
ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

# AUTOMATIC SUMMARIZATION FOR AMHARIC TEXT USING OPEN TEXT SUMMARIZER

ADDIS ASHAGRE TEKLEWOLD

JUNE, 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

# AUTOMATIC SUMMARIZATION FOR AMHARIC TEXT USING OPEN TEXT SUMMARIZER

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

ADDIS ASHAGRE TEKLEWOLD

JUNE, 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

# AUTOMATIC SUMMARIZATION FOR AMHARIC TEXT USING OPEN TEXT SUMMARIZER

By

ADDIS ASHAGRE TEKLEWOLD

Name and signature of Members of Examining Board for Approval

| **Name** | **Title** | **Signature** | **Date** |
|---|---|---|---|
| | Chairperson, | | |
| **Martha Yifiru (PhD)** | Advisor, | | |
| **Ermias Abebe** | Examiner, | | |

# DECLARATION

THIS THESIS IS MY ORIGINAL WORK AND HAS NOT BEEN SUBMITTED FOR AS A PARTIAL REQUIREMENT FOR A DEGREE IN ANY UNIVERSITY.

---

ADDIS ASHAGRE TEKLEWOLD

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

---

DR. MARTHA YIFIRU

JUNE, 2013

# Dedication

This thesis is dedicated to my loving family and friends. Thank you all for your unfailing support in every way I needed.

## ACKNOWLEDGEMENT

Thank you GOD for YOUR uplifting Sprit, love, grace and mercy in my life. I know I won't make it without YOU!

I would like to thank my mum, W/ro Muluwork, sisters: Yodit, Tigist, Feven, Hanna, my one and only brother Yidnekachew, my cousin Mihret for your unfailing love, encouragement and support. Miss you dad! My nephew Nazri and niece Susu I love you so much!

I would like to express my deepest gratitude to my advisor Dr. Martha Yifiru, for her commitment, support, constructive comments and guidance throughout this research.

I would also like to thank Jenefir Costa and Susan P.D. for your financial support to sponsor my education. Thank you Barbara Patton, for your input and encouraging support to continue my study.

I would like to thank those who participated in the manual summary preparation and evaluation process: Ashenafi W., Biruk F., Tigist A. and Solome G.. Also thank you Fikadu B., Girma D., Yeshinigus G. and Melese T. for your support and sharing your experiences.

Last and definitely not least, all my friends really thank you.

## WHERE THERE IS LOVE, THERE IS GOD AND WHERE THERE IS GOD, THERE IS LIFE!

Addis Ashagre

June 2013

# Table of Contents

## List of Tables

## List of Figures

## List of Algorithms

# LIST OF ACRONYMS

AAU – Addis Ababa University

Adj – Adjectives

AdjP – Adjectival Phrase

Adv – Adverbs

AdvP – Adverbial Phrase

C – Consonants

E1 – Experiment one

E2 – Experiment Two

EPA – Ethiopian Press Agency

F-m – F-measure

GUI – Graphic User Interface

HTML – Hyper Text Markup Language

LA – Large Size Articles

LSA – Latent Semantic Analysis

MA – Medium Size Articles

N – Nouns

NLP – Natural Language Processing

NP – Noun Phrase

O – Object

OTS – Open Text Summarizer

P – Precision

PP – Prepositional Phrase

R – Recall

RANP – Reporter Amharic news paper

RQ – Research Question

S – Subject

SA – Short Size Articles

TF – Term Frequency

V - Vowels

Vb - Verbs

VP – Verb Phrase

WIC – Walta Information Center

## ABSTRACT

Information overload is a problem in this information era due to the mass production of information in many formats which is enhanced by the internet technology. Amharic text documents are part of this mass production. In order to extract the useful information from a given text document with in short time, automatic text summarization plays a decisive role. There are quite a few researches done for Amharic text summarization but still more research needs to be done to accomplish better result achieved in other languages like English.

The objective of this study is therefore to investigate the applicability of the open text summarizer for Amharic news text summarization. The system is an open source, language independent single document text summarization tool. It uses combinations of term frequency and sentence position methods to rank the sentences of the article. 40 news articles on different issues are gathered from EPA, WIC and RANP web pages from which a corpus containing 30 news articles is prepared for the experimentation.

Some modifications were made on the interface of the tool that was designed in C# programming language. The OTS tool is customized in two ways for performing the two experiments. The first one is done without changing the code of the tool significantly, but with few modifications on the punctuation rules and by preparing the dictionary file that holds the Amharic language lexicons. The system uses language specific lexicons which include list of affixes, abbreviations, stop words, synonyms, compound words and other rules. The second one is done by changing the Porter stemmer of the tool with an Amharic stemmer. The experiment is done on both systems by generating 90 summaries for each news article at 10%, 20% and 30% extraction rates.

The performance of the two systems is evaluated using subjective and objective evaluation. Subjective evaluation is done for 45 summaries extracted in experiment one and good result is obtained. Objective evaluation is done for all the summaries generated in both experiments by comparing them with an ideal manual summary using F-measure. The highest 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 system with Amharic stemmer gave better performance than the other regardless of the size of the original article in a given extraction rate with better average corpus score at 20% and 30%. The system also showed regularity in performance improvement as the extraction rate increases.

# CHAPTER ONE

## 1.1. INTRODUCTION

In this information era documents are produced and disseminated through printed formats, digital formats and the internet every day. To benefit from this mass production of information, there are many ways of document retrieval systems and search engines developed. These systems give access to and make information available for the user to choose from based on their requirement or need. In accessing these available digital documents, the major challenge is information overload which occurs due to the hundreds or thousands of relevant documents presented to the user for one query.

As the quantity of digital data increases causing more problem of information overload, the interest in automatic summarization has also increased. This is to address the need by the user to get the main idea of the document without reading the whole document so that the document that meets the user needs can be chosen within short time. The technology of automatic text summarization is becoming indispensable solution for dealing with the information overload problem. Early experimentation in the late 1950s and 1960s suggested that text summarization by computer was possible though it was not straight forward (Luhn, 1958).

Text summarization is one of the natural language processing (NLP) application that propose to extract the most important information from a source to produce a condensed version for a particular user or task. Automatic text summarization is the process of reducing a text document with a computer program in order to create a summary that retains the concept of the original document. In order to generate a summary of a document, we have to identify key pieces of information existing in a document or a sentence, omitting the redundant information and reducing details. The goal of automatic text summarization is to take an information source, extract content from it and present the major content to the user in a condensed form and in a manner sensitive to the user's or application's need (Lloret, 2008).

Based on the quantity of the input document used there are two types of document summarization. These are single document summarization and multi document summarization.

Single document summarization, as the name implies takes a single document and presents the most important content in a condensed manner for the needs of further task. Whereas multi document summarization extracts information from many text documents written about the same topic. The complexity involved with the integration of different documents about a topic to extract a single summary, makes multi document summarization more complex and wider concept than the single document summarization.

There are two approaches to automatic summarization of text documents which are extraction and abstraction. Extractive methods work by selecting a subset of existing words, phrases or sentences in the original text to form the summary. Whereas abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original document. Extraction approach is more practicable in many applications of text summarization.

In applying the extraction approach, there are different methods for measuring the importance of a sentence to be used in the summary. To choose the concept that should be included in the summary, some algorithms calculate a weight for each sentence, taking into account the position of the sentence and word frequencies, while others use semantic information in order to find the hierarchy of concepts (Girma, 2012).

The major application areas, where automatic text summarization is used, are:-

- Search engines – to present compressed descriptions of the search results to the user so that the user can read and understand the content of the retrieved documents.
- Document summarization – to store only the summarized version rather than the whole document.
- Text to Speech application – a text to speech application can use summaries rather than the whole document, since written text can be too long to listen to and time taking while the main information can be transmitted to the listener from the summary of the text.
- Small Screen Devices – summaries best fit small screen applications like in smart phones and iPods rather than having the whole document which might be too long in the small screen devices.

## 1.2. STATEMENT OF THE PROBLEM

Naturally human beings use only few of the information they perceive from their sense organs. This nature of humans creates the desire of abstraction in everyday life. Due to developments in technology and change in life style, people have less time to spend reading huge amounts of information available digitally to satisfy their considerably less information need. Users therefore tend to focus on condensing the information provided to them by taking abstractive perception and ignoring the details if they are not interested.

Amharic being the official spoken and written language of Ethiopia is used to produce text documents for readers. These text documents are available digitally and the amount is highly increasing every day. Even Google now supports searching for documents with Amharic fonts using queries in Amharic font. Interested users are now spending time in searching for Amharic documents online, which provides lots of text documents creating user information overload.

In addition to these, nowadays many local companies and service providers are developing bilingual or multilingual web pages for their customers/users. News providing organizations are among the major providers of Amharic texts for online readers. Blogs, Articles, laws, proclamations, health related printouts, religious teachings, awareness creation broachers and educational printed materials also contribute to the production of information in Amharic.

To handle all these available text documents and make use of its information content, it is necessary to consider the use of text summarization technology for Amharic text documents. This in effect helps users to get the major concept of the document and choose the relevant document based on their interest without reading the whole text within short time. Indirectly it also increases the usability of the available information in Amharic text documents.

There are few researches conducted locally in the area of automatic text summarization applying different methodologies. To find the best one that is compatible, effective and efficient for Amharic text more research has to be done. The first research "Amharic News Text Summarization" was done by Kamil Nuri, in 2004. The system was developed by using natural language processing techniques and statistical methods. Title words, head sentences, head sentence words, paragraph starting sentence, cue phrases and high frequency key words are used as extraction features. The system evaluation of the research showed 74.4% precision and 58%

recall (Kamil, 2004). He used nine news articles that were in printed format, since by the time the research was conducted there were no web based Amharic news service providers. However, since he used Perl programming language which does not support Amharic alphabets, he transliterated the nine news articles. That means he did not use Amharic alphabets directly to the system which makes the system not to be directly implemented to the real world Amharic text summarizing application.

The second work, "The Application of Machine Learning Technique for Automatic Text Summarization – The Case of Amharic News Text", is done by Teferi Andargie in 2005. He used predefined features like location of a sentence in a document, title words occurring in the sentence, and cue words occurring in the sentence that are found to be a good indicator in giving an optimum summary. He used a corpus of 480 news articles in the experiment which was used by Saba (2001), Theodros (2003) and Samuel (2004) for testing different retrieval models. A manual summary at 30% extraction rate was prepared for the 480 articles. The naïve Bayes algorithm is used to classify sentences as a summary or not a summary based on the feature vectors (Teferi, 2005). A prototype is developed which extracts sentences to a desired compression level. The result of the experiment shows that the location features gives the best result in the classification of sentences when using individual features.

The other study, "Automatic Text Summarization for Amharic Legal Judgments" is a work done by Helen Adane in 2006. The study focuses on producing prototype system of text summarization on Amharic legal judgments using Python programming tool (Helen, 2006). The methodology employed is an extraction technique on Amharic legal judgments rendered by the supreme court of Ethiopia. To evaluate the performance of the system a random automatic summary is generated using the same extraction rates (10% and 20%) by the system for the selected legal judgments. Using extrinsic evaluation technique, the performance of the system summary and the random summary were compared with an ideal (manual) summary which is manually prepared summary by legal experts. The result showed that the sentences extracted by the system summary using different extraction features are much closer to the manual summary (Helen, 2006). The domain of the research, legal judgment, writing style is inverted pyramid or downward triangular which states the most important part of the paragraph on the last sentence of the paragraph.

Another study, "Automatic Amharic Text Summarization Using Latent Semantic Analysis (LSA)" is done by Melese Tamiru in 2009. The study proposed two new generic text summarization methods, topic LSA and LSA-graph, which create text summaries by ranking and extracting sentences from the original documents (Melese, 2009). Based on these two methods, a prototype Amharic news text summarization system is built. A manually ranked summary of the tested 50 news articles are used to evaluate the performance of the proposed system. The two methods gave comparatively closer results despite the difference in the procedure followed. The system performance is also compared with other graph based summarization methods, and claims to perform better than all previously developed graph based methods (Melese, 2009).

The only text summarization research conducted for Oromifa language was the one done by Girma Debele in 2012 under the title "Afan Oromo News Text Summarizer". It is developed based on the Open Text Summarizer which uses the combinations of term frequency and sentence position methods with language specific lexicons in order to identify the most important sentence for extractive summary (Girma, 2012). Three methods were developed and tested; 1st method uses term frequency and position methods, 2nd using Afan Oromo stemmer and language specific lexicons with additional feature on the first method, 3rd with improved position method and term frequency on the 2nd method. The evaluation showed that the 3rd method with improved position method and term frequency as well as the stemmer and language specific lexicons like synonyms and abbreviations outperformed the other two methods. This research had an advantage on the written Oromifa language. Since Oromifa uses "Qube" which is written using the English alphabets in Latin writing style, he had the advantage of directly using the text news documents into the open text summarizer.

The latest local study is "Topic based Amharic Text Summarization with Probabilistic Latent Semantic Analysis" done by Eyob Delele and Dejene Ejigu in 2012. It investigates the problem of building a concept based single document Amharic text summarization system. The proposed algorithms are language and domain independent and hence can also be used for other local languages, more specifically it propose to use the topic modeling approach of probabilistic latent semantic analysis (Eyob & Dejene, 2012). However, the probabilistic topic model that is used in their approach do not use term weighting (Eyob & Dejene, 2012), which is highly important to identify the most representative words of the text document after removing stop words. They

used news articles and tested the system using six algorithms. The study claims that the result obtained is encouraging hence the systems gave improved results than previous summarization approaches.

This research is part of the effort that should be done to fill the gap in the area. It aims to investigate the applicability of the open source system called open text summarizer (OTS), to single document Amharic text summarization using news domain. In line with the problem statement this research tries to address the following research questions.

RQ1. What are the characteristics of Amharic language relevant for text summarization?

RQ2. How can the open text summarizer be used for Amharic news text summarization?

The open text summarizer is proved to be effective for more than 25 languages with highest evaluation scores (Rotem, 2003). It has also been proved that the summarizer works for one of the local languages, Oromifa (Girma, 2012), which is written using the Latin alphabets. This research, therefore, investigates the possibility of customizing OTS for Amharic which has its own writing system. OTS takes the most frequent words by removing stop words and gives high score for the first sentence of each paragraph. It also uses Porter stemmer which is 90% accurate (Rotem, 2003). Moreover the open text summarizer implements term weighting by giving higher score for the first sentence of each paragraph which will increase the effectiveness of the system in providing a well structured text summary.

## 1.3. OBJECTIVES OF THE RESEARCH

The main objective of this research is to investigate the applicability of the open source system, open text summarizer for Amharic news text summarization.

To achieve the above mentioned general objective of the study, the following specific objectives are set:-

- Review related works on text summarization.
- Review the characteristics of Amharic language and news writing techniques that are relevant to text summarization.
- Review text summarization techniques and algorithms developed for text summarization.

- Customize the open text summarizer.
- Collect/ develop the language resources (stemmer, stop word list. Abbreviations, synonyms list, affixes, etc).
- Collect Amharic news texts that are available in digital format for doing the experiment.
- Test the performance of the two customized text summarization systems.
- Draw conclusions based on the results and make recommendations for further research.

## 1.4. SIGNIFICANCE OF THE STUDY

This research is part of the effort to develop Amharic text summarizer which can produce a human quality summary of Amharic text to best satisfy users need. It can serve as an input for other researches to be done on text summarization for other Ethiopian languages especially for those that use the Geez alphabets, by changing the language lexicons and stemming rules. This research is done using news domain text documents. However, since the OTS tool is domain and language independent, the customized systems can also be used for other domain areas with few additions or modifications on the prepared stop word list and language lexicons. It can also be used as a source for application development for Amharic text summarization program.

## 1.5. SCOPE AND LIMITATION OF THE RESEARCH

This research is limited to single document Amharic text summarization for the news domain. It focuses on investigating the applicability of an open source system for text summarization on Amharic text documents called open text summarizer. The text documents used for testing the summarizer are sourced from online available Amharic news web pages only.

The study is not aimed to validate the applicability of the output of the study in other domains other than Amharic news text documents. It also does not aim to produce Amharic stemmer. Lack of standard corpus and standard ideal summary hindered to exactly measure the performance and compare it with other developed summarization systems.

## 1.6. ORGANIZATION OF THE THESIS

The organization of this thesis is as follows. In chapter one the general introduction, the statement of the problem, objective of the study, scope and significance of the research are presented. In chapter two, the major concepts related to automatic text summarization, local and global related works are discussed. In chapter three the Amharic language writing system is discussed. The alphabets, Amharic numeric, punctuation marks, words, sentence structure and news writing technique are described. In chapter four the methodologies used to conduct the research are discussed. In chapter five the tasks and concepts related to the implementation of Amharic OTS are discussed, which includes the description and requirements of the OTS, preprocessing, architecture and the summarization process are described. In chapter six, the experimentation using the OTS with the Porter stemmer and with Amharic stemmer, results of comparative subjective and objective evaluation measures are presented. Finally in chapter seven the conclusion from the findings of the research and the recommendations are presented.

# CHAPTER TWO

## 2. LITERATURE REVIEW

In this chapter literature that are relevant to this research are revised and presented in a summary. Moreover relevant concepts that are related to the research are also discussed. The chapter presents explanation on the types of summaries, stages in the summarization process, selection criteria for summarization, techniques of summarization, evaluation methods and different tools available in text summarization. There are many research works conducted on text summarization for foreign languages specially the English language in the past six decades. Among the few researches done on text summarization written in local language, all are done for documents written in the Amharic language except one which is done for Oromifa language. In order to avoid repetition, the local researches done on text summarization are not discussed in this chapter, hence can be referred on section 1.2 in chapter one.

## 2.1. TEXT SUMMARIZATION

The internet made a wide availability of information which is caused by many factors including the rapid digitalization of paper documents, rapid internet expansion, mass production of digital information, availability of different simple means of digitization and technological advancements in the production and dissemination of information. The abundance in global information raised a need for alternative ways of presenting the most important parts in a condensed form to the users in the form of a summary.

A summary is a text produced from one or more texts. It contains a significant portion of the information in the original text and is no longer than half of the original text (Hovy, 2002). The two requirements a summary must fulfill are:

i. It must be shorter than the original input text and
ii. It must contain the important information of the study, as defined by the user.

As discussed in the introduction part of chapter one of this research, there are two types of summaries based on the number of texts used as an input to the summarization process, Single document summary and multi document summary. A summary is single document

summary if a summary is prepared from one document only. And it is multi document summary when a summary about one topic is prepared from many different documents.

## 2.2 AUTOMATIC TEXT SUMMARIZATION

The process of preparing a text summary using a computer technique in the form of abstract or summary is referred as Automatic Text Summarization. There are many efforts done in automatic text summarization in the past six decades. From 1955 to 1979 simple extraction and linguistic approaches were applied (Basagic, Krupic, & Suzic, 2009). The inventor of IBM Hans Peter Luhn was the first to use computers for information retrieval in the 1950s. He developed many information retrieval applications including KWIC (Keyword in Context) using three fundamental information retrieval elements: keyword, title and context. New methods were introduced by Edmundson (1968) in automatic extracting of documents for screening purposes. In his paper in 1968 he described pragmatic (Cue) words, title and heading words and structural indicators as three additional components in automatic extraction. These components dominated the frequency components in the creation of better extracts. Edmundson (1968) also tried to develop an algorithm for automatic extraction of summaries from a corpus (Basagic, et al., 2009).

The second categories in automatic text summarization are the 1980's and 1990's which brought about the artificial intelligence approaches and renaissance (Basagic, et al., 2009). During this time the interest of researchers shifted to multi document summarization and extraction of summary from multimedia documents and a series of knowledge based text summarization systems also evolved. These systems use only informal heuristics to determine the salient topics from the text documents. The second generations of summarization systems adapted more mature knowledge representation approach which is based on the evolving methodological framework of hybrid classification based knowledge representation languages like SUSY, SCISOR and TOPIC (Basagic, et al., 2009).

The initial interest in automatic shortening of texts was started during the sixties in American research libraries where a large amount of scientific papers and books were to be digitally stored and made searchable (Hassel, 2004). Due to the limited available memory space the summary of the books has to be stored in the database and for those which have no summary

on the book itself, the automatic summarization was applied to create a summary. Early experimentation in the late 1950s and 1960s suggested that text summarization by computer was feasible though not straight forward (Luhn, 1958).

The growth in technology both in the hardware and software has contributed greatly to the production of huge amount of text documents. The World Wide Web coupled with the technology advancements initiated the concept of automatic summarization again to handle the huge corpus available digitally.

### 2.2.1 TYPES OF SUMMARIES

Effective summarizing requires an explicit and detailed analysis of context factors (Lloret, 2008). Sparck Jones (1998) distinguished three categories of context factors: input, purpose and output. Each of these context factors has three groups. The features of the input text to be summarized crucially determine the way the summary can be obtained. The input factors are categorized further into three groups as text form, subject type and unit; the purpose factors are categorized into three as situation, audience and use; the output factor has also three groups: format, material and style (Jones, 1998). Each of the group is discussed in the following sections.

1. **Input Characters:**

   The first group in the input factors is based on ***Text form/document structure*** of the input used, as *Genre Vs Scale*. The typical input ***Genres*** include novels, short stories, non-fiction books, newspaper articles, editorials or opinion pieces, progress reports, business reports and so on. The ***Scale*** may vary from book length to paragraph length. In this categorization different techniques may apply to some genres and scales and not on others (Hovy & Lin, 1998).

   The second group in the input characters is based on the ***Specificity or subject type*** of the domain of the input document which categorizes summaries as *domain specific Vs general*. ***Domain specific*** summarization derives summary from documents that have a theme related to a restricted single domain. It focuses on specific content and outputs specific formats. It can assume less term ambiguity, idiosyncratic word and grammar usage, specialized formatting, etc, and can reflect them in the summary

(Hovy & Lin, 1998). A general domain summarization on the contrary, derives summary from input texts in different domains. It makes no assumptions unlike the domain specific summary.

The third and last categorization in the input characters in based on the ***Unit*** or ***size of the input*** document used in the summarization which can categorize summaries into *single document Vs multi document summarization*. A ***single document*** summary is a summary that is derived from a single input document and a ***multi document*** summary is a summary that covers the content of more than one input documents and is usually used only when the input texts are thematically related.

## 2. Purpose Characteristics

The first group of the purpose characteristics category is based on ***expansiveness*** of the summary as *background Vs just the news* summary. A ***background*** summary assumes the prior knowledge of the reader on the general setting of the content of the input document is poor. Therefore, it includes explanatory material like circumstances of place, time and actors. A ***just the news*** summary contains just the new or principal themes and assumes that the reader knows enough background to interpret the summary in the context of the source document.

The second group of the purpose characteristics of summary usage is based on the ***readers*** of the summaries, which can be categorized as *generic Vs user focused*. ***Generic*** summaries do not target any particular group and does not focus on a special user need. Rather address a broad community of readers. It gives the author's point of view of the input document by giving equal value of importance to all major themes in the document. On the other hand ***user focused*** summaries are modified to the specific need of an individual or a particular group and represent only a given topic. It is also referred as topic focused or query oriented summary. It favors specific aspects of the document and is built on top of the previously submitted user query. It favors specific themes explicitly by highlighting pertinent themes or implicitly by omitting themes that do not match with the user's query. It is gaining importance due to the prevalence of full text searching and personalized information filtering.

The third group in the purpose characteristics of summaries is based on their ***function*** as *indicative, informative and critical*. ***Indicative*** summaries provide enough content to alert users to relevant source document so that users can select and read in depth. It gives an indication of the principal subject matter or domain of the input document excluding the contents. ***Informative*** summaries assemble relevant factual information in a concise structure and act as substitutes for the source document. From reading informative summary, one can explain what the input document is about but not what it contains. And ***critical*** summaries incorporate opinion statements on content and add value by brining expertise to bear that is not available from the source alone.

**3. <u>Output Characters</u>**

The first group in the output category of context feature is based on the characteristics of the summary as a text ***derivation/format*** which categorizes as *extract Vs abstract*. ***Extraction*** methods select the original pieces from the source document and concatenate them to produce a shorter text as a summary. It is mainly concerned with what the summary content should be (Dipanjan, 2007). Extracts can contain phrases and even paragraphs in addition to a complete sentence from the document. The drawback of this method is the lack of the balance and cohesion that is the sentences could be extracted out of the context and anaphoric references can be broken (Karel & Josef, 2007).

Whereas ***abstraction*** methods paraphrase in more general terms what the text is about. The sentences are built from the existing content using more advanced natural language processing methods after analysis of the input. It puts strong emphasis on the form aiming to produce a grammatical summary (Dipanjan, 2007). It is difficult for a computer program to successfully solve and provide solution to the requirements of such approach with many limitations including language generation and human language complexity. Most of the tools available today use the extraction method. However the summary produced using the extraction method may not be logically correct.

The second group in the output category is based on ***coherence/content-material*** which categorizes summaries as *fluent Vs disfluent* summary. In ***fluent*** summary the

sentences are related and follow one another according to the rules of coherent discourse structure. It is written in full grammatical sentences. Whereas, a ***disfluent*** summary is fragmented into individual words or text portions that are either not composed into grammatical sentences or not composed into coherent paragraphs (Hovy & Lin, 1998).

The third group in the output category applies principally when the input material is subject to opinion or bias based on ***partiality.*** It categorizes summaries as *neutral Vs evaluative*. ***Neutral*** summary reflects the content of the input text partial or impartial, where as an ***evaluative*** summary includes some of the system's own bias. This can be either implicitly through inclusion of material with one bias and omission of material with another or explicitly using statements of opinion.

The last group in the output category is based on ***conventionality/style*** which categorizes summaries as *fixed Vs floating*. A ***fixed*** situation summary is a summary that is created for a specific user, purpose or situation and conform to appropriate in-house conventions of highlighting and formatting. A ***floating*** situation summary on the other hand cannot assume fixed conventions but it is created and displayed in different settings to a various readers for different applications.

### 2.2.2. STAGES OF AUTOMATIC TEXT SUMMARIZATION

Researchers in automated text summarization have identified three distinct stages topic identification, interpretation or topic fusion and summary generation. Most systems today embody the first stage only (Hovy, 2002).

Topic is a particular subject to write about. Topic identification, therefore, produces the simplest type of summary. Once the system has identified the most important units, which might be words, sentences or paragraphs, it lists them creating extract or display them diagrammatically creating a schematic summary in whatever kind of criteria used. The next step interpretation is fusion of concepts, evaluation and other processing. The result of this stage is something new not explicitly contained in the input. It, therefore, requires the system to have access to knowledge separate from the input.

Given the difficulty of building domain knowledge, few existing systems perform interpretation, and no system includes more than a small domain model (Hovy, 2002). Systems include the third stage, generation to produce human readable text since the results of interpretation are usually unreadable abstract representations and extracts are incoherent due to omitted discourse linkages and repeated or omitted material. Each stage is discussed in detail below.

**Stage 1- Topic Identification**

There are different independent modules used to perform topic identification. A score is assigned to each unit of input text by each module. A combination module then combines the scores for each unit to assign a single integrated score to it. According to the summary length requested by the user, the system finally returns the nth highest scoring units. Most systems take a sentence as a scoring unit but taking sub-sentence size is more effective in producing summaries with more information.

The performance of topic identification modules is usually measured using recall and precision scores by quantifying how closely the system extracted summary corresponds to the human extracted summary. Precision shows how many of the sentences extracted by the system were good. It can be defined as the ratio of the number of correctly extracted sentences in both the system and human summary to the total number of sentences extracted by the human and the system. Recall on the other hand shows how many good sentences the system missed or failed to extract to include in the summary. It can be defined as the ratio of the number of correctly extracted sentences in both the system and human summary to the total number of correct sentences that are either extracted or not by both the human and the system.

Precision = Correctly extracted/Total extracted

Recall= Correctly extracted/ Total Correct

**Stage 2- Interpretation or Topic Fusion**

In this second stage of summarization the topics identified as important are fused, represented in new terms and are expressed using a new formulation with words that are

not found in the original document. This stage differentiates extract type of summarization from abstract type summarization systems. Since it must interpret the input text in terms of something extraneous to the text, no system can perform this stage without prior knowledge about the domain area of the document which is so difficult that summarizers to date have only attempted in small way. The difficulty of building such structures and filling them makes large scale summarization impractical.

Hahn and Reimer (2000) developed operators that condense knowledge representation structures in a terminological logic through conceptual abstraction. To date, no parser has been built to produce the knowledge structures from text, and no generator to produce language from the results (Hovy, 2002). Hovy and Lin (1998) used topic signatures to perform topic fusion. These signatures are sets of words and relative strengths of association in which each set is related to a single headword. By automatically constructing these signatures they overcome the knowledge paucity problem using thirty thousand texts from the Wall Street journal and TF*IDF to identify for each topic the set of words most relevant to it. They used these topic signatures both during topic identification and topic interpretation to score sentences by signature overlap and to substitute the signature head for the sentences containing enough of its words respectively.

However, the effectiveness of signatures to perform interpretations has not yet been shown and interpretation remains blocked by the problem of domain knowledge acquisition. This problem will have to get a solution before summarizers can produce abstracts.

**Stage 3- Summary Generation**

Summary generation is the last stage of summarization and extract summarization do not require generation stage. When the summary content has been created, it exists in internal notation within the computer. It, therefore, requires the natural language generation techniques like text planning, sentence planning and sentence realization. When extracted units are simply extracted in order of importance score or in text order, disfluencies like repetition of clauses, named entities and inclusion of less important material may result

which may be identified and repaired by a process of smoothing. The repair for repetition of clauses is to aggregate the material into a conjunction, for the repletion of named entities is to pronominalize and for inclusion of less important material is elimination.

### 2.2.3. AUTOMATIC TEXT SUMMARIZATION SELECTION CRITERIA

A text document contains different word categories like stop words, domain specific words, content bearing words, etc. Stop words are words that appear so frequently in a text document but they do not carry any message of the document. Whereas domain specific words are words that appear rarely and are also referred as 'read only once'. Content bearing words are the words which carry the most important part of the document. The summary that is extracted from the document has to contain much of the content bearing words and has to be more representative of the document. In order to select these words for extraction and include them in the summary, there are different criteria that are to be considered in automatic text summarization techniques. The seven major criteria are discussed below.

1. **Position Criteria**:– it is based on the fact that the position of certain text in a fixed position like heading, title, first sentences of a paragraph and first paragraphs contain important information depending on the text structure of different genres. The early studies by Luhn (1959) and Edmundson (1968) presented some of these positions like the first and last sentences to be important representative of the text. Taking the first paragraph as a summary often gives good performance especially with news paper articles. The first and last paragraphs of a text usually contain the introduction and conclusion parts of the document which contains highly content bearing important sentences. They can therefore be essential for extraction of good summary.

   In order to automatically determine the best positions and quantify their utility, Lin and Hovy (1998) define the genre and domain oriented optimum position policy (OPP). It is a ranked list of positions of content bearing sentences that on average produce the highest yields for extracts and describe an automated procedure to create OPPs for given texts and extracts. The training phase of their method calculates the product of each sentence position by comparing the similarity between the contents of the sentence in each ordinal position with the human created summary for a given text.

By the summation of large collection of text and abstract pairs of a given text and normalizing it appropriately, the OPP is created. In operation, the position module first selects an appropriate OPP for the input text. Then it assigns a score to each sentence in order of the OPP and computes the OPP keyword list for the text. It then assigns additional scores to sentences according to how many OPP keywords they contain (Hovy & Lin, 1998).

2. **Cue Phrase Indicator Criteria**:- sentences containing certain words and phrases that show importance in some genres should be extracted. Baxendale (1958) identified two sets of phrases which are bonus phrases which indicate if a sentence is likely to be included in a summary and stigma phrases which indicate that the sentence should not be a candidate to be included in the summary. Bonus phrases like in a summary, in a conclusion and superlatives such as: the best, the most important, can be taken as good indicators of significance content. The cue phrase module gives significance for sentences containing cue phrase with an appropriate score and sentences that contain stigma phrase are penalized, during processing. Identifying cue phrase is a major problem since they are genre dependent. Conclusion and abstracts, for example, are found in scientific literature than in newspaper articles. The common method is to identify high yield sentences in texts, compare them to the human made abstracts and identify common phrases found in both.

   Edmundson (1969) is the first to use this criterion for text analysis (Kamil, 2004). Edmundson (1969) also observed effects of stigma words that decrease the significance of a sentence for summary extraction and concluded that sentences containing them should not be extracted since they are not important. Using a manually built list of 1,423 cue phrases in a genre of scientific texts, Teufel and Moens (1997) reported 54 percent of joint recall and precision. They have assigned a goodness score which is positive or negative for each cue phrase manually and explained these cue phrases signal the nature of the multi sentence rather than single sentence rhetorical blocks of text in which they occur (Hovy & Lin, 1998).

3. **Word and Phrase Frequency Criteria**:- Using Zipf's Law of word distribution that states: "a few words occur very often, fewer words occur somewhat often and many words occur infrequently"; Luhn (1959) developed extraction criteria. The extraction criterion is: if a text contains some words unusually frequently, then sentences containing these words are probably important. However taking only the words with highest frequency makes the system to take stop words as important units for the extraction process. Luhn (1959) used fixed threshold to remove stop words which in effect increased the risk of losing content bearing words and inclusion of more stop words in the extraction process.

   A common method is to create list of stop words using TF*IDF measure. It rewards more frequent words in one text than on average across the corpus. This method was first used in information retrieval systems to achieve query term expansion by Salton (1988). This concept can be used in topic identification assuming that semantically related words co-occur. The frequency of the words that occur together can be counted rather than each word frequency, indicating the importance of its common semantic notion in the document. In order to implement this idea, topic signature was defined as a topic word or the head word together with a list of associated keyword together with a list of associated pairs (Hovy & Lin, 1998).

4. **Query and Title Overlap Criteria**:- to score each sentence by the number of desirable words it contains like those in the text's title or headings and in the user's query, is a simple but useful method. The query method is a direct descendant of IR techniques and is used for query based text summarization system. The sentences that contain more of the query words get higher score than those containing a single or less query words. The sentences with the highest score are selected to be included in the summary together with their structural context.

5. **Cohesive or Lexical Connectedness Criteria**:- repetition, co-reference, synonymy, and semantic association as expressed in thesauri are the various ways that words can be connected. Based on the degree of connectedness of their words, sentences and paragraphs can be scored; hence more connected sentences are assumed to be more important. Mani and Bloedorn (1997) represent the text as a graph in which words are

nodes and arcs represent adjacency, co-reference and lexical similarity. Cohesive methods are based on internal structure of the input text. This lexical cohesion arises from semantic relationships between words which allow the parts of the text to function as a whole. The connectedness of the words can be used for summary extraction through different techniques like:-

- **Word Co-occurrence:**- words that occur in common context can be related to each other.
- **Connectedness:**- the text structure is represented in terms of coherence and cohesion relations like proper name, anaphora, reiteration, synonymy and hyponymy. Using statistical metrics the salience of words and sentences is determined.
- **Co-reference:**- important words are identified in co-reference chains detected in sentences within the document and between query and document.
- **Local Salience and Grammatical Relations:**- key phrases can be identified in combination of grammatical, syntactic and contextual parameters.
- **Lexical Chains:**- occurs between pairs of words and over sequences of related words. Strong chains can be created using lexical relations which is determined from lexical database. These strong chains help to identify the important sentences.

6. **Discourse Structure Criteria**:- a sophisticated variant of connectedness involves producing the underlying discourse structure of the text and scoring sentences by their discourse centrality (Lin & Hovy, 2002). Using an algorithm to learn the optimal combination of scores from centrality, based on the shape and content of the discourse tree, Marcu's (1998) system does almost as well as human summarization for Scientific American Texts. Marcu's (1998) system is based on the fact that texts are not simply flat lists of sentences; they have a hierarchical structure, one in which certain clauses are more important than others (Hovy & Lin, 1998). After parsing the

hierarchical structures and identifying the important clauses, and kept the most important clauses discarding unimportant ones (Hovy & Marcu, 2005).

By adopting rhetorical structure theory which postulates approximately 25 relations that are signaled by cue phrases and binds clauses together if they have certain semantic and pragmatic properties, Marcu (1998) produced the text's discourse structure (Hovy & Lin, 1998). Most of the relations are binary and can be nested recursively. One of the binary relations is that having a principal component which is the nucleus and a subsidiary component which is the satellite. A text is only coherent if all its clauses can be linked together. The link is under a single relation, first in local sub-trees and then in progressively larger ones.

A constraint satisfaction algorithm is used by Marcu (1998) to assemble all the trees and employ several heuristics to choose among the trees. He then discards the least salient material by traversing the discourse structure top down to produce an extract summary. He combined the discourse structure paradigm with other surface based methods like cue phrases, discourse tree shape, title words, position method and so on. Using an automated coefficient learning method, he finds that the best linear combination of values for all these methods. Evaluation of the system on newspaper articles and on scientific American texts shows that the resulting extracts approximate human performance.

7. **Combination of Various Module Scores**:- Researchers have found that no single method of scoring performs as well as humans do to create extracts (Hovy, 2002). Combining methods improves scores significantly, since they rely on different kinds of evidence. There are different methods of finding a combination of function automatically, however there is no obvious best strategy. Kupiec, Pedersen and Chen (1995) trained a Bayesian classifier by computing the probability that any sentence will be included in a summary, given the features paragraph position, cue phrase indicators, word frequency, upper-case words and sentence length. In their experiment they found that the paragraph position feature gives 33% precision and the cue phrase indicators 29% individually but gave 42% when combined.

### 2.2.4. TECHNIQUES OF AUTOMATIC TEXT SUMMARIZATION

As discussed in the previous section, the classical method is based on the selection of statistically the most frequent words in the input document to be included in the summary. Another method is based on the position in the text, paragraph and so on. The other methods use outstanding parts of the text like titles, subtitles and leads with the assumption that sentences containing these words are better candidates for the summary. Methods based on entities are grounded on techniques allowing us to acknowledge linguistic units amount the mass of alphanumeric symbol strings which need lemmatizers, morphological parsers and part of speech taggers (Sharan, Jain, & Imran, 2008). For example the same string might belong to different parts of speech like 'plant as noun or verb', and different strings might be instances of the same part of speech as 'plant and planted'.

The techniques in automatic summarization are classified in three major families. These are: 1) Based on the surface where no linguistic analysis is performed; 2) Based on Named Entities in the input text where there is some kind of lexical acknowledgment and classification and 3) Based on discourse structures where some kind of structural processing (like linguistic) of the input document is required. The simple methods based on structure take advantage of the hyper textual scaffold of an HTML page. More complex methods using linguistic technology resources and techniques might build structure that will allow the relevant parts to be extracted. It is clear that when creating a text using fragments of a previous original, reference chains and, in general, text cohesion, is easily lost (Sharan, et al., 2008). A brief discussion of each method is presented below.

1. **Cut and Paste Method**

   It is a domain independent single document summarization method developed by Jing (2000). The most important sentences from the input document are extracted in the first stage, to give an output of a list of key sentences. These are edited by the cut and paste generation component by simulating two revision operations. The first one is sentence reduction which removes nonessential phrases from an extracted key sentence which gives a shortened version of the extracted sentence. It uses multiple sources of knowledge like lexical information, syntactic information from linguistic

databases and statistical information from corpus to enhance the selection of the sentence that has to be removed or has to be left in the extraction. The second one is the sentence combination module which merges the output of the sentence reduction module. Based on the phrases and sentences to be combined the rule based combination module can apply various combination operations. By using an automatic summary sentence decomposition program, the training corpora for the two revision operations is constructed.

**2. Extracting Sentence Segments for Text Summarization**

The method extracts sentence segments to produce a summary, as proposed by Chuang (2000). The method first breaks sentences into segments using special cue markers. Then the summarizer using an algorithm is trained to extract important sentences and finally a sentence is segmented by a cue phrase. In a complex sentence with two clauses, the main segment is called the nucleus and its subordinate segment is called satellite and is connected to the main segment by some kind of rhetorical relation (Sharan, et al., 2008).

**3. Lexical Chains for Text Summarization**

It describes lexical chain identification in a given text by extracting sentences to form summaries. It was suggested by Barzilay (1997), Chan (2005) and Lei (2007). A lexical chain is a set of semantically related words in a text. These relations between words are stored in WorldNet database and words that appear as nous in the WorldNet entry are chosen. Three kinds of relations are defined: 1) Extra strong relation that is between a word and its repetition, 2) Strong relation which is among two words that are connected by a WorldNet relation, and 3) Medium Strong relation which is when the link between the sunsets of the two words is longer than one.

Three steps are followed to construct lexical chains. The first one is to select a set of candidate words. Then the second one is to find an appropriate chain for each candidate word. And the third one is to insert the word in the chain and update it if it is found in the previous step. The strongest chains among those produced by the algorithm have to be identified in order to obtain the summary for a given input text. Then the full sentences are extracted from the original text.

**4. The Pyramid Method**

This is a new approach developed by Nenkova (2004) and is based on summarization content units (SCUs). When making a summary as different people include various information SCUs highlights what people agree on. These units are sub-sentential content units not bigger than a clause and are taken from the human made summary. It contains a label that shows the meaning of the content units in a concise sentence. The SCU is defined by identifying semantically similar snippets and grouping them into common label. Each SCU has a weight corresponding to the number of summaries in which it appears and the SCUs of the same weight are partitioned in a pyramid. Each pyramid has as many tiers as the number of the model summaries. The score of a summary is calculated as the ratio of the sum of the weights of its SCUs to the sum of the weights of an optimal summary with the same number of SCUs (Sharan, et al., 2008).

**5. Trainable Summarizer**

It was developed by Kupiec (1995) and also Jen Yuan (2005) and was influenced by Edmundson (1969) extraction system. It extracts the sentence based on a discrete feature set as sentence length cut off feature, fixed phrase feature, paragraph feature, thematic word feature and uppercase word feature. Let s be a particular sentence, S the set of sentences that are in the summary and F1, F2…, Fk the features. The probability of being included in the summary for each sentence is calculated based on the k given features Fj: j=1….k which can be expressed using the naïve - Bayes rule as (Sharan, et al., 2008):

$$P(s \in S \mid F_1, F_2 \ldots F_k) = \frac{P((F_1, F_2 \ldots F_k) \mid s \in S)\, P(s \in S)}{P(F_1, F_2 \ldots F_k)}$$

Assuming statistical independence of the features

$$P(s \in S \mid F_1, F_2 \ldots F_k) = \frac{\pi_{j=1}^{k} P(F_j \mid s \in S)\, P(s \in S)}{\pi_{j=1}^{k} P(F_j)}$$

Where P(s Є S) is a constant and the probability of a feature to be in a sentence given by $P(F_j | s Є S)$ and the probability of a given feature given by $P(F_j)$ can be estimated directly from the training set by counting occurrences.

This gives a simple Bayesian classification function that assigns for each sentence a score that can be used to select sentences. Paragraph feature, Sentence length cut off feature, fixed phrase feature and thematic word feature are the heuristics utilized in this method.

### 2.2.5. EVALUATION TECHNIQUES

Summaries are task and genre specific and so user oriented that no single measurement covers all cases (Hovy, 2002). Evaluating summaries and automatic text summarizations systems is not a straightforward process. A good extractive summarizer has to find the relevant information for which the user is looking as well as to eliminate the irrelevant one. It is then crucial to evaluate the system on the way it is able to identify how well it can extract the pieces of articles that are relevant to a user (Massih, Nicolas, & Gallinari, 2005).

Summary evaluation methods attempt to determine how adequate or how useful a summary is relative to its source document. There are two types of summary evaluations as defined by Sparck Jones and Galliers (1996), which are intrinsic and extrinsic. Intrinsic evaluation can be defined as measuring output quality only in which users judge the quality of summarization by directly analyzing the summary. Extrinsic evaluation, on the other hand, is defined as measuring user assistance in task performance. In this type of evaluation, users judge a summary's quality according to how it affects the completion of some other task. Most of the existing evaluations of summarization systems are intrinsic.

Donaway, Drummey and Mather (2000) showed how summaries receive different scores with different evaluation measures or when compared to different but equivalent ideal human summaries. By comparing several intrinsic and extrinsic evaluation methods on the same extract summaries, Jing and Mckeown (1999) found high consistency in the news genre for fixed length short summary. By calculating n-gram overlaps between systems generated summaries and human summaries, Recall Oriented Understudy for

Gisting Evaluation (ROUGE) is also used for summary evaluation. The high level of overlap indicates a high level of shared concepts between the human summary and the system generated summary. Extractive approach of summarization enables the use of precision, recall and F-measure as an evaluation measures for automatic text summaries.

During the evaluation of summaries and summarization systems, there are two properties of summary that must be quantitatively measured. One of these measure is the compression ratio (CR) which analyses how much shorter the summary is than the original input text. CR is calculated as the ratio of the length or summary to length of full text. The other measure is the retention ratio (RR) which analyses how much information is retained in the summary. RR is also referred as omission ratio and is calculated as the ratio of information in the summary divided by the information in the full input text. A good summary has small CR closer to zero and large RR values closer to one.

CR=Length of the summary / Length of the Input text

RR=Information in the summary / Information in the Input text

In the following paragraph the different kinds of evaluation techniques are briefly discussed based on the two summarization evaluation categorizations intrinsic and extrinsic evaluations.

#### 2.2.5.1 Intrinsic Evaluation

Intrinsic evaluation is done by creating an ideal summary for each of the input texts and comparing it with the summary of the summarizing system. It is done either by measuring the similarity of the content by sentence or phrase recall and precision measures or by simple word similarity (Pembe, 2010). Some evaluators use more than one measure and take the average of the scores obtained from each. However none of the measures are entirely satisfactory since there is no only one correct ideal summary for any given document. There are seven different intrinsic evaluation methods and are discussed as follows.

1. **Summary Coherence**

   The extraction based methods that use cut and paste operations on phrases, sentences or paragraph might result in a summary part extracted out of context which results in coherence problem. This can be measured by having humans rank summary sentences for coherence and compare the ranks with the scores for reference summaries or source sentences.

2. **Summary Informativeness**

   By comparing the system generated summary with the input text being summarized analyzing how much of information from the input is represented in the summary, infomativeness of the summary can be measured. Or in an alternative, the generated summary can be compared with a reference summary measuring how much information in the reference summary is present in the generated summary.

3. **Sentence Precision and Recall**

   Precision and Recall are standard measures for information retrieval in terms of sentences. Precision measures how many of the sentences in the system generated summary are in the reference human summary. And recall for sentences measures how many of the sentences in the human summary are extracted in the system summary. The main problems with these measures for text summarization is that they are not capable of distinguishing between many possible, but equally good summaries and that summaries that differ quite a lot content wise may get very similar scores (Hassel, 2004).

4. **Sentence Rank**

   Sentence rank is a more fine grained approach than precision and recall (Hassel, 2004). The sentences in the input text are ranked so that those with the highest rank could be worthy to be included in the construction of the reference summary. Then by using correlation measure the system generated summary can be compared with the constructed reference summary. Sentence rank is mainly applied to extraction based summaries.

5. **The Utility Method (UM)**

The UM allows reference summaries to have extraction units like sentences, paragraphs and so on with fuzzy membership in it. All the units of the input text and the confidence value for each of the units for their inclusion in the summary are kept in a reference summary. It can be further expanded to allow extraction units to apply negative support on one another by the units which are useful in evaluating multi document summaries. In the case of one sentence making another redundant, it can automatically penalize the evaluation score, i.e. a system that extracts two or more equivalent sentences get penalized more than a system that extracts only one of the aforementioned sentences and a less informative sentence like a sentence that has a lower confidence score (Hassel, 2004). Unlike precision and recall measures, this method allows summaries to be evaluated at different text compression rates and it is useful mainly for evaluating extraction type of summaries.

6. **Content Similarity**

This method can be applied to evaluate the semantic content in both extraction summaries and abstracts. The vocabulary test is one content similarity measure that uses standard information retrieval methods to compare term frequency vectors calculated over stemmed or lemmatized both extraction based summaries or true abstracts and reference summaries of the same sort (Hassel, 2004). Controlled thesauri and synonym sets created with latent semantic analysis or random indexing can be used to reduce terms in the vectors by combining the frequencies of terms deemed synonymous thus allowing for greater variation among summaries which is especially useful when evaluating abstracts (Hassel, 2004). These methods are very sensitive to negation and word order differences which can be mentioned as their disadvantage.

7. **BLEU Scores**

The concept in this evaluation method is that there may be many equally good summaries for a given source text document. They may be different in the unit choice or order of presentation even if they use the same unit. Lin and Hovy

(2003) have applied this method using automatically computed accumulative n-gram matching scores between ideal summaries and system summaries as performance indicator. Position independent content words were used in forming the n-grams and n-gram matches between the compared summaries. Direct application of this method, however, does not always give good results (Hassel, 2004).

#### 2.2.5.2 Extrinsic Evaluation

Extrinsic evaluation measures the efficiency and acceptability of the system summary to achieve some task. For a summary that has instructions, the extent of following the instruction and the result achieved can be measured. Information gathering in large corpus, the time and effort required to post edit the system generated summary for specific task, and summarization system's impact on a system of which it is part of, are some other possible measurable tasks. Measures like time spent and accuracy obtained during these tasks are used to assess the quality of the summary (Pembe, 2010). Though the extrinsic evaluation is easy to motivate, it has a problem. The major problem in is to ensure that the applied metric correlates well with the task performance efficiency.

There are many game-like scenarios that have been proposed as methods of evaluation of summaries such as the Shannon Game, the Question Game, the Classification Game and Keyword Association which are discussed below.

1. **The Shannon Game**

   It is the variant of Shannon's (1948) measures in information theory which attempts to quantify information content by guessing the next token to recreate the original text. Three groups of informants are asked to reconstruct important passages from the input text having seen either the whole text, generated summary or nothing from the text in each of the three tests. Using the number of wrong guesses wg and total guesses tg, the ratio R can be calculated as: R=wg/tg. The information retention is measured in number of keystrokes it takes to recreate the original passage (Hassel, 2004). This evaluation is dependent on the person doing the guessing. This makes it implicitly related to the reader's knowledge and

can be taken as a disadvantage of the evaluation. The measure increases with increased knowledge of the domain, language, and so on.

2. **The Question Game**

   It measures approximates the information content of the summary by determining how well it allows readers to answer questions drawn up about the input text (Hovy, 2002). The purpose of this evaluation is to test the readers' understanding of the summary and the summary's performance to convey key factors of the source article. Set of questions are created based on what is considered principal content of the text by one or more people before proceeding with the game process. It has two steps incurring out the evaluation task. First the testers read and mark key passages in the input text. Then they create questions that relate to the factual statements in the key passages they have identified. Assessors answer these questions three times: 1) without reading any related text, 2) after reading a system summary and 3) after reading the input text. At the completion of each round, the questions answered correctly are tallied and compared to compute the quality of the summary. A summary that includes the key concepts of the input text should enable to answer more questions better. That is, it will have closer score for both the summary and the input text.

3. **The Classification Game**

   In this evaluation classifiability of the system is compared by asking assessors to categorize the input text documents (testers) or the summaries of the text (informants) into one of N categories (Hovy, 2002). Then the correspondence of classification of summaries to source text document is measured. A good summary should be classified into the same class as its source.

4. **Keyword Association**

   This method relies on keywords associated manually or automatically to the input documents. It is a shallower evaluation approach and is inexpensive. A good summary contains the central concept of the text being summarized since the keywords associated to the article are content indicative. The main advantage of this evaluation is that it does not require bulky manual representation.

## 2.3 GLOBAL RESEARCH

Research in ATS begun in the 1950s, Luhn (1958), following the surface level approach and in the 1960s, by Edmondson (1969), following entity level approach. In the 1970s the first commercial applications were developed and discourse based approaches begun to be used. Artificial intelligence concept based entity level approaches begun to appear in the 1980s. All the three approaches were taken into consideration in the 1990s both for research and commercial purpose. Research on multi-document, multilingual and multimedia summarization systems started during this decade. The highly grown public access to texts on the web and better evaluation programs for summarization systems are the great impacts of the 2000s. There are many automatic text summarization systems developed either for commercial purpose or found online for free. The major systems are presented as follows.

The first automatic text summarizer is SweSum, which is developed for Swedish language that summarizes Swedish news text in HTML format on the internet. It is a system to system comparison against a model summary and based on statistical, linguistic and heuristic methods. The system calculates the frequency of the key words in the text, in which sentences they appeared, and the location of these sentences in the text. It is mainly a Swedish language automatic text summarizer, also supports plug in lexicons and heuristics for other languages which includes English, Danish, Norwegian, German, Spanish, French, Italian, Greek and Farsi languages.

The other major system is Summarist, which is an attempt to develop robust extraction technology and then continue research and development of techniques to perform abstraction. It produces extract summaries in five languages. The MEAD summarizer is also a summarization system available now in the public domain. It is developed for query based application. The relative ease of adapting scripts and modules to customize it to new summarization system can be taken as the advantage of Perl. The open text summarizaer is among the major smmarizers. It is an open source tool for summarizing text. Since it is used in this research, it is discussed in detail in chapter 5.

Kify online text summarizer uses text semantic indexing and some mathematics. The summarizer decides which parts of a document are important by analyzing the content. Automatic text summarizer is also another online text summarization tool. Intellexer

Summarizer 3.1 is a commercial summarization tool and is an innovative program for computers that it creates a short summary from any document or a browsed web page. Pertinence summarizer is also a commercial summarization tool that performs linguistic processing over a document and evaluates the pertinence of its sentences. Quicklist summarizer 1.2 is another summarization tool that is capable of highlighting summaries of web pages directly on the browser windows.

## 2.4 AUTOMATIC TEXT SUMMARIZATION FOR LOCAL LANGUAGES

There are quite a few researches conducted locally at the school of graduate studies, department of information science at the Addis Ababa University. These researches are briefly discussed in chapter 1 on section "1.2 statement of the problem".

# CHAPTER THREE

## 3. AMHARIC WRITING AND MORPHOLOGY

In this chapter the Amharic language writing and morphology are discussed. The Amharic alphabets, numbers, punctuation marks are briefly explained. The word classes, formation and morphological structure of Amharic words are discussed. Phrases and sentence structures and types are also described. Finally the different ways of writing Amharic news are discussed.

## 3.1. INTRODUCTION

Amharic is one of the "Ethio-Semitic" languages belonging to the Semitic branches of Afro-Asiatic languages and is related to Hebrew, Aramaic, Arabic and Syrian. It is the second most populous Semitic language next to Arabic and has its own writing alphabets and numeral symbols. Amharic has been a written language for roughly 600 years (Adane, 2011). A wide variety of Amharic literatures including books, religious writings, fiction, poetry, plays, and magazines are available both in printed and machine readable format (Solomon, 2011).

Like other Semitic languages such as Arabic, Amharic exhibits a root pattern morphological phenomenon and also it uses different affixes to create inflectional and derivational word forms (Martha, 2010). The Amharic word classes are nouns, pronouns verbs, adjectives, adverbs, prepositions, and conjunctions (Mersehazen, 1934). Interjections and numerals are also part of the Amharic writing and are also discussed in this chapter. The Amharic language has phrases like verb phrases, adjectival phrases, adverbial phrases and prepositional phrases (Mesfin, 2001).

## 3.2. AMHARIC WRITING SYSTEM

Writing system is the use of symbols to represent the sound units of a language systematically (Daniel, 2003). The writing system for the Amharic language was adopted from the Ge'ez writing system, which belongs to the class of the Semitic language family and used to be the literature language of Ethiopia in earlier times (Bender & Hailu, 1978). The

widely used Ethiopic alphabet, which is derived from the writing system of ancient South Arabian inscriptions, is used for Ge'ez, Amharic, Tigrigna and other Semitic languages (Solomon, 2011). The Amharic writing system includes different characters for alphabets, numbers and punctuation marks (Abiyot, 2000), which are discussed below.

### 3.2.1 AMHARIC ALPHABETS

The Amharic alphabet called "Fidel", has thirty three characters which occur in one basic form. Each of these characters has six additional forms referred to as orders which follow regular pattern of vowel usage. Each of these orders uses a consonant and a vowel to represent a phoneme, which makes Amharic syllabic writing system as explained by Bender (1976). Hence such languages require as much number of characters as the basic sounds used in the language, thus Amharic has 233 different characters and nearly 40 more other characters. Most labialized consonants are basically redundant, and there are actually only 39 context independent phonemes (mono-phones); of the 275 symbols of the script, only about 233 remain if the redundant ones are removed (Asker, Atelach, Gambäck, & Sahlgren, 2007).

The 233 Amharic characters are comprised of all the alphabets from Ge'ez language and additional some new alphabetic characters that represent sounds not found in Ge'ez like "ቾ, ጨ, ጇ, ኘ, ቨ, ሸ, ኸ, and ዠ" (Solomon, 2011). The 40 additional characters contain special feature representing labialization like "ጕ/ gwe" from "ገ/ gä", "ቋ/ qwe" from "ቀ/ qä", "ኗ/ luä" from "ላ/ lä" and "ጓ/ guä" from "ገ/ gä" (Bender, 1976). However only around 20 of the 40 more characters are common and are listed as an appendix on the character list called "Yefidel Gebeta". Each symbol in the alphabet represents a consonant together with its vowel (Bethlehem, 2002).

There are seven vowels in Amharic alphabets "ኧ/ä, ኡ/u, ኢ/i, አ/a, ኤ/e, እ/_, and ኦ/o" which are based on their point of articulation grouped as Peripherals "ኡ/u, ኢ/i, አ/a, ኤ/e, and ኦ/o" and Central vowels "ኧ/ä and እ/_" which are mostly used than the peripherals. The other idea worth mentioning is that two consonants can appear in the middle or at the end of words in a cluster whereas clusters at the beginning of the word are very restricted in which "እ/ _" is used.

### 3.2.2 AMHARIC PUNCTUATION

There are 17 punctuation marks in the Amharic writing system (Abiyot, 2000). However, only few of them are practically used, especially in Amharic computer writing system (Theodros, 2003). For example the punctuation that separates words in Amharic text "ሁለት ነጥብ/ hulet netib" which looks like a colon '፡' is hardly used in current Amharic writings being replaced by an empty space. The sentence separator for Amharic text writing is four dots arranged in a square sequence as '።' and is referred as "አራት ነጥብ/ arat netib". The comma equivalence punctuation in Amharic is "ነጠላ ሰረዝ/ netela serez" which is symbolized as '፣' used to separate lists. The equivalence of the semi colon which is used to separate phrasal lists or compound sentences is referred as "ድርብ ሰረዝ/ dereb serez" which is denoted by the symbol '፤'. Punctuation marks like the question mark '?' and the exclamation mark '!' are borrowed from the English language and used in Amharic language for the same purpose as they are used in other foreign languages (Bender, 1976). The lists of punctuation marks that are used in the Amharic language which is adopted from (Ager, 1998-2013) is presented in Table 3.1 below.

| Punctuation Mark | Description | | Punctuation Mark | Description |
|---|---|---|---|---|
| ፣ | Comma | | ፧ | Question mark (no longer used) |
| ። | Full stop | | ፡ | Word Separator (Empty space is mostly used now) |
| ፤ | Semi-colon | | « » | Quotation Mark |
| ፥ | Colon | | ! | Exclamation Mark |
| ፦ | Preface colon | | ? | Question Mark |

**Table 3.1 Punctuation marks used in Amharic language**

### 3.2.3 AMHARIC NUMBERS

The number system in Amharic writing has 20 single characters which represent numbers: ones (1/፩ up to 9/፱), tenths (ten/፲ to ninety/፺), hundred (፻) and ten thousand (፼). These characters are derived from Greek letters and are modified to look like the Amharic

characters by adding a horizontal line on top and bottom of each character (Bender, 1976). The Amharic number system, however, does not have any representing symbol for zero value and it does not use any decimal points and commas. As a result arithmetic computation is very complicated using the Amharic number system. It is mainly used in calendar dates to show the dates in the Ethiopian calendar. Using the Amharic suffix "-ኛ/ -ägna" on the cardinal numbers, their ordinal number equivalents are formed like "ሁለት/ hulät"/ 'two' forms "ሁለተኛ/ hulätäna"/ 'second'. There are numerals that are used to indicate distribution and partitions of something whole like "ሁለት ሁለት/ hulät hulät"/ 'two two' and "ግማሽ/ gemash"/ 'half', "ሲሶ/ siso"/ 'one third', "እሩብ/ rub"/ 'quarter' (Daniel, 2003).

## 3.3. AMHARIC WORD CLASSES

There are different researches done to study the linguistic characteristics of Amharic which categorize the words differently. Mesfin (2001) grouped the researches in to two as early and recent works. The early works such as Mersehazen (1934) categorized the Amharic words into the following eight parts of speech (POS): Noun, Verb, Adjective, Adverb, Preposition, Pronoun, Conjunction and Interjection (Daniel, 2003). Whereas recent works like (Baye, 1998) classified the words under five parts of speech: Noun, Verb, Adverb, Adjective and Preposition. The pronouns and conjunctions are included under noun and preposition categories, respectively. Interjections are not considered as part of speech at all, since they are words without syntactic functions (Daniel, 2003).

Any Amharic word can be categorized in any one of the above recent classes. However one has to consider the context of some words to properly categorize them into the word classes. For example the word "አበበ/ abäbä" which means "it flourished" can be either a proper name and categorized as a noun or a verb representing its meaning in the context of the sentence it is used. Other names like "አየለ/ Ayälä", "ከበደ/ Käbädä", "በቀለ/ Bäkälä", etc. can also be taken as an example with the same feature. Prefixes like "በ/ bä", "ስለ/ s_lä", etc. can be considered as both conjunctions and prepositions.

According to Williams (1981) there are two methods that can be used to identify lexical classes of a word to solve the problem of the prefix categorization. One is by Morphological

analysis which considers the type of suffixes a given word takes. Amharic nouns, for example, do not take "-ኦች/ -äch" as a suffix while perfective verbs do. The other one is by Syntactic analysis which is considering the position of a word in a given sentence. From the sentence structure of the Amharic language we can see that verbs do not come at the beginning of a sentence as the first word and nouns do not come at the end, provided that the sentence is a declarative sentence (Baye, 2000 E.C.).

### 3.3.1 NOUN

These are words used to identify or name a class of people, places, things, ideas, etc. or any member of these classes (Daniel, 2003). It usually consists of common pronouns and proper nouns that usually take the "-ኦች/ -och" and "-ne" as a suffix for plural and for accusative case marker. A noun can be identified either morphologically if it can take the above suffix or syntactically if it comes after demonstrative pronouns or modified by adjectives. Amharic nouns can be either primitive or derived. Primitive nouns are words that exist in their original form. And derived nouns are nouns that originated from different state or completely different categories. For example, the noun "ጎረቤት/ goräbet" which means 'neighbor' can be taken as primitive noun and "ጎረቤትነት/ goräbetnät" which means 'being neighbor' can be taken as derived noun.

### 3.3.2 VERB

These are words that usually come at the end of a declarative sentence which is grammatically complete in Amharic language. This is an important property of the Amharic verbs (Baye, 2000 E.C.). In order to grammatically agree with the subject of the sentence, they usually take subject maker suffixes as "-ሁ/ -hu" for 'I', "-ህ/ -h" for 'He', "-ች/ -ch" for 'She', and so on.

### 3.3.3 ADVERB

Adverbs refer to places, time, circumstances, etc. of the action which is represented in the verb of the sentence (Daniel, 2003). In Amharic, adverbs that are found in their primitive form are very few and they do not attach any kind of affixes. Therefore, phrases that are combinations of prepositions and some other words like nouns (e.g. "በኃይል/ bä-hay_l" which means 'by force') or prepositions and adjectives (e.g. "በደህና/ bädähna" which

means ‘well’) perform the functions of adverbs in the language. Moreover there are adverbs of position which are known as Noun Adverbs that function either as a noun or as an adverb depending on the context they are used in (Daniel, 2003). For example, the word “ውስጥ/ w_st” in “ወደ ውስጥ ገባ/ wädä w_st gäba” which means ‘He entered inside’ is used as a noun while in “በውስጥ ይሰራል/ bä-w_st yäsäral” which means ‘He works in’ it is used as an adverb (Atelach, 2002). The most common adverbs in Amharic include “ገና/ gäna” which means ‘still’, “ቶሎ/ tolo” which means ‘quickly’, “እንደገና/_ndägäna” which means ‘again’, “ክፉኛ/ k_fugna” which means ‘severely’ and “መናኛ/ menagna” which means ‘foolish’ (Getahun, 1998).

### 3.3.4 ADJECTIVE

Adjectives in Amharic sentence come before nouns and serve as modifiers of nouns which follow them. For example, consider the sentence “እሱ በጣም ጎበዝ ልጅ ነው/ _su bätam gobäz l_j näw”/ ‘he is a very cleaver boy’, where there are two adjectives “በጣም/ bätam”/ ‘very’ and “ጎበዝ/ gobäz”/ ‘cleaver’. The adjective “በጣም/ bätam”/ ‘very’ modifies the adjective “ጎበዝ/ gobäz”/ ‘cleaver’ and the adjective “ጎበዝ/ gobäz”/ ‘cleaver’ in turn modifies the noun “ልጅ/ l_j”/ ‘boy’ (Tesfaye, 2002). Just like nouns, adjectives also can be primary which are very small in number or derived.

### 3.3.5 PREPOSITIONS

These are words that will have meanings only when they are attached to another word or are used together with other word classes like nouns, verbs and adjectives. The Amharic prepositions includes affixes like “ከ/ kä”, “ለ/ lä”, “የ/ yä”, “በ/ bä”, “ስለ/ slä” and so on which stands for ‘from’, ‘for’, ‘that or of’, ‘with’ and ‘about or for’, respectively. They are used to form adverbial phrases that appear preceding nouns and take various forms. They can appear as simple preposition that stands alone as a word or as a prefix to a word. For example, in the phrase “ስለ ሙዚቃ/ s_lä musica” which means ‘about music’ the preposition “ስለ/ s_lä” appears as a word. Whereas, the preposition “በ/ bä” appears as a prefix to the noun “ፈረስ/ färäs”/ ’horse’ in the phrase “በፈረስ/ bä-färäs” which means ‘by horse’ and as a prefix to a verb in the phrase “በመጣ/ bä-mäta” which means ’wish he comes’. However, when they are prefixed to verbs they form different types of subordinate clauses that are typical components of complex sentences (Daniel, 2003).

Prepositions can also appear as compound prepositions consisting prepositional prefixes and post positions with a noun in the middle as in "ከአባቱ ጋር/ kä-abatu-gar" which means 'with his father'.

### 3.3.6 CONJUNCTIONS

Both subordinating and coordinating conjunctions are included in this class and are used to harmonize words, phrases, clauses and sentences. The major examples are "ወይንም/ wäyn_m"/ 'or' and "እና/ e_na"/ 'and'. Some words or affixes like "በ/ bä", "ስለ/ s_lä", "ከ/ kä" might be used both as conjunctions and prepositions which creates a problem to take the two classes as separate POS. however it can be resolved by carefully analyzing the context of the words used.

### 3.3.7 INTERJECTIONS

These are words or phrases that can stand alone or appear anywhere in a sentence. They are used to express emotions like anger, sudden surprise, happiness, annoyance, pleasure, satisfaction and so on (Atelach, 2002). "ጎሽ/ Gosh!"/ 'Good job!', "ዋ/ wa"/ which is used as 'becareful', "ውይ/ w_y"/ 'ouch' can be mentioned as examples.

## 3.4. AMHARIC WORD FORMATION

Amharic words are believed to have a Mono-. Bi-, Tri-, Quadric-, Quiniqui- or Sexi- radicals consonantal roots that carry their semantic information or meaning (Bender & Hailu, 1978). For example: "ša"/ 'desire' which is Mono- radical; "ተው/ täw"/ 'leave' which is Bi-radical; "ሰብር/ säbr"/ 'break' which is a Tri- radical; "ገልብጥ/ gälb_t"/ 'turn over' which is a Quadric-radical; "አሽቀንጥር/ -š_känätt-r-" / 'throw away violently' which is a Quinqui- radicals (Tesfaye, 2002). However, like most Semitic languages, tri-radical verbs are the most common basic types in Amharic language.

According to Mullen (1988), there are three processes in Amharic word formation. 1) The creation of verbal stems which are formed by infixing vocalic melodies into consonantal roots. 2) The production of fully inflected word forms from stems is the next process. This formed verbal stem goes through inflectional process for different grammatical contexts like

gender, tense, number and so on. 3) Cliticization in which fully inflected verbal stems attach Clitics on the word (Mullen, 1988).

The following example shows the three steps in word formation (Tesfaye, 2002):
Consonantal root: "ስብር/ sbr"/ 'break';
Stem formation: "ስብር/ sbr" + "ኧ/ ä" = "ሰበር/ säbbär-"/ 'broke'
Inflection: "ሰበር/ säbbär-" + 2nd person, singular, masculine ("ክ/ -k") = "ሰበርክ/ säbbärk"/ 'you broke'
Cliticization: "ሰበርክ/ säbbärk" + 1st person, singular, object maker ("-ኝ/ -änn") = "ሰበርክኝ/ säbbärkänn"/ 'you broke me'.

The Amharic language has a highly complex morphology (Nega & Willett, 2003). This is due to the different morphological and phonological processes involved with the formation of words. However, observing the pattern in the morphs makes it easy to understand the morphological structure of the language. Different affixes are used in the creation of inflectional and derivational word forms. However, from the Amharic parts of speech verbs, nouns and adjectives are important and undergo both derivation and inflection. Most function words in Amharic, such as conjunction, preposition, article, relative marker, pronominal affixes, negation markers are bound morphemes, which are attached to content words resulting in complex Amharic words composed of several morphemes (Sisay, 2005). These morphological patterns are discussed in the following sections.

### 3.4.1 DERIVATIONAL MORPHOLOGY

The derivational morphological process is the formation of words from words that belong to other parts of speech. As discussed in section 3.4 of this chapter, Amharic has one up to six radicals or roots that are the set of consonants that carry the basic information or meaning. A pattern which is also referred as the vocalic element consists of vowels which are inserted among the consonants of the root (Martha, 2010). According to Bender (1976) affixes are combined with this pattern to form a single grammatical form, which also could be another stem as described by Baye (2000 E.C.). Usually the form taken by the stems of a particular class is indicated by a sequence of Cs and Vs, called template like CVCCVC, CCVC and CVCC which represent the perfective, jussive and

imperfective stems of a tri-radical verb respectively (Martha, 2010). The derivation of verbs, nouns and adjectives is discussed below.

1. **Verbs** – can be either simple verbs which are derived from roots by intercalating vowel patterns or derived verbs which are derivatives of simple verbs. It is not common to derive verbs that carry the basic lexical meaning from other part of speech. However, verbs can be derived from nouns by taking the consonants only that is intercalating vowel patterns and taking them as a root word (Anbessa & Hudson, 2007). For example, the verb "መገበ/ mägäbä"/ 'he fed' is derived from the noun "ምግብ/ m-g-b"/ 'food'.

   The penultimate radical of a verbal stem where gemination pattern takes place, is very important part of the verb. Based on the gemination patter of this radical, verbs are grouped into three types. In type A verbs, the penultimate radical geminates in perfect tense only; in type B, it geminates irrespective of the verb forms; and in type C, it geminates in both perfect and imperfect verb forms (Martha, 2010). Except in type C verbs which use the vowel "a" after the first radical, in all Amharic verbs the vowel "ä/ ä" is intercalated among the consonants. The appearance of any other vowel in the root of a verb shows the reduction of weak radicals and glides. The use of "አ/ a" vowel in root forms shows the reduction of the consonants "ሀ/ h" and "ዀ/ ?", "ኦ/ o" vowel shows the reduction of "ወ/ w" and "እ/ e" vowel shows the reduction of "ይ/ y". When weak radicals "ወ/ w, ይ/ y, ሀ/ h,ዀ/ ?" and "ብ/b and ር/r" appear in verbal roots it shows existence of a vowel "ä/ ä".

2. **Nouns** – can be either primitive nouns which are words that exist in their original form or derived nouns which originate from different state or completely different categories. Nouns can be derived from stems, other nouns, adjectives or infinitive form of a verb by adding affixes or intercalation.

   i. To derive nouns from other nouns the following affixes are used. Suffixes: "-ነት/ -nät", "-ኛኛ/ -_gna", "-አት/ -ät", "-ተኛ/ -tägna", "-ኛ/ -gna", "-አዊ/ -awi" and prefix: "ባለ-/ balä-".

   ii. To derive nouns form adjectives, the suffixes used are: "-ነት/ -nät" and "-ኛት/ -ät".

iii. To derive nouns from verbal roots is complex than the others which involves intercalation and affixation. The intercalation of the vowel "I" after the first or among the root consonants derives a noun which can also be a bound morpheme that can form other nouns by adding affixes. Agent nouns are derived using "ኧ/ä-ኣ/a" pattern and "I" as a suffix; Nouns of manner can be derived by prefixing "ኣ-/ a-" to the stem which is formed by duplicating the penultimate radical and intercalating the pattern "ኧ-ኣ-ኧ/ ä-a-ä"; Infinitive/ Verbal noun is derived by prefixing the morpheme "መ-/ mä-" to the jussive verb stem; and the instrumental noun is derived by suffixing "-እያ/ -iya" to the infinitive (Martha, 2010).

iv. The last way of forming nouns is by the compounding words either without the use of any morpheme as in the word "እንጀራ እናት/ _njära _nat"/ 'stepmother' which combines "እንጀራ/ _njära"/ 'Ethiopian bread' and "እናት/ _nat"/ 'mother' or using the morpheme "ኧ/ ä" as in the word "ቤተ መጻሐፍት" betä mätsähaft"/ 'library' which combines "ቤት/ bet"/ 'house' and "መጻሐፍ/ mätsähaf"/ 'book'.

3. **Adjectives** – just like the formation of nouns, adjectives can be derived from stems, nouns or verbal roots by adding affixes and by intercalation.

i. To derive adjectives from nouns the following suffixes are used: "-ኣም/ -am", "-እኛ/ -_gna", "-ኣዊ/ -awi" and "-ኣማ/ -ama".

ii. To derive adjectives form verbal roots there are two ways; one by intercalation of the vocalic elements "ኧ-ኣ/ ä-a", "ኧ-ኧ/ ä-ä", "ኧ-ኢ/ ä-i" and "ኡ/ u" to the root consonants and the second is attaching the suffix "-ኣ/ -a" to bound stems.

iii. By compounding other parts of speech like nouns and other adjectives using the compounding morpheme "ኧ/ ä", adjectives can be formed.

### 3.4.2 INFLECTIONAL MORPHOLOGY

1. **Verbs** – according to Baye (2000 E.C.) verbs are inflected for person, gender, number, tense, aspect and mood. Amharic verbs are grouped in to two based on

aspect as perfect and imperfect. The following table summarizes the subject marker morphemes that are used on verbs.

| Inflectional Marker for | Subject Marker Morpheme for | |
|---|---|---|
| | **Perfective Verb** | **Imperfective Verb** |
| 1st Person | "-ኩ/ -ku" or "-ሁ/ -hu" | "እ-/ _-" |
| 2nd Person | "-ክ/ -k" or "-ህ/ -h" | "ት-/ t-" |
| 3rd Person | "-ኧ/ -ä" | "ይ-/ y-" |
| Gender | "-ህ/ -h" or "-ሽ/ -sh" | "-ኢ/ -ei" |
| 1st Person Plural | "-ን/ -n" | "-ን/ -n" |
| 2nd Person Plural | "-ኡ/ -u" | "-ኡ/ -u" |

**Table 3.2 Subject Marker Morphemes**

The third person subject marker morpheme "-ኧ/ -ä" always comes after stem verbs which might be followed by object marker morphemes. When both subject and object markers come together the order is stem - subject marker -object marker. These object markers can be direct or prepositional like "ለ-/ lä-" and "በ-/ bä-". For example "አመጣልኝ/ amätal**l**gn"/ 'he brought for me' and "አመጣብኝ/ amätal**b**gn"/ 'he brought on me, without my will' shows the use of the propositional object markers.

Amharic verbs can show four moods which are declarative, interrogative, negative and imperative. On the other hand, the Amharic tense can be categorized as past and non-past. The Past Amharic tense is of three types. 1) Simple past which indicates that an action is completed in the past without specifying the time of completion; 2) Recent past which is formed by using the auxiliary verb suffix "-አል/ -al" on a stem to show a completive aspect of the verb; and 3) Remote past is formed when the auxiliary verb "ነበር/ näbär"/ 'was' is used.

The Non-Past Amharic tense in turn is formed by using the auxiliary verb suffix "-አል/ -al" attached with the imperfective stem. Due to the time ranges from present to future in these kinds of sentences, the tense of the verb can be either of the two. This can be clearly presented by using adverbs of time in the sentence. The use of the prefix "እየ-/ _yä-" attached to a verb indicates a continuous tense either past or present.

2. **Nouns** – in Amharic nouns are inflected for case, number, definiteness and gender marker affixes (Martha, 2010). Three cases for nouns are identified: nominative which is identified by its location in a sentence since it always precedes accusatives and the subject markers attached on the verb; accusative which is formed using the suffix "-ን/ -n"; and genitive which is formed using the case maker suffixes shown in Table 3.3 below or using the prefix "የ-/ ye-".

| Person | Singular | | Plural |
|---|---|---|---|
| | Vowel Ending | Consonant Ending | |
| 1st | -ዬ/ -ye | -ኤ/ -e | -ኣችን/ -achn |
| 2nd | | | |
| Masculine | -ህ/ -h | -እህ/ -Ih | -ኣችሁ/ -achu |
| Feminine | -ሽ/ -sh | -እሽ/ -Is | |
| Polite | -ዎ/ -wo | -ዎ/ -wo | |
| 3rd | | | |
| Masculine | -ው/ -w | -ኡ/ -u | -ኣቸው |
| Feminine | -ዋ/-wa | -ዋ/ -wa | |
| Polite | -ኣቸው/ -achaw | -ኣቸው/ -achw | |

**Table 3.3 Case Marker Suffix for Nouns (Martha, 2010)**

Amharic numbers are of two types singular and plural and the suffixes used to inflect nouns for numbers are "-ኦች/ -och", "-ኣን/ -an" and "-ኣት/ -att". In Amharic some words use the plural form of the Geez language, which is another Semitic language spoken mostly in the Ethiopian Orthodox church mostly in the northern part of Ethiopia. The plural forms of the Amharic words inherited from the Geez language are formed either by vocalic changes or reduplication. For example the plural form of the noun "አምላክ/ amlak"/ 'god' is "አማልክት/ amalkt" and "አንበሳ/ anbässa"/ 'lion' is "አናብስት/ anabest" which are formed by vocalic changes and the plural form of the noun "ቁስ/ qus"/ 'thing' is "ቁሳቁስ/ qusaqus" which is formed by reduplication.

Amharic nouns can be grouped as definite and indefinite where the latter is sometimes expressed using the numeral "አንድ /and"/ 'one' since there is no special indefiniteness marker. However, bound morphemes are used as definiteness markers for nouns. These are "-ኡ/ -u" for masculine nouns ending with consonant and "-ዋ/ -w" those ending with vowels. Whereas for feminine nouns the morphemes used are "-ዋ/ -wa", "-ኢቱ/ -itu" and "-ኢትዋ/ -itwa" or "-ዋ/ -wa", "-ዩቱ/ -yitu" and "ዩትዋ/ -yitwa" for

those ending with consonant and vowel, respectively. For plural nouns the definiteness marker morpheme used is "ኡ/ -u" by coming after the plural marker of the noun. However, when an adjective comes before a definite noun, the definiteness marker is attached on the adjective.

Two genders are represented in Amharic that differentiates in the second and third person. These are masculine which has no gender marker morpheme and Feminine for which the gender marker morphemes "-ኢት/ -it" and "-ኢቱ/ -itu" are used. Additional gender markers for nouns are also used such as the definite article "-ው/ -w" for masculine and "-ዋ/ -wa" for feminine; demonstrative pronoun like "ይህ/ yeh" for masculine and "ይህች/ yehech" for feminine; the verb which refers to a noun by using the suffix "-ል/ -l" for masculine and "-ች/ -ch" for feminine; and gender specifying words like "ወንድ/ wänd", "ተባዕት/ täbaät", "አውራ/ awra" and "አባ/ aba" as in "አባወራ/ abawära " indicates masculine and "ሴት/ set", "እንስት/ _n_s_t", and "እማ/ ema" as in "እማወራ/ emawära " indicate feminine gender in a sentence.

3. **Adjectives –** just like the way nouns inflect, adjectives also inflect for case, number, definiteness and gender. Since the same inflectional character is observed there is no need to repeat what have been discussed above. Please refer to the discussion under inflection for nouns.

## 3.5. AMHARIC PHRASE CATEGORIES

A Phrase is constructed from a combination of two or more words that convey a complete meaning in a given language. There are five Amharic phrase categories: Noun Phrase (NP), Verb Phrase (VP), Adverbial Phrase (AdvP), Adjectival Phrase (AdjP) and Prepositional phrase (PP) (Baye, 2000 E.C.). Based on the word class a given phrase contains, each of these phrase categories can be of two types: simple and complex. Simple phrases are those that are constructed from one word class and Complex phrases are those that are constructed from two or more word classes.

### 3.5.1. NOUN PHRASE (NP)

In Amharic phrases, NP is a phrase where the head word is a noun or a pronoun. The simple NP contains a single noun or pronoun and a complex NP contains a noun referred as head and other constituents which can be complements, specifiers, adverbial and adjective modifiers (Baye, 2000 E.C.). Let us, for example, take the NP "ያ የትናንቱ ትልቅ የሳር ቤት/ ya yät_nan_tu t_l_k yäsar bet" which means "that yesterday's big thatched house". In this phrase "ya"/ 'that' is a specifier, "የትናንቱ/ yät_nan_tu"/ 'yesterday's' is an adverbial modifier, "ትልቅ/ t_l_k"/ 'big' is an adjectival modifier, "የሳር/ yäsar"/ 'thatched' is also an adjectival modifier specifying the material from which the object named by the "ቤት/ bet"/ 'house' is made of and "ቤት/ bet"/ 'house' is the head noun (Daniel, 2003). In this NP the word "የሳር/ yäsar"/ 'thatched' is also referred as a complement of the head noun which makes it not to be treated like the rest of the modifiers. For the above example we can formulate the grammar rule as: NP_Specifier - AdvP – AdjP which can be rewritten as NP =NP-N.

### 3.5.2. VERB PHRASE (VP)

Verb phrase is a phrase where the head is a verb and contains other constituents like complements, modifiers and specifiers (Atelach, 2002). For example, consider the phrase "በፈረስ ወደ ሃገሩ ገባ/ bä färäsä wädä hageru gäba" which means '(he/it) entered to (his/its) country by horse'. In the phrase "በፈረስ/ bä färäsä" / 'by horse' is a prepositional phrase that modifies the verb to show manner, "ወደ ሃገሩ/ wädä hageru"/ 'to (his/its) country' is also a prepositional phrase that modifies the verb to show place and "ገባ/ gäba"/ 'entered' is a verb. The grammar rule for the given example is: VP=PP- PP- V.

### 3.5.3. ADVERBIAL PHRASE (AdvP)

Adverbs in Amharic are used in many constructs such as indicators of degree, in location phrases, to show the manner in which something is done, the time of something, or the frequency of something, etc (Atelach, 2002). Adverbial phrase is composed of one or more adverbs as a head. Consider the following example: "በኃይል ተዋጉ/ bä_hayil täwagu"/ 'they fight strongly'. The word: "በኃይል/ bä_hayil"/ (strongly) is the head adverb which form the adverbial phrase and "ተዋጉ/ täwagu"/ 'they fight' is a verb. The grammar rule

for the example is AdvP_Adv-V. If the phrase contains two adverbs like "ፈጠን ፈጠን/ fätäne fätäne"/ 'quickly quickly', the grammatical rule is: AdvP=Adv-Adv.

### 3.5.4. ADJECTIVAL PHRASE (AdjP)

Adjectival phrases are composed of head adjective and constituents like complements, modifiers and specifiers. Consider the phrase "ያ በጣም ኃይለኛ ልጅ/ ya bätam hay_legna lij"/ 'that very strong boy' as an example. The component "ያ/ ya"/ 'that' is a specifier, "በጣም/ bätam"/ 'very' is a modifier that modify the level of the head adjective "ኃይለኛ/ hay_legna"/ 'strong' and "ልጅ/ lij"/ 'boy' is a noun. The grammatical rule for the phrase is: AdjP_Spec-Adv-Adj-noun.

### 3.5.5. PREPOSITIONAL PHRASE (PP)

Prepositional phrases are composed of a head preposition and constituents like nouns, noun phrases, verbs, verb phrases, and so on. Consider the phrase "እንደ ወፍ በሰማይ/ endä wäf bäsämay"/ 'like a bird on the sky' for example. In the phrase, "እንደ/ endä"/ 'like' and "በ/ bä"/ 'on' are both prepositions that form prepositional phrases by joining with the nouns "ወፍ/ wäf"/ 'bird' and "ሰማይ/ sämay"/ 'sky' that follow each of them. The grammatical rule for the given phrase in the example is: PP_PP-PP, since the whole PP is a combination of two PPs which are constructed from preposition and noun PN_P-N. The structure, however, is dependent on whether the preposition is a prefix or a suffix. For example, consider the phrase "ወረቀት ላይ/ wäräqät lay"/ 'on paper' where the preposition "ላይ/ lay"/ 'on' comes after a noun "ወረቀት/ wäräqät"/ 'paper' which makes the structure PP_N-P.

## 3.6. AMHARIC PREFIXES ON VERBS IN CLAUSE FORMATION

Amharic has a number of prefixes and suffixes that attach on stems in a concatenating manner and even two or more of affixes could also come together. Amharic verbs thus take affixes in either end before or after the verb as prefix or suffixes. Various suffixes on verb stems are usually subject and object makers and represent grammatical categories of nouns like number, gender, aspect, person and tense. Based on the function they perform Amharic verbal affixes can be categorized as clause markers, negative markers, subject markers or the imperfect stem and derived verb prefixes (Daniel, 2003). Clause marker prefixes are used in

clause formation as indirect object indicator “ለ-/ lä-”/ ‘to’, relative clause marker “የ/ yä-”/ ‘’s’, reason clause “ስለ-/ s_lä-”/ ‘for’, time clause “እስከ-/ eskä-”/ ‘until’ and conditional clause “ከ-/ kä-”/ ‘from’.

## 3.7. AMHARIC SENTENCE STRUCTURE

A sentence is a set of words or phrases that combine to form a bigger phrase that convey a complete meaning. Amharic the basic characteristic of the syntactic structure of Amharic is Subject (S) – Object (O) – Verb (V) which gives it a characteristic that modifiers in Amharic generally precede the words they modify (Atelach, 2002). A sentence usually is a combination of two phrasal structures a noun phrase and a verb phrase that follows it. All the remaining phrase types are found within one of these two main phrasal components. Based on the number of verbs that are used in the sentence construction, sentences can be classified in to two groups as simple sentences and complex sentences.

### 3.7.1. SIMPLE SENTENCES

Simple sentences are those that contain a noun phrase as the subject of the sentence and a verb phrase which contains the predicate. For example, consider the sentence “ያ ቤት ትልቅ ነው/ ya bet tilk näw” which means ‘That house is big’, where “ያ ቤት/ ya bet”/ ‘that house’ is a noun phrase and “ትልቅ ነው/ tilik näw”/ ‘is big’ is a verb phrase. When the morphemes like “-ዉ/ -wu” comes with a verb, it refers to definiteness and indicates the number of the object and subject of the sentence (Daniel, 2003). The morpheme that indicates the subject always precedes the morpheme that indicates the object. For example, in the sentence “ሐና ልጁን አመጣችው/ hanna lijun amätachw”/ ‘hanna brought the boy’, the morpheme “-ች/ -ch” indicates “hanna” the subject is a girl and the morpheme “-ው/ -w” indicates the object “ልጁን/ lijun”/ ‘the boy’ of the sentence.

According to Baye (2000 E.C.), simple sentences are classified in to four groups as declarative sentences, interrogative sentences, negative sentences and imperative sentences (Baye, 2000 E.C.).

1. Declarative sentences are used to show feelings and ideas that the constructor of the sentence reflects about physical things, mental feelings, real or imaginary things, etc.

2. Interrogative sentence are those that question about the subject, the complement or what action the verb specifies in sentence using Amharic pronouns like "ስንት/ s_nt"/ 'how many', "ምን ያህል/ men yahel"/ 'how much', "ማን/ man"/ 'who', "የት/ yät"/ 'where', "ምንድን/ m_nden"/ 'what' and "መቼ/ mäche"/ 'when'. These interrogatives can further be joined with prepositions to give interrogative prepositional phrases. Example: "ከ/ kä"/ 'from', "ለ/ lä"/ 'to', "ለምን/ lä_mn"/ 'why', etc.

3. Negative sentences are created when a declarative statement that was made about something is negated. The prefix morpheme "አል-/ al-" when used on verbs of the sentence, it makes it negative. For example, in the sentence "ሳሙኤል አልመጣም/ Samuel al_mätam"/ 'Samuel did not come', "አል-/ al-" makes the negation on the verb "መጣ/ mäta"/ 'he comes' and the suffix "-ም/ -m" is used to indicate the gender and number of the subject on the negated sentence which is masculine third person singular.

4. Imperative sentences have a second person pronoun as a subject which is represented by the suffix on the verb of the sentence. Usually imperative sentences communicate instructions to the subject of the sentence.

### 3.7.2. COMPLEX SENTENCES

Complex sentences are sentences constructed using any combination of the simple or complex phrases like noun phrase, verb phrase or adjective phrase but not all of the phrases pattern can be of simple phrase type in a sentence. They are used to convey complete ideas like simple sentences but unlike simple sentence they are composed of complex phases (Atelach, 2002).

## 3.8. AMHARIC WRITING STRUCTURE

So far we have seen the structure of Amharic language parts and their construction to form a sentence. In this section the style of Amharic writing are discussed.

Proper Amharic writings have "አርእስት/ ar_st"/ 'Title', "መግቢያ/ mägbia"/ 'Introduction', "ሃተታ/ hatäta"/ 'Body or Detail' and "መደምደሚያ/ mädämdemia"/ 'Conclusion' parts. Except the title, the rest are presented in one or more paragraphs "አንቀጽ/ ankäts". Each paragraph usually discusses one principal idea in detail. The principal idea "ኃይለ-ቃል/ hailä-kal" may be

found either in the first, middle or last sentence of the paragraph. Based on the location of the principal idea, there are three kinds of paragraph writings. 1) Upright triangular paragraph - is when the principal idea is located at the first sentence of the paragraph and the discussion follows. 2) Inverted triangular paragraph - is when the principal idea is located at the last sentence of the paragraph after the discussion is made. 3) Diamond shaped paragraph type - is when the principal idea is stated at both the beginning and ending sentences and the discussion in the middle sentences of the paragraph.

News articles tend to put the most important information in the beginning to grab the attention of readers and a good ending. A good ending gives a rounded out feeling to a story especially if the end makes reference to information in the lead. Amharic news articles then have a title which is a sentence or two and different paragraphs that are introduction, body and conclusion. They may or may not have subtitles and encourages the use of the title words in the body since the news article concentrates on the title. News articles are usually written in the third kind, Diamond shaped, writing style. The structure follows uniform pattern which encourages the use of location heuristics in extraction of representative elements.

# CHAPTER FOUR

## 4. RESEARCH METHODOLOGY

In this chapter the methodology used to understand the concept behind, customize the system, test corpus preparation, implementation of the system to Amharic news article summarization and evaluate the output of the systems are discussed. The type of research methodology followed is Quantitative method of research. To achieve the objectives of the research the following methodologies have been used in each major tasks involved in doing the study.

### 4.1 LITERATURE REVIEW

Relevant literature that focus on text summarization especially open source system of summarization, automatic text summarization, Amharic text summarization and Amharic language writing techniques specifically news writing techniques, Amharic stemming rules and other useful documents were reviewed.

### 4.2 AMHARIC LANGUAGE LEXICON GATHERING

The OTS tool requires a dictionary file to be prepared for each language and saved in .xml format as a rule for processing by the tool in preparing the automatic summary. The Amharic language lexicon is therefore prepared by gathering from different articles, books and online sources as explained in chapter 5 of this thesis. The prepared list includes punctuation marks, numbers, synonyms, stop words, abbreviations, compound words, affixes, same sounding alphabets and list of words inherited from foreign language that are written in different forms. The Amharic synonym list is prepared using systematic random sampling, by taking two words from each alphabet in Amharic. In addition to the stop words gathered from different articles, 43 more stop words are extracted from the news corpus using a tool written in Python programming language that counts the frequency of occurrence of words.

### 4.3 CORPUS PREPARATION

To prepare the Amharic news corpus different Amharic news websites were checked. Forty news articles are gathered from the Ethiopian Press Agency – Addis Zemen newspaper, Walta Information Center and The Reporter – Amharic newspaper web pages. The selected articles are from different news categories that include economy, politics, sport, social life,

health and entertainment. From these news articles thirty are selected for the purpose of this study. The selection is made by taking those having sentences from 30 to 127, so that it is not too short or too long but can be relatively compared with the manual summary. Then the thirty articles are grouped in to three, based on the number of sentence each contains, as: short article (SA), medium article (MA) and long article (LA). Each of these groups contains 10 articles. The selected articles are preprocessed which includes removing tables, pictures, web links, changing the font types to Geez and saving each article in a .txt file.

## 4.4 CUSTOMIZATION OF THE OTS

The customization process of the OTS for Amharic news article summarization has three stages: requirement analysis, system design and implementation. Each of these are discussed as follows.

### 4.4.1 Requirements of the system

Since the open text summarizer tool is a language independent tool, the major requirements of the tool are the language lexicons that are incorporated in to the dictionary file. These lexicons are the list of Amharic punctuation marks, abbreviations, compound words, list of Amharic synonyms, list of Amharic stop words and Amharic affixes. Moreover, the writing rules for parsing and tokenization are also identified for Amharic and integrated into the OTS tool.

### 4.4.2 System Design

The OTS is an open source tool and does not have an installable format. It however has a simple interface that runs on C# or C++. For this research, the C# version is downloaded. Visual studio 2010 is installed to run and customize the tool. First the user interface is customized by adding some additional features and Amharic labels. Then the code is made to include the Amharic language writing rules like end of sentence mark, "።". Then the tool customization process is continued in two sessions for each of the two experiments carried in this research. For the first experiment the OTS is customized to include the Amharic language writing rules and all the prepared lists mentioned above in section 4.2 in the rules list inside the dictionary file. For the second experimentation, an

Amharic stemmer designed by Tessema (2007) is integrated in to the OTS tool in place of the Porter stemmer which makes use of the dictionary file.

### 4.4.3 Implementation

The customized two systems using the two different stemming algorithms are implemented separately. Using these systems, 90 summaries are extracted from the 30 selected news articles in 10%, 20% and 30% extraction rate. A total of 180 automatic summaries were extracted for the 30 articles which are separately saved and evaluated using the techniques discussed in section 4.7 of this chapter.

## 4.5 MANUAL SUMMARY PREPARATION

For the purpose of evaluation an ideal summary is prepared manually by experts in the news domain based on a guideline prepared to rank the sentences of a given article. Four experts were involved in this process, one from WIC, one from FBC and two from AAU journalism masters of art (M.A) students. The guideline for the summary preparation is composed from (Melese, 2009) and (Girma, 2012). The summary is prepared on different compression rate at 10%, 20% and 30%. This is used during the evaluation process to measure the performance of the customized systems by comparing the automatically generated summary with the corresponding manual summary. The same ideal summary is used in both experiments to evaluate the summaries generated by the systems.

## 4.6 SUMMARIZATION TECHNIQUE AND TOOLS USED

As described in the literature review part of this paper, there are different approaches to text summarization. In this research a single document text summarization using extraction technique is applied. The Windows version of the open source summarization tool, open text summarizer, written in C# which was modified by (Gohr, 2013) is used in this research. The original tool was developed by (Rotem, 2003). Since the windows version of the OTS is designed in C#, visual studio 2010 is installed and used to run the tool. The two experiments were conducted using different stemmers: the Porter stemmer inside the OTS and Amharic stemmer developed by (Tessema, 2007). In order to conduct the objective evaluation of these experiments, a simple evaluation tool that implements F-m is developed. For the table and graph preparations MS-excel and for the report writing MS-word are used.

## 4.7 EVALUATION TECHNIQUE

Once the customization of the Open Text Summarizer tool is finalized and automatic summaries are generated for Amharic text documents, evaluation of the output of the automatic Amharic text summarizer is done in two ways. The first one is subjective evaluation which involves checking whether the automatic summary makes sense linguistically, grammatical correctedness, coherence and structure. For this evaluation a guideline is prepared for the evaluators to follow while evaluating the summaries. The evaluation was done by two experts and the average of the two scores is taken. The second evaluation is done by comparing the automatically generated summaries with the manually prepared summaries and calculating the standard effectiveness measure of evaluation, F-measure. The evaluation tool designed using C# is used to evaluate the summaries objectively F-m as a performance measure. The score is given in percentage from hundred for each summary. The evaluation observations are discussed and conclusions are drawn based on the findings.

# CHAPTER FIVE

## 5. IMPLEMENTATION OF AMHARIC OPEN TEXT SUMMARIZER

In this chapter, the open text summarizer and its customization to support Amharic text summarization are discussed. The open text summarizer (OTS) is an open source tool which can be downloaded for free and used for different purposes related to summarization. The description of this tool, the procedures it follows in the summarization process, its effectiveness in summarizing text documents, preprocessing involved to make it work for Amharic language, the process based system architecture and the summarization process are discussed in this chapter. In order to accomplish the task of summarizing news text document, the tool has different requirements that must be specified in an ".xml" file format. These requirements and the procedures involved with their preparation are also presented in this chapter. The tool is available in a C# library format and has a simple user interface which is modified using visual studio 2010 C# programming language. Finally the description of the summarization process involved with the execution of the OTS for Amharic news articles is explained step by step.

## 5.1 INTRODUCTION

The Open Text Summarizer is chosen for this research as a tool to make automatic summary of Amharic news articles that are gathered from Walta Information Center, Ethiopian News Agency and Reporter news paper. The OTS which is an open source summarizing tool is customized to support the Amharic text documents by incorporating some features in the ".xml" file to make the Amharic text summarization more effective and efficient. These modifications include: providing the rule based list of Amharic stop words, abbreviations, synonyms, as well as stemming and punctuation rules. The OTS lacks documentation written in booklet, journal article, book or even "help file" form. A brief description about OTS is available only at http://libots.sourceforge.net/ and the readme file of the OTS library function, which are both used as a source for the description about OTS in this paper.

## 5.2 DESCRIPTION OF OTS

### 5.2.1 What Is OTS?

The Open Text Summarizer is an open source tool for summarizing text documents that is available on the internet and can be downloaded for free. The free software foundation gives unlimited permission to copy, distribute and modify the OTS library functions (Dom Lachowicz & Rotem, 2008). The original OTS is a language independent tool designed for the UNIX platform and only supports the ".txt" file format. According to Rotem (2003), OTS ships with Ubuntu, Fedora and other Linux versions. For Windows a modified version of OTS source code is available in visual C++ and visual C#. For this experimentation, the OTS written in visual C# is used. The original OTS does not have a standard user interface but it can be designed by a user (Yatsko & Vishnyakov, 2007), hence the modified tool by Gohr (2013) has a simple interface that runs on visual studio.

OTS is both a library and a command line tool where word processors like Abiword and KWord can link to the library. The command line tool lets to summarize text on the console. The original OTS supports summary extraction for 25 languages and also it works with UTF-8 code (Rotem, 2003). However, the dictionary folder of the web interface version of the OTS tool now contains ".xml" files for 37 languages, which includes 12 more languages than the original version which are not mentioned on the webpage. In order to add a new language, a dictionary of stop word list containing up to 150 or more words has to be prepared and included in the "/dics" folder of the OTS tool.

The web-interface version [1] of the OTS supports 37 different languages and 9 summarizing ratio percentages (Gohr, 2013). It has a better user interface than the original OTS where a user can copy and paste a text to be summarized in the input space provided or insert a URL link of the location of the text to be summarized. It also gives the option for the output to be either in a summary or keyword listed form which is displayed to the user in an output box on the same page with efficient speed. This page is tasted for English documents during this experimentation and it was found to be effective and satisfactory in producing summaries for English text documents.

[1] Available on the URL http://www.splitbrain.org/services/ots

### 5.2.2 How OTS Works?

Since the tool is language independent, it uses different general rules to produce a good summary for a given text. The major assumption in the OTS is that the important ideas in a text are described with many of the same words whereas redundant information uses less technical terms and is not related to the main subject of the article (Rotem, 2003). Sentences that are related to the subject of the article are considered to be important where the subject of the article is the list of ideas that are most discussed in the article. In addition to statistical word frequency and sentence position methods for sentence grade scoring, it incorporates NLP techniques via language specific lexicons with synonyms, abbreviations, punctuations, stop words and rules of stemming and parsing for a given language.

OTS uses sentence extraction mechanism considering the key word frequency and sentence position methods. It uses statistical word frequency methods to score sentences that are beforehand parsed and incorporates an English language lexicon with synonyms and cue terms (Boudin, Torres-Moreno, & Velázquez-Morales, 2008). This, however, can be modified for another language that is designed to use this tool. OTS does not work on books because they cover too many topics, however each page may be fed individually, this will produce a good result (Rotem, 2003). Moreover, because of their writing structure and content, OTS cannot produce a good summary for poems and other kind of literatures which are not standard text documents.

The size of the summary generated by OTS is determined by the percentage of the sentences in the input text document. Once the summarization process begins, OTS does not supply the user with an opportunity to input or change the vocabulary (list of the key words) on which the summarization is based (Yatsko & Vishnyakov, 2007). In the summarization process, the program reads a text document and decides which of the sentences are important to represent the document and which are not. The most representative sentences are then selected for the summary of the text which can be either printed in a separate text document or in HTML form where the sentences that are important are highlighted in different color.

The tool scans an article and parses the words using the “parser”, list of rules that are listed in the dictionary file. The parser rules include rules for “line break” to indicate the end of a sentence and “line don’t break” to indicate the rules of punctuation that does not show the end of a sentence. It includes another list of rules to handle compound words. The list under the “depreciate” makes compound words to be considered as one word by deleting the space in between and consider them as one word. Based on these rules the tokens of the text document are created. These tokens are then matched with the stop word list in the dictionary file which helps to remove all non content bearing tokens from the list, since they are not representative of the subject of the text document.

The list of tokens after the removal of stop words are passed to the stemmer, which using the stemming rules of the language listed in the dictionary, derives the stem of each token. The stemmer rules include the removal of punctuation marks from the list of the tokens using the “pre” and “post” list of punctuations that are used or attached before and after the tokens. It also includes list of synonym words which show a list of words corresponding to another word which are similar words in meaning and only one of each could be used in the token list.

This is not done in the English OTS but abbreviations of words and their expanded form can also be listed in the “synonym” rule list, so that the expanded form of each of the abbreviations are used in the token list. The major lists in the stemming rules are the list of affixes which are used to extract the stem of a word by removing suffixes and prefixes. The replacing characters during the removal of some of the affixes are also listed corresponding to the affix to be removed. Words that are exceptions to stemming rule are also listed under the “manual” list of rules with their corresponding stem word.

The stemmed words are then used to count the term frequency for each word. After calculating term frequencies, the tool stores the list of all words corresponding to their frequency of occurrence separately. This list is sorted in a descending order of frequency of occurrences. From this list, high frequency words are taken as key words that are more representative of the document. Then grades are given to the sentences that contain these key words, those having many of the key words get higher grades. The sentences are then ranked based on their grade for importance to be included in the summary. Finally, those

sentences having highest grades are selected based on the percentage of summarizing the text document. That means, to produce 15% summary, for instance, the top 15% of the total sentences with high grade score are taken.

According to Rotem (2003), Porter stemming is about 90% accurate for the English language which increased the effectiveness of the OTS. Stemming makes the summarization process more effective by enabling the OTS to consider the stem of the different word forms. That is, it helps to take the stem word for all the different derivatives of that word and consider them as the stem word in the sentence where the derivatives are found. So that those sentences containing the derivatives could also get score of containing the key words.

### 5.2.3 Effectiveness of OTS

The quality of a summary can be interpreted in terms of its Coherence, Informativeness, Non redundancy, Referential integrity and Relevance. Coherence is the connectedness of the summary sentences. Informativeness is the extent of the key ideas in the manual summary that are being included in the machine summary. Non-Redundancy is avoiding repetition of ideas using different sentences. Referential Integrity shows the ease of identifying the referential words, that is when reading the summary in the ranked order of the sentences, those sentences that have referential words like pronouns and noun phrases should be easy for the reader to identify to whom or what they are referring to in the sentence. Relevance is the extent of the key ideas in the source text that are being covered in the automatic summary.

In the first evaluation method by Yatsko and Vishnyakov (2007) OTS has the best performance compared with other popular summarizing tools, as show in Table 5.1.

| Name of Summarizing Tool | R 5% | R10% | R15% | R20% | R25% | $D_{Score}$ |
|---|---|---|---|---|---|---|
| ESSENCE (ESS) | 8.75 | 10.38 | 11 | 10.267 | 9.3158 | 49.707 |
| Subject Search Summarizer (SSS) | 14.5 | 14.5 | 13.64 | 12.533 | 12.263 | 67.432 |
| COPERNIC (COP) | 8.75 | 7.25 | 6.364 | 6.333 | 6.2632 | 34.960 |
| Open Text Summarizer (OTS) | 15.75 | 14.88 | 13.64 | 13.133 | 12.158 | 69.552 |

**Table 5.1 Relevance (R) and Quality (D) (taken from (Yatsko & Vishnyakov, 2007))**

In the second evaluation method used by Yatsko and Vishnyakov (2007) measuring efficiency of the summarizing tools, OTS scored **100%** with more refined testing criteria. While ESS, SSS and COP scored 71%, 97% and 50% respectively. In their finding OTS is more effective with higher quality for small summaries like 5% extraction. They also found that the reference dictionary methodology which is used by the OTS proved to be flexible, cheap and effective in producing summary for text documents.

## 5.3 REQUIREMENTS OF THE SYSTEM

In order to customize the OTS for Amharic text summarization, the list of Amharic punctuation marks, abbreviations, compound words, list of Amharic synonyms, list of Amharic stop words and Amharic affixes for the stemmer are required to be integrated to the dictionary files of the OTS library functions. These required components are discussed briefly as follows.

### 5.3.1 List of Amharic Punctuation Marks

There are many unique punctuation marks in the Amharic language. However, quite few of them are used in the computer writing system. Those that are never used in the computer system include "፡-, ፥, ፧, ፨, ※". The detail about punctuation marks is discussed in chapter three, section 3.2.2. The list of punctuation marks which includes marks like "፡, ።, ፣, ፤, «, », ?, !", is included in the dictionary rules, so that they can be removed from the text document. And the rules involved with these punctuation marks are used in tokenization of words and parsing of sentences in the summarization process.

### 5.3.2 List of Amharic Abbreviations

The languages which use a 'dot' to indicate the end of a sentence, suffer from the effects of abbreviations written using 'dots' to separate the characters that represent the abbreviation of a word or a phrase. This creates conflict when identifying sentence boundaries. Since Amharic uses four dots "አራት ነጥብ" as "።" for sentence separator, it does not suffer from this type of ambiguity. However, list of abbreviations need to be included in the dictionary to increase the word frequency of occurrence that is used in the identification of key words and stop word removal.

Some words which are representative of a given document might be presented in abbreviation and in normal form in the same text. As a result they will be counted separately as different words during the word frequency count. The list provided is used to identify such words in Amharic to consider them as one when they appear in either abbreviation or normal form in the word frequency of occurrence count process. The list of abbreviations which is presented in appendix IV of this paper is taken from (Melese, 2009) and (ELSRC, 2001 E.C.). This list containing 49 abbreviations and their corresponding expanded form, are introduced to the dictionary under the "synonyms" rule list, since it is used to show the similar form of one word in short and expanded form. In the English XML dictionary of the OTS, this is listed under lines don't break to prevent the abbreviating dots from indicating end of sentence. The list of some of the most common Amharic abbreviations is presented in Table 5.2 as follows.

| Abbreviation | Expanded Form | | Abbreviation | Expanded Form |
|---|---|---|---|---|
| ፕ/ት | ፕሬዝዳንት | | አ.አ. | አዲስ አበባ |
| መ/ቤት | መስሪያ ቤት | | እ. ኤ. አ. | እንደ ኤሮፓውያን አቆጣጠር |
| ት/ቤት | ትምህርት ቤት | | ዓ.ም. | ዓመተ ምህረት |
| ክ/ከ | ክፍለ ከተማ | | አ.ሕ. | አፍሪካ ሕብረት |

Table 5.2 List of some of the Amharic Abbreviations

### 5.3.3 List of Amharic Synonyms

Different words could be used to express the same thing or concept in a text document. Synonyms are words or phrases that mean exactly or nearly the same as another word or phrase in the same language. Using list of synonyms helps to identify those words which have the same meaning and used interchangeably. By using one selected word and identifying its synonyms as the same word; the word frequency count is increased for the selected word. This in effect helps words to have higher value of frequency of occurrence so that they could be used as best candidate representative words of the text document. The list of Amharic synonyms is therefore prepared by randomly selecting words that are more related to news writing from Amharic dictionary (ELSRC, 2001 E.C.) and annexed list on the poem book (Yohanes, 1990 E.C.). The words are selected from the dictionary using a statistical random sampling, by taking two commonly used words from each

Amharic alphabet. A list containing a total of 91 pairs of words and their synonyms is prepared. This list is then added to the Amharic dictionary file of the OTS. The list of some common Amharic synonyms is shown in Table 5.3 and the whole list that was prepared for this study is presented in appendix V.

| Word | Synonym | | Word | Synonym |
|---|---|---|---|---|
| ቀድሞ | በፊት | | አቢሲንያ | ኢትዮጵያ |
| መምህር | አስተማሪ | | አንጋፋ | ታላቅ |
| ምግባር | ባህሪ | | መሰናዶ | ዝግጅት |
| ክንውን | አፈፃፀም | | ዳኛ | ፈረደ |

Table 5.3 List of Amharic Synonym words

### 5.3.4 List of Amharic Stop Words

Stop words are non content bearing words that do not convey the conceptual or meaningful part of the text document. They, therefore, cannot be taken as representative words for a given text document. To generate a generic summary, nonstop words that occur most frequently in the document may be taken as the query words, since these words represent the theme of the document (Gupta & Lehal, 2010). Common words with no semantic and which do not aggregate the relevant information, are eliminated using the stop word list.

There is no standard list of stop words prepared for Amharic that could be directly used as an input list. It is even difficult to have a list that exhaustively includes all Amharic stop words, since stop words can be either domain dependent or independent. The domain independent stop words can be identified from few text documents in the language but the domain dependent ones are extracted by considering many text documents in one domain and taking those that most frequently occur in most of the text documents that are not unique to each document.

Since this experiment is on NEWS domain, Amharic stop words that are domain dependent are considered in addition to the domain independent stop words of Amharic text. For this experiment, the list of Amharic stop words for NEWS domain containing 166 words is gathered from different articles and also by testing the news articles corpus (Which is discussed in section 6.1 of chapter 6) and taking the most frequently occurring

words in the news articles 43 more words are added. A total of 209 stop words are therefore included in the list of stop words.

A code written in Python 3.2 programming language is used for the purpose of identifying the common words in the news article corpus. This is incorporated to the list of stop words prepared for the experimentation by avoiding redundancy with the list gathered from other literatures. In order to select these words a threshold of 20 frequency of occurrence is used. These resulted in 155 words which contain 97 non content bearing words and 58 content bearing words. From these non content bearing words, 43 additional stop words were found which are not included in the stop word list that is gathered from other sources. Some of the most common and news domain based stop words are presented in Table 5.4 and the full list of stop words used for this experimentation is presented in appendix VI.

| **Common Amharic Stop Words** | | | **NEWS Domain Stop Words** | |
|---|---|---|---|---|
| ነው | ሌላ | | ተገለፀ | አስገንዝበዋል |
| ነገሮች | ማለት | | እንዳሉት | አብራርተዋል |
| በታች | ነበሩ | | አስገነዘቡ | አስታወቀ |
| በጣም | ናቸው | | እንደአስረዱት | አሳስበዋል |

**Table 5.4 List of the common Amharic and news domain stop words**

### 5.3.5 Stemmer for Amharic

Stemming is the computational process of removing any affixes present in a word and creating the stem of the word. The goal of stemming is to reduce variant inflectional word forms and identify derivationally related forms of a word to a common base root form. Stemming is an important analysis used in a number of areas such as information retrieval (IR), machine translation (MT) and text classification (Atelach & Asker, 2007). It is language dependent and should be tailored for each language since languages have a varying degree of differences in their morphological properties (Asker, Atelach, Gambäck, Samuel, & Lemma, 2009). Moreover, Stemming has to be tailored to the specific task it is going to be implemented for in any of the application areas mentioned above.

Like most Semitic languages Amharic is rich in word variants and derivative formation. The Amharic word formation involves using prefix, suffix, infix, reduplication and Semitic stem interdigitating (Sisay, 2005). As discussed in chapter three of this research, most function words in Amharic like Conjunctions, Prepositions, Articles, Relative Markers, Pronominal Affixes and Negation Markers are bound morphemes resulting in complex words containing many morphemes.

The extreme stemming process involves removing all affixes from a word and creating the root word from the stem obtained by taking only the consonants of the stem word. The stemming process for Amharic words employs a minimum stem length condition (Nega & Willett, 2002). Even if the smallest Amharic word has Mono radical consonantal roots as we have seen in chapter 3, only words that have Tri- radicals and more are to be stemmed so that the words can remain carrying their semantic meaning.

As we have seen in chapter 3 of this thesis, the most common prefixes in Amharic are "የ"/ 'yä', "ከ"/ 'kä', "ለ"/ 'lä', "ስለ"/ 'slä' and "እንደ"/ '_ndä', which are mostly used in both verbs and nouns. In Amharic prefixes can be formed by combination of other prefixes. For example the prefix "እንደየ"/ '_ndäyä' is a combination of the individual prefix "እንደ"/ '_ndä' and "የ"/ 'yä' which might be used with words like "እንደየቦታው"/ '_ndäyäbotaw'. This means 'from place to place' as in "ባህል እንደየቦታው ይለያያል።"/ 'Tradition differs from place to place'. The individual and combined prefixes list is prepared and included in the dictionary .xml file for the stemmer to remove matching prefixes from the tokens of the Amharic text to be summarized. This list which is presented in appendix II.A is compiled from different sources.

The other word variant forming affixes are suffixes. The most common suffixes in Amharic nouns are the possession marker "-ን"/ '-ne', the emphasis marker "-ም"/ '-me', the object marker "-ና"/ '-na' and the plural marker "-ዎች"/ '-woch'. These suffixes like the prefixes discussed above can be used either individually or in combination as shown in Table 5.5 which is taken from (Zelalem, 2001).

| Individual Suffixes | Combined Suffixes |
|---|---|
| ዎች + ን | ዎችን |
| ዎች + ን + ና | ዎችንና |
| ዎች + ን + ም | ዎችንም |
| ዎች + ና | ዎችና |
| ዎች + ና + ም | ዎችናም |
| ዎች + ና + ን | ዎችናን |
| ዎች + ም | ዎችም |
| ዎች + ም + ና | ዎችምና |

**Table 5.5 Formation of a suffix by combining other suffixes (Zelalem, 2001)**

This list together with the list of the individual suffixes is presented in appendix II.B which is used in the .xml dictionary file of the OTS to formulate the rules of removing suffix from the tokens of the text to be summarized. The prefix list is also used in the same manner as the suffixes in forming the dictionary rules for the Porter stemmer in the OTS tool.

By considering the different requirements of the OTS to carry out the stemming process, different list of rules are prepared and incorporated into the dictionary .xml file. The rules include the list of punctuation marks that are attached with a word either at the beginning or the end of the word, the list of prefixes, suffixes and words with their corresponding stems for those words that deviate from the rules of stemming. However a detailed list of such words that are exceptional to stemming could not found during this research. A sample of few words is therefore used just to show the possibility of including such words into the dictionary.

This rule based dictionary file is used by the stemming algorithm which is saved under the “libxml2.lib” file. This stemmer takes a word, removes all affixes and punctuation marks using the list in the dictionary file, and returns the stem of the word as an output. Words whose stem deviates from the rule based stemming process are listed manually in the dictionary along with their stem and the stem is used directly without applying the stemming rules on them.

The detail description of the rules and processes involved with the dictionary file is discussed in section "5.2.2 How OTS works" above. And the description of the general rules of the stemmer algorithm, for rules of stemming Amharic words to their stem words in a rule based stemming approach is presented in section 5.4.4 of this chapter. The complete lists of the Amharic affixes that are used as an input in this experimentation for the stemming algorithm are presented in appendix II.

## 5.4 PREPROCESSING INVOLVED

### 5.4.1 Amharic News Article Gathering and Cleaning

The corpus of Amharic news articles is collected from the official web pages of The Reporter newspaper – Amharic page, Walta Information Center (WIC) and Ethiopian Press Agency (EPA) – Addis Zemen newspaper. The articles are selected from different content categories like politics, social life, technology, business and economy, health, entertainment and sport. This corpus is then analyzed before it is tested on the customized Amharic OTS tool. From the total of 40 collected articles, 30 are selected to be used for this experimentation. The selection is made based on the size (in terms of number of sentences) of the article. Articles that have more than 30 sentences and less than 120 sentences are selected. This lower boundary is set because it does not make sense to summarize small size text documents and automatic summarization is not effective on small size text. And article with more than 100 sentences is difficult for manual summary preparation. The content of the articles is also considered as another selection criteria. Concepts reported by more than one newspaper are considered only once in order to avoid redundancy.

These selected articles are then preprocessed. Preprocessing includes removal of web content noise like links, pictures and tables. Moreover, the articles are checked for typographical errors by the assigned people to do the manual summary. This is done so that the subjective evaluators do not consider such errors as if it is made by the system in the summarization process. Then the cleaned articles are normalized before they are used in the automatic summary generation by the Amharic OTS tool. The normalization process is discussed in section 5.4.3 of this chapter. After the normalization is done they

are used by the Amharic OTS for automatic summary extraction. The detail content description of the selected articles is presented in Table 6.1 together with the brief discussion in chapter 6 of this research.

### 5.4.2 Manual Summary Preparation

In order to evaluate the performance of the automatic Amharic summaries extracted by the Amharic OTS in this experiment, summaries that are prepared manually by experts are needed. The automatic summary is checked with the human summary using the selected evaluation criteria. The cleaned articles were given to experts from WIC, Fana Broadcasing Corporation (FBC) and Addis Ababa University (AAU) journalism postgraduate students to prepare a manual summary by ranking the sentences of the news article in the order of importance with numbers from 1 up to N, where N is 35% of the total number of sentences in a given article under consideration. Since the maximum summary extraction percentage selected for this experiment is 30%, the manual ranking of sentences is made to be up to 35% of the total number of sentences. This is to have some extra ranked sentences, in case there are errors in ranking like ranking phrases rather than a sentence or omission of sequential number. By concatenating the ranked sentences in the ascending order of their rank a meaningful summary is extracted.

A guideline is prepared based on five different criteria to rank the sentences of a given news article for ideal summary preparation, which is attached in appendix VIII of this thesis. These criteria are Informativeness of the sentence by including the core ideas of the news article; Coverage which is the scope of the sentence in covering the most important topics presented in the news article; Non-Redundancy which is avoiding repetition of ideas using different sentences so that consecutive sentences would not carry the same meaning; Referential Integrity which means that those sentences that have referential words like pronouns and noun phrases should be easy for the reader to identify to whom or what they are referring to in the sentence; and Coherence which is the smooth flow of information about a topic in consecutive sentences rather than having unrelated sentences.

The sentences ranked according to these criteria are arranged in ascending order to prepare a summary. Based on the selected percentage of the total number of sentences in

the article, a summary of M% is extracted. For this experiment M is selected to be 10%, 20% and 30% in both the automatic and manual summary. That is if a news article to be summarized has 50 sentences, a 10% summary is composed of the top 5 ranked sentence, a 20% summary is composed of 10 top ranked sentences and so on. These manual summaries prepared are then used to evaluate the accuracy of the automatic summary extracted by the Amharic OTS as discussed in chapter 6.

### 5.4.3 Normalization of Amharic Characters

To make the summarization process effective by enhancing the text processing involved in this experimentation, normalization plays an important role. It increases the effectiveness of the process of stop word removal, matching of abbreviations and synonyms, stemming, word frequency of occurrence counting and the key stem words matching in sentence selection. According to Seid and Mulugeta (2009), normalization of Amharic characters has improved the Amharic Question Answer System they developed by 12%. This shows the importance of normalization in Amharic text processing. The preprocessing includes normalization of the Amharic characters which include same sounding characters, numbers and special characters like punctuation marks.

1. The first normalization done on the collected news article corpus is the use of the Amharic sentence ending punctuation mark '።' which is referred as "አራት ነጥብ/ Arat Netib". Most of the writers of the news articles use a mix of two English colons as "::" and the proper sentence end punctuation mark '።'interchangeably. This inconsistency in using this punctuation mark needs to be normalized by changing it to the correct Amharic sentence end punctuation mark '።' in all the articles. Doing this enhances the system to identify the end of a sentence with the correct end of sentence marking character, since it is used as one of the sentence delimiter by computer systems in Amharic text documents. This is included in the "am.xml" file under the synonym list of rules.

2. Amharic has different characters that represent the same sound phoneme. These are grouped in to two as: different alphabets that share the same sound and the same sound for the first and fourth order alphabets (Bethlehem, 2002). The sound 'ha' can be taken as an example for the two groups. It has six different representing characters

“ሀ፣ሃ ፤ ሐ፣ሓ ፤ ኀ፣ኃ”. Each pair shows the first and fourth same sounding characters, while the three different alphabets also represent the same sound. It has also been found that the second order of the consonant “ወ”/ ‘wä’ which is “ዉ”/ ‘wu’ and its sixth order “ው”/ ‘w’ are interchangeably used and there is no consistency in their use (Zelalem, 2001). This therefore needs normalizing the second order character “ዉ”/ ‘wu’ to the sixth order “ው”/ ‘w’, since the later one is mostly used.

These same sounding and interchangeably used characters create inconsistency in writing of Amharic texts. Therefore normalization is needed to create consistency throughout the news articles. In the news articles for example, the word ‘mengist’ which means government was written in two forms as “መንግስት” and “መንግሥት”. During the normalization process, the same sounding characters are normalized to one mostly used character like for example all the “ሥ”/ ‘s’ in the above example “መንግሥት” to “መንግስት”. And “ha” sounding characters “ሀ፣ሃ፤ሐ፣ሓ፤ኀ፣ኃ” are normalized to “ሀ” and the second order character “ዉ”/ ‘wu’ is normalized to the sixth order “ው”/ ‘w’. The complete list of such characters and the corresponding normalized character used in this experiment is presented in appendix III.B. It is included in the am.xml file under the synonym list of rules, which includes the different characters that need normalization together with the orders for some of the alphabetical characters that are normalized.

3. Numbers in Amharic text writing are usually represented in three ways using Alphabetical words, Arabic numerals or Ethiopic numerals (Seid & Mulugeta, 2009). Ethiopic numerals are not usually used due to the lack of some characters like a character to represent zero value as we have seen in chapter three of this thesis. Ethiopic numerals are usually used in texts to write calendar years, biblical verses or sequential numbers in a list. This can be normalized to Arabic numerals expressions which are mostly used in the Amharic text writing. Numbers written in Ethiopic numerals are, therefore, converted into Arabic numerals. The list of Ethiopic numerals and their respective normalizing Arabic numerals that is included in the am.xml dictionary file, under the list of synonym rules is presented in appendix I.E of this thesis.

4. The other normalization task was converting the abbreviations in to their respective detail form. Amharic abbreviations are written using the characters "/" forward slash, "." dot, " " empty space or "" no character at all in between the abbreviated characters. This inconsistency in writing abbreviations is handled by normalization process introduced to the tool in the dictionary file together with the stemming rules. The list of abbreviations and their expanded form that are used to formulate the rules in the dictionary file under the synonym list for this experiment is presented under appendix IV.

5. The other normalization process involves compound words. Compound words might be written as one word or separated by space in texts that do not use the Amharic word separator "፡". Those using the separator in writing put compound words as one word or separated by a space without the use of the separator in between, this avoids any complication related to compound words. Based on the list provided in appendix III.A, rules of normalizing compound words that are written as separate words into one word is included in the dictionary file under the "depreciate" list of rules. This rule deletes the space between two words in a compound word.

   For example compound words like "ቤተ ክርስትያን"/ 'betä krestian' which means church, "ቤተ መንግስት"/ 'betä mengist' which means palace and "ወጥ ቤት"/ 'wät bet' which means kitchen, are normalized into one word as "ቤተክርስትያን", "ቤተመንግስት" and "ወጥቤት" respectively. Normalizing compound words helps to avoid problems of dropping part of these words or considering them as separate words during the process of word count and stemming which may result in the change of the meaning of the words completely.

6. The last problem is related to writing of affixes in different formats, as in "ዘግቦአል፣ ዘግቦዋል"/ 'zegboal, zegbowal' which all mean that 'he has reported'. This is solved during the stemming process, as the suffixes "አል" and "ዋል" are removed the same root word is extracted for both words. Related to this, words that are inherited in Amharic from other languages like English are other problems in Amharic writing. Mostly such words are written in the way they are pronounced by the writer and do not seem to have standard form. For example the word 'television' can be written in

Amharic as "ቴሌቪዥን፣ቴሌቭዥን፣ቴሌቭዝን". However due to lack of a detailed list containing more of such words, normalization is handled in this experimentation with few words that are adopted from (Surafel, 2012), (Hassen, Tessema, & Solomon, 2009) and few other additions. The following Table 5.6 shows some of the list of such words that are used in this experimentation by including them under the synonyms list of rules in the dictionary file. The complete list of these words and some Amharic words with the same writing pattern are presented in appendix III.C of this thesis.

| English Word | Transliteration Form 1 | Transliteration Form 2 |
|---|---|---|
| Sponsor | ስፖንሰር | እስፖንሰር |
| Sport | ስፖርት | እስፖርት |
| Television | ቴሌቪዥን | ቴለቭዥን፣ቴሌቭዝን |
| European | ኤሮፓውያን | አውሮፓውያን |
| Automobile | አውቶሞቢል | ኦቶሞቢል |
| Encyclopedia | ኢንሳይክሎፒድያ | ኢንሳይክሎፕድያ፣እንሳይክሎፔድያ |
| Amharic word with different writing pattern | ሚያዚያ | ሚያዝያ |

**Table 5.6 List of Borrowed words with transliteration problem**

### 5.4.4 Rules for the Stemmers Used

As we have seen in chapter 3, affixes are usually used for changing and forming tense, number, gender and case of a word. A simplified version of the Porter stemmer is used in the OTS, which only removes prefixes and suffixes. However, Amharic uses infixes and other word derivatives which need to be considered in stemming a word. An Amharic stemmer that includes all Amharic word stemming rules developed by (Tessema, 2007) is therefore used in the second experimentation of this research. This algorithm is developed based on the Porter stemmer and includes other Amharic stemming rules. It removes affixes like the Porter stemmer and handles exceptions to removing affixes like in names and titles. Moreover, it removes infixes and changes the last character to "sades"/ 'the sixth order of the alphabet', if the last character of the stemmed word is a vowel.

The first experimentation is done using the Porter stemming algorithm of the OTS tool. The stemming algorithm of the OTS tool removes affixes using the dictionary file to form the stem of the word. Affixes are removed by iterative procedures that employ a minimum stem length, recoding and context sensitive rules; with the prefixes being removed prior to the suffixes (Nega & Willett, 2002). In the extreme word stemming, the root word which is the consonant sequence of the stem, is created by removing all vowels from the stem of the word that is obtained by removing affixes. As discussed in section 5.3.5 of this chapter, the smallest stem size is two, and from the affix list the count shows the maximum prefix size is four and the maximum suffix size is five characters.

The stemming algorithm of the OTS uses the reference dictionary for stemming words in a manner described in this and the following paragraphs. Most of the time in Amharic writing, spacing between a word and a punctuation mark is not used. Hence, the symbols remain attached to the word during tokenization. These characters are removed by referring to the list of Amharic punctuation marks in the dictionary file under the stemming rules. The stemming algorithm first checks the token word in both ends for the existence of attached symbol using the “pre and post” list of rules containing punctuation marks. If a match is found it removes any symbol that is attached to a word to be stemmed. Then the word without punctuation marks is stemmed using the consecutive rules.

There is a list of exceptional words to stem formation and corresponding to their stem words that are included in the stemming rules in the dictionary file. The stemming algorithms checks for this list before proceeding to affix removal. If a word to be stemmed is found in this list, the corresponding stem is taken. Otherwise the word will be stemmed using affix removal set of rules using the affix list of rules provided in the dictionary file.

The affix removal is done by removing prefixes first and then suffixes using the set of rule list provided to the stemmer under “pre and post” list of rules in the dictionary. The stemming algorithm cuts the possible prefixes from each word (on the left side of the word) and checks its presence by using exact matching in the prefix list, from the longest to the shortest size of prefixes. Removing prefixes from a word follows the same style as

removing suffixes but removing suffixes in Amharic involves changing the remaining last character in to the sixth order of the alphabet if it ends with a vowel. There might be over stemming the word when the one character prefixes are removed.

For example removing "ከ"/ 'kä' from the word "ከሰሰ"/ 'käsäsä' which means 'he sued' will form a meaningless group of characters "ሰሰ"/ 'säsä'. This however can be controlled by using Amharic dictionary. The Amharic dictionary can be checked for the word after the removal of single character affixes as suggested by Atelach and Asker (2007). If the word is not found in the dictionary, then the stemming is stopped for that word and its previous form is taken as a stem word. This approach has resulted in a 40.09% precision which was better than 36.15% that is observed without applying it (Atelach, Asker, Cöster, & Karlgren, 2005). However, since the scope of this experimentation is based on the open text summarizer tool, this approach is not incorporated to control over stemming.

The rules driving the prefix removal in the Porter stemming algorithm are:-

1. *Read a word from the list that holds the representative unique words of the article after preprocessing, normalization and stop words removal processes are completed.*
2. *If the size of the word is less than three, then return the word as a word without prefix and pass to the next word in the list.*
3. *If the word has M number of characters (where M is greater than two), strip the left most M-2 characters from the word.*
4. *Check if the striped group of characters exists in the prefix list. If it matches with a prefix, then do step 7 and return the word without prefix and pass to the next word in the list.*
5. *Else, if the M-1 group of characters does not match with the prefixes in the prefix list, then strip the left most M-2 characters from the word and repeat step 4 above.*
6. *Else, if the M-2 characters are not found in the prefix list, step 4 and 5 are repeated iteratively by decreasing the number of characters in the left striped from the word by one in each loop until one character (M-(M-1)) is striped and checked in the prefix list.*
7. *If the prefix removed has only one character, then return the word after the prefix removal.*

8. *Else, if all the characters are checked and none of the striped group of characters have match in the prefix list, then return the word with no prefix.*

**Algorithm 5.1 Stemming rules for prefix removal**

In the same manner as prefix removal, the stemmer cuts the maximum possible size of suffixes from each word (from the right side) and checks its appearance in the suffix list from the longest to the shortest size of suffixes. The rules for the iterative process of stripping suffixes in the Porter stemmer algorithm are:-

1. *Read a word from the list that holds the list of words after all prefixes are removed.*
2. *If the size of the word is less than three, then return the word as a word without suffix and pass to the next word in the list.*
3. *If the word has N number of characters (where N is greater than two), strip the right most N-1 characters from the word.*
4. *Check if the striped group of characters exists in the suffix list. If it matches with a suffix, then do step 7 and return the word without suffix and pass to the next word in the list.*
5. *Else, if the N-1 group of characters does not match with the suffixes in the suffix list, then strip the right most N-2 characters from the word and repeat step 4 above.*
6. *Else, if the N-2 characters are not found in the suffix list, step 4 and 5 are repeated iteratively by decreasing the number of characters in the right striped from the word by one in each loop until one character (N-(N-1)) is striped and checked in the suffix list.*
7. *If the suffix removed has one character, then return the word after the suffix removal.*
8. *Else, if all the characters are checked and none of the striped characters have match in the suffix list, then return the word with no suffix.*

**Algorithm 5.2 Stemming rules for suffix removal**

The stem word is then used in calculating the word frequency to select the key most representing word to rank the sentences of the news article to extract the summary. In this research, therefore, the experimentations are conducted using the Porter stemmer inside the OTS tool and an Amharic stemmer that includes the Amharic stemming rules. The

impact of these two stemmers on the summarization process is evaluated and reported in chapter 6 of this thesis.

### 5.4.5 Customization of the OTS to Amharic News Summarization

OTS is independent to any language to be summarized, however it has to be modified to enable it to work for a given language. The freely available library function that is downloaded from the OTS website has a simple user interface that needs modification to make it suitable for the purpose of this experiment. The first thing in the customization procedure is to redesign the user interface to make the tool interactive with a user. The interface is designed using visual studio C# programming language which is explained in section 5.4.6. The next procedure includes the preparation and integration of the different lists mentioned in this chapter into the dictionary file by creating a rule based set of xml code.

Due to the difference in the rules of the language and the morphological structure of Amharic the derivatives of stem words are different from the English or other languages. A separate dictionary file in the ".xml" format is therefore prepared for Amharic just like the other languages that are so far incorporated in the OTS. Some of the rules that are included in the Amharic dictionary "am.xml" file and used in the summarization process of Amharic news articles are attached in appendix VII. The procedure involved is discussed as follows.

1. **Integration of the Prepared Lists**

   As it was discussed in section 5.3 of this paper, the OTS requires list of abbreviations, synonyms, affixes, exceptional stem words, compound words, stop words, numbers, punctuation marks and special characters. These are gathered from different sources and are also compiled by the researcher. For example, as we have seen in section 5.3.4, the list of additional stop words was collected from the corpus of news articles.

   Each of these prepared lists are then integrated and written in an xml file in HTML rule based code. These rules include: - 1) 29 prefix and 95 suffix rules for stemming; 2) 16 punctuation marks forming 41 rules to remove punctuation marks from the beginning and end of a word; 3) 12 sentence delimiter rules to indicate end of a

sentence and 14 exception rules to this under the “line don’t break” rule set; 4) 91 pair of synonym words; 5) 18 words with different writing patterns; 6) 41 same sounding characters for normalization to the selected character; 7) 20 Amharic numerals and their respective normalizing Arabic numerals; 8) 48 pairs of abbreviations and their expanded form; 9) 26 compound words and 10) 211 stop words. The file is then saved as “am.xml” in the dictionary folder “/dic” of the tool. This file is used by the OTS as a reference library to perform the different tasks involved with the summarization process.

## 2. Description of the Rules Sets functionality

The list prepared are formulated in to set of rules that carry functionality that include 1) The rules for parsing in Amharic where a sentence boundary in Amharic could be identified by either one of the punctuation marks “።, !, ?”,

2) The rules for tokenization in Amharic where a word might be separated using either one of the punctuation marks “-, ፣, ፡, ፤, or “ ” empty space”,

3) The rules for removing special characters and punctuation marks that are attached to a word in either end of the word,

4) The expanded forms corresponding to Amharic abbreviations in order to increase the word frequency count of those words which are written in abbreviation or in expanded form as one word,

5) The list of common words for removal of stop words so that only key words could be used to identify and rank sentences,

6) The normalization of Amharic compound words, inherited words, numbers and same sounding characters,

7) The list of similar words for matching synonyms in word frequency counts,

8) The list of affixes for removing any prefix or suffix in stem word formation, and

9) The list of exceptional words to the rules of stemming together with their respective stem word. These set of formulated rules that are listed in the dictionary file are the base for the OTS in extracting the Amharic text summary for the news articles that are tasted in this experiment.

### 5.4.6 Redesigning Graphic User Interface for AOTS

The OTS is an open source text summarization tool with a simple user interface available that comes with the source code. The tool is in library format and does not have executable file to install it directly. In order to make use of this tool on MS-Windows operating system, one has to install Visual studio. For this experiment visual studio 2010 is installed and C# programming language is used from it, since the OTS source code written in C# is used in this study. However, the web interface available online has a better format for users to extract summaries using the application in any one of the 37 languages supported by the tool. The output could be in two options either in a highlighted format on the original text with different color or in a summary form based on the percentage selected by the user.

The original user interface is redesigned for this experimentation using visual studio 2010 C# programming language. The original interface has three text boxes; one for text document to be summarized, one for summary and one for inserting number of sentences to be included in the summary. And it has only one button for summarization. The redesigned interface includes three additional buttons: for saving the extracted summary in the given number of sentences to a text file to any location specified by the user; for clearing the two text boxes for inserting and displaying the extracted new text document; and for exiting and closing the form.

An attempt was made to include a forth button to open a text file directly to the article text box, but it resulted in none functionality of the original tool. Therefore, copy and paste mechanism is used to insert the news article to be summarized in the text box as the original interface forces to use. Figure 5.1 shows the screen shot picture of the GUI redesigned for this experimentation.

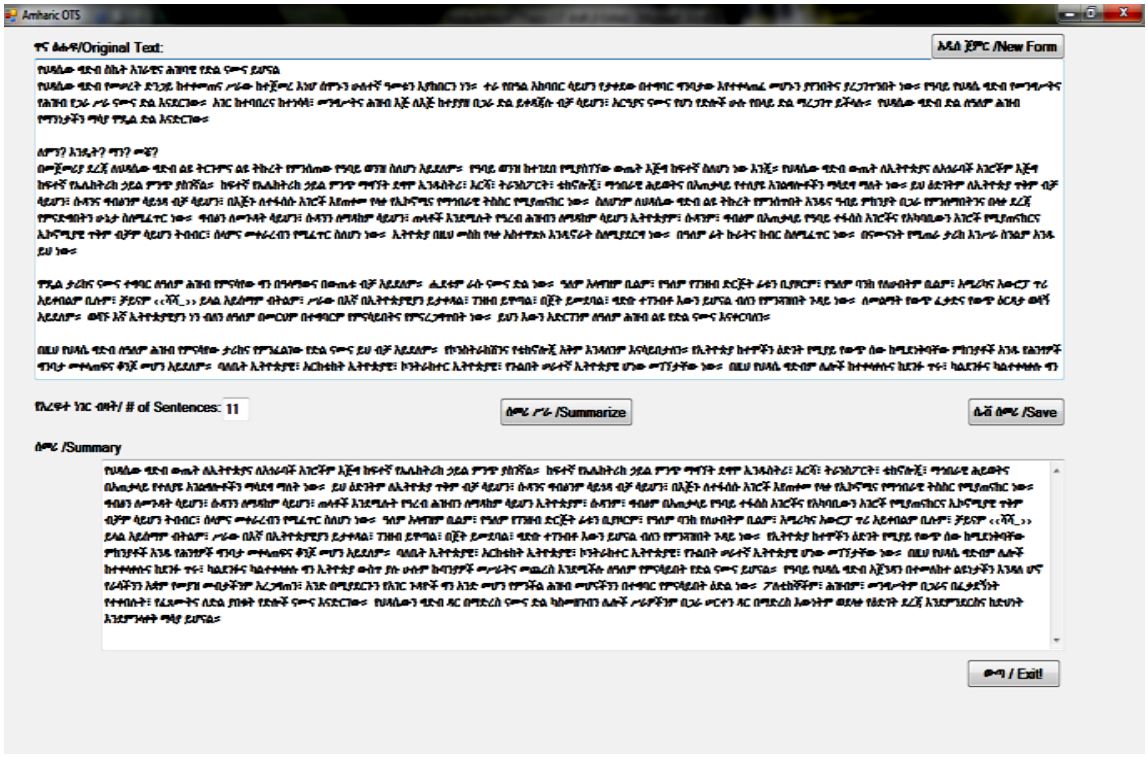


Figure 5.1 GUI redesigned for Amharic news text summarization using OTS

## 5.5 ARCHITECTURE OF THE SYSTEM

The summarization process has three phases: analyzing the source text, determining its salient points and synthesizing an appropriate output (Hahn & Mani, 2000). The Amharic text summarization system using the open text summarizer, therefore, has three components architecture like most summarization systems: input, process and output components. The input component involves the corpus prepared for the Amharic news articles that are to be summarized and tested in the system. The processing component has two sub components. It includes the OTS tool itself and the preprocessing components that are introduced to the OTS using both the dictionary file and additional external modules. The last component is the output of the process which is the automatic summary of the news articles which are evaluated to check the system performance. Figure 5.2 shows the diagrammatical representation of this system.

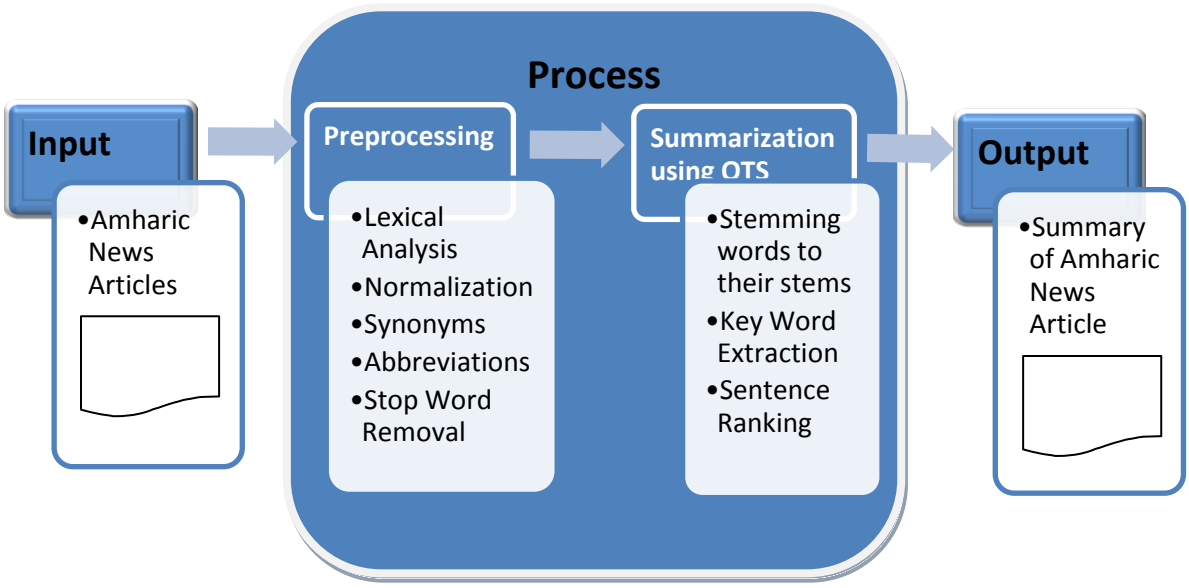


**Figure 5.2 Process based architecture of Amharic news article summarization using OTS**

## 5.6 DESCRIPTION OF THE SUMMARIZATION PROCESS

The automatic text summarization process is a step by step extraction of the most representative sentences of a given article to generate a summary. The OTS uses sentence ranking method for selecting these representative sentences to extract summary from a single text document. There are three steps involved in summary extraction using the OTS. These are 1) Preprocessing, 2) Sentence ranking and 3) Summary extraction.

1) Preprocessing – As we have seen in this chapter there are different lexical analysis that were performed before the ranking of the sentences is done. This step includes the formation of tokens of key representative terms of the text to be summarized. First the text document is read in to a variable and using the dictionary file the sentence and word boundaries are identified. Then normalization of punctuation marks, characters, numbers and compound words is performed. In parallel, synonyms of words and expanding of abbreviations is done. After the stop words are removed from the list of the tokens, they are stemmed using the stemming algorithms (either the Porter stemmer for the first experiment or the Amharic

stemmer for the second experiment). This list of stemmed words is used as an input to the next step in the summarization process.

2) Sentence Ranking – The output of the first step, the list of stemmed words is used as an input in this step to calculate the frequency of occurrence for each of the stemmed words. By arranging them in the decreasing order of the frequency of occurrences, key representative words are selected. Using the dictionary file the sentences of the text are parsed based on the sentence delimiter set of rules. The words that are found in the title of the article are given higher grade by the OTS tool, to penalize the other less representative words of the text. Based on the term frequency of occurrence list, sentences that contain most of the top frequent words are given ranks. The sentence ranking is based on the term frequency (TF) and sentence position in the text. For each sentence the sum of the term frequency of the words that it contains is calculated as:-

$$SS=\sum TF,$$

Where SS is the sentence score in terms of the sum of the word frequency value for each word it contains and TF is the term frequency of each word in a sentence.

The positional grading of the sentences is computed by giving constant multiplicative value "K" where the first sentence of a document gets the highest value which is 2 and the rest of the sentences get the lowest value which is 1.6 in the OTS. This is to penalize the other sentences by giving the title more score than the rest.

Based on the positional grading and sentence score of each sentence the rank of each sentence is calculated. The computational formula is:-

$$SR=SS * K,$$

Where SR is the rank of a given sentence, SS is the sentence score in terms of word frequency and K is the constant multiplicative factor for sentence position.

3) Summary Extraction – This is the last step in the automatic summarization of text documents. A summary is generated based on the rank given to each of the sentences in the previous step. The sentences are arranged in the descending order of their rank. The

summary is extracted based on the number of sentences to be included in the summary, N which is set by the user in the user interface form of the OTS tool while providing input requirements. The top N ranked sentences are displayed to the user concatenated in the order of their appearance in the original text document rather than their rank order. This is to increase the structure and coherence of the summary.

# CHAPTER SIX

# 6. EXPERIMENTATION AND EVALUATION

## 6.1 INTRODUCTION

In this chapter the data set, the experiments conducted, results and the evaluation of the systems are discussed. The selected 30 news articles are saved separately in a ".txt" file for each article after being cleaned from any noise or other non text content like picture, web link and so on. As mentioned in chapter 5, the articles are selected based on their sentence number ranging from 30 to 127. These articles are then grouped into three different categories based on range of the number of sentences they contain. In order to classify each article into one of the three groups, ranges of number of sentences are used. These groups are short articles (SA) containing 30-55 sentences, medium articles (MA) containing 55-80 sentences and long articles (LA) containing 80-127 sentences. All except one of the articles (which has 127 sentences) have less than 100 sentences. Those articles below the minimum range (grouped as very short articles - VSA) and above the maximum range (grouped as very large articles - VLA), are both dropped from the corpus. Each of the three groups consists of 10 different articles which are saved under a file name for each group as SA, MA and LA, and sequential number like "SA1.txt".

As discussed in chapter 5 the sentence ranking in the manual summary is made to contain 35% of the sentences. Since the maximum summary extraction percentage is 30%, the 35% of the total sentences ranked manually are enough to get the top ranked sentence to be included in the summary extraction. The ranked number of sentences in each article is presented in Table 6.1 corresponding to each of the article considered for this experimentation. The Table shows the detail description of the 30 news articles in terms of number of words, total number of sentences in an article, total number of manually ranked sentences, their file name and title of the article.

| File Name | Article Title | # of Words | # of Total Sentences | # of Manually Ranked Sente. |
|---|---|---|---|---|
| SA1.txt | ትርጉም ያላቸው ትራሶች- በካፊኾ | 358 | 32 | 12 |
| SA2.txt | የሕንጻ መጠቀሚያ ፈቃድ ማረጋገጫው መልካም ዜና ነው | 488 | 30 | 11 |
| SA3.txt | ጠቅላይ ሚኒስትር ኃይለ ማርያም ሹማምንቱን አስጠነቀቁ | 721 | 38 | 14 |
| SA4.txt | አሥሩ ዋንጫዎች፣ አሥሮቹ ጀግኖችና አሥሩ ዓመታት በኦልድትራፎርድ | 1,049 | 50 | 18 |
| SA5.txt | የቶዮታ ብቃት ምስጢር | 613 | 37 | 13 |
| SA6.txt | ኑሮ እየተባባሰ ነው! መፍትሔውስ? | 461 | 53 | 19 |
| SA7.txt | በራሪው አማካይ «ፍራንክ ላምፓርድ» | 765 | 38 | 14 |
| SA8.txt | «በትክክለኛው መንገድ ከተሰራ የብቃት መለኪያው በጣም ውጤታማ ይሆናል»-ጠቅላይ ሚኒስትር ኃይለማርያም ደሳለኝ | 485 | 44 | 15 |
| SA9.txt | እየሰመጡ ያሉ ኢዲሞክራሲያዊያን! | 745 | 45 | 16 |
| SA10.txt | ድህነት የማስወገድ ጉዞአችንን በስኬት ለመደምደም | 1,219 | 51 | 18 |
| | **Short Articles Average** | **702** | **43** | **16** |
| MA1.txt | ያልበላውን ማከክ የሚወደው ሂዩማን ራይትስ ዎች | 1,254 | 64 | 23 |
| MA2.txt | ምርጫውና የሴራው መክሸፍ | 957 | 60 | 21 |
| MA3.txt | ከቸኮለው የነዳጅ ተገኘ መግለጫ... እስከ ጥቃቅንና አነስተኛ ተግዳሮት | 870 | 73 | 26 |
| MA4.txt | የውሳኔዎች ሁሉ የበላይ ሕገ መንግሥቱን ተግባራዊ እናደርጋለን የሚለው ውሳኔ ነው | 610 | 57 | 20 |
| MA5.txt | ለሥርዓትና ለተቋማት ግንባታ ልዩና አስቸኳይ ትኩረት ይሰጥ! | 608 | 61 | 22 |
| MA6.txt | ምን ዓይነት ፕሬስ ይበጀናል?(ክፍል ሁለት) | 1,305 | 77 | 27 |
| MA7.txt | የህዳሴው ግድብ ስኬት አገራዊና ሕዝባዊ የድል ናሙና ይሆናል | 605 | 56 | 20 |
| MA8.txt | ‹‹በሕዝብ የተመረጥኩ ነኝ›› ብሎ የሚያምን መንግሥት «ሕዝብ አለቃዬ ነው›› ብሎ ማመንም አለበት | 552 | 56 | 20 |
| MA9.txt | ሙስናን ይዋጉታል እንጂ አያሙትም! | 529 | 70 | 25 |
| MA10.txt | ወደፊት መራመድ ቢያቅታቸው ወደኋላ መንሸራተት የለባቸውም | 596 | 65 | 23 |
| | **Medium Articles Average** | **788** | **66** | **24** |
| LA1.txt | የቤት ፍላጎትን የማሟላት ቀሪ የቤት ሥራዎች | 1,081 | 82 | 29 |
| LA2.txt | «የቁጠባ አቅማችንን የማሳደጉ ጥረት ይበልጥ ተጠናክሮ ሊቀጥል ይገባል» -አቶ ኃይለማርያም ደሳለኝ የኢፌዴሪ ጠ. ሚ. | 1,027 | 95 | 34 |
| LA3.txt | የካፋዎች መገለጫ - ቡና | 1,102 | 95 | 34 |

| | | | | |
|---|---|---|---|---|
| LA4.txt | በሕፃናት ክትባት ዙሪያ | 1,279 | 96 | 34 |
| LA5.txt | አንድ ለአምስት የልማት ወይስ የፖለቲካ ሥራዊት? | 1,423 | 99 | 35 |
| LA6.txt | በማርፊድ አንደኛ | 1,172 | 95 | 34 |
| LA7.txt | ኢኮኖሚውና ቢዝነሱ ፊዚል! እናነቃቃው እንጂ! | 631 | 80 | 28 |
| LA8.txt | ሕዝብ እየጠበቀ ነው! | 603 | 86 | 30 |
| LA9.txt | « ኢንፌክሽን በመጀመሪያዎቹ የሕፃንነት ዕድሜ ሲከሰት ገዳይ ይሆናል» | 892 | 87 | 31 |
| LA10.txt | በታላቁ የህዳሴ ግድብ ላይ የተደመጠው ባእድ ድምፅ | 2,386 | 127 | 45 |
| | **Large Articles Average** | **1,265** | **68** | **33** |

**Table 6.1 Description of the corpus used in the experiment**

## 6.2 AUTOMATIC SUMMARIZATION

Two summarization experiments are conducted: by summarizing news articles using the OTS before integrating Amharic stemmer and using Amharic stemmer with the OTS tool. To differentiate the two experimentations and the summaries extracted under each experiment, the two experiments are referred as E1 and E2. Both of these experiments are conducted for all the articles in the corpus with summary extraction percentage of 10%, 20% and 30%. These summary extraction percentages are selected to observe the effect of the summarization process on different ranges of summarization which are not closer to each other. The total number of automatic summaries is 180, having 90 automatic summaries in each experiment.

The number of sentences extracted for an automatic summary from each article using the selected percentage is determined by multiplying the number of sentences in the article by the percentage and rounding it to the nearest zero. That is if the original article has 44 sentences, the 10% summary has 4 sentences, for 57 sentences a 10% summary has 6 sentences and so on. This list of the number of sentences extracted using each percentage for each article is presented in Table 6.2.

| Original File Name | Number of Total Sentences | Number of Sentences in the Ideal and Automatic Summary under E1 & E2 | | |
|---|---|---|---|---|
| | | 10% | 20% | 30% |
| SA1.txt | 32 | 3 | 6 | 10 |
| SA2.txt | 30 | 3 | 6 | 9 |
| SA3.txt | 38 | 4 | 8 | 11 |
| SA4.txt | 50 | 5 | 10 | 15 |
| SA5.txt | 37 | 4 | 7 | 11 |
| SA6.txt | 53 | 5 | 11 | 16 |
| SA7.txt | 38 | 4 | 8 | 11 |
| SA8.txt | 44 | 4 | 9 | 13 |
| SA9.txt | 45 | 5 | 9 | 14 |
| SA10.txt | 51 | 5 | 10 | 15 |
| **Average** | **43** | **4** | **9** | **13** |
| MA1.txt | 64 | 6 | 13 | 19 |
| MA2.txt | 60 | 6 | 12 | 18 |
| MA3.txt | 73 | 7 | 15 | 22 |
| MA4.txt | 57 | 6 | 11 | 17 |
| MA5.txt | 61 | 6 | 12 | 18 |
| MA6.txt | 77 | 8 | 15 | 23 |
| MA7.txt | 56 | 6 | 11 | 17 |
| MA8.txt | 56 | 6 | 11 | 17 |
| MA9.txt | 70 | 7 | 14 | 21 |
| MA10.txt | 65 | 7 | 13 | 20 |
| **Average** | **66** | **7** | **13** | **20** |
| LA1.txt | 82 | 8 | 16 | 25 |
| LA2.txt | 95 | 10 | 19 | 29 |
| LA3.txt | 95 | 10 | 19 | 29 |
| LA4.txt | 96 | 10 | 19 | 29 |
| LA5.txt | 99 | 10 | 20 | 30 |
| LA6.txt | 95 | 10 | 19 | 29 |
| LA7.txt | 80 | 8 | 16 | 24 |
| LA8.txt | 86 | 9 | 17 | 26 |
| LA9.txt | 87 | 9 | 17 | 26 |
| LA10.txt | 127 | 13 | 25 | 38 |
| **Average** | **94** | **9** | **19** | **28** |
| **Corpus Average** | **68** | **7** | **14** | **20** |

**Table 6.2 The number of sentences extracted using the selected percentages for each article**

A respective ideal summary is also prepared manually for each article in the above extraction percentages. This is done based on the rank given to the sentences of each article by the experts. To prepare the manual summary under each summary extraction percentage, the following steps are followed. The ranked sentences are first extracted and arranged in the ascending order of rank in to a ".txt" text file. Based on the corresponding value in Table 6.2 for a given summary extraction percentage, sentences are selected to appear in the summary. Then it is saved in a text file separately to be used in objective evaluation. A total of 90 manual summaries are created for the thirty articles at the three extraction percentages. These manually prepared ideal summaries are used to automatically evaluate the quality of the automatic summary in both experiments.

### 6.2.1 E1 – Using OTS's Porter Stemmer

The first experiment is done by using the OTS directly without additional Amharic stemmer, by implementing the dictionary file which contains the Amharic language lexical rules and list of necessary words for the summarizer as discussed in chapter 5. This experiment implements the term frequency and sentence ranking methods. It also uses the Porter stemmer inside the OTS tool which makes use of the Amharic dictionary prepared that lists the rules for prefix and suffix removal and list of other language lexicons that includes stop word, synonyms, abbreviations, punctuation and so on which help to increase the effectiveness of the term frequency method.

An automatic summary of 10%, 20% and 30% was generated for each of the thirty news articles. A total of 90 automatic summaries are generated at these three percentage rates. These summaries are saved in a text file separately under each percentage rate. The evaluation of these summaries and results obtained are discussed in section 6.3 of this chapter.

### 6.2.2 E2 – After Introducing Amharic Stemmer

The second experiment is conducted by integrating the Amharic stemming algorithm into the OTS tool. It uses the same dictionary file that is used in experiment one which contains the Amharic language lexical rules and list of necessary words as discussed in chapter 5. This experimentation also implements the term frequency and sentence ranking

methods. But instead of the Porter stemmer, it uses the Amharic stemming algorithm that is taken from Tessema (2007). This is to observe the effect of the Amharic stemmer on the summarization process by increasing the effectiveness of the term frequency method.

Like experiment one, an automatic summary of 10%, 20% and 30% was prepared for each of the thirty news articles. A total of 90 automatic summaries are extracted at these percentage rates for the thirty articles. The evaluation of these summaries and results obtained are discussed in section 6.3.

## 6.3 EVALUATION AND RESULTS DISCUSSION

As we have seen in chapter two of this thesis there are two major types of evaluation methods related to text summarization. These are Intrinsic and extrinsic evaluation methods. Each of them has its own number of methods which are mostly used in combination in text summary evaluations. Intrinsic evaluation is done by comparing the automatic summary with the ideal summary prepared manually. It measures the automatic systems performance on its own. Whereas extrinsic evaluation is done based on evaluating how the automatic summaries are good enough to accomplish the purpose of some other specific task. For the purpose of evaluating summaries in this study, intrinsic evaluation is implemented in two ways. The first one is Manual evaluation which is done by selected two experts, who have also prepared the manual summary before, based on chosen evaluation criteria. And the second one is Automatic summary evaluation which compares the automatic summary with an ideal summary prepared by experts.

1. The manual evaluation is done by rating the automatic summaries based on the evaluation criteria prepared for this experiment which is attached in appendix IX. These evaluation criteria used are based on readability, fluency, level of information carried, flow, coverage and overall observation. These are grouped into four major quality evaluation criteria with six measures. The fluency and readability of the summaries is measured in terms of the linguistic quality. In this criterion, the summaries are checked for and given grades based on grammatical correctness, non redundancy and referential clarity. The informativeness of the summaries measures the coverage and level of information carried in the automatic summaries. The flow of the automatic summaries is

measured in terms of the structure and coherence of the sentences in the summary. The final grade is given based on the overall observation and expectation of the user from the automatic summary. The evaluator grades the summary as, to what level the summary is found to be satisfactory in the overall examination of the summary.

A grade level of 1 up to 5 (where 1 represents very poor, 2 – poor, 3 – fair, 4 – good and 5 – very good) is used for each of the above six criteria measuring the four quality evaluations for the automatic summaries. The grading is given from a total of 30 for each automatic summary. The grades given under each criterion for each automatic summary are summed to get the total grade for the summary. This total grades are then converted to percentage to show the evaluation results in percentage.

2. The objective evaluation is done by comparing the automatic summary with a reference ideal summary. Sentence extraction method of text summarization allows the use of precision (P), recall (R) and F-measure (F-m) as an evaluation measures for automatic text summaries. The similarity of the content is measured by considering the extraction unit, which is 'sentence' in the case of this experiment. This similarity of sentences among the automatic and ideal summaries for a given article is determined using these three evaluation measures.

   In order to accomplish this evaluation task automatically, a simple sentence similarity evaluating tool that implements the above three measures is designed in C# programming language. The tool receives two summary text files, in this case the automatic summary and the ideal summary for a given article, and checks for the existence of similar sentences in both files. Based on the number of similar sentences precision, recall and F-measure are calculated using the formula shown in the next paragraphs. It finally displays the output in a tabular form as shown in Figure 6.1, the screen shot taken after execution. It gives the option to save the evaluation results in to an MS-excel file.

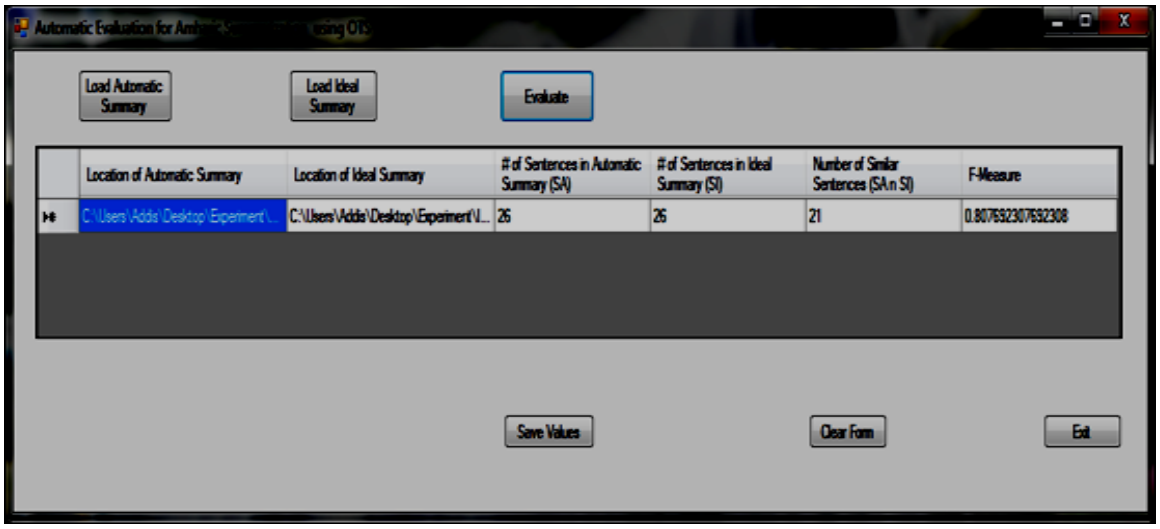


**Figure 6.1 Automatic / Objective evaluation tool screen shot**

This evaluation is performed using the co-selection measures that are mentioned previously, which assess the quality of the summary based on the number of sentences that commonly appear in the ideal and automatic summary. The most common of evaluation measures P, R and F-m are computed as follows. In this experiment F-m is used to report the evaluation results. SI and SA are sentences in the ideal summary and automatic summary for a given article respectively.

1) Precision reflects how many of the system's extracted sentences are good. It is computed as the ratio of the relevant sentences extracted in the automatic summary (which is the intersection of the sentences that appear in both automatic and ideal summary) to the total relevant sentences,

$$\boldsymbol{P = \frac{SA \cap SI}{SI}}$$

2) Recall reflects how many good sentences the automatic summary has not included or how many the system has missed. It is computed as the ratio of the relevant sentences extracted in the automatic summary to the total sentence extracted, and

$$\boldsymbol{R = \frac{SA \cap SI}{SA}}$$

3) F-measure is a good tradeoff between P and R since it gives the same importance for both precision and recall, hence it is important to use it as an evaluation measure. It is computed as:-

$$Fm = \frac{2PR}{P + R}$$

The objective evaluation is done for both of the experiments on the 180 summaries extracted at the three summarization percentage level. However due to the fact that manual evaluation is time and labor intensive, it is done for the first experiment only for 15 article summaries, at the three different extraction rates. This includes all the 10 articles from the short article group and 5 articles chosen by taking the file name with odd numbers from the medium article group. These articles are chosen systematically by taking the 1st, 3rd, 5th, etc. articles. That is a total of 45 summaries for the fifteen articles in the short and medium article group are evaluated using manual evaluation method. However, objective evaluation is done for all the automatic summaries generated using both of the experimentations E1 and E2. The discussion of the evaluations and the results observed are presented in the following sections.

### 6.3.1 Manual Evaluation for Experiment One

The manual evaluation is done for 45 summaries generated from the 15 articles in the short and medium article groups at the three summary extraction percentages used in this experiment. As described in the previous section, six evaluation criteria are used to grade the summaries on the scales 1up to 5. The grades given show to what level the criteria under consideration is achieved in each summary. The evaluation guideline is attached in appendix IX. An average of the grade values given by the two evaluators is calculated for each of the measure and are converted to percentage values to show the results in percent from hundred. The comparison is done considering the size of the original article using the two small and medium article groups (SA and MA) and considering the size of the summary for a given article using the three extraction percentages.

1. Is the summary grammatically correct?

   This is one of the linguistic quality measure used to evaluate the automatic summary. The experts check any grammatical error which occurred during the summarization process like fragments of sentences. What was observed was, in some of the summaries sub titles were concatenated to the sentences or they were followed by sentences from another subtitle.

   Since the experimentation uses sentence extraction methods, most of the results obtained are accurate in terms of grammatical correctness. The summary is composed of directly extracted sentences in their rank order; therefore by this criterion a good result is obtained in almost all the automatic summaries. From the 45 automatic summaries, 3 of them got 60%, 18 got a score of 80% and the rest 24 summaries got 100%. It is also observed that the grammatical correctness is not related to the size of the summary or the size of the original article; hence each summary got the same grade in all extraction rates. However, this criterion is dependent on the writing style of the original article. We can see that the summaries in the MA group got better score than those in the SA group. This shows that the structure of the summaries of the original articles in the MA group has less sub titles (which create the fragment sentence in the automatic summary) than articles in the SA group. The table showing the grades given in percentage is presented in appendix XI.A.

2. Is there redundancy in the summary?

   The second linguistic quality measure used is non redundancy in the extracted sentences of the automatic summary. Here the experts check for any repetition of ideas in sentences. As much as possible the summary should contain one sentence to express one idea, rather than having more than one sentence to express the same concept with different words. Having non redundancy in the summary makes the summary more readable and interesting for readers.

   Based on this criterion the automatic summaries are given scores: 10 of the summaries got 60%, 11 of the summaries got 80% and 24 of the summaries got 100%. From the 15 summaries in the medium article group (MA files), 9 got 60% and 6 got 80%. It has also been observed that none of the summaries in this group got 100%.

The average rank for the articles in the SA group at 10% and 30% extraction rate is 96% and at 20% it is 94%. Where as, in the MA group it is 68% for all the three extraction rates. This means there is more redundancy in the MA size article summaries than those in the SA article summary group. This shows that as the size of the original text increases, writers tend to write the same idea in different sentences repeatedly, resulting in more redundancy. Due to this, the extracted summary then has repeated sentences since the system uses word frequency to rank the sentences.

For all the summaries except for the SA6.txt at 20% extraction rate, the average score given is the same under all the three extraction rates. This criterion, therefore, is not related the size of the summary or the extraction rate. It is also observed that the average scores for the SA group summaries are 94% at 20% and 96% at 10% and 30% extraction rates and for the MA group summaries, it is 68% under all the three extraction rates. This shows that as the size of the original article increases summary extracted increases, redundancy also increases. The overall results show that the system summaries avoid redundancy very well for short articles. The grades given for the summaries based on this criterion is shown in appendix XI.B.

3. Is referential clarity maintained in the summary?

   The last linguistic quality measuring criterion is referential clarity which measures the easiness to identify to whom or what the referential pronouns or words used refer to and if the reference is correct in consecutive sentences. Here the experts check that if questions like “who? Or what?” can easily be answered in the summary for the referential words used.

   More references are used as the size of the article increases. Among the 45 summaries evaluated using this criterion, 6 got 60%, 19 got 80% and 20 got 100%. It is observed that this criterion is maintained more in the summaries for the MA articles than the SA articles except at 30% extraction rate, where the average is 90%. More than half (9 out of 15) of the medium size article summaries got 100% and one third of the small size article summaries got 100%. The average referential clarity is lower for small size summaries at 10% and 20% extraction rate, in the SA group of articles (which is 80% and 86% respectively) than those in the MA (which is 88% at both

extraction rates). In addition to this, the average score for the SA group increased as the extraction rate increases.

Among the three summary extraction rates, better results are obtained at the 30% rate for the small article group, showing increase for each rate as 80%, 86% and 90% on the average. This shows that as the size of the summary increases for a given article, more sentences are included in the summary resulting in the increase of referential clarity. That is as the summary size increases the referential clarity is maintained more than it is in the small size summary for a given article. The average grade values for the evaluation based on this criterion are shown in a table in appendix XI.C.

4. To what level is the summary informative?

   It measures the information content of the automatic summary regarding the topic of the article. It helps to understand which of the summaries include the most important information of the article. The summary should include the main information that relates to the topic of the article rather than sentences that are written about none related details.

   The grades given for the automatic summaries under this criterion ranges from 40% to 100%, where 1 got 40%, 2 got 60%, 18 got 80% and 24 got 100%. Since 42 of the summaries got 80 and 100%, it shows that the summaries are more informative of the topic of the article. As we have seen in chapter 5, this is because the system gives more rank for the topic sentences, by multiplying the frequency of the topic sentence by 2. It is observed that the summaries in the SA group got average rank of 80% for the small size rate (10% summary) and 88% for the 20% and 30% summaries. The average for the MA summaries is 96% for all the three summary extraction rates. This shows that as the size of the original article increases, the informativeness of the summary extracted is high since the number of sentences to be included in the summary is large.

   The average in the MA is the same for the three extraction rates which is 96%. However for the small size summaries it is 80% at 10% and 88% for the 20% and 30% rates. This shows that as the number of sentences in a summary for a given article

increase, the informativeness of the summary also increases, showing a direct relationship between summary size and information content. The overall rank which is presented in appendix XI.D shows the automatic summaries are informative regarding the topic of the article.

5. How well is the summary structured and organized?

   This criterion measures the connectedness of the sentences in the summary and the extraction structure of the summaries. The extracted summaries should not be a heap of concatenated sentences but rather should build up the topic of the article in the consecutive sentences to a well structured coherent summary.

   The automatic summaries are given grades under this criterion as, 16 of them got 60%, 19 of them got 80% and 10 of them got 100%. This shows that one third of the summaries are less structured and organized. It is observed that the average grade in the SA summaries varies from 72% in the 10 and 20% extraction rates to 80% in the 30% extraction rate. Where as in the MA group summaries it is 80% in the 10 and 20% extraction rates to 88% in the 30% extraction rate. This shows that as the size of the summary increases, the coherence and structuredness increases, resulting in a more attractive, interesting and readable summary.

   It is also observed that the average of the SA group of summaries (72% to 80%) is lower than the average grade value of the MA group of summaries (80% to 88%). The maximum average grade for the SA group (80%) is the minimum average grade for the MA group. This shows that the structuredness and coherence of a summary increases as the size of the article increases, since it results in large size summaries. The grades given under this criterion are presented in the table in appendix XI.E.

6. From your overall reading of the article, how do you rate the extracted summary?

   After reading the article and the corresponding extracted summary, the evaluators have some subjective judgment regarding the automatic summary. Based on their expectation and satisfaction, each automatic summary is given a grade value.

   The grading values in this criterion range from 20% to 100% as; 1 summary got 20%, 5 summaries got 40%, 11 summaries got 60%, 16 summaries got 80% and 12

summaries got 100%. The average grade value for the summaries in the SA group are 54%, 72% and 90% at the extraction rates of 10%, 20% and 30%, respectively. For the MA group of summaries the average grade values at the three extraction rates (10%, 20% and 30%) are 64%, 84% and 92%, respectively.

This shows that as the size of the extracted summary increases, the summaries are more satisfactory. It also shows that the small size (10%) summaries for either SA or MA group summaries are not good enough to meet the reader's expectation. It is observed that the average grades for the two summary groups show different values at the three different extraction rates, increasing in parallel as the extraction rate increases. The values show that the summaries extracted using higher extraction percentage, i.e. having more sentences in the summaries increase the achievement of reader's expectation and satisfaction, regardless of the size of the original article. The grade values given under this criterion by the evaluators is shown in appendix XI.F of this thesis.

The grade values given under each criterion for each of the automatic summaries are added to determine the total grade score. The total score is then divided by 30 (the total of the maximum rank given under each criterion) and multiplied by hundred to change it to percentage. The grade score value, the total grade rank and the average rank calculated for each summary is presented in appendix X. The average of these percentage values for the SA group summaries are 78.33%, 83.33% and 86.67% and for the MA group of summaries are 81.33%, 84.67% and 88% in the three extraction rates 10%, 20% and 30% respectively.

These average evaluation grades of the automatic summaries using the manual evaluation shows that the quality of the automatic summary increases as the size of the extracted summary increases regardless of the size of the original article. Based on the manual evaluation, the more sentences are included in the extracted summary, the better the quality is for a given article of any size. The following Table 6.3 shows the total evaluation grade given to each of the evaluated article summaries based on the six manual evaluation criteria chosen for this experimentation. The grade values are

presented for each of the 45 summaries selected for manual evaluation which are extracted from the 15 news articles.

| Original File Name | Number of Total Sentences | Total %age | | |
|---|---|---|---|---|
| | | 10% Summary | 20% Summary | 30% Summary |
| SA1.txt | 32 | 56.67% | 70.00% | 66.67% |
| SA2.txt | 30 | 80.00% | 83.33% | 86.67% |
| SA3.txt | 38 | 83.33% | 86.67% | 90.00% |
| SA4.txt | 50 | 80.00% | 83.33% | 90.00% |
| SA5.txt | 37 | 76.67% | 86.67% | 96.67% |
| SA6.txt | 53 | 76.67% | 80.00% | 93.33% |
| SA7.txt | 38 | 73.33% | 80.00% | 86.67% |
| SA8.txt | 44 | 80.00% | 83.33% | 86.67% |
| SA9.txt | 45 | 93.33% | 90.00% | 96.67% |
| SA10.txt | 51 | 83.33% | 90.00% | 93.33% |
| **Average** | **42** | **78.33%** | **83.33%** | **88.67%** |
| MA1.txt | 64 | 86.67% | 90.00% | 90.00% |
| MA3.txt | 73 | 76.67% | 80.00% | 90.00% |
| MA5.txt | 61 | 83.33% | 80.00% | 83.33% |
| MA7.txt | 56 | 80.00% | 86.67% | 90.00% |
| MA9.txt | 70 | 80.00% | 86.67% | 86.67% |
| **Average** | **65** | **81.33%** | **84.67%** | **88.00%** |

**Table 6.3 The average score of subjective evaluation given to the 45 automatic summaries**

### 6.3.2 Objective Evaluation for E1

The 90 automatically generated summaries under the first experiment are evaluated using the evaluation tool developed for this purpose in C# programming language. It is done by checking the existence of similar sentences that are found in both the automatic and ideal summary prepared for each article. Therefore the 90 automatically generated summaries are compared with the corresponding 90 ideal summaries. As we have seen in section 6.3, F-measure is the weighted average of precision and recall; hence it is used as evaluation measure. Table 6.4 shows the summary of the evaluation values for the 90 automatic summaries under each extraction rate in percentage.

The average performance measure at the three extraction rates 10%, 20% and 30% are: for the articles in the SA group 48.17%, 49.71% and 56.24%; for the articles in the MA group 52.8%m 62.17% and 75.65%; and for the articles in the LA group 37.71%, 49.70% and 66.80% respectively. When we examine the increase rate of performance as the extraction rate changes from 10% to 20% and from 20% to 30%; for the SA group it is

1.54% and 6.53%; for the MA group it is 9.37% and 13.48%; and for those in the LA group it is 11.99% and 17.10% respectively. The corpus average performance measure at the 10%, 20% and 30% extraction rates are 46.22%, 53.86% and 66.23% with an average rate of increase 7.64% and 12.37% as the extraction rate changes from 10% to 20% and 20% to 30% respectively.

This shows that as the extraction rate increases the performance of the system also increases, showing a direct relationship between size of summary and performance of the system. It also shows that the increase rate of the performance measure at the different extraction rates is high for MA and LA size articles than small size articles. The change in the extraction rate brought a comparatively minimum effect on performance, (1.54% as the rate changes from 10% to 20%) in the SA news article groups.

When we observe the performance measure among original document size, the highest result is scored by the middle size articles. The average of this group is 52.80%, 62.17% and 75.65% at 10%, 20% and 30% extraction rates, respectively. The highest average performance of all the 9 group (10 summaries with three extraction rates for the three groups) of summaries is 75.65% observed for the middle group summaries at the 30% extraction rate. The next two good average performance measures are observed for the LA group of article summaries 66.80% at the 30% extraction rate and for the MA group of article summaries 62.17% at the 20% extraction rate. The lower average performance measure 37.71% is observed for the large article summaries at the 10% extraction rate. It is so visible that the automatic summaries are of good performance rate for the middle size article summaries, since the score for all the extraction rates are more than 50%. Whereas the SA and LA group articles scored less than 50% for the 10% and 20% extraction rates and scored more than 50% at the 30% extraction rates.

From the evaluation results, we can see that the summaries get better as the extraction rate that determines the number of sentences extracted, increases. This proves the basic principle that the average performance increased as the extraction rate increased, hence the summary at higher extraction rates gets more similar to the original article. However the increasing rate of performance showed irregularity. The average of the corpus result for the three extraction rates 46.22%, 53.86% and 66.23% shows that the concept of a

given article is best covered at the higher extraction rates. The increase in performance of the system is 7.64% and 12.37% as the rate increases from 10% to 20% and then to 30% respectively. This also shows that the performance also increases at an increased rate as the extraction rate increases by a fixed percent. That is the F-m increased at an increasing rate as the extraction percentage increases constantly. Table 6.4 shows the summary of the objective evaluation results of the 90 summaries extracted using experiment one. The detail result observed for each summary is presented in appendix XII (A, B and C for each rate).

| Original File Name | F-Measure in % | | |
|---|---|---|---|
| | at 10% | at 20% | at 30% |
| SA1.txt | 100.00% | 50.00% | 30.00% |
| SA2.txt | 66.67% | 50.00% | 66.67% |
| SA3.txt | 25.00% | 62.50% | 72.73% |
| SA4.txt | 20.00% | 50.00% | 40.00% |
| SA5.txt | 75.00% | 57.14% | 54.55% |
| SA6.txt | 60.00% | 36.36% | 37.50% |
| SA7.txt | 50.00% | 50.00% | 45.45% |
| SA8.txt | 25.00% | 44.44% | 76.92% |
| SA9.txt | 40.00% | 66.67% | 78.57% |
| SA10.txt | 20.00% | 30.00% | 60.00% |
| **Average** | **48.17%** | **49.71%** | **56.24%** |
| MA1.txt | 66.67% | 46.15% | 73.68% |
| MA2.txt | 50.00% | 75.00% | 77.78% |
| MA3.txt | 42.86% | 53.33% | 68.18% |
| MA4.txt | 83.33% | 72.73% | 88.24% |
| MA5.txt | 66.67% | 66.67% | 72.22% |
| MA6.txt | 37.50% | 53.33% | 73.91% |
| MA7.txt | 33.33% | 63.64% | 82.35% |
| MA8.txt | 33.33% | 72.73% | 88.24% |
| MA9.txt | 57.14% | 64.29% | 61.90% |
| MA10.txt | 57.14% | 53.85% | 70.00% |
| **Average** | **52.80%** | **62.17%** | **75.65%** |
| LA1.txt | 37.50% | 75.00% | 60.00% |
| LA2.txt | 20.00% | 26.32% | 58.62% |
| LA3.txt | 40.00% | 57.89% | 72.41% |
| LA4.txt | 30.00% | 36.84% | 62.07% |
| LA5.txt | 30.00% | 45.00% | 70.00% |
| LA6.txt | 20.00% | 31.58% | 62.07% |
| LA7.txt | 50.00% | 62.50% | 87.50% |
| LA8.txt | 44.44% | 41.18% | 61.54% |
| LA9.txt | 66.67% | 64.71% | 65.38% |
| LA10.txt | 38.46% | 56.00% | 68.42% |
| **Average** | **37.71%** | **49.70%** | **66.80%** |
| **Corpus Average** | **46.22%** | **53.86%** | **66.23%** |

**Table 6.4 Summary of objective evaluation results for experiment one**

### 6.3.3 Objective Evaluation for E2

The evaluation for experiment two on the 90 automatically generated summaries, extracted after including Amharic stemmer on the OTS, is done using the same tool as experiment one. Each of the automatic summaries are checked with the corresponding ideal summary used for evaluating experiment one. F-measure is used here to measure the performance of the system and to examine the effect of using Amharic stemmer in the OTS tool. The evaluation measure results and the observations are discussed in the following paragraphs and Table 6.5 shows the summary of the performance evaluation measure values for the 90 automatic summaries under each extraction rate in percentage.

The observed average performance measure at the three extraction rates 10%, 20% and 30% are: for the articles in the SA group 30.67%, 61.26% and 72.12%; for the articles in the MA group 38.75%, 50.18% and 72.14%; and for the articles in the LA group 33.85%, 58.58% and 72.83% respectively. When we examine the increase rate of performance as the extraction rate changes from 10% to 20% and from 20% to 30%; for the SA group it is 30.60% and 10.86%; for the MA group it is 11.43% and 21.96%; and for those in the LA group it is 24.73% and 14.25% respectively. The corpus average performance measure under the extraction rates 10%, 20% and 30% are 34.42%, 56.68% and 72.37% respectively. It shows an average rate of increase of 22.25% as the extraction rate changes from 10% to 20% and 15.69% when it changes from 20% to 30%.

The above performance figures shows that the system in the second experiment (E2) better performs at the 30% extraction rates for all the three groups of news article summaries. It also shows that the system performance is minimum for the 10% extraction rate and average at the 20% extraction rate. As it is expected, the performance of the system increased as the extraction rate increases, showing a direct relationship between performance and extraction rate. The average rate of increase, as the extraction rate increases, is significantly increasing for all the summaries of all the news articles in the three groups. We can also observe that the average performance increase rate for the corpus increased at a decreasing rate as the extraction rate changes from 10% to 20% and 20% to 30% (that is 22.25% and 15.69%, respectively).

When we examine the performance measure among the original document size for a given extraction rate, at the 10% extraction rate the MA group of summaries, at the 20% extraction rate the SA group of summaries and at the 30% extraction rate the LA group of summaries scored the highest average F-m. Moreover, at the 30% extraction rate the average F-m for each group of summaries is not significantly different. That is, the average F-m for the SA, MA and LA group of summaries under the 30% extraction rate are 72.12%, 72.14% and 72.83% showing a variance of 0.71%. This shows a closely similar performance of the system irrespective of the size of the original size of the article within the same extraction rate. The variance among the group of summaries is higher at the 20% extraction rate giving a variance of 11.08% (61.26%-50.18%) among the average performance measure for the three groups.

The highest average F-m of all the 9 group of summaries is 72.83% observed for the LA group summaries at the 30% extraction rate. The next two good average performance measures are also observed at the 30% extraction rate, for the MA group of article summaries 72.14% and for the SA group of article summaries 72.12%. The lower average performance measure 30.67% is observed for the SA group of summaries at the 10% extraction rate. It is so visible that the automatic summaries are of good performance for all the groups at 30% extraction rate. From the evaluation results, we can see that the performance measure for the summaries increases as the extraction rate increases. This proves the basic principle that the average performance increased as the extraction rate increased.

During this experiment, the average performance measure of the corpus increased by 22.25% and 15.69% as the extraction rate changes from 10% to 20% and then to 30%. That is the F-m increased at a decreasing rate as the extraction percentage increases constantly. The average of the corpus result for the three extraction rates 10%, 20% and 30% gave 34.42%, 56.68% and 72.37% respectively. This shows that the concept of a given article is best covered at the higher extraction rates. Table 6.5, shows the summary of these discussed F-m scores observed during the objective evaluation of the 90 summaries extracted using experiment two by introducing Amharic stemmer into the OTS tool. The detail result observed for each summary during this experimentation is

presented in appendix XIII (A, B and C for 10%, 20% and 30% extraction rate respectively).

| Original File Name | F-Measure in % | | |
|---|---|---|---|
| | at 10% | at 20% | at 30% |
| SA1.txt | 33.33% | 50.00% | 30.00% |
| SA2.txt | 33.33% | 66.67% | 77.78% |
| SA3.txt | 25.00% | 62.50% | 81.82% |
| SA4.txt | 20.00% | 80.00% | 86.67% |
| SA5.txt | 50.00% | 71.43% | 63.64% |
| SA6.txt | 40.00% | 54.55% | 68.75% |
| SA7.txt | 40.00% | 87.50% | 90.91% |
| SA8.txt | 25.00% | 44.44% | 76.92% |
| SA9.txt | 20.00% | 55.56% | 71.43% |
| SA10.txt | 20.00% | 40.00% | 73.33% |
| **Average** | **30.67%** | **61.26%** | **72.12%** |
| MA1.txt | 66.67% | 38.46% | 73.68% |
| MA2.txt | 16.67% | 41.67% | 66.67% |
| MA3.txt | 14.29% | 26.67% | 54.55% |
| MA4.txt | 33.33% | 72.73% | 70.59% |
| MA5.txt | 66.67% | 58.33% | 72.22% |
| MA6.txt | 37.50% | 40.00% | 69.57% |
| MA7.txt | 50.00% | 72.73% | 82.35% |
| MA8.txt | 16.67% | 54.55% | 70.59% |
| MA9.txt | 28.57% | 42.86% | 76.19% |
| MA10.txt | 57.14% | 53.85% | 85.00% |
| **Average** | **38.75%** | **50.18%** | **72.14%** |
| LA1.txt | 25.00% | 75.00% | 76.00% |
| LA2.txt | 40.00% | 63.16% | 72.41% |
| LA3.txt | 40.00% | 57.89% | 62.07% |
| LA4.txt | 20.00% | 57.89% | 75.86% |
| LA5.txt | 60.00% | 55.00% | 76.67% |
| LA6.txt | 20.00% | 47.37% | 58.62% |
| LA7.txt | 25.00% | 50.00% | 79.17% |
| LA8.txt | 33.33% | 70.59% | 80.77% |
| LA9.txt | 44.44% | 52.94% | 73.08% |
| LA10.txt | 30.77% | 56.00% | 73.68% |
| **Average** | **33.85%** | **58.58%** | **72.83%** |
| **Corpus Average** | **34.42%** | **56.68%** | **72.37%** |

**Table 6.5 Objective evaluation performance measure results for experiment two**

### 6.3.4 Comparison of the two Experiments

The two experiments were implemented on 90 news articles in three groups of size. The major difference between the two experiments is the stemmer used in the OTS tool. For the first experiment, the Porter stemmer which is designed inside the original OTS is used by developing Amharic dictionary file that this Porter stemmer makes use of. The detail of this dictionary file is discussed in chapter 5 of this thesis. The second experiment on the other hand makes use of Amharic stemmer which is designed by Tessema Mindaye (2007). After including this stemmer, the Porter stemmer is removed together with the dictionary file from the OTS system. After generating 90 summaries for the 30 news articles selected, each summary is evaluated using a respective ideal summary. The same ideal summary corpus is used in both experiments to evaluate the summaries generated in each experiment.

During the experimentation the two systems (in the two experiments) showed different performance on efficiency and effectiveness. For efficiency, when we compare the time it take to extract a summary, the system in E1 is more efficient than the system in E2. For example, to extract summary for the news article under the file name “MA7”, it took 3 seconds in E1 and 16 seconds in E2 at 30% extraction rate.

The system in the first experiment is effective for middle size (MA) articles irrespective of extraction rates whereas the system in the second experiment is effective for 30% extraction rate irrespective of original size. The experiment two evaluation results show that the system using the Amharic stemmer performs with better consistency in a higher extraction rate irrespective of the size of the original article. The first experiment on the other hand is dependent on the size of the original article size and performed well for middle size articles. It also showed irregularity in performance as the extraction rate and the summary group changes. The average of the results obtained in E1 vary significantly as the size of the original article changes in a given extraction rate, which is closer among the results of E2.

As we can see in Table 6.6, at 10% extraction rate E1 performed better in all the article groups as 17.5%, 14.05% and 3.85% more than E2 for SA, MA and LA size articles. At 20% extraction rate E2 performed better, 11.55% for SA and 8.88% for LA more than E1

and E1 outperformed E2 for MA size articles by 11.99%. At 30% extraction rate E2 outperformed E1 for SA by 15.89% and for LA by 6.03%. While E1 performed better than E2 for MA group articles by 3.51% more. The corpus average shows that E1 scored better result at 10%, and E2 outperformed E1 at 20% and 30% extraction rate. The graphical representation of Table 6.6 that shows the line graph for the average performance results of the two experiments at the three extraction rates is presented in Figure 6.2.

| File Group | Average F-Measure in % for E1 and E2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | at 10% | | | at 20% | | | at 30% | | |
| | E1 | E2 | Difference | E1 | E2 | Difference | E1 | E2 | Difference |
| SA | 48.17% | 30.67% | -17.50% | 49.71% | 61.26% | 11.55% | 56.24% | 72.12% | 11.89% |
| MA | 52.80% | 38.75% | -14.05% | 62.17% | 50.18% | -11.99% | 75.65% | 72.14% | -4.51% |
| LA | 37.71% | 33.85% | -3.85% | 49.70% | 58.58% | 8.88% | 66.80% | 72.83% | 6.03% |
| Corps Average | 46.22% | 34.42% | -11.80% | 53.86% | 56.68% | 2.82% | 66.23% | 72.37% | 6.14% |

**Table 6.6 Average performance result comparison for E1 and E2**

Graphical representation of E1 and E2 at 10%, 20% and 30% extraction rates

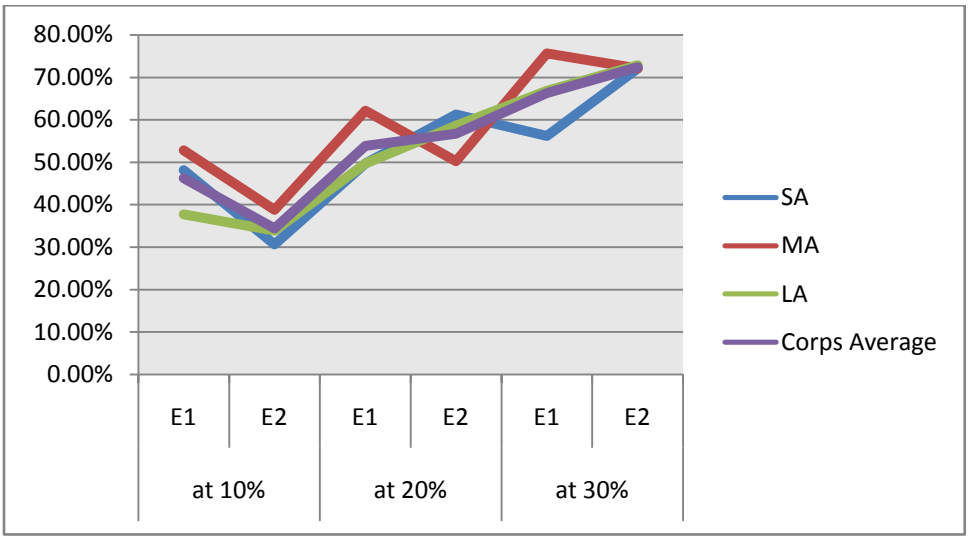


**Figure 6.2 Graph showing the performance measures comparison for E1 and E2**

From the graph we can see that all the three groups of summaries results for E2 are lower than E1 at 10% extraction rate, showing a downward slope. MA results show a chain saw under all extraction rates, due to the higher performance by E1 than E2 for that group. The graph at the tip merges since the performance results of E2 at 30% are closer to each other with lower variance while it is wider for E1 at 30%.

# CHAPTER SEVEN

# 7. CONCLUSION AND RECOMMENDATION

## 7.1 CONCLUSION

In this information era there is huge production of information in many formats. One of the medium to disseminate information is using text through digital and printed medium. The internet fueled up the mass production of information by providing easier technology to produce and disseminate information worldwide. This caused information overload on users. In order to get information from this mass production of text documents and reduce the information overload text summarization plays a major role. Summarizing the produced information manually is hard if not impossible due to the flooding production over the already available text document. Designing an automatic system to produce summaries for text documents has been attempted for many decades and successful results are obtained. Nowadays there are many text summarization systems developed for different languages. The researches done for the Amharic language text summarization also show promising results, proving the possibility of summarizing Amharic texts automatically.

In this research one of the well know text summarization system for single document summarization, the open text summarizer (OTS) is tested for its performance for Amharic text document summarization in the news domain. The system is tested with two different stemming approaches. Since Amharic is morphologically rich language, stemming is important process to achieve better performance. The first experiment is done using the Porter stemmer inside the OTS tool which uses a dictionary file that contains the Amharic language lexicons and punctuation rules. The different lists that are needed for the dictionary file are gathered from books and research papers. The second experiment is conducted by incorporating Amharic stemmer developed by Tessema (2007) in place of the Porter stemmer.

Amharic news article corpus containing 40 articles is gathered from news providing agencies (WIC, EPA and RANP), from which 30 are selected for the experimentation. These are grouped in to three based on their size in terms of sentences as SA, MA and LA. From the

test corpus of 30 news articles, 180 automatic summaries are generated using the two systems, 90 for each experiment at the three extraction rates 10%, 20% and 30%. The performance of the systems is evaluated by comparing the generated summaries with an ideal summary prepared manually by experts for objective evaluation. The same ideal summary is used to evaluate both experiments. For measuring the performance of the systems, F-measure is used as objective evaluation measures for the two experiments. Subjective evaluation using standard criteria is also done for fifteen summaries (10 from SA and 5 from MA group at 10%, 20% and 30%, with a total of 45 summaries) generated using the first system in experiment one. Though evaluation of text summarization systems is difficult, these commonly used subjective and objective evaluation methods are used in this research. The evaluation is difficult due to the fact that there is no perfect ideal summary to evaluate text summarization performance, since there could be many ideal summaries that can be prepared for a given text.

In experiment one (E1) the results obtained did not have pattern but showed dependency on both extraction rate and size of original article. The system is more effective for MA size articles having sentences in the 55 to 80 sentence range. The system should not be dependent on size of the article but rather on the extraction rate. In experiment two (E2), the system gave a proper pattern for the extraction. When we see the evaluation results at 30% extraction rate, the measures are closer to each other irrespective of the size of the original article. The system performance also increased as the extraction rate increases. These show that the performance of the system is better than E1, since any summarization system should perform irrespective of the size of the original article size and its performance should improve as the extraction rate increases.

However, the Amharic stemmer decreased the efficiency of the summarizer in terms of time it takes to generate a summary compared to the Porter stemmer using the Amharic lexicons. That is the system in E1 generates summaries much faster than the system in E2, making it more efficient in terms of time. This is due to the process involved with the Amharic stemmer used in the second experiment (E2). The Porter stemmer in E1 uses the dictionary file with language lexicons, removing affixes only in the stemming process. This takes a lot much shorter time than the stemming process in the Amharic stemmer.

When we see the corpus average F-measure, E1 scored 46.22%, 53.86% and 66.23% for 10%, 20% and 30% extraction rate whereas E2 scored 34.42%, 56.68% and 72.37%. From these scores we can observe that E2 outperformed E1 at 20% and 30% extraction rates whereas E1 performed better at 10% extraction rate. This shows using the Amharic stemmer resulted in better performing system that the Porter stemmer that uses the dictionary file. Moreover, the pattern observed in the performance measures by the second system in E2 (specially for the 30% extraction rates, the average scores 72.12%, 72.14% and 72.83% for the three groups SA, MA and LA, respectively) is found satisfactory to conclude that it performed better than the system in E1. The observed patters are, as the rate of extraction increases E2 showed better performance, giving highest performance at the 30% rate of extraction irrespective of the size of the original article. Whereas E1 performed better only for news articles in the MA group, E2 performed well for all size news articles. The system should be dependent on extraction rate but not on document size, where E2 gave better performance result than E1.

In order to answer the research questions, all the formulated objectives were achieved. The first research question: what are the characteristics of Amharic language relevant for text summarization? , was addressed by achieving the first three objectives. Different ambiguous characters of Amharic writings were identified and addressed using normalization as shown in section 5.4.3 of chapter 5. Moreover, since Amharic is a morphologically rich Semitic language, it responds highly to stemming. Therefore two types of stemming algorithms were used in the two experiments and the evaluation showed the system with the Amharic stemmer performed better regardless of the size of the original article, than the Porter stemmer version used originally inside the OTS. And also the Amharic news writing structure has been identified as diamond shape, having the most representative sentences at the beginning and at the end. This matches with the OTS since the system penalizes the sentences that are in the middle and give higher rank for the topic sentence. All these characteristics of Amharic writing have been implemented.

The second research question: how can the open text summarizer be used for Amharic news text summarization? , was answered by achieving the next three more objectives of the research. After identifying the requirement of the system, the tool was customized to support Amharic text summarization by including the Amharic writing systems for sentence parsing

and tokenization, the different language lexicons and normalization characters. As discussed above two different ways of stemming were also used. The customized systems were used to extract summaries for the 30 articles at different extraction rates. The evaluation of the systems showed very good results.

The automatic Amharic text summarization developed by Melese (2009) scored 47% F-measure using topic-LSA method and The customized OTS for Oromifa by Girma (2012) scored 81% F-measure (though it was tested using very short 8 news articles only). The OTS customized using Amharic stemmer scored average F-measure of 72.37% for the 90 summaries extracted. Even though the same corpus and evaluating summary has to be used to compare the different summarization systems developed, the results obtained show that very good performance has been obtained by the Amharic OTS.

## 7.2 RECOMMENDATION

In order to achieve better results in automatic summarization and enhance future researches in this area, the following recommendations are made on the basis of the observations in the process of conducting the research.

- A complete list of Amharic lexicons should be prepared and made available. The list of stop words, synonyms, abbreviations and compound words should be prepared in a standard manner for Amharic language. The dictionary file in the OTS tool makes use of these lists which improve the term weighting and sentence ranking processes of the system that is tested in E1.

- There are different researches done in Amharic language stemming but there is still no standard stemmer available to be used. For E2 a stemming algorithm developed by Tessema (2007) is used. Including this stemmer in the OTS tool made it perform better as observed in the evaluation results, performing independent of the size of the original article. However it was slower in terms of time compared to the Porter stemmer used in E1.

- This research experimented text summarization for Amharic text using OTS, further studies can be done for other Ethiopian written languages, other than Oromifa which is already done by Girma (2012), as well using this tool.
- Better results can be achieved for this experiment by including more Amharic lexicon rules and lists in to the dictionary file and using a better stemmer that performs Amharic stemming efficiently and uses dictionary to control over stemming.
- The exact performance of the two systems tested in this research could be measured in better quality, if a standard Amharic news text corpus and ideal summary is available. Therefore, text corpus and ideal summaries should be prepared for future research.
- An evaluation research can be done to compare the different systems developed so far for Amharic text summarization using the same text corpus and ideal summary to know which system performs better.
- Until the completion of this research, multi document text summarization for Amharic text is not attempted. This can be a future research area.
- Another future area can be multi language text summarization for local (Ethiopian) languages or local language with foreign languages like English.

# BIBLIOGRAPHY


Abiyot, B. (2000). Developing automatic word parser for Amharic verbs and their derivation. *Masteres thesis, Addis Ababa University, Addis Ababa*.

Adane, L. (2011). Feature Extraction and Matching in Amharic Document Image Collections.

Ager, S. (1998-2013). Omniglot - The Online Encyclopedia of Writing Systems and Languages Retrieved April 28,, 2013, from http://www.omniglot.com/

Anbessa, T., & Hudson, G. (2007). *Essentials of Amharic* (Vol. 18): Rüdiger Köppe.

Asker, L., Atelach, A. A., Gambäck, B., & Sahlgren, M. (2007). *Applying Machine Learning to Amharic Text Classification.* Paper presented at the Proceedings of the 5th World Congress of African Linguistics.

Asker, L., Atelach, A. A., Gambäck, B., Samuel, E. A., & Lemma, N. H. (2009). Classifying Amharic webnews. *Information retrieval, 12*(3), 416-435.

Atelach, A. A. (2002). *Automatic Sentence Parsing for Amharic Text: An Experiment using Probabilistic Context Free Grammars.* Master Thesis, Addis Ababa University.

Atelach, A. A., & Asker, L. (2007). *An Amharic stemmer: Reducing words to their citation forms.* Paper presented at the Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources.

Atelach, A. A., Asker, L., Cöster, R., & Karlgren, J. (2005). Dictionary-based Amharic–English information retrieval *Multilingual Information Access for Text, Speech and Images* (pp. 143-149): Springer.

Basagic, R., Krupic, D., & Suzic, B. (2009). Automatic Text Summarization. *Graz University of Technology*.

Baye, Y. (1998). *Root Reduction and Extension in Amharic'*. Paper presented at the Tenth Annual Conference of the Institute of Language Studies, Addis Ababa University.

Baye, Y. (2000 E.C.). *YäamarIña Säwasäw* (2nd ed.). Addis Ababa: EMPDE.

Bender, M. L. (1976). *Language in Ethiopia* (illustrated ed.). London: Oxford University Press.

Bender, M. L., & Hailu, F. (1978). *Amharic verb morphology: A generative approach*: African Studies Center, Michigan State University.

Bethlehem, M. (2002). *N-Gram-based Automatic Indexing for Amharic Text.* Masters Thesis. Addis Ababa University: Addis Ababa.

Boudin, F., Torres-Moreno, J.-M., & Velázquez-Morales, P. (2008). An efficient statistical approach for automatic organic chemistry summarization *Advances in Natural Language Processing* (pp. 89-99): Springer.

Daniel, G. A. (2003). An Integrated Approach to Automatic Complex Sentence Parsing for Amharic Text. *Unpublished Master'' s Thesis, School of Graduate Studies, Addis Ababa University*.

Dipanjan, D. (2007). A Survey on Automatic Text Summarization.

Dom Lachowicz, M. G., & Rotem, N. (2008). Open Text Summarizer, version 0.4. 2, libots group.

ELSRC, E. L. S. a. R. C. (2001 E.C.). YeAmargna Mezgebe Qalat - "የአማርኛ መዝገበ ቃላት": Addis Ababa University.

Eyob, D. Y., & Dejene, E. (2012). *Topic-based Amharic text summarization with probabilistic latent semantic analysis.* Paper presented at the Proceedings of the International Conference on Management of Emergent Digital EcoSystems.

Getahun, A. (1998). Zamanawi yaamarEna Sawasaw baqalal aqaararab: Commercial Printing Press: Addis Ababa.

Girma, D. D. (2012). *Afan Oromo News Text Summarizer.* Addis Ababa University, School of Graduate Studies, School of Information Science, Addis Ababa.

Gohr, A. (2013). Webinterface for Open Text Summarizer, April 11, 2013, from http://www.splitbrain.org/services/ots

Gupta, V., & Lehal, G. S. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence, 2*(3), 258-268.

Hahn, U., & Mani, I. (2000). The challenges of automatic summarization. *Computer, 33*(11), 29-36.

Hassel, M. (2004). Evaluation of Automatic Text Summarization. *Licentiate thesis, School of Computer Science and Communication, Royal Institute of Technology (KTH), published*.

Hassen, R., Tessema, M., & Solomon, A. (2009). *Search Engine for Amharic Web Content.* Paper presented at the IEEE AFRICON, Nairobi, Kenya.

Helen, A. (2006). *Automatic Text Summrizerfor Amharic Legal Judgemenet.* Addis Ababa University, School of Graduate Studies,, Addis Ababa.

Hovy, E. (2002). *Text Summarization*.

Hovy, E., & Lin, C.-Y. (1998). *Automated text summarization and the SUMMARIST system.* Paper presented at the Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998.

Hovy, E., & Marcu, D. (2005). Automated text summarization. *The Oxford Handbook of computational linguistics*, 583-598.

Jones, K. S. (1998). Automatic Summarizing Factors and Directions Advances in Automatic Text Summarization. *Cambridge MA: MIT Press, 1*(998), 675-685.

Kamil, N. (2004). *Automatic Amharic News Text Summarizer.* Master's Thesis, Addis Ababa University, Faculity of Informatics, Department Information Science,, Addis Ababa.

Karel, J., & Josef, S. (2007). Automatic Text Summarization (The State of the art 2007 and new challenges).

Lin, C.-Y., & Hovy, E. (2002). *Automated multi-document summarization in neats.* Paper presented at the Proceedings of the second international conference on Human Language Technology Research.

Lloret, E. (2008). Text Summarization: An Overview: Dept. Lenguajes y Sistemas Inform_aticos Universidad de Alicante Alicante, Spain.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development, 2*(2), 159-165.

Martha, Y. T. (2010). *Morphology-based language modeling for Amharic.* Doctoral Thesis, University of Hamburg, Hamburg, Germany.

Massih, R. A., Nicolas, U., & Gallinari, P. (2005). Automatic Text Summarization Based on Word - Clusters and Ranking Algorithms.

Melese, T. (2009). *Automatic Amharic Text Summarization using Latent Semantic Analysis.* Master Thesis, Addis Ababa University.

Mersehazen, W. (1934). *Yäamargna Säwasäw*. Addis Ababa: Birhanena Selam Printing Press.

Mesfin, G. (2001). *Automatic part of speech tagging for Amharic language: An experiment using stochastic HMM.* Master Thesis, Addis Ababa University.

Mullen, D. S. (1988). *Issues in the morphology and phonology of Amharic: The lexical generation of pronominal clitics*: University of Ottawa.

Nega, A., & Willett, P. (2002). Stemming of Amharic words for information retrieval. *Literary and Linguistic computing, 17*(1), 1-17.

Nega, A., & Willett, P. (2003). The effectiveness of stemming for information retrieval in Amharic. *Program: electronic library and information systems, 37*(4), 254-259.

Pembe, F. C. (2010). Automated Query-Biased and Structure-Preserving Document Summarization for Web Search Tasks.

Rotem, N. (2003, July, 2006). Open text summarizer (ots). *Retrieved July* 3. Retrieved March 3, 2013, from http://libots.sourceforge.net/

Seid, M. Y., & Mulugeta, L. (2009). TETEYEQ: Amharic Question Answering For Factoid Questions. *IE-IR-LRL, 3*(4), 17.

Sharan, A., Jain, N., & Imran, H. (2008). Automatic Text Summarization:by Extracting Significanant Sentence.

Sisay, F. A. (2005). *Part of speech tagging for Amharic using conditional random fields.* Paper presented at the Proceedings of the ACL workshop on computational approaches to semitic languages.

Solomon, A. (2011). Unsupervised Machine Learning Approach for Word Sense Disambiguationn to Amharic Words.

Surafel, T. (2012). Automatic Categorization Of Amharic News Text: A Machine Learning Approach.

Teferi, A. (2005). *The Application of Machine Learning Technique (Naive Bayes) for Automatic Text Summarization [The Case of Amharic News Texts]*. Addis Ababa University, School of Graduate studies, Faculty of Informatics, Department of Information Science, Addis Ababa.

Tesfaye, B. B. (2002). *Automatic morphological analyzer for Amharic: An experiment employing unsupervised learning and autosegmental analysis approaches.* Master's thesis, School of Graduate Studies of Addis Ababa University, Addis Ababa.

Tessema, M. (2007). *Design and Implementation of Amharic Search Engine.* Addis Ababa University, Addis Ababa, Ethiopia.

Theodros, G. (2003). Amharic text retrieval: An experiment using latent semantic indexing (LSI) with singular value decomposition (SVD). *Master of Science thesis, School of Information Studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia*.

Yatsko, V., & Vishnyakov, T. (2007). A method for evaluating modern systems of automatic text summarization. *Automatic Documentation and Mathematical Linguistics, 41*(3), 93-103.

Yohanes, A. (1990 E.C.). *Eski Teteyequ - "እሰኪ ተጠየቁ.."* Addis Ababa: Mega Printing Enterprise.

Zelalem, S. S. (2001). *Automatic classification of Amharic news items: The case of the Ethiopian News Agency.*

# APPENDIX

## I. List of Amharic characters, libiliazed characters and numbers

### A. Fidel

| | ā/ä [a] | u [u] | ī/i [i] | a [a] | ē/e [e/ɛ] | (i)/(ə) [ə] | 0 [o/ɔ] |
|---|---|---|---|---|---|---|---|
| H | ሀ | ሁ | ሂ | ሃ | ሄ | ህ | ሆ |
| [h] | ha | hu | Hi | ha | he | h(ə) | ho |
| L | ለ | ሉ | ሊ | ላ | ሌ | ል | ሎ |
| [l] | le | lu | Li | la | le | l(ə) | lo |
| h/ḥ | ሐ | ሑ | ሒ | ሓ | ሔ | ሕ | ሖ |
| [h] | ha | hu | Hi | ha | he | h(ə) | ho |
| M | መ | ሙ | ሚ | ማ | ሜ | ም | ሞ |
| [m] | me | mu | Mi | ma | me | m(ə) | mo |
| s/ś | ሠ | ሡ | ሢ | ሣ | ሤ | ሥ | ሦ |
| [s] | se | su | si | sa | se | s(ə) | so |
| R | ረ | ሩ | ሪ | ራ | ሬ | ር | ሮ |
| [r] | re | ru | ri | ra | re | r(ə) | ro |
| S | ሰ | ሱ | ሲ | ሳ | ሴ | ስ | ሶ |
| [s] | se | su | si | sa | se | s(ə) | so |
| sh/š | ሸ | ሹ | ሺ | ሻ | ሼ | ሽ | ሾ |
| [ʃ] | ʃe | ʃu | ʃi | ʃa | ʃe | ʃ(ə) | ʃo |
| k'/q | ቀ | ቁ | ቂ | ቃ | ቄ | ቅ | ቆ |
| [k'] | k'e | k'u | k'i | k'a | k'e | k'(ə) | k'o |
| B | በ | ቡ | ቢ | ባ | ቤ | ብ | ቦ |
| [b] | be | bu | bi | ba | be | b(ə) | bo |
| T | ተ | ቱ | ቲ | ታ | ቴ | ት | ቶ |
| [t] | te | tu | ti | ta | te | t(ə) | to |
| ch/č | ቸ | ቹ | ቺ | ቻ | ቼ | ች | ቾ |
| [ʧ] | ʧe | ʧu | ʧi | ʧa | ʧe | ʧ(ə) | ʧo |
| h/ḫ | ኀ | ኁ | ኂ | ኃ | ኄ | ኅ | ኆ |
| [h] | ha | hu | hi | ha | he | h(ə) | ho |
| n | ነ | ኑ | ኒ | ና | ኔ | ን | ኖ |
| [n] | ne | nu | ni | na | ne | n(ə) | no |
| ny/ñ | ኘ | ኙ | ኚ | ኛ | ኜ | ኝ | ኞ |
| [ɲ] | ɲe | ɲu | ɲi | ɲa | ɲe | ɲ(ə) | ɲo |
| ’/’ | አ | ኡ | ኢ | ኣ | ኤ | እ | ኦ |
| [ʔ] | (ʔ)a | (ʔ)u | (ʔ)i | (ʔ)a | (ʔ)e | (ʔ)(ə) | (ʔ)o |
| k | ከ | ኩ | ኪ | ካ | ኬ | ክ | ኮ |
| [k] | ke | ku | ki | ka | ke | k(ə) | ko |

| | ā/ä [a] | u [u] | ī/i [i] | a [a] | ē/e [e/ɛ] | (i)/(ə) [ə] | o [o/ɔ] |
|---|---|---|---|---|---|---|---|
| h/ḵ | ኸ | ኹ | ኺ | ኻ | ኼ | ኽ | ኾ |
| [h] | he | hu | hi | ha | he | h(ə) | ho |
| w | ወ | ዉ | ዊ | ዋ | ዌ | ው | ዎ |
| [w] | we | wu | wi | wa | we | w(ə) | wo |
| ‘/‘ | ዐ | ዑ | ዒ | ዓ | ዔ | ዕ | ዖ |
| [ʔ] | ʔa | ʔu | ʔi | ʔa | ʔe | ʔ(ə) | ʔo |
| z | ዘ | ዙ | ዚ | ዛ | ዜ | ዝ | ዞ |
| [z] | Ze | zu | zi | za | ze | z(ə) | zo |
| zh/ž | ዠ | ዡ | ዢ | ዣ | ዤ | ዥ | ዦ |
| [ʒ] | ʒe | ʒu | ʒi | ʒa | ʒe | ʒ(ə) | ʒo |
| y | የ | ዩ | ዪ | ያ | ዬ | ይ | ዮ |
| [j] | Je | ju | ji | ja | je | j(ə) | jo |
| d | ደ | ዱ | ዲ | ዳ | ዴ | ድ | ዶ |
| [d] | De | du | di | da | de | d(ə) | do |
| j/ǧ | ጀ | ጁ | ጂ | ጃ | ጄ | ጅ | ጆ |
| [ʤ] | ʤe | ʤu | ʤi | ʤa | ʤe | ʤ(ə) | ʤo |
| g | ገ | ጉ | ጊ | ጋ | ጌ | ግ | ጎ |
| [g] | Ge | gu | gi | ga | ge | g(ə) | go |
| t'/ṭ | ጠ | ጡ | ጢ | ጣ | ጤ | ጥ | ጦ |
| [t'] | t'e | t'u | t'i | t'a | t'e | t'(ə) | t'o |
| ch'/ç | ጨ | ጩ | ጪ | ጫ | ጬ | ጭ | ጮ |
| [ʧ'] | ʧ'e | ʧ'u | ʧ'i | ʧ'a | ʧ'e | ʧ'(ə) | ʧ'o |
| p'/ṗ | ጰ | ጱ | ጲ | ጳ | ጴ | ጵ | ጶ |
| [p'] | p'e | p'u | p'i | p'a | p'e | p'(ə) | p'o |
| ts'/ṣ | ጸ | ጹ | ጺ | ጻ | ጼ | ጽ | ጾ |
| [ts] | ts'e | ts'u | ts'i | ts'a | ts'e | ts'(ə) | ts'o |
| ts'/ś | ፀ | ፁ | ፂ | ፃ | ፄ | ፅ | ፆ |
| [ts] | ts'e | ts'u | ts'i | ts'a | ts'e | ts'(ə) | ts'o |
| f | ፈ | ፉ | ፊ | ፋ | ፌ | ፍ | ፎ |
| [f] | Fe | fu | fi | Fa | fe | f(ə) | fo |
| p | ፐ | ፑ | ፒ | ፓ | ፔ | ፕ | ፖ |
| [p] | Pe | pu | pi | Pa | pe | p(ə) | po |
| v | ቨ | ቩ | ቪ | ቫ | ቬ | ቭ | ቮ |
| [v] | Ve | vu | vi | Va | ve | v(ə) | vo |

B. libiliazed characters

| ቀ፟ | ቋ | ቈ | ቀ፟ | ኈ | ኋ | ኌ | ኈ | ኩ | ካ | ኴ | ኸ | ጐ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k’ʷi | k’ʷa | k’ʷe | k’ʷi | hʷi | hʷa | hʷe | hʷi | kʷi | kʷa | kʷe | kʷi | gʷi |
| ጓ | ጔ | ጕ | ሏ | ቧ | ዟ | ጧ | ሟ | ቷ | ኟ | ቈ | ሯ | ቿ |
| gʷa | gʷe | gʷi | lʷa | bʷa | zʷa | t’ʷa | mʷa | tʷa | ʒʷa | ɡ’ʷa | rʷa | ɡʷa |
| ጇ | ጟ | ሷ | ኗ | ዷ | ፏ | ኍ | ኯ | ሯ | ሻ | ሬ | ኧ | |
| ʤʷa | ts’ʷa | sʷa | nʷa | dʷa | fʷa | ʃʷa | ɲʷa | rʲa | nʲa | fʲa | E | |

C. Amharic Pronunciation

| ሀ | ለ | ሐ | መ | ሠ | ረ | ሰ | ሸ | ቀ | ቈ | በ | ተ | ቸ | ኀ | ኈ | ነ | ኘ | አ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| h | l | h | m | ś | R | s | š | q | qu | b | t | č | ẖ | ẖu | n | ñ | ʾ |
| [ h ] | [ l ] | [ h ] | [ m ] | [ s ] | [ r ] | [ s ] | [ ʃ ] | [ k’ ] | [ kʷ ] | [ b ] | [ t ] | [ ʧ ] | [ h ] | [ hʷ ] | [ n ] | [ ɲ ] | [ ʔ ] |
| | | | | | | | | | | | | | | | | | |
| ከ | ኸ | ወ | ዐ | ዘ | ዠ | የ | ደ | ጀ | ገ | ጐ | ጠ | ጨ | ጰ | ጸ | ፀ | ፈ | ፐ |
| k | h | w | ʿ | z | Ž | y | d | ǧ | g | gu | ṭ | č̣ | ṗ | ṣ | ẓ | f | P |
| [ k ] | [ h ] | [ w ] | [ ʔ ] | [ z ] | [ ʒ ] | [ j ] | [ d ] | [ ʤ ] | [ g ] | [ gʷ ] | [ t’ ] | [ ʧ’ ] | [ p’ ] | [ ts ] | [ ts ] | [ f ] | [ p ] |

D. Amharic Vowels

| ā | U | i | A | e | ə | o |
|---|---|---|---|---|---|---|
| [ a ] | [ u ] | [ i ] | [ a ] | [ e/ɛ ] | [ ə ] | [ o/ɔ ] |

E. Amharic Numbers

| ፩ | ፪ | ፫ | ፬ | ፭ | ፮ | ፯ | ፰ | ፱ | ፲ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | | | | | | | | | |
| ፳ | ፴ | ፵ | ፶ | ፷ | ፸ | ፹ | ፺ | ፻ | ፼ |
| 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 10,000 |

## II. List of Affixes

A. Prefix

| Individual Prefixes | | | Combined Prefixes | | |
|---|---|---|---|---|---|
| በ | ለ | እየ | እንደ | እስክት | በየ |
| ከ | ይ | እን | እንዲ | እስክን | ከየ |
| የ | መ | የሚ | እስክ | እንደየ | የእነ |
| ሲ | ስለ | ክነ | እንድ | እንድን | በእነ |
| ን | እነ | | እስኪ | ለእነ | ከእነ |

B. Suffix

| Individual Suffixes | | | Combined Suffixes | | |
|---|---|---|---|---|---|
| ን | ያል | ላችሁ | ዎች | ዎቹን | ምና |
| ና | የዋ | ላቸው | ዎችን | ናም | ምናም |
| ም | ቸው | ባቸው | ዎችንና | ናምና | ምናን |
| ች | ተው | ባችሁ | ዎችንም | ናምን | ምቹ |
| ዬ | የው | | ዎችና | ናን | ምን |
| ዎ | በት | | ዎችናም | ናንና | ምንና |
| ህ | ባት | | ዎችናን | ናንም | ምንም |
| ሽ | ነት | | ዎችም | ናንን | ምንን |
| ዋ | ለት | | ዎችምና | ናው | ምዉና |
| ቹ | ላት | | ዎቹና | ናዉና | ምዉም |
| ው | ይቱ | | ዎቹናም | ናውና | ምውም |
| ቱ | ያዊ | | ዎቹናን | ናዉም | ምዉን |
| ሁ | ኞች | | ዎቹም | ናውም | ዉም |
| ኻ | ዎቹ | | ዎቹምና | ናዉን | ዉምና |
| ኛ | አዊ | | ዎቹምን | ንና | ያውያኑ |
| ዊ | | | ዉናን | ዉና | ያውያንና |
| ት | | | ዉናዉ | ዉናም | ያውያን |
| | | | ንም | ንን | ንናም |
| | | | ንምን | ንምና | ንምዉ |
| | | | ችን | ችው | ናና |

## III. Normalization List

A. List of compound words

| Compound Words | Normalized Form | | Compound Words | Normalized Form |
|---|---|---|---|---|
| ቤተ ክርስቲያን | ቤተክርስቲያን | | ወጥ ቤት | ወጥቤት |
| መካነ እየሱስ | መካነየሱስ | | መፀሃፍ ቅዱስ | መፀሃፍቅዱስ |
| ቤተ መንግስት | ቤተመንግስት | | ድረ-ገፅ | ድረገፅ |
| ቤተ ውበት | ቤተውበት | | መካነ መቃብር | መካነመቃብር |
| ቤተ መፀሃፍት | ቤተመፀሃፍት | | ቀይ መስቀል | ቀይመስቀል |
| ህገ መንግስት | ህገመንግስት | | በትረ መንግስት | በትረመንግስት |
| ኃይለ ማርያም | ኃይለማርያም | | ኮልፌ ቀራንዮ | ኮልፌቀራንዮ |
| ኃይለ ስላሴ | ኃይለስላሴ | | ንፋስ ስልክ | ንፋስስልክ |
| ክፍለ ከተማ | ክፍለከተማ | | ክፍለ ሃገር | ክፍለሃገር |
| አዲስ ከተማ | አዲስከተማ | | አፍሪካ ህብረት | አፍሪካህብረት |
| አዲስ አበባ | አዲስአበባ | | ስነ ፍጥረት | ስነፍጥረት |
| ስነ ስርዓት | ስነስርዓት | | ስነ ምግባር | ስነምግባር |

B. List of Normalizing Characters

| Same Sound Characters | Normalizing Character |
|---|---|
| ሀ፣ሐ፣ኅ ፤ ሁ፣ሑ፣ኁ ፤ ሂ፣ሒ፣ኂ ፤ ሃ፣ሓ፣ኃ ፤ ሄ፣ሔ፣ኄ ፤ ህ፣ሕ፣ኅ ፤ ሆ፣ሖ፣ኆ | ሀ ፤ ሁ ፤ ሂ ፤ ሀ ፤ ሄ ፤ ህ ፤ ሆ |
| ሠ፣ሰ ፤ ሡ፣ሱ ፤ ሢ፣ሲ ፤ ሣ፣ሳ ፤ ሤ፣ሴ ፤ ሥ፣ስ ፤ ሦ፣ሶ | ሰ ፤ ሱ ፤ ሲ ፤ ሳ ፤ ሴ ፤ ስ ፤ ሶ |
| አ፣ዐ ፤ ኡ፣ዑ ፤ ኢ፣ዒ ፤ ኣ፣ዓ ፤ ኤ፣ዔ ፤ እ፣ዕ ፤ ኦ፣ዖ | አ ፤ ኡ ፤ ኢ ፤ ኣ ፤ ኤ ፤ እ ፤ ኦ |
| ጸ፣ፀ ፤ ጹ፣ፁ ፤ ጺ፣ፂ ፤ ጻ፣ፃ ፤ ጼ፣ፄ ፤ ጽ፣ፅ ፤ ጾ፣ፆ | ፀ ፤ ፁ ፤ ፂ ፤ ፃ ፤ ፄ ፤ ፅ ፤ ፆ |
| ኮ፣ኰ | ኮ |
| ጎ፣ጐ | ጐ |
| ዉ ፤ ው | ው |
| ። ፤ ፡ | ፡ |
| “ ” Double space between two consecutive words in one sentence | “ ” Changed to Single Space |

C. List of words with different writing pattern

| English Word | Transliteration Form 1 | Transliteration Form 2 |
|---|---|---|
| Oxford | አክስፎርድ | አክስፈርድ |
| Sponsor | ስፖንሰር | እስፖንሰር |
| Stadium | ስታድዮም | እስታድየም |
| Sport | ስፖርት | እስፖርት |
| Television | ቴሌቪሾን | ቴለቭሾን፣ቴሌቭዢን |
| European | ኤሮፓውያን | አውሮፓውያን |
| Automobile | አውቶሞቢል | ኦቶሞቢል |
| Automatic | አውቶማቲክ | አቶማቲክ |
| Encyclopedia | ኢንሳይክሎፒድያ | ኢንሳይክሎፕድያ፣እንሳይክሎፔድያ |
| Technology | ቴክኖሎጂ | ቴክኒዎሎጂ |
| Millennium | ሚሊኒየም | ሚሊንየም፣ሚለኒየም |
| Computer | ኮምፒዩተር | ኮምፒውተር |
| Email | ኢሜል | ኢሜይል |
| Some Amharic Words with the same behavior | ሰማኒያ | ሰማንያ |
| | ሚያዚያ | ሚያዝያ |

## IV. List of Common Amharic Abbreviations and their expanded forms

| Abbreviation | Expanded Form | Abbreviation | Expanded Form |
|---|---|---|---|
| ት/ቤት | ትምህርት ቤት | ተ/ሀይማኖት | ተክለ ሀይማኖት |
| ት/ርት | ትምህርት | ሚ/ር | ሚኒስትር |
| ት/ክፍል | ትምህርት ክፍል | ኮ/ል | ኮለኔል |
| ኃ/አለቃ | ሃምሳ አለቃ | ሜ/ጀነራል | ሜጀር ጄኔራል |
| ኃ/ስላሴ | ኃይለ ስላሴ | ብ/ጄኔራል | ብርጋዴር ጄኔራል |
| ደ/ዘይት | ደብረ ዘይት | ሌ/ኮለኔል | ሌተናል ኮለኔል |
| ደ/ታቦር | ደብረ ታቦር | ሊ/መንበር | ሊቀ መንበር |
| መ/ር | መምህር | አ/አ/ | አዲስ አበባ |
| መ/ቤት | መስሪያ ቤት | ር/መምህር | ርእሰ መምህር |
| መ/አለቃ | መቶ አለቃ | ፕ/ት | ፕሬዝዳንት |
| ክ/ከተማ | ክፍለ ከተማ | ዓ.ም | ዓመተ ምህረት |
| ክ/ሀገር | ክፍለ ሀገር | ፍ/ቤት | ፍርድ ቤት |
| ወ/ር | ወታደር | ፅ/ቤት | ፅህፈት ቤት |
| ወ/ሮ | ወይዘሮ | ሲ/ር | ሲስተር |
| ወ/ት | ወይዘሪት | ፕ/ር | ፕሮፌሰር |
| ኃ/ማርያም | ኃይለማርያም | መ. ቅ. | መፅሃፍ ቅዱስ |
| ተ/ወልድ | ተክለወልድ | አ.አ | አዲስ አበባ |
| ወ/ስለሴ | ወልደ ስለሴ | ዶ.ር | ዶክተር |
| ፍ/ስላሴ | ፍቅረ ስላሴ | ዓ. ዓ. | ዓመተ ዓለም |
| ጠ/ሚኒስትር | ጠቅላይ ሚኒስትር | እ. ኤ. አ. | እንደ አውሮፓውያን አቆጣጠር |
| ዶ/ር | ዶክተር | አ.ህ. | አፍሪካ ህብረት |
| ገ/ጊዮርጊስ | ገብረ ጊዮርጊስ | የተ.መ.ድ. | የተባበሩት መንግስታት ድርጅት |
| ቤ/ክርስትያን | ቤተ ክርስትያን | ተ.እ.ታ. | ተጨማሪ እሴት ታክስ |
| ም/ስራ | ምክትል ስራ | | |
| ም/ቤት | ምክር ቤት | | |

## V. List of Some Common Amharic Synonyms

| Synonym | Word | Synonym | Word | Synonym | Word | Synonym | Word |
|---|---|---|---|---|---|---|---|
| መሰናዶ | ዝግጅት | ተሃድሶ | መታደስ | ተደመደመ | ተፈፀመ | ፕላን | እቅድ |
| አንጋፋ | ታላቅ | መምህር | አስተማሪ | ጨበጠ | ያዘ | ፈነጠዘ | ተደሰተ |
| ተጓዘ | ሄደ | ትንታግ | ፍም | አስገነዘበ | አሳወቀ | ጫረ | ለኮሰ |
| ቀድሞ | በፊት | አማለለ | አታለለ | አፈነገጠ | ተለየ | ወንጀል | ጥፋት |
| መንታ | ጥንድ | አገተ | ከለከለ | ፅንስ | ውጥን | ተጣመረ | ተባበረ |
| ረግረግ | ማጥ | አፍላ | ጎረምሳ | ጫና | ጭቆና | ገለልተኛ | ብቸኛ |
| ሙግት | ክርክር | እንግጫ | ሳር | ጠፈር | ሰማይ | ዶሴ | ማህደር |
| መፃተኛ | እንግዳ | የሳርቤት | ጎጆቤት | ተጎነጨ | ቀመሰ | አጠናቀረ | አዘጋጀ |
| ሙስና | ጥፋት | ወየበ | ገረጣ | ጀምበር | ፀሃይ | ዘገበ | አቀረበ |
| ምስካይ | መጠጊያ | ዋጀ | አዳነ | ዳኘ | ፈረደ | ውል | ስምምነት |
| ምናብ | ልቦና | ብሄር | ዘር | ዘለፈ | ነቀፈ | ወጠምሻ | ደንዳና |
| ዜማ | ሙዚቃ | ገረሰሰ | ጣለ | ተዋሃደ | ተቀላቀለ | ወሰለተ | ዋሸ |
| ጣዕም | ቃና | ፀዳል | ብርሃን | ተወዳጅ | ተፈቃሪ | አለኝታ | መመኪያ |
| ግዳይ | ሰለባ | ፅሞና | ዝምታ | ኮበለለ | ከዳ | አወከ | ረበሸ |
| ሰረፀ | ገባ | መንበር | መቀመጫ | ከተተ | ዘመተ | አከለ | ጨመረ |
| ሰቆቃ | እንባ | ሞገደኛ | ሃይለኛ | ታፈነ | ተዘጋ | አወጀ | ነገረ |
| ሰናይ | መልካም | አሻ | ፈለገ | ሁናቴ | አዝማሚያ | ቄብሳይት | ድረገፅ |
| ሸረበ | ገመደ | ነገረፈጅ | ጠበቃ | አወላወለ | አመነታ | ትርኢት | መዝናኛ |
| ቀመረ | ቆጠረ | ተከራካሪ | ተሟጋች | ማብራሪያ | ገለፃ | አቢሲንያ | ኢትዮጵያ |
| ቀነጨበ | ቆነጠረ | ናኘ | ታወቀ | ከንውን | አፈፃፀም | | |
| በኩር | መጀመሪያ | ረታ | አሸነፈ | ዱር | ጫካ | | |
| ብሂል | ትርጉም | ምግባር | ባህሪ | አበቃ | አለቀ | | |
| ባዘነ | ባከነ | ተፈጠፈጠ | ወረደ | ተተረከ | ተነገረ | | |
| ብላቴና | ወጣት | ገሃድ | ግልፅ | በረበረ | ፈተሸ | | |

## VI. List of stop words

| | | | |
|---|---|---|---|
| ሁሉ | ሲሉ | ተከናውኗል | አስረድተዋል |
| ሁሉም | ስለ | ችግር | ምክንያት |
| ኋላ | ቢቢሲ | ታች | እባክህ |
| ሁኔታ | ቢሆን | ትናንት | እባክሽ |
| ሆነ | ብለዋል | ነበረ | እባክዎ |
| ሆኑ | ብቻ | ነበረች | አንድ |
| ሆኖም | ብዛት | ነበሩ | አንፃር |
| ሁሌ | ብዙ | ነው | እስኪደርስ |
| ሁልጊዜ | ቦታ | ነይ | እንኳ |
| ሁሉንም | በርካታ | ነገር | እንኳን |
| ላይ | በሰሞኑ | ነገሮች | እስከ |
| ሌላ | በታች | ናት | እዚሁ |
| ሌሎች | በኋላ | ናቸው | እና |
| ልዩ | በኩል | አሁን | እንደ |
| መሆኑ | በውስጥ | አለ | እንደገና |
| ማለት | በጣም | አስታወቀ | እንደገለፁት |
| ማለቱ | ይህን | አስታውቀዋል | እንደተገለፀው |
| መካከል | በተለይ | አስታውሰዋል | እንደተናገሩት |
| የሚገኝ | በተመለከተ | እስካሁን | እንደአስረዱት |
| የሚገኙ | በተመሳሳይ | አሳሰበ | እንደቀረበው |
| ማድረግ | የተለያዩ | አሳስበዋል | እንደታየው |
| ማን | የተለያየ | አስፈላጊ | እንደተነገረው |
| ማንም | ተባለ | አስገነዘቡ | እንደተብራራው |
| ስሞኑን | ተገለፀ | አስገንዝበዋል | እንዳሉት |
| ሲሆን | ተገልጿል | አብራርተዋል | ወቅት |
| ሲል | ተጨማሪ | አብራርተው | እንዲሁም |

እንጂ ያሉ ይገልፃል ከፍተኛ

እዚህ ይገባል ሲሉ በከፍተኛ

እዚያ የኋላ ብለዋል የሚል

እያንዳንዱ የሰሞኑ ስለሆነ በላይ

እያንዳንዳቸው የታች አቶ ሳይሆን

እያንዳንዷ የውስጥ ሆኖ ሆነው

ከዚህ የውጭ መገለፁን ምን

ከኋላ የጋራ መግለፁን ምንም

ከላይ ያ አመልክተዋል ለምን

ከመሃል ይታወሳል ይናገራሉ ዓይነት

ከመካከል ይህ ይህም ይሁን

ከሰሞኑ ደግሞ በአንድ በዚህ

ከታች ድረስ ጉዳይ አለበት

ከውስጥ ስለዚህ ነበር ያህል

ከጋራ ግን አይደለም እነዚህ

ከፊት ገልጿል ደረጃ በእዚህ

ጋራ ገልፀዋል ሥራ ይችላል

ወዘተ ግዜ የሥራ አንዳንድ

ወይም ጊዜ ዙሪያ ታዲያ

ወደ ጥቂት እናም እንዲህ

ዋና ፊት ያለው ይቻላል

ወደፊት እንደሆነ ያለውን በሚል

ውስጥ ዛሬ መሆኑን የሆነ

ውጭ ጋር በመሆኑ በፊት

ውጪ ተናግረዋል መሆን ስለሆነም

ያለ የገለፁት ብሎ በማለት እንኳን

## VII. Dictionary Rules – “am.xml”

```
<?xml version="1.0" encoding="UTF-8"?>
<dictionary lang="Amharic">
  <stemmer>
  <step1_pre>
        <rule>"|</rule>
        <rule>(|</rule>
        <rule>'|</rule>
        <rule>፡|</rule>
        <rule>\s|</rule>
        <rule> |</rule>
        <rule>«|</rule>
        <rule>|</rule>
        <rule>-|</rule>
        </step1_pre>
    <step1_post>
        <rule>።"|</rule>
        <rule>,"|</rule>
        <rule>."|</rule>
        <rule>"።|</rule>
        <rule>፡"|</rule>
        <rule>'።|</rule>
        <rule>።'|</rule>
        <rule>፡'|</rule>
        <rule>.'|</rule>
        <rule>,'|</rule>
        <rule>፤|</rule>
        <rule>፡-|</rule>
        <rule>፣|</rule>
        <rule>፥|</rule>
        <rule>"|</rule>
        <rule>'|</rule>
        <rule>»|</rule>
        <rule>.|</rule>
        <rule>..|</rule>
        <rule>...|</rule>
        <rule>,|</rule>
        <rule>"|</rule>
        <rule>")|</rule>
        <rule>)|</rule>
        <rule>?|</rule>
        <rule>፡|</rule>
        <rule> |</rule>
        <rule>\s|</rule>
        <rule>;|</rule>
        <rule>!|</rule>
        <rule>-|</rule>
        <rule>--|</rule>
        </step1_post>
    <manual>
<rule>ከሰከሰው|ከሰከሰ</rule>
<rule>ከፋፈለ|ከፍል</rule>
<rule>ከፍልፋይ|ከፍል</rule>
<rule>ከመረ|ከምር</rule>
<rule>መግለፅ|ግልፅ</rule>
<rule>የሚገኝ|ግኝት</rule>
<rule>የሚገኙ|ግኝት</rule>
</manual>
<pre>
  <rule>|</rule>
        <rule>እንደየ|</rule>
        <rule>እንድን|</rule>
        <rule>እስኪ|</rule>
        <rule>እንድ|</rule>
        <rule>እንዲ|</rule>
        <rule>እስከ|</rule>
        <rule>እስክት|</rule>
        <rule>እስክን|</rule>
        <rule>እንደ|</rule>
        <rule>ከእነ|</rule>
        <rule>የእነ|</rule>
        <rule>በእነ|</rule>
        <rule>ለእነ|</rule>
        <rule>እየ|</rule>
        <rule>የሚ|</rule>
        <rule>ስለ|</rule>
        <rule>በየ|</rule>
        <rule>ከነ|</rule>
        <rule>እን|</rule>
        <rule>እነ|</rule>
        <rule>ከየ|</rule>
        <rule>ከ|</rule>
        <rule>መ|</rule>
        <rule>ን|</rule>
        <rule>በ|</rule>
```

<rule>ለ|</rule>
<rule>ይ|</rule>
<rule>ሲ|</rule>
<rule>የ|</rule>
</pre>
<post>
<rule>|</rule>
<rule>ቹ|</rule>
<rule>ም|</rule>
<rule>ት|</rule>
<rule>ን|</rule>
<rule>ዬ|</rule>
<rule>ዎ|</rule>
<rule>ህ|</rule>
<rule>ዋ|</rule>
<rule>ችን|</rule>
<rule>ተው|</rule>
<rule>ወ|</rule>
<rule>ውና|</rule>
<rule>ውናው|</rule>
<rule>ውናን|</rule>
<rule>ውናም|</rule>
<rule>ውምና|</rule>
<rule>ውም|</rule>
<rule>ሀ|</rule>
<rule>ና|</rule>
<rule>ናም|</rule>
<rule>ናምና|</rule>
<rule>ናምን|</rule>
<rule>ና|</rule>
<rule>ናና|</rule>
<rule>ናንም|</rule>
<rule>ናንን|</rule>
<rule>ናና|</rule>
<rule>ናው|</rule>
<rule>ናውና|</rule>
<rule>ናንም|</rule>
<rule>ናውም|</rule>
<rule>ናዉና|</rule>
<rule>ናዉም|</rule>
<rule>ናውን|</rule>
<rule>ንን|</rule>
<rule>ንና|</rule>
<rule>ንናም|</rule>
<rule>ንምና|</rule>
<rule>ንምን|</rule>
<rule>ንም|</rule>
<rule>ንምዉ|</rule>
<rule>ምዉና|</rule>
<rule>ምዉም|</rule>
<rule>ምዉን|</rule>
<rule>ምንን|</rule>
<rule>ምና|</rule>
<rule>ምናም|</rule>
<rule>ምናን|</rule>
<rule>ምቹ|</rule>
<rule>ምን|</rule>
<rule>ምንም|</rule>
<rule>ምንና|</rule>
<rule>ነት|</rule>
<rule>ለት|</rule>
<rule>ላት|</rule>
<rule>ላቸዉ|</rule>
<rule>በት|</rule>
<rule>ባት|</rule>
<rule>ባቸዉ|</rule>
<rule>ባችሁ|</rule>
<rule>ቱ|</rule>
<rule>ይቱ|</rule>
<rule>ዎቹ|</rule>
<rule>ዎቹን|</rule>
<rule>ዎቹና|</rule>
<rule>ዎቹናም|</rule>
<rule>ዎቹናን|</rule>
<rule>ዎቹም|</rule>
<rule>ዎቹምና|</rule>
<rule>ዎቹምን|</rule>
<rule>ዎች|</rule>
<rule>ዎችን|</rule>
<rule>ዎችም|</rule>
<rule>ዎችናን|</rule>
<rule>ዎችንና|</rule>
<rule>ዎችንም|</rule>
<rule>ዎችና|</rule>
<rule>ዎችናም|</rule>
<rule>የዉ|</rule>
<rule>ኞች|</rule>

<rule>ያል|</rule>
<rule>ኛ|</rule>
<rule>ችዉ|</rule>
<rule>የዋ|</rule>
<rule>ች|</rule>
<rule>ሽ|</rule>
<rule>ላችሁ|</rule>
<rule>ዊ|</rule>
<rule>አዊ|</rule>
<rule>ያዊ|</rule>
<rule>ያውያን|</rule>
<rule>ያውያንና|</rule>
<rule>ያውያኑ|</rule>
<rule>ችው|</rule>
<rule>ኝ|</rule>
</post>
<synonyms>
<rule>ዌብሳይት|ድረገፅ</rule>
<rule>መሰናዶ|ዝግጅት</rule>
<rule>አንጋፋ|ታላቅ</rule>
<rule>ተጓዘ|ሄደ</rule>
<rule>ቀድሞ|በፊት</rule>
<rule>መንታ|ጥንድ</rule>
<rule>ረድኤት|በረከት</rule>
<rule>ሙግት|ክርክር</rule>
<rule>መፃተኛ|እንግዳ</rule>
<rule>ሙስና|ጥፋት</rule>
<rule>ምስካይ|መጠጊያ</rule>
<rule>ምናብ|ልቦና</rule>
<rule>ዜማ|ሙዚቃ</rule>
<rule>ጣዕም|ቃና</rule>
<rule>ግዳይ|ሰለባ</rule>
<rule>ሰረፀ|ገባ</rule>
<rule>ሰቆቃ|እንባ</rule>
<rule>ሰናይ|መልካም</rule>
<rule>ሸረበ|ገመደ</rule>
<rule>ቀመረ|ቆጠረ</rule>
<rule>ቀነጨበ|ቆነጠረ</rule>
<rule>በኩር|መጀመሪያ</rule>
<rule>ብሂል|ትርጉም</rule>
<rule>ባድመ|ባዶ</rule>
<rule>ተሃድሶ|መታደስ</rule>
<rule>መምህር|አስተማሪ</rule>
<rule>ትንታግ|ፍም</rule>
<rule>አማለለ|አታለለ</rule>
<rule>አገተ|ያዘ</rule>
<rule>አፍላ|ጎረምሳ</rule>
<rule>እንግጫ|ሳር</rule>
<rule>የሳርቤት|ጎጆቤት</rule>
<rule>ወየበ|ገረጣ</rule>
<rule>ዋጀ|አዳነ</rule>
<rule>ብሄር|ዘር</rule>
<rule>ገረሰሰ|ጣለ</rule>
<rule>ፀዳል|ብርሃን</rule>
<rule>ፅሞና|ዝምታ</rule>
<rule>መንበር|መቀመጫ</rule>
<rule>ሞገደኛ|ሃይለኛ</rule>
<rule>አሻ|ፈለገ</rule>
<rule>ነገረፈጅ|ጠበቃ</rule>
<rule>ተከራካሪ|ተሟጋች</rule>
<rule>ናኘ|ታወቀ</rule>
<rule>ረታ|አሸነፈ</rule>
<rule>ምግባር|ባህሪ</rule>
<rule>ተፈጠፈጠ|ወረደ</rule>
<rule>ገሃድ|ግልፅ</rule>
<rule>ተደመደመ|ተፈፀመ</rule>
<rule>ጨበጠ|ያዘ</rule>
<rule>አስገነዘበ|አሳወቀ</rule>
<rule>አፈነገጠ|ተለየ</rule>
<rule>ፅንስ|ውጥን</rule>
<rule>ጫና|ጭቆና</rule>
<rule>ጠፈር|ሰማይ</rule>
<rule>ተጎነጨ|ቀመሰ</rule>
<rule>ጀምበር|ፀሃይ</rule>
<rule>ዳኘ|ፈረደ</rule>
<rule>ዘለፈ|ነቀፈ</rule>
<rule>ተዋሃደ|ተቀላቀለ</rule>
<rule>ተወዳጅ|ተፈቃሪ</rule>
<rule>ኮበለለ|ከዳ</rule>
<rule>ከተተ|ዘመተ</rule>
<rule>ታፈነ|ተዘጋ</rule>
<rule>ሁናቴ|አዝማሚያ</rule>
<rule>አወላወለ|አመነታ</rule>
<rule>ማብራሪያ|ገለፃ</rule>
<rule>ከንውን|አፈፃፀም</rule>
<rule>ዱር|ጫካ</rule>
<rule>አበቃ|አለቀ</rule>

<rule>ተተረከ|ተነገረ</rule>

<rule>በረበረ|ፈተሸ</rule>

<rule>ፕላን|እቅድ</rule>

<rule>ፈነጠዘ|ተደሰተ</rule>

<rule>ጫረ|ለኮሰ</rule>

<rule>ወንጀል|ጥፋት</rule>

<rule>ተጣመረ|ተባበረ</rule>

<rule>ገለልተኛ|ብቸኛ</rule>

<rule>ዶሴ|ማህደር</rule>

<rule>አጠናቀረ|አዘጋጀ</rule>

<rule>ዘገበ|አቀረበ</rule>

<rule>ውል|ስምምነት</rule>

<rule>ወጠምሻ|ደንዳና</rule>

<rule>ወሰለተ|ዋሸ</rule>

<rule>አለኝታ|መመኪያ</rule>

<rule>አወከ|ረበሸ</rule>

<rule>አከለ|ጨመረ</rule>

<rule>አወጀ|ነገረ</rule>

<rule>ትርኢት|መዝናኛ</rule>

<rule>ብላቴና|ወጣት</rule>

<rule>አቢሲንያ|ኢትዮጵያ</rule>

<rule>አቶሞቢል|አውቶሞቢል</rule>

<rule>አቶማቲክ|አውቶማቲክ</rule>

<rule>እስፖርት|ስፖርት</rule>

<rule>እስታዲየም|ስታዲየም</rule>

<rule>አክስፈርድ|አክስፎርድ</rule>

<rule>እስፖንሰር|ስፖንሰር</rule>

<rule>ኢንሳይክሎፕድያ|ኢንሳይክሎፒድያ</rule>

<rule>ኢንሳይክሎፔድያ|እንሳይክሎፒድያ</rule>

<rule>ቴለቭዥን|ቴሌቪዥን</rule>

<rule>ቴሌቭዠን|ቴሌቪዥን</rule>

<rule>አውሮፓውያን|ኤሮፓውያን</rule>

<rule>ሚሊንየም|ሚሊኒየም</rule>

<rule>ሚለኒየም|ሚሊኒየም</rule>

<rule>ሚያዚያ|ሚያዝያ</rule>

<rule>ሰማኒያ|ሰማንያ</rule>

<rule>ቴክኒዎሎጂ|ቴክኖሎጂ</rule>

<rule>ኮምፒዩተር|ኮምፒውተር</rule>

<rule>ኢሜል|ኢሜይል</rule>

<rule>ሐ|ሀ</rule>

<rule>ኀ|ሀ</rule>

<rule>ሐ|ሁ</rule>

<rule>ኍ|ሁ</rule>

<rule>ሒ|ሂ</rule>

<rule>ኂ|ሂ</rule>

<rule>ሃ|ሀ</rule>

<rule>ሓ|ሀ</rule>

<rule>ኃ|ሀ</rule>

<rule>ሔ|ሄ</rule>

<rule>ኄ|ሄ</rule>

<rule>ሕ|ህ</rule>

<rule>ኅ|ህ</rule>

<rule>ሓ|ሆ</rule>

<rule>ኖ|ሆ</rule>

<rule>ሠ|ሰ</rule>

<rule>ሡ|ሱ</rule>

<rule>ሢ|ሲ</rule>

<rule>ሣ|ሳ</rule>

<rule>ሤ|ሴ</rule>

<rule>ሥ|ስ</rule>

<rule>ሦ|ሶ</rule>

<rule>ዀ|አ</rule>

<rule>዁|ኡ</rule>

<rule>ዂ|ኢ</rule>

<rule>ዓ|አ</rule>

<rule>ኣ|አ</rule>

<rule>ዔ|ኤ</rule>

<rule>እ|ዕ</rule>

<rule>ዖ|አ</rule>

<rule>ጸ|ፀ</rule>

<rule>ጹ|ፀ</rule>

<rule>ጺ|ፂ</rule>

<rule>ጻ|ፃ</rule>

<rule>ጼ|ፄ</rule>

<rule>ጽ|ፅ</rule>

<rule>ጾ|ፆ</rule>

<rule>ኰ|ኮ</rule>

<rule>ጎ|ጐ</rule>

<rule>ዉ|ው</rule>

<rule>::|፡፡</rule>

<rule>፩|1</rule>

<rule>፪|2</rule>

<rule>፫|3</rule>

<rule>፬|4</rule>

<rule>፭|5</rule>

<rule>፮|6</rule>

<rule>፯|7</rule>

<rule>፰|8</rule>

<rule>፱|9</rule>

<rule>፲|10</rule>

<rule>፳|20</rule>

<rule>፴|30</rule>

<rule>፵|40</rule>

<rule>፶|50</rule>

<rule>፷|60</rule>

<rule>፸|70</rule>

<rule>፹|80</rule>

<rule>፺|90</rule>

<rule>፻|100</rule>

<rule>፼|10000</rule>

<rule>ት/ቤት|ትምህርት ቤት</rule>

<rule>ት/ርት|ትምህርት</rule>

<rule>ት/ክፍል|ትምህርት ክፍል</rule>

<rule>ኃ/አለቃ|ሃምሳ አለቃ</rule>

<rule>ኃ/ስላሴ|ኃይለ ስላሴ</rule>

<rule>ደ/ዘይት|ደብረ ዘይት</rule>

<rule>ደ/ታቦር|ደብረ ታቦር</rule>

<rule>መ/ር|መምህር</rule>

<rule>መ/ቤት|መስሪያ ቤት</rule>

<rule>መ/አለቃ|መቶ አለቃ</rule>

<rule>ከ/ከተማ|ክፍለ ከተማ</rule>

<rule>ከ/ሀገር|ክፍለ ሀገር</rule>

<rule>ወ/ር|ወታደር</rule>

<rule>ወ/ሮ|ወይዘሮ</rule>

<rule>ወ/ት|ወይዘሪት</rule>

<rule>ኃ/ማርያም|ኃይለማርያም</rule>

<rule>ተ/ወልድ|ተክለወልድ</rule>

<rule>ወ/ስለሴ|ወልደ ስለሴ</rule>

<rule>ፍ/ስላሴ|ፍቅረ ስላሴ</rule>

<rule>ጠ/ሚኒስትር|ጠቅላይ ሚኒስትር</rule>

<rule>ዶ/ር|ዶክተር</rule>

<rule>ገ/ጊዮርጊስ|ገብረ ጊዮርጊስ</rule>

<rule>ቤ/ክርስትያን|ቤተ ክርስትያን</rule>

<rule>ም/ስራ|ምክትል ስራ</rule>

<rule>ም/ቤት|ምክር ቤት</rule>

<rule>ተ/ሀይማኖት|ተክለ ሀይማኖት</rule>

<rule>ሚ/ር|ሚኒስትር</rule>

<rule>ኮ/ል|ኮለኔል</rule>

<rule>ሜ/ጀነራል|ሜጀር ጀኔራል</rule>

<rule>ብ/ጄኔራል|ብርጋዴር ጀኔራል</rule>

<rule>ሌ/ኮለኔል|ሌተናል ኮለኔል</rule>

<rule>ሊ/መንበር|ሊቀ መንበር</rule>

<rule>አ/አ|አዲስ አበባ</rule>

<rule>ር/መምህር|ርእሰ መምህር</rule>

<rule>ፕ/ት|ፕሬዝዳንት</rule>

<rule>ዓ.ም|ዓመተ ምህረት</rule>

<rule>ፍ/ቤት|ፍርድ ቤት</rule>

<rule>ፅ/ቤት|ፅህፈት ቤት</rule>

<rule>ሲ/ር|ሲስተር</rule>

<rule>ፕ/ር|ፕሮፌሰር</rule>

<rule>መ. ቅ.|መፅሃፍ ቅዱስ</rule>

<rule>አ.አ.|አዲስ አበባ</rule>

<rule>ዶ.ር|ዶክተር</rule>

<rule>ዓ. ዓ.|ዓመተ ዓለም</rule>

<rule>እ. ኤ. አ.|እንደ አውሮፓውያን አቆጣጠር</rule>

<rule>አ.ህ.|አፍሪካ ህብረት</rule>

<rule>የተ.መ.ድ.|የተባበሩት መንግስታት ድርጅት</rule>

<rule>ተ.እ.ታ.|ተጨማሪ እሴት ታክስ</rule>

</synonyms>

</stemmer>

<parser>

<linebreak>

<rule>።"</rule>

<rule>።»</rule>

<rule>።'</rule>

<rule>?"</rule>

<rule>!"</rule>

<rule>?»</rule>

<rule>!»</rule>

<rule>?'</rule>

<rule>!'</rule>

<rule>።</rule>

<rule>?</rule>

<rule>!</rule>

</linebreak>

<linedontbreak>

<rule>.</rule>

<rule>;</rule>
<rule>፣</rule>
<rule>፥</rule>
<rule>፡</rule>
<rule>፡-</rule>
<rule>፤</rule>
<rule>አ.አ.</rule>
<rule>ዓ.ዓ.</rule>
<rule>ዓ.ም.</rule>
<rule>a.m.</rule>
<rule>p.m.</rule>
<rule>G.C.</rule>
<rule>E.C.</rule>
</linedontbreak>
</parser>
<grader-syn>
<depreciate>
<rule> </rule>
<rule>ቤተ መንግስት</rule>
<rule>ቤተ ክርስትያን</rule>
<rule>መካነ እየሱስ</rule>
<rule>ቤተ ውበት</rule>
<rule>ቤተ መፀሃፍት</rule>
<rule>ህገ መንግስት</rule>
<rule>ኃይለ ማርያም</rule>
<rule>ኃይለ ስላሴ</rule>
<rule>ከፍለ ከተማ</rule>
<rule>አዲስ ከተማ</rule>
<rule>አዲስ አበባ</rule>
<rule>ስነ ስርዓት</rule>
<rule>ወጥ ቤት</rule>
<rule>መፀሃፍ ቅዱስ</rule>
<rule>ቀይ መስቀል</rule>
<rule>በትረ መንግስት</rule>
<rule>ኮልፌ ቀራንዮ</rule>
<rule>ንፋስ ስልክ</rule>
<rule>ክፍለ ሃገር</rule>
<rule>አፍሪካ ህብረት</rule>
<rule>ስነ ፍጥረት</rule>
<rule>ስነ ምግባር</rule>
<rule>መካነ መቃብር</rule>
<rule>ድረ ገፅ</rule>
<rule>ቄብ ሳይት</rule>
<rule> </rule>
</depreciate>
</grader-syn>
<grader-tc>
<word>ሁሉ</word>
<word>ሁሉም</word>
<word>ኋላ</word>
<word>ሁኔታ</word>
<word>ሆነ</word>
<word>ሆኑ</word>
<word>ሆኖም</word>
<word>ሁሌ</word>
<word>ሁልጊዜ</word>
<word>ሁሉንም</word>
<word>ላይ</word>
<word>ሌላ</word>
<word>ሌሎች</word>
<word>ልዩ</word>
<word>መሆኑ</word>
<word>ማለት</word>
<word>ማለቱ</word>
<word>መካከል</word>
<word>የሚገኝ</word>
<word>የሚገኙ</word>
<word>ማድረግ</word>
<word>ማን</word>
<word>ማንም</word>
<word>ሰሞኑን</word>
<word>ሲሆን</word>
<word>ሲል</word>
<word>ሲሉ</word>
<word>ስለ</word>
<word>ቢቢሲ</word>
<word>ቢሆን</word>
<word>ብለዋል</word>
<word>ብቻ</word>
<word>ብዛት</word>
<word>ብዙ</word>
<word>ቦታ</word>
<word>በርካታ</word>
<word>በሰሞኑ</word>
<word>በታች</word>
<word>በኋላ</word>
<word>በኩል</word>

<word>በውስጥ</word>
<word>በጣም</word>
<word>ይህን</word>
<word>በተለይ</word>
<word>በተመለከተ</word>
<word>በተመሳሳይ</word>
<word>የተለያዩ</word>
<word>የተለያየ</word>
<word>ተባለ</word>
<word>ተገለፀ</word>
<word>ተገልጿል</word>
<word>ተጨማሪ</word>
<word>ተከናውኗል</word>
<word>ችግር</word>
<word>ታች</word>
<word>ትናንት</word>
<word>ነበረ</word>
<word>ነበረች</word>
<word>ነበሩ</word>
<word>ነው</word>
<word>ነይ</word>
<word>ነገር</word>
<word>ነገሮች</word>
<word>ናት</word>
<word>ናቸው</word>
<word>አሁን</word>
<word>አለ</word>
<word>አስታወቀ</word>
<word>አስታውቀዋል</word>
<word>አስታውሰዋል</word>
<word>እስካሁን</word>
<word>አሳሰበ</word>
<word>አሳስበዋል</word>
<word>አስፈላጊ</word>
<word>አስገነዘቡ</word>
<word>አስገንዝበዋል</word>
<word>አብራርተዋል</word>
<word>አብራርተው</word>
<word>አስረድተዋል</word>
<word>ምክንያት</word>
<word>እባክህ</word>
<word>እባክሽ</word>
<word>እባክዎ</word>
<word>አንድ</word>
<word>አንፃር</word>
<word>እስኪደርስ</word>
<word>እንኳ</word>
<word>እንኳን</word>
<word>እስከ</word>
<word>እዚሁ</word>
<word>እና</word>
<word>እንደ</word>
<word>እንደገና</word>
<word>እንደገለፁት</word>
<word>እንደተገለፀው</word>
<word>እንደተናገሩት</word>
<word>እንደአስረዱት</word>
<word>እንደቀረበው</word>
<word>እንደታየው</word>
<word>እንደተነገረው</word>
<word>እንደተብራራው</word>
<word>እንዳሉት</word>
<word>ወቅት</word>
<word>እንዲሁም</word>
<word>እንጂ</word>
<word>እዚህ</word>
<word>እዚያ</word>
<word>እያንዳንዱ</word>
<word>እያንዳንዳቸው</word>
<word>እያንዳንዷ</word>
<word>ከዚህ</word>
<word>ከኋላ</word>
<word>ከላይ</word>
<word>ከመሃል</word>
<word>ከመካከል</word>
<word>ከሰሞኑ</word>
<word>ከታች</word>
<word>ከውስጥ</word>
<word>ከጋራ</word>
<word>ከፊት</word>
<word>ጋራ</word>
<word>ወዘተ</word>
<word>ወይም</word>
<word>ወደ</word>
<word>ዋና</word>
<word>ወደፊት</word>
<word>ውስጥ</word>

<word>ውጭ</word>
<word>ውጪ</word>
<word>ያለ</word>
<word>ያሉ</word>
<word>ይገባል</word>
<word>የኋላ</word>
<word>የሰሞኑ</word>
<word>የታች</word>
<word>የውስጥ</word>
<word>የውጭ</word>
<word>የጋራ</word>
<word>ያ</word>
<word>ይታወሳል</word>
<word>ይህ</word>
<word>ደግሞ</word>
<word>ድረስ</word>
<word>ስለዚህ</word>
<word>ግን</word>
<word>ገልጿል</word>
<word>ገልፀዋል</word>
<word>ግዜ</word>
<word>ጊዜ</word>
<word>ጥቂት</word>
<word>ፊት</word>
<word>እንደሆነ</word>
<word>ዛሬ</word>
<word>ጋር</word>
<word>ተናግረዋል</word>
<word>የገለፁት</word>
<word>ይገልፃል</word>
<word>ሲሉ</word>
<word>ብለዋል</word>
<word>ስለሆነ</word>
<word>አቶ</word>
<word>ሆኖ</word>
<word>መገለፁን</word>
<word>መግለፁን</word>
<word>አቶ</word>
<word>ይናገራሉ</word>
<word>አመልክተዋል</word>
<word>ይህም</word>
<word>በአንድ</word>
<word>ጉዳይ</word>
<word>ነበር</word>
<word>አይደለም</word>
<word>ደረጃ</word>
<word>ሥራ</word>
<word>የሥራ</word>
<word>እናም</word>
<word>ያለው</word>
<word>ያለውን</word>
<word>መሆኑን</word>
<word>በመሆኑ</word>
<word>መሆን</word>
<word>ብሎ</word>
<word>በማለት</word>
<word>ከፍተኛ</word>
<word>በከፍተኛ</word>
<word>የሚል</word>
<word>በላይ</word>
<word>ሳይሆን</word>
<word>ሆነው</word>
<word>ምን</word>
<word>ምንም</word>
<word>ለምን</word>
<word>ዓይነት</word>
<word>ይሁን</word>
<word>በዚህ</word>
<word>አለበት</word>
<word>ያህል</word>
<word>እነዚህ</word>
<word>በእዚህ</word>
<word>ይችላል</word>
<word>አንዳንድ</word>
<word>ታዲያ</word>
<word>እንዲህ</word>
<word>ይቻላል</word>
<word>በሚል</word>
<word>የሆነ</word>
<word>በፊት</word>
<word>ስለሆነም</word>
<word>እንኳን</word>
<word>ዙሪያ</word>
<word>ብለው</word>
</grader-tc>
</dictionary>

## VIII. Guidelines for Manual Summary Preparation

### Guidelines on How to Prepare a Manual Summary

### Composed from (Girma, 2012) and (Melese, 2009)

The purpose of this research is to prepare an Automatic Amharic Text Summarizer using the Open Text Summarizer by incorporation additional Natural Language Processing techniques related to the Amharic language. The performance of this summarizer will be evaluated by comparing it with a human prepared manual summary. Then you are kindly requested to prepare a summary based on the following guidelines for the articles provided to you. The major action that you will be taking is ranking each sentence of the articles based on their importance with number 1 (for the most important sentence) to N (for the least important sentence), where N is the total number of sentences in a single news article. A well structured and well organized meaningful summary can be created combining these ranked sentences one after the other in their order of rank. The sentence ranking guidelines are based on the decision on criteria that includes Informativeness, Non-Redundancy, Coverage, Coherence and Referential Integrity. These are described as:-

1. Informativeness – the sentence should contain the core idea of the article. That is in order to get high rank, a sentence should be more informative and containing the main idea of the article.
2. Coverage – a summary should cover all of the important topics presented in the news article. That is the sentence that contain main topic of discussion should not be left out or given a lesser rank and the highly ranked sentences should contain the maximum possible information about the topic of the article.
3. Non-Redundancy – a repetition of ideas using different sentences should completely be avoided. All the main topics of the news should be covered without repeating the same idea with different words in another sentence. Therefore, if a sentence contains the same information as one of the sentence that is already ranked, it should be given the least rank than the other sentences whose information content is not ranked.
4. Referential Integrity – when reading the summary in the ranked order of the sentences, those sentences that have referential words like pronouns and noun phrases should be easy for the reader to identify to whom or what they are referring to in the sentence.

5. Coherence – is the smooth flow of information about a topic in consecutive sentences. When the sentences are read together, the idea should have a flow rather than having unrelated sentences.

THANK YOU VERY MUCH FOR TAKING YOUR TIME TO PREPARE THE SUMMARIES!!!

## IX. Guidelines for Summary Subjective Evaluation

### Subjective Summary Evaluation Guideline

### Composed from (Girma, 2012)

This Guideline is prepared to help you, the evaluator, in the evaluation of the quality of the automatic summary prepared for the selected news articles. Since you have read the article well while preparing an ideal summary your participation in the evaluation of the automatic summary is important. Therefore you are provided with the automatic summary equivalent of some of the articles that you and others have summarized to evaluate the system's performance subjectively. You are provided a summary of few articles at the three different summarizing rates (10%, 20% and 30% for each of the article) that are randomly selected. The evaluation criteria that you should follow are described as follows.

1. Linguistic Quality – how readable and fluent are the summaries?

   A summary should have a good quality when it comes to linguistics. The focus here is on readability, fluency and quality of the summary. The evaluation includes criteria like grammatical correctness, non redundancy and referential clarity. In grammatical correctness you have to check the occurrence of grammatically incorrect sentences like fragments or missing words in the sentences of the summary. In non redundancy checking, you should look for repeated ideas in sentences to avoid unnecessary repetition. Referential clarity measures the easiness to identify answers to "who or what" in the sentences. Based on the quality level of the summary, please give rank to each of the summaries that you are provided with a grade level from 1 up to 5 (which will be converted to 20-100%).
2. Informativeness – which of the summaries include the most important concept of the article?

   It is measured in terms of the information content of the automatic summary to provide enough information regarding the topic concept of the article. The level of satisfying the

information need by each summary and containing the topic concept is different. You are asked therefore to identify this level. Please read the topic statement and the automatic summary extracted for each article. Based on the informative level of the summary, please give rank to each of the summaries that you are provided with a grade level from 1 up to 5.

3. Coherence and Structure - how well is the summary structured and organized?

   The summary should not be a heap of concatenated sentences. It should build up the topic of the article in the consecutive sentences to a coherent summary. You should therefore look at the buildup of the topic in consecutive sentences that comprise the automatic summary. Based on the coherence level of the summary, please give rank to each of the summaries that you are provided with a grade level from 1 up to 5.

4. How do you rate the summary?

   That is to say, to what level of measure does the summary qualify in your observation? After reading the automatic summary and the news article, from the overall observation what do you think the automatic summary is from the five level of gradding – 1. Very Poor? 2. Poor?, 3. Fair?, 4. Good?, or 5. Very Good?

Please rank each automatic summary as:-

| Quality Measure | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1. Linguistic Quality | | | | | |
| i. Grammatical correctness | | | | | |
| ii. Non redundancy | | | | | |
| iii. Referential Clarity | | | | | |
| 2. Informativeness | | | | | |
| 3. Coherence and Structure | | | | | |
| 4. How do you rate the summary | | | | | |

THANK YOU VERY MUCH FOR TAKING YOUR TIME TO EVALUATE THE SUMMARIES!!!

## X. Grade Score Given Under Subjective Evaluation for 45 Selected Summaries

| Original File Name | Number of Total Sentences | Evaluation Grade Given | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Grammatical Correctness | Non - Redundancy | Referential Clarity | Informativeness | Coherence and Structure | Overall Observation | Total Grade Score from 30 | Total Grade Value in %age |
| **At 10%** | | | | | | | | | |
| SA1.txt | 32 | 3 | 5 | 3 | 2 | 3 | 1 | 17 | **56.67%** |
| SA2.txt | 30 | 4 | 4 | 5 | 5 | 3 | 3 | 24 | **80.00%** |
| SA3.txt | 38 | 5 | 5 | 4 | 3 | 5 | 3 | 25 | **83.33%** |
| SA4.txt | 50 | 5 | 5 | 4 | 4 | 3 | 3 | 24 | **80.00%** |
| SA5.txt | 37 | 4 | 5 | 3 | 4 | 4 | 3 | 23 | **76.67%** |
| SA6.txt | 53 | 4 | 4 | 5 | 4 | 4 | 2 | 23 | **76.67%** |
| SA7.txt | 38 | 4 | 5 | 4 | 4 | 3 | 2 | 22 | **73.33%** |
| SA8.txt | 44 | 5 | 5 | 4 | 4 | 3 | 3 | 24 | **80.00%** |
| SA9.txt | 45 | 5 | 5 | 5 | 5 | 4 | 4 | 28 | **93.33%** |
| SA10.txt | 51 | 5 | 5 | 3 | 5 | 4 | 3 | 25 | **83.33%** |
| **Average** | **42** | **4.40** | **4.80** | **4.00** | **4.00** | **3.60** | **2.70** | **23.50** | **78.33%** |
| MA1.txt | 64 | 4 | 4 | 5 | 5 | 4 | 4 | 26 | **86.67%** |
| MA3.txt | 73 | 5 | 4 | 3 | 5 | 3 | 3 | 23 | **76.67%** |
| MA5.txt | 61 | 5 | 3 | 5 | 4 | 4 | 4 | 25 | **83.33%** |
| MA7.txt | 56 | 5 | 3 | 4 | 5 | 5 | 2 | 24 | **80.00%** |
| MA9.txt | 70 | 4 | 3 | 5 | 5 | 4 | 3 | 24 | **80.00%** |
| **Average** | **65** | **4.60** | **3.40** | **4.40** | **4.80** | **4.00** | **3.20** | **24.40** | **81.33%** |
| **At 20%** | | | | | | | | | |
| SA1.txt | 32 | 3 | 5 | 4 | 4 | 3 | 2 | 21 | **70.00%** |
| SA2.txt | 30 | 4 | 4 | 5 | 5 | 3 | 4 | 25 | **83.33%** |
| SA3.txt | 38 | 5 | 5 | 4 | 4 | 5 | 3 | 26 | **86.67%** |
| SA4.txt | 50 | 5 | 5 | 4 | 4 | 3 | 4 | 25 | **83.33%** |
| SA5.txt | 37 | 4 | 5 | 4 | 5 | 4 | 4 | 26 | **86.67%** |
| SA6.txt | 53 | 4 | 3 | 5 | 4 | 4 | 4 | 24 | **80.00%** |
| SA7.txt | 38 | 4 | 5 | 5 | 4 | 3 | 3 | 24 | **80.00%** |
| SA8.txt | 44 | 5 | 5 | 4 | 4 | 3 | 4 | 25 | **83.33%** |
| SA9.txt | 45 | 5 | 5 | 4 | 5 | 4 | 4 | 27 | **90.00%** |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SA10.txt | 51 | 5 | 5 | 4 | 5 | 4 | 4 | 27 | **90.00%** |
| **Average** | **42** | **4.40** | **4.70** | **4.30** | **4.40** | **3.60** | **3.60** | **25.00** | **83.33%** |
| MA1.txt | 64 | 4 | 4 | 5 | 5 | 4 | 5 | 27 | **90.00%** |
| MA3.txt | 73 | 5 | 4 | 3 | 5 | 3 | 4 | 24 | **80.00%** |
| MA5.txt | 61 | 5 | 3 | 4 | 4 | 4 | 4 | 24 | **80.00%** |
| MA7.txt | 56 | 5 | 3 | 5 | 5 | 5 | 3 | 26 | **86.67%** |
| MA9.txt | 70 | 4 | 3 | 5 | 5 | 4 | 5 | 26 | **86.67%** |
| **Average** | **65** | **4.60** | **3.40** | **4.40** | **4.80** | **4.00** | **4.20** | **25.40** | **84.67%** |
| **At 30%** | | | | | | | | | |
| SA1.txt | 32 | 3 | 5 | 4 | 3 | 3 | 2 | 20 | **66.67%** |
| SA2.txt | 30 | 4 | 4 | 5 | 5 | 3 | 5 | 26 | **86.67%** |
| SA3.txt | 38 | 5 | 5 | 4 | 4 | 5 | 4 | 27 | **90.00%** |
| SA4.txt | 50 | 5 | 5 | 4 | 4 | 4 | 5 | 27 | **90.00%** |
| SA5.txt | 37 | 4 | 5 | 5 | 5 | 5 | 5 | 29 | **96.67%** |
| SA6.txt | 53 | 4 | 4 | 5 | 5 | 5 | 5 | 28 | **93.33%** |
| SA7.txt | 38 | 4 | 5 | 5 | 4 | 3 | 5 | 26 | **86.67%** |
| SA8.txt | 44 | 5 | 5 | 4 | 4 | 3 | 5 | 26 | **86.67%** |
| SA9.txt | 45 | 5 | 5 | 5 | 5 | 4 | 5 | 29 | **96.67%** |
| SA10.txt | 51 | 5 | 5 | 4 | 5 | 5 | 4 | 28 | **93.33%** |
| **Average** | **42** | **4.40** | **4.80** | **4.50** | **4.40** | **4.00** | **4.50** | **26.60** | **88.67%** |
| MA1.txt | 64 | 4 | 4 | 5 | 5 | 4 | 5 | 27 | **90.00%** |
| MA3.txt | 73 | 5 | 4 | 3 | 5 | 5 | 5 | 27 | **90.00%** |
| MA5.txt | 61 | 5 | 3 | 4 | 4 | 4 | 5 | 25 | **83.33%** |
| MA7.txt | 56 | 5 | 3 | 5 | 5 | 5 | 4 | 27 | **90.00%** |
| MA9.txt | 70 | 4 | 3 | 5 | 5 | 4 | 5 | 26 | **86.67%** |
| **Average** | **65** | **4.60** | **3.40** | **4.40** | **4.80** | **4.40** | **4.80** | **26** | **88.00%** |

## XI. Manual Evaluation Results

**A.** Linguistic Quality - i) measured in terms of Grammatical Correctness

| Original File Name | 10% Summary | | 20% Summary | | 30% Summary | |
|---|---|---|---|---|---|---|
| | Grade | in % | Grade | in % | Grade | in % |
| SA1.txt | 3 | 60% | 3 | 60% | 3 | 60% |
| SA2.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA3.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA4.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA5.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA6.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA7.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA8.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA9.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA10.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| **Average** | **4.40** | **88.00%** | **4.40** | **88.00%** | **4.40** | **88.00%** |
| MA1.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| MA3.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| MA5.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| MA7.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| MA9.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| **Average** | **4.60** | **92.00%** | **4.60** | **92.00%** | **4.60** | **92.00%** |

**B.** Linguistic Quality - ii) measured in terms of non redundancy

| Original File Name | 10% Summary | | 20% Summary | | 30% Summary | |
|---|---|---|---|---|---|---|
| | Grade | in % | Grade | in % | Grade | in % |
| SA1.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA2.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA3.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA4.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA5.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA6.txt | 4 | 80% | 3 | 60% | 4 | 80% |
| SA7.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA8.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA9.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA10.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| **Average** | **5** | **96.00%** | **5** | **94.00%** | **5** | **96.00%** |
| MA1.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| MA3.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| MA5.txt | 3 | 60% | 3 | 60% | 3 | 60% |
| MA7.txt | 3 | 60% | 3 | 60% | 3 | 60% |
| MA9.txt | 3 | 60% | 3 | 60% | 3 | 60% |
| **Average** | **3** | **68.00%** | **3** | **68.00%** | **3** | **68.00%** |

**C.** Linguistic Quality - iii) measured in terms of Referential Clarity

| Original File Name | 10% Summary | | 20% Summary | | 30% Summary | |
|---|---|---|---|---|---|---|
| | **Grade** | **in %** | **Grade** | **in %** | **Grade** | **in %** |
| SA1.txt | 3 | 60% | 4 | 80% | 4 | 80% |
| SA2.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA3.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA4.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA5.txt | 3 | 60% | 4 | 80% | 5 | 100% |
| SA6.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA7.txt | 4 | 80% | 5 | 100% | 5 | 100% |
| SA8.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA9.txt | 5 | 100% | 4 | 80% | 5 | 100% |
| SA10.txt | 3 | 60% | 4 | 80% | 4 | 80% |
| **Average** | **4** | **80.00%** | **4** | **86.00%** | **5** | **90.00%** |
| MA1.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| MA3.txt | 3 | 60% | 3 | 60% | 3 | 60% |
| MA5.txt | 5 | 100% | 4 | 80% | 4 | 80% |
| MA7.txt | 4 | 80% | 5 | 100% | 5 | 100% |
| MA9.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| **Average** | **4** | **88.00%** | **4** | **88.00%** | **4** | **88.00%** |

**D.** Informativeness

| Original File Name | 10% Summary | | 20% Summary | | 30% Summary | |
|---|---|---|---|---|---|---|
| | **Grade** | **in %** | **Grade** | **in %** | **Grade** | **in %** |
| SA1.txt | 2 | 40% | 4 | 80% | 3 | 60% |
| SA2.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA3.txt | 3 | 60% | 4 | 80% | 4 | 80% |
| SA4.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA5.txt | 4 | 80% | 5 | 100% | 5 | 100% |
| SA6.txt | 4 | 80% | 4 | 80% | 5 | 100% |
| SA7.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA8.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA9.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA10.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| **Average** | **4** | **80.00%** | **4** | **88.00%** | **4** | **88.00%** |
| MA1.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| MA3.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| MA5.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| MA7.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| MA9.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| **Average** | **5** | **96.00%** | **5** | **96.00%** | **5** | **96.00%** |

**E.** Coherence and Structure

| Original File Name | 10% Summary | | 20% Summary | | 30% Summary | |
|---|---|---|---|---|---|---|
| | **Grade** | **in %** | **Grade** | **in %** | **Grade** | **in %** |
| SA1.txt | 3 | 60% | 3 | 60% | 3 | 60% |
| SA2.txt | 3 | 60% | 3 | 60% | 3 | 60% |
| SA3.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| SA4.txt | 3 | 60% | 3 | 60% | 4 | 80% |
| SA5.txt | 4 | 80% | 4 | 80% | 5 | 100% |
| SA6.txt | 4 | 80% | 4 | 80% | 5 | 100% |
| SA7.txt | 3 | 60% | 3 | 60% | 3 | 60% |
| SA8.txt | 3 | 60% | 3 | 60% | 3 | 60% |
| SA9.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| SA10.txt | 4 | 80% | 4 | 80% | 5 | 100% |
| **Average** | **4** | **72.00%** | **4** | **72.00%** | **4** | **80.00%** |
| MA1.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| MA3.txt | 3 | 60% | 3 | 60% | 5 | 100% |
| MA5.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| MA7.txt | 5 | 100% | 5 | 100% | 5 | 100% |
| MA9.txt | 4 | 80% | 4 | 80% | 4 | 80% |
| **Average** | **4** | **80.00%** | **4** | **80.00%** | **4** | **88.00%** |

**F.** Over all observation - how do you rate the summary

| Original File Name | 10% Summary | | 20% Summary | | 30% Summary | |
|---|---|---|---|---|---|---|
| | **Grade** | **in %** | **Grade** | **in %** | **Grade** | **in %** |
| SA1.txt | 1 | 20% | 2 | 40% | 2 | 40% |
| SA2.txt | 3 | 60% | 4 | 80% | 5 | 100% |
| SA3.txt | 3 | 60% | 3 | 60% | 4 | 80% |
| SA4.txt | 3 | 60% | 4 | 80% | 5 | 100% |
| SA5.txt | 3 | 60% | 4 | 80% | 5 | 100% |
| SA6.txt | 2 | 40% | 4 | 80% | 5 | 100% |
| SA7.txt | 2 | 40% | 3 | 60% | 5 | 100% |
| SA8.txt | 3 | 60% | 4 | 80% | 5 | 100% |
| SA9.txt | 4 | 80% | 4 | 80% | 5 | 100% |
| SA10.txt | 3 | 60% | 4 | 80% | 4 | 80% |
| **Average** | **3** | **54.00%** | **4** | **72.00%** | **5** | **90.00%** |
| MA1.txt | 4 | 80% | 5 | 100% | 5 | 100% |
| MA3.txt | 3 | 60% | 4 | 80% | 5 | 100% |
| MA5.txt | 4 | 80% | 4 | 80% | 5 | 100% |
| MA7.txt | 2 | 40% | 3 | 60% | 4 | 80% |
| MA9.txt | 3 | 60% | 5 | 100% | 4 | 80% |
| **Average** | **3** | **64.00%** | **4** | **84.00%** | **5** | **92.00%** |

## XII. Objective Evaluation Results for Experiment One

A. At 10% Extraction Rate

| Original File Name | Experiment 1 Summary File Name | Number of Total Sentences | Number of Similar Sentences | F-Measure |
|---|---|---|---|---|
| SA1.txt | E1SA1P10.txt | 3 | 3 | 100.00% |
| SA2.txt | E1SA2P10.txt | 3 | 2 | 66.67% |
| SA3.txt | E1SA3P10.txt | 4 | 1 | 25.00% |
| SA4.txt | E1SA4P10.txt | 5 | 1 | 20.00% |
| SA5.txt | E1SA5P10.txt | 4 | 3 | 75.00% |
| SA6.txt | E1SA6P10.txt | 5 | 3 | 60.00% |
| SA7.txt | E1SA7P10.txt | 4 | 2 | 50.00% |
| SA8.txt | E1SA8P10.txt | 4 | 1 | 25.00% |
| SA9.txt | E1SA9P10.txt | 5 | 2 | 40.00% |
| SA10.txt | E1SA10P10.txt | 5 | 1 | 20.00% |
| **Average** | | **4** | **2** | **48.17%** |
| MA1.txt | E1MA1P10.txt | 6 | 4 | 66.67% |
| MA2.txt | E1MA2P10.txt | 6 | 3 | 50.00% |
| MA3.txt | E1MA3P10.txt | 7 | 3 | 42.86% |
| MA4.txt | E1MA4P10.txt | 6 | 5 | 83.33% |
| MA5.txt | E1MA5P10.txt | 6 | 4 | 66.67% |
| MA6.txt | E1MA6P10.txt | 8 | 3 | 37.50% |
| MA7.txt | E1MA7P10.txt | 6 | 2 | 33.33% |
| MA8.txt | E1MA8P10.txt | 6 | 2 | 33.33% |
| MA9.txt | E1MA9P10.txt | 7 | 4 | 57.14% |
| MA10.txt | E1MA10P10.txt | 7 | 4 | 57.14% |
| **Average** | | **6** | **3** | **52.80%** |
| LA1.txt | E1LA1P10.txt | 8 | 3 | 37.50% |
| LA2.txt | E1LA2P10.txt | 10 | 2 | 20.00% |
| LA3.txt | E1LA3P10.txt | 10 | 4 | 40.00% |
| LA4.txt | E1LA4P10.txt | 10 | 3 | 30.00% |
| LA5.txt | E1LA5P10.txt | 10 | 3 | 30.00% |
| LA6.txt | E1LA6P10.txt | 10 | 2 | 20.00% |
| LA7.txt | E1LA7P10.txt | 8 | 4 | 50.00% |
| LA8.txt | E1LA8P10.txt | 9 | 4 | 44.44% |
| LA9.txt | E1LA9P10.txt | 9 | 6 | 66.67% |
| LA10.txt | E1LA10P10.txt | 13 | 5 | 38.46% |
| **Average** | | **9** | **4** | **37.71%** |
| **Corpus Average** | | **7** | **3** | **46.22%** |

B. At 20% Extraction Rate

| Original File Name | Experiment 1 Summary File Name | Number of Total Sentences | Number of Similar Sentences | F-Measure |
|---|---|---|---|---|
| SA1.txt | E1SA1P20.txt | 6 | 3 | 50.00% |
| SA2.txt | E1SA2P20.txt | 6 | 3 | 50.00% |
| SA3.txt | E1SA3P20.txt | 8 | 5 | 62.50% |
| SA4.txt | E1SA4P20.txt | 10 | 5 | 50.00% |
| SA5.txt | E1SA5P20.txt | 7 | 4 | 57.14% |
| SA6.txt | E1SA6P20.txt | 11 | 4 | 36.36% |
| SA7.txt | E1SA7P20.txt | 8 | 4 | 50.00% |
| SA8.txt | E1SA8P20.txt | 9 | 4 | 44.44% |
| SA9.txt | E1SA9P20.txt | 9 | 6 | 66.67% |
| SA10.txt | E1SA10P20.txt | 10 | 3 | 30.00% |
| **Average** | | **8** | **4** | **49.71%** |
| MA1.txt | E1MA1P20.txt | 13 | 6 | 46.15% |
| MA2.txt | E1MA2P20.txt | 12 | 9 | 75.00% |
| MA3.txt | E1MA3P20.txt | 15 | 8 | 53.33% |
| MA4.txt | E1MA4P20.txt | 11 | 8 | 72.73% |
| MA5.txt | E1MA5P20.txt | 12 | 8 | 66.67% |
| MA6.txt | E1MA6P20.txt | 15 | 8 | 53.33% |
| MA7.txt | E1MA7P20.txt | 11 | 7 | 63.64% |
| MA8.txt | E1MA8P20.txt | 11 | 8 | 72.73% |
| MA9.txt | E1MA9P20.txt | 14 | 9 | 64.29% |
| MA10.txt | E1MA10P20.txt | 13 | 7 | 53.85% |
| **Average** | | **13** | **8** | **62.17%** |
| LA1.txt | E1LA1P20.txt | 16 | 12 | 75.00% |
| LA2.txt | E1LA2P20.txt | 19 | 5 | 26.32% |
| LA3.txt | E1LA3P20.txt | 19 | 11 | 57.89% |
| LA4.txt | E1LA4P20.txt | 19 | 7 | 36.84% |
| LA5.txt | E1LA5P20.txt | 20 | 9 | 45.00% |
| LA6.txt | E1LA6P20.txt | 19 | 6 | 31.58% |
| LA7.txt | E1LA7P20.txt | 16 | 10 | 62.50% |
| LA8.txt | E1LA8P20.txt | 17 | 7 | 41.18% |
| LA9.txt | E1LA9P20.txt | 17 | 11 | 64.71% |
| LA10.txt | E1LA10P20.txt | 25 | 14 | 56.00% |
| **Average** | | **19** | **9** | **49.70%** |
| **Corpus Average** | | **13** | **7** | **53.86%** |

C. At 30% Extraction Rate

| Original File Name | Experiment 1 Summary File Name | Number of Total Sentences | Number of Similar Sentences | F-Measure |
|---|---|---|---|---|
| SA1.txt | E1SA1P30.txt | 10 | 3 | 30.00% |
| SA2.txt | E1SA2P30.txt | 9 | 6 | 66.67% |
| SA3.txt | E1SA3P30.txt | 11 | 8 | 72.73% |
| SA4.txt | E1SA4P30.txt | 15 | 6 | 40.00% |
| SA5.txt | E1SA5P30.txt | 11 | 6 | 54.55% |
| SA6.txt | E1SA6P30.txt | 16 | 6 | 37.50% |
| SA7.txt | E1SA7P30.txt | 11 | 5 | 45.45% |
| SA8.txt | E1SA8P30.txt | 13 | 10 | 76.92% |
| SA9.txt | E1SA9P30.txt | 14 | 11 | 78.57% |
| SA10.txt | E1SA10P30.txt | 15 | 9 | 60.00% |
| **Average** | | **13** | **7** | **56.24%** |
| MA1.txt | E1MA1P30.txt | 19 | 14 | 73.68% |
| MA2.txt | E1MA2P30.txt | 18 | 14 | 77.78% |
| MA3.txt | E1MA3P30.txt | 22 | 15 | 68.18% |
| MA4.txt | E1MA4P30.txt | 17 | 15 | 88.24% |
| MA5.txt | E1MA5P30.txt | 18 | 13 | 72.22% |
| MA6.txt | E1MA6P30.txt | 23 | 17 | 73.91% |
| MA7.txt | E1MA7P30.txt | 17 | 14 | 82.35% |
| MA8.txt | E1MA8P30.txt | 17 | 15 | 88.24% |
| MA9.txt | E1MA9P30.txt | 21 | 13 | 61.90% |
| MA10.txt | E1MA10P30.txt | 20 | 14 | 70.00% |
| **Average** | | **19** | **14** | **75.65%** |
| LA1.txt | E1LA1P30.txt | 25 | 15 | 60.00% |
| LA2.txt | E1LA2P30.txt | 29 | 17 | 58.62% |
| LA3.txt | E1LA3P30.txt | 29 | 21 | 72.41% |
| LA4.txt | E1LA4P30.txt | 29 | 18 | 62.07% |
| LA5.txt | E1LA5P30.txt | 30 | 21 | 70.00% |
| LA6.txt | E1LA6P30.txt | 29 | 18 | 62.07% |
| LA7.txt | E1LA7P30.txt | 24 | 21 | 87.50% |
| LA8.txt | E1LA8P30.txt | 26 | 16 | 61.54% |
| LA9.txt | E1LA9P30.txt | 26 | 17 | 65.38% |
| LA10.txt | E1LA10P30.txt | 38 | 26 | 68.42% |
| **Average** | | **28** | **19** | **66.80%** |
| **Corpus Average** | | **20** | **13** | **66.23%** |

## XIII. Objective Evaluation Results for Experiment Two

A. At 10% Extraction Rate

| Original File Name | Experiment 2 Summary File Name | Number of Total Sentences | Number of Similar Sentences | F-Measure |
|---|---|---|---|---|
| SA1.txt | E2SA1P10.txt | 3 | 1 | 33.33% |
| SA2.txt | E2SA2P10.txt | 3 | 1 | 33.33% |
| SA3.txt | E2SA3P10.txt | 4 | 1 | 25.00% |
| SA4.txt | E2SA4P10.txt | 5 | 1 | 20.00% |
| SA5.txt | E2SA5P10.txt | 4 | 2 | 50.00% |
| SA6.txt | E2SA6P10.txt | 5 | 2 | 40.00% |
| SA7.txt | E2SA7P10.txt | 4 | 2 | 40.00% |
| SA8.txt | E2SA8P10.txt | 4 | 1 | 25.00% |
| SA9.txt | E2SA9P10.txt | 5 | 1 | 20.00% |
| SA10.txt | E2SA10P10.txt | 5 | 1 | 20.00% |
| **Average** | | **4** | **1** | **30.67%** |
| MA1.txt | E2MA1P10.txt | 6 | 4 | 66.67% |
| MA2.txt | E2MA2P10.txt | 6 | 1 | 16.67% |
| MA3.txt | E2MA3P10.txt | 7 | 1 | 14.29% |
| MA4.txt | E2MA4P10.txt | 6 | 2 | 33.33% |
| MA5.txt | E2MA5P10.txt | 6 | 4 | 66.67% |
| MA6.txt | E2MA6P10.txt | 8 | 3 | 37.50% |
| MA7.txt | E2MA7P10.txt | 6 | 3 | 50.00% |
| MA8.txt | E2MA8P10.txt | 6 | 1 | 16.67% |
| MA9.txt | E2MA9P10.txt | 7 | 2 | 28.57% |
| MA10.txt | E2MA10P10.txt | 7 | 4 | 57.14% |
| **Average** | | **6** | **3** | **38.75%** |
| LA1.txt | E2LA1P10.txt | 8 | 2 | 25.00% |
| LA2.txt | E2LA2P10.txt | 10 | 4 | 40.00% |
| LA3.txt | E2LA3P10.txt | 10 | 4 | 40.00% |
| LA4.txt | E2LA4P10.txt | 10 | 2 | 20.00% |
| LA5.txt | E2LA5P10.txt | 10 | 6 | 60.00% |
| LA6.txt | E2LA6P10.txt | 10 | 2 | 20.00% |
| LA7.txt | E2LA7P10.txt | 8 | 2 | 25.00% |
| LA8.txt | E2LA8P10.txt | 9 | 3 | 33.33% |
| LA9.txt | E2LA9P10.txt | 9 | 4 | 44.44% |
| LA10.txt | E2LA10P10.txt | 13 | 4 | 30.77% |
| **Average** | | **9** | **3** | **33.85%** |
| **Corpus Average** | | **7** | **2** | **34.42%** |

B. At 20% Extraction Rate

| Original File Name | Experiment 2 Summary File Name | Number of Total Sentences | Number of Similar Sentences | F-Measure |
|---|---|---|---|---|
| SA1.txt | E2SA1P20.txt | 6 | 3 | 50.00% |
| SA2.txt | E2SA2P20.txt | 6 | 4 | 66.67% |
| SA3.txt | E2SA3P20.txt | 8 | 5 | 62.50% |
| SA4.txt | E2SA4P20.txt | 10 | 8 | 80.00% |
| SA5.txt | E2SA5P20.txt | 7 | 5 | 71.43% |
| SA6.txt | E2SA6P20.txt | 11 | 6 | 54.55% |
| SA7.txt | E2SA7P20.txt | 8 | 7 | 87.50% |
| SA8.txt | E2SA8P20.txt | 9 | 4 | 44.44% |
| SA9.txt | E2SA9P20.txt | 9 | 5 | 55.56% |
| SA10.txt | E2SA10P20.txt | 10 | 4 | 40.00% |
| **Average** | | **8** | **5** | **61.26%** |
| MA1.txt | E2MA1P20.txt | 13 | 5 | 38.46% |
| MA2.txt | E2MA2P20.txt | 12 | 5 | 41.67% |
| MA3.txt | E2MA3P20.txt | 15 | 4 | 26.67% |
| MA4.txt | E2MA4P20.txt | 11 | 8 | 72.73% |
| MA5.txt | E2MA5P20.txt | 12 | 7 | 58.33% |
| MA6.txt | E2MA6P20.txt | 15 | 6 | 40.00% |
| MA7.txt | E2MA7P20.txt | 11 | 8 | 72.73% |
| MA8.txt | E2MA8P20.txt | 11 | 6 | 54.55% |
| MA9.txt | E2MA9P20.txt | 14 | 6 | 42.86% |
| MA10.txt | E2MA10P20.txt | 13 | 7 | 53.85% |
| **Average** | | **13** | **6** | **50.18%** |
| LA1.txt | E2LA1P20.txt | 16 | 12 | 75.00% |
| LA2.txt | E2LA2P20.txt | 19 | 12 | 63.16% |
| LA3.txt | E2LA3P20.txt | 19 | 11 | 57.89% |
| LA4.txt | E2LA4P20.txt | 19 | 11 | 57.89% |
| LA5.txt | E2LA5P20.txt | 20 | 11 | 55.00% |
| LA6.txt | E2LA6P20.txt | 19 | 9 | 47.37% |
| LA7.txt | E2LA7P20.txt | 16 | 8 | 50.00% |
| LA8.txt | E2LA8P20.txt | 17 | 12 | 70.59% |
| LA9.txt | E2LA9P20.txt | 17 | 9 | 52.94% |
| LA10.txt | E2LA10P20.txt | 25 | 14 | 56.00% |
| **Average** | | **19** | **11** | **58.58%** |
| **Corpus Average** | | **13** | **7** | **56.68%** |

C. At 30% Extraction Rate

| Original File Name | Experiment 2 Summary File Name | Number of Total Sentences | Number of Similar Sentences | F-Measure |
|---|---|---|---|---|
| SA1.txt | E2SA1P30.txt | 10 | 3 | 30.00% |
| SA2.txt | E2SA2P30.txt | 9 | 7 | 77.78% |
| SA3.txt | E2SA3P30.txt | 11 | 9 | 81.82% |
| SA4.txt | E2SA4P30.txt | 15 | 7 | 46.67% |
| SA5.txt | E2SA5P30.txt | 11 | 7 | 63.64% |
| SA6.txt | E2SA6P30.txt | 16 | 11 | 68.75% |
| SA7.txt | E2SA7P30.txt | 11 | 10 | 90.91% |
| SA8.txt | E2SA8P30.txt | 13 | 10 | 76.92% |
| SA9.txt | E2SA9P30.txt | 14 | 10 | 71.43% |
| SA10.txt | E2SA10P30.txt | 15 | 11 | 73.33% |
| **Average** | | **13** | **9** | **68.12%** |
| MA1.txt | E2MA1P30.txt | 19 | 14 | 73.68% |
| MA2.txt | E2MA2P30.txt | 18 | 12 | 66.67% |
| MA3.txt | E2MA3P30.txt | 22 | 12 | 54.55% |
| MA4.txt | E2MA4P30.txt | 17 | 12 | 70.59% |
| MA5.txt | E2MA5P30.txt | 18 | 13 | 72.22% |
| MA6.txt | E2MA6P30.txt | 23 | 16 | 69.57% |
| MA7.txt | E2MA7P30.txt | 17 | 14 | 82.35% |
| MA8.txt | E2MA8P30.txt | 17 | 12 | 70.59% |
| MA9.txt | E2MA9P30.txt | 21 | 16 | 76.19% |
| MA10.txt | E2MA10P30.txt | 20 | 15 | 75.00% |
| **Average** | | **19** | **14** | **71.14%** |
| LA1.txt | E2LA1P30.txt | 25 | 19 | 76.00% |
| LA2.txt | E2LA2P30.txt | 29 | 21 | 72.41% |
| LA3.txt | E2LA3P30.txt | 29 | 18 | 62.07% |
| LA4.txt | E2LA4P30.txt | 29 | 22 | 75.86% |
| LA5.txt | E2LA5P30.txt | 30 | 23 | 76.67% |
| LA6.txt | E2LA6P30.txt | 29 | 17 | 58.62% |
| LA7.txt | E2LA7P30.txt | 24 | 19 | 79.17% |
| LA8.txt | E2LA8P30.txt | 26 | 21 | 80.77% |
| LA9.txt | E2LA9P30.txt | 26 | 19 | 73.08% |
| LA10.txt | E2LA10P30.txt | 38 | 28 | 73.68% |
| **Average** | | **28** | **21** | **72.83%** |
| **Corpus Average** | | **20** | **14** | **70.70%** |