ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCES

Afan Oromo news text summarizer

BY

GIRMA DEBELE DINEGDE

A THESIS SUBMITED TO THE SCHOOL OF GRADUTE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCINECE IN INFORMATION SCIENCE

ADDIS ABABA, ETHIOPIA

June, 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUTE STUDIES

SCHOOL OF INFORMATION SCIENCES

DEPARTMENT OF INFORMATION SCIENCE

Afan Oromo news text summarizer

BY

GIRMA DEBELE DINEGDE

Name and Signature of the Board of Examiners for Approval

Chairman, Department Examination Board: _____ _____

Examiner: Dr. Dereje Teferi _____

Advisor: Dr. MarthaYifiru _____

Dedication

This work is dedicated to my father, Ato Debele Dinegde who was unfortunate to reap a fruit of his own.

Declaration

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

_____

Girma Debele

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

_____

Dr. MarthaYifiru

June, 2012

Acknowledgment

First of all, I would like to thank **God** for helping me to finalize my thesis work.
My deepest heartfelt gratitude also goes to my advisor Dr. MarthaYifiru for her critical comments on my work and helpful advice, without whom this work was impossible.

I would like to thank journalists of Oromia Radio and Television Organization Ato Tolosa Mideksa , Alemayehu H/Mariam , Mekonin Alemu and Deraje Geda for their helpfulness during evaluation of the system.

I am also grateful to my collogues Ato Belayneh Mengistu, Ato Ketema Adare, and other people who have supported me in moral, providing me good working environment and materials.

Last not least, I would like to thank Ayela Gonfa and Birhanu Wakjira who have helped me in many situations.

Table of contents

Contents **Page**

Abstract

Information overload is a global problem that requires solution. Automatic text summarization is one of the natural language processing technologies that have got researchers focus to help information users. It is a computer program that summarizes a text. A summarizer removes redundant information from the input text and produces a shorter non-redundant output text. In this study, a generic automatic text summarizer for Afan Oromo news text has been developed based upon the Open Text Summarizer (OTS). OTS summarizes texts in English, German, Spanish, Russian, Hebrew, Esperanto and other languages. For this master's thesis most of the work done is customizing the OTS code so that it can make use of the Afan Oromo lexicons and work for the Afan Oromo language. The summarizer basically 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.

In this study we have developed three methods for Afan Oromo news text summarization and tested their performance both objectively and subjectively. These three summarizers are: M1 that uses term frequency and position methods without Afan Oromo stemmer and other lexicons (synonyms and abbreviations), M2 is a summarizer with combination of term frequency and position methods with Afan Oromo stemmer and language specific lexicons (synonyms and abbreviations) and M3 is with improved position method and term frequency as well as the stemmer and language specific lexicons (synonyms and abbreviations).

The performance of the summarizers was measured based on subjective as well as objective evaluation methods. The result of objective evaluation shows that the three summarizers: M1, M2 and M3 registered f-measure values of 34%, 47% and 81% respectively i.e. M3 outperformed the two summarizers (M1 and M2) by 47% and 34 % . Moreover, the subjective evaluation result shows that the three summarizers' (M1, M2 and M3) performances with informativeness, linguistic quality and coherence and structure are: (34.37 %, 37%, and 62.5%), (59.37%, 60% and 65%) and (21.87%, 28.12% and 75%) respectively as it is judged by human evaluators. In both subjective and objective evaluation, the results are consistent. Summarizer M3 that uses the combination of term frequency and improved position methods outperform other summarizers followed by M2.

List of Tables

List of Figures

List of Abbreviations

Sg. 1.p.	1st person singular
Sg. 2.p.	2nd person singular
Sg. 3.p.m.	3rd person singular masculine
Sg. 3.p. f.	3rd person singular feminine
Pl. 1.p.	1st person plural
pl. 2.p.	2nd person plural
pl. 3.p.	3rdperson plural
ATS	Automatic Text Summarization
XML	Extensible Markup Language
HTML	Hyper Text Markup Language
NLP	Natural Language Processing
OOTS	Open Oromo Text Summarizer
OTS	Open Text Summarizer
ORTO	Oromia Radio and Television Organization
VOA	Voice of America
WWW	World Wide Web

CHAPTER ONE

1. INTRODUCTION

1.1 Background

As the amount of information available increases, systems that can automatically summarize one or more documents become increasingly desirable (Radev, 2001). Document summarization is the creation of a shortened version of a text by the use of computer program (Park, 2004). Automatic summarization has attracted attention both in the research community and commercially as a solution for reducing information overload and helping users to scan a large number of documents to identify documents of their interest (Khoo and Goh, 2007). It has been a research topic since the 1950s. Nowadays, it is becoming more and more significant that attracts many research groups around the world (Park, 2004).

Document summarization can be categorized into two types with different techniques: single-document summarization and multi-document summarization. Single-document summarization is aimed at obtaining a source text and presenting the most important content in a condensed form in a manner sensitive to the needs of the further task while multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. It has turned out to be much more complex than summarizing a single document, even a very large one (Helen, 2006). It consists of computing the summary of a set of related documents such that they give the user a general view of the events in the documents (Khoo and Goh, 2007).

Most research on summary generation techniques still relies on extraction of important sentences from the original document to form a summary. There are several methods for measuring the importance of a sentence. Some algorithms calculate a weight for each sentence, taking into account the position of the sentence and word frequencies (Dalianis et al, 2003), while others use semantic information in order to find the hierarchy of concepts.

1.2 Statement of the problem and justification of the study

These days, documents in paper and electronic format are growing dramatically. As a result, the users (readers) are facing information overload problem with vast quantities of text. In almost all languages in the world, texts in any domain are written in detail and readers are forced to see unwanted detail without being interested in it unless it is summarized to save the readers' time.

Afan Oromo text readers are not exceptional to suffer from this problem. There are many domain areas that produce large content of textual information which needs summarization to save the time of readers. Some of the textual information are large volumes of legal judgments which is very essential if they are used by the experts (for timely justice) and by law students for their study, newspaper texts and online news articles produced by media agencies, criminal investigation document produced by polices at different level, reports from government offices, etc.

Textual information both printed and in digital form, in Afan Oromo is increasing highly from time to time since the language became official language in Oromia regional state. News items comprise a certain part from these outputs. Currently newspapers and other news releases in the language reach the readers from many sources. There are a number of media agencies and presses releasing news in electronic and non-digital format. There are a number of newspapers publishers that produce news articles. Some of such sources of newspaper are: *Barriisa*, *Kallacha Oromiya* and *Oromiya*. *Bariisa* is a weekly newspaper, whereas the rest two come out once in two weeks. There are also radio broadcasts in Afan Oromo by Ethiopian Radio and Radio Fana for 14 and 30 hours weekly, respectively. Moreover, Oromia Radio and Television Organization found in Adama releases daily news through radio and television broadcast and on its official website. On the other hand, magazines, judiciary documents and office reports also constitute some portion of the documents produced in the language.

Though it is becoming more important to read the daily news in ones preference area, due to time shortage and other workloads , reading a news articles about a given topic fully is not always possible .

With the absence of automatic text summarization services that can potentially reduce the readers' browsing and reading time, it can be said that readers have been and being spending more time than they should browsing over the content that they are not interested in.

Automatic Afan Oromo text summarizer, especially for large amount of news releases by newspapers and online news agencies, could then be justified as it is very essential to save the readers' time. Therefore, it is advisable to employ a powerful computational tool to do the task of text summarization in news domain. As far as my knowledge is concerned, there is no attempt on automatic text summarization for Afan Oromo.

To this end, the purpose of this study is to explore appropriate statistical approaches for developing and implementing an automatic news text summarizer for Afan Oromo that generate extract summary to satisfy readers' requirements.

Currently, a few - researches in automatic text summarization have been commenced for Ethiopian languages, particularly for Amharic text in different domains by adopting different techniques. The present work is a contribution towards developing natural language processing applications for Ethiopian Languages. Specifically it increases the scope of the text summarization research by investigating its application for Afan Oromo language. The techniques used in this study is term frequency and sentence position methods with language specific lexicons (synonyms and abbreviations) to assign weights to the sentences to be extracted for the summary.

1.3 Objectives of the study

The general objective of the study is to build up a single document automatic summarizer for Afan Oromo news text.

The specific objectives of the study set to achieve the general objective are:

- To review related research works in the area of text summarization
- To review algorithms and techniques that have been used in the area of text summarization
- To investigate existing summarization methods and techniques in view of Afan Oromo news structure and select and use the feasible best combination of them
- To develop a prototype summarizer as a framework that will serve as a model for Afan Oromo news text summarization
- To test and evaluate the summarizer
- To draw conclusions based on experimental result and recommend further research works

1.4 Significance of the study

This thesis can serve as an input to the development of a complete Afan Oromo news text summarizer and has the importance to initiate further research in the area of document summarization for Afan Oromo language. Moreover, it can also help to initiate text summarization researches in other Ethiopian languages.

1.5 Research methodology

To achieve the objectives stated in Section 1.3, the researcher made use of the following methods.

Primarily, literatures related to automatic text summarization have been reviewed. As the study is conducted on Afan Oromo news text summarization, the nature of the language and the structure of the documents to be summarized for testing were investigated. To carry out this task, books, journal articles, and relevant websites are consulted.

1.5.1 Corpus Preparation

A corpus to evaluate the summarizer (Afan Oromo news articles) was selected and prepared as there is no previous research and corpora in Afan Oromo for evaluating summarizer. The prepared corpus consists of 8 news items from Oromia Radio and Television Organization (ORTO) [1] as well as Voice of America (VOA) [2] Afan Oromo official websites written on different topics .While selecting from news archives, longer articles (at least one page or more than 200 words) are considered due to the fact that as the text itself gets shorter summarizing it becomes unnecessary. The average length of news items, in the corpus, is approximately 277 words or 11 sentences as shown in Table 1

Text ID	News size in words	News size in sentences
Test 1	250	11
Test 2	403	14
Test 3	250	9
Test 4	231	10
Test 5	290	13
Test 6	295	14
Test 7	232	11
Test 8	269	13
Average	**277.5**	**11.875**

Table 1: Prepared corpuses for the study

1.5.2 Summary Generation

For the purpose of manual summary generation, the corpus was provided to the human subjects together with the corresponding guideline. The four available experts ranked the sentences based on their ability of providing salient information for the reference summary. For a sentence, an average rank was calculated as the sum of its four ranks divided by four. The sentences have then been ordered according to their average rank. Finally, reference summaries were produced from the top ranking sentences at 10 %, 20%, 30 % and 40% of the original text's word length (compression rate) of randomly selected test sets (See Section 4.4.2) .

[1]See: http://www.orto.gov.et

[2]See: http://www.voanews.com

1.5.3 Summarization technique and tools used

Most research on summary generation techniques still relies on extraction of important sentences from the original document to form a summary (Kaili and Pilleriin, 2005). There are several ways in which one can characterize different approaches to text summarization.

The technique proposed for this study is extraction technique for single news text. Using extraction technique most important sentences from the document are extracted and displayed to the reader. To create a summary by this technique there is no need of rewriting the document by making linguistics analysis. To extract important sentence from a text to be summarized, sentence can be weighted based on cue phrases it contains, location of the sentence, sentence containing most frequent words in the document. Then sentences with the highest weight obtained by efficient combination of extraction features will be selected and a summary is written.

This work is based upon the Open Text Summarizer (OTS) (Rotem, 2001), an open source tool for summarizing texts. The program reads a text and decides which sentences are important and which are not. It ships with Ubuntu, Fedora and other Linux distributions. OTS supports many (more than 25) languages which are configured in XML[3] files. OTS incorporates natural language processing (NLP) techniques via an English language lexicon with synonyms and cue terms as well as rules for stemming. These are used in combination with a statistical word frequency based method for sentence scoring. Therefore, the source code available in C# has been used and the XML file has been configured with Afan Oromo rule of stemming, stop list, synonyms and abbreviations such that it can support Afan Oromo news text summarization. The summarizer prototype is therefore customized from the existing OTS. Moreover, the researcher developed and integrated a tool for objective evaluation (compute standard recall and precision) with the summarizer.

[3]XML: stands for *Extensible Markup Language*, and it is used to describe documents and data in a standardized, text-based format that can be easily transported via standard Internet protocols.

1.5.4 Evaluation technique

After configuring and developing the prototype text summarizer based on OTS, two forms of summaries prepared (system summary and reference summary) are used to evaluate the performance of the system. The evaluation process was conducted using an intrinsic[4] method. It comprised of both subjective (qualitative) and objective (quantitative) evaluation methods. For both measures the four human subjects (expert journalists) are involved (see Section 4.5).

Subjective evaluation was used to measure the linguistic quality, informativeness and coherence of the automatically generated summaries. The linguistic quality is basically aimed to measure the readability and fluency of the summary. We adopted subjective summarization techniques used by Greek text summarizer (Pachantouris, 2004). On the other hand, objective evaluation was basically used to measure the summarizer's performance in identification and extraction of salient sentences. This performance is measured by the standard recall and precision measures. Given an input text, human's (reference) summary and summarizer's extract, it measures how close the extracts are to the reference summary.

The standard recall and precision measures is calculated as follows:

- $Recall = correct\ /\ (correct + missed)$
- $Precision = correct/\ (\ correct + wrong)$

Where:

- **Correct** = the number of sentences in both the summarizer's summary and the reference summary ,
- **Wrong** = the number of sentences in the summarizer's summary but not in the reference summary,
- **Missed** = the number of sentences in the reference summary but not in the summarizer's summary.

[4] Intrinsic: a method of summary evaluation that concentrates on the summary itself, trying to measure its cohesion, coherence and informativeness, usually in comparison with other summaries of the same text ("gold standard")

1.6 Scope and limitations of the study

This research focuses on single document summarization for Afan Oromo news articles. Therefore, the experimentation has dealt with Afan Oromo news texts only, excluding the summarization of information in other types or format.

On the other hand, the absence of standard test corpus and evaluation tool for Afan Oromo language was a limitation though the researcher prepared the small corpus for the experimentation and has developed a tool for evaluation and integrated with a summarizer. However, the amount of corpus prepared for this study is relatively small and requires further development.

1.7 Organization of the thesis

This thesis report is organized into five chapters. The first chapter talks about the motivation behind conducting the research and discusses: background of the study, statement of the problem, the objectives, methodology and scope and limitations of the study.

The second chapter presents the basic concepts and related works on automatic text summarization. Concerning the basic concepts of text summarization it discusses the process, types, approaches, techniques and evaluation methods of text summarization. Furthermore, it reviewed history of automatic text summarization and global works and research works on automatic text summarization on local languages.

Chapter three discusses Afan Oromo language features such as Afan Oromo writing system and punctuations, morphology, word and sentence boundary and describe news writing style.

Chapter four describes the practical activities carried out to implement the prototype summarizer, corpus preparation for the experimentation and evaluation and discussion of the result.

Finally chapter five gives conclusions and recommendations based on the findings of the study.

CHAPTER TWO

2. REVIEW OF RELATED LITERATURE

2.1 Introduction

The advancement of information and communication technologies (ICT) has simplified the production, collection, organization, storage, and dissemination of information. On the other hand, especially with advent of internet and World Wide Web (WWW), information users are facing challenge in evaluating, filtering and selecting information that meet their information needs.

The rapid growth of the web and online electronic information services, that have supported the availability of large amount of information in a variety of format, highly initiated researches in natural language processing (NLP) field. So far, different technologies have been devised to help users to manage the problem of information overload and able to access information in multi-source, multi-format and multi language. Automatic text summarization is one of these technologies that help in condensing primarily textual information from one or more sources to present the most relevant information to the user.

There are many uses of summarization. It is essential for instance in order to be able to keep up with what is happening in the world. The following are some examples of uses of summarization in everyday life (Pachantouris and Dalianis, 2005):

- Headlines of the news
- Table of contents of a magazine
- Preview of a movie
- Abstract summary of a scientific paper
- Review of a book
- Highlights of a meeting

The remaining sections of this chapter are intended to present: the basic concepts, processes, types, approaches and techniques of automatic text summarization and review of related abroad and local research woks.

2.2 Basic Concepts of Automatic Text Summarization

According to (Hennig et al,2008) Automatic Text Summarization(ATS) is defined as the task of creating a document from one or more textual sources that is smaller in size but retains some or most of the information contained in the original sources. It is a task of producing summary using computer where digital format text entered in to a computer and a summarized text which is the most relevant parts of a document are extracted is returned. Moreover, ATS is aimed at reducing the complexity and length of texts, while retaining the most important information (Luhn, 1958).

The need to automatic summarization of document is increasing due to the fact that: it dramatically reduces the time required to produce a summary or abstract by experts; it enables a readers to quickly revise a content they have already seen and it enables one to create certain standard or consistent summary format etc. Moreover, automatic text summarization systems can be applied in: summarizing news articles of newspapers and online news; can be embedded in large systems like search engines and in extracting key word and summaries of e-mail for SMS in mobile phones etc.

Though ATS is becoming a very interesting and useful task that serves the above mentioned purposes and gives support for many other tasks, it is still a challenging work (Lloret, 2008). Though early experiments in the field of automatic text summarization have showed the possibility and viability of creating text summary, it is not simple (Luhn, 1958) and (Edmundson, 1969). In creating document summary automatically, one of the challenges is determining what information from the source text to be included in the summary. According to (Mani et al ,1998) the task of determining how important information to be included to the summary needs to consider several factors such as nature and genre (domain) of the source text, compression rate desired , the user's information need etc.

The next subsections discuss the basic process, types, approaches and techniques of automatic text summarization.

2.2.1 Process of Automatic Text Summarization

According to (Alguliev and Aliguliyev, 2009) and (Moens, 1997) the process of text summarization can be decomposed in to three phases: analysis of source text, transformation, synthesis of output text.

Analysis of the source text is to identify the essential content to build an internal representation. The techniques used for this task ranges from statistical methods that search for specific key content for extraction to complex techniques that employ natural language understanding. The statistical approaches in general concerned for identification of important topic terms and the extraction of contextual sentences that contain them. On the other hand, other approaches for source analysis needs the complete understanding of the source text i.e. each sentence is processed into its propositions representing the meaning of the sentence.

The second step in automatic text summarization process is transformation of the internal representation into summary representation. This stage requires additional knowledge about the task and audience of the summary to guide the selection of the information as well as about the subject domain to conduct and accurate generalization of the information.

The synthesis phase takes the summary representation, and produces an appropriate summary corresponding to users' needs. This last step is concerned with the organization of the content and essential for abstract type of summary [5].

[5]Abstract type summary: uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

2.2.2 Types of Summaries

The uses of text summarization vary with users' need and its applications. Therefore, while designing automatic text summarization systems, one should take into account the intended purpose of the summary produced by the system. Different types of summaries have been classified based on different scenarios like: the nature of input text to be summarized, purpose of the summary, output of the summary, etc. The following listed are some of the types of summaries (Ganapathiraju, 2002), (Schlesinger and Baker, 2001) and (Manabu and Hajime, 2000):

- Single-document vs. Multi-document:- The input document for the summarizer can be one (single-document) or a set of multiple similar documents (multi-document). Accordingly, the summaries can be categorized as single-document and multiple-document summaries.
- Extract vs. Abstract: - An extract is a summary created by taking parts of the original text at a certain granularity such as key words, cue phrases, sentence or paragraph positions. On the other hand, an abstract is a summary created by regenerating text units that could convey the main concepts of the original text.
- Indicative vs. Informative: - an indicative summary provides an idea of what the text is about. While an informative summary tries to provide some shortened version of the content
- Generic vs. Query-based :- a generic summary is an objective summary (author's view) of a text while that of query-based one tends to reflect the user's information need
- Just-the-News vs. Background: - just-the-news summary presents the newest facts about a topic by assuming that the reader has prior knowledge of the past event, whereas background summary offers the whole story of the event briefly.

Generally, a summary can be one or combination of types discussed above having different features. Each type (or combination of types) needs different methods and techniques to be created and evaluated differently. According to the above-mentioned types and sub-types of automatic text summarization, the summarization technique presented in this thesis can be called sentence extraction-based single document informative summarization in news domain.

2.2.3 Approaches to Text Summarization

The approaches to text summarization based on the form of summary to be produced can be categorized into two: extractive and abstractive. Extractive summarization methods simplify the problem of summarization into the problem of selecting a representative subset of the sentences in the original documents. This approach produces summaries completely consisting from the sentences or word sequences contained in the original document (Alguliev and Aliguliyev, 2009). Besides the complete sentences, extracts can contain phrases and paragraphs. Problem with this approach is usually lack of the balance and cohesion. Sentences could be extracted out of the context and anaphoric references can be broken (Rejhan et al, 2009). On the other hand, abstractive summarization may compose novel sentences, unseen in the original sources. They are usually built from the existing content but using advanced methods. However, abstractive approaches require deep NLP such as semantic representation, inference and natural language generation, which have yet to reach a mature stage nowadays (Alguliev and Aliguliyev, 2009). It is generally hard for computer to successfully solve the requirements of such approach as of many limitations, including the state of the art in language generation and human language complexity (Rejhan et al, 2009).

Moreover, (Alguliev and Aliguliyev, 2009) based on processing level involved in the creation of document summaries, summarization approaches can be grouped as: surface level approach and deeper level approach. In the case of surface level approach, information is represented from the point of shallow features. These include different types of terms, e.g. statistically and positional salient ones, terms from cue phrases or domain specific and user inserted terms. Usually this approach produces extraction based summary as an output. Deeper level approach may involve sentence generation. Advanced semantic analysis is necessary in order to accomplish tasks require deeper level approach. The output of this approach may be in form of abstracts or extracts.

2.2.4 Techniques of Text Summarization

The most important concept useful to create a summarizer is to understand and decide appropriate technique to be used for creating it. To decide and identify the most important text units for the required summary, different researchers have been using one or a combination of different extraction features and weighting techniques to determine the summary to be produced. A number of methods have been employed for automatic text summarization. Commonly, summarization systems use several methods in independent modules. Each module assigns a weight to each unit of the text (such as key word, sentence, cue phrase etc). An integrator module combines the scores for each unit to get a single score. Finally, the system returns the first *N* highest-scoring text units, based on the extraction rate (summary length) (Hassel, 1999). The following discussion presents some of the techniques used and corresponding works that apply the technique is reviewed.

i. Position method

Certain locations of the text to be summarized (like heading, titles, first sentences, first paragraphs, etc) likely contains important information (Ishikawa et al, 2007). As newspapers articles are written in inverted pyramid style, the first (lead) sentence is the best single sentence summary. More generally, taking the lead, sentences or paragraph as summary often outperforms other methods (Hovy and Lin, 1999).

ii. Cue Word or phrase Method

In some genres certain words and phrases such as ‘significant ‘and ‘in conclusion’ explicitly signal importance. Sentences containing these cue words or phrases worth to be extracted.

In the work of (Edmundson,1969) , three types of cue words used for the experiment: 783 bonus words (positively affecting the relevance of a sentence e.g. “Significant”, “Greatest”), 73 stigma words (negatively affecting the relevance to a sentence e.g. “Impossible”, “Hardly”) and 139 null words (irrelevant). Then, he computed the cue weight of each sentence as the summation of weight of each cue word in the sentence.

Teufel and Moens (1997) also applied the technique. After their experimentation they reported as cue phrase method was their best single feature, 54 percent joint recall and precision achieved, using a manually built list of cue phrases in a domain of scientific texts. In order to distinguish the level of contribution of each cue phrase to the relevance to the text unit they assigned a 'goodness score' from -1 to +3.

iii. Query method

Query method is used for query based text summarization system (Pembe and Güngör 2007); the sentences in a given document are scored based on the frequency counts of terms (words or phrases). The sentences containing the query phrases are given higher scores than the ones containing single query words. Then, the sentences with highest scores are incorporated into the output summary together with their structural context. Portions of text may be extracted from different sections or subsections. The resulting summary is the union of such extracts. The number of extracted sentences and the extent to which their context is displayed depends on the summary frame size which is fixed to the size of the screen that can be seen without scrolling. In the sentence extraction algorithm, whenever a sentence is selected for the inclusion in the summary, some of the headings in that context are also selected (Hovy and Lin, 199).

iv. Word and Phrase Frequency Method

Luhn (Luhn, 1958) used Zipf's Law of word distribution (a few words occur very often, fewer words occur somewhat often, and many words occur infrequently) to develop the following extraction criterion: if a text contains some words that are unusually frequent, then sentences containing these words are probably important.

The systems of Luhn (1958), Edmundon (1969), Teufel and Moens (1997), and others employ various frequency measures, and report performance of between 15 percent and 35 percent recall and precision (using word frequency alone).

v.Title Method

The title method is similar to the query method except the desirable words are those in the text's titles or headings. In combination with word and phrase frequency method in Edmundson's work (1969), each title word was given the same score and the scores were summed within text units, but the score was the mean frequency of title word occurrences in the sentence in Teufel and Moens (1997) work.

vi.Cohesive or Lexical Chain Method

Within a text, words can be connected in various ways such as: co- reference, synonymy, and semantic association as expressed in thesauri. Sentences and paragraphs can be scored based on their words degree of connectedness; more-connected sentences are assumed to be more important.

Cohesive methods are based on internal text structure, which is a text feature that allows different parts of a text to function as a whole. This lexical cohesion arises from semantic relationships between words. The most relevant sentences in a text are the highest connected entities in this semantic structure. The connection between these entities can be exploited for text summarization purpose through different techniques including the following.

- **Word co-occurrence:-** words can be related if they occur in common contexts. Some uses word similarity (repetitions, synonyms) measures to establish links between the text units (Abracos and Lopes, 1997);
- **Local salience and grammatical relations :-** important phrasal expressions are given by combination of grammatical , syntactic and contextual parameters (Booguraev and Kennedy,1997) ;
- **Co-reference:-** the more important sentences are traversed by co-reference chains (noun, event identity, part-whole relations) detected between query and document; and sentences within a document (Mani et al ,1998) ;
- **Lexical chains: -** the lexical cohesion can occur between pairs of words and over sequences of related words. Using lexical databases to determine the lexical relations it is possible to create strong chains. The most important sentences are traversed by strong chains (Manabu and Hajime, 2000);

- **Connectedness: -** the text structure is represented in terms of cohesion relations (proper name, anaphora, reiteration, synonymy and hyponymy) and coherence. The text is mapped in a graph, whose nodes represent word instances and links represent adjacency, grammatical, co-reference and lexical similarity relations. The salience of works and sentences is calculated by applying statistical metrics (Mani et al, 1998).

vii.Discourse structure criteria

A variant of connectedness involves producing the underlying discourse structure of the text and scoring sentences by their discourse centrality, as shown in (Marcu, 1998).
This method is based on the rhetorical structure theory. The central idea is that the notion of rhetorical relation, which is a relationship between two text spans called nucleus and satellite. This rhetorical relation can be assembled into rhetorical structure of tree. A rhetorical parser is used to build this discourse representation structure and the centrality to the textual units (Marcu, 1998).

viii.Machine learning techniques

With the advent of machine learning techniques in NLP in the 1990s, a series of influential publications appeared that employed statistical techniques to produce document extracts (Dipanjan,2007). While initially most systems assumed feature independence and relied on naive-Bayes methods, others have focused on the choice of appropriate features and on learning algorithms that make no independence assumptions. Other significant approaches involved hidden Markov models and log-linear models to improve extractive summarization. A very recent paper, in contrast, used neural networks and third party features (like common words in search engine queries) to improve purely extractive single document summarization.

a. Naive-Bayes Methods

Kupiec et al. (1995) describe a method derived from Edmundson (1969) that is able to learn from data. The classification function categorizes each sentence as worthy of extraction or not, using a naive-Bayes classifier. Let s be a particular sentence, S the set of sentences that make up the summary, and F1, F2 …, Fk the features.

Assuming independence of the features:

$$P(s \in \mathcal{S} \mid F_1, F_2, ..F_k) = \frac{\prod_{i=1}^{k} P(F_i \mid s \in \mathcal{S}) \cdot P(s \in \mathcal{S})}{\prod_{i=1}^{k} P(F_i)}$$

Aone et al. (1999) also incorporated a naive-Bayes classifier, but with richer features. They describe a system called DimSum that made use of features like term frequency (tf and inverse document frequency (idf) to derive signature words. The idf was computed from a large corpus of the same domain as the concerned documents. Statistically derived two-noun word collocations were used as units for counting, along with single words. A named-entity tagger was used and each entity was considered as a single token. They also employed some shallow discourse analysis like reference to same entities in the text, maintaining cohesion. The references were resolved at a very shallow level by linking name aliases within a document like “U.S." to “United States", or “IBM" for “International Business Machines". Synonyms and morphological variants were also merged while considering lexical terms, the former being identified by using Wordnet (Miller, 1995). The corpora used in the experiments were from newswire, some of which belonged to the TREC[6] evaluations.

[6] TREC: See http://trec.nist.gov/

b. Rich Features and Decision Trees

Lin and Hovy (1997) studied the importance of a single feature, sentence position. Just weighing a sentence by its position in text, which the authors term as the “position method", arises from the idea that texts generally follow a predictable discourse structure, and that the sentences of greater topic centrality tend to occur in certain specifiable locations (e.g. title, abstracts, etc). However, since the discourse structure significantly varies over domains, the position method cannot be defined as naively as in (Baxendale, 1958).

c. Hidden Markov Models

In contrast with previous approaches that were mostly feature-based and non-sequential, Conroy and O'leary (2001) modeled the problem of extracting a sentence from a document using a hidden Markov model (HMM). The basic motivation for using a sequential model is to account for local dependencies between sentences. Only three features were used: position of the sentence in the document (built into the state structure of the HMM), number of terms in the sentence, and likeliness of the sentence terms given the document terms.

d. Neural Networks and Third Party Features

This method involves training the neural networks to learn the types of sentences that should be included in the summary (Gupta and Lehal, 2010). This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The neural network (Kaikhah, 2004), learns the patterns inherent in sentences that should be included in the summary and those that should not be included. It uses three-layered Feed forward neural network, which has been proven to be a universal function approximate. The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The neural network learns patterns inherent in sentences that should be included in the summary and those that should not be included.

ix.Combinations of Various Methods

The predominant tendency in current systems is to adopt a hybrid approach and combine and integrate some of the techniques mentioned before. In many cases, researchers have found that no single method of scoring perform as good as humans do to create extracts. However, since different methods rely on different kinds of evidence, combining function have been tried; all seem to work, and there is no obvious best strategy. In their land marks work , (Kupiec et al, 1995) train 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 (since short sentences are generally not included in summaries). They find that, individually, the paragraph position feature gives 33 percent, the cue phrase indicators 29 percent. But combinations of the two methods give 42 percent.

Using SUMMARIST, (Lin, 1999) compares eighteen different features, a naïve combination of them, and an optimal combination was obtained using machine learning algorithm. These features include most of the above mentioned ones, as well as features signaling the presence of proper names, dates, quantities, pronouns, and quotes in sentence. The top best method was the learned combination function.

The second-best score is achieved by query term overlap method. The third best sore (up to 20 percent length) is achieved equally by word frequency, the lead method, and the naïve combination function. The other important point forwarded by (Lin, 1999) was that to be most useful, summaries should not be longer than about 35 percent and not shorter than about 15 percent.

2.2.5 Evaluation Methods of Automatic Text Summarization

The issue of how to evaluate computer produced summaries has been a topic of research in the field of automatic summarization. The absence of exact definition for “ideal” summary, either an automatically generated summary or manually constructed summary by professional abstractors, makes evaluation technique a hot issue. Techniques for automatic summaries evaluation have been a hot topic for as long as automatic summarization has been started.

According to (Inderjeet, 2001) there are two types of summary evaluation: extrinsic and intrinsic. An extrinsic method of evaluation is where the quality of the summary is judged on how well it helps a person performing other task such as information retrieval .Whereas an intrinsic evaluation is where humans judge the quality of summarization directly on an analysis of the auto-generated summary.

In intrinsic evaluation an ideal summary is created for each test text, and then the summarizer's output is compared to it. The method measures content overlap often by sentence or phrase recall and precision, but sometimes by simple word overlap. Since there is no 'correct' summary, some evaluators use more than one ideal summary per test text, and average the score of the system across the set of ideals.

Comparing system output to some ideal summary was performed in works of (Edmundson, 1969), (Marcu, 1998) and (Kupiec et al, 1995). To simplify evaluating extracts, Marcu (1998) independently developed and automated method to create extracts corresponding to abstracts (ideal summary). The other way to use intrinsic method is to have evaluators rate systems' summaries responsiveness and /or linguistic quality using some scale (readability, grammar, informativeness, fluency, coverage, redundancy) (Brandow, 1995) .

Extrinsic evaluation is easy to motivate. The major problem is to ensure that the metric applied perfectly correlates with task performance efficiency. One of the largest extrinsic evaluation experiment was the TPSTER-SUMMAC study (Farmin and Chrzanowski, 1999), involving some eighteen systems (research and commercial), in three tests. In the categorization task testers classified a set for Text REtrieval Conference (TREC) texts and their summaries created by various systems .After classification, the agreement between the classification of texts and their corresponding summaries is measured. The greater the agreement, the better the summary has captured the information that caused the full text to be classified as it did.

2.3 Review on Related Automatic Text Summarization Studies

We have given a general overview of the classical techniques used in summarization in the previous section and there are a large number of different techniques and systems. We are going to describe in this section research focusing on single document in news domain applying different techniques. In this section, we first focus on reviewing some earliest works, related global researches and then review all local works in the area of text summarization.

2.4 History of Automatic Text Summarization and Global Related Works

The research work on text summarization can be traced back to 1950's when the first extractive system developed by (Luhn, 1958). He proposed that words appearing many times in a text furnish good idea about the content of the document though there are words that appear very frequently but not content bearing. As a result, he tried to cut off these words by determining a fixed threshold. The idea of Luhn was acknowledged and used in many automatic information processing systems. The system developed takes single document as input. It is domain specific to summarizing technical articles and the system used features like term filtering and word frequency (low-frequency terms are removed). Sentences are weighted by the significant terms they contained and sentence segmentation and extraction is performed.

Edmundson (1969) expanded the work of Luhn. He carefully outlined the human extracting principles and noticed that the location of a sentence in a text gives some clue about the importance of the sentence. Thus, he suggested word frequency, cue phrases, title and heading words and sentence location as an extraction feature. Like the work of Luhn, Edmundson's system is a single document and domain specific (that deals with technical articles). Moreover, the output of the system is an extract summary.

Since then many systems have been developed in the area of automatic text summarization both on single and multi-documents. The researchers in the field of automatic text summarization have been using both statistical and machine learning techniques to create either abstract or extract summaries.

SweSum (Dalianisi, 2000) is the first automatic text summarizer for Swedish language. It summarizes Swedish news text in HTML/text format on the WWW. It is also available for Danish, Norwegian, English, Spanish, French, Italian, Greek, Farsi (Persian) and German texts. It is 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 considers if the text is tagged with bold text tag, first paragraph tag or numerical values. During the summarization 5-10 key words- a mini summary is produced. Performance evaluation shows that accuracy of 84% at 40% summary of news with an average original length of 181 words achieved.

SUMMARIST (Hovy, 1999), is a single-document genre specific to news text. It combines concept-level world knowledge with NLP processing techniques to generate a summary. Stages for summarization are divided in: topic identification, interpretation and generation. It is a multi-lingual system and an attempt to develop robust extraction technology as far as it can go and then continue research and development of techniques to perform abstraction. This work faces the depth vs. robustness tradeoff: either system analyze/interpret the input deeply enough to produce good summaries, or they work robustly over more or less unrestricted text (but cannot analyze deeply enough to fuse the input into a true summary, and hence perform only topic extraction).

LAKE (D'Avanzo et al, 2004), a summarization system developed in 2004 for DUC (Document Understanding Conference). It is single-document domain specific to news summarization. It exploits key phrase extraction methodology to identify relevant terms in the document. It is based on a supervised learning approach and considers linguistic features like name entity recognition or multi words. The system works in two phases. It first considers a number of linguistic features to extract a list of more motivated candidate

key phrases and then it uses machine learning framework to select significant key phrases for that document.

Net-Sum (Svore et al, 2007) is a summarization system developed in 2007 by Microsoft Research Department and focused on single document instead of multi-document summarization. The system produces fully automated single-document extracts of newswire articles based on neuronal nets. It uses machine learning techniques in this way: a train set is labeled so that the labels identify the best sentences. Then a set of features is extracted from each sentence in the train and test sets, and the train set is used to train the system. The system is then evaluated on the test set. The system learns from a training set the distribution of features for the best sentences and outputs a ranked list of sentences for each document.

GreekSum (Pachantouris, 2004) is a master's thesis with aim of building an automatic text summarizer for the Greek language. It is built based on the algorithms developed and used for the SweSum (Dalianis, 2000), text summarizer for Swedish. According to Pachantouris (2004) several changes needed to be made to support the differences of the Greek language from the Swedish already implemented in SweSum. A version of SweSum which is language independent called Generic (without Greek keyword dictionary) and the customized version of the summarizer for Greek language called GreekSum is compared. Subjective evaluation was carried out were they found that using the Greek keyword dictionary in GreekSum made the summarizer 16 percent better that not using a dictionary.

FariSum (Hassel, 1999) is an attempt to create an automatic text summarization system for Persian language. The system is implemented as a HTTP client/server application written in Perl. It is a web-based text summarizer for Persian based upon SweSum. It summarizes Persian newspaper text/HTML in Unicode format. FarsiSum uses the same structure used by SweSum (Dalianisi, 2000), with exception of the **lexicons**, but some modifications have been made in SweSum in order to support Persian texts in Unicode format. The current implementation of FarsiSum is still a prototype. It uses a very simple stop-list in order to filter and identify the important keywords in the text. Persian acronyms and abbreviations are not detected by the current tokenizer.

Among the related works discussed above GreekSum (Pachantouris,2004) and FariSum (Hassel, 1999) shows the possibility of developing a text summarizer for another language based upon the earlier development for other language which uses the advantage of not to reinvent the wheel. These works are the main motivations for our work to be based upon an Open Text Summarizer (the open source toolkit for text summarization).

2.4.1 Local Works on Automatic Text Summarization

Regarding local works in the area of automatic text summarization, student researchers have conducted study in the school of graduate studies, department of information science at Addis Ababa University (AAU). These works are reviewed in terms of problem addressed, techniques (methods) used, finding of the study and performance of result achieved.

The first Amharic news summarization research is conducted by Kamil Nuru (2004). The study addressed the problem of news articles releases from different sources in Amharic language causes information overload. The system was developed by integrating selected statistical and natural language processing techniques .The extraction feature used are title words, head sentences, head sentences words, paragraph starting sentences, cue phrases and high frequency key words. Performance evaluation result shows that the system registers 74.4% and 58% precision and recall respectively with 38.5% condensation rate. Beside on his finding , the researcher recommended development of good stemmer, availability of standard Amharic corpus, exhaustive lists of stop words, and the inclusion of more NLP, statistical and heuristic parameters.

The research work by Teferi Andargie (2005) is on the same language, genre and similar problem as in the previous work by (Kamil, 2004). This study, however, employed machine learning technique (naïve Bayes). In this study, title, location, cue words and content words features are examined. The results of the analysis shows that precision of 75.00%, recall 74.90 % and classification accuracy of 86.03% in predicting the summary sentences. The researcher recommends availability of standard Amharic corpus, analysis

of each single feature like cue words didn't help in the prediction of sentences for the summary and availability of standard stop-list.

Helen Adane (2006) studied "Automatic Text Summarization for Amharic Legal Judgments". The study addressed the problem that legal experts in Ethiopia has been forced to spend their time on reading large volume documents and find relevant judgments for their cases which results in too delay of decision on cases and proposed text summarization as a solution. The researcher employed statistical extraction techniques. Weight is assigned to each sentence based on its location and the cue words/ phrases that it contains to extract the highest weighted sentences .The system is tested for sample text and precision and recall measure is used for 20 % and 10% compression rate. The system calculates precision and recall. The system summary is compared against the human (ideal) summary. As a result, precision of the system summary is 33.9% and 39%; Precision of the random summary is 23% and 27%; recall of system summary is 57% and 50.5 %; recall of random summary is 46% is 38% for 20% and 10 % compression rate respectively.

Unlike the above mentioned works, this study focuses on Afan Oromo language text in the news domain. The purpose of this thesis work is to build an automatic text summarizer for Afan Oromo language news text. It is based upon the open source system developed known as Open Text Summarizer (OTS) (Rotem , 2001).
OTS is an open source tool for summarizing texts. The program reads a text and decides which sentences are important and which are not. It ships with Ubuntu, Fedora and other Linux distributions. OTS supports more than 25 languages which are configured in XML files. (see Section 4.2 for detail).

CHAPTER THREE

3. AFAN OROMO LANGUAGE

3.1 Introduction

Afan Oromo is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of other neighbor countries like Kenya and Somalia (Abera, 1988) and (Grage and Kumsa, 1982). It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 34.5% of the total population. Besides first language speakers, a number of members of other ethnicities who are in contact with the Oromos speak it as a second language, for example, the Omotic-speaking Bambassi and the Nilo-Saharan-speaking Kwama in northwestern Oromia (Tilahun, 1993). Currently, Afan Oromo is an official language of Oromia regional state (which is the largest Regional State among the current Federal States in Ethiopia). Being the official language, it has been used as medium of instruction for primary and junior secondary schools of the region. Moreover, the language is offered as a subject from grade one throughout the schools of the region. Few literature works, a number of newspapers, magazines, educational resources, official credentials and religious documents are published and available in the language.

In general, Afan Oromo is widely used as written and spoken language in Ethiopia and neighboring courtiers like Kenya and Somalia .With regard to the writing system, "**Qubee**" (a Latin-based alphabet) has been adopted and become the official script of Afan Oromo since 1991(Abera, 1988).

The remaining sections of this chapter discusses: Afan Oromo Alphabet and writing system, punctuation marks and usage, Afan Oromo morphology, Afan Oromo word and sentence boundaries and news writing structure.

3.2 Afan Oromo Alphabets and Writing System

According to (Taha, 2004), Afan Oromo is a phonetic language, which means that it is spoken in the way it is written. The writing system of the language is straightforward which is designed based on the Latin script. Unlike English or other Latin based languages there are no skipped or unpronounced sounds/alphabets in the language. Every alphabet is to be pronounced in a clear short/quick or long /stretched sounds. In a word where consonant is doubled the sounds are more emphasized. Besides, in a word where the vowels are doubled the sounds are stretched or elongated.

Like in English, Afan Oromo has vowels and consonants. Afan Oromo vowels are represented by the five basic letters such as a, e, i, o, u. Besides, it has the typical Eastern Cushitic set of five short and five long vowels by doubling the five vowel letters: 'aa', 'ee', 'ii', 'oo', 'uu' (Abera, 1988).

Consonants, on the other hand , do not differ greatly from English, but there are few special combinations such as "**ch**" and "**sh**" (same sound as English),"**dh**" in Afan Oromo is like an English "*d*" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins. Another Afan Oromo consonant is "**ph**" made when with a smack of the lips toward the outside "**ny**" closely resembles the English sound of "gn". We commonly use these few special combination letters to form words. For instance, **ch** used in **barbaa<u>ch</u>isaa** '*important'*, **sh** used in **<u>sh</u>amarree** ***'****girl'*, **dh** use in **<u>dh</u>a<u>dh</u>aa** '*butter'* , **ph** used in **buu<u>ph</u>aa** ***'****egg'*, and **ny** used in **nyaata** '*food'* .

In general, Afan Oromo has 36 letters (26 consonants and 10 vowels) called "**Qubee".** In general, all letters in English language are also in Afan Oromo except the way it is written. Table 2 shows Afan Oromo alphabet.

Afan Oromo Consonants

		Bilabial/ Labiodental	Alveolar/ Retroflex	Palato-alveolar/ Palatal	Velar/Glottal			
Stops	Voiceless	(p)	t	k	'			
	Voiced	b	d	g				
	Ejective	ph	x	q				
	Implosive	dh						
Affricates	Voiceless	ch						
	Voiced	j						
	Ejective	c						
Fricatives	Voiceless	f	s	sh	h			
	Voiced	(v)	-	nasals		m	n	ny
Approximants		w	l	y				
Flap/Trill		R						

Afan Oromo vowels

	Front	Central	Back
High	i , ii	u , uu	
Mid	e , ee	o , oo	
Low	a	aa	

Table 2 : Afan Oromo Alphabet (source: Debela (2010))

3.3 Punctuation Marks in Afan Oromo

Punctuation is placed in text to make meaning clear and reading easier. Analysis of Afan Oromo texts reveals that different punctuation marks follow the same punctuation pattern used in English and other languages that follow Latin Writing System (Diriba, 2002). Similar to English, the following are some of the most commonly used punctuation marks in Afan Oromo (Gumii, 1995):

i. **Tuqaa** *Full stop* (.): is used at the end of a sentence and in abbreviations.
ii. **Mallattoo Gaafii** *Question mark* (?): is used in interrogative or at the end of a direct question.
iii. **Rajeffannoo** *Exclamation mark* (!): is used at the end of command and exclamatory sentences.
iv. **Qooduu** *Comma* (,): it is used to separate listing in a sentence or to separate the elements in a series.
v. **Tuqlamee** *colon* (:): the function of the colon is to separate and introduce lists, clauses, and quotations, along with several conventional uses, and etc.

3.4 AFAN OROMO MORPHOLOGY

Morphology is a branch of linguistics that studies and describes how words are formed in a language (Debela, 2010). There are two types of morphology: inflectional and derivational. Inflectional morphology is concerned with the inflectional changes in words where word stems are combined with grammatical markers for things like person, gender, number, tense, case and mode. Inflectional changes do not result in changes of parts of speech. On the other hand, derivational morphology deals with those changes that result in changing classes of words (changes in the part of speech). For instance, a noun or an adjective may be derived from a verb.

3.4.1 Types of morphemes in Afan Oromo

A morpheme is the smallest semantically meaningful unit in a language. A morpheme is not identical to a word, and the principal difference between the two is that a morpheme

may or may not stand alone, whereas a word, by definition, is a freestanding unit of meaning. Every word comprises one or more morphemes. In Afan Oromo, there are two categories of morphemes: free and bound morphemes. Free morpheme can stand as word on its own where as bound morpheme does not occur as a word on its own (Meiws, 2001). In Afan Oromo roots (stems) are bound as they cannot occur on their own Example: "**dhug-**" (*drink*) and "**beek-**" (*know*), which are pronounceable only when other completing affixes are added to them (Gumii, 1995).

Similarly an affix is also a bounded morpheme that cannot occur independently. It is attached in some manner to the root, which serves as a base. These affixes are of three types – prefix, suffix and infix. The first and the second types of affixes occur at the beginning and at the end of a root respectively in creating a word whereas the infix occurs in between characters of the word. In **dhugaatii '***dirink',* for instance,-**aatii** is a suffix and **dhug-** is a stem. Moreover, an infix is a morpheme that is inserted within morpheme. In the work of (Debela, 2010) it is discovered that Afan Oromo does not have infixes like English.

There are many ways of word formation in Afan Oromo. These morphological analyses of the language are organized in six categories (Debela, 2010). The categories are: nouns, verbs, adjectives, adverbs, functional words, and conjunctions. Almost all Afan Oromo nouns in a given text have person, number, gender, and possession markers which are concatenated and affixed to a stem or singular noun form. Afan Oromo verbs are also highly inflected for gender, person, number and tenses. Adjectives in Afan Oromo are also inflected for gender and number. Moreover, adverbs can be categorized into: adverb of time, adverb of place, and adverb of manner in which some of the adverbs are affixed. Furthermore, functional words can be classified as prepositions; postpositions and articles markers which are often indicated through affixes in Afan Oromo. Lastly, conjunctions can be separate words (subordinating or coordinating), and some of them are affixed. Since Afan Oromo is morphologically very productive, derivation, reduplication and compounding are also common in the language (Gumii, 1995). The following is detail descriptions and examples of word formation process of Afan Oromo based on the works of (Debela, 2001), (Meiws, 2001) and (Gumii, 1995).

3.4.1.1 Nouns

i. Gender

Afan Oromo has a two gender system (feminine and masculine). Most nouns are not marked by gender affixes. Only a limited group of nouns differ by using different suffixes for the masculine and the feminine form. The language use **-ssa** for masculine and **-ttii** for feminine.

Obbolee<u>ssa</u> *brother* **- obbolee<u>ttii</u>** *sister*

Ogee<u>ssa</u> *expert* (m.) **- ogee<u>ttii</u>** *expert* (f.)

Natural female gender corresponds to grammatical feminine gender. The sun, moon, stars and other astronomic bodies are usually feminine. In some Afan Oromo dialects geographical terms such as names of towns, countries, rivers, etc are feminine, in other dialects such terms are treated as masculine nouns. It is due to this fact that there are different subject forms for the noun **biyya** '*country'*. Example: **biyy**i(*m.)* or **biitti** (*f.*) .

There are also suffixes like **-a, -e** that indicate present and past form of masculine markers respectively. **-ti** and **-tii** for present feminine marker and *-te* past tense marker, *-du* for making adjective form (Debela, 2010). **Biiftuun baate** '*the sun rose'*. The word **baate** takes **-te** to show feminine gender. We can see that **-tii** can also show feminine gender in the following statement. **Adurreen maal ariitii?** *What does the cat run after?* (Mewis, 2001).

ii. Number

Afan Oromo has different suffixes to form the plural of a noun. The use of different suffixes differs from dialects to dialects. In connection with numbers the plural suffix is very often considered unnecessary: **harka ishee lamaaniin** with *her two hand(s).*

According to (Meiws, 2001) the majority of plural nous are formed by using the suffixes – **oota**, followed by **–lee**, **-wwan** , **-een**, **-olii**/ **-olee** and –**aan**. Other suffixes like –**iin** in **sar<u>iin</u>** *dogs* are found rarely.

-oota	**hiriy<u>oota</u>**	*friends*	**jechoota**	*words*
-lee	**gafi<u>lee</u>**	*questions*	**kitabilee**	*books*
-wwan	**saa<u>wwan</u>**	*cows*	**hojiiwwan**	*works*
-olii/-olee	**gango<u>olii</u>**	*mules*	**jarsolii/jarsolee**	*elders*
-een	**fard<u>een</u>**	*horses*	**mukkeen**	*trees*
-aan	**ilm<u>aan</u>**	*children*		

iii. Definiteness

Afan Oromo language does not possess a special word class of articles. Instead demonstrative pronouns are used to express definiteness.

kitaabni kun *this/ the book (Subject)*

kitaaba kana *this/ the book (Object)*

kitaabni sun *that/ the book (Subject)*

kitaaba sana *that / the book (Object)*

To express indefiniteness emphatically the Oromo speaker my use numerical **tokko** *one*, Example: **namni tokko** *one / a man*.

In some Afan Oromo dialects the suffix **-icha** (m.), **-ittii(**n**)**(f.) which usually has a singularize function is used where other languages would use a definite article.

Example:

jaars<u>ichi</u>	*the old man (Subject*)	**jars<u>icha</u>**	*the old man (Object)*
jaart<u>ittiin</u>	*the old women (Subject)*	**jaart<u>ittii</u>**	*the old lady (Object)*

iv. Derived noun forms

Afan Oromo is very productive in word formation by different means. The most common word formation methods are derivational and compounding (Mewis, 2001).

a. Derivation

Derivational suffixes are added to the root or stem of the word. From derived verbal stem and adjectives may be formed by means of derivational suffixes. The following suffixes play an important role in Afan Oromo word derivation. They are **-eenya, -ina, -ummaa, -annoo, -ii, -ee, -a, -iinsa, -aa,-i(tii), -umsa, -oota, -aata,** and **–ooma.**

Examples:

jab<u>aa</u>	*strong*	**jab<u>eenya</u>**	*strength*
jab<u>ina</u>	*strength ,hardiness*	**jab<u>ee</u>**	*intensive*
jab<u>ummaa</u>	*strength*	**jab<u>aachuu</u>**	*to be strong*
jab<u>aachisuu</u>	*to make strong*	**jab<u>eessuu</u>**	*to make strong*
<u>ja</u>jab<u>aachuu</u>	*to be consoled*	**jab<u>eefachuu</u>**	*to make strong for one self*

b. Compound words

On the other hand, it seems that the use of genitive constructions is a very old method of forming compound nouns, as traditional titles shown.

abbaa gadaa	*traditional Oromo president*
abbaa caffee	*chairman of the legislative assembly*
abbaa dubbii	*chief speaker of the caffee assembly*
abbaa duulaa	*traditional Oromo minister of war*

3.4.1.2 Verbs

Verbs are a content word that denotes an action, occurrence, or state of existence. Afan Oromo has base (stem) verbs and four derived verbs from the stem. Moreover, verbs in Afan Oromo are inflected for gender, person, number and tenses

i. Derived stems

The four derived stems the formation of which is still productive in Afan Oromo are:

Autobenefactive	(AS)
Passive	(PS)
Causative	(CS)
Intensive	(IS)

Passive, causative, and autobenefactive are formed with addition of a suffix to the root, yielding the stem that the inflectional suffixes are added to. The personal terminations according to different conjunctions are added to these affixes.

The intensive stem is formed by reduplicating the first consonant and vowel of the first syllable of the stem. The derived stems may be formed from all verbs the meaning of which permits it (Mewis, 2001).

a. Autobenefactive

The Afan Oromo autobenefactive (or "middle" or "reflexive-middle") is formed by adding **-(a)adh**, **-(a)ach** or -(a)at or sometimes **-edh**, **-ech** *or* –**et** to the verb root.

This stem has the function to express an action done for the benefit of the agent himself. Example:

bitachuu *to buy for oneself* the root verb in this case is **bit-**

The conjugation of a middle verb is irregular in the third person singular masculine of the present and past (**-dh** in the stem changes to **-t**) and in the singular imperative (the suffix is **-u** rather than **–i**).

Examples:

bit- *buy*

bitadh- *buy for oneself*

Infinitive and participles are always formed with **-(a)ch**, while the imperative forms have -*(***a)(a)dh** instead of ***-(a)ch.***

Infinitive	imperative sg.	Imperative pl.	English
argachuu	**argadhu**	**argadhaa**	*to find / get*

Argachuu	to find /get	waammachuu	(to call up on)
Sg. 1.p.	argadha	waammadha	
Sg. 2.p.	argatta	waammatta	
Sg. 3.p.m.	argata	waammata	
Sg. 3.p. f.	argatti	waammatti	
Pl. 1.p.	arganna	waammanna	
pl. 2.p.	argattani	waammattani	
pl. 3.p.	argatani	waammatani	

Table 3 : Examples conjugated forms that have -dh only in the first person singular

b. Passive

The Oromo passive corresponds closely to the English passive in function. It is formed by adding **-am** to the verb root. The resulting stem is conjugated regularly.

Example:

beek- *know* **beekam**- *be known*

c. Causative

The Afan Oromo causative of a verb corresponds to English expressions such as 'cause ', 'make ', 'let '.With intransitive verbs, it has a transitiviving function. It is formed by adding **-s, -sis, or -siis** to the verb root example:

deemuu to go **deemsisuu** to cause to go

A second causative of an intransitive verb would create a real causative.

Base stem	causative I	causative II
Agarsiisuu to show	waamsiisuu	(to cause to call)
Sg. 1.p.	n agarsiisa	n waamsiisa
Sg. 2.p.	agarsiifta	waamsiifta
Sg. 3.p.m.	agarsiisa	waamsiisa
Sg. 3.p. f.	agarsiifti	waamsiifti
Pl. 1.p.	agarsiifna	waamsiifna
pl. 2.p.	agarsiifti	waamsiiftu
pl. 3.p.	agarsiisu	waamsiisu

A base (root) stem terminating in **l-** will get a causative stem formed by means of **-ch**, example:

galuu *to enter, return home* **galchuu** *to take home, let enter*

Verbs whose roots end in ′ drop this consonant and may lengthen the preceding vowel before adding -*s*. Example:

ka`uu *to rise /get up* **kaasuu** *to lift up/arouse*

d. Intensive

It is formed by duplication of the initial consonant and the following vowel, geminating the consonant. Example:

waamuu *to call, invite* **wawwaamuu** *to call intensively*

ii. Simple tenses

a. Infinite forms

i) Infinitive

Infinitive is an uninflected form of the verb. In Afan Oromo infinitive form of verbs terminates in -uu. Examples:

arguu *to see* **deemuu** *to* go

On the other hand, the infinitive forms of autobenefactive verbs terminate in **-chuu**. Example:

jiraachuu *to live* **bitachuu** *to buy for oneself*

ii) Participle/ gerund

Participle is a non-finite form of the verb whereas a gerund is a noun formed from a verb (in English the '-ing' form of a verb when used as a noun). In Afan Oromo a participle is formed by adding **-aa** to the verb stem (Mewis, 2001).

Example:

deemaa *going* **jiraachaa** *living*

According to the meaning of the verb these forms may serve as agent nouns.

barsiisaa *teacher* **gaafatamaa** *responsible person*

For these agent nouns feminine forms are used according to the pattern of feminine adjective formation.

barsiiftuu *teacher* **gaafatamtuu** *responsible person*

On the other hand, a gerund is formed by adding **-naan** to the verb stem.

deemnaan *after having gone* **nyaannaan** *after having eaten*

b. Imperative

Imperative singular of base stems and all derived stems beside autobenefactive stems is formed by means of the suffix **-i**. Example:

deemi! *go!* **argi**! *look !*

The imperative singular of autobenefactive stems is formed by means of the suffix **-u**. Example:

jiraadhu! *live!*

Imperative plural of all stems is formed by means of *-aa*.

Example:

deemaa! *go!* **argaa!** *see!*

Negative imperatives are formed by means of **-(i)in** for singular and **-(i)inaa** for plural. Example:

Qubaan jechoota irra hin deemiin. *Don't point on the words with your finger.*

c. Finite forms

The Afan Oromo language uses different conjugations for the verbs in main clauses and in subordinated clauses for actions in present or near future. The first person singular is differentiated from the third person masculine by means of an *-n* that normally is suffixed to the word preceding the verb (Oromoo, 1995).

i) Present tense main clause conjugation

The present tense main clause conjugation is characterized by the vowel **-a**:

deemuu	*to go*
sg. 1.p.	**deema**
2.p.	**deemta**
3.p.m	**deema**
3.p.f	**deemti**
pI. l.p.	**deemna**
2.p. and polite form	**deemtu/deemtan(i)**
3.p. and polite form	**deemu/deeman(i)**

Examples:

gara mana yaalaan deema. *I go to the laboratory.*

ii) Past tense conjugation

The past tense conjugation is characterized by the vowel ***-e:***

deemuu	*to go*
sg. 1.p.	**deeme**
2.p	**deemte**
3.p.m	**deeme**
3.p.f	**deemte**
pI. l.p.	**deemne**
2.p. and polite form	**deemtani**
3.p. and polite form	**deemani**

Example

Kumsaan gara mana barumsaa deeme. *Kumsa went to the school.*

iii) Subordinate conjugation

The subordinate conjugation is used in affirmative subordinated clauses and in connection with the particle **haa** for the jussive. Beside this the subordinate conjugation is used to negate present tense actions.

Deemuu *to go*

sg. l.p	**akkan deemu**
2.p.	**akka deemtu**
3.p.m.	**akka deemu**
3.p.f.	**akka deemtu**
pI. l.p.	**akka deemnu**
2.p. and polite form	**akka deemtani**
3.p.and polite form	**akka deemani**

Examples:

Akka<u>n</u> yaad<u>u</u>tti biqiltootni guutaniiru. *As I thought there are many plants.*

iv) Contemporary verb conjugation

The contemporary verb conjugation is used only in connection with the temporal conjunction **-odoo,-otoo,-osoo,-otuu** or **-utuu** that being connected with this conjugation means '*while'* .The contemporary verb conjugation is a kind of subordinated conjugation with lengthened final vowels(Mewis, 2001).

Example:

*"***<u>Otuun</u> isin waamuu maaliif deemta ?" jedhe**. *"While I was calling you (pI.) why do you go?" he said.*

v) Jussive

To form the jussive in Afan Oromo the particle haa has to be used in connection with the subordinate conjugation. Example:

Isaan haa deemani *they shall go*

vi) Negation

Present tense main clause actions are negated by means of the negative particle **hin** and the verb in subordinate conjugation.

Example:

Maannaaloon <u>hin</u> jir<u>u</u>. *Menelow is not present.*

Present tense actions in subordinated clauses are negated by means of the negative particle **hin** and a suffix -**ne** that is used for all persons. Past tense actions are negated in the same way using the particle **hin** and the suffix -**ne**.

Example:

Sinbirroon halkanii bakka namni arguu <u>hin</u> dand<u>eenve</u> jiraatu.

Bats live in places that people cannot see.

iii. Verb derivation

Some Afan Oromo verbs are derived from nouns or adjectives by means of an affix -**oom**. These verbs usually express the process of reaching the state or quality that is expressed by the corresponding noun or adjective. From these process verbs causative and autobenefactive stems may be formed. Examples:

danuu *much, many, a lot* **guraacha** *black*

danoomuu *to become much* **gurraachomuu** *to become black*

Causative verbs, however, can also be derived directly from adjectives or nouns by suffixing a causative affix -**eess** to the stem of the noun or adjective, example:

danuu *much* **dan<u>eessuu</u>** *to increase, multiply*

Another means to derive process verbs from adjectives in Afan Oromo is to form an autobenefactive stem,

Example:

Adii *white* **addaachuu** *to become white*

iv. Compound verbs

In addition to the above discussed derived verbs, compound verbs can be formed by means of pre-/postpositions, pronouns and adverbs in Afan Oromo such as **ol** *above,* **gad** *below* , **wal, waliin, walitti, wajjin** *together* , **keessa** *in* ,, **jala** *under ;* they precede different verbs and express a broad variety of meanings (Debela, 2010). Examples:

gadi dhiisuu *to let go of* **gaddhiisuu** *to let* go *of*

Compound verbs can also be formed with **jechuu** or **gochuu**.

Example:

With **jechuu**	with **gochuu**
cal **jechuu** (*to be quiet, silent)*	*cal* **gochuu** (*to make quiet silent)*

v. 'To be' and 'to have'

Afan Oromo has different means to express *'to be'*. One of them are copulas, other means are the verbs **ta'uu**, **jiruu** and **turuu** (Mewis, 2001).

The morphemes *(-)***dha** and *(-)***ti** (suffixed or used as independent words) serve as affirmative copulas as well as the vowel **-i** that is added to nouns terminating in a consonant. The copula **dha** is used only after nouns terminating in a long vowel. Negative copula is **miti**, irrespective of the termination of the noun.

Examples:

Present tense:

Atis jabaa dha. *You are strong, too.*

Nouns terminating in a short vowel do not take any copula.

Example:

Isheen durba. *She is a girl.*

Nouns and pronouns terminating in a consonant are combined with the copula.

Example:

Kuni bisbaani. *This is water.*

In all utterances related to possession only the copula *-ti* may be used.example:

Hojiin hundee guddinaa ti! *Work is the basis of development.*

Present progressive:

Waa'een jarreen Axaballaa warra isaaniitiif qofa otuu hin taane uummata naannoofiyyuu hibboo ta'aa iira.

The life of Axaballaa is like a mystery not only, for his family, but also for the people around him.

vi. Past tense:

Sangaan kan <u>eenvuu ture</u>? *Whose ox was it?*

The forms of the verb **qabuu** '*to have'* are overlapping with the forms of the verb ***qabuu*** '*to grasp', 'keep'.*

The verb **qabuu** appears with the meaning '*to have'* only in the present tense and one past tense form. In present tense conjugation both verbs have the same form.

3.4.1.3 Adjectives

An adjective is a word which describes or modifies a noun or pronoun. A modifier is a word that limits, changes, or alters the meaning of another word. Unlike English adjectives are usually placed after the noun in Afan Oromo. For instance, in **Tolaan farda <u>adii</u> bite** *"Tola bought <u>white</u> horse"* the adjective **adii** comes after the noun **farda.** Moreover, in Afan Oromo sometimes it is difficult to differentiate adjective from noun (Meiws, 2001).

Example: **dhugaa** *truth, reality, true, right*

dhugaa keeti *your truth/ you are right* (truth served as noun)

obboleessi hiriyaa dhugaati *brother is the friend for truth / brother is a true friend* (true served as adjective)

i. Gender

In Afan Oromo adjectives are inflected for gender. We can divide adjectives into four groups with respect to gender marking. These are:

a. In the first group the masculine form terminates in –**aa**, and the feminine form in **–oo**.

Example:

guddaa (*m***.)** **nama guddaa** *a big man*

guddoo(*f*.) **nama guddoo** *a big woman*

b. In the second group the masculine form terminates in –**aa**, the feminine form in –**tuu** (with different assimilations).

Example:

dheeraa(*m.)* **nama dheeraa** *a tall man*

dheertuu(*f.)* **intal dheertuu** *a tall girl*

c. Adjectives that terminate in –**eessa** or **–(a)acha** have a feminine form in –**eettii** or –**aattii**.

Example:

dureessa (*m.)* **nama dureessa** *a rich man*

dureettii (*f.)* **nitii dureettii** *a rich woman*

d. Adjectives whose masculine form terminates in a long vowel other than **–aa** as in short vowel **–a** (but not of the suffix –**eessa**/-**aacha**) are not differentiated with respect to their gender.

collee(*m*.) **farda collee** *an active horse*

collee(*f.)* **gaangee collee** *an active mule*

ii. Number

There are four groups of adjectives with respect to number. These are:

a. Most of the adjectives form the plural by reduplication of the first syllable masculine and feminine adjectives differ in plural as they do in singular (Meiws, 2001):

Example:

Singular	Plural
guddaa(*m***.)**	**gugguddaa(***m.)*
guddoo(*f***.)**	**gugguddoo(***f.)*
xinnaa(*m***.)**	**xixinnaa(***m***.)**
xinnoo	**xixinnoo**

pl.f. **lageewwan gu<u>gu</u>ddoo** *big rivers*

pl.m. **qubeewwan gugguddaa fi xixiqqaa** *big and small letters*

b. There is a further plural form which is gender neutral for adjectives of this group beside a special masculine and feminine plural. This plural form

terminates in **-oo**, and is sometimes used with reduplication and sometimes without. Table 4 shows examples of plural adjectives formed by reduplication which are gender neutral

Singular		plural		plural
M	F	M	f	Gender neutral
dheeraa jabaa	Dheertuu Jabduu	Dhedheeraa Jajabaa	Dhedheertuu Jajjabduu	Dhedheertuu Jajjaboo

Table 4 : Examples of gender neutral adjectives

c. Adjectives which may function as nouns as well form the plural only by using noun plural suffixes. Table 5 shows examples of plural adjectives formed using noun plural suffixes

Singular		Plural	
M	F	m	F
dureessa	Dureettii	Dureeyyii/dureessota	dureettiwwan

Table 5 : Examples of plural adjectives

d. Adjectives of the fourth group form the plural without marking the gender, very often by reduplication of the first syllable. Sometimes adjectives of this group form the plural by using a noun plural suffix (Mewis, 2001). Table 6 shows examples of plural adjectives formed by reduplication of the first syllable or using noun plural suffixes.

Singular	Plural	English
Adii	a`adii/adaadii	White
Collee	Colleewwan	Active

Table 6: Examples of plural adjectives formed plural suffixes

iii. Definiteness

The demonstrative pronouns that express definiteness in Afan Oromo follow the adjective if the noun is qualified by an adjective and a demonstrative pronoun as well.

Example:

Namicha dheeraa sana argitee?

Did you see that tall man?

The suffix **–icha** that sometimes has a definite function normally is suffixed to nouns, but it can be suffixed to adjectives or numerals, too,

Example

Lagni guddichi *the big river* **namichi tokkichi** *a single man*

iv. Compound adjectives

In the new terminology of Afan Oromo compound adjectives play a growing role.

Example:

afrogaawaa **afur + rogaawaa** *rectangular four + angled*

sibilala **sibila + ala** *non-metal metal + outside*

3.4.1.4 Adverbs

Adverbs have the function to express different adverbial relations such as relations of time, place, and manner or measure

Some examples of adverbs of time:

amma *now*

booda *later*

Some examples of adverbs of place:

achi(tti) *there*

ala *outside*

Some examples of adverbs of manner:

saffisaan *quickly*

sirritti c*orrectly*

Some examples of adverbs of measure:

baay'ee , danuu *much , many , very*

duwwaa *only, empty*

3.4.1.5 Pre-, Post, and Para-positions

Afan Oromo language uses prepositions, postpositions and para-positions (Meiws, 2001):

i. Postpositions

Postpositions can be grouped into suffixed and independent words.

a. Suffixed postpositions

-tti *in, at, to*

-rra/irra *on*

-rraa/irraa *out of, from*

The post position –**tti** is used to form the locative. The postposition **-rraa/irra** may be used to express a meaning similar to ablative .

Example: **Adaamaa<u>tti</u> yoom deebina ?** *When shall we go back to Adama*?

Gammachuun sireer<u>ra</u> ciise . *Gemachu lay down on bed.*

b. Post position as independent words

ala *outside* **wajjiin** *with , together with*

bira *beside* **teellaa** *behind*

Example: **Namoota nu bira jiraniis hin jeeqnu.** *We don't hurt people who are with us.*

ii. Prepositions

akka *like, according to*

gara *to, in the direction of*

hanga/hamma *until, up to*

karaa *along, the way of , through*

The prepositions **gara**, **hanga**, and **waa'ee/waayee** are still treated as nouns and therefore are used in a genitive construction with other noun they belong to, expression: the direction to, the matter of, etc.

Example:

Namni akka harkaan waa hojjechuuf fayyadamu argi maalitti fayyadamaa?

As people use hands to work something what does the elephant use?

iii. Para-positions

Gara… tti *to*

Gara… tiin *from the direction of*

Example: **Lukkichi rifatee jeedaloo dheesuuf gara manaatti gale.**

The cock was scared and went home to take refuge from the fox.

3.4.1.6 Conjunctions

Conjunctions are unchanging words which coordinate sentences or single ports of a sentence. The main task of conjunctions is to be a syntactical formative element that establishes grammatical and logical relation between the coordinated constituents. According to (Meiws, 2001) the main functions of conjunctions are indentified as: the function of coordinating clauses (coordination), the function of coordinating parts of sentence (coordination) and the function of coordinating syntactical unequal clauses(subordination) . On the other hand, with regard to their form we can subdivide the conjunctions of Afan Oromo into:

i. Independent Conjunctions

a. Coordinating

Example: **garuu** *but*

Hoolaan garuu rooba hin sodaattu. *But the sheep is not afraid of rain.*

b. subordinating

E.g **akka** *that, as if , as whether*

Maaliif akka yaada dhuunfaa yookaan yaada haqaa akka ta'e adda baasii barreessi.

Write separately why it is an individual opinion or that it is an opinion about justice

ii. Suffixed Conjunctions

Example: **–f/ -fi/ -dhaaf** *and, that, in order to, because, for*

Loon horsiisuuf bittee?

Did you buy the cattle for breeding?

iii. Conjunction consisting of one , two or more parts

Conjunctions consisting of two parts can be formed by two independent words or two enclitics or one independent word plus enclitic. They can be formed made up of two single conjunctions that are used after each other in order to give more detailed information about the logical relation or to intensify it.

Example: **akkam akka** *how, that*

Dura namni tokko beekumsa mammaaksaa akkam akka jabeeffatu ilaaluu nu barbaachisa.

At first we have to see how a person extends the knowledge of proverbs

iv. Conjunctions consisting of several segments

Conjunctions consisting of several segments are copulative or disjunctive conjunctions which –as they stand separately from each other –are to emphasize the segments of a parallel construction. These are stable, stereotyped constructions the first segment of which has to be followed by a certain second segment:

Example: **–s… -s ,** *as well as*

Jechoota hudhaa wajjiiniis, hudhaa malees karaa lamaan barreeffaman

Words with glottal stop as well as without glottal stop are written in two ways.

The complexity of Afan Oromo like other morphologically reached languages increases the load on professionals working in the field of natural language processing (NLP). Morphology adds a burden to NLP works. For the purpose of text summarization and also other NLPs, the variant words of a morpheme should be reduced to their root so that they can be counted as one while calculating term frequency, thereby increase the performance of the summarizer. Using stemmer is believed to minimize the difficulty of dealing with different forms of a word (Debala, 2010). Stemming is the process for reducing inflected or derived words to their root. Stemmer is software that does this process automatically. There have been efforts of developing stemming algorithm for Afan Oromo. We used algorithm developed by (Debala, 2010) for our work.

3.5 WORD AND SENTENCE BOUNDARIES

In Afan Oromo, like in other languages, the blank character (space) shows the end of one word. Moreover, parenthesis, brackets, quotes, etc are being used to show a word boundary. Furthermore, sentence boundaries punctuations are almost similar to English language i.e. a sentence may end with a period (.), a question mark (?), or an exclamation point (!) (Taha, 2004).

3.6 NEWS WRITING STRUCTURE

News is an account of what is happening around us. It may involve current events, new initiatives, or ongoing projects or other issues. News writing structure or style is the way in which elements of the news are presented based on relative importance, tone and intended audience. In addition, it is also concerned with the structure of vocabulary and sentences (Parks, 2009).

News writing attempts to answer all the basic questions about any particular event - who, what, when, where and why (the Five Ws) and also often how - at the opening of the article. This form of structure is sometimes called the "inverted pyramid", to refer to the decreasing importance of information in subsequent paragraphs (Parks, 2009). The most important structural element of a story is the lead which is contained in the story’s first sentence. The lead is usually the first sentence, or in some cases the first two sentences, and is ideally 20-25 words in length (Parks, 2009).

CHAPTER FOUR

4. IMPLIMENTATION, EXPERMANTATION AND EVALUATION

4.1 INTRODUCTION

The aim of this chapter is to present how Afan Oromo news text summarizer is implemented based upon the well known Open Text Summarizer (OTS) (Rotem, 2001). Test set has been prepared to conduct an experiment to see the performance of the system with different methods. The application has been tested both objectively using the tool we have developed and subjectively by human evaluators.

4.2 THE OPEN TEXT SUMMARIZER

The Open Text Summarizer (OTS) is an open source tool for summarizing texts. The program reads a text and decides which sentences are important and which are not. It is based on sentence extraction using key term frequency and sentence position methods to calculate sentence importance.

OTS ships with Ubuntu, Fedora and other Linux distributions. OTS Windows version source code is also available in visual C++ and visual C#, etc. It supports more than 25 languages which are configured in XML files. OTS summarizes texts in English, German, Spanish, Russian, Hebrew, Esperanto and other languages. According to Rotem (2001) supporting more languages or tweaking existing languages can be done by editing an XML file of rules. OTS incorporates NLP techniques via language specific lexicons with synonyms and abbreviations in specific language as well as rules for stemming and parsing. These are used in combination with statistical word frequency and sentence position methods for sentence scoring.

The latest version of this open source (toolkit) which has been used as a base for this study is available in C#. C# version of the open source has been selected as it is familiar to the researcher and therefore easier to customize in order to support Afan Oromo text summarization. With OTS, adding new (human) languages is relatively easy; especially for Afan Oromo that use common character sets with English language.OTS lacks documentation even if its source code is readable to understand how it works.

4.2.1 How OTS Works

The English version of OTS that has been used as a benchmark of this study removes common words(stop-words), such as articles like "the" or "a" or conjunctions like "and" and "but," from consideration by using a dictionary list maintained in XML file that accompanies the utility. Words that occur most frequently in the text are assumed to be content bearing and therefore, the sentences that have the highest percentage of the most frequently occurring words are the ones that are used in the output. Like other single-document summarizers, it is based on the idea that the most relevant sentences are those containing the largest number of the most frequent words in the document (stop-words excluded). The most frequent words are usually the ones that better describe the topics of the documents. Besides, English version OTS exploits a simpler grading function which involves a constant multiplicative factor based on the “structure” of the document (e.g., the leading sentence of a new paragraph). For news text summarization the total score of a sentence to be extracted is based on the weight obtained by term frequency multiplied by the constant multiplicative factor. As such a constant number 2 is multiplied by sentence score weight of term frequency for the first sentence of first paragraph, 1.6 is multiplied for every first sentence of other paragraphs. This grading function is effective in producing a summary which is easily readable by humans (Rotem, 2001). Therefore, the grading function of OTS can be represented as:

$$TIVs = \sum tf * c$$

Where, TIVs is Total Importance Value of sentence *s* in a given news item

∑tf is summation of keywords (content bearing term) frequency in sentence *s*

c is a constant multiplicative factor base on sentence position . The value of c is 2 for first sentence of first paragraph and 1.6 for every first sentence of other paragraphs.

For greater accuracy, OTS also references grammatical rules, so that it does not assume, for instance, that the period used to indicate an abbreviation marks the end of a sentence. Similarly, OTS uses the Porter stemming algorithm[7] so that variants of the same word, such as "run," "ran," and "running," are grouped together in the frequency count.

[7] See: http://tartarus.org/~martin/PorterStemmer

According to Rotem (2001), Porter stemming is about 90% accurate, which in turn makes OTS more accurate. Furthermore, collections of synonyms are integrated to enhance term frequency based method.

4.2.2 Performance of OTS

OTS is a single-document summarizer whose implementation was proved to be particularly efficient by recent studies. It is referenced in several academic publications, including reputable journals. In publications such as Oisin and Barry (2007) and Viatcheslav and Timur (2007) OTS is used as a benchmark for other text summarizers or for human summary. In all publications OTS scored very well. According to Yatsko and Vishnyakov (2007), OTS outperformed Subject Search (SSS), Copernic (COP) and Essence (ESS) summarizers. The performance of summarizers is estimated in percentage from the best D-score to find out that OTS outperforms other systems (Yatsko and Vishnyakov, 2007).As it is depicted by Figure 1,among four automatic summarization systems (including the OTS) OTS scored 100% followed by subject search system scoring 97% .

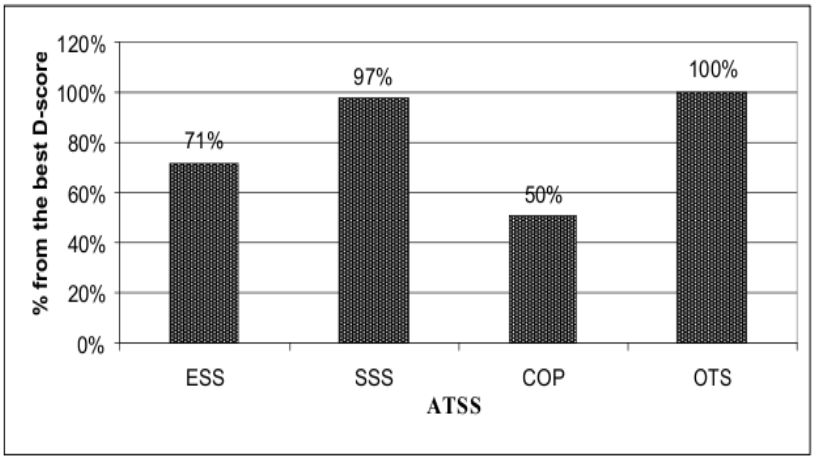

Figure 1: Comparison of performance of OTS with other summarizers. Source: from Yatsko and Vishnyakov(2007)

4.3 IMPLEMENTATION OF AFAN OROMO NEWS TEXT SUMMARIZER

We named our customized summarizer Open Oromo Text Summarizer (OOTS), the version based upon OTS which summarizes Afan Oromo news texts. It is open because we planned to make it open to the public to serve as a framework that can be used for other Latin based Ethiopian languages. The basic principles of OOTS are the same with OTS, but adjustments have been made in order to support Afan Oromo language. Every modification considering the specific rule of the language has been done by creating XML file: oro.xml by modifying the English dictionary: en.xml.

We modified the English mode of XML file and configured the rules of Afan Oromo lexicons[8]**.** The adjustments made to the original OTS in English mode to support Afan Oromo news text summarization are changing the rule of stemming as well as compiling and integrating stop word list, synonyms and abbreviations. In general, for this master's thesis most of the work done in adjusting the OTS code so that it can make use of the Afan Oromo lexicon and actually work for the Afan Oromo language.

4.3.1 Resources required for the OOTS

To customize OTS so as to support Afan Oromo text summarization, we required to have some lexicons and natural language processing tool. These are: Afan Oromo stop-word list, Afan Oromo abbreviation list and list of synonyms as well as the rules for stemming. We found all the components required by the original OTS system for supporting Afan Oromo language even if all are not complete.

i. Afan Oromo stop-word list

Stop-word list are a list of words that should not be stemmed by the stemmer as they are non-content bearing words. Commonly, stop-word list consists of prepositions, conjunctions, articles and particles. The stop-word list compiled by Debela (2010) has been used. Besides, stop-words found in the book entitled: *"A Grammatical sketch of written Oromo"* by Mewis (2001) has been added to enhance term frequency method as Debela's (2010) stop word list is not complete. The total number of stop-words reached 124 that are still incomplete.

[8]Lexicons: The stock of words used in a language or by a person or group of people

Randomly selected sample stop-words are shown in Table 7 and the entire list is available in Appendix-I.

Word	Meaning
Ammo	*however, but*
Garuu	*But*
Bira	*beside, at, near of*
Ala	*outside, out*
Akka	*such as, like, according to*

Table 7 : Sample Afan Oromo Stop-words

ii. Afan Oromo abbreviations

The aim of tokenization is to split the text into sentences, a seemingly trivial task, but which can be complicated by the fact that punctuation marks also serve other purposes, for example, in abbreviations. A language-dependent list of abbreviations is therefore used to prevent false detection of sentence boundaries. We compiled common abbreviations available in different literature (grade 9 to 12 Afan Oromo student text books). Some samples of abbreviations with full meaning are shown in table 8 and the remaining in Appendix-IV.

Abbreviations	Full meaning
k.k.f	Kan kana fakkaatan
w.k.f	Waan kana fakkaatan
Fkn.	Fakkeenyaaf
Hub.	Hubachiisaa

Table 8: Sample Afan Oromo abbreviations

iii. Afan Oromo synonyms

Even if term frequency method is very important to text summarization, it alone is not enough to produce a good quality summary (Edmundson, 1969). It has been criticized for the reason that there may be more than one word to express the same thing which is termed as synonyms. With synonyms one concept can be expressed by different words. For example **waangoo**‘*fox’* and **jeedala** ‘*fox’* refer to same kind of animal.

A list of available Afan Oromo synonyms are prepared for Afan Oromo dictionary and configured to **oro.xml** file to enhance the term frequency based method we compiled the list of synonyms from Afan Oromo dictionary entitled, "Galmee jechoota Afaan Oromoo". Table 9 below contains some of Afan Oromo synonyms. The complete list that we used in this work is found in Appendix III.

Term	Synonymy	Meaning
Tolchuu	**Gochuu**	*Make*
Dhibamuu	**Dhukkubsachuu**	*Sick*
Qooduu	**Hiruu**	*Share*
Jijjiiruu	**Diddiiruu**	*Change*
Herreguu	**Yaaduu**	*Think*

Table 9: Sample synonyms words

iv. Afan Oromo Stemmer

In our work, we have used lightweight stemmer rules for Afan Oromo that strips the suffixes using a predefined suffix list using the algorithm developed by Debela (2010). This system takes as input a word and removes its suffixes according to a rule based algorithm. The algorithm follows the known Porter algorithm for the English language and it is developed according to the grammatical rules of the Afan Oromo. According to Debela (2010) an evaluation of the system showed the algorithm accuracy giving 96 percent correct results. Therefore, for our system, we compiled lists of affixes integrated to: oro.xml file to apply the rule of stemming to our OOTS similar to the Porter's stemmer used by OTS. The complete list of suffixes is available in Appendix II.

4.3.2 Summarization process and techniques used

The adopted summarization method is sentence extraction based. It has three major steps: (i) preprocessing, (ii) sentence ranking and (iii) summary generation.

i. Preprocessing

As is in other ATS systems, preprocessing step includes tokenizing, stop-word removal, stemming and parsing (breaking the input document in to a collection of sentences). For stop word removal, we have used the Afan Oromo stop-word compiled from different literature in addition to the stop-word list prepared by Debela (2010).

Furthermore, using stemmer, a word is split into its stem and affix after stop-word removal. Affixes striped can be replaced by another affix or replaced by white space as per the rule it matches with. The design of a stemmer is language specific, and requires some significant linguistic expertise in the language. A typical simple stemmer algorithm involves removing suffixes using a list of frequent suffixes, while a more complex one would use morphological knowledge to derive a stem from the words. Since Afan Oromo is a highly inflectional language, stemming is necessary while computing frequency of a term.

ii. Sentence Ranking

After an input document is formatted and stemmed, the document is broken into a collection of sentences and the sentences are ranked based on two important features: term frequency (TF) and sentence position.

TF is frequency of keyword appearance in an article. This method is the earliest known method to be used for automatic text summarization since research began in this area. It is based on the idea that the most relevant sentences are those containing the largest number of the most frequent words in the document (stop-words excluded) (Luhn, 1958). With the *tf (term frequency) method*, the importance value (score) of a sentence *s* (IV*s*) is given by: $$IVs = \sum tf$$

Where, IV is Importance Value based on term frequency

tf, is Term frequency

On the other hand, positional value **(**score**)** of a sentence *s* is computed in such a way that the first sentence of a document gets the highest score and the last sentence gets the lowest score in news domain as the original OTS uses constant multiplicative factor of term frequency score calculated. The positional value for the sentence *s* is computed using the following formula by combining two parameters for sentence ranking. Therefore, the total importance value (score) of a given sentence *s* (TIVs)

$$TIVs = IVs * c$$

Where, c is constant multiplicative factor. The value of c is 2 for first statement of first paragraph, 1.6 for first sentences of all other paragraphs. All other sentences are weighed only by their term frequency score.

TIVs, is total score of importance value of a sentence based on term frequency and position value

iii. Summary Generation

A summary is produced after ranking the sentences based on their scores and selecting N-top ranked sentences, where the value of *N* is set by the user. To increase the readability of the summary, the sentences in the summary are reordered based on their appearances in the original text; for example, the sentence which occurs first in the original text will appear first in the summary.

4.3.3 Architecture of OOTS

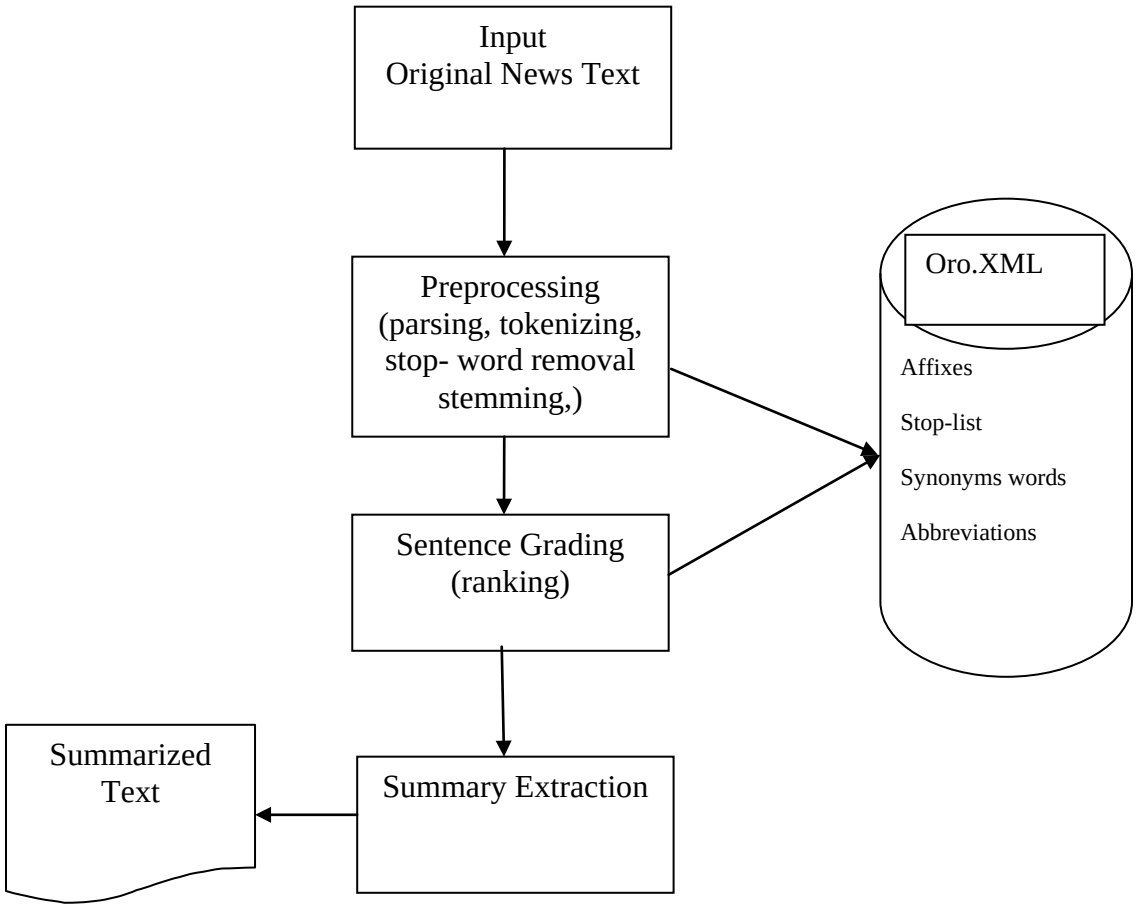

Figure 2 : Architecture of the summarizer

4.3.4 User Interface of the summarizer

Using our customized summarizer (OOTS) the summary sentences are re-arranged in their natural order in the news and presented to the user. Figure 3 shows user interface of the summarizer used for experimentation.

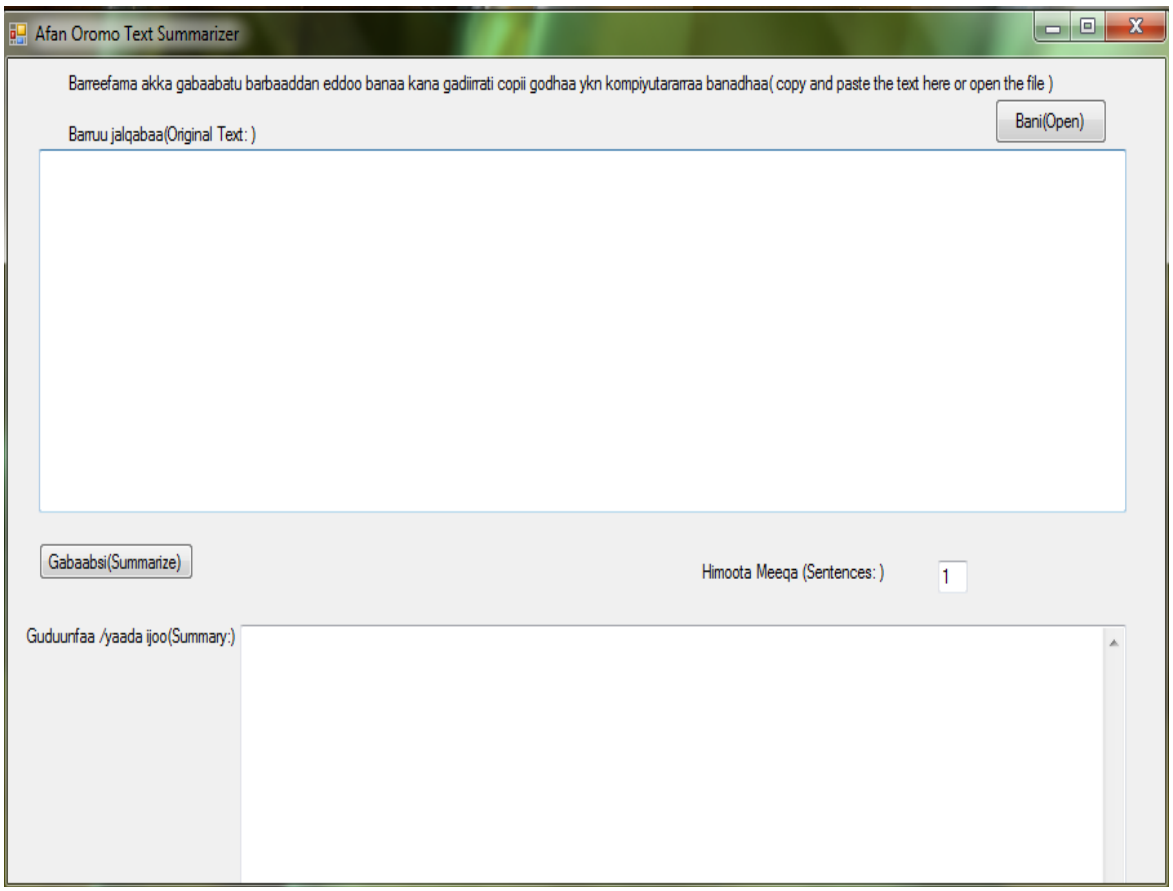

Figure 3: User interface of the summarizer

4.4 EXPERIMENTATION

4.4.1 Corpus preparation

As it has been discussed in Section 1.5.1 of this study, the corpus is prepared from scratch as there is no previous work in the area of text summarization on Afan Oromo language. The sources of the test set, where news items for experimentation selected, are Oromia Radio and Television Organization (ORTO) and Voice of America (VOA) Afan Oromo official websites written on different topics. During selection, news articles are considered to be on different topics of the community: social, economical, technological and political issues so that they are a potential source for collecting balanced corpus for the task of news text summarization. Table 10 shows the statistics of the experimentation corpus. The collection wide average length of news items is approximately 277 words or 11 sentences. Moreover, the table depicts the compression rate in % of the summary tested for experimentation.

Text ID	News size in words	News size in sentences	Compression rate (% of summary tested)
Test 1	250	11	10
Test 2	403	14	10
Test 3	250	9	20
Test 4	231	10	20
Test 5	290	13	30
Test 6	295	14	30
Test 7	232	11	40
Test 8	269	13	40
Average	**277.5**	**11.87**	_

Table 10 : Statistics of the experimentation corpus

4.4.2 Summary preparation

4.4.2.1 Reference Summary

In the manual summary generation, four journalists from Oromia Radio and Television Organization are involved. The experts were provided with eight news items with a guideline (Reference summary preparation guideline is available in Appendix-V). The guideline was required to avoid misunderstanding. In addition to the guideline, introduction about automatic summarization was given to the experts. Furthermore, they were instructed to consider the summary's linguistic qualities (Informativeness, Non-redundancy, Referential clarity, Grammar and coherence and structure) while ranking sentences. Accordingly, they ranked the sentences of the news from the most important to the least important based on the sentences' relevance for generic summary.

The reference summary is generated based on the four human summarizers' average rank of sentences. The average rank for a sentence is computed as the sum of the four ranks divided by four. The sentences are sorted by their average rank in descending order and the top *N* numbers of sentences having maximum score are selected at a required compression rate for the reference summary. The prepared reference summary for each test set is required to be compared against summaries generated by the summarizer for the purpose of performance evaluation (see Section 4.5.1)

4.4.3 Experimentation methods

For each news item, three experiments have been conducted with different methods (M1, M2 and M3). All the summaries extracted by each experiment at a given extraction rate are compared against one reference summary. The three experiments with different methods are:

i. M1: using original English language mode of the summarizer with term frequency and sentence position methods but, without Afan Oromo stemmer and other language specific lexicons (synonyms, stop-word list and abbreviations)

In this case, the original English version summarizer has been used to summarize news items in Afan Oromo given the English language version summarizer, as the rules for stemming, synonyms and abbreviations do not match for Afan Oromo; it can be considered as a summarizer for Afan Oromo without stemmer and other language specific lexicons (synonyms, stop-word list and abbreviations).

ii. M2: using term frequency and position method with Afan Oromo stemmer and other language specific lexicons.

In this case, we planned to test the performance of the summarizer with the term frequency and position methods. To enhance the effect of term frequency method on selecting informative sentences for extraction and consequently improve the performance of the summarizer, we included rule of stemming and language specific lexicons of Afan Oromo. This mode of the summarizer directly access Afan Oromo language specific dictionaries developed with file name "oro.xml" (consisting of Afan Oromo stemming rule, synonyms, stop-word list, and abbreviations).

iii. M3: using combinations of term frequency with improved position method with stemmer and other lexicons (synonyms and abbreviations)

In this case, we planned to test the performance the summarizer by doubling the score of the position method. As it has been discussed in Section 4.3.2, we use the sentence position method which is the constant multiplicative factor of the term frequency method score to grade (rank) sentences. In order to produce good summary of news items the position method should be given emphasis based on the inverted pyramid structure of the news text. We, therefore, improved the position method by doubling the value of the constant multiplicative factor so as to increase its effect on the total score of a sentence which is given by the following formula:

$$\text{TIV}s = \sum tf * c$$

Where TIV, is total importance value of sentence *s*

tf, term frequency of content bearing terms

c, constant multiplicative factor . In this case, the value of *c* is 4 for the first sentence of first paragraph and 3.2 for every first sentence of other paragraphs instead of the commonly used values (2 for the first sentence of the first paragraph and 1.6 for the first sentences of other paragraphs).

Therefore, for each news article discussed in Section 4.4.1, three summaries (using the three methods: M1, M2 and M3) at a given extraction rate has been generated by the summarizers. The following section discuses the performances of the summarizers.

4.5 EVALUATION AND DISCUSSION OF RESULTS

One of the usual and challenging tasks to be carried out in any research is evaluation and discussion of the result. For this study, the summarizers are evaluated using objective and subjective methods. Both subjective and objective evaluation methods used are intrinsic to the summary.

4.5.1 Subjective evaluation

We adapted the subjective evaluation technique used by GreekSum (Pachantouris, 2004), a text summarizer for Greek language in order to evaluate the summaries generated by our customized summarizers (OOTS). The eight Afan Oromo news items, of various contents, used in the objective evaluation have also been used for this evaluation. In our evaluation process three different system summaries (generated by the three different methods discussed in Section 4.4.3) have been evaluated by experts. The three system summaries created with different methods are compared according to the following three check points.

i. In which summary the most important information is kept?
   The experts (evaluators) check the summary's informativeness. Specifically, they check whether the summary includes best sentences that contain the most important information about the topic and satisfies information need of readers or not.
ii. Out of a scale from 1-5, where 5 is the best, what score would you assign to each summary?
   This question checks the linguistic quality. It includes the assessment of grammar, non-redundancy and referential clarity.

iii. Which summary is more coherent?

The evaluators check whether the summary has smooth transition of sentences. While reading the sentences in their rank order, it should not just be a heap of related information, but also should build a coherent body of information.

4.5.1.1 Results of subjective evaluation and discussion

In this section we present results of the subjective evaluation based on the three points and interpretation of the results (Evaluation guideline and result are available in Appendix VI and VII).

i. In which summary the most important information is being kept?

The informativeness of the summary created by one of the three methods for each test item is scaled to 100 out of the expected total 4 votes by the four evaluators. For instance, if M1 is selected by two of the evaluators, the percentage of the informativeness of the summary is measured as 2/4= 0.5 i.e. 50%. Table 11 depicts the percentages and average performance of the three methods as it is judged by evaluators. As it is the case in objective evaluation, result of this evaluation shows that the summary created by M3 is the most informative as compared to other methods. With informativeness, M3 outperformed M1 and M2 by 28.13% and 25% respectively. Due to the improved position method, the result of M3 is satisfactory with the most important information being kept in the summarized text. Additionally, the content is clear and the basic meaning of the original text is kept. Surprisingly, the average performance result by M1 is almost the same as M2 i.e. the use of Afan Oromo stemming rule and language specific lexicons did not bring much improvement. On the other hand, though in principle the informativeness of the summaries increases as the extraction rate increases (as we go down form test1 – test 8), the evaluation result rarely shows the improvement of the informativeness of the summary as human evaluation result is highly subjective.

Text ID	System Summary with different methods		
	M1	**M2**	**M3**
Test 1	25%	25%	75%
Test 2	25%	25%	100%
Test 3	25%	50%	25%
Test 4	50%	50%	50%
Test 5	50%	50%	50%
Test 6	25%	25%	75%
Test 7	50%	25%	50%
Test 8	25 %	50%	75%
Average	**34.37 %**	**37.5 %**	**62.5 %**

Table 11 : Information preserved analysis result

ii. Out of a scale from 1-5, where 5 is the best, what score would you assign to each summary?

This assessment method is intended to evaluate the overall language qualities of the summary such as grammar, non-redundancy and referential qualities (see Appendix-VI). The results from the users are turned into statistics based on the added score of the four results and compared on a scale of 100. For instance, if the summary of Test1 using M1 is scored 3 by evaluator1, 4 by evaluator2 , 5 by evaluator3 and 2 by evaluator4 , the percentage of the overall linguistic quality of the summary produced by M1 for Test1 is the average of the sum of the scores in percentage i.e. 3 + 4 + 5 + 2 =11/20 = 55%. As depicted in Table 12, the same is true as it is the case in objective evaluation that; M3 outperforms other methods M1 and M2 by 5.63% and5% respectively as it is judged by evaluators. The average performance of the three methods shows that the overall linguistics quality of the summaries is almost the same. This implies that system summaries produced by the three methods, they are non-redundant, do not contain illegal breaks and use the rules of the punctuation properly.

On the other hand, it is evident that the summarizers do not resolve referential integrity (pronoun resolution) that is perhaps a factor that diminishes the overall performance of the summarizers.

Text ID	System Summary with different methods		
	M1	**M2**	**M3**
Test 1	55 %	65%	75%
Test 2	45 %	55%	90%
Test 3	55%	55%	80%
Test 4	70%	60%	75%
Test 5	60%	60%	65%
Test 6	55%	55%	60%
Test 7	70%	70%	80%
Test 8	65%	60%	85%
Average	**59.37%**	**60 %**	**65 %**

Table 12: Linguistics quality rating result table

iii. Which summary is more coherent?

Like in the previous cases, the average performance for coherence depicted by Table 13 shows that M3 performed better than M1 and M2 by 53.13% and 46.88%. The use of improved position method that gives more weight for the first sentence of first paragraph as well as first sentence of all other paragraphs able to create more coherent and well structured summary than the other methods . In all methods the summarizer gave better results as summary sentences are re-arranged in their natural order in the news.

Text ID	System Summary with different methods		
	M1	**M2**	**M3**
Test 1	25%	25%	75%
Test 2	25%	25%	100%
Test 3	25%	25%	75%
Test 4	25%	50%	50%
Test 5	50%	0%	50%
Test 6	0%	25%	100%
Test 7	25%	25%	75%
Test 8	0%	50%	75%
Average	**21.87%**	**28.12%**	**75%**

Table 13: Coherent information analysis result

4.5.1.2 Limitations of subjective evaluation method

As discussed in previous section the performance result is satisfactory with the most important information as first sentence being kept in the summarized text through method M3. In all cases, the system summary content is clear and the basic meaning of the original text is kept. Moreover, no illegal breaks were found and the rules of the punctuation were used.

The overall performance evaluation result of subjective evaluation shows that the summarizer denoted by M3 is the best performing even if the results are totally subjective and according to every evaluator's personal judgments. For further investigation, objective evaluation is carried out where we found that the result of objective evaluation is also consistent with these findings.

4.5.2 Objective evaluation

It is one of the evaluation methods employed for this study to measure effectiveness of the summarizers. The objective evaluation assumes one and only one best (reference) summary and compares the system summary against the reference summary. It is intended to measure the system's summary approximation to the reference summary on the basis of standard recall (R), precision (P) and F-measure (F).

The standard recall and precision measures are calculated as follows and f-measure is calculated based on the values of precision and recall:

- $Recall(R) = correct / (correct + missed)$
- $Precision(P) = correct/ (correct + wrong)$
- $F - measure(F) = 2 * R * P/ (R + P)$

Where:

- **Correct** = the number of sentences in both the summarizer's summary and the reference summary,
- **Wrong** = the number of sentences in the summarizer's summary but not in the reference summary,
- **Missed** = the number of sentences in the reference summary but not in the summarizer's summary.

Summaries are required to be generated at four compression rates 10%, 20%, 30% and 40%. From a total of 8 news articles prepared for experimentation, a pair of news items randomly selected to be input to the system for a given extraction rate. For instance, the first two news items extracted at compression rate of 10% and the second two news items at 20% compression rate, etc. the system summaries of all news items are compared against one best reference summary created by expert summarizers. The tool used to compute the standard recall, precision and f-measures has been developed and integrated with the tool.

4.5.2.1 Results of objective evaluation and discussion

The results of the three experimentation methods (M1, M2 and M3) have been compared against the reference summary using the tool we developed and integrated with the tool, we computed the standard precision and recall as well as F-measure. As it has been discussed in previous section, **precision** is the ratio of correct sentences available in both summarizer's summary and reference summary to instances that are generated by the summarizer, while **recall** is the ratio of correct sentences available in both summarizer's summary and reference summary to instances that are available in reference summary. Moreover, both precision and recall are 0 when none of the sentences in system's summary exists in reference summary and 1 if all of the sentences in system's summary are available in the reference summary. On the other hand, since recall and precision are highly sensitive to extraction rate, **F-measure** (score) can be interpreted as a weighted average of the precision and recall, where an F-measure reaches its best value at 1 and worst score at 0. The value of F-measure is 0 if either precision or recall is 0, between 0 and 1 for none 0 values of precision and recall and undefined if both precision and recall is 0.

For the sake of discussing evaluation result, we used F-measure as it is the weighted average of precision and recall. Table 14 summarizes the evaluation results. As can be seen from the table, f-measure is undefined for the first two test sets: Test1 and Test2 for which the summaries are generated using M1 and M2 at a compression rate of 10%. This is because the reference summary and system summary have no common sentences. On the other hand, as it can be observed in the table, the average F-measures of the three methods M1, M2 and M3 are 34%, 47% and 67% respectively, where M3 outperformed the two methods M1 and M2 by 32% and 20% that shows the improvement of the summarizer with improved position method. Moreover, the gap between the performances of M1 and M2 shows the improvements of the summarizer with improved term frequency method using the language specific lexicons.

Surprisingly, even if the average performance by principle increases with an increase in extraction rate, it did not hold for this experimentation result as it can be observed, better or similar result can be attained a 20% than 30% or 40% extraction rates. This can be due to the limitation of reference summary created by human experts which is highly subjective as there is no correct 'reference summary'.

Text ID	Compression rate in %	M1			M2			M3		
		P	R	F	P	R	F	P	R	F
Test 1	10	0%	0%	-	0%	0%	-	100%	100%	100%
Test 2	10	0%	0%	-	0%	0%	-	100%	100%	100%
Test 3	20	50%	50%	50%	100%	100%	100%	100%	100%	100%
Test 4	20	50%	50%	50%	50%	50%	50%	50%	50%	50%
Test 5	30	25%	25%	25%	50%	50%	50%	50%	50%	50%
Test 6	30	25%	25%	25%	50%	50%	50%	75%	75%	75%
Test 7	40	50%	50%	50%	50%	50%	50%	75%	75%	75%
Test 8	40	75%	75%	75%	75%	75%	75%	100%	100%	100%
Average score	**_**	**34%**	**34%**	**34%**	**47%**	**47%**	**47%**	**81%**	**81%**	**81%**

Table 14 : Objective evaluation result

Figure 4 compares the performance of the summarizers with the three methods (M1, M2 and M3) for a given test set. As it can be observed, M3 is best performing experiment followed by M2 and M1 respectively for most of the test sets.

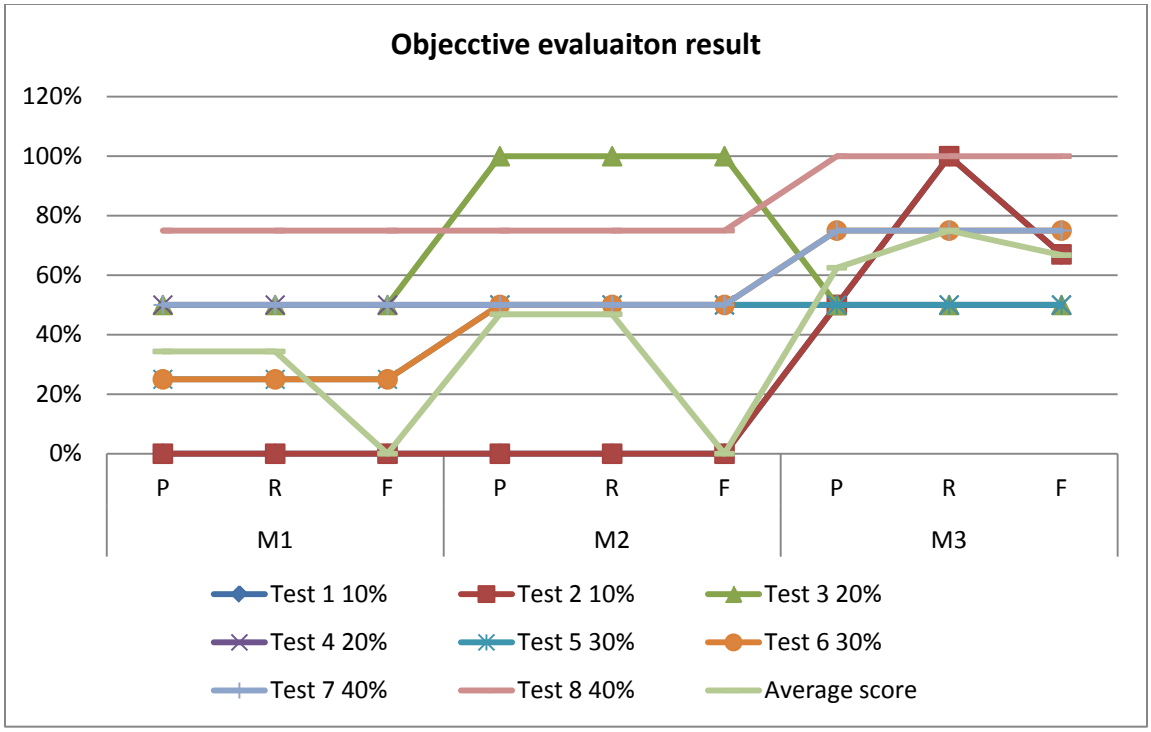

Figure 4: Comparison of performance results of the three methods (M1, M2 and M3)

4.5.3 Comparison of objective and subjective evaluation results

Both subjective and objective evaluation has been conducted for a summary extracted by three different methods. The results of both evaluations support each other; in all cases the summary using M3 outperformed the other methods as it has been discussed in previous section, the performance of the summarizers for subjective evaluation shows for the three methods (M1, M2 and M3) the average informativeness of the summary is (34.37%, 37.5% and 62.5%); the average language quality is (59.37%, 60% and 65%) and the average coherence and structure is (21.87%, 28.12% and 75%) respectively. On the other hand the objective evaluation result shows that the average f-measure score for (M1, M2 and M3) is (34%, 47% and 81%).

CHAPTER FIVE

5. CONCULUSIONS AND RECOMMENDATIONS

This study deals with the development and evaluation of the first automatic text summarizer for Afan Oromo news text, named OOTS (Open Oromo Text Summarizer) which is based upon the Open Text Summarizer (OTS). We customized OTS by modifying the code so that it can support Afan Oromo language. Both subjective and objective evaluations were carried out where we found that using the combination of improved position and term frequency methods showed a promising result. This chapter gives conclusions and recommendations based on the findings of the study.

5.1 Conclusions

The following concluding remarks were made based on the findings:

- For this master's thesis the most work done were adjusting the OTS code so that it can be used of the Afan Oromo language and developing and integrating the automatic evaluation tool to evaluate the performance the summarizer objectively using the standard precision, recall and F-measure.

- The overall subjective and objective evaluation results show the effect of the included Afan Oromo stemming rule, stop-word list and synonyms brought some improvement for the summarizer and are not as expected. However, as compared with subjective evaluation, objective evaluation result showed that there is a benefit from including the stemmer and language specific lexicons.

- The current implementation of OOTS is a prototype and it uses a simple stop-list, stemming rule and synonyms in order to filter and identify the important keywords in the text for term frequency based method.

- We modified the position method and it really improved the performance of the summarizer as it gives higher weight for first sentences of the news item following the news writing structure. The results obtained while improving the position method have proved the claimed writing style of the news. The news items used for

experimentation were written using the inverted pyramid writing style. Hence, the combination of improved position method and term frequency with Afan Oromo stemmer and other lexicons (synonyms, abbreviations and stop-word list) improved the performance of the summarizer that can serve as a model.

- Even if the preliminary results by OOTS (as indicated by M3 in the experimentation) is relatively good as compared to other two methods described in this paper, the evaluation was carried out on relatively small data sets and, therefore OOTS needs a further development and testing. The results of both objective and subjective evaluations have shown relatively harmonized result about the effectiveness of the summarizer.

5.2 Recommendations

Based on the findings and knowledge acquired from literature the following recommendations are forwarded:

- We strongly recommend a balanced well prepared corpus is an essential part for further evaluation of the performances of the summarizers.

- We also recommend complete stop-word list, synonyms, and abbreviations are very useful to enhance term frequency based method.

- Beside this implementation, more languages can be included in the OTS system especially Latin-based Ethiopian languages can use the OOTS system as a framework to develop a summarizer. The algorithm to be used is basically similar and can be easily adjusted to serve the needs of different languages.

- From the evaluation, it is evident that the summarizer with method (M3) i.e. improved position method with term frequency based works well for news text summarization. Though it is not perfect, but it can still add much to an existing gap in the Afan Oromo reader community to access a summarized news texts to save their time. Being the first tool for this language, further task is required to make it freely accessibly on the Internet for everyone to use.

- Like other extraction based summarizers, the result of this study’s summarizers lack coherence, therefore more advanced method to implement in future version is the usage of abstract summarization method that the resulting summary is an interpretation of the original text. The results will be much more coherent but this method is not easy to implement.

References

1. Abera N. (1988), "Long *vowels in Afan Oromo: A generic approach*" , Master's thesis , School of graduate studies, Addis Ababa University, Ethiopia.

2. Abracos J. and Lopes G.P. (1997), "*Statistical methods for retrieving most significant paragraphs in newspaper articles"* ,In proceedings of ACL/EACL Workshop on Intelligent Sealable Text Summarization,.
3. Alguliev R. and Aliguliyev R .(2009) ," *Evolutionary Algorithm for Extractive Text Summarization*," pp.128-138.
4. Booguraev B. and Kennedy C. (1997), "*Salience-based content characterization*", In Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization.

5. Brandow R., Mitze K.and Rau L. (1995), "*Automatic condensation of electronic publications by sentence selection*", Information processing and Management,

6. D'Avanzo E., Magnini B., and Vallin A. (2004) ,"*Keyphrase Extraction for Summarization Purposes*", *The LAKE System at DUC-2004 , Boston, USA.*

7. Dalianis, H., M. Hassel, J. Wedekind, D. Haltrup, K. de Smedt and T.C. Lech. (2003),"*Automatic text summarization for the Scandinavian languages*." In Holmboe, H.ed.) Nordisk Sprogteknologi .

8. Dalianis H. (2000), "*SweSum - A Text Summarizer for Swedish*", Interaction and Presentation Laboratory (IPLab).

9. Debela T. (2010), "*Designing a Stemmer for Afan Oromo Text: A hybrid approach*", Master's thesis, School of graduate studies, Addis Ababa University, Ethiopia.

10. Dipanjan D. (2007), "*A Survey on Automatic Text Summarization Single-Document Summarization*," *Language,* pp. 1-31.

11. Diriba M. (2002), "*An automatic sentence parser for Oromo language using supervised learning techniques***",** Master's thesis, School of graduate studies, Addis Ababa University, Ethiopia.

12. Edmundson H. P(1969), "*New Methods in Automatic Extracting*," Computing, vol. 16, no. 2, pp. 264-285.

13. Farmin T.and Chrzanowski M.J (1999), "*An evaluation of text summarization systems*", 1999.

14. Ganapathiraju M.(2002), "*Relevance of Cluster size in MMR based Summarizer*" , A report in proceedings of the Second NTCIR Workshop.
15. Grage G. & Kumsa T.(1982), "*Oromo dictionary*", African studies center. Michigan state University.
16. Gupta V. and Lehal G. S.(2010), "*A Survey of Text Summarization Extractive Techniques*," *Text,* vol. 2, no. 3, pp. 258-268.
17. Hassel M.(1999) "*FarsiSum - A Persian text summarizer*," Cognitive Science, pp. 2-4.

18. Helen A. (2006), "*Automatic Text Summarization for Amharic Legal Judgments*", Master's Thesis, Faculty of Informatics, Addis Ababa University. Addis Ababa.

19. Hennig L., Umbrath W. and Wetzker R.(2008), "*An Ontology-based Approach to Text Summarization*," pp. 1-4.
20. Hovy E. and Lin C. (1999), "*Automated Text Summarization in SUMMARIST*," Advances in Automatic Text Summarization. MIT Press.

21. Inderjeet M.(2001), "*Summarization Evaluation: An Overview"*

22. Ishikawa K., Ando S., and Okumura A. (2001), "*Hybrid Text Summarization Method based on the TF Method and the LEAD Method*," Language, no. 1.
23. Kaikhah k. (2004), "*Automatic Text Summarization with Neural Networks*", in Proceedings of second international Conference on intelligent systems, IEEE, 40-44, Texas, USA.

24. Kaili M. & Pilleriin M. (2005), "*ESTSUM - Estonian newspaper texts summarizer",* Proceedings of The Second Baltic Conference on Human Language Technologies Pp. 311-316

25. Kamil N.(2005).," *Automatic Amharic News Text Summarizer*", Master's Thesis, Faculty of Informatics, Addis Ababa University, Addis Ababa.

26. Khoo S., and Goh D.H.(2007), "*Automatic Multi-document Summarization of Research Abstracts : Design and User Evaluation,"* Journal of the American Society for Information Science, vol. 58, pp. 1419-1435.

27. Kupiec J., Pedersen J. and Chen F. (1995), " *A trainable document summarizer*", Proceedings of the 18th Annual International ACM Conference On Research and Development in Information Retrieval (SIGIR).
28. Lawrence H. , Hyoil H. and Ari D. (2007) , " *The use of domain-specific concepts in biomedical text summarization* " , Information Processing and Management: an International Journal.
29. Lin C.(1999), "*Training a selection function for extraction* ", Proceedings of the 8th International Conference on Information and Knowledge Management .

30. Lloret E.(2008) ,"*Text summarization: an overview*" ,Dept. Lenguajes y Sistemas Inform_aticos Universidad de Alicante Alicante, Spain.
31. Luhn H. P. (1958), "*The Automatic Creation of Literature Abstracts*", pp. 159-165.

32. Manabu O. and Hajime M. (2000), "*Query-Biased summarization based on lexical chaining*", Computational Linguistics, 16, 578-585.

33. Mani I., Bloedorn E.and Gates B.(1998), " *Using cohesion and coherence models for text summarization*" , In AAAI 98 Spring Symposium on Intelligent text summarization.

34. Marcu D. (1998), "*The rhetorical parsing, summarization, and generation of natural language texts*" , Proceedings of the COLING-ACL Workshop on Very Large Corpora.

35. Meiws C.G.(2001), "*A grammatical sketch of Written Oromo*", ISBN 3- 89645- 039-5.

36. Moens M. (1997), "*Automatic indexing and Abstracting of Document texts*", Belgium, Kluwer Academic Publisher.

37. Oisin B. and Barry S. (2007) , "*From social bookmarking to social summarization: an experiment in community-based summary generation*" , International Conference on Intelligent User Interface

38. Pachantouris G. and Dalianis H. (2005), "*GreekSum A Greek Text Summarizer*," Word Journal of the International Linguistic Association, pp. 1-45.

39. Park S.C.(2004), "*Generation of Non-redundant Summary Based on Sum of Similarity*," Computing, pp. 3-4.

40. Pembe F. C. and T. Güngör (2007), "*Automated Query- biased and Structure-preserving Text Summarization on Web Documents*", Proceeding Symposium on Innovations in Intelligent Systems and Applications, İstanbul.

41. Radev D.R.(2001), "*Introduction to the Special Issue on Summarization*," Computational Linguistics, pp. 1-11.

42. Rejhan B., Damir K., and Bojan S. (2009), "*Automatic Text Summarization*," Information Systems.

43. Rotem N.(2001), " *Open Text Summarizer*" . Available at: http://libots.sourceforge.net/

44. Schlesinger J., and Baker D. (2001), "*Using Document features and Statistical Modeling to Improve Query-Based Summarization*", DUC.

45. Svore, K., Vanderwende, L., Burges, C. (2007), "*Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources*", In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

46. Taha R.(2004), "*MODERN AFAAN OROMO GRAMMAR*" ISBN: 9781468515060.

47. Teferi A.(2005), " *The application of Machine learning Technique (NAÏVE BAYES) for Automatic Text Summarization the Case of Amharic News Text*", Master's Thesis. Faculty of Informatics , Addis Ababa University, Ethiopia.

48. Teufel S. and Moens M.(1997), " *Sentence extraction as a classification task*", Proceedings of the ACL Workshop on Intelligent Text Summarization ,Madrid.

49. Tilahun G.(1993), " *Qubee Afan Oromo : Reasons for choosing the Latin script for developing an Afan Oromo Alphabet*" Journal of Oromo studies .

50. Viatcheslav Y., Timur V. (2007), "*Evaluating contemporary automatic text summarization systems: an experiment*", Computational Linguistics Laboratory (CLL).

51. Yatsko V. A. and Vishnyakov T. N. (2007 **),** "*A method for evaluating modern systems of automatic text summarization*" , Automatic Documentation and Mathematical Linguistics

52. Gumii Qormaata Afan Oromoo,(1995) " *Caasluga Afan Oromo , Jildi I*", Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee.

53. Parks B. (2009), "*BASIC NEWS WRITING*", united states. Available at http://www.ohlone.edu/people/bparks/.../basic**newswriting**.**pdf** accessed on April 12,2012

List of Appendixes

1. Appendix-I: Afan Oromo stop-word

waan	ofii	akka	Kun	sun	an	kan	inni
isheen	isaan	nu	nuyi	keenya	keenyaa	koo	kee
sun	ani	ini	Isaan	iseen	isaa	akka	kan
koo	kee	Ammo	Garuu	yookaan	yookiin	akkasumas	Booda
Erga	Eega	kanaaf	kanaafi	kanaafuu	tanaaf	tanaafi	tanaafuu
Fi	Immoo	Moo	Illee	akka	jechuu	jechuun	jechaan
Osoo	Odoo	Ituu	Akkum	akkuma	booda	booddee	Dura
Kanaafi	Saniif	tanaaf	tanaafi	tanaafuu	waan	itumallee	otumallee
Ituullee	Otuullee	ennaa	Henna	innaa	hoggaa	oggaa	hogguu
Yeroo	Yommuu	yammuu	Yemmuu	yommii	simmoo	oo	Woo
Akka	Ituu	Odoo	Silaa	yeroo	hanga	erga	Osoo
ishee	kan	kun	eegasii	yookinimoo	utuu	kanaaf	tahullee
Akkam	Otoo	iseen	Keetii	yoom	eegana	silaa	eega
Nuti	tawullee	Isee	Keeti	otuu	utuu	otuma	ka
Yoo	akkasumas	ofii	Malee	erga	erga	waggaa	oggaa

2. Appendix-II: Afan Oromo suffixes

olee	olii	oolii	ota	oolee	oota	icha	ichi	oma
oma	fis	siis	ooma	siif	fam	ata	ittii	dha
ttii	irra	tii	rra	eenya	ina	offaa	annoo	umsa
ummaa	insa	am	ni	affaa	aa'	uu'	ee'	suu
dud	did	dand	wwan	Een	an	tet	tut	tit
teet	tuut	tanu	taanuut	tant	tanit	nu	na	nne
nnu	nna	dhaa	tiift	chaaf	dhaaf	ach	adh	chuu
at	ch	e	u	s	suu	Si	ssi	sse
ssa	nye	nya	lee					

3. Appendix-III: Afan Oromo synonyms

aaddachiisuu | haadhachiisuu	aaduu | haaduu	ajjaa | omborii	aankoo | jaldeessa	aantii | aanaa
aarii | haarii	aayyoo | aayyaa	abaabayyuu | habaabayyuu	ababoo | habaaboo	ilillii | habaaboo
dararaa | habaaboo	abadan | siruma	abashaa | habashaa	abbaagadaa | luba	abbala | hawwa
abboomama | adabamaa	abboomuu | adabuu	abishii | sunqoo	ablee | hablee	alalee | halalee
alamii | addunyaa	ankarsaa | dhulaandhula	anqaaquu | hanqaaquu	buphaa | hanqaaquu	killee | hanqaaquu
arcumee | harcumee	asimii | asmaa	kudhaama | asmaa	axawuu | haruu	qulqulleessuu | haruu
atamtama | harifannaa	sardama | harifannaa	muddama | hariifannaa	jarjara | hariifannaa	awaalama | hacuucama
cunqursaa | hacuucama	gidiraa | hacuucama	baaduu | areera	hareera | areera	baallama | beelama
baasaa | riqicha	baashee | beela	hoongee | beela	bantii | qarree	dubrummaa | qarree
bara | beela	barchaa | ganboo	bareeda | miidhaga	barraaqa | barii	beekuma | barumsa
beenyaa | gumaa	ciciwwii | cuucii	cilee | cilaattii	cimoo | cimaa	coxee | catee
cufantaa | cufaa	da'a | daha	daaktuu | daattuu	dabarsaa | dabaree	da'umasa | daha
digdama | diddama	dirredawaa | dirreedhawaa	eega | eegee	billaa | halbee	bisaan | bishaan
biyyoo | biyyee	bokkaa | rooba	boollo | boolla	bukkee | maddii	eebba | heebba
foonaa | mooraa	fooyuu | foowuu	gaadduu | keettoo	gaachana | gaalee	gaddii | milkii
geedala | sardiida	waangoo | sardiida	habbayyii | abbayyii	ja'a | jaha	kaawoo | surraa
keenya | keenna	kofla | kolfa	| milkii	geedala | sardiida	waangoo | sardiida
habbayyii | abbayyii	ja'a | jaha	kaawoo | surraa	keenya | keenna	kofla | kolfa
macuree | mar'imaan	harcummee | shaxxee	dhiluu | foowuu	nahuu | rifachuu	obboroo | subii
qoonqoo | beela	raajjuu | raagduu	reettii | re'ee	nasuu | rifachuu	siddisa | hamaaqixa
sooressa | dureessa	taa'aa | hudduu	xiqqoo | bicuu	yemmuu | yeroo	yeella'aa | qaanii
makoodii | handarii	gugee | handarii	jalqabuu | eegaluu	dhaanuu | reebuu	tumuu | reebuu
horii | loon	beeylada | loon	wayyaa | uffata	kafana | uffata	mi'a | meeshaa
miya | meeshaa	qodaa | meeshaa	baallii | angoo	tayitaa | angoo	muudama | aangoo
nafa | qaama	dhaqna | qaama	jismii | qaama	funyoo | haada	warra | maatii
lukkuu | handaaqqoo	waaqa | rabbi	marga | citaa	mayra | citaa	gadda | boo'a
taziyaa | boo'a	naasuu | boo'a	callaa | qofaa	kophaa | qofaa	dhibamuu | dhukkubsachuu
jijjiiruu | diddiruu	geeddaruu | diddiruu	herreguu | yaaduu	xiinxaluu | yaaduu	

4. Appendix –IV: Afan Oromo Abbreviations

k.k.f	Kan kana fakkaatan	Obb.	Obboo
Add.	Addee	Bil.	Biliyoona
fkn.	Fakkeenyaaf	hub.	Bubaachiisaa
w.k.f	Waan kana fakkaatan	mil.	Miliyoona
ful.	Fulbaana	Sad.	Sadaasa
Mr.	Mister(in some cases)	Ama.	Amajjii
Onk.	Onkololeessa	Bit.	Biteetossa
Mud.	Muddee	Wax.	Waxabajjii
Gur.	Guraandhala	Hag.	Hagayya
Ebl.	Ebla	W.B.	Waree booda
Ado.	Adoolleessa		
W.D.	Waaree dura		

5. Appendix-V: Manual summary creation guideline

The purpose of this guideline is to enable you (the human summarizers) to create an extract summary by ranking sentences. The summary will be used as a reference to evaluate our generic automatic news text summarizer.

Dear evaluator! You are expected to read the original news items carefully until you understand the concepts. Then you are going to rank the sentences according to their importance for generic summary.

You rank sentences based on the number of sentences available in the news item as such the most important sentence will be ranked with **N/ N** , where **N** is the number of sentences available in news item and **N-1/ N** for the second important sentence , **and N-2 / N** for third important sentence etc .

While ranking the sentences for extraction base your decision on the criteria such as Informativeness, coverage, non-redundancy, referential integrity, focus and coherence.

i. Informativeness :- the best sentences are that contain the most important information of the topic sentence

ii. Non-redundancy :- a summary should not contain unnecessary repetition of whole sentences or ideas/facts

iii. Referential integrity: - while reading the sentences according to their rank order it should be easy to identify who or what the pronouns and nouns phrases in each sentence are refereeing to.

iv. Coherence: - there should be a smooth transition of sentences. While reading the sentences in their rank order it should not just be a heap of related information, but also should build a coherent body of information about a topic/s.

6. Appendix-VI: Manual summary evaluation guideline

With this guide line you are going to evaluation the following three qualities of the system sample summaries to evaluate the performance to the system subjectively.

i. Informativeness (In which one of the summaries the most important information is being kept?)

It is measured in terms of the amount of information in the summary that actually helps to satisfy the information need expressed by the topic statement. Given the topic statement and a number of summaries that are supposed to contribute towards satisfying the information need expressed in the topic statement, some of the summaries may be more responsive to the topic than others. Therefore, your task is to help us understand how relatively well each summary respond to the topic.
Read the topic statement and all associated summaries generated with different methods. Then, mark the method / methods that create more informative summary.

ii. Linguistic quality (Out of a scale from 1-5, where 5 is the best, what score would you assign to each summary?)

The linguistic quality assessment is targeted to evaluate how readable and fluent the summaries are, and it measures quality of the summary. It includes: grammaticality, non-redundancy and referential clarity among others. By grammaticality we mean summary should have no capitalization errors or ungrammatical sentence (e.g., fragments, missing components etc). Non-redundancy is to mean there should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of noun or noun phrase (e.g., “Obama”) when a pronoun (“he”) would suffice. Referential clarity measures whether it is easy to identify who or what
The linguistic quality is assessed on a five-point scale from “1” to “5” , where “5” indicates that the summary is good , “1” indicates that the summary is bad , and “2” to “4” show the grades in between.

iii. Coherence and structure (Which summary is more coherent?)

The summary should be well-structured and well-organized. It should not just be a heap of related information, but should build from sentences to sentence to a coherent body of information about a topic. Coherence and structure can also be treated as one of linguistic quality. However, we treated separately as coherence needs more emphasis to produce meaningful summary.
You are required to read the original news item and system summaries and mark which of the method/methods produce coherent summaries

7. Appendix-VII: Subjective summary evaluation result

Table 1: In which one the most important information is being kept?

TID	E1	E2	E3	E4	Total Result
Test1	M3	M3	M1	M2, M3	M1(1),M2(1) , M3(3)
Test2	M3	M3	M1, M3	M2, M3	M1(1),M2(1) , M3(4)
Test3	M2	M3,	M2, M3	M1, M3	M1(1),M2(2), M3(3)
Test4	M2, M3	M2	M1, M3	M1	M1(2),M2(2), M3(2)
Test5	M1	M1, M3	M2, M3	M3	M1(2),M2(2) , M3(3)
Test6	M3	M3	M2, M3	M1	M1(1),M2(1), M3(3)
Test7	M3	M1	M2	M1, M3	M1(2),M2(1), M3(2)
Test8	M2	M3,	M1, M3	M2, M3	M1(1),M2(2), M3(3)

Table 2: Which summary is more coherent?

TID	**E1**	**E2**	**E3**	**E4**	**Result**
Test1	M3	M3	M2	M1, M3	M1(1),M2(1), M3(3)
Test2	M3	M1,M3	M3	M2, M3	M1(1),M2(1), M3(4)
Test3	M1	M3	M3	M2, M3	M1(1),M2(1), M3(3)
Test4	M2	M2	M3	M1, M3	M1(1),M2(2), M3(2)
Test5	M1	M1	M3	M3	M1(2),M2(0), M3(2)
Test6	M3	M3	M3	M2, M3	M1(0),M2(1), M3(4)
Test7	M3	M2	M1, M3	M3	M1(1),M2(1), M3(2)
Test8	M2	M3	M3	M2, M3	M1(0),M2(2), M3(3)

Table 3: Out of a scale from 1-5, where 5 is the best, what score would you assign to each summary?

TID	E1	E2	E3	E4	Total Result
Test1(M1)	3	3	2	3	11
Test1(M2)	3	4	2	4	13
Test1(M3)	4	3	3	5	15
Test2(M1)	2	2	3	2	9
Test2(M2)	3	3	2	3	11
Test2(M3)	5	5	4	4	18
Test3(M1)	1	4	2	4	11
Test3(M2)	2	4	2	3	11
Test3(M3)	3	4	4	5	16
Test4(M1)	2	4	4	4	14
Test4(M2)	2	3	4	3	12
Test4(M3)	3	5	5	2	15
Test5(M1)	3	4	2	3	12
Test5(M2)	3	4	3	3	12
Test5(M3)	4	4	3	2	12
Test6(M1)	2	3	2	4	11
Test6(M2)	3	4	2	2	11
Test6(M3)	1	4	5	2	12
Test7(M1)	4	3	4	3	14
Test7(M2)	2	4	4	4	14
Test7(M3)	4	4	4	4	16
Test8(M1)	5	3	3	2	13
Test8(M2)	2	4	3	3	12
Test8(M3)	4	5	4	4	17

8. Appendix-VIII: Sample news and system summary for better performing method (M3)

Original news text
TID: Test2
Source: ORTO

Abbaan Taayitaa daandiiwwan Itiyoophiyaa daandii aspaaltii km 247 Asallaa, Dodolaa fi Gobbaa walqunnamsisu birrii Bil. tokko fi mil.200'n hojjechiisaa ture tajaajila eegale .

Daandiin aspaaltii Asallaa, Dodolaa, fi Gobbaa walqunnamsiisu km 247 birrii Bil. tokkoo fi mil. 200'n waggoota 3'n darbaniif ijaaramaa ture xumuramee tajaajila kennuu eegaluusaa Abbaan Taayitaa daandiiwwanii Itiyoophiyaa beeksise.

Hojjatamuun daandii kana, rakkoolee hawaas- diinagdee ummataa Godinaalee Arsii, Arsii Lixaa fi Baalee furuu irraa darbee, qabeenya aadaa fi tuurizimii Godinaalee kanneenii daawwachiisuun galii biyyattiin damee kanarraa argattus guddisuuf gahee ol'aanaa akka qabus himameera. Godinaaleen Arsii, Arsii lixaa fi Baalee qabeenya uumama fi omishtummaan lafa isaaniitiin adda dureewwanii, keessattuu oomisha midhaaniitiin. Qabeenya uumamaa godinaaleen kunneen qaban dawwachuufis ta'e, oomishaa gabaatti dhiyeeffachuuf rakkoon daandii godinaalee kanneen giddu galeessa biyyattiin walqunnamsiisu gaaffii ummataa baroota dheeraat .

Rakkoolee kanneen furuudhaaf sagantaa misooma daandiitiif qabameen, abbaan taayitaa daandiiwwan Itiyoophiyaa daandii aspaaltii km 247 Asallaa, Dodolaa fi Gobbaa walqunnamsisu birrii Bil. tokko fi mil. 200'n hojjechiisaa tureera . Daandiin kuni yeroo qabameef keessatti akka hin xumuramneef rakkooleen muraasni mudatanis waayita ammaa xumuramee tajaajila kennuu eegaleera jedhan Daareektoreetii Komunikeeshinii Abbaa taayitaa Daandiiwwan Itiyoophiyaatti Dursaan garee dhimmoota komunikeeshinii obbo Darajjee Hayiluu himan.

Daandii kuni Kontiraaktaroota Chaayinaa lamaan bakka lamatti qoodamee hojjetamaa kan ture yoo ta'u Asallaa hanga Dodolaatti km 117 dhaabbata ijaarsaa Siinoo Haayidiroo jedhamutu birrii miliyoona 500 olii, Dodolaa hanga Gobbaatti kna jiru km 130 ammoo dhaabbata ijaarsaa"CGC over CC" jedhamutu birrii miliyoona 600 fi miliyoona 20'n hojjetan . Ijaarsa piroojektoota kanneenii dhaabbileen biyya keessaa fi alaa hojii to'annoo fi gorsaa irratti qooda fudhataniiru .

Ijaarsa piroojektoota kanneenii irratti rakkooleen gara garaa mudachuu isaaniitii yeroo jedhame keessatti xumuramuu baatus qulqullina isaa eegme xumuruun tajaajilaaf oolchineerra jedhan Injinara ol'aanaan piroojektichaa Mr. Moo-Guatan . Jaarsoliin biyya godinaalee kanneenii gama isaaniitiin, gaaffiin daandii ummataa godinaalee Baalee fi Arsii lamaanii waggoota dheeraaf ture xiyyeeffannoo argatee deebi'uu isaatiin gammanneerra jedhan . Keessattuu godinni Baalee qabeenya aadaa fi dhaabbilee Tuurizimii hedduu waan qabduuf daawwattoonni daandii kanatti fayyadamuun gara godinichaa dhufan galii biyyattiin damee tuurizimmii irraa argattu ni guddisa jedhan Bulchaan Godina Baalee obbo Sisaay Hurrisaa .

Sadarkaa naannoottis, sagantaa waliin gahaansa daandii baadiyyaatiin daandiin Aanaalee godinichaa walqunnamsisu km 627 barana hojjetamaa akka jirus himan . Aanaalee godina Arsii keessattis daandiin km 586 hojjetamaa jiraachuu Itti gaaffataman waajjira daandiiwwannii godinichaa obbo Dassaalany Dibaabaa himan . Daandii hawaasni hirmaannaa isaatiin hojjetuun cinaatti daandii godinaalee kanneen keessa qaxxamuruun giddu gala biyyattiin walqunnamsiisu kana hawaasnii kunuunsee itti fayyadamuu akka qabus dhaamameera .

Summary with 10% compression rate

Abbaan Taayitaa daandiiwwan Itiyoophiyaa daandii aspaaltii km 247 Asallaa, Dodolaa fi Gobbaa walqunnamsisu birrii Biliyoona tokko fi miliyoona 200'n hojjechiisaa ture tajaajila eegale.

Summary with 20% compression rate

Abbaan Taayitaa daandiiwwan Itiyoophiyaa daandii aspaaltii km 247 Asallaa, Dodolaa fi Gobbaa walqunnamsisu birrii Biliyoona tokko fi miliyoona 200'n hojjechiisaa ture tajaajila eegale.
Hojjatamuun daandii kanaa, rakkoolee hawaas-diinagdee ummataa Godinaalee Arsii, Arsii Lixaa fi Baalee furuu irraa darbee, qabeenya aadaa fi tuurizimii Godinaalee kanneenii daawwachiisuun galii biyyattiin damee kanarraa argattus guddisuuf gahee ol'aanaa akka qabus himameera.
Daandii kuni Kontiraaktaroota Chaayinaa lamaan bakka lamatti qoodamee hojjetamaa kan ture yoo ta'u Asallaa hanga Dodolaatti km 117 dhaabbata ijaarsaa Siinoo Haayidiroo jedhamutu birrii miliyoona 500 olii, Dodolaa hanga Gobbaatti kna jiru km 130 ammoo dhaabbata ijaarsaa&quot CGC over CC"

Summary with 30% compression rate

Abbaan Taayitaa daandiiwwan Itiyoophiyaa daandii aspaaltii km 247 Asallaa, Dodolaa fi Gobbaa walqunnamsisu birrii Biliyoona tokko fi miliyoona 200'n hojjechiisaa ture tajaajila eegale. Hojjatamuun daandii kana, rakkoolee hawaas-diinagdee ummataa Godinaalee Arsii, Arsii Lixaa fi Baalee furuu irraa darbee, qabeenya aadaa fi tuurizimii Godinaalee kanneenii daawwachiisuun galii biyyattiin damee kanarraa argattus guddisuuf gahee ol'aanaa akka qabus himameera. Daandii kuni Kontiraaktaroota Chaayinaa lamaan bakka lamatti qoodamee hojjetamaa kan ture yoo ta'u Asallaa hanga Dodolaatti km 117 dhaabbata ijaarsaa Siinoo Haayidiroo jedhamutu birrii miliyoona 500 olii, Dodolaa hanga Gobbaatti kna jiru km 130 ammoo dhaabbata ijaarsaa&quot CGC over CC" Keessattuu godinni Baalee qabeenya aadaa fi dhaabbilee Tuurizimii hedduu waan qabduuf daawwattoonni daandii kanatti fayyadamuun gara godinichaa dhufan galii biyyattiin damee tuurizimmii irraa argattu ni guddisa jedhan Bulchaan Godina Baalee obbo Sisaay Hurrisaa.

Summary with 40% compression rate

Abbaan Taayitaa daandiiwwan Itiyoophiyaa daandii aspaaltii km 247 Asallaa, Dodolaa fi Gobbaa walqunnamsisu birrii Biliyoona tokko fi miliyoona 200'n hojjechiisaa ture tajaajila eegale.
Hojjatamuun daandii kana, rakkoolee hawaas-diinagdee ummataa Godinaalee Arsii, Arsii Lixaa fi Baalee furuu irraa darbee, qabeenya aadaa fi tuurizimii Godinaalee kanneenii daawwachiisuun galii biyyattiin damee kanarraa argattus guddisuuf gahee ol'aanaa akka qabus himameera.
Daandiin kuni yeroo qabameef keessatti akka hin xumuramneef rakkooleen muraasni mudatanis waayita ammaa xumuramee tajaajila kennuu eegaleera jedhan Daareektoreetii Komunikeeshinii Abbaa taayitaa Daandiiwwan Itiyoophiyaatti Dursaan garee dhimmoota komunikeeshinii obbo Darajjee Hayiluu himan. Daandii kuni Kontiraaktaroota Chaayinaa lamaan bakka lamatti qoodamee hojjetamaa kan ture yoo ta'u Asallaa hanga

Dodolaatti km 117 dhaabbata ijaarsaa Siinoo Haayidiroo jedhamutu birrii miliyoona 500 olii, Dodolaa hanga Gobbaatti kna jiru km 130 ammoo dhaabbata ijaarsaa&quot CGC over CC"

Ijaarsa piroojektoota kanneenii irratti rakkooleen gara garaa mudachuu isaaniitii yeroo jedhame keessatti xumuramuu baatus qulqullina isaa eegme xumuruun tajaajilaaf oolchineerra jedhan Injinara ol'aanaan piroojektichaa Mr. Moo-Guatan.

Keessattuu godinni Baalee qabeenya aadaa fi dhaabbilee Tuurizimii hedduu waan qabduuf daawwattoonni daandii kanatti fayyadamuun gara godinichaa dhufan galii biyyattiin damee tuurizimmii irraa argattu ni guddisa jedhan Bulchaan Godina Baalee obbo Sisaay Hurrisaa.