ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

Query-based Automatic Summarizer for Afaan Oromo Text

By: Asefa Bayisa Kabeta

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABAUNIVERSITY IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTERS OF SCIENCE ININFORMATION SCIENCE

March, 2015

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

Query-based Automatic Summarizer for Afaan Oromo Text

By: Asefa Bayisa Kabeta

Approved by the Examining board:

Advisor: **Dr. Martha Yifiru**	_____	_____
	Signature	Date
Examiner:	_____	_____
	Signature	Date
Examiner:	_____	_____
	Signature	Date
Chairperson:	_____	_____
	Signature	Date

Acknowledgements

First and foremost, I praise the almighty God for his love and support to accomplish my thesis work. Next, I would like to thank my advisor, Dr. Martha Yifiru, for her unreserved support, guidance, understanding and motivation throughout the course of this thesis.

My thanks also go to Oromia Culture and Tourism Bureau, language Department employees, Ato Alemu Senbato, Ato Shiferaw Tadesse and Aster Tulu who assisted me with the preparation of manual summary. I also would like to extend my heartfelt gratitude to Ato Girma Mengistu, Ato Bikila Ashenafi, Ato Yihunbelay Teshome and Ato Girma Hundessa for their patience and willingness in evaluation of the system.

I am deeply indebted to Girma Dabele, kifle Derese and Gezehagn Gutema for their valuable advice and support during this thesis work. Moreover, I want to express my special appreciation for my friends Abe (filfilu), Abrish, Kasish, Biruke and Babi for their ideas, comments and encouragements.

Last but not least, I owe a great debt of gratitude to my beloved wife, Addis Daba, my daughter Urji Asefa and my brothers Sami, Dani and Tsinu for their prayer, encouragement and support. God bless you all!

Table of Contents

Contents **Page**

Abstract

Text summarization is the most challenging task in information retrieval. It is an outcome of electronic document explosion and can be seen as the condensation of the document collection. Automatic text summarization can be generic or query specific. In query-focused or query-oriented summarization a query is provided to a summarizer in addition to the source documents. The summarizer is supposed to construct a summary that contains information requested by the query. A document retrieval system together with a query-oriented summarization system is potentially a very powerful combination, which might be much more effective than a document retrieval system alone. Thus, this thesis focused on the possibility of developing query-based, single document, extractive summarization system.

In this thesis, two methods that create text summaries by extracting and ranking sentences from the original documents are proposed. The first method is based on the most commonly used IR model called vector space model (VSM) for finding the most important sentences related to the query given by the user. The second method is the position method which is used in attempting to improve the quality of the summary along with VSM. The sentence ranking algorithm performs based on the sentence score to rank sentences in the order of their importance and finally summary is produced by selecting the top N sentences, where the value of N is set by the user.

Experiments were conducted using40 Afaan Oromo news contained in the corpus. Three language experts of Oromia Culture and Tourism Bureau, language department were employed to conduct manual summarization which serves as the ideal summary. Intrinsic evaluation technique is used for evolution purpose. It involves both objective and subjective evaluation. The objective evaluation evaluate the performance of the system using standard Information Retrieval (IR) evaluation metrics (Precision, Recall and F-measure) and the subjective evaluation evaluate the linguistic quality such as informativeness and coherence using the scores on five scale measures by human evaluators. The results of the evaluations showed that the proposed system registered f-measure of 82%, 78% and 82% at summary extraction rate of 10%, 20%, and 30% respectively when VSM is used along with position method. Moreover, the informativeness and coherence of the proposed system also registered its best performance summary of 59%, 77% and 91% average score on five scale measures at extraction rate of 10%, 20%, and 30% respectively when both methods used together.

The challenging task in the study includes lack of query expansion tools which help to obtain more clues in finding important sentences in the final summary and some final summaries contain unresolved references that may cause difficulties in understanding. These will be the future research directions in this area which contribute in the improvement of the proposed system.

List of Tables

Page

List of Figures

Page

List of Algorithms

Page

List of Acronyms

AI	Artificial Intelligence
ATS	Automatic Text Summarization
CLASSY	Clustering Linguistics, And Statistics for Summarization Yield
DUC	Document Understanding Conference
HMM	Hiden Markov Model
ICT	Information and Communication Technology
IE	Information Extraction
IR	Information Retrieval
LSA	Latent Semantic Analysis
NLP	Natural Language Processing
ORTO	Oromia Radio and Television Orgnization
OTS	OpenSource Text Summarizer
QA	Question and Answering
ROUGE	Recall Oriented Understudy for Gisting Evaluation
RST	Rhetorical Structure Theory
TF	Term Frequency
TFIDF	Term Frequency Inverse Document Frequency
TFISF	Term Frequency Inverse Document Frequency
VSM	Vector Space Model

CHAPTER ONE

INTRODUCTION

1.1 Background

A huge amount of on-line information is available on the web, and is still growing. While search engines were developed to deal with this huge volume of documents, even they give large number of documents for a given user's query [1]. Under these circumstances it became very difficult for the user to find the document he actually needs, because most of the users are reluctant to make the cumbersome effort of going through each of these documents. Therefore systems that can automatically summarize one or more documents are becoming increasingly desirable to cope with the information overload.

Automatic text summarization is the task of generating a summary by a computer. It is the process of extracting key phrases/sentences from a given text and presenting them in a readable format to the user [2]. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning. Although automatic text summarization is a hot topic of research nowadays, only very few software tools are available to the end users and none of them are particularly popular because of the low quality of automatically produced summaries [3].

A summary can be defined as a text that is produced from one or more texts, that contain a significant portion of the information in the original text(s), and that is no longer than half of the original text(s) [2]. It can be an indicative as a pointer to some parts of the original document, or an informative which covers all relevant information of the text [4]. In both cases the most important advantage of using a summary is its reduced reading time. Summary can be generated for an individual (single) document or for a collection (multiple) of documents [5]. Single-document summarization is aimed at extracting and presenting the most salient information from a given text.

Multi-document summarization, on the other hand, deals with summarizing a collection of thematically related documents. Summarizing a single document is a challenging task but it is even more intricate with multi-document summarization [5]. Depending on what the

summarization program focuses to make the summary of the text, summaries can be generic or query-based. Generic summaries are those that give general overview of the text content. Query-based summaries, on the other hand, presents the contents of the document that are closely related to users query and it is often achieved by extending conventional IR[1] technologies[4]. On the basis how summary is generated, summaries can be broadly divided as abstractive and extractive [4]. Abstractive summary need not have exact sentences that present in the original document. In extractive summary, important sentences are extracted from the document based on different methods. Extractive summary contains all such extracted sentences arranged in a meaningful order. Most research on summary generation techniques including ours still relies on extraction of important sentences from the original document to form a summary.

In light of this, the summarization system that we aim to implement in this thesis is a single document query oriented extractive summarization system.

1.2. Statement of the Problem

The recent growth of ICT infrastructure in Ethiopia is resulting in an exponential increase of information being produced. Currently, there is a large amount of information produced both in digital and printed format available in local languages [6]. The continual increase of locally available data will increase the demand for tools that summarize relevant information. Afaan Oromo text readers are those who need such tools for many domain areas that produce large content of textual information.

Afaan Oromo is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of others neighbor countries like Kenya and Somalia [7]. Currently, it is an official language of Oromia state (which is the largest Regional State among the current Federal states in Ethiopia). It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts 34.5% of the local population [8].

With regard to the writing system, "**Qubee**" (a Latin-based alphabet) has been adopted and become the official script of Afaan Oromo since 1991. Many domain areas produce large content of textual information in Afaan Oromo both in printed and digital format especially since the

[1] *the activity of obtaining information resources relevant to an information need from a collection of information resources.*

language became the official language of the state [8]. Some of these domains include 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.

The news domain is one among those required an appropriate summarization tool to save the time of the reader. 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 [8]. There are a number of newspaper 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 Afaan 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.

A few researches in automatic text summarization have been commenced for Ethiopian languages. There are attempts to develop the generic text summarization system for Amharic language [30],[31],[32]and [33], the working language of the federal government, in different domains by adopting different techniques. Girma [8] tried to develop a single document generic text summarizer for Afaan Oromo news text which has been developed by customizing the (Open Text Summarizer) OTS[2]. Details found in sections 2.6.2.

Until recently, generic summaries were more popular, but with the prevalence of full-text searching and personalized information filtering, query based summaries are gaining importance [9].

Researches [1] [43] revealed that in the information retrieval scenario, if users could see the sentences in which their query words appeared, they could better judge the relevance of the documents, considered generating the query based summaries. Query-based summarization focuses on a user of information than the author of the document in that it generates summary

[2]*an open source tool for single English document summarization.*

based on the user's need. The user's information need is basically provided in the form of query. Query-based summarization has become especially important as the amount of information available online is growing exponentially [44]. Therefore, query based automatic text summarization could be justified as very essential tool for a very large volume of news release to help users grasp the main information related to a query. This formed the motivation for researchers to employ such tool to do the task of text summarization in news domain and as far as my knowledge is concerned, there is no attempt on query based automatic text summarization for Afaan Oromo and other local languages.

Hence the aim of this study is to explore the possibility of developing a query-based automatic text summarizer for Afaan Oromo news text using vector space model (VSM[3]), one of the oldest and most extensively studied models for IR; and sentence position method.

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

1. To what extent vector space model enables to design effective query based Afaan Oromo text summarizer?
2. Does the proposed system results in better performance when paired with the location method?
3. Does the proposed system generate a summary that is coherent and informative?

1.3. Objectives

The general objective of this research work is to explore the possibility of developing query-based, single document, extractive summarization system for Afaan Oromo news text.

In order to achieve the general objective, the following specific objectives are identified.

- Conduct literature review on relevant works in the area;
- Select and customize or develop a query-based summarization algorithm for use with the Afaan Oromo language;
- To develop a prototype summarizer as a framework that will serve as a model for query based Afaan Oromo news text;

[3] *an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings.*

- To prepare reference summaries and example queries used in the experiments in order to evaluate the proposed system;
- Evaluate the proposed system;
- Draw conclusions based on experimental result and forward recommendations.

1.4 Methodology

To achieve the objectives stated in section 1.3, a quantitative research methodology is employed as our work involves a lot of numerical data and experimentation. More over, the following methods and tools are also used in the implementation of the proposed system.

1.4.1 Literature Review

Extensive literature review is conducted to get a deeper understanding of automatic text summarization. In addition, as the study is conducted on Afaan Oromo, the nature of the language and the structure of the news documents have been investigated. To carry out this task, books, journal articles, and relevant websites are consulted.

1.4.2 Data Collection

The corpus used for evaluating the proposed system contains 40 Afaan Oromo news items whose lengths are in the range of 20 to 66 sentences. News items with less than 20 sentences were not used in the evaluation due to the fact that summarizing short articles does not make much sense in real applications [10]. Though evaluating the system with news stories containing more than 66 sentences is very desirable, it was very difficult to obtain such news stories in digital format and converting from hardcopy format takes considerable budget and time. These news items were collected from ***Kallacha Oromiyaa*** and ***Bariisaa*** newspapers.

1.4.2.1 Manual Summary Preparation

Three experts from Oromia Culture and Tourism Bureau, language department were engaged to conduct manual summarization on the 40 news contained in the corpus. For query based single document summarization the title of the document can be used as a query and simplify the process of evaluation [4][5][28].Thus, the title (query) and the document were given to the experts with the corresponding guideline modified from [9] and they selected the sentences

mostly related to the title (query) at 10%, 20% and 30% extraction rate. The guideline and lists of queries used are available in Appendix I and II, respectively.

1.4.3 Implementation

1.4.3.1 Methods

In this thesis we used two methods to extract the most important sentences from the document: The first one is one of the oldest and most extensively studied models of IR called vector space model (VSM). In VSM both the sentences in document and query are arranged in vectors [5].The angle between the document vector and the query vector is computed using cosine similarity measures. Hence, the sentences returned as an answer to a user query are those geometrically closest to the query according to the value obtained using the cosine similarity measure. Secondly, in attempt to take document genre information into consideration; the position of each sentence in the document is used as additional features to compute the significance of a sentence. Since news articles in general are written in such a way that the information presented is arranged in descending order of importance [15].

Thus, based on the value obtained from both methods (VSM and Position method) the most important sentence ranked first and other sorted depending on the score from very important to least important. Finally, we extracted top N sentences to form a summary. To improve the readability of the system extracts we re-arrange the selected sentences to their original position in the text.

1.4.3.2 Tools

Python4 2.7 is used to build the query-based automatic summarization for Afaan Oromo News since it is dynamic programming language that is used in a wide variety of application domains.

1.4.4 Evaluation

An intrinsic evaluation method which evaluate the performance of the proposed system objectively and the linguistic quality (informativness and coherence) subjectively has been employed. Performance evaluation was conducted by comparing the system summaries with

[4] *a widely used general-purpose, high-level programming language.*

their corresponding manual summaries using co-selection evaluation metrics. These are Precision (P), Recall (R), and F-Score (F). For each document, the manually extracted sentences are considered as the manual summary and the summary produced by the automatic summarization system is system summary. The formulas for the measures P, R and F score are as follows [2].

Precision = (reference summary) ∩ (system summary)/ system summary

Recall = (reference summary) ∩ (system summary)/ Reference summary

F-score = (2*P*R) / (P+R)

In addition, the informativeness and coherence of the system is also evaluated. For these purpose five graduate students from Addis Ababa University, Afaan Oromo department was selected using purposive sampling technique because we believe that they know the language very well and able to evaluate the summary produced by the system. We could not get more than five evaluators as the evaluation process taking time.

The original document and the system's summary were given to the evaluators and they were asked to read both documents (the original document and its summary) then they score based on five scale measures adapted from [11] as shown in the table 5.2. Finally the interpretation is done based on the scores given by each evaluator.

1.5 Scope and Limitation of the Study

This thesis focuses on developing query-based extractive summarizer for Afaan Oromo news texts. It doesn't employ an abstractive summarization since it requires deep linguistic analysis and difficult to implement with current state of the art of the field. Moreover, the developed summarizer uses only a single document.

1.6 Significance of the Study

The researcher believes that the outcome of the research is crucial in many aspects.

- It contributes towards the possibility of developing query-based Afaan Oromo news text summarizer.
- It has the importance to initiate further research in the area of query based summarization for Afan Oromo language.
- It contributes in improving the development of the Afaan Oromo language.

- It helps the researcher to investigate problems and solving it in scientific manner.

1.7 Organization of the Thesis

This thesis report is organized into six chapters. The first chapter explains about the motivation behind conducting this research and discusses: background of the study, statement of the problem, the objectives, methodology and scope and limitations of the study. Chapter two presents a comprehensive literature review on automatic text summarization regarding: background of automatic text summarization including definition, types, Query-Based Automatic Summarization and related concepts and state of the arts; theoretical background of VSM, cosine similarity measure, position method and ranking algorithms. The evaluation techniques of text summarization and critical review of related and/or previous works are also presented under this chapter. Chapter three introduces the linguistic aspects of Afaan Oromo including the writing systems, language specific markers, and morphological analysis. Chapter four presents detailed description of the design and implementation of the proposed system. Chapter five demonstrates the experiments and performance evaluation for the proposed system. Findings of our experiments and challenges that affect the performance are also presented under this chapter. Finally, Chapter six concludes our work and outlines the directions for future research.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter, we provided a brief overview of the field of automatic summarization. There are several types of summarization and some of them are explained in this chapter. This chapter also describes the most important stages of automatic text summarization. Furthermore, we investigate state of the art techniques used in the area of text summarization. Automatic text summarization is an active field of research and hence, there are a lot of works in the area. Here, we presented some of those works whose contribution made a great progress to the field.

We believe that background information on the theory of vector space mode is necessary as our proposed query based summarization system is based on it. Thus, this chapter explains the vector space model ranking algorithms from information retrieval and summarization perspectives. Evaluating the performance of a summarization system is a very challenging task and over the years various methods have been proposed. Principal methods of evaluating summaries are also reviewed in this chapter.

2.2 Basic Notion of Automatic Text Summarization

Luhn [12] who is a pioneer researcher in the area defined text summarization as the process of distilling the most important information from a source text to produce an abridged version for a particular user/users and task/tasks. It is also defined as a process by which the most important concepts in a document are identified and then presented in a condensed and human-readable form [3]. Its main objective is to reduce the complexity and length of the original text while retaining the main ideas of the text. The produced summary should be non-repetitive and should give as much precise information as possible. In short, the summary should allow the reader to answer questions about the main topics in the given text or work as a reference pointer to parts of the original text.

Text summarization is in general a hard task, because we have to characterize the source text as a whole, we have to capture its important content, where content is a matter of both information and its expression, and importance is a matter of what is essential as well as what is salient. Summarization explores different levels of analysis of the text to help determine what information in the text is salient for a given summarization task [1]. This determination of the salience of information in the source depends on a number of interacting factors, including the nature and genre of the source text, the desired compression, and the application's information needs [9]. The information must be able to address a user's needs effectively, and an abridgment of a text according to a user's interests and expertise must be considered, i.e., user-focused summary versus generic summary which will be discussed in the subsequent section.

Though automatic text summarization has been around for more than 40 years now, it has become an important and timely tool for assisting and interpreting text information in today's fast-growing information age [3]. It is very difficult for human beings to manually summarize large documents of text.

There is an abundance of text material available on the internet and usually the Internet provides more information than is needed. Therefore, a twofold problem is encountered. These are searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning.

The application areas of automatic text summarization are extensive. In general, according to [1] the functions of a summary include:

- Announcement: announce the existence of the original document
- Screening: determine the relat0iveness of the original document
- Substitution: replace the original document
- Retrospection: point to the original document

Depending on the length and requirement of the summary some of these can be included while discarding others. Text summarization is used to generate an abstract of a scientific article, a summary of email threads, a headline for a news article, or generate the short snippets that search engines like Google return to the user to describe each retrieved document [10].

2.2.1 Types of Summary

There is no a unit classification for summarized texts and summaries could be categorized based on different criteria. Different papers discussed some of these criteria. However, in this research we tried to cover the most comprehensive criteria for classifying summaries as follows.

Summary construction methods: a summary can be an Abstractive or Extractive. An Abstractive summarization attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language [13].

It 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. Extractive summaries are formulated by extracting key text segments (sentences or passages) from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency, location or cue words to locate the sentences to be extracted[13].The “most important” content is treated as the “most frequent” or the “most favorably positioned” content. Such an approach thus avoids any efforts on deep text understanding. They are conceptually simple and easy to implement.

Number of Sources for the summary: a summary can be classified as single document summaries and multi-document summaries based on the number of sources. The difference between the two is mainly in the input document. In case of single document summarization, a single document is condensed to form an abridged version of it. On the other hand, multi-document summarization deals with summarizing a collection of thematically related documents [14]. Summarizing a single document is a challenging task but this is even more with multi-document summarization. This is attributed to the fact that in case of summarizing multiple documents, repetitions and inconsistencies between documents have to be well accounted for [15]. Due to this reason, multi-document summarization is much less developed than single document.

Purpose: An indicative summary gives the user an overview of the content of a document or document collection, whereas an informative summary's purpose is to contain the most relevant information, which would allow the user to extract key information. An informative summary's purpose would be to act as a replacement for the original text [13].

Genre: The information contained in the genres of documents can provide linguistic and structural information useful for summary creation [16]. Different genres include news documents, opinion pieces, letters and memos, email, scientific documents, books, web pages, legal judgments and speech transcripts (including monologues and dialogues).

Summary trigger: a summary can be Generic or Query based. Automatic text summarization systems often produce generic summaries that highlight the most salient points of a given text. However in the online search and retrieval context, a summarization system has access to the query entered by the user and should adapt its output to suit the user's information need. A query-oriented summary presents the information that is most relevant to the given queries, while a generic summary gives an overall sense of the document's content [13].

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 system proposed in this thesis can be called an extractive single document informative query-based summarization in news domain.

2.2.2 Main Steps in Text Summarization

According to [11] and [12] in order to better understand the operation of summarization systems and to emphasize the design choices system developers need to make, we distinguish three relatively independent tasks performed by virtually all summarizers: creating an intermediate representation of the input which captures only the key aspects of the text, scoring sentences based on that representation and selecting a summary consisting of several sentences. These steps are more elaborated in the following subsequent paragraphs.

In the first step even the simplest systems derive some intermediate representation of the text they have to summarize and identify important content based on this representation.

There are various approaches used in this step such as topic representation, lexical chain, and latent semantic analysis. Topic representation approaches convert the text to an intermediate representation interpreted as the topic(s) discussed in the text. Some of the most popular summarization methods rely on topic representations and this class of approaches exhibits an impressive variation in sophistication and representation power. They include frequency, TF.IDF and topic word approaches in which the topic representation consists of a simple table of words and their corresponding weights, with more highly weighted words being more indicative of the topic; lexical chain approaches in which a thesaurus such as WordNet[5] is used to find topics or concepts of semantically related words and then give weight to the concepts; latent semantic analysis in which patterns of word co-occurrence are identified and roughly construed as topics, as well as weights for each pattern; full blown Bayesian topic models in which the input is represented as a mixture of topics and each topic is given as a table of word probabilities (weights) for that topic. Indicator representation approaches represent each sentence in the input as a list of indicators of importance such as sentence length, location in the document, presence of certain phrases, etc. In graph models, such as LexRank[6], the entire document is represented as a network of inter-related sentences.

Once an intermediate representation has been derived, the next step is sentence scoring in which each sentence is assigned a score which indicates its importance.

For topic representation approaches, the score is commonly related to how well a sentence expresses some of the most important topics in the document or to what extent it combines information about different topics. For the majority of indicator representation methods, the weight of each sentence is determined by combining the evidence from the different indicators, most commonly by using machine learning techniques to discover indicator weights. In LexRank, the weight of each sentence is derived by applying stochastic techniques to the graph representation of the text.

[5]*is a lexical database for the English language. It groups English words into sets of synonyms called* synsets*, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members.*

[6]*is the method for document ranking and text summarization developed at U. Michigan by Gunes Erkan and Dragomir Radev.*

Finally, the summarizer has to select the best combination of important sentences to form a paragraph length summary. In the best n approaches, the top n most important sentences which combined have the desired summary length are selected to form the summary. In maximal marginal relevance approaches, sentences are selected in an iterative greedy procedure.

At each step of the procedure the sentence importance score is recomputed as a linear combination between the original importance weight of the sentence and its similarity with already chosen sentences. Sentences that are similar to already chosen sentences are preferred. In global selection approaches, the optimal collection of sentences is selected subject to constraints that try to maximize overall importance, minimize redundancy, and, for some approaches, maximize coherence.

2.3 Query-Based Automatic Summarization and related concepts

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. Some these tools include information retrieval, information extraction, question answering etc. Query-based 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. Here, we explained the basic relationship between query-based automatic summarization and other tools mentioned above.

The query-based summarization task is to generate a summary of single/multiple documents which is focused towards the user's query [16]. A query score is calculated for each sentence based on the distribution of query terms and added to its overall score obtained by sentence extraction methods. The top scoring sentences were used as a summary for each of the retrieved document. In the Information Retrieval scenario, if users could see the sentences in which their query words appeared, they could better judge the relevance of the documents, considered generating the query based summaries.

The goal of Information Retrieval (IR) systems is to extract documents that best match a query [17]. The two main tasks in IR are document indexing and searching. Indexing is the task of representing a document by its key features for the purpose of speeding up its finding when a query is invoked. There were many indexing schemes explored, but the most commonly used are based on word stemming and its enhancements. Term weighting is an indexing technique that gives a degree of importance to a word in a description.

Word proximity, especially adjacency, is frequently used to capture some of the linguistic relations between words. The goal of searching is to locate the most relevant documents and to rank them in Order of decreasing match with a query. Thus, a similarity measure needs to be introduced for this. In query-based summarization some of the IR techniques are used with some modification in extracting the most relevant sentence that match with the user's query [1]. This further discussed in section 2.4.

Information Extraction (IE) systems attempt to pull out information from documents by filling out predefined templates [17]. The information typically consists of entities and relations between entities like who did what to whom, where, when and how.

The task of Question Answering (Q&A) is to return an actual answer, rather than a set of ranked sentences, in response to a question [1]. A particularly useful complementarily exists between summarization and question answering systems. From the viewpoint of summarization, question answering is one way to provide focus for summarization. From the viewpoint of question answering, summarization is a way of extracting and fusing just the relevant information from a heap of text as an answer to a specific type of non-factoid questions. However, both question answering and summarization include aspects that are unrelated to the other. Sometimes, an answer to a question cannot be summarized; either it is a brief factoid (capital of India is Delhi) or the answer is complete in itself (give me the text of pledge of India) [17]. Likewise generic summaries (author's point of view) do not involve a question, they reflect the text as it stands without any user's input.

2.4 Text Summarization Approaches

Since the introduction of the field of automatic text summarization by Luhn [13], various approaches of summarization have been proposed and most are based on sentence extraction or selection. Different ways of classifying these approaches can be found in the literature and here, we will make use of the classification method presented in [8] to offer a brief overview of the traditional techniques used in automatic text summarization. This classification is based on the level of processing required to build a summary. Accordingly, three categories are identified as surface, entity, and discourse levels.

2.4.1 Surface Level Approaches

The earliest works in automatic summarization used surface level approaches to decide which parts of a text are important. The oldest known automatic summarization is that of Luhn [13], wherein term frequencies were used to measure sentence relevance. The basic assumption here is that the most frequent words in a text are the most representative of its content, and hence fragments of text containing them are more relevant.

However, not all words in a text are important and to this end, words beyond high and low frequency as well as those words contained in a stop word list are left out of consideration. The remaining words in the document are then sorted alphabetically so that pairs of succeeding words can be compared letter by letter. This allows for similar words to be found (e.g., differ, difference, different). These similar words thereby attained are called significant words. Then, sentences with the highest significance factor are extracted to produce a summary or what Luhn calls as “auto-abstracts”. The results obtained by Luhn were neither good condensations nor very coherent texts though he believed his “auto-abstracts” were satisfactory indicative abstracts for papers within the science and technology fields.

The work of Luhn was further extended to include three new parameters for calculating the weights of sentences. These were sentence position in a text, cue phrases, and title and heading words [8].

The position of a sentence in a text, in general, is believed to reflect the importance of the sentence. For example, newspaper articles have the most important sentences at the beginning of the article while technical documents have the most important sentences in the conclusion

section. In fact, a very simple and surprisingly successful method for summarization is the selection of the first sentences in a text. Various researchers have reported that this simple method of taking the lead (first sentence in the first paragraph) as summary often outperforms other methods, especially with newspaper articles [15].

Cue phrases are words or phrases that signal whether a sentence is important or not. In general there are three categories of cue phrases: bonus, stigma, and null phrases. Bonus phrases are used to emphasize the importance of a sentence in a text while stigma phrases reflect that the sentence is not important. Null phrases are neutral phrases and are not considered when the weight of a sentence is computed. Few examples of bonus phrases are "significantly", "in conclusion", "in this paper we show", etc. whereas "hardly" and "impossible" are examples of stigma phrases. Thus, each cue phrase is assigned a positive or negative relevance. The weight of each sentence is then the sum of the weights of the words in it [2, 7]. Sentences can also be scored for containing words that appear in the text's title or headings, or in the user's query, for a query based summary. The basic idea behind this assumption is that authors usually use informative titles which could reveal the subject matter of the document.

Using the combination of the features cue-words, title words, and the position of a sentence to generate summaries were shown to produce successful results in various studies [7, 12].

Surface level approaches can also be tailored to a specific domain or corpus to give corpus based approach for text summarization [15]. Corpus based approaches try to determine the relevance of words by first assigning the document to a particular domain. This helps to determine certain common terms in a given field that do not carry salient information and hence, their relevance can be reduced. It was further proved that the relevance of a term in a document is inversely proportional to the number of documents in the corpus containing the term [14, 15]. The formula to compute term relevance is given by:

$$tf_i \; x \; idf_i \qquad \textbf{(2.1)}$$

Where tf_i is the frequency of term i in the document and
idf_i is the inverted frequency of documents containing this term.

Sentences can then be scored by the sum of term relevance in the sentence [15]. Term relevance can also be measured by counting concepts rather than mere term counting.

By making use of an electronic thesaurus or WordNet, each word in the text is associated to a more general concept, and frequency is computed on concepts instead of particular words. For instance, the occurrence of the concept “bicycle” is counted when any of the words “bicycle”, “bike”, “pedal”, or “brake” is found [13, 14]. Furthermore, surface level approaches in combination with machine learning algorithms have resulted in more advanced summarization systems. Particularly, such systems use a Bayesian classifier algorithm to compute the probability of a sentence in a source document being included in a summary. The classifier is first trained using a corpus of several pairs of full documents/summaries. These summaries are abstracts created by professional abstractors. The Bayesian formula uses different statistical features to compute the probability of a sentence being relevant. Statistical features that the Bayesian formula uses are usually sentence length, cue phrases, position of a sentence in a paragraph, most frequent words (thematic words) and proper names [15].

2.4.2 Entity Level Approaches

Entity level approaches model text entities and their relationships to capture patterns of connectivity in a text which may be used to determine salient information [12]. In general, words can be connected in various ways, including repetition, co-reference, synonymy, and semantic association as expressed in thesauri. The degree of connectedness of words can then be used to score sentences and paragraphs. In essence, more connected sentences are assumed to be more important. With this approach, it has been shown that the main drawback of surface level approaches, which is lack of coherence and cohesion, can be resolved [7]. Various automatic summarization methods employ text connectivity to summarize a document and of these, lexical chain which is first introduced by Barzilay and Elhadad [18], is to be mentioned. Summarization based on lexical chains first selects a set of candidate words and then, for each candidate word an appropriate chain is computed. Cohesive relation (i.e., repetition, synonymy, antonymy, hypernymy, and holonymy) between terms is a criterion for chain formation. A lexical chain, therefore, is a chain of words in a text such that each word in the chain bears some kind of cohesive relationship to a word that is already in the chain [18]. Once the source text is represented using lexical chains, the strength of each lexical chain is determined on the basis of the number and type of relations in the chain.

A summary is then built by selecting sentences where the strongest chains are highly concentrated. This is in an assumption that picking sentences represented by strong lexical chains gives a better indication of the central topic of a text than simply picking the most frequent words in the text. Another way of using text connectivity for summarization is based on phrasal analysis and anaphoric relations in text. Here, the main aim is to identify those phrasal units across the entire span of the document that best function as representative highlights of the document's content. One way to achieve this is by using co-reference resolution system. [7]Co-reference resolution is the process of determining if two expressions in natural language refer to the same entity [17]. Once the desired phrasal units are identified, they are combined to form "capsule overviews". The capsule overview is not a sequence of sentences as is expected in a summary but it is a list of key-phrases preserving the flow narration in the original document.

2.4.3 Discourse Level Approaches

Discourse level approaches exploit the discursive organization of a text to improve the relevance and quality of summaries. The discursive organization of a text implies the global structure of the text and it includes format of a document, threads of topics in the text and rhetorical structure of the text [19]. This approach asserts that texts are not just a linear sequence of clauses and sentences but they are a cluster of clauses and sentences, called discourse segments that are related pragmatically to form a hierarchical structure. Thus, in this approach, a text is first divided into discourse segments and based on these segments, the discourse structure of the text as intended by its author is re-constructed. This discourse structure or discursive representation of the text has been shown to be one way of determining the most important units of a text [19]. There have been various text summarization works that exploit the discursive representation of text. The most popular theory of text organization used for summarization has been the Rhetorical Structure Theory (RST)[8] [18]. According to Mann and Thompson [18], one can associate a rhetorical structure tree to any text. RST is a binary tree representing rhetorical relations between text units.

---

[7]occurs when two or more expressions in a text refer to the same person or thing; they have the same referent, e.g. $Bill_i$ said he_i would come; the proper nun Bill and the pronoun he refer to the same person, namely to Bill.
[8]addresses text organization by means of relations that hold between parts of text. It explains coherence by postulating a hierarchical, connected structure of texts.

The relations tie together two non-overlapping pieces of text spans: the nucleus which is central to the writer's goal, and a satellite which is less central or a marginal part. Rhetorical relations reflect semantic, intentional, and textual relations that hold between text spans. These text spans could be clauses or sentences extracted from the original text. Text spans could be related in RST in such a way that one text span may elaborate on another text span or one text span may provide background information for another text span [19]. In one application of discourse structure for text summarization importance score is associated to each clause in a text; the closer a clause is to the root of the tree, the higher is the score. Then clauses of highest score are extracted to produce a summary [19].

2.5 Text Summarization and Information Retrieval Model

Text summarization, which is one of the application areas of natural language processing, is intimately connected with information retrieval [20]. In this Section, the basic concept of information retrieval model called Vector Space Model (VSM) is discussed considering its relevance to this thesis.

2.5.1 The Vector Space Model

The Vector Space Model (VSM) is one of the oldest and most extensively studied models for text mining [21]. In this model, a set of documents are conceptualized as a two-dimensional co-occurrence matrix, where the columns represent the documents and the rows represent the unique terms (usually words or short phrases) occurring in the documents. For this study we modified the classical VSM using a sentence instead of document and a document in place of corpus or document collection as our summarization system is based on single document. Each dimension of our vector corresponds to a sentence that is present in the document or query. The value in each cell of the matrix reflects the importance of the term in representing the semantics of the sentence in the document or query.

2.5.2 Term weighting

The idea behind term weighting is to assign a weight to represent the importance of a term. The raw frequency of a term only states how often a term occurs in a document without measuring the importance of that term within the sentence or within the document. In literature different weighting schemes are available.

The most common and popular one is the TF-IDF (Term Frequency-Inverse Document Frequency) weighting scheme, which combines local and global weighting of a term [15]. The weight of term i in sentence j is defined as a local weight (Lij) where $L_{ij} = tf_{ij}$ and tf_{ij} is term frequency or the number of times term i occurs in sentence j. However, this way of representing terms has some serious drawbacks [15]. One limitation is that long sentences are favored for inclusion in a summary simply because they tend to have more words, not because they are relevant.

Furthermore, the relative global frequency of terms across the entire document is ignored and thus, it is not possible to determine the importance of each term in describing the topics of the document.

Thus, it is necessary to represent terms according to their distribution in the entire document. Therefore the global weight which reflects the term's importance in the document is used. More specifically, the weight of term j in sentence i, A_{ij}, is calculated as follows [15]:

$$A_{ij}=L_{(tij)} \cdot G_{(tij)} \qquad \textbf{(2.2)}$$

Where, A_{ij} denotes the frequency with which term j occurs in sentence i,

$L_{(tij)}$ is the local weight for term j in sentence i, and

$G_{(tij)}$ is the global weight for term j in the whole document.

The local weight is the number of times a term occurred in a sentence is called as the 'Term frequency' which is represented as ***"tf"*** and the global weight scheme is ***tdf*** (term document frequency) which is the logarithm of inverse of the document frequency.

In this thesis, we used this weighting scheme with some adoption called TF-ISF (Term Frequency-Inverse sentence Frequency) that is computed as:

$$W_{ij} = tf_{ij} * isf_i \equiv \frac{freq_{ij}}{max} * log\left(\frac{n}{n_i}\right) \qquad \textbf{(2.3)}$$

Where denotes the normalized weight of the term i in document j, is said to be 'term frequency' of i_{th} index term in the jth sentence, and is 'inverse sentence frequency' of ith index term, where **N** is the number of all sentences and is the number of sentences in which the term occurs.

In general, the term weight for the document and the query would represent by term vector of the form: D = (Wt1, Wt2,… Wtk) Where each Wt1, Wt2…..Wtk identifies a content term assigned to some sample document D. Analogously, the information requests, or queries, would be represented in vector form. Thus, a typical query Q might be formulated as: Q=(Wq1, Wq2…..Wqk) Where Wq1, Wq2 ….. Wqk represents weights of content term in query.

2.5.3 Cosine similarity measures

There is many different ways to measure how similar two documents are to each other, or how similar a document is to a query. The cosine measure is a very common similarity measure. Using a similarity measure, a set of documents can be compared to a query and the most similar document returned [3].

If the query term does not occur in the sentence, similarity measure score should be 0, else the more frequent the query term in the sentence, the higher the score (should be). Query and sentence similarity is based on length and direction of their vectors.

We calculate Wti, Wqi which are indexing terms given to a sentence in the document and the query respectively. The similarity is calculated by widely used cosine measure.

$$S_{(Q,D)} = \frac{\sum i^{t} = 1 \;\; wqi * wdi}{\sqrt{\sum_{t=}^{i} 1 \; (wqi)^2} * \sqrt{\sum_{t}^{i} = 1 \; (wdi)^2}} \qquad \textbf{(2.4)}$$

2.6 Evaluation Methods

Evaluating the quality of a summary has proven to be a difficult problem, principally because there is no obvious "ideal" summary [1].

Even for relatively straightforward news articles, human summarizers tend to agree only (approximately) 60% of the time, moreover humans agreed with their own judgments in only 82% of the cases [22]. And there is always a possibility of system generating a good summary that is quite different from any human summary used as approximation to the correct output

summary. The output summaries are generated under some constraints such as compression ratio and user's application needs; this increases the scale and complexity of the evaluation.

According to [22] methods for evaluating text summarization can be broadly classified into two categories. The first is intrinsic evaluation, tests the summarization system itself. The second method is an extrinsic evaluation which tests the summarization system based on how it affects the completion of some other task. In this thesis we used an intrinsic evaluation method.

Intrinsic evaluation metrics test the summarization task itself, in terms of the coherence and content of the generated summaries. Two broad classes of metrics have been developed: form metrics or coherence and content metrics or informativeness. Form metrics focus on, grammaticality, overall text coherence, and organization. These metrics are assessed by humans with grading of the summaries on the basis dangling anaphora, gaps in rhetorical structure etc. [22]. Content or informativeness of the summary is more difficult to measure. There are two types of content evaluation measures: co-selection and content-based [23]. Since Co-selection is used in this thesis it is discussed in detail as follows. Co-selection evaluation metrics are suitable for summaries produced by sentence selection from the document to be summarized and where the ideal summaries are made from sentence extracts from the original document. These include precision, recall and F-measure. Recall (R) is the fraction of sentences chosen by the person that were also correctly identified by the system.

$$R = (\text{reference summary}) \cap (\text{system summary}) / \text{Reference summary} \quad (2.5)$$

Precision (P) is the ratio of the number of sentences occurring in both system and ideal summaries to the number of sentences in the system summary.

$$P = (\text{reference summary}) \cap (\text{system summary}) / \text{system summary} \quad (2.6)$$

F-Score (F) is a composite measure that combines precision and recall, and computed as:

$$F\text{-score} = (2*P*R)/(P+R) \quad (2.7)$$

Co-selection evaluation metrics by far are the simplest. However, despite their simplicity they are easily susceptible to human variation. Human summarizers do not often agree on what constitutes an ideal summary [23]. As a result a good quality system summary may get low evaluation score.

2.7 Reviews of Related Works on Automatic Text Summarization

In this section some of the international and local works on automatic text summarization is reviewed by focusing on query-based single/multiple document summarization. We will also review and identify gaps that exist in previous summarization works for local language.

2.7.1 Global works

The history of automatic text summarization began 50 years ago when the first extractive system developed by Luhn [13]. Luhn's method uses term frequencies to appraise eligibility of sentences for the summary.

Its main idea is based neither on knowledge that significant words carrying most information are not too frequent nor too seldom in the text. Establishing boundaries of words significance by the help of their frequency would be a matter of experience. The next step is ranking of sentences, reflecting the number of significant words and their distance in a sentence. After it remains only to choose one or several highly ranked as a result. The idea of Luhn was acknowledged and used in many automatic information processing systems. The system developed takes single document as input.

The next remarkable progress was done ten years later by Edmundson [18]. His work introduced hypothesis concerning high information value of title phrases, sentences from the beginning and from the conclusion of the article, sentences containing cue words and phrases as "important", "results are", "paper introduces", etc. 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.

The renaissance of this field and remarkable progress came in 90's due to broader use of artificial intelligence methods in this area and combination of various methods in hybrid systems [24]. Since then many systems have been developed in the area of automatic text summarization both on single and multi-documents. Most of the automatic summarizer developed during this time focused on generic summarization system. As mentioned in the work of [1] While there have been attempts to build statistical models for generic summaries and learning their parameters automatically from the data, there are less number of attempts towards generating

query based summaries. Partly this is because of lack of training data needed to learn the parameters.

Berger and Mittal [25] discusses a statistical modeling approach for query based summarization and learns the parameter values from the frequently asked questions from UseNet archives and question and answer pairs from call centers.

CLASSY [26] query-based multi document summarization based on an HMM [9](Hidden Markov Model) for sentence selection within a document and a pivoted QA algorithm to generate a multi document summary in news domain. The system, called "CLASSY" (Clustering, Linguistics, And Statistics for Summarization Yield), preprocesses each document, applying word- and phrase-elimination techniques. It classifies sentences into two sets: those ones belonging to the summary and those ones which not. The overall results indicate the method scored within the top group of systems for both ROUGE [10]and pyramid evaluation.

There were 10 human summarizers (labeled A-J), one baseline (labeled 1), and 31 machine systems. CLASSY ranked 10th out of the 31 machine systems. The performance of CLASSY, and all the other systems listed, was approximately midway between the baseline and human performance. In addition the system performance was measured by the modified pyramid score. The scores are sorted by medians. CLASSY ranked third among the peer systems.

Humans A and B, who were evaluated on 25 and 24 summaries, respectively, in lieu of the 26 for the machine systems (25 peers + baseline), had median pyramid scores of 0.46 and 0.44, respectively, which is far above the best peer performance of 0.20.

QASUM-TALP [27] A query-driven multi document summarizer. This system generates a set of new queries from the complex one given by the user. Then a question answering system is used to detect relevant information. After that the semantic representation of all relevant sentence are contrasted in order to avoid redundancy when selecting the summary content. The system is presented and evaluated in DUC 2005.

The evaluation metrics is based on Pyramid information manually created for DUC 2005. The system shows a very good correlation to manual Pyramid scores.

9is a statisticalMarkov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states.

10is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing.

Jagadeesh et.al. [1] developed Query-based multi-document summarization using language modeling. In this work, Relevance Based Language Modeling along with the representation of words in higher dimensional semantic spaces is used to calculate the explicit relevance score of sentences towards the information need.
With the theoretical support from document risk minimization framework they have defined an entropy based measure called Information Measure to capture the notion of importance or prior of a sentence.

External information sources like WWW and Wikipedia2 were explored for the necessary information to compute Information Measure. Finally these two scores are combined using a weighted linear combination technique, which followed from the earlier proof. Experiments conducted on DUC3 2005 and 2006 data sets show that the final system has generated the best answer summaries in comparison with the rest of the DUC participant systems.

Ramakrishna et. al. [28] presented a method to create query-specific summaries by identifying the most query-relevant fragments and combining them using the semantic associations within the document. In particular, they first add structure to the documents in the preprocessing stage and convert them to document graphs. Then, the best summaries are computed by calculating the top spanning trees on the document graphs. They present and experimentally evaluate efficient algorithms that support computing summaries in interactive time. Furthermore, the quality of their summarization method is compared to current approaches using a user survey. Accordingly, the experimental result show that their approach performs better than other state of the art approaches.

Ibrahim et.al.[29] introduced a query based, Arabic text, single document summarization using an existing Arabic language thesaurus and an extracted knowledge base. They used an Arabic corpus to extract domain knowledge represented by topic related concepts/ keywords and the lexical relations among them. The user's query is expanded once by using the Arabic WordNet thesaurus and then by adding the domain specific knowledge base to the expansion. For the summarization dataset, Essex Arabic Summaries Corpus was used. It has many topic based articles with multiple human summaries.

The performance appeared to be enhanced when using their extracted knowledge base than to just use the WordNet.

2.7.2 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). Even though all of the local works are focused on generic automatic text summarization, they are reviewed in terms of problem addressed, techniques (methods) used, finding of the study and performance of result achieved.

The first Amharic summarization system was developed by Kamil [30], a single document summarizer. He used surface level statistical features to assign weights to sentences.The highest scoring sentences were extracted to form the summary. Seven features are used: title words, cue phrases, first sentence of the document (header), words in a header, first sentence of a paragraph, paragraph end sentences and high frequency words (keywords). Each feature has an associated weight, obtained from training with manual summaries of four news articles, and the weights are combined linearly to produce an overall score for a sentence. As part of a learning phase, the stop-words and cue phrases lists of the system can be updated by a user that is generating a summary. The system is tested with five news articles at 38.5% extraction rate. The average sentence length of the five news articles is 17.4 sentences, the lowest length being 13 sentences and the highest 23. The system achieves a precision of 0.704, recall of 0.58 and an F-score of 0.636. The most dominant features in the experiment are title words and keywords. For the feature calculated using keywords, the percentage of keywords used for scoring and the weight given to this feature gave significant variations in system performance. Kamil needed to use 33% of the keywords identified to achieve the best result. He points out that because no stemmer is used for preprocessing the document, the frequency assigned to keywords is not accurate and hence less percentage of keywords could have been used if a stemmer was included. He also points out that the use of exhaustive stop-words could improve system performance.

Teferi [31] uses Naïve Bayes classifier (using the machine learning tool WEKA[11] (Hall et al. 2009) to create a single document summarization system. 480 news articles are used for experimentation, where 20 of them are used for testing and the rest for training. The test set has an average sentence number of 9.2 sentences. The 480 articles have an average of at most 6.7 sentences. Each document has a corresponding manual summary created by three different summarizers. The features used for training are presence of title words, location feature (first, middle, and last) and keyword (frequent words). Summarization is performed at 30% extraction rate. The system has a precision of 0.8, recall of 0.833 and an F-score of 0.816. These high values could be attributed to the very short news articles considered for testing purposes. For the keyword feature, the use of the top three keywords has been shown to be sufficient. This seems to be a reasonable quantity due to the use of very short articles for the experimentation. Teferi reports that the use of a single feature for summarization gives a very poor result while the use of all features results in the best system performance.

Helen [32] produced a single document summarizer for legal judgments. She uses two statistical features to score sentences: cue phrases and sentence location. Documents to be summarized are manually segmented into five thematic areas: introduction, reason, fact, judicial analysis and decisions. A summary is generated for each segment at a given extraction rate and combined to give a single summary.

Sentences with the highest weight are selected at 20% and 10% extraction rates from each segment and are merged together to serve as a single summary for thedocument. The system has the following performance: at 20% extraction rate, precision = 0.339, recall = 0.576, F-measure = 0.427 and at 10% extraction rate, precision = 0.392, recall = 0.506 and F-measure = 0.441. These performance values might decrease if the segmentation was made automatically.

Addis [33] customized and tested an open source tool, OTS, for summarizing Amharic news texts. In his study the researcher followed pure statistical approach. To determine the importance of the sentence he used frequency of terms. Addis conducted two experiments, which he called

[11] *is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License.*

them E1 and E2. For testing the OTS's performance, he used 30 news texts which were collected from different sources along with 90 ideal summaries prepared by two independent human evaluators three for each news text with extraction rates of 10%, 20%, and 30%.

Girma [8] has been developed automatic text summarizer for Afan Oromo news text based upon the Open Text Summarizer (OTS). 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 he 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 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.

This study also focuses on Afan Oromo language text in the news domain. However, it is different from previous local works in that it is Query specific text summarization which is based on the calculation of the relationship between sentences in the text document and the query given. 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. It employed one of the most extensively used information retrieval (IR) model and position method for finding the most similar sentences to the given query. The sentence ranking algorithm performs based on the sentence score to rank sentences in the order of their importance and finally summary is produced by selecting the top N sentences, where the value of N is set by the user.

CHAPTER THREE

AFAAN OROMO LANGUAGE

3.1 Background

Afaan Oromo is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of others neighbor countries like Kenya and Somalia [6]. It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts 34.5% of the local population [5]. 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. Currently, it is an official language of Oromia sate (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 in some universities with Bachelor and Masters Degree levels.

Few literature works, a number of newspapers, magazines, educational resources, official credentials and religious documents are published and available in the language [8]. 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.

3.2 Afaan Oromo Writing System

With regard to the writing system, "***Qubee"*** (a Latin-based alphabet) has been adopted and become the official script of Afaan Oromo since 1991 [34].The Oromo writing system is a modification to Latin writing system. Thus, the language shares a lot of features with English

writing with some modification. The writing system of the language is known as "***Qubee Afaan Oromoo***" is straightforward which is designed based on the Latin script. Thus letters in English language are also in Oromo except the way it is written. According to [8], Afaan Oromo is a phonetic language, which means that it is spoken in the way it is written. Unlike English or other Latin based languages there are no skipped or unpronounced sounds/alphabets in the language.

3.3 Afaan Oromo Alphabets: Vowels and Consonants

The "***Qubee***" writing system has a total of 33 letters that consists of all the 26 English letters with an addition of 7 combined consonant letters which are known as "***Qubee Dachaa***" [35]. These include ***ch, dh, sh, ny, ts, ph*** and ***zy***. All the vowels in English are also vowels in "***Qubee***". These are ***a, e, o, u*** and ***"i"***. Vowels have two natures in the language and they can result in different meaning. The natures are short and long vowels. A vowel is said to be short if it is one. If it is two, which is the maximum, then it is called long vowel. Consider these words: ***lafa*** (ground), ***laafaa*** (soft). In a word where consonant is doubled the sounds are more emphasized. For example, ***dammee*** (sweety), ***damee*** (branch). Afaan Oromo vowels and consonants are depicted in the tables 3.1 and 3.2 respectively. The International Phonetic Alphabet (IPA)[12] symbol for a phoneme is shown in brackets where it differs from the Oromo letter.

Vowels			
	Front	**Central**	**Back**
close	i /ɪ/, ii /i:/		u /ʊ/, uu /u:/
Mid	e /ɛ/, ee /e:/		o /ɔ/, oo /o:/
Open		a /ʌ/	aa /ɑ:/

Table 3.1: Afaan Oromo Vowels

[12] *is an alphabetic system of phonetic notation based primarily on the Latin alphabet.*

Consonants						
		Bilabial/ Labiodental	**Alveolar/ Retroflex**	**Palato-alveolar/ Palatal**	**Velar**	**Glottal**
Stops and affricates	Voiceless	(p)	T	ch /tʃ/	k	' /ʔ/
	Voiced	b	D	j /dʒ/	g	
	Ejective	ph /p'/	x /t'/	c /tʃ'/	q /k'/	
	Implosive		dh /ɗ/			
Fricatives	Voiceless	f	S	sh /ʃ/		h
	Voiced	(v)	(z)			
Nasals		m	N	ny /ɲ/		
Approximants		w	L	y /j/		
Rhotic			R			

Table 3.2: Afaan Oromo Consonants

3.4 Punctuation marks in Afaan Oromo

Punctuation is placed in text to make meaning clear and reading easier. Analysis of Afaan Oromo texts reveals that different punctuation marks follow the same punctuation pattern used in English and other languages that follow Latin Writing System [36]. Similar to English, the following are some of the most commonly used punctuation marks in Afaan Oromo.

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

Unlike English language apostrophe (‘) is not punctuation mark in Afaan Oromo, rather it is part of words. For example ***har'a*** (today), ***sa'a*** (time) etc.

3.5 Afaan Oromo Morphology

Morphology is the study of the meaning of individual units or morphemes of language and is concerned with the structure of words. Words are the fundamental building blocks of a language. Every human language, spoken, signed, or written, is composed of words. Every area of speech and language processing, from speech recognition to machine translation to information retrieval on the web, requires extensive knowledge about words [34].

The basic building blocks in morphology are morphemes. Morphemes are defined as the smallest units in a language to which a meaning may be assigned or, alternatively, as the minimal unit of grammatical analysis [34]. The form of a morpheme may be free or bound. A free morpheme occurs relatively freely within other words or morphemes.

In other words a free morpheme may form a word on its own, e.g., /door/. We call such words monomorphemic because they consist of a single morph. Bound morphemes, on the other hand, occur only in combination with other forms. The majority of affixes are bound morphemes. For example, the word /dogs/ consists of the free morph /dog/ and the bound morph /-s/ which is an affix [34]. 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 [35].

Morphemes in a language are composed of affixes and stems. 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 English, re-, en-, in-, as in reemploy, endanger, inaccessible, are examples of prefixes. In Afaan Oromoo, ***hin***, ***ni*** as in /***hinnyannu***/ (we don't eat) and ***nideemna*** (let us go) are examples of prefixes. A suffix, on the other hand, is an affix that is attached after a base. The plural markers ***-s*** and ***-oota*** of English and Afaan Oromo, respectively, are examples of suffixes, as in ***plants*** and ***namoota*** (men). An infix is a morpheme that is inserted within morpheme. In the work of [34] it is discovered that Afaan Oromo does not have infixes like English. There are two productive ways to form words from morphemes: inflection and derivation [36].

Inflectional Morphology: deals with the combination of a word with a grammatical morpheme, usually resulting in a word of the same class as the original stem, and serving some syntactic function, for example plurals of nouns.

They do not change the part-of-speech category but the grammatical function (also called morphosyntactic information) is changed. The different forms of a word are produced by inflection.

Derivational Morphology: creates new words (i.e., words with a different part-of-speech category) by adding a bound morpheme to a stem. Derivation can be applied recursively, i.e., words that are already the product of one derivation process can undergo the process again. The following is an example from English: large (adj.) = enlarge (en- + large) (v) = enlargement (enlarge + -ment) (noun), and from Afaan Oromo: ***bar-*** to know(v) **_** ***barumsa*** education (***bar-***+-***umsa***)(noun).

There are many ways of word formation in Afaan Oromo. These morphological analyses of the language are organized in six categories [36]. The categories are: nouns, verbs, adjectives, adverbs, prepositions, and conjunctions. The following is detail descriptions and examples of word formation process of Afaan Oromo based on the works of [8], [34] and [36].

3.5.1 Noun Inflection

Afaan Oromo nouns are words used to name any of categories of things, people, places or ideas. Nouns are inflected to indicate different grammatical functions such as number, gender, definiteness and case. Inflectional suffixes are combined with stem usually resulting in a word of the same class as the original stem.

i. Number

There are singular and plural numbers in Afaan Oromo and it has different suffixes to form the plural of a noun. The use of different suffixes differs from dialects to dialects. According to the work of [35] these suffixes include ***-oota, -oolii, -een, -lee, -wwan, -yyii, -eetii, -ii, and –oo.***

Even though the usage of some suffixes is not common, they can be used and accepted by the speakers of the language. Linguists agree that some groups of suffixes are most preferably applied to almost all nouns, and the others are used with only some words. Following are some examples corresponding to the common plural marker suffix.

Singular		Plural	
barsiisaa	teacher	***Barsiisota***	teachers
waraabessa	hyena	***Waraabeyyii***	hyenas
gaaffii	question	***gaffiiwwan***	questions
jabbii	calf	***jabbiilee***	calves

ii. Gender

Oromo has two grammatical genders, masculine and feminine in similar way to other Afro-Asiatic language. There is no neutral gender as in other language like English. Small number of nouns ends with ***-eessa***(m.) and ***-eettii***(f.), as do adjectives when they are used as nouns: ***obboleessa***'brother', ***obboleettii***'sister', ***dureessa***'the rich one (male.)', ***hiyyeettii***'the poor one (female). Grammatical gender normally agrees with biological gender for people and animals; thus nouns such as ***abbaa***'father', ***ilma***'son', and ***sangaa***'ox' are masculine, while nouns such as ***haadha***'mother' and ***intala***'girl, daughter' are feminine. However, most names for animals do not specify biological gender. Names of astronomical bodies are feminine: ***aduu***'sun', ***urjii***'star'. The gender of other inanimate nouns varies somewhat among dialects.

iii. **Definiteness**

Afaan Oromo has no indefinite articles (corresponding to English a, an, or some), but it indicates definiteness (English the) with suffixes on the noun: ***-icha*** for masculine nouns and ***-ittii*** for feminine nouns. Vowel endings of nouns are dropped before adding these suffixes: ***karaa*** 'road', ***karicha*** ' the road', ***nama*** 'man', ***namicha*** 'the man', ***haroo*** 'lake', ***harittii*** 'the lake'. Note that for animate nouns that can take either gender, the definite suffix may indicate the intended gender: ***qaalluu*** 'priest', ***qaallicha*** 'the priest (masculine)', ***qallittii*** 'the priest (feminine)'. The definite suffixes appear to be used less often than the in English, and they do not to co-occur with the plural suffixes.

iv. Cases

Case is a grammatical category of nouns that indicates the nature of their relationship to the verb in sentences. The number of cases varies from language to language. In this regard, nouns in Afaan Oromo are inflected for nominative, ablative, instrumental and locative cases. Each case is described and a summary is given in Table 3.3.

English	Base	Subject	Dative	Instrumental	Locative	Ablative	Possessive adjectives
I	*ana*, *na*	*ani*, *an*	*naa*, *naaf*, *natti*	*naan*	*natti*	*narraa*	*koo*, *kiyya* [*too*, *tiyya* (f.)]
you (sg.)	*si*	*ati*	*sii*, *siif*, *sitti*	*siin*	*sitti*	*sirraa*	*kee* [*tee* (f.)]
he	*isa*	*inni*	*isaa*, *isaa(tii)f*, *isatti*	*isaatiin*	*isatti*	*isarraa*	*(i)saa*
she	*isii*, *ishii*, *isee*, *ishee*	*isiin*, etc.	*ishii*, *ishiif*, *ishiitti*, etc.	*ishiin*, etc.	*ishiitti*, etc.	*ishiirraa*, etc.	*(i)sii*, *(i)shii*
we	*nu*	*nuti*, *nu'i*, *nuy*, *nu*	*nuu*, *nuuf*, *nutti*	*nuun*	*nutti*	*nurraa*	*keenna*, *keenya* [*teenna*, *teenya* (f.)]
you (pl.)	*isin*	*isini*	*isinii*, *isiniif*, *isinitti*	*isiniin*	*isinitti*	*isinirraa*	*keessan(i)* [*teessan(i)* (f.)]
they	*isaan*	*isaani*	*isaanii*, *isaaniif*, *isaanitti*	*isaaniitiin*	*isaanitti*	*isaanirraa*	*(i)saani*

Table 3.3 Personal pronoun with cases

3.5.2 Noun Derivation

In Afaan Oromo, derivational suffixes enable a new word, often with a different grammatical category to be built from stem/root of other words. But the distribution of suffixes is unpredictable since some nouns are formed with different suffixes. There are three derivational processes that create nouns in Afaan Oromo.

Abstract nouns are derived from other nouns by adding the suffix ***-ummaa,-eenya*** or ***-ooma*** to the noun stems. Thus, when these abstract noun formative morphemes are added to nouns, the final vowels of these words are deleted as the following set of examples illustrate.

bilisa	free	***bilisummaa***	freedom
fira	relative	***firummaa***	relationship
nagaa	peace	***nageenya***	peaceful

In Afaan Oromoo the nominal can be derived from the verb stem by suffixing the morphemes like ***-aa,-eenya, -tuu, -ina, -noo, -ii, -ee, -iinsa, -iisa, -umsa, -maata, -aatii***. [36] The following examples indicate the derivation of such nouns.

qab-	to have	***qabeenya***	property
rak-	to suffer	***rakkina***	problem
hubat-	to understand	***hubannomoo***	understanding
falm-	to argue	***falmii***	argument
tiks-	to shepherd	***tiksee***	shepherd/guardian
barsiis-	to teach	***barsiisaa***	teacher
bulch-	to govern	***bulchiinsa***	government
lol-	to fight	***loltuu***	soldier

3.5.3 Verb

Verbs are words that tell us the state of doing or being. Verbs are morphologically the most complex POS in Afaan Oromo, with many inflectional forms; numerous words with other POS are derived primarily from verbs.

3.5.3.1 Verb inflaction

A verb in Afaan Oromo can take various form while inflecting for persons. For example the root form of the verb '***duf***' (to come) inflected for person, number and gender as shown in the table.

<table>
<tr><td rowspan="3">Gender</td><td colspan="6">Persons</td></tr>
<tr><td colspan="2">First person</td><td colspan="2">Second Person</td><td colspan="2">Third person</td></tr>
<tr><td>Singular</td><td>Plural</td><td>Singular</td><td>Plural</td><td>Singular</td><td>Plural</td></tr>
<tr><td>Masculine</td><td rowspan="2">dhufe
I came</td><td rowspan="2">dhufne
We came</td><td rowspan="2">dhufte
You came</td><td rowspan="2">dhuftan
you came</td><td>dhufe
He came</td><td rowspan="2">dhufan
They came</td></tr>
<tr><td>Femanine</td><td>dhufte
She came</td></tr>
</table>

Table 3.4 verb inflactation for person, gender and number

3.5.3.2 Verb Derivation

An Afaan Oromo verb stem can be the basis for three derived voices, passive, causative, and autobenefactive, each formed with addition of a suffix to the stem, yielding the stem that the inflectional suffixes are added to [36].

a) Passive voice

The Afaan Oromoo passive corresponds closely to the English passive in function. It is formedby adding ***-am*** to the verb root. The resulting stem is conjugated regularly. For instance, ***beek-*** 'know', ***beekam-*** 'be known', ***beekamani*** 'they were known'; ***jedh-*** 'say', ***jedham-*** 'be said'**,** ***jedhama*** 'it is said'

b) Causative voice

The Afaan Oromoo causative of a verb V corresponds to English expressions such as 'cause V', 'make V', 'let V'. With intransitive verbs, it has a transitivizing function. It is formed by adding --***s***, ***-sis***, or ***–siis*** to the verb stem, except that stems ending in ***-l*** add ***-ch***. For instance, beek- 'know', ***beeksis-*** 'cause to know, inform', ***beeksifne*** 'we informed'; ***ka-*** 'go up, get up', ***kaas-*** 'pick up', ***kaasi*** 'pick up (sing.)!'; ***gal-*** 'enter', ***galch-*** 'put in', ***galchiti*** 'she puts in'; ***bar-*** 'learn', ***barsiis-*** 'teach', ***nan barsiisa*** 'I teach'.

c) Autobenefactive voice

The Afaan Oromo autobenefactive (or "middle" or "reflexive-middle") voice of a verb V corresponds roughly to English expressions such as 'V for oneself' or 'V on one's own', though the precise meaning may be somewhat unpredictable for many verbs. It is formed by adding ***-adh*** and ***-at*** to the verb stem. 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***). For instance, ***bit-*** 'buy', ***bitadh-*** 'buy for oneself', ***bitate*** 'he bought (something) for himself', ***bitadhu*** 'buy for yourself (sing.)!'; ***qab-*** 'have', ***qabadh-*** 'seize, hold (for oneself)', ***qabanna*** 'we hold'. Some autobenefactives are derived from nouns rather than verbs, for example, ***hojjadh-*** 'work' from the noun ***hojii***'work'.

3.5.4 Adjectives

Adjectives are very important in Afaan Oromo because its structure is used in every day conversation. Oromo Adjectives are words that describe or modify another person or thing in the sentence [40]. Unlike English adjectives are usually placed after the noun in Afan Oromo.

For instance, in ***Caalaan mana guddaa bite*** "Chala bought a big house" the adjective ***guddaa*** comes after the noun ***mana***. Moreover, in Afan Oromo sometimes it is difficult to differentiate adjective from noun.

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)

Adjectives in Afaan Oromo can be classified based on their function like English such as those express color, size, shape, quality and taste. Some of these adjectives are described below.

Halluwwan (Colors)hamma(size)boca (shape)

gurraacha (Black)	***gadifagoo*** (deep**)**	**geengoo** (circular)
baluu (Blue)	***dheeraa*** (long)	**sirrii** (straight)
bifa bunaa (Brown)	***dhiphaa*** (narrow)	**golarfee** (square)
daalacha (Gray)	***yabbuu*** (thick)	**golsadee** (triangle)
magariisa *(*Green)	***gabaabaa*** (short)	
diimaa *(Red*)	***xinnaa*** (small)	
adii(White*)*		

Tests Quality

hadhaawaa (bitter)	***hamaa*** (bad)
asheeta (fresh)	***qulqulluu*** (clean)
sogidaawaa (salty)	***dukkanaawaa*** *(*dark)
oganaawaa (sour)	***ulfaataa*** *(*difficult)
kaurgoo baayyatu (spicy)	***xuraawaa*** (dirty)
mi'aawaa (sweet)	***salphaa*** (easy)

3.5.5 Adverbs

Oromo adverbs are part of speech. Generally they are words that modify any part of language other than a noun. Adverbs can modify verbs, adjectives (including numbers), clauses, sentences and other adverbs [37]. The four categories of adverbs in Afaan Oromo: *Adverb of time, adverbs of place, adverbs of manner and adverbs of frequency*. Some Examples of Adverbs in Afaan Oromo are:

Adverb of Time

Kaleessa (yesterday)

harr'a (today)

bor (tomorrow)

amma (now)

gaafas (then)

booddee/ eger (later)

suuta (slowly)

Adverb of Manner

baayyee (very)

baayyee (quite)

dhugumaan (really)

dafee (fast)

gaarii (well)

dafee (quickly)

Adverb of place

As (here)

Achi (there)

gar asana (over there)

iddoo hunda (everywhere)

fagoo (away)

ala (out)

3.5.6 Prepositions

Prepositions in Afaan Oromo, links nouns, pronouns and phrases to other words in a sentence. The word or phrase that the preposition introduces is called the object of the preposition [37].

Example

In gara Finfinnee deeme. He went to Addis.

Hangan dhufutti naheegi. Wait for me until I came back.

Namni akka harkaan waa hojjechuuf fayyadamu argi maalitti fayyadamaa?

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

Here are some lists of prepositions in Afaan Oromo.

waa'ee	about
naannoo	around
dudduuba	behind
garuu	but
f	for
irraa	from
ittaanee	next
hamma	till
jala/ gajjallaa	under
karaa	via
wajjin	with
kanaaf	due to

3.5.7 Conjunctions

A word that can be used to join or connect two phrases, clauses and sentences is known as a conjunction [38]. Conjunctions can be divided into coordinating and subordinating conjunctions. Coordinating conjunctions are used to connect two independent clauses. Mostly these conjunctions are used when the speaker needs to lay emphasis on the two sentences equally. Some of these conjunctions in Afaan Oromo include: ***garuu*** 'but' ,***moo***' or', ***kanaafuu***' therefore',***haata'u malee***'however/so', ***tu'ullee***'even though' etc. Consider the following example: ***Hawiin siriitti hinqo'atu garuu qormaata dabartee jirti.*** This means 'Hawi is not studying hard but she passed the exm'. '***garuu***' in this sentence is coordinating conjunction. It is used to join the two sentences ***Hawiin siriitti hinqo'atu'***and ***qormaata dabartee jirti*** which are independent.

Subordinating conjunctions are those conjunctions that are used to join main clause with subordinate clause. A subordinating conjunction is always followed by a clause [39]. Afaan Oromo subordinating conjunctions include ***yoo***'if', ***akka waan***'as if', ***wayta/yeenna***'when', ***hamma***'until', ***erga***'after', ***dursa***'before' etc.

The following example illustrates the above case: ***Akka waan nabeeku fakkeesse*.** This means "He acts as if he knows me". ***Akka waan***in this sentence is used as subordinating conjunction. It joins one subordinating clause that is ***waan nabeeku*** **'**As if he knows me' and ***fakkeesse*** **'**He acts'.

3.6 Word and Sentence Boundaries in Afaan Oromo

Most Latin languages a word is separated from other words by white space character [46]. Afaan Oromo also uses white space to separate words from each other. 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 (!) [39].

3.7 News Writing Style

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. Over a century the inverted pyramid style has been the defacto of writing [47]. This refers to the decreasing importance of information in subsequent paragraphs. 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 [48]. The all articles used in our corpus follows this writing style.

CHAPTER FOUR

DESIGN AND IMPELEMATATION

This Chapter describes the design and implementation of query based automatic text summarizer for Afaan Oromo. The system named as Query Based Automatic Text Summarizer for Afaan Oromo (QBATSAO).

4.1 System Architecture

The design and implementation process of the proposed summarizer consists of four major modules. These are preprocessing, index term construction, sentence ranking and summary extraction. The preprocessing module takes a raw text as an input and applies some basic routines to eliminate textual elements that are not useful in further processing of textual data. After an input document is formatted and stemmed, the second module is responsible for constructing term by sentence matrix. Where the columns represent the sentences and the rows represent the unique terms (usually words) occurring in the documents. Then in the third module, the significance score of sentences are computed which help us to rank sentences in the document. Finally, the summary extraction module is responsible for selecting best candidate sentences to form as summary.

The general architecture of the system is shown in Figure 4.1 and the methods and techniques used in each module of the summarization system are explained in detail in the succeeding sections.

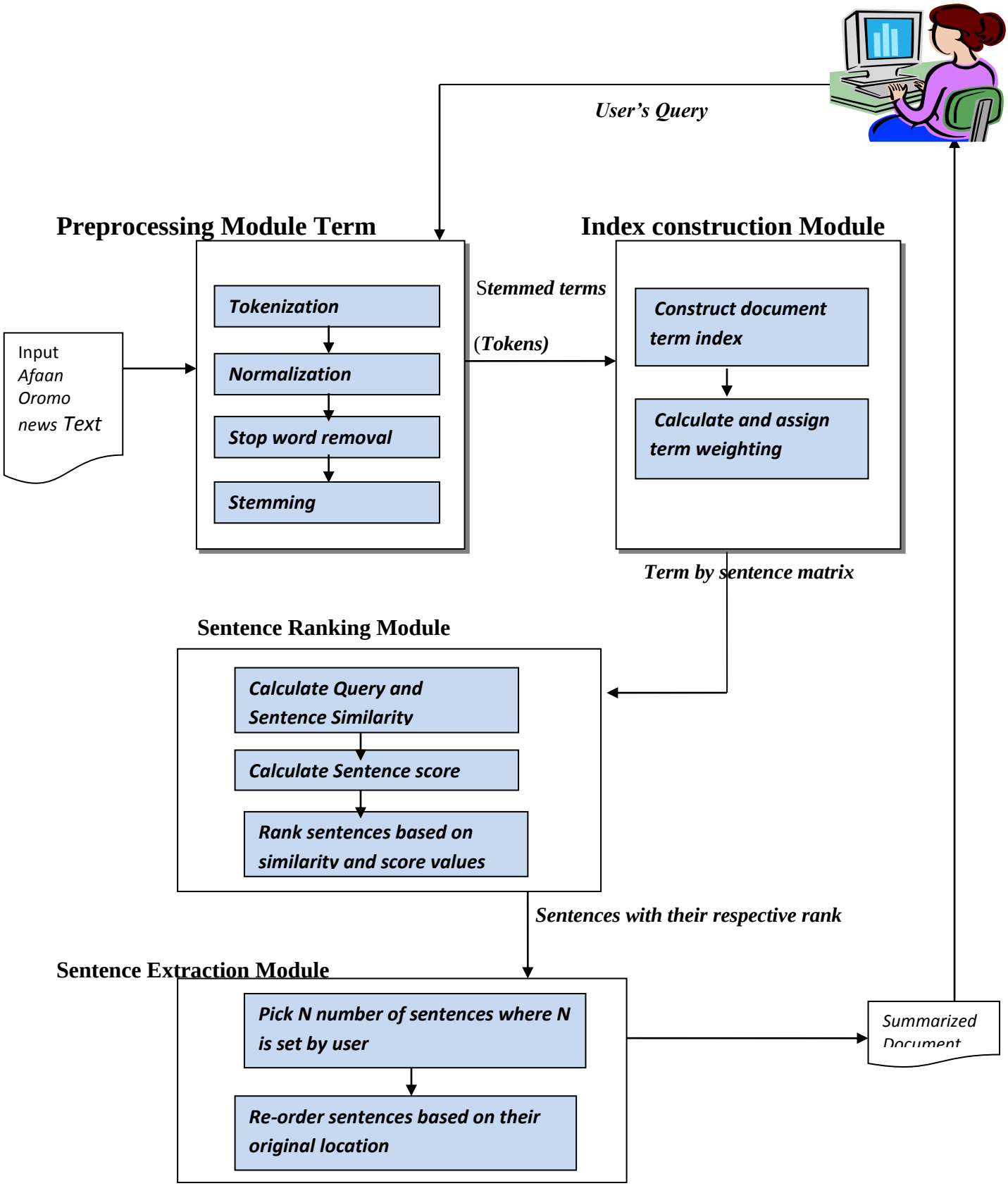

*Figure 4.1 Architecture of the proposed summarize*r

4.1.1 Preprocessing

This module consists of four components including tokenization, normalization stop words removal, and stemming. In extraction based summarization the output of this step is often a set of sentences where non-content words and punctuations are removed and content words are reduced to their base forms [45]. We discussed each component in the subsequent section.

4.1.1.1 Tokenization

The generation of a summary depends on the computed score of each sentence, and these scores depend on the individual words that constitute the sentences [9]. Hence, tokenization in text summarization involves text splitting into words and sentences. Extractive summarization requires high accuracy sentence boundary detection methods because even a single incorrectly separated incomplete sentence will greatly reduce the coherence and readability of a summary [45]. This process highly relies on language specific markers. In this thesis, as we discussed in chapter 3 section 3.4, Afaan Oromo texts follow the same punctuation pattern used in English. Thus, periods (.), question marks (?) and exclamation marks (!) are used to identify sentence boundaries. One of the challenges related to identifying sentence boundaries was punctuational ambiguities that periods can also used for abbreviations. For instance, **W.B.*Waree booda*** (After noon), **K.K.F*Kan kana fakkaatan*** (etc..). To overcome such challenges some of pre defined abbreviations taken from the previous studies [8] and identified from the documents, were given to the system with their expanded form. Lists of abbreviations used are available in Appendix V.

Following the sentence splitting, the individual words are extracted from the sentences using white space. The challenges related to extracting words was identifying compound words such as ***Mana Barumsaa*** (School), ***Mana Murtii*** (Courts), ***Qonnaan Bulaa*** (farmer). Predefined compound words are identified from the articles in the corpus and an attempt was made to treat such compound words as they are (without splitting them using white space). Lists of compound words identified from the documents are available in appendex VI. Then, cleaning of non-language specific symbols or characters such as $, @, %, &, #, etc...and punctuation marks have been performed.

4.1.1.2 Normalization

Normalization involves process of handling different writing system. Primarily it is handling problem related with variation of cases (UPPER CASE, or lower case or Mixed Cases). So the good way to handle this problem is converting the whole document in to similar case. Often convert to lowercase everything, since most of the time users use lowercase regardless of correct capitalization. As discussed in sections 3.2 Afaan Oromo uses the same capitalization system with English. Here are some examples: ***Dimokraasii*** vs. ***dimokraasii,Caalaa***vs. ***caalaa*** *vs.* ***CAALAA***, and etc. In the following algorithm we tried to how tokenization and normalization are implemented.

Input*: Afaan Oromo news text, abbreviation lists*

Output*: tokenized and normalized terms and sentences*

```
Let  sen be a string which set to be null
Let wrd be a list which set to be null
        Read file
       Convert each character in file to lower case
For lists in abbrlists
     Read list
     If  file contains list
         Replace it with the corresponding expansion form
     If file in read endswith '.', '!', or '?'
        Put each sentence in sen
Return sen
       For file in sen
       Split each word using white space
      Put each word in wrd
If wrd contains "?,$#@^&*1234567890
   Replace wrd with white space
Return wrd
```

Algorithm 4.1 normalization and tokenization

4.1.1.3 Stop words removal

Some words, referred to as "***stop-words***" are either words that do not contribute significantly to the overall topic of the document such as conjunctions, articles, pronouns or are words that appear in many sentences, and thus do not serve to topically distinguish one sentence from another. Such words can be identified and removed by using a predefined list of stop-words.

In this thesis, we have used some common stop-words which are taken from previous studies [6] and [38] and those specific to the corpus. See Appendix III for the list of stop-words.

Input: *tokenized text; a list which contains stop words*
Output: *tokenized stop word free text*
Let wrdstrm be a list which set to be null
Open stopwordlists
Read file in wrd
For stopword in wrd
If stopword in wrd
Replace with white space
Put wrd in wrdstrm
Return wrdstrm

Algorithm 4.2 Removing Stop Words

4.1.1.4 Stemming

Once the process of removing stop-words is completed, the next task in the preprocessing of the input document is stemming. Stemming is the process of reducing the inflected forms of a word to its root form by stripping off the affixes. It is commonly used in information retrieval tasks to reveal semantic similarity between different morphological variants of a word and makes the data less sparse[38]. In this thesis, we employ the stemming algorithm developed by [41]. It takes as input Afaan Oromo 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 [41] an evaluation of the system showed the algorithm accuracy giving 96 percent correct results. See Appendex IV for the list of affixes (suffixes and prefixes).

Input: *tokenized stop words free text, Aaan Oromo prefix and suffix lists*
Output: *stemmed text*
Let stemwrd be a list which set to be null
Let wrdstrm be a list contains tokenized stop words free text
Open affixlists
for file in wrdstrm
If wrdstrm matches with one of the rules
Remove the suffix and do the necessary adjustments
ELSE
Stop processing
IF there is no applicable condition and action exist
Remove vowel and return the result

Algorithm 4.3 Stemming Algorithm

4.1.2 Index Term Construction

Vector Space Model, requires both the documents and the queries (document title in our case) to be represented mathematically as vectors in some vector space. To achieve this, a python code is written that first identifies unique terms from the output of the first module and computes the frequency of each unique term in the sentences they are contained in. Then a term by sentence matrix is constructed which has as many rows as the number of unique terms in the document and as many columns as the number of sentences in the document. Thus, the entry at row m and column n of the matrix is filled with the TFISF of the term m in sentence n. The query is also represented by one column in the term by sentence matrix using the database of terms identified in the document.

Terms	*Sent1*	*Sent2*	*Sent3*	*Sent3*	*Sent4*	*Sent5*	*Sent6*	*Sent7*	*Sent8*
godinaal	0.069	0.125	0.000	0.000	0.034	0.150	0.000	0.000	0.000
oromiy	0.069	0.000	0.000	0.333	0.034	0.000	0.000	0.066	0.000
Nam	0.000	0.000	0.000	0.0250	0.033	0.000	0.000	0.012	0.000
malaammal	0.000	0.000	0.000	0.083	0.033	0.000	0.090	0.000	0.000
himatam	0.000	0.135	0.000	0.000	0.034	0.150	0.000	0.000	0.000
adamuusaa	0.000	0.000	0.000	0.016	0.033	0.000	0.000	0.066	0.333
komiish	0.000	0.000	0.000	0.083	0.033	0.000	0.000	0.000	0.000
naamusaa	0.000	0.135	0.000	0.000	0.034	0.150	0.000	0.000	0.000
Beek	0.000	0.135	0.000	0.000	0.034	0.000	0.000	0.000	0.000
adeem	0.000	0.000	0.000	0.016	0.033	0.000	0.000	0.066	0.333
hoj	0.000	0.000	0.000	0.083	0.033	0.000	0.090	0.000	0.000
dhimm	0.000	0.135	0.000	0.000	0.034	0.150	0.000	0.000	0.000
mootum	0.000	0.000	0.000	0.333	0.034	0.000	0.000	0.066	0.333
barum	0.000	0.000	0.000	0.083	0.033	0.000	0.000	0.000	0.000

Table 4.1: Sample of term by sentence matrix

As shown in the table above the unique terms found in 8 sentences are taken and the TFISF of each term is calculated and filled in each cell.

4.2.3 Sentence Ranking

As the primary goal of extractive summarization is to determine the best candidate sentences for summary generation, the sentence ranking module is the core of the summarization system. The sentence ranking algorithm performs based on the sentence score.

Thus, to compute the significance score of a sentence we employed in this thesis two features: cosine similarity to the title vector, and sentence position in the document.

4.2.3.1 Similarity measures

There aremany different ways to measure how similar two documents are to each other, or how similar a document is to a query. The cosine measure is a very common similarity measure. Using a similarity measure, a set of documents can be compared to a query and the most similar document returned. If the query term does not occur in the sentence, similarity measure score should be 0, else the more frequent the query term in the sentence, the higher the score (should be). Query and sentence similarity is based on length and direction of their vectors.

As we discussed in chapter two equation 2.3, we first calculated W_{ti}, W_{qi} which are indexing terms given to a document. Then the similarity is calculated by using the widely used cosine similarity measure (presented in chapter 2, equation 2.4). The following table shows sample sentences rank using cosine similarity measure.

Rank	Sentences	similarity
1	sentenc1:	1.112
2	sentenc2:	0.501
3	sentenc3:	0.337
4	sentenc4:	0.252
5	sentenc5:	0.203
6	sentenc10:	0.178
7	sentenc6:	0.167
8	sentenc7:	0.164
9	sentenc9:	0.126
10	sentenc8:	0.125
11	sentenc12:	0.108
12	sentenc11:	0.092
13	sentenc13:	0.080
14	sentenc14:	0.077
15	sentenc18:	0.073
16	sentenc15:	0.068
17	sentenc16:	0.064
18	sentenc21:	0.062
19	sentenc17:	0.059
20	sentenc19:	0.054
21	sentenc20:	0.050

Table 4.2 Rank of sentences with cosine similarity measure score

As shown in the table above the sentences are ranked based on the similarity value of the sentence and query vectors.

4.2.3.2 Sentence position method

In an attempt to take document genre information into consideration, the position of each sentence in the document is used as additional features to compute the significance of a sentence. News articles in general are written in such a way that the information presented is arranged in descending order of importance. The most important information is placed at the beginning of the news article followed by less important information [15].

The position score of a sentence is calculated using the formula 1/n where n is the position of the sentence in the document as adopted from [9]. Thus, the first sentence in the document gets the highest score equal to 1 and the last sentence in 0the document gets the lowest score.

4.2.3.3 Sentence Score Calculation

Since our thesis is focused on query based extraction summarization, for the evaluation purpose we used the title of the document as a query and provide to the system to get a summary at different extraction rates. The titles of news articles are usually very informative about the content of the article [9]. Thus we used cosine similarity measure to compute the similarity between the sentences in the document and the query (title). In addition to this, we assume that summarization results can further be improved if sentence position feature is used along with the similarity value. Therefore, in order to compute the significance score of a sentence, we define a formula adopted from [9] that combines the two features as follows:

$$S(si)=\alpha Sim(i) + \beta Pos(i) \quad (4.1)$$

Where, **S(si)** is the significance score of sentence i,

Sim is the cosine similarity of sentence i and the query vector,

Pos is the position score of sentence i.

The constant **α** and **β** represents the weighting value given to the cosine similarity and position method respectively. We gave higher weighting value for the cosine similarity than the position

method. Since our first aim is extracting the sentences which are similar to the query (title). The weighting value given to both methods was determined though experiments.Table4.3 shows the rank of sentences based on the sum cosine similarity and position methods values.

Rank	Sentences	similarity
1	sentenc1:	0.112
2	sentenc10:	0.078
3	sentenc12:	0.025
4	sentenc7:	0.021
5	sentenc18:	0.017
6	sentenc9:	0.015
7	sentenc21:	0.015
8	sentenc14:	0.005
9	sentenc3:	0.004
10	sentenc5:	0.003
11	sentenc13:	0.003
12	sentenc4:	0.002
13	sentenc19:	0.002
14	sentenc11:	0.002
15	sentenc15:	0.001
16	sentenc2:	0.001
17	sentenc16:	0.001
18	sentenc8:	0.000
19	sentenc6:	0.000
20	sentenc20:	0.000
21	sentenc17:	0.000

Table 4.3 Rank of sentences with cosine similarity measure and position method Score

As shown in the table above the position method changed the rank of some sentences when compared with the rank of the sentences depicted in the table 4.2.

4.2.4 Summary Extraction and Re-ordering

A summary is produced by selecting the top N sentences, where the value of N is set by the user. Then our summarizer extracts the most salient sentences from the original document sequentially until the desired length of the summary is reached. For instance, as shown in the table 4.2 if 20% extraction rate is required by the user our summarizer selects [sentence1, sentence10, sentence12, sentence7].

To generate a readable coherent summary it is very important to order the sentences correctly. For a single document the order of sentence in the original document can be used as order to generate summaries [4]. In this study the sentences in original document have been given serial number in ascending order and the selected sentences are reordered accordingly. Thus, the above selected sentences are reordered and generated as [sentence1, sentence7, sentence10, sentence12].

CHAPTER FIVE

EXPERIMENTATION AND DISCUSSION

5.1 Introduction

In this Chapter, we describe the performance of the summarization system presented in Chapter four. Though evaluating the performance of the summarization system is an important part of the study, this was not a straightforward task. This is because there is no standard method or well defined criteria for summary evaluation. In the subsequent pages of this thesis, we describe the set of procedures that have been used to conduct the experiment.

5.2 Experimental Settings

5.2.1 Data Corpus

As indicated in chapter one section 1.4.2, for this study we used 40 Afaan Oromo news items whose lengths are in the range of 20 to 66 sentences were used. The news items were collected from ***Kallacha Oromiyaa*** and ***Bariisaa*** newspapers. ***Kallacha Oromiyaa*** is prepared by Oromia Regional State government published once in two weeks and ***Bariisaa*** is prepared by Ethiopian press agency and published once in a week. The reason why these two newspapers were selected is as a large collection of news items on different domains is provided on these newspapers and they provide up-to-date news articles in digital format. Prior to experiment each of these articles was manually cleaned by removing information in non text format such as tables, figures and document source. The title of each document also removed and used as a query for their respective documents in the experimentation. The title can serve as a query in some query-based text summarizer to incorporate sentences which are highly related to the title in final summary [4].

Document Attributes	Values
Number of documents	40
Maximum number of sentences per document	66
Minimum number of sentences per document	20
Maximum number of words per document	729
Minimum number of words per document	322
Average words per document	38

Table 5.1: Particulars of the evaluation data set

As shown in the table 5.1 the news articles with less than 20 sentences were not used in the evaluation due to the fact that summarizing short articles does not make much sense in real applications [42]. Though evaluating the system with news stories containing more than 66 sentences is very desirable, it was very difficult to obtain such news stories.

5.2.2 Creating Manual Summary

Three Experts from Oromia Culture and Tourism Bureau, language department were employed to prepare manual summarization on the 40 news contained in the data corpus. The title (query) and the document were given to the human summarizers with the corresponding guideline adapted from [11]. This guideline was prepared for generic based evaluation which orders the summarizer to summarize the document to identify the most representative of the sentences and incorporate in manual summary. But for this study we customized the guideline so that it guides the summarizer to select the sentences mostly related to the title (query) at 10%, 20% and 30% extraction rate.

Because of the disparities in the human summarizer's sentence ranking, sentences in each news text are further re-ranked based on the average rank of the ranks assigned by the three summarizers. For instance, if a sentence is ranked as 2nd, 4th and 6th by the first, second and third evaluator respectively, then the average rank of this sentence will be the 4th.

Extraction rate is calculated as multiplying the total number of sentences in a document by corresponding extraction rate and rounding off the result to the nearest integer. For instance, to prepare the manual summary of a given news text containing 27 sentences at 30% extraction rate, the top ranking sentences are extracted until the total number of sentences in the summary reaches 9 (27x0.3≈8).

5.3 Experiments

As we discussed in Chapter 4 section 4.3, of thesis to compute the significance score of a sentence we employed two features: cosine similarity to the title vector, and sentence position in the document. As a result, we have two distinct experiments (EXP 1 and EXP 2). All the summaries extracted by each experiment at a given extraction rate are compared against the corresponding reference summary. The two experiments with different methods are:

EXP1: The significance scores of sentences are calculated by cosine similarity to the title vector to rank the sentences in the document.

EXP2: The significance scores of sentences are calculated using both cosine similarities to the title vector and position method to rank the sentence in the document.

5.3.1 Experiments to Determine Weighting Values

In an attempt to find the best weights that are to be used in equation 4.1, we performed experiments on 15 news articles randomly selected from our dataset with different weights constrained between 0 and 1. We confined to 15 news articles because it was time taking and tedious to perform the experiments on all news articles in the dataset. We started the experiment by giving equal weight for both features (cosine similarity and position method) then higher value for cosine similarity through the experiments. For all the weights, the manual summary and the system summary are produced at 10%, 20% and 30% extraction rate. Thus, 5 experiments conducted to determine the best weights as follows:

1. ***0.5 sim(i) + 0.5 pos(i)***
2. ***0.6 sim(i) + 0.4 pos(i)***
3. ***0.7 sim(i) + 0.3 pos(i)***
4. ***0.8 sim(i) + 0.2 pos(i)***
5. ***0.9 sim(i) + 0.1 pos(i)***

For each experiment, the F-Score obtained using the above weighting values is shown in Figure 5.1.

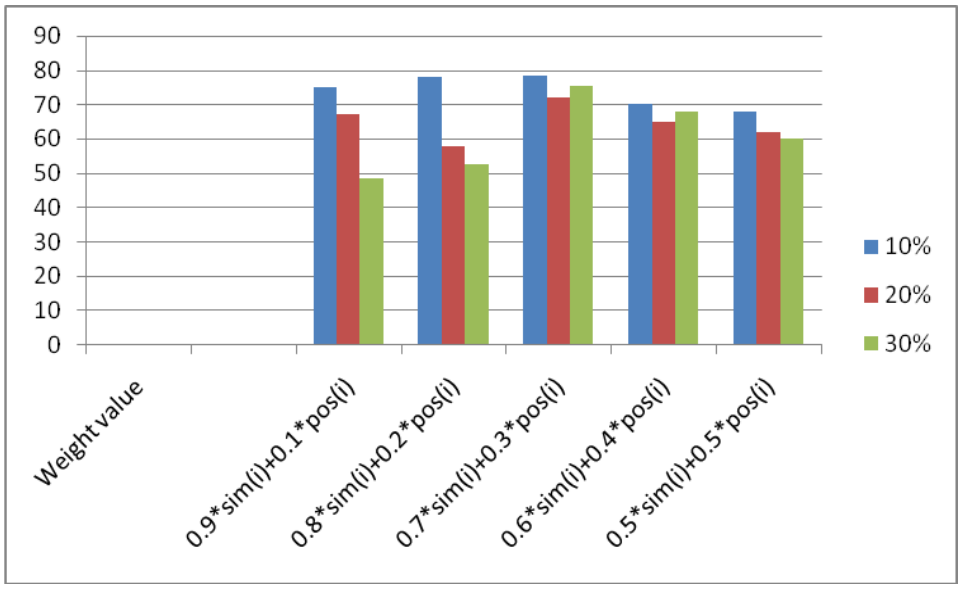

Figure 5.1 Average F-measures of EXP2 for different weight value

As shown in figure 5.1 the weighting value used with both features are registered different scores. In all experiments, the extraction rates at 10% exceed the rest (20% and 30%). But an experiment with 0.7*sim(i) and 0.3*pos(i) registered the highest value at 10%, 20 % and 30% extractions rate. Threrefore we used 0.7*sim(i) + 0.3*pos(i) in EXP2 to extract the most salient sentences in the document as the final summary. Finally we compared the result obtained by both Experments (EXP1 and EXP2) with the manual summary created by experts.

5.4 Evaluation and Discussion of the Result

As indicated in the methodology section of this research for evaluating the proposed system we employed an intrinsic evaluation method which evaluate the performance of the proposed system objectively and the linguistic quality (informativness and coherence) subjectively.

5.4.1 Performance evaluation and discussion

It is one of the evaluation methods employed for this study to measure performance of the summarizer. As described in Chapter 2, performance evaluation metrics assess the quality of the system summary based on the number of sentences that are common to the system summary and

the manual summary. The most common performance evaluation metrics are recall (R), precision (P), and F-Score (F). The standard definitions of R, P and F are given in chapter 2, section 2.6

Experimentation is done on the total of 40 news articles in the dataset (corpus). Hence, for each of the two experiments (EXP1 and EXP2) 120 automatic summaries, three for each news articles, have been generated at 10%, 20%, and 30% extraction rate.

The results of the two experimentation methods (EXP1 and EXP2) have been compared against the manual summary created by experts. Accordingly, both the value of recall and precision are 100% if all of the sentences in system's summary are available in the manual summary and 0% if none of the sentences in system's summary exists in manual summary. In addition, F-measure (score) is used to compute a weighted average of the recall and precision, where an F-measure reaches its best values at 100% and worst score at 0%.

The value of F-measure is 0% if either precision or recall is 0%, between 0% and 100% for non-zero values of precision and recall and undefined if both recall and precision is 0%. Since F-Score gives the same importance to precision and recall, using it as an evaluation measure is a good trade-off between precision and recall. Hence, for the sake of discussing evaluation result, we used F-measure in this thesis. The evaluation results of both experiments (EXP1 and EXP2) are depicted in tables 5.2.

Extraction Rate %	F-scores Result of the Experiments	
	EX1	EX2
10%	88%	82%
20%	70%	78%
30%	73%	82%

Table 5.2 F-measure results of EXP1 and EXP2

As shown in table 5.2, the average of F-measures (scores) of the EXP1 at extraction rate of 10%, 20% and 30% is 88%, 70% and 73% respectively and EXP2 at the same extraction rate is 82%, 78% and 82% respectively. See Appendixes VII and VIII for detail. EXP2 outperforms EXP1 at 20% and 30% extraction rates by 8% and 9% respectively. For 10% extraction rate EXP1 outperforms EXP2 by 6%.

The improvement occurred at 20% and 30% extraction rate is due to the position method used along with cosine similarity measure to incorporating the most salient sentences in the final summary. The average performance increases with an increase in extraction rate for 20% and 30% but it did not hold for this experimentation at 10% extraction rate. This can be due to the fact that most of the terms in the title of news articles are appeared in the sentences found at beginning lines than the last ones[48].

Therefore these sentences have high chance to be incorporated in the final summary at 10% extraction rate than 20% and 30% extraction rates.

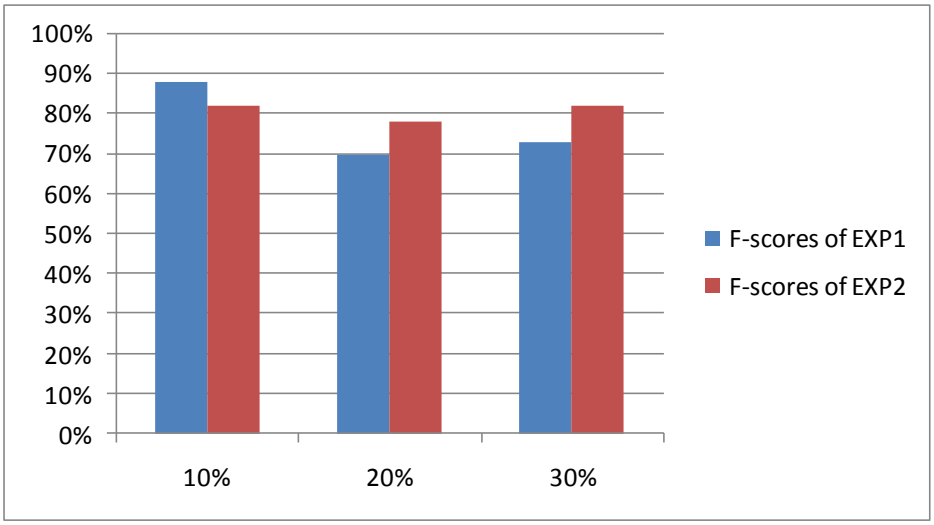

Figure 5.2 system performance comparisons EXP1 and EXP2

5.4.2 Subjective Evaluation and Discussion

In this evaluation system we tried to evaluate the linguistic quality (informativeness and coherence) of the summary generated by the system. Five evaluators who are graduate students from Addis Ababa University, Afaan Oromo Department were selected using purposive sampling techniques to evaluate the system. These students are selected because we believe that they know the language very well and able to evaluate the summary produced by the system.

We could not get more than five evaluators because they complain that the process might take much of their time and they are not willing to do so. We customized the subjective evaluation technique used by [11] an automatic text summarizer for Arabic in order to evaluate our summarizer.

40 news items which are used for objective evaluation were also used in EXP1 and EXP2 for this evaluation. However, we divided the 40 news items in to three to test for each extraction rate. Accordingly the first 13th news articles have been evaluated at 10% extraction rate, the second 13th at 20% and the last 14th at 30% extraction rates. The two system summaries created with different methods are compared according to the following scale measures shown in table 5.4. 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 score value for Exp1 by the first evaluator is 2, the second is 3, the third is 2, the fourth is 3 and the fifth is 4 then the percentage of the total result for Exp1 will be the sum of the score values divided by 20(since the sum of the maximum score of all evaluators is 20) multiply by 100. **i.e. (2+3+2+3+4)/20*100 = (14/20)*100 = 70%.**

Scale Measures	Score Value	Interpretation of the measure
Very poor	0	The summary is very poor and not related to the core meaning of the document at all
Poor	1	The summary is poor as the core meaning of the document is missing
Fair	2	The user is somehow satisfied with the result, but expected more
Good	3	The summary is coherent and it carries the main idea
V.good	4	The summary is of much coherent and focuses more on the core meaning of the document and the user is happy with the result.

Table 5.3 Five Scale measures for subjective evaluation

Extraction Rate %	Average Score Result of the Experiments	
	EX1	EX2
10%	55%	59%
20%	67%	77%
30%	79%	91%

Table 5.4 the average scale measures result of EXP1 and EXP2

As shown in table 5.5, the percentage average of EXP2 exceeds EXP1 by 4%, 10% and 12% at extraction rate of 10%, 20% and 30%, respectively. See Appendix IX for the detail of subjective evaluation result.

This reveals that the evaluators judge that the summaries of the EXP2 are more informative than EXP1. This is due to that the position method used in EXP2 improves the informativeness of the system generated summary. Even though unresolved anaphora (reference) exits in some system generated summary, the reordering of the sentences to their original position after extracting the top N sentences also contributes for summary to be coherent.

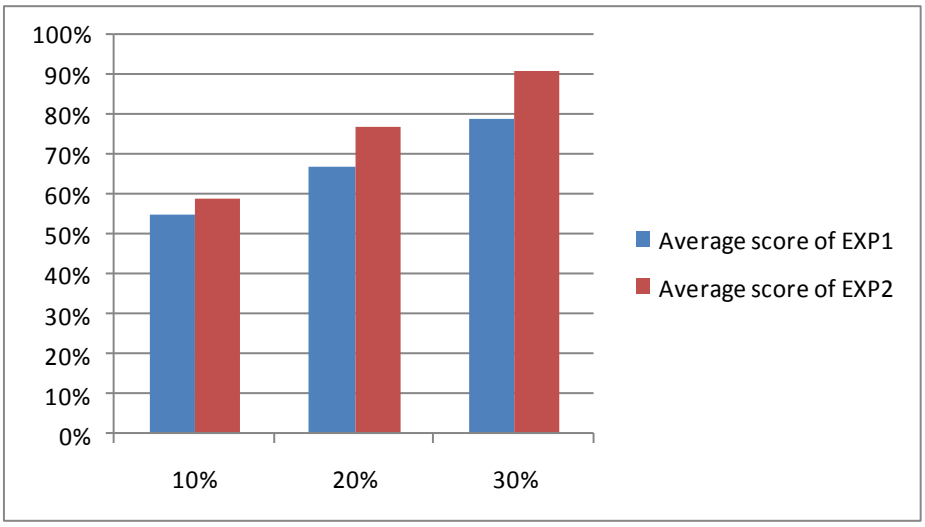

Figure 5.3 linguistic quality comparison of EXP1 and EXP2.

We can also observe that summaries informativeness and extraction rates are directly related, that is the extent of automatically generated summaries in representing the main theme of the input article improves when the extraction rate is increases.

5.5 Challenges

In the course of conducting this research we faced a number of challenges that need to be resolved. One of the challenges was that some summary sentences contain unresolved references that may cause difficulties in understanding. Another challenge was the lack of standardized and well prepared Afaan Oromo Corpus which required conducting conclusive experimentation of the proposed system.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

As the amount and availability of textual information increases, techniques to help people obtain and digest the important information in these sources become increasingly important. Today there are numerous documents, papers, reports and articles available in digital form but the information in them is often too abundant to manually search and choose which knowledge one should acquire. Thus, this information must first be automatically filtered and extracted.

In this thesis, we tried to develop a prototype of query based extractive summarization system for texts written in Afaan Oromo. For this purpose we used an IR model known as VSM to compute the significant sentence score and rank the sentences in the document. We also used position method along with VSM in attempt to improve the quality of the produced final summary. In both cases, a summary is produced by selecting the top N sentences, where the value of N is set by the user. After the N-top sentence is selected the final summary is generated based on the sequence they have in the original document.

We have used 40 Afaan Oromo news articles which are collected from both ***Kallacha Oromiyaa*** and ***Bariisaa*** newspapers. Two distinct experiments were conducted: EXP1 (the significance scores of sentences are calculated by cosine similarity to the title vector to rank the sentences the document.) and EXP 2 (the significance scores of sentences are calculated using the sum of both cosine similarities to the title vector and position method with given weighting value obtained through experiments).

Then, we evaluated system generated summaries using intrinsic evaluation method which evaluate the performance of the proposed system objectively and the linguistic quality (informativeness and coherence) subjectively for three extraction rate (10%, 20% and 30%).

The results of our objective evaluation have shown that the use of combined score (cosine similarity and position score) led to better performance than the use of a single score (only cosine similarity) for three extraction rates of (10%, 20% and 30%).

In addition to this, the results of our subjective experiment evaluation have also shown that EXP2 is more informative than EXP1 for the three extraction rates. This is due to the use of position method along with cosine similarity to the title vector value produced more informative summaries. The reordering of the extracted sentences in their original position also contributes for the system generated summary to be coherent. But unresolved references occurred in some system generated summaries.

The average performance of the proposed system increases with an increase in extraction rate, for both objective and subjective evaluation except that in 10% extraction rate for the objective evaluation. This can be due to the fact that most of the terms in the title of news articles are appered in the sentences found at beginning lines than the last ones[48]. Therefore these sentences have high chance to be incorporated in the final summary at 10% extraction rate than 20% and 30% extraction rates.

6.2 Recommendation

The study has shown that query based summarization can be done automatically for Afaan Oromo single document. However, further research and developmental effort is needed to apply these summarization approaches in a full-fledged automatic summarization system for Afaan Oromo documents. Additional features that can be added to increase the performance of the proposed summarization system and future research directions are outlined below:

- Even though an attempt was made to improve the coherence of the system extracts by re-ordering sentences to their original position in the text, still there is a room for further improvement by applying other appropriate NLP methods such as anaphora resolution.

- In query-oriented summarization the content of a query provides valuable clues for deciding which parts of the documents are important. Afaan Oromo thesaurus required to expand the query so that more salient sentences are included in the final summary.

- Standardized and well prepared Afaan Oromo text corpus is required to make a conclusive performance evaluation for further studies.

- Query based automatic summarization of multi documents for local languages are also another research direction for those interested researcher.

REFERENCES

[1] Jagadeesh J. et.al., "Query-Based Multi-Summarization Using Language Modeling", Master's Thesis, Department of Computer Science and Engineering, Gachibowli, Hyderabad, India, July 2006

[2] Elena Lloret, "Text summarization : an Overview", Department of Lenguajes y Sistemas Informaticos Universidad de Alicante Alicante, Spain,2006

[3] Jie Tangy et.al., "Multi-topic Based Query-Oriented Summarization", Department of Computer Science and Technology Tsinghua University, China, 2005

[4] Murali Krishna et.al., "A Hybrid Method for Query Based Automatic Summarization System", International Journal of Computer Applications (0975 – 8887) Volume 68– No.6, April 2013

[5] Mariana Damova and Ivan Koychev, "Query-Based Summarization: A survey", Faculty of Mathematics and Informatics, University of Sofia, Bulgaria,2010

[6] Samuel Eyassu and Björn Gambäck, "Classifying Amharic News Text Using Self-organizing Maps". In Proc. of ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan, Jun. 2005.

[7] Census Report. "Ethiopia's population now 76 million, 2008 available at : http://ethiopoltics.com/news

[8] Girma D., "Afan Oromo News Text Summarizer", Master's Thesis, Faculty of Informatics , Addis Ababa University, Ethiopia,2013

[9] Melese T., "Automatic Amharic Text Summarization Using Latent Semantic Analysis", Addis Ababa University school of graduate, October, 2009

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

[11] Mahmoud El-Haj et.al "Experimenting with Automatic Text Summarization for Arabic", University of Essex School of Computer Science and Electronic Engineering, 2010

[12] H.P. Luhn, "The Automatic Creation of Literature Abstracts", In IBM Journal of Research and Development, pp. 159-165, 1958.

[13] Vishal Gupta, “A Survey of Text Summarization Extractive Techniques", Journal Emerging Technology in Web Intelligence, Vol. 2, No. 3, August,2010

[14] Robin Cohen, “Advances in Artificial Intelligence” 15th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2002 Calgary, Canada, May 27-29, 2002Proceedings

[15] M. Saravanan, B. Ravindran, and S. Raman, ”Improving Legal Informatio Retrieval Using Ontological Framework”, communicated to International Journal of Artificial Intelligence and Law, Springer Verlag,2005.

[16] Anastasios Tombros and Mark Sanderson, “Advantages of Query Biased Summaries in Information Retrieval”. In Research and Development in Information Retrieval, pages 2-10, 1998.

[17] Dan Moldovan and Mihai Surdeanu, “On the Role of Information Retrieval and Information Extraction in Question Answering Systems”, Human Language Technology Research Institute, University of Texas at Dallas, 2003.

[18] Edmundson H. P(1969), “New Methods in Automatic Extracting,” Computing, vol. 16, no. 2, pp. 264-285.

[19] Hovy E. and Lin C., “Automated Text Summarization in SUMMARIST,” Advances in Automatic Text Summarization. MIT Press, 1999.

[20] Manning, C. “*An Introduction to Information Retrieval.”* Cambridge, UK: Cambridge University Press, 2007

[21] Laura Alonso, Irene Castellon and Salvador Climent, “Approach to text summarization”: Questions and answers

[22] Fuentes, M. et al. “QASUM-TALP at DUC 2005 Automatically Evaluated with a Pyramid

[23] Based Metric”. In the Document Understanding Workshop (presented at the Annual Meeting), Vancouver, B.C., Canada, 2005.

[24] početní techniky, FAV, ZČU – Západočeská, “Automatic Text Summarization (The state of the art 2007 and new challenges), Katedra informatiky a výpoèetní techniky, FAV, 2007

[25] Adam L. Berger and Vibhu O. Mittal. “Query-Relevant Summarization UsingFAQs”, Proceedings of Association for Computational Linguistics ACL, pages 294-301, 2000.

[26] John M. et.al. “CLASSY Query-Based Multi-Document Summarization”, Jade Goldstein Stewart, Department of Defense, 2005

[27] Saeedeh G. et.al “A Comprehensive Survey on Text Summarization Systems”, Islamic Azad University Islamic and Institute of Higher Education of Mashhad, 2009

[28] Ramakrishna Varadarajan and Vagelis Hristidis, “A System for Query-Specific Document Summarization”, Arlington, Virginia, USA, November 5–11, 2006.

[29] Ibrahim Imam, Nihal Nounou, et. al. “Query Based Arabic Text Summarization”, Dept. of CS, Arab Academy for Science, Tech. and Maritime Transport, Cairo, Egypt, 2013.

[30] Kamil N., “Automatic Amharic News Text Summarizer”, Master’s Thesis, Faculty of Informatics, Addis Ababa University, Addis Ababa,2005

[31] Teferi A., “The application of Machine learning Technique (NAÏVE BAYES) for Text Summarization the Case of Amharic News Text”, Master’s Thesis, Faculty of Informatics , Addis Ababa University, Ethiopia,2005.

[32] Helen A. (2006), “Automatic Text Summarization for Amharic Legal Judgments”, Master’s Thesis, Faculty of Informatics, Addis Ababa University. Addis Ababa.

[33] Addis, A. “Automatic Summarization for Amharic Text Using Open Text Summarizer.” Master Thesis. Addis Ababa, Ethiopia: Addis Ababa University, 2013

[34] Abebe Abeshu Diro, “AUTOMATIC MORPHOLOGICAL SYNTHESIZER FOR AFAAN OROMOO”, Master’s Thesis, Department of computer Science , Addis Ababa University, Ethiopia, June 2010.

[35] G.Q.A. Oromoo, Caasluga Afan Oromoo Jildi I” Komishinii Aadaaf Turizimii Oromiyaa, Finfinnee, Ethiopia,1995

[36] Mandefro Legesse Kejela, “Named Entity Recognition for Afan Oromo”, Master’s Thesis, Department of computer Science , Addis Ababa University, Ethiopia,2010.

[37] Gumii Qormaata Afan Oromoo, “ Caasluga Afan Oromo , Jildi I”, Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee, 1995

[38] Gezehagn Gutema Eggi, “Afaan Oromo Text Retrieval System”, Master’s Thesis, Department of computer Science , Addis Ababa University, Ethiopia,2012.

[39] Irresoo Nagi “Caassefama Afan oromo”, Kuraz International, Addis Ababa, 2006.

[40] Hssel, M. [45]. “Evaluation Automatic Text Summarization”: Apractical Implementation. Licentiate thesis,. Stockholm: Royal Institute of Technology, Department of Numerical Analysis and Computer Science,2007.

[41] Debela T., “Designing a Stemmer for Afan Oromo Text: A hybrid approach”, Master’s thesis, School of graduate studies, Addis Ababa University, Ethiopia, 2010.

[42] Welgama, W. “Atutomatic Text Summarization for Sinhala.”, Master thesis of Philosopy. University of Colombo, School of Computing.

[43] Anastasios Tombros and Mark Sanderson. “Advantages of Query Biased Summaries in Information Retrieval.” In Research and Development in Information Retrieval, pages 2{10, 1998.

[44] Carbonell, J., and J. Goldstein. 1998. “The use of MMR, diversity-based re ranking for reordering documents and producing summaries.” In Proc. ACM SIGIR.

[45] Gleb Sizov. “Extraction-Based Automatic Summarization.” Master of Science in Computer Science , Norwegian University of Science and Technology, June 2010

[46] Meyer, C. “On Improving Natural Language Processing through Phrase-based and one-to-one Syntactic Algorithm”, Msc. Thesis, Kansas State University Manhatan, Kansas, 2008

[47] Park S.C. “Generation of Non-redundant Summary Based on Sum of Similarity,” Computing, pp. 3-4,2004

[48] Kifle D. “Graph-based Automatic Text Summarizer for Amharic” Master’s thesis, School of graduate studies, Addis Ababa University, Ethiopia, 2014

Appendixes

Appendix I: Ideal summary construction guideline

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

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

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

Appendix II: Lists of document titles used as a query

Oduu1: Ayyaanni Caamsaa 20 waggaa 22ffaan kabajame

Oduu2: Hojiin qonna gannaa akkaataa karooraan raawwatamaa jiraachuu ibsame

Oduu3: Barattoonni dhaabbilee barnoota ol'aanoorraa ba'an hojii uummachuun jiruufi jireenyasaanii fooyyeffachaa jiru

Oduu4: Imaammanniifi Tarsiimoon qonnaa bu'aa qabatamaa fidaa jiraachuunsaa ibsame

Oduu5: Wareegamni waggaa 1ffaa ministira muummee duraanii kabajamoo obbo Mallas Zeenaawii yaadatame

Oduu6: Guyyaan Ayyaana Dargaggoota Oromiyaa 7ffaan dhaadannoo "Mul'ata Jaallee Mallas dhugoomsuun hirmaannaafi fayyadamummaa dargaggootaa ni mirkaneessina" jedhuun kabajame

Oduu7: Pirojektiin Daandii Asfaaltii Baalee eebbifame

Oduu 8: Qabeenya uumamaa kunuunsuun dhimma jireenya har'aafi eegreef

Oduu9: Oomishaafi Oomishtummaa qonnaa dabaluuf hudhaalee jiran haala hiikuun danda'amurratti leenjiin kennamaa jira

Oduu 10: Piroojektoonni bishaanii deggersa mootummaa Jaappaaniin ijaaraman ummataaf dabarfaman

Oduu11: Karoorri hojii bara 2005 milkaa'inaan raawwatamuunsaa ibsame.

Oduu12: Wareegamni waggaa 1ffaa ministira muummee duraanii kabajamoo obbo Mallas Zeenaawii yaadatame

Oduu13: Giddugalichi hawaasa fayyadamaa taasisaa jira

Oduu14: Piroojektoonni garaa garaa qarshii miliyoona 105 oliin hojjetamaa jiru

Oduu15: Imaammanniifi Tarsiimoon qonnaa bu'aa qabatamaa fidaa jiraachuunsaa ibsame

Oduu16: Sosochii finiinsa hawaasaa keessatti gaheen dubartootaa olaanaadha jedhame

Oduu17: Intarpiraayizichi kaayyoo dhaabbateef galmaan gahuuf raayyaa jijjiramaatiin socha'uu qaba jedhame

Oduu18: Kunuunsa qabeenya uumamaatiin bu'aan gaariin argamaa akka jiru ibsame

Oduu19: Leenjiin sosochii bulchiinsa gaarii marsaa 2ffaa gaggeefamaa jira

Oduu20: Qabeenya uumamaa kunuunsuun dhimma jireenya har'aafi eegreef

Oduu21:WALQOn ji'oottan jahan darbanitti qarshii mil. 415 ol liqii kenne

Oduu22:Rifoormii Waldaalee Hojii Gamtaarratti leenjiin kenname

Oduu23:Balaa tiraafikaa xiqqeessuuf meeshaaleen ammayyaa hojiirra oolan

Oduu24: Talaalliin dhukkuba manjalloo har'aa jalqabee akka kennamu ibsame

Oduu25: Qulqullina Barnootaa mirkaneessuuf qaamni hundi tumsuu qaba jedhame

Oomisha dammaarraa galiin qar bil 1.8 ol argame

Oduu26:Hawaasinni odeeffannoo mootummaa dhihoosaatti argachuun misooma keessaatti akka hirmatu taasisuun murteessaadha

Oduu27:Guyyaan ayyaana hoosiisee bulaa Oromiyaa 14ffaan kabajame.

Oduu28:Karoorri hojii bara 2005 milkaa'inaan raawwatamuunsaa ibsame.

Oduu29:Sassaabbin galii mootummaa fooyya'aa akka jiru ibsame

Oduu30:Warshaa huccuu AKPER ijaaruuf dhagaan bu'uraa kaawwame

Oduu31:Godina shawaa Kibba Lixaafi Bulchniisa magaalaa Sabbataatti galmeen filattootaa haala gaariin gaaggeeffamaa jira

Oduu32:Qabeenya Tuurizimii Naannichaa bu'aa barbaadameef akka oolu hojjetamaa jira

Oduu33:Du'aatii haadholeefi daa'immanii hir'isuuf ammas bal'inaan hojjetamuu qaba

Oduu34:Qulqullina barnootaa mirkaneessuuf barsiisota adda dureef leenjiin kennamaa jira

Oduu35:Namoonni horsiisa lukkuurratti bobba'an bu'a qabeessa ta'aa jiru

Oduu36:Namoonni waldaaleen gurma'anii hojjechaa jiran kaappitaala Qar.Mil.45 caalu horatan

Oduu37:Lafti heektaara kuma 20tti siqu jallisiin Misoomsamuufi

Oduu38: Naannichatti nageenyi amnasiisaafi itti fufiinsa qabu mirkanaa'aa jira

Oduu39:Aanaa Yaayyaa Gullaalleetti ummanni qabeenya uumammaa misoomsaa jirurraa fayyadamuu eegale

Oduu40: Irreechi Malkaa Ateetee kabajame

Appendix III: Afaan Oromo stop words

aanee	erga	irraa	isinirraa
agarsiisoo	ergii	irraan	isinitti
akka	f	isa	ittaanee
akkam	faallaa	isaa	itti
akkasumas	fagaatee	isaaf	itumallee
akkum	fi	isaan	ituu
akkuma	fullee	isaani	ituullee
ala	fuullee	isaanii	jala
alatti	gajjallaa	isaaniitiin	jara
alla	gama	isaanirraa	jechaan
amma	gararraa	isaanitti	jechoota
ammo	garas	isaatiin	jechuu
ammoo	garuu	isarraa	jechuun
an	giddu	isatti	kan
ana	gidduu	isee	kana
ani	gubbaa	iseen	kanaa
ati	ha	ishee	kanaaf
bira	hamma	ishii	kanaafi
booda	hanga	ishiif	kanaafi
booddee	henna	ishiin	kanaafuu
dabalatees	hoggaa	ishiirraa	kanaan
dhaan	hogguu	ishiitti	kanaatti
dudduuba	hoo	ishiitti	karaa
dugda	hoo	isii	kee
dura	illee	isiin	keenna
duuba	immoo	isin	keenya
eega	ini	isini	keessa
eegana	innaa	isinii	keessan
eegasii	inni	isiniif	keessatti
ennaa	irra	isiniin	kiyya

koo
kun
lafa
lama
malee
manna
maqaa
moo
na
naa
naaf
naan
naannoo
narraa
natti
nu
nu'i
nurraa
nuti
nutti
nuu
nuuf
nuun
nuy
odoo
ofii
oggaa
oo
osoo
otoo
otumallee
otuu
otuullee
saaniif
sadii
sana
saniif
si
sii
siif
siin
silaa
silaa
simmoo
sinitti
siqee
sirraa
sitti
sun
tahullee
tana
tanaaf
tanaafi
tanaafuu
ta'ullee
ta'uyyu
ta'uyyuu
tawullee
teenya
teessan
tiyya
too
tti
utuu
waa'ee
waan
waggaa
wajjin
warra
woo
yammuu
yemmuu
yeroo
yommii
yommuu
yoo
yookaan
yookiin
yoolinimoo

Appendix IV: Afaan Oromo affixes

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

Appendix V: Afaan Oromo abbreviations

k.k.f	Kan kana fakkaatan
Obb.	Obboo
Add.	Addee
Bili.	Biliyoona
fkn.	Fakkeenyaaf
hub.	Bubaachiisaa
w.k.f	Waan kana fakkaatan
mil.	Miliyoona
ful.	Fulbaana
Sad.	Sadaasa
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
Bil.	Bilbila
G/L/Sha	Godina Lixa shawaa
G/ad/F	Godina addaa Finfinnee

Appendix VI: Afaan Oromo compound word lists & their normalized form

Mana barumsaa

Mana kitabaa

godina Shwaa Kibba Lixaa

aanaa Goorootti

Manni Murtii

Iluu Abbaa Booraa

mana mootummaa

Caffeen Oromiyaa

Biiroon Qonna Oromiyaa

Qotee Bulaa

nyaata horii

bona dabree

biqiltuu bunaa

Godina Arsii

Dhaabbilee Barnootaa ol'aanaa

muummee Saayinsii

Muummee Karooraafi Ooggansa barnootaa

Ministira Muummee

farra hiyyummaa

Appendix VII: Objective evaluation result of EP1

Extraction rate in %									
Doc No.	**10%**			**20%**			**30%**		
	R	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**
Doc 1	50%	50%	50%	60%	60%	60%	37%	37%	37%
Doc2	50%	50%	50%	57%	57%	57%	80%	80%	60%
Doc3	50%	50%	50%	50%	50%	50%	75%	75%	33%
Doc4	50%	50%	50%	50%	50%	50%	33%	33%	33%
Doc5	50%	50%	50%	80%	80%	80%	75%	75%	57%
Doc6	50%	50%	50%	50%	50%	50%	50%	50%	50%
Doc7	50%	50%	50%	60%	60%	60%	96%	96%	96%
Doc8	60%	60%	60%	60%	60%	60%	67%	67%	67%
Doc9	100%	100%	100%	50%	50%	50%	66%	66%	66%
Doc10	33%	33%	33%	80%	80%	80%	92%	92%	92%
Doc11	50%	50%	50%	75%	75%	75%	75%	75%	75%
Doc12	100%	100%	100%	100%	100%	100%	75%	75%	75%
Doc13	50	50%	50%	75%	75%	75%	75%	75%	75%
Doc14	0%	0%	-	67%	67%	67%	80%	80%	80%
Doc15	50%	50%	50%	82%	82%	82%	82%	82%	82%
Doc16	0%	0%	-	67%	67%	67%	50%	50%	50%
Doc17	100%	100%	100%	50%	100%	100%	100%	100%	100%
Doc18	75%	75%	75%	60%	60%	60%	82%	82%	82%
Doc19	50%	50%	50%	66%	66%	66%	82%	82%	82%
Doc20	80%	80%	80%	44%	44%	44%	71%	71%	71%
Doc21	82%	82%	82%	67%	67%	67%	75%	75%	75%
Doc22	100%	100%	100%	100%	100%	100%	96%	96%	96%
Doc23	50%	50%	50%	50%	50%	50%	67%	67%	67%
Doc24	91%	91%	91%	80%	80%	80%	92%	92%	92%
Doc25	0%	0%	-	50%	50%	50%	50%	50%	50%
Doc26	50%	50%	50%	67%	67%	67%	71%	71%	71%
Doc27	50%	50%	50%	80%	80%	80%	71%	71%	71%
Doc28	100%	100%	100%	100%	100%	100%	66%	66%	66%
Doc29	50%	50%	50%	80%	80%	80%	90%	90%	90%
Doc30	50%	50%	50%	75%	75%	75%	67%	67%	67%
Doc31	67%	67%	67%	80%	80%	80%	90%	90%	90%
Doc32	50%	50%	50%	50%	50%	50%	75%	75%	75%
Doc33	100%	100%	100%	100%	100%	100%	91%	91%	91%
Doc34	0%	0%	-	67%	67%	67%	75%	75%	75%
Doc35	100%	100%	100%	80%	80%	80%	94%	94%	94%
Doc36	0%	0%	-	67%	67%	67%	100%	100%	100%
Doc37	100%	100%	100%	100%	100%	100%	50%	50%	50%
Doc38	100%	100%	100%	60%	60%	60%	62%	62%	62%
Doc39	100%	100%	100%	66%	66%	66%	80%	80%	80%
Doc40	50%	50%	50%	50%	50%	50%	76%	76%	76%
Average	**88%**	**88%**	**88%**	**70%**	**70%**	**70%**	**73%**	**73%**	**73%**

Appendix VIII: Objective evaluation result of EP2

Extraction rate in %									
Doc No.	**10%**			**20%**			**30%**		
	R	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**
Doc 1	100%	100%	100%	50%	50%	50%	75%	75%	75%
Doc2	66%	66%	66%	60%	60%	60%	80%	80%	80%
Doc3	100%	100%	100%	75%	75%	75%	75%	75%	75%
Doc4	50%	50%	50%	75%	75%	75%	84%	84%	84%
Doc5	40%	40%	40%	50%	50%	50%	46%	46%	46%
Doc6	67%	67%	67%	67%	67%	67%	75%	75%	75%
Doc7	50%	50%	50%	67%	67%	67%	50%	50%	50%
Doc8	67%	67%	67%	71%	71%	71%	71%	71%	71%
Doc9	100%	100%	100%	67%	67%	67%	80%	80%	80%
Doc10	100%	100%	100%	75%	75%	75%	67%	67%	67%
Doc11	50%	50%	50%	67%	67%	67%	75%	75%	75%
Doc12	100%	100%	100%	100%	100%	100%	100%	100%	100%
Doc13	50%	50%	50%	80%	80%	80%	92%	92%	92%
Doc14	50%	50%	50%	85%	85%	85%	94%	94%	94%
Doc15	100%	100%	100%	80%	80%	80%	90%	90%	90%
Doc16	87%	87%	87%	100%	100%	100%	90%	90%	90%
Doc17	67%	67%	67%	78%	78%	78%	70%	70%	70%
Doc18	100%	100%	100%	75%	75%	75%	75%	75%	75%
Doc19	100%	100%	100%	80%	80%	80%	80%	80%	80%
Doc20	75%	75%	75%	75%	75%	75%	80%	80%	80%
Doc21	100%	100%	100%	78%	78%	78%	91%	91%	91%
Doc22	75%	75%	75%	75%	75%	75%	90%	90%	90%
Doc23	100%	100%	100%	75%	75%	75%	75%	75%	75%
Doc24	100%	100%	100%	100%	100%	100%	100%	100%	100%
Doc25	100%	100%	100%	100%	100%	100%	100%	100%	100%
Doc26	100%	100%	100%	100%	100%	100%	93%	93%	93%
Doc27	75%	75%	75%	80%	80%	80%	85%	85%	85%
Doc28	88%	88%	88%	80%	80%	80%	80%	80%	80%
Doc29	91%	91%	91%	67%	67%	67%	84%	84%	84%
Doc30	75%	75%	75%	67%	67%	67%	67%	67%	67%
Doc31	50%	50%	50%	67%	67%	67%	87%	87%	87%
Doc32	100%	100%	100%	75%	75%	75%	86%	86%	86%
Doc33	75%	75%	75%	80%	80%	80%	80%	80%	80%
Doc34	67%	67%	67%	87%	87%	87%	95%	95%	95%
Doc35	91%	91%	91%	91%	91%	91%	91%	91%	91%
Doc36	100%	100%	100%	87%	87%	87%	91%	91%	91%
Doc37	100%	100%	100%	100%	100%	100%	100%	100%	100%
Doc38	100%	100%	100%	75%	75%	75%	75%	75%	75%
Doc39	75%	75%	75%	75%	75%	75%	78%	78%	78%
Doc40	100%	100%	100%	67%	67%	67%	82%	82%	82%
Average	**91%**	**91%**	**91%**	**81%**	**81%**	**81%**	**75%**	**75%**	**75%**

Appendix IX: Subjective Evaluation Results of EXP1 and EXP2

Doc ID	Extraction Rate (%)	Experiments	Ev 1	Ev2	Ev3	Ev4	Ev5	Total %
Doc1	**10**	Exp1	1	2	2	1	2	**40%**
		Exp2	1	3	3	2	2	**55%**
Doc2	**10**	Exp1	3	3	2	3	2	**65%**
		Exp2	3	3	2	3	2	**65%**
Doc3	**10**	Exp1	2	3	2	2	1	**50%**
		Exp2	2	3	2	2	1	**50%**
Doc4	**10**	Exp1	3	3	2	1	1	**50%**
		Exp2	2	3	3	2	2	**60%**
Doc5	**10**	Exp1	2	3	3	2	3	**70%**
		Exp2	2	3	2	3	3	**65%**
Doc6	**10**	Exp1	3	2	2	3	3	**65%**
		Exp2	3	3	3	2	4	**75%**
Doc7	**10**	Exp1	2	3	4	3	2	**65%**
		Exp2	3	2	3	4	4	**80%**
Doc8	**10**	Exp1	3	3	3	3	3	**75%**
		Exp2	4	3	3	4	2	**80%**
Doc7	**10**	Exp1	3	3	2	2	3	**65%**
		Exp2	3	4	3	3	3	**80%**
Doc8	**10**	Exp1	2	3	3	3	2	**65%**
		Exp2	3	2	3	3	3	**70%**
Doc9	**10**	Exp1	2	3	3	3	3	**70%**
		Exp2	3	4	4	3	3	**85%**
Doc10	**10**	Exp1	3	3	4	4	2	**80%**
		Exp2	3	3	3	4	4	**85%**
Doc11	**10**	Exp1	4	4	3	3	3	**85%**
		Exp2	4	4	4	4	4	**100%**
Doc12	**10**	Exp1	3	3	3	3	3	**75%**
		Exp2	3	4	4	4	3	**90%**
Doc13	**10**	Exp1	3	4	4	4	4	**95%**
		Exp2	4	3	3	4	3	**85%**
Doc14	**20**	Exp1	1	1	2	2	3	**45%**
		Exp2	2	3	1	2	3	**55%**
Doc15	**20**	Exp1	2	3	3	1	2	**55%**
		Exp2	3	3	3	3	3	**70%**
Doc16	**20**	Exp1	3	2	3	1	2	**55%**
		Exp2	3	2	3	1	2	**55%**
Doc17	**20**	Exp1	3	4	3	2	3	**75%**
		Exp2	3	3	3	2	3	**70%**
Doc18	**20**	Exp1	3	2	3	3	3	**70%**
		Exp2	4	4	3	3	3	**85%**
Doc19	**20**	Exp1	1	2	2	1	2	**40%**
		Exp2	1	3	3	2	2	**55%**
Doc20	**20**	Exp1	3	3	3	3	1	**65%**

		Exp2	3	3	2	3	2	**70%**
Doc21	**20**	Exp1	2	3	2	2	1	**50%**
		Exp2	2	3	2	2	1	**50%**
Doc22	**20**	Exp1	3	3	2	1	1	**50%**
		Exp2	2	3	3	2	3	**70%**
Doc23	**20**	Exp1	2	3	3	3	3	**70%**
		Exp2	2	3	2	3	3	**65%**
Doc24	**20**	Exp1	3	2	2	3	3	**65%**
		Exp2	3	3	3	2	2	**65%**
Doc25	**20**	Exp1	2	3	4	3	2	**70%**
		Exp2	3	2	3	4	4	**80%**
Doc26	**20**	Exp1	3	3	3	3	3	**75%**
		Exp2	4	3	3	4	2	**80%**
Doc27	**30**	Exp1	3	3	2	2	3	**65%**
		Exp2	3	4	3	3	3	**80%**
Doc28	**30**	Exp1	2	3	3	3	2	**65%**
		Exp2	3	2	3	3	3	**70%**
Doc29	**30**	Exp1	2	3	3	3	3	**70%**
		Exp2	3	4	4	3	3	**85%**
Doc30	**30**	Exp1	3	3	4	4	2	**80%**
		Exp2	3	3	3	4	4	**85%**
Doc31	**30**	Exp1	4	4	3	3	3	**85%**
		Exp2	4	4	4	4	4	**100%**
Doc32	**30**	Exp1	3	3	3	3	3	**75%**
		Exp2	3	4	4	4	3	**90%**
Doc33	**30**	Exp1	3	4	4	4	4	**95%**
		Exp2	4	3	3	4	3	**85%**
Doc34	**30**	Exp1	3	3	3	3	3	**75%**
		Exp2	3	3	3	3	3	**75%**
Doc35	**30**	Exp1	2	1	2	3	3	**55%**
		Exp2	2	4	4	3	4	**85%**
Doc36	**30**	Exp1	3	4	3	2	2	**70%**
		Exp2	3	4	3	2	2	**75%**
Doc37	**30**	Exp1	3	3	4	3	2	**75%**
		Exp2	3	3	4	4	3	**85%**
Doc38	**30**	Exp1	3	3	3	3	3	**75%**
		Exp2	3	4	4	4	4	**95%**
Doc39	**30**	Exp1	2	3	2	3	3	**65%**
		Exp2	2	4	3	3	2	**70%**
Doc40	**30**	Exp1	4	3	3	2	3	**75%**
		Exp2	4	4	3	4	4	**95%**

Appendix X: Sample Summary

Doc No: 8

Source: Kalachaa Oromiyaa

Title (Query): Imaammanniifi Tarsiimoon qonnaa bu'aa qabatamaa fidaa jiraachuunsaa ibsame

Imaammanniifi Tarsiimoon misooma bahee hojiirra oolaa jiru bu'aa qabatamaa fidaa jiraachuunsaa Ministirri Muummee Obbo Hayilamaaraam Dassaaleny ibsan.Jilli ooggantoota olaanoo Mootummaa Federaalaa fi Naannoo Oromiyaa Ministira Muummee obbo Haylamaaraam Dassaalenyiin durfame godina Arsiitti oyiruu qonnaan bultoota aanaalee Diiggaluufi xiyyoo fi muuneessaa qamadii teekinooloojii BBM fi sararaan facaasuun milkaa'anii daawwateera. Qonnaan bultoota oyiruun isaanii jila kanaan daawwataman keessaa qonnaan bulaa aanaa diigaluufi xiyyoo ganda Qacamaa Murqichaa kan ta'e Abiyoot Wuubee isa tokko. Qonnaan bulaan kun teekinooloojii BBM fayyadamee hektaara lafa walakkaa kana dura calla kuntaalli 15rra darbe argachuu hindandeenyerraa bara kana calla harka dacha oliin argachuuf abdateera. Qonnaan bulaan Tasfaayee Abbabaatiis Jiraatuma ganda kanaa yoo ta'u, kanaan dura osoo hubannoo hin argatiin teekinooloojii qonnaatti fayyadamee akka hin beeknee fi oomisha gadi aanaa argachaa akka ture hin dhoksine. Wayita leenjiin TMG kennamu, irratt hirmaachu dhaan godinuma kanatti bara darbe lafa heektaara tokkorraa caallaan qamadii kuntaala 101 argamuusaa xiyyeeffannoo itti kennuun hubate. Bara oomishaa 2005/2006tti muuxannoo kana hojiirra oolchee bu'aa kana caalu argachuufis onnatee ka'e. Ijaarsa raayyaa misoomaa tokko-shaneen ijaaramee, hojjetaa dirree misoomaatiin deeggaramee oomisha guddaa argachuuf qamadii lafa heektaara tokkorratti facaase. Lafti kuni kana dura calla kuntaalli 20 caalu irraa akka hin argamne dubbata, Bara kana garuu bu'aa olaanaa bara darbe godinuma kanatti galmaa'ee ture kan callu kuntaala 120 argachuuf abdii dhugoomurra gaheera. Jila kana kan hoogganan Ministirri Muummee Obbo Hayilamaaraam Dassaaleny yaada qonnaan bultootaa kana erga dhaggeeffataniifi oomisha oyiruurra jirus erga daawwatanii booda haasawa taasisaniin, Immaammannii fi Tarsiimoon qonnaan durfamu bu'aa qabatamaa fiduunsaa mirkanaa'eera jedhan. Biyyi teenya kaleessa qamadii kadhachaa kan turte, har'a seenaan suni jijjiiramuun qonnaan bulaan keenya kan nugaafatu qamadiin akka dhiyaatuuf osoo hin taane gabaa oomishaasaa itti gurguratu ta'eera jedhan Obbo Hayilaamaaram. Lafa ciccitaarratti oomisha

oomishamuun jijjiirama guddaa fiduun hindanda'amu Ilaalcha jedhus qonnaan bultoonni kunneen qabatamaahdaan kan fashaleessan t'uu mirkanaa'eera. Lafa hektaara tokkorraa calla kuntaalli 101 bara darbe argame suni dhaabbata qorannoo fi qo'annoo qonnaa keessatti kan eegamu malee maasii qonnaan bulaarratti kan eegamu hin turre. Hara'a garuu qonnaan bultoonni teekinooloojii qonnaa oomishasaanii dabalu uumuu fi fayyadamuutti ce'anii heektara tokkoorraa calla hanga kuntaala 120 argachuun dhaabbata qorannoo fi qo'annoo qonnaa caaluu danda'aniiru jedhan Obbo Hayilamaaraam. Ammaan booda qonnaan bultoonni biyya keenyaa hundi muuxannoo kana walirraa fudhachuun hojiirra yoo oolchan biyyi keenya qamadii biyya alaatti kan galchitu osoo hin taane alatti kan ergitu akka taatus ibsaniiru. Sadarkaa itti aanaa Pirezidaantiitti Oogganaan Biiroo Qonnaa oromiyaa Obbo Zalaalem Jamaanee gamasaatiin karoora guddinaa fi tiraansfoormeeshinii bu'uura godhachuun hojii hojjetamaa tureen oromiyaa keessatti oomishaafi oomishitummaan sadarkaa olaanaadhaan guddachaa dhufeera.Godinaalee Oromiyaa oomsishaa fi oomishtummaa dachaan dabaluun milkaa'inni keessatti galmaa'ee keessaa Godinni Arsii isa tokkodha.Muuxannoo asitti argame kana naannoo keenya hundaa fi naannoolee olla keessatti babal'isuuf daawwannaan kuni gahee guddaa akka qabus- ibsaniiru. Godinichatti teekinooloojii qonnaa guututti fayyadamuun harka 70n hojiirra akka oolee fi oomisha bara 2005/2006 kanarraas callaa kuntaalli miliyoona 19 fi kuma 700 ol akka argamu kan himan ammoo Bulchaa Godina Arsii Obbo Saadaat Nashaati. Akka ibsa Obbo Saadaatitti, godinichatti lafti hektaarri 599, 401,kan misoome yoo ta'u, calla guddistuu xaa'oo kuntaalli 344,974 fi sanyi filatamaa kuntaalli 152,939 fayyadamuun danda'ameera. Bara kana oomisha akka biyyatti harka 20n dabaluuf qabamee keessaa godinichatti harka 31n dabaluun karoorfamee %32n dabaluun dandameeraas jedhan.

Summary with 30% extraction rate

Imaammanniifi tarsiimoon misooma bahee hojiirra oolaa jiru bu'aa qabatamaa fidaa jiraachuunsaa ministirri muummee obbo hayilamaaraam dassaaleny ibsan. bara oomishaa 2005/2006tti muuxannoo kana hojiirra oolchee bu'aa kana caalu argachuufis onnatee ka'e. lafti kuni kana dura calla kuntaalli 20 caalu irraa akka hin argamne dubbata, bara kana garuu bu'aa olaanaa bara darbe godinuma kanatti galmaa'ee ture kan callu kuntaala 120 argachuuf abdii dhugoomurra gaheera. jila kana kan hoogganan ministirri muummee obbo hayilamaaraam dassaaleny yaada qonnaan bultootaa kana erga dhaggeeffataniifi oomisha oyiruurra jirus erga

daawwatanii booda haasawa taasisaniin, immaammannii fi tarsiimoon qonnaan durfamu bu'aa qabatamaa fiduunsaa mirkanaa'eera jedhan. lafa ciccitaarratti oomisha oomishamuun jijjiirama guddaa fiduun hindanda'amu ilaalcha jedhus qonnaan bultoonni kunneen qabatamaahdaan kan fashaleessan t'uu mirkanaa'eera. muuxannoo asitti argame kana naannoo keenya hundaa fi naannoolee olla keessatti babal'isuuf daawwannaan kuni gahee guddaa akka qabus- ibsaniiru.

Summary with 20% extraction rate

Imaammanniifi tarsiimoon misooma bahee hojiirra oolaa jiru bu'aa qabatamaa fidaa jiraachuunsaa ministirri muummee obbo hayilamaaraam dassaaleny ibsan. bara oomishaa 2005/2006tti muuxannoo kana hojiirra oolchee bu'aa kana caalu argachuufis onnatee ka'e.

jila kana kan hoogganan ministirri muummee obbo hayilamaaraam dassaaleny yaada qonnaan bultootaa kana erga dhaggeeffataniifi oomisha oyiruurra jirus erga daawwatanii booda haasawa taasisaniin, immaammannii fi tarsiimoon qonnaan durfamu bu'aa qabatamaa fiduunsaa mirkanaa'eera jedhan. lafa ciccitaarratti oomisha oomishamuun jijjiirama guddaa fiduun hindanda'amu ilaalcha jedhus qonnaan bultoonni kunneen qabatamaahdaan kan fashaleessan t'uu mirkanaa'eera.

Summary with 10% extraction rate

Imaammanniifi tarsiimoon misooma bahee hojiirra oolaa jiru bu'aa qabatamaa fidaa jiraachuunsaa ministirri muummee obbo hayilamaaraam dassaaleny ibsan. jila kana kan hoogganan ministirri muummee obbo hayilamaaraam dassaaleny yaada qonnaan bultootaa kana erga dhaggeeffataniifi oomisha oyiruurra jirus erga daawwatanii booda haasawa taasisaniin, immaammannii fi tarsiimoon qonnaan durfamu bu'aa qabatamaa fiduunsaa mirkanaa'eera jedhan.