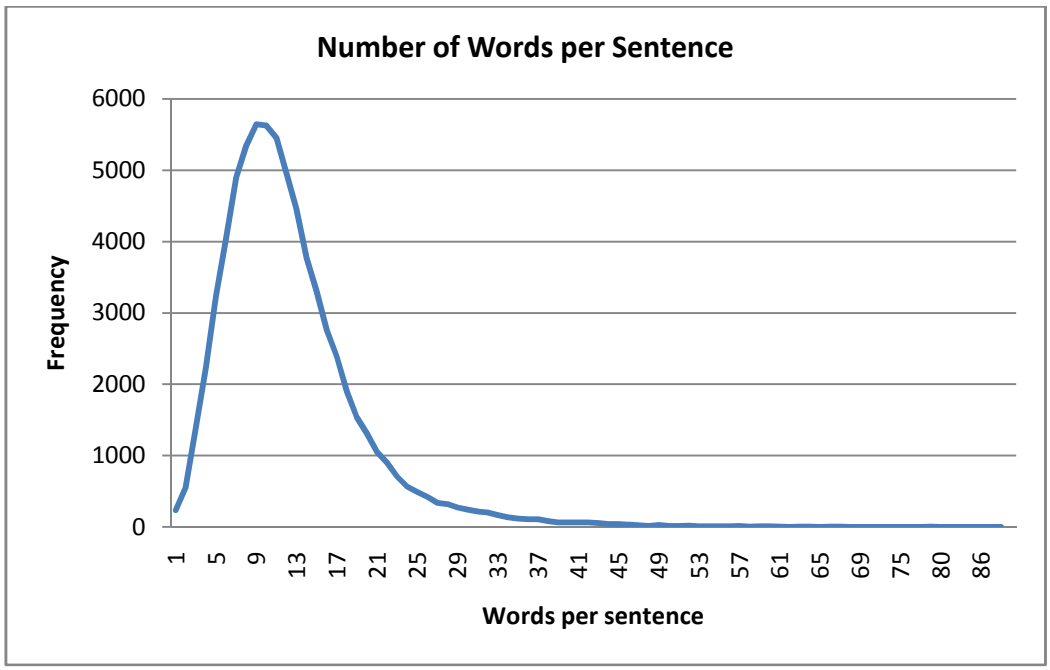

Figure 3.9: Distribution of number of words per sentence

According to the graph shown in figure 3.9, the distribution of number of words per sentence followed a (skewed) poison distribution. This indicates that the data-set prepared for the research is unbiased and fitting for a research like automatic summarization.

Experiments carried out with the data-set using the designed methodology in detail along with the experimental results and then the evaluation of the results will be described in the next chapter.

Chapter 04 – Experiments and Results

This chapter explains the experiments carried out using the designed methodology for the data-set which was explained in the previous chapter. Results of those experiments along with their evaluation are also described with related assumptions and hypotheses.

4.1 Creating Manual Summaries

The proposed methodology of the research in automatic text summarization for Sinhala is evaluated against the manually selected extracts, which is marked by the language experts. 120 articles were randomly selected among 1,400 articles of the corpus based on the algorithm defined in figure 4.1.

.

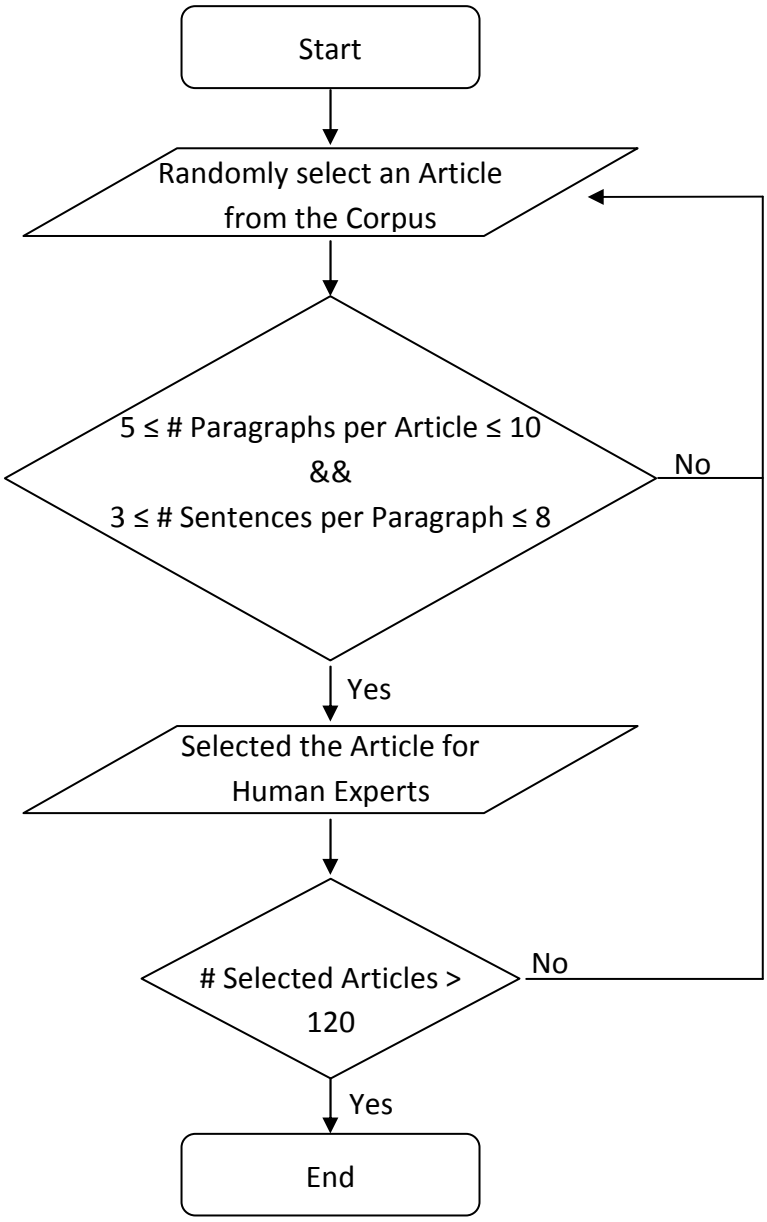

Figure 4.1: Algorithm for selecting articles for human annotation

This algorithm selects only the articles which have reasonable number of paragraphs per article and reasonable number of sentences per paragraph by pruning the other articles. Even though the average number of paragraphs per article is 13 (Table 3.2), most of the articles contain 5 to 10 paragraphs. Figure 4.2 illustrates the frequency distribution of number of paragraphs per article over 1,400 articles and the range marked in between two horizontal lines indicates the range which is selected to create manual extractions.

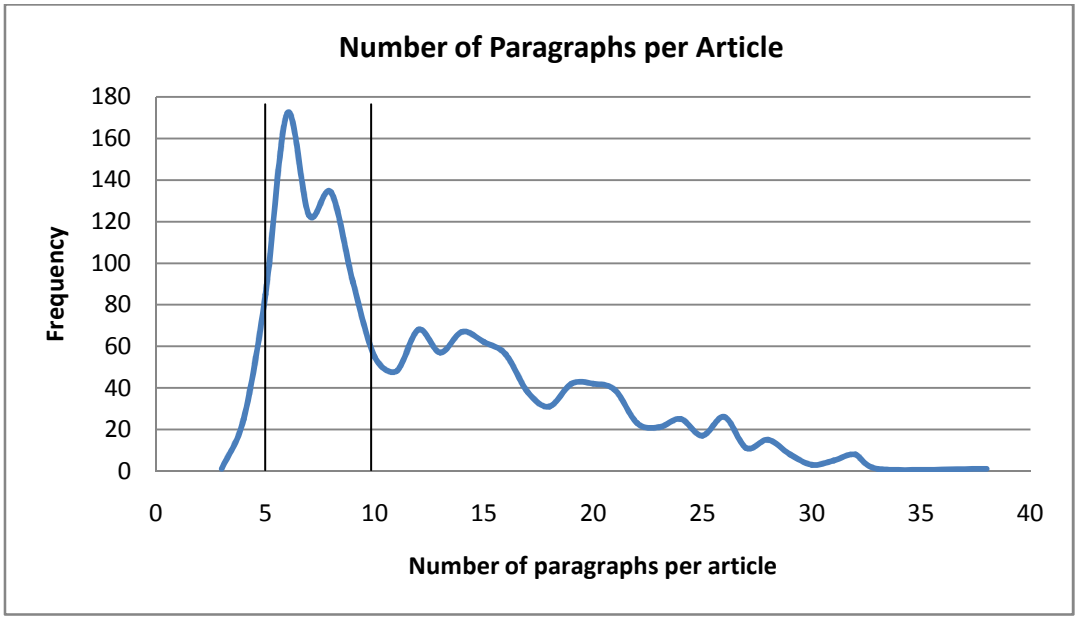

Figure 4.2: Distribution of number of paragraphs per article

The approach defined for selecting articles for manual annotation prevented summarizing biased and mis-formatted articles by humans. It helped to get the maximum outcome from the human effort and provide unbiased and solid data for estimating the parameters defined in the equation 09.

However, the number of articles with human annotations was limited only for 120 articles due to the lack of human resources. It was assumed that the 100 articles for training and the 20 articles for testing are sufficient to evaluate the performance of the summarizer.

Selected 120 articles were randomly allocated for three native language experts who have professional experience in writing and summarizing articles in different domains. They were asked to highlight the 10 most important sentences of each article, which could be described as the intention of the articles. They were instructed to select only the 10 most important sentences of the article, (which have 50 sentences in average) rather than ranking them based on the importance, mainly to reduce their work load and the complexity of the work. However,

it restricts the evaluator to perform on different compression rates and the evaluation is performed only for 20% compression rate. According to Mani (2001: 14), the compression ratio of a summary usually runs anywhere between 5% and 30% and therefore the 20% of compression rate is a reasonable ratio to evaluate the performance of the summarizer.

The designed approach for automatically summarizing text was kept away from human annotators and they were allowed to mark the most important sentences according to their intuitive knowledge, without any specific guidelines. Once they returned all 120 articles after highlighting 10 most salient sentences, paragraph numbers and sentence numbers of those sentences for each article were entered to the computer and the corpus of *Manually Extracted Summaries* was created.

4.2 Defining the Evaluation Criteria

To evaluate the quality of computer extracted summaries against the manually extracted summaries, the *Precision* and *Recall* were calculated for the computer extracted summaries. Calculating the Precision and Recall to measure the relevance of a set of machine generated data against to a real data-set is a well established technique, especially in the domain of pattern recognition and information retrieval. Precision is defined as the fraction of retrieved instances that are relevant, while Recall is defined as the fraction of relevant instances that are retrieved. Equation 11 and 12 show the exact mathematical definitions of Precision and Recall respectively.

$$Precision = \frac{|\{relevant\ instances\} \cap \{retrieved\ instances\}|}{|\{retrieved\ instances\}|} \quad ---- \ (11)$$

$$Recall = \frac{|\{relevant\ instances\} \cap \{retrieved\ instances\}|}{|\{relevant\ instances\}|} \quad ---- \ (12)$$

As it can be seen in equation 11 and 12, if it attempts to increase the Recall by retrieving more instances, it will cause to decrease the Precision and vice versa. Therefore, to get the maximum values for both of these measures, the harmonic mean of the Precision and Recall, called *F-Score* is calculated. F-Score reaches its best value at 1 and worst score at 0. Even though there are some variations of the definition for the F-Score, the traditional definition which was used to evaluate these experiments is shown in equation 13.

$$F - Score = 2.\frac{Precision \times Recall}{Precision + Recall} \quad - - - - \ (13)$$

This F-Score measure was calculated for each computer generated and manually extracted summaries to evaluate the performance of the proposed methodologies.

4.3 Identifying the Best Approach for each Individual Feature

As explained in the previous chapter, several possible approaches were experimented for selecting the best single approach to define each of four features. Out of 120 articles which were selected for human annotated summaries, randomly selected 100 articles were used as the training set to estimate these best individual approaches and to tuning the parameters (which are defined in equation 09) while the rest of 20 articles were kept as the test set for the final evaluation.

4.3.1 Weighting for Keywords

As explained in section 3.2.1, two different stemming approaches were defined to stem the words before getting their frequencies, namely Knowledge-Based Approach (KBA) and Data-Driven approach (DDA). In addition to the above two approaches, experiments are carried out without stem the words as well (hereafter this approach is referred as WSW, where WSW stands for Without Stem the Words), to measure the impact on stemming for Sinhala text Summarization. Three different sets of *tf.idf* values were calculated for each of 1,400 articles of the main corpus and then all the sentences were weighted in each 100 articles in the training set separately as defined in equation 04 (in page 28).

100 summaries were created only using the keyword feature based on DRA by extracting 10 most weighted sentences of each 100 articles. The original sentence order of the source article was retained to maintain the flow of information of the extracted summary. Then the Precision, Recall and then the F-Score for each 100 articles were calculated with respect to their corresponding manually extracted summaries. Same steps were followed to find the F-Scores of each 100 articles based on KBA and WSW as well. Table 4.1 shows a sample of different F-Score values calculated based on KBA, DRA and WSW for 10 articles.

To compare these three approaches, the mean and the standard deviation of these three values was calculated. The mean of the DRA is slightly better than WSW and KBA, but its standard deviation is higher than other two approaches. Then, the coefficient of variation (relative

standard deviation) was calculated as defined in equation 14 to compare these three approaches.

$$C_v = \frac{\sigma}{\mu} \times 100 \quad - - - - \ (14)$$

Where,

C_v is the coefficient of variation (CV),

σ is the standard deviation and

μ is the mean

Table 4.1: F-Score values calculated based on KBA, DRA and WSW for 10 articles

Article Name	DRA	KBA	WSW
DN_002.txt	0.4	0.4	0.3
DN_058.txt	0.3	0.2	0.3
DN_097.txt	0.4	0.4	0.2
DN_1007.txt	0.3	0.2	0.2
DN_1012.txt	0.2	0.3	0.2
DN_1015.txt	0.2	0.4	0.2
DN_1018.txt	0.2	0.2	0.2
DN_1020.txt	0.3	0.3	0.3
DN_1028.txt	0.3	0.2	0.3
DN_1033.txt	0.4	0.3	0.4

Table 4.2 shows the values for all three variables for DRA, KBA and WSW approaches.

Table 4.2: Mean, Standard Deviation and CV values for DRA, KBA and WSW

Approach	Mean	Stdev	CV %
DRA	0.334	0.131	39.28
KBA	0.328	0.127	38.78
WSW	0.330	0.120	36.42

According to the table 4.2, WSW gives best performance because its CV measure is lower than that of the other two approaches. This indicates that the stemming the words before calculating the frequencies do not affect for Sinhala text summarizing, if we only consider the keyword feature. However, the experiments were carried out further with and without stemming the words before coming to a final conclusion, because the final approach for Sinhala text summarization is dependent on three other features as well. Since the difference

of DRA and KBA for stemming Sinhala words is considerably low, it was concluded that the two approaches used to stem the words are giving the same performance and hence, the simple lightweight algorithm used in DRA can be successfully used for stemming Sinhala words for the applications such as automatic summarizers. Using such lightweight approaches is more efficient, especially for less resourced languages such as Sinhala since it does not require any rich linguistic resources such as morphological parsers.

4.3.2 Weighting for the Sentence Location

As explained in section 3.2.3, three different equations (equation 06, 07 and 08) were defined to weight sentences based on its location. Each sentence in all 100 articles weighted based on these three questions separately and generated 300 computer extracted summaries by using the sentence location as the only feature. The F-Score for each file was calculated based on the equation 13 and table 4.3 shows those F-Score values for sample of 15 articles. As each column title indicates, three columns show F-Score values for three different equations.

Table 4.3: F-Score values generated by the sentence location feature, calculated based on three defined equations for 15 articles

Article Name	Linear Function	Hyperbolic Function	Quadratic Function
DN_002.txt	0.6	0.6	0.2
DN_058.txt	0.4	0.4	0.2
DN_097.txt	0.5	0.5	0.3
DN_1007.txt	0.4	0.4	0.4
DN_1012.txt	0.6	0.6	0.1
DN_1015.txt	0.8	0.8	0.2
DN_1018.txt	0.4	0.4	0.1
DN_1020.txt	0.5	0.5	0.3
DN_1028.txt	0.5	0.5	0.0
DN_1033.txt	0.4	0.4	0.1
DN_1038.txt	0.6	0.6	0.1
DN_1049.txt	0.2	0.2	0.3
DN_1051.txt	0.2	0.2	0.3
DN_1056.txt	0.3	0.3	0.2
DN_268.txt	0.4	0.5	0.0

To compare the defined three equations based on these F-Score values, the mean, standard deviation and the confident of variation of three data-sets were calculated. Table 4.4 shows the values gained for these three measurements.

Table 4.4: Mean, Standard Deviation and CV values for three equations based on sentence location feature

Approach	Mean	Stdev	CV %
Linear Function	0.449	0.137	30.61
Hyperbolic Function	0.450	0.129	28.69
Quadratic Function	0.348	0.166	47.71

As the data shown in table 4.4, quadratic function gives the considerably low mean value with compared to the other two approaches. Also, it has the highest deviation from the mean, which has caused to get higher coefficient of variation. Therefore, it can be easily conclude that a quadratic function is not suitable to explain the distribution of the information in Sinhala sentences over a paragraph.

Linear function and hyperbolic function give almost similar mean values and it indicates that there is no significant difference in these two variations. However, since the F-Score values of linear function has a lower mean value, hyperbolic function was selected as the most suitable approach to weight sentences based on its location within a paragraph.

The results in the table 4.4 indicate the behavior of Sinhala language, especially on editorials. The hypothesis made by assuming the quadratic behavior of Sinhala sentences over a paragraph has failed while the other two hypotheses are equally true. The evidences are not sufficient to generalize this behavior for Sinhala language since the genre of the data-set is specific and follows a particular style. However, it can clearly claim that, the style of the Sinhala editorials is to present the most informative sentence at the very beginning of a paragraph and then rest of sentences are in the paragraph is used to describe it.

4.3.3 Weighting for the Paragraph Location

Three equations defined to find the distribution of information over a paragraph were also used to measure the distribution of information over an article by paragraphs. Sentences within a paragraph are weighted only based on their paragraph location separately for three defined equations (equation 06, 07 and 08) and 300 computer extracted summaries were generated by extracting 10 most weighted sentences from each article. Finally, F-Score values for each article were calculated according to the equation 13 and table 4.5 shows the calculated F-Score values based on three different equations for sample of 15 articles.

Table 4.5: F-Score values generated by the paragraph location feature, calculated based on three defined equations for 15 articles

Article Name	Linear Function	Hyperbolic Function	Quadratic Function
DN_002.txt	0.2	0.2	0.1
DN_058.txt	0.3	0.3	0.2
DN_097.txt	0.2	0.2	0.2
DN_1007.txt	0.3	0.3	0.2
DN_1012.txt	0.3	0.3	0.2
DN_1015.txt	0.1	0.1	0.2
DN_1018.txt	0.3	0.3	0.3
DN_1020.txt	0.4	0.4	0.2
DN_1028.txt	0.4	0.4	0.2
DN_1033.txt	0.1	0.1	0.3
DN_1038.txt	0.3	0.3	0.3
DN_1049.txt	0.5	0.5	0.6
DN_1051.txt	0.1	0.1	0.2
DN_1056.txt	0.1	0.1	0.2
DN_268.txt	0.3	0.3	0.1

To compare the defined three equations based on these F-Score values, the mean, standard deviation and the confident of variation for three data-sets were calculated. Table 4.6 shows the values obtained for these three measurements.

Table 4.6: Mean, Standard Deviation and CV values for three equations based on paragraph location feature

Approach	Mean	Stdev	CV %
Linear Function	0.320	0.138	43.30
Hyperbolic Function	0.320	0.138	43.30
Quadratic Function	0.369	0.166	45.03

Even though the mean value gained from the quadratic function is higher than the other two approaches (as shown in table 4.6), the distribution of the data-set has deviated from the mean than the other two approaches and that caused to gain a higher coefficient of variation. Therefore, the quadratic function was deselected to assign weights for the paragraph location of an article. However, both linear function and hyperbolic function give the same performance and therefore it was concluded that the difference between the linear function behavior and the hyperbolic function behavior does not effect for the distribution of information through paragraphs of an article. Finally the linear function was selected to assign weights for the paragraph location due to simplicity.

Based on the selection criteria explained above, final selected approaches to define each of the four features can be summarized as follows:

1. **Keywords:** *if.idf* values will be calculated after stemming each word using the data-driven approach and without stemming the words. The weights related to keywords will be calculated as defined in the equation 04 (in page 28).

2. **Title Words:** A weight will be assigned to the sentences which contain title words as defined in the equation 05 (in page 28).

3. **Sentence Location:** Weights will be assigned for sentences based on its location within a paragraph using a hyperbolic function as defined in the equation 07 (in page 30).

4. **Paragraph Location:** Weights will be assigned for sentences based on its paragraph location in the article using a linear function as defined in the equation 06 (in page 29).

Approaches used to identify each of the above features were evaluated based on the F-Score values generated by comparing machine extracted summaries against the human extracted summaries. It was considered only a single feature at a time to generate those machine extracted summaries. However, to use these four features for a summarization application, it has to find the best possible proportions which these four features can be combined to give the maximum performance. Experiments were carried out to find the best possible proportions of defined four features by tuning parameters defined in the equation 09.

4.4 Tuning the Parameters

The summarizer designed to summarize Sinhala editorial text is defined by a linear combination of described four features as defined in the equation 09 (in page 33). Experiments were carried out to find the best possible combination of defined four features by assigning and testing all possible values for the constants α, β, γ and δ. Algorithm defined to assign and test all possible combinations for the parameters α, β, γ and δ is defined as figure 4.3.

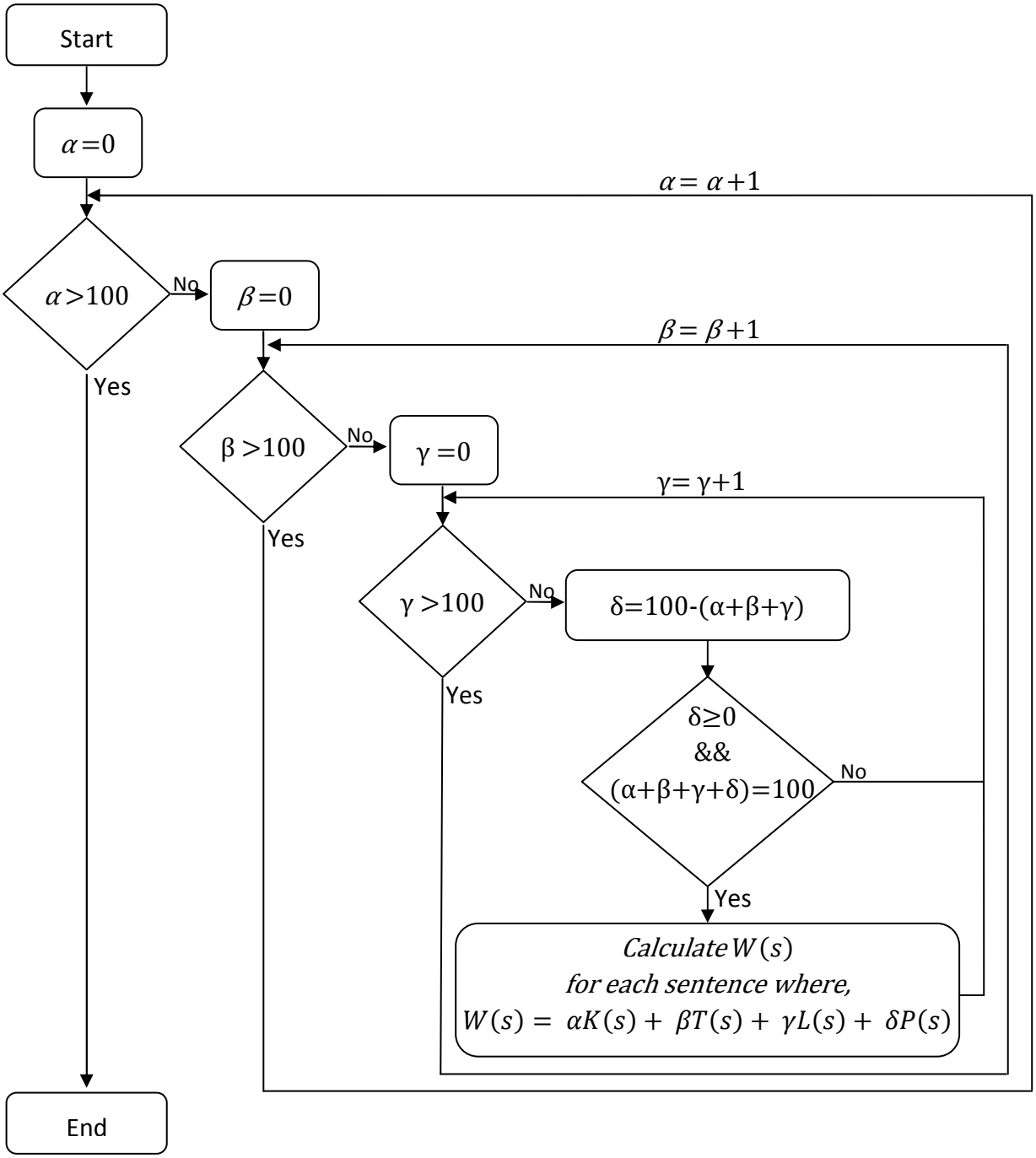

Figure 4.3: Algorithm defined to assign all possible values for parameters

According to the algorithm defined in figure 4.3, 176,851 possible combinations will be generated for α, β, γ and δ parameters. The value assigned for each parameter varies from 0 to 100 and it is considered as the percentage of the feature which will be assigned for the final weight calculation for a sentence. Table 4.7 shows the first 20 combinations for α, β, γ and δ parameters out of 176,851 possibilities, generated from the above algorithm.

Table 4.7: First 20 combinations generated from the algorithm defined in figure 4.3

Case	α	β	γ	δ
1	0	0	0	100
2	0	0	1	99
3	0	0	10	90
4	0	0	100	0
5	0	0	11	89
6	0	0	12	88
7	0	0	13	87
8	0	0	14	86
9	0	0	15	85
10	0	0	16	84
11	0	0	17	83
12	0	0	18	82
13	0	0	19	81
14	0	0	2	98
15	0	0	20	80
16	0	0	21	79
17	0	0	22	78
18	0	0	23	77
19	0	0	24	76
20	0	0	25	75

As explained above, these values for α, β, γ and δ parameters are considered as their percentage for the final equation. For example, in *case 10*, summaries will be generated for all 100 articles (which are in the training set) by assigning weights for sentences as, 0% from Keywords, 0% from Title Words, 16% from the Sentence Location and 84% from the Paragraph Location.

Once it generates summaries for all 100 articles for a given combination, F-Score values for all 100 articles will be calculated as defined in the equation 13. Then, a single F-Score value will be assigned for a given combination by computing the mean value for all F-Scores of 100 articles as defined in the equation 15.

$$F_{(\alpha,\beta,\gamma,\delta)} = \frac{\sum_{i=1}^{n} f_{i(\alpha,\beta,\gamma,\delta)}}{n} \times 100\% \quad - - - - (15)$$

Where,

$F_{(\alpha,\beta,\gamma,\delta)}$ is the mean F-Score value calculated for all the *n* articles for the given combination of α, β, γ and δ,

$f_{i_{(\alpha,\beta,\gamma,\delta)}}$ is the F-Score value of the i^{th} article for the given combination of α, β, γ and δ and,

n is the total number of articles which is generated summaries for the given combination of α, β, γ and δ.

Table 4.8 shows the averaged F-Score values (using the equation 15) for the first 20 combinations for α, β, γ and δ. According to the table 4.8, some combinations give higher averaged F-Score values while other combinations generate lower average.

Table 4.8: Averaged F-Score values for first 20 combinations of α, β, γ and δ

Case	α	β	γ	δ	F-Score
1	0	0	0	100	32.0
2	0	0	1	99	35.3
3	0	0	10	90	35.3
4	0	0	100	0	44.9
5	0	0	11	89	35.3
6	0	0	12	88	35.3
7	0	0	13	87	35.4
8	0	0	14	86	35.5
9	0	0	15	85	35.8
10	0	0	16	84	35.9
11	0	0	17	83	35.9
12	0	0	18	82	36.0
13	0	0	19	81	36.0
14	0	0	2	98	35.3
15	0	0	20	80	36.0
16	0	0	21	79	35.9
17	0	0	22	78	36.2
18	0	0	23	77	36.5
19	0	0	24	76	36.5
20	0	0	25	75	36.7

By analyzing all of these 176,851 combinations, three different combinations were identified to get the maximum averaged F-Score values. Table 4.9 shows the results of these best combinations with the averaged F-Score values obtained for each individual feature alone for the comparison. The first 100 best possible combinations and first 100 worst possible combinations for α, β, γ and δ parameters have been attached as the Appendix B.

Table 4.9: Averaged F-Score values for each individual feature and for best three combinations

Case	α	β	γ	δ	A. F-Score
1	**100**	0	0	0	33.4
2	0	**100**	0	0	41.4
3	0	0	**100**	0	44.9
4	0	0	0	**100**	32.0
5	**34**	**34**	**32**	**0**	**47.0**
6	**50**	**25**	**25**	**0**	**46.6**
7	**10**	**45**	**45**	**0**	**46.6**

As it can be seen in the data on table 4.9, the linear combination of features gives better performance than any of individual feature. However, the results revealed that the feature introduced based on the paragraph location is not significant at all for extracting most important sentences from an article. The best most 110 combinations have the value zero for δ parameter, which is defined to assign weights for the paragraph location. The averaged F-Score value obtained by considering the paragraph location as the only feature (case 4 in table 4.9) gives the lowest F-Score value among all possible 176,851 combinations. The F-Score value obtained only from keyword feature (case 1 in table 4.9) is the second lowest value among the results. Conclusions based on these experimental results will be discussed in detail in the next chapter.

Based on the results obtained from the above experiment, the equation 09 (in page 33) defined to assign weights for the sentence *s* can be re-defined as in equation 16, to get the maximum performance for summarizing Sinhala text.

$$W_{max}(s) = \alpha K(s) + \beta T(s) + \gamma L(s) \quad - - - - (16)$$

Where,

$W_{max}(s)$ is the maximum weight can be obtained for the sentence *s,* and

α = 34, β = 34 and γ = 32 or,

α = 50, β = 25 and γ = 25 or,

α = 10, β = 45 and γ = 45

Finally, to measure the impact of stemming Sinhala words before counting their frequencies, the F-Score value was calculated based on equation 16, without stemming the words. Table 4.10 shows the F-Score values with and without stemming the words.

Table 4.10: Averaged F-Score values for the best three combinations with and without stemming the words

Case	α	β	γ	Average F-Score Value	
				With Stemming	Without Stemming
01	34	34	32	47.0	43.5
02	50	25	25	46.6	46.3
03	10	45	45	46.6	46.1

According to the values in Table 4.10, there is a slight improvement of the performance when the words are stemmed before calculating the frequencies. Therefore stemming the words is helpful to improve the performance of Sinhala Text Summarizer.

4.5 Experiments with the Test Data

After identifying the best possible combinations for the defined four features using the training data (100 articles), the averaged F-Score values were calculated for the test data (20 articles), which were unseen during the parameter tuning process. F-Score values which were calculated for the best three combinations based on the equation 16 are shown in the table 4.11 for both the training and the test data.

Table 4.11: Averaged F-Score values for the best three combinations

Case	α	β	γ	Average F-Score Value	
				Training Data	Test Data
01	34	34	32	47.0	45.0
02	50	25	25	46.6	45.0
03	10	45	45	46.6	45.0

According to the values shown in table 4.11, the best three combinations give the same performance with the test data. However, the averaged F-Score values for the test data are lower than the values obtained from the training data.

Table 4.12 shows the averaged F-Score values obtained for the three worst possible combinations for both the training and the test data. Values of the table 4.12 (case 01) prove that the paragraph location feature is the worst individual feature even with the test data. However, the F-Score value gained only from the keyword feature shows a significant improvement with the test data even though it was the second lowest value with the training data. Manual observation of the test data revealed that this is due to highly subject specific

four articles in the test data which have gained over 60 F-Score value using the keyword feature alone.

Table 4.12: Averaged F-Score values for the worst three combinations

Case	α	β	γ	δ	Average F-Score Value	
					Training Data	Test Data
01	0	0	0	100	32.0	33.0
02	100	0	0	0	33.4	42.0
03	82	2	0	13	33.6	36.5

Sample of machine extracted summaries and human extracted summaries along with their source articles have been attached as the Appendix A.

Experiments explained above were carried out to identify the most suitable approach to define each individual feature and then to identify the best possible combination of them for weighting sentences to create a summary. Three best combinations to weight Sinhala sentences were identified and those combinations were verified by testing them against the test data. Conclusions come up with these results and the possible future works will be discussed in the next chapter.

Chapter 05 – Conclusion and Future Works

This chapter is summing up the outputs of the designed research and the achievements gained through it. Author's view of the achieved results is discussed in detail with the possible future works which can be carried out to improve the quality of a summarizer, which is specially designed for summarizing Sinhala text.

5.1 Conclusion

As the title indicates, this research was carried out to automatically summarize the text written in Sinhala. There are no any previous attempts recorded in the literature to identify possible approaches for summarizing Sinhala text. However, a vast amount of research has been carried out and many different approaches have been tried out over the last six decades to identify the best possible approaches to automatically summarize human languages. Most of these approaches and the linguistic resources developed to aid them were built for the languages such as English, which shows different linguistic behavior than Sinhala.

This research was mainly focused on finding the most suitable approaches for automatically summarizing Sinhala texts with minimum linguistic resources. Therefore, the experiments were carried out based on classical approaches used in automatic text summarization. Experimental results prove that some thematic features which researchers have identified for the languages such as English can be used for Sinhala language as well for the same objectives. Features which have identified at the earlier stages of the field have been modified by successive research and they were experimented by this research to find out how such improved features work for Sinhala language. The results of the research proved that the same performance can be achieved for Sinhala language as well, after identifying each individual feature's behavior separately.

Evaluation is an essential part of a practical discipline like automatic summarization. However, it is crucial to say one summary is better than another summary even though it can be easily said if it is a bad summary. Researchers have been launched over the last six decades to find out most accurate ways to evaluate machine generated summaries. Since humans need to be involved to judge the machine outputs for giving a perfect evaluation of a summary and although it is expensive, scoring programs are preferred to evaluate machine generated summaries. Therefore, the technique used to evaluate the output of this research was selected by considering the factors such as cost effectiveness and repeatability.

Evaluation of this research is entirely based on a single hypothesis, which is *the human annotated summaries are perfect*. However, it need not be true for all the cases since creating

summaries are highly subjective and it depends on many parameters such as user audience, genre of the source text and compression ratio among others. There is no deep theory of summarization. Two summaries of a same source, but made by two human experts need not to be identical. This phenomenon is significant for abstracts, but it can be trivial for making extracts from a 54 sentenced article. Therefore, it was attempted to achieve the best score by comparing the machine extracted summaries against human extracted summaries and it was able to achieve 47% Precision and 47% Recall with the training data while it achieved 45% Precision and Recall with test data for 20% compression rate.

Kupiec, Pedersen and Chen (1995) work on trainable document summarizer (KPC approach) had been achieved only 42% Recall as the overall performance. They managed to peak 84% sentence Recall by lengthening the summary, but they have not calculated the Precision, which may have gone down while the Recall is increased. Teuful and Moens (1999) work based on the KPC approach have been reported 54.4% Precision for their Indicator Quality feature which they define to denote the presence of cue phrases. They have gained 66% Precision for the combination of Indicator Quality, Location, Sentence Length, Title, Header and Thematic words. Teuful and Moens (1999) have used some discourse analysis of text to identify the rhetorical roles of the text and have been reported 56.3% Precision for a single feature and 64.2% Precision for the collective performance. Jeganathan (2005) also reported 47% Precision for Tamil text extraction using *tf.idf* based keywords.

Achieving 45% Precision and Recall for unseen test data is sensible with compared to the above results gained for previous works, especially since this research is designed based on minimum linguistic resources. Similar methods can be applied for many other such less resourced languages even though they do not have basic linguistic resources to process their languages. Results of this research lead to build a useable automatic text summarizer for Sinhala using minimum linguistic resources which was the primary objective of the research.

Another main objective of this research was to identify the best approaches to define each individual feature. Even though the stemming of words is not significant for weighting sentences only based on keywords, stemming has an impact for the overall performance of the summarizer. That may be due to title word feature, because title words were stemmed before looking for them in the sentences. Two stemming algorithms were defined to stem Sinhala words before calculating their *tf.idf* values for each word, and the results revealed that there is no significant difference between the proposed two approaches. Therefore, the proposed lightweight algorithm (defined in Figure 3.1) can be successfully used for inflectional languages such as Sinhala for stemming their words, especially if they do not have such rich

linguistic resources described in Weerasinghe, Herath and Welgama 2009. The proposed algorithm can be enhanced using linguistic knowledge to improve the accuracy achieved.

The results of the experiments carried out to find the distribution of information in Sinhala articles show that the information distribution of a Sinhala article can be described using a linear function or a hyperbolic function, but it cannot be explained using a quadratic function. That is the sentences which carry most salient information are laid on the beginning of a paragraph while the importance of them is desecrated along with its location gradually or significantly. It can be concluded that the most informative sentences do not likely occur at the end of the paragraph. However, Edmundson (1969) has used a quadratic function's behavior to assign weights for the sentence location based on the work carried out by Baxendale (1958). Nevertheless the Baxendale (1958) revealed that the most salient sentences were likely to occur as the first sentence of a paragraph 85% of the time. These results on information distribution over an article will be helpful in future research which is intended to process Sinhala language to gain the information.

Final scores gained by evaluating the machine extracted summaries against the human extracted summaries confirm that there are three best possible ways to linearly combine the four features. All three best combinations explored that it is not worth to consider the paragraph location as an individual feature. According to the experimental results, paragraph location is the worst individual feature and the top most 110 possible combinations have zero weight for the paragraph location. These results indicate that the combination of Keywords, Title words and the sentence location is sufficient to achieve the best performance for a Sinhala summarizer.

According to the experimental results in table 4.9, the sentence location is the best individual feature while the keywords alone is the worst for Sinhala language as well (since the paragraph location is no longer considered as a feature for scoring sentences). The worst 1300 combinations have zero value for the sentence location feature and it strongly proves that how significant the sentence location feature to weight most informative sentences. This result has been proved for the English language by Edmundson (1969) by using a similar experiment. However, the final results in table 4.10 show that the equal contribution of each of three features will also cause for the maximum performance. It always tends to keep both title feature's and sentence location feature's contribution equally while the contribution of the keyword feature may get lower or higher for achieving best results.

One of the objectives of this research is to provide a benchmark for the future research on automatic text summarization in Sinhala. Experimental results which have been reported in Chapter 04 can be used by successive researchers to compare their results which will be taken

from new approaches or by applying same techniques for a different domain rather than news editorials.

5.2 Future Works

This research was carried out based on the classical approaches used in automatic text summarization. The main objective on selecting such approaches is to carry out the research with minimum available linguistic resources for Sinhala language. Also, it was aimed to ensure the adaptability of such approaches for languages like Sinhala. Future extensions of this research can be carried out in many directions and this section is intended to describe some of these in detail.

As explained in Chapter 3.3, the data-set used to carry out this research is the editorials of three national daily newspapers. Virtually, the content of an editorial can be anything and therefore it is hard to define a certain domain for the data-set rather than categorizing them as "news editorials", which is too broad to identify domain specific features such as cue words. Edmundson (1969) and successive researchers such as Pollock and Zamora (1975) have successfully used domain specific cue phrases as a feature for their classical approaches in summarization. Therefore, future researchers who will be working on summarizing Sinhala text can use such domain specific cue phrases as the forth individual feature instead of the paragraph location and can experiment for a better performance.

The experimental results on various methods of assigning weights show that the quadratic distribution does not work for Sinhala editorials. Further experiments can be done in different domains and may be with a different application to generalize this as a behavior of Sinhala language. This will eventually help structural linguists to identify the structural behavior of Sinhala language.

As explained in section 5.1, the evaluation of this work is based on an assumption that is the human summaries are perfect. Then the F-Score values were calculated by comparing the machine extracted summaries against the human extracted summaries. A future research can be carried out to compare both of these machine and human extracted summaries using humans to get some gray scale values as the comparison and that will give more accurate evaluation for the machine generated summaries. However, the evaluation techniques have to be designed carefully by considering the cost effectiveness and the repeatability of the process, especially when humans involve to the process.

Further research can be carried out using the same methodology and linguistic resources for evaluating the performances in different compression rates. However, to evaluate the performances in different compression rates, the human extracted summaries have to be annotated with the ranking. To make such annotated summaries, human summarizers need to be advised to select the most important and most informative sentence first and then rank the rest of sentences sequentially according to their importance. This resource can be used to evaluate machine extracted summaries in different compression ratios.

Once it is available such resource (human summaries with ranked sentences) as a linguistic resource, further research can be carried out based on the method used in Pardo, Rino and Nunes (2003) for developing their GIST SUMMarizer. They have identified most important sentence of an article (the gist sentence) based on some classical features and then have used some similarity measures for extracting rest of most important sentences which related to the gist sentence. Further research can be carried out to compare both these approaches using the same data-set. The results can be compared against the Jeganathan (2005) work carried out for Tamil language as well.

Future research on automatic text summarization can be enhanced with development of Sinhala linguistic resources. The techniques used in corpus based approaches can be applied for Sinhala language when a corpus of human created summaries with their source texts is available. Corpora can be used to train the features and that lead to give better performance for a summarizer. Other linguistic resources such as sentence parsers, taggers, named entity recognizers and WordNet will be greatly helpful to identify the information in a sentence and then to extract them and represent in a machine understandable format. Finally, such resources can be used to regenerate Sinhala language texts which are essential to present the processed summaries which are in machine readable format. The ability of generating Sinhala language text is vital to create abstract summaries which are more closed for human summaries.

Finally, the future researchers who work on automatic summarization in Sinhala can expand their work for the new areas on summarization such as abstraction, multi-document summarization and multi-media summarization with sufficient linguistic resources. Further, the hybrid approaches such as integrating statistical models with other information such as shallow features, discourse structure and thesauri for generalization can be experimented to gain improved summary extracts and abstracts. New methods of evaluations will also be a challenging future area because the field of automatic summarization is still suffering from some generic, cost-effective and user-centered evaluation techniques.

References

ANSI, A. N. (1997). *Guidelines for Abstracts.* Bethesda, Maryland: National Information Standards Organization (NISO) Press.

Aone, C., Okurowski, M. E., Gorlinsky, J., & Larsen, B. (1999). A trainable Summarizer with Knowledge Acquired from Robust NLP Techniques. In I. Mani, & M. T. Maybury, *Advances in Automatic Text Summarization* (pp. 71-80). Cambridge, Massachusetts: MIT Press.

Banu, M., Karthika, C., Sudarmani, P., & Geetha, T. V. (2007). Tamil Document Summarization Using Semantic Graph Method . *International Conference on Computational Intelligence and Multimedia Applications*, (pp. 128-134). Sivakasi, Tamil Nadu.

Barzilay, R., & Elhadad, M. (1999). Using Lexical Chains for Text Summarization. In I. Mani, & M. T. Maybury, *Advances in Automatic Text Summarization* (pp. 111-121). Cambridge, Massachusetts: MIT Press.

Baxendale, P. B. (1958). Man-made index for technical literature: an experiment. *IBM Journal of research and Development 2* , 354-361.

Borko, H., & Bernier, C. L. (1975). *Abstracting Concepts and Methods.* San Diego, California: Academic Press.

CDAC-Noida. (2006). *Text summarization system for Hindi*. Retrieved 2011, from Centre for Development of Advanced Computing, Noida: http://www.cdacnoida.in/snlp/digital_library/text_summ.asp

Climenson, W. D., Hardwick, N. H., & Jacobson, S. N. (1961). Automatic syntax analysis in machine indexing and abstracting. *American Documentation* , 178-183.

Correira, A. (1980). Computing Story Trees. *American Journal of Computational Linguistics* , 135-149.

Edmundson, H. P. (1969). New methords in automatic abstracting. *Journal of the Association for Computing Machinery 16* , 264-285.

Gupta, V., & Lehal, G. (2011). Preprocessing Phase of Punjabi language Text Summarization. *Information Systems for Indian Languages* (pp. 250-253). Patiala, India: Springer Berlin Heidelberg.

Hahn, U., & Reimer, U. (1999). Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. In I. Mani, & M. T. Maybury, *Advances in Automatic Text Summarization* (pp. 215-232). Cambridge, Massachusetts: MIT Press.

Hovy, E., & Lin, C.-Y. (1999). Automated Text Summarization in SUMMARIST. In I. Mani, & M. T. Maybury, *Advances in Automatic Text Summarization* (pp. 81-94). Cambridge, Massachusetts: MIT Press.

Hovy, E., & Marcu, D. (1998, August). Automated Text Summarization Tutorial — COLING/ACL'98. Montreal, Quebec, Canada.

Jeganathan, T. (2005). *Automatic Text Summarizer for Tamil Using Sentence-Extractive Approach.* Colombo.

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. *In Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'95)*, (pp. 68-73). Seattle, WA.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IRE National Convention*, (pp. 159-165). New york.

Mani, I. (2001). *Automatic Summarization.* Philadelphia: John Benjamins Publishing.

Mani, I., & Maybury, M. T. (1999). *Advances in Automatic Text Summarization.* Cambridge, Massachusetts: MIT Press.

McKeown, K., Robin, J., & Kukich, K. (1995). Generating Concise Natural Language Summaries. *Information Processing & Management 31* , 703-733.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM* , 39-41.

Myaeng, S. H., & Jang, D.-H. (1999). Development and Evaluation of a Statistically-Based Document Summarization System. In I. Mani, & M. T. Maybury, *Advances in Automatic Text Summarization* (pp. 61-70). Cambridge, Massachusetts: MIT Press.

Nakao, Y. (2000). An Algorithm for One-Page Summarization of a Long Text Based on Thematic Hierarchy Detection. *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)* (pp. 1071-1075). New Brunswick, New Lersey: Association for Computatinal Lingustics.

NIE, N. I. (1989). *Sinhala Lekhana Reethiya.* Colombo: Government Printer, Sri Lanka.

Pardo, T. A., Rino, L. H., & Nunes, M. G. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken – PROPOR*, (pp. 210-218). Faro, Portugal.

Patel, A., Siddiqui, T., & Tiwary, U. (2007). A language independent approach to multilingual text summarization. *RIAO '07 Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, (pp. 123-132). France.

Pollock, J. J., & Zamora, A. (1975). Automatic Abstracting Research at Chemical Abstracts Service. *Journal of Chemical Information and Computer Science* , 226-232.

Rath, G. J., Resnick, A., & Savage, T. R. (1961). The Formation of Abstracts By the Selection of Sentences. *American Documentation*, (pp. 139-141).

Spärck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28 (1)* , 11-20.

Teufel, S., & Moens, M. (1999). Argumentative classification of extracted sentences as a frist step towards flexible abstracting. In I. Mani, & M. T. Maybury, *Advances in Automatic Text Summarization* (pp. 155-171). Cambridge, Massachusetts: MIT Press.

Weerasinghe, R., Herath, D., & Welgama, V. (2009). Corpus-based Sinhala lexicon. *ALR7 Proceedings of the 7th Workshop on Asian Language Resources* (pp. 17-23). Singapore: Association for Computational Linguistics.

Wikipedia. (n.d.). *Sinhala language*. Retrieved November 27, 2011, from Wikipedia: http://en.wikipedia.org/wiki/Sinhala_language

Appendix A – Samples of Source Article, Human Extracted Summaries and Machine Extracted Summaries

Source Article – DN_867.txt

අසම්මත වාර්තාවට තවත් ප්‍රතිචාර !
දරුස්මාන් වාර්තාව සම්බන්ධයෙන් තවදුරටත් විවිධ මත පළ වේ. ශ්‍රී ලංකාවේ රුසියානු තානාපති විලැදිමීර් මිහායිලෝව් මහතා කියා සිටින්නේ එය අතාර්කික වාර්තාවක් බව ය. එහි ඇතුළත් දත්ත හා තොරතුරු තහවුරු නොකළ ඒවා බවත් එහි ම භාෂා විලාසය ගැන තමන් පුදුමයට පත් වන බවත් රුසියානු තානාපතිවරයා කියා සිටී. රටක පැවති බරපතළ ප්‍රශ්නයක් නිරීක්ෂණය කොට ඒ පිළිබඳ වාර්තාවක් සකස් කිරීමේ දී අනුගමනය කළ යුතු විශ්වසනීය උපාය මාර්ග හා නිර්ණායක විද්වත් කමිටුවට අමතක වූ සෙයක් දකින්නට ලැබේ.
මේ වාර්තාවේ සංගෘහීත කරුණු මුළුමනින් ම කොටි සංවිධානයට පක්ෂපාත ලෙස ගොනු කර ඇති බව බොහෝ දෙනා පෙන්වා දෙති. ශ්‍රී ලංකාවේ ආරක්ෂක අංශ කොටි ග්‍රහණයේ සිටි දෙමළ ජනතාව බේරා ගත් ආකාරය ජන මාධ්‍ය මගින් ප්‍රකාශයට පත් කැරිණි. හමුදා නිලධාරීන් දරුවන් හා දෙමළ වැඩිහිටියන් ඔසොවා ගෙන යන අයුරු. රෝගීනට ප්‍රතිකාර ලබා දෙන අයුරු දකින කෙනකුගේ දෑසට කඳුළු පුරයි. එබඳු සංවේදී අවස්ථා දරුස්මාන් වාර්තාවේ නැත. ආරක්ෂක අංශවල මානව දයාව නිරූපණය කරන කිසිදු තොරතුරක් හෝ ඡායාරූපයක් හෝ වාර්තාවට ඇතුළත් නැත. එබැවින් එය ඒකපාර්ශ්වීය වාර්තාවක් ලෙස සැලකීම යුක්ති සහගත ය.
ශ්‍රී ලංකාවට ආර්ථික ආධාර සපයන කණ්ඩායමේ නායකයා වූ යසුසි අකාෂි මහතා යුද ජයග්‍රහණයෙන් පසු ද ඊට පෙර ද ශ්‍රී ලංකාවට පැමිණියේ ය. උතුරට ගිය අකාෂි මහතා කියා සිටියේ යුද්ධයෙන් පසු එම පළාත ඉතා ශීඝ්‍ර සංවර්ධනයක් අත්පත් කර ගනිමින් සිටින බව ය. එය පුදුම සහගත බව ද ඒ මහතා පෙන්වා දුන්නේ ය. බෑන් කී මූන් මහතා ද යුද්ධයෙන් පසු උතුරට ගියේ ය. එදා මූන් මහතා කියා සිටියේ උතුරේ සිදු වන සංවර්ධන ක්‍රියාවලිය ගැන තමන් බෙහෙවින් තෘප්තිමත් වන බව ය. උතුරු වසන්තය සංවර්ධන වැඩසටහන සඳහා ශ්‍රී ලංකා රජය විශාල මුදලක් යොදවා තිබේ.
උතුරු වසන්තය හා නැහෙනහිර නවෝදය යන සංවර්ධන වැඩසටහන් අදටත් සාර්ථක ව ක්‍රියාත්මක වේ. නිවාස සංවර්ධනය, නැවත පදිංචි කිරීම, මං මාවත් ඇතුළු පොදු පහසුකම් ඉදිකිරීම් ඉතා සාර්ථක ය. මේ හැරෙන්නට අඩපණ ව තිබූ කෘෂිකර්මය, ධීවර කර්මාන්තය හා සංචාරක කර්මාන්තය යළි යථා තත්ත්වයට පැමිණ තිබේ. මේ වැඩසටහන් දෙක මගින් නිරූපණය වනුයේ ආණ්ඩුව දෙමළ ජනයාට දක්වන ආදරය, කරුණාව හා අනුග්‍රහය යි. ත්‍රස්තවාදී ක්‍රියාකාරකම් පැවැති ප්‍රදේශයක ට ඉහළින් ම සලකන ආණ්ඩුවක් පවතිනු ඇත්තේ ශ්‍රී ලංකාවේ ම පමණි. එහෙත් මේ වෙනස දරුස්මාන් වාර්තාවේ සඳහන් නොවේ.
විපක්ෂ නායක රනිල් වික්‍රමසිංහ මහතා එක්සත් ජාතික පක්ෂයේ ස්ථාවරය පැහැදිලි කරමින්; මේ වාර්තාවේ ඇතැම් කරුණු සම්බන්ධයෙන් ඉතා තදින් ප්‍රතිචාර දැක්විය යුතු බව කියා තිබේ. ඇතැම් කරුණු නොව සමස්තයක් ලෙස මුළු වාර්තාවට ම විරෝධය දැක්විය යුතු ය. අව ම ලෙස වික්‍රමසිංහ මහතා එබඳු ප්‍රතිචාරයක් දැක්වීම ගැන සතුටු විය හැකි ය. මෙතෙක් විපක්ෂය වෙතින් අපට දකින්නට ලැබුණේ වාර්තා ව පසෙක තබා ආණ්ඩුව විවේචනය කිරීමක් පමණි. ආණ්ඩුව විවේචනය කිරීමේ ප්‍රජාතන්ත්‍රවාදී නිදහස රට තුළ තිබේ. තත්ත්වය එසේ වුව බුද්ධිමත් දේශපාලකයන් අවස්ථාවෝචිත ලෙස හැසිරිය යුතු ය. අදහස් දැක්විය යුතු ය.
දරුස්මාන් වාර්තාව මගින් සිදු වන හානිය සමස්ත ජාතියටත්, රටටත් බලපායි. ආණ්ඩුවට පක්ෂපාතී නොවන ස්වාධීන පුද්ගලයන් ද, දෙමළ ජාතික දේශපාලකයන් ද මේ තත්ත්වය පැහැදිලි කර දී තිබේ. ඊනියා වාර්තාව සකස් කිරීම සඳහා ගෙන ඇති සෑම පියවරක් ම නීත්‍යානුකූල නොවන අතර දෝෂ සහිත බව ද විශ්ලේෂකයෝ පෙන්වා දෙති. මූන් මහතා පවා සිටින්නේ වැට උඩ ය. ඇමෙරිකන් හමුදාව ඔසාමා බින් ලාඩන් ඝාතනය කිරීමත් සමඟ තත්ත්වය වඩාත් සංකීර්ණ වී තිබේ. බින් ලාඩන් ඝාතනය කළ ආකාරය හා ඔහුගේ සිරුර මුහුදට වීසි කළ ආකාරයත් වැඩි දෙනකුගේ විවේචනයට ලක් ව තිබේ. ඔසාමා බින් ලාඩන්ගේ මරණය ගැනත්; ඇමෙරිකා හමුදා ප්‍රමුඛ නෙටෝ සේනාංක ඇෆ්ඝනිස්ථානය හා පාකිස්ථානය තුළ සිදු කළ මෙහෙයුම් ගැනත් විමර්ශනය කළහොත් සිදු වන්නේ කුමක් ද ? ඒ සඳහා මූන් මහතා සූදානම් ද ? ඇෆ්ඝනිස්ථානය, පාකිස්ථානය, ඉරාකය වැනි රටවල සිදු වූ මිනිස් ඝාතන ගැන විමර්ශනය කළහොත්; ශ්‍රී ලංකාවේ ත්‍රස්ත මර්දන මෙහෙයුම් වල සාධාරණත්වය හා යුක්ති සහගත බව තේරුම් ගත හැකි ය. දරුස්මාන් වාර්තාව සකසන විට ලෝකයේ වෙනත් ත්‍රස්ත මර්දන ව්‍යාපාර ක්‍රියාත්මක වූ ආකාරය සමඟ ශ්‍රී ලංකාවේ මානුෂික මෙහෙයුම් ගැන සංසන්දනාත්මක ලෙස කරුණු විග්‍රහ කරන්නට ඉඩ තිබිණි. එහෙත් ඒ කිසිවක් සිදු වී නැත.
ත්‍රස්තවාදය මර්දනය කිරීම සම්බන්ධයෙන් පමණක් නොව ජාතීන් අතර සමගිය ගොඩනැඟීම සම්බන්ධයෙන් ද ලෝකයට පාඩමක් කියා දීමට ශ්‍රී ලංකාව සමත් විය. රටවල් ගණනාවක් ශ්‍රී ලංකාවේ අත්දැකීම් හා උපාය මාර්ග මේ වන විට ද උකහාගෙන තිබේ. ජුනි මාසයේ රටවල් තිස් ගණනක නියෝජිතයන්ගෙන් සමන්විත මහා සමුළුවක් පැවැත්වීමට ද නියමිත ය. එහි අරමුණ ශ්‍රී ලංකාවේ ත්‍රස්ත මර්දනය පිළිබඳ සාර්ථක අත්දැකීම් ලෝකයේ වෙනත් රටවල් සමඟ බෙදා ගැනීම ය. ලොව පුරා රටවල් අසූ ගණනක ත්‍රස්තවාදී ගැටුම් පවතින බව අමතක නොකළ යුතු ය. එම රටවලට ශ්‍රී ලංකාවෙන් ලබා ගත හැකි පාඩම් තිබිය දී ඊනියා වාර්තාවෙන් අපට මහත් අගෞරවයක් සිදු කර තිබේ.

Human Extracted Summary – DN_867.txt

අසම්මත වාර්තාවට තවත් ප්‍රතිචාර !
දරුස්මාන් වාර්තාව සම්බන්ධයෙන් තවදුරටත් විවිධ මත පළ වේ.
ශ්‍රී ලංකාවේ රුසියානු තානාපති විලැදිමීර් මිහායිලෝව් මහතා කියා සිටින්නේ එය අතාර්කික වාර්තාවක් බව ය.
මේ වාර්තාවේ සංගෘහිත කරුණු මුළුමනින් ම කොටි සංවිධානයට පක්ෂපාත ලෙස ගොනු කර ඇති බව බොහෝ දෙනා පෙන්වා දෙති.
උතුරට ගිය අකාෂි මහතා කියා සිටියේ යුද්ධයෙන් පසු එම පළාත ඉතා ශීඝ්‍ර සංවර්ධනයක් අත්පත් කර ගනිමින් සිටින බව ය.
එදා මූන් මහතා කියා සිටියේ උතුරේ සිදු වන සංවර්ධන ක්‍රියාවලිය ගැන තමන් බෙහෙවින් තෘප්තිමත් වන බව ය.
උතුරු වසන්තය සංවර්ධන වැඩසටහන සඳහා ශ්‍රී ලංකා රජය විශාල මුදලක් යොදවා තිබේ.
දරුස්මාන් වාර්තාව මගින් සිදු වන හානිය සමස්ත ජාතියටත්, රටටත් බලපායි.
ත්‍රස්තවාදය මර්දනය කිරීම සම්බන්ධයෙන් පමණක් නො ව ජාතීන් අතර සමගිය ගොඩනැඟීම සම්බන්ධයෙන් ද ලෝකයට පාඩමක් කියා දීමට ශ්‍රී ලංකාව සමත් විය.
ලොව පුරා රටවල් අසූ ගණනක ත්‍රස්තවාදී ගැටුම් පවතින බව අමතක නො කළ යුතු ය.
එම රටවලට ශ්‍රී ලංකාවෙන් ලබා ගත හැකි පාඩම් තිබිය දී ඊනියා වාර්තාවෙන් අපට මහත් අගෞරවයක් සිදු කර තිබේ.

Machine Extracted Summary – DN_867.txt

අසම්මත වාර්තාවට තවත් ප්‍රතිචාර !
දරුස්මාන් වාර්තාව සම්බන්ධයෙන් තවදුරටත් විවිධ මත පළ වේ.
ශ්‍රී ලංකාවේ රුසියානු තානාපති විලැදිමීර් මිහායිලෝව් මහතා කියා සිටින්නේ එය අතාර්කික වාර්තාවක් බව ය.
මේ වාර්තාවේ සංගෘහිත කරුණු මුළුමනින් ම කොටි සංවිධානයට පක්ෂපාත ලෙස ගොනු කර ඇති බව බොහෝ දෙනා පෙන්වා දෙති.
ශ්‍රී ලංකාවට ආර්ථික ආධාර සපයන කණ්ඩායමේ නායකයා වූ යසුසි අකාෂි මහතා යුද ජයග්‍රහණයෙන් පසු ද ඊට පෙර ද ශ්‍රී ලංකාවට පැමිණියේ ය.
උතුරු වසන්තය හා නැඟෙනහිර නවෝදය යන සංවර්ධන වැඩසටහන් අදටත් සාර්ථක ව ක්‍රියාත්මක වේ.
විපක්ෂ නායක රනිල් වික්‍රමසිංහ මහතා එක්සත් ජාතික පක්ෂයේ ස්ථාවරය පැහැදිලි කරමින්; මේ වාර්තාවේ ඇතැම් කරුණු සම්බන්ධයෙන් ඉතා තදින් ප්‍රතිචාර දැක්විය යුතු බව කියා තිබේ.
දරුස්මාන් වාර්තාව මගින් සිදු වන හානිය සමස්ත ජාතියටත්, රටටත් බලපායි.
ඔසාමා බින් ලාඩන්ගේ මරණය ගැනත්; ඇමෙරිකා හමුදා ප්‍රමුඛ නෙටෝ සේනාංක ඇෆ්ඝනිස්ථානය හා පාකිස්ථානය තුළ සිදු කළ මෙහෙයුම් ගැනත් විමර්ශනය කළහොත් සිදු වන්නේ කුමක් ද ? ඒ සඳහා මූන් මහතා සූදානම් ද ? ඇෆ්ඝනිස්ථානය, පාකිස්ථානය, ඉරාකය වැනි රටවල සිදු වූ මිනිස් ඝාතන ගැන විමර්ශනය කළහොත්; ශ්‍රී ලංකාවේ ත්‍රස්ත මර්දන මෙහෙයුම් වල සාධාරණත්වය හා යුක්ති සහගත බව තේරුම් ගත හැකි ය.
දරුස්මාන් වාර්තාව සකසන විට ලෝකයේ වෙනත් ත්‍රස්ත මර්දන ව්‍යාපාර ක්‍රියාත්මක වූ ආකාරය සමඟ ශ්‍රී ලංකාවේ මානුෂික මෙහෙයුම් ගැන සංසන්දනාත්මක ලෙස කරුණු විග්‍රහ කරන්නට ඉඩ තිබිණි.
ත්‍රස්තවාදය මර්දනය කිරීම සම්බන්ධයෙන් පමණක් නො ව ජාතීන් අතර සමගිය ගොඩනැඟීම සම්බන්ධයෙන් ද ලෝකයට පාඩමක් කියා දීමට ශ්‍රී ලංකාව සමත් විය.

Source Article – DN_980.txt

මල් බඳු දරුවනට නෙත් බඳු පොත්

අධ්‍යාපන විද්‍යාවේදී සමාජා'නුයෝගය නම් ශිල්පීය වචනයක් භාවිත කරනු ලැබේ. දැනුම, ආකල්ප, කුසලතා, සමාජ සාරධර්ම උකහා ගනිමින් සමාජයීය ගැටලුවලට සාර්ථකව මුහුණ දෙන ශක්තිමත් මිනිසකු බිහිකිරීම සමාජා'නුයෝගයෙහි ඉලක්කය යි. එය මිනිස් ජීවිතය මෙහෙය වනු ලබන ප්‍රධාන බලවේගයෙකි. ඉන් ප්‍රධාන කරුණු දෙකක් ඉටුවේ. පළමුවන්න; සමාජයේ සෙසු සාමාජිකයන් සමඟ සහයෝගයෙන් ජීවත්වීමට පුහුණු කිරීම ය. දෙවන්න; සමාජය වෙත ගලාඑන අභියෝග ජයගැනීම සඳහා පුද්ගලයාගේ අත්දැකීම් හා බුද්ධිය මෙහෙයැවීම ය. මේ දෙකම අධ්‍යාපනය මගින් ඉටුවේ.

අධ්‍යාපනය ගැන කියන්ම ලෙනින්ගේ ප්‍රකාශයක් අප සිහියට නැඟේ. සාර් පාලනය යටතේ පැවැති රුසියාව තුළ අධ්‍යාපනය යනු වරප්‍රසාද ලත් පන්තියේ සම්පතක් විය. ඒ අනුව එදා රුසියාවේ මුළු ජනගහනයෙන් 90%ක් පමණ අකුරු කියැවීමට නොදත්හ. සාර් පාලනය යටතේ රුසියානුවන් අන්ධකාරයේ ජීවත්වූ බවත් එම අන්ධකාරය යුරෝපයේ අන් රටවලට වඩා තම රටෙහි භයානක ලෙස පැවැති බවත් ලෙනින් පෙන්වා දුන්නේ ය. 1917 විප්ලවයෙන් පසු රුසියාවේ සියලු දෙනාට ම සාර්ථක අධ්‍යාපනයක් ලබාදෙන ක්‍රමයක් නිර්මාණය විය. ඉන් නො නැවතුණු සෝවියට් දේශය තම මිත්‍ර රටවලට ද නොමිලේ අධ්‍යාපනය ලබාදීමට ක්‍රියා කළේය.

තුන්වන ලෝකයේ රාජ්‍යයන් අතර සාර්ථක ම අධ්‍යාපන ක්‍රමවේදයක් පැවැති රට ලෙස ටැන්සානියාව හැඳින්වේ. අප්‍රිකාව පුරා නූගත්කම පැතිර යද්දී; ටැන්සානියාව අධ්‍යාපනය අතින් පෙරමුණට පැමිණියේ ය. ඊට මූලිකත්වය දුන් නායකයා වූයේ ජුලියස් නියරේරේ ය. ඔහු 1967 දී හඳුන්වා දුන් අරුෂා ප්‍රකාශනය මගින් රටේ සමස්ත අධ්‍යාපනය පරිවර්තනයකට ලක් කළේ ය. වැඩිහිටි අධ්‍යාපනයට ද අරුෂා ප්‍රකාශනය මගින් ප්‍රමුඛ තැනක් ලැබිණි. දේශපාලන නායකත්වය අධ්‍යාපනයේ බර කරට ගත යුතු ය. රටේ සංවර්ධනය, ඉදිරිගමන, නිදහස යන සියලු කරුණු තීන්දු කරන ප්‍රධාන බලවේගය අධ්‍යාපනය යැයි කිව හැකි ය.

ශ්‍රී ලංකාව යනු දකුණු ආසියාවේ හොඳම අධ්‍යාපනයක් පවතින නිදහස් රාජ්‍යයෙකි. හෝඩියේ සිට විශ්වවිද්‍යාල උපාධිය තෙක් නොමිලේ අධ්‍යාපනය ලබාදෙන එකම රාජ්‍යය ශ්‍රී ලංකාව වන්නට පුළුවන. මුළු ජාතික ආදායමෙන් සැලකියයුතු ප්‍රතිපාදනයක් අධ්‍යාපනය වෙනුවෙන් වෙන් කෙරේ. මෙරට නිදහස් අධ්‍යාපනයේ පියා ලෙස සලකනු ලබනුයේ කන්නන්ගර මැතිතුමන් ය. 1945 දී එතුමා හඳුන්වා දුන් ප්‍රතිපත්ති මෙරට ධනපති පන්තියේ ගර්භාවට ලක්විය. බූර්ෂුවා පන්තිය ඊට විරුද්ධ ලෙස කටයුතු කළ ආකාරය ඉතිහාසයේ සඳහන් වේ. එහෙත් 1956 ස්වභාෂා අධ්‍යාපන ක්‍රමය නිසා නැවතත් නිදහස් අධ්‍යාපනය ස්ථාපිත විය. අද අප පැමිණ සිටිනුයේ නිදහස් අධ්‍යාපනයේ තවත් ස්වර්ණමය යුගයකට ය.

ජාතියේ අනාගතය භාරගැනීමට නියමිත ලක්ෂ හතළිහකටත් අධික දූ – දරුවන් උදෙසා නොමිලේ නිල ඇඳුම්, නොමිලේ පෙළ පොත් ලබාදුන්නේ කවුද? ඉතා කෙටි කාලයක් තුළ පාසල් දහසක් සංවර්ධනය කිරීමට තීන්දු කළේ කවුද? වත්මන් ආණ්ඩුවට ජාතියේ අධ්‍යාපනය සම්බන්ධයෙන් පැහැදිලි සැලැසුමක් හා දැක්මක් ඇත'යි නිසැක ලෙසම කිව හැකි ය. අධ්‍යාපන ලේකම්වරුන්ගේ අභිමතයට අනුවද, අමාත්‍යවරුන්ගේ අභිමතයට අනුවද හිතුවක්කාරී ලෙස අධ්‍යාපන ප්‍රතිපත්ති වෙනස් කොට අනාගත පරපුර අසහනයට පත් කළ කාලයක් තිබිණි. සියලු විෂයයන් ඉංග්‍රීසියෙන් ඉගැන්විය යුතු යැයි එක්තරා කාලයක ගත් තීන්දුව අප මතකයට නැඟේ. එබඳු තීන්දු ගත හැක්කේ රට හා ජාතිය ගැන අල්පමාත්‍ර හෝ හැඟීමක් නැති අයටය.

ශ්‍රී ලංකාවේ වත්මන් රාජ්‍ය නායකයා දරුවන්ගේ අධ්‍යාපනය සම්බන්ධයෙන් ඉතා වුවමනාවෙන් කටයුතු කරනු දක්නට ලැබේ. ශ්‍රී ලංකාව බහුජාතික හා බහු ආගමික රාජ්‍යයක් ලෙස සලකා සියලු දරුවනට ස්වභාෂාවෙන් අධ්‍යාපනය ලැබීමටත්; ආගමික අධ්‍යාපනය හැදෑරීමටත් අවශ්‍ය පසුබිම දැන් සකස් වී තිබේ. අධ්‍යාපන අමාත්‍යවරයා මෙන් ම උසස් අධ්‍යාපන අමාත්‍යවරයා ද තම කාර්යයන් වගකීමෙන් යුතුව ඉටුකරමින් සිටිති. කලා හා සෞන්දර්යය අංශයේ උසස් අධ්‍යාපනය සඳහා ගණිතය විෂයයෙන් සමත්වීම අනිවාර්ය නොවන බවට තීන්දුවක් අධ්‍යාපන අමාත්‍යවරයා විසින් ගෙන තිබේ. එය ඉතා බුද්ධිමත් තීන්දුවක් සේ සැලැකිය හැකි ය. පොත් බෑගය සම්බන්ධයෙන් ද පාසල් වෑන් රථ ධාවනය සම්බන්ධයෙන් ද එබඳු බුද්ධිමත් තීරණ අධ්‍යාපන අමාත්‍යවරයා විසින් ගෙන තිබේ.

ශ්‍රී ජයවර්ධනපුර විශ්වවිද්‍යාලයෙහි ශාඛාවක් මාලදිවයිනෙහි පිහිටුවීමට උසස් අධ්‍යාපන අමාත්‍යවරයා තීන්දු කර තිබේ. මෙවැන්නක් මීට පෙර සිදුවී නැත. මෙරට විශ්වවිද්‍යාල අන්තර්ජාතික මට්ටමට ගෙනඒමේ දැඩි උත්සාහයක් උසස් අධ්‍යාපන අමාත්‍යවරයාට තිබේ. එතුමාගේ වැඩපිළිවෙළ නිවැරැදිය. ශ්‍රී ලංකාවේ උසස් අධ්‍යාපනය අන්තර්ජාතික මට්ටමට ගෙන යා යුතුය.

අපගේ අධ්‍යාපනය කොදෙව් මානසිකත්වයෙන් මුදාගත යුතුය යන්න මෙරට ඇතැම් දේශපාලන ව්‍යාපාර තවම තේරුම් ගෙන නැත. ඒ අනුව අධ්‍යාපනයේ පුළුල් දැක්මට එරෙහිව සටන් වදින කණ්ඩායමක් බිහිව සිටිති. අධ්‍යාපනය ඉදිරියට ගමන් කරනුයේ එම බාධාකිරීම්ද සමඟ ය.

අධ්‍යාපනයේ මූලික අරමුණ යහපත් පුරවැසියන් බිහිකිරීම ය. ඒ සමඟ රටට අවශ්‍ය මානව සම්පත නිර්මාණය කිරීම ය. මේ සාධක දෙක රටක ඉදිරිගමන හා සංවර්ධනය සමඟ අත්‍යන්තයෙන් බැඳී පවතී. දරුවනි ! හොඳින් ඉගෙනගන්න ! රටට හා ජාතියට ආඩම්බර විය හැකි දූ – පුතුන් වන්න ! හෙට දවසේ අනාගතය ඔබ අතට පත්කරමු ! වැරැදි වර්යා ඉවත් කිරීමට ද නිවැරැදි වර්යා මතුකර ගැනීමට ද අවශ්‍ය ඥානය අපේ දරුවන් වෙත පහළ වේවා ! ආණ්ඩුවේ ද අපගේ ද ප්‍රාර්ථනය එය වේ.

Human Extracted Summary – DN_980.txt

මල් බඳු දරුවනට නෙත් බඳු පොත්.
අධ්‍යාපන විද්‍යාවේදී සමාජා'නුයෝගය නම් ශිල්පීය වචනයක් භාවිත කරනු ලැබේ.
දැනුම, ආකල්ප, කුසලතා, සමාජ සාරධර්ම උකහා ගනිමින් සමාජයීය ගැටලුවලට සාර්ථකව මුහුණ දෙන ශක්තිමත් මිනිසකු බිහිකිරීම සමාජා'නුයෝගයෙහි ඉලක්කය යි.
ශ්‍රී ලංකාව යනු දකුණු ආසියාවේ හොඳම අධ්‍යාපනයක් පවතින නිදහස් රාජ්‍යයෙකි.
හෝඩියේ සිට විශ්වවිද්‍යාල උපාධිය තෙක් නොමිලේ අධ්‍යාපනය ලබාදෙන එකම රාජ්‍යය ශ්‍රී ලංකාව වන්නට පුළුවන.
ශ්‍රී ලංකාවේ වත්මන් රාජ්‍ය නායකයා දරුවන්ගේ අධ්‍යාපනය සම්බන්ධයෙන් ඉතා වුවමනාවෙන් කටයුතු කරනු දක්නට ලැබේ.
මෙරට විශ්වවිද්‍යාල අන්තර්ජාතික මට්ටමට ගෙනඒමේ දැඩි උත්සාහයක් උසස් අධ්‍යාපන අමාත්‍යවරයාට තිබේ.
අපගේ අධ්‍යාපනය කොදෙව් මානසිකත්වයෙන් මුදාගත යුතුය යන්න මෙරට ඇතැම් දේශපාලන ව්‍යාපාර තවම තේරුම් ගෙන නැත.
අධ්‍යාපනයේ මූලික අරමුණ යහපත් පුරවැසියන් බිහිකිරීම ය.
ඒ සමඟ රටට අවශ්‍ය මානව සම්පත නිර්මාණය කිරීම ය.
මේ සාධක දෙක රටක ඉදිරිගමන හා සංවර්ධනය සමඟ අත්‍යන්තයෙන් බැඳී පවතී.

Machine Extracted Summary – DN_980.txt

මල් බඳු දරුවනට නෙත් බඳු පොත්.
අධ්‍යාපන විද්‍යාවේදී සමාජා'නුයෝගය නම් ශිල්පීය වචනයක් භාවිත කරනු ලැබේ.
අධ්‍යාපනය ගැන කියත්ම ලෙනින්ගේ ප්‍රකාශයක් අප සිහියට නැඟේ.
තුන්වන ලෝකයේ රාජ්‍යයන් අතර සාර්ථක ම අධ්‍යාපන ක්‍රමවේදයක් පැවැති රට ලෙස ටැන්සානියාව හැඳින්වේ.
ශ්‍රී ලංකාව යනු දකුණු ආසියාවේ හොඳම අධ්‍යාපනයක් පවතින නිදහස් රාජ්‍යයෙකි.
ජාතියේ අනාගතය භාරගැනීමට නියමිත ලක්ෂ හතළිහකටත් අධික දූ – දරුවන් උදෙසා නොමිලේ නිල ඇඳුම්, නොමිලේ පෙළ පොත් ලබාදුන්නේ කවුද? ඉතා කෙටි කාලයක් තුළ පාසල් දහසක් සංවර්ධනය කිරීමට තීන්දු කළේ කවුද? වත්මන් ආණ්ඩුවට ජාතියේ අධ්‍යාපනය සම්බන්ධයෙන් පැහැදිලි සැලැසුමක් හා දැක්මක් ඇතැ'යි නිසැක ලෙසම කිව හැකි ය.
ශ්‍රී ලංකාවේ වත්මන් රාජ්‍ය නායකයා දරුවන්ගේ අධ්‍යාපනය සම්බන්ධයෙන් ඉතා වුවමනාවෙන් කටයුතු කරනු දක්නට ලැබේ.
ශ්‍රී ලංකාව බහුජාතික හා බහු ආගමික රාජ්‍යයක් ලෙස සලකා සියලු දරුවනට ස්වභාෂාවෙන් අධ්‍යාපනය ලැබීමටත්; ආගමික අධ්‍යාපනය හැදෑරීමටත් අවශ්‍ය පසුබිම දැන් සකස් වී තිබේ.
ශ්‍රී ජයවර්ධනපුර විශ්වවිද්‍යාලයෙහි ශාඛාවක් මාලදිවයිනෙහි පිහිටුවීමට උසස් අධ්‍යාපන අමාත්‍යවරයා තීන්දු කර තිබේ.
අපගේ අධ්‍යාපනය කොදෙව් මානසිකත්වයෙන් මුදාගත යුතුය යන්න මෙරට ඇතැම් දේශපාලන ව්‍යාපාර තවම තේරුම් ගෙන නැත.
අධ්‍යාපනයේ මූලික අරමුණ යහපත් පුරවැසියන් බිහිකිරීම ය.

Source Article – LN_144.txt

සිරගෙදර උමඟක් කපද්දී නොදැක නොදැන හිටියේ කෙසේද?
කළුතර බන්ධනාගාරයෙහි රඳවනු ලැබ සිටි කොටි සැකකරුවන් පිරිසක් ඔවුන්ගේ සිර මැදිරිය තුළ වසර තුනකට වැඩි කාලයක් තිස්සේ උමඟක් හෑරීමෙහි යෙදී ඇති බවට තොරතුරු අනාවරණය වී ඇතැයි නොබෝදා ලංකාදීපයෙන් වාර්තා විය. සිර මැදිරියක වැසිකිළි වළක් උමං කට වශයෙන් යොදා ගෙන එම කොටි සැකකරුවන් සිර මැදිරියේ සිට කළු ගඟ දෙසට මීටර පන් සියයක් පමණ හාරා තිබුණු මෙම උමඟ පසු ගිය විසි එක්වැනිදා සොයා ගන්නා ලද්දේ කොළඹින් එහි ගිය ත්‍රස්ත විමර්ශන ඒකකයේ නිලධාරීන් පිරිසක් විසින් බවද ලංකාදීප වාර්තාවෙන් ප්‍රකාශ විය. හිර ගෙදර හොර උමඟ පිළිබඳ තොරතුරු අනාවරණය වී ඇත්තේ කොළඹදී අත්අඩංගුවට ගැනුණු කොටි සැකකරුවකුගෙන් බවද ඉන් කියැවිණි. ඒ සැකකරු අදාළ නිලධාරීනට හසු නොවූයේ නම් උමඟ හෑරූ කොටින්ගේ අරමුණ සර්වප්‍රකාරයෙන්ම ඉෂ්ට විය හැකිව තිබිණැයි අවුරුදු තුනකට වැඩි කාලයක් තිස්සේ හිර ගෙදර කිසිදු නිලධාරියෙකුට හසු නොවන සේ උමඟ හාරන්නට ඔවුනට පුළුවන් වී ඇති ආකාරයෙන්ම නිගමනය කළ හැකිය.
උමං හාරා ගෙන හිර ගෙදරින් පැන යන්නට තැත් කළ සිරකරුවන් ගැන වාර්තා මීට පෙරද අසන්නට ලැබිණි. ඔවුන්ගේ ඒ ප්‍රයත්න බන්ධනාගාර නිලධාරීනට දැන ගන්නට ලැබීම නිසා ලත් තැනම ලොජ් වී ගිය බවද ඒ ඒ අවස්ථාවල වාර්තා විය. එහෙත් කළුතර බන්ධනාගාරයේ කොටි සැකකරුවන් එහි නිලධාරීනට නොදැනෙන සේ තුන් අවුරුද්දකටද වැඩි කලක් තිස්සේ සිර මැදිරියක් තුළ උමඟක් හාරන්නට සමත්වීම සැබැවින්ම පුදුමයට හේතුවකි. වාර්තා වී ඇති පරිදි කොටි සැකකරුවන් හාරන්නට පටන් ගෙන ඇත්තේ මිනිහකු දණ බඩගාගෙන යා යුතු උමඟක් නොව අඩි දහයක විෂ්කම්භයක් සහිත විශාල උමං මාර්ගයකි. ඒ සඳහා ඔවුන් විදුලිය බලයද හොරෙන් ලබා ගෙන ඇතැයි කියැවේ. මෙවැනි බරපතළ කාර්යයක් තුන් අවුරුද්දකට වැඩිකලක් සිර මැදිරියක් තුළ සිදු කර ගෙන යන්නට පුළුවන්වීමෙන් ඉදුරාම පැහැදිලි වන්නේ හිර ගෙදර අදාළ නිලධාරීන්ගේ නිසි අවධානය මේ කොටි සැකකරුවන් කෙරෙහි යොමු වී නැති බවය. උමඟ හාරා ගෙන පලා යන්නට කොටින්ට ඉඩදීම සඳහා නිලධාරීන් අහක බලාගෙන සිටියහයි කෙනකු තුළ සැකයක් ඇති වුවහොත් එය අසාධාරණයැයි කිව හැකි නොවේ.
සිර මැදිරියක් තුළ තිබෙන වැසිකිළියක් උමං කටක් වශයෙන් යොදා ගැනීමට එම වැසිකිළිය පාවිච්චියෙන් තොර විය යුතුය. ඒ වෙනුවට ඔවුන් වෙනත් වැසිකිළියක් පාවිච්චියට ගන්නට ඇති බවද නිසැකය. මිනිහකුට කිසිදු අපහසුවක් නැතිව ඇවිද ගෙන යා හැකි අන්දමේ අඩි දහයක විෂ්කම්භයෙන් යුත් උමඟක් හෑරීම සඳහා ඔවුනට අවශ්‍ය උපකරණ හා මෙවලම්ද පස් ඉවත් කිරීම සඳහා යොදා ගෙන ඇතැයි කියන ප්ලාස්ටික් බාල්දි විදුලිය බලය ලබා ගැනීමට අවශ්‍ය රැහැන් හා වෙනත් විදුලි උපකරණ හා මෙවලම් යනාදියද හිර ගෙදරින්ම ඔවුනට ලබාගත හැකි වූයේ නම් ඒ එහි කිසිවකුගේ හෝ කීප දෙනකුගේ හෝ උදව් උපකාර ඇතිවම බවට කිසිම සැකයක් නැත. ඔවුනට අවශ්‍ය වූ එවැනි දෑ බැහැරින් ඔවුන් අතට පත්වූයේ නම් එයද ඒ කොටි සැකකරුවන් පිළිබඳ වගකීම හිමි බන්ධනාගාර නිලධාරීන්ගේ අනුදැනුම හා අනුග්‍රහය නැතිව සිදුවිය හැක්කක් නොවේ. මීටර පන් සියයක් වන තෙක් උමඟ කැණීම, එම පස් පිටතට ගෙන ඒම, එම පස් දියකර සිර මැදිරියේ භූගත ජලාපවහන මං ඔස්සේ යැවීම යනාදී වශයෙන් කාර්ය බහුල වැඩ පිළිවෙළක් අවුරුදු තුනක් තිස්සේ සිර මැදිරියක් තුළ සිදු වෙද්දී එය අදාළ නිලධාරීනට ඉව නොවැටිණියි කිසිවෙක් නොපිළිගන්නවා ඇත.
කළුතර බන්ධනාගාරයෙහි රැඳවුණු මෙකී කොටි සැකකරුවන් දරුණු ගණයේ කොටි පිරිසක් බවද වාර්තා වී තිබේ. ඔවුන්ගේ මේ වැඩ පිළිවෙළෙන්ද ඒ බව සනාථ වෙයි. එවැනි පිරිසක් කෙරෙහි බන්ධනාගාර නිලධාරීන්ගේ විශේෂ අවධානයක් මෙන්ම නිරන්තර සෝදිසියක්ද යෙදෙන අන්දමේ ආරක්ෂක වැඩපිළිවෙළක් කළුතර බන්ධනාගාරයෙහි ක්‍රියාත්මක නොවූයේ නම් එය බලවත් අඩුපාඩුවකි. කොටි සැකකරුවන් බන්ධනාගාරවලින් පැන ගිය හා පැනයන්නට තැත් කළ අවස්ථා කීපයක්ම වාර්තා විය. බන්ධනාගාර දෙපාර්තමේන්තුවට එවැනි අත්දැකීම්ද තිබියදී දරුණු ගණයේ යැයි සැලැකෙන කොටි පිරිසකට මෙවැනි මහා පරිමාණ රහස් වැඩපිළිවෙළක් සඳහා ඉඩකඩ ලැබෙන තරමට ඔවුන් පිළිබඳ නිලධාරීන්ගේ අවධානය හීනවූයේ නම් එය සැබැවින්ම ඉතා බරපතළ ඌනතාවකි. මුළු බන්ධනාගාරයම පුපුරුවා හැරීම සඳහා කටයුතු යොදාගන්නටද එවැනි ඌනතාවකින් ඔවුනට ඉඩකඩ සැලැසෙන්නට පිළිවන.
කොටි සැකකරුවන් රඳවනු ලැබ සිටි සිරමැදිරි ඒ කාලය තුළ බන්ධනාගාර නිලධාරීන්ගේ නිසි පරීක්ෂාවට ලක්වී නැති බව පෙනී යතැයිද උමඟ හෑරීමේ කටයුත්තට ඇතැම් නිලධාරීන්ගේද සහායක් ලැබුණේද යන්න ගැන සැකයක් මතුවී ඇතැයිද බන්ධනාගාර කොමසාරිස් ජනරාල් වජිර විජේගුණවර්ධන ප්‍රකාශ කළහයි වාර්තා වී තිබේ. සිද්ධිය පිළිබඳ සොයා බැලීම සඳහා විශේෂ පරීක්ෂණ මණ්ඩලයක් පත් කළ බවද ඒ මහතා ප්‍රකාශ කර ඇත. සිර මැදිරිය තුළ මෙතරම් විශාල උමඟක් කැණීම ඇතැම් බන්ධනාගාර නිලධාරීන්ගේ අනුදැනුම හා අනුග්‍රහය නැතිව කිසිසේත්ම කළ හැක්කක් නොවේය යන සැකය මේ සිද්ධිය පිළිබඳ තොරතුරු කියවන කවරම කෙනකු තුළ වුවද නිතැතින්ම ඇතිවීම ස්වාභාවික සිද්ධියකි. නොවැළැක්විය හැක්කකි.
එහෙයින් ඒ සිද්ධිය පිළිබඳ සොයා බැලීම සඳහා විශේෂ පරීක්ෂණ මණ්ඩලයක් පත් කිරීම කීප අතකින් වැදගත් වෙයි. උමඟ කැණීම සඳහා බන්ධනාගාර නිලධාරීන් කිසිවෙකුගේ සහාය කොටි සැකකරුවන්ට ලැබී නැති බව ඒ පරීක්ෂණයේදී සියලු සැකයෙන් තොරව සනාථ වුවහොත් බන්ධනාගාර නිලධාරීන් වෙත එල්ල වන සැකය තුරන් වෙයි. කිසියම් නිලධාරියකුගේ සහාය කොටි සැකකරුවන්ට ලැබී ඇතැයි සනාථ වුවහොත් ඔවුනට එරෙහිව නෛතික පියවර ගැනීමේ මඟ පෑදෙයි. කොටි සැකකරුවන් කෙරෙහි නිලධාරීන්ගේ අවධානය ප්‍රමාණවත් නොවීම නිසා කොටි සැකකරුවන්ට මේ උමං කැණීමට ඉඩකඩ සැලසිණැයි පෙනී ගිය හොත් කොටි සැකකරුවන් රඳවනු ලැබ සිටින සියලු බන්ධනාගාරවල නිලධාරීනට එයින් පාඩමක් උගත හැකිවෙයි.

Human Extracted Summary – LN_144.txt

සිරගෙදර උමහක් කපද්දී නොදැක නොදැන හිටියේ කෙසේද?
කළුතර බන්ධනාගාරයෙහි රඳවනු ලැබ සිටි කොටි සැකකරුවන් පිරිසක් ඔවුන්ගේ සිර මැදිරිය තුළ වසර තුනකට වැඩි කාලයක් තිස්සේ උමහක් හෑරීමෙහි යෙදී ඇති බවට තොරතුරු අනාවරණය වී ඇතැයි නොබෝදා ලංකාදීපයෙන් වාර්තා විය.
හිර ගෙදර හොර උමඟ පිළිබඳ තොරතුරු අනාවරණය වී ඇත්තේ කොළඹදී අත්අඩංගුවට ගැනුණු කොටි සැකකරුවකුගෙන් බවද ඉන් කියැවිණි.
උමං හාරා ගෙන හිර ගෙදරින් පැන යන්නට තැත් කළ සිරකරුවන් ගැන වාර්තා මීට පෙරද අසන්නට ලැබිණි.
එහෙත් කළුතර බන්ධනාගාරයේ කොටි සැකකරුවන් එහි නිලධාරීනට නොදැනෙන සේ තුන් අවුරුද්දකටද වැඩි කලක් තිස්සේ සිර මැදිරියක් තුළ උමහක් හාරන්නට සමත්වීම සැබැවින්ම පුදුමයට හේතුවකි.
වාර්තා වී ඇති පරිදි කොටි සැකකරුවන් හාරන්නට පටන් ගෙන ඇත්තේ මිනිහකු දණ බඩගාගෙන යා යුතු උමහක් නොව අඩි දහයක විෂ්කම්භයක් සහිත විශාල උමං මාර්ගයකි.
මීටර පන් සියයක් වන තෙක් උමඟ කැණීම, එම පස් පිටතට ගෙන ඒම, එම පස් දියකර සිර මැදිරියේ භූගත ජලාපවහන මං ඔස්සේ යැවීම යනාදී වශයෙන් කාර්ය බහුල වැඩ පිළිවෙළක් අවුරුදු තුනක් තිස්සේ සිර මැදිරියක් තුළ සිදු වෙද්දී එය අදාළ නිලධාරීනට ඉව නොවැටිණියි කිසිවෙක් නොපිළිගන්නවා ඇත.
කොටි සැකකරුවන් රඳවනු ලැබ සිටි සිරමැදිරි ඒ කාලය තුළ බන්ධනාගාර නිලධාරීන්ගේ නිසි පරීක්ෂාවට ලක්වී නැති බව පෙනී යතැයිද උමඟ හෑරීමේ කටයුත්තට ඇතැම් නිලධාරීන්ගේද සහායක් ලැබුණේද යන්න ගැන සැකයක් මතුවී ඇතැයිද බන්ධනාගාර කොමසාරිස් ජනරාල් වජිර විජේගුණවර්ධන ප්‍රකාශ කළහයි වාර්තා වී තිබේ.
එහෙයින් ඒ සිද්ධිය පිළිබඳ සොයා බැලීම සඳහා විශේෂ පරීක්ෂණ මණ්ඩලයක් පත් කිරීම කීප අතකින් වැදගත් වෙයි.
උමඟ කැණීම සඳහා බන්ධනාගාර නිලධාරීන් කිසිවෙකුගේ සහාය කොටි සැකකරුවන්ට ලැබී නැති බව ඒ පරීක්ෂණයේදී සියලු සැකයෙන් තොරව සනාථ වුවහොත් බන්ධනාගාර නිලධාරීන් වෙත එල්ල වන සැකය තුරන් වෙයි.
කිසියම් නිලධාරියකුගේ සහාය කොටි සැකකරුවන්ට ලැබී ඇතැයි සනාථ වුවහොත් ඔවුනට එරෙහිව නෛතික පියවර ගැනීමේ මඟ පෑදෙයි.

Machine Extracted Summary – LN_144txt

සිරගෙදර උමහක් කපද්දී නොදැක නොදැන හිටියේ කෙසේද?
කළුතර බන්ධනාගාරයෙහි රඳවනු ලැබ සිටි කොටි සැකකරුවන් පිරිසක් ඔවුන්ගේ සිර මැදිරිය තුළ වසර තුනකට වැඩි කාලයක් තිස්සේ උමහක් හෑරීමෙහි යෙදී ඇති බවට තොරතුරු අනාවරණය වී ඇතැයි නොබෝදා ලංකාදීපයෙන් වාර්තා විය.
සිර මැදිරියක වැසිකිළි වළක් උමං කට වශයෙන් යොදා ගෙන එම කොටි සැකකරුවන් සිර මැදිරියේ සිට කළු ගඟ දෙසට මීටර පන් සියයක් පමණ හාරා තිබුණු මෙම උමඟ පසු ගිය විසි එක්වැනිදා සොයා ගන්නා ලද්දේ කොළඹින් එහි ගිය ත්‍රස්ත විමර්ශන ඒකකයේ නිලධාරීන් පිරිසක් විසින් බවද ලංකාදීප වාර්තාවෙන් ප්‍රකාශ විය.
ඒ සැකකරු අදාළ නිලධාරීනට හසු නොවූයේ නම් උමඟ හෑරූ කොටින්ගේ අරමුණ සර්වප්‍රකාරයෙන්ම ඉෂ්ට විය හැකිව තිබිණැයි අවුරුදු තුනකට වැඩි කාලයක් තිස්සේ හිර ගෙදර කිසිදු නිලධාරියෙකුට හසු නොවන සේ උමඟ හාරන්නට ඔවුනට පුළුවන් වී ඇති ආකාරයෙන්ම නිගමනය කළ හැකිය.
උමං හාරා ගෙන හිර ගෙදරින් පැන යන්නට තැත් කළ සිරකරුවන් ගැන වාර්තා මීට පෙරද අසන්නට ලැබිණි.
එහෙත් කළුතර බන්ධනාගාරයේ කොටි සැකකරුවන් එහි නිලධාරීනට නොදැනෙන සේ තුන් අවුරුද්දකටද වැඩි කලක් තිස්සේ සිර මැදිරියක් තුළ උමහක් හාරන්නට සමත්වීම සැබැවින්ම පුදුමයට හේතුවකි.
සිර මැදිරියක් තුළ තිබෙන වැසිකිළියක් උමං කටක් වශයෙන් යොදා ගැනීමට එම වැසිකිළිය පාවිච්චියෙන් තොර විය යුතුය.
කළුතර බන්ධනාගාරයෙහි රඳවුණු මෙකී කොටි සැකකරුවන් දරුණු ගණයේ කොටි පිරිසක් බවද වාර්තා වී තිබේ.
කොටි සැකකරුවන් රඳවනු ලැබ සිටි සිරමැදිරි ඒ කාලය තුළ බන්ධනාගාර නිලධාරීන්ගේ නිසි පරීක්ෂාවට ලක්වී නැති බව පෙනී යතැයිද උමඟ හෑරීමේ කටයුත්තට ඇතැම් නිලධාරීන්ගේද සහායක් ලැබුණේද යන්න ගැන සැකයක් මතුවී ඇතැයිද බන්ධනාගාර කොමසාරිස් ජනරාල් වජිර විජේගුණවර්ධන ප්‍රකාශ කළහයි වාර්තා වී තිබේ.
එහෙයින් ඒ සිද්ධිය පිළිබඳ සොයා බැලීම සඳහා විශේෂ පරීක්ෂණ මණ්ඩලයක් පත් කිරීම කීප අතකින් වැදගත් වෙයි.
උමඟ කැණීම සඳහා බන්ධනාගාර නිලධාරීන් කිසිවෙකුගේ සහාය කොටි සැකකරුවන්ට ලැබී නැති බව ඒ පරීක්ෂණයේදී සියලු සැකයෙන් තොරව සනාථ වුවහොත් බන්ධනාගාර නිලධාරීන් වෙත එල්ල වන සැකය තුරන් වෙයි.

Appendix B – Parameter Tuning Data

Best 100 Possible Combinations for α, β, γ and δ for the equation
$W(s) = \alpha K(s) + \beta T(s) + \gamma L(s) + \delta P(s)$

Case	α	β	γ	δ	F-Score
1	8	47	45	0	47
2	7	48	45	0	47
3	6	48	46	0	47
4	51	24	25	0	47
5	5	49	46	0	47
6	49	25	26	0	47
7	47	26	27	0	47
8	45	27	28	0	47
9	4	49	47	0	47
10	37	31	32	0	47
11	36	33	31	0	47
12	35	33	32	0	47
13	35	32	33	0	47
14	34	34	32	0	47
15	33	34	33	0	47
16	33	33	34	0	47
17	32	35	33	0	47
18	31	35	34	0	47
19	31	34	35	0	47
20	30	36	34	0	47
21	3	50	47	0	47
22	29	36	35	0	47
23	28	37	35	0	47
24	27	37	36	0	47
25	26	38	36	0	47
26	25	38	37	0	47
27	24	39	37	0	47
28	23	39	38	0	47
29	22	40	38	0	47
30	21	40	39	0	47
31	20	41	39	0	47
32	19	41	40	0	47
33	18	42	40	0	47
34	17	42	41	0	47
35	16	43	41	0	47
36	15	43	42	0	47
37	14	44	42	0	47
38	13	44	43	0	47
39	12	45	43	0	47
40	10	46	44	0	47
41	1	51	48	0	47
42	9	47	44	0	46.9
43	9	46	45	0	46.9
44	7	47	46	0	46.9
45	53	23	24	0	46.9
46	5	48	47	0	46.9
47	43	28	29	0	46.9
48	41	30	29	0	46.9
49	41	29	30	0	46.9
50	39	31	30	0	46.9
51	39	30	31	0	46.9
52	37	32	31	0	46.9
53	3	49	48	0	46.9
54	29	35	36	0	46.9
55	27	36	37	0	46.9
56	25	37	38	0	46.9
57	23	38	39	0	46.9
58	21	39	40	0	46.9
59	2	50	48	0	46.9
60	19	40	41	0	46.9
61	17	41	42	0	46.9
62	15	42	43	0	46.9
63	13	43	44	0	46.9
64	11	46	43	0	46.9
65	11	45	44	0	46.9
66	1	50	49	0	46.9
67	19	45	36	0	46.8
68	10	50	40	0	46.8
69	1	55	44	0	46.8
70	9	48	43	0	46.8
71	9	45	46	0	46.8
72	8	48	44	0	46.8
73	8	45	47	0	46.8
74	7	49	44	0	46.8
75	7	46	47	0	46.8
76	6	50	44	0	46.8
77	6	49	45	0	46.8
78	6	46	48	0	46.8
79	55	22	23	0	46.8
80	5	50	45	0	46.8
81	5	47	48	0	46.8
82	48	26	26	0	46.8
83	46	27	27	0	46.8
84	46	26	28	0	46.8
85	44	28	28	0	46.8
86	44	27	29	0	46.8
87	43	29	28	0	46.8
88	42	29	29	0	46.8
89	42	28	30	0	46.8
90	40	30	30	0	46.8
91	40	29	31	0	46.8
92	4	51	45	0	46.8
93	4	50	46	0	46.8
94	4	47	49	0	46.8
95	38	31	31	0	46.8
96	36	32	32	0	46.8
97	34	33	33	0	46.8
98	34	32	34	0	46.8
99	32	34	34	0	46.8
100	32	33	35	0	46.8

Worst 100 Possible Combinations for α, β, γ and δ for the equation
$W(s) = \alpha K(s) + \beta T(s) + \gamma L(s) + \delta P(s)$

Case	α	β	γ	δ	F-Score
1	0	0	0	100	32
2	100	0	0	0	33.4
3	85	2	0	13	33.6
4	84	2	0	14	33.7
5	83	3	0	14	33.7
6	81	2	0	17	33.7
7	80	2	0	18	33.7
8	79	4	0	17	33.7
9	79	2	0	19	33.7
10	78	2	0	20	33.7
11	96	0	0	4	33.7
12	77	3	0	20	33.7
13	76	3	0	21	33.7
14	75	3	0	22	33.7
15	74	3	0	23	33.7
16	73	3	0	24	33.7
17	72	3	0	25	33.7
18	71	3	0	26	33.7
19	70	3	0	27	33.7
20	69	3	0	28	33.7
21	68	3	0	29	33.7
22	67	3	0	30	33.7
23	92	0	0	8	33.8
24	87	2	0	11	33.8
25	84	3	0	13	33.8
26	80	4	0	16	33.8
27	80	3	0	17	33.8
28	79	3	0	18	33.8
29	78	3	0	19	33.8
30	76	4	0	20	33.8
31	75	4	0	21	33.8
32	72	4	0	24	33.8
33	71	4	0	25	33.8
34	70	4	0	26	33.8
35	69	4	0	27	33.8
36	68	4	0	28	33.8
37	67	4	0	29	33.8
38	66	4	0	30	33.8
39	66	3	0	31	33.8
40	65	4	0	31	33.8
41	65	3	0	32	33.8
42	64	4	0	32	33.8
43	64	3	0	33	33.8
44	63	4	0	33	33.8
45	63	3	0	34	33.8
46	62	4	0	34	33.8
47	0	9	0	91	33.8
48	0	8	0	92	33.8
49	0	7	0	93	33.8
50	0	6	0	94	33.8
51	0	5	0	95	33.8
52	0	4	0	96	33.8
53	0	3	0	97	33.8
54	0	2	0	98	33.8
55	0	16	0	84	33.8
56	0	15	0	85	33.8
57	0	14	0	86	33.8
58	0	13	0	87	33.8
59	0	12	0	88	33.8
60	0	11	0	89	33.8
61	0	10	0	90	33.8
62	0	1	0	99	33.8
63	91	0	0	9	33.8
64	87	0	0	13	33.8
65	86	2	0	12	33.8
66	86	0	0	14	33.8
67	85	0	0	15	33.8
68	84	0	0	16	33.8
69	83	2	0	15	33.8
70	83	0	0	17	33.8
71	82	3	0	15	33.8
72	82	2	0	16	33.8
73	82	0	0	18	33.8
74	81	3	0	16	33.8
75	78	4	0	18	33.8
76	77	4	0	19	33.8
77	77	2	0	21	33.8
78	76	2	0	22	33.8
79	75	5	0	20	33.8
80	75	2	0	23	33.8
81	74	5	0	21	33.8
82	74	2	0	24	33.8
83	73	2	0	25	33.8
84	72	2	0	26	33.8
85	95	0	0	5	33.9
86	94	0	0	6	33.9
87	93	0	0	7	33.9
88	89	2	0	9	33.9
89	86	1	0	13	33.9
90	76	5	0	19	33.9
91	92	1	0	7	33.9
92	24	1	0	75	33.9
93	23	1	0	76	33.9
94	22	1	0	77	33.9
95	21	1	0	78	33.9
96	20	1	0	79	33.9
97	19	1	0	80	33.9
98	0	21	0	79	33.9
99	0	20	0	80	33.9
100	0	19	0	81	33.9