ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

AUTOMATIC AMHARIC NEWS TEXT SUMMARIZER

By

KAMIL NURU

A thesis submitted to

the School of Graduate Studies of Addis Ababa

University

in partial fulfillment of the requirements for the Degree

of Master of Science in Information Science

June 20, 2004

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
Faculty of Informatics
Department of Information Science

AUTOMATIC AMHARIC NEWS TEXT SUMMARIZER (EXTRACTION)

BY

KAMIL NURU

Name and Signature of Members of the Examining Board

Ato Getachew Jemaneh, Chairman, Examining Board _____
Dr. Björn Gambäck, Advisor _____
Dr. Nega Alemayehu , Examiner _____

_____ _____ _____
Chairman, Faculty Signature Date

_____ _____ _____
Chairman, Graduate Council Signature Date

ACKNOWLEDGEMENTS

I would like to thank my advisors Dr. Björn Gambäck and Mr. Gunnar Eriksson for every contribution they made to the success of this study.

My warm gratitude mainly goes to Mr. Yacob who helped me a lot in the area of perl programming work for the development of the model.

My gratitude also goes to my friends Ato Eskindir Belyneh and W/rt Meron Seid who helped me a lot in editing, data collection and preprocessing.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ..... **I**

TABLE OF CONTENTS ..... **II**

LIST OF TABLES ..... **IV**

ABBREVIATIONS ..... **V**

ABSTRACT ..... **VI**

CHAPTER ONE **1**

INTRODUCTION **1**

1.1 BACKGROUND 1
1.2 STATEMENT OF THE PROBLEM AND JUSTIFICATION OF THE STUDY 4
1.3 OBJECTIVES OF THE STUDY 7
1.3.1 General objectives *7*
1.3.2 Specific objectives *7*
1.4 METHODOLOGY 8
1.4.1 Data collection *8*
1.4.2 Adopting extraction techniques *8*
1.4.3 Design of the system *9*
1.4.4 Adjusting weights *9*
1.4.5 Evaluation *10*
1.5 SCOPE AND LIMITATIONS OF THE STUDY 11
1.6 ORGANIZATION OF THE THESIS 11

CHAPTER TWO **13**

AUTMATIC TEXT SUMMARIZATION **13**

2.1 INTRODUCTION 13
2.2 BASIC CONCEPT 13
2.3 APPROACHES TO ATS 14
2.3.1 Abstraction approaches *16*
2.3.2 Extraction approaches *17*
2.4 REVIEW OF TEXT MINING ELEMENTS 22
2.4.1 High frequency words *22*
2.4.2 Cue phrases *23*
2.4.3 Titles and headings *23*
2.4.4 Location heuristics *23*
2.4.5 Linguistic analysis *24*
2.4.6 Text formats *24*
2.4.7 Word multiples *25*
2.5 WEIGHT ASSIGNING TECHNIQUE 25
2.6 REVIEW OF ATS EVALUATION 26
2.7 REVIEW OF SAMPLE SUMMARIZATION SYSTEMS 27
2.7.1 Luhn's frequency method *27*
2.7.2 Edmundson's extraction system *28*
2.7.3 Salton's vector space model *28*
2.7.4 SweSum summarizer *29*
2.7.5 Kupiec's document summarizer *30*
2.7.6 Using cohesive properties of text for Automatic summarization *30*
2.7.7 The LAKE System at DUC2004 ..... *31*
2.7.8 The copy-and-paste system (1999) *31*
2.7.9 Workshops and conferences *32*

CHAPTER THREE **34**

STRUCTURE OF AMHARIC TEXT **34**

3.1 INTRODUCTION 34
3.2 ORIGIN 34

3.3 WRITING SYSTEM 35
3.3.1 The nature of Amharic script 35
3.3.2 Text parts 35
3.3.3 Grammar 37
3.3.4 Punctuation marks 38
3.3.5 Homophonic characters 39
3.3.6 Abbreviations 39
3.3.7 Word forms 39
3.3.8 Numbers 40
3.4 AMAHRIC NEWS ITEMS 40

CHAPTER FOUR **42**

ARCHITECTURE OF THE SYSTEM **42**

4.1 INTRODUCTION 42
4.2 DESCRIPTION OF THE SYSTEM 42
4.2.1 Extraction features 43
4.2.2 Learning features 43
4.3 DESCRIPTION OF THE ALGORITHM 44
4.4 DATA REQUIREMENTS 48
4.5 PREPROCESSING REQUIRED 48
4.5.1 Segmenting 50
4.5.2 Data format 51
4.6 EXPERIMENTATION 52
4.6.1 Why perl is preferred 52
4.6.2 Adjusting weights of diagnostic units 53
4.6.3 Evaluation of the performance 63
4.7 DISCUSSION 63

CHAPTER FIVE **65**

CONCLUSION AND RECOMMENDATIONS **65**

5.1 CONCLUSION 65
5.2 RECOMMENDATIONS 67

REFERENCES: **68**

APPENDIX ONE **73**

APPENDIX TWO **78**

APPENDIX THREE **79**

APPENDIX FOUR **84**

APPENDIX FIVE **89**

APPENDIX SIX **92**

LIST OF TABLES

Table 4.1 Starting weights 55
Table 4.3 Human vs. system summary with 41% condensation factor..... 56
Table 4.3 summary of adjusting weights set one at 41%..... 57
Table 4.4 Summary of adjusting weights set two at 39% condensation 60
Table 4.5 summary of adjusting weights set three at 42% 61
Table 4.6 Summary of adjusting weights set four at 36% condensation..... 62
Table 4.7 final weights assigned 62
Table 4.8 Summary of model performance on test news items 63

ABBREVIATIONS

AAAI - American Association for Artificial Intelligence

ACL - Association for Computational Linguistics

ATS - Automatic Text Summarization

DUC – Document Understanding Conference

EACL - European chapter of the Association for Computational Linguistics

HTML – Hyper Text Markup Language

ID – Identification

IR – Information Retrieval

LAKE – Learning Algorithm for Key phrase Extraction

NAACL - North American chapter of the Association for Computational Linguistics

NLP - Natural Language Processing

OCR – Optical Character Recognition

SGML - Standard Generalized Markup Language

SMS - Short Message Service

XML - eXtensible Markup Language

ABSTRACT

It is visible that the amount of textual information output is highly increasing from day to day. Compared to the text output the human capacity of reading is almost negligible. This big difference creates a problem in communicating information to the best possible extent. Managing the output also becomes very difficult. Tasks of sorting, searching through and categorizing are turning out to be cumbersome. The limited carrying capacities of the communication channels also require huge reduction in size.

The focus of this research is on development of a mechanism for shortening Amharic news texts and for producing concise summaries of them. The system tries to pin point the most important sentences of the original text and extract them as a summary of the news. Thus the extract is a lot shorter and painless to handle.

The proposed summarizer utilizes several statistical techniques, location heuristics and diagnostic units to determine the parts of the text to be extracted. Selected information retrieval and text mining techniques are adopted to build a model for the proposed system.

The application of the system after adjusting the weight of its diagnostic units by using four Amharic news items in 124 different ways reveals a promising result in automating the task of generating news summaries. Human generated summaries are used for adjusting weight and evaluating the system. Finally 58% Recall and 70.4% Precision values are attained. Based on this result, further work is recommended for future improvements of this system and studies in the area of automatic Amharic text summarization.

CHAPTER ONE

INTRODUCTION

1.1 Background

The explosively growing amount of textual information from day to day is hindering efficient information gathering and distribution. The limited human reading capacity worsens the problem.

Wu and Liu (2003) showed that instead of requiring readers to go through all articles, providing summaries of the articles is one way to save readers' time. However the work of generating summaries is far away from being easy and requires enormous amount of time. An automatic generation system helps in accelerating the generation of summaries.

Automatic text summarization is a technique where a computer program generates summaries of texts. It is a process of reducing the size of a text yet preserving its important information content.

The origin of automatic text summarization dates back to the start of automatic text analysis and information retrieval. According to van Rijsbergen (1979) automatic text analysis started to show a promising result in the late fifties. It is at this time that scientists felt confident to put their hands to every NLP problem, including automatic text summarization.

The first abstracting system based on word frequency was developed by Luhn (1959). He proposed that the frequency of word occurrence in an article furnishes a useful measurement of the significance of the word in the article. Luhn (1959) further proposed that the relative position of words within a sentence gives us a useful measurement for determining the significance of the sentences.

Therefore sentence importance could be measured based on the mentioned two qualities.

With the blink processing speed of the current machines and their large memory different natural language processing techniques have been developed. Nowadays, depending on the processing tools, the available corpus and the research motives one can choose between pure NLP, Statistical NLP and hybrid methods. In order to select a method we need to know what kind of text to summarize and what tools we have.

Whichever method is used, automatic text summarization is a way of expressing the main points of a long text in a shorter way by extracting the core points that are contained in the original text. The summarization can be done either by an abstracting technique or by sentence extraction. In the abstracting system the entire content of the text is analyzed (language structure, context, semantics, etc) and new sentences which are shorter and more expressive are generated. Whereas in the extraction system the sentences in the original text are weighed based on different weighing parameters and the high scoring ones above a determined threshold are taken out as the summary of the text.

Summarizers provide the spinal idea of a text and readers could save their time by going through the summary. Summaries help not only to have a general overview of the full document, but also in making decisions on whether the full material is relevant enough to be read fully or not.

Automatic text summarizers help a lot in the process of information retrieval by narrowing the search space. Unquestionably assessing a smaller space is easier than a larger space. The information gathered from a well extracted summary is almost equivalent (not equal) to the information in the main

document. Hence search is easier in the summaries which are shorter in size but representatives of large documents.

Several text summarizers have been developed for different languages using different techniques. Among the many summarization systems the SweSum summarizer for Scandinavian languages (Dalianis and Hassel, 2000) and the Copernic summarizer (Copernic, 2003) are some examples in the area of automatic text summarization. When we come to Amharic language, automatic Amharic text analysis has been one of the research areas conducted at Addis Ababa University, Faculty of Informatics, and Department of Information Science. See for example, Atelach (2002) on automatic Amharic sentence parsing, Mesfin (2001) on part of speech tagging, and Surafel (2003) on Automatic categorization of Amharic news items. Amharic text summarization has not been studied so far. The automatic Amharic news text categorizer by Surafel (2003) and Classification of Amharic news items by Zelalem (2001) are attempts to point out the category in which a news text falls. This can tell the center around which the news is spinning about. These systems can be taken as good starting points to build an automatic news text summarizer for Amharic news items.

To exploit the advantages of text summarization, the development of automatic Amharic text summarizer is of vital importance. This study is an attempt to develop Amharic text summarizer specific to news texts, as task which is a part of building a general text summarizer.

A number of press organizations bring news to the daily Amharic news desk. As a result considerable numbers of news items are released daily requiring readers to pay more time.

In response to the above mentioned amount of the outputs, this study tries to find a means of extracting summaries of Amharic news items that are free of redundancy and details. Hence the extracts are short enough to take less time to go through them.

Most news items reach the reader in printed form but automatic summarization systems work on digital form. To bridge this gap Optical Character Recognitions (OCR) systems could be used to convert the news items to digital form. From the point of view of saving readers time the use of OCR systems is not recommended. Parallel release (that is news with its summary may be at the beginning) or prior release of news summaries (i.e. releasing news summary first and then the full news) seems better way.

1.2 Statement of the problem and justification of the study

The textual information output is highly increasing from day to day while the human capacity of reading is limited. News items comprise a certain part from these outputs. With the current absorbing pace one can not catch up with the information output. Hence to mediate the gap the need for automated systems that generate summaries are of great benefit.

Though it is becoming more important to listen to or read the daily news in our preference area, due to shortage of time and other workloads, listening to full news attentively or reading a newspaper about a topic start-to-end is not always possible. In fact, nowadays most people seem to have the will of reading newspapers, magazine articles, specialized literature and listen to daily news. But this is not always successful. News summarizers therefore, create concise news summaries, enabling people to absorb more information in less time.

Short message service (SMS) on mobile phones needs a lot of reduction in size. Summarizers creates convenience to release news summaries over these cellular devices by making the news items short enough to be piped.

With the limited time people have, sorting through all the information available today does not seem possible. Devising an automatic system to take care of the summarizing work could provide some way out. The development of Automatic Amharic news summarizer can at least help with part of the problem of information overload. And also it puts the corner stone for the development of general Amharic text summarizer.

Rather than the time consumed to read a material searching the material is becoming more difficult in the current days as a result of the big amount of information that is published everyday. Even in preparing some material it is very important to know what has been said about what we are intending to write or say. There are a lot of redundant things that are released everyday as a result of the difficulty to assess what already exists. A summary provides short overviews to observe the general idea of what has been touched so far. Hence summary reduces the chance of redundancy by narrowing the amount of text to assess.

Summaries help not only to know the existence of some material but also they provide some overview of the materials. By going through each and every detail, covering what is already written has become a dream compared to the massive outputs. Therefore a summary gives a suitable way of assessing the previous deeds and saves time for the next work to do.

News is produced in such an attractive way to capture interest of the listeners/readers. To do so a lot of unnecessary ways of expressions are added to the original content of the news. Of course to appreciate the way of presentation

and author expressions, the actual substance is mandatory. But observed from the point of view of the content delivered in reference to the time function there is no need to waste time on details. Summarized news is preferred in this case which tells about the main point only.

Currently newspapers and other news releases reach the reader in a non digital format. Of course most of the news items are processed using digital processing systems before they reach the reader in a printed or other form. The automatic Amharic news text summarizer to be developed essentially needs news items to be in a digital format, but most of the news items are found in printed forms. Commercial OCR (Optical Character Recognition) could be used to convert the items in hard copy to their softcopy form. As readers' time is tried to be saved the use of OCR is another burden. A better solution is prior release of the summaries of the news items by the news agency. This can function as an advertisement to the detail news. Parallel release (detail news with its summary may be at the beginning) is also another alternative.

For the purpose of emphasis and giving weights to important issues short items are easy to manage. The amount to retain decreases enormously when a lot of repetitive details with little changes are encountered. Even the theme of news may get blurred due to details coming one after the other and hiding the core message. Original texts are full but in reference to time short and precise things seem to be attractive and preferable.

Another additional advantage of a news summary is the enhancement of search. Korfhage (1997) described the contribution of summaries in searching. Searching is faster in short summaries as compared to full documents. Since the summary contains most of the important issues, finding in the summary is easier.

To the best of the researcher's knowledge there are no previous works done to handle automatic Amharic text summarization. There are research works done to take care of other Amharic language processing areas in the Faculty of Informatics, Department of Information Science Addis Ababa University. Most of the works addressed areas like parsing, tagging, stemming . . . etc. which will be good aids and starting points to the development of this automatic news text summarizer. Automatic Amharic news categorizer by Surafel (2003) and Classification of Amharic news items by Zelalem (2001) lay a firm ground to the construction of this system.

1.3 Objectives of the study

1.3.1 General objectives

The general objective of this research is to develop a system which prepares short summary of a news text by extracting the sentences with more vital substances in them. This is a means of taking out the sentences with many key points of the theme of the news.

1.3.2 Specific objectives

1. Identify different diagnostic units to point out the sentences with many crucial points.
2. Adapt available techniques and algorithms to rank the sentences in a text to Amharic news items.
3. Develop an algorithm for extraction.
4. Build and train a computer model.
5. Test and evaluate the system by comparing output of the model to human summaries made by sentence selection.

6. Place a corner stone for future development of Amharic text summarizer.

1.4 Methodology

Available related literature is reviewed to understand the concept of automatic text summarization. Several techniques of text mining are reviewed. History of text summarization is discussed by taking sample summarization systems. Different workshops and conference held on text analysis in reference to time are assessed. The nature of Amharic language and its text structure are also reviewed to build foundations of this work as it works on Amharic items.

1.4.1 Data collection

Nine news items from newspapers (Reporter and Addiszemen) and a news agency (Radio Fana) are collected. The news items taken from news papers which are found in printed form are first converted to digital form. The news items are kept separately. Then they pass though a pre-processing step. This step makes the items ready to be feed to the system.

As the system to be developed handles textual part of the news, pictures and other drawings are not included in the news that is selected for adjusting weight, testing and evaluating the system to be developed.

Variation of content is tried to be considered in the process of data collection so that the system is trained to several domains.

1.4.2 Adopting extraction techniques

Different techniques that are useful for this system are adopted and an extraction algorithm is developed for the system. The extraction techniques of Edmundosn (1969) are adapted to this system by taking some of the outlines he

made. Key method, cue method, title method and location methods are taken. Linguistic and structural heuristics are not considered because the system to be developed is basically statistical.

The basic foundation laid by Luhn (1959), the idea of word frequency is taken and adopted to this system by integrating it with the involvement of title sentence and header sentence words.

A modification is made to Luhn's approach of removing stopwords by a threshold. Stop words are removed from the frequency list of the developed system by descriptively comparing the list with a predetermined list of non content bearing words.

1.4.3 Design of the system

After understanding the concept of automatic text summarization and identifying the techniques of application to news items a model is developed to test the components of the proposed system. The developed model is intended to utilize concepts of extraction. The system makes use of different diagnostic units to point out the important sentences of the news text. Some of the diagnostic units may have a fixed weight while the others change dependently. The system is going to be designed to entertain users' need of size reduction. A dynamicity component is added to the system so that it updates its predetermined tools of text analysis by interacting with domain expert user. The learning feature is going to include the updating of cue phrases and updating of list of stop words.

1.4.4 Adjusting weights

Summaries of the collected news items are manually generated by two different persons. Ranking of sentences by relevance and removing relatively

irrelevant sentence is used in the process of manual extraction of the summaries. The amount to reduce is also determined by the human summarizers. Controversial points in the extraction of the summaries are resolved by removing the item form the summary to be used as a target extract.

Starting weights are assigned to the diagnostic units assuming their relative importance. By afftecting the starting weights one by one and observing their effect on one news item an attempt is made to make the system extract the sentences that are judged as a target extract. The process continued till the human summary and the system summary for the news item matched.

The adjusted weights of the diagnostic units using the fist news item are carried to the second news item and the same adjustment as the first news item is made for the second. The result obtained from the second adjustment is then compared to the first and an average of the two is carried to the third.

Doing the same adjustment for the third and fourth news items, average weights are assigned to the text mining tools.

1.4.5 Evaluation

After the system outputs are tried to be pulled to the target extracts of the four news items, by affecting the weights of the diagnostic units, final weights are assigned in the weights adjustment process.

With the final weights assigned to the diagnostic units the system is given five news items to extract the summary of each. The system is evaluated using manually generated summaries of the five news items as target extracts. Precision/Recall evaluation is used to measure the performance of the system.

The ability of the system to extract useful sentences and the ability of the system to reject unnecessary sentences is tested on the five news items. Finally the average of the attained value is placed as the performance of the system.

1.5 Scope and limitations of the study

This research focuses on developing news text summarizer by extraction technique. By news summary the important things said in the news will be extracted, leaving the explanations and details. There is no concern of summarizing every text like educational articles, text books, novels and other texts talking about different issues which need a deep linguistic analysis.

For the reason that the system employs extraction technique, beautiful makeup in the flow of idea may not be kept in the summary made by the system. Coherence is sacrificed for size reduction. The original taste of the sentences will be maintained but the expressions may not be as sweet as the original text due to the fact of chopping preceding or following sentences to the sentences extracted.

1.6 Organization of the thesis

This thesis is divided into five chapters. The first chapter is an introduction to the research and discusses the background of the problem, the objectives, and the methodology of conducting the study.

The second chapter presents the basic concepts behind automatic text summarization. The different approaches to automatic text summarization and the history of text summarization are discussed in this chapter. Different text diagnostic units are also discussed.

The third chapter is about Amharic writing system in view of automatic text summarization. Amharic linguistic structures that create problems to the system are discussed and alternative solutions are pointed out in this chapter.

Chapter four is about the architecture of the system and development of a model. The conducted adjusting weight test and analysis of the results are presented in this chapter.

The final chapter contains the conclusion drawn from the test results and recommendations forwarded for further development and researches.

The appendix part contained the perl code used in the model and some of the news items used in adjusting weights and evaluation. Manually generated summaries of the news items by two professionals and the system extracts are attached in the appendix. Hence the reader can compare the differences between the summaries and perceive the model performance.

CHAPTER TWO

AUTMATIC TEXT SUMMARIZATION

2.1 Introduction

This chapter tries to address the main concepts behind automatic text summarization. Different approaches to automatic text summarization and the available diagnostic units of text analysis for automatic text summarization are also discussed. A number of text mining techniques from different disciplines are examined in relation to what they contribute to text summarization systems. A short review of the history of automatic text summarization is tried to be addressed by describing some summarization systems. Major conferences and workshops held on the topic of summarization are presented graphically.

2.2 Basic concept

Automatic text summarization (ATS) is a way of expressing a long text shortly by extracting the main points that are contained in the original text. This is based on the assumption that there exists parts of a text which tell the important points in the text and extracting these units will convey the important message intended to be delivered by the original text.

According to Dalianis (2002) in summarization process the most relevant parts of a document are extracted and put together in the summary. The summary must be shorter than the original document, and the summary must reduce redundancy. He also pointed out that one approach to text summarization is extraction of sentences that contain much of the principal points in the document. Another approach is the construction of new sentences based on the core idea of the document.

The extraction of summary may be from a single document where the summary of a document is generated independent of other documents. In another hand different documents can be condensed to a single summary. The former is single document summarization while the later is termed as multi-document summarization. Dalianis (2002) referred multi-document summarization as a more advanced form of summarization where several texts are summed into one summary.

2.3 Approaches to ATS

Basically there are two approaches to produce the summary out of a text. One is the creation of summaries using terms, phrases and sentences pulled out directly from the source text using different measures. The other is formulating a new sentence after a thorough analysis of the text. The methods are called extraction and abstraction respectively.

Karen Sparck Jones (1998) drew a critical distinction between extraction and abstraction. She quoted that, "What you see is what you get," incase of extraction because part or parts of source text are extracted. But in abstuction "What you see is what you know" that is abstraction can be referred as fact extraction.

As to Fuentes (2001) initially summarization was reduced to textual monolingual single-document condensation task, but afterwards it has evolved covering a wide spectrum along several dimensions: extraction vs. abstracting indicative vs. informative. Generic vs. query based, background vs. getting the new, restricted domain vs. unrestricted domain, textual vs. multimedia, single document vs. multiple document summarizers have been developed. The summarizers can be applied on generating biographical summaries, medical

patient summaries, e-mail summaries, WebPages summaries, and news summaries and so on.

Dalianis & Hassel (2001) devised three basic steps to do the summarization task automatically. Understanding of the topic is shown to be the first and then extraction of important parts of the text on the basis of the topic selected is next. Finally generation of the summary is discussed to be the third step.

As stated by Dalianis & Hassel (2001) Topic detection, or detection of important parts of a text, can be done by the following parameters:

Baseline: Sentence order in text gives the importance of the sentences. First sentence usually has highest ranking and last sentence lowest ranking. Important concepts are placed at the beginning to capture the attention of the reader and introduce what is coming next.

Title: Words in title and in the immediately following sentences are given high score.

Term frequency (tf): Open class terms that are frequent in the text are more important than the less frequent ones.

Position score: The assumption is that certain genres put important sentences in fixed positions. For example, newspaper articles usually have most important terms in the 4 first paragraphs. Reports on the other hand have many important sentences at the end of the text.

Query signature: The query of the user can be used to affect the summary in the way that the extract will contain these words. If such words are found, the summary will be slanted to them.

Sentence length: The sentence length may imply which sentence is the most important.

Average lexical connectivity: Number of terms shared with other sentences. The notion of average lexical connectivity is that sentences sharing more terms with other sentences are more important.

Numerical data: Sentences containing numerical data are scored higher than the ones without numerical values.

It is also noted that all the above parameters are normalized and put in a simple combination function with modifiable weighting. The idea is that high scoring sentences in the original text are kept in the summary. The scores are calculated according to the criteria discussed above. Lastly a summary could be produced with the selected sentences.

2.3.1 Abstraction approaches

Abstraction approach is a method of generating the summary of a text by first understanding the text totally and then producing sentences that are short and more representative. This approach seems closer to the human way of generating summaries. The implementation of abstraction systems requires a very deep understanding and analysis of the language structures, grammar and semantics. Modeling the human dimension which includes common sense and enormous experience is not that easy. Abstraction approaches to text summarization is tough because it needs deep linguistic analysis. Statistical methods are does not seem much helpful in this approach.

Maybury and Mani (2001) mentioned that an abstract is a summary at least some of whose materials is not present in the main document. They also stated that abstracts can result in shorter summaries than extracts. This is because

compact sentences which express more points can be constructed in abstracting. Generating sentences and language analysis is not the only task that makes the abstraction approach difficult but also identification of the sentence making elements is also a problem.

As to Maybury and Mani (2001) the abstract can be constructed by template extraction or concept abstraction. After identification of the basic template, sentences are built around the template according to the instances of their occurrences. The system by Paice and Jones (1983) is an example of abstraction by template extraction. While this method provides a good capacity for abstracting semantic content, it requires customization for specific type of input.

Hahn & Reimer (1999) developed concept abstraction system. Their system captures the content of a document in terms of abstract categories. Abstract categories are defined as sets of terms or topics from the document for the system knowledgebase. Based on the abstract categories summary sentences are constructed incorporating the idea of the abstract category.

2.3.2 Extraction approaches

The main theory behind sentence extraction is that there are sets of sentences which present most of the key ideas of the text. Extracting theses sentences results in the summary of the text. But extracted sentences may not be coherent. Although the extracted sentences probably contain unresolved logical order they do provide the core content of the text. Extracts also keeps original articulation and ways of expression.

The Copernic summarization technology (2003) stated three points in the summarization task that are very important when extraction is used. The first is the extraction of the concepts that are associated with the documents main

content. This can be taken as identifying the information in the document at the atomic level. This tells that a good summarizer just picks core information contents and produces the summary containing theses concepts.

The second important point is breaking a long document into several sub parts. This is because a document may contain subparts that are quite different in content. These parts are expected to be summarized without the influence of one another and the total summary must be the integration of these parts. The consideration of these points allows summarizers to perform regardless of the length of the document.

The third important point is the selection of sentences according to their relative importance. Sentences that are less informative will be given less weight and those sentences with many of the key concepts identified in the first step are given more weights. The summary is then constructed by selecting the sentences with more substances in them.

As this study is concerned with the summarization of news text by sentence extraction the concept of long document segmentation may not have remarkable value for this work. But the summary text will be formed by selected sentence/s from the original text. This sentence/s may need some reorganization to improve flow of idea and textual coherence.

Several rule based and statistical natural language processing techniques are required to lay a background for the extraction system to be developed. Following are some of the techniques.

A) Rule-based techniques

This technique is a method of imposing governing rules to extract sentences as summary of the text. Some of the rules are:

- **Extract first sentences:** in a text it is common to put the theme containing sentence at the beginning therefore taking first sentences to the summary could make the summary state the core of the text to be summarized. But this is not common to all written materials. Important sentences of news may appear at the beginning whereas reports contain important things around the end of text.
- **Pick sentences containing title words:** a text is about to be organized around the title and this method is a way of extracting sentences that contain words in the title. The assumption behind is that sentences containing title words are more probable to mention the main points in the text.
- **Pick sentences containing cue phrases:** a text usually tries to express itself. There are phrases which indicate a certain sentence is important. In English phrases like "this paper explains", "The general idea is" indicates that the sentences are important. Sentences with cue phrases in them are good enough to be included in the summary.

Generally rule-based approaches work on the basis of if . . . then . . . type of instruction. They are easy to implement. But rules are commonly domain dependent. A rule may be very efficient in one type of document. But when applied on another type of document it may not be as efficient as the first type.

B) Statistical techniques

This study is mainly concerned with statistical approaches to automatic text summarization by sentence extraction. Statistical principles, rules and laws that are useful for the purpose of sentence extraction are dealt with.

In order to extract sentence which are eligible to the summary, there must be some mode of ranking the sentences. Statistical weights could be assigned to sentences based on several parameters. Some of the diagnostic units are position and key terms.

Considering terms as a measure to assign scores to sentences Robertson and Spark Jones (1994) proposed three term weighting techniques

- **Term frequency:** the number of times a term appeared in a document is taken as the weight of the term. Terms may be stemmed to reach their root word so that minor changes and additions due to language necessities are eliminated.
- **Collection frequency(CFW):** if **n** is the number of documents term **t** appears in and **N** is the total number of documents

  $$\mathbf{CFW = \log N - \log n = \log (N/n)}$$

  This method selects terms which occur often in a certain document but rarely in the rest. This can be applied to the system under development by considering each element as a single document.
- **Term frequency inverse document frequency (TFIDF):** this technique expresses the weight of a term by relating its frequency to the number of documents it appeared in and the total number of documents under consideration.

  $$\mathbf{Wt = TF_i * \log(N/n_i)}$$

Where: n_i is the number of documents term t_i appear in and N is the total number of documents. TF_i is the frequency of the term and **Wt** is the **TFIDF** weight. This is a measure that can be applied when thinking of generating summaries of multiple documents. This method is found to fever terms which are highly frequent in a document but rare in the rest of the documents.

The relative importance of sentences in a text can be measured by the weight of the terms they contain. By fevering the sentences that are containing key terms in the text, it is possible to assign relative weights to sentences. In a deeper analysis sentences can be given weights by considering the value of each of the terms in them. Some terms may increase the weight of the sentence while the others decrease.

Unlike rule based techniques statistical techniques tries to draw out parameters of extraction from the text itself. Sentences in a text are given scores that represent the degree of goodness of a sentence to be included in the summary. These methods are not influenced by predetermined qualities of goodness.

C) Rule-based vs. Statistical techniques

Statistical approaches do not totally contradict the rule-based approaches. Instead the rules could be used as one measure of assigning weights to sentences.

The main advantage of statistical approaches is the drawing out of basic sentence weighting tools from the text. This point makes the approach domain independent and trainable as long as it can learn the statistics of any text from any domain.

In contrary to the above mentioned advantage statistical methods take words as independent units. Considered from the language aspect there are words

which do not appear independently. Such words must be treated as one but statistical methods for open word classes do not take this aspect into account.

2.4 Review of text mining elements

Text is having different elements and absolutely all the elements are not equally important. Simply there are stop words which are not that much content bearing but important to manage grammatical and linguistic structures of expressions. On the other hand there are subjects of which the whole text is about. There are also cue phrases which points the relevance of sentences in a text in relation to the text. Therefore it can be perceived that different measuring units can be used for text analysis. Some of the diagnostic units are as follows:

2.4.1 High frequency words

This diagnostic unit, suggested by Luhn (1959) assumes that if a word appears many times in a text it is more likely that the text is about the word. Logically high frequency words which are not concepts but appear a lot of times in a text will have higher values in this mere context. But in a natural way of writing key ideas of a paper are mentioned now and then in the text. Observed from different perspectives and accompanied with explanations the coins of the text are mentioned over and over again.

By taking high frequency words only, it is impossible to point out the major concerns of a text. The very first problem that can be told blindly is the way of removing stop words (language necessities). Stop words are words which do not convey any message but appear many times in a text. These words should not be used as tools of summarization. Luhn (1959) came up with an idea of removing these words by the use of a fixed threshold. Removing theses words by using a threshold increases the risk of losing very important words or the inclusion

of stop words in the diagnosis process. A list of stop words is developed for this system and stored in a file. The developed system consults the file of stop words before using any word as analysis tool. Words found to match with the existing list are automatically removed. As a consequence there will be less chance of incorporating stop words as key words.

2.4.2 Cue phrases

Edmundson (1969) who first observed and used these heuristic for text analysis mentioned that there are phrases which highlight the importance of a sentence. So the worthiness of sentences may be measured from the bonus words which are typical marks of importance. On the contrary to this Edmundson also observed words which decrease the significance of a sentence. He called these words Stigma words. Sentences containing these words should not be included in the summary because there is already a clue that they are not important.

2.4.3 Titles and headings

Undoubtedly the title and the header of a text provide a good amount of information about its content. By extracting the content words from titles and headers we can create a scoring formula which increases the score of sentence that contain these words. We can also put headers and titles directly into the extract thus creating a more coherent and structured output. Finally the system can be guided to select at least one sentence from every major section. So given this sentence is good, the extract is more likely to maintain all the basic subjects.

2.4.4 Location heuristics

The position of a sentence within a paragraph and the position of a paragraph within a text can sometimes be indicative of significance. Though this

is dependent on specific practices of the author it happens usually with the first sentence of a paragraph which supposedly gives a very good clue of what is going to follow. Another similar assertion is that good proportion of important information about a paper is to be found in the first and last paragraphs (they indeed contain introductory and conclusive materials) so their sentences should be favored more than the sentences of the other paragraphs.

2.4.5 Linguistic analysis

Linguistic information can also prove usefulness on the basis of looking for strings of words that form a syntactic structure. Extending the idea of high frequency words we can assume that noun phrases form more meaningful concepts, thus getting closer to the idea of terms. This overcomes several problems of the first single word methods because it can utilize compound nouns and terms which consists of adjective and noun. There is also a possibility of analyzing the use of one term with more than one noun phrase.

2.4.6 Text formats

Commonly, while writing a text emphasis is can be made to important points by changing the format of the text say like font changes, bolding, italicizing and underlining. This shows that these formats could be given more weights than the others which are in favor of the sentences containing such formats. But the main problem in using text formats as diagnosis tools is standard of use. No fixed standard for emphasizing is used through out writing texts. This creates difficulty in tracing text formats from different domains.

2.4.7 Word multiples

The idea behind the concept of word multiples (usually pairs) is that there are words which appear together many times. If words occur frequently, either as a specific term and mainly together, they contribute some weight of importance. Good example of word pairs are noun compounds like the word "yepres higu" equivalent to "the press law" in the adjusting weights example used for this study (see section 4.6.3). This is suggested from the ground of the writing styles and the author preference of use. Like most other pure NLP techniques this technique is very difficult to apply in statistics based approaches when there is a need of figuring out semantics. But from statistical considerations it is of much benefit to analyze such multiple usages. 2.6 Review of ATS Evaluation

2.5 Weight assigning technique

The final weights of the sentences are given by sum of the weights contributed by each of the text mining elements discussed in section 2.4. The weight of the sentences starts with zero. The weight that corresponds to a mining element will be added to the sentence weight if that mining element is found in the sentence.

The following formula is used to calculate the weights of the sentences

$$\text{Swt} = \sum_{i=1}^{n} wh_i$$

Where Swt is the sentence weight and *wh* is the weight assigned to a diagnostic unit that is found in the sentence. For a diagnostic unit that is not existing in the sentence the corresponding weight will be zero.

2.6 Review of ATS Evaluation

Evaluating NLP systems in an objective way is not an easy task. Researchers have just recently begun to define and systematize it. Precision/recall, similarity, and subjective evaluation measures are reviewed for adoption to this system.

- **Precision and recall**

The classical information retrieval measures of success (precision and recall) can be used to measure the performance of a Natural Language Processing system. Recall measures the ability of the system to extract useful elements. On the contrary precision measures the ability of the system to reject unnecessary elements.

$$\text{Recall} = \frac{\text{Number of correct items extracted}}{\text{Number of target extract}}$$

$$\text{Precession} = \frac{\text{Number of correct items extracted}}{\text{Total number of extracted items}}$$

To apply these measures to the current system basically there must be some target extract with which we compare the performance of the system. These measures try to figure out the matching points of the system extract with human extracts. It should be noted that human extracts must be real extracts not abstracts. Comparison also tells how much the system resembles the human act of summarizing.

- **Similarity measure**

Using vector representation of the document and the summary we can compare the similarity between the manual extracts and the system extracts. The dot product of the document vector and the extract vector can be used as a measure of the system performance.

$$S(D_i, D_j) = \sum_{k=1}^{t} W_{ik} . W_{jk}$$

Where S is the similarity, $\mathbf{D_i}$ and $\mathbf{D_j}$ are the document vector and the extract vector respectively. W_i and W_j are weighted terms in the corresponding documents.

- **Subjective evaluation**

Subjective evaluation is a method where a professional human summarizer reads the system extracts and judges for the performance of the system. Of course this is done by consulting the original text and measuring the extract in qualitative as well as quantitative views. A noteworthy feature of this evaluation is its unbiased ness; it is not influenced by a pre-determined output like the human extract.

2.7 Review of Sample Summarization systems

2.7.1 Luhn's frequency method

The first work on automatic text summarization was done by Luhn (1959) he laid a corner stone for the development of automatic text summarization. Luhn showed that words appearing many times in a text furnish good idea about the content of the document. He also proposed that there are words that appear very frequently but bear no content. He tried to cut off these words by determining a fixed threshold.

The idea of Luhn was acknowledged and used in many automatic information processing systems. This well known idea is used as one of the extraction parameters in the development of the Amharic news text summarizer which is the objective of this study. High frequency words in the Amharic news

text are used as key words of search. Of course, attempts are made to remove stop words from the list. The Luhn's threshold method is not used in the system under study. Rather a descriptive technique of removing such stop words by consulting a file containing list of no content bearing words is used.

2.7.2 Edmundson's extraction system

The next prominent work in the history of automatic text summarization after Luhn was the work by Edmundson (1969). With the starting point by Luhn, Edmundson tried to utilize the use of cue phrases, titles and location heuristics.

Edmundson carefully outlined the human extracting principles. He observed that the location of a sentence in a text tells some clue about the importance of the sentence. For example title sentences and mere paragraph sentences can not be measured to be equally important. On the other hand there are phrases which mark out the value of the sentence in a specific text.

After observing the text features in a text Edmundson finally suggested four tools of diagnosis (location, cue phrases, key words and title) which can be applied independently and collectively. These tools are directly adapted to the Amharic news text summarizer under this study by determining location of important elements in Amharic news and by identifying most common cue phrases in Amharic news text.

2.7.3 Salton's vector space model

SMART developed by Salton et al (1993) is not strictly an extraction system. But it deals with word ambiguity and similarity between sentences in the area of automatic extracting.

It was shown that retrieving passage is a right step towards better response to user quarries. The construction of SMART was based on vector space model. Every document is represented as vectors of weighted terms. By finding the similarity between the user query and the documents the items with high similarity will be extracted. The major idea emphasized in Smart is the vector representation of a document. By finding the similarity between the summary and the main document it is possible to see how close the document and the summary are. This view seems applicable on the evaluation of summaries.

In addition to measuring the closeness of the summary and the main document, what to extract can be determined by applying the measurement at sentence level or so. In this way the vector representation can be used as extraction mechanism.

2.7.4 SweSum summarizer

SweSum is a summarizer developed for Swedish by Dalianis (2000). SweSum is built on statistical and linguistic methods. The system makes use of word stemming to decrease the dissimilarity of words due to modifications to fulfill language necessities. The frequency count of a word is made after all words are brought to their stem.

The main idea in SweSum is extracting high scoring sentences evaluated both on statistical and linguistic diagnosis systems. Sentence scores are determined by the statistical techniques like giving more scores to sentences containing highly frequent content bearing words and linguistic diagnosis techniques like giving more scores to nouns.

HTML tags are used to control the format of the page. A different feature in SweSum than other previous systems is its consideration to boldfaced text and

numerical data. In SweSum sentences containing numerical data are given more weight and sentences containing bold faced items are considered as more important and hence given more weight.

2.7.5 Kupiec's document summarizer

The kupiec (1993) system was highly influenced by Edmundson extraction system. Document extraction is viewed as a statistical classification problem in the Kupiec (1993) system. I.e. for every sentence its score means the probability that it can be included in a summary

The heuristics, called features in the paper, utilized were sentence length cut off feature, fixed phrase feature, paragraph feature and thematic word feature. Upper case word features are also considered as important in the kupiec approach. In addition to the devised techniques most of the tools in preceding systems are included in the Kupiec (1993) system.

2.7.6 Using cohesive properties of text for Automatic summarization

The SUMMAC and more recently the DUC competition are some active lines of research in the area of automatic text summarization. One of the many systems that appeared on the DUC competiton is the summarization system Using cohesive properties of text by Fuentes (2001)

This system uses cohesive properties of the text for selecting the most informative fragments included in the summary. The system principally uses lexical chains as indicative of lexical cohesiveness. Complementary tool named co-reference chains are also used in this system.

After deciding the informative elements and assigning weights on the basis of the cohesiveness what to extract is determined in the system. The product of the

system is also mentioned to be a paragraph extracted from the main text. Though not stated in the Fuentes (2001) paper it is mentioned that finer grained units, sentences and clauses were tested during the experimentation. Several compression rates were also used during experimentation.

2.7.7 The LAKE System at DUC2004

The LAKE which is Learning Algorithm for Key Phrase Extraction is a supervised approach that makes use of linguistic processing approaches to document summarization. D'Avanzo etal (2004) prepared this system for the DUC competition. The system works in two phases. First it considers a number of linguistic features to extract a list of more motivated candidate key phrases and then it uses machine learning framework to select significant key phrases for that document.

The so called candidate phrases in the system are sequences of words that match a set of manually define linguistic patterns. The supervised learning algorithm is used to score the head of each phrase according to TF/IDF and the position heuristic. Finally the score of the head is assigned to the whole candidate phrases and the best scored phrases that fill the 75 characters extract are given as output of the system.

It can be mentioned that the LAKE system uses deeper linguistic analysis than the simple statistical systems.

2.7.8 The copy-and-paste system (1999)

The Copy and Paste system by Mckeown (1999) presented a framework for fast construction of concise and coherent summaries of single documents in any domain. Extracted document sentences are not directly used for producing

summaries in this system. The cut and paste system formulated editing method of extracted sentences. This is because simple sentence extracts are not coherent..

The cut and paste system is designed to take the results of a sentence extraction summarizer, and extract key concepts from these sentences. These concepts are then combined to form new sentences. The system thus, copies the surface form of these key concepts and pastes them into the new sentences. This is done by first reducing the sentence removing any extraneous information. This step uses probabilities learnt from a training corpus, and lexical links. The reduced sentences are merged by using rules such as adding extra information about speakers, adding conjunctives and merging common elements. The cut and Paste system is a single document summarizer that is domain independent. In addition the cut and paste system can be used with any single-document summarizer, serving as summary generation component.

2.7.9 Workshops and conferences

Maybury and Mani (2001) summarized conferences and workshops held on summarization issues. The diagrammatically presented history put the conferences and workshops in reference to the time axis.

CHAPTER THREE

STRUCTURE OF AMHARIC TEXT

3.1 Introduction

As this system is developed for Amharic language text examining the different parts of Amharic text, with emphasis on news items, contributes a lot in the identification of the valuable elements of text.

It is obvious that a text is constructed following grammatical rules in the language. And also there are common practices of organizing the parts of a text. Tracing this path of construction allows the use of location heuristics in identifying the important elements of the text.

Text parts, linguistic characteristics, alphabets, numerals, grammatical rules, punctuation principles and basic text organization techniques in the language that are very useful in building the system are discussed in this chapter. The origin of the language is used as a kicker to the topic.

After a through inspection of the language fundamentals, in view of the summarizer, alternative ways of resolving difficulties in development of the system are pointed out. The innate architectures of the language which are like a bottleneck to automatic text summarization systems are also tried to be given solutions.

3.2 Origin

Bender et al. (1976) stated that the Ethiopic script developed from the Ge'ez. Ge'ez is a language derived from Sabean script that has been used since 4th century A.D. Amharic which is one of the Semitic languages, is the national language of Ethiopia.

Through a long period of use Amharic has undergone changes. The shape of some of the letters has been changed. The number of the symbols in the language also changed. Some of the symbols are dropped and some are added. (Bender and et al1976).

3.3 Writing system

3.3.1 The nature of Amharic script

The script of Amharic language is phonetic in nature. It has 32 consonants and 7 vowels. Each of the 32 consonants has seven orders. Sebsibe et et al (2004) marked that the total number of symbols of the language exceeds 230.

The Amharic script is written from left to right. Each symbol in Amharic represents a syllable consisting of a consonant plus a vowel. Simon (1998) has clearly stated the basic symbols in the writing system and the way how the basic symbols are affected in several ways to include the vowels in the basic symbols.

The Amharic script uses no uppercase and lowercase letters. All the letters are just in the same case. The beginning letters of sentences and nouns are the same with all other letters.

3.3.2 Text parts

Like the texts in most other languages, Amharic text is also divided into parts. These partitions allow grouping of related items together. These groups of related items are placed in a controlled order so that the reader can reach the intended message. Some common parts of a document are shortly discussed below.

Title: the topic, which is termed as "res" in Amharic, is the subject about which the other parts of the text are constructed around. The title is usually a

single word or phrase. In rare cases the title contains a number of sentences. But most often the title is presented in concise language. Solomon (1991)

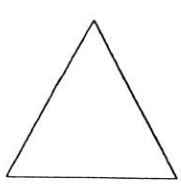
Introduction: "megbia" the Amharic equivalent of introduction commonly contains what is intended to be elaborated in the main body of the text. The start to end idea of the body is highlighted in the "megbia". Though not identified by a subtitle in news item, the first paragraphs of news contain introductory sentences.

Detail: the detail called "hateta" in Amharic includes all the details and explanations given for the concern of the subject matter that the text addresses. All attempts of explanations are placed in this section. Different illustrations like diagrams and examples are included in this part of the text. Hence when the reader goes through these elements he will have a better understanding of the matter by observing the topic from various perspectives.

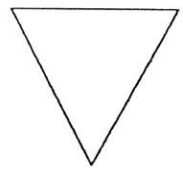
Conclusion: the conclusion called "medemdemia" is a portion where the generalization of the explanations is placed. After assessment of topics the finally reached stage is placed in the conclusion. The summary is a piece where the important points assessed by the text are placed. The main ntention of the summary is to rehearse the very important topics in the body so that the reader grasps the conveyed message.

Paragraph: the "ankets" is a sub division in the "megbia", "hateta" or "medemdemia". This division strictly addresses a specific bit of idea. According to Solomon (1991) the paragraph is a group of sentences that form a distinct unit developing one major idea. The principal idea in the paragraph is called "hile kal". The sentence containing the "haile kal" is called topic sentence. The "hilekal" may be found in the first sentence of the paragraph. In such a case the

paragraph is said to be upright triangular. The paragraph will be diamond shaped when the "hilekal" is mentioned at the beginning and end of the paragraph. Inverted triangular paragraphs start with the explanation and come to the "haile kal" at the end. The structure of the paragraph ("ankets") can be described diagrammatically as follows:



a) Up right triangular paragraph



b) inverted triangular paragraph

c) diamond shaped paragraph

Sentence and its parts: the sentence is called "Areftenegeer" in Amharic. The sentence is a smaller division in the paragraph. Like in most other language the Amharic sentence is has go subject which is the "Balebet", verb that is "gis" and object "

As the major concern of this study on news items it is of much benefit to look the structure of Amharic news items. The news items usually have one title and no subtitles in them. There may be a number of paragraphs and sentences. As this system deals with single news item and the Amharic news items are having one title, the concept of long document segmentation technique of breaking the text into different parts is not a problem in this study. Appendix [3, 4, 6] are typical examples of Amharic news items taken form Reporter (2004). And appendix [5] is examples of radio news items taken from Radio Fana.

3.3.3 Grammar

According to Gasser M (2003) An Amharic verb root consists of a set of consonants. Depending on the tense, and other grammatical features, the

consonants may be separated by particular vowels and possibly geminated (doubled). A verb form normally has one or more suffixes and possibly one or more prefixes as well, agreeing with the subject and sometimes the direct or indirect object of the verb. There are at least ten different classes of verbs, each modifying its stem in a different way for the different forms. Amharic nouns are relatively simple by comparison, though they may take suffixes indicating possession ('my', 'his', etc.), plural, and a few other grammatical functions. These additions and modifications to words in the text are somewhat a problem in counting the frequency of words. A stemmer reduces this problem providing roots of words.

3.3.4 Punctuation marks

Like the punctuation marks in English and other languages Amharic also uses symbols to control the grammatical setup and flow of the text. Some of these are borrowed from English (/, ?, \) and some are innate to the language. Examples are (:, ::, :-)

In current Amharic writing system words are delimited by space but previously ':' Amharic 'hulet netib' was used to delimit words. The comma is denoted by 'netela serez' '፣' and the end of a sentence is marked by '::' 'arat netib. Paragraphs are separated with horizontal space. Pages proceed from left to right.

For the purpose of statistical counting it is meaningless to count the number of these punctuation marks in the text to be summarized as these are mere language necessities without any content to convey by themselves. Therefore the system is designed in such a way to ignore punctuation marks for the purpose of ranking the words in the text. But the punctuation marks are used to define the

different parts of the text. For instance the '::' denotes the end of a sentence and this automatically substituted to '</S> which is used as the mark for sentence end.

3.3.5 Homophonic characters

In Amharic writing system there are letters which correspond to the same sound. Letters like "ሀ" "ሐ" Correspond to the same sound "ha". Any word containing such sounds could be written in more than one way. This practice in the writing system affects the system under study. As the system is based on statistical techniques if a word is having two or more forms in a text it is considered as different words. This action decreases the frequency count of the word. Low frequency words are treated by the system as specific words and non-eligible to be key words of extraction.

3.3.6 Abbreviations

Abbreviations must be taken into account while pre-processing the text to be summarized by the system. If concepts appear once abbreviated and another time fully, the word or phrase is going to be considered by the system as two different words or phrases which decreases the frequency count of the word.

3.3.7 Word forms

As a result of language grammatical rules words are modified to express different situations like tenses, plurality and possession. There are additions or modification (sometimes whole changes) that we make to words to incorporate additional descriptions to the same word. This is usually observed on verbs while making changes for tenses.

Statistical systems consider exact matches for counting the frequency of a word. Unless a resolving system is used to bring the different forms to the same

root the statistical system is about to treat the words differently which is against the frequent appearance of a word.

Stemming is a technique which tries to remove the inflections and infixing made to a word. The words that existed differently will be counted as one if the root of these words is considered. This phenomenon increases the frequency of such words (words in different forms) by helping to count them as one word. Therefore use of morphological analyzer in the designed system is about to contribute at this spot.

3.3.8 Numbers

Amharic has got its own way of representing numbers. Though the Hindu Arabic numbers are used commonly in Amharic texts, the original Amharic numerals also appear in some texts. It is important for summarization systems to follow a uniform way of representing the numbers.

Numeric data is given more value in systems like the SweSum by Dalianis, H. et al (2000). Likewise in the system under this study a very small additional weight is given to numeric data. In addition to the fixed small weight, numbers are fevered like any other word in the text, for their frequent appearance.

3.4 Amahric news items

Commonly Amharic news items are organized having a title which is not more than a sentence or two. The body has got different paragraphs. Sub titles are not that much observed in news items. This character of the news items drops the use of long document segmentation. On the other hand it points title reflection in the body. Hence the use of title words is encouraged.

Apart from concentrating on the title, the structural organization of the news items is of uniform pattern. This encourages the use of location heuristics in

identification of the important elements. The very first and the very end sentences seem to encompass the explanations and make most of the news items diamond shaped in their appearance.

In addition to the common text mining tools, the uniform like format and expression of the news items invites the concentration on cue phrases as diagnosis elements.

CHAPTER FOUR

ARCHITECTURE OF THE SYSTEM

4.1 Introduction

This chapter presents the structure of the developed Amharic news text summarizer. The way how the developed extraction tools are organized and integrated is discussed in this chapter. The model is built having two features. One is extraction and the other is leaning feature.

Linguistic parameters, location heuristics and statistical measures are selected to be included for the design of the model. Based on the adopted techniques an algorithm of extraction is developed and using perl programming language the algorithm is coded for practical application of the extraction. The different language features used, basic functioning principle of the system, the conducted tests with the results obtained and the evaluation of the system are included in this chapter.

4.2 Description of the system

It is mentioned earlier that the system to be constructed by this study is an extraction system. By utilizing the reviewed diagnostic units the summary of a news text is generated as an extract from the original news. Potentially useful techniques are used to design and build the model of the automatic Amharic news text summarizer.

The model has got two basic features the first is an extraction feature where the system applies sentence weighting formula shown in section 2.5 and assign weights to the sentences in the text. The high scoring sentences will be extracted as summary of the news text. The second feature is the learning component of the system. The system tries to update some of the tools it uses for

further specialization and good performance. Hence when the system is kept in use, it improves the quality of its performance. At least, the chance of using stop words as text mining elements will be decreased.

4.2.1 Extraction features

Most of the previously reviewed diagnostic units are included in the model developed under this study. Measured on this units sentences with high score comprise the summary. It should be noted that there must be a reduction threshold which will be determined by the user as a percentage of condensation. Reduction factor allows the user to adjust the size of the news item to his/her preferred size. The following is list of the diagnostic units used for the purpose of extraction. These items are weight assigning tools. For each existence of these items the weights of sentences are increased by the corresponding amount.

- Titles
- Title words
- Cue phrases
- Headers
- Paragraph initials
- Words in the header sentences
- Paragraph end sentences
- High frequency words

4.2.2 Learning features

An attempt has been made to make the developed model dynamic. Using specific number of tools is believed to make the system static and limited to a fixed performance level. But if the system tries to develop and improve some of

the tools by interacting with the user, the system will at least improve its performance. In this case the user must be an expert user who can comment on the words. To the minimum the user must be able to decide the items that must be added to the stop words list of this system. Through time the system enriches its experience and takes the human dimension of discarding stop words.

The learning feature that is tried to be included in this model is the appending of list of stop words and cue phrases in the language. Therefore whenever the system summarizes one news text the user is prompted to update the tools for further use. This results in rich elements of the tools that exist in the system.

4.3 Description of the algorithm

It is stated earlier that there are a number of parameters (diagnostic units) that can be used for extraction purpose. Depending on the nature of the language in use and the type of the text to be summarized, the diagnostic units to be applied differ. We may need long document segmentation for long texts. We may also need inter-document analysis while working with many documents.

This system is concerned with single document Amharic news text summarization. The title, title words, cue phrases, header sentences, paragraph ends and key words are used for analyzing the text and decide the parts to be extracted. The following algorithm is developed to build the system prototype:

1. Take input file name and condensation ratio
2. Try to open the file

   If open able open it and go to the next step otherwise die with "alkefet ale"
3. analyze condensation ratio

If condensation ratio is between zero and hundred go to next step otherwise die with "kezero – meto bicha"

4. go through the sentences in the file start to end for each of the following and do:

   If tag <RES> found give the sentence title weight and make a list of title words else go to next sentence up to the last line.

   If tag <MEG> found give the sentence header weight and make a list of header words else go to next sentence till the last sentence is found

   If tag<x> found give the sentence paragraph initial weight else go to next sentence up to the last sentence.

   If tag </x> found give the sentence paragraph end weight else go to next line up to the last sentence.

   If tag <S> found give the sentence starting weight and make a list of sentence words else go to next line till the last sentence is found.

5. Sort lists of words uniquely starting with the highest frequency first.

6. open stopwords file and compare list of words with list of stopwords

   If equal remove the word from list of words else keep the word

7. Determine the keywords to use by multiplying the number of words with percentage of words to use and cut words below the result from the list of words.

8. Go through each sentence of the text and increment weight of sentences for each keyword, title word, header word and cue phrases according to the pre-assigned weight of these elements.

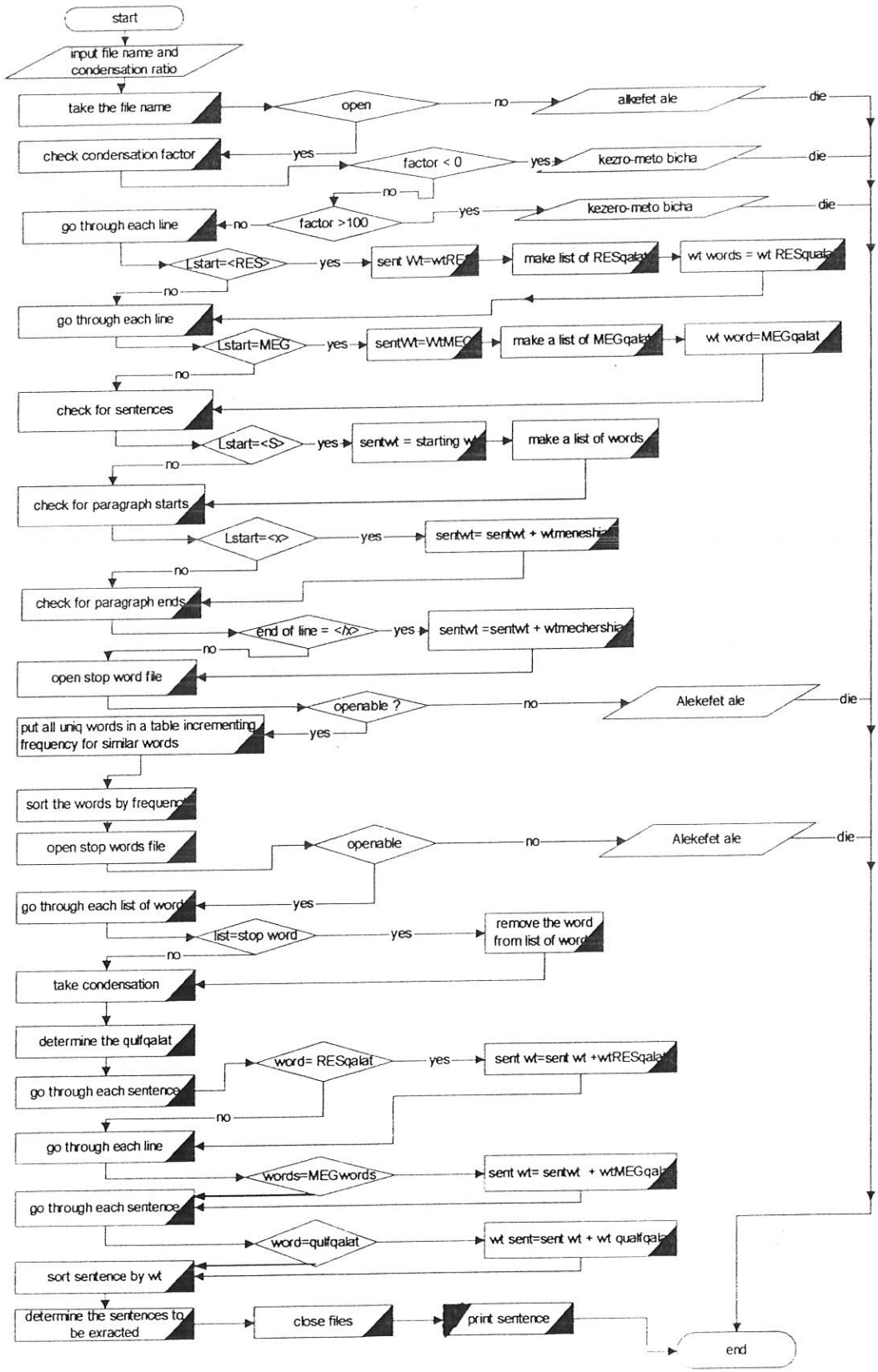


Figure 4.3 the structure of the system

4.4 Data requirements

Obviously the corpus must be in an electronic format to be processed by the system. The nature and the characteristics of the corpus have direct relationship with the system design. The text which this system is designed for is a news text. The text is expected to have title and body. The title may contain sentences and words. The body may contain paragraphs sentences and words. Because the system is single document concerned each news item must be kept in a separate file. There is no cross document analysis that the system takes care of. Thus the summary of a single news item with out any comparison to other news items is expected form the system at one time run.

4.5 Preprocessing required

The digitally obtained data can not be directly fed to the system. The collected news texts are first pre processed to make them useable by the system.

First the text has to be Latinized. This is because the programming language used (perl) does not support Amharic fonts. Amharic font mapping also required another module in the program which requires UTF8 input output at each transaction. Therefore every input output is carried out by using the Latin alphabets.

The most important preprocessing is segmenting the corpus. Using a local markup technique developed for this study the different elements to be considered are given an ID. Searching through the text is mainly dependent on the identifier given for that part of the text at a word level.

Like all other texts Amharic news text is produced with the appropriate punctuations and grammatical constructions. Even if there are spelling, punctuation and other grammatical mistakes in a text, readers can tolerate making

corrections by common sense and understand the theme of the news text. Some times humans make corrections based on the context of sentences while reading. When we come to the system under this study corrections are not expected to be made by the system. Every task to be carried out should be notified earlier.

There are different ways of assembling the components of a text in different languages. Suppose in English the end of a sentence is marked by a period and the beginning is an upper case letter. This clearly identifies the start and end of a sentence for the human reader. But in view of the machine processing a lot more things should be analyzed, for example the capital letters of names (proper nouns) would mean the start of a sentence to the machine. This creates one sentence in the text which is endless another sentence without starting capitalization. Another difficulty is space. There is no definite number of spaces to follow a paragraph. Identifying the beginning and end is also a difficulty if the space is followed. It is possible to site more problems like this which makes the natural language processing task to be more complex in machine aspect.

In Amharic writing system there is no upper case thing. Therefore a sentence starts with the same letter as any other word begins. Following natural language punctuation marks as the readers do for the purpose of automatic text analysis requires deep considerations of punctuation marks in the language. Especially for searching through texts, it is far from being easy. Marking the required elements by special identifiers designed for this system is preferred. This process of marking is termed as segmenting. The following are two important preprocessing steps stated with their rational:

4.5.1 Segmenting

For the rationale of making search easier, the parts of the text are given their own ID. And while searching the system is going to follow this tags. Titles, sentence, paragraphs and words are given independent identifiers for both the category and their start/end.

For the purpose of this study the corpus text is required to be segmented into titles, paragraphs, header sentences, sentences and words.

Major advantages of tokenizing:

- **Uniformity of search**: whenever there is a need to search a diagnostic unit the system is going to search the tag. Anything under the specified tag is a unit that the search is fired for regardless of the various text formats.
- **Independence**: The system is going to look for a defined tag not a text, position of a text or punctuation mark. This makes the system independent of the text under consideration and the format of the text.

It should be noted that this approach is very advantageous in Amharic texts for there are no special identifications like capitalizing the beginning of a proper noun or the beginning of a sentence (i.e. no uppercase and lowercase letters are found in the writing system of Amharic).

- **Easy maintenance:** the use of identifiers to the parts of the text to be analyzed makes the maintenance of the system easier by making the elements mutually independent of each other. Any action taken on one element remains to that element only. Independent actions can be taken on any part of the text. Any action on one element persists to that element only.

In the system developed under this study every entity that is to be considered as diagnostic unit is identified and given a tag as follows:

- <RES> beginning of title
- </RES> end of title. taken from the Amharic word "RES" to mean title
- <MEG> beginning of the header sentence
- </MEG> end of header sentence taken from the Amharic word "MEGEMERIA" to mean Starting.
- <S> beginning of a sentence
- </S> end of a sentence
- <Q> beginning of a word
- </Q> end of word taken from the Amharic word "QAL" to mean word.

Words are given specific start and end tags to differentiate them form the spaces between paragraphs and multiple spaces while processing in a stack. The approach is similar to the SGML (Standard Generalized Markup Language) environment. But in this system – since the above mentioned structures only are required for the purpose of weighing sentences, the SGML tool cannot be used as it is. The data is segmented by another local program designed for this study (MYXML).

The HTML format is also a segmented format but not for words which makes it unfit to the system in this research. The local markup language attaches an ID to all the elements the system needs.

4.5.2 Data format

There are some important points to be raised about the data format of the text to be fed to the system. It is unquestionable that the data must be available digitally. The system is designed in such a way that it is not sensitive to the size

and color of the fonts used. Italics, underlines and other reader attention halts are not at all recognized by the developed system. Once tags are given to the parts, the position of any part of the text doesn't matter. But the file must be in plain text form and the availability of one sentence per line is a must. This is thoroughly taken care of by the MYXML parser

4.6 Experimentation

4.6.1 Why perl is preferred

The perl programming language is used to develop the source code of the system model under study. According to Larry (1996) who is accounted to be the inventor of perl programming language "computer languages differ not so much in what they make, but in what they make easy. Perl is designed to make the easy jobs easy without making the hard jobs impossible."

Perl is stated to be an easy language to learn and use. Larry (1996) mentioned that in perl we just say what we want to say, we don't have to do some prior declarations like in other languages. The completeness of the perl is symbolized by the camel which evolved to be a relatively self-sufficient animal.

This research is mainly concerned with text processing and perl is basically designed for that. The programming language is selected by the researcher because of the aforementioned reasons and the researcher's preference. Not only for the language is fit for text processing but also in consideration to its easiness for coding and the time allotted for this study perl is selected for designing this system.

4.6.2 Adjusting weights of diagnostic units

Four news items were used to train the model. By changing one of the weights of the parameters (title words, paragraph starts/ends, header sentence, keywords, header sentence words and percentage of key words to use) at a time, 124 iterations were made. The summary of the iterations is extracted independently. Because the system is a single document summarizer, cross-document analysis is not considered.

The following is one of the four news items used to train the model:

የፕሬስ ሕጉ የህዝቡንና የሚመለከታቸውን አካላት አስተያየት ማካተት አለበት
አዲሱ የፕሬስ ህግ ረቂቅ የህዝቡንና የሚመለከታቸውን አካላት አስተያየት ያካተተ ሊሆን እንደሚገባ እውነተኛ ዲሞክራሲ ባለባቸው ሀገራት ጥሩ ህጎች የሚወጡት በሕዝብ አስተያየት ስለሚዳብሩ መሆኑን ሚስተር ሀርሙት ሑስ አመለከቱ። እውነተኛ ዲሞክራሲ ባለበት ሀገር የፕሬስ ህግ እንደማያስፈልግ ጠቅሰዋል።

በፍሬደሪክ ኤበርት ስቲፍቱን ፋውንዴሽን የኢትዮጵያ ተወካይ የሆኑት ሚስተር ሀርሙት ይህን የገሌት የመገናኛ ብዙሀን ባለሙያ ሴቶች ማህበር ባለፈው ሳምንት ባዘጋጀው ስብሰባ የነፃ ፕሬስ አስፈላጊነት አስመልክተው በሰጡት አስተያየት ነው።

ሚስተር ሀርሙት በዚሁ ወቅት እንደገለፁት አዲሱ የፕሬስ ህግ ለመፅደቅ በሂደት ላይ ነው። ለሚኒስትሮች ምክር ቤት ቀጥሎም ለተወካዮች ምክር ቤት ቀርቦ እንደሚፀድቅም ሲነገር ቆይቷል። ይህን አይነቱ ህግ በነፃ ፕሬስ ዋነኛ ተጠቃሚ ለሚሆነው ህዝብና ለሚመለከታቸው አካላት ቀርቦ ጥልቅ ወይይትና አስተያየት እንዲሰጥበት ማድረግ የመልካም አስተዳደርና የእውነተኛ ዲሞክራሲ መኖር ምልክት ነው። ይህን ማድረግም የፓርላማውን የመወሰንና ረቂቁን አሻሽሎ ህግ አድርጎ የማውጣት ሥልጣን መጋፋት አይደለም።

ይህ እንዳለ ሆኖ እውነተኛ ዲሞክራሲ ባለበት ሀገር የፕሬስ ህግ ማውጣት አስፈላጊ አለመሆኑን ሌሎች ህጎች ለፕሬሱም እንደሚውሉ ሚ/ር ሀርሙት ተናግረዋል። "ይልቁንም ከእንዲህ አይነቱ ህግ ይልቅ ከአሳታሚዎችና ከጋዜጠኞች እንዲሁም ከሲቪል ማህበረሰቡ የተውጣጣ የሚዲያ ካውንስል ማቋቋም ሚዲያውን ለማሻሻልና ጥሩ ለመሥራት እንዲችል ማድረግ ያስችላል" ብለዋል ሚስተር ሀርሙት።

አብዛኛውን ህዝቡ ሊያስማሙ የሚችሉ ሀሳቦች የሚፈልቁትና የሁሉም የህብረተሰብ ክፍል አስተያየት የሚደመጠው ብሎም የሚካተተው ግልፅ ውይይት ሲኖር ብቻ መሆኑንም ሚ/ር ሀርሙት።

ለሁለት ቀናት የቆየውን ስብሰባ በንግግር የከፈቱት የማስታወቂያ ሚ/ሩ አቶ በረከት ስምኦን በበኩላቸው ዮፕሬስ ህጉ በመጪው ጁላይ (ከሰኔ 24 እስከ ሐምሌ 24) ለውይይት በድጋሚ እንደሚቀርብ ገልፀው የማፅደቁ ሂደት የሁሉንም የህብረተሰቡ ክፍል እንደማያሳትፍ አመልክተዋል። ከዚህ በፊት በዚሁ ረቂቅ ህግ ላይ ለመወያየት ሁለት ወርክሾፖች መካሄዳቸውንም አስታውሰዋል።
የኢትዮጵያ መገናኛ ብዙሃን ባለሙያ ሴቶች ማኅበርን የተለያየ የፖቴቲካ አመለካከት ያላቸው ሴቶች በአባልነት ይዞ መንዙን አድንቀው ሌሎች ማህበራት ይህን ማድረግ አልቻሉም ብለዋል።

The above Amharic news text is Latinized in order to be processed by the developed model. The sentences are numbered just for reader's comfort.

Ye pres higu yehizbunia yemimeleketachewn akalat asteyayet makatet alebet

1. Adisu yepres hig rekik yehizbunina yemimeleketachewn akalt asteyayet yakatete lihon endemigeba ewnetegna dimokarsi balachew hagerat tiru higoch yemwetut mhizbu asteyayet selemdabiru mehonum mister harmut hes ameleketu
2. ewnetegna dimokrasi balebet hager yepres hig endemayasfelig teksewal
3. Befrederik erbert stiftun fawundeshin yeityopia tewekay yehonut mister harmut yihin yalut yemegnagna bezuhan balemuya setoch mahber balefew samint bazegajew siseba yenetsa pres asfelaginet asmelktew besetut asteyayet new
4. Mister harmut bezihu wekt endegeletsut adisu yepres hig lemetsdeq behidet lay new
5. leministroch mikir bet ketlom ltewekayoch mikit bet kerbo endemitsediq sineger qoyitowal
6. yihin aynetun hig benetsa wanegna teterkami lemihonew hizbna lemimeleketachew akalat qerbo tilk wuyiyitna asteyayet endisetibet madreg yemelkam astedaderna ewnetegan dimokrasi menor milikit new
7. yihin madregm yeparlamawn uemewosenna rekikun ashashilo hig adirgo yemawtat siltan megafat aydelem
8. Yih endale hono ewnetegna democracy balebet hager yepres hig mawtat asfelagi alemehonun lenoch higoch lepresum endemiwulu mister harmut tenagrewal
9. yilikunm keindih aynetu hig yikik keastamiwochna kegazetegnoch endihum kesivil mahberesebu yetewtatu yemidiya kawunsl maquaquam midiyawun lemashashalna turu lemesrat endichal madreg yaschlal blewal mister harmut
10. Abzagnawn hizb liyasmamu yemichlu hasaboch yemifelkutna yehulum yehebreteseb kifl asteyayet yemidemetew blom yemikatetew gilts wuyiyit sinor bicha mehonunm mister harmut tenagrewal
11. lehulet kenat yeqoyewn sibseba benigigir yekefetut ymastaweqiya ministru ato bereket simon bebekulachew yepres higu memechiw julay kesene 24 eske hamle 24 lewuyiyit bedigami endemiqerb geltsew yematsdequ hidet yehulunm yehibretesebu kifl endemiyasatf amelkitewal
12. kezih befit bezihu rekik hig lay lemeweyayet hulet workshopoch mekahedachewnm astawkewal
13. yeetiopia megenagna bizuhan balemuya setoch mahbern yeteleyaye yepoletica amelekaket yalachew setoch abalat yizo meguazun adreneqew leleloch mahberat yihn madreg alchalum bilewal

The summary of the above news is manually generated by two linguists.

Full sentence extraction technique is used for the summay generation. The following five sentences are extracted with the title. (It should be noted that the controversial points between the two human extracts are resolved by removing the sentence).

Ye pres higu yehizbunia yemimeleketachewn akalat asteyayet masakat alebet

1- Adisu yepres hig rekik yehizbunina yemimeleketachewn akalt asteyayet yakatete lihon endemigeba ewnetegna dimokarsi balachew hagerat tiru higoch yemwetut mhizbu asteyayet selemdabiru mehonum mister harmut hes ameleketu

11- lehulet kenat yeqoyewn sibseba benigigir yekefetut ymastaweqiya ministru ato bereket simon bebekulachew yepres higu memechiw julay kesene 24 eske hamle 24 lewuyiyit bedigami endemiqerb geltsew yematsdequ hidet yehulunm yehibretesebu kifl endemiyasatf amelkitewal

6- yihin aynetun hig benetsa wanegna teterkami lemihonew hizbna lemimeleketachew akalat qerbo tilk wuyiyitna asteyayet endisetibet madreg yemelkam astedaderna ewnetegan dimokrasi menor milikit new

8- Yih endale hono ewnetegna democracy balebet hager yepres hig mawtat asfelagi alemehonun lenoch higoch lepresum endemiwulu mister harmut tenagrewal

3- Befrederik erbert stiftun fawundeshin yeityopia tewekay yehonut mister harmut yihin yalut yemegnagna bezuhan balemuya setoch mahber balefew samint bazegajew siseba yenetsa pres asfelaginet asmelktew besetut asteyayet new

The news is given to the system with the following weights assigned where six of them are variable for adjusting weights:

Unit ID	**Diagnostic unit**	**Weight**	**Reason**
D1	Weight of tile	100,000	must be extracted always(fixed value)
D2	Weight of title words	4	Title words are important **(variable)**
D3	Header sentence	3	important next to the title **(variable)**
D4	header sentence words	2	At least more than ordinary words **(variable)**
D5	Key words	5	If above the others it is most probably important **(variable)**
D6	Percentage of keywords to use	0.05	This is the amount of key words to use **(variable)**
D7	Cue phrases	25	Important if it is found in the sentence(fixed)
D8	Paragraph start/end sentences	2	Additional weight to paragraph starts and ends. **(variable)**

Table 4.1 Starting weights

The weights are assigned considering the theoretical relative importance of the identified elements of news items as starting for adjusting wwights in the

model. The title is given very high weight because it must be extracted always. Next to that cue phrases are given more weight because they are useful clues about the importance of a sentence. Key words (words with high frequency) are given the next weight. Title words are given a starting weight of four. Header sentence is also given extra weight. Words appearing in header are given bonus weight. Sentences are also given extra weight for appearing at the beginning or end of a paragraph.

Condensation factor is determined by the user. For this weight adjustment the condensation factor is adjusted so that the extracted number of sentences by the system is equal to the manually extracted sentences.

The following is the training conducted using the first news article. A value is fixed if it is having high effect in the adjustment of weights and the other values are affected. This is done to see the effect of each in reference to the other.

Adjusting weights using first news article

Sentences	Title	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	13
Human	✓	✓		✓			✓		✓			✓		
Model	✓	✓	✓	✓			✓		✓					

Table 4.3 Human vs. system summary with 41% condensation factor

Precision = 5/6 = 0.83

Recall =5/6 = 0.83

1.Adjusting weights one: weight of title words is changed and no effect is observed.

2.Adjusting weights two: weight of header sentence is affected and no effect is observed.

3.**Adjusting weights three:** weight of header words is changed still no effect is observed.

4.**Adjusting weights four**: weight of key words is changed and brought no positive effect.

5.**Adjusting weights five**: by affecting the amount of key words to use (from 0.05 to 0.2) 100 percent recall and 85.7% precision is attained.

6.**Adjusting weights six**: weight of paragraph starts and ends are affected and no better result is observed.

The above described are six samples (one set of adjusting weights) of the 31 adjusting weights iterations made for the first news article. The next adjusting weights for this news proceeded by keeping the maximum result achieved in the first six experiments.

The total adjusting weights (including the first) using this news item can be summarized as follows:

Adjusting weights sets	Diagnostic units affected	Highly Effective diagnostic unit	Max. Recall achieved	Max Precision achieved
1	6	D6 = 0.2 fixD6	100	85.7
2	5	D5 = 7 fixD7	100	85.7
3	5	D6 = 0.301 fix D6	100	100
4	5	No effect		
5	5	No effect		
6	5	No effect		

Table 4.3 summary of adjusting weights set one at 41%

What was actually done inside the model can better be illustrated by procedural overview of the steps of extraction. Note that the title, the header sentence, paragraph starts/ends and words are given specific identification tags by MYXML parser. Each sentence is also put on one line by the parser as follows:

<RES><Q>ewenet</Q><Q>ehadeg</Q><Q>yemeret</Q><Q>policy</Q><Q>lewet</Q><Q>yekebelal</Q></RES>

<X><MEG><S><Q>g8</Q><Q>eyetebalu</Q><Q>yemitawekut</Q><Q>yealemachen</Q><Q>sement</Q><Q>yebeletsegu</Q><Q>ageroch</Q><Q>meriwoch</Q><Q>bekirbu</Q><Q>badere

gut</Q><Q>sebseba</Q><Q>yeafrican</Q><Q>guday</Q><Q>besefat</Q><Q>teweyayetewbetal</Q></S></MEG>

<S><Q>kerehabuna</Q><Q>kedirku</Q><Q>gar</Q><Q>beteyayaze</Q><Q>ajendam</Q><Q>leafrica</Q><Q>kend</Q><Q>huneta</Q><Q>tekuret</Q><Q>setew</Q><Q>akababiwen</Q><Q>lemerdat</Q><Q>mereha</Q><Q>geber</Q><Q>awetetewal</Q></S></X>

<X><S><Q>beyazenw</Q><Q></Q><Q>haya</Q><Q>andegnaw</Q><Q>kefle</Q><Q>zemen</Q><Q>rehab</Q><Q>lemekelakel</Q><Q>endemichalna</Q><Q>lezihum</Q><Q>asfelagi</Q><Q>erdata</Q><Q>endemiyadergu</Q><Q>geltsewal</Q></S>

<S><Q>yeh</Q><Q>bergetem</Q><Q>mehon</Q><Q>yalebet</Q><Q>guday</Q><Q>sihon</Q><Q>asfelagi</Q><Q>metebaberna</Q><Q>metegagez</Q><Q>binor</Q><Q>noro</Q><Q>beyazenew</Q><Q>haya</Q><Q>andegnaw</Q><Q>kefle</Q><Q>zemen</Q><Q>sayhon</Q><Q>balefew</Q><Q>hayagnaw</Q><Q>kefele</Q><Q>zemen</Q><Q>rehaben</Q><Q>lemekelakel</Q><Q>masweged</Q><Q>yechalen</Q><Q>neber</Q></S></X>

<X><S><Q>simintu</Q><Q>yebeletesgu</Q><Q>hageroch</Q><Q>betedegagami</Q><Q>sele</Q><Q>ityopiya</Q><Q>yanesut</Q><Q>guday</Q><Q>lalefut</Q><Q>haya</Q><Q>ametat</Q><Q>keselsa</Q><Q>sement</Q><Q>million</Q><Q>yeityopia</Q><Q>hezeb</Q><Q>gimashu</Q><Q>berehab</Q><Q>bemegeb</Q><Q>wastena</Q><Q>matat</Q><Q>mesekayetun</Q><Q>neber</Q></S>

<S><Q></Q><Q>ahunem</Q><Q>bihon</Q><Q>amest</Q><Q>million</Q><Q>yeitiopia</Q><Q>hezeb</Q><Q>ejig</Q><Q>kebad</Q><Q>behone</Q><Q>yemegeb</Q><Q>etot</Q><Q>chegir</Q><Q>enedmisekay</Q><Q>askemetewal</Q></S>

<S><Q>enedezih</Q><Q>ayenetu</Q><Q>adega</Q><Q>lemasweged</Q><Q>keitiopia</Q><Q>mengest</Q><Q>gar</Q><Q>kealem</Q><Q>bank</Q><Q>keleloch</Q><Q>legash</Q><Q>mengestatna</Q><Q>betaweku</Q><Q>alem</Q><Q>akef</Q><Q>mengestawi</Q><Q>yalhonu</Q><Q>dirijitoch</Q><Q>gar</Q><Q>honew</Q><Q>begara</Q><Q>enseralen</Q><Q>belom</Q><Q>aketachawen</Q><Q>asawekewal</Q></S></X>

<X><S><Q>yemiserut</Q><Q>sira</Q><Q>min</Q><Q>endehonem</Q><Q>zerzerew</Q><Q>dingetegna</Q><Q>erdata</Q><Q>lay</Q><Q>kematekor</Q><Q>yilik</Q><Q>lezeleketa</Q><Q>yemibej</Q><Q>lela</Q><Q>amarach</Q><Q>lay</Q><Q>lemageth</Q><Q>keetyopia</Q><Q>mengist</Q><Q>gar</Q><Q>abrew</Q><Q>endemiseru</Q><Q>yegeter</Q><Q>meseret</Q><Q>limat</Q><Q>lemasfafat</Q><Q>endemiredu</Q><Q>yeteyayaznewn</Q><Q>yeseftinet</Q><Q>program</Q><Q>endemidegfu</Q><Q>etyopia</Q><Q>yenedefechewn</Q><Q>"yedihinet</Q><Q>neqesa"</Q><Q>strategy</Q><Q>ewun</Q><Q>lemadreg</Q><Q>kemengist</Q><Q>gar</Q><Q>honew</Q><Q>endemisru</Q><Q>hurlu</Q><Q>geltsewal</Q><Q>qualm</Q><Q>gebtewal</Q></S>

<S><Q>tegbar</Q><Q>lay</Q><Q>endiyawlut</Q><Q>kememegnet</Q><Q>wuchi</Q><Q>yeminlew</Q><Q>yelemn</Q></S>

<S><Q>bequal</Q><Q>yetekemetewn</Q><Q>betegbar</Q><Q>tergumew</Q><Q>endiyasayun</Q><Q>yebereta</Q><Q>mignot</Q><Q>alen</Q></S></X>

<X><S><Q>"yemeret</Q><Q>yizota</Q><Q>lewt</Q><Q>geberewoch</Q><Q>bemeretachew</Q><Q>lay</Q><Q>yibelt</Q><Q>genzebna</Q><Q>gulbet</Q><Q>endiyafesu</Q><Q>yemiyadergna</Q><Q>ye</Q><Q>ersha</Q><Q>mirtinim</Q><Q>endiyadig</Q><Q>yemiyaderg</Q><Q>new"</Q><Q>bilew</Q><Q>begiltse</Q><Q>asfirewutal</Q></S></X>

<X><S><Q>bezih</Q><Q>huneta</Q><Q>temesritew</Q><Q>end</Q><Q>awropawyan</Q><Q>akontater</Q><Q>2006</Q><Q>amete</Q><Q>mihret</Q><Q>dires</Q><Q>bebizu</Q><Q>kililoch</Q><Q>geberew</Q><Q>beyazew</Q><Q>meret</Q><Q>lay</Q><Q>yemulu</Q><Q>tetekaminet</Q><Q>mebtun</Q><Q>yemiyaregagit</Q><Q>lewt</Q><Q>endimeta</Q><Q>endemidegifu</Q><Q>geltswal</Q></S>

<S><Q>yemayashama</Q><Q>ena</Q><Q>gilts</Q><Q>yehone</Q><Q>yetetekaminet</Q><Q>mebt</Q><Q>yemiyaregagt</Q><Q>yemeret</Q><Q>yizota</Q><Q>lewt</Q><Q>"</Q><Q>land</Q><Q>reform"</Q><Q>endiyaderg</Q><Q>megefafat</Q><Q>bicha</Q><Q>sayhon</Q><Q>bemerhagibrachew</Q><Q>antsar</Q><Q>ende</Q><Q>awropawyan</Q><Q>akotater</Q><Q>be</Q><Q>2004</Q><Q>be</Q><Q>hulet</Q><Q>kililoch</Q><Q></Q><Q>be</Q><Q>2005</Q><Q>betechemari</Q><Q>beleloch</Q><Q>sost</Q><Q>kililoch</Q><Q>eske</Q><Q>2006</Q><Q>degmo</Q><Q>bekerut</Q><Q>hulet</Q><Q>kililoch</Q><Q>yih</Q><Q>lewt</Q><Q>endikahed</Q><Q>endemiyaberetatu</Q><Q>ena</Q><Q>lezih</Q><Q>lewt</Q><Q>mengistn</Q><Q>endemidegifu</Q><Q>asmirewbetal</Q></S></X>

<X><S><Q>bezih</Q><Q>yemeret</Q><Q>polisi</Q><Q>lewt</Q><Q>zuria</Q><Q>yeminesaw</Q><Q>kedamiwna</Q><Q>tiliku</Q><Q>tiyake</Q><Q>yebeletsegut</Q><Q>hageroch</Q><Q>yemeret</Q><Q>yizota</Q><Q>lewt</Q><Q>yemilutn</Q><Q>neger</Q><Q>ehadeg</Q><Q>yikebelewal</Q><Q>wey</Q><Q>yemil</Q><Q>new</Q></S>

<S><Q>ehadeg</Q><Q>ende</Q><Q>partim</Q><Q>endemengistm</Q><Q>betedegagami</Q><Q>betshufm</Q><Q>beqalm</Q><Q>yemigeltsew</Q><Q>aquam</Q><Q>ale</Q></S>

<S><Q>meret</Q><Q>bemengist</Q><Q>kutitir</Q><Q>enji</Q><Q>begil</Q><Q>balebet net</Q><Q>endemayiyaz</Q><Q>meret</Q><Q>endemayshet</Q><Q>endemaylewet</Q><Q>beted egagami</Q><Q>yinageral</Q></S>
<S><Q>be</Q><Q>atsnot</Q><Q>arat</Q><Q>netib</Q><Q>eyale</Q><Q>yemiyasemrebe t</Q><Q>amelekaketna</Q><Q>program</Q><Q>new</Q></S></X>
<X><S><Q>silezih</Q><Q>yemeret</Q><Q>polisi</Q><Q>lewt</Q><Q>woym</Q><Q>ye meret</Q><Q>yizota</Q><Q>lewt</Q><Q>woym</Q><Q>benesu</Q><Q>quwanquwa</Q><Q>/la nd</Q><Q>reform/</Q><Q>sibal</Q><Q>min</Q><Q>aynet</Q><Q>lewt</Q><Q>yemiyamelekit< /Q><Q>new</Q></S>
<S><Q>ehadeg</Q><Q>amelekaketun</Q><Q>keyirual</Q><Q>mallet</Q><Q>new</Q><Q >weys</Q><Q>meseretawi</Q><Q>yalhone</Q><Q>amelekaket</Q><Q>new</Q><Q>yemeret</Q> <Q>polisi</Q><Q>weym</Q><Q>yemeret</Q><Q>yizota</Q><Q>lewt</Q><Q>yetebalew</Q></S ></X>
<X><S><Q>benegerachin</Q><Q>lay</Q><Q>ehadeg</Q><Q>yemeret</Q><Q>polisiwn</ Q><Q>keyro</Q><Q>meret</Q><Q>begil</Q><Q>ayshetim</Q><Q>aylewetim</Q><Q>yilew</Q> <Q>yeneberewn</Q><Q>tito</Q><Q>begil</Q><Q>yizota</Q><Q>endegena</Q><Q>ayfekdm</Q> <Q>meret</Q><Q>endishetim</Q><Q>endilewetim</Q><Q>ayadergm</Q><Q>maletachin</Q><Q> aydelem</Q></S>
<S><Q>yeyazewn</Q><Q>aquam</Q><Q>kehizb</Q><Q>tikim</Q><Q>antsar</Q><Q>ayt o</Q><Q>yemastekakel</Q><Q>chigir</Q><Q>alebet</Q><Q>maletachin</Q><Q>endji</Q><Q>si ltan</Q><Q>lay</Q><Q>lemekoyetena</Q><Q>siltan</Q><Q>tebek</Q><Q>lemadreg</Q><Q>esk etekemew</Q><Q>dres</Q><Q>yemeret</Q><Q>polisiwn</Q><Q>bilewit</Q><Q>yigermal</Q>< Q>maletachin</Q><Q>aydelem</Q></S>
<S><Q>beteley</Q><Q>bahunu</Q><Q>gize</Q><Q>yemitayew</Q><Q>yeehadeg</Q><Q >akahed</Q><Q>ende</Q><Q>eqa</Q><Q>gizhi</Q><Q>ena</Q><Q>shiyach</Q><Q>mastawekia </Q><Q>"yeteshale</Q><Q>menged</Q><Q>ketegegne</Q><Q>becheretaw</Q><Q>aygededim"</ Q><Q>yemil</Q><Q>mehonu</Q><Q>yetaweke</Q><Q>new</Q></S></X>

Taking the content of the tokenized file, the model measured the value of the sentences as: (note that every dimension of measuring are used)

abzagnawn hizb liyasmamu....110
ewnetegna dimokrasi bale.....76
yilikunm keindih aynetu hig...116
befrederik erbert stiftun faw....150
yihin madregm yeparlamawn.....80
mister harmut bezihu wekt 82
kezih befit bezihu rekik hig.....60
ye pres higu yehizbunia 150124
adisu yepres hig rekik.....154
yih endale hono ewnetegna 122
yeetiopia megenagna bizuhan.....104
leministroch mikir bet ketlom 48
yihin aynetun hig benetsa wanegna.....144
lehulet kenat yeqoyewn sibseba.....178

After ranking the sentence, what was done in the system is just giving out high scoring sentences till the condensation ratio relative to the size of the item is reached. Hence it gives out the title which scored the highest, then "lehulet kenat yeqoyewn sibseba.....178", "adisu yepres hig rekik.....154" etc until the condensation ratio is used. The full extract is as follows:

Amhasum output

ye pres higu yehizbunia yemimeleketachewn akalat asteyayet makatet alebet

lehulet kenat yeqoyewn sibseba benigigir yekefetut ymastaweqiya ministru ato bereket simon bebekulachew yepres higu memechiw julay kesene 24 eske hamle 24 lewuyiyit bedigami endemiqerb geltsew yematsdequ hidet yehulunm yehibretesebu kifl endemiyasatf amelkitewal

adisu yepres hig rekik yehizbunina yemimeleketachewn akalt asteyayet yakatete lihon endemigeba ewnetegna dimokarsi balachew hagerat tiru higoch yemwetut mhizbu asteyayet selemdabiru mehonum mister harmut hes ameleketu

befrederik erbert stiftun fawundeshin yeityopia tewekay yehonut mister harmut yihin yalut yemegnagna bezuhan balemuya setoch mahber balefew samint bazegajew siseba yenetsa pres asfelaginet asmelktew besetut asteyayet new

yihin aynetun hig benetsa wanegna teterkami lemihonew hizbna lemimeleketachew akalat qerbo tilk wuyiyitna asteyayet endisetibet madreg yemelkam astedaderna ewnetegan dimokrasi menor milikit new

yih endale hono ewnetegna democracy balebet hager yepres hig mawtat asfelagi alemehonun lenoch higoch lepresum endemiwulu mister harmut tenagrewal

- **Observation**: six sentences out of the given 14 sentences including the title are extracted by using a condensation factor 41%. 100% precision and recall is attained by assigning weights as stated above.

The result obtained by adjusting weights in the system using the first news item is carried to the next adjusting weights set (D6 = 0.301 and D5 = 7).

The carried weights from the first adjusting weights set resulted in a maximum of 50% recall and 33.3% precision at 39% condensation. This will be improved by changing the most effective diagnostic units in the next adjusting weights.

Summary of adjusting weights using second news item

Adjusting weights set	Diagnostic units affected	Highly Effective diagnostic unit	Max. Recall achieved	Max Precision achieved
1	6	D5 = 8 fix D5	41.6	66.6
2	5	D6 = 0.32 fix D6	50	83.3
3	5	D5 =10 fix D5	66.6	83.3
4	5	D6=0.35 fix D6	100	100
5	5	No effect		
6	5	No effect		

Table 4.4 Summary of adjusting weights set two at 39% condensation

- **Observation:** The system is trained to extract 12 sentences out of 28 sentences at a condensation factor of 43% percent.
- The maximum result obtained in these adjusting weights is back referenced to the first adjusting weights set and the average value is carried to the next adjusting weights. I.e. D6=0.3255, D5=8.5

The carried weights from the second adjusting weights responded 83% precision and 83% recall at 30% condensation when applied to the third news at 34% condensation factor.

Summary of adjusting weights using third news item

Adjusting weights number	Diagnostic units affected	Highly Effective component	Max. Recall achieved	Max Precision achieved
1	6	D5 = 10 fixD5	100	83
2	5	D6 = 0.34 fix D6	100	100
3	5	No effect		
4	5	No effect		
5	5	No effect		
6	5	No effect		

Table 4.5 summary of adjusting weights set three at 42%

- **Observation:** The system is trained to extract 6 sentences out of 18 sentences at a condensation factor of 34% percent.
- The result obtained is back referenced to the previous adjusting weights and the average value is carried to the next adjusting weights. I.e. D6=0.333, D5=9.25

The carried weights from the third adjusting weights responded 60% precision and 60% recall at 60% condensation when applied to the third news at a condensation factor 36%

Summary of adjusting weights using fourth news item

Adjusting weights set	Diagnostic units affected	Highly Effective component	Max. Recall achieved	Max Precision achieved
1	6	D3=5 fix D3	80	60
2	5	D2= 5 fix D2	100	100
3	5	No effect		
4	5	No effect		
5	5	No effect		
6	5	No effect		

Table 4.6 Summary of adjusting weights set four at 36% condensation

- **Observation:** The system is trained to extract 5 sentences out of 14 sentences at a condensation factor of 36% percent.
- The result obtained is back referenced to the previous adjusting weights and the average value is carried to the next adjusting weights. I.e. D2=4.5, D3=4

After adjusting weights in the model developed, the weights are fixed to the values that showed better performances in the adjusting weights conducted. These values are going to be used by the system to generate summaries of news items.

Finally the weights of diagnostic units are set to:

Unit ID	Diagnostic unit	Weight
D1	Weight of title	100,000
D2	Weight of title words	4.5
D3	Header sentence	4
D4	header sentence words	2
D5	Key words	9.25
D6	Percentage of keywords to use	0.333
D7	Cue phrases	25
D8	Paragraph start/end sentences	2

Table 4.7 final weights assigned

The system with the final weight is run on each of the adjusting weights news items again and average Precision 80.4% and average Recall 64.5% is achieved at 38.5% condensation.

high frequency. The less frequency of a word forced the use of more number of words. As a result of this more number of words is used to determine the sentences to be extracted. This dissimilarity is observed because no stemmer is used in the system. The use of good stemmer module may reduce this problem and decrease word dissimilarity.

The conducted test reveals that the model did not reach the saturation point on the adjusting weights. A better adjustment can be done by incorporating a stemmer module and exhaustive list of stop words.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This research has proposed a system that automatically extracts summary of Amharic news items. The system uses the extraction approach to summarization, where the summary is generated by picking out the most important point containing sentences of a text.

The system was developed by integrating selected statistical and natural language processing techniques adopted from prominent works in the area of text processing, information retrieval and automatic text summarization for other languages.

The model developed was trained with four news articles in 124 independent iterations by changing one of the five mining units at a time. The diagnostic units used are titles, title words, head sentences, head sentences words, end sentences, paragraph starting sentence, cue phrases and high frequency key words. The title is extracted as it is and the body of the news is tried to be condensed.

Evaluation of the system after adjusting weights resulted in 70.4% precision and 58% Recall while condensing the size of the news to 38.5%. These results are obtained after adjusting weights of the diagnostic units in the model using four news items in 124 iterations by affecting six of the eight diagnostic units used in the extraction system. This investigation revealed that the use of title words, percentage of key words and weight assigned to key words are the top hits in capturing the core points of a news text.

The obtained result shows that by using sentence extraction approach of generating summaries it is possible to automate the task of generating summaries of news texts. Further NLP diagnostic units like nouns and adjectives can be added to this system to improve the performance.

5.2 Recommendations

The following recommendations are forwarded for further study to the improvement and upgrading of the proposed system to its uppermost and the development of general Amharic text summarizer.

1. The application of a good stemmer module will help in the reduction of words to their root. Word variations due to language necessity could be minimized and the frequency concept of Luhn (1959) could be used wisely.
2. The availability of a standard Amharic corpus for testing the text processing systems will help a lot in measuring the performances of text processing systems in an objective way. Hence a standard Amharic corpus should be developed.
3. Exhaustive lists of language necessity words – sometimes called stop words – will help in easy removal of these words from frequency lists in statistical approaches to text processing. Hence its development would worth a lot.
4. Further study could be conducted as a continuation of this work to include more NLP, statistical and heuristic parameters with thorough list of stop words and good stemmer module.
5. Based on this foundational study, further research can be conducted to develop a full-fledged Amharic document summarizer.

References:

1. Alex, A. etal (2000) A trainable algorithm for summarizing news stories. Brazil URL=http://www.cs.kent.ac.uk/people/staff/aaf/pub_papers.dir/PKDD-Ws-2000.ps
2. Atelach A. (2002) automatic Amharic sentence parsing, Masters Thesis Addis Ababa University. Addis Ababa. Ethiopia
3. Bender, M. et al (1976) Languages in Ethiopia. Oxford university press. London
4. Buyukkokten, O. et al. Text Summarization for Web Browsing on Handheld Devices: URL=http://www-db.stanford.edu/~orkut/papers/autosumm.pdf
5. Coperinc (2003) Copernic Summarizer. URL=http://www.copernic.com/en/products/summarizer/
6. D'Avanzo E. et al (2004) Keyphrase Extraction for Summarization Purposes:The LAKE System at DUC2004 URL=http://www-nlpir.nist.gov/projects/duc/pubs/2004papers/itc-irst.magnini.pdf
7. Dalianis, H. et al (2000) SweSum. Text summarization for Swedish URL=http://www.nada.kth.se/~hercules/Textsumsummary.html
8. Dalianis, H. et al (2002) From SweSum to ScandSum-automatic text summarization for Scandinavian languages. URL=http://www.dsv.su.se/~hercules/scandsum/ScandSumArsbog2002.pdf
9. Edmundson, H. P. (1969) New methods in automatic extracting, Journal of the Association for computing Machinery 16(2)

10. Fuentes, M. (2001) Using cohesive properties of text for automatic summarization.
URL=http://www.nlpif.nist.gov/projects/due/2001.html

11. Gasser M (2003).(111) Some differences between Amharic and European languages How the language works. Indiana University and Michael Gasser.
URL=www.indiana.edu/~hlw/Meaning/differences.html

12. Goldstien, J. et al Summarizing text Documents
URL=http://citeseer.nj.nec.com/cache/papers/cs/26885/http:zSzzSzranger.uta.eduzSz~alpzSzixzSzreadingszSzGoldsteinSigir99NewsSummarization.pdf/goldstein99summarizing.pdf

13. Hahn, U. & Mani, I (1999) The goal of automatic summarization
URL=http://www.dsv.su.se/ijcai-99/tutorials/e3.html

14. Hassel, M. (2000) Pronominal resolution in automatic text summarization
URL=http://www.members.tripos.com/~mhassel/curicit/

15. Hassen, A. et al (2003) structured and unstructured document summarization: design of commercial summarizer using lexical chains.
URL=http://www-nlpir.nist.gov/projects/duc/pubs/2004papers/itc-irst.magnini.pdf

16. Jones, S.K. (1981) Information Retrieval expriment. Butter worths, London

17. Khorfhage, R. R. (1997) Information storage and retrieval. Whyley computer publishing, Singapore.

18. Kupiec, J. etal (1993) A trainable document summarizer. Xerox Palo Alto research center California.
URL=http://www.dcs.shef.ac.uk/~mlap/teaching/kupiec95trainable.pdf.

19. Larry, W. et al (1996) Programming perl. O'Reily & associates, Inc.

20. Luhn, H. P. (1959) the automatic creations of literature abstracts, IBM J.

21. Maybury, M. T. and Mani, I. (2001) Automatic summarization. American/European conference on computational linguistics (ACL/EACL) Toulous, France. URL=www.mitre.org/resources/centers/it/maybury/summarization/summarization.htm

22. Mckeown, K.R et al (2000) Tracking and summarizing news on a daily basis with colombia's newsblaster URL=http://www.newsblsater.cs.colonbia.edu/papers/hlt-blaster

23. Mckeown, K.R et al (1999) Summary Generation through Intelligent Cutting and Pasting of the Input Document. URL=http://www1.cs.columbia.edu/~hjing/sumDemo/CPS/

24. Meadow, T.C. et al (2000) test information retrieval systems. Second edition. Academic press. Tokyo

25. Mesfin Getachew (2001), Automatic Part of Speech Tagging for Amharic: An Experiment Using Stochastic Hidden Markov (HMM) Approach, Masters Thesis Addis Ababa University. Addis Ababa. Ethiopia

26. Paice, C. D. and Jones, A. P. (1993). The identification of important concepts in highly structured technical papers. In Proceedings of the Sixteenth Annual International ACM SIGIR conference on research and development in IR.

27. Robertson, and Jones(1994)

28. Salton, G (1993) approaches to passage retrieval in full text information systems, technical report, Department of computer science, Cornel University.

29. Salton, G. (1989) automatic text processing: the transformation, analyses, and retrieval of information by computer. Addison-Wesly

30. Sebsibe etal (2004) Unit Selection Voice For Amharic Using Festvox International Institute of Information Technology, Hyderabad Language Technologies Institute, Carnegie Mellon University URL:http://www-2.cs.cmu.edu/~awb/papers/ssw5/amharic.pdf

31. Simon Ager (1998-2004) Omniglot a guide to writing system URL=http://www.omniglot.com/writing/index.htm

32. Solomon, G. (1991) Writing for academic purpose. Volume(1) foreign language and literature department. Institute of language studies. Addis Ababa University. Addis Ababa. Ethiopia.

33. Surafel, T. (2003) Automatic categorization of Amharic news text, Masters Thesis Addis Ababa University. Addis Ababa. Ethiopia

34. van Rijsbergen, C.J. (1979) Information retrieval URL=www.dcs.gla.ac.uk/Keith/Preface.html

35. Wu, C. and Liu, C. (2003) Ontology based text summarization for business news articles. URL=http://www.cs.nccu.edu.tw/~chaolin/papers/wi03

36. Zelalem (2001) Automatic Amharic News Text categorizer Masters Thesis Addis Ababa University. Addis Ababa. Ethiopia

37. McCargar, V. (2004) Statistical Approaches to Automatic Text Summarization Summarization. Bulletin of the American Society for Information Science & Technology, Apr/May 2004

URL=http://www.findarticles.com/p/articles/mi_qa3991/is_200404/ai_n9397126#continue

38. Sparck Jones, K. (1998) Automatic summarizing: factors and directions. In Advances in automatic text summarization, Mani and Maybury, eds., MIT Press. URL= http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sum-mani.pdf

Conferences/Workshops

1. AAAI Spring Symposium on Intelligent Text Summarization, Stanford, 1998 (Dragomir Radev, Eduard Hovy)
   URL=www.cs.columbia.edu/~radev/aaai-sss98-its
2. ACL/EACL Workshop on Intelligent Scalable Text Summarization, Madrid, 1997 (Inderjeet Mani, Mark Maybury)
   URL=www.cs.columbia.edu/~radev/ists97/program.html
3. ANLP/NAACL Summarization Workshop, Seattle, 2000 (Udo Hahn, Chin-Yew Lin, Inderjeet Mani, Dragomir Radev)
   URL=www.isi.edu/~cyl/was-anlp2000.html
4. Dagstuhl Seminar, 1993 (Karen Spärck Jones, Brigitte Endres-Niggemeyer)
   URL=www.ik.fhhannover.de/ik/projekte/Dagstuhl/Abstract

Appendix one

Amharic news text summarizer (Amhasum) model system in perl

Written by kamil Nuru
June 20, 2004 Addis Ababa
Faculty of Informatics, Department of Information Science, AAu

Requirements:

- Input file to this system must be tokenized for titles, words, sentences, paragraph initials and paragraph ends. Refer the MYXML parser in appendix two.
- A file containing list of stop words must be placed right under the C directory by the name stop_words.txt.

How to start it:

- After properly installing perl on your system copy the code to notepad and save it.
- You can then run the file both from the perl window and from the command prompt.

```
################################ START ######################################
 print "yezenaw simna adrashia yet new?\n\n"; $ARGV[0]=<STDIN>;
       print "\n\nbemin yakl lasatirew?\n\n";      $ARGV[1]=<STDIN>;
#SUMMARIZATION PARAMETERS_____________________________________________
  %tequami = ('\btedergual\b', 25, '\btegeltsual\b', 25, '\astaweke\\b',25,);
  $wt{RES} = 1000000; $wt{yeRES_qalat} = 4.5;       # title/title word
  $wt{megbia} = 4;  $wt{yemegbia_qalat} = 2;     # beginning of paragraph/ph_words
  $wt{ankets_mejemeria} = 2;                  # start of a paragraph
  $wt{ankets_mechereshia} = 2;                 # end of paragraph
  $wt{qulf_qal} = 9.25;                      # keyword
  $wt{yekulfqalt_meten}  = 0.333;              #no_of key words
#ANALYZER USER SPECIFICATIONS_____________________________________________
        $ARGV[0]="c:\\perl\\training1.txt";
        $ARGV[1]=38;
  if (!defined $ARGV[0]) {print "yetal zenaw?\n"; exit;}
    elsif (!defined $ARGV[1]) {print "yemasatiribetin meten alawekum.\n";  exit;}
  if ($ARGV[1] > 100 || $ARGV[1] < 0){ print "kezero-meto bicha .\n";   exit;}
 open FILE, $ARGV[0] or die "alkefet ale$ARGV[0] lenibab :$!"; #Try to open the file
  $condensation  = $ARGV[1]/100;                    # calculate atirara meten
  while($_ = <FILE>){                               # for all line in the newS
#ASSIGN WT TO PARAGRAPH STARTS_____________________________________________
 if($_ =~ /<x/) {$weight2 = $wt{ankets_mejemeria};}   # Assign wt to megbia areftеneger
 else      { $weight2  = 0,}          print " YEANKETS MEJEMERIAWOCH =$weight2 \n";
#otherwiseweight is zero#CHECK TIELES_____________________________________
```

```
    if($_ =~ /<RES>/) {                          # check for title
     $rest = $';                                 # discard rest of line
     @res = split(/<\/RES>.*/, $rest);           # discard after </RXS>
     $rest = "@res";                             # convert back to string
     @res = split(/^.*?<Q>|<\/Q>.*?<Q>|<\/Q>.*/, $rest);   #print @res; prints the title words
     push(@titles, "@res");                      # create a list of title words
     unshift(@res, "");                          # create title index (no detailed index needed)
     $sentences{"@res"} = $wt{RES};}
  #NALYZE MEG SENTENCES_____________________________________________
    if($_ =~ /<MEG.*?>/) {                       # check for header
     $_ =~ /<MEG ID='(.*?)'/;                    # extract header id - result is in $1
     $rest = $';                                 # discard rest of line
     @header = split(/<\/MEG>.*/, $rest);        # discard after </MEG>
     $rest = "@header";                                  # convert back to string
     @header = split(/^.*?<Q>|<\/Q>.*?<Q>|<\/Q>.*/, $rest);   #print @header; #
     push(@headers, "@res");                             # create a list of header words
     unshift(@header, $1);                       # place header id at beginning
     $sentences{"@header"} = $wt{megbia}; }      ########print $sentences{"@header"};
  #IDENTIFY SENTENCES AND MAKE A LIST OF WORDS_____________________
    if($_ =~ /<S.*?>/) {                                 # check for sentence
     $_ =~ /<S ID='(.*?)'/;                              # extract sentence id - result is in $1
     $rest = $';                                         # discard rest of line
     @sentence = split(/<\/S>.*/, $rest);                # discard after </S>
     $rest = "@sentence";                                # convert back to string
     @sentence = split(/^.*?<Q>|<\/Q>.*?<Q>|<\/Q>.*/, $rest);   # print @sentence;  ##all the
sentennces
     unshift(@sentence, $1);                             # GIVE ID TO SENTENCES
     $sentences{"@sentence"} = $weight2;};  ;            #print "\n",print $sentences{"@sentence"};
  #ADD TO THE WT OF A SENTENCE IF IT IS AN END SENTENCE_____________
    if($_ =~ /<\/x>/) { $sentences{"@sentence"} += $wt{ankets_mejemeria};} }
                                 # if end of paragraph increase weight of last sentence
close(FILE);      # CLOSE THE FILE HANDLE
#print "\nnews title=@titles ";                  #########print "wt of title sentence=";print
$sentences{"@res"}; print "\ntitle words\n";
  while(($_, $wt) = each(%sentences)) {
     @word_list = split;                         # convert to array to put in hash
  while($word = pop(@word_list)) {               # go thru each word and put in hash  # $word
= &stem($word);
     $words_freq{$word} += 1; }}
```

```
    @sorted_words = sort { $words_freq{$a} <=> $words_freq{$b} } keys %words_freq;
####print @sorted_words;
open(FILE, 'C:\\stop_words.txt');                    # OPEN THE FILE OF STOP WORDS
    $_ = "@sorted_words";                            # MAKE LIST OF STRINGS
  while($word = <FILE>) {chop $word;                 # MAKE THEM ON A LINE
    s/\b$word\b//g; }                                # SEPARATE WORDS
close FILE;                                          # CLOSE THE FILE OF STOP WORDS
#DETERMINE KEY WORDS TO USE AND ASSIGN qulf_qal WT TO THE KEY WORDS_
    @sorted_words = split;                           # convert string back to a list of words
$w = $wt{yekulfqalt_meten} * @sorted_words."";   # no. of qulf_qalat = % * total words
  for($i=1;$i<$w;$i++) {
    $qulf_qalat{pop(@sorted_words)} = $wt{qulf_qal}; }
#REMOVE STOP WORDS FROM THE TITLE SENTENCE_____________________
  foreach (@titles) {                                # go through every title
open(FILE, 'stop_words');                        # OPEN THE FILE OF STOP WORDS AGAIN
    while($word = <FILE>) {
      chop $word;                                    # remove newline from word
    s/\b$word\b//g; }
close FILE;                                      # CLOSE THE FILE OF STOP WORDS
#MAKE A LIST OF TITLE WORDS AND ASSIGN WT yeRES_qalat TO THE WORDS
    @words = split;                                  # split into seperate words
  foreach $word (@words) {                           #print "$word\n";  # for all the words
    $qulf_qalat{$word} += $wt{yeRES_qalat}; }}
  foreach (@headers) {
open(FILE, 'stop_words');                        # OPEN THE FILE OF STOP WORDS AGAIN
    while($word = <FILE>) {
      chop $word;                              #print $word;# removed words
# remove newline from word
    s/\b$word\b//g;    }
close FILE;                                      # CLOSE THE FILE OF STOP WORDS
    @words = split;                                  #print @words;# split into seperate words
#ADD THE WT OF A WORD IF IT IS HEAD WORD, KEY WORD, CUE WORD_____
  foreach $word (@words) {                           #print @words;  # for all the words
    $qulf_qalat{$word} += $wt{yemegbia_qalat}; }}  #ADD TO THE WT IF HEADER WORD
  foreach $word (keys %qulf_qalat) {         #print  %qulf_qalat; #key words used
  foreach (keys %sentences) {
  if (/$word/) { $sentences{$_} += $qulf_qalat{$word}; } }}
#ADD TO THE WT OF A SENTNENCE WITH KEY WORD
  foreach (keys %sentences) {
  foreach $tequamioch (keys %tequami) {
```

```
    if(/$tequamioch/) {
      $sentences{$_} += $tequami{$tequamioch};} }}
      @sorted_sentences = sort { $sentences{$a} <=> $sentences{$b} } keys %sentences;
      $w = $condensation  * @sorted_sentences.""; #APPLY CONDENSATION FACTOR
    for($i=1;$i<$w;$i++) {
      push(@atirara, pop(@sorted_sentences)); }          # print "\n\n@sorted_sentences";
    foreach (@atirara) {                                 # extract every item from atirara
      @sentence=split;                                   # convert element to array
      $id=shift(@sentence);                              # extract first id element
      push(@ids, $id); }
      $ids = "@ids";                                     # convert back to string
open FILE, $ARGV[0];
   while($line = <FILE>) {
   if($line =~ /<RES>/) {                                # check for title
     @title = split(/<\/RES>.*/, $line); # REMOVE END TITLE TOKEN MARK </RES>
     $xml = "@title";                                    # convert back to string
     @title = split(/^.*?<RES>/, $line);     # REMOVE BIGING TOKEN MARK <RES>
     unshift(@title, 'RES:');                            # put title id at front
     $xml = "@title";                                    # convert back to string
     &parse_xml_sentence($xml)    }
  if($line =~ /<MEG.*?>/) {                              # check for header
     $line =~ /<MEG ID='(.*?)'/;                         # extract header id - result is in $1
     $id = $1;                                           # backup $1
  if($ids =~ /$1/) {          # if id is in our targeted list then output sentence
     @header = split(/<\/MEG>.*/, $line);     # REMOVE END MARKER </MEG>
     $xml = "@header";                                   # convert back to string
     @header = split(/^.*?<MEG.*?>/, $line);REMOVE TOKEN MARKS <MEG....>
     unshift(@header, "$id:");                           # put id at front
     $xml = "@header";                                   # convert back to string
    &parse_xml_sentence($xml) } }
 if($line =~ /<S.*?>/) {                                 # check for sentence
  $line =~ /<S ID='(.*?)'/;                              # extract sentence id - result is in $1
  $id = $1;                                              # backup $1
  if($ids =~ /$1/) {                                     # if id is in our targeted list then output sentence
   @sentence = split(/<\/S>.*/, $line);                  # discard after </S>
   $xml = "@sentence";                                   # convert back to string
   @sentence = split(/^.*?<S.*?>/, $line);               # discard before <S....>
   unshift(@sentence, "$id:");                           # put id at front
   $xml = "@sentence";                                   # convert back to string
   &parse_xml_sentence($xml)       } } }                 #print  "\n@sorted_words";
```

```
#OPEN A NEW FILE AND PUT THE SUMMARY IN THE FILE_____
    print "\n\n summary\n\n\t @atirara"; RINTS THE SUMMARY TO STANDARD OUTPUT
    #print "\nTOP SCORING SENTENCE=@header";  # THE TOP SCORING SENTENCE
open(OUTFILE, ">D:\\ATIRAR\\output.txt") or die "output.txt n MEKFET ALTERCHALEM: $!";
print OUTFILE "atirara\n\n";print OUTFILE @atirara;
close (OUTFILE);
close(FILE);  # CLOSE THE INPUT FILE
#THIS IS THE END OF THE PROGRAM YOU CAN FIND THE OUTPUT OF THE ##PROGRAM IN OUTPUT.TXT
sub parse_xml_sentence {
 $_ = $_[0];  s/<Q>//g; s/<\/Q>//g;s/<BR>\(<\/BR> /\(/g; s/<QUOT>`<\/QUOT> /'/g; s/ <.*?>//g; s/<.*?>//g;}
#####################  end of code  ########################
```

Appendix two

MYXML parser developed for Amhasum in perl

Written by kamil Nuru
June 20, 2004 Addis Ababa
Faculty of Informatics, Department of Information Science, AAu

Note: This local parser is developed for Amhasum and it is quite different from the SGML. The tags are also local to Amharic language. The parser tokenizes the file for title, paragraph start and ends, sentences, and words.

Requirements:

- The Amharic input file containing the news item to be to be tokenized must be fully Latinized without affecting the punctuation marks in Amharic.

How to start it:

- After properly installing perl on your system copy the code to notepad and save it.
- You can then run the file both from the perl window and from the command prompt.
- You will get the output in the C directory by the name "myxml"

```
################################## START ########################################
  print "yezenaw simna adrashia yet new?\n\n";
  $ARGV[0]=<STDIN>;
  #$ARGV[0]="c:\\yepress.txt"; #open the comment for direct feed of file
   open FILE, $ARGV[0] or die "Cannot open $ARGV[0] for read :$!";
while($_ = <FILE>){
 tr/[A-Z]/[a-z]/; s/,//g;  s/          /<RES><Q>/;
 s/ /<\/Q><Q>/g;
 s/\?/\./g;
 if($_=~/<RES>/){
 s/\./<\/Q><\/RES>/;}
 s/\n/\n  /g;  s/  /<X><S><Q>/g;
 if($_=~/<X><S><Q>/){
    s/\.\n/<\/Q><\/S><\/X>\n/, }
 s/\.<\/Q><Q><\/Q><Q><\/Q><Q><\/Q><Q><\/Q><Q>/<\/Q><\/S><\/X>/;
 s/\.<\/Q><Q>/<\/Q><\/S>\n<S><Q>/g;
 s/<\/Q><\/RES>\n<X><S><Q>/<\/Q><\/RES>\n<X><MEG><S><Q>/;
 print $_;
 $parsed = $_;}
open(OUTFILE, ">c:\\output.txt") or die "output.txt n MEKFET ALTERCHALEM: $!";
  print OUTFILE  $parsed;
close (OUTFILE);
#####################  end of code  ##########################
```

Appendix three

Amharic news item Taken form Reporter news paper (May 2004)

ጥበብ ለማንበረስብ ዕድገት ፫ÀĮውል ያሳዩ ምሳሌዎችየገጠር አድማጮች::

በተለይ በእንደኛ ዓይነቱ ታዳጊ ሀገር ጥበብ ከማዝናኛነቱ ይልቅ ለማስተማሩ ብንጠቀምበት የጐላ እንደሚሆን ግልጽ ነው:: አሳሳቢ በሆነበን ጉዳይ በኤች. አይ ቪ/ ኤድስ ላይ በተለይ በኪነ-ጥበብ ተጠቅመን ህብረተሰቡን ማስተማር መቻላችን ታውቆ ጥረቶች ይደረጋሉ:: በአብዛኛው «ሰባኪ» ዓይነት ባህሪ የሚታይባቸው ድራማዎችና ጭውውቶች እንዲሁም ቲአትሮቻችን ምን ያህል ግብ መትተዋል የሚለውን በትክክል ለመናገር ጥናት ያስፈልግ ይሆናል:: በደፈናው መናገር የምንችለው ግን ብዙዎቹ አሰልቺና ተደጋጋሚ መሆናቸውን ነው:: እንደሚንሰብክውና እንደምንለው ሰዎቻችን ያልተለወጡት የምንነግረውን ስላልሰሙ ሳይሆን መልዕክቶቻችን ተግባር ላይ ለማዋል የሚሆን የባህሪ ለውጥ ማምጣት ስላልቻሉ ነው:: አንዱ ለዚህ ተጠቃሽ የሚሆነው ነጥብ የምንሠራቸው የኪነጥበብ ውጤቶች የሚፈለገውን የባህሪ ለውጥ እንዲያመጡ በምን መልኩ መሠራት አበባቸው? የሚለው ጥያቄ ተመልሶ፣ በጥናት ተደግፈው የሚሠሩ ባለመሆናቸው ነው::

ከዚህ በተለየ መልኩ ደግሞ እንደ «ማፍቀር ነው መሰልጠን» አይነት የተሳካላቸው፣ አስተማሪ መሆናቸውና ለዚሁ ተግባር መዘጋጀታቸው እስኪረሳ ከሰዎች አፍ ያልጠፉ ሥራዎች መኖራቸው ይታወቃል:: በፖፑሌሽን ሚዲያ ሴንተር ኢትዮጵያ የተዘጋጁት «የቀን ቅኝት» እና «ዲምቢባ» የሬዲዮ ደራማዎች እንዲሁ ገጸ ባህርያቶቻቸው ለህዝብ ልቦና ቦታ እስኪያገኙ፣ በስማቸው ተራራ እስኪሰየም ተወዳጅ መሆናቸው ይነገራል:: በተለይ ከከተማ ወጣ ላሉት የህበረተሰብ ክፍሎች በስነ ተዋልዶና በኤች አይ ቪ ዙሪያ እንደመረጃ ምንጭ በመሆነን እያገለገሉ መሆናቸውን በየጤና ጣቢያዎች የተደረጉ ጥናቶችን ጠቅሰው በማዕከሉ ከፍተኛ የጥናት ኦፊሠር የሆኑት አቶ አበባው ፈረደ ይገልጻሉ::

በተለይ የቀን ቅኝት «እና «ዲምቢባ» የሬዲዮ ድራማዎችና የሳቅ ጀምበር ቲያትር እንዲሁም «ማለዳ» በፔጅ የተቀረጸ ተከታታይ ማስተማሪያ ፣መሠረቱ ልማሃሔሑላሑሀሥ፣ በተሰኘው የተከታታይ የሬዲዮ ድራማዎች መሥሪያ መንገድ መሠረት የተሠሩ መሆናቸውን የፖፑሌሽን ሚዲያ ሴንተር የኢትዮጵያ ተጠሪ ዶ/ር ንጉሴ ተፈራ ይገልጻሉ:: ዶ/ር ንጉሴ ይህን የገለጹት የሬዲዮ ድራማዎቹ ሁለተኛ ዓመታ ባለፈው ሳምነት መጀመሪያ በተከበረበት ወቅት ነው:: ይኸው ዘዴ ሚጉዊል ሳቢዲ በተሰኘ ሜክሲኳዊ የኮሚኒኬሽንስ ምሁር የተቀመረ ነው::

በኤፍ ኤም ሬዲዮ በየቁኑ 9 ሰዓት ላይ የሚተላለፈውና በኢትዮጵያ ሬዲዮ ማታ ሥርጭት የሚደገመው «የቀን ቅኝት» ሁለት ዓመት ያህል ስለኤች አይ ቪ፣ ስለተዋልዶ ጤና፣ ስለጎጀ ባህል፣ ስለሴቶች ጭቆና፣ በትዳር መካከል ሊኖር ስለሚገባው ግልጽነት ወዘተ... ሲያስተምር ቆይቷል:: በትረካ ታጅቦ መቅረቡ በእኩይና ሰናይ ገጸ ባህርያቱ መካከል ያለው ግብግብ አስተማሪ መሆኑን ደብቆት በጉጉት የሚደመጥ አድርጎታል::

በአለም አቀፍ ፖፑሌሽን ሚዲያ ሴንተር የዓለም አቀፍ ፕሮግራሞች ምክትል ፕሬዚዳንት ሚስ ክሪስ ባርከር ስለዚሁ ዘዴ ሲያስረዱ ሳቢዶ ዘዴውን የቀመረው ከአምስት ዓይነት የኮሙኒኬሽንና የባህርይ ለውጥ ሀግጋት መሆኑን ይገልጻሉ:: ከዚሁ ህግጋት የተወጣጣው ዘዴ ለማስተላለፍ የሚፈለገውን የተመረጠ መልዕክት ከሌሎች የማህበረሰቡ ጉዳዮች ጋር ተቀናብሮ እንዴት መቅረብ እንዳለበት የሚገልጽ ነው::

በዚሁ ዘዴ የሚቀርቡ ተከታታይ ደራማዎች መሠረታቸውን የሚያደርጉት ጠለቅ ብለው በተካሄዱ ጥናቶች ላይ ነው ጥናቶቹ አድማጭ/ተመልካቾች ያሉበት አገርና ባህል ጠንቅቆ ለማወቅ በዚያ ላይ ያተረኮረም ስራ ለመስራት ይረዳል:: ጐጂ የሆኑ/ ሊወገዱ የሚገቡ ልማዳዊ ደርጊቶች ካሉ እነኝህን ለማስወገድ ሁኔታዎችን ያመቻቻሉ:: ጥናቱ በሚገባ ከተከናወነ እሱን መሠረት አድርገው የሚጽፉት ደራሲዎች አይቸገሩም::

እንደምሳሌ የሚወሰዱና አድማጭ/ተመልካቾች በቀላሉ የሚለይዋቸው ገጽ ባህርያት ይፈጠራሉ:: ይኸውም ጥናት ደግሞ በማህበረሰቡ ዘንድ ያሉትን አዎንታዊና አሉታዊ እሴቶች ለይቶ ለማውጣትም ያግዛል:: ይህ ደግሞ እኩይና ሰናይ ተብለው (እንደ የድርጊታቸው) የሚቀረጹ ገጸ ባህሪያትን ለመሳል ይረዳል:: በሁለቱ መካከል የሚደረገው ጠንካራ ፍጭት የተመልካቹን ቀልብ ይዞ ይሄዳል::

የ«ቀን ቅኝት» እና «ዲምቢባ» ሬዲዮ ድራማዎች ከመጻፋቸው በፊት በትወናና በድርሰት ሥራ ላይ ለተሰማሩ የኪነጥበብ ሰዎች ስለ ቋሮሯሸሳቁ ቂስቋሻቁሳቁቁቁሸሸ እና ሊተላለፍ ስለሚገባው መልዕክት ለሁለት ሳምንታት ስልጠና ተሰጥቷቸዋል:: አሰልጣኞቹ ከአሜሪና ከኬንያ የመጡ ነበሩ:: የቀን ቅኝት በኢትዮጵያ ጠለቅ ብሎ በተሠራ ጥናት ላይ ያተረኮረ የመጀመሪያው የአማርኛ ቋንቋ ድራማ መሆኑ ይነገራል:: ከሥልጠናው በኋላ የ«ቀን ቅኝት» ተወለዶ ከአንዱ ወደ ሌላው እየተሸጋገረ በመሀል በሚደረጉ ትረካዎች ቀጠለ::

Latinized form of the above news item

Tibeb lemahberseb ediget endemiwul yasayu misalewoch yegeter admachoch.

Beteley be endegna aynetu tadagi hager tibeb kemaznagnanetu yilik lemastemaru binitekembet yegola endemihon giltse new. Asasabi behonebin guday be ech ay v aids lay beteley bekine-tibeb tetekmen hibertesebun mastemar mechalachin tawko tiretoch yderegalu. Beabzagnaw "sebaki" aynet bahiri yemitaybachew dramawochina chiwuwutoch endihum tiatrochin min yahl gib metitewal yemilewn betikikl lemenager tinat yasfelig yihonal. Bedefenaw menager yeminichilew gin bizuwoch aselchina tedegagami mehonachewn new. Endeminisebkewna endeminilew sewochachin yaltelewetut yemininagerewn silalsemu sayhon meliektochchin tegbar lay lemawal yemihon yrbahiri lewt mamtat silalchalu new. Andu lezih tetekach yemihonew netib yeminiserachew yekinetibeb wutetoch yemifelegewun yebahri lewt endiyametu bemin melku meserat alebachew? Yemilew tiyake temeliso betinat tedegfew yemiseru balemehonachew new.

Kezih beteleye melku degmo ende "mafker new meselten" aynet yetesakalachew, astemari mehonachewna lezihu tegbar mezegajetachew eskiresa kesewoch af yaltefu sirawoch menorachew yitawekal. Bepopulation media senter ityopiya yetezegaju "yken kignit" ena "dimbiba" yeredio dramawoch endihu getsebahiriatochachew lehizb libona bota eskiagegnu, besimachew terara eskiseyem tewedajmehonachew yinegeral. Beteley keketema weta lalut yehibreteseb kifloch besine tewaldona be ech ay v zuria endemereja minch bemehon eyagelegelu mehonachewn betena tabiyawoch yetederegu tinatochin teksew bemaekelu kefitegna yetinat ofiser yehonut ato abebaw freed yigeltsalu.

Beteley "yeken kingnt" ena "dembiba" yeredio dramawoch ysak jember tiyatir endihum "maleda" betep yekeretsa teketatay mastemaria "sabido mehondology" betesegnew yeteketatay yeredio dramawoch mesria menged meseret yeteseru mehonachewn yepopulation media senter yeityopiya teteri doctor nigusie tefera yigeltsalu. Doctor nigusie yihin yegeletsut yeradio dramwochu whuletegna amet balefew samint

Summary generated by the developed model during training

tibeb lemahberseb ediget endemiwul yasayu misalewoch yegeter admachoch

beteley yeken kingnt ena "dembiba" yeredio dramawoch ysak jember tiyatir endihum maleda betep yekeretsa teketatay mastemaria sabido mehondology betesegnew yeteketatay yeredio dramawoch mesria menged meseret yeteseru mehonachewn yepopulation media senter yeityopiya teteri doctor nigusie tefera yigeltsalu

beteley keketema weta lalut yehibreteseb kifloch besine tewaldona be ech ay v zuria endemereja minch bemehon eyagelegelu mehonachewn betena tabiyawoch yetederegu tinatochin teksew bemaekelu kefitegna yetinat ofiser yehonut ato abebaw freed yigeltsalu

be e fem radiyo beyekenu 9 seat lay yemitelalefewna beityopiya radiyo mata sirichit yemidegemew yeken kignit whulet amet yahil sile eh ay v sine tewaldo tena silegoji bahil sile setoch chikona betidar mekakel linor silemigebaw gitsenet wezete siyastemir koyitwal

yeken kignit ena dembiba yeredio dramawoch kemetsafachew befit betiwenana bedirset sira lay letesemaru yekinetibeb sewoch sile sabido methodology ena litelalef silemigebaw melkt lehulet samintat siltena tesetwachewal

bealem akef population midia senter yealem akef programoch mikitil prizedant mis kris barner silezihu zede siyasredu sabido zedewun yekemerew keamist aynet yekominikation ena yebahiryi lewt higigat mehonun yigeltsalu

kezih beteleye melku degmo ende "mafker new meselten" aynet yetesakalachew astemari mehonachewna lezihu tegbar mezegajetachew eskiresa kesewoch af yaltefu sirawoch menorachew yitawekal

bezihu zede yemikerbu teketatay dramawoch meseretachewn yemiyadergut telek bilew betekahedu tinatoch lay sihon temelkachoch yalubet agerna bahil tenkiko lemawok beziya lay yatekorem sira lemesrat yiredal

bepopulation media senter ityopiya yetezegaju "yken kignit" ena "dimbiba" yeredio dramawoch endihu getsebahiriatochachew lehizb libona bota eskiagegnu besimachew terara eskiseyem tewedajmehonachew yinegeral

beabzagnaw "sebaki" aynet bahiri yemitaybachew dramawochina chiwuwutoch endihum tiatrochin min yahl gib metitewal yemilewn betikikl lemenager tinat yasfelig yihonal

betireka tajibo mekrebu beikuyna senay getse bahiriyatu mekakel yalew gibgib astemari mehonun debiko begugut yemidemet adirgotal

andu lezih tetekach yemihonew netib yeminiserachew yekinetibeb wutetoch yemifelegewun yebahri lewt endiyametu bemin melku meserat alebachew

Appendix four

Amharic news item Taken form Reporter news paper (May 2004)

እውነት ኢሕአዴግ የመሬትፖሊሲ ለውጥ ይቀበላልን?

G-8 እየተባሉ የሚታወቁት የአለማችን ስምንት የበለጸጉ አገሮች መሪዎች በቅርቡ ባደረጉት ስብሰባ የአፍሪካን ጉዳይ በስፋት ተወያይተውበታል። ከረሀብና ከድርቅ ጋር በተያያዘ አጀንዳም ለአፍሪካ ቀንድ ሁኔታ ትኩረት ሰጥተው አካባቢውን ለመርዳት መርሃ ግብር አውጥተዋል።

በያዝነው ሃያአንደኛው ክፍለ ዘመን ረሀብን ለመከላከል እንሚቻልና ለዚሁም አስፈላጊ እርዳታ እንደሚያደረጉ ገልፀዋል። ይህ በርግጥም መሆን ያለበት ጉዳይ ሲሆን አስፈላጊው መተባበርና መተጋገዝ ቢኖር ኖሮ በያዝነው ሃያአንደኛው ክፍለ ዘመን ላይሆን ባለፈው ሃያኛው ክፍለ ዘመን ረሀብን ለመከላከልና ማስወገድ ይቻል ነበር።

ስምንቱ የበለጸጉ ሀገሮች በተደጋጋሚ ስለ ኢትዮጵያ ያነሱት ጉዳይ ላለፉት ሃያ አመታት፣ ከስልሳ ስምንት ሚሊዮን የኢትዮጵያ ህዝብ ግማሹ በረሀብ በምግብ ዋስትና ማጣት መሰቃየቱን ነበር። አሁንም ቢሮን አምስት ሚሊዮን የኢትዮጵያ ሕዝብ እጅግ ከባድ በሆነ የምግብ እጦት ችግር እንደሚሰቃይ አስቀምጠዋል። እንደዚህ አይነቱ አደጋ ለማስወገድ ከኢትዮጵያ መንግስት ጋር ከአለም ባንክ፣ ከሌሎች ለጋሽ መንግስታትና በታወቁ አለም አቀፍ መንግስታዊ ያልሆኑ ድርጅቶች ጋር ሆነን በጋራ እንሰራለን፤ ብለውም አቅጣጫቸውን አሳውቀዋል።

የሚሰሩት ስራ ምን እንደሆነ ዘርዝረው ድንገተኛ እርዳታ ላይ ከማተኮር ይልቅ፣ ለዘለቄታ የሚበጅ ሌላ አማራጭ ላይ ለማገዝ በኢትዮጵያ መንግስት ጋር አብረው እንደሚሰሩ፣ የገጠር መሰረተ ልማት ለማስፋፋት እንደሚረዱ፣ የተያዘነው የ«ሴፍቲ ኔት» ፕሮግራም እንደሚደግፉ፣ ኢትዮጵያ የነደፈችው «የድህነት ቅነሳ አስትራቴጂ» እውን ለማድረግ ከመንግስት ጋር ሆነው እንደሚጥሩ ሁሉ ገልጸዋል፣ቃልም ገብተዋል። ተግባር ላይ እንዲያውሉት ከመመኘት ውጭ የምንለው የለንም። በቃል የተቀመጠውን በተግባር ትርጉመው እንዲያሳዩን የበረታ ምኞት አለን።

«የመሬት ይዞታ ለውጥ ገበሬዎች በመሬታቸው ላይ ይበልጥ ገንዘብና ጉልበት እንዲያፈሱ የሚያደርግና የእርሻ ምርትም እንዲያድግ የሚያደርግ ነው» ብለው በግልጽ አስፍረውታል።

በዚህ ሁኔታቸው ተመስርተው እንደ አውሮፓዊያን አቆጣጠር 2006 ዓ.ም. ድረስ በብዙ ክልሎች ገበሬው በያዘው መሬት ላይ የሙሉ ተጠቃሚነት መብቱን የሚያረጋግጥ ለውጥ እንዲመጣ እንደሚደግፉ ገልጸዋል። የማያሻማና ግልጽ የሆነ የተጠቃሚነት መብት የሚያረጋግጥ የመሬት ይዞታ ለውጥ ("Land Reform") እንዲደረግ መገፋፋት ብቻ ሳይሆን በመርሃ ግብራቸው አንጻር ኤ.አ በ2004 በሁለት ክፍሎች፣ በ2005 በተጨማሪ በሌሎች ሶስት ክልሎች እስከ 2006 ዓ.ም ደግሞ በቀሩት ሁለት ክልሎች ይህ ለውጥ እንዲካሄድም እንደሚያበረታቱ እና ለዚህ ለውጥ መንግስትን እንደሚደግፉ አስምረውበታል።

በዚህ የመሬት ፖሊሲ ለዉጥ ዙሪያ የሚነሳው ቀዳሚውና ትልቁ ጥያቄ የበለጸጉት ሀገሮች የመሬት ይዞታ ለውጥ የሚሉትን ነገር ኢሕአዴግ ይቀበለዋል ወይ የሚል ነው።ኢሕአዴግ እንደ ፓርቲም፣ እንደመንግስትም በተደጋጋሚ በጽሁፍም በቃልም የሚገልጸው አቋም አለ። መሬት በመንግስት ቁጥጥር እንጂ በግል ባለቤትነት እንደማይያዝ፣ መሬት እንደማይሸጥ እንደማይለወጥ በተደጋጋሚ ይናገራል። በአጽንኦት አራት ነጥብ እያለ የሚያሰምርበት አመለካከትና ፕሮግራም ነው።

ስለዚህ የመሬት ፖሊሲ ለውጥ ወይም የመሬት ይዞታ ለውጥ ወይም በነሱ ቋንቋ /Land Reform/ ሲባል ምን ዓይነት ለውጥ የሚያመለክት ነው? ኢሕአዴግ አመለካከቱን ቀይሯል ማለት ነው? ወይስ መሠረታዊ ያልሆነ አመለካከት ነው የመሬት ፖሊሲ ወይም የመሬት ይዞታ ለውጥ የተባለው?

በነገራች ላይ ኢሕአዴግ የመሬት ፖሊሲውን ቀይሮ መሬት በግል አይሸጥም፣ አይለወጥም ይለው የነበረውን ትቶ በግል ይዞታ እንደገና አይፈቅድም፣ መሬት እንዲሸጥም፣ እንዲለወጥም አያደርግም ማለታችን አይደለም። የያዘው አቋም ከሕዝብ ጥቅም አንጻር አይቶ የማስተካከል ችግር አለበት ማለታችን እንጂ ስልጣን ላይ ለመቆየትና ሥልጣን ጠበቅ ለማድረግ እስከጠቀመው ድረስ የመሬት ፖሊሲውን ቢለውጥ ይገርማል ማለታችን አይደለም። በተለይ በአሁኑ ጊዜ የሚታየው የኢህአዴግ አካሄድ እንደ የዕቃ ግዢና ሽያጭ ማስታወቂያ «የተሻለ መንገድ ከተገኘ በጨረታው አይገደድም» የሚል መሆኑ የታወቀ ነው።

Latinaized form of the above news item that is put into the parser and then to the system after the necessary pre-processing

Ewenet ehadeg Yemeret policy Lewet Yekebelal?

G-8 eyetebalu yemitawekut yealemachen sement yebeletsegu ageroch meriwoch bekirbu baderegut sebseba yeafrican guday besefat teweyayetewbetal. Kerehabuna kedirku gar beteyayaze ajendam leafrica kend huneta tekuret setew akababiwen lemerdat mereha geber awetetewal.

beyazenw haya andegnaw kefle zemen rehab lemekelakel endemichalna lezihum asfelagi erdata endemiyadergu geltsewal. Yeh bergetem mehon yalebet guday sihon asfelagi metebaberna metegagez binor noro beyazenew haya andegnaw kefle zemen sayhon balefew hayagnaw kefele zemen rehaben lemekelakel masweged yechalen neber.

simintu yebeletesgu hageroch betedegagami sele ityopiya yanesut guday lalefut haya ametat keselsa sement million yeityopia hezeb gimashu berehab, bemegeb wastena matat mesekayetun neber. Ahunem bihon amest million yeltiopia hezeb ejig kebad behone yemegeb etot chegir enedmisekay askemetewal. Enedezih ayenetu adega lemasweged keltiopia mengest gar kealem bank, keleloch legash mengestatna betaweku alem akef mengestawi yalhonu dirijitoch gar honew begara enseralen belom aketachawen asawekewal.

Yemiserut sira min endehonem zerzerew dingetegna erdata lay kematekor yilik lezeleketa yemibej lela amarach lay lemageth keetyopia mengist gar abrew endemiseru, yegeter meseret limat lemasfafat endemiredu, yeteyayaznewn yeseftinet program endemidegfu etyopia yenedefechewn "yedihinet neqesa" strategy ewun lemadreg kemengist gar honew endemisru hurlu geltsewal, qualm gebtewal. tegbar lay endiyawlut kememegnet wuchi yeminlew yelemn. bequal yetekemetewn betegbar tergumew endiyasayun yebereta mignot alen.

"yemeret yizota lewt geberewoch bemeretachew lay yibelt genzebna gulbet endiyafesu yemiyadergna ye ersha mirtinim endiyadig yemiyaderg new" bilew begiltse asfirewutal.

Bezih huneta temesritew end awropawyan akontater 2006 amete mihret dires bebizu kililoch geberew beyazew meret lay yemulu tetekaminet mebtun yemiyaregagit lewt endimeta endemidegifu geltswal. yemayashama ena gilts yehone yetetekaminet mebt yemiyaregagt yemeret yizota lewt " land Reform" endiyaderg megefafat bicha sayhon bemerhagibrachew antsar ende awropawyan akotater be 2004 be hulet kililoch, be 2005 betechemari beleloch sost kililoch, eske 2006 degmo bekerut hulet kililoch yih lewt endikahed endemiyaberetatu ena lezih lewt mengistn endemidegifu asmirewbetal.

Bezih yemeret polisi lewt zuria yeminesaw kedamiwna tiliku tiyake yebeletsegut hageroch yemeret yizota lewt yemilutn neger ehadeg yikebelewal wey yemil new. Ehadeg ende partim, endemengistm betedegagami betshufm beqalm yemigeltsew aquam ale. meret bemengist kutitir enji begil balebetnet endemayiyaz, meret endemayshet endemaylewet betedegagami yinageral. Be atsnot arat netib eyale yemiyasemrebet amelekaketna program new.

Silezih yemeret polisi lewt woym yemeret yizota lewt woym benesu quwanquwa /land reform/ sibal min aynet lewt yemiyamelekit new? Ehadeg amelekaketun keyirual mallet new Weys meseretawi yalhone amelekaket new yemeret polisi weym yemeret yizota lewt yetebalew?

Benegerachin lay ehadeg yemeret polisiwn keyro meret begil ayshetim, aylewetim yilew yeneberewn tito begil yizota endegena ayfekdm, meret endishetim, endilewetim ayadergm maletachin aydelem. yeyazewn aquam kehizb tikim antsar ayto yemastekakel chigir alebet maletachin endji siltan lay lemekoyetena siltan tebek lemadreg esketekemew dres yemeret polisiwn bilewit yigermal maletachin aydelem. beteley bahunu gize yemitayew yeehadeg akahed ende eqa gizhi ena shiyach mastawekia "yeteshale menged ketegegne becheretaw aygededim" yemil mehonu yetaweke new.

Manually generated extract of the above news item by two professionals (linguists) which is used as a target extract. Note that sentence extraction is used.

ewenet ehadeg yemeret policy lewet yekebelal

g-8 eyetebalu yemitawekut yealemachen sement yebeletsegu ageroch meriwoch bekirbu baderegut sebseba yeafrican guday besefat teweyayetewbetal

yemiserut sira min endehonem zerzerew dingetegna erdata lay kematekor yilik lezeleketa yemibej lela amarach lay lemageth keetyopia mengist gar abrew endemiseru yegeter meseret limat lemasfafat endemiredu yeteyayaznewn yeseftinet program endemidegfu etyopia yenedefechewn "yedihinet neqesa" strategy ewun lemadreg kemengist gar honew endemisru hurlu geltsewal qualm gebtewal

yemayashama ena gilts yehone yetetekaminet mebt yemiyaregagt yemeret yizota lewt " land reform" endiyaderg megefafat bicha sayhon bemerhagibrachew antsar ende awropawyan akotater be 2004 be hulet kililoch be 2005 betechemari beleloch sost kililoch eske 2006 degmo bekerut hulet kililoch yih lewt endikahed endemiyaberetatu ena lezih lewt mengistn endemidegifu asmirewbetal

simintu yebeletesgu hageroch betedegagami sele ityopiya yanesut guday lalefut haya ametat keselsa sement million yeityopia hezeb gimashu berehab bemegeb wastena matat mesekayetun neber

"yemeret yizota lewt geberewoch bemeretachew lay yibelt genzebna gulbet endiyafesu yemiyadergna ye ersha mirtinim endiyadig yemiyaderg new" bilew begiltse asfirewutal

bezih yemeret polisi lewt zuria yeminesaw kedamiwna tiliku tiyake yebeletsegut hageroch yemeret yizota lewt yemilutn neger ehadeg yikebelewal wey yemil new

yeh bergetem mehon yalebet guday sihon asfelagi metebaberna metegagez binor noro beyazenew haya andegnaw kefle zemen sayhon balefew hayagnaw kefele zemen rehaben lemekelakel masweged yechalen neber

Extracted by:

_____ _____

Date _____ *Date* _____

Summary generated by the developed model during evaluation

ewenet ehadeg yemeret policy lewet yekebelal

yemayashama ena gilts yehone yetetekaminet mebt yemiyaregagt yemeret yizota lewt " land reform" endiyaderg megefafat bicha sayhon bemerhagibrachew antsar ende awropawyan akotater be 2004 be hulet kililoch be 2005 betechemari beleloch sost kililoch eske 2006 degmo bekerut hulet kililoch yih lewt endikahed endemiyaberetatu ena lezih lewt mengistn endemidegifu asmirewbetal

yemiserut sira min endehonem zerzerew dingetegna erdata lay kematekor yilik lezeleketa yemibej lela amarach lay lemageth keetyopia mengist gar abrew endemiseru yegeter meseret limat lemasfafat endemiredu yeteyayaznewn yeseftinet program endemidegfu etyopia yenedefechewn "yedihinet neqesa" strategy ewun lemadreg kemengist gar honew endemisru hurlu geltsewal qualm gebtewal

benegerachin lay ehadeg yemeret polisiwn keyro meret begil ayshetim aylewetim yilew yeneberewn tito begil yizota endegena ayfekdm meret endishetim endilewetim ayadergm maletachin aydelem

yeyazewn aquam kehizb tikim antsar ayto yemastekakel chigir alebet maletachin endji siltan lay lemekoyetena siltan tebek lemadreg esketekemew dres yemeret polisiwn bilewit yigermal maletachin aydelem

"yemeret yizota lewt geberewoch bemeretachew lay yibelt genzebna gulbet endiyafesu yemiyadergna ye ersha mirtinim endiyadig yemiyaderg new" bilew begiltse asfirewutal

yeh bergetem mehon yalebet guday sihon asfelagi metebaberna metegagez binor noro beyazenew haya andegnaw kefle zemen sayhon balefew hayagnaw kefele zemen rehaben lemekelakel masweged yechalen neber

Appendix five

Amharic news item Taken form Reporter news paper (May 2004)

የወሊድ ችግርን ለመፍታት የሰለጠኑ አዋላጅ ነርሶችን ማፍራት::

በኢትዮጵያ በመውለድ እድሜ ክልል ካሉት እናቶች አብዛኞቹ በእርግዝና ከወዲድ ጋር ተያያዥነት ባላቸው ሕመሞች ሕይወታቸውን ያጣሉ:: ከሚወልዱ አንድ መቶ ሺህ እናቶች መካከል ስምንት መቶ ሰባ አንዱ የዚህ ችግር ሰለባ እንደሆኑ ጥናቶች ያሳያሉ:: በሕይወት ከሚወለዱ 1000 ሕፃናት መካከልም 96.8 ያህሉ እንደተወለዱ ይሞታሉ:: ይህን ችግር ለማስቀረት በተቻለ መጠን የሰለጠኑ አዋላጅ ነርሶችን በጥራትና በብዛት ማፍራት የግል ይላል:: በዚህ ሙያ የተሰለፉ ነርሶች መኖር ማለት ደግሞ በእርግዝናና በወሊድ ምክንያት የሚሞቱትን ብዙ እናቶች የሚታደግ ኃይል እንደማደራጀት ይቆጠራል::

አዋላጅ ነርሶች የእናቶችንና ሕፃናትን ጤንነት ማሳደግ፣ ለጉዳትም በጥብቅና የቆመና በመታገልም ላይ ያሉ ሲሆኑ ይገባል:: የእናቶችን ጤና ከመጠበቅ አኳያም ለወጣት ሴቶች መውለድ ዕድሜ ክልል ያሉ የማዋለጃ አካላትን ጤንነት መጠበቅና መውለድን ለመወሰን የሚያስችል ትምህርት የማስጨበጥ ኃላፊነት ከአዋላጅ ነርሶች ይጠበቃል:: ጥናቶች እንዳመለከቱት በኢትዮጵያ የእናቶችና የሕፃናት ጤና አገልግሎት በአብዛኛው የሚያተኩረው የሚወለደውን ሕፃን ወደ ጐን በመተው በወላጆች እናት ላይ ያዘነበለ ነው::

ሲስተር ኪሮስ ከበደ የኢትዮጵያ አዋላጅ ነርሶች ማህበር ፕሬዚዳንት አብዛኞቹ ወላጆች የሚገላገሉት በልምድ አዋላጆች እንደሆነ ያስረዳሉ:: በሠለጠነ ባለመያና በጤና ተቋማት የሚወልዱት የእናቶች ቁጥር 10 ከመቶ አይበልጥም:: ከዚህ አኳያ ለልምድ አዋላጆች የሙያ ማሻሻያ ሥልጠና መስጠቱ በእጅጉ ወሳኝ እንደሆነ ነው የሚናገሩት::

የዓለም ጤናና የሕፃናት መርጃ ድርጅቶች ለልምድ አዋላጆች የሙያ ማሻሻያ ሥልጠና በመስጠት ከፍተኛ እገዛ ማበርከታቸው አይዘነጋም:: ሆኖም እነዚህ ሁለት የተባበሩት መንግሥታት ድርጅት አካላት ይህ ዓይነቱን ሥልጠና የመስጠት ተግባር ካቆመ ዓመታት ተቆጥረዋል:: ለዚህም የሚሰጡት ምክንያት «የእናቶች ሞት ብዙም አልቀነሰም በአጠቃላይ ውጤት አልታየበትም» የሚል መሆኑን ሲስተር ኪሮስ ይገልፃሉ::

ሲስተር ኪሮስ እንደገለፁት በኢትዮጵያ ውስጥ 862 አዋላጅ ነርሶች አሉ:: ከእነዚህም መካከል 350 ያህሉ የማኅበሩ አባላት ናቸው:: የቀሩትንም በአባልነት ለማቀፍ የሚያስችል ሥራ እየተከናወነ ነው:: በሙያተኞቹ አልፎ አልፎ የሥነ-ምግባር ጉድለቶች እንደሚታይና ይህም ሊሆን የቻለው የሕክምና ተቋሙ ለሙያተኞቹ የሚያስፈልገውን የሥራ መገልገያ ካለሟሟላት፣ ካለማደራጀትና ካለማቅረብ የተነሳ ነው::

Latinized form of the above news item

yewelid chigirn lemeftat yeseletenu awalaj nersochn mafrat.

beetiyopia bemewled edme killil kalut enatoch abzagnochu beergizina kewelid gar teyayazinet balachew himemoch hiywotachewn yatalu. kemiwoldu and meto shih enatoch mekakel simint meto seba andu yezih chigir seleba endehonu tinatoch yasayalu. behiywot kemiwoledu 1000 hitsanat mekakel 96.8 yahilu endetewoledu yimotalu. yihn chigir lemaskeret betechale meten yeseletenu awalaj nersoch betiratina bebezat mafrat yegid yilal. bezih muya yeteselefu nersoch menor malet degmo beergizinana bewolid miknyat yemimotutn bizu enatoch yemitadeg hayl endemafrat yikoteral.

awalaj nersoch yenatochinena yehitsanatn teninet masadeg legudatim betibkina yekomena bemetagelm lay yalu lihonu yigebal. yenatochin tena kemetebeq aquwyam lewetat setoch mewled edmie killil yalu yemawaleja akalatin teninet metebeqena mewledin lmewesen yemiyaschil timihirt yemaschebet halafinet keawalaj nersoch yitebeqal. tinatoch endemiyamelekitut beetiyopia yeenatochina yehitsanat tena agelgilot beabzgnaw yemiyatekurew yemiweledewn hitsan wede gon bemetew bewelaj enat lay yazenebele new.

Sister kiros kebede ye ityopia awalaj nersoch mahber presedant abzagnochu welajoch yemigelagelut belimd awalajoch endehone yasredalu. Beseletene balemuyana betena tequamat yemiweledut yehitsanatoch kutir asir kemeto ayberltim. Kezih akuaya lelimd awalajoch yemuya mashashaya siltena mestetu be ejigu wesagn endehone new yeminagerut.

Yealem tenana yehitsanat merja dirijitoch lelimd awalajoch yemuya mashashaya sitena bemestet kefitegna egeza maberketachew ayzenegam. Honom enezih hulet yetebaberut mengistat dirijit akalat yih aynetun siltena yemestet tegbar kakomu ametat tekotrewal. Lezihm yemisetut mikniyat "yeenatoch mot bizum alkenesem Beatekalay wutet altayebetim" Yemil mehonun sister kiros yigeltsalu.

Sister kiros endegeletsut beityopya wust 862 awalaj nersoch alu. Kenezihm mekakel 350 yahlu yemahberu abalat nachew. Yekerutinm beabalnet lemakef yemiyaschil sir eyetekenawene new. Bemuyategnoch alfo alfo yesine-megbar gudletoch yemitayena yihm lihon yechalew yehikimna tekuamu lemuyategnochu yemiyasfeligewn yesira megelgeya kalemamualt, kalemaderajet ena kalemakreb yetenesa new.

DECLARATION

This thesis is my original work, has not been presented for a degree in any other university and all sources of material used for the thesis have been duly acknowledged.

Kamil Nuru

THE THESIS HAS BEEN SUBMITTED FOR EXAMINATION WITH MY APPROVAL AS AN ADVISOR

Dr. Björn Gambäck