S	
E1	
E2	
For Office Use Only	

Masters Project Final Report
(MCS)
2019

Project Title	Automatic Text Summarization for Sinhala
Student Name	O.S Wimalasuriya
Registration No. & Index No.	2016/MCS/116 16441165
Supervisor's Name	Dr. Ruvan Weerasinghe

For Office Use ONLY

Automatic Text Summarization for Sinhala

A dissertation submitted for the Degree of Master of Computer Science

O.S Wimalasuriya
University of Colombo School of Computing
2019

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 acknowledged in the text.

Student Name: Oshitha Sewmal Wimalasuriya

Registration Number:2016/MCS/116

Index Number: 16441165

Signature ..... Date .....

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

Certified by

Supervisor Name: Dr. Ruwan Weerasinghe

Signature: ..... Date: .....

Abstract

According to the people life style, people are surrounded with vast amounts of information albeit with less and less time or ability to make sense of it. Automatic summarization first started as early as in the 1950s.In the modern world due to lack of the time, generating an accurate and intelligent summary for a long document or text pieces has become a popular research as well as an industry problem. This research is carried out to find the suitable approaches to address the above mention issue with minimum linguistic resources.

This research proposes a solution for summarizing Sinhala text by identifying the most important and relevant sentences based on linguistic and statistical features of a given text, using an unsupervised extractive summarization approach. In order to generate a better summary, keyword and sentence extraction is manipulated by using a graph based TextRank algorithm.

The proposed method was evaluated by comparing the machine generated summaries and human generated summaries based on the assumption that human generated summaries are perfect. The critical evaluation was done by using ROUGE-n and F1 Score to ensure the proposed method usability, performance and efficiency. According to the ROUGE-n values it gives more than 60% of recall rate and more than 42% of precision rate. Further, this research provides a benchmark for future research on Sinhala automatic text summarization.

Acknowledgement

I would not have been possible to complete this project successfully, without the kind support, assistance and courage of many people. I would like to thanks to all of them.

I would like to thank to my supervisor Dr. Ruwan Weerasinghe, Senior Lecturer of University of Colombo School of Computing.

I would like to thanks for everyone who participating online surveys, interview and contributing their ideas for critical evaluation.

Finally, I would like to thank my parents and family for their support and encouragement through the duration of the project.

Table of Content

List of Figures

List of Tables

List of Abbreviations

Abbreviations	Definition
CRF	Conditional Random Fields
HMM	Hidden Markov model
LTRL	Language Technology Research Laboratory
ME	Maximum Entropy
NER	Name Entity Recognition
NLP	Natural Language Processing
POS	Part-Of-Speech
SVM	Support Vector Machine
UCSC	University of Colombo School of Computing

Chapter 1 – Introduction

1.1 Chapter Overview

The purpose of this chapter is to provide an overview of the Sinhala Text Summarization. It commences with a brief introduction along with the current limitations faced by Sinhala Text Summarization. Furthermore, this chapter provides the aim, objectives, scope and feature of proposed application. Additionally, previous works have been done is evaluated. Also proposed project plan is included, which will be followed throughout the entire project. Eventually, concludes with an overview of the chapters to follow.

1.2 Problem Domain

With today's immense and rapid growth in the use of smartphones, tablets and electronic document readers, there is a well-known need to develop tools to summarize and visualize documentary content. [17] The summaries are a process in which the most important essential information of the source text is obtained by presenting the information so that the people can understand the idea in a shorter time.

Memory works in a mysterious way. Cognitive scientist Dan Willingham gave a good explanation about memory and how it works. "Since you cannot save everything, your memory system registers your bets: if you think carefully and repeatedly something, you'll probably have to think again, so you should save it." If you do not think of something so fast, you probably do not want to think about it anymore, so it does not have to be stored, your memory is a product that you think about more carefully, and what students think more carefully is what they will remember."[3]

In this new era, it is of utmost importance to provide an improved mechanism to extract the most important information and data from documents or the Internet in the most efficient and fastest manner. Sometimes it is difficult to extract the summary of a large text document for people. To solve the problem, the automatic text summary is needed.

Sinhala is one of the native and national languages that is used mainly in Sri Lanka. Although Sinhala is the national language of Sri Lanka, today there are still a handful of applications and websites that use Sinhalese as the main language. Although the studies are done to summarize the

text in English, there are no attempts at the Sinhala language. This research focuses on the problem of summarizing text in the Sinhalese language.

With the growing amount of data on the Internet, such as websites, news, e-books, publications and the like, one of the biggest challenges for all is access to reliable and accurate data. However, with hectic schedules of people and the large amount of data available, there is a need for abstraction or summary of information. [7]
In a well-summarized text, it is easy to understand the most important concepts and the essential information of the original document. Sinhala is the main language of Sri Lanka and more than 19 million people use it. Although it is used by many people, Sinhala has done very little research in computational linguistics. [4]

Although the text summary is available for languages such as English, Punjabi, Hindi and etc., [21],[14] There have been limited studies to carry out the text summary in the Sinhalese language and it is one of the biggest problems facing the Sinhala community on a daily basis. To address the above problem, this research is done for the Sinhalese language.

1.3 Problem Definition

From the above problem domain, it is derived that there is no easy way for Sinhala text summarization.

1.4 Significance

This is one of the research to do the text summarization in Sinhala Language. In order to the summarization, huge amount of effort should be put because the less availability of the resources for Sinhala Language. If the desired accuracy level is achieved, it'll be a landmark in natural language processing in Sinhala language.

1.5 Project Aim

The aim of the project is to summarize a long text in Sinhala language by identifying the most important sentences.

1.6 Project Objectives

In Order to achieve the project aim, following project objectives have been identified.

The major objective of this research project is to find the most suitable approach to summarize the Sinhala text. When considering Sinhala language, there are limited number of NLP resources are available. One of the goal in this project is to apply appropriate techniques, methods and architecture to find the adaptability of them. Because Sinhala language has limited number of resources than other rich languages (English, Spanish etc.). The challenging part of this project is to find an approach which doesn't need many linguistic resources in order to achieve accuracy and performance. There were limited number of previous attempts are available in the Sinhala text summarizer. Also those studies have been different constraints.
Finding out how to identifying the most important and relevant sentences based on linguistic and statistical features of the text and how key word and sentence extraction is manipulated, those are the supplementary objectives of this research.

In future there will be enough linguistics resources for Sinhala language. Also new researches will be able find the advanced techniques, method and architecture for Sinhala language. It will be more accurate and more efficient.

However, those research will be based on previous research and previous results were able to compare with the newest ones. Also it will help to improve the new approach. When considering further goal of this project, this research project is provided a new method for Sinhala text summarization. It will be help for future researches on Sinhala language.

1.7 Scope

Sinhala sentences can have may different structures and more complex sentence structures. Also Sinhala has specific type of policies for written grammar and talking grammar. But writers can be used their own policies. Hence, the sentence structures of an arbitrary text would be more complex to analyze using programmatically especially without some linguistic resources such as taggers, parsers and other tools. To overcome this issues, this research used national newspapers articles. It was assumed that those articles were written by the professionals. Those articles have more than 20 sentences in length which is more suitable for the proposed design of the research. There are few type of articles was selected which is too length, shot, medium and Sinhala and English mixed articles. 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 Text Summarization. Hence, the scope of this research was limited to apply latest technologies

applicable to low resourced languages, to find out the most suitable factors for achieving accurate summaries automatically.

1.8 Features of Prototype

1. Summarize a long document/text/paragraph in Sinhala
2. Identify the important sentences in the source text and sort based on the important level
3. Identify the relevance between the sentences and the keywords

1.9 Project Deliverables

1. Literature Review
2. Interim report
3. Evaluation

1.10 Outline of Chapters to Follow

The next chapter of the thesis will discuss the literary feature of the problem domain. Also it will have discussed approaches, techniques and technologies of the particular area. Also norms of summarization will be explained with different approaches and resources used over the past few years. Methodology chapter will be discussed resources used to carry out the research. Also it will be covered the high level design, low-level design and model the proposed application by using different design methods. Testing and evaluation chapter will provide the appropriate test plan details, function and non- functional testing, module testing details. Also evaluation will be done by using quantitative and qualitative methods. Conclusion chapter will discuss the problem and challenges face during the project, the learning out comes, limitation of the project and the future enhancements. References which were used to carry out the research will be listed at the end of the thesis.

Chapter 2 - Literature Review

2.1 Chapter Overview

Previous chapter contained the problem domain and problem definition of the application. Also project aim, objectives, scope and project features were discussed.

This chapter will examine on the current literature available on generating text summarization from extractive text summarization using natural language processing approaches for Sinhala Language. Identifying and evaluating the existing technologies, methods which are used in previous works and determining the most appropriate approach for the solution is the purpose of the literature review. Also, the limitation, advantages and disadvantages of the current approaches have been discussed.

2.3 Automatic Text summarization Overview

In the modern world due to lack of the time, generating accurate and intelligent summaries for a long document or text pieces has become a popular research as well as an industry problem. This has increased within the last few years because of the growth of the internet and people are overwhelmed by the huge amount of online documents and the information. [2]

Due to this massive development of the data and information, top notch universities like National Centre for Text Mining Manchester University, Aarhus University-Denmark, etc. have been continuously working for its improvements. [12]

According to researches [13] a summary is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text and usually, significantly less than that". By prevailing the key concepts and the information, producing a good and precise summary is called automatic text summarization. [2] Automatic text summarization has been developed and applied in numerous domains in recent years. As an example, summarization techniques are used in search engines to preview the summary of the particular web site, generate news headlines based on the content and etc.

Automatic text summarization is very difficult and challenging because when a human summarizes a text, the normal procedure would be reading the whole text or the document and writing down the key concepts out of it to develop the understanding.

Even though the quality of the human generated summary might be good, it would take a quite a bit of time. [7] Since the computer lacks the human skills like thinking and language capabilities summarization would be a difficult task [2]

To do the automatic text summarization there are three key aspects of research which are delineated by the definition.

1. Summaries which are produced from a single document or multiple documents
2. Summaries should preserve important information
3. Summaries should be short [7]

Mainly there are two fundamental approaches for automatic text summarization which are abstractive and extractive. Extractive approach extract phrases and words from the original text to create the summary while the abstractive approach learn the language representation to generate summaries more like human generated using linguistics methods. [12]

2.4 Abstractive Text Summarization

The generation of a short summary composed of a few sentences or a title capturing the main idea of the text of the document is called abstract abstract text. Usually, abstraction is performed by mapping an input word sequence into a source document with a target word sequence. [13] Compared to extractive synthesis, abstract synthesis is an effective way to create accurate summaries of information because it retrieves information from multiple documents. [12] An abstract abstract displays summarized data in an easily readable and grammatically correct manner. As indicated below, abstract synthesis can be divided into two main parts, such as the structure-based approach and the semantic approach.

2.4.1 Structure Based Approach

Through psychological schemes such as models, ontology, lead and body, rule extraction and alternative structure such as tree and graph structure, this structure can still be categorized.

1. Tree based method

   Depending on the context of the document or text, this technique creates a dependency tree. This used a language generator or an associated degree algorithm to create the outline of the document.

By using a local multi-sequence alignment from bottom to top, this technique should be able to identify common information sentences. As you enter multiple documents or texts and process these entries, the central theme is identified and, once the theme is finalized, using text grouping algorithms, the ranking of sentences is complete.

2. Template based method

   Template based method is used for guidance to represent the document or the text. To map the text to guide slots linguistics patterns extraction rules are used here. In this technique a novel approach called “GISTEXTER” is presented and it will identify the topic related information from the input document and then translate it into the database. After that based on user requests the sentences are added from the database.

3. Ontology based method

   Representing its own ontology, each domain has its own information structure. Most documents available on the Internet are connected to a domain. One of the main models to describe this method is "Fuzzy". In fuzzy ontology, the fuzzy inference phase generates degrees of membership. For each fuzzy idea, various ontology events are associated.

4. Lead and body phase method

   This method relies heavily on the functioning of sentences such as substitution and insertion. By identifying common phrases in the body and lead, followed by insertion and substitution of sentences, this method is developed.

5. Rule based method

   In terms of listing of aspects and classes the document will be summarized in this technique. By selecting the most effective crucial part among the generated data, this data extraction rule answer one or lot of aspects of a category. One approach is to find semantically related nouns and verbs from the suggested information. Another approach is doing the abstractive text summarization using word graph, discourse rules and syntactical constraints. Based on different aspects such as keywords, syntactic constraints and the input sentences, the sentence reduction step is done. Doing text summarization based on random forest classification and feature score, is another approach for abstractive text summarization. In this technique, to pre-process the data it will calculate the feature score by training the cross-validation classifier.

This classifier will decide whether the sentence belongs to the summary or not.

Based on the minimum redundancy and maximum relevance all the selected sentences are generated.

6. Graph based method

   To represent the text of the language, many will use the graphical data structure called "Opinosis-Graph". In this method, each node represents a unit of words representing the structure. For this method, there are several approaches. One approach is to evaluate the text using average score, edge count, median aggregation, average ranking, and so on. To generate the results, a new weighted consensus scheme is proposed. Another approach is to generate an abstract abstract from the appropriate subgraph encoding and using high redundancy scores. Path notation and meaningful sentences are important components of this approach. Jaccard method used to eliminate duplicate paths and all paths are ranked in descending order.

2.4.2 Semantic Based approach

In the semantic approach, all data is introduced into the natural language generation system (NLG). Nominal expressions and verbal expressions are mainly identified by treating linguistic data in this approach. The diagram below represents an overview of the semantic approach.

1. Multimodal semantic method

In this approach using a chosen measurement all the important ideas and the relationship among the ideas are captured and rated using a score. Then the by ordering the chosen ideas the summary will be created.

2. Information item-based method

Instead of sentences from the supplied documents, in this method, summary is generated from the abstract representation of the supplied documents.

For a better summary identifying all text entities, attributes, predicate between them and predicate characteristics are important.

3. Semantic graph method

For the source text, the generation of a semantic graph called Enhanced Semantic Graph (RSG) is used in this approach to perform the synthesis. Abstract final abstract will be generated from the reduced semantic chart.

4. Semantic text representation model

Rather than analyzing the syntax or the structure of the text, in this approach context selecting is done by ranking the most significant predicate argument. Using a language generation tool final summary will be generated.

Due to the lack of a generalized framework, abstractive text summarization would be a major issue because parsing and alignment of a parse tree is difficult. [12] Extracting important information and doing the sentence ordering is still an open issue in abstractive text summarization. Furthermore, paraphrase and reformulation, compression involving lexical substitution is also difficult with abstraction summarization. Most of the times extractive summarization gives better results compared to the automatic abstractive summarization. [2] This is because when it comes to data driven approaches like sentence extraction, abstractive text summarizations techniques will face problems such as inference, semantic representation and natural language generation which are relatively harder.

2.5 Extractive Text Summarization

The main approach in extractive text summarization is to select the most important sentences or words from a given text and then compose a summary [10] In this approach, first give an important score to each and every word in the source text and then sort the sentences and words from that given important score. All the words and sentences are directly copied from the source resource. In this method, it won't be able to generate some new words which do not appear in the source text. Extractive summarization can be mainly divided into three categories.

- To express the main aspects of the text, construct the intermediate representation of the source text
- Based on the representation score the sentences
- Select a summary from the best scored sentences [2]

2.5.1 Intermediate Representation

By using intermediate representation, every summarizer intends to find the source text salient content mainly based on this representation. Topic representation and indicator representation are the main approaches which are used in this representation.

Topic representation interpret the text which is discussed in the source text by using topic word approaches, Bayesian topic models, latent semantic analysis and etc. Indicator representation annotated every sentence as a list of features based on having certain phrases, sentence length and etc.

2.5.2 Sentence Score

After generating the intermediate representation, assign score to each sentence based on the level of importance to the source text. For topic representation, the score is generated by calculating how well the main topics are explained by the generated topic by aggregating the evidence from different indicators.

2.5.3 Select the sentences for summary

There are many approaches to selecting the best phrases and producing the summary such as K's most important algorithm, greedy algorithms and others. and etc. Below table shows a comparison between the abstractive and extractive approaches.

Extractive approach	**Abstractive approach**
Create the summary from phrases or sentences in the source document	Express the ideas in the source document using different words
Widely studied for years	Limited researches are available
Fast with less resources	Slow with a lot of resources such as CPU and GPU
Can be simply applied to many statistical languages	Hard to be implemented

Not suitable for summarizing highly redundant text	Language dependent

Table 1- Comparison between Extractive approach and abstractive approach

2.6 Extractive Text Summarization Technique

Following text extractive text summarization methods are used in today.

2.6.1 Text summarization in Genism

This module implements the TextRank, an unsupervised algorithm which is based on weighted graph. [11] This module is built on popular page ranker algorithm that Google is using for ranking the pages. This module consists of five steps.

1. Preprocessing the word – removing stop word and stemming
2. Generate a graph for sentence and vertices
3. Connect every sentence by using the edges
4. Execute the page rank algorithm
5. According to the page rank score select most suitable sentence

2.6.2 PyTeaser

PyTeaser is heuristic method for extractive text summarization which is based on python and Scala project called TextTeaser. TextTeaser assigns a score for every sentence and this is a linear combination of features that are extracted from that particular sentence. It has a title feature to count the words which are common to the document. There is another module to calculate the ideal summary length and sentence position module to normalize the sentence number. After removing the stop words there is another feature to calculate the key word frequency.

2.6.3 LexRank

LexRank is an unsupervised graph-based method which use IDF-modified Cosine to measure the similarity between two sentences.

This similarity is used as weight of the graph edge between two sentences. LexRank would be able to select the top sentences for the summary that are not too similar to each other.

2.7 Previous work

Automated Text Summarization for Sinhala

In this research project was used classical approach for automated text summarization. It attempted to identify the most salient information of an article using some thematic features. This research was identified those features for the Sinhala language. Also it helps to archive accurate summaries. A part from that that research proposes a best possible liner combination of identified features. [21]

Extractive text summarization in Hindi

In this research article they have proposed two ways to extract information in Hindi language. Linguistic extracted from a simplified argumentative structure of the text and statistical based on the frequency are those two approaches. When processing the text in Hindi they have broken down the text into three categories such as preprocessing, processing and post processing. Word level features are sorted by using frequency-based approach, length of the word and occurrence in the heading of articles. Length of the sentence, position, presence of verb in sentence, similarity to headline of article, referring pronouns and cohesions similarity score of the sentence are considered to calculate the sentence level features. [20]

Automatic Punjabi text extractive summarization system

The Punjabi text summary has been developed based on the statistical and linguistic characteristics of the text. It has two main components, such as roughing and processing.

The preprocessing phase includes identifying Punjabi sentence borders, identifying word borders, eliminating paused words, elector for proper names with nouns and eliminating double sentences. Using the weight equation of the function, the score of each sentence is calculated in the processing phase.

The identification function of the following Punjabi line, the title identification function and the number identification function were also considered in this phase. Finally, the best-ranked sentences are arranged according to the specified score. [5]

Text Summarization for Bengali Documentation

The proposed solution for the Bengali text summary consists of several steps. Pre-processing language text together with tokenization is the first step.

During this step, the Bengali text will consist of tinted and derived words and sentences, followed by a final word discrimination process. Subsequently, the main key phrases of the source text are identified by word analysis and the sentence analysis process. The process of word analysis is done by identifying the frequency of the words, the identification of the numerical value, the distance of the repeated words and the analysis of the keywords. The sum of frequent words, the length of the sentence, the position of the sentence, the analysis of identical sentences and the identification of imitation sentences are some of the techniques to generate a score for the sentences.

In the final process, the most important identified sentences are re-evaluated with the help of global agreements, final analysis and sentence classification. [1]

Following table shows the comparison between the existing solutions.

Research	Summary
Automated Text Summarization for Sinhala	• Using classical approach. • It attempted to identify the most salient information of an article using some thematic features.
Extractive text summarization in Hindi	• Using extractive summarization approach to create the summary • Supports only Hindi language • Supports only Hindi language • Application consisting modules such as preprocessing, processing and post processing
Automatic Punjabi text summarization system	• supports only extractive approach • The system has component such as preprocessing and processing • Feature weight algorithm is used to generate a score for each sentence
Text summarization for Bengali	• Support only for Bengali language • Extractive approach is used to summarize the text • Scoring methods are implemented based on different association and linguistic rules

Table 2-Comparison of existing solutions

By analyzing the previous works, the author was able to identified the pros and cons of each system. Furthermore, it was identified that how the relevant algorithms are selected and how the systems are implemented. Researches carried out in Hindi and Punjabi language is more similar to the proposed project. Therefore, the techniques used in pre-processing are taken into the consideration and thoroughly evaluated.

2.8 Text Rank Algorithms

Within the citation analysis, social networks and world wide web link structure, graph-based analysis like Google's PageRank algorithm and Kleinberg's HITS algorithm made a huge success. By using a graph based algorithm able to decide the important of a vertex within a graph by extracting the information from the entire paragraph. [11]

This technique could be applied for generating an extractive summary, extracting important keywords from the given paragraph and for the word disambiguation process as well.

The basic idea behind the TextRank model is "Voting" and "Recommendation". When one vertex links to another vertex its defined as "Voting". If a particular vertex obtains a higher number of votes it would become a more important.

Where d is the damping factor which can vary between 0 and 1. The factor is normally set to 0.85. After assigning an arbitrary value for each node in the graph, the computation iterates until convergence below a given threshold is achieved. [11]

In this model the graph is built from the natural language text and it includes multiple or partial links between the vertices that are extracted from the text. TextRank algorithm has the following main steps in the procedure.

1. Identify the text units using tokenization and added as vertices in the graph.
2. Identify the relations between the vertices in the graph and draw edges between the vertices.
3. Until convergence ,iterate the graph-based Text Rank algorithm.
4. Based on the final achieved score, sort the vertices

2.8.1 Keyword extraction using TextRank

Automatically identifying the best keywords which describe the content in the document is called keyword extraction. In this keyword extraction method, first the given text is tokenized into words.

Then removing the stop words process is carried out. Next all the lexical units are passed for identifying the lemmas for a given word.

Finally, all the simplified texts are added to the undirected and unweighted graph. Each vertex of the graph represents one or more lexical unit from the original source. [11]

Any relation between the two lexical units is called connection (edge). After creating the graph each vertex is initiated with an initial value of 1.0. The algorithm is run for about 20-30 iterations until it converges at a given threshold of 0.0001.

After receiving the final score for each node, vertices are sorted. The top keywords are selected upon the length of the original document. If the original document has a larger length, more keywords are selected and if the original document has a lesser length, less keywords are selected.

2.8.2 Sentence Extraction using TextRank

Sentence extraction process is more similar to the keyword extraction process, because both processes are trying to identify the most representative key phrases for the given text.

In this method, for each sentence in the source text, a dedicated vertex is added to the weighted graph. Based on the similarity between the two sentences, the score is calculated. [11]

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

Where Si and Sj are the given two sentences and words that appear in the sentence are Si = w1i, w2i, w3i... wnii. In the weighted graph, the assigned score for each node indicates the strength of the connections established between various sentences in the source text. After the algorithm is executed, the top sentences are selected based on the final generated score for each one.

2.8.3 TextRank usage in other languages

As stated by the Liu [10] they have used an enhanced TextRank algorithm to do the text summarization in Tibetan language.

$$S(v) = d + (1-d) \sum_{u \in B(v)} \frac{S(u)}{\sum_{j \in F(u)} W_{uj}}$$

Where the S(v) represents the weight of each node, B(v) is a node set which point to node V. “d” is the damping factor which vary between 0 to 1 and F(u) is node set which is pointed by node v. [10] The following figure displays the basic theory of “vote” and “recommend”. When there is a link between point A and point B, then the point A is recommended to point B. If the point B gets more votes it is considered that the point B is more important.

To extract the keywords from the given Tibetan text, they have used TF/IDF algorithm. In this method each word is assigned with a weight and according to the weight, top ranked words are selected. Following steps are executed for generating an automatic abstract in Tibetan language.

1. Extract the most important top 3 keywords by comparing the important of the sorted keywords

2. Identify the sentences which contain above selected keywords. If a sentence is found which includes all the above three selected keywords that sentence is selected as the abstract.
If not, search for sentences containing top 2 candidate keywords. Otherwise, search for sentences which contain first number one keyword.

2.9 Sinhala WordNet

A large lexical database which is developed for a specific language is called WordNet. All the nouns, adjectives, adverbs, verbs synonyms are grouped together in a WordNet. By defining the conceptual semantic relation and lexical relation, synonyms are interlinked within the WordNet. By extracting the common words in Sinhala from a corpus and getting the expert opinions, there is a WordNet built for the Sinhala language by relating the merge approach. [22]

A novel Sinhala WordNet has been introduced recently. This Sinhala WordNet is based on English (Princeton) WordNet. In this approach, the research is carried out using the Hindi WordNet. [24] Developing a fully functional and completed Sinhala WordNet could be considered as a landmark of NLP approaches in Sinhala in such cases like information retrieval systems, Sinhala text summarization systems, Sinhala text classifiers and Sinhala translators. This WordNet observed the words which are subtle but there is an important difference between the written words and spoken form of the words. For some words, form of the gender is not specified in Sinhala and commonly genders are masculine and feminine. The gender of the noun is used to decide the morphological form of a verb. [24]

The Sinhala language word formation can be divided into three categories such as native words, words which came from another language without any modification (තත්සම - tatsama) and words which came from another language with modification (තද්භව - tatbawa). Mainly the inherited words inherit from English, Pali, Hindi, Tamil and Portuguese languages. The origin of the words should be considered when constructing phrases in Sinhala.

There are nine morphological formations for a verb in Sinhala called 'vibhakthi' (විභක්ති). Additionally, there are forming compound words called 'sandi'(සන්ධි) and 'samasa' (සමාස). The root of the word is the base for formation of these forms. Therefore, when storing the words in the WordNet, along with the word root, most common morphological form is also stored. As a resource for aiding language processing task, building a word net for Sinhala is very important and it requires significant expert knowledge and resource allocation. [22]

2.10 Named Entity Recognition in Sinhala

Identifying named entities is one of the major subtasks when it comes to natural language processing. Thus, it is a very difficult task to build a NER for an Indic language like Sinhala because it lacks the features like capitalization [9] Named entity doesn't have any agreed definition, yet it is explained according the occasion it is used in. Basically, person names, countries, cities, organization names, street names and etc. are considered as named entities.

There are many techniques which are implemented in English language for NER, but unfortunately those cannot be applied for Sinhala language. Ambiguities about the words, free word-order, lack of capitalization, lack of resources and agglutinative nature are some of the reasons behind this. NER approaches can be categorized into main three groups. Rule based method, statistical or data-driven method and hybrid methods are those three classifications. Accurate linguistics knowledge and thorough analysis is needed for rule-based approach and an annotated corpus is needed for statistical approach. But the hybrid approach is using best points from both methods.

According to the researches [9] corpus is the main backbone for any data-driven technique. At that time there were no NE tagged corpus for Sinhala language which can be used directly. For the Indic language like Sinhala there are data-driven approaches like Hidden Markov model (HMM), Maximum Entropy (ME), Support Vector Machine (SVM) and Conditional Random Fields (CRF). For lot of Indic languages CRF and ME approaches work more precisely.

For the Sinhala language there is a POS tagged corpus which was created by Language Technology Research Laboratory (LTRL) of University of Colombo School of Computing (UCSC). It has about 75000 Sinhala tagged words which are collected from different articles. There are no features that can be identified directly or clearly for the Sinhala language. [9] In this research paper there are three types of features considered such as context word feature, word suffix feature and language dependent feature. Altogether, this approach is to create the Maximum Entropy model. When testing the Conditional Random Fields additionally Bi-gram feature has taken into consideration.

Researches has stated [8] that the statistical modelling is the best for applying in the NER for the Sinhala language. Bi-gram and suffix information are the best when detecting the feature set in Sinhala. Finally, the research stated that CRF has outperformed the Maximum Entropy model.

2.11 Chapter summary

The chapter examined several topics relevant in text summarization to identify several key findings. The study commences with text summarization approaches. There are two main approaches for text summarization. Those two main approaches are compared. Also every approaches/ techniques are compared and discussed.

Chapter 3 – Methodology

This chapter examines the methodology to carried out the research on Sinhala text summarization. It is intended to generate the abstract as an abstract instead of abstract due to some limitations of past infrastructure resources. It is required natural language processing technique which require linguistic resources.

Since the most popular approach for text summarization is extractive approach., the author intends to use that in this research with the help of text rank algorithm for word sentences scoring.

3.1 Tokenization

Tokenization is chopping the text into pieces and removing certain characters such as punctuations and special characters. There are two types of tokenization such as sentence and word tokenization.

3.2 Stop word removal

In language there are high frequency words which are used to complete grammatical rules and the tone. According to the researches those words are not relevant and those words do not have impact to generating a summary.

According to the Language Technology Research Laboratory (LTRL), there are 425 stop words. LTRL has identified some of if the Sinhala stop words such as ඔව, ම, නිසා, දී, අනුව, ඉතා. For this research this list is used to filter stop words.

3.3 Stemming

Stemming is the required process of NLP. It will be reduced the inflected form of a word to its stem. Lot of NLP applications which use words as its basic elements. It's vey efficient and light weight approach compared to morphological parsing. According to the research there are some advanced stemmers for English language which do not work Sinhala. Sinhala is a highly inflectional language. It has many word forms to denote a single concept. According to the previous research, there are two types of algorithms.

3.3.1Knowledge-Based Approach

This list consists of word-stem pairs were used as a lookup.

3.3.2 Data-Driven Approach

A simple lightweight algorithm is used to identify the stems of a list of words. The list of words extracted from the Sinhala News Editorials Corpus. This list is sorted by alphabetically order.

3.4 Named entity recognition

In NLP, named entity recognition is a major concern as well as a primary task which is executed in preliminary stages. There are two popular data driven ways of identifying name entities for Sinhala such as Maximum entropy model and Conditional random fields. This process is directly applied for name entity recognition.

3.5 Text Rank Algorithm

The basic idea behind the Text Rank Algorithm is "Voting" and "Recommendation". If one vertex links to another vertex its defined as "Voting". The score of a vertex (Vi) is defined as follows. [11]

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

d is a damping factor which can be vary between 0 and 1. This factor is normally set in to 0.85. Each node is assigned is assigned with a value and after that iterate until convergence. The convergence is calculated, the stability of each node in the graph represents the center of the word related to the node.

Following procedures are applied for Sinhala text.

1. Identify the Sinhala text which define the task.
2. Identify the relation between Sinhala texts.
3. Iterate graph base ranking algorithm, until convergence.
4. Sort the vertices base on the values.

3.6 Keyword Extraction

The best keywords which describe the content in the document is called the keywords. According to the keyword extraction process, first tokenize the Sinhala text. Afterwards remove the stop word from the text and stemming the Sinhala texts.

Eventually, all the process texts are added to the undirected weighted graph. Each vertex of the graph represents one or more lexical units. Edges are representing as relations between vertices. After received the final value for each node, vertices are sorted.

3.7 Sentence Extraction

It's basically similar to the keyword extraction process. According to this method, for each Sinhala text, a dedicated vertex is added to the weighted graph. Based on the similarity between sentences, value is calculated.

Following algorithm is used for similarity calculations. More details are examined in chapter -2.

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

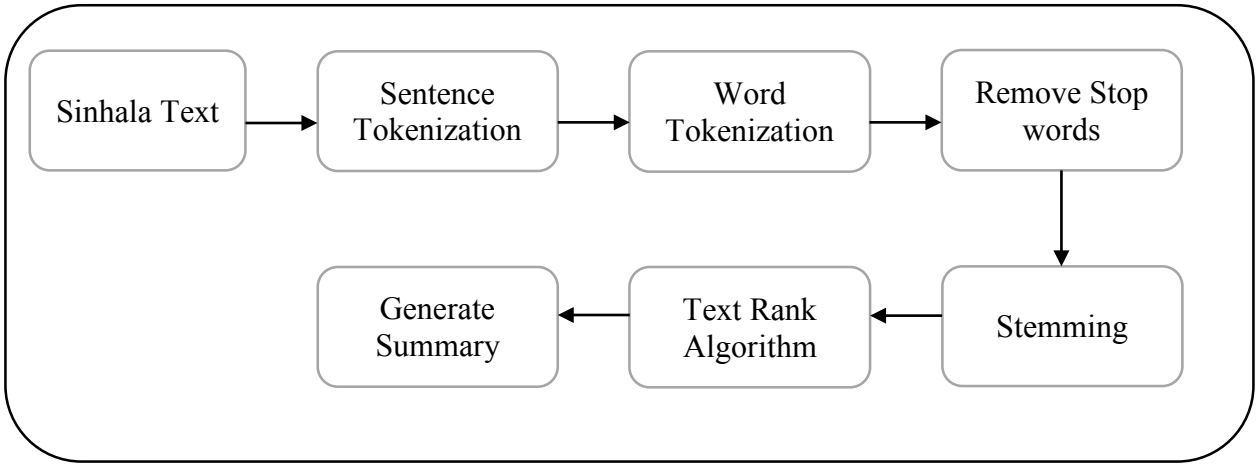

Figure 1- Proposed solution

Chapter 04 - Testing and Evaluation

This chapter will critically evaluate the implemented system in different aspects such as technical, usability and concept wise. Furthermore, this chapter thoroughly discuss about the evaluation criteria, evaluation methodologies and selection of the evaluators. At the end of the chapter self-evaluation of the project is carried out.

4.1 Objectives and Goals

The main objectives are:

1. Verify and validate functional requirements.
2. Verify and validate non-functional requirements.
3. Identify errors and fixed those errors.
4. Future enhancements

4.2 Testing Criteria

1. Functional Quality

   Functional quality will verify whether the system has satisfied the functional requirements. Furthermore, it will minimize the defects and maximize the ease of use.

2. Non-Functional Quality

   This manly focus on efficiency, reliability, durability etc.

4.3 Functional Testing

Functional requirements were testing by using functional testing. It was very important because it provides an entire impression of the implementation The test results are shown in Table-3

Functional Requirement	Test Status
Enter small Sinhala text	Tested
Enter medium Sinhala text	Tested
Enter long Sinhala text	Tested
Sinhala text with special symbols	Tested
Create summarized text	Tested

Table 3- Tested Functional Requirements

4.4 Non-Functional Testing

According to the System Requirement Specification (Non-functional requirements), set of non-functional requirements were defined. Those functional requirements reflect the quality of the project.

4.4.1 Accuracy and Performance testing

Accuracy is a major non-functional requirement. The generated summary should be more accurate. Text summarization has two standard evaluation methods:

1. Extrinsic Method

   Using recall rate, accurate rate and F-measure
2. Intrinsic Method

   Measure the quality of the abstract by using generated abstract complete specific tasks.

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially of a set of metrics for evaluating automatic summarization of texts as well as machine translation. It works by comparing an automatically produced summary or translation against a set of reference summaries.

Precision and recall in the context of ROUGE is calculated as follows ,

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$$

There are different type of ROUGE measurement available. (ROUGE-N, ROUGE-L and ROUGE-S)

Bigrams, unigrams, trigrams and higher order of n-gram overlaps are measured using ROUGE-N measurement. Using LCS, ROGUE-L measures the longest matching sequence of words. LCS matches the sequence that reflects sentence level word order. ROUGE-S measures any pair of word in a sentence in order. This is also called skip-gram co-occurrence.

4.4.2 Scalability

It's used to measure whether the system could functional well, when a large data set is inserted.

4.5 Module Testing

Module	Sentence/ Word Tokenizer
Input	Sinhala text
Expected Result	Tokenized text
Actual Result	Tokenized text
Status	Pass

Table 4- Sentence tokenizing module test

Module	Stop word remove
Input	Sinhala text with stopwords
Expected Result	Text without stopwords
Actual Result	Text without stopwords
Status	Pass

Table 5- Stopword module test

Module	Stemming
Input	Sinhala text
Expected Result	Stemming text
Actual Result	Stemming text
Status	Pass

Table 6- Steaming module test

4.6 Evaluation Methodology

Evaluation process provides the feedback for the entire project. The evaluation is done by using both qualitative and quantitative methods. Evaluation process carried through the questionnaire and interview. According to the time constraint more priority is given to the questionnaire.

4.6.1 Evaluation by External Evaluators

External Evaluation is done by using questionnaire. The questionnaire is focused on several areas such as the concept of the solution, used technology and tools, usability, drawbacks and future enhancements. According to the external evaluators ranking 52% of people are more like machine generated summaries.

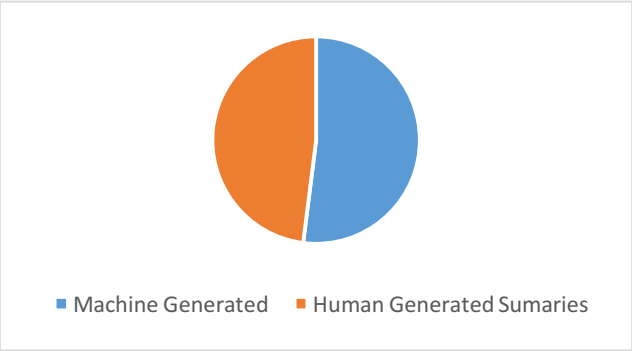

Figure 2- Ranking percentage

4.6.2 Quantitative Evaluation

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.

Precision = |{relevant instances} ∩ {retrieved instances}| / |{retrieved instances }|

Precision - It is the number of correct positive results divided by the number of positive results predicted by the classifier.

Recall = |{relevant instances} ∩ {retrieval instances}| / |{ relevant instances}|

Recall - It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).According to the above equations, if it attempts to increase the recall rate by retrieving more instances, it will cause to decrease the Precision rate. Also vice versa Will be occurred. To get the maximum values for both Precision and recall, *F-Score* is calculated. F-Score reaches its best value at 1 and worst score at 0.

F-Score = 2 (Precision * Recall) / (Precision + Recall)

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

		rouge-l	rouge-2	rouge-3
p		0.4444444444444444	0.3738738738738739	0.4444444444444444
r		0.9620253164556962	0.9431818181818182	0.9620253164556962
f		0.490899194465383	0.5354838669019772	0.60799999567712
p		0.385	0.9871794871794872	0.41872042321482716
r		0.28321678321678323	0.8901098901098901	0.4297082191493643
f		0.385	0.9871794871794872	0.5539568304953161
p		0.17223650385604114	0.7613636363636364	0.17897882220516836
r		0.10481099656357389	0.5754716981132075	0.1773255787887034
f		0.17480719794344474	0.7727272727272727	0.2851153009742055
p		0.24825174825174826	0.9466666666666667	0.26062366612749255
r		0.17298578199052134	0.8021978021978022	0.2846003869451189
f		0.24825174825174826	0.9466666666666667	0.3933517972621451
p		0.4430379746835443	0.9859154929577465	0.4882010997564946
r		0.36231884057971014	0.9036144578313253	0.5172413752244948
f		0.4430379746835443	0.9859154929577465	0.6113537075120612

Table 7-Rouge Values

ROUGE-1 refers to overlap of unigrams between the system summary and reference summary. ROUGE-2 refers to the overlap of bigrams between the system and reference summaries.

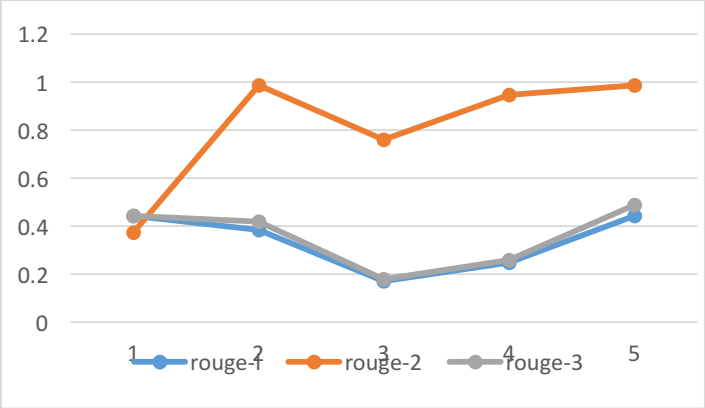

Figure 3- Precisions rate

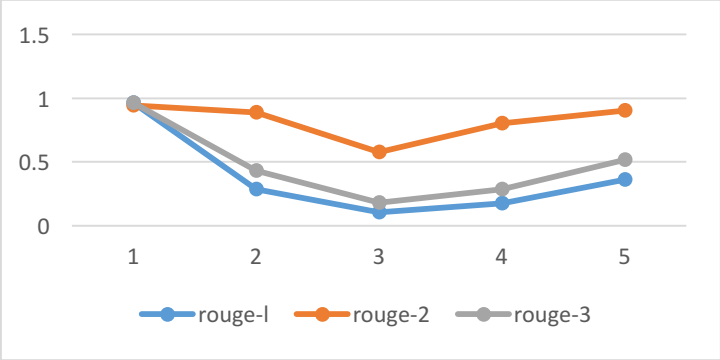

Figure 4- Recall rate

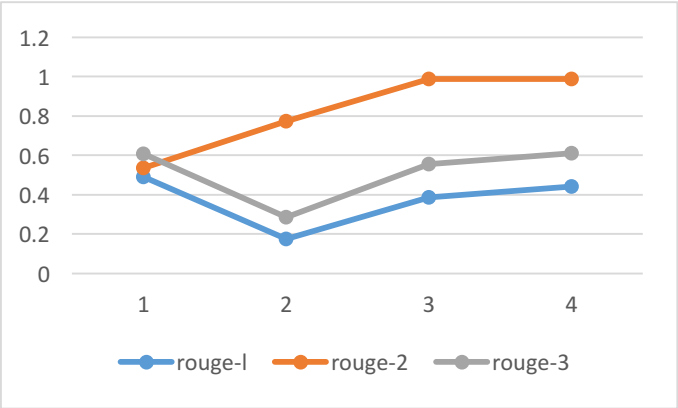

Figure 5-F-Score Values

Interpretation –

ROUGE-n recall=96% means that 96% of the n-grams in the reference summary are also present in the generated summary.

ROUGE-n precision=44% means that 44% of the n-grams in the generated summary are also present in the reference summary.

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially of a set of metrics for evaluating automatic summarization. It works by comparing an automatically produced summaries and human produced summaries

F1 Score is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. F1 Score tries to find the balance between precision and recall.

Chapter 5 – Conclusion

The previous chapter was discussed the testing and evaluation of the research application.
This chapter will conclude on the design, development and evaluation of the prototype. Also it will discuss the problem and challenges faced during the project development stage. The learning outcomes and limitation of the proposed method and future enhancements will discuss. Finally, this chapter concludes with a reflection on the project.

5.1 Problems and Challenges Faced During the Project

During the project implementation phase and entire project there were several problems and challenges which were faced. The most critical problem is Time Management, because time is very limited constraint. Also there were several problems that were faced in during the implementation stage. Those problems were resolved by through self-study and research. The Natural Language Processing was very deep and wide area. According to the problem, chose relevant parts and solve the problem. There are limited resources and researches carried out for Sinhala language in NLP. With the help of previous researches grain the domain knowledge and methods for sinhala linguistic text processing.

5.2 Learning Outcomes

This project improves the technical and soft skills. According to the soft skills, problem solving, time management, critical analysis, creativity, innovative and work under pressure were improved. When considering the technical skills, grain the knowledge of natural language processing, text summarization and Sinhala language in natural language processing.

5.3 Future Enhancement

This research was carried out based on the classical approaches used in automatic text summarization. Also this research was carried out with minimum available linguistic resources for Sinhala Language.

ID	Future Enhancement	Description	Priority
FE1	Improve the text rank algorithm.	Current research is carried with using base Text Rank algorithm. Text Rank algorithm could be improved by tuning the relevant parameters.	High
FE2	Try with abstractive approach	Current research is carried out by using extractive approach, due to the time and resources constraints. Abstractive approach will create text summary more simulator to human generated one.	High
FE3	Try with another algorithm	There is other algorithms such as Lex Rank. Those algorithms could fit to the extractive Sinhala text summarization.	Medium

Table 8-Future enhancements

5.4 Contribution and Conclusion

There were many researches carried out for other languages such as English, Hindi, French etc. There were few research attempts recorded in the Sinhala text summarization. This research was carried out to find the suitable approach for automatically summarizing Sinhala texts with minimum linguistic resources. The research was carried out on text rank algorithm base approach.

Evaluation is the most important part. Also it is crucial to evaluate one summary is better than other summary. Researches have been researched past few years to find out most suitable and accurate ways to evaluate summaries. Once machine was generated summary, human need to involved and evaluated the summary. This process is very expensive. Hence, this research is based on a hypothesis which is the human generated summaries are perfect. But this hypothesis is not true at all, because summaries are subjective and also it's depend on the text.

If two summaries of same source, will be made with two different humans, those summaries will not be identical. Sometimes human made an abstract summary. But human also generated average defined length of summaries. According to the test result, it was given many different F-score values (Precession, Recall). With high precision but low recall, hence classifier is extremely accurate. Precision can be thought of as a measure of a classifiers exactness.
Recall can be thought of as a measure of a classifiers completeness. If recall values closer to 1.0, (94%,89%,57%80%,90%, etc - test data) According to the test scenario it gives more than 60 % average. it's really good for a text summarization system.

However, it does not tell the other side of the story. A machine generated summary can be extremely long, capturing all words in the reference summary. But, much of the words in the system summary may be useless, making the summary unnecessarily verbose. This is where precision comes into play. In terms of precision, what you are essentially measuring is, how much of the system summary was in fact relevant or needed?

The precision here tells us that out of all the system summary bigrams, there is a 37%,98%,76%,94%,98%, etc. (testing data) Precision is also gives more than 60% average. overlap with the reference summary. This is not too bad either. Note that as the summaries (both system and reference summaries) get longer and longer, there will be fewer overlapping bigrams especially in the case of abstractive summarization where human made summaries are not directly re-using sentences for summarization.

The precision aspect becomes really crucial when you are trying to generate summaries that are concise in nature. Therefore, it is always best to compute both the Precision and Recall and then report the F-Measure.

This research is a landmark in natural language processing for Sinhala Language. Research can be further enhanced and developed with the growth of Sinhala linguistic resources. Using the new methods and technologies, the text summarization can be taken into a next level in near future.

6. References

[1] Abujar, S., Hasan, M. and Hossain, S.A. (2017). A Heuristic Approach of Text Summarization for Bengali Documentation.

[2] Allahyari, M., Trippe, E.D. and Gutierrez, J.B. (no date). Text Summarization Techniques : A Brief Survey(1)

[3] "Articles", *Daniel Willingham--Science & Education*, 2018. [Online]. Available: http://www.danielwillingham.com/articles.html. [Accessed: 22- Jul- 2018].

[4] Arukgoda, J. et al. (2015). A Word Sense Disambiguation Technique for Sinhala. *Proceedings - 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, ICAIET 2014*,207–211.

[5] Gupta, V. and Singh, G. (2012). Automatic Punjabi Text Extractive Summarization System. *Proceedings of 24th International Conference on Computational Linguistics*, 2 (December 2012), 191–198.

[6] I.Wijesiri, B. Gunathilaka,D.C Wimalasuriya.R.Paranavithana, M. Gallage, M.Lakjeewa, G. Dias and N. de Silva, "Building a WordNet for Sinhala" GWC 2014 *Proc. 7th Glob. Wordnet Conf.*, no.June 2014.

[7] Hingu, D., Shah, D. and Udmale, S.S. (2017). Automatic Text Summarization of Wikipedia Articles23–29.

[8] K.Tldx and D.F Lu, "An Overview on Extractive Text Sumarization" pp. 54-62

[9] J.K. Dahanyaka, A.R.W. (2014). Named Entity recognition for sinhala language1–6.

[10] Liu, J. (2017). A Multi-Level Encoder for Text Summarization0–5.

[11] Mihalcea, R. and Tarau, P. (1998). TextRank: Bringing Order into Texts 85 .

[12] Moratanch, N. (2016). A Survey on Abstractive Text Summarization.

[13] Nallapati, R. et al. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. Available from http://arxiv.org/abs/1602.06023

[14] *Nlp.lacasahassel.net*, 2018. [Online]. Available: http://nlp.lacasahassel.net/publications/hasselthesis04lic.pdf. [Accessed: 22- Jul- 2018].

[15] Perera, Y. et al. (2017). Intelligent Mobile Assistant for Hearing Impairers to Interact With the Society in Sinhala Language.

[16] P. Mehta, "From Extractive to Abstractive Summarization: A Journey," *Proceedings of the ACL 2016 Student Research Workshop*, 2016.

[17] Tldx, K. and Lu, D.F. (2017). An Overview on Extractive Text Summarization54–62.

[18] Van Rijsbergen, C. J. (1979). Information Retrieval, 2nd edition. Dept. of Computer Science, University of Glasgow.

[19] V.Gupta and G.Singh, “Automative Punjab Text Extractive Sumarixation System”, *Proc. 24th Int. Conf. Comput.Linguist., vol.2,* no. December 2012,pp. 191-198, 20212.

[20] Vijay, S. et al. (2017). Extractive Text Summarization in Hindi318–321.

[21] V. Welgama, "Automatic Text Summarization for Sinhala", Degree of Master of Philosophy, University of Colombo School of Computing, 2012.

[22] Welgama, V. et al. (2011). Towards a Sinhala Wordnet. Preceedings of Conference on Human Language Technology for Development, HHLTD 2011,(May), 39–43.

[23] 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.

[24] Wijesiri, I. et al. (2014). Building a WordNet for Sinhala. GWC 2014: Proceedings of the 7th Global Wordnet Conference,(June) .

Appendix A –

Samples of Source Article, Human Extracted Summaries and Machine Extracted Summaries

Source Text - 1

සේනා දළඹුවන් මර්දනය කළ හැකි බව කියන විශේෂ වෛරස් කාණ්ඩයක සාම්පල් විදෙස් රටකින් මෙරටට ආනයනය කර පරීක්ෂණ පැවැත්වීමට බලාපොරොත්තු වන බව උද්‍යාන භෝග පර්යේෂණ හා සංවර්ධනය කිරීමේ ආයතනයේ කීට විද්‍යා අංශයේ ප්‍රධානී එස්.එස්.වැලිගමගේ මහතා පවසයි. ඔහු වැඩිදුරටත් ප්‍රකාශ කලේ බැක්ටීරියාවන් සහ ක්ෂුද්‍ර ජීවීන්ද භාවිත කරමින් දළඹු උවදුර පාලනය කිරීමට මේ වනවිටත් පරීක්ෂණ සිදුකරන බවයි. විදේශයකින් මෙරටට ගෙනා බඩඉරිඟු ඇට තොගයක තිබූ වෙනත් බීජ වර්ග කිහිපයක් තවාන් දැමීමෙන් පසු සේනා දළඹුවාට සමානකම් ඇති වෙනත් දළඹු විශේෂයක් එම තවානේ සිටියදී හමුවූ පුවතක් මාරවිල ප්‍රදේශයෙන් වාර්තා විය. හලාවත ,මාරවිල ප්‍රදේශයේ ගොවිපලකට ඉකුත් සතියේදී මෙම බඩඉරිඟු ඇට තොගය ගෙන්වා තිබූ අතර එය පරීක්ෂා කිරීමේදී බඩඉරිඟු ඇට සමඟ බඩඉරිඟු ශාක කොටස් කිහිපයක් ද වෙනත් බීජ වර්ග කිහිපයක්ද හමු වී ඇති අතර ඔහු එම වෙනත් බීජ වර්ග කිහිපය ද ගෙන බඩ ඉරිඟුවලින් වෙන්කොට වගා කර තිබේ. මින් එක් බීජ වර්ගයක් පැළ වී ඇති අතර අනෙක් වර්ගය පැළ නොවීම නිසා ඒ පිළිබඳ සොයා බැලීමට උරය තුළ ඇති පස් ඉවත් කළ විට එහි බීජ නොතිබී ඇති අතර ඒ වෙනුවට දළඹුවකු දක්නට ලැබුනු බවයි ගොවිපලේ හිමිකරු පැවසුවේය. පසුව ගොවිපලේ හිමිකරු විසින් මේ සම්බන්ධයෙන් අදාල ආයතනවලට දැනුම් දීමක්ද සිදුකර තිබුනු අතර ඒ අනුව අදාල බලධාරින් පැමිණ එහි සාම්පල පරීක්ෂණ සඳහා රැගෙන ගොස් තිබේ. මේ අතර අනුරාධපුර , නොච්චියාගම , දොඹවලගම ප්‍රදේශයේ වී වගාව සේනා දළඹුවාට සමාන දළඹු විශේෂයක් ආක්‍රමණය කර ඇති බවට ඊයේ(29) අද දෙරණ වාර්තා කිරීමෙන් පසු එම වගාව නිරීක්ෂණය කිරීමට මහවැලි එච්. කලාපයේ කෘෂිකර්ම නිලධාරින් පිරිසක් පැමිණියේය. මේ අතර සේනා දළඹුවන් මර්දනය කළ හැකි බව කියන විශේෂ වෛරස් කාණ්ඩයක් විදේශ රටකින් මෙරටට ආනයනය කර තිබේ."

Human Extracted summary - 1

සේනා දළඹුවන් මර්දනය කළ හැකි බව කියන විශේෂ වෛරස් කාණ්ඩයක සාම්පල් විදෙස් රටකින් මෙරටට ආනයනය කර පරීක්ෂණ පැවැත්වීමට බලාපොරොත්තු වන බව උද්‍යාන භෝග පර්යේෂණ හා සංවර්ධනය කිරීමේ ආයතනයේ කීට විද්‍යා අංශයේ ප්‍රධානී එස්.එස්.වැලිගමගේ මහතා පවසයි. එසේම බැක්ටීරියාවන් සහ ක්ෂුද්‍ර ජීවීන්ද භාවිත කරමින් දළඹු උවදුර පාලනය කිරීමට මේ වනවිටත් පරීක්ෂණ සිදු කරයි. මේ අතර අනුරාධපුර , නොච්චියාගම , දොඹවලගම ප්‍රදේශයේ වී වගාව සේනා දළඹුවාට සමාන දළඹු විශේෂයක් ආක්‍රමණය කර ඇති බවට ඊයේ(29) අද දෙරණ වාර්තා කිරීමෙන් පසු එම වගාව නිරීක්ෂණය කිරීමට මහවැලි එච්. කලාපයේ කෘෂිකර්ම නිලධාරින් පිරිසක් පැමිණියේය."

Machine Extracted summary - 1

සේනා දළඹුවන් මර්දනය කළ හැකි බව කියන විශේෂ වෛරස් කාණ්ඩයක සාම්පල් විදෙස් රටකින් මෙරටට ආනයනය කර පරීක්ෂණ පැවැත්වීමට බලාපොරොත්තු වන බව උද්‍යාන භෝග පර්යේෂණ හා සංවර්ධනය කිරීමේ ආයතනයේ කීට විද්‍යා අංශයේ ප්‍රධානී එස්.එස්.වැලිගමගේ මහතා පවසයි. මේ අතර සේනා දළඹුවන් මර්දනය කළ හැකි බව කියන විශේෂ වෛරස් කාණ්ඩයක් විදේශ රටකින් මෙරටට ආනයනය කර තිබේ. හලාවත ,මාරවිල ප්‍රදේශයේ ගොවිපලකට ඉකුත් සතියේදී මෙම බඩඉරිඟු ඇට තොගය ගෙන්වා තිබූ අතර එය පරීක්ෂා කිරීමේදී බඩඉරිඟු ඇට සමඟ බඩඉරිඟු ශාක කොටස් කිහිපයක් ද වෙනත් බීජ වර්ග කිහිපයක්ද හමු වී ඇති අතර ඔහු එම වෙනත් බීජ වර්ග කිහිපය ද ගෙන බඩ ඉරිඟුවලින් වෙන්කොට වගා කර තිබේ. විදේශයකින් මෙරටට ගෙනා බඩඉරිඟු ඇට තොගයක තිබූ වෙනත් බීජ වර්ග කිහිපයක් තවාන් දැමීමෙන් පසු සේනා දළඹුවාට සමානකම් ඇති වෙනත් දළඹු විශේෂයක් එම තවානේ සිටියදී හමුවූ පුවතක් මාරවිල ප්‍රදේශයෙන් වාර්තා විය. මින් එක් බීජ වර්ගයක් පැළ වී ඇති අතර අනෙක් වර්ගය පැළ නොවීම නිසා ඒ පිළිබඳ සොයා බැලීමට උරය තුළ ඇති පස් ඉවත් කළ විට එහි බීජ නොතිබී ඇති අතර ඒ වෙනුවට දළඹුවකු දක්නට ලැබුනු බවයි ගොවිපලේ හිමිකරු පැවසුවේය.

Source Text - 2

Apple කිව්වම බොහෝ දෙනෙකුට එකවරම සිතට නැගෙන්නේ අනෙකුත් දුරකථන වලට සාපේක්ෂව වෙනස්ම වූ නිර්මාණ රටාවක් සහ නවම වූ තාක්ෂණයෙන් පරිපූර්ණ දුරකථනයක්. මෙයට හේතුව මින් පෙර ඇමරිකාවේ Apple සමාගම ඔවුන්ගේ දුරකථන තුළ පවත්වාගෙන පැමිණී එකින් එකට වෙනස් සුපිරි සහ ගුණාත්මක බවයි. එම නිසාම Apple දුරකථන වටා විශාල පාරිභෝගිකයන් රැසක් එකතු වූ නමුත් මෑත කාලයේදී එම කරුණු විශාල වශයෙන් වෙනස් වී තිබෙනවා. අපේ සමාජය තුළ Apple දුරකථන භාවිතය ඉහළ සමාජ මට්ටම පෙන්නුම් කරන සාධකයක් බව බොහෝදෙනා ප්‍රකාශ කර සිටින අතර එයට විරුද්ධ පිරිස් ද දැක ගැනීමට පුළුවන්. මින් පෙර Apple දුරකථන තුල පැවති සුපිරි ක්‍රියාකාරිත්වය, පෙනුම සහ සරල බව ආදී කරුණු හේතුවෙන් අනෙකුත් දුරකථන වලට සාපේක්ෂව Apple දුරකථන භාවිතය වඩාත් වටිනාකමකින් යුක්ත විය. නමුත් වර්තමානයේ මේ තත්වයන් වෙනස් වීම නිසා Apple දුරකථන වලට ප්‍රබල අභියෝගයක් දීමට අනෙකුත් දුරකථන වලට හැකි වී තිබෙනවා. එම නිසා මේ වනවිට Apple දුරකථනයකට වැය කරන මිලට වඩා ඉතාමත් අඩුවෙන් වෙනත් සුපිරි දුරකථන මිලදී ගැනීමට හැකියාව පවතිනවා. මෑත කාලයේ නිකුත් වූ Apple දුරකථන වෙන් වෙන් ව ගෙන බැලූ විට එම දුරකථන වල එතරම් සැළකිය හැකි වෙනස්කමක් තාක්ෂණය අතින් හෝ පෙනුම අතින් දැකගැනීමට නොහැකි වූ නමුත් මිල අතින් විශාල වැඩි වීමක් දැකගැනීමට පුළුවන්. මෙම හේතූන් නිසාම Apple සමාගමේ විකුණුම් පහත වැටී තිබෙන බව විදෙස් වෙබ් අඩවි පසුගිය කාලයේදී විටින් විට වාර්තා කර තිබුනා. විශේෂයෙන් Apple දුරකථන වල මිල ගණන් ඇමරිකානු වෙළඳපොල පදනම් කරගෙන තීරණය කරන බැවින් මෙම දුරකථන වෙනත් රටවලට පැමිණීමේදී අධික මිල වැඩිවීමක් දකින්නට ලැබෙන අතර එම එක් එක් රටවල වෙළඳපොළ ඩොලරයට ගෙවන මිල වැඩිවීම මෙම දුරකථන වල මිල අධික ලෙස වැඩිවීමට හේතුවක් ලෙස දැකිය හැකිය. උදා :- Apple දුරකථන ඉන්දියාව වැනි වෙළඳපොළකට හඳුන්වාදීමේදී එම වෙළඳපොළට ගැළපෙන අයුරින් මිලගණන් වෙනස් නොවීමට අමතරව අතිරේක බදු ගෙවීමට සිදුවීම නිසාත් මෙම දුරකථන අධික මිලක් වීමට හේතු වෙනවා. මෙම හේතූ කිහිපය නිසා බොහෝ රටවල පාරිභෝගිකයන් Apple දුරකථන මිලදී ගැනීමේදී රුචිකත්වය පසෙකලා වටිනාකමට මුල්තැන දෙන බවට මෙමගින් පැහැදිලි වී තිබෙන අතර මේ අනුව Apple දුරකථන වලට වඩා Samsung , Huawei ,Xiaomi වැනි විකල්ප සඳහා පාරිභෝගිකයන් පෙළඹී තිබෙනවා."

Human Extracted summary - 2

Apple දුරකථන අනෙකුත් දුරකථන වලට සාපේක්ෂව වෙනස්ම වූ නිර්මාණ රටාවක් සහ නවම වූ තාක්ෂණයෙන් පරිපූර්ණ දුරකථනයක්. මෙයට හේතුව ඔවුන්ගේ දුරකථන තුළ පවත්වාගෙන පැමිණී එකින් එකට වෙනස් සුපිරි ක්‍රියාකාරිත්වය, පෙනුම,ගුණාත්මක බව සහ බවයි.වර්තමානයේ Apple දුරකථන වලට ප්‍රබල අභියෝගයක් දීමට අනෙකුත් දුරකථන වලට හැකි වී තිබෙනවා. එම නිසා Apple දුරකථනයකට වැය කරන මිලට වඩා ඉතාමත් අඩුවෙන් වෙනත් සුපිරි දුරකථන මිලදී ගැනීමට හැකියාව පවතිනවා. මේ අනුව Apple දුරකථන වලට වඩා Samsung , Huawei ,Xiaomi වැනි විකල්ප සඳහා පාරිභෝගිකයන් පෙළඹී තිබෙනවා. මෙම හේතූන් නිසාම Apple සමාගමේ විකුණුම් පහත වැටී තිබෙන බව විදෙස් වෙබ් අඩවි පසුගිය කාලයේදී විටින් විට වාර්තා කර තිබුනා."

Machine Extracted summary - 2

මින් පෙර Apple දුරකථන තුල පැවති සුපිරි ක්‍රියාකාරිත්වය, පෙනුම සහ සරල බව ආදී කරුණු හේතුවෙන් අනෙකුත් දුරකථන වලට සාපේක්ෂව Apple දුරකථන භාවිතය වඩාත් වටිනාකමකින් යුක්ත විය. එම නිසාම Apple දුරකථන වටා විශාල පාරිභෝගිකයන් රැසක් එකතු වූ නමුත් මෑත කාලයේදී එම කරුණු විශාල වශයෙන් වෙනස් වී තිබෙනවා. මෑත කාලයේ නිකුත් වූ Apple දුරකථන වෙන් වෙන් ව ගෙන බැලූ විට එම දුරකථන වල එතරම් සැළකිය හැකි වෙනස්කමක් තාක්ෂණය අතින් හෝ පෙනුම අතින් දැකගැනීමට නොහැකි වූ නමුත් මිල අතින් විශාල වැඩි වීමක් දැකගැනීමට පුළුවන්. විශේෂයෙන් Apple දුරකථන වල මිල ගණන් ඇමරිකානු වෙළඳපොල පදනම් කරගෙන තීරණය කරන බැවින් මෙම දුරකථන වෙනත් රටවලට පැමිණීමේදී අධික මිල වැඩිවීමක් දකින්නට ලැබෙන අතර එම එක් එක් රටවල වෙළඳපොළ ඩොලරයට ගෙවන මිල වැඩිවීම මෙම දුරකථන වල මිල අධික ලෙස වැඩිවීමට හේතුවක් ලෙස දැකිය හැකිය. නමුත් වර්තමානයේ මේ තත්වයන් වෙනස් වීම නිසා Apple දුරකථන වලට ප්‍රබල අභියෝගයක් දීමට අනෙකුත් දුරකථන වලට හැකි වී තිබෙනවා. අපේ සමාජය තුළ Apple දුරකථන භාවිතය ඉහළ සමාජ මට්ටම පෙන්නුම් කරන සාධකයක් බව බොහෝදෙනා ප්‍රකාශ කර සිටින අතර එයට විරුද්ධ පිරිස් ද දැක ගැනීමට පුළුවන්. මෙම හේතූ කිහිපය නිසා බොහෝ රටවල පාරිභෝගිකයන් Apple දුරකථන මිලදී ගැනීමේදී රුචිකත්වය පසෙකලා වටිනාකමට මුල්තැන දෙන බවට මෙමගින් පැහැදිලි වී තිබෙන අතර මේ අනුව Apple දුරකථන වලට වඩා Samsung , Huawei ,Xiaomi වැනි විකල්ප සඳහා පාරිභෝගිකයන් පෙළඹී තිබෙනවා.