Addis Ababa University
Faculty of Natural Sciences
Department of Computer Science

Afaan Oromo Automatic News Text Summarizer Based on Sentence Selection Function

BY:
Fiseha Berhanu Tesema

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

November 2013

Addis Ababa University
Faculty of Natural Sciences
Department of Computer Science

Afaan Oromo Automatic News Text Summarizer Based on Sentence Selection Function

BY:
Fiseha Berhanu Tesema

Signature of the Board of Examiners for Approval

	Name	Signature
1.	Sebsibe HaileMariam, PhD (Advisor)	_____
2.	Dejene Ejigu, (PhD) (Examiner)	_____
3.	_____	_____

Dedicated to

My dad, Berhanu Tesema Guta and my mom, Askale Gobena Wirtu

Acknowledgment

First of all, I would like to thank the Almighty God for His help in all aspects of my life. Then, there are a number of people who had helped me directly or indirectly in the development of my thesis. On the top of these people is **Dr. Sebsibe HaileMariam (PhD)**, who was my advisor and whose advice and support help me to shape this thesis and pass through all ordeals. Many thanks!

Girma Debele is the other person who deserves my heartfelt gratitude, his kindness and support in giving me the all document and systems without any hesitation. Dilla University Afaan Oromo department, Adama University Afaan Oromo department, **Tadese (PhD Candidate at AAU, Specializing in linguistic and culture)**, **Dr. Berhanu (PhD)** and **Tsegaye (MA, Literature, English Department**, Dilla university), your contributions to this work is great. **Dr. Tadese J. (PhD),** thank you for your help in editing my document. I also pay big thank to all my respondents.

Dad and mom, I dedicate this work to you with great pleasure and honor. My Dad, **Berhanu Tesema**, thank you for your astonishing follow-up since my childhood. I appreciate your advice through which you planted in my mind the passion for education and knowledge. My Mom **Askale Gobena**, I do not know how I can express your endless love, which lets you drop out everything to look after and educate your children.

My wife **Hawi**, I thank you very much as you let me focus on my study by caring many burdens that only you and I know. My **Son Falmi**, who was born during the development of this study, welcome to new world and I wish for you a bright future.

Mullu, I thank you very much for you serve me without loss of patience and being tedious to come in the morning and cook for me a meal every day. Finally, my thank goes to my friends who had directly or indirectly contributed in development of this thesis, without which it would not have reached this status.

Table of Contents

Content **Page**

Chapter Two: Review of Related Literature

List of Figures

Figures **Page**

List of Tables

Tables **Page**

Acronyms and Abbreviation

AO	Afaan Oromo
AOTS	Afaan Oromo Text Summarizer
AI	Artificial Inelegancy
Cat. 1	Category 1
Cat. 2	Category 2
Cat. 3	Category 3
Cat. 4	Category 4
C	Centrality
CR	Compression Ratio
DSOP	Descriptive sentence of all Paragraphs
PhD	Doctor of philosophy
Exp.	Experiment
Fn	False Negative
Fp	False Positive
Fm	F-measure
FSBP	Frist Sentence of Body of Paragraphs
FSFP	Frist Sentence of Frist Paragraphs
FSLP	Frist Sentence of Last Paragraphs
HTML	Hyper Text Transfer Protocol
IDE	Integrated Development Environment
KP	Key phrase
KB	knowledge Base
LSA	Latent Semantic Analysis
NLP	Natural Language processing
OOP	Object Oriented Programming
OOTS	Open Oromo Text Summarizer
OTS	Open Text Summarizer
P1	Paragraph 1
P	Precision

PRM	Pronominal Resolution Module
R	Recall
ROUGE	Recall Oriented Understudy for Gisting Evaluation
SDSOP	Selected Descriptive sentence of all Paragraphs
SFSBP	Selected Frist Sentence of Body of Paragraphs
SFSFP	Selected Frist Sentence of Last Paragraphs
SCR	Sentence Compression Ratio
SP	Sentence Position
SS	Sentence Score
SWBCP	Sentence Weighting Based on Cue Phrase
SWBEv	Sentence Weighting Based on Events
SWBKW	Sentence Weighting Based on Key Word
SWBNum	Sentence Weighting Based on Numbers
SWBSP	Sentence Weighting Based on Sentence Position
Subj 1	Subject one
SSG	Summary Sentence Generator
TF	Term Frequency
Tn	True Negative
TP	True Positive
WCup	Weight of Cue Phrase
WEv	Weight of Event
WKw	Weight of Keyword
Wmum	Weight of number
WSp	Weight of Sentence Position
Wo Ev	Without Event
Wo Num	Without Number
Wo SL	Without Sentence length
Wo Stem & LSL	Without stemmer and Language specific Lexicon

Abstract

The existence of the World Wide Web and advancement in digital device has caused an information explosion. Readers are overloaded with lengthy text where a shorter version would suffice. This abundance of information needs efficient tools to handle. Automatic text summarizer is one of the various tools used for the purpose of shortening lengthy documents, and alleviating the type of problem.

This work focuses on developing efficient extractive Afaan Oromo automatic news text summarizer, through systematic integration of features: sentence position, keyword frequency, cue phrase, sentence length handler, occurrence of numbers and events like: - time, date and month in sentences. The data that aids for the system development are like: abbreviation, synonym, stop word, suffix, numbers, and name of: (time, date and month) collected from both secondary and primary sources. In addition, 350 English cue phrases are collected and translated to 729 Afaan Oromo cue phrases. For validation and testing 33 different newspaper topics are collected, of these, 20 of them have been used for validation while the rest 13 employed for testing purpose. The Total numbers of respondents who have participated in the validation ad testing data corpus preparation are 110. Besides, Open text summarizer C# version open source has been selected as a tool to develop the system

The system has been evaluated based on seven experimental scenarios and evaluation is made both subjectively and objectively. The subjective evaluation focuses on evaluation of the structure of the summary like referential integrity and non-redundancy, coherence and informativeness of the summary. The objective evaluation uses metrics like precision, recall and F-measure for evaluation. The result of subjective evaluation is 88% informativeness, 75% referential integrity and non-redundancy, and 68% coherence. Because of the added features, different techniques and experiment applied to this work the system gave 87.47%fm and outperform by 26.95% than the previous work.

Keywords: *Afaan Oromo, Automatic news text summarizer, Cue Phrase, Sentence Selection Function*

Chapter One: Introduction

1.1 Background

According to Edward Hovy [64] a summary is 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). McKeown, K. R et al [25] also defined a summary as a text that produced from one or more texts that convey important information in the original text(s), and that is not longer than half of the original text(s) and usually significantly less than that. Hence, from the definitions given, one can draw that summary of a text is half minimized of the original text which consists only the main points of the text.

"Currently, the world is all about information, most of it is in digital format. The World Wide Web contains billions of documents and it is growing at an exponential pace."[1] The contributing factors are the availability and accessibility of rapid growth of broadcast systems, Internet and online information services [27]. Obviously, this explosion of information has caused a well-recognized information overload problem. For this matter, it is obvious that there is no time to read everything to make critical decisions based on available information. Therefore, for this problem there must be tools that provide timely access to, and digest of, various sources are necessary in order to alleviate the information overload people are facing.

Similarly, in the book entitled *"Advances in Automatic Text Summarization"* Inderjeet Mani and Mark T. Maybury [36] point out the fact that more and more information has become available in the last decades requires tools to handle it. These concerns have sparked interest in the development of automatic summarization systems.

To this end, they defined automatic summarizer as a system, which designed to take a voice, single article, a cluster of news articles, a broadcast news show, or an email thread, as input and produce a concise, and fluent summary of the most important information.

Furthermore, among different tools, automatic text summarizer is a tool used to identify and select the central content or user inquired content from the given original texts to form the summarized output. The output sentence might be identical to the original input or newly generated sentences [20].

According to Mani et al [18] the goal of a summary is to give the reader an accurate and complete idea of the contents of the source. Hence, this simple definition captures three important aspects that characterize research on automatic summarization:

- Summaries may be produced from a single document or multiple documents,
- Summaries should preserve important information,
- Summaries should be short.

History evidenced that there was no research carried out on automatic text summarization until 1958 by Luhn's [49]. Although, researches in the late 1950s and early 1960s suggested that computer based text summarization was feasible, though not straightforward [7] [22]. However, the progress in language processing, coupled with great rises of computer memory and speed, and the growing presence of online text-in corpora and especially on the web-renewed lay ground to the need and study automatic text summarization. After a decade, from the 70s to the early 80s, Artificial Intelligence (AI) takes over the classical text summarization system (i.e. an approach that extracts the main sentence using statistical analysis) system. AI employs knowledge representations, such as frames or templates, to identify conceptual structures of a text and find salient concepts by inference [3][28].

However, the main drawback of AI is that limited templates make conceptual structures incomplete. Therefore, since early 90s, Information Retrieval (IR) has been introduced [2] [14] [18] which is used as one of the best approach in dealing with automatic text summarization. IR considers text summarization as how to find significant sentences in a document. IR technique in text summarization focuses on symbolic-level analysis; while it does not consider semantic issues. For this reason, many researchers' deals with automatic text summarization prefer this approach.

As discussed by Girma [10], the pioneer of Afaan Oromo automatic news text summarizer in Ethiopia, there are few researches conducted in automatic text summarization in Ethiopian languages, particularly in Amharic text at various domains by adopting different techniques. Regarding to Afaan Oromo, there were no study conducted on automatic news text summarization before.

Thus, based on the forth-mentioned work on automatic news text summarization, this study attempted to fill the gaps identified and develop robust Afaan Oromo news text summarizer using different techniques, and algorithm proposed so as to come up with a quality summarizer.

1.2 Motivation

The population of Oromo is around 40 million in Ethiopia and 3rd largest single nationality group in Africa [9] [10] [15]. The Oromo nation has a single common mother tongue, called the Oromo language or Afaan Oromo or Oromiffa. It is the third most-widely spoken language in Africa as a mother tongue, next to Hausa and Arabic. Today, Afaan Oromo is serving as an official language of Oromia regional state (which is the largest regional state among the current federal states of Ethiopia). Being an official language, it is uses as medium of instruction for primary and junior secondary schools of the region [20]. It is also a field of specialization at Diploma, Bachelor Degree, and Master's Degree levels at various universities in Ethiopia.

Beside this, a number of: literature works, newspapers, magazines, education resources, official credentials and religious documents are published and available in the language [10]. Hence, above all these facts initiate us to conduct this work.

On the other hand, in his work, Sisay Adugna [35] presented that Afaan Oromo is one of the most resource scarce languages in context of NLP. Even though the language is spoken and serves as an official language for more than 40 million people, there was only one study that was conducted last year by Girma [10]. This is another motivating factor to conduct this research.

The other factor is that today the improvement in modern technology raises the availability of digital information on the Internet, which is written by Afaan Oromo. Due to this, two basic problems are encountered, searching for relevant documents from an overwhelming number of documents, and absorbing a large quantity of relevant information from these abundance documents. Finally, the knowledge gap found in Girma's [10] work also initiate us.

In general, lack of active research on the automatic text summarization and a dramatic growth of electronic document from time to time are a motivating factor for this work to come up with modules that can alleviate or minimize these problems.

1.3 Statement of the problem and justification of the study

Afaan Oromo is one of the most widely used languages in Ethiopia, and also spoken in Kenya, Somalia, and Djibouti. Nowadays, textual information is highly increasing from time to time both in softcopy and hardcopy format. The textual information disseminated to the populations located at various countries is provided by various sources, such as; Internet, media, presses, books, journal articles, newspaper, etc.

Specially, digitally information, it is available in abundance and in a myriad of forms to an extent of making it near impossible to search manually, sift and choose relevant information.

Therefore, this information must instead be filtered, and extracted to avoid drowning in it. Otherwise, the users of the language forced to spend more of their energy, resource and time by reading or processing unnecessary information.

Therefore, in order to alleviate this problem automatic text summarization in which computers automatically create an abstract, or summary, is an answer. Hence, automatic Afaan Oromo news text summarizer, especially for large amount of news releases by newspapers and online news agencies, could then be justified, as it is very essential to save the readers' time, space, energy and resources [10]. Consequently, this study answers these problems by providing an efficient automatic Afaan Oromo news text summarizer that can handle the abundance with short summary.

Despite these general problems mentioned above, this work is based on research gap shown in previous work. Girma [10] tried to come up with automatic news text summarizer, using extractive method of generating the news text summary. He used only two features term frequency and sentence position to extract the sentence. However, the work has many research gaps, which include:

- lack of scientific justification during weighting of sentence based on the two features stated in his work.
- lack of scoring/weight adjustment mechanism.
- lack of computation of summary compression ratio.
- lack of sentence length, cue phrases , name of events, and number handling mechanism.

These problems decrease the quality of the generated summary. Hence, this work would explore previous work in detail, as well as examine the quality and performance of automatic Afaan Oromo news text summarizer by adding; name of numbers, cue phrase, name of times, days and months and sentence length mechanism futures, including the two features that has been used in previous work.

Therefore, the researchers set the following research question to examine the problem in automatic Afaan Oromo news text summarization and to develop efficient automatic news summarizer.

1. Which compression ratio is relevant for automatic Afaan Oromo news text summarizer?
2. How and when did the performance of Afaan Oromo text summarizer increases?
3. To what extent, additional features incorporated in this study, affect the performance and quality of the summarizer?
4. Which feature contributes more, and less, to the performance of the summarizer?

1.4 Objective of the Research

1.4.1 General Objective

The general objective of this study is to investigate the way of designing and developing Afaan Oromo news text summarization based on sentence selection functions.

1.4.2 Specific Objective

- ☞ To study linguistic aspect of Afaan Oromo language
- ☞ To identify techniques, methods and define new equations to assign weight for selected features.
- ☞ Conduct experiments to choose better techniques and methods for Afaan Oromo text summarization.
- ☞ Conduct experiment to choose relevant compression rate
- ☞ To design, adopt and develop a suitable algorithm based on the identified techniques and defined equations.
- ☞ To develop a prototype of automatic Afaan Oromo news text summarizer based on the newly proposed algorithm.
- ☞ To evaluate and test the performance of the summarizer

1.5 Scope and Limitation of the Research

This study focus on the development of automatic Afaan Oromo news text summarization from a document organized around topics using features like: cue phrase, sentence position, keyword frequency, name of numbers, sentence length handling mechanism, and name of weeks, days and time.

1.6 Methodology

Under this section, the methodology how Afaan Oromo news text summarizer designed and developed is discussed.

1.6.1 Literature review

In order to know the subject matter in detail, extensive literature review will be conducted on automatic text summarization, language technology, and Afaan Oromo language.

1.6.2 Data Gathering

During the study and system development, different data will be collected from different source based on their contribution to the system development and study. In this paper, three kinds of data will be gathered, lexicon data, validation data and test data. Below, the methodology how these three kinds of data is prepared is discussed.

1.6.2.1 Lexicon data gathering

To build a lexicon of Afaan Oromo cue phrase, synonyms, abbreviations, stop words and name of weekdays and months will be collected from various secondary source data. Those sources are Afaan Oromo textbook, newspaper, fictions, journal articles, dictionaries, media documents and web sites. Secondary source data is preferred, because it contains standard format of words in Afaan Oromo language. Therefore, it facilitates the work by eliminating the time spent for standardizing the collected data.

Among lexicon data, cue phrase preparation is not straightforward like other lexicon data, i.e. the preparation task is not only gathering, it also includes translation. Hence, Standard English cue phrases will be collected from various literatures and translated to Afaan Oromo language by experts or professionals. Therefore, questioner and focus group discussion will be conducted to translate.

1.6.2.2 Validation data gathering

During system development, to design new system or to tune a parameters validation data will be collected. Its sole purpose is to design a summarizer which resembles to human generated summary, and used to check whether the system is appropriately performing according to the proposed approach.

Hence, to prepare validation data, different newspaper articles will be compiled from Afaan Oromo news portals. These articles are from different topics such as politics, entertainment, technology, social, business, economy, agriculture and sport. The main reason that different topics are selected is, to robust the quality and performance of the summarizer. Then, 20 different topics will be set for validation data.

Then, to prepare validation data these topics will be distributed in the form of questioner to 60 subject respondents. The subjects are 53 Dilla University Afaan Oromo 3rd year students and 7 Afaan Oromo instructors. The subjects are requested to underline the candidate sentences that incorporated in the summary.

In this work, because of the complexity in preparation of validation data, only three different subjects are expected to underline each topic. Besides, only 20 topics are selected and 60 subject respondents are taken. On the other hand It is possible to take any universities in Ethiopia those have Afaan Oromo department, but here Dilla University is selected purposely. In addition, the students are taken purposely only from 3rd year, because they are better in proper judgment and preparation of summary than both 1st and 2nd year students; that is because of the course they offer and seniority.

1.6.2.3 Test Data Gathering

To evaluate the performance of the system it is mandatory to prepare the test data corpus. Therefore, to test the performance of the system like a validation data test data will be also gathered from the same sources. Accordingly, 13 different topics are reserved for testing purpose. Like validation data preparation, again our subjects are another 58 Dilla University Afaan Oromo 3rd year students and 7 Afaan Oromo instructors. Hence, 65 subjects will take part in preparation of test data.

The number of subject and a kind of respondents that are participated in test data preparation is limited, because of the same reasons that have been mentioned for validation data preparation.

1.6.3 Experimental Evaluation

Methods of evaluation of automatic text summarization are classified into two major categories [38]. The first one is an intrinsic evaluation, which tests the summary by itself. The second, an extrinsic evaluation, tests the summary based on how it affects the completion of some other task. Intrinsic evaluations have assessed mainly the coherence and informativeness of summaries based on human generated summary. On the other hand, extrinsic evaluations test the impact of summary on tasks like relevance assessment, reading comprehension, etc.

Intrinsic evaluation method will be chosen in this work, since it is most widely used method for evaluation of text summarization. The evaluation will be undertaken in to two ways subjective and objective ways. The subjective evaluation will focuses on evaluation of the structure of the summary, such as summary referential integrity and non-redundancy, coherence and informativeness. On the other hand, the objective evaluation will uses metrics like precision[1], recall[2] and F-measure. F-measure will be selected to determine the performance of the summarizer.

---

1. Sentence recall measures how many of the sentences in the reference summary that are present in the generated summary and in a similar manner precision can be calculated.[34]
2. Precision is the number of sentences in the generated summary that are present in the reference summary.[34]

1.6.4 Development Tool

Open Text Summarizer (OTS) [3] is chosen as development tool in this work. It is an open source tool for summarizing texts. It is a program that reads a text and decides which sentences are important and which are not. OTS is based on sentence extraction using only keyword frequency and sentence position methods to calculate sentence importance [10]. It supports more than 25 languages, which configured in XML[4] files. OTS ships with Ubuntu, Fedora and other Linux distribution OTS, Windows's version source code available in visual C++ and visual C#, etc.; C# version of the open source selected in this study. The main reason that C# version is choose is that it is object oriented programming; it support reusability, inheritance and easy to expand.

1.7 Significance of the Research

With the advent of the information age, people are puzzled with the problem of finding relevant information efficiently and effectively. Today, search engine and automatic text summarizer are two essential technologies that solve these problems. Text search engines serve as information filters that retrieve an initial set of relevant documents, and text summarizers play the role of information spotters to help users locate a final set of desired documents [12]. Text search and summarizer are technologies to reduce the access time for information [5]. Similarly, text summarizer generates summary of document that enables users quickly identify the content of the text to determine final set of relevant documents [11]. Therefore, without any doubt automated Afaan Oromo news text summarization provides these benefits.

---

3. http://libots.sourceforge.net/

4. XML: eXtensible Markup Language.

Having these facts in general, particularly a number of contributions are provided by this study, including-

- ☞ Various organizations/individuals like news agency, embassy, businessperson, politician, and language departments in educational institutions has benefited from the result of the research.
- ☞ It contributes towards the realization of robust Afaan Oromo summarizer.
- ☞ A number of applications can benefited, such application are:-
    - To summarize news down to SMS[5] or WAP[6], for Mobile phones
    - To make computers read summarized text during language translation, because written text can be too long to listen too
    - For search engines, to present short description of matching text
    - The application that provides short translated text of a summarized news text in foreign language

1.8 Thesis outline

This thesis contains seven chapters. Chapter 2 gives a comprehensive literature review of text summarization. Chapter 3 discuss about the related work done in the area of extractive automatic text summarization. Chapter 4 discusses overview of Afaan Oromo and newspaper. Chapter 5 is the broad and the crucial part of the research, which discuss about architecture, design, and implementation of the proposed system. Chapter 6 discusses AOTS evaluation techniques, evaluation result and discussion, the performance gap between this work and Open Oromo Text Summarizer (OOTS). The last chapter discusses the conclusion, recommendation and future work forwarded.

---

5. SMS: Short Message Service a secure specification that allows users to access information instantly via handheld wireless devices such as mobile phones, pagers and communicators.
6. WAP: Wireless Application Protocol the transmission of short text messages to and from a mobile phone, fax machine and/or IP address. Messages must be no longer than 160 alpha-numeric character

Chapter Two: Review of Related Literature

2.1 Overview of Automatic Text Summarization

Obviously, necessity is a mother of innovation. In similar manner, the need of automatic text summarizer calls for attention because of managing and searching relevant information becomes difficult from day to day. This is, because of the dramatic raise of information. The consequence of this problem leads the users to find a solution for quickly locating desired information. As a result, the idea of automatic text summarization came to ground in the 1950s by Luhn [49].

2.1.1 Types of summaries and their properties

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user or task [48]. The summary may generated from different source: email, legal proceedings, customer reviews, search results, meetings, and videos. However, the goal of each type of summary is the same, which is to convey the most important information from a set of documents within a length constraint using natural language processing.

There are several distinctions typically made in summarization, here in this section terminology that has often mentioned in different summarization literature seen. As presented in [1] [9] summaries may be classified into several types based on the following criteria.

2.1.1.1 Indicative vs. informative

Summaries may distinguished by their content. A summary that enables the reader to determine about-ness has often called an indicative summary, while one that can be read in place of the document called an informative summary [7]. An indicative summary may provide characteristics such as length, writing style, etc., while an informative summary will include facts that reported in the input document(s).

2.1.1.2 Extractive vs. Abstractive

The summary can also classified based on the sentence similarity with the original input document. It can be extractive summary (extracts), which produced by concatenating several sentences taken exactly as they appear in the materials that summarized. On the other hand, it can be abstractive summary (abstracts) which is a summary that written to convey the main information in the input and reuse phrases or clauses from it, but the summaries are overall expressed in the words of the summary author.

Abstractive and extractive text summarizations are popular known method of text summarization, which has identified by different researchers; however, both types of summaries have a several problems [38] that stated as follows:

Problems with Extractive methods:

- ☞ Extracted sentences usually tend to be longer than average. Due to this, parts of the segments that are not essential for summary also get included, consuming space.
- ☞ Important or relevant information may usually spread across sentences, and extractive summaries cannot capture them (unless the summary is long enough to hold all those sentences).
- ☞ Conflicting information may not present accurately.

Problems with abstractive methods:

- ☞ The users prefer extractive summaries instead of glossed-over abstractive summaries [40]. This is because extractive summaries present the information as is by the author, and would allow the users to read between the lines information.

- ☞ Sentence synthesis is not a well-developed field yet, and hence the machine generated automatic summaries would result incoherence even within a sentence. In case of extractive summaries, incoherence occurs only at the border of two sentences.
- ☞ It is generally hard for computer to successfully solve the requirements of such approach as of many limitations, including the state of the art in language generation and human language complexity.

Nowadays, most of automated text summarization systems use extraction methods to produce summary. Even though extract summarization is easy to implement and popularly used method, it has three major difficulties [31] [32] those are:

- ☞ Finding out which are the most important sentences to use on the summary
- ☞ How to generate a coherent summary
- ☞ Remove all redundancies in the summary

2.1.1.3 Generalized vs. Query-based

Based on user need the summary classified as generic and query based summary. Generic summarization makes few assumptions about the audience or the goal for generating the summary [12]. Typically, it had assumed that the audience is a general one: anyone may end up reading the summary. Furthermore, no assumptions has made about the genre or domain of the materials that need to be summarized. In this setting, importance of information is determined only with respect to the content of the input alone. It is further assumed that the summary will help the reader quickly determine what the document is about, possibly avoiding reading the document itself.

In contrast, [12] in query-focused summarization, the goal is to summarize only the information in the input document(s) that is relevant to a specific user query. For example, in the context of information retrieval, given a query issued

by the user and a set of relevant documents retrieved by the search engine, a summary of each document could make it easier for the user to determine which document is relevant. To generate a useful summary in this context, an automatic summarizer needs to take the query into account as well as the document. The summarizer tries to find information within the document that is relevant to the query or in some cases, may indicate how much information in the document relates to the query. According to [12] much of the work to date has been in the context of generic summarization.

2.1.1.4 Single vs. Multi document summarization

The summary is also classified based on the input document in to two; single and multi-document summary. Single document summarization is systems produced a summary of one document, whether a news story, scientific article, broadcast show, or lecture. Most of early work in summarization dealt with single document summarization [38].

Due to improvement on text summarization, a new type of summarization task emerged: Multi-document summarization motivated by use cases on the web. Given the large amount of redundancy on the web, summarization was often more useful, that provides a brief digest of many documents on the same topic or the same event. In the first deployed online systems, multi-document summarization applied to clusters of news articles on the same event and used to produce online browsing pages of current events [40].

2.1.1.5 Shallow vs. Deeper Summarization

The methods of summarization classified in two broad groups based on the level in the linguistic space [1]. (a) shallow approaches, which are restricted to the syntactic level of representation and try to extract salient parts of the text in a convenient way; and (b) deeper approaches, which assume a semantics level of representation of the original text and involve linguistic processing at some level.

2.1.2 Summary Evaluation Strategies and Metrics

After a number of reviews, Martin Hassel [34] found that there are at least two properties of the summary that measured when evaluating summaries, and summarization systems: the Compression Ratio (how much shorter the summary is than the original); and the Retention Ratio (how much information is retained).

Compression ratio (CR) computed:

$$CR = \frac{Length\ of\ the\ Summary}{Length\ of\ Full\ Text} \qquad (2.1)$$

Retention Ration computed:

$$RR = \frac{information\ \ in\ Summary}{information\ in\ Full\ Text} \qquad (2.2)$$

Despite the summary properties, Martin Hassel [7] and [43] found the first broad division methods for evaluation automatic text summarization systems, as well as many other systems. They classified into intrinsic and extrinsic evaluation methods.

2.1.2.1 Intrinsic Evaluation Method

The intrinsic evaluation involves assessing the quality of a summary by comparing it to an ideal summary (summary produced by humans). This is often done by comparison to some gold standard, which can be made by a reference summarization system or, more often than not, is man-made using informants. Intrinsic evaluation has mainly focused on the coherence and informativeness of summaries.

This evaluation method has different criteria to measure the quality of the summary. Those criteria discussed as follows;

This evaluation method computed by two different mechanisms [38]; those mechanisms are:

A. Defining criteria
B. Comparing the summary against reference output

A. Defining criteria

I. Summary Coherence

Summaries generated through extraction-based methods sometimes suffer from parts of the summary being extracted out of context, resulting in coherence problem (e.g. dangling anaphors or gaps in the rhetorical structure of the summary).

This can be can be assessed by having humans grade summaries for coherence based on specific criteria [38]. For example, as presented in Minel et al. [46] for extractive summary, subjects' grade readability of summaries based on the presence of dangling anaphors, lack of preservation of the integrity of structured environments like lists or tables, choppiness of the text, presence of autologous statements such as predicting the future is difficult, etc. Robin, J. [61] explore that abstracts summary, like extracts, can also be incoherent, especially when natural language generation is used. He had judges grade the acceptability of abstracts produced by cut-and-paste operations on the source text, based on general readability criteria such as good spelling and grammar, clear indication of the topic of the source document, impersonal style, conciseness, readability and understandability, acronyms being presented with expansions, etc.

II. Summary Informativeness

The main reason of measuring informativeness of a summary is to assess the summary's information content regarding to the original input document. Mani [38] present in his study that as a summary of a source becomes shorter, there is less information from the source that preserved in the summary. Therefore, one measure of the informativeness of a summary is to assess how much information from the source preserved in the summary.

Another measure is how much information from a reference summary covered by information in the system summary. In other words, as in the case of coherence, comparisons made between system summaries, the source, reference summaries, and scores for other summarization systems. However, while subjective grading used for informativeness, informativeness is more amenable than coherence to automatic scoring.

B. Comparing the summary against reference output

The idea of a reference summary, against which machine output compared, is a very natural one. A varies of different measures used based on [34]; however, in this paper two of them discussed:

I. Sentence Precision and Recall

Sentence recall measures how many of the sentences in the reference summary that are present in the generated summary and in similarly precision calculated. Precision and Recall are standard measures for Information Retrieval and are often combined in a so-called F-score. In their study, Inderjeet [43] describe the method of text summarization that adopted from IR literature as follows:

Where TP= true positive FP= False Positive, TN = True negative, FN= false negative

TP: Manually generated intersection with machine generated summary

TP + FP: Machine generated summary

TP + FN: Manually generated summary

Precision: The ratio of sentence that exist both in manually generated summary and machine generated summary per machine generated summary.

$$Precision = \frac{TP}{(TP+FP)} \qquad (2.3)$$

Recall: The ratio of sentence that exists in both manual generated summary and machine generated summary per manual generated summary.

$$Recall = \frac{TP}{(TP+FN)} \quad (2.4)$$

F- Measure: This is harmonic mean of precision and recall.

$$Fmeasure = \frac{2*Precision*Recall}{(Precision + Recall)} \quad (2.5)$$

However, the main problems with these measures for text summarization is that they are not capable of distinguishing between many possible, but equally good, summaries and that summaries that differ quite a lot content wise may get very similar scores.

II. Sentence Rank

In this mechanism, the reference summary has constructed by ranking the sentences in the source text by worthiness of inclusion in a summary of the text. Correlation measures applied to compare the generated summary with the reference summary. As in the case of P&R this method mainly applies to extraction based summaries, even if standard methods of sentence alignment with abstracts applied.

2.1.2.2 Extrinsic evaluation

As presented in [38] the idea of an extrinsic summarization evaluation is to determine the effect of summarization on some other task. There have been a number of extrinsic evaluations involving question-answering and comprehension tasks [43], as well as tasks, which measure the impact of summarization on determining the relevance of document to a topic.

Mani [38] discussed a two selected tasks to convey an idea of the type of evaluation carried out in extrinsic evaluation method.

1. Relevance Assessment

In the task of relevance assessment, a subject presented with a document and a topic, and asked to determine the relevance of the document to the topic. The influence of summarization on accuracy and time in the task is then studied.

2. Reading Comprehension Tasks

In reading comprehension tasks, the human first reads full sources or summaries assembled from one or more documents. The human then answers multiple-choice test. The system then automatically scores the answers, measuring the percentage of correct answers. Thus, a human's comprehension based on the summary objectively compared with that based on the source. The reasoning here is that if reading a summary allows a human to answer questions as accurately as he would reading the source, the summary is highly informative.[62] carried out on an extrinsic evaluation of the impact of summarization in a task of question-answering.

2.1.2.3 Intrinsic compared to Extrinsic

The choice of an intrinsic or extrinsic method depends very much on the goals of the developers, funders, and consumers of the summarization technology [38]. In general, at early stages of the technology cycle, intrinsic evaluations recommended, with an emphasis on evaluation of summarization components; as the technology matures, more situated, task-based tests of the system as a whole involving real users become more important.

Extrinsic evaluations have the advantage of assessing the utility of summarization in a task, so they can be of tremendous practical value to a funder or consumer of summarization technology. However, they are less useful to developers in terms of offering feedback as to how they might improve their systems. This can be somewhat alleviated when a systems performance in

an intrinsic evaluation predicts performance in an extrinsic one. In addition, developers need repeatable, less expensive, automatically scorable evaluations.

2.1.3 Challenges in evaluating automatic text summarization

According to [38] identified several serious challenges in evaluating automatic text summarization, which makes summarization evaluation a very interesting problem. Those challenges are:

Summarization involves a machine producing output that results in natural language communication. In cases where the output is an answer to a question, there may be a correct answer, but in other cases, it is hard to arrive at a notion of what the correct output is. There is always the possibility of a system generating a good summary that is quite different from any human summary used as an approximation to the correct output. (Similar problems occur with machine translation, NLP.)

Since humans may be required to judge the systems output, this may greatly increase the expense of an evaluation. An evaluation method, which uses a scoring program instead of human judgments, is preferable, since it is easily repeatable.

Summarization involves compression, so it is important to be able to evaluate summaries at different compression rates. This increases the scale and complexity of the evaluation.

Since summarization involves presenting information in a manner sensitive to a user's or applications needs, these factors need taken into account. This in turn complicates the design of an evaluation.

2.1.4 Features for extractive text summarization

Extractive text summarization method based on sentence extraction methods, which normally work by scoring each sentence in a document as a candidate to be part of summary, and then selecting the highest scoring subset of sentences. A number of NLP researchers [30, 31, 32, 33] found some features that often increase the candidacy of a sentence for inclusion in summary. Those futures are:

- **Keyword-occurrence**: Selecting sentences with keywords that are most often used in the document usually represent theme of the document
- **Title-keyword**: Sentences containing words that appear in the title are also indicative of the theme of the document
- **Location heuristic**: In newspaper articles, the first sentence is often the most important sentence; in technical articles, last couple of sentences in the abstract or those from conclusions is informative of the findings in the document [42].
- **Cue phrases**: Agustín Gravano et al. [39] states in their study cue phrases are linguistic expressions that used to convey explicit information about the discourse or dialogue, or to convey a more literal, semantic contribution. They aid speakers and writers in organizing the discourse, and listeners and readers in processing it. Sentences containing key phrases like "*this report", "in conclusion", "this letter", "this report", "summary", "argue", "purpose", "develop", "attempt" etc."*.
- **Short-length and long-length cutoff**: Short sentences and long sentence are usually not included in summary. Because there is always a risk that long, sentences will be ranked higher.
- **Pronouns:** A pronoun can replace a noun or another pronoun. You use pronouns like "he," "which," "none," and "you" to make your sentences less cumbersome and less repetitive. Pronouns such as "she, they, it" cannot be included in summary unless they are expanded into corresponding nouns.
- **Weekdays, times and Months:** Sentences that include days of the week and months scored higher.

- **Quotation:** Sentences containing quotations might be important for certain questions from user.
- **Numerical data:** Sentences containing numerical data scored higher than ones without numerical values are.

Different literature [30, 31, 32, 33] proved that the features that are mentioned above cannot alone produce high quality extracts. Therefore, to get high quality extracts using two or more features in combination.

2.1.5 Method and Algorithms of summarization

Like other NLP application, automatic summarizer has its own method and algorithm that has done so far. There are a number of algorithm and methods developed by scholars of NLP; off a number of algorithms [31, 32] totally nine approaches will be discussed in this paper ; seven of them was identified by [31,32] and Jen-Yuan et al. [26] at university of National Chiao-Tung found two novel approach ; approach VIII and IX that are listed below.

I. Sentence selection function for Extraction
II. Knowledge-Based Concept Counting
III. Lexical Chain Methods
IV. Vector-Based Semantic Analysis using Random Indexing
V. Latent Semantic Analysis (LSA)
VI. Pronoun Resolution
VII. Machine Learning Techniques
VIII. Corpus based approach using features analysis.
IX. Text Summarization Using a Text Relationship Map

2.1.5.1 Sentence Selection Function for Extraction

In order to extract the sentence from the document, this summarization method uses several features, which listed, on section 2.1.5. Hence, the score of each feature combined to create a weight for the individual sentence. However, it is not quite clear how to combine these several different scores [31] [32], but from the several approaches that have been described on different literature, the most common point of them is that coefficients assign various weights to the individual scores, those are added together. It is important to mention that those coefficients are depended on the language of the text. Because of lack of standard in assigning weight, Simple combination function proposed: This is a linear combination in which the parameters specified manually by experimentation. These coefficients can be, as described before first sentence, numerical data etc. The score calculated according to the following equation:

$$Sentence\ score\ = \sum_{j}^{n} CjPj \qquad (2.6)$$

Where **Cj:** is the jth parameter coefficient**, Pj:** is the jth parameter**, n:** is the number of parameter.

Therefore, to assign coefficient for each features conducting the experiments is necessary thing, having this fact simple combination function needs high effort in conducting the experiments for each features.

2.1.5.2 Knowledge-Based Concept Counting

This new method first presented by Lin [4], knowledge-based concept counting is a method for automatically identifying the central ideas in a text, based on a knowledge-based concept counting paradigm. According to Lin concepts is generalized using concept generalization taxonomy (WordNet). Figure 2.1 shows a possible hierarchy for the concept "World country" For example, the main topic in the sentence "*Chaltu bought some pen, pencil, exercise book, and binder.*" those items should be stationary, but we cannot make any conclusion about the topic of this sentence by using word counting methods or other type

of summarization approach. Because, word-counting methods miss the important concepts behind those words: *pen*, *pencil*, *etc*. relates to *stationary* at the deeper level of semantics.

Based on this hierarchy, if we find France, Spain, Germany, in a text, we can infer that the text is about *Europe* Country. In addition, if the text also mentions about Ghana, Sudan and Somalia, it is reasonable to say that the topic of the text related to Africa. Using a hierarchy, the question is now how to find the most appropriate generalization in the taxonomy hierarchy. According to this method, the nodes in the middle of the taxonomy are most appropriate, since the top concept is always, a thing (everything is a thing) and using the leaf concepts give us no power from generalization. Therefore, let us discuss on how to find most appropriate generalization based on the hierarchy given above on (Fig. 2.1).

Ratio (R) is a way to identify the degree of summarization. The higher the ratio, the more it reflects only one child. The ratio defined with the following formula:

$$R = \frac{MAX(W)}{SUM(W)} \qquad (2.7)$$

, when W = the weight of all the direct children of a concept. The weight of a parent concept defined as the frequency of occurrence of a concept C and its sub concepts in a text. Moreover, in this hierarchy the weight given randomly to describe this approach. For example the Ratio (R) for the parent's concept in the (Fig. 2.1) (b) is 10/ (2+3+6+10+4+1) = 0.5 while it is 0.2 in the Figure 2.1 (c).

For determination of the degree of generalization, the branch ratio threshold (Rt) is defined. Rt serves as a cutoff point for the interestingness. If a concept's ratio R is less than Rt, it is an interesting concept. For example consider in case (Fig. 2.1) (b) if the Rt = 0.4, we should choose Spain as the main topic instead of its parent since Rt < R. In contrast, in case (c) we should use the

parent concept Europe as the concept of interest, because no R which is less than the given threshold Rt.

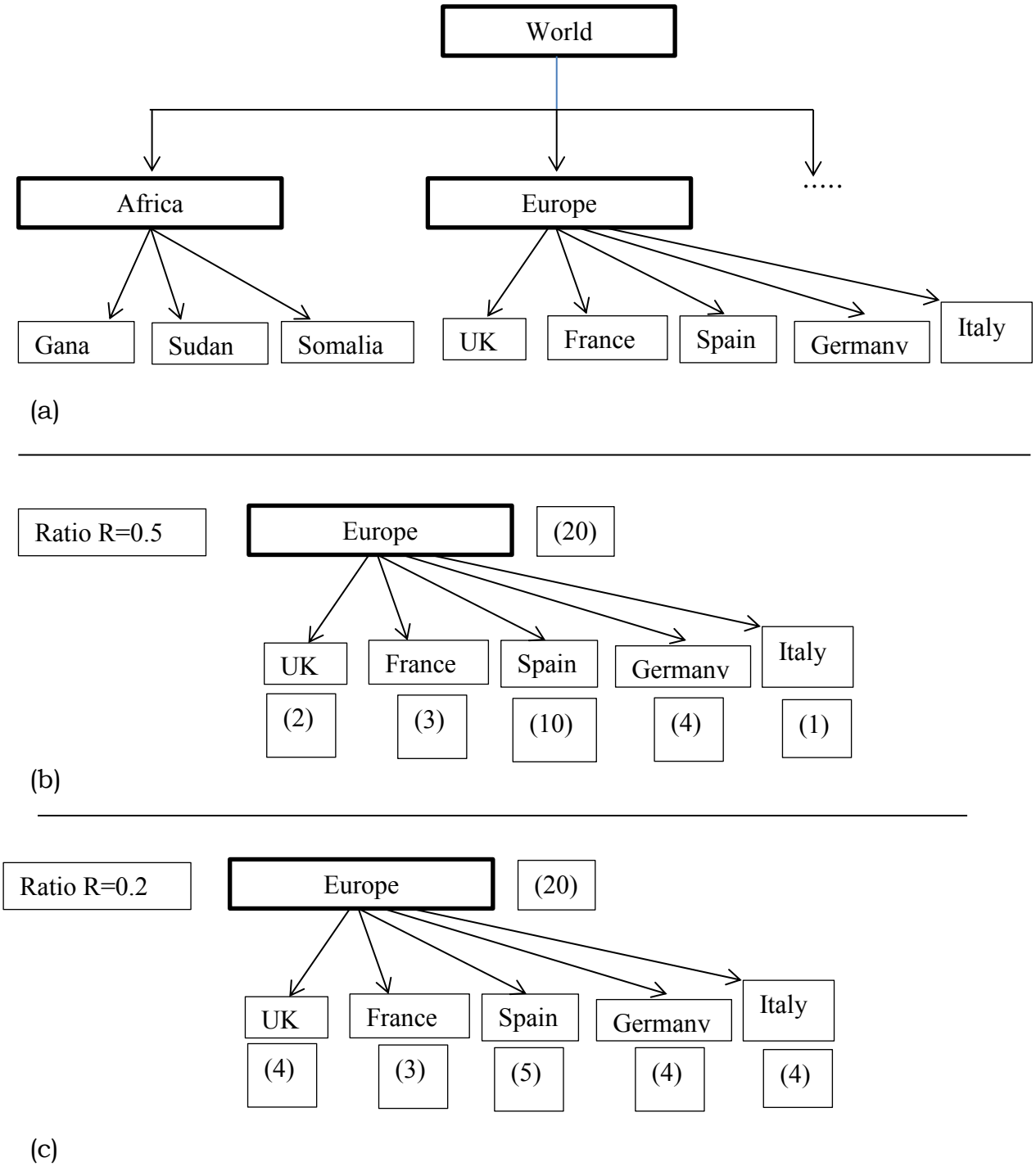

Figure 2.1: *A sample hierarchy for world countries*

2.1.5.3 Lexical Chain Methods

This is a statistical, corpus-based text comparison mechanism [31]. According to [13] a lexical chain is a set of words in a text that related to each other. In the beginning, it was launched as a task for information retrieval, but in the preceding years, it showed remarkably human-like abilities in several language tasks [4]. It used in sentence extraction methods and it helps reducing the summary [31].

The relation between the words can be found using lexical lists such as thesaurus or WordNet. With this method, the most important concepts can be found statistically, by looking the structure of the passage rather than deep semantic meaning. To calculate this, all required is a generic knowledge base containing nouns and their associations.

The general algorithm for computing word lexical chain is the following:

- ☞ Make a list of candidate words from the passage
- ☞ For each of the candidate words, find an appropriate lexical chain to get a candidate word, relying on the relatedness criterion between members of the lexical chain and the candidate words.
- ☞ If such a chain found, insert the candidate word in the lexical chain and update it accordingly or else create a new chain.

Chains scored depending on a number of heuristics, some of which are their length, the kind of relation between their words, the position they hold in the passage, etc. The ones that are the mostly connected to lexical chains are those that are being extracted.

However, lexical chain method has one major drawback; they are insensitive to the non-lexical structure of passages, such as their rhetorical, argumentative or document structure [30, 31]. For instance, they do not take into account the position of elements of a chain within the argumentative line of the discourse.

2.1.5.4 Latent Semantic Analysis (LSA)

Makbule Gulcin Ozsoy et al. [51] defined Latent Semantic Analysis (LSA) as an algebraic statistical method that extracts meaning of words and similarity of sentences using the information about the usage of the words in the context. It keeps information about which words are used in a sentence, while preserving information of common words among sentences. The more common words between sentences mean that those sentences are more semantically related.

LSA method can represent the meaning of words and the meaning of sentences simultaneously. It averages the meaning of words that a sentence contains to find out the meaning of that sentence. It represents the meaning of words by averaging the meaning of sentences that contain this word. LSA method uses Singular Value Decomposition (SVD) for finding out semantically similar words and sentences. SVD is a method that models relationships among words and sentences. It has the capability of noise reduction, which leads to an improvement in accuracy.

They also identified that LSA has three main limitations. The first limitation is that it uses only the information in the input text, and it does not use the information of world knowledge. The second limitation is that it does not use the information of word order, syntactic relations, or morphologies. Such information is used for finding out the meaning of words and texts. The third limitation is that the performance of the algorithm decreases with large and inhomogeneous data. The decrease in performance is observed since SVD which is a very complex algorithm used for finding out the similarities.

2.1.5.5 Vector-Based Semantic Analysis using Random Indexing

As George Pachantouris [31] referred from Karlgren and Sahlgren's study Vector-Based Semantic Analysis using Random Indexing is a technique to extract, from a text, semantically similar terms by observing the distribution and collection of terms inside the text. The result of running a vector-based semantic analysis on a text is a thesaurus: an associative model of term meaning. Random Indexing (RI) uses sparse, high-dimensional random index vectors to represent documents. Based on the hypothesis that any document has assigned a random index vector the term similarities can be calculated by computation of terms-by-contexts co-occurrence matrix. Each of the rows of it represents a term, and the term vectors are of the same dimensionality as are the random vectors assigned to texts. Every time a term is found in a text, that text's random index vector is being added to the row for the term in question. With this method, terms are represented in the matrix be high-dimensional semantic context vectors that contain traces of every context the specific has been observed in. The assumption behind this theory is that semantically similar terms will show up in similar contexts and therefore their context vectors will be quite similar. In this way, by calculating similarities between context vectors, it should be possible to calculate the semantic similarity between any given terms. This similarity measure will reflect the distributional or contextual similarity among different terms.

2.1.5.6 Pronoun Resolution

The consequence of performing an automatic summary without deeper linguistic analysis cause the resulting text can often result in broken anaphoric references [31]. For example in a sentence "***Chaltu*** *arrived to* ***Nekemte****. She has lived* ***there*** *for 3 months*." If "*Chaltu*" and "*Nekemte*" are mentioned nowhere in the previous text; it is impossible to understand that "She" refers to "Chaltu" and "there" refers to "Nekemte".

Hence, Pronoun resolution is text summarization is a module that can point is pronouns to their exact reference by using deep linguistic analysis. In Swesum [34], they are tried to resolve pronoun by only building reference for gender and noun lexicon.

2.1.5.7 Machine Learning Techniques

In this technique by giving a set of training documents and their extractive summaries, the summarization process is modeled as a classification problem: sentences are classified as summary sentences and non-summary sentences based on the features that they possess [32]. The classification probabilities are learnt statistically from the training data, using Bayes' rule[1]:

Nima [32] adopt the Bayesian rule as follows:

P (s ∈ < S | F1, F2… FN) = P (F1, F2… FN | s ∈ S) * P (s ∈ S) / P (F_1, F_2… FN)

where, s is sentences from the document collection, F_1, F_2…FN, are features used in classification and

P (s ∈ < S | F_1, F_2… F*n*) is the probability that sentence s will be chosen to form the summary S given that it possesses features F_1, F_2…F_n.

---

[1] Bayes' theorem/ rule is named after Thomas Bayes, a nonconformist English clergyman who did early work in probability and decision theory during the 18th century. In Bayesian terms, X is considered "evidence." As usual, it is described by measurements made on a set of n attributes. Let H be some hypothesis, such as that the data tuple X belongs to specified class C. For classification problems in data mining P (H/X), the probability that the hypothesis H holds given the "evidence" or observed data tuple X [27].

$$P(H/X)\frac{P(X/H)P(H)}{P(X)}$$

2.1.5.8 Corpus based approach

Jen-Yuan et al. [26] propose the corpus based text summarization approach possible by exploiting technologies of machine learning, which lets the machine to learn rules from a corpus of documents and the corresponding summaries. They decomposed the process into two phases: the training phase and the test phase.

In the training phase, the system extracts particular features from the training corpus and generates rules by a learning algorithm. In the test phase, the system applies rules learned from the training phase to the test corpus to generate the corresponding summaries; and measure the performance.

2.1.5.9 Text Summarization Using a Text Relationship Map

The idea is first introduced by Salton et al., the construction of a text relationship map (T.R.M.) is to link similar paragraphs [26]. In the map, each node stands for a paragraph and is represented by a vector of weighted terms. A link is created between two nodes if the two corresponding paragraphs have strong relevance. The relevance between two paragraphs is determined by their similarity that is typically computed as the inner product between the corresponding vectors. When the similarity between two paragraphs is larger than a predefined threshold, the link is constructed.

They defined bushiness to measure the significance of a paragraph. The bushiness of a paragraph is the number of links connecting it to other paragraphs. They also proposed three heuristic methods to generate a summary: global bushy path, depth first path, and segmented bushy path. Since a highly bushy node is linked to many other nodes (i.e. it has many overlapping vocabularies with others), it is likely to discuss main topics covered in many other paragraphs.

2.3.6 Stage of Text Summarization

Different scholars agreed on that there are three basic stage of automatic text summarization. The stages are listed on figure 2.2 [1]:

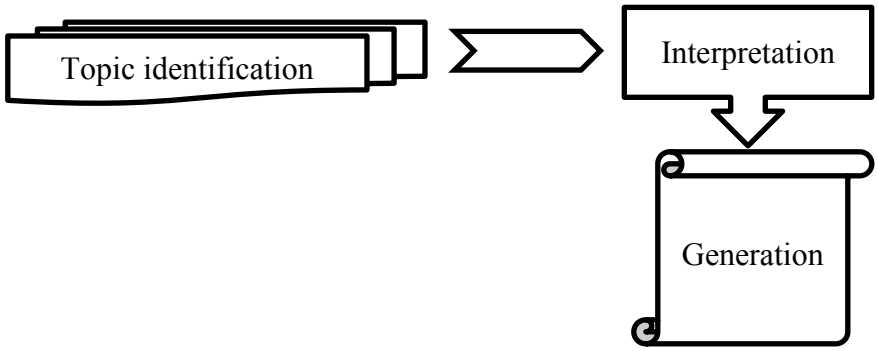

Figure 2. 1: *Stages in automatic text summarization*

I. Topic Identification:

The purpose is to filter the input to retain only the most important, central, topics. Typically, topic identification can be achieved using various complementary techniques, including those based on stereotypical text structure, cue words, high frequency indicator phrases, and discourse structure.

II. Interpretation:

Once the desired central topics have been identified, they can simply be output, to form an extract. In human summaries, however, a process of interpretation is usually performed to achieve further compaction. To remove redundancies, rephrase sentences to pack material more densely and, importantly, to merge or fuse related topics into more general ones. The various types of fusion are not yet known, but they include at least simple concept generalization

(Tola ate Mango, Orange, and bananas → he ate fruit) and script identification (Chaltu sat down; pay a money, receive ticket, take a picture of animals, and left → she visited the zoo).

III. Generation:

The goal is to reformulate the extracted and fused material into a coherent, densely phrased, new text. If this stage is skipped, the output is a verbatim quotation of some portion(s) of the input, and is not likely to be high-quality text (although this might be sufficient for the application).

2.3.7 Key Challenges in Text Summarization

As Oi Mean Foong et al. [41] found in their study entitled Challenges and Trends of Automatic Text Summarization they examined three basic challenges in the development of automatic text summarization:

- ☞ Firstly, they faced difficulty to select the important features of a text summarization system that extracts the main ideas from original documents. Because of that unlike the single document summary, there exists an inherent problem of overlapping of themes, i.e. sets of similar text units or paragraphs. Similarly, documents, which contain long sentences, are still a problem because the abstraction process requires further knowledge in NLP to perform sentence reduction. Another challenge in text summarization is how to interpret jargon accurately, or close to human linguists' summarized output.
- ☞ The second challenge is on how to address ambiguous sentences in the original documents, if any. Which means the challenge is to find the know-how to resolving the word ambiguity with different meanings depending on the context.
- ☞ The last difficulty is after generating the summary, how to evaluate the text summarization system. The problem is that matching a system summary against the ideal summary is very difficult to establish

Chapter Three: Related Works

3.1 Introduction

Because of dramatic growth in both IT device and information exchange mechanisms, automatic text summarization has become one of the most popular research areas in the field of Natural Language Processing starting since 1950's. Besides, it has drawn a lot of interest in the natural language processing and information retrieval communities in the recent years. Some of the research works are done in different languages; however, most of the works done on English language than the rest of the languages in the world. Most of them have employed extractive based text summarization. In this section the work done on area of automatic text summarization shown roughly; the one that most related to the work briefly reviewed. The related work that discussed under this section categorized into three. Those are foreign language text summarizer from a global context, Amharic text summarizers and Afaan Oromo text summarizer.

3.2 Text Summarization on foreign language

It has now been more than 50 years since the publication of Luhn's influential paper on automatic text summarization. During these years, the practical need for automatic summarization has become increase dramatically, for this reason, urgent and numerous papers published on this topic. Among those papers, Oi Mean Foong et al. [41] listed internationally published automatic text summarization by different scholars shown on the table 3.1 according to their category and technique they used.

Table 3.1: *List of existing research so far done on English text summarization*

Author/Year	Existing Work		
	Category	**Techniques**	**Journal/Proceedings**
Luhn, 1958	Statistical Approach	Word Frequency	IBM Journal of Research and Development
Baxendale, 1958		Position in Text	IBM Journal of Research and Development
Edmunson, 1969		Cue words and Heading	Journal of ACM
Kupiec, 1995		Naïve Bayes	SIGIR 1995
Miller, 1995		WordNet Lexical Terms	Communication ACM
Lin & Hovy, 1997		Sentence Position	5th Conference on Applied Natural Language Processing
Lin, 1999		Decision Tree	CIKM
Conroy & O'leavy, 2001		Hidden Markov Method & Sequential Model	SIGIR 2001
Osborne, 2002		Maximum Entropy	ACL 2002 Workshop on Automatic Summarization
Lin, 2004		Similarity of Sentences	ACL2004 Workshop
Nenkova, 2005		Proper ranking of sentences	AAAI 2005
Yong, 2005		Neural Network	International

			Conference on Data Mining
Svore, 2007		Neural Network algorithm (RankNet)	EMNLP-CoNLL
Aone, 1990	Natural Language Processing(NLP)	Inverse Term Frequency & NLP technique	Advances in Automatic Text Summarization
Barzilay, 1997		Deep NLP	ISTS 1997
McKeown, 1997		Lexical Chains	AAAI
Marcu, 1998		Rhetorical Structure Theory (RST)	6th Workshop on Very Large Corpora
Carbonell & Goldstein, 1998		Maximal Marginal Relevance (MMR)	SIGIR 1998
Daume & Marcu,2002, 2004		Log Probability & Rhetoric Structure Tree (RST)	ACL 2002, DUC 2004
Kaustubh Patil, 2007	Semantic Analysis Approach	Graph Theory, Latent Semantic Analysis (LSA), Node Centrality	International Journal on Computer Science and Information Systems (IADIS)
Zhan, 2007		Info Extraction of salient topics from online reviews	IEEE International Conference on Computer Science and Information Technology.
Verma, 2007		Ontology Knowledge (e.g. WordNet & UMLS)	Document Understanding Conference DUC

		in Medical field	2007
Bawakid, 2008		Semantic Analysis (sentence location, named entities, semantic similarity between user query & sentences)	1st Text Analysis Conference (TAC) 2008
Liu, 2009		Query-based Words Extraction & New Sentence Ranking Formula	ICCPOL 2009, LNAI 5459, Spinger-Verlag
Troels Andreasen, 2009	Fuzzy Logic	Conceptual Clustering & Semantic Similarity Measure	Springer-Verlag
Hamid Khosravi, 2008			
Mohammed Salem BinWahlan, 2009	Swarm Intelligence	Fuzzy Swarm Based Text Summarization	Journal of Computer Science

3.2.1 Generic Text Summarization for Turkish

Generic text summarization is the first Turkish summarization system. Celal Cığır et al. [5] propose a generic text summarization method that generates summaries of Turkish texts by ranking sentences according to their scores calculated using their surface level features and extracting the highest ranked ones from the original documents. In order to extract sentences that form a summary with an extensive coverage of main content of the text and less redundancy, they use the features such as term frequency, key phrase, centrality, title similarity and position of the sentence in the original text.

Sentence rank computed using a score function that uses its feature values and the weights of the features. The best feature weights are learned using machine learning techniques with the help of human constructed summaries. Performance evaluation conducted by comparing summarization outputs with manual summaries generated by 25 independent human evaluators. They used ROUGE evaluation technique to compare summarization outputs with human generated summaries.

Table 3.2: *ROUGE results for recall and precision values of applying all features and all quadruple and all quadruple combinations of features.*

Features	**Recall results**	**Precision Results**	**F-measure**
All Features	0.540	0.809	0.648
Without TF	0.540	0.809	0.648
With TS	0.534	0.789	0.640
With KP	0.543	0.805	0.649
Without SP	0.540	0.770	0.635
Without C	0.540	0.809	0.648

Celal Cığır et al. [5] by applying five features they have found 80.9% precision, 54.0% recall and 64.80 f-measure precision results, which is very inspiring result.

3.2.2 SweSum

SweSum [69] is a web-based text summarizer developed at Royal Institute of Technology (KTH); it is developed for Swedish language [17]. It uses text extraction method based on statistical and linguistic as well as heuristic methods to build text summarization system. The input of the model is Swedish HTML-tagged newspaper text. SweSum is currently available for different language like; in Danish, Norwegian, English, Spanish, French, Italian, Greek, Farsi (Persian) and German texts.

SweSum Lexicon

In SweSum, the lexicon is a data structure used for storing key pairs root table. It is two column file, containing words where the key is the inflected word and the value is the root of the word. The Swedish version contains around 40,000 words and 700,000 different inflections.

SweSum Architecture

Swesum architecture contains three basic stages, which are preprocessing, sentence ranking and sentence extraction.

I. Preprocessing

This phase involves three basic operations: tokenization, keyword extraction and scoring

a. Tokenizing

In this first step the tokenize goes through the input text, and output is the tokenized text. The tokenize perform the following task; removes new line character "\n" , remove abbreviations , it invokes pronominal resolution and it searches for symbols that mark the sentence boundaries like “.” , ”,” , ”!” , ”?” , “<” , “>” , “:” .

b. Keyword Extraction

SweSum uses nouns, adjectives and adverbs to count their occurrences in the document and the most frequently occurred word is considered as a keyword.

c. Scoring

The scoring mechanism that is used in Swesum based on: Sentence position, numerical value, bold text, keywords, and user keywords features. The function that contains the addition of the all feature is known as Simple combination function.

II. Processing

In this stage, the score of every sentence is being identified and sorted according to their occurrence in the document.

III. Generate summary

In this third and final pass, the final summary file is created and which contains: All non HTML lines, from the sorted text value, all the lines that have been ranked high and some statistical information like the percentage of the summary, the keywords, number of lines, words etc.

SweSum has been evaluated and its performance for English, Danish, Norwegian and Swedish is considered to be state-of-the-art. The French, German and Spanish versions are in prototype states. In order to measure and evaluate the performance of the system, they select nine students to carry out the test. The student's carried out the test by first reading the text to be summarized and then gradually lowering the size of the summary giving SweSum the amount of the original text they would like in the summary, noting in a questionnaire when coherence was broken and when important information was missing. This procedure repeated for each of the 10 texts. Finally, [17] they used median as a statistical measurement to get the results, which collected from the 10 respondent. They first calculated the total amount of summarized text (given in percent) and the result is on table 3.3.

Table 3.3: *result from the field test*

	Information	**Coherence**
Total median	30%	24%
Total average	31%	26%

The SweSum research groups are currently working on pronoun resolution to make the summarized text more coherent. One possibility they propose when doing topic identification or keyword extraction is to use a synonym term by using some sort of Swedish Ontology similar to Wordnet, namely Swedish *Wordnet-Swordnet*. Currently it is under development at the department of Linguistics at Lund University, Sweden.

3.2.3 Open Text Summarizer (OTS)

3.2.3.1 About OTS

AOTS developed based on the architecture of Open text summarizer (OTS). OTS is an open source tool for summarizing texts; it reads a text stream and decides which sentences are important and which are not and generate the summary. It is used by external applications, such as word processors such as AbiWord[2] and KWord[3], can link to the library, while the command line tool summarizes text on the console. It can output as either plain text or HTML, when outputs as HTML, the important sentences highlighted; it is also multilingual and support UTF-8[1] encoding. OTS works for 37 languages the list of languages that OTS support shown on Appendix M.

3.2.3.2 How it works

OTS is programmed to accept the input from the user through the provided interface. The inputted text is first segmented in to sentence and the next step is to identify the important ideas in an article and remove stop words. However, before the summarizer starts counting the frequency of words, the words are stemmed.

The researchers distinguished from the example that the algorithm used in OTS is Porter algorithm[4] stemmer; because, the stemmer uses the suffix removal method.

---

1. See: *http://www.fileformat.info/info/unicode/utf8.htm*
2. See: *www.abisource.com*
3. See*: http://www.kde.org/applications/office/kword/*
4. Porter algorithm [52] defines five successively applied steps of word transformation. Each step consists of set of rules in the form <condition> <suffix> -> <new suffix>. For example, a rule (m>0) EED -> EE means "if the word has at least one vowel and consonant plus EED *ending*, change the ending to EE". So "agreed" becomes "agree" while "feed" remains unchanged.

Example: *OTS stemmed words*

Inflected words	***stem of words***
[dam] | [dams]	*stem[dam]*
[asking]| [asked]	*stem[ask]*
[build]| [building]	*stem[build]*
[accidents]|[accident]	*stem[accident]*

After stemming the next phase is keyword detection, the summarizer count the total number of words in the text sort the list of words with their frequency in the text and cross check the words synonym from static lexicon, if it exist or not. If the synonym of word does exist, the static lexicon helps not to count the same word as a different word.

Table 3.4: *List of OTS words*

Words	Frequency
are	11
is	17
a	16
Harry	14
Sally	13
Love	11
such	11
an	4
taxi	2
university	1
Chicago	1
…	

The idea behind OTS is that the important concept in an article are described with many of the same words, while redundant information uses less technical terms and is not related to the main subject of the article.

Once the word is sorted and counted as shown on table 3.4, the next process is to remove all the words that are common in the English language or stop words using a dictionary file. Example of such words: *"The", "a", "since", "after", "will".*

These words are words that don't inform us anything about the subject of the article. For instance, knowing that the word "School" is in a given text can tell us that the article might talk about the school, however knowing that the word "will" is in the text cannot teach us anything about the subject of the text. After removing the redundant words the new keyword list will look like lists on table 3.5:

Table 3.5: *List of OTS keywords*

Words	Frequency
Harry	14
Sally	13
Love	11
taxi	2
university	1
Chicago	1

From the list shown in table 3.5 they assume that the text talks about "Harry, Sally, and Love". So an important sentence in the text will be, a sentence that talks about Harry, Sally and Love and the others are ignored. Each sentence is given a grade based on the keywords in it. A line that holds many important words will be given a high grade. In addition, sentences got different score based in their position in document. To produce a 20% a summary the system print the top 20% sentences with the highest grade.

3.2.3.3 Performance of OTS

Because of a good performance of OTS, its approach has been citied in several academic publications and reputable journals. Among those publications, Yatsko and Vishnyakov [59] investigated the performance of OTS with other automatic text summarizer using D-score[1] like: - Subject Search Summarizer (SSS)[2], Copernic (COP) [3] and Essence (ESS) summarizers. As it is illustrated on (Fig. 3.1), among four automatic summarization systems (including the OTS) OTS scored 100% followed by subject search system scoring 97% based [59] or see [4].

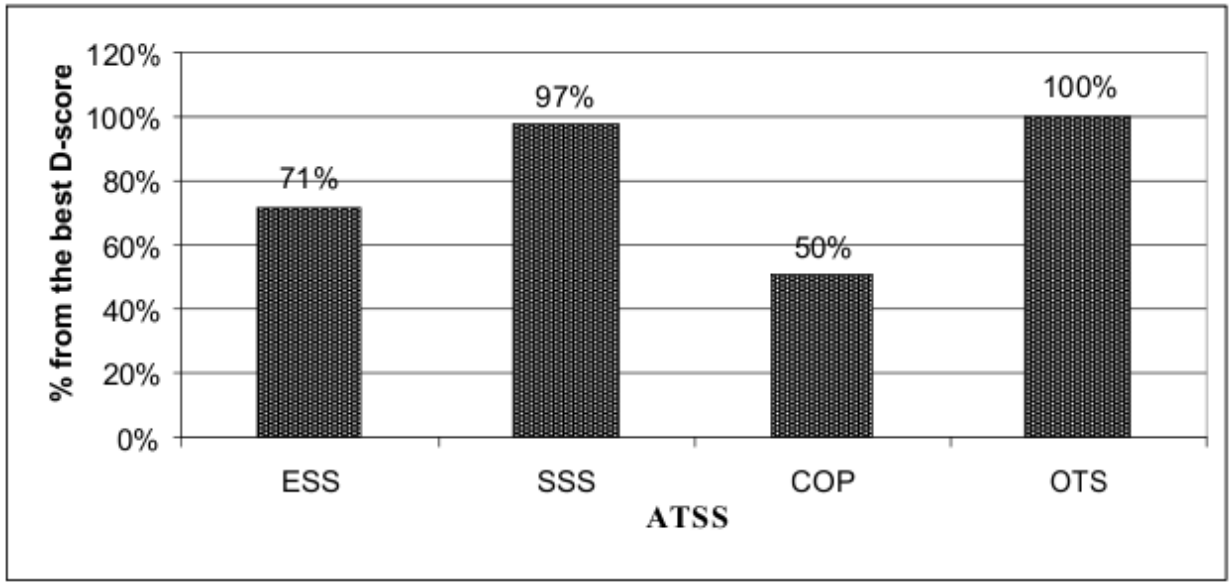

Figure 3.1: Comparison of performance of OTS with other summarizers. Source: from [59]

---

1. D-score: The difference score indicates the amount of change between two testing's. It is computed by subtracting the score on the first testing from the score on the second where d is the difference score (sometimes called change score or gain score), X is the first test score (sometimes called the baseline or pretest score), and Y is the second test score (sometimes called the posttest score). [60]
2. See: *http://www.freenew.net/windows/subject-search-summarizer-401/43113.htm*
3. *See: http://www.copernic.com/en/products/summarizer/*
4. *http://harvest.sourceforge.net/harvest-1.4.pl2-docs/node53.html*

3.2.4 Other Works in Automatic Extracting

The work entitled "New method in automatic extracting", describes new methods of automatically extracting documents for screening purposes, while in his previous work Edmundson [7] focused only on one component of sentence significance for extraction, namely, the presence of high-frequency content words (key words). This work is the extension of his previous work by adding features to extract the sentence. Therefore, the methods described here treat three additional components: pragmatic words (cue words); title and heading words; and structural indicators (sentence location). He selects those four features based on the following hypotheses:

- **Cue words:** Relevance of the summary is affected by pragmatic words ("significant", "hardly")
- **Key method:** High-frequency content words are positively relevant to extract the sentence.
- **Title method:** Author conceives the title as circumscribing the subject matter. Words of the title and headings are positively relevant.
- **Location Method:** Sentences occurring under headings are positively relevant.

The objective of this study was twofold: first, an extracting system to produce indicative extracts, and second, a research methodology to handle new text and new extracting criteria efficiently. In addition, the objective of the study is to make the result of the study more readily available.

The research methodology included a study of the abstracting behavior of humans, a general formulation of the abstracting problem and its relation to the problem of evaluation, a mathematical and logical study of the problem of assigning numerical weights to sentences, and a set of extracting experiments employing cyclic improvement.

Three different evaluation schemes applied to the resulting automatic extracts. Comparison of the automatic extracts and corresponding "target" extracts of 40 documents, which had not been used in the developmental research, showed that a mean of 44 percent of the sentences were co-selected. In addition, the mean similarity rating, in terms of a subjective evaluation of content by information type, was 66 percent. These are to be compared with a mean of 25 percent co- selected sentences and a mean of 34 percent similarity rating between target extracts and random extracts, respectively. Statistical comparison of the automatic and the corresponding target extracts for the documents used in the developmental phase showed a mean of 57 percent co-selected sentences with a standard deviation of 15 percent. A sentence-by-sentence analysis of the corresponding automatic and target extracts of 20 of these documents resulted in a judgment that 84 percent of the computer-selected sentences could be classified as extract worthy.

3.2.5 Summary

Throughout the world there are number of works has been done so far by employing different techniques and approaches. Form these approaches, statistical approach combined with IR is most employed one. This approach is preferred by different scholars because of it is not put in to consideration the sematic issues, i.e. it is based on defined features to extract the summary sentence. Hence, to gain better summarizer by employing different feature researcher got different performances. Of these features, sentence position, keyword frequency, cue phrase, title, name of number and events are most widely selected features.

3.3 Amharic Text Summarization

3.3.1 Automatic Amharic News Text Summarization

It is the first Amharic summarization system, which is a single document summarizer. Kamil [23] used surface level statistical features to assign weights to sentences. The highest scoring sentences extracted to form the summary. Seven features 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). The Performance evaluation result shows that the system registers 74.4% precision and 58% recall and respectively with 38.5% condensation rate. Beside on his finding , the researcher recommended development of good stemmer, availability of standard Amharic corpus, exhaustive lists of stop words, and the inclusion of more NLP, statistical and heuristic parameters.

3.3.2 The application of Machine learning Technique (Naive Bayes)

Using machine-learning technique Teferi [24] conducted his study based on Naïve Bayes classifier to create a single document summarization system. In his study, features like title, location, cue words and content words features examined, 480 news articles were used for experimentation. In his approach, the articles are categorized into training and testing/evaluation, 20 of news articles used for testing and the rest for training. The results of the analysis shows that precision of 75.00%, recall 74.90 % and classification accuracy of 86.03% in predicting the summary sentences. Teferi finally examined 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.

3.3.3 Automatic Text Summarization for Amharic Legal Judgments

The problem that legal experts in Ethiopia has been forced to spend their time on reading large volume documents to find relevant judgments for their cases which results significant delay of decision on cases, motivates Helen [29] to conduct her study on automatic text summarization for Amharic legal judgments. This study made her the first person who conducts the study on legal Amharic legal document. She employed statistical extraction techniques, by assigning weight to each sentence based on its location and the cue words/ phrases that it contains, to extract the highest weighted sentences .The system is tested for sample text and precision and recall measure is used for 20 % and 10% compression rate. The system calculates precision and recall. The system summary is compared against the human (ideal) summary. As a result, precision of the system summary is 33.9% and 39%; Precision of the random summary is 23% and 27%; recall of system summary is 57% and 50.5 %; recall of random summary is 46% is 38% for 20% and 10 % compression rate respectively.

3.3.4 Topic-based Amharic Text Summarization

This thesis deals with Amharic text summarization system based on statistical approaches called topic modeling to create the text summary. The proposed algorithms are language and domain independent and hence can be used for other local languages. Specifically, Eyob [19] propose to use the topic modeling approach of probabilistic latent semantic analysis (PLSA).

He shows that a principled use of the term by concept matrix that results from a PLSA model can help produce summaries that capture the main topics of a document, and propose six algorithms to help explore the use of the term by concept matrix. All of the algorithms have two common steps. In the first step, keywords of the document are selected using the term by concept matrix. In the second step, sentences that best contain the keywords are selected for

inclusion in the summary. To take advantage of the kind of texts he experiment with (news articles) the algorithms always select the first sentence of the document for inclusion in the summary.

He evaluated the proposed algorithms for precision/recall for summaries of 20%, 25% and 30% extraction rates. The best precision results he achieved are as follows: 0.45511 at 20%, 0.48499 at 25% and 0.52012 at 30% extraction rate. He also compared his systems with previous summarization methods that have been developed for other languages based on topic modeling approaches using his Amharic data set. The result shows that the proposed algorithms perform better at all extraction rates.

3.4 Afaan Oromo Text Summarization

3.4.1 Afaan Oromo news text summarizer

The first Afaan Oromo text summarizer study was conducted by Girma Debele [10] entitled Automatic Afaan Oromo news text summarizer. In his study, a generic automatic text summarizer for Afaan Oromo news text has been developed based upon the Open Text Summarizer (OTS) open source. OTS summarizes texts for different languages like; English, German, Spanish, Russian, Hebrew, Esperanto and other languages. Due to that, the approach of OTS and Girma's work is familiar in their approach; Girma was expected to adjust the OTS to let it work for Afaan Oromo, spent most of his time on adjusting the OTS code to make use of the Afaan Oromo lexicons and work for the Afaan 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. The components to generate summary of the document is discussed below. The system contains three basic subsystems and with additional one XML based lexicon.

1. Prepossessing

The preprocessing subsystem includes stop-word removal, stemming and parsing (breaking the input document in to a collection of sentences). For stop word removal, he used the Afaan Oromo stop-word compiled from different literature in addition to the stop word list prepared by Debela [54] using stemmer, a word is split into its stem and affix after stop-word removal. Affixes can be replaced by another affix or replaced by white space as per the rule that matches with it.

2. Sentence Ranking

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

With the tf (term frequency) method, the importance value (score) of a sentence (IVs) is given by:

$$IVs = \sum tf \qquad (3.1)$$

Where, IV is Importance Value based on term frequency and tf is Term frequency. To compute the positional value **(**score**)** of a sentence k , the first sentence of a document gets the highest score and the last sentence gets the lowest score. Therefore, to compute positional value of a sentence *s*, Girma used the total importance value (score) of a given sentence *s* (TIVs) by combining two parameters for sentence ranking.

$$TIVs = IVs * C \qquad (3.2)$$

Where, C is constant multiplicative factor for a sentence position. The value of C is 2 for first statement of first paragraph, 1.6 for first sentences of all other paragraphs. The rest all other sentences are weighed only by their term frequency score. TIV*s*, is total score of importance value of a sentence based on term frequency and position value.

3. Summary Generation

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

Girma used intrinsic summary evaluation method to evaluate the performance of his system. He proposed three methods to test the performance of his system using both objective and subjective manner. These e methods are: M1 that uses term frequency and position methods without Afaan Oromo stemmer and other lexicons (synonyms and abbreviations), M2 is a summarizer with combination of term frequency and position methods with Afaan 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 scores a values of 34%, 47% and 81% respectively i.e. M3 outperformed the two summarizers (M1 and M2) by 47% and 34 % . On the other way, the subjective evaluation result shows that the three summarizers' (M1, M2 and M3) performances of 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.

3.4.2 Critics

After reviewing the work, the following drawbacks/holes were identified:

1. Whenever newspaper is written to report the measurement or value of anything the author of the newspaper use number or digits. For instance; the authors of newspaper in order to report about: the height, distance, currency, year, date, month, statistical value in percentage, price of things, quantity of things , speed, weight or anything in the world; they used digits or number for indication. This indicates that the sentence, which contains number, is the one that caught newspaper readers' attention.

   However, Girma's works do not consider the sentence that contains numbers. This makes the generated summary loss the basic objective of the summary, which is the summary, must give the reader an accurate and complete idea of the contents of the source [18].

2. In addition, the second basic objective of the summary is to generate a short summary that should not be too short [18]. However, the previous work only considers sentence position and keyword frequency to extract the sentence from the document but he didn't consider the length of the sentence during extraction. For instance: too short sentence like proverbs may occur in the beginning of the paragraph, because of this the sentence gains highest score than other sentence. Consequently, his system extracts short phrase as a summary. In the same way, there is no mechanism to handle too long sentences. Therefore, the consequence of extracting long sentence as a summary may probably give a summary that has the same length to the original document.

3. Clearly newspaper reports the new events that happen across the world in different time and places or When did something happen? This indicates that the sentence that reports when things happened attracts reader

attention. However, Girma do not compute any mechanism to include the sentence that reports about events.

4. Edmunson [7] found that by adding only cue words on his previous two works which is done based on the sentence position and keyword frequency and during evaluation he found about 44% of the auto-extracts matched the manual extracted summary. This indicates that the advantage of adding cue phrase increases the quality of generated summary. But, Girma's work does not provide any mechanism to include the sentence which contains cue phrases.

5. Besides not adding, the features that have been mention in these critics, during the review of this document, no experimental method has been employed to assign weight for the two features used by him.

6. The compression ratio for the summary is not computed. Rather the previous summarizer generates the summary by using the compression ration inputted by the user.

3.5 Summary

Summarization of natural language texts involves identification of important information in the input text and generation of a paragraph or more that summarizes that information. Based on the critics that are addressed, this work focuses on summarization of different news articles using five features. Those features are including the two features, which are already incorporated in previous work; the sentence position and keyword frequency, and the additional four features that are incorporated in this study such as cue phrase, sentence length, name of numbers and name of date month and year. As we tried to review the most related works, features that we selected for this study increased the performance of the summarizer. We also strongly believed that these feature could also boosts the performance of Afaan Oromo news text summarizer. As weakens we did observe two major weakens. The first one is using small number of features for extraction degraded the performance of most the summarizers. Secondly, most of the works are not explicitly put the impact of each feature on the overall performance of the system. Hence, this work going to address features that contributes more to Afaan Oromo news text summarizer by conducting different experiments. On the other hand, it is also going to address suitable algorithms and techniques that can increases the performance of the Afaan Oromo news summarizer. Despite this, this work also reveals the performances difference between this work and the previous work.

Chapter Four: Afaan Oromo Language and overview of newspaper

4.1 Introduction

This section will discuss the overview of Afaan Oromo language such as: the writing system, sentence and word boundaries, numbers, short form of compound words and finally morphology seen at glance. At the end of this chapter about the newspaper have been discussed.

4.2 Afaan Oromo alphabet/Qube Afaan Oromo

The Oromo writing system is a modification to Latin writing system. Thus, the language shares many features with English writing system except some modification [54]. Until 1991, Geez script was used for writing Afaan Oromo documents. A Latin-based alphabet called Qubee Afaan Oromo has been adopted and became the official script of Afaan Oromo since 1991[1] , it consists of twenty-nine basic letters of which five are vowels (a, e, i, o, u), twenty-four are consonants, out of which five are pair letters and fall together (a combination of two consonant characters such as "*ny*"). The basic alphabet in Afaan Oromo does not contain p, v and z. This is because there are no native words in Afaan Oromo that are formed from these characters. However, in writing Afaan Oromo language they are used to refer to foreign words such as "polisii" (police) [54].

A a	B b	C c	CH ch	D d	DH dh	E e	F f	G g	H h	I i
[a]	[b]	[ʧʼ]	[ʧ]	[d]	[ɗ]	[e]	[f]	[g]	[h]	[i]
J j	K k	L l	M m	N n	NY ny	O o	P p	PH ph	Q q	R r
[ʤ]	[k]	[l]	[m]	[n]	[ɲ]	[o]	[p]	[pʼ]	[kʼ]	[r]
S s	SH sh	T t	U u	V v	W w	X x	Y y	Z z		
[s]	[ʃ]	[t]	[u]	[v]	[w]	[tʼ]	[j]	[z]		

Figure 4.1: *Afaan Oromo Alphabet/ Qube Afaan Oromoo*

4.3 Word and sentence boundaries

4.3.1 Words

The word is the smallest unit of a language. There are different methods for separating words from each other. However, most of the world language including English use the blank character (space) shows the end of a word. Some long words are been cut in written form (abbreviation), with the symbols "/",".", and therefore this symbol should not determine a word boundary. The usual parenthesis, brackets, quotes, all kinds of marks, are being used to show a word boundary in Afaan Oromo [54].

4.3.2 Sentence

Afaan Oromo sentence is terminated like English and other languages that follow Latin writing system [54]. That means, the full stop (.) in statement, the question mark (?) in interrogative and the exclamation mark (!) in command and exclamatory sentences marks the end of a sentence and the comma (,) which separates listing in a sentence and the semi colon is to mark a break that is stronger than a comma but not as final as a full stop balance [16].

4.4 Afaan Oromo numbers

As many of world language Afaan Oromo Language use European Numbers/Indian-Arabic numbers. In Afaan Oromo numbers can occur in the form of cardinal numbers, ordinal numbers and nominal numbers [63].

a. Afaan Oromo cardinal number convey the "how many" they're also known as "counting numbers," because they show quantity. Here are some examples:

Table 4.1: *Cardinal numbers*

English	Afaan Oromo
Zero	zeroo
One	tokko
Two	lama
Three	sadii
Four	afur
Five	shan
Six	Ja'a
Seven	torba
Eight	saddet
Nine	sagal

b. An Ordinal Number is a number that tells the position of something in a list. Some examples are listed below. In Afaan Oromo, the ordinal numbers are formed from the cardinal numbers by suffixing the suffix {affaa}[54].

Table 4.2: *Ordinal numbers*

English	Afaan Oromo
1st | first	1ffaa | tokkoffaa
2nd |Second	2ffaa | lammaffaa
3rd |Third	3ffaa | sadaffaa
4th |Fourth	4ffaa | arfaffaa
5th | Fifth	5ffaa | shanaffaa

c. A Nominal Number is a number used only as a name, or to identify something (not as an actual value or position. For nominal number representation Afaan Oromo uses decimal digits from 0,1,2,3,4,5,6,7,8,9 like English language [63].

For Example:

- The number on the back of a footballer ("10")
- a postal code ("91210")
- a model number ("380")

4.5 Short forms of compound words

Abbreviation or Short form representation of compound words is common in Afaan Oromo similar to other languages like English and other world languages. In such representation, the use of forward slash (/) is much common although a dot or period (.) can be used alternatively. For example below there are list Afaan Oromo short form.

k.k.f *"Kan kana fakkaatan" or*

k/k/f *"Kan kana fakkaatan"*

4.6 Morphology

Since the automatic text summarizer needs the stemmer, the researchers investigate that it is necessary to study about the morphological structure of the language. Because morphology is a branch of linguistic, that studies and describes the internal structure of words and how words formed in a language.

There are two branches of morphology: Inflectional and derivational. Inflectional morphology deals with combination of a word stem with a grammatical morpheme in the same word class. In inflectional morphology, inflectional morphemes, morphemes that serve a purely grammatical function, which never create a new word but only a different form of the same word, are added in words. However, derivational morphology deals with combination of a word stem with a grammatical morpheme that yields different word class. Thus, in derivational morphology, there are methods of forming new lexemes from already existing ones by affixing derivational morphemes, morphemes that change the meaning or lexical category of the words to which they are attached Afaan Oromo morphology. Like in a number of other African and Ethiopian languages, Afaan Oromo has a very complex and rich morphology [10, 54, 56].

Morphemes is the smallest individually meaningful elements in the utterances of language. They are the morphological building blocks of words. In general, morpheme is smallest linguistic unit that has a meaning or grammatical function. It is also called minimal meaning-bearing unit in a language. Basically, language morphemes are categorized in to two: free and bound morphemes. Free morpheme can stand as a word on its own whereas bound morpheme does not occur as a word on its own.

There are wide ranges of word formation processes in Afaan Oromo; the major categories are nouns, pronouns and determinants, case and relational concepts, functional words, verb and adverbs. Almost all Oromo nouns in a given text have person, number, gender and possession markers that are concatenated and affixed to a stem or singular noun form. Likewise, Afaan Oromo pronouns and determinants have number, gender, adjectives, and quantifier markers in Oromo nouns. The researcher has thoroughly discussed Afaan Oromo nouns in detail. Because this study more concerned on noun, since it deals one name of events, cue phrase and numbers.

4.6.1 Noun

4.6.1.1 Gender

Gender is one category of nouns, pronouns and adjectives into masculine and feminine and some language neuter based on whether a noun considered as male, female or without sex respectively.

Gender is of two types: natural and grammatical. Natural gender refers to natural sex of animate things while grammatical indicate morphological marked gender In Afan Oromo nouns, both natural and grammatical gender are marked as follow.

1. Natural gender

Masculine	Transliteration	Feminine	Transliteration
/abbaa/'	'father'	/haaɗa/	'mother'
/ɗhirsa/	'husband'	/niitii/	'wife'

2. Grammatical gender

The suffix / ttii/ and /tuu/ are used as feminine gender marker whereas /-aa/ marks masculine gender in Afaan Oromo .

Masculine	Transliteration	Feminine	Transliteration
/Barataa /	'student'	'Bara tuu'	'student'
/jaldeessa/	'male monkey'	' jaldeettii'	' female monkey'

Limited group of noun differ by using different suffixes for mascular and feminine form. Frequent gender markers in Afaan Oromo include -eessa/-eettii, -a/-ttii or aa/tuu.

Example:

Afaan Oromo	Construction	Transliteration
obboleessa	obbol + eessa	male, brother
obboleettii	obbol + eettii	female, sister
beekaa	beek + aa	male, knowledgeable
beektuu	beek + tuu	female, knowledgeable

There are also suffixes like -a, -e that indicate present and past form of masculine markers respectively. -ti and -tii for present feminine marker and -te past tense marker, -du for making adjective form) [65]. For Example *Chaltu demte* (Chalu gone) and *-tii* can also show feminine gender in the following statement. *Isheen malif artii?* (Why does she is angry?).

4.6.1.2 Number

Number is one of the grammatical categories for which nouns are inflected. Nouns have the capacity to be inflected for number. Number indicates a singular and plural form of noun.

Singular noun	**Transliteration**	**plural**	**Transliteratio**
/barataa/	'student'	/baratt-oota	'students'
/farda/	'horse '	/farde-en/	' horses
/gangee/	'mule'	/ gaango olii/	'mules'
/kitaaba/	' book'	/' kitaabo lee/	'books'
/bineensa/	' animal'	/ bineensa wwan/	'animals'

Afaan Oromo has different suffixes to form the plural of a noun. The use of different suffixes differs from dialects to dialects. Majority of noun plural froms were formed by using the suffix -

(o)ota, followed by -lee,-wwan,-een,-olii,-olee and -a(n) [65].

Afaan Oromo	**Transliteration**	**Afaan Oromo**	**Transliteration**
-o(ota) hiriyoota	*Ferinds*	*-aan ilmaan*	*Children*
-wwan hojiiwwan	*Works*	*-olii/olee Jaarsolii/jaarsolee*	*Elders*
-lee gaaffilee	*Questions*	*-een fardeen*	*Horses*

4.6.1.3 Definiteness

Definiteness is a grammatical category used for distinguishing noun phrases according to weather their reference in a given context is presumed to uniquely identifiable. In Afan Oromo language demonstrative pronouns like *"kun"* (this), sun (that) are used to express definiteness.

Afaan Oromo	**Transliteration**
mani kun	*this house (Subject)*
mana kana	*this house (Object)*
mani sun	*that house (Subject)*
mana sana	*that house (Object)*

In Afaan Oromo dialects the suffix -icha for male and -ittii(n) for female and for undermining that usually has a singularize function is used where other languages would use a definite article.

For example:

Afaan Oromo	**Transliteration**	**Afaan Oromo**	**Transliteration**
Afaanichi	*this language*	*Afaanicha*	*the/this language*
Jaartittiin	*that old lady*	*Jaarsichi*	*the old man*
jaarsicha	*the old man*	*Jaartittii*	*the old lady*
Re`ettiin	*that got*	*Re`ettii*	*that/the got*

4.6.1.4 Derived noun forms

Derivation is the morphological process by which new words formed from other words or roots. According to [67] states that derivation as a morphological process that changes one word (or lexeme) into another. Nominal are derived from other nominal, adjectival stem and verbal stems. Afan Oromo is very productive in word formation by different means. One method is the use of different derivational suffixes. The other method is the formation of compounds [65].

i. Derivational suffixes

Derivational suffixes are added to the root or stem of the word. From derived verbal stem and adjectives may be formed by means of derivational suffixes.

The following suffixes play an important role in Afaan Oromo word derivation. They are -eenya, -ina, -ummaa, -annoo, -ii, -ee, -a, -iinsa, -aa,-i(tii), -umsa, -oota, -aata, and –ooma.

Examples:

-eenya jabeenya(strength)

Fakkeenya(example)

-ina guddina(growth)

Dheerina(length)

-ummaa haxxummaa(haxxummaa)

-(t)tii fakkaattii(image)

Butii(hijacking)

-ee dhibee(problem)

-aa amantaa(belief)

-iinsa bulchiinsa(administration)

-(o) oma firooma (friendship)

-tuu furtuu(key)

barattuu(student)

-oota barnoota(education)

-annoo yaadannoo(rememberance)

-umsa barumsa(science of)

-ii hawwii(ambition)

Beekumsa(knowledge)

Abbaltii(intension)

-a yaada(thought)

ii. Compound words

The use of genitive constructions is a very old method of forming compound nouns, as traditional titles shown.

abbaa duulaa *traditional Oromo minister of war*

Similarly, Afaan Oromo adjectives have case, person, number, gender, and possession markers similar to Afaan Oromo nouns. Afaan Oromo verbs are also highly inflected for gender, person, number, tenses, voice, and transitivity. Furthermore, prepositions, postpositions and article markers are often indicated through affixes in Afaan Oromo [58] [59]. Due to the extensive inflectional and derivational features of Afaan Oromo are presenting various challenges for text processing and NLP (Natural Language Processing) in the language.

4.7 Overview of newspaper

A newspaper is a publication that is issued daily or weekly and includes local and international news stories, advertisements, announcements, opinions, cartoons, sports news and television listings. The advancements of computer technology let the newspapers continue to be an important aspect of everyday life. Newspapers can be published as often as they like via the Web or electronic editions, with limited added costs. The first online newspapers appear in 1994 [60].

The Basic functions of newspaper are;

- **To inform:-** designed to make its reader aware of contemporary issues and events, a newspaper reports significant, interesting, and exciting events to enhance the people's level of awareness and understanding by pointing out the What, the Who, the When, the Where , the Why, and the how of a particular event.
- **To educate: -** it provides instructional and educational articles that are meant to teach and make the readers understand the significant events as well as the issues resulting from them.
- **To entertain: -** for a lot of people around the world, this function ranks first. Many say that they read the comic page first for a very simple reason- to have some fun and forget the usual worries on meets daily. Special articles and human-interest features stories are all meant for light reading and enjoyment of readers.

4.7.1 Properties of newspaper

Newspaper has structure that differs from other type of articles or written documents. Those properties are;

- ☞ The first paragraph gives the answers to the most important of the 5 W's and H. The second paragraph tells the rest of the 5 Ws if they were not included in the lead [14].
- ☞ The rest of the paragraphs elaborates on the information given in the opening and gives more information and details. [14]
- ☞ Background information is included if it is giving new information to a story that had been printed previously in the newspaper. Sometimes it gives information which is necessary to understand the story.
- ☞ A quotation or a statement about the news story is often included in order to explain the importance of the story.
- ☞ Details are provided about the story and are organized into paragraphs. Each paragraph provides one aspect of the story and the details are arranged in order of importance.

4.7.2 Summarizing newspaper article

To summarize the newspaper article it has its own rule and procedure therefore any author of the newspaper must put in consideration points that are listed below [58]:

1. Find the "5 W's": who, what, when, where and why the detail shown under section 4.7.1 before. Hence, remember to put these facts into your own words.

2. Add the main idea(s). The author of the newspaper article wrote the article to get a message across and to create a sentiment among readers and that message is the main idea. The main idea has a direct correlation with the "why" of the article because it is an extension of it. No more than three sentences should be needed to summarize the main idea. Sometimes a

newspaper article may have multiple main ideas and if that is the case, keep the description of each brief.

3. Include supporting details. Once you have read the newspaper article over at least twice, you should have an understanding of the information that is essential and which details were just added for creative effect. The details that first must be added are those that are imperative to the understanding of the article, like the job position of the subject or how many years of research has gone into a new discovery. Next, those details that give help with imagery can be added.
4. Finish your summary with a concluding sentence. You do not have to end where the article ends, just where the story ends.

Chapter Five: Design and Implementation

5.1 Introduction

This chapter consists three major sections. The first section with deals the design of a system. It explores the architecture of proposed system, design requirements and techniques applied to build design requirements in detail. The second section, discusses the implementation of the system. Accordingly, this section illustrates on a workflow of the system and shows how each module in a system has been build. Additionally, the techniques and algorithms to build the system are also covered in this section. The last section, shows the scree shot of a user interface of the developed system.

5.2 Design

The design of Afaan Oromo text summarizer is partially adopted from OTS and the work of Girma [10]. Moreover, the architecture of this work is familiar with most of automatic text summarization system; it shares the three stages of automatic text summarization phases: the preprocessing, processing and summarizer phase (Shown on Fig. 5.1).

This, architecture is named Afaan Oromo text Summarizer (AOTS). It contains different modules; these modules are categorized into two; knowledge base modules and process modules. Knowledge base modules are a technology used in this work to store complex structured and unstructured information about Afaan Oromo language, that are accessed by process module. Those modules are; abbreviation, stop word, affixes, synonym, keyword, cue phrase, name of events and name of numbers knowledge base. Process modules, are those modules that accept input from other modules or knowledge base and give the output result. Those modules are; tokenize, sentence length handler, stop word remover, stemmer, keyword counter, SWBSP, SWBCP, SWBKW, SWBEv, SWBNum, SS, SCR and SSG (Fig. 5.1).

To develop the system, two basic categories of design requirements are specified in this work, knowledge base and validation data. These, requirements are the prior task to be prepared, before to start to develop the system. The detail of these design requirements discussed, under section 5.2.1 and 5.2.2.

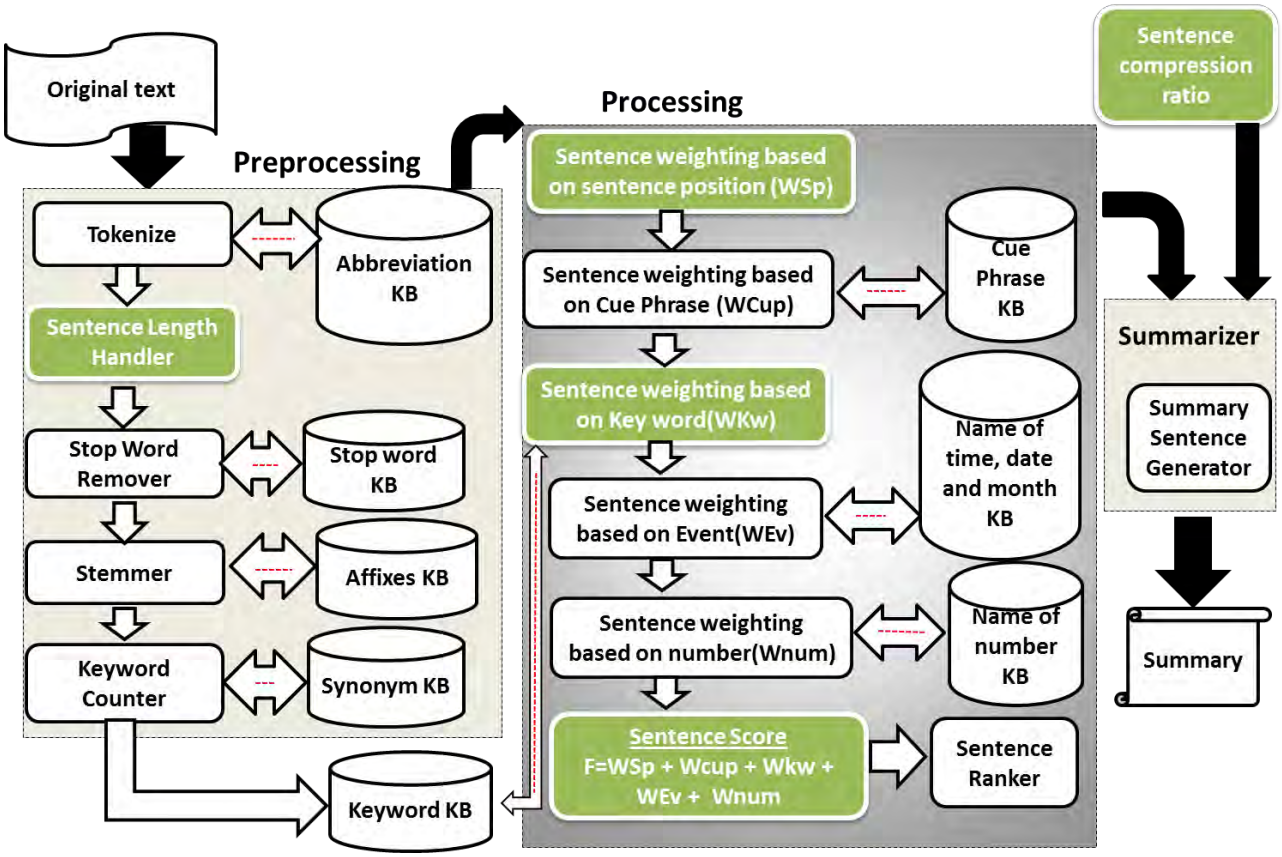

Figure 5.1: AOTS Architecture

5.2.1 Knowledge base module Preparation

Under this section, the preparation of information, which stored in each knowledge base, has been discussed. Information that stored in KB discussed as corpus. Thoroughly, the methods of preparation of these corpuses are also discussed.

5.2.1.1 Corpus of Afaan Oromo cue phrase

Cue phrases are words and phrases that are used to explicitly signal discourse structure in both text and speech [45]. As Edmunson [7] presented in his study, phrases such as: *"in summary", "in conclusion"* and superlatives such as *"the best"*, *"the most important"* are a good indicators of important content. He named these phrases a cue phrases. In his work, to prepare these phrases, he manually collected it and the sentence that contains this phrase given more weight than other sentences.

In similar way, for current study English cue phrase was manually collected from different sources and translated to Afaan Oromo. Because of Afaan Oromo has different dialects, context and interpretation, a single English cue phrases have more than one equivalent Afaan Oromo meaning. Therefore, to handle this, six subjects are chosen from Addis Ababa University and Adama University linguist department. Out of six subjects, two of them were PhD candidate at Addis Ababa University and the others 4 are instructors from Adama University. The collected English words are distributed in a form of questioner for all subjects. All subjects are expected to provide a meaning of Afaan Oromo word for each English word.

The translation process was not stop on this, to make the translation more standard and meaningful using focus group discussion 7 Dilla University Afaan Oromo department instructors translate each English cue phrases and check the previously translated phrase by the former subject's trough focus group discussion. Finally, the translated cue phrases are combined together and checked by professional language translator. From those results, the collected 350 English cue phrases were translated to 729 Afaan Oromo cue phrases. From 729 translated phrase listed on Appendix A, some of the sample of cue phrase is show below in table 5.1. After the cue phrases were compiled in to Afaan Oromo; it is stored into cue phrase knowledge base (Fig. 5.1) for implementation.

Table 5.1 : *Sample Afaan Oromo cue phrase*

Terms	Transliteration
suddenly	akkatasa|akka daguu|battalumatti
summing up	walitti qabaattii|dimshaashumatti
In short	gababumatii
in the first place	Hunda dura/hunda caalaa| Tokkoffarratti
consequently	kanaafuu|waan ta'eef

5.2.1.2 Corpus of Afaan Oromo stop-words

In computing, stop words are words that are filtered out prior to, or after, processing of natural language system or text documents. R.V.V Murali Kirshna and Ch. Satyananda [52] stated in their study that stop word remover play important role in building search engines and text summarization system, they help in filtering useful information from original text, i.e it is also necessary in text summarization.

In his work, Girma [10] used a list of stop words, which was compiled by Debela [54]. Debela collected and compiled lists of stop word based on information in "A Grammatical sketch of Written Oromo "[66] and "*Caasluga Afaan Oromoo, Jildi I* "[65], he found 99 stop words. However, Girma [10] found that list of stop words complied by Debela was not enough, he adds additional stop-words which are not included by Debela from the book titled: *"A Grammatical sketch of written Oromo"* [66] and added to the list and the total number of stop-words reaches 120. Hence, in this work list of stop words collected and complied by Girma is used. The lists of stop words are stored into stop word knowledge base (Fig. 5.1), for implementation. Randomly selected sample stop-words are shown in table 5.2 and the entire list is available in Appendix-D.

Table 5.2: *Sample Afaan Oromo stop words*

Terms	Transliteration
Waan	*Because cause*
Garuu	*But*
Sun	*That*
Ituu	*If*
Akka	*such as, like,*

5.2.1.3 Corpus of Afaan Oromo abbreviations

Abbreviation is a shortened form of words, etc. The main purpose of building the corpus of Afaan Oromo abbreviations is to prevent false detection of sentence boundaries. During segmentation (tokenization) and splitting text into words (tokens) in order to correctly identify the boundaries between clauses, phrases or sentences abbreviation corpus is necessary. In this study list of abbreviation that was collected by Girma [10] and additional abbreviations that are collected from textbooks and Afaan Oromo news portal is used.

For implementation list of abbreviation are integrated into abbreviation knowledge base (Fig. 1), after they are collected. Sample list of abbreviation from 22 abbreviations listed on Appendix C is shown in table 5.3 listed.

Table 5.3: *Sample Afaan Oromo abbreviations*

Abbreviations	Transliteration
Add.	addee
Mil.	Miliyoona
Sad.	Sadaasa
W.B.	Waree booda

5.2.1.4 Corpus of Afaan Oromo synonyms

A synonym is a word or phrase that has the same or nearly the same meaning. One of the limitations of extractive text summarization method as investigated by Edmonson [7] and different scholars is a challenge of handling of synonym words. Therefore, they recommend that there must be some mechanism to handle synonyms words during keyword counting. Storing list of synonym has been taken as one of the solution to handle synonym words.

Furthermore, during keyword frequency counting, no word which has the same meaning counted as different word. For instance; "**keenya**" *ours* and "**keenna**" *ours* refers to the same kind possessiveness pronoun in Afaan Oromo, therefore this word is counted as one word.

Due to this reason, corpus of Afaan Oromo synonyms were collected from Afaan Oromo Dictionary entitled "Galmee Jechota Afaan Oromo" and literatures (includes list of synonym list collected by Girma [10]); integrated in to synonyms knowledge base (Fig 5.1). Sample list of synonym that are collected is shown on appendix B and sample list of them is shown on table 5.4.

Table 5.4: *Sample Afaan Oromo synonym*

Term	Synonymy	Translit
Aankoo	*Jaldeessa*	*Monkey*
Anqaaquu	*Hanqaaquu*	*Egg*
Mi'a	*Meeshaa*	*Material*
Jijjiiruu	*Diddiiruu*	*Change*
Herreguu	*Yaaduu*	*Think*

5.2.1.5 Corpus of Afaan Oromo suffix

OTS and its customized OOTS implement suffix stripping stemmer; similarly, in current study we used the same approach.

Girma [10] used the suffixes that are collected by Debela [54] for suffix stripping. The number of suffixes gather by Debele is only 76. However, Abebe Abeshu [56] in his work titled "automatic morphological synthesizer for Afaan Oromo" he found three categories of Afaan Oromo suffixes: derivational, inflectional and attached suffixes. He grouped these three suffixes into to verb and noun suffixes, and he founded 249 suffixes. This is greater in number than the suffixes gathered by Girma. Hence, in this work suffixes gathered by Girma and additional suffix gathered by Abebe [56] used and integrated in affix knowledge base (Fig 5.1). From 161 verb suffixes and 88 noun suffixes listed in Appendix G and H, sample of them are listed in table 5.5.

Table 5.5: *Verb and noun suffix*

Verb suffix	Noun suffix
adhee	illee
adhuu	irraa
adhuu	irraahille
ama	irraahis
amaa	irraahuu

5.2.1.6 Corpus of name of Events in Afaan Oromo

Like any other languages, Afaan Oromo language has name for time, dates and months. In order to extract sentence which contain name of time, date and month such a corpus is necessary. These lists are collected from different Afaan Oromo books, news portal and websites; and integrated in name of time, date and moths knowledge base (Fig. 5.1). From the list of time, date and month listed on appendix E, sample of them are listed below in table 5.6.

Table 5.6*: Sample Afaan Oromo time, date and month*

Afaan Oromo names	Transliteration
Amajjii	January
Kamisa	Thursday
Ganama	Morning
Jimaata	Friday

5.2.1.7 Corpus of Afaan Oromo numbers

The sole purpose of this corpus is to give greater weight for a sentence that contains number. As discussed in chapter four, Afaan Oromo numbers exist in three different format cardinal, ordinal and numeral. Hence, list of those numbers are collected from textbooks, journals and news portal; they are stored in name of number knowledge base (Fig. 5.1). List of those numbers are listed on appendix F and sample of them is show in table 5.7.

Table 5.7: *Sample list of Afaan Oromo number*

Name of number in Afaan Oromo	Transliteration
Tokko	one
Kudhan	Ten
Digdama	Twenty
Soddoma	Thirty
Miliyoona	Million

5.2.2 Validation data preparation and analysis

In order to tune a parameter validation data is necessary for extractive type of text summarizer. Table 5.8 shows the detail description about 20 topics gathered from different sources. Then, to prepare validation data after the topics are gathered and compiled; it is distributed to 60 subjects in a form questioner as shown on appendix I. After the articles are processed by the subjects, we organized data that obtained from the subject as show on appendix J. In this appendix if the sentence is, underline "1" represents it and if the sentence is not underlined "0" represents it.

From tables listed in appendix J, topic 1 results is taken for illustration, and shown in table 5.9. In this table 5.9, the row heading "subject" shows list of subjects those underline each topic and the row heading "sentence in paragraph" indicate the location of the sentence located in the in a paragraph. The row heading "paragraph", indicates the paragraph location in a document.

For example, let take a look the result of sentence that is completed by subj2. For instance, the result of Subject 2, paragraph 1, sentence location in paragraph 2 is 0.

Table 5.8: *Statistics of the training corpus*

Topics	Number of words	Number of sentence	Number of paragraph	CR
Topic 1	266	13	5	41.73%
Topic 2	214	12	4	35.23%
Topic 3	256	9	8	22.27%
Topic 4	275	8	6	48.5%
Topic 5	248	10	3	42.74%
Topic 6	162	8	2	53.01%
Topic 7	244	9	4	41.80%
Topic 8	255	10	3	40.07%
Topic 9	247	7	7	36.03%
Topic 10	433	13	6	22.6%
Topic 11	172	8	3	43.02%
Topic 12	251	14	6	67.20%
Topic 13	276	14	7	65.90%
Topic 14	272	15	2	45.89%
Topic 15	321	15	3	67.93%
Topic 16	263	9	9	49.80%
Topic 17	276	9	3	34.30%
Topic 18	261	10	6	37.55%
Topic 19	237	9	4	67.17%
Topic 20	248	7	7	55.65%
Average	**258.85**	**10.47**	**4.9**	**45.92%**

The main reason to adjust and fill the underlined sentence in this manner (Table 5.9) is to make it more flexible during the experiment, especially during computing of precision and recall. In similar way, such way of keeping data collected from the respondents are also used during test data preparation.

Table 5.9 *Sample of sentences underlined by the subjects in a document*

	Sentence Position													
	Paragraph	**1**		**2**		**3**				**4**			**5**	
	Sentence location in paragraph	**1**	**2**	**1**	**2**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**1**	**2**
Subjects	**1**	1	0	1	1	0	1	0	1	1	1	0	0	0
	2	1	0	1	1	0	1	0	0	0	0	0	0	1
	3	1	0	1	0	0	0	0	1	1	1	0	1	0
	Decision	*1*	*0*	*1*	*1*	*0*	*1*	*0*	*1*	*1*	*1*	*0*	*0*	*0*

1= underlined *0=not underlined*

The row heading "Decision" shows the decision whether the sentence is taken as selected or not selected. Therefore, when the sentence is selected by two or three subjects the sentence is considered as a selected sentence and the decision is 1. On the other hand, if the sentence is not selected by two or three subject the sentence considered as not selected, and the decision is 0. For example, the decision of sentence at position paragraph 1, sentence position 1 is 1, i.e. the sentence is selected.

5.2 Implementation

To mention it again, the development tool that used in this experiment is OTS C# version. In order to, work with this tool the first task is customizing and modifying OTS to support for Afaan Oromo language, i.e. during modification the rule are constructed and corpuses are stored in each KB (Fig. 5.1). The next task is to adjust OTS according to newly proposed algorithms, and then modification to the source code is accomplished. Finally, new source code generated and incorporated as new component to the tool; at the end, AOTS is developed.

Under this section, the workflow of proposed architecture (Fig 5.1), algorithms and different experiments that were accomplished to build each module has been discussed in detail. The workflow the system discussed as phase I, phase II and Phase III.

5.3.1 Phase I: Preprocessing

In the preprocessing phase, a structured representation of the original text is obtained [53]. Before, the summarizer start any kind of task the system performs a preprocessing task on the original input document (Fig. 5.1). Therefore, several preprocessing modules are activated. Each module performs certain preprocessing tasks such as; sentence length handling, tokenizing or sentence segmenting, stemming, stop word removing and key word counting.

Finally, this phase generate structured representational format of the original text. In short, the input text is made ready to be processed by the next phase. Thus, in this section, the detail how each module built is discussed.

5.3.1.1 Sentence Segmentation (Tokenize) Module

In other word, sentence segmentation is also known as sentence boundary identification or Tokenize. This module is considered as the first preprocessing module in many NLP applications. This module identifies sentence boundaries between clauses, phrases or sentences, by first splitting the all text into words (tokens) using the list and rules stored in abbreviation knowledge base (Fig. 5.1); and generate array of sentence and token of words. This module perform sentence segmentation task based on the algorithm that show below (Fig. 5.2);

```
Start
        Input text document
        detect punctuation symbol ()
         Split the document in to words()
         if the word is abbreviation
                Remove the word from the list()
         Create array of a sentence()
        Output
         list of each sentence
        list of words
Stop
```

Figure 5.2: *Segmentation algorithm*

5.3.1.2 Sentence Length Handler Module

Even if one of the major goals of automatic summarizer is to generate short form of the given document, there is always a risk that too long and too short sentences ranked higher and extracted. As a result, these two problems reduce the quality of the summarizer. Hence, this module handle too short and too long sentence, by accepting array of sentence from tokenize (Fig 5.1).

Therefore, in order to build this module small experiments were conducted in this study. The experiment is conducted based on the 20 topics that are prepared for system validation purpose. These topics are already mentioned in table 5.8. In order to detect too short and too long sentence, sentence length was clustered into equal range as shown in table 5.10. The range was made into seven difference based on the validation data. The validation data has the maximum sentence length size, 70 words and minimum 5 words. Based on this fact ten different ranges where identified.

In table 5.10, the column heading "*Range*" indicates the defined range of a sentence. For instance: a range (1, 7] indicate the sentence that contain a number of words from one to seven. The column heading "*Number of sentence in each range*", indicates the total number of sentence that have the same length with in a given range and the column heading "*Probability*", indicates the probability of sentence that exist in each range from a total of 209 sentence and it is computed based on equation 5.1.

$$probability = \frac{Number\ of\ sentence\ in\ each\ range}{Total\ number\ of\ sentence} * 100\% \qquad (5.1)$$

For example, the probability of the sentence exist in a range of [1-7] is (1/209) * 100% = 0.48%. Based on the fact show in table 5.10 to detect too short and too long sentence, the researchers ignore the sentence, which has a probability result below 5%. The sentences that has this result are ignored based on the

hypothesis that too short and too long sentence occur less frequently than normal length sentence [30, 31, 32, 33].

Table 5. 10: *Sentences with in a range. Column "Percent" shows the percentage in each range*

Range	Number of sentence in each range	Probability
[1,7]	1	0.48%
(7,14]	36	17.22%
(14,21]	60	28.71%
(21,28]	45	21.53%
(28,35]	39	18.66%
(35,42]	15	7.18%
(42,49]	7	3.35%
(49,56]	2	0.96%
(56,63]	3	1.44%
(63,70]	1	0.48%
Total	**209**	**100%**

Therefore, from a result shown in Table 5.10, we found that the sentences, which are below seven words, are too short sentences and which are greater than 42 words are too long sentence. Based on this result the following algorithm (Fig. 5.3) was designed, and the sentence length handling module is design accordingly.

```
Start
    Input array of sentence
    For each sentence in a paragraph
        If sentence length <= 7 or sentence length is >= 42
            Remove the sentence form the list ()
Output normal sentence
End
```

Figure 5.3: *Sentence length handling algorithm*

5.3.1.3 Stop Word Remover module

Stop word filtering/removal is a common technique used to counter words that does not contribute to the description of the documents. Furthermore, as presented in many IR and NLP studies a stop words is a word that occurs frequently and might miss lead the keyword counter. For example: in Afaan Oromo, words like: *"kun", "fi"* and *"koo"* contribute very little value to the description of the document, and in many cases instead of a description, they add noise.

Therefore, this module removes the stop words from the input document after accepting list of words from tokenize module. This module removes stop words from segmented sentence by finding the match with list of stop word stored in stop word knowledge base (Fig. 5.1). The algorithm defined in this study to handle stop words shown on Fig 5.4.

```
Begin
    Input segmented text
    For each sentence in each paragraph
        If a word in each sentence is stop word
            Remove the word from the sentence
    End for
    Output list of sentences without stop word
Stop
```

Figure 5.4: *Stop word removal algorithm*

5.3.1.4 Stemmer Module

Stemmer in NLP and IR is an attempt to reduce a word to a common root or stem form. Subsequently, the words in from a given document represented by one lexical string (term) rather than by the original word forms. During counting of frequency of words in sentences, the same word that has several morphological variants miss guides the keyword counter. Hence, a stemmer is necessary to split words into their stem and affix.

The main reason to stem words after stop word removal is to reduce the computational time of AOTS, by stemming only nonstop words rather than stemming the whole words in the document.

In OOTS, Girma adopt Debela's [54] stemmer rules. The stemmer named porter algorithm for Afaan Oromo by Debela and later Girma named it lightweight algorithm. Since AOTS developed based on OTS and OOTS, this algorithm selected as stemmer in this work (Fig 5.5). Then, the algorithm is adopted and integrated in to AOTS. The main reason that this algorithm is preferred in this work is, for two reasons stated by Debela [54]:

- ☞ The first reason is he achieves 96 percent accuracy. Because, this stemmer algorithm is not only removes the suffix from the words, it replaces by another affix or replaced by white space as per the rule it matches with the rule. The detail description if this algorithm discussed in Debelas [54] Work.
- ☞ The second reason is porter algorithm is well known algorithm, which is frequently used in IR systems.

```
Start
do
READ the next word to be stemmed
If word matches with one of the rules
        Remove the suffix and do the necessary adjustments
Else
Stop processing
End if
While not end of words
Output stemmed word
Stop
```

Figure 5.5: *Stemmer pseudo-code adopted from Debela*

5.3.1.5 Keyword Counter Module

This module count nonstop words like; noun, verb, adverb and adjective in each sentence and the most frequent one is considered as a keyword. This module, accept two inputs, one from stemmer module and the other is from synonym knowledge base (Fig. 5.1). From the stemmer module, it gets stemmed word and from synonym KB, it gets a synonym of a word whether synonym of word exists or not. This mechanism is to handle not the word that has the same meaning counted as a different word. Finally, after counting the frequency of each word in document, this module generates another knowledge base known as keyword knowledge base, which contains list of keywords with their frequency in descending order. The same algorithm that was applied in OOTS and OTS for keyword counting is also used in this work shown on (Fig. 5.6).

```
Start
    Input stemmed and nonstop words
    For each word in document
            If word has synonym
               count = count
            else
               count ++
     End for
     Output list words with their frequency
Stop
```

Figure 5.6: *key word counting algorithm*

5.3.2 Phase II: Processing

In this phase, an algorithm transforms the input document to candidate summary sentence [53].This phase is broad and complex phase than the phases that exist in AOTS. In this phase a sentence, got score, based on the seleted features stated in this work. In order to determine the scoring function,

different algorithm are proposed and sentence weight function is defined by using different experimentation technique. Based on this defined function, this phase adjusted and generates list of sentences with their total scores. The detail description of experiment and procedure, which was under taken to define this scoring function, is shown under this section.

5.3.2.1 Sentence Weighting Based on Sentence Position Module

The researchers in SweSum [17] examined that in newspaper the first line is the most important sentence and gets the highest score, followed by the others. Endmunson [7] also studied that the first sentence of each paragraph contains topic information; he also examined that the strategy of taking of the first paragraph works only for the newspaper and news magazine genres.

Kamal [55] strengthen this idea that if the position increment the importance of the sentence decreases to some degree. Additionally, Jun-Jie Li and Key-Sun Choi [20] in similar way, they presented that sentence in privilege locations (first paragraph, or immediately following section headings Introduction, purpose, conclusion, etc.) contain the topic.

Besides, developer in OTS and OOTS also gave highest privilege for the first sentence of the first paragraph [10]. Having all this in consideration, in order to give weight for a sentence according to their position, in this work experiment are conducted using twenty topics that are shown before in table 5.8. The result of the each twenty topics, which filled in a form that stated in table 5.9 was used in this experiment. Hence, to conduct this experiment based on sentence position, two basic procedures is under taken:

I. Defining equation for sentence according to their position
II. Computing the result based on defined equation and assigning weight

I. Defining Equation for a sentence according to their position

The sentence weighting equation that has been used in this study was coefficient assign method, which also has been used in [10, 31, 32]. Similarly, to assign weight for a sentence based on their position, four category of the sentence are identified according to their position in a document:

- **Category 1:** First sentence of first paragraph (FSFP)
- **Category 2:** Descriptive sentence of all paragraphs (DSOP)
- **Category 3:** Frist sentence of body of paragraphs (FSBP)
- **Category 4:** Frist sentence of last paragraph (FSLP)

In order to compute a value for each category from all topics a number of equation was defined and discussed as follows.

Notes for the equations:

- ***Descriptive sentence:*** *all sentence of the paragraph except the first sentence of all paragraphs.*
- ***Topics (N):*** *is the number of topics selected in this experiment is 20.*
- ***np****: number of paragraph*

Category 1

This category computes the probability of first sentence of first paragraph of all topics that has been selected by the subjects.

Count (Cat. 1) = count (FSFP of all Topic) *(5.2)*

Count_selected (Cat. 1) =Count (Selected(S) FSFP of all topic) *(5.3)*

$$FSFP = \sum_{i=1}^{N} S1i \quad (5.4)$$

$$SFSFP = \sum_{i=1}^{N} Sis1 \quad (5.5)$$

Where:

Sis1: *Selected first sentence of i^{th} topic*

S1i: *Frist sentence i^{th} topic*

Finally after we compute equation 5.2 and 5.3, the probability of category 1 is computed using equation 5.18.

Category 2

In similar way in this category the probability of descriptive sentence of all paragraphs of all topics that has been selected by the subjects.

Count (Cat. 2) = Count (DSOP) *(5.6)*

Count_selected = Count (Selected/ SDSOP) *(5.7)*

$$DSOP = \sum_{i=1}^{N} \sum_{j=1}^{np} \sum_{k=2}^{m} Sijk \quad (5.8)$$

$$SDSOP = \sum_{i=1}^{N} \sum_{j=1}^{np} \sum_{k=2}^{m} Ssijk \quad (5.9)$$

Where:

Sijk: *sentence of i^{th} Topic j^{th} paragraph k^{th} sentence position*

Ssijk: *Selected sentence of i^{th} Topic j^{th} paragraph k^{th} sentence position*

Like Category 1 to compute the probability once equation 5.6 and 5.7 is computed, then using equation 5.18 the probability of category 2 is computed.

Category 3

After computing equation 5.10 and 5.11, this category is used to compute the probability of first sentence of body of paragraphs using equation 5.18.

Count (Cat. 3) = Count (FSBP of all Topic) *(5.10)*

Count_selected = Count (Selected/ SFSBP of all topic) *(5.11)*

$$FSBP = \sum_{i=1}^{N} \sum_{j=2}^{np-1} Sij1 \quad (5.12)$$

$$SFSBP = \sum_{i=1}^{N} \sum_{j=2}^{np-1} Ssij1 \quad (5.13)$$

Where:
Sij1: Sentence at topic i^{th} j^{th} paragraph first sentence
Ssij1: Selected sentence at topic i^{th} j^{th} paragraph first sentence

Category 4

Finally, this category is used to compute the probability of first sentence of last paragraph using equation 5.18 after equation 5.14 and equation 5.15 is computed.

Count (Cat. 4) =Count (FSLP of all Topic) *(5.14)*

Count_selected =Count(Selected/FSLP of all topic) *(5.15)*

$FSLP = \sum_{i=1}^{N} LpiS1$ *(5.16)*

$SFSLP = \sum_{i=1}^{N} SLpiS1$ *(5.17)*

Where:
LpiS1: *Frist sentence of last paragraph of i^{th} topic*

SLpiS1: *Selected first sentence of last paragraph of i^{th} topic*

II. Computing the result based on defined equation and assign weight

Based on the above equation the result of each category filled in table 5.11. The equation that helps to compute the probability of each selected sentence in each category is defined as follows;

$$P(Cat.i) = \frac{Count\_selecetd(Cat.i)}{Count(Cat.i)} \quad (5.18)$$

Weight for sentences that exist in each category is assigned based on defined range shown on (Fig. 5.7), i.e. if the value of P (Cat. i) is between in each range the weight of a sentence is assigned accordingly for each categories. Where WSp (weight of sentence position) is a constant weight value given for a sentence position feature. The constant value of WSp is known after weight adjustment that will be computed for each feature under section 5.3.2.6 .

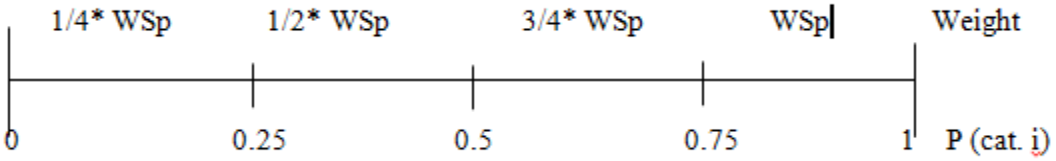

Figure 5.7: *Defined range to assign weight*

Hence, based on the defined weight and the result of each category defined in table 5.11 the weight is assigned for each category. This result justify that the first sentence of first paragraph has more weight than another sentence.

Table 5.11: *Result sentence weight based on their position*

List of category	Selected	Not Selected	Total	Result	Weight
P(Cat. 1)	55	5	60	0.92	WSp
P(Cat. 2)	146	230	376	0.39	WSp/2
P(Cat. 3)	56	91	147	0.38	WSp/2
P(Cat. 4)	27	33	60	0.45	WSp/2
Total	**284**	**359**	**643**	**2.14**	**----**

Based on the founded result for illustration the new designed algorithm is defined as follows (Fig. 5.8):

```
Start
    Input processed text
    For each sentence in a paragraphs
        If sentence position is in the FSFP
            Then sentence score += WSp
        If sentence position is a DSOP
            Then sentence score +=WSp/2
        If sentence position is the FSLP or the FSBP
            Then sentence score += WSp/2
    End for
    Output Sentence score based on position (SSP)
Stop
```

Figure 5.8: *Algorithm for a weighting sentence based on sentence position*

5.3.2.2 Sentence Weighting Based on Key Word Module

According to Jun-Jie Li [20] keyword frequency refers to the frequencies calculated both in the form of context and corpus. In addition, [17][31][49] they found that the sentences that contain more keywords gate a higher score than those that contain fewer or none. The keyword frequency here in this study is based on the input corpus. The list of keyword that was generated by keyword counter module is used by this module. The Keyword detection algorithm in OOTS that was adopted directly from OTS was also used in this study with slight deference. The difference is Girma [10] considers only one word as a keyword but not in this work.

However, in this paper to determine the number of relevant keywords an experiment was conducted. To undertake this experiment, three topics were randomly selected from the twenty topics listed on table 5.8. Precision, recall and f-measure equation that has been listed in section 2.2.2 was used, to measure the performance of the system using different number of keyword.
The system f-measure was computed based on keywords starting from 1 to 10. The average f-measure each topic was taken and plotted as shown on (Fig. 5.9).

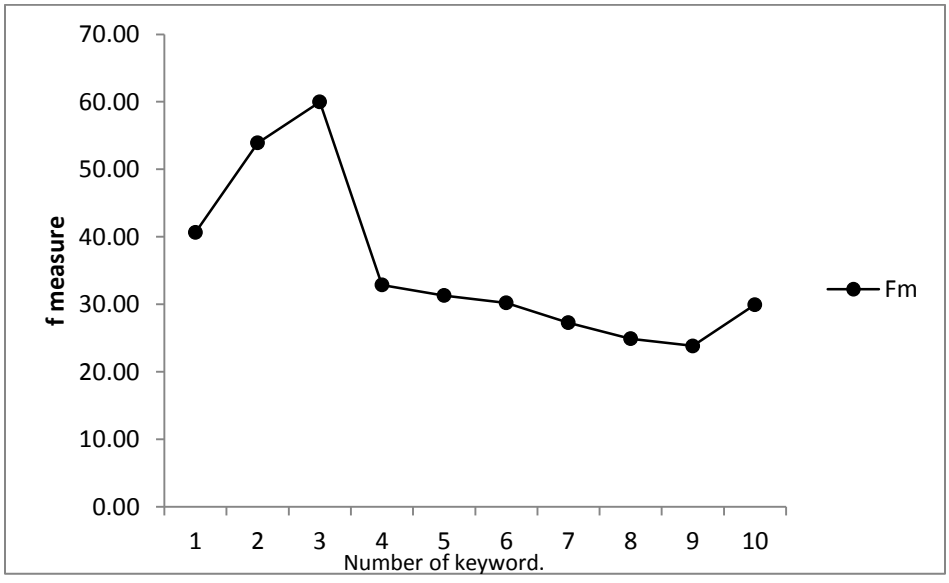

Figure 5.9: *The Fm of AOTS as the number of keywords increases*

When the number keyword reaches 3 the f measure of the system scores 59.89% (Fig. 5.9).On the other hand as the number of keyword exceeds 3 the f-measure of the system become decrease. This implies that as the number of keyword increase the possibility that all sentences get the same score is high, i.e. the possibility of a sentence containing one of the keyword is high, and this leads to that all sentence get equal weight and equal chance to be extracted. That is way the average f-measure drops after a number of keyword exceed three. Based on this experiment, the number of keyword of AOTS is adjusted to 3. The following algorithm was proposed and this module is built (Fig. 5.10). This module access the keyword knowledge base (Fig. 5.1) to get high ranked words. Accordingly, this module takes the top ranked 3 keywords from the keyword knowledge base. The algorithm assigns a constant value *Wkw (Weight for a keyword)* to a sentence that contains this keyword.

```
Start
Input processed sentence
For each sentence in the paragraphs
        For each word in each sentence
                if Sentence contain keywords
                        Sentence score +=WKw
                Else
                        Sentence score +=0
                End if
        End for
End for
Output Sentence score which contains keyword (SSKW)
Stop
```

Figure 5.10: *Sentence weighting algorithm based on number of keyword*

5.3.2.3 Sentence Weighting Based on Cue Phrase Module

During processing, the cue phrase module recognizes the occurrence of cue phrases and assigns a sentence that contains cue phrase a constant value *WCup (weight for cue phrase).* Where *WCup* is a constant value assigned to sentence that contain cue phrases, it will be discussed under sentence score module (section 5.3.2.6).

The algorithm that shows how a sentence that contains a cue phrase is handled is shown on (Fig. 5.11).The list of cue phrase used by this module is stored in cue phrase knowledge base (Fig.5.1).

```
Start
    Input Preprocessed text document
    For each sentence in each paragraph
        If word in each sentence contain cue phrase that stored in lexicon
            Sentence score += WCup
        else
            Sentence score += 0
    End for
Output Sentence score which contains cue phrase (SSCP)
Stop
```

Figure 5.11: *Sentence weighting algorithm based on cue phrases*

5.3.2.4 Sentence Weighting Based on Number Module

In this section, sentence that contains ordinal, nominal or cardinal number get higher weight than a sentence that does not contain numbers. Therefore, sentences that contain a name of a number get a constant value *Wnum (weight for number).* Where *Wnum*: is constant value that going to be discussed under sentence score module. List of name of numbers accessed by this module was stored in name of number knowledge base. The algorithm that is designed for this module is shown on (Fig. 5.12)

```
Start
    Input Processed text
    For each sentence in the paragraph
        For each word in the sentence
            If the sentence contain a numbers
                Sentence score += Wnum;
        End for
    End for
    Output Sentence score based on numbers (SSN)
Stop
```

Figure 5.12: *Sentence weighting algorithm based on name of number*

5.3.2.5 Sentence Weighting Based on Event Module

Oxford dictionary defines a newspaper is a printed publication (usually issued daily or weekly) consisting of folded unstapled sheets and containing news, articles, advertisements, and correspondence: And every newspaper story has to answer when? - When does the story take place? Which means it must be new information as underlined in the definition. Therefore, similarly summary must incorporate sentence that contain name of event in order to share a definition term of newspaper. Because of this fact, sentence that contain name of events got weight than a sentence that do not contain. Constant coefficient value *WEv* (Weight for Event) that will be computed during weight adjustment is added if the sentence contains name of event according to defined algorithm (Fig. 5.13). List of events that is stored in name of time, date and month knowledge base (Fig. 5.1) is accessed by this module, i.e. this module cross check if the sentence contains name of events or not from name of time, day and month KB.

```
Start
    Input Processed text
    For each sentence in the paragraph
        For each word in the sentence
            If the sentence contain a name of events
                Sentence score += WEv;
        End for
    End for
    Output Sentence score (SSE)
Stop
```

Figure 5.13: *Sentence weighting algorithm based on name of events*

5.3.2.6 Sentence Score Module

Under this section, the coefficient values like: - *WEv, WCup, Wnum, WSp* and *WKw*, which are discussed in this phase has been computed, based on defined function that will be seen under this section. This function is known as linear combination function. It contains the sum of features that are used to construct extractive text summarization; in which the parameters are specified manually by experimentation [30, 31, 32]. In similar manner, for this work the total score of all features is calculated based to the equation 5.19 defined in this study. The experiment is all about specifying the coefficient value for each feature.

The major intention of this experiment is to find equation, which gives the best f-measure, when the weight varies. Accordingly, the best equation will be taken for this study and AOTS is adjusted accordingly.

$$Sentence\ score = \sum_{i=1}^{5} Fi \qquad (5.19)$$

Where, Fi is *Wsp=F1, WKw=F2, WCup= F3, WEv=F4 and Wnum=F5,*

or

$$F = F1 + F2 + F3 + F4 + F51 \quad (5.20)$$

To adjust weight for each five features based on the above mentioned equation, we did an experiment using validation data corpus that stated in table 5.8. From equation, 5.20 there are infinite possible coefficient can be assigned to these features. However, in this research, we choose purposely only six possible functions and we call these functions cases. Those, cases are categorized in to two , when all feature has equal weight and when one feature get highest value and the other features has equal value.

To conduct the experiment five topics where randomly selected from validation data corpus listed in table 5.8. In the experiment to choose best equation Precision (P), Recall(R), F-measure of each topic was computed. The F-measure value is considered as a comparison variable to differentiate the best equation among six cases.

During P, R and Fm computing compression ratio of the summary is computed. As mentioned in section 2.2.2 about the compression ratio, the same compression ratio formula was used in this study. CR is used in this paper to compute the CR of human generated summary and system generated summary, CR value of the topics is vary from topic to topic.

$$CR = \frac{length\ of\ the\ generated\ Summary}{Length\ of\ Full\ Text} * 100\% \quad (5.21)$$

The detail score of each six cases was computed and filled accordingly in table 5.12 - 5.17, in these tables the column "compression of reference summary" indicates the compression rate of the summary generated by the subjects manually. This column is used to adjust the CR of AOTS according to the CR of the topic of reference summary. During computation of the f-measure of AOTS, the system CR was adjusted based on the CR generated by the subject's respondents manually.

For example: in table 5.12 when the CR Topic 1 is 41.73%, the AOTS CR is also adjusted to the same CR, i.e. by using the same CR, manually generated summary and system generated summary, the system performance was measured and the best equation was selected. To compute P, R and f-m, the equation that listed in section 2.2.2 was adopted, assume that *Ma* is a manual summary; is a sentence which is underlined by the subjects and *Mc* is a machine generated summary, is a sentence which is generated by AOTS as a summary.

- ☞ Precision is the fraction of retrieved sentence which are relevant.

$$Precision: P = \frac{Ma \cap Mc}{Mc} * 100\% \quad (5.22)$$

- ☞ Recall is the fraction relevant sentences that are retrieved.

$$Recall: R = \frac{Ma \cap Mc}{Ma} * 100\% \quad (5.23)$$

- ☞ Harmonic mean of precision and recall.

$$Fm = \frac{2PR}{P+R} * 100\% \quad (5.24)$$

The experiments of each the six cases are shown below:

Case 1: When each features has equal weight = 1/5.

$$F = \frac{1}{5Sp} + \frac{1}{5Kw} + \frac{1}{5Cup} + \frac{1}{5Ev} + \frac{1}{5Wnum} \quad (5.25$$

Note: Since there are five features, when one is equally distributed for each features, each feature got a value of 1/5.

Table 5. 12 *P, R and Fm with given CR when each features has equal weight*

Topics	Text length	Selected sentence length	CR ratio reference summary	P	R	Fm
Topic 1	266	111	41.73%	20%	14%	16.47%
Topic 3	256	57	22.27%	50%	20%	28.57%
Topic 7	244	102	41.80%	50%	40%	44.44%
Topic 9	247	89	36.03%	100%	33.33%	49.99%
Topic 10	433	98	22.6%	75%	75%	75%
Average	**289.2**	**91.4**	**32.89%**	**59%**	**36%**	**42.90%**

From this table 5.14 we can easily understood that when all features has equal coefficient which is equally divided, the system scores 42.90% average f-measure.

Case 2: When *WSp =1/2* and other four has equal weight.

$$F = \frac{1}{2Sp} + \frac{1}{8Kw} + \frac{1}{8Cup} + \frac{1}{8Ev} + \frac{1}{8Wnum} \quad (5.26)$$

Note: Since there are five features, when WSp got 1/2, other features share equally ½ in to four. Hence, each feature gets a value 1/8; similarly, other four cases (case 3-6) the difference is only the feature that got ½ value.

Table 5.13: *P, R and Fm with given CR when WSP has large weight*

Topics	Text length	Selected sentence length	CR ratio reference summary	P	R	Fm
Topic 1	266	111	41.73%	40%	28.57%	33.33%
Topic 3	256	57	22.27%	50%	25%	33.33%
Topic 7	244	102	41.80%	50%	40%	44.44%
Topic 9	247	89	36.03%	50%	33.33%	40%
Topic 10	433	98	22.6%	33.33%	33.33%	33.33%
Average	**289.2**	**91.4**	**32.89%**	**45%**	**32%**	**36.89%**

In this table 5.15 when WSp has highest coefficient value than other features AOTS scores 36.89% f-measure.

Case 3: When *WKw=1/2* and other four has equal weight.

$$F = \frac{1}{8Sp} + \frac{1}{2WKw} + \frac{1}{8Cup} + \frac{1}{8Ev} + \frac{1}{8Wnum} \quad (5.27)$$

Table 5.14: *P, R and Fm with given CR when Wkw has large weight*

Topics	Text length	Selected sentence length	CR ratio reference summary	P	R	Fm
Topic 1	266	111	41.73%	40%	28.57%	33.33%
Topic 3	256	57	22.27%	33.33%	20%	24.91%
Topic 7	244	102	41.80%	75%	60%	66.67%
Topic 9	247	89	36.03%	50%	33.33%	40%
Topic 10	433	98	22.6%	75%	100%	85.71%
Average	**289.2**	**91.4**	**32.89%**	**55%**	**48%**	**50.12%**

In this table 5.16 when WKw has highest coefficient value than other features AOTS scores 50.12% f-measure.

Case 4: When *WCup=1/2* and other four has equal weight.

$$F = \frac{1}{8Sp} + \frac{1}{8Kw} + \frac{1}{2Cup} + \frac{1}{8Ev} + \frac{1}{8Wnum} \quad (5.28)$$

Table 5.15*: P, R and Fm with given CR when WCup has large weight*

Topics	Text length	Selected sentence length	CR ratio reference summary	P	R	Fm
Topic 1	266	111	41.73%	20%	14%	16.47%
Topic 3	256	57	22.27%	50%	20%	28.57%
Topic 7	244	102	41.80%	50%	20%	28.57%
Topic 9	247	89	36.03%	100%	33.33%	50%
Topic 10	433	98	22.6%	66.67%	66.67%	66.67%
Average	**289.2**	**91.4**	**32.89%**	**57%**	**31%**	**38.06%**

In this table 5.17 when WCup has highest coefficient value than other features AOTS scores 38.06% f-measure.

Case 5: When *WEv=1/2* and other four has equal weight.

$$F = \frac{1}{8Sp} + \frac{1}{8Kw} + \frac{1}{8Cup} + \frac{1}{2Ev} + \frac{1}{8Wnum} \quad (5.29)$$

Table 5.16: *P, R and Fm with given CR when WEv has large weight*

Topics	Text length	Selected sentence length	CR ratio reference summary	P	R	Fm
Topic 1	266	111	41.73%	25%	14.28%	18.18%
Topic 3	256	57	22.27%	50%	20%	28.57%
Topic 7	244	102	41.80%	25%	20%	22.22%
Topic 9	247	89	36.03%	100%	33.33%	50%
Topic 10	433	98	22.6%	75%	75%	75%
Average	**289.2**	**91.4**	**32.89%**	**55%**	**33%**	**38.79%**

In this table 5.18 when WEv has highest coefficient value than other features AOTS scores 38.79% f-measure.

Case 6: When *Wnum=1/2* and other four has equal weight.

$$F = \frac{1}{8Sp} + \frac{1}{8Kw} + \frac{1}{8Cup} + \frac{1}{8Ev} + \frac{1}{2Wnum} \quad (5.30)$$

Table 5.17: *P, R and Fm with given CR when Wnum has large weight*

Topics	Text length	Selected sentence length	CR ratio reference summary	P	R	Fm
Topic 1	266	111	41.73%	50%	28.57%	36.36%
Topic 3	256	57	22.27%	50%	20%	28.57%
Topic 7	244	102	41.80%	50%	40%	44.44%
Topic 9	247	89	36.03%	100%	33.33%	50%
Topic 10	433	98	22.6%	66.67%	66.67%	66.67
Average	**289.2**	**91.4**	**32.89%**	**63%**	**38%**	**45.21%**

In this table when WSp has highest coefficient value than other features, AOTS scores 45.21% f-measure. For clear illustration, the performance difference between those 6 cases is shown on figure 5.14. From the figure, case three

shows greater Fm. Therefore, case 3 (*F= 0.125Sp + 0.5Wkw + 0.125Cup + 0.125Ev + 0.125)* is chosen as AOTS sentence selection function (Scoring).

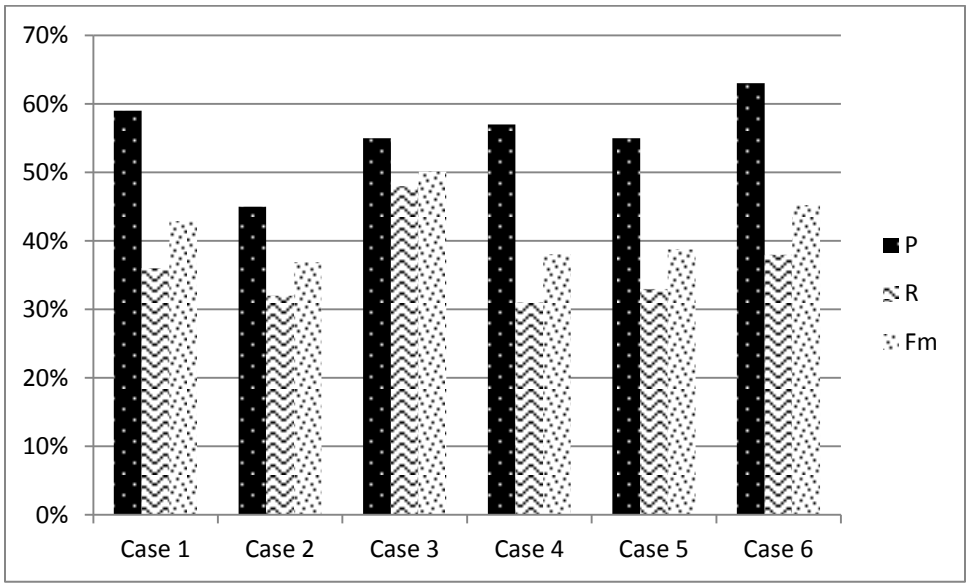

Figure 5.14: *The f-measure performance difference between cases*

5.3.2.7 Sentence Rank Module

After each sentence is scored based on the features used, this module rank sentences based on their total score. For instance, table 5.20 shows the result of topic number 2 of validation data. The table 5.18 shows how the system ranks the sentence.

Table 5.18: *Sentence rank based on their total weight*

No	Sentence Name	Total Weight	Rank
1	Sent1	0.925	2
2	Sent 2	0.993	1
3	Sent 3	0.366	5
4	Sent 4	0.888	3
5	Sent 5	0.672	4

5.3.3 Phase III: Summarizer

In the first two phases, the input document adjusted for processing in preprocessing phase and processed by finding scores of all sentences.
Finally, sentences are ranked according to their scores. In this phase, the summary sentence generator module is activated to generate summary sentence.

This phase is different from other phase by number of inputs it accepts. It accepts two inputs, one from the processing phase and one from the compression ratio module (Fig 5.1). Hence, in this section how these two modules is designed and developed is discussed.

5.3.3.1 Sentence Compression module

Before the automatic text summarizer generates the summary the system generate the CR of the original document. This module served as input for summary sentence generator module. Therefore, in order to choose compute the best compression ratio for AOTS six topics are selected randomly from validation data corpus listed in table 5.8. And, to choose best compression ratio 9 categories of compression ratio are proposed within five difference, 5%, 10%, 15%, 20%, 25%, 30%,35%,40% and 45%. The compression ratio limited to only from 5% to 45%, because of that fact that the summary of the document must be less than half of the original document [8].

After compression ratio was categorized the summary of each topic were generated by AOTS based on the adjusted CR, and the precision, recall and f-measure of the topics are computed. The result of each topic with a given CR is show in table 5.19. The CR, precision, recall and f-measure computed using the equation 5.21, 5.22, 5.23 and 5.24.

Table 5.19: *P, R and Fm of topics with given CR*

Topics	CR		P	R	Fm
Topic 1	5%		100%	14.28%	25%
Topic 4			-	-	-
Topic 8			-	-	-
Topic 14			100%	11.11%	20%
Topic 15			-	-	-
Topic 18			-	-	-
		Average	**100%**	**12.70%**	**22.53%**
Topic 1	10%		100%	14.28%	25%
Topic 4			50	12.5%	20%
Topic 8			-	-	-
Topic 14			100%	11.11%	20%
Topic 15			-	-	-
Topic 18			-	-	-
		Average	**83.33%**	**12.63%**	**21.94%**
Topic 1	15 %		50%	14.28%	22.22%
Topic 4			66.67%	25%	36.36%
Topic 8			100%	25%	40.00%
Topic 14			50%	11.11%	18.18%
Topic 15			50%	16.67%	25.00%
Topic 18			-	-	-
		Average	**52.78%**	**15.34%**	**28.53%**
Topic 1	20 %		50%	14.28%	22.22%
Topic 4			66.67%	25%	36.36%
Topic 8			100%	25%	40.00%
Topic 14			20%	11.11%	14.28%
Topic 15			66.67%	33.33%	44.44%
Topic 18			100%	50%	66.67%
		Average	**67.22**	**26.45%**	**37.33%**
Topic 1	25%		33.33%	14.28%	19.99%
Topic 4			75%	37.50%	50.00%
Topic 8			100%	55%	70.97%
Topic 14			25%	51.11%	33.58%
Topic 15			66.67%	33.33%	44.44%
Topic 18			100%	50%	66.67%
		Average	**67%**	**40%**	**48%**
Topic 1			50%	30.57%	37.94%
Topic 4			75%	40.50%	52.60%
Topic 8			100%	25%	40.00%

Topic 14	30%		60%	51.11%	55.20%
Topic 15			66.67%	53.33%	59.26%
Topic 18			75%	50%	60.00%
		Average	**71%**	**42%**	**51%**
Topic 1			40%	54%	45.96%
Topic 4			60%	57.50%	58.72%
Topic 8			100%	50%	66.67%
Topic 14	35%		20%	45.11%	27.71%
Topic 15			75%	50%	60.00%
Topic 18			50%	50%	50.00%
		Average	**58%**	**51%**	**52%**
Topic 1			40%	65.57%	49.69%
Topic 4			66.67%	50%	57.14%
Topic 8			57%	100%	72.61%
Topic 14	40%		60%	53.33%	56.47%
Topic 15			66%	60%	62.86%
Topic 18			45%	100%	62.07%
		Average	56%	71%	60%
Topic 1			33.33%	80.00%	47.06%
Topic 4			66.67%	80%	72.73%
Topic 8			45%	100%	62.51%
Topic 14	45%		57.14%	84.44%	68.16%
Topic 15			70%	100%	82.35%
Topic 18			60%	80%	68.57%
		Average	**56%**	**87%**	**67.01%**

From the result shown in table 5.19 above, we take the best f-measure and the best result is 45% compression ratio. Because, it scores an average of 67.01% f-m and accordingly AOTS is adjusted with 45 % CR. In other word, the average human generated summary CR of 20 topics listed in table 5.8 is 45.92% and the result gained from experiments 45% CR. This indicates that the CR found from the experiments and average CR taken from validation is almost resembles to each other. Hence, based on this fact and the hypotheses that the summary must be short, the researchers decided to take 45% CR as a parameter tuning. For clear illustration among different CR, average P, R, and f-m is shown on (Fig. 5.15).

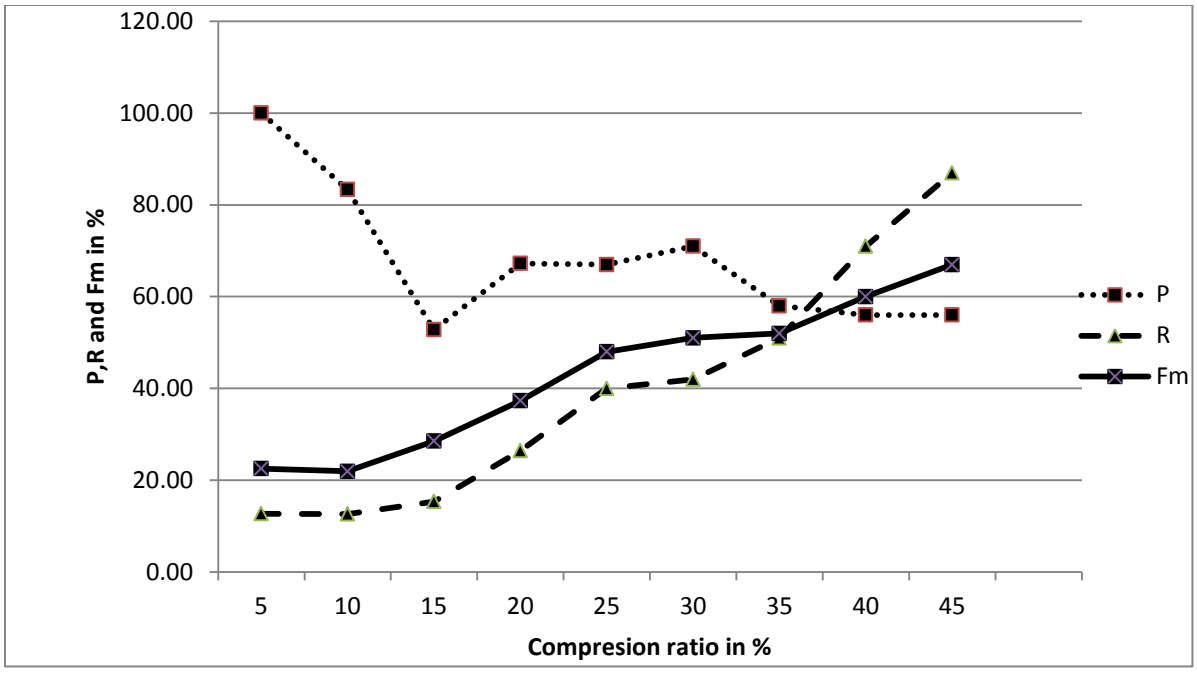

Figure 5.15: *The performance gap between different CR*

5.3.3.2 Summary Sentence Generator module

This module accepts the CR and sentence rank as input, and after accepting these two, it resort sentences that will be generated based on the specified sentence compression ratio and according to their order in the document and finally generate the summary sentence. However, during sentence generation based on their compression the last sentence may be out of the generated sentence due to the specified compression. On the other hand, the generated summary may be under specified compression ratio. The algorithm that describes this module is shown in (Fig. 5.16.)

Begin

Rank=1

CR= Sentece_CR//Get compression ratio from compression ratio module

generated_Sentence=null

For Each sentence in Sentencerank //Sentence in sentence ranker

If (genereated_Sentence.words < Totall_words_in_CR)

genereated_Sentence = Sentecnce(rank)

Else

If(generated_Sentence.words(rank)/2 < (Totall_words_in_CR-Totall_words_insentence_generatedsofar())

genereated_Sentence=Sentence

Else

Stop

End If

rank++

End For

Stop

Figure 5.16: *The algorithm that shows the how sentence generators generate a sentence*

For clear illustration of this algorithm, the example is described below.

Let For example:

1. The compression ratio 30%
2. 30% CR of the document contains 50 words
3. The numbers of words in each sentence are given as follows in table 5.20.

Table 5.20: *Sentence with its number of words*

Sentences	Number of words in a sentence
1st	13
2nd	17
3rd	11
4th	12

Based on the given value, until to 3rd sentence the sum of the sentence is less than 30% of the document, which is 41 the space left 9 words but the next sentence length is 12, in normal condition the sentence is out of extracted sentence. Because the next sentence exceeds the space left by (12-9) 3 words, this leads to under compersion[1] by 9 words.

Therefore, to handle this when the last sentence extracted, its half-length is compared with the space left. In table 5.20 last sentence length (4th sentence) is 12/2 =6, if half the sentence is less than space left the sentence is included in extracted sentence (equation 5.18)

$$\frac{Length\ of\ last\ sentece}{2} < Total\ num.\ def\ \ CR - \sum_{i=1}^{n-1} SLi \quad = \text{True} \qquad (5.18)$$

Where, *Total num. def CR :- is total number of words exist in the summary generated based on defined CR, n number of sentence, SLi i^{th} sentence length.*

Therefore, based on the given data on table 5.20 the 4th sentence is included in list of extracted sentence. But, if the size of the 4th sentence were 20 words the sentence will be ignored because the 20/2 = 10 and 10 is not less than 9, in other means more than half of the sentence is out of the range. This solution helps to fit the generated summary by specified CR.

1. Under compression ratio: in this paper under compression ratio is when the machine generated summary below the given compression ratio.

5.3 The prototype

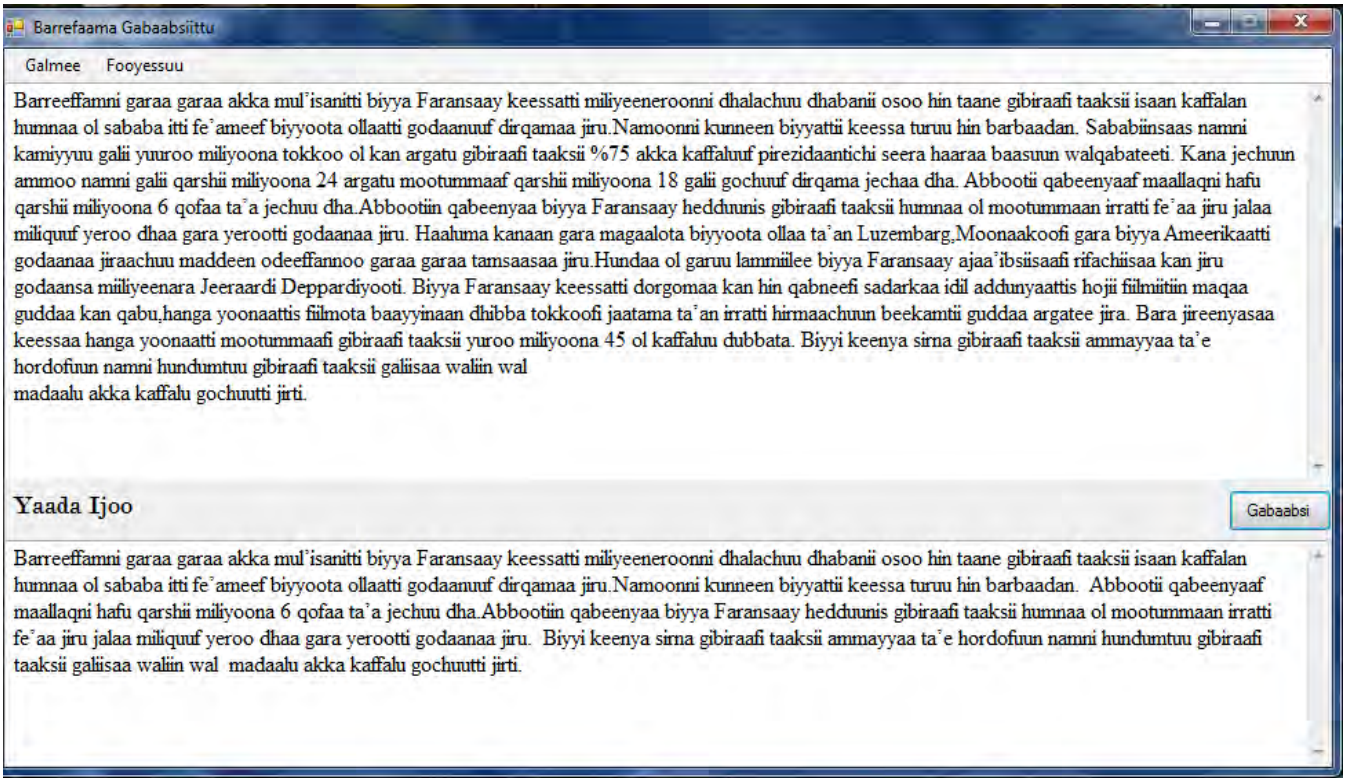

Figure 5.17:*The screenshot of prototype's user interface*

5.4 Summary

This chapter presented design requirements that are necessary to build a knowledge base for Afaan Oromo. Architecture of AOTS discussed by categorizing it in to three phases: the preprocessing, processing and the summarizer phase. In each phase, the detail approaches and techniques that show how modules are build and the function of each module are discussed; and the experiment that under taken to take best equation was also discussed. The workflow of AOTS also discussed in detail, AOTS takes input text, and pass to the preprocessing phase. The preprocessing phases first detect the boundary of sentence and words in a sentence. By using sentence length handler module, too short and long sentence is detected and removed from the list. The stop words are then removed from each sentence using stop word remover module.

The stemmer module stems words into their stem/root. Finally, in this phase a keyword counter module count list of words and store in it keyword knowledge base. In the second phase, weight is assigned for each features and a score is calculated for each sentence by defined sentence selection function. The sentence ranker module then ranks sentences according to their score and the last phase the summarizer generate the summary by accepting and input from sentence compression module the compression ratio and the sentence rank from processing phase. Finally, AOTS generate the summary sentence(Fig 5.17).

Chapter Six: Experimental Result and Analysis

6.1 Introduction

Evaluating summaries and automatic text summarization systems is not a straightforward, it needs techniques and method. Hence, this chapter deals details about evaluation techniques, results and analysis conducted

6.2 Test Data Preparation and Analysis

The preparation of test data corpus consist two steps. The first step is gathering different newspaper topics. A detail statistic of these 13 topics selected for testing purpose is shown in table 6.1. After those topics gathered from different news portal, the next step is preparing the summary for testing purpose. Test data was prepared into two forms:

1. Reference Summary Test Data

The reference summary test data is a summary prepared by the subject evaluators to test the performance of the system. Like validation data, 8 topics were selected purposely according to their test number order (from test # 6 to test #13, that are shown in table 6.1) and distributed to subject respondents in a form of questioner. The respondents were expected to underline the main idea of the document, as shown in Appendix I. It takes less effort for subject respondent when compared to system summary test data preparation. The result of reference summary was used for objective evaluation. The way we keep the collected result is the same as validation data representation (Appendix L).

2. System Summary Test Data

The system summary is a summary generated by AOTS and evaluated by human. Here in system summary data preparation the subject are expected to judge the output of AOTS according to a guide line provided on the questioner (Appendix K). The judgment of system summary by subjects is critical and little bit need attention for judgment than reference summary test data only

preparation. Therefore, out of 13 topics available for test data only five topic were selected purposely (from test#1 to test#5, shown in table 6.1).

Table 6.1: *Statistics of test data corpus*

Test ID	# of Words	# of Sentences	# of Paragraphs
Test #1	263	9	9
Test #2	379	14	9
Test #3	424	20	4
Test #4	315	16	9
Test #5	340	10	10
Test #6	266	13	5
Test #7	214	12	4
Test #8	248	10	3
Test #9	251	14	6
Test #10	276	14	7
Test #11	263	9	9
Test #12	276	9	3
Test #13	248	7	7
Average	**289.46**	**12.08**	**6.54**

6.3 Afaan Oromo Text Summarizer Performance Evaluation

6.3.1 Experimentation Technique

In this work, seven experimentation techniques were proposed to observe the strength of AOTS from deferent angle. The development tool selected was object-oriented programming. Subsequently, among different benefits of OOP in comparison of other system development that it is easy: to develop, manipulate, test and understand. Because, OOP clusters things in terms of class and objects. Therefore, the procedure to undertake the whole experiment is the same, i.e. the experiment was undertaken by accessing or not to accessing different module according to the given experimentation techniques.

For example: experiment 3 was conducted by the procedure of not to access the sentence length class and accessing other class. In short, the Procedure used in all experiment is almost the same; the only difference is the class they access. List of the seven proposed experiments and their objectives is show below.

Experiment 1

Investigating the performance of AOTS

Objective:

- ☞ To measure the system when all features are incorporated.

Experiment 2

Investigating the effect of stemmer and other language specific lexicon

Objective:

- ☞ Measure the performance of AOTS without stemmer, synonym and stop word remover
- ☞ To observe, the capability of the keyword extractor, when stemmer, synonym, and stop word remover module not incorporated.

Experiment 3

Investigating the effect of Sentence Length

Objective:

- ☞ Measure the performance of AOTS without sentence length feature.

Experiment 4

Investigate the effect of Cue phrase

Objective:

- ☞ Measure the performance of AOTS Without cue phrase

Experiment 5

Investigate the effect of name Events on ATOS

Objective:

- ☞ Measure the performance of AOTS Without events

Experiment 6

Investigate the effect of numbers

Objective:

- ☞ Measure the performance of AOTS Without numbers

Experiment 7

Investigate the performance of AOTS having only keyword and sentence position features (without Cue phrase, name of Event, name of Numbers and Sentence Length are not incorporated)

Objective:

- ☞ Measure the performance of AOTS with only sentence position and keyword frequency.
- ☞ To observe the impact of additional feature on the performance of the summarizer.
- ☞ To measure the performance gap between the current system and OOTS have the same feature.

6.3.2 Evaluation and Discussion of Result

Based on the defined experimentation techniques the experiment was conducted, the result found and the inferred result will be discussed in this section in both subjective and objective manner.

6.3.2.1 Subjective Evaluation

This is traditional evaluation method of automatic text summarizer which involves human judgments in categories like informativeness, non-redundancy and referential clarity and coherence [7, 42]. Similarly, in this paper the summary which is generated by the system based on different experimentation

method was evaluated by the subject evaluators. The subject evaluators evaluate the performance of AOTS based the following three check points.

- ☞ The summary informativeness
- ☞ Non-redundancy and referential clarity
- ☞ Coherence

As mentioned before in section 6.2 the subject evaluators are requested to read the original documents and evaluate summaries without respect to any particular task or goal. In order to resolve miss understanding during evaluation the description about these three check point has been stated on the questioner as shown on Appendix k. On this Appendix summary 3: is equivalent to *Exp. 1*, summary 2 is equivalent to *Exp. 2* and Summary 1 is equivalent to *Exp. 7*. The summary is coded as summary 1, 2 and 3, is to minimize bias.

Subjective Evaluation Result and Discussion

The intention of subjective evaluation in this study is to measure the performance of the system from human perspective. Therefore, based on the 3 check points stated in this section, the result that was collected from 5 subjects explained as follows. For subjective evaluation only three experiments were purposely selected (*Exp.1*, *Exp. 2* and *Exp. 7*).

It is chosen from the seven experiments listed in section 6.3.1. These three experiments were chosen by researcher for 3 reasons;

1. Because it is tedious for the subject evaluator to evaluate the summarizer for the seven experimentation techniques.
2. The result from the evaluators will be biased if other five experiments will be incorporated, because the generated summary has small difference for human evaluators for other experimentation technique, except for *Exp 1*, *Exp 2* and *Exp 7*.
3. To observe the influence of the additional features on the performance of the system from human perspective.

A. Informativeness of the Summary

As we discussed in chapter two one way of evaluation method is measuring how much information in the reference summary is present in the generated summary which is known as summary informativeness [7, 43]. The summary of the original document was generated by AOTS based on three experiments as shown in appendix k, the result of each summary according to 3 experiments collected from the subject evaluators interpreted as follows. The scoring mechanism used in this study is based on 5 different point; Very Good= 5, Good=4, Not bad= 3, Poor= 2 and Very Poor =1. Each topic is evaluated by five different subjects evaluators the result of the informativeness of the summary show in table 6.2. It is out of 100%, the percentage of the informativeness of the summary for each experiment is the sum of the score given by each subject evaluators for each topic, divided by the sum of maximum score.

$$\frac{\sum_{i}^{5} Ri}{25} * 100 \qquad (6.1)$$

When, *Ri* is result scored by each subjects evaluators, as shown on Appendix L. For example, using this equation the result of ***Test #9 of Exp. 1*** shown on table 6.2 is $(5 + 5 + 5 + 3 + 4) = 22$, $^{22}/_{25} * 100 = 88\%$

Table 6.2: *Informativeness evaluation result*

Test ID	Exp.1	Exp. 2	Exp. 7
Test #9	88%	52%	72%
Test #10	80%	32%	76%
Test #11	84%	36%	60%
Test #12	92%	40%	56%
Test #13	96%	48%	64%
Average	**88%**	**42%**	**66%**

Compared to *Exp2* and *Exp7* the information preserved in *Exp1* is scores higher than both. Without stemmer and other language specific lexicon the summarizer informativeness is 42% when these feature is added the summary informativeness is improved in 46%. On the other hand, when cue phrase, name of events, name of numbers and sentence length handler is combined the informativeness is increases in 22%. From this result we can deduce that the informativeness of a summary improved when the pre-process and all features were combined together. And presence of additional features proposed in this paper play a big role in the improving the performance of the summarizer.

B. Referential integrity and Non Redundancy Summary Result

This check point let the respondent give a score based on two points. The first one is non-redundancy, that enables the respondent to observe the summary weather it contain unnecessary repetition of whole sentences or ideas or not. The second point is referential integrity, while reading the sentences according to their generated order, the respondent identify who or what the pronouns and nouns phrases in each sentence are refereeing to. The same scoring mechanism that was used in measuring the informativeness of the summary is also used in this check point.

Table 6.3: *Referential integrity and non-redundancy evaluation result*

Test ID	Exp.1	Exp. 2	Exp. 7
Test #9	68%	48%	60%
Test #10	76%	24%	64%
Test #11	76%	32%	60%
Test #12	80%	20%	56%
Test #13	76%	40%	60%
Average	**75%**	**33%**	**60%**

Referential integrity and non-redundancy of the summarizer is also better when all features are combined together, it scores 75% for Exp1, 33% for Exp2 and 60% for Exp7. From this checkpoint result the summarizer non redundancy and referential integrity is poor when stemmer and language specific lexicon are exclude from the summarizer.

C. Coherence Summary Result

Finally in the subjective evaluation the coherence of the summary was measured. In this check point the respondents measured the smooth transition of sentence from sentence for generated summary. The same scoring mechanism that was applied to measure the informativeness of the summary is also applied in this evaluation mechanism and the result of the respondents is shown below in table 6.4.

Table 6.4: *Coherence evaluation result*

Test ID	Exp.1	Exp. 2	Exp. 7
Test #9	68%	36%	56%
Test #10	60%	32%	56%
Test #11	76%	36%	60%
Test #12	64%	48%	52%
Test #13	72%	32%	56%
Average	**68%**	**37%**	**56%**

As far as, the summarizer extractive type of summary clearly the coherence of the summary is poor compared to the other two check points. In addition, the coherence of the summary increased in 12% when: cue phrase, sentence length handler, name of events and numbers are incorporated in the summarizer. And the summary coherence shows very poor result when stemmer and language specific lexicon is excluded, this implies the importance of this feature on the performance of the summarizer. However, the coherence of the summary is better when all features are combined together. Generally, AOTS scores 88% informativeness, 75% referential integrity and non-redundancy, and 68% coherence.

6.3.2.2 Objective Evaluation

Among different methods that are used to evaluate the performance of a text summarization system, objective evaluation play great role in measuring the effectiveness, quality and performance of summarizer [38]. Since intrinsic evaluation method was used in this paper; which is a method of evaluating the quality of a machine generated summary based on the correspondence between

the system generated summary and the human generated summary. Then, Precision (P), recall (R) and f-measure (Fm) were used as computation mechanism. Equations that are discussed in chapter five were used again to compute P, R and Fm. Where Equation 5.22 used to calculate P, Equation 5.23 used to calculate R and equation 5.24 used to calculate Fm.

Objective Evaluation Result and Analysis

For objective evaluation all seven experimental techniques, which have been proposed under section 6.3.1. The main reason to take all experimentation techniques is to answer the research question. However, since it was manually evaluated it takes a lot of effort and time from the researchers. Hence, the result for seven different experiment techniques and inferred analysis seen under this section one by one.

A. Experiment 1 Result and Discussion

The objective of this experiment is to discuss the performances of the system when all features are involved for extraction of a sentence. Hence, in this experiment when all features are used, the average f-measure result of the system gives 87.47%.

Table 6.5: *Experimental result when all features is used*

Test ID	P	R	Fm
Test #6	86.43%	82.5%	84.42%
Test #7	83.33%	100%	90.91%
Test #8	86.67%	80%	83.20%
Test #9	100%	80%	88.89%
Test #10	100%	86.43%	92.72%
Test #11	80%	85%	82.42%
Test #12	80%	100%	88.89%
Test #13	86.67%	90%	88.30%
Average	**87.89%**	**87.99%**	**87.47%**

The result from this experiment shows very promising result (Table 6.5). This result implies directly the strength and efficiency of the techniques, parameter adjustment and algorithm used in this study.

B. Experiment 2 Result and Discussion

The objective of this experiment is to show the performance of AOTS without stemmer, synonym and stop word remover. On the other hand, to observe, the capability of the keyword extractor extract the keywords, when stemmer, synonym, and stop word remover module not incorporated. From the result of this experiment, that is shown in table 6.6 the performance of the system decreases when the stemmer, stop word and synonym are not incorporated in the summarizer. This is because of that the keyword extractor not accomplishes its counting properly; since, the key word extractor is based on stemmer, synonym and stopword remover module. The F-measure score decrease in (87.47% - 77.99%) 9.48 %.

Where, 87.47 is the average F-measure score, when all feature (Exp. 1) incorporated in the summarizer shown in table 6.6 and 77.99 is the average F-measure score of this experiment (Exp2.) shown in table 6.6.

Table 6.6: *Experimental result , when stemmer and other language specific lexicon are not incorporated.*

Test ID	**P**	**R**	**Fm**
Test #6	76.43%	72.5%	74.41%
Test #7	73.33%	80%	76.52%
Test #8	76.67%	75%	75.83%
Test #9	80%	80%	80.00%
Test #10	100%	86.43%	92.72%
Test #11	75%	85%	79.69%
Test #12	80%	75%	77.42%
Test #13	76.67	60%	67.32%
Average	**79.76%**	**76.74%**	**77.99%**

C. Experiment 3 Result and Discussion

In this experiment, the system performance was measured, with the absence of handling of sentence length and the result gained is 83%. This is the system loss 4.47% performance.

Table 6.7: *Experimental result, when: Sentence Length is not incorporated.*

Test ID	P	R	Fm
Test #6	76%	60%	67%
Test #7	80%	67%	73%
Test #8	83.33%	76%	79%
Test #9	87%	100%	93%
Test #10	100%	85%	92%
Test #11	86.67%	86%	87%
Test #12	100%	67.50%	81%
Test #13	80%	100%	89%
Average	**87%**	**80%**	**83%**

From this experiment, we observed that lack of sentence length feature let too length and too short sentence degrade the performance of the system.

D. Experiment 4 Result and Discussion

The system loss 9.74 % f-m performance, if the cue phrase feature were not incorporated in the weighting of the sentence.

Table 6.8: *Experimental result when cue phrase is not incorporated*

Test ID	P	R	Fm
Test #6	76.43%	82.5%	79.35%
Test #7	83.33%	60%	69.77%
Test #8	86.67%	80%	83.20%
Test #9	80%	80%	80.00%
Test #10	76%	66.43%	70.89%
Test #11	60%	85%	70.34%
Test #12	66.67%	100%	80.00%
Test #13	86.67%	90%	88.30%
Average	**76.97%**	**80.49%**	**77.73%**

E. Experiment 5 Result and Discussion

From this experiment 5.68% performance loss gained when the name of event is not added as feature.

Table 6.9: *Experimental result, when name of event is not incorporated*

Test ID	P	R	Fm
Test #6	76.43%	82.5%	79.35%
Test #7	83.33%	75%	78.95%
Test #8	85%	77.67%	81.17%
Test #9	80%	80%	80.00%
Test #10	76%	86.43%	80.88%
Test #11	75%	80%	77.42%
Test #12	85.47%	85%	85.23%
Test #13	84%	100%	91.30%
Average	**80.65%**	**83.33%**	**81.79**

F. Experiment 6 Result and Discussion

This experiment investigated that the performance of the system was decreased in 4.12% if the system misses the name of numbers from the system.

Table 6.10: *Experimental result when name number is not incorporated*

Test ID	P	R	Fm
Test #6	80.43%	82.5%	81.45%
Test #7	75.33%	84%%	79.43%
Test #8	86.67%	85%	85.83%
Test #9	80%	80%	80.00%
Test #10	76%	100%	86.36%
Test #11	70%	85%	76.77%
Test #12	85.65%	100%	92.27%
Test #13	80%	90%	84.71%
Average	**78.01%**	**88.31%**	**83.35%**

G. Experiment 7 Result and Discussion

The major intention of this experiment is to measure the system performance using only keyword frequency and sentence position based on the weighting method used in this study. Moreover, to show the performance difference between Girma's work and this work, using the same features. In addition, this experiment is conducted to see the impact of additional features on the performance of AOTS. As a result the system scores 66%Fm, shown on table 6.11. In other word the system performance increases in 21.27% when cue phrase, sentence length handler, name of numbers and name of events is added.

Table 6.11: *Experimental result, when only keyword frequency and sentence position are used*

Test ID	**P**	**R**	**Fm**
Test #6	60%	50%	55%
Test #7	80%	60%	69%
Test #8	63.33%	65%	64%
Test #9	70%	80%	75%
Test #10	100%	50%	67%
Test #11	86.67%	50%	63%
Test #12	100%	57.50%	73%
Test #13	80%	59%	66%
Average	**80%**	**59%**	**66%**

The following indication inferred from objective evaluation experiments:

1. A feature that more contributes to the performance of the system was identified. Hence, (Fig. 6.1) the absence of sentence length handling and name of number has small impact on the performance of the system. On the other hand, stemmer and language specific lexicon, name of event and cue phrase has high impact on the performance of the summarizer. Fig 6.1 shows the impact of each features in percentage when they are not incorporated in the AOTS.

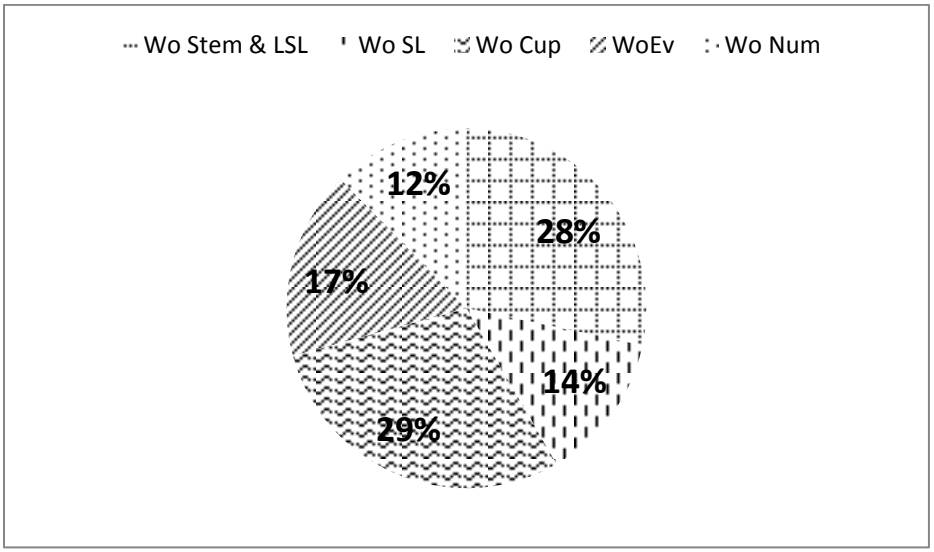

Figure 6.1: *The performance of AOTS without different features*

2. As the number of features increases the performance of the system increases.
3. According to our experiment the weighting/scoring technique used in this study shows consistency in the all experiments. In other word, the F-measure of different topics, for different experiment does not show huge gap among each other.
4. The CR founded in this paper let the system provide better summary result.
5. The additional features incorporated in the current study increases the performance of the summarizer in 21.47%fm.

6.4 Open Oromo Text Summarizer vs. Afaan Oromo Text Summarizer

Since both of the work has been done on the same language and have in common features, hence it is necessary to see the performance gap created among OOTS and AOTS. To measure the gap similar corpus, CR and similar techniques were used. (Fig. 6.2) illustrate that the F-measure performance difference among OOTS and AOTS. In all experiments, AOTS perform better than OOTS.

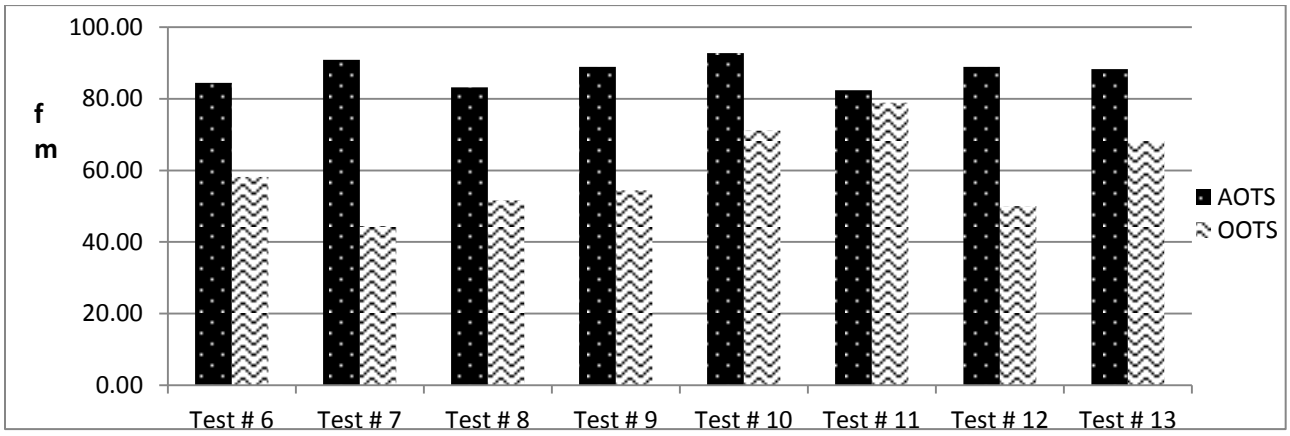

Figure 6.2: *The f-measure of AOTS and OOTS*

Fig. 6.3 illustrates the average f-measure performance difference between AOTS and OOTS, and the result shows that AOTS was outperform in 26.95% than OOTS. In addition, by using only keyword frequency and sentence position AOTS (Table 6.11) scores 66%. This result indicate that the experiment that were conducted to specify number of keyword used and the experiment done to adjust weight based on sentence position improved the performance of the summarizer from the former work of Girma in (66% - 60.52%) = 5.48%.

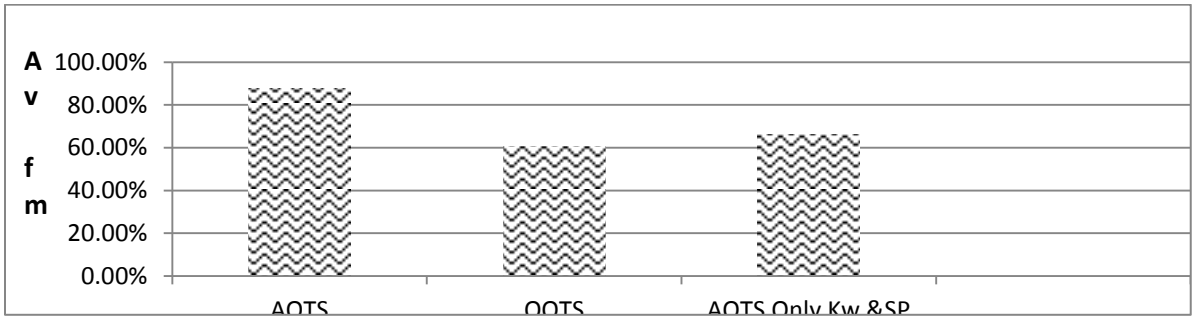

Figure 6.3: *The Average fm Measure gap between AOTS and OOTS*

Chapter seven: Conclusion, Recommendation and Future Work

7.1 Conclusion

As stated earlier, this study was to design and develop Afaan Oromo news text summarization based on sentence selection functions. The function uses features like; sentence position, keyword frequency, sentence length handling mechanism, cue phrase, name of numbers and events. Even if, there are a number of features exist to extract a summary sentence using sentence selection function, features that are selected in this work is for three reasons;

- ☞ To observe the performances Afaan Oromo automatic text summarizer performance by increasing the number features from two to five.
- ☞ Based on the properties of the newspaper investigated during the review, i.e. the newspaper contains some of the features that used in this study [14].
- ☞ Most of the paper reviewed [5][7][10][17][23][24][29] in this study achieves good performance by incorporating some of the feature selected in this work.

Hence, the system was developed based on selected features, by using OTS C# version open source as a tool. During evaluation seven different experimentation scenarios has been used to evaluate the summary of the system from different angle. Form this experiment; features that are selected in this work improved the performance and quality of the summarizer. In addition, the degree of consistent of the summarizer shows best consistency from experiment to experiment. Beside this, by using similar corpus and evaluation method the current summarizer outperform by 26.96% Fm than the previous work by. This all result implies that, techniques, algorithms, defined equation, conducted experiments and especially the added features let the system improve the performance and quality of the summary and show consistency among experiments.

Furthermore, the study sought to answer four questions and founded result has been discussed accordingly as follows:

1. Which compression ratio is relevant for automatic Afaan Oromo news text summarizer?

The relevant CR explored through the experiments was 45% CR. This taken as an ideal CR, because the summarizer achieves nice performance than the other CR. Similarly, using this CR the result found in both subjective and objective manner shows promising result.

2. How and when did the performance of Afaan Oromo text summarizer increases?

To answer how question the selected feature provide an answer for that, i.e the features selected in this study increased the performance of the summarizer. Moreover, the experiments that were undertake to assign weight, finding relevant number of keyword show great improvement on the quality of the summarizer. In addition, when the number of features increased from two to five the performance of the summarizer is increased. Besides, the discovered performance gain when compared to the previous work is a witness for this research question.

3. To what extent, additional features incorporated in this study, affect the performance and quality of the summarizer?

☞ **Subjective Evaluation result**

Subjectively the summarizer achieves good result based on the three checkpoints that has been conducted in the study. The results of these three experiments are: the summarizer gives 88% informative, 78% referential integrity and non-redundancy and 68% coherence. This result implies that, the contribution of the additional features incorporated in the summary let the system provides improved result.

☞ **Objective Evaluation result**

Objective evaluation result also shows better when this work compared with a system, which has two features. This means, the system gave 87.47% Fm when

all features are incorporated. This indicates, that the added features let the summarizer outperform in 21.47%Fm.

4. Which feature contributes more, and less, to the performance of the summarizer?

Based on conducted seven experimentation scenarios, the absence of sentence length handler and name of number has small impact on the performance of the summarizer. On the other hand, stemmer and language specific lexicon, and name of event and cue phrase has high impact on the performance of the summarizer if they were not incorporated in the system. The detail of their effect on the performance of the system is showed before on Fig. 6.1.

7.2 Recommendation

- ☞ From the experiments that have been conducted, in this work absence of stemmer degrades the performance of the summarizer by 9.48%Fm. The researchers have also seen the presence of stemmer used in this paper increases the performance of the summarizer. This indicates that to gain better performance and efficient AOTS the presence of the stemmer that can handle suffix, infix, and prefix increases the performance of the summarizer more than the current study. Having this in mind better stemmer algorithm that can handle highly morphologically inflected word is recommended by the researchers.

- ☞ A standard well prepared corpus is an essential part for future evaluation of the performance of the summarizer and development of the AOTS. In addition, the presence of such a data set would encourage more research to carry out on the mentioned types of summarization.

- ☞ In this paper, the evaluation was carried out on small data sets; however, for further study and to gain better performance increasing the number of experiment can robust more the work.

- By increasing possible weight adjustment function; there will be high possibility to acquire high performance summarizer.

- The researcher was found many holes in the linguistic study of Afaan Oromo language in general, and in morphology in particular. Therefore, linguists should give outstanding consideration to intensively study the language structure and make it available for use in developing computational models.

- During objective summary evaluation, there was a difficulty of computing precision, recall and f-measure. Therefore, the researcher recommends a tool that can automatically evaluate summary system.

- From the result of two intrinsic evaluation types used in this study: subjective and objective evaluation, as number of feature increase the performance of the summarizer also increases. This implies that to get better summarizer, adding additional features can give better result than the current system.

7.3 Future Work

To enhance more the quality and performance of the summarizer as a future work, two basic ideas is forwarded. The first one is, to let the summarizer to answer more about why, where. In future work will incorporate named entity recognition systems, which identify all the names of people, places, organizations, date, etc.. Secondly, this work does not deal with how to handle the redundant sentences, hence for further improvement of the performance, in future work we will work on how to handle the redundant sentence in the preprocessing to enhance the quality of the generated summary.

Reference

[1] Mani, I., "Automatic Summarization" ,John Benjamin's Publishing Company, 2001

[2] Aone, C., Okurowski, M.E., Gorlinsky, J., Larsen, B.," A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques", 1999.

[3] Azzam, S., Humphreys, K., Gaizauskas, R., "Using Coreference Chains for Text Summarization.", Processing of the ACL'99 Workshop on Coreference and its Applications ,ACL, Baltimore , 1999.

[4] Wiemer-Hastings, P. "How Latent is Latent Semantic Analysis?", in Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99), Stockholm, 1999.

[5] CelalCığır and et al., "Generic Text Summarization for Turkish", Department of Computer Science, Bilkent University, Ankara, Turkey,2008.

[6] Morns, J and G Hzrst," Lexical cohesion computed by thesaural relations as an indicator of the structure of the text Computational Linguistics", 17(1) pp 21-45,1991

[7] Edmundson, H. P., "New methods in automatic extraction", Journal of the ACM, 16(2),264–85. Also in Mani and Maybury (1999), 23–42, 1969.

[8] Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto, "Automated Summarization Evaluation with Basic Elements", In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006.

[9] Gadaa Malbaa, "Oromia" ,Sudan:Khartoum, 1988.

[10] Girma Debele 2012, "Afaan Oromo news text summarizer.", MSc thesis, Addis Ababa University, Addis Ababa, Ethiopia.

[11] Gong, Y., Liu, X., "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis.", SIGIR. ACM, New Orleans Louisiana, 2001.

[12]Gong, Y. & Liu, X. ,“Generic text summarization using a trainable summarizer and Latent Semantic Analysis”, In Proceedings of 24th annual international ACM SIGIR conference on research and development in information retrieval (SGIR’01) (pp. 19-25), New Orleans, LA, USA, 2001.

[13]Brunn, M., Chali, Y. and Pincha, C.J. 2001., "Text summarization using lexical chains, in Document Understanding Conference (DUC)", New Orleans, Louisiana USA, September 1314, 2001.

[14]Hovy, E., Lin, C.Y.," Automated Text Summarization in SUMMARIST.", In: Mani, I., Maybury, M. (eds.): Advances in Automated Text Summarization. MIT Press (1999) 81-94

[15]Available: http://oromodictionary.com/aboutOromo.php last [Accessed: October 25, 2012]

[16]Available: http://www.oxforddictionary.com/words/punctution [Accessed :April 25 2013]

[17]Dalianis H, "SweSum – A text summarizer for Swedish", Technical report TRITANAP0015, IPLab-174, NADA, KTH, 2000.

[18]In Mani, I. and Maybury, M., eds., "Proceedings of the A CL/EA CL '97 Workshop on Intelligent Scalable Text Summarization.", Myaeng, S.H., Jang, D.: Development and Evaluation of a Statistical Based Document System.

[19]Eyob Delele Yirdaw, " Topic-based Amharic Text Summarization", Master's thesis, Faculty of Computer and Mathematical Science, Addis Ababa University, 2011

[20]Jun-Jie Li and Key-Sun Choi., "Corpus Based Chines Text summarization System.", CSLab, Ceneter for AI Research, Korea Advanced Institute of Science and Technology Taejon, Republic of Korea, page 1.

[21]Halliday, Michael and Ruqaiya Hasan ," Cohesion in English Longman, London, 1976.

[22]Lin, C.-Y. ,"Training a selection function for extraction", Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM) (Kansas City), 1–8, 1999.

[23]Kamil N., “Automatic Amharic News Text Summarization.”, MSc thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2004.

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

[25]McKeown, K. R. and Radev, D. R. "Generating summaries of multiple news articles", In Proceedings of SIGIR '95, pages 74-82, Seattle, Washington, 1995.

[26]Jen-Yuan Yeh et al., "Chinese Text Summarization Using a Trainable Summarizer and Latent Semantic Analysis", E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 76-87, 2002.

[27]Jiawei Han and Micheline Kamber, " Data Mining Concepts and Techniques Second Edition", Morgan Kaufmann Publisher: 2006. pp 300 - 301

[28]Lin, C.Y. and Hovy, E., "Identify Topics by Position", Proceedings of the 5th Conference on Applied Natural Language Processing, March, 1997.

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

[30]Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization.", Extractive Techniques Journal of Emerging Technologies in web intelligence, Vol. 2, No. 3, August 2010

[31]George Pachantouris, " GreekSum", Master Thesis, DSV,2005

[32]NimaMazdak; A Persian text summarizer Master Thesis , Stockholm University

[33]Joel LaroccaNetoand et al., "Automatic Text Summarization using a Machine Learning Approach", Pontifical Catholic University of Parana (PUCPR) RuaImaculadaConceicao, 1155

[34]Martin Hassel," Evaluation of Automatic Text Summarization", Licentiate Thesis, Stockholm, Sweden, 2004.

[35]Sisay Adugna, " English-Oromo Machine Translation: An Experiment Using a Statistical Approach, Master's thesis", Faculty of Informatics, Addis Ababa University, 2009.

[36]Inderjeet Mani, Mark T. Maybury," Advances in Automatic TextSummarization", [E book] pp 1,1999

[37]Endres-Niggemeyer, B., "Human-Style Www Summarization.", 2000.

[38]Inderjeet MANI, "Summarization Evaluation: An Overview", The MITRE Corporation, W640 11493 Sunset Hills Road Reston, VA 20190-5214, USA

[39]Agustín Gravano et al., " Affirmative Cue Words in Task-Oriented Dialogue", Association for Computational Linguistics: 2011

[40]Ani Nenkova and Kathleen McKeown. Automatic Summarization: Foundations and Trends in Information Retrieval Vol. 5, Nos. 2–3 (2011) 103–233

[41]Oi Mean Foong et al., "Challenges and Trends of Automatic Text Summarization", International Journal of Information and Telecommunication Technology, Vol. 1, Issue 1, 2010 ISSN: 0976–5972,2011

[42]Goldstein, J., "The Use of Genre in Summarization",2009

[43]Inderjeet Mani et al. SUMMAC: a text summarization evaluation: Natural Language Engineering 8 (1): 43-68: 2002

[44]Available: http://www.cs.otago.ac.nz/staffpriv/alik/papers/apps.ps [Accessed: December 8 2012]

[45]Diane J. Litman: Cue Phrase Classification Using Machine Learning: Journal of Artificial Intelligence Research 5 (1996) 53-94

[46]Minel, J-L., Nugier, S., and Piat, G. "How to appreciate the quality of automatic text summarization. In Mani, I. and Maybury, M., eds., Proceedings of the ACL/EACL97 Workshop on Intelligent Scalable Text Summarization, pp. 2530, 1997

[47]Meiws C.G.(2001), “A grammatical sketch of Written Oromo”, ISBN 3-89645- 039-5.

[48]Lawrence Wong, "Automatic News Summarization and Extraction System, MEng Computing Imperial College Dept. of computing http://www.doc.ic.ac.uk/~lkhw98/project

[49]H. P. Luhn, “The automatic creation of literature abstracts,” IBM Journal of Research and Development, vol. 2, no. 2, pp. 159–165, 1958.

[50]Halliday, M.A.K and Ruqayia Hasan, "Cohesion in English. London", Longman, 1976

[51]Makbule Gulcin Ozsoy et al.: Text Summarization of Turkish Texts using Latent Semantic Analysis

[52]Willett, P. "The Porter stemming algorithm: then and n", Program: electronic library and information systems , 40 (3). pp. 219-223, 2006.

[53]Joel Larocca Neto et al,"Automatic Text Summarization using a Machine Learning Approach", Pontifical Catholic University of Parana (PUCPR) Rua Imaculada Conceicao, 1155

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

[55]Kamal Sarkar, "Using Domain Knowledge for Text Summarization in Medical Domain", International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009

[56]Abebe Abeshu, "Automatic morphological synthesizer for Afaan Oromo", Master's thesis, Faculty Natural science, Addis Ababa University, 2010

[57]Diriba M., “An automatic sentence parser for Oromo language using supervised learning techniques”, Master’s thesis, School of graduate studies, Addis Ababa University, Ethiopia, 2002.

[58]Available:http://www.ehow.com/how_7706132_summarize-newspaper-article.html [Accessed: April 30, 2013]

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

[60]Available:http://srmo.sagepub.com/view/encyclopedia-of-measurement-and-statistics/n135.xml [Accessed : May 9, 2013]

[61]Robin, J. ,"Revision-based generation of natural language summaries providing historical background: corpus-based analysis, design and implementation. ", Ph.D. Dissertation, Columbia University, 1994.

[62]Morris, A., Kasper, G., and Adams, D. 1992. The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance. Information Systems Research, 3(1), pp. 17-35. Reprinted in Mani, I., and Maybury, M., eds., Advances in Automatic Text Summarization, MIT Press, pp. 305-323.

[63]Available: http://www.internetlooks.com/onnumbers.html [Accessed: April 29 2013]

[64]Available: http://www.isi.edu/~hovy [Accessed: June 20 2013]

[65]Komishinii. Aadaaf Turizmii Oromiyaa, *"Caasluga Afaan Oromoo"*, *Jildi I*, Finfinnee, Ethiopia. 1995

[66]Catherine Griefenow-Mewis, Wilhelm J.G. Möhlig and Bernd Heine," A Grammatical Sketch of Written Oromo (Grammatical Analyses of African Languages vol. 16)", 2001

[67]Kroeger Paul, Analyzing, "Grammar:An IntroductionCambridge",University Press, May 5, 2005

[68]Available:www.sas.upenn.edu/African_Studies/Hornet/Afaan_Oromo_19777.html[Accssessed: May 20 2013]

[69]Available:http://people.dsv.su.se/~hercules/textsammanfattningeng.htm [Accssessed: May 1 2013]

List of Appendixes

Appendix A. The corpus of Afaan Oromo cue phrases

English	Afaan Oromo Meaning
above all	Hunda calaa| Hundaarra| hunda dura|hundaanolitti
accordingly	Haluma kanaan|kanaafuuhaaluma sanaan| sana waliin
actually	dhugumaan| qabatamaan
admittedly	Shakkii maleehaqaaf| walirraa fuudhuu| amanuu
after	duuba|maayii| Booda
after all	Kanan booda| Sirumaa| hunda caalaa|hundaa ol| Hunda booda|dhumarratti
after that	Sana booda| sana boda| achhin boda
afterwards	Isa booda| iiti aansee| Boodarra
again	amoo|ammas|lammessa| itti dabalees
all in all	Dhibbaa dhibbatti/ guutummaatti| guutuun| Guutumaan guututti| dimshaashummatti
all the same	Hata'u malee| yeroo hunda| Hundu walfakkaata| gama hundaanuu
also	dabalataan|cinaan |innis|sunis|kunis
alternatively	Karaa biraa| Gamaa Birrattin| Akka Fillanotti|filannof|carraa biroo| yookan
although	ta,us| ta'uu illee
always assuming that	Yeroo mara yaaduun| Yeroo hundaa yoo akkasitti yaadame
and	fi
and/or	Fi/ykn/ kana ykn sana|yookin
another time	yeroo kan birra| booda|yeroo biroo|maa'essa
anyway	Karaa kamuu/ hata'u malee|ta'us| Yaa ta'u| sanas ta'e kana
apart from that	Sana malee| isaa as| sanas| Kana malees| kanaan alatti
as	akka| haala| sababa kanaaf
as a consequence	kanaafuu| akka itti aanutti|bu'aa| Kana irra kan ka'e
as a corollary	Dhiibbaa|bu'aa
as a result	kanaafuu| kana irra kan ka'e
as it happened	Akka ta'etti/ akka tasaa|osoo hin yaadin
as it is	Akka jirutti
as it turned out	Osoo hin yaadin
as long as	Yoo|sun yoo ta'e

as luck would have it	Akka carraa/ akka tasaa
as soon as	Amma/ booda/ yeroo| Bakkumatti| akkuma sanaan
as well	Dabalataan| wajjin|waliin| Dabalatamaan| akkasuma
at any rate	Karaa kamuu/ haala kamiinuu| Daddafina kaminu
at first	jalqabarratti| jalqabaaf|yeroo duraa| tokkoffarratti| dursa
at first sight	Irra keessa/ osoo gadi hin fagaatin| Ilalch durarratti|mil'uu duraan
at first view	Ilaalcha jalqabaarratti| Millandhaa dura
at last	dhumarratti
at least	Yoo xiqqaate| yoo xiqqaatee xiqqaate
at once	Al tokkotti| si'a tokkotti| Yeroo tokkoon
at that time	Yeroo sanatti
at the moment	Amma |yeroo kanatti| yeroos|yammus
at the outset	jalqabarratti
at the same time	Walfaana/yeroo tokkotti
at which point	Sababa kamiif| bakka kamitti
back	duuba|teella
because	sababa
before	dursee| dura
before then	san duraa| achiin dura
before long	Yeroo xiqqoo booda|osoo hin turin|dhiheenyaan
Before now	kana dura/yeroo darbe|amman dura
before ever	Hunda dura| kamiyuu dura
besides	bira|cinaa| Dabalatamaan| itti dabalees
but	garuu
but then	yeroos garuu|yaggus garuu
by all means	Sirridhumatti|dhugumatti|Karaalee hundaa|yaalii huundaan|gama kamiinuu
by and by	Yerootti/ osoo hin turin| Ammas ammas| yeroodhaa yerootti
by comparison	Wal dorgomsiisuun/walbira qabuun| Wal cinaa qabuudhan
by contrast	Karaa biraa|wal faallessuun| Garaagarumaa
by the same token	Karaa wal fakkaatuun
by the time	Yerootti|yeroo sanatti| yommus|yeroos|akka sanaan
by the way	Osoo jennu|gidduutti|osoo dubbannuu| ani kanan jedhu
certainly	Dhugumaatti/shakkii malee|Haqumman| qabataman
clearly	Ifatti| dhoksa malee| mulinatti| ifa qabessan

come to think of it	yaadati seenu| Mee itti yaadii
consequently	kanaafuu|waan ta'eef| kana irra kan ka'e
considering that	Kanaafuu| sababa kanaan| Ilaalcha kessa galchun| Yaada keesa galchuudhan
conversely	Karaa biraa/karaa faallaa ta'een
correspondingly	Bifa walfakkaatuun| walqabatee
despite this	yoo kana |ta'e iyyuu| ta'ullee
despite the fact that	Hata'u malee| Dhugaan akkasiis ta'u
each time	Yeroo hunda| yeroo mara
earlier	abboroo|subii| Durarratti
either	kan yookiin sana
else	kana malee
equally	walqixhummaa
especially because	Hunda caalaa sababnisaa| Keessattu sababniisaa| sababni addumaan
especially if	Hunda caalaa yoo| Keessattu yoo| yoo addumaan
especially when	Hunda caalaa yeroo| Keessatuu yeroo
essentially then	Barbaachisaa,yoosan| yeroos faayid-qabeessumman
essentially	Bu'uuraan|dhugumaan|barbaachisummadhaan
even	Ammayyuu/ hanga-|iyyuu|isaa|isii|sana iyyuu
even after	Isa boodayyuu
even before	Isa durayyuu
even if	Hata'u malee |ta'u malee|yoo ta'e iyyuu|ta'us
even so	Hata'u malee|ammayyuu
even though	Hata'u malee| Yaa ta'uu malee
even when	Yeroo kamuu
eventually	dhumarratti
ever since	hanga ammaatt|yoonaatti
every time	Yeroo kamuu/ yoomiyyuu| Yeroo hunda
everywhere	Eessayyuu/bakka hunda
except	sanamalee, iboo
except after	Booda malee
except before	Dura malee
except when	yeros malee| yeroo sana irra kan hafe
failing that	bakka hin jirrettti| sana osoo hin darbin
finally	Dhumarratti| Xumurarratti

first	tokkoffaa|duree
first of all	Hunda dura| Hunadura
firstly	Dursee| Jalqabaaf| Tokkoffarratti
following this	Itti aansee|kanatti aanee| Ittiaansuun
for	fi|akka|dhaaf|sababa kanaaf
for a start	jalqabaaf|eegaluuf
for another thing	Kan biraaf|gama biraan| Wanta biraatiif
for example	fakkeenyaaf
for fear that	Yoo/soda sanaa| Isa sodaaf|kana sodaachuun
for instance	fakkeenyaaf
for one thing	Tokkoffaa|gama tokkoon| Sababni tokko
for one,	tokkoof|jalqabuuf
for that matter	sanaaf|sababa kanaaf
for the simple reason	Sababa salphaa sanaaf| Sababni xiqaan
for this reason	Sababa kanaaf| waan kana ta'eef
fortunately	Carraa ta'ee| akka carraa| Akka carraa
from then on	Sana booda/ sanatti aansee| yerosi kasee| yerosi Jalqabee| Yeroo sani kaasee
further	garas|fagoo| Dabalataaf
furthermore	dabalataan| gadii fagenyan | Kanbiraa|Irra guddessaan|dabalatan
given that	Yoo ta'e| isa kennun| Yoo akkas ta'ee
having said that	jedhee| akka sana jechuun| Kana akkas jedhee ergan fixee |hagana erga jennee
hence	kanaafuu|yeroos
however	Garuu/ ta'uyyuu|haa ta'u malee
I mean	jechuun|kanan jedhu|kan ani jedhu|kanaan jechuu fedhe
if	yoo
if ever	yoomiyyu
if not	Yoo hin ta'u ta'e| Yoo ta'uu baate|ta'uu baatu
if only	Yoo ta'e malee| yoo ta'e qofa
if so	Erga ta'ee| yoo ta'e malee
in a deferent vein/way	Karaa biraa| karra addaa
in actual fact	dhugaaf|akka dhugaatti|Haala qabatamaan
in addition	dabalataan| akka dabalaatti| Itti dabaluun

in any case	Karaa kamuu| Haaluma huundaanuu| sababa kamiinuu
in case	yoo| sababa kanaan| Tari|Yoo sana|kana|akkas ta'ee
in conclusion	dhumarratti|xumerrarratti|akka xumuraatti|walumaagalatti
in contrast	Karaa biraa| akka faallaatti
in doing this	Kana hojjachuun| Kana gochuudhan
in fact	dhugaaf|dhugaatti|dhugumman
in other respects	haala biraattiin|gama biraan|ilaalcha biraan
in other words	Jecha biraan| karaa biraa| Kana jechuun
in particular	qofaatti|addumaan
in short	gababumatii
in so doing	Akkasitti/ haala kanaan| akkas gochuun
in spite of that	Ta'uyyuu/garuu/hanga ammaatti|sanuma cinaatti|ta'ullee| Sana ta'uus
in sum	waliigala|ida'amaan|dimshaashumatti
in that	sababni| achi kessatti|sana keessatti
in that case	Karaa kana/kanaafuu|sababa sana keessatti|sababa sanaan
in that respect	haala sana keessatti
in the beginning	jalqabarratti|dursa
in the case of	Karaa kanaa|sababa kanaan
in the end	dhumarratti| xumurarratti
in the event	Akka jirutti/osoo hin yaadin/ta'umsa keessa|ta'insicha kessatti
in the first place	Hunda dura/hunda caalaa| Tokkoffarratti
in the hope that	Abdiin| Abdidhan | sana abdachuu keessatti
in the meantime	Wayitii sana/yeroo sana|yerooma sana|gidduu sanatti|gidduutti
in this way	Karaa kana| haaluma kanaan
in turn	Dabareen/ tokko tokkoon| wal duraa duubaan
in which case	Karaa kamiinuu| sababa kamiinuu|sababa sana kessatti
in as much as	hanga| baayinuma kanaan|hammanatti
incidentally	Utuu hin irraanfatin/akka tasaa| Tasumatti|akka carra
indeed	Isa dhugaa/ shakkii malee|hojiin|dhugumaan
initially	jalqabarratti|akka ka'umsaatti| Dursa irratti
insofar as	Hamma/hanga| sanaan olitti|hanga kanatti|sanatti
instantly	battalumatti|yerosuma
instead	Bakkasaa/iddoosaa|sanaa| irra|kana irra
it follows that	Itti aanee/iti fufee|sana hordofuun| Sanitti fufuun|kana irra kan

	maddu
it is because	sababiibsaa|sababa kanaaf
it is only because	Sababii qofaa|sababniisa|wan kana qofa ta'eef
it might appear that	Akka jedhametti/fakkaachuu|sana ta'uu danda'a
it might seem that	Tarii fakkaachuu
just	Amma|reefuu| amma kana
just then	Yeroo sanatti| yerosuma|akkuma sanaan
largely because	Bal'inaan sababni| caalmatti sababni| sababni guddan
last	hafaa|dhuma|mayii
lastly	Dhumarratti
later	boodaan| Gulana
lest	ta'uu baannaan
let us assume	hajennu|jenne haayaannu| Akkas jenne yaa yaadnu|akkasitti haa fudhannu
like wise	yeroo akkasii|yoos|akkasumas
luckily	Carraa ta'ee| akka carraa
mainly because	Hunda caalaa sababni|Sababni ijoon|sababni guddaan
meanwhile	Yeroo sanatti|akkuma ta'een|akkuma sanaan
merely because	Salphaatti sababni|sababa kanaan xiqqo|sababa kana qofaaf
mind you	hubadhu| qalbeffadhu
more	dabalee|caalaa
moreover	Dabalataan|sanaaol| Kana caalaa|irra caalaan
most	baayyinaan|caalmaan|irra guddeessaan|hunda caaalaatti
much as	hanga danda'ame
much later	Boodarra|duubarra
much sooner	Bay'ee dafee|arittiin akkumasanaan|hatatamaan|ariitiin
naturally	Uumamaan|akka eegametti
neither is it the case	sababa ta'u dhiisuu| innnis sababa miti
nevertheless	Garuu/ karaa biraa|kamiin gaditti
not	hinta'u |hin taane/miti
not because	Sababa hin taane
not only	qofa osoo hintaane
not that	sana mitii| Isa mitii
notably	kan bekamu|beekamtumman
notwithstanding that	sana osoo hin mormin|osoo hin faallessin

now	yeroo ammaa, ammma
now that	sanaan as| achiin as
obviously	ifatti
of course	dhugumaatti
on balance	madaalli kanaan|Qixxumman
on condition that	haala kanaan| yoo akkas ta'ee
on one hand	gama tokkon
on one side	gama biraatin
on the assumption that	haa jennuu|tolmaama|yaada sanarratt
on the contrary	faallaa kanaatiin, faallaa sanaatiin
on the grounds that	bu'uurra kanaatiin|sababa sanaan
on the one hand	karaa tokkoon
on the one side	gama tokkon
on the other hand	gama birootin
on the other side	karaa biraatin
on top of this	kanaa ol
once	al tokko
once again	ammalle|irra deebi'ee|ammas|itti dabalee|irra deebii agrsiis
once more	tokko caalaa
only	qofa
only because	qofa waan ta'eef|sababa kana qofaaf
only before	dura qofa|sana dura qofa
only if	yoo ta'e|yoo ta'e qofa
only when	yoon sana qofa
or	yookin
or again	yookiin ammas
or else	kan biroo yookiin
originally	ka'umsarra|asliirraa|uumamumaan
otherwise	gama biraatiin|kana ta'uu baannaan
overall	jimlaa|dimshaasha
particularly when	yeroo addumaan
plainly	haala ifaa ta'een
presently	amma|dhiyootti|si'ana
presumably because	sababnisaa akka yadamutti

previously	dura|sila
provided that	yoo akkana|akkasana ta'e
providing that	sana gumaachuudhaan
put another way	karaa biraatiin|gama biraatiin yoo ibsamu
rather	kan ta'uuf malu|kana irra
reciprocally	faallaan
regardless of that	ilaalcha kessa osoo hin galchin|sanaan alatti
regardless of whether	sana yoo ta'een ala
second	lamma
secondly	lammaffaa|2ffaa
seeing as	akkanatti|yoo ilaalame
similarly	halumma wal-fakkatun
simply because	salphamatti sabaaba|salphaadhumatti sababnisaa
simultaneously	walfaana|ergaan takkaa
since	hanga-eega|sababa
so	wantata'eef
so that	kanaafuu
soon	ammuma|amma kana
specially	haala addaatiin|adduman
still	hanga ammaatti
subsequently	sana booda|itti aansee
such that	kanneen jedhaman
suddenly	akka tasa|akka daguu|battalumatti
summarizing	cuunfuuf|goolabuuf
summing up	walitti qabaattii|dimshaashumatti
suppose	haa jenu
suppose that	sana jenne haa yaadnu|kana jenne haa yaadnu
supposing that	akkasitti yaaduun
sure enough	dhugaadhumatti
surely	mirkanaan
that is	inni sun|sunis|inni sunis
that is to say	akkas jechun| kan jechuu barbaade|akkas jechuun
that's all	kan jechuun barbaade kanuma
that's how	akks jechuu
that's when	yammus jechuu

that's why	sababnisaa
the fact is that	dhugaan
the first time	si'aa durattif|jalqabarratti
the moment	si'a sanatti
the more often	caalmaati deddebi'ee kan mullatu
the next time	yeroo ittaanuu
the one time	yeroo tokko|yeroo sanatti
the thing is	wantichi
then	sana boodaa|ittiaanuun|achumaan
then again	saniin booda|itti aansee|achiin booda
thereby	achi|achumarran
therefore	kanaafuu
third	sadii
thirdly	sadaffaa
this means	kana jechuun
this time	yeroo kana
though	ta'ullee|yaa ta'u malee
thus	kanaafuu|achirran
to be precise	ifa tasiisuuf
to be sure	mirkaneeffachuuf|dhugoomsuuf
to begin with	itti eegaluuf|itti calqabuuf|jalqabarratti
to conclude	goolabuuf
to make matters worse	hammeessuuf|wantoota hammataa taasisuuf
to start with	ittiin eegaluuf| ittiin jalqabuuf|dursa
to sum up	walitti qabuuf|
to summaries	cuunfuuf
to take an example	fakkenya kaa'uuf|fakkeenya dhiyeessuf|akka fakkeenyaatti
to the degree that	hanga sanatti
too	baayyee|akkasuma|walfakkaataa
true	dhugaa|haqa
ultimately	kan dhumaa|kan xumuraa|olaanaa|daraan olaanaa
undoubtedly	shakkii malee
unfortunately	akka carraa
unless	ta'uu baannaan

until	hamma
until then	hamma sanatti
we might say	tarii kana jechuun dandeenyu
well	gaarii|tole
what is more	kana caalaa|caalaan|irra guddessaan
when	yoom
whenever	yeroo kamuu|yoomiiyyuu
where	eessa
whereas	gama biraatiin
where in	sana keessatti|achii keessatti
whereupon	achirratti
wherever	essattuu
whether or not	ta'us ta'uu baatuus
which is why	sababanisaas|kun sababnisaa
which means	kana jechuun
which reminds me	yaada kana sammuutti qabachun|kun kan naadachiisu
while	yeroo sana
whilst	nearl same
with that	gama sanaan|sana waliin
yet	hanga ammattuu|hanga yoonaatti|ammayyuu
you know	beektee|bartee

Appendix B. The corpus of Afaan Oromo synonyms

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

Appendix C. The corpus of Afan Oromo abbreviation

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

Appendix D. The corpuses of Afan Oromo stop words

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

Appendix E. The Corpus of Afaan Oromo time, Date's, Month's

Maqaa guyyaa (maqaa guyyaa torbanii)/Name of Weeks		Maqaa Baatii (ji'a)/ Name of Month	
Afaan Oromo	English	Afaan Oromo	English
Wixata|Dafinoo|Hojjaduree|Isinina	Monday	Amajjii |Ama.	January
Kibxata |Facaasaa| Lammaffo|Balloo|Salaasa	Tuesday	Guraandhala|Gur.	February
Roobii |Arbii	Wednesday	Bitootessa|Bit.	March
Kamisa	Thursday	Caamsaa |Cam.	April
Jimaata	Friday	Ebla|Ebl.	May
Sambata|Sambat'xinaa|Sambataduraa|Sabtii	Saturday	Waxabajjii|Wax.	June
Sambata-guddaa| Aalaada	Sunday	Adoolessa|Ado.	July
		Hagayya|Hag.	August
		Birraa |Fulbaana |Bir.	September
		Onkoloolessa|Onk.	October
		Sadaasa|Sad.	November
		Mude|Arfaasaa |Afraasaa| Arf.	December
Maqaa Yeroo/ Names of Time			
Ganama		Saafaa	
Waaree		Darbamtuu	
Guyyaa		Galgala	
Iyyalukku		Waarii	
Obboroo		Halkanii	
Barraqa		Subii	

Appendix F. The Corpus of Afaan Oromo numbers

0	Zeroo	Kudhaa	Jatamii	Billiyonaa
1	tokko	Digdama	Torbaatama	Triliyoona
2	lama	Digdamii	Torbatamii	Kutriliyoona
3	sadii	Soddoma	Saadeettama	zeeroo|duwwaa
4	afur	Sodomii	Sadetamii	
5	Shan	Afurtama	Sagaltama	
6	Ja'a	Afurtamii	Sagaltamii	
7	torba	Shantam	Dhibaa	
8	saddet	Shantamii	Kuma	
9	sagal	Jahaatama	Miliyoona	

Appendix G: The corpus of Afaan Oromo verb suffixes

a	adhee	amtu	atte	dani	ine	itani	nnu	ta	uuttan
aa	adhuu	amtuu	atti	de	inu	ite	nu	tani	uutti
aaf	adhuu	amu	attu	dha	is	iti	ooftna	taniittu	xa
aas	ama	amuu	atu	dhe	isa	itu	oofta	te	xani
aat	amaa	amuudhaa	chiisa	dhu	isan	ja	ooftan	tetta	xe
aatii	aman	amuudhaf	chiisan	di	ise	jani	oofte	teetii	xi
aatu	amani	amuuf	chiise	du	isisa	je	oofti	ti	xu
achisa	amanii	amuun	chiisna	duu	isise	ju	ra	tu	xuu
achiisan	ame	ani	chiisne	e	isisna	la	re	tuu	
achiise	amne	aniiru	chiiste	eera	isista	le	ru	u	
achisna	amni	anna	chisiisa	ees	isistan	lu	se	ulle	
achistan	amoo	anne	chisiisan	eet	isiste	na	sisa	umsa	
achiste	amta	annu	chisiista	eeti	isna	naan	sisan	uu	
achuu	amtan	ata	chisiistan	i	istan	ne	sise	uuf	
achuuf	amtani	atani	chisiiste	ifna	iste	neerra	sisna	uufan	
adha	amte	ate	chisiistu	ifte	isu	ni	sisne	uufi	
adhe	amti	atini	da	ina	ita	nna	siste	uufii	

Appendix H: The Corpus of Afaan Oromo noun suffixes

aa	eeyyii	iinis	irratti	llee	ooliin	s
aaf	f	iis	irrattillee	n	ooliiwwan	tii
an	I	illee	irrattis	ni	ooma	tiin
aniif	icha	irraa	irrattuu	oolee	oota	tu
aniin	ichi	irraahille	itti	ooleedhaan	ootaaf	uma
arraa	ii	irraahis	ittii	ooleef	ootaan	umaa
atti	iif	irraahuu	ittiin	ooleen	ootadhaan	umaaf
dhaa	iifis	irraan	ittillee	ooleewwan	ootawwan	umaafillee
dhaaf	iifuu	iraannille	ittis	oolii	ootni	umaanille
dhaan	iin	iraanis	ittuu	ooliidhan	oottan	umaanis
een	inille	irraanuu	lee	ooliif	rra	umaanuu

Appendix I: Validation Summary Preparation Guideline

Addis Ababa University
College of Natural Science
Department of Computer Science

Dear respondent,

The purpose of this questioner is to design Automatic Afaan Oromo news text summarizer. The system generates extractive type of summary for each of the input text. An extractive summary is created by selecting a certain number of sentences that are judged to be the most important out of the original text.

Hence, this appendix describes guideline and instructions that you follow to prepare summay of a given topic. You are requested to form an extractive summary for each of the text you are given. An extractive summary is created by selecting a certain number of sentences that are judged to be the most important out of the original text. When sentences are selected for inclusion in the summary all that needs to be considered is the importance of the sentences. Furthermore, you can underline a sentence for inclusion in a summary and leave out a sentence which is not that important but contains information that describes another selected sentence.

Thank you for your cooperation!

Topic # 1
Source: http://www.voaafaanoromoo.com news last accessed date: March 19, 2013

Addunyaa irratti du'a dhukkubi kaansarii fidu sadii keessaa harka tokko ittisuun kan danda'amu ta'uu jaarmayaaan fayyaa addunyaa gabaasee jira. Guyyaa kaansarii addunyaa sababeeffachuun gabaasaan sadarkaa addunyaatti ba'e akka jedhutti biyyoota addunyaa irra jiran keessaa walakkaa caalaan sagantaa ittiin dhukkuba kana ittisan ykn wal'aanan waan hin qabaanneef jecha dandeenye.

Jaarmayaan fayyaa addunyaa akka gabaasetti bara 2008 keessa uummanni miliyoona 7.6 sababaa dhukkuba kaansariin du'an. Waggaa waggaanis namoonni miliyoona 13 ta'an haaraa dhukkuba kanaan qabamu. Jaarmayaan kun akka jedhutti warri kaansariin haaraa qabamanis haa ta'u warri dhukkuba kanaan du'aa jiran 3 keessaa harki 2 biyyoota guddataa jiran keessatti argamu.

Aadaan sochii qaamaa ykn spoortii hojjechuu lafa hin baratamnetti furdinnis hammaatee kanneen tamboo xuuxanii fi dhugaarii alkoolii dhugan dabalaa yeroo jiran kanatti kaansariif saaxilamuun dabalaa deema. Kunis keessumaa magaalaalee biyyoota guddataa jiran adda addaa keessatti faca'ina kaansarii hammeessa. Gama kaaniin ijoolleen durbaa biyyoota guddataa jiran keessa jiraatan dhukkuba kaansarii afaan gadameessaan (Servical Cancer) akka hin qabamneef talaalii ittisu argachuuf jiru. Kaansariin Afaan gadameessaa miidhu kun Hiyumaan Paapiloomaa vaayras ykn HPV jedhamuun kan dhufu yoo ta'u kan daddarbus quunnamtii saalaan ta'uun ibsamee jira.

Addunyaa irrattis sababaa kaansarii kanaan daqiiqaa lama keessatti dubartii tokkotu du'a. Waggaa waggaan dubartoota kuma 200 fi kuma 75tu dhukkuba kanaan dhuma. Kanneen keessaa harka 85 kan ta'an biyyoota guddataa jiran keessa jiraatu. Talaaliin kaansarii afaan gadameessaa ittisuuf kannemu ijoollee durbaa umuriin isaanii waggaa 9 fi 13 gidduu jiraniif kennama.

Gamtaan Gaavii Alaayaans jedhamu talaalii farra kaansarii kana kennuuf shamarran biyyoota guddataa jiran 8 filate. Biyyoota Afriikaa uffee sahaaraa gadii keessaa Gaanaa, keeniyaa, madagaaskaar, maalaawwii, Nigeer, Seeraaliyoon fi Taansaaniyaa akkasumas biyyoota Eshiyaa keessaa laaoos ta'uun baakemee jira.

Appendix J: Result of underlined summary sentence by subjects

Topic #1

Sentence Position														
	Paragraph	1		2		3				4			5	
	Sentence in paragraph	1	2	1	2	1	2	3	4	1	2	3	1	2
Subjects	1	1	0	1	1	0	1	0	1	1	1	0	0	0
	2	1	0	1	1	0	1	0	0	0	0	0	0	1
	3	1	0	1	0	0	0	0	1	1	1	0	1	0
	Decision	1	0	1	1	0	1	0	1	1	1	0	0	0

Topic #2

Sentence Position												
	Paragraph	1				2		3			4	
	Sentence in paragraph	1	2	3	4	1	2	1	2	3	1	2
Subjects	1	1	0	0	0	1	1	0	0	0	1	0
	2	1	0	0	0	0	1	1	0	0	1	0
	3	1	0	0	0	1	0	1	0	0	1	1
	Decision	1	0	0	0	1	1	1	0	0	1	0

Topic #3

Sentence Position														
	Paragraph	1		2		3			4		5	6	7	8
	Sentence in paragraph	1	2	1		1	2	3	1	2	1	1	1	1
Subjects	1	1	1	1	1	0	0	0	0	0	0	1	0	0
	2	1	1	0	1	0	0	0	0	1	1	1	0	0
	3	1	0	0	0	0	1	0	0	1	0	1	1	0
	Decision	1	1	0	1	0	0	0	0	1	0	1	0	0

Topic #4

Sentence Position								
	Paragraph	1		2		3	4	5	6
	Sentence in paragraph	1	2	1	2	1	1	1	1
Subjects	1	1	1	0	0	1	0	0	1
	2	1	1	1	0	1	0	0	1
	3	1	1	0	0	1	0	0	1
	Decision	1	1	0	0	1	0	0	1

Topic #5

Sentence Position										
	Paragraph	1			2		3			
	Sentence in paragraph	1	2	3	1	2	1	2	3	4
Subjects	1	1	0	0	1	1	0	1	0	0
	2	1	0	0	1	1	0	1	0	0
	3	1	0	0	0	0	0	1	0	0
	Decision	1	0	0	1	1	0	1	0	0

Topic #6

Sentence Position								
	Paragraph	1					2	
	Sentence in paragraph	1	2	3	4	5	1	2
Subjects	1	1	0	0	1	0	1	0
	2	1	0	0	1	1	1	0
	3	1	0	1	0	1	1	0
	Decision	1	0	0	1	1	1	0

Topic #7

Sentence Position									
	Paragraph	1			2		3		p4
	Sentence in paragraph	1	2	3	1	2	1	2	1
Subjects	1	1	0	1	1	0	1	0	1
	2	1	0	1	1	0	1	0	0
	3	1	0	1	1	0	1	0	1
	Decision	1	0	1	1	0	1	0	1

Topic # 8

Sentence Position														
	Paragraph	1						2			3			
	Sentence in paragraph	1	2	3	4	5	6	1	2	3	1	2	3	4
Subjects	1	1	1	1	1	0	1	0	0	1	1	0	0	0
	2	0	1	0	0	0	1	0	1	1	0	0	0	0
	3	1	0	0	0	0	1	0	1	0	1	0	0	0
	Decision	1	1	0	0	0	1	0	1	1	1	0	0	0

Topic #9

Sentence Position								
	Paragraph	1		2	3	4	5	6
	Sentence in paragraph	1	2	1	1	1	1	1
Subjects	1	1	1	0	0	1	0	0
	2	1	0	0	0	1	0	0
	3	1	1	0	0	1	0	1
	Decision	1	1	0	0	1	0	1

Topic # 10

Sentence Position													
	Paragraph	1		2			3				4		5
	Sentence in paragraph	1	2	1	2	3	1	2	3	4	1	2	1
Subjects	1	1	0	0	0	1	0	0	0	0	0	0	1
	2	1	0	0	0	1	0	0	0	1	0	0	1
	3	1	0	0	1	1	0	0	1	0	1	1	1
	Decision	1	0	0	1	1	0	0	1	0	0	0	1

Topic # 11

Sentence Position								
	Paragraph	1	2					3
	Sentence in paragraph	1	1	2	3	4	5	1
Subjects	1	1	0	1	0	1	0	1
	2	1	0	0	0	0	0	1
	3	1	0	1	0	0	0	1
	Decision	1	0	1	0	0	0	1

Topic #12

Sentence Position														
	Paragraph	1		2		3		4		5				6
	Sentence in paragraph	1	2	1	2	1	2	1	2	1	2	3	4	1
Subjects	1	1	0	0	0	0	0	1	0	1	0	0	1	0
	2	1	0	0	1	0	1	1	0	1	0	0	1	0
	3	1	0	0	1	0	0	1	0	0	0	1	1	0
	Decision	1	0	0	1	0	0	1	0	1	0	0	1	0

Topic #13

Sentence Position														
	Paragraph	1		2			3	4		5	6			7	
	Sentence in paragraph	1	2	1	2	3	1	1	2	1	1	2	3	1	2
Subjects	1	1	0	0	0	1	0	1	1	1	1	1	1	1	0
	2	1	0	0	0	1	0	1	1	1	1	1	0	1	0
	3	1	1	0	0	1	0	0	0	0	1	0	0	0	0
	Decision	1	0	0	0	1	0	1	1	1	1	1	0	1	0

Topic #14

Sentence Position																
	Paragraph	1							2							
	Sentence in paragraph	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8
Subjects	1	1	0	0	1	1	1	1	0	0	0	0	0	0	1	0
	2	1	0	0	1	1	1	0	0	1	1	0	0	0	1	0
	3	1	0	0	0	0	0	1	0	1	1	0	0	0	1	0
	Decision	1	0	0	1	1	1	1	0	1	1	0	0	0	1	0

Topic #15

Sentence Position																		
	Paragraph	1				2					3							
	Sentence in paragraph	1	2	3	4	1	2	3	4	5	1	2	3	4	5	6	7	8
Subjects	1	1	0	0	1	0	1	1	1	1	0	0	1	1	1	1	1	0
	2	1	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0
	3	1	0	0	0	0	0	1	1	1	0	0	1	1	1	1	0	1
	Decision	1	0	0	1	0	0	1	1	1	0	0	1	1	1	1	0	0

Topic #16

Sentence Position										
	Paragraph	1	2	3	4	5	6	7	8	9
	Sentence in paragraph	1	1	1	1	1	1	1	1	1
Subjects	1	1	0	1	0	1	1	0	0	0
	2	1	0	1	0	1	1	1	0	1
	3	1	0	0	0	1	0	0	0	0
	Decision	1	0	1	0	1	1	0	0	0

Topic # 17

Sentence Position									
	Paragraph	1				2		3	
	Sentence in paragraph	1	2	3	4	1	2	1	2
Subjects	1	1	0	0	1	0	0	1	0
	2	1	1	1	0	0	1	0	0
	3	1	0	0	1	0	1	0	0
	Decision	1	0	0	1	0	1	0	0

Topic #18

Sentence Position												
	Paragraph	1	2	3		4		5		6		
	Sentence in paragraph	1	1	1	2	1	2	1	2	1	2	3
Subjects	1	0	0	0	1	0	0	0	0	1	0	1
	2	1	0	0	1	0	0	1	0	1	0	0
	3	1	0	0	0	0	1	0	0	1	0	1
	Decision	1	0	0	1	0	0	0	0	1	0	1

Topic #19

Sentence Position										
	Paragraph	1	2		3		4			
	Sentence in paragraph	1	1	2	1	2	1	2	3	4
Subjects	1	1	0	1	0	0	0	1	0	1
	2	1	0	0	1	1	0	1	0	0
	3	1	0	1	0	1	0	1	0	0
	Decision	1	0	1	0	1	0	1	0	0

Topic #20

Sentence Position											
	Paragraph	1	2	3	4		5	6	7		
	Sentence in paragraph	1	1	1	1	2	1	1	1	2	3
Subjects	1	1	0	0	1	0	1	0	1	1	0
	2	1	1	0	1	0	0	0	0	1	0
	3	1	1	0	0	0	1	0	1	1	0
	Decision	1	1	0	1	0	1	0	1	1	0

Appendix K: System summary evaluation guide line

Addis Ababa University
College of Natural Science
Department of Computer Science

Dear respondent,
The purpose of this questioner is to evaluate the performance of Automatic Afaan Oromo news text summarizer. The system generates extractive type of summary for each of the input text. An extractive summary is created by selecting a certain number of sentences that are judged to be the most important out of the original text.

Three different summary is generated at the end of the topic after you read the summary evaluate the summary based on the three question listed below. Fill the number given for the summary: for example if the summary informativeness of summary 1 is Good write *2* in the box provided under choice b.

1. **The summary informativeness:**

   a. Very good b. Good c. Not Bad d. Poor e. Very Poor

   ☐ ☐ ☐ ☐ ☐

2. **Grammar, non-redundancy and referential clarity.**

   a. Very good b. Good c. Not Bad d. Poor e. Very poor

   ☐ ☐ ☐ ☐ ☐

3. **Coherence: - Which summary is more coherent?**

   a. Very good b. Good c. Not Bad d. Poor e. Very Poor

   ☐ ☐ ☐ ☐ ☐

Note:

- **Informativeness**: the best sentences are that contain the most important information of the topic sentence
- **Non-redundancy:-** a summary should not contain unnecessary repetition of whole sentences.
- **Referential integrity:** - while reading the sentences according to their rank order it should be easy to identify who or what the pronouns and nouns phrases in each sentence are refereeing to.
- **Coherence: -** there should be a smooth transition of sentences. While reading the sentences in their rank order it should not just be a heap of related information, but also should build a coherent body of information about a topic/s.

Thank you for your cooperation!

Source*: http://www.voaafaanoromoo.com/content/article/1645099.html accessed on April 26 2013*

Angawoonni Federaalaa Ameericaa har'a suraa fi video shakkamtoota haleellaa Boston raawwatan jedhanii ifa taasisaniiru. Shakkamaan inni duraa kofiyyaa gurraachaa keeyyatee jira, inni lammataa immoo kofiyaa adii gara boodaatti gara galchee keeyyatee jiru agarsiisan.

Angawoonni akka jehdantti shakkamaan koofiyaa adii keeyyatuu kan haleellaa boobmii sanaan shakkamuuf shakkamaa lammataa amma iyyuu hin qabamin jira, meeshaa hidhatee waan jiruuf sodaachisaa dha jedhan.

Qorattonni namoonni lamaan konkolaataa tokko hatanii waaltaa Technology Massachusetts kan bakka Watertown jiruu utuu hin baqatin dura angawaa police waaltaa sanaa tokkotti dhukaasuun ajjeesan. Yeroo sana ergaa poolisiin walitti dhukaasuun sun vidiyoo ogeessotaan hin ta'in ykn vidiyoo amateriin kan waraabame.

Poolisiin yeroo faana dha'aa turetti shakkamtoonni sun konkolaataa isaanii keessaa dhuka'aa darbaa turan. Shakkamaan inni tokko suraa FBIn ka'e irraa akka argametti koofiyyaa gurraacha kaayyatee kan ture si'a ta'u dhukaasa poolsiin walitti banameen madaa'ee hospitaala Beth Israel-tti erga geessamee booda achitti du'uu agawoonni ibsanii jiran.

Dr. David Shoenfeld Hospitaala sana keesaa doktota yoo ta'an jiraataa bakka dhukaasi poolisii fi shakkamtoota gidduutti itti geggeessame bakka Watertown jiraatu. Dhukaasa sana dhaga'een jira jedhu.

Anawoonni Hospitaalaa akka jedhanitti shakkamaan madaa'e sun yeroo hospitaala ga'u onneen isaa dadhabee ture, bakka hedduu rasaasaan rukutamee jira. Madaa hamaa kan dhuka'aan dha'ames qaama isaa irratti ni mul'ata ture. Jimaata har'aa ganama barii dubbi himaan Police kutaa Massachusetts Timothy Alden haala isaa ibsanii turan.

Naannoo shakkamaan lammataa tarii keessa dahatee jira jedhame bakka Watertown jedhamu marsuu dhaan Poolisiin barbaacha isaa itti fufee jira.

Eessummni shakkamtoota kanaa Mr. Gasrniin tuuta oduuf ibsa kennaniin ijoolleen obboleessa koo tii waan sukaneessaa akkasii raawwachuu isaanii hedduu nu qaanesse jedhan. shakkamaan hanga yoonaa hin qabamin jiru mucaan obboleessa isaanii lubbuun jira taanaan harka akka kennu maatii namooni duraa miidhamanis dhiifama akka gaafatu waamicha dabarsanii firoottan namoota du'anii fi madaa'anii argee maatii koo bakka bu'ee jilbeenfadhee dhiifama gaafachuun fedha jedhan.

Shakkamtoonni kun kan obbolaa yoo ta'an umruiin waggaa 19- Dzhokhar Tsarnaev fi kan umuriin waggaa 26-Tamerlan Tsarnaev lammiiwwan chechenya ti.

Summary one

Angawoonni Federaalaa Ameericaa har'a suraa fi video shakkamtoota haleellaa Boston raawwatan jedhanii ifa taasisaniiru. Shakkamaan inni duraa kofiyyaa gurraachaa keeyyatee jira, inni lammataa immoo kofiyaa adii gara boodaatti gara galchee keeyyatee jiru agarsiisan. Qorattonni namoonni lamaan konkolaataa tokko hatanii waaltaa Technology Massachusetts kan bakka Watertown jiruu utuu hin baqatin dura angawaa police waaltaa sanaa tokkotti dhukaasuun ajjeesan. Yeroo sana ergaa poolisiin walitti dhukaasuun sun vidiyoo ogeessotaan hin ta'in ykn vidiyoo amateriin kan waraabame. Shakkamaan inni tokko suraa FBIn ka'e irraa akka argametti koofiyyaa gurraacha kaayyatee kan ture si'a ta'u dhukaasa poolsiin walitti banameen madaa'ee hospitaala Beth Israel-tti erga geessamee booda achitti du'uu agawoonni

ibsanii jiran. Dr. David Shoenfeld Hospitaala sana keesaa doktota yoo ta'an jiraataa bakka dhukaasi poolisii fi shakkamtoota gidduutti itti geggeessame bakka Watertown jiraatu. Jimaata har'aa ganama barii dubbi himaan Police kutaa Massachusetts Timothy Alden haala isaa ibsanii turan.

Summary two

Angawoonni Federaalaa Ameericaa har'a suraa fi video shakkamtoota haleellaa Boston raawwatan jedhanii ifa taasisaniiru. Shakkamaan inni duraa kofiyyaa gurraachaa keeyyatee jira, inni lammataa immoo kofiyaa adii gara boodaatti gara galchee keeyyatee jiru agarsiisan. Angawoonni akka jehdantti shakkamaan koofiyaa adii keeyyatuu kan haleellaa boobmii sanaan shakkamuuf shakkamaa lammataa amma iyyuu hin qabamin jira, meeshaa hidhatee waan jiruuf sodaachisaa dha jedhan. Qorattonni namoonni lamaan konkolaataa tokko hatanii waaltaa Technology Massachusetts kan bakka Watertown jiruu utuu hin baqatin dura angawaa police waaltaa sanaa tokkotti dhukaasuun ajjeesan. Shakkamaan inni tokko suraa FBIn ka'e irraa akka argametti koofiyyaa gurraacha kaayyatee kan ture si'a ta'u dhukaasa poolsiin walitti banameen madaa'ee hospitaala Beth Israel-tti erga geessamee booda achitti du'uu agawoonni ibsanii jiran. shakkamaan hanga yoonaa hin qabamin jiru mucaan obboleessa isaanii lubbuun jira taanaan harka akka kennu maatii namooni duraa miidhamanis dhiifama akka gaafatu waamicha dabarsanii firoottan namoota du'anii fi madaa'anii argee maatii koo bakka bu'ee jilbeenfadhee dhiifama gaafachuun fedha jedhan.

Summary three

Angawoonni Federaalaa Ameericaa har'a suraa fi video shakkamtoota haleellaa Boston raawwatan jedhanii ifa taasisaniiru. Angawoonni akka jehdantti shakkamaan koofiyaa adii keeyyatuu kan haleellaa boobmii sanaan shakkamuuf shakkamaa lammataa amma iyyuu hin qabamin jira, meeshaa hidhatee waan jiruuf sodaachisaa dha jedhan. Qorattonni namoonni lamaan konkolaataa tokko hatanii waaltaa Technology Massachusetts kan bakka Watertown jiruu utuu hin baqatin dura angawaa police waaltaa sanaa tokkotti dhukaasuun ajjeesan. Shakkamaan inni tokko suraa FBIn ka'e irraa akka argametti koofiyyaa gurraacha kaayyatee kan ture si'a ta'u dhukaasa poolsiin walitti banameen madaa'ee hospitaala Beth Israel-tti erga geessamee booda achitti du'uu agawoonni ibsanii jiran. Jimaata har'aa ganama barii dubbi himaan Police kutaa Massachusetts Timothy Alden haala isaa ibsanii turan.shakkamaan hanga yoonaa hin qabamin jiru mucaan obboleessa isaanii lubbuun jira taanaan harka akka kennu maatii namooni duraa miidhamanis dhiifama akka gaafatu waamicha dabarsanii firoottan namoota du'anii fi madaa'anii argee maatii koo bakka bu'ee jilbeenfadhee dhiifama gaafachuun fedha jedhan.

Appendix L: Subjective test data evaluation result

I. Informativeness

	Exp 1					Exp 2					Exp 3				
Tests	S1	S2	S3	S4	S5	S1	S2	S3	S4	S4	S1	S2	S3	S4	S5
Test #9	5	5	5	3	4	2	1	4	3	3	4	2	3	4	5
Test #10	4	5	3	5	2	2	2	2	1	1	5	3	2	4	5
Test #11	5	4	5	3	4	2	3	1	2	1	3	4	2	3	3
Test #12	5	5	4	4	5	3	2	2	1	2	2	3	3	3	3
Test #13	5	5	4	5	5	3	2	2	2	3	3	2	3	3	2

II. Referential integrity and Non redundancy summary result

	Exp 1					Exp 2					Exp 3				
Tests	S1	S2	S3	S4	S5	S1	S2	S3	S4	S4	S1	S2	S3	S4	S5
Test #9	3	3	4	4	3	3	2	2	2	3	3	2	4	3	3
Test #10	4	4	5	3	3	1	1	1	1	2	4	4	3	3	2
Test #11	5	3	4	3	4	2	1	2	1	2	2	3	3	4	3
Test #12	5	3	3	5	4	1	1	1	1	1	3	3	3	2	3
Test #13	3	3	4	5	4	2	2	2	1	3	3	2	3	3	4

III. Coherence Summary result

	Exp 1					Exp 2					Exp 3				
Tests	S1	S2	S3	S4	S5	S1	S2	S3	S4	S4	S1	S2	S3	S4	S5
Test #9	3	3	4	4	3	3	2	2	2	3	3	2	4	3	3
Test #10	3	2	4	3	4	1	1	1	1	2	4	4	3	3	2
Test #11	3	5	4	4	3	2	1	2	1	2	2	3	3	4	3
Test #12	3	3	4	4	2	1	1	1	1	1	3	3	3	2	3
Test #13	4	4	3	3	4	2	2	2	1	3	3	2	3	3	4

Appendix M: List of languages that OTS supports

Basque, Bulgarian, Catalan, Czech, Danish, Dutch, English, Esperanto, Estonian, Finnish, French, Galician, German, Greek, Hebrew, Hungarian, Icelandic, Indonesian, Interlingua, Irish (Gaelic), Italian, Latvian, Malaysian, Maltese, Maori, Norwegian Nynorsk, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, Tagalog (Pilipino), Turkish, Ukrainian, Welsh, Yiddish

Declaration

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

Fiseha Berhanu Tesema

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

Sebsibe Hailemariam, PhD

Addis Ababa, Ethiopia June 2013