ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Automatic Text Summarizer for Tigrinya Language

By

Guesh Amiha Birhanu

A THESIS SUBMITTED TO COLLEGE OF NATURAL SCIENCE OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

ADDIS ABABA, ETHIOPIA

February, 2017

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Automatic Text Summarizer for Tigrinya Language

By

Guesh Amiha Birhanu

APPROVED BY THE EXAMINING BOARD:

Advisor: Dr. Wondwossen Mulugeta _____ _____

Signature Date

Examiner: Dr. Martaha Yiferu _____ _____

Signature Date

Examiner: Ato Ermias Abebe _____ _____

Signature Date

Chair person: _____ _____

Signature Date

Acknowledgments

First and for most I am truthfully grateful to the enormous God, for giving me the strength and perseverance in my life. I owe my deepest thankfulness to my advisor Dr. Wondwossen Mulugeta (PHD) for his patience putting up with my flaws, unnerved encouragement and enlightening comments and discussion throughout the course of this thesis. Of all the qualities of Dr. Wondwossen, his kindness touched me most. Thanks Dr. Wondwossen for being who you are.

I know I tend to be an irritation when things are not going well in my way and putting up with this is very difficult. Fortunately I am blessed with a wonderful family who has a limitless tolerance. I really appreciate this with lots of love and respect. My special thanks also go to my wife Mlashu Gidey, my lovely child Yeabsra Guesh and my bother Halefom Amiha for their moral support and encouragement during my study.

It is an honor for me to give my sincere gratitude to Mr. Tewelde Tesfay who supports me to assign who voluntarily stood to be human annotator. This research is not possible without their expert knowledge on human summarizing.

Finally, I extend my heartfelt thanks and respect to my friends and all those people who were not mentioned here but their contributions have been inspiring for the completion of this work.

Contents

Abstract

With the continous increase in the number of electronic documents the need for faster techniques to asses the relevance of documents emerges. An ideal summary is one that conveys to the reader the main themes of the document and consequently the rader can determine weather the complete document does have any relevance. Automatic text summarization is a technique where a program summarizes a longer text to a shorter and non redundant extract of the original text.

In this thesis, two generic text summarization methods that create text summaries by ranking and extracting sentences from the original documents are proposed. The first method is term frequency that employs frequency of word to identify the relevant sentences that contains the frequent words. The frequent words are top frequent words from the original document and sentences intersections of the top frequent words are important sentences for summary generation. The second method title words identify the title words of the document and extract sentences that contain the title words to include in the summary.

For experimenting purpose we have used 30 news articles, which are collected from the sources of aiga forum and dmtsi woyane tigray web sites. Evaluation of the summarization system is then conducted by comparing the the system's summaries with manual summaries that are generated by human evaluators. According to the experimentation done the system registered 0.46(46%), 0.47(46%) and 0.46(46%) for recall, precision and F-Score respectively for the feature of term frequency. In the case of title word the registered recall, precision and F-Score values were 0.46(46%), 0.50(50%) and 0.48(48%) respectively which shows the improvement of the summarizer with this method. In general according to the experiment results show the best performer feature was title word than term frequency in both subjective and objective evaluations.

The challenging task in the study was lack of standardized and well prepared Tigrinya corpus which required conducting conclusive experimentation of the proposed system and these will be future research directions in this area which contribute in the improvement of the system.

List of Tables

List of Figures

List of Algorithms

List of Acronyms

ANSI American National Standards Institute

AW Augmented weight

BW Binary weight

EF Entropy frequency

FW Frequency weight

GFID Global frequency inverse document

IDF Invert document frequency

IR Information Retrieval

LW Logarithm weight

NLP Natural Language Processing

OTS Open Text Summarizer

P Precision

R Recall

ROUGE Recall Oriented Understudy for Gisting Evaluation

SF significance factor

TF term frequency

W weight

CHAPTER ONE

1. Introduction

1.1 Background

The speedy growth of information in broadcasting, online information service and the internet in general makes the accessibility of information very easy for us. Due to this every time someone searches something on the internet, the response obtained is lots of different WebPages with a lot of information which is impracticable for a person to read completely the whole that is written. That's why a new technology is needed to manage the vast amount of information within a short period of time that is automatic text summarization as a solution.

The technology of automatic text summarization is one tool that can help the easy capturing of the information available. That is the summarized document can give information to the readers in a very precise and concise form without losing the theme of the document. Basically text summarization can be accomplished manually with professionals and automatically generating the important sentences. A summary is a text produced from one or more texts that contain the important portion of the information from the original text and that is no longer than half of the original document [1]. According to [2] text summarization is the process of extracting the most relevant information from a source document to produce condensed information for a particular user or for particular task. When this is done automatically through computer assisted it is called automatic text summarization.

Automatic text summarization systems can be categorized in to two types that are abstraction based summarization and extractive based summarization [3]. Abstractive summarization tries to suggest summarized information by selecting the information from the most important section of the document and rephrase it possibly including new words that are not found on the original document. Extractive summarization on the other hand summarizes the document by using the words, sentences and paragraphs that exist and organizing them in a way to produce an articulate summery. Extractive summarization is easier than abstractive summarization and currently it is the general practice among researchers in the area of automatic text summarization [4]. This research is also aiming at extractive summarization.

Text summarization can also be classified single document, preparation of summary from one document or multi document, which is preparing the summary from different documents. In this research we are also going to concern a single document summarization.

1.2 Statement of the problem and justification

The continuous flow of information through broadcastings, World Wide Web and online text collections is available for every user in today's environment. As a result of this continuous flow of information users or readers are in front of an information overload problem because of written in un wanted details that is difficult to cover in short period of time.

Tigrinya text readers are also suffering this kind of problem. A growing number of Tigrinya news service providers are publishing their content online. To point out some of them like *Dmtsi Woyane Tigray, Aigaforum and FM Mekelle* have been updating their content website regularly. Due to this high amount of information flowing through the digital environment there is a need of summarizer that gives a condensed and precise type of information for the expected readers.

To the researcher knowledge there is no text summarization done for Tigrinya language before. But there are a lot of researchers done in other local languages like Amharic and Afan Oromo and their central idea is the problem of flowing vast amount of information accessibility which is an information overload that needs automatic summarization system to minimize the time and cost of the reader. Some of the recent works done on the area of automatic text summarization are [5, 6, 7, 8, 9, and 10] for Amharic and [11 and 12] for Afan Oromo. So there is a need to develop automatic text summarization system for Tigrinya language also. The method is language independent but in the experiment we use stemming, stop word list and normalization which are language specific tools or resources [51].

Therefore the aim of this study is to explore and design an automatic text summarizer for Tigrinya language that process texts to extract the most important information from a source (or sources) to produce an abridged version for a particular users using extraction method.

To this end, this study aims to answer the following research questions.

- Does the proposed system generate summary that has well in structure and coherence of sentences?
- Which of the proposed approach is the best performer in selecting the important sentences either term frequency based or title words based?

1.3 Objective of the study

The general objective of the study is to explore and develop an automatic text summarizer for Tigrinya news articles.

In order to achieve the main objective and answer the research questions the following are the specific objective of the study.

- ✓ To review related works journals, books, conference papers, articles in the area of automatic text summarization and especially on the area of title based and term frequency based salience sentences and prepare data corpus.
- ✓ To design a generic model for term frequency and title based sentence extractive summarizer for document written in Tigrinya
- ✓ To develop a prototype summarizer that accepts Tigrinya text as an input and produce extracted output
- ✓ To prepare data from different socurces for the experiment
- ✓ To conduct experiments to evaluate the performance of the proposed system

1.4 Scope of the study

The study focuses on developing salient sentence based text summarization based on term frequencies and title words of the document for Tigrinya language. This thesis further focuses on the particular nature of news texts to enhance the use of term frequency and title word for summarization. The summarizer doesn't process document with various styles such as table, graphs, image and figures are out of focus of the research.

1.5 Methodology

This research would be in order to figure out challenges of implementing an automatic text summarization for Tigrinya language. Experimental research methodology has been used

for this study. To accomplish this and to meet the prescribed objectives the following sub tasks have been carried out in detail.

1.5.1 Literature Review

In order to achieve the objective of the research books, journals, articles and conference papers on automatic text summarization especially which use statistical term frequencies of sentences that give thematic assumption of a document would be reviewed. In addition to that as there is no previous work done on Tigrinya text summarization we were review other related works done on other local languages like Amharic and Afan oromo.

1.5.2 Data Corpus

The dataset that we use for evaluating the proposed summarization system contains 30 Tigrinya news articles whose lengths could be ranging from 16-42 sentences. These news articles have been collected from the web site of *Dmtsi Woyane Tigray, Aigaforum and FM Mekelle*. Short news items with less than 16 sentences were not used in the evaluation due to the fact that summarizing short news articles doesn't make much sense in real applications [50]. Though evaluating the system with news containing more than 42 sentence is very desirable, but it is difficult to obtain such news articles. In addition to that the previous researchers done on this area were used around this number of news articles. Furthermore, the news articles used for the evaluation were on different domains (politics, sport, society) which help to evaluate the performance of the system for different domains. Before experimentation, manual cleaning, removing irrelevant information such as tables and figures has been performed on the articles.

1.5.3 Manual Summary Preparation

For manual summery preparation three independent human summarizers who are working in Mekelle University, could be employed for preparing manual summary for the 30 news articles. And for each news articles three summaries were prepared by ranking the whole content of the document.

1.5.4 Implementation

1.5.4.1 Methods

In this research we have used an automatic text summarization based on excretive approach that are title identification in order to measure with the whole content of the document and term

frequency throughout the document could be applied to extract the summary. Based on those two mechanisms lastly we compare and contrast which way is the best for the creation of the summary of a given document.

1.5.4.2 Development Tools

Python programming language is used to build the summarization system. The reason we selected this programming language is the researcher's familiarity with the language and Python has different natural language toolkits that can be easily imported.

1.5.5 Evaluation

Performance evaluation were conducted by comparing the system summaries with their corresponding manual summaries. The evaluation metrics that was used are the well known IR metrics, precision, recall, and F-Score. We would also use the comparison method of summary that is based title words and term frequency of the document. We would also evaluate the system summary by human judgment.

1.6 Significance of the study

The main contribution of this research focuses on finding an efficient method of automatic text summarization system for Tigrinya news texts. It can also be useful for initiating text summarization researchers using other approaches for the Tigrinya language. To do this those are the significances of the research.

- ✓ It motivated other researchers in applying other summarizing techniques to Tigrinya language and finding their applicability to Tigrinya text.
- ✓ Could benefit in information acquisition tasks such as to promote current awareness and save readers time.
- ✓ It can serve as an input for other researches to be done on text summarization for the Tigrinya language.

1.7 Organization of the Thesis

This thesis is organized in five chapters. The first chapter, presents out the background, statement of the problem, and the general and specific objectives of the study together with scope, limitations study and significance of the study are included.

. Chapter two is literature review and it involves two main topics, related work and conceptual review. Conceptual review is review on basic norms of text summarization, concepts of automatic text summarization, process of automatic text summarization and related topics and related work involves work done so far on the research topic. Chapter three discusses about the Tigrinya language in general.

Chapter four discusses about the system architecture used in this research. Chapter five discusses the experiment and result of the study. In this part corpus selections and preparations, implementations of the proposed work, experimentations, findings of the study, and issues in implementations are discussed in detail.

Finally in chapter six conclusions and works identified as future work and needs to get attention of other researchers are listed in recommendation section.

CHAPTER TWO

2. Literature review

2.1 Introduction

In this chapter we have been explained the overall varieties of an automatic text summarization in terms of its type, the different approaches and the state of art in general. As text summarization is an active research in today's activities there are a lot of works done on this area. To this end we were presented the works from the starting until now in which how different researchers done their research and what they gain after all.

As the proposed work is on the statistical approach specifically extraction of salient sentences based on their term frequency and title of words of the sentences. For the summarization technique we were reviewed related works to those techniques and other approaches of automatic text summarization in general. Beyond that evaluating the performance of the system is also very challenging job and various evaluation mechanisms have been proposed. The principal methods of evaluating summaries are also reviewed in this chapter as well.

2.1.1 Basic notions of automatic text summarization

Automatic text summarization is a procedure by which the most significance concept in a document are identified and then obtainable in a condensed way. Its main objective is decreasing the complexity and length of the original text at the same time as keeping the most important information of the text. To this extent the produced summary should be non repetitive and should give as much as possible accurate and relevant information of the original document. In short the summary should allow the reader to answer questions about the main ideas in the given text or work as a reference pointer to parts of the original text.

A summary is a text that is created from one or more texts that contains the relevant terms of the original document and the summary couldn't be more than half page [1]. The text could be including of multimedia documents, online documents and hypertexts etc. text summarization is the process of condensing the most significant information from the source to produce an abridged version for a particular user [2]. When this type of process is done through computer or automatic system it is called an automatic text summarization. Examples of naturally occurring

summaries include news headlines, scientific abstracts, movie previews and reviews, meeting minutes, TV guides, weather bulletins, drastically condensed books etc.

Since abridgment is crucial, that is an important parameter to summarization it is the level of compression (ratio of summary length to source length) desired. According to Mani [2] as the quality of the summary can be measured by the amount of relevant information present and omitted the length of the summary presents an issue. The higher the compression ratio the more probable important information was omitted. For example a journal paper of 20 pages which is summarized in only half a page, has a compression ratio of 2.5% , while a newswire article compressed to the same size probably has compression ratio between 20% and 30% extraction rate could be contain more relevant information.

2.1.2 History of automatic summarization

Experiments on summarizing texts using computers were begun in the late 1950 by characterizing surface level approach. Lhun [13] is the opening researcher on the area of automatic text summarization that applies statistical approach measuring significance of individual words and then sentence. Based on this the high frequent sentences can make the summary. Whereas the first entity level approach of text summarization is based on syntactic analysis appeared in the early 1960 [15]. The other researcher who is called Edmonson [14] uses different relevant characteristics of the text and program the computer to recognize and weight them. The automatic extraction method uses four methods that are cue, key, title and location. In the early 1970 there was renewed interest in the field, with extension being developed to the surface level approach to include the use of cue phrases (bonus verses stigma items), and which resulted in the first commercial application of automated abstracting [16]. Progressively in the late 1960s the field of automatic summarization grew aggressively with all type of approaches being explored already due to the government and commercial interest for the application domains.

Currently the research works have exclusively focused on extracts rather than abstracts along with the renewed interest in earlier surface-level approaches. However, more natural language generation works have been begun to focus on automatic summarization and the field is now exploring new areas such as multi document summarization, multi lingual summarization and multimedia summarization rather than focusing on single document summarization.

2.1.3Types of summary

There are different types of summaries identified which are generally fall in to either inductive or informative groupings. An indicative summary is one that provides an idea of what the document is about or it indicates the document's relevance to the reader. That gives condensed information on the main topics of the document by preserving the most important passages of the document. Indicative summaries are often returned by search engines as a response to user queries. Hence they are only meant to help the user decide whether or not to read the full document. On the other hand the purpose of informative summaries is to deliver as much information as possible to the user and to serve as a substitute to the full document. The typical length of an indicative summaries ranges between 5 to 10 % of the full document and that of informative summaries ranging between 20 to 30 % of the complete text [3].

Furthermore text summarization can be also classified in to extractive and abstractive approaches. In extractive summarization the technique involves assigning scores to some units of the document text for example words, sentences and paragraphs of the documents and extracting those words, sentences or paragraphs based on their highest score on their importance or significance measure. Abstractive summarization on the other hand deals with techniques usually needs information fusion, sentence compression and reformulation and the words or phrases that makes the summary may not be include on the original document. The enormous majority of the current summarization systems are focusing on the extractive summarization. This is because abstractive summarization relies on a deep understanding of natural languages which is a very challenging task that is not successfully achieved yet [3].

We can also distinguish two types of summarization based on the volume text to be summarized as single document and multi document summarization. If the summarization is performed for a single text document than from different documents it is called single document text summarization. On the other hand if the summarization is to be performed for multiple text documents it is called multi document summarization. Summarizing a single document is a challenging task but the multi document summarization poses several additional challenges. In order to avoid repetitions, it should have to identify and locate thematic overlaps and has also inconsistencies between the documents. For this reason multi document summarization is not as much developed as single document summarization [1].

Another criterion for the classification of summarization is based on the purpose of the summarization task. Due to this summary could be generic which tries to represent all relevant features of the source text or it can be query based that the summaries center of attention is on the user's query. In short generic summaries are text driven where as query driven or user focused ones rely on the specification of the user's information need like a question or keyword. Until recently generic summaries were more popular one but with the prevalence of full text searching and personalized information filtering, user focused summaries are gaining importance. Query based summaries on the other hand is a simple method to tailor the summary to a user specified subject. Sentences which match the specified query could get a higher score than sentences which do not match the specified query. The summarizer is supposed to construct a summary that contains information requested by the query. In this thesis the researcher aimed generic summarization which is informative enough to be used as a substitute to a single document.

2.1.4 The stages of automatic text summarization

According to Hovey [1] there are three distinct stages of an automatic text summarization. Those are **topic identification**, **interpretation** and **generation**. The **topic identification** stage is performed by assigning a score for the different units of the input text such as words, sentences or paragraphs and selecting the high scoring using some threshold value according to the summery length requested. The second stage of **interpretation** shows what extractive summarization differs from abstractive summarization systems. During interpretation the topics identified are represented in new terms that are not found in the original text. No system can perform interpretation without prior knowledge of the domain specific because it needs an input in terms something unconnected to the text. But due to the difficulty of building enough domain knowledge very few summarizers have performed interpretation. The third stage of summarization which is **generation** was done after the summary is created weather in extractive or abstractive summarization it exists within the computer in internal notation and thus requires the techniques of natural language generation, naming text planning, sentence planning, and sentence realization.

In the consequence we describe basic operations that are performed in the topic identification stage of the summarization considering its relevance to our thesis. Concepts related to the

creation of automatic text summarization and how representation of the documents and queries could be processed such as term extraction, sentence extraction, and term weighting could be discussed briefly.

The most essential operation required in information retrieval or in natural language processing at large is assigning appropriate term and identifiers capable of representing the content of a collection of documents. This task which is known as indexing can be performed manually by trained experts or it can be performed automatically in modern environments. The process of automatic indexing is a composed of two major tasks. The first task is extracting terms or concepts from each document which are capable of representing the content of the document. The second task is assigning each term a weight or value that signifies its importance for the purpose of content description [17].

A. Index Term Extraction

The task of index term extraction follows a series of activities. Each of them is explained below.

Lexical analysis

Lexical analysis starts with tokenization which is the identification of all the individual words that make up the input text. That is given a character sequence and a defined document unit; tokenization is the task of chopping it up into pieces, called tokens. Tokenization can occur at different levels: a text could be broken up into paragraphs, sentences, words, syllables or phonemes. Punctuation marks and spaces are usually used to infer the beginning and the end of a token. For instance the procedure for identifying words in Tigrinya documents makes use of the Tigrinya word separators such as single space, (netsela serez) (□), colon (klte netbi) (:), dereb serez (□), period (arbate netbi) (::), carriage return, line feed, tab etc.

Lexical analysis also incorporates a sort of text cleaning process in addition to tokenization. The text cleaning process removes numbers and symbols such as 2000, @, %, #, etc. that don't make a good index terms. It also converts abbreviations and acronyms in their full text, and merged hyphenated words. For instance, the hyphenated word □/□□□ could be treated as single word "□□□□□□" and the abbreviation □.□ is expanded to □□□□□□□. The text cleaning process helps to avoid errors which are against the syntactical rules of the language under consideration [17, 18, and 19].

Normalization

There are different Tigrinya letters that have the same sound and such letters have been used interchangeably in a given word. These different symbols must be considered as equivalent because they do not cause changes in meaning. As a result in this research all different symbols of the same sound were converted to one common form. For examples the characters □ and □ have similar sound (with the sound se). These two characters with equivalent sound are converted to □ (se). There are also other characters that have same sound different structure like □ and □ are characters with the same sound and are changed to □ (tse). Normalization handles such type of inconsistency in writing words by changing characters of the same sound to a common form. This avoids the unnecessary representation of a given word in different forms [13]. This is especially useful in keyword identification which is employed in this thesis to identify important sentences in a given document.

Stop-word removal

The words of a document text do not have equal value for indexing purpose. Some words are lexical devices that serve grammatical purposes and do not refer to object or concept. The common words in English such as of, a, and the, are stop-words and such words are generally used to "glue" sentences together but usually do not carry meanings. Thus such words could be removed from the text by comparing each term in the text with a list of common words developed for a particular language and sometimes for a particular domain.

Tigrinya domain independent stop words include prepositions, conjunctions, and articles. Tigrinya languages have also its own stop words such as the articles of Tigrinya "iti", "ita" and "itom" and the conjunctions "kab", "ab" and "nab" are examples of such highly frequent words. So removing such kind of stop words can be also reducing file size and processing time. From summarization point of view, removing such words from the document guarantees that sentences would not be favored for inclusion in a summary just because they contain highly occurring stop-words. Stop-word removal also reduces the complexity of the document representation and the number of tokens to be processed [20].

Stemming

After removing the stop-words in a document, that remaining words are stemmed to their root form if they have morphological variants. This is based on the assumption that words with the same stem are semantically related and have the same meaning to the user of the text. Furthermore, bringing verities of a word to a common form reduces the number of different terms needed for representing a document which saves storage space and processing time. For example, the following variants □□□, □□□, □□□□, □□□□, □□□□□, □□□□□ and □□□□ are changed in to their stem word □□□ [20].

Various attempts have been made to develop a stemming algorithm for the Tigrinya language.

B. Term Weighting

After performing tokenization, normalization, stop-word, removal, and stemming, the next step is to find the weight of the terms according to their importance in representing a document. Not all terms are equally important in reflecting the content of a specific text and thus, an importance indicator or a term weight should be associated with each index term. There are many weighting functions and most of them rely upon the distribution pattern of the terms with in a document as well as in the document collection as whole. The weighting functions use these distribution statistics to compute the local (with in a document) and the global (with in a document collection) weight of each term. The weight of a term is then found by talking the product of the term's local weight and global weight.

When term weighting is applied in text summarization, the local weight of a term reflects the importance of the term in the sentence containing the term and the global weight reflects the term's importance in a document. More specifically the weight of term j in sentences i, $a_{ij,}$ is calculated as follows [20]:

$$aij = L(tij).G(tij)$$

Where, t_{ij} denotes the frequency with which term j occurs in sentences I,

$L(t_{ij})$ is the local weight for term j in sentence I, and

$G(t_{ij})$ is the global weight for term j in the whole document.

Major global weighting functions that are used in information retrieval are described below [21, 20].

Local weighting

Local weighting has the following four alternatives:

- ✓ Frequency weight (FW): $L(t_{ij}) = tf_{ij}$ is the number of times term j occurs in sentence I.
- ✓ Binary weight (BW): $L(t_{ij}) = 1$, if term j appears at least once in sentence I; $L(t_{ij}) = 0$ otherwise
- ✓ Augmented weight (AW): $L(t_{ij}) = 0.5+0.5\ (tf_{ij}/tfmax_i)$, where $tfmax_i$ is the frequency of the most frequently occurring term in the sentence.
- ✓ Logarithm weight (LW): $L(t_{ij}) = \log(1+ tf_{ij})$.

The most common local weighting function is frequency weight. It is based on the assumption that the importance of a content term (after stop-word removal) in describing the topic of a document is determined by the frequency of term in the document. That is a content term that appears more frequency in a text is more important than a rarely appearing term. Raw frequency weight doesn't give any distinction between the occurrences of a rare term in a short sentence (document, in the context of information retrieval) and in a long sentence. However the occurrence of a rare term in a short sentence is more significant than its occurrence in a long sentence. Hence the algorithm, binary and augmented weight are often used to smooth this bias [21].

Global weighting

Global weighting has the following four possible alternatives:

- ✓ Invert document frequency (IDF): $G(t_{ij}) = \log (N/n_j) + 1$, where N is the total number of sentences in the document, and n_j is the number of sentences that contain term j.
- ✓ Global frequency inverse document frequency (GFID): $G(t_{ij}) = gf_j/sf_j$, where the sentence frequency sf_j is the number of sentences in which term j occurs and the global frequency gf_j is the total number of times that term j occurs in the whole document.
- ✓ Entropy frequency (EF): $G(t_{ij}) = 1 - \sum pijlog(pij)/\log(nsent)$ where pij tf_{ij}/gf_j and nsent is the number of sentences in the document.

All of the global weighting functions basically give less weight to terms that occur frequently or in many sentences. IDF GFIDF are closely related, both assign a high degree of importance to terms occurring in only a few sentences of a document. However GFIDF increases the weight of frequently occurring terms. In addition neither weighting function considers the distribution of terms over sentences. EF makes use of information theory to measure the importance of a term. It assigns minimum weight to terms that are equally distributed over sentences and maximum weight to terms which are concentrated in a few sentences. EF, unlike IDF and GFIDF, takes in to account the distribution of terms over sentences [22].

2.2 Approaches to text summarization

As the introduction of the field automatic text summarization is introduced by Luhn [13] who is the first researcher in this area. After that a lot of researchers have been proposed different approaches to text summarization that most of them are based on sentence extraction or selection. On useful way is to examine the level of processing. There are different ways of classifying the approaches can be found in the literature and here we have been used the tradition method presented in [23] to offer a brief overview of the traditional techniques used in automatic text summarization this classification is based on the level of processing required to build a summary. Based on this three approaches are identified as a **surface**, **entity**, and **discourse** levels.

2.2.1 Surface level approach

The most primitive works in automatic text summarization was used surface level approaches for deciding which parts of the text are important. This approach have a tendency to represent information in terms of shallow features which are then selectively combined together to yield a salience function used for the extraction of the important information. These important features include thematic features (presence of statistically salient terms, based on term frequencies statistics), location (position in text, position in paragraph, section depth, and particular sections), background (presence of terms from the title or headings in the text), cue words and phrases (for example, in text summary cues such as “in summary”, “our investigation”, emphasizers such as “important”, “in particular”, as well as domain specific ‘bonus’ and ‘sigma’ terms [23].)

The pioneer research for this automatic text summarization based on this approach is luhn [13] that suggests term frequency was important for the sentence relevance to produce the

summarization that represents the whole document. The basic postulation here is the most frequent words in a text are the most representative of its content and the remaining of text containing them are more relevant. However not all text of the words are important and to this end words beyond high and low frequencies as well as those words contained in a stop word list are left out of consideration. This rather unrefined argument on "significance" keeps away from such linguistic implications as grammar and syntax. In general, the method does not even propose to differentiate between word forms. Thus the variants differ, differentiate, different, differently, difference and differential could ordinarily be considered identical notions and regarded as the same word. The remaining words in the document are then sorted alphabetically so that pairs of succeeding words can be compared letter by letter. Therefore wherever the greatest number of frequently occurring different words are found in greatest physical proximity to each other the probability is very high that the information being conveyed is the most representative of the article.

Based on this the significance factor of the sentence is computed using the formula.

$$sf = \frac{(the\ number\ of\ significant\ words)2}{the\ total\ numberofwords}$$

After that sentences that have highest significance factor are extracted to produce a summary or what Luhn calls as "auto-abstracts". The results obtained by Luhn were neither good condensations nor very coherent texts though he believed his "auto-abstracts" were satisfactory indicative abstracts for papers within the science and technology fields [2].

The work of Lhun [13] was further extended through different techniques for extracting the important sentences to formulate the summary. While the previous work was on one feature of sentence significance namely the presence of highly frequent words for extracting the significant sentence that contain those words. But work of Edmonson adds other important features for the selection of the significant words that contain the significant sentence. In general the features for the selection of the salience sentences are like sentence position in a text, cue phrases, key words, title and heading words.

The position of the sentences in a text, in general, is believed to reflect the importance of a sentence. To this end for example news paper articles have the most important sentence at the

beginning of the article while technical documents have the most important sentences in the conclusion section. In fact a very simple and surprisingly successful method for summarization is the selection of the first sentences in a text. Various researchers have reported that this simple method of taking the lead (first sentence in the first paragraph) as summary often out performs other methods, especially with news paper articles [14].

The cue method is based on the hypothesis that the probable relevance of a sentence is exaggerated by the presence of pragmatic words such as "significant", "impossible" and "hardly". The cue method uses a presorted cue dictionary of selected words of the corpus. Basically cue phrases are words or a phrase that indicates weather a sentence is important or not. According to Edmonson the cue dictionary contains three grouping of cue phrases. Those are bonus, sigma and null phrases. Bonus phrases are used to give emphasis to the importance of a sentence in text while sigma phrases reflect that the sentences is not significant. Null phrases are neutral phrases and are not measured when the weight of the sentences is computed. Some few examples of bonus phrases are 'significantly', 'in conclusion', 'in this paper we show', etc. whereas words such like 'hardly' and 'impossible' are some examples of sigma phrases. Thus each cue phrase is assigned a positive or negative relevance. The weight of each sentence is then the sum of the weights of the words in it.

Sentences can also be achieved for containing words that appear in the text's title or headings, or in the user's query, those are specifically for a query based summaries. The important idea behind this assumption is authors can use an informative title that could make known the subject matter of the document. Also when the author partitions the body of the document in to major sections he summarizes it by choosing appropriate headings. Using the combination of the features cue-words, title words, and the position of the sentence for generating summaries were shown to produce successful results in various studies [14, 24].

Surface level approaches can also be customized to a specific domain or corpus to give corpus based approach for text summarization. Corpus based approach tries to determine the important words by assigning the document to a particular domain. This helps to determine certain common terms in a given field that do not carry salient information and hence their relevance can be reduced. It was further proved that the relevance of a term in the document is an inversely proportional to the number of documents in the corpus containing the term. The normalized

formula for computing the term relevance in a given corpus is given by $(tfi * idfi)$ where tfi is the frequency of the term i in the document and $idfi$ is the inverted frequency of a documents containing this term. Sentences score can then be measured by the sum of term relevance in the sentences [25, 14].

In general term relevance is measured by counting concepts rather than counting the terms only. By making use of an electronic thesaurus or WordNet, each word in a text is associated to a more general concept and the frequency must be computed based on the concept than the frequency of particular words. For instance the occurrence of a concept "bicycle" is counted when any of the words "bicycle", "bike", "pedal", or "brake" is found [15, 16].

Additionally surface level approaches with combination of machine learning algorithms have resulted in more advanced summarization systems. Such kinds of systems use a Bayesian classifier algorithm for computing the probability of sentences to be included in the summary document. The classifier is trained first with a corpus of several documents with their respective summaries. The summaries are abstracts that are created by professionals who are abstractors. The Bayesian formula uses different statistical features to compute the probability of a sentence being relevant. Statistical features that the Bayesian formula uses are usually sentence length, cue phrases, position of the sentences in a paragraph, most frequent words(thematic words) and proper names [26, 24]. In this thesis we use this approach for the term frequency feature.

2.2.2 Entity level approaches

Entity level approaches model text entities and their relationships to capture patterns of connectivity in a text which could be used to determine salient information. In general words can be connected in various ways, including repetition, co-reference, synonymy, and semantic association, as expressed in thesauri. The degree of connectedness words can then be used to score sentences and paragraphs. In this context more connected sentences are more important. With this approach it has been shown that the main drawbacks of surface level approaches, which are lack of coherence and cohesion, can be resolved [25].

Various automatic text summarization methods employ text connectivity to summarize a document and of these, lexical chain which is first introduced by Barzilay and Elahadad [27], is to be mentioned. Summarization based on lexical chains first selects a set of candidate words and

then for each candidate word an appropriate chain relying on a relatedness criterion among the numbers of chains is computed and lastly inserts the word in the chain and updates it accordingly. Cohesive relation (i.e. repetition, synonymy, antonyms, hypernymy, and holonymy), between terms is a criterion for chain information. A lexical chain therefore is a chain of words in a text such that each words in the chain beers some kind of cohesive relationship (hyponymy, meronymy, etc) to a word that is already in the chain [28].

Once the source text is represented using lexical chains, the strength of each lexical chain is determined on the basis of the number and type of relation in the chain. A summary is then built by selecting sentences when the strongest chains are highly concentrated. This is in an assumption that picking sentences represented by strong lexical chains gives a better indication of the central topic of a text than simply picking the most frequent words in the text.

Another way of using text connectivity for summarization is based on phrasal analysis and anaphoric relations in text. Here the main aim is to identify those phrasal units across the entire span of the document that best function as representative highlight of the document's content. One way to achieve this is by using co-reference resolution system. Co-reference resolution is the process of determining if two expressions in natural language refer to the same entity [25]. Once the desired phrasal units are identified, they are combined to form "capsule overviews". The capsule overview is not a sequence of sentences as is expected in a summary but it is a semi formal (normalized) representation of the document derived after the process of data reduction over the original text. Indeed, by adopting better granularity of representation (below that of sentences) consciously trade in "readability" (or narrative coherence) for tracking of detail [29].

2.2.3 Discourse level approaches

Before we are going to see what are the different summarization approaches based on discourse structure let's first define what is discourse itself. Discourse is referring to any form of language based communication involving multiple sentences or utterances such as text and dialogue [30] and the most important forms discourse of interest to computerized natural language processing are text and dialogue. Normally discourse such as written text appears to be a linear sequence of clauses and sentences, it has long been recognized by linguists that these clauses and sentences tend to cluster together in to units, and called discourse segments that are related pragmatically to form hierarchical structure.

Discourse level approaches exploit the discursive organization of a text and its relation to communicative goals for the improvement of the relevance and quality of summaries. The discursive organization of a text implies the global structure of the text and it includes format of a document, threads of topics in the text and rhetorical structure of the text such as argumentation and narrative structure [23].

This approach asserts that discourse analysis goes beyond the levels of syntactic and semantic analysis, which typically treats each sentence as an isolated, independent unit. Thus in this approach a text is first divided into discourse segments and based on these segments, the discourse structure of the text as intended by its author is reconstructed. This discourse structure or discursive representation of the text has been shown to be one way of determining the most important units of a text [30].

There have been various text summarization works that exploit the discursive representation of text to improve the relevance and quality of final summaries. There have been quite some approaches basing summaries on a representation of the discursive aspect of texts. Some of those approaches take advantage of deep understanding of texts; others are based on superficial evidence seems more adequate to address the task of text summarization. The most popular theory of text organization used for summarization has been the rhetorical structure theory (RST) [31]. According to Mann and Thompson [32], one can associate a rhetorical structure tree to any text. RST is a binary tree representing rhetorical relations between text units and it also provides combinations of features that have turned out to be useful in several kinds of discourse studies. The key elements of RST are relations and spans [33]. Relation is the relationship that can hold between two text spans. Text spans on the other hand are any portion of the text that have RST structure and thus have also functional integrity from a text organizational point of view or that is realized by a unit. The relations tie together two non overlapping pieces of text spans: Each field specifies particular judgments that the text analyst must make in building the RST structure. It considers the nature of the text analysis; these are judgments of plausibility rather than certainty [32].

Rhetorical relations reflect semantic, intentional, and textual relations that hold between text spans. These text spans could be clauses or sentences extracted from the original text. Text spans could be related in RST in such a way that one text span may elaborate on another text span or

one text span may provide background information for another text span [33]. In one application of discourse structure for text summarization importance score is associated to each clause in a text; the closer a clause is to the root of the tree, the higher is the score. Then clauses of highest score are extracted to produce a summary [30].

2.3 Evaluation of an automatic text summarization

As the research area were implemented using different approaches there must be an evaluation mechanism how the systems performs in terms of identifying the approach which is performing well and the other which is worse pursuing. Many summarization evaluation methods have been proposed since the up rise of automatic summarization techniques, a recent overview of which is given by Mani [2]. Those methods can be roughly categorized in to two categories. Those are intrinsic and extrinsic evaluations. An intrinsic evaluation tests the summarization system on itself measures how well the summary compares with an ideal summary written by the author of the source text or human abstractor. It usually involves human judges who determine the quality of the summary by directly analyzing it that could be text quality refers to some aspect of the text such as grammatically, non redundancy, reference clarity and coherence.

Whereas extrinsic evaluation is concerned with testing how the summarization effects the completion of some other task (e.g. human assessing a document's relevance). That is the quality of the summary is judged by users accordingly such as how well it helps them determine the source's relevance to topics of interest or how well they can answer certain questions relative to the full source text [34].

However both intrinsic and extrinsic evaluations are not as much satisfactory that have their own problems. In the case of an intrinsic evaluation the main problem is difficulty of constructing unique ideal summary for a given document or set of documents. As because of there are different ways for describing an event, it is expected that users can generate more than one summary to a particular document. That is the reason agreement between human judgments becomes a concern. In addition to that this manual evaluation is too expensive. Extrinsic evaluation on the other hand is time consuming, expensive and requires a considerable amount of careful planning [35,36]. There various approaches to intrinsic and extrinsic evaluation methods on the literature here, we focus on two approaches of an intrinsic evaluation method which were applied in our thesis. Those are co-selection and content based measures or informativeness.

Co-selection measure

Most summarization systems select the most important sentences from the input document to generate the extractive summary. In such cases the quality of the generated summary is usually decided using co-selection measures which finds out how many of the sentences in the automatic summary are included in the ideal or manual summary. The main evaluation metrics of co-selection measure are precision, recall and F-measure which are the commonly used information retrieval metrics. Precision (P) is the ratio of the number of sentences occurring in both system and ideal summaries to the number of sentences in the system summary. Recall (R) is the ratio of the number of sentences occurring in both system and ideal summaries to the number of sentences in the ideal summary. F-measure on the other hand is calculated as follows [15, 35].

$$F = \frac{2 * P * R}{P + R}$$

The main advantage of co-selection measure is that once human judge define the gold-standard summary, it can be repeatedly used to evaluate automatic summaries by a simple comparison. Unfortunately there are also disadvantages in terms of defining gold-standard summary. It has been shown that the difference in recall measure of a summary may range from 25 % up to 50 % depending on which of two available human extracts are used for evaluation. Thus, using co-selection measure creates the possibility that two equally good extracts are judged very differently.

Many of the subsequently developed evaluations measure were designed to address the problems facing with precision and recall. For instance it has been suggested that more emphasis be given to recall than precision. That is because precision might be too strict in that some of sentences chosen by the system might be good though they have not been chosen by the gold-standard. However recall measures the overlap with already observed sentences choices. It has also been suggested to use multiple human judges rather than a single person's judgment [36].

The other evaluation metric that can be used in co-selection measure is relative utility and it was introduced as an improvement to precision and recall. The method involves multiple judges who score each sentence in the input text with confidence values for their inclusion in the summary. The principle suggested that high score sentences have more possibility to be included in the

summary and low score sentences have very low or no possibility to be included. Hence, each possible selection of sentences by a system can be assigned a score showing how good a choice of sentences it represents. Other than requiring a good deal of manual effort in sentence tagging, this approach offers a simple and easy way of evaluating summaries [15, 36].

Content-based measures

Co-selection measure can only determine the quality of a summary based on the number sentences that are common to ideal and automatic summaries. However tow sentences can can contain the same information even if they are written differently. That is the weakness of co-selection measure that could be addressed in content-based similarity measure. The advantage of using content-based similarity measure is that two summaries can be compared at amore fine grained level that just sentences. There are several content-based similarity measures that take in to account different properties of the text such as cosine similarity, word over lap, longest common subsequence and ROUGE (Recall Oriented Understudy for Gisting Evaluation) are some of them [15, 35]. Among those cosine similarity and ROUGE are the most common one that we discuss them in detail now.

ROUGE calculates scores candidate summary based on the n-gram overlap between the candidate and reference summaries. It takes as input pairs of auto-generated summaries and their corresponding reference summaries, and determines their similarity based on different features [15, 36]. The use of a generally on and automatic metric such as ROUGE allows cheap evaluation and ease in comparing results from different research efforts. Thus ROUGE has become a de-facto standard evaluation method in the field of automatic text summarization. For instance, Document Understanding Conference (DUC) which is a series of evaluation workshops held each year to evaluate summarization systems, has been using ROUGE since 2003 [15]. In this thesis we use F-score to evaluate the performance of our summarization system because F-score gives the same importance to precision and recall, thus using it as an evaluation measure is a good trade-off between precision and recall.

2.4 Review Related works on automatic text summarization

There were a lot of researches done on the area of an automatic texts summarization from the starting up to know. On behalf of this research we were reviewed different researches whose

approaches of that summarization were dealing with the extraction of an important sentence of content usually at the sentence level. As of the summarization techniques are systems that take one or more documents as an input and attempt to produce a concise and fluent summary of the most important information in the input.

In this chapter we were critically review previous summarization works based on these sentence extraction techniques. That extractive approach summarization focuses research on the key issues how can the system identify which sentences are important for the construction of the summary.

2.4.1 Global works

The very opening work on the area of an automatic text summarization was done by Lhun[13] that was in the 1950s set the tradition for sentence extraction. His approach was executed on the technical papers that are specifically magazine articles. Luhn put forward a simple that twisted much of later research, namely that some words in a document are descriptive of the whole content and the sentences that convey the most important information in the document are the one that contain such descriptive words close to each other. He also uses frequency of occurrence to identify the descriptive words of the topic of the document. Based on his assumption words that are frequently occurring on the given document are likely to be the main topic of the document. For using the word frequency Lhun identified two requirements for the extraction of the most frequent occurring words. He used thresholds for removal of most frequent words that are stop words and removing words that are common for a particular domain. Sentences described by high compactness of descriptive words, measured as clusters of five consecutive words by Luhn are the most significant words that should be included in the summary.

The other researcher that focused on the sentence extraction were the work of Edmundson [14] that was the foundation of several other trends in summarization research which finally lead to machine learning approach in summarization. He expanded his research based on Luhns [13] approach by suggesting other multiple features for weighting sentences that indicate sentence importance. The proposed features were number of times a word appears in the article, the number of words in the sentences that also appear in the title of the article or in the section headings, position of sentence in the article and in the section and the number of sentences words matching a pre-compiled list of cue words such as “In sum”. He used the corpus both to

determine weights on the four features and to do evaluation. His result fascinatingly suggested that word frequency is the least important from the four classes of features for his specific task and corpus. The other features take the advantage of the knowledge domain and genre of the input to the summarizer.

In other comparatively early and seminal work, Piace shifted the research work towards the need for language generation technique in summarization. The researcher focused on the problem in extractive summarization of unintentionally selecting sentences that contain unanswered references to sentences not included in the summary or not explicitly included in the original document. He says the problem can arise not only the reason of the presence of the a pronouns but it is also due to a wide variety of other phrases such as "our investigations have shown this to be true." And there are three distinct methods to be considered." Piace built an extractive summarizer which uses the presence of phrases from a list that he compiled such as "the main goal of our paper", to determine an initial set of beginning sentences that should be selected. Then an aggregation procedure adds sentences preceding or following the beginnings until all phrases are resolved. Piace also suggested modifying sentences to resolve phrases when the reference can be found but didn't implement an actual system for doing this. His research was the first to point out the problem of unintentionally including phrases in extractive summaries, but the solution of simply adding more sentences until the antecedent is found is not satisfactory and much letter research on using language generation for summarization has revisited the problem.

2.4.2 Automatic text summarization systems for local language

Though as the field of an automatic text summarization has enjoyed a lot of research for many languages. As the other local languages like Amharic and afan oromoo was go forward in some extent but not this in the case Tigrinya language. So in this section the researcher was reviewed the other local languages that the researchers have attempted to develop automatic summarization system. Such efforts are reviewed below giving due attention to the techniques employed.

The first automatic text summarization research for Amharic was conducted by Kamil [37]. He proposed a summarization system for Amharic news items based on extraction approaches. The basic features researcher used for the extraction approaches in order to produce the summary was

title words, head sentences; head sentences words, paragraph starting sentences, cue phrase and high frequency key words appearing on the text. Each feature has assigned weight obtained from training with manual summaries of four news articles, and the weights are combined linearly to produce an overall score for a sentence. He used five news articles with an average length of 17.4 numbers of sentences. The performance evaluation of his approach shows 74.4% and 58% precision and recall respectively at 38.5% extraction rate. For the creation of the summary the most dominant features are title words and key words. Lastly Kemil [37] recommend that development of good stemmer, preparing of standard Amharic corpus, organizing exhaustive list of stop words, and the inclusion of more NLP, statistical and heuristic parameters could improve the system performance.

Kifle [5] conducted a research on Graph-based automatic text summarizer for Amharic through an approach of graph based with combines sentence centrality measures: cumulative sum (MI) and discounted cumulative sum (MII), which are useful in exploiting the relation between sentences in a text, with graph-based algorithms: PageRank and HITS. The researcher used a corpus of 30 news articles ranging from 17-70 numbers of sentences in different domains of economic, politics, society, and sport. The performance Evaluations of the summaries were done using an intrinsic evaluation technique, which involves evaluating linguistic and content quality of system generated summaries, using 10%, 20%, and 30% extraction rates. The results of the evaluations showed that the proposed system registered its best linguistic and content quality of summaries of 83.23% and 75.02% respectively when MII is paired with HITS at 30% extraction rate. The researcher used different domains in his summarization and the overall results of the evaluations showed that linguistic and content qualities are not dependent on domain.

Addis [6] customized and tested an open source tool, OTS, for its performance in summarizing Amharic news texts. In his study the researcher followed pure statistical approach. To determine the importance of the sentence he used frequency of terms. Addis conducted two experiments, which he called them E1 and E2. In the first Experiment he adapted the Porter stemmer to fit his purpose. In E2, on the other hand he adopted a stemmer from the work of [Tesema, 2007]. For testing the OTS's performance, he used 30 news texts which were collected from different sources along with 90 ideal summaries prepared by two independent human evaluators three for

each news text with extraction rates of 10%, 20%, and 30%. F-measure score for the first experiment is 75.65% at the 30% extraction rate for middle size articles and a corpus average score of 66.23% has been achieved whereas for experiment two it is 72.83% at the extraction rate of 30% for the large size news articles and a corpus average score of 72.37%. The performance evaluation of his work showed E2 out performed E1. This is mainly due to the Amharic stemmer being used that improves the frequency of terms. However, in terms of efficiency E1, which used Porter stemmer, outperformed E2.

Melese [7] proposed two approaches for Amharic text summarization. He followed an algebraic approach, specifically LSA. The two approaches proposed by him are TopicLSA and LSAGraph. The LSAGrap is a mixture of graph-based and LSAbased approaches. To test his system, the researcher used 50 Amharic news texts with their corresponding ideal summaries for extraction rates of 20% and 30%, which were prepared by six human evaluators. Performance evaluation showed LSAGraph + PageRank outperformed LSAGraph + HITS for 20% extraction rate, while LSAGraph + HITS outperformed LSAGraph + PageRank for 30% extraction rate. Furthermore, the researcher claimed that his approaches give better results than previously done works.

Helen [8] Automatic Text Summarization for Amharic Legal documents used in Judgments. The researcher first manually segmented a given legal judgment in to five pre-defined themes: introduction, reason, fact, judicial analysis, and decision.. The statistical extraction techniques were carrying out for the research. Weight is assigned to each sentence based on its location and the cue words/phrases that it contains to extract the highest weighted sentences. To see the performance of the system Precision and recall measure is used for 20% and 10% compression rate. The system summary is compared against the human (ideal) summary. As a result, precision of the system summary is 33.9% and 39%; Precision of the random summary is 23% and 27%; recall of system summary is 57% and 50.5 %; recall of random summary is 46% is 38% for 20% and 10 % compression rate respectively.

Teferi [9] The application of Machine Learning Technique for Automatic Text Summarization-The Case of Amharic News Texts. This study, however, employed machine learning technique (naïve Bayes). In this study, title, location, cue words and content words features are examined. In his experimentation, the researcher used about 480 news articles. Of

which 20 were used as a test set while the remaining articles used to train the system. The results of the analysis shows that precision of 75.00%, recall 74.90 % and classification accuracy of 86.03% in predicting the summary sentences. The researcher recommends availability of standard Amharic corpus, analysis 26 of each single feature like cue words didn't help in the prediction of sentences for the summary and availability of standard stop-list.

Asefa [10] conductd a research that is query based automatic summarizer for afaan oromo text. In doing so the researcher used two methods that are the information retrieval model which is vector space model (VSM) and sentence position method. The data corpus used for the research was 40 afaan oromo news articles ranging from 20 to 66 numbers of sentences and used 10%, 20% and 30% extraction rates for the evaluation. He also uses two experiments that are EXP1 (the significance score of sentence using cosine similarity to the title vector) and EXP2 (the significance score of sentence using cosine similarity to the title vector and position method) and compared those two against the reference summary.

The researcher evaluated the performance of the system using standard Information Retrieval (IR) evaluation metrics (Precision, Recall and F-measure) and the subjective evaluation evaluate the linguistic quality such as informativeness and coherence using the scores on five scale measures by human evaluators. The results of the evaluations showed that the proposed system registered f-measure of 82%, 78% and 82% at summary extraction rate of 10%, 20%, and 30% respectively when VSM is used along with position method. Moreover, the informativeness and coherence of the proposed system also registered its best performance summary of 59%, 77% and 91% average score on five scale measures at extraction rate of 10%, 20%, and 30% respectively when both methods used together.

Girma [11] on Afan Oromo news text summarizer based upon the Open Text Summarizer (OTS). His work was done on customizing the OTS code. The summarizer basically uses the combinations of term frequency and sentence position method with language specific lexicons in order to identify the most important sentence for extractive summary. The researcher used a corpus of 8 news articles averagely 11 sentences and 277 words. The performance of system was evaluated using three methods. M1 (term frequency and position method without stemmer & lexicon, M2 (with stemmer & lexicon) and M3 improved all term frequency, position method, stemmer and lexicon. The performance of M1, M2 and M3 registered f-measure values of

34%, 47% and 81% respectively i.e. M3 outperformed the two summarizers (M1 and M2) by 47% and 34 % . The subjective evaluation result shows that the three summarizers' (M1, M2 and M3) performances regarding informativeness, linguistic quality and coherence structure are: (34.37 %, 37%, and 62.5%), (59.37%, 60% and 65%) and (21.87%, 28.12% and 75%) respectively as it is compared the developed system with human evaluators.

CHAPTER THREE

3. The Tigrinya Language

3.1. Introduction

Tigrinya is a member of the Ethio-Semitic languages, which belong to Afro-Asiatic super family [38]. Tigrinya is spoken primarily in Eritrea and Ethiopia. There are more than 6 million Tigrinya speakers worldwide. According to the 2007 population and housing census Ethiopia, there are over 4.3 million Tigrinya speakers in Tigray [39] and according to Ethnologue there are 2.4 million Tigrinya speakers in Eritrea [40]. Tigrinya is written in the Geez script which these days is called Ethiopic and originally developed for Geez language. In Tigrinya each symbol represents a consonant and vowel combination and the symbols are organized in groups of similar symbols on the basis of both the consonant and the vowel. For each consonant in each symbol, there is an unmarked symbol representation of that consonant followed by a canonical or inherent vowel [41].

Tigrinya like other Semitic languages such as Arabic and Amharic exhibits a root pattern morphological phenomenon. In addition, it uses different affixes to create inflectional and derivational word forms.

3.2. Tigrinya writing system

The Tigrinya writing system is one variant of what is often referred to as the “Ethiopic” writing system or “Ethiopic syllabary”. It is slight variant of writing system used for Amharic and for Ge’ez, The classical language still in use as the liturgical language of Ethiopian and Eritrean orthodox Christians. The most salient graphical units in this writing system represent as a consonant followed by vowel. Characters representing the same consonant followed by different vowels are similar in shape. For example here are the character representing

□	□	□	□	□	□	□
he	Hu	hi	ha	hie	h	ho
□	□	□	□	□	□	□
me	Mu	mi	ma	mie	m	mo

Table 3.1 character representation

As a result the writing system is usually displayed as a two dimensional matrix in which the rows contain units beginning with the same consonant and the columns contain units encoding in the same vowel. The columns are traditionally known as “orders”. The first order in Tigrinya represents the vowel /e/, the second the vowel /u/, the third the vowel /i/, the fourth the vowel /a/, the fifth the vowel (really diphthong) /ie/, and the seventh the vowel /o/, the sixth order represents the consonant alone or followed by the vowel.

3.2.1 Punctuation Marks in Tigrinya

Punctuation marks are symbols that are placed in different position of the txt for making clear understanding and easy reading [11]. In Tigrinya there are different punctuation marks. Some of them are (□□) □□□□□□□□/arbaet netbi/ period, full stop which is used to explicitly express sentence boundary; (□) □□□□□□/ ntsela serez/ comma that is used to separate the elements in a series (three or more things), including the last two; and (□) □□□□□□/drb srez/semi colon which allows the writer to imply a relationship between nicely balanced ideas are the most commonly used in both hand and type written Tigrinya texts. In addition to these, Tigrinya has some borrowed punctuation marks from Englishlanguage like (?) □□□□□□/ hito milkt/ question mark which is used in interrogative or at the end of the direct question, and (!) □□□□□□/ kal aganino/ exclamation mark, which is used at the end of an imperative sentences from foreign languages.

Punctuation marks	Meaning
□□	End of sentences
□	Word separator
□	Sentence connector
□	List separator marks
:-	Beginning of the list mark
?	End of question
!	End of n emphatic declaration, or command
“	Quote some words or sentences taken from other

Table 3.1 list of punctuation marks in Tigrinya language

1.3.Word classes

Word of Tigrinya in general can be classified in to two categories: which are open and closed. When we say open class because of new members are added are always added to the former and are unlimited in number; where as members of the closed classes are relatively fixed and few in number [41].

As far as the number and types of classes are worried, languages differ from one another in closed than in open classes. Some language may have a dozen or more closed classes where as others may have tremendously few in number [41].

The first effort on Tigrinya word classification was PHD work by Tesfaye in 2002 [36]. The researcher classified Tigrinya words as open and closed categories and in to eight classes in particular. The classes of noun, verb and adjective are grouped under open classes where as the other five namely pronouns, determiners, adverbs, prepositions and conjunctions are in the group of closed classes. Interjections are words without syntactic functions and in this category they are not considered as word classes.

In the second effort on Tigrinya word classification was the work done by Daniel [41]. Accordingly he reduced the categorization in to five word classes. Those are preposition, noun, verb, adjective and adverb. Pronouns and conjunctions are place under noun and preposition categories correspondingly. In this categorization interjections are also not considered as word classes as of Tesfaye's classification. A brief description of each of the classes in Tesfaye's and Daniel's classification is explained as the following.

Nouns

Noun is one from the word classes that can be given for humans, animals, events, places, situations, visualizations and phenomena to be described. Noun also contains three things in it. Those are pronoun, common noun and concrete noun. Tigrinya nouns are either primary or derived. If they are primary nouns are not formed from verbs or other nouns. In other way if they are related to the root consonants (radicals) and meaning to verbs, adjectives, other nouns, or

other part of speech through intercalation and addition of suffix like "-□□/-eya", "-□□□/-nhna" and prefixes "-□□/-sle", "-□□□/-ende" are derived.

In Tigrinya nouns often inflect for gender (feminine/masculine), number (singular/plural) and species (definite/indefinite). Feminine noun is mostly formed using a marker

Verbs

Verbs are words that indicate an action (e.g. □□□/keydu/'go', □□□/beliu/'eat', etc.) or state of being or subsistence (e.g. □□□□/tehagusu/'become happy'/, □□□/regudu/'become fat', etc.) in a complete grammatical declarative sentence and often come at the end. Like nouns Tigrinya verbs can be either simple or derived. A could have various forms depending on the tense aspect mood, and root structure while inflecting for person (first, second, third), gender (masculine/feminine), and number (singular/plural). For instance the root form of the verb □□□/ble/'-eat' inflected for person, number and gender as shown in the following table:

	Persons					
	First person		Second person		Third person	
	Singular	Plural	Singular	Plural	Singular	Plural
Masculine	□□□ Belia I eat	□□□□ Beliena We eat	□□□□ Belieka You eat	□□□□□ Beliekum You eat	□□□ Beliu He eat	□□□□ Beliom They eat
Feminine			□□□□ Beliki You eat	□□□□□ Belikin You eat	□□□ Belia She eat	□□□□ Belien They eat

Table 3.2: verb inflection for person, number and gender

Adjectives

Adjectives are words that usually came before nouns to serve as modifiers to the nouns they precede. In addition to that Tigrinya adjectives can indicate amount, structure, color and situation.

For example adjectives that indicate structure are like □□□/kebib/'circle', □□□□/molmal/'oval, etc), adjectives that indicate color are also like □□□/tseda/'white', □□□/tselim/'black' and □□□/Qyh/'red, etc)and adjectives that indicate situation are □□□/rsn/'hot', □□□/dkm/'weak and □□□/hyl/'powerfull, etc).

Adverbs

Adverbs are words that generally come before verbs and express different adverbial functions such as time, place and manner or circumstance, degree, measure etc. modifiers of adjectives and adverbs commonly express degree while adverbs functioning as a sentence modifiers usually the speakers attitude regarding the event spoken.

E.g 1. □□□ □□□□ □□□□□ /tmali qelTifu metSiu. (yesterday he came quickly)
□□□ = Adverb of time □□□□ = adverb of manner □□□ = verb
E.g 2. □□□□ □□ □□□□□/ntSegam getS kydu. (he go to left) □□□□ = adverb of place □□ = Noun □□□ = verb
E.g 3. □□□ □□□ □□□□ □□□□□/tmali azyu qelTifu metSiu. (Yesterday he came very quickly.)
□□□ = adverb of modifying the adverb □□□□/ qelTifu (quickly)

E.g 4. □□□□□ □□□ □□□□□ □□□□ □□□ □□□□□□/bagaTami abta mekina znberu sebat moytom. (Fortunately, the people in the car were died)
□□□□□ = a senctence modifier

Table 3.3 different adverbs of Tigrinya

Conjunctions

Conjunctions are the linkage between words, phrases and clauses to create larger grammatical units. Conjunctions were treated as a separate category [41]. However, conjunctions are also treated as a sub class of prepositions [41]. Conjunctions can also be classified as two categories which are coordinating or subordinating. Coordinating conjunctions can make coordination between words, phrases, clauses and sentences.

E.g. □□- □□□- □/gal-n wd-n (girl and boy)

Most of the items that function as prepositions in Tigrinya can also function as conjunctions. Moreover the coordinating conjunctions such as - □/n (and), □□/wey (or) can form structural units with the nominal and prepositional phrases they precede or follow.

Preposition

Prepositions are small set of words, that they have meaning only if they are attached or used together with other words such as nouns, verbs, pronouns and adjectives. They can express relationship between person, thing, or event etc. Tigrinya prepositions have the following central properties.

- ✓ They take nouns and or adjectives as their complements: e.g. □-□□□□□ /b-handebet(suddenly)
- ✓ They can stand alone as a separate word: e.g. □□□□□□□□□□ /nab bEt me^ShafTi (to library)

E.g. □□ □□□□ □□ /^ab ^afdge geza (outside the house)
□□ = preposition

□□□□ = noun □□ = noun

Tale 3.4 preposition as a separate word [41]

In general they are not inflected for gender, person and number etc.

Pronoun

According to Daniel [41] pronoun is a word which is regarded as a sub class of noun. It could be taken as noun because it can function as a noun and can take the position of a noun. However pronouns are closed classes for two main reasons: first they are few in number and second the number of their numbers doesn't increase [41]. Examples of Tigrinya pronouns are: □□/nsu (He), □□/^ane (I), □□□ /nsKa (you (masculine)), □□□/nsKi (you (feminine)), □□/nsa (she) etc.

3.4. Morphology of Tigrinya language

Introduction

Morphology is the branch of linguistics that deals with internal structure of words and word formation, including affixation behavior, roots, and pattern properties [42]. Morphology is the main source of difference in natural language text, with suffixing and prefixing being the most common ways creating a word variant. Generally morphology can be classified as either inflectional or derivational. Inflection is variation or change of form that words undergo to mark distinctions of case, gender, number, tense, person, mood, voice, comparison. Inflectional morphology is applied to a given stem with predictable formation. It doesn't affect the word's grammatical category, such as noun, verb, etc. case, gender, number, tense; person, mood, and voice are some examples of characteristics that might be affected by inflection. Derivational morphology, on the other hand, concatenates to a given word a set of morphemes that may affect the grammatical and syntactic category of the word.

A word can have several word forms, e.g. the word "write" can take the forms "writes", "wrote" and "written", usually called inflected forms. The root is the original form of the word before any transformation process, and it plays an important role in language studies. The root is the form of

a word from which the other forms can be derived using the morphological rules of a language. A morpheme is the smallest unit of a language that has a meaning and can't be broken down further into meaningful or recognizable parts and should impart a function or a meaning to the word which they are part of. An affix is a morpheme that can be added before (prefix) or after (suffix), or inserted inside (infix) a root or a stem to form new words or meanings [43]. Morphological information of a language is useful for several natural language applications such as stemming, morphological analysis, text generation, machine translation, document retrieval, etc.

3.4.1. Tigrinya morphological system and word formation

85% of the words in Tigrinya are created from a root of three radicals (trilateral words) and to a lesser extent there are also quad literal, pen-literal, or hexa-literal words [44]. Each word group generates an increased verb forms and noun forms by the addition of derivational and inflectional affixes. Words in Tigrinya are built from the roots by means of a variety of morphological operations such as compounding, affixation, and reduplication [45].

An affix in Tigrinya is a morpheme that can be added before or after, or inserted inside a root or a stem as a prefix, suffix or infix, respectively to form new words or meanings. Tigrinya affixes have the feature of concatenating with each other in predefined linguistic rules. This feature increases the overall numbers of affixes [44]. There are also some prefixes and suffixes which determine whether a word is a subject marker, pronoun, preposition, or definite article. Tigrinya is highly productive, both derivationally and inflectionally. Definite articles, conjunctions, particles, and other prefixes can attach to the beginning of a word, and large numbers of suffixes can attach to the end. A given headword can be found in huge number of different forms.

Tigrinya concatenative morphology regulates how a stem and affixes glue together, while non-concatenative one combines morphemes in more complex ways. Affixes in Tigrinya can be classified as four categories [44]. Prefixes precede the base form, such as 'Intey-,'ay-,kemz-, ktete-, sle-, zte-. Suffixes follow the base form, i.e. –kum, -tat, -tatat, -net, -awi, and infixes are inside the base form. Circumfixes are affixes attached before and after the base form at the same time. While circufixes formally are combination of allowed prefixes and suffixes, they have to be treated as discontinuous units form for semantic and grammatical reasons. Tigrinya non-concatenative morphology refers to reduplicated morpheme forms. Reduplicated words based on

morpheme regularity are grouped into full reduplication (the word □□□□□ derived from the stem □□□) and practical reduplication of different kinds. The letter includes reduplicated stems with affixes (e.g. word □□□□ is derived from stem □□□ sebere, □□□□□ 'teregagemu' is derived from stem □□□ 'rgeme', the word □□□□□□ 'Gel Tem Tem' is derived from the word □□□□ 'Gel Tem') and there also various irregular reduplications.

3.4.2 Derivational and inflectional morphology

There are five parts of speech in Tigrinya that are adjectives, nouns, verbs, adverbs, and prepositions [41]. Prepositions and conjunctions are totally unproductive. Adverbs are few in number and are less productive. Therefore, the discussion of derivational and inflectional morphology concentrates on the remaining three parts of speech, namely verbs, nouns and adjectives.

1. Inflectional morphology of Tigrinya

As Tigrinya language is highly inflectional language definite articles, conjunctions, particles and other prefixes can attach to the beginning of a word, and large numbers of suffixes can attach to the end. A given root of word can be found in huge number of different forms.

1.1.Inflection of verbs

This section presents the inflection of verbs. It is compiled for the purpose of the study from Tigrinya grammar books by [41] and [44].

A significantly large part of the vocabulary consists of verbs, which exhibit different morph syntactic properties based on the arrangement of the consonant-vowel patterns. For example, the root sbr, meaning 'to break' can have the perfect form sebere with the pattern CVCVCV, imperfect form tsebr with the pattern CCVCC, gerund form sebirka with the pattern CVCVCCV, imperative form sber with the pattern CCVC, causative form 'asbere with the pattern as-CVCV, passive form tesebere with the pattern te-CVCVCV, etc. subject, gender, number, etc are also indicated as bound morphemes on the verb, as well as objects and possessions markers, mood and tense, transitive, dative, negative, etc, producing complex verb morphology.

The simplest form of the verb is the third person masculine singular of the perfect tense. In most Tigrinya dictionaries, all the words derived from a trilateral root are entered under the third

person masculine singular form of the verb. Each three-consonant (or “trilateral”) root belongs to one of three conjugation classes, conventionally known as A, B, and C. this division is a basic feature of Ethiopian Semitic Languages.

Most three-consonant roots are in the A class. In the citation form (perfect), these have no germination but the vowel ‘e’ appears between both pairs of consonants. Examples are: □□□ derefe “he sung”, □□□ deyebe “he climbed”, □□□ seteye, “he drunk”. The B class is distinguished by the germination of the second consonant in all forms. Some examples are: □□□ deqqese ‘sleep’ □□□ wesseKe ‘add’. The relatively few members of the C class take the vowel a between the first and second consonants. Examples are □□□ bareKe ‘bless’ and □□□ nafeke ‘long for, miss’.

1.2. Derivation of Verbs

Unlike the other word catagories such as nouns and adjectives, the derivation of verbs is not common form the other parts of speech. Almost all Tigrinya verbs are derived from root consonants [44]. Traditionally a distincition is made between simple and and derived forms.

Simple verbs are verbs derived from roots by intercalating vowel patterns wheras derived verbs are considered as derivatives of simple verbs. The derivation process can be an internal one in which consonant-vowel patterns are changd, and external one where derviational affixes are attached to the simple derived verb or a combination of the internal and external derviational process. The derivations of causative, passive, repetitive and reciprocal verbs are presented below.

1. Causative: causative verbs are derived by adding the derivational morphemes ‘a- and to the verb stem. For example □□□ /betsHe/ ‘arrive’, □□□□ /abtsHe/, ‘cause to arrive’. In most the ‘a- morpheme is used to form causative of intransitive verbs, transitive ones and verbs of state.
2. Passive/Reflexive: the passive verbs are derived using the derivational morpheme □ /te/. This derivational morpheme is realized as □ /te/ befor consonants and as □-/t-/ before vawoles. More over in the imperfect jussive and in derived nominals like verbal noun, the derivational morpheme □ /t/ is used. In this case it assimilates to the first consonant of the verb and as a result the first radical of the verb gemnates. Some exceptions are intransitive

verbs like □□□ /feliHu/ ‘it boiled’ that form their passive forms using the perifix □ -/te-/ as in □□□□ /tefeliHu/ ‘it was bolied’. Such kind of verbs can derive their passive from their causative form (□□□□ /afliHu/ ‘he boiled’).

3. Reduplicative/repetitive: Reduplicative stems indicate an action which is performed repeatedly. For tri-radical verbs, such stems are formed by duplicating the second consonant of the root and using the □-/a-/ after the duplicated consonant as in □□□□ /se-ba-bere/ ‘he broke repeatidely’ derived from the root □□□/sbr/ break. All verb types, Type A, B and C have the same reduplicative forms.
4. Reciprocal: Reciprocal verbs are derived by prefixing the derivational morphemes □ -/te-/ either to the derived type C forms (that use the vowel a after the first radical) or to the reduplicative stem. For example, reciprocal forms of □□□□ /teqatelu/ ‘killed each other’ and □□□□□ /teqetatelu/ ‘killed one another’ are derived from the derived type C stem qetelu- and reduplicative stem qetatelu-, respectively. The causative reciprocal varbs are fromed by adding the causative prefix ‘a- to the reciprocal verb forms. How ever, the reciprocal verb prefix t- or ‘a- assimilates to the stem-initial consonant (thus causes the first radical of the stem to geminate) and doesn’t show up in the surface from of the reciprocal causative.

3.5 Styles of news writing

News is the communication of different events through different communication mechanisms and it is shared in various ways. It could be among individuals and small group weather it could by word of mouth or newsletters; with wider audiences such as publishing, either it is in print or online, broadcasting, it could on television or radio, or it could be in social media sharing among individuals but goes viral. News writing structure or style is the way in which elements of the news are presented based on relative importance, tone and intended audience [46]. Over a century the inverted pyramid style has been the dominant style of writing. But recently there is other style which is called diamond shape writing style, which shows the main ideas at the beginning and at the end. This style of news writing is more suitable for writing news briefs and news events in any language. And all news articles used in this thesis as a data set sre using this news style.

CHAPTER FOUR

4. Implementation

4.1 System architecture

The proposed system is purely extractive type of summary; means that is the process of selecting an important sentence based on the frequency of individual words and title words. Based on this the sentences that contain the frequent words are ranked using the intersection of the frequent words. The primary goal of our system is selecting the most frequent words and which sentence should be included in the summary. Figure 4.1 shows the architecture of our system that contains three modules. That are (1) preprocessing: this module consists of four components which are tokenization, stop word removal, stemming and normalization and their function is efficiently represent the input text in to a suitable format for further text summarization process and maintain the quality of the summary. (2) Scoring: this module was doing the individual score of the term and the sentence that contains the term itself. It gives a score for terms that appear in different sentences through out the whole document. (3) Ranking: in this module sentences are ranked based on the intersection of the most frequent words. Sentence that contains the most frequent words should have ranked first. (4) Summary generation: this module is responsible for selecting best candidate sentences to form the summary.

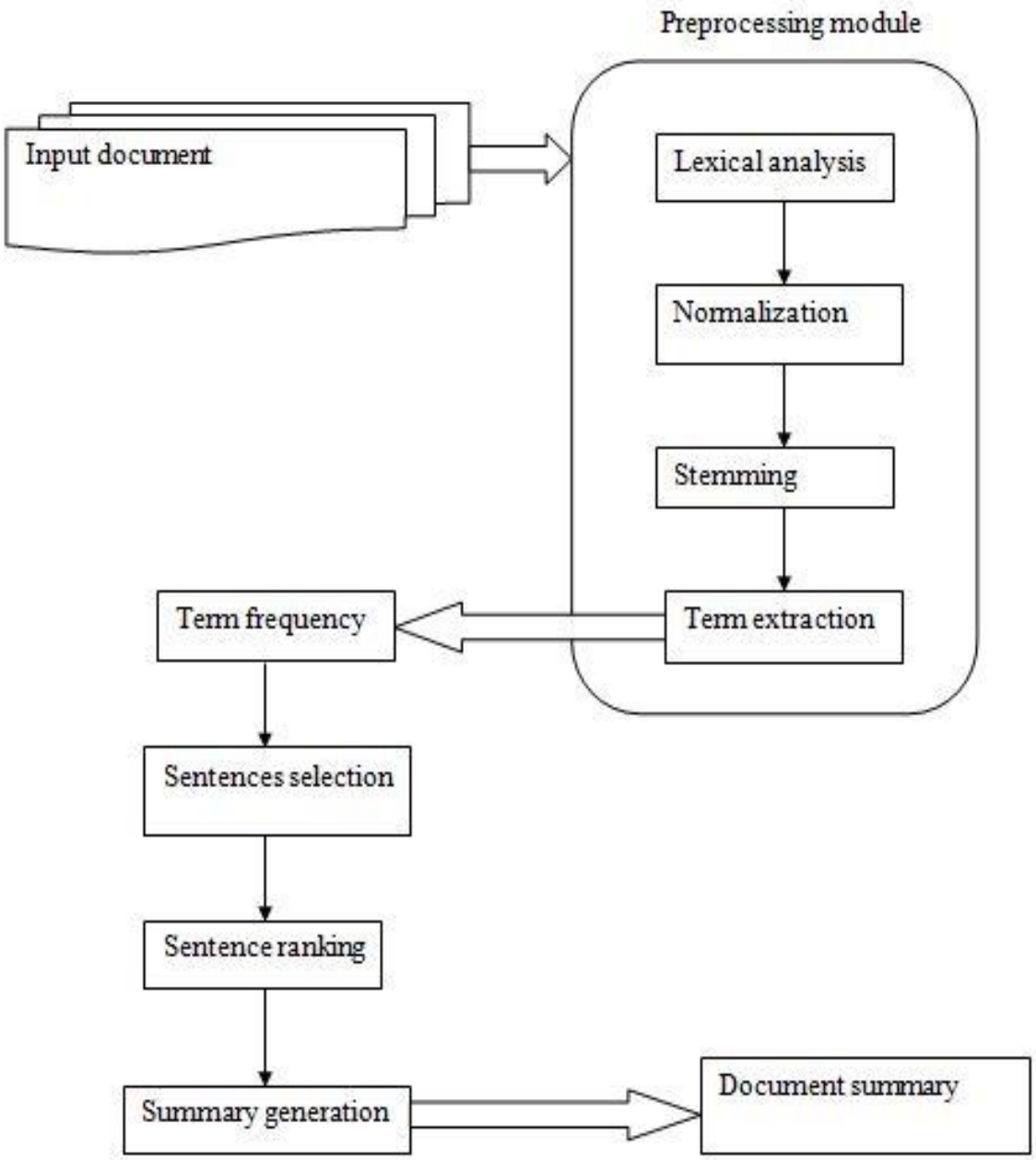

Figure 4.1.The general architecture of automatic Tigrinya text summarizer

4.2.1 Text preprocessing

The preprocessing step is perhaps the most important in the area of natural language processing because of the quality of the obtained summary depends on how efficient is the representation of the text that makes the summary. In order to achieve this preprocessing stage that carries out the different steps of lexical analysis, normalization, stop word removal, stemming and term extraction.

1. Lexical analysis

The first step in preprocessing of the input document is lexical analysis which is also known as tokenization. As the generation of the summary depends on the computed score of each sentence from the given document and these scores also depends on the individual words that contains the sentence. Hence lexical analysis in automatic text summarization involves text splitting into words and sentences. In this thesis the extraction of the sentences out of the text is based on the sentences delimiter, □□□□□ □□□ (□□). Following the sentences splitting, the individual words are extracted from the sentences by scanning each sentence for predefined word delimiter such as new line and space etc.

In the tokenization process punctuation marks and digits are irrelevant component of the text documents. In this situation the punctuation marks and digits are separated from the sentence and pass it.

The words are considered as relevant for the given documents and they are separated by space. Algorithm 4.1 illustrates the tokenization process of the given document and removal of the punctuation marks as well as numbers. First the content of the file is read line by line. Second, split them by space in to list of words. Third check whether the word within the list contains punctuation marks of Tigrinya language; if punctuation marks exist within the word replace it with space. This step continues until end of line reaches.

Algorithm 4.1 word tokenization and punctuation separation

Input: document

Output: word tokens

```
Open the file for processing
Do
        Read contents of the file at sentence level
        Assign the content to string
        Tigrinya punctuation list = []
        For word in string split by space
                If word not in Tigrinya punctuation list
                        Append the word
                Otherwise word has digit
                        Pass
                End for
While end file
```

Algorithm 4.1 word tokenization and removal of punctuation marks

2. Normalization

As explained in chapter 2 there are quite a few of Tigrinya characters that have similar functionality but different symbols. Such characters are normalized by the most widely used character for common understanding of the system. For instance the different form of the word "Nigus" which are □□□ and □□□ are all converted to the common form □□□ by changing the last character of the word.

Furthermore the normalization stage includes the expansion of words that are written in a short form using "/" or "." into one word or two words. For example □/□□□□□ is expanded to □□□□ □□□□□ and □/□ is expanded to □□□□. For achieving this each of the words obtained after tokenization is checked for its presence in a list of common short words and if a word is found on that it is expanded to its common form.

Algorithm 4.2 normalization

Input: tokenized text, normslization list, expansion word list

Output: normalized text

```
Open the file for processing
Do
    Read the document at token level
    Assign the content to string
    For word in string
        If str[i] in expansion word list
            Expand the word and do for the same
    For character in word
        If character in normalization list
            Replace character in to their common standard form
and do for the same
        Else
            Continue
        End if
    End for
While end file
```

Algorithm 4.2 the algorithm first reads the input text at token level and converts each token in to array of character. Then until it reaches the size of the array the algorithm checks if each element of the array is equal to any of the character variants. If that is the character, it will be replaced by the corresponding order. Finally it returns the normalized token.

3. Stop word removal

Some words that are called stop words are words that don't contribute significantly to the overall idea of the document. Such words are like conjunctions, articles, pronouns or are words that appear in many sentences that don't have an influence to distinguish one sentence from the other sentences. Such words could be identified and removed by using a predefined list of stop words. Here are some examples of stopwords:

Stop word	Meaning
□□	To
□□	At
□□	I
□□□	Are
□□	The
□□□	You

Table 4.1 sample stop word list of Tigrinya document [20]

Algorithm 4.3 Removing non content bearing words from a text

Input: tokenized text; a list which contains stop words

Output: tokenized; stop word free text

```
Read stop word list
Open the file for processing
Do
    Read the document at token level
    Assign the content to string
    For word in string
        If word is stop word list
            Remove word from the index term
        Else if word not in stopword list
            Append them for further assesement
        End if
    End for
While end file
```

Algorithm 4.2 first reads a tokenized text and list of stop words from a document and checks weather the input document has the list of stop words if has remove otherwise append it the token.

4. Stemming

After the stop word removal is completed the next step under the preprocessing of the input document is stemming. As explained in chapter 2 the stemming process is a shallow stemming which is simply removal of prefixes and suffixes from the input document. To this end there is prepared list of suffixes and prefixes for Tigrinya texts.

Algorithm 4.4 Removal of suffix and prefix from the input document

Input: tokenized stop word free text, prefix list, suffix list

Output: stemmed text

```
Read prefix and suffix list file
Open the file for processing
Do
    Read the document at token level
    Assign the content to string
    For word in string
        If length of word is greater than two
            If word starts with prefix
                Remove prefix from the word
            Else if word ends with suffix
                Remove suffix from the word
                Continue
        Eend if
    End for
While end file
```

Algorithm 4.3 shows the algorithmic description of stemming that accepts the tokenized free stop word texts and checking whether they have suffixes and prefixes or not and if they have exclude them otherwise append it the token.

As shown in algorithm 4.4, the affix removal algorithm checks the existence of the prefix and suffix and then removes them from the word. The prefix striping algorithm removes the prefix which is placed before the root word. For example "□□□□□□" (slez-gebere) contains "□□□" (slez-) prefix. As a result it is stemmed to "□□□" (gebere). Similarly the suffix striping algorithm removes the suffix which is written at the end of the root word. For instance "□□□□" (geza-wti) has "□□" (-wti) suffix at the end. So it is stemmed in to "□□" (geza). In Tigrinya language prefix and suffix also exists in a single word. For example the word "□□□□□□" (mewerwerya) has a prefix "□" (me-) and suffix "□" (-ya). First the prefix "□" (me-) is removed using algorithm 4.3. As a result the word "□□□□□□" (mewerwerya) is stemmed in to "□□□□□" (werwerya) but it has a suffix "□" (-ya). Then the suffix "□" (-ya) is removed from "□□□□□" (werwerya) and it is stemmed in to "□□□□" (werwer) using algorithm 4.4. As a result the stemming helps to define words in the same context with the same term and consequently reduce their dimensionality.

4.2.2 Term Frequency

As a result of completing the preprocessing stages of lexical analysis, normalization, stop word removal and stemming we now have an index of terms that are capable of representing the content of the document. Text representation techniques based on the extraction of terms of a text or documents which consist in choosing terms that are frequently occurring and then selecting sentences contains these terms to make the summary. As the thesis was carried out by term frequency and the individual words are counted throughout the document and give score to them. After counting of the word is completed identification of the most frequent words have been done.

Algorithm 4.3 Term frequency counting algorithm

Input: stemmed text;

Output: Most frequent words

```
Do
    For word in stemmed string
        For word not in punctuation_list or word is different from
numerical values
            If word is not in dictionary keys
                Dictionary Keys [word] =1
                Continue
            Else if word in Dictionary keys
                Dictionary keys [word]+=1
            Return the most three frequent word from the dictionary
            Goto final document
            End if
    End do
```

4.2.3. Sentence Ranking

The sentence ranking is done through sentences which contain the most frequent words of the input document. The ranking have been done using these most common frequent words and the intersection between them in finding the sentences that contains these frequent words. For example if we have top five most common words then the ranking done through the top five, top four, top three and top two. So in ranking the important sentence that makes the summary intersection of the most common words is checked to each sentence and make them to include in the summary.

Algorithm 4.5 Sentence ranking

Input: frequency based selected sentences

Output: Ranked sentence based on relevance

Do

If sentences that contain the three most frequent words found

Return the sentence and rank it as very relevant sentences
continue

Else if no sentence was found try with the two most frequent words

Return sentence and rank it as very relevant sentences next to the above found sentence

Continue

Else if no sentence was found try with the one most frequent word

Return sentence and rank it as very relevant sentences next to the above sentence

Continue

End if

End for

4.2.4. Sentence generation

A summary is produced based on the above algorithm as the sentences are ranked based on the most frequent words which are intersection of the senentences. The summarizer extracts sentences that contain the most frequent words in the text document.

4.2 Evaluation criteria

To evaluate the quality of the system extracted summaries adjacent to the human or manual extracted summaries, the precision, recall and F-score were calculated for the system summaries. Calculating the precision and recall is used to measure the relevance of a set system summary with reference summaries. That shows how the developed system extracts the most important sentences to create the summary.

$$\text{Precision} = \frac{\#of\ sentence\ in\ the\ automatic\ extract\ and\ also\ in\ the\ human\ extract\ sentence}{total\ \#of\ sentence\ in\ the\ automatic\ extract\ sentence}$$

$$\text{Recall} = \frac{\#of\ sentence\ in\ the\ automatic\ extract\ and\ aslo\ in\ the\ human\ extract\ sentence}{total\ \#of\ sentences\ in\ the\ human\ extract\ sentence}$$

$$\text{F-Score} = \frac{2\ Precision\ X\ Recall}{Precision + Recall}$$

4.3 The features used for Tigrinya language text summarization

4.3.1 Summarization features

This research attempted to design a model using two different thematic features for assigning the weights of important sentences to be included in the summary. This research is carried out for the different domains for data sets like the news items of sport, social, politics to be included. It was also assumed that the document has only textual documents to be summarized. Generally the research was conducted based on two thematic features.

1. **Identifying term frequency**

Keywords of the document are primarily identified based on the term frequency. The main assumption of this program is called “Thematic Term Assumption” that is relatively more frequent terms are more salient [2]. A weight is assigned to each sentence according to the term frequencies in the text. As the document is preprocessed stop words and other common words are filtered by Tigrinya stop words list. The sentences which have the highest weight value are extracted to produce the summary.

2. **Title words**

The title words features it is assumed that authors always used contents related to the title for filtering the article. Therefore the title could be considered as the essential part of the document. Edmundson [14] has defined title words as a feature and that is used to assign a weight to the sentences based on the terms in it that are also present in the title.

Edmundson [14] has used the title, subtitle and heading to identify the title words and has manually assigned weight as it leads to the best performance. The selected corpus doesn’t have any subtitle and heading by each document consists of an appropriate main title.

$$W(s) = \frac{\text{No of title words in the sentence s}}{\text{total number of words in the sentences s}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \; 3.2$$

Where, W(s) is the weight assigned for the sentence s based on title word.

Equation 3.2 which is defined to assign weight for the sentence s due to title words, always gives a value.

CHAPTER FIVE

5. Experiments, Results and Analysis

In this chapter we present the results of the experiments of the system as explained in the previous chapter. Table 5.1 presents the experimental setting of the corpus that already use in terms of the process of ideal summary construction as well as minimum and maximum length of sentences that we use in the experimental setting. Table 5.2 shows the distribution of the data set in different domains.

5.1 Experimental Settings

5.1.1 Data Set

For this thesis we only use 30 news articles due to lack of human resource for evaluating the manual summary. The news articles are from different domains such as politics, society, economy and sports were used. Those articles were collected from Dmtsi Woyane Tigray and Aiga Forum websites. The reason that we use those websites they provide a well structured and up-to-date news articles in an electronic format. For the experimental setup each of those articles was free from tables and figures in the document sources. The following tables show the details of our data set.

Data set attributes	Values
Number of news articles	30
Min sentences per document	16
Max sentences per document	42
Average sentences per document	26
Min words per document	420
Max words per document	1213

Average words per document	728

Table 5.1: Statistics of the Data Set

As indicated in table 5.1 the articles have 16 and 42 minimum and maximum number of sentences respectively.

Table 5.2, below shows distributions of news articles based on the different domains or topics.

Domains/Topics	Number Articles
Economic	6
Politics	12
Society	7
Sport	5
Total	30

Table 5.2: distribution of the data set based on domain

5.2. Experiments

As discussed in the earlier chapter in this thesis we use two kinds of features: which are most frequent words based sentence selection and title words based sentence selection for determining the significant part of the sentences in building the summary. For each Tigrinya text document three experiments have been used with different extraction rate in doing the summary. After the preprocessing phase sentences are ranked based on the weight of individual words for the generation of the summary. The final score of the sentence marks the importance of the sentence in Tigrinya text are scored using term frequency and title word with 10%, 20%, 30% extraction rate for term frequency of the data set.

5.2.1. Identification of frequent Tigrinya word

The words which are frequently occurring in every sentence after the removal of stop words are considered to be the important words of the document. In this feature we used a technique that finds the most frequent words in the document. For example selecting most three frequent words or selecting most five frequent words and checking these words in the sentences. Based on this the sentences containing such words are considered to be the important and could be provided in

the summary. To be included in the summary the sentence containing the frequent words are ranked based on the common words or intersections of words in each sentences throughout the document

.

For instance: the below document is the original document

□□□ □□□□□ □□□□ □□□□ □□□ □□□□□

□□□ □□□□ □□ □□□□ □□□ □□□□□ □□□□□ □□□□□ □□□□□□ □□□ □□ □□□□ □□□□ □□□□ □□ □□□□ □□□□ □□□□□□ □□□□□ □□□□□ □□□□□□ □□□□□□ □□□□□□ □□ □□ □□□□ □□ □□□□□□ □□□□ □□□□ □□□□□ □□□□ □□□ □□ □□ □□ □□□□□□ 11 □□□□ □□□□□ □□ □□□□□ □□□ □□□ □□□ □□□□□ □□□ □□□ □□□□ □□□□□ □□□ □□□□ □□□□ □□□□□□ □□ □□□□ □□□□ □□ □□□ □□□□□□ □□□□□ □□□ □□ □□□□□□□ □□□□□□□□ □□□ □□□ □□□ □□□□□ □□□ □□□□ □□□ □□□ □□ ‘□□ □□□ 11 □□□□ □□□□□ □□□□□□□ □□ □□□□□ □□□ □□□ □□□□□□ □□□□□ □□□ 60 □□ □□□ □□□ □□□□□ □□□□□ □□□□ □□□□□ □□□□□□ □□ □□□□□□□ □ □□ □□□□□□ □□□□□□□ □□□□ □□□□ □□□□□□□ □□□□□□ □□ □□ □□□ □□□ □□ □□□ □□□□ □□□□□□□ □□□ □□ □□ □□□□ □□□□ □□□□□ □□□□□ □□ □□□□□□ □□□ □□□□□

□□ □□□□□ 11 □□□□ □□□□ □□□□ □□□□□ □□□ □□□□ □□□□ □□□□ □□□□□□ □□□ □□□□□ □□□□□ □□□□□□□ □□□□□□ □□□□□ □□ □□ □□□□ □□ □□□□□□ □□□ □□□ □□ □□ □□□ □□ □□□□ □□ □□ □□□ 11 □□□□ □□□□□ □□□ □□□ □□□ □□□ □□□□□’’ □□□□□□□□ □□□□□ □□ ’’ □□□ □□□□ □□□□ □□□□□□ □□ □□□□□ □□□ □□□□□□□ □□□ □□□ □□□□ □□□□□ □□□□□ □□□ □□□□ □□ □□□□ □□□ □□□□ □□□□□ □□□ □□□□□□□ □□□□□□ □□□□□□ □□□□□□□□ □□□□ □□□□□ □□□ □□□□□ □□□□□ □□□ □□□ □□ □□□ □□□□□□□ ‘□□□ □□□□□□□ □□□□□□□ □□ □□□ □□ □□□□ □□ □□□□ □□□□□ □□□□ □□□ □□□□ □□□ □□□ □□□□□ □□□□ □□

□□□□□□ □□□□□□□□ □□□□□ □□□□□□□□□ □□□□□□□□□□ □□□□□□□ □□□□□□ □□□□□□ □□□□ □□□□□□ □□□□□ □□□□□□□□□ □□ □□□□□□ □/□□□□□□ □□□□ □□□□ □□□ □□ □□□□□□□□□ □□□□□ □□□□ □□□□□□ □□ □□□□ □□□□ □/□□□□□□ □□□□ □□□□ □□□□ □□□□ □□□□□ □□□□□ □□□□ □□□□ □□□□□□ □□ □□□□□ □□□□□ □□□□ □□ □□□□□ □□□□ □□□□□ □□□□ □□□□□ □□□□□□ □□□□□□ □□□□□□ □□□□□ □□□□ □ □□□□ □/□□□□□□ □□□□ □□□□ □□□□□□□ □□□□□□□□□□ □□□□□ □□□□□□□□ □□□□□□□ □□□□□ □□□□ □□□□□□□ □□□□ □□□□□ □□□□□□ □□□□□□□ □□□□□□□□□ □□□□□□ □□□□□□ □□□□□□□□ □□ □□□□□ □□□□ □□□□□ □□□□□ □□□□□ □□ □□□□ □□ □□□□□□ □□□□□□□ □□□□□ □□□□ □□□□□□□ □□□□ □□□□□ □□□□□ □□□□□□ □□□□□□□ □□□□□ □□□□□□□□ □□□□□ □□□□□□ □□□□□□□ □□□□□□ □□□□ □□□□ □□□□□ □□□□□ □□□□ □□□□□ □□ □□□□ □□□□ □□□□□ □□□□□□ □□ □□□□□ □□□□□ □□□□□ □□□□ □□□□□ □□□□□□□ □ □□ □□□□□ □□□□ □□□□ □ □□ G-8 □□□□ G-20 □ □□ □□□□□□□ □□□□ □□□□ □□□□□□ □□□□□□ □□□□□ □□□□□□□ □□□□□□□□ □□□□□ □□ □□□□□□□□□ □□□□□ □□□□□□□□ □□□□□□□□ □□□□□□□ □□□□□□□□□ □□□□ □□□□ □□□□□ □□ □□□□□ □□□□ □□□□□ □□□□ □□□□□ □□□□□ □□□□□□ □□□□□ □□□□ 11 □□□□□ □□□□□□ □□□□□□□□ □□'□□ □□□□□□ □□ □□□□ □□□ □□ □□□□ □□ □□□□ □□□□□ □□□□□ □□□□ □□□□□□ □□□□□□ □□□□ □□□□□ □□□□□ □□□□ □□□□□□ □□□□□□ □□□□□□□□ □□□□ □□□□ □□□□□□□□ □□□□ □□□□□ □□□□ □□□□□□□ □□□□□□□ □□□□□ □□□□□ □□□□ □□□□□ □□□□□□ □□□□ □□□□□ □□□□□□ □□□□□□ □□'□□ □□□□□□□ □□ □□□□ 11 □□□□□ □□□□ □□□□□ □□□□□□ □□□□□□□ □□□□□□ □□□□□□ □□□□□ □□□□□ □□ □□□□ □□□□□□ □□□□□ □□□□□ □□□□ □□□□□□ □□□□□□ □□ □□□□ □□□□□ □□□□□□ □□□□□□ □□□□□□□ □□ □□□□ □□□□□□ □□□□□□ □□□□ □□□□ □□□□□□ □□□□□ □□□□□□ □□□□□□ □□□□□□ □□□□□□ □□□□ □□ □□□□□□ □□ □□□□□ □□□□□□□ □□ □□□□ □□□□□□□ □□□□□ ‘’ □□□□□ □□ □□□□ □□□□ “ □□□□□ □□□□□□ □□□□□------‘ □□□□□□□□ □□□□ □□□□□□ □□□□□ □□ □□□□ 11 □□□□□ □□□□□ □□□□□ □□□□ □□□□□ □□□□□□ □□□□□ □ □□□□□ □□ □□□□ □□□□□ □□□□□ □□□□□ □□□□ □□□□□□ □□ □□□□□□ □□□□□□ □□ □□□□□ □□□□□□□ □□□□□□□ □□□□□□□□ □□□□□ □□□□□□ □□□□□ □□□□□□□ □□□□□□ □□□□□ □□□□□ □ □□□□□□□ □□□□□□□ □□□□□□ □□ □□□□□ □□□□□ □□□□□□ □ □□ □□□□□□□□ □□ □□□□□ □ □□ □□□□□ □□□□□ □□□□□ □□□□□ □□□□□ □□□□□ □ □□□□□□□□□ □□□□□□□□□□ □□ □□ □□□□□ □□□□□□□□ □□ □□□□□□

□□□ □□□□□ □□ □□□□□□ □□□□ □□ □□□□ □□□□ □□□□□ □□ □□□□□ □□□□□□ □□ □□□□□ □□□□□□ □□□□ □□□□□ □□□ □□□□□ □□□ □□□□□ □□□□□ □□□□ □□□□ □□□□□ □□□□□□□ □□□ □□□□ □□□□ □□□□□ □□□ 11 □□□□ □□ □□□ □□□□□ □□□ □□□ □□□□□□□ □□□□□ □□□□□ □□□□□ □□□□□ □□ □□□□□ □□□□□□ □□□□ □□□□□□ □□□□ □□□□□□□ □□□□□ □□□□□ □□□□□□□□ □□ □□□ □□□□□ □□□□ □□□□ □□□□□ □□□□□□ □□□ □□□□ □□□□□ □□□ □□□□ □□□□□□ □□□□ □□ □□□□□□ □□□□□ □□□□□ □□□□□ □□ □□□ □□□□□ □□□□□ □□□□□ □□□ □□□□□ □□□□□ □□□□ □□ □□□ □□ □□□ □□□□□ □□□ □□□□ □□□□ □□□□□ □□□□□ □□□□□ □□□ □□□□□□□ □□□□ □□□□□ □□ □□□ □□□□□□□ □□□□□ □□□□ □□ □□□□ □□□□□ □□□□ □□□□□ □□□□□ □□ □□□□□ □□□□□ □□□□□ □□□□ □□□□□□ □□□□ □□□□□ □□□□□□□□ □□ □□□□ □□□□ □□□□□□□ □□□□ □□□□ □□□□ □□□□□ □□□□ □□□□

From the above original document after preprocessing here are the most three common words which have appeared frequently more than 9 times in the source documentas shown in the below figure 5.1.

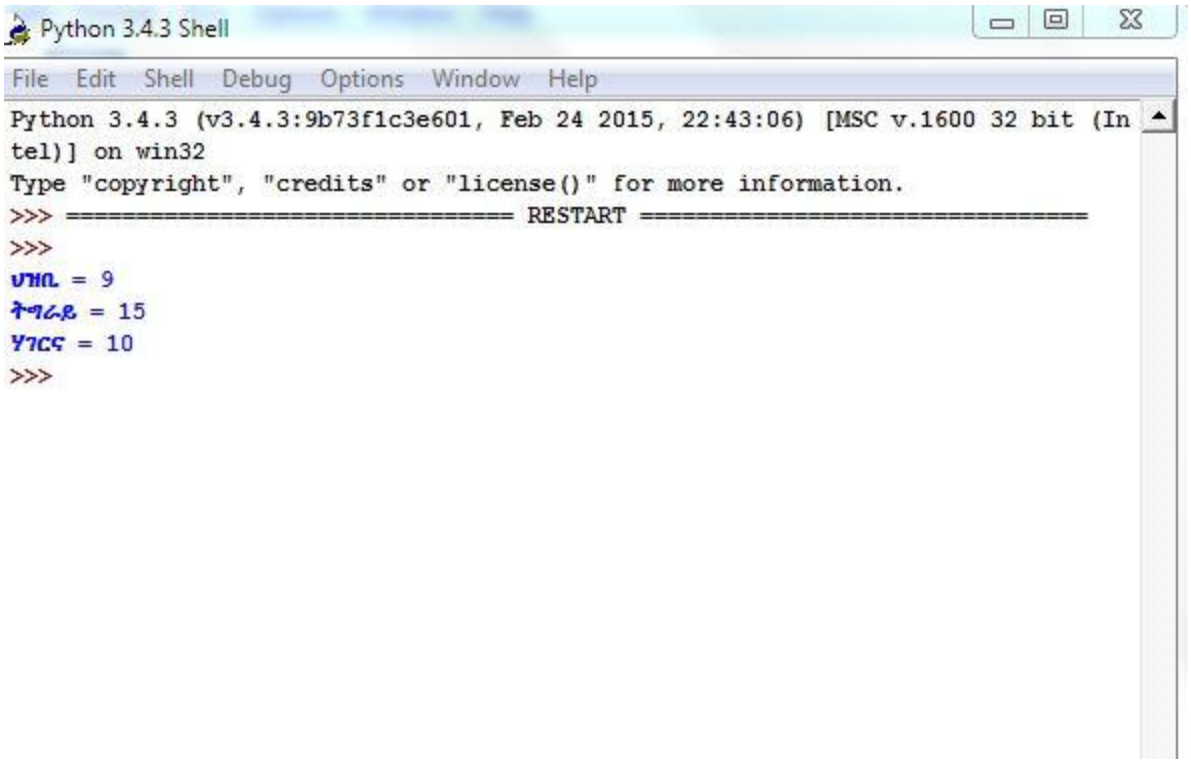

Figure 5.1 selection of most common word

Based on the above program after identified the frequent words here is the summary which consists of selected sentences that have selected words in it. And the program checks all of those words in different sentences and makes the summary.

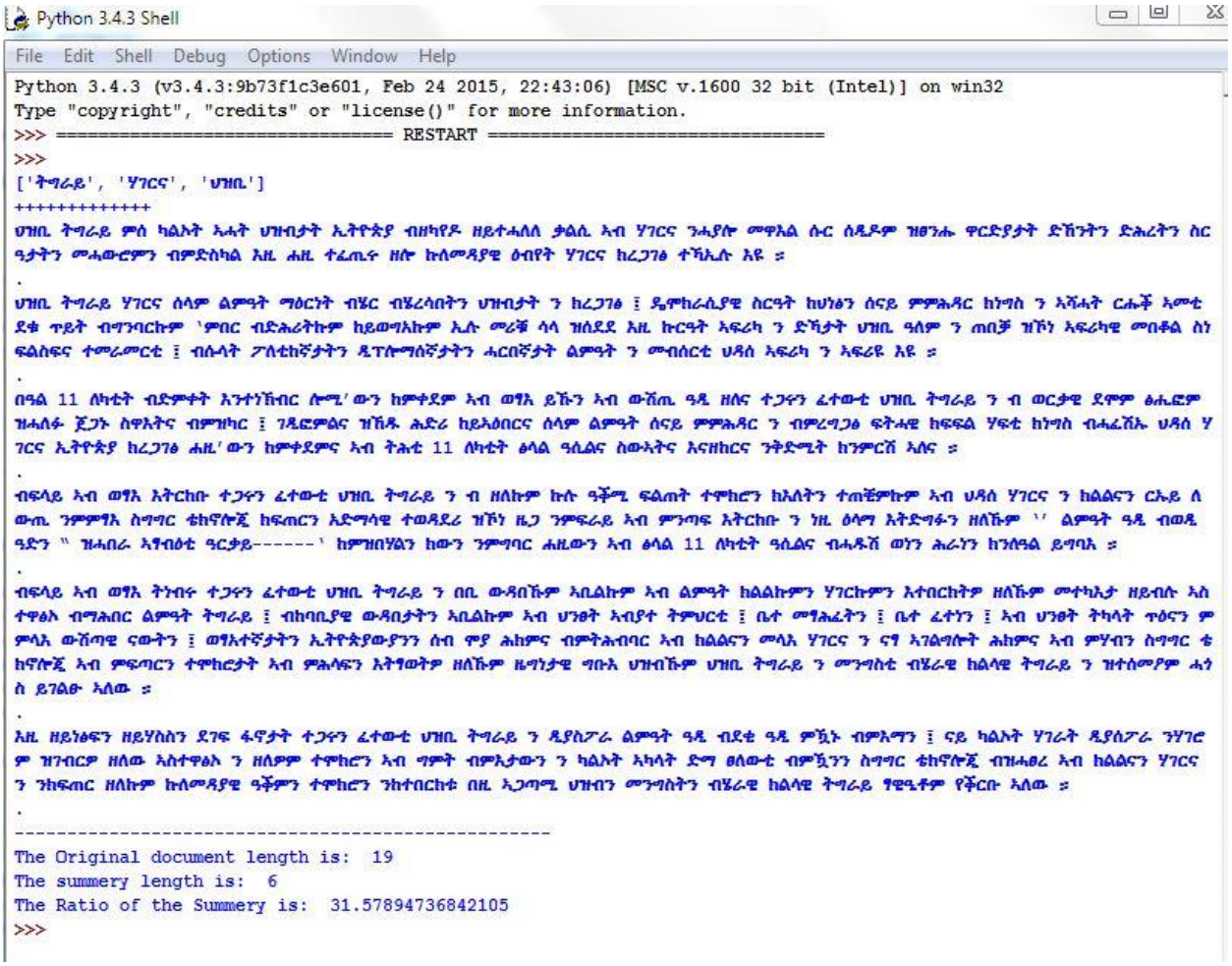

Figure 5.2 selection of the important summary

5.2.2. Identification of Tigrinya title word

The title words features it is assumed that authors always used contents related to the title for filling the article. Therefore the title can be considered as the essential part of the document. Edmundson [14] has defined title words as a feature and that is used to assign a weight to the sentences based on the terms in it that are also present in the title. The researcher has used the title subtitle and heading to identify the title words and has manually assigned weight as it leads to the best performance. In this thesis the selected corpus doesn’t have subtitle and heading but each article consists of an appropriate main title. Sentences containing the title word are considered as important for the summary. All the title words are checked in every sentence throughout the whole document.

As shown in the below figure 5.2 here is the summary of the above original document that contains the title words in the sentence and rank it.

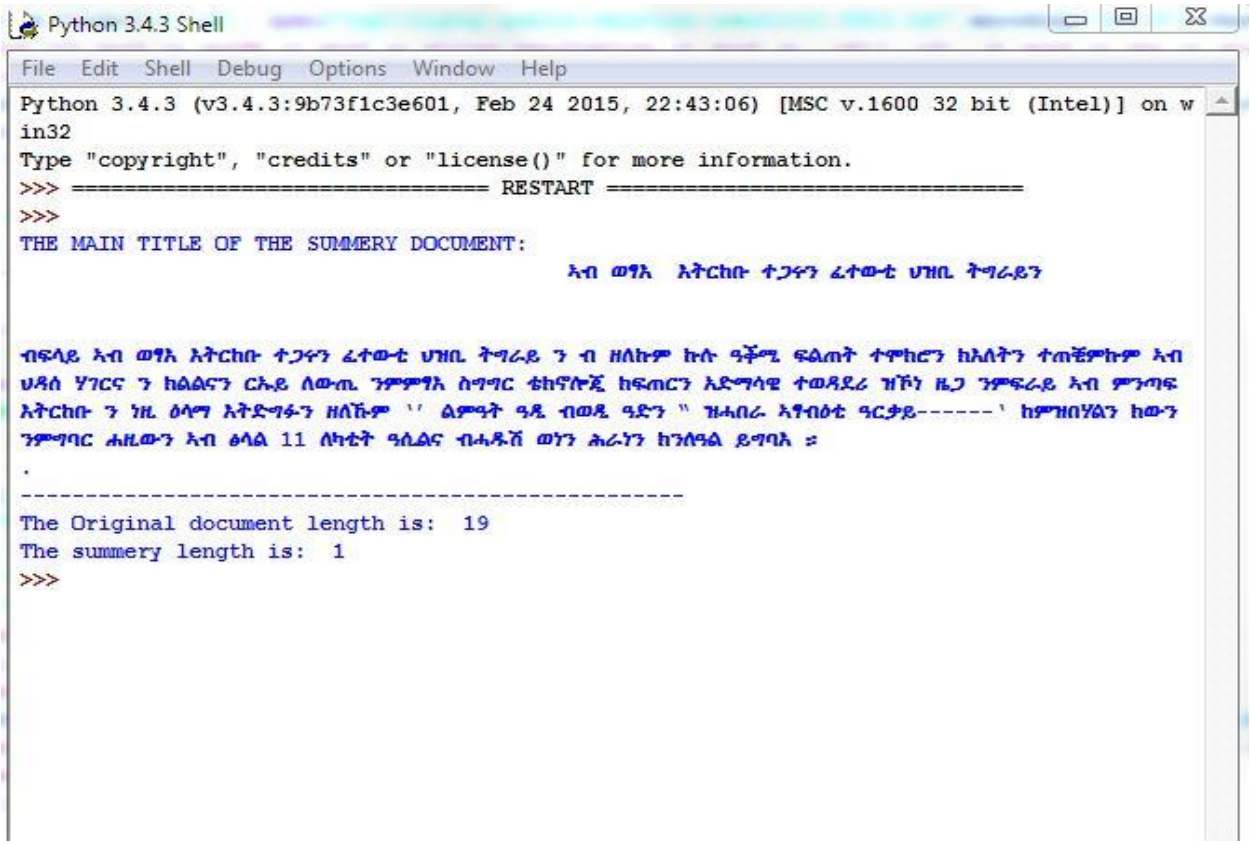

Figure 5.3 title base summary selections

5.3. Text summarization evaluation measures and discussion

An important phase of the development of any system, method or methodology is the evaluation and validation of the task. Natural Language Processing (NLP) systems have also different evaluation mechanisms. This is because a summary's quality can be evaluated in different dimensions such as selected contents importance, which is an objective evaluation, and presentation or linguistic quality etc, which is a subjective evaluation. Both evaluations are as intrinsic evaluation. In this type of evaluation user's judge the quality of the summarization by directly analyzing the summary. Users can judge fluency, how well the summary covers specified key ideas, or how it compares to an ideal summary written by the author of the source text or a human abstractor. None of these measures are entirely satisfactory. The ideal summary, in particular is hard to construct and rarely unique. In most cases there is no only one correct

ideal summary for a given document. For this study the summarizers are evaluated using objective and subjective methods. For both subjective and objective evaluation methods used are intrinsic to the summary.

5.3.1. Subjective evaluation

In order to establish criteria for evaluating automatic summary, 30 automatic summaries were evaluated by three human subjects who are Mekele University staff Journalism departement. The summaries were evaluated in terms of ease of understanding and appropriateness as summaries in three levels: 2- poor, 3-fair, 5- good. The result of the subjective evaluation based evaluation point; results are available in (appendix IV). The subjective evaluation results were converted in to factor scores using factor analysis in order to normalize subjective differences. The evaluation checks weather the summary has smooth transition of sentence, linguistic quality includes non redundancy and referentially and check the best sentence that contain the most important information of the topic sentence.

5.4. Result and discussion of subjective evaluation of system summary

5.4.1. Content of summaries created by the system.

In this section we present the result of subjective evaluation based on evaluation criteria as explained in the appendix II. For content of the summaries created for each text item is scaled out of 100 if the expected total score is 15 scales (2-Poor, 3-Fair, and 5-Good) by three human subject evaluators. The results from the evaluators are turned in to statistics based on the added score of the three results and compared on a scale out of 100. For instance if the summary test1 score 2 by evaluator 1, score 3 by evaluator 2, and score 3 by evaluator 3 have been generated by the three evaluators and the percentage of the overall grading for informativeness and content of sentence generated by the machine was the average value of the sum of the scores given by the evaluators in percentage i.e. 2+3+3=8. The total score out of 15 that mean 8/25=0.53(53%). For more detail of test document in terms of how much the machine summaries covers the important content of the original document and informativeness summaries measure best sentence that contain the most important information of the topic. For detail see the next table. And results are available in (appendix IV).

Document	System summary			
	Term frequency			Title word
	Extraction rate			
	10 %	20 %	30 %	
DocP	0.39	0.48	0.57	0.63
DocSp	0.37	0.48	0.58	0.60
DocSc	0.36	0.46	0.53	0.58
DocE	0.42	0.47	0.58	0.63
Average	**0.39**	**0.47**	**0.57**	**0.61**

Table 5.3 Content of system summaries result

As shown in the above table the subjective evaluation result of the system summary evaluated by the human evaluators were averagely 0.39 (39%), 0.47 (47%) and 0.57 (57%) for 10%, 20% and 30% extraction rates of the term frequency and 0.61 (61%) for the title word which is good parameter for the summary generation as compared to term frequency in representing the important sentences.

5.4.2. Coherence

In this section the evaluation is on how the summary structures the sentences and how they are coherently arranged. The results from the evaluator are turned in to statistics based on the values what the evaluator give for coherence of the sentence are added the score of the three human subject evaluator and compared to the scale of 100. For example if the summary test1 score 2 by evaluator 1, score 5 by evaluator 2 and score2 by evaluator3 and the percentage of the overall grading for the coherence generated was the average of the sum of the scores in percentage i.e. for example 2+5+2= 9 the total score out of 15 that means 9/15=0.60(60%).

Document	System summary			
	Term frequency			Title word
	Extraction rate			
	10 %	20 %	30 %	
DocP	0.43	0.45	0.52	0.55
DocSp	0.47	0.48	0.56	0.59
DocSc	0.42	0.49	0.60	0.63
DocE	0.48	0.55	0.62	0.65
Average	**0.45**	**0.50**	**0.58**	**0.60**

Table 5.4 coherence of system summaries result

As it shows in table 5.4 the obtained result from three human subject evaluator average are 0.45 (45%), 0.50 (50%) and 0.58(58%) for 10%, 20% and 30% extraction rates for the term frequency feature and 0.60% (60%) for title word that was selected the important sentence from a document. The system summary was coherent in terms of the title word features and was better than the performance of term frequency feature.

5.5. Objective evaluation

5.5.1. Precision, recall and F-score

The evaluation of a summary quality is a very ambitious task. Serious questions stay behind concerning the appropriate methods and types as well the types of evaluation. There are a variety of possible bases for the compression of summarization systems performance. In achieving that it could be compare a system summary to the source text, to a human generated summary. In evaluating the quality of the extracted system summary in opposition to the manually extracted summaries, the precision and recall were calculated for the summary which extracted by the computer.

Calculating the precision and recall in measuring the relevance of a set of machine generated output with reference summary is a well established technique for evaluation. Precision is defined as the number of sentences occurring in both system and ideal summaries divided by the number of sentences in the system summary, while recall is defined as the number of sentences

occurring in both system and ideal summaries divided by the number of sentences in ideal summary. Accordingly here is the result of the precision, recal and F-Score.

Document	Term frequency									Title word		
	Precision			Recall			F-Score			Precision	Recall	F-Score
	Extraction rate			Extraction rate			Extraction rate					
	10 %	20 %	30 %	10 %	20 %	30 %	10 %	20 %	30 %			
DocP	0.25	0.38	0.42	0.10	0.25	0.42	0.15	0.30	0.42	0.43	0.42	0.42
DocSp	0.50	0.50	0.57	0.15	0.30	0.57	0.23	0.38	0.57	0.58	0.54	0.56
DocSc	0.33	0.34	0.45	0.11	0.22	0.44	0.17	0.27	0.44	0.44	0.45	0.44
DocE	0.25	0.38	0.42	0.10	0.25	0.42	0.15	0.30	0.42	0.53	0.44	0.48
Average	**0.33**	**0.40**	**0.47**	**0.16**	**0.26**	**0.46**	**0.16**	**0.31**	**0.46**	**0.50**	**0.46**	**0.48**

Table 5.5 objective evaluation result

5.5.2. Results of Objective Evaluation and Discussion

The main thing to notice from the above table is title word is the best parameter than the term frequency in the summary evaluation. The results of the experimentation have been compared with gold standard summary. We compute the recall, precision and F-Score. As it has been discussed in section 5.5.1 Recall (R) is defined as the number of sentences occurring in both system summary and ideal summaries divided by the number of sentences in ideal summary. Whereas precision (P) is defined as the number of sentences occurring in both system and ideal summaries divided by the number of sentences in the system summary. The F-Score is the composite measure that combines both of precision and recall. As it is shown in the above table 5.5 the results obtained for recall, precision and F-Score were 0.47 (47%), 0.46 (46%) and 0.46(46%) respectively at 30% extraction rate for the method of term frequency. On the other hand using the method of title word 0.50 (50%), 0.46 (46%) and 0.48 (48%) has been registered for the recall, precision and F-Score which shows an improvement of the summarizer in this method. According to table 5.5 there is also there is a small percent of recall in the term frequency feature at the extraction rate of 10% which indicates system summary and reference summaries have small number of common sentences. But at the extraction rate of 20% and 30% the summarizer has a better performance.

5.6. Subjective Vs Objective Evaluation Result

In this study the summarizer in both features which are term frequency and title word are evaluated using subjective and objective methods of evaluation. The results of these subjective and objective evaluation shows that title word is the best feature than term frequency in extracting the important sentences that makes the summary. The obtained result of the created summary for the features of term frequency and title word were 57% and 61% respectively in the case of content and informativeness.

In measuring how the created summary is structure and coherent the evaluation shows that 58% for the term frequency and 60% for the title word which is still the best performer in generation of the summary.

5.4.2.1 Domain based evaluation

As shown in the above table 5.3 summaries generated using term frequencies and title word from the economic domain are good in their content and structure at 10 % and 30 % extraction rates than the other domains. Documents on the society domain have the least informativeness at 10%, 20% and 30% in the case of tem frequency as well as title words. In the case of coherence system summaries generated using termfrequency and title word have registered low coherence at 20% and 30% extraction rates a swell as title words.

In general as the experiment shows that the title word feature was the best performer in selecting the important sentences that are good representative of the extracted summary.

CHAPTER SIX

6. Conclusions and Recommendations

6.1. Conclusion

As the amount of textual information available electronically grows rapidly, it becomes more difficult for users to deal with entire texts. To that challenge an automatic document summarization method is progressively more important task. As the document summarization is condensing the source document in to shorter version of that document by preserving its information content. The extractive summarization type is one method that extracts the most significant sentences from the source document to form the summary.

In this thesis we explore an extractive based automatic text summarization for Tigrinya language news texts. However, a vast amount of research has been carried out and many different approaches have been tried out over the last decades in different languages. To this end we explored the statistical based algorithm for selecting the frequent words for Tigrinya language news texts. Two kinds of features were used for selecting the sentences to be included in the summary that are term frequency and title word.

For experimental evaluations a data set consisting of 30 news texts which were collected from fro the web site of Aiga forum and Dimtsi Woyane Tigray was prepared. Three evaluators were employed to prepare three manual summaries for each news text contained in the evaluation data set. Performance evaluations of the two summarization approaches were conducted by comparing the system generated summaries with manual summaries using IR common metrics precision, recall and F-Score which combines both.

The results of the experiments which we have conducted for the news articles considering their average scores have shown that summaries created by proposed system have comparatively closer to informativeness and coherence based on the human evaluators. The obtained result of the created summary for the features of term frequency and title word were 57% and 61% respectively in the case of content and informativeness.

In measuring how the created summary is structure and coherent the evaluation shows that 58% for the term frequency and 60% for the title word which is still the best performer in generation of the summary.

In general as the experiment shows that the title word feature was the best performer in selecting the important sentences that are good representative of the extracted summary.

6.2. Recommendations

This thesis was carried out based on extraction approaches used in automatic text summarization for Tigrinya Language. And we believed that future extensions of this research can be carried out in many directions and this section is intended to address the feature works.

- One of the challenges in conducting this research was the absence of standardized linguistic resource for Tigrinya text summarization like corpus, which is crucial in making a convincing performance evaluation.
- Combination of term frequency and title word features will be good research direction because it has impact in selecting the important sentences. Combination of several scoring metrics outperforms individual metrics [49].
- As we have seen in the result to improve the performance of the summarizer a good developed stemmer is recommended becaue we use shallow stemming and it has an effect in selecting the frequent words.
- As in the experiment shows the title words are good in selecting the important sentences than term frequency it will be also better if the tilte is supported by sub topics of the in put document [14].

References

[1] Hovy, E. H. Automated Text Summarization. In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, chapter 32, pages 583{598.Oxford University Press, ,2005.

[2] Mani, I., House, D., Klein, G., et al .The TIPSTER SUMMAC Text Summarization Evaluation. In Proceedings of EACL, 1999

[3] Jones, Karen Spärck. "Automatic summarizing: a review and discussion of the state of the art." Technical Report UCAM-CL-TR-679, University of Cambridge, 2007.

[4] Karel Ježek and Josef Steinberger, "Automatic Text Summarization (The state of the art 2007 and new challenges)", In Proceedings of Znalosti 2008, pp. 1–12, 2008.

[5] Kifle Derese. "Graph based Amharic text summarizer" Master's Thesis, Graduate Studies, Addis Ababa University, 2014.

[6] Adis, Ashagre. "Automatic Summarization for Amharic Text Using Open Text Summarizer". Master thesis. Graduate Studies, Addis Ababa University, 2013.

[7] Melese Tamru. "Amharic Text Summarizer Using Latent Semantic Analysis" Master's Thesis, Graduate Studies, Addis Ababa University, 2009.

[8] Helen, A. "Text Summarization on Amharic Legal Judgments". Master thesis. Graduate Studies, Addis Ababa University, 2006.

[9] Teferi, A. "The Application of Machine Learning Technique (Naïve Bayes) For Automatic Text Summarization": The Case of Amharic News Texts. Master thesis. Graduate Studies, Addis Ababa University, 2005.

[10] Asefa Baysa. "Query Based Automatic Summarizer for Afaan Oromo Text" Graduate Studies, Addis Ababa University, 2015.

[11] Girma Debele. "Afan Oromo text summarizer" Master's Thesis, Graduate Studies, Addis Ababa University, 2012.

[12] Eyob Delele. "Topic Based Amharic Text Summarizer" Master's Thesis, Graduate Studies, Addis Ababa University, 2012.

[13] Luhn, H. P. The Automatic Creation of Literature Abstracts. IRE National Convention, New York, 1958.

[14] Edmundson, H. P. New methods in automatic extracting. Journal of the Association for comuputing machinery, vol.16, No. 2, PP.264-285, 1969.

[15] Climenson, W. D., Hardwick, N. H., & Jacobson, S. N. Automatic syntax analysis in Machine indexing and abstracting. American Documentation, 178-183, 1961.

[16] Pollock, J. and Zamora, A. Automatic Abstracting Research at Chemical AbstractsService. Journal of Chemical Information and Computer Sciences, 1975.

[17] Meron Sahlemariam, "Concept-Based Automatic Amharic Document Categorization" Master"s Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2009.

[18] Yohannes Afework, "Automatic Amharic Document Categorization: the Case of Ethiopian News Agency", Master"s Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2007.

[19] Tessema Mindaye, "Design and Implementation of Amharic Search Engine", Master"s Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2007.

[20] Yonas Fisseha, "Development of Stemming Algorithm for Tigrinya Text", Master"s Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2011.

[21] Tewodros Hailemeskel, "Text Retrieval: An Experiment Using Latent Semantic Indexing (LSI) with Singular Value Decomposition (SVD)", Master"s Thesis, Addis Ababa University, 2003.

[22] Susan T. Dumais, "Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval", Technical Report Technical Memorandum, Bellcore, 1992.

[23] Lawrence Wong, "ANSES: Automatic News Summarization and Extraction System", Imperial College, Department of Computing, 1998.

[24] Teferi Andargie, "The Application of Machine Learning Technique (NAÏVE BAYES) for Automatic Text summarization (The Case of Amharic News Texts)", Master's Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2005.

[25] Karel Ježek and Josef Steinberger, "Automatic Text Summarization (The state of the art 2007 and new challenges)", In Proceedings of Znalosti 2008, pp. 1–12, 2008.

[26] Julian Kupiec, Jan Pedersen, and Francine Chen, "A Trainable Document Summarizer",

Xerox Palo Alto Research Center, 1995.

[27] Regina Barzilay and Michael Elhadad, “Using Lexical Chains for Text Summarization”, In Proceedings of the ACL/EACL‟97 Workshop on Intelligent Scalable Text Summarization, pp. 10–17, 1997.

[28] Tristan Miller, “Generating Coherent Extracts of Single Documents Using Latent Semantic Analysis”, Master‟s Thesis, Graduate Department of Computer Science, University of Toronto, 2003.

[29] B. Boguraev and C. Kennedy, “Salience Based Content Characterization of Text Documents”, In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, 1997.

[30] Samuel W. K. Chan, Tom B, Y. Lai, W. J.Gao, and Benjamin K. Tsou, “Mining Discourse Markers for Chinese Textual Summarization”, Language Information Sciences Research Center, City University of Hong Kong, 2000.

[31] Laura Alonso, “Representing Discourse for Automatic Text Summarization Via Shallow NLP Techniques”, PhD Thesis, Department of Linguistica General, University of Barcelona, Barcelona, Spain, 2005

[32] W. C. Mann and S. A.Thompson, “Rhetorical Structure Theory: Toward a Functional Theory of Text Organization”, Text, Vol. 8, No. 3, pp. 243-281, 1988.

[33] Waleed Al-Sanie, “Towards an Infrastructure For Arabic Text Summarization Using Rhetorical Structure Theory”, Master‟s Thesis, King Saud University, 2005.

[34] Luis Perez-Breva and Osamu Yoshimi, “Model Selection in Summary Evaluation”, Massachusetts Institute of Technology, Cambridge, 2002.

[35] Ani Nenkova, “Summarization Evaluation for Text and Speech: Issues and Approaches”, Stanford University, Interspeech 2006 – ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 2006.

[36] Dragomir R. Radev, Wai Lam, Arda C. elebi, Simone Teufel, John Blitzer, Danyu Liu, Horacio Saggion, Hong Qi, and Elliott Drabek, “Evaluation challenges in large-scale document summarization”, In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 375-382, July 2003.

[37] Kamil N.,“ Automatic Amharic News Text Summarizer”, Master’s Thesis, Faculty of Informatics, Addis Ababa University, Addis Ababa, 2004.

[38]R.M.Voigt.The classification of central Semitic. Journal of Semitic Studies,(32):1–21,33, 1987.

[39] The 2007 Population and Housing Census of Ethiopia: Statistical Report forTigray Region, CSA 2007 National Statistics, Table 2.1, 2007.

[40] Lewis, M. Paul (ed.). Ethnologue: Languages of the World, Sixteenth edition.Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/, 2009.

[41] Daniel Teklu. "Zebenawi sewasw quanqua Tigrigna", Mekelle: Mega printing enterprise, 2008.

[42] Andrew Spencer. "Morphological Theory: An introduction to word structure ingenerative grammar". Oxford: Blackwell Publishers, 1991.

[43]Gregory T.Stump. Inflectional and Derivational morphology, Kentuckyuniversity,USA: Cambridge University Press, 2001.

[44] Kassa G., Daniel G. Siwasw Tigrigna.Addis Ababa:Mega printing enterprise,2004.

[45] Amanuel Sahle. sewasew Tigrigna bsefiḥu. Lawrencevill, NJ, USA: Red Sea Press, 1998.

[46] Salton, G., Amit, S., Mandar, M., & Buckley, C. Automatic text structuring and summarization. *Information Processing and Management, (33) 3*, 193-207, 1997.

[47] Julian Kupiec, Jan Pedersen, and Francine Chen, "A Trainable Document Summarizer", Xerox Palo Alto Research Center, 1995.

[48] Josef Steinberger, "Text Summarization within the LSA Framework", PhD Thesis, University of West Bohemia in Pilsen, Czech Republic, 2007.

[49] Mark LAST and Martina LITVAK, "Language-independent Techniques for Automated Text Summarization", Ben-Gurion University of the Negev, Beer-Sheva, Israel

[50] Daniel Jurafsky & James H. Martin, "Speech and Language Processing", An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, October 5, 2007

[51] Oskar Gross & Antoine Doucet, "Language-independent Multi-document text Sumarization with Document-Specific Word Associations" Pisa, Italy, 2016

Appendix I: guide line for manual summaries

Dear evaluator you are kindly requested to read the articles provided to you and ranks sentences based on their relative importance in expressing the main theme of the article. The rank that you provide for sentences will be used to construct an ideal summary. Ideal summary is an extract prepared by human-evaluators selecting sentences that are judged to be most important to test the performance of automatic text summarization system. Therefore, while ranking sentences please take the following points in to consideration:

- ✓ Consider the informativeness of the sentence: to what extent the sentence represents the main theme of the article.
- ✓ Try to avoid redundancy: if two or more sentences that represent the same idea but written differently exist in the article, pick one which you judge is best represents the theme of the article.
- ✓ If possible try to be coherent in your selection so that a well-organized ideal summary can be formed by concatenating the sentences together.

Based on these points rank sentences in such a way that the most important sentence of the article should be given a rank of 1, 2 for the second most important, etc., and the least important sentence should be given a rank of N. where, N is the total number of sentences in the article.

Sample of ideal summary

□□ □□ □□□ □□ □□□□ □□□□□ □□□□

4) □□□□□□ □□□□ □□□□□ □□□□□□ □□ □□□□□ □□□□ □□ □□□□□ □□□□□ □□□□ □□□ □□□ □□□□□ □□□□□ □□□□□ □ □□□□ □□□□ □□□□□ □□□ □□□□□ □□□ □□□□□ □□□□□□ □□ □□□□□ □□ □□□□□ □□□□□ □□□□ □□ □

7) □□□□□ □ □□□□□□□ □□□□□ □□□□□□ □□□□ □□□ □□ □□□ □ □□□□□ □□□□ □□□ □□□ □□□□ □□□□□ □□ □□□ □□ □□□□ □□□ □ □□□□□ □ □□□□□□□ □□□□ □□□□ □□□ □□□□ □□□ □□□ □□□ □□□□ □□□□□□□□□□ □□□□□□ □□□□ □

5) □□□□□ □□ □□ □□□□□ □□□□□ □□□□□ □□□ □□□□□ □ □□□□ □□□□□□ □□ □□ □□□ □□□□ □ □□□□ □□□□ □□□□□□ □□□□□ □□ □□□□□ □□□□□ □□ □□□ □□□□□□ □□□ □□□□ □□□ □□ □□□□ □□ □□ □□□□ □□□□□□ □□□□□□ □□ □□□□□ □□□□□ □□□□□ □□□□□ □□ □□□ □□□□□ □□□□□ □□ □□□□ □□□ □□□□□ □□ □

6) □□□□□ □□ □□□□□ □□□□ □□□□□ □□□□ □□□□ □□ □□□ □ □□□□ □□□□□ □□□□□□□ □□□ □□□□□ □□□□ □□□ □□□□ □□□□ □□□□□ □ □□□□□ □□□ □

8) □□□□ □□□□□ □□□□□ □□□□□ □□□□ □□ □□□ □□ □□□□□ □□□□ □□□□□ □□□ □□ □□□□□□ □□□ □□□ □□□□ □□□□□□ □□□□ □□ □□□ □□□ □□□□□ □□□ □□□□□□ □□□ □□□□ □□□□□ □□ □

11) □□ □□□□□ □□□□ □□□□□□□ □□□□□ □□□□□□ □□□□ □□□□□ □□□□ □□□□□□ □□□□ □□□□ □□□□ □□□□□ □□□□□□ □□ □□□□□ □□ □□□□□□ □□□□ □□□□□ □□□□ □□ □

13) □□□ □□□□□ □□ □□□□□□ □□□ □□□□ □□□□ □ □□□□□ □□□□□ □□□□□ □ □□□□ □□ □□□□□□ □□□□ □□□□□ □□ □□□ □□□ □□□□ □□□□ □□□□□□□□□ □□□□ □□□ □□□□ □□□□ □□□□□□ □□□□ □□□ □

12) □□ □□□□ □□ □□□ □□□ □□□□ □□□□□□ □□□□□□□□ □□□□ □□□ □□□□ □□□□□ □□ □□□ □□ □□□□ □□□□□ □□□ □ □□□ □□□□□ □□□□ □□□□□□□ □□□□□□□ □□□□ □□ □

9) □□□ □□□□□□ □□□□ □□□□□□□□ □□□□□□ □□□□□ □□□ □□□□□ □□□□ □□ □□□ □□□ □□□□□ □□□ □□ □□□□ □ □□□ □ □□□□ □□□□ □□ □□□□□ □□□□ □□ □□□□ □□□□□ □□□□□□ □□□□□ □□□ □ □□□□ □□□□□ □□□ □□□□ □□ □□ □□ □□□□□□□ □□□ □□□□□ □□□□ □□□□□□□ □ □□□□ □□□□□ □□□□□ □□ □

14) □□ □□□□□□□ □□ □□□□□□□ □□□□□□ □□□□□□□□ □□□□□ □□□□□□ □□□□□ □□□□□□ □□□□ □□□□ □□□□□ □□□□□□ □□□□□□ □□□□□ □□ □□□□□ □□□□□□ □□□□□ □□□□□□ □□□□□□ □□□□□ □□□□□□ □□□□ □□□□□□□ □□ □

10) □□□□ □□□□□ □□ □ □□□□□ □□ □□□□□ □□□□ □□□□□ □□□□ □□□□ □□□□□□ □□ □□□□□□ □□□□ □□□□ □□□□□ □□□□ □□□□□ □□□□□□ □□□□ □□ □□□□□□□ □□□□ □□□□ □□□□□□ □□□□□□□ □□ □□□□□□ □□□□□ □□□□□ □□□□ □

3) □□□□ □□□□□ □□□□□□ □□□□□ □□ □□□□□□ □ □□□□□ □□ □□□□□ □□□□□□ □□□□ □□□□ □□ □□□□□ □□□□□ □

15) □□ □□□□□□ □□□□ □□□□ □□□□□ □□□□ □□ □□□□ □□□□□ □□□□□□□ □□□□ □□□□□□ □□ □

17) □□□□ □□□□ □□□□ □□□□□ □□□□□ □□□□□ □□□□ □□□□□ □□□□□□ □□□□□□ □□□□ □□□□ □□□□□ □□□□ □

2) □□□□□ □□□□□ □ □□□□□ □□□□□ □□□□□ □□□□ □□□□ □□□□□ □□□□ □□□□□□ □□□□□ □□□□ □□ □

21) □□□□□□□ ?

18) □□ □□□□□□ □□□□□□ □□□□□□ □□□□□ □□□□□□ □□□□□□□ □□□□□□ □□ □□□□□□ □□□□□□ □□ □□□□□□□ □□□□□□ □□□□ □□□□□□ □□□□□□ □

20) □□□□□□□ ?

19) □□□□ □□□□□□ □□□□□□ □□ □□□□□ □□□□□ □□□□□□ □□ □□□□□□□□□ □□□□□ □□□□□□□ □□□□ □□□□□ □□□□ □□□□ □□□□□ □□□□□□ □

16) □□□□□□□ □□□□□□□ □□□□□ □□□□□ □□□□□ □ □□□□□ □□□□□ □□□□□□□ □□□□□ □□□□□ □□□□□ □□□□ □□□□□ □□□□□□ □□□□□□ □□□□□□□□□□□□□ □

1) □□□□ □□□□□ □□□□□□ □□□□ □ □□□□□□□ □□□□□ □□ □□□□□ □□□□□□ □□□□ □□□□□ □ □□□□ □□ □□□□ □ □□□□□ □□□□□ □□□□□ □□□□□□□□ □□ □□ □

Appendix-II: guideline for subjective evaluation

Dear evaluator, you are kindly expected to read the document cautiously and then you are going to evaluate system summaries according to evaluation scale from 2-5. This evaluation scale is give based on text quality measure such as informativeness and content of the summaries and coherence of the sentences.

1. Is the summary including best sentences of the document as well as the most important information of the topic sentence?
2. Does the summary have good structure and the sentences are coherent?

You can give score from 2 to 5 scales; the scales are represented 2 for poor, 3 for fair and 5 for good and you can select one of them.

Appendix III: Tigrinya stop word list

□□□	□□□
□□□□	□□□
□□	□□□□
□□□	□□□
□□□	□□
□□□□	□□'□□
□□	□□
□□	□□□
□□□	□□□□
□□□	□□□□
□□□	□□□
□□	□□

List of stop words [20]

Appendix IV Prefix list

□ □

□ □

□□ □□□

□□ □□□□□

□□ □□

□□ □

□ □

□□□ □□

□□ □□□

□ □□

□ □□□

□ □□□

Appendix V suffix list

□ □

□ □□

□□ □

□ □

□□

□ □□

□ □

□ □

□ □

□ □

□□ □□□

□□ □□

□ □□

□

□

List of suffix and prefix [20]

Appendix VI subjective summary evaluation

Content of system summary evaluation using term frequency				
Document	E1	E2	E3	Total
DocP	2	2	2	6
DocP	2	3	2	7
DocP	3	2	3	8
DocSp	2	2	2	6
DocSp	3	2	2	7
DocSp	2	3	3	8
DocSc	2	2	2	6
DocSc	3	2	2	7
DocSc	2	3	2	7
DocE	2	2	2	6
DocE	2	3	3	8
DocE	3	3	2	8
Content of system summary evaluation using title word				
Document	E1	E2	E3	Total
DocP	3	2	2	7
DocP	2	3	3	8
DocP	2	3	5	10
DocSp	3	2	2	7
DocSp	5	2	2	9
DocSp	5	3	2	10
DocSc	3	3	2	8
DocSc	3	5	2	9
DocSc	3	5	2	10
DocE	3	2	2	7
DocE	3	3	3	9
DocE	5	2	3	10

Coherence system summary result using term frequency				
Document	E1	E2	E3	Total
DocP	2	2	2	6
DocP	2	2	2	6
DocP	3	2	3	8
DocSp	2	2	2	6
DocSp	3	2	2	7
DocSp	3	3	3	9
DocSc	2	2	3	7
DocSc	3	2	2	7
DocSc	3	3	3	9
DocE	2	2	3	7
DocE	2	3	3	8
DocE	3	3	3	9
Coherence system summary evaluation using title word				
Document	E1	E2	E3	Total
DocP	3	2	3	8
DocP	2	3	5	10
DocP	3	3	5	11
DocSp	3	2	2	7
DocSp	3	2	2	8
DocSp	5	3	3	9
DocSc	2	3	3	8
DocSc	3	5	3	10
DocSc	3	5	2	10
DocE	3	2	2	7
DocE	3	3	3	9
DocE	3	2	3	9

Appendix VII sample of source document machine extracted summaries

Sample source: Doc_E

□□ □□□□□□ □□□□□ □□□□ □□□□ □□□□ 2008 □.□ □□ □□□□ □□□□ □□ □□□□ □□□□ □□□□□ □□□□ □□□□□ □□□□□ □□□□□□□□□ □□□□□□□ □□□□□□□□□ □□ □□□□ □□ □□□□□□□ □□□□□□ □□□□□□ □□□□□□ □□□□ □□□□□□□ □□ □□□□□□ □□□□ □□□□□ □□□□ □□□□□□□ □□□□□ □□□□□ □□ □□□□□□□□ □□□□ □□□□□ □□□□□ □□□□□ □□□□□□ □□□□ □□□□□□□□□□ □□□□□ □□□□□□□ □□□□ □□□□□□ □□□□ □□□□□□ □□□□ □□□□□□□ □□□□□ □□ □□□□ □□□□□ □□ □□□□ □□ □□□□□□ □□□□ □□□□□□□ □□□□□□□□□□ □□□□□ □□□□ □□□□□□ □□□□□ □□ □□□□ □□□□ □□□□□□□ □□□□□□ □□□□□ □□□□ □ □□□□□ □□□□□□□ □□□□□□ □□□□□□□ □□□□□□□□□□ □□□□□ □□□□□□ □□□□□□ □□□□□□□ □□□□□ □□□□□ □□□□ □□ □□□□□□□□□□ □□□□□□□ □□□□□ □□□□□□ □□ □□□□□□□ □□□□□ □□ □□□□□□□ □□ □□□□ □□□□□ □□ □□□□□ □□ □□□□ □□□□ □□□□□□ □□□□□ □□□□ □ □□□□□ □□□□□□ □□□□ □ □□□□□ □□□□□ □□□□□ □□□□□□□ □□□□□□ □□ □□□□ □□□□□ □□ □□□□□ □□□□□ □□□□□ □□□□□□ □□□□ □□□□ □□□□□□ □□□□ □□□□□□ □□□□ □

□□’□□ □□□□□□ □□□□□□ □□□□□□□ □□□□□□ □□□□□□ □□□□□□□ □□□□□□□ □□□□□ □□□□□□□ □□□□□□□□□ □□□□□□□ □□□□ □□□□ □□□□ □□□□□□ □□□□□ □□ □□□□□□ □□□□ □ □□ □□□□□ □□□□□ □□□□□ □□□□□□□□ □□□□□□□□□ □□□□□□□□ □□□□ □□□□□□□ □□□□□ □□□□□ □□□□□□ □□□□□ □ □□□□□□ □□□□□□□ □ □□□□□ □□□□□□ □□□□□□□□ □□□□□□□ □□□□□ □□’□ □□□□ □□□□□‘□ □□□□□□□ □□ □□□□□ □□□□□ □□□□ □□□□□□ □□□□□ □□ □□□□ □□□□□□ □□ □□□□ □□□□ □□□□□□□□ □□□□ □□□□ □□□□□ □□ □□□□□ □□□□□□ □□□□□□□ □□□□□□□ □□□□□□□ □□ □□□□ □□□□□ □□ □□ □□□□ □□□□□ □□□□□ □ □□□□□ □ □□□□ □□□□ □ □□□□□□□ □□□□□ □□□□ □□□□ □□□□□□□□ □□□□ □□□□□ □□ □□□□□□□□ □□□□□□□ □□□□□□ □□□□□ □□□□□ (80%) □□ □□□□ □□ □□□□□ □□ □□□□□□ □□ □□□□□ □□ □□□□ □□□□ □□□□□□□ □□□□□□ □□□□□ □□□□□ □□□□□□ □□□□□□ □□□□□□□ □□□□□□ □□□□□ □□□□□□ □□ □□□□ □□□□ □□□□□□ □□□□□□□

□□□□ □□□□□ □□ □□□□□ □□ □□□□ □□□□□□□ □□□□ □□□□□ □□□□□ □□ □□□ □□ □□□ 2018 □□□□□ □□□□ □□□ □□ □□□□ □□□ □□□ □□□□□ □□□□□ □□□□□ □□ □

□□□□□ □□□□□□ □□ □□□ □□□ □□ □□□□□ □□□□□ □□□□□□ □□□□□ □□□□□ □□ □□□□□□□□ □□□□□□ □□□□ □32 □□□□ □□ □□□□ □□□□□ □□□□□□ □□ □ □□□□ □□□ □ □□□□□ □□□□□ □□□□□ □□□□ □□□□ □□□□ □□□□ □□’□ □□ □□□□ □□□□ □□ □□ □□□□□ □□□ □□□□□□□□ □□□□□□ □□□ □□□□□ □□□ □□ □□□□□ □□□□□ □□□□ □□□□ □□□ □□□□□ □□□□□□ □□□□□□ □□□ □□□□□ □□□□□ □□□□□ □□□ □□□□ □□ □□□ □□□ □□□□ □□□□□ □□□□□□ □□□□□ □□□□□ □□ □□□ □□□□ □□□□ □□□’□ □□□ □□ □□□ □□□□□ □□□□ □□□ □□□□□ □□□ □□□ □□□ □□□ □□□□ □□□□□□ □□□ □□ □□□ □□□ □□ □□□□□□□□□ □□□□□□ □□□□ □□ □□□ □□□ □□□□ □□□□□ □□□□□□ □□□□ □□□ □□□□ □□□□ □□□□□□ □ □□□□ □ □□□□□ □□□□ □□□□□ □ □□□□ □□□□ □□□□ □□ □□□ □□□ □□ □□□ □□□ □□□□ □□ □□□□□□□□□ □□□□□□ □□□□□ □□□□□ □□□ □□□□□ □□ □□□□□ □□□ □□ □□□ □□□□ □□□ □ □□□ □□□□ □□□□ □□ □□□□□□□□□ □□□□□□ □□□ □□ □□□□ □□□ □□□ □□ □□□ □□□ □□□□ □□□□□□□□□ □□□□□□ □□□□ □□□□ □□□□□□ □□□□□ □□□ □

□□□□□ □□□ □□□□ □□ □□□□ □□□ □□□□ □□□□□ □□□ □□□ □□□□□ □□□□ □□□ □□□□□ □□□□□ □□□□□ □□□ □□□ □□□□ □□□ □□ □□□□ □□□□□ □□□□ □□□□□ □□□□ □□□□□□□□□ □□□□□□□ □□□□□□□ □□□ □□□□□□□ □□□ □□□□ □□□□□ □□□ □□□ □□□ □□□□ □□□□ □□□□□ □□□□ (95%) □□□ □□□ □□ □□□□ □□ □□□ □□/value / □□□□ □□□□□□□ □□□ □□□ □□□ □□□□□ □□□□□□□ □□ □□□□ □□□ □□□□□□□□□□□□ □□□ □□□□□ □□□□□ □□□ □□□□□ □□□□ □□□ □□ □□□□□□ □□□□ □□□ □□□ □□ □□□□ □□ □□□□□ □□□ □□□□ □□/value / □□□□□ □□ □□□ □□□□□ □□□ □□□ □□ □□□□ □□□ □□□□□ □□□□□□□ □□ □□□□ □□□ □□□□ □□□□□ □□□□ □□□□□ □□□□□ □□ □□□’□ □□□□□ □□□□□□ □□□ □□□ □□□ □□□□ □□□□ □□□□□ □ □□□ □ □□□□ □ □□□□

[illegible] 100 [illegible] (rural ransformation center) [illegible]

[illegible] 10 [illegible] 15 [illegible] (export) [illegible] 15 [illegible] 85 [illegible]

Summary with 30% extraction rate

[illegible] 32 [illegible] 15 [illegible]

□□
□□
□□□□□□□□□□□□□□□□□□□□□□□ □□□ □□ □□□□ □□□ □□□ □□ □□□□□□ □□□□□ □□□□ □□□ □□□□□□ □□□□ □□□□□□ □□ □□□ □□□ □□□ □□□□□ □ □□□ □□□□□□□□□□□□ □□□□ □□□□ □□ □□□□□ □□□ □□□□□□ □□ □□□ □□□ □□□ □

Summary with 20% extraction rate

□□32
□□ 15
□□

Summary with 10% extraction rate

□□32
□□

Declaration

The undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all sources of materials for the thesis have been suitably acknowledged.

Guesh Amiha

This thesis has been submitted for examination with my approval as an advisor.

WONDWOSSEN MULGETA (PHD)