Afaan Oromo News Text Summarization Using Sentence Scoring Method

A Thesis Presented

by

Gammachiis Temesgen Olana

to

The Faculty of Informatics

of

St. Mary's University

In Partial Fulfillment of the Requirements
for the Degree of Master of Science

in

Computer Science
January 2021

ACCEPTANCE

Afaan Oromo News Text Summarization Using Sentence Scoring Method

By

Gammachiis Temesgen Olana

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

____

Internal Examiner

____

External Examiner

____

Dean, Faculty of Informatics

January 2021

DECLARATION

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

Gammachiis Temesgen Olana

____

Signature

Addis Ababa

Ethiopia

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

Dr. Michael Melese

____

Signature

Addis Ababa

Ethiopia

January 2021

Acknowledgment

Before every think I would like to give thanks to my Almighty God for his mercy, help, care, guidance to do this work. Next to God I also say thank you to my advisor for his advice.

On the other hand there are many people who has contribution to my work. Among them Mr. Takele Kebede, Ms. Niya Boja and Mr. Galana Senbeta who help me by manual summary preparation I have many thanks for them. And also Mr. Liul who help me in different direction starting from the beginning to the end of my thesis work I have many thanks for him, Rev. Dr. Gofare Andisho and Dr. Lemi Tesfa who help me by reading my thesis and all my staff who help me by different direction such as giving permission to do my thesis especilly Dr. Ebise Gudeta I would like to say thank you.

And my aunt Rev. Shingule Olana I thank you very much for you help me with many direction such as praying, encouraging me by advice and lastly my sister Talile Temesgen who is always with me by cooking for me, generally I would like to say God bless all your time.

Contents

Acronyms

- OBN : Oromia broadcast network
- ETV: Ethiopian Television
- AAOTS: Automatic Afaan Oromo Text summarizer
- VOA: Voice of America
- NLTK: Natural language toolkit
- NLP: Natural language processing
- SVD: singular value decomposition
- RST: rhetorical structure theory
- NAACL: The North American Chapter of the Association for Computational Linguistics
- ACL: anterior cruciate ligament
- LSTM: long short term memory

Lists of Figures

List of Tables

Abstract

Nowadays information is available in both electronic (soft copy) and hard copy format. Due to presence of huge amount of electronic format information it needs lot of time and money to access information. So, to get information in short period of time with minimum amount of money it needs a system which summarize and present it for readers. Therefore, this research attempt on the Afaan Oromo News Text Summarization Using Sentence Scoring Method. The researcher used features like thematic words, word frequency, title words, term weight, cue phrases, name of numbers and sentence position in this work to achieve the study of way of designing and developing single document summarizer for Afaan Oromo news text. So, using extractive method the researcher did experiments on ten selected topics out of 30 gathered topics. Manual summary is prepared by three Afaan Oromo speaker domain expert. The system is developed by NLTK using python programming language. The developed system calculates the score of the sentence by adding the score of each individual words and the score is computed for sentence. The system generates the summary by extracting n top scored sentences at three extraction rate i.e. at 20%, 30% and 40%.

The system was evaluated based on the nine experimental situations both subjectively and objectively. Subjective evaluation focused on the structure of the summary referential clarity, to check as there is any redundancy or not, in-formativeness, grammatical correctness and coherence of the summary. So, at 20%, 30% and 40% extraction rate grammatical correctness is 90%, 90% and 92% respectively, concerning redundancy at 20%, 30% and 40% extraction rate performance of the summarizer system is 72%, 82% and 84% respectively. And at 20%, 30% and 40% extraction rate performs 66%, 74% and 86% in concerning referential clarity. Coherency of the summary evaluation performed at 20%, 30% and 40% extraction rate 62%, 66% and 72% respectively. And concerning informativeness at 20%, 30% and 40% extraction rate the performance of automatic summary was 74%, 78% and 86%. And with that of objective evaluation the three metrics recall, precision and F-score computed and 86.1% was performed by the system.

Key Words: Single document, Text Summarization, Sentence scoring, Extractive summarization.

Chapter One

Introduction

1.1 Background

Today information in our world stored in digital formats mostly, which can be accessed by anyone at any time from anywhere [1]. And because of growth of technology from time to time documents growth on intranet and web is also increasing the information amount accessibility in semi structured and unstructured format [1]. Structured data resides in fixed fields of spreadsheets, databases, etc., while unstructured data refers to free-form texts as in text documents. Whereas, Semi-structured data are data neither reside in a relational database nor just plain textual content, but with some process the data can be transformed and stored in a structured manner [75]. Some examples of unstructured data web pages and free texts (such as News articles), example of semi-structured data is social Medias and hand written HTML and of structured data is database. Many information represented by unstructured textual data available in large size of document that has diverse information is used in security, business and education. Also, fast growth of Internet and its easily accessibility, large amount of document shared on the web has increased exponentially [2]. All this indicates that as the life is become difficult for those who use web to summarize text data manually large size data available of the Internet. Receiving the correct information for decision making from existing abundant unstructured text and non-readiness of tools for extracting applicable information which is effective enough to fulfil the consumers requirement have been a major problem for years [1].

More than 80 different languages in Ethiopia enables people to communicate each other are there [40]. From these more than 80 languages Oromo people who has large size in Ethiopia use a language called Afaan Oromo which is categorized under Cushitic family [1]. This language is the third most widely spoken language in Afrika as a mother tongue, next to Hausa and Arabic [3]. The largest regional state of Ethiopia which is Oromia use this language as an official working language.

Recently, Afaan Oromo language is decided to be the second official working language of Ethiopian federal language by Ethiopian parliamentary this year [49]. Alongside to this, the language is also used in Oromia region including Addis Ababa as a media transmission, medium of instruction at different level such as at primary, secondary and teritary level of education [1] [3].

These days various Afaan Oromo language documents such as journals, magazines, newspapers, online education, books, entertainment media and videos being produced in digital format [11]. Moreover, Medias that use Afaan Oromo language such as Kallacha Oromiya, Oromia Television and Radio (Web news), Bariisaa, Voice of America (VOA) Afaan Oromo (web news), Ethiopian Television and different academic and spiritual Medias are sources of articles used in this work [4]. These sources can be present either as unstructured, semi-structured or structured formats of textual information.

Peoples use as a communication tools to exchange information or to deliver information and arguments within each other in their day-to-day life a language [50]. Natural language processing is area of study that gives education calculating natural language to deliver a possible way of gaining access to data available through intranet and the Internet [5]. And this field of study has an application on area such as: machine translation, information retrieval, text summarization, speech recognition, question answering and etc. [62].

From these applications, one that processes and extracts the most important information in a text to produce a reduced version of the given document is called text summarization [12]. And it gives to the reader an accurate and a complete idea of the contents of the source.

Text summarization method is classified in to abstractive and extractive summarization [19]. An extractive method pick out main sentences, paragraphs from an original text and form summary. These type of summaries are formulated by taking out key text from the document, based on an assembly of fragments and statistical analysis from the original text [19]. Correspondingly, a summary prepared by abstractive tries to develop an understanding of the core concepts in a text and then define those concepts by using another words or its own words in clearly [19]. Thus, there is a requirement to develop an automatic text summarization for Afaan Oromo language.

1.2 Motivation of the study

Afaan Oromo is one of the languages in Ethiopia which has largest number of speakers. The size of Oromo people who use this language is around 40 million and 3rd and 1st ranked single ethnic group in Africa and in Ethiopia respectively [10]. This language is serving as an official language in the largest regional state of Ethiopia called Oromia [3] and also as a medium of instruction from primary to higher education in the country.

Today, online contents of Afaan Oromo news documents are growing very rapidly and mass people are reading this regularly. But it consume time, resource and budget to select necessary information from huge amount of data manually. So it need a system which save time, resource and money to get necessary information. But the unavailability of tools for extracting and exploiting the valuable information, which is effective enough to satisfy the users for Afaan Oromo language has also been a major problem. Hence, above all these facts initiate me to conduct this research.

1.3 Statement of the Problem

Today, the accessibility of digital information is extremely growing from time to time in different formats and the spreading of these resources were over newspaper, Internet, books, media, journal articles and others [10]. Because of growth of technology and change in life style the main communication channel is Internet [62]. Besides, users have no more time to find the main idea of the document from large number of information given to them by taking abstractive perception ignoring the details if they are not interested [11]. Even though working on text summarization is the most motivating, it has difficulties to give to the reader accurate and all ideas of the original text [6]. Since there are a lot of information on the Internet it is difficult for the reader concerning with time to read everything available on the Internet, and people become careless to read long articles and began avoiding it [8]. Therefore, developing a good automated text summarization system is an important solution for this problem which is also one of the problems for under resourced language like Afaan Oromo.

Now a days, many area use large amount of documents that need text summarization to save their time [11]. Some of these ares are such as law students in their education, legal judgments to give justice timely, police sectors to make criminal examination document at different levels, for different offices to write report and etc. [11].

Ethiopia is a multilingual country with more than 80 different spoken language among 7117 registered language in the world [51], and they are drawn into three families [52]. Afro-asiatic family is the first standing family consisting of Cushitic and Sematic languages by size of population use, and the Omotic family language is the second out standing family next to Afro-asiatic family, whereas Nilo Saharan family is the last in Ethiopia. The number of people who speak Afaan Oromo languages is the first with amounts 34.5% [29].

Moreover, this language is used most widely in Ethiopia and the neighbouring countries including Kenya, Somalia and Djibouti [11]. And too much information available in this language is in the form of digital which is difficult to easily search manually and take relevant information [10]. Therefore, the documents should be summarized to avoid sinking in it. Otherwise, users are going to kill their time, lose their source and energy by reading not important part of the material. So, to give solution for this difficullty Automatic Afaan Oromo news summarization using extractive method can be taken as one part of solution by creating automatically summary of the large document.

Understanding all these problems mentioned above, the researcher did in this thesis to improve the performance of the automatic Afaan Oromo news text summarization by adding some new features to the work of the previous researchers work. As Fiseha who did his thesis on title “Afaan Oromo Automatic News Text Summarizer Based on Sentence Selection Function” [10] improved the work of Girma who did on title “Afan Oromo news text summarizer“[11] by adding some features. In this work the researcher added some features to work of Fiseha and Girma by applying extractive method for summary generation. However, their work has a research gap such as: how to handle problems related to thematic words, title words and numerical data.

So, the features the researcher added in this work to improve the previous works are title words, thematic words, capital letters, numerical data and term weight. Thus, the central target of this research work is to design and develop an automatic News Text Summarization for Afaan Oromo language.

In order to design and develop the summarizer system for Afaan Oromo by extractive method the researcher used the following research questions:

- To what extent the features incorporated, affect the performance and quality of the summarizer?
- Which feature is more important and less important, to contribute to the performance of the summarizer?

1.4 Objective of the study

1.4.1 General objective

The general objective of this research study is to design and develop a single document summarizer for Afaan Oromo news text.

1.4.2 Specific objectives

The specific objectives of this research work the researcher used to achieve the general objective are:

- To review literature on text summarization, techniques and algorithms to get the state-of-the-art of text summarization.
- To collect news articles data used to test the system.
- To prepare a corpus for Afaan Oromo.
- To design and develop Afaan Oromo news text summarizer model.
- Conduct different extractive summarization using different compression ratio so as to select the appropriate compression ratio.
- To test and evaluate the performance of Afaan Oromo text summarizer.
- To report the finding of the study based on experimental result and recommend for the future research works.

1.5 Scope and Limitations

The focus of this work is developing automatic Afaan Oromo single document summarizer using extractive method with logical combination of features: title words, thematic words, capital letters, term weight, sentence position, cue phrases, word frequency and name of number handling mechanism.

This study further focuses on the particular nature of news texts to enhance the use of similarity with the title words, thematic words, capital letters, term weight, name of number, cue phrase, keyword frequency, sentence position, sentence length handling mechanism and name of weeks, days and time for summarization. The summarizer doesn’t process document with non-text such as figures, table, image and graphics were not included in the scope of the research.

The researcher gathered from online available sources 30 Afaan Oromo news which have 8-83 number of sentences in different domains such as politics, economic, sports and health news.

From these document, 10 news text documents for testing the summarizer are randomly selected. This work is aimed only to validate the applicability of the output of the study for Afaan Oromo news text documents not other domains.

For evaluating the summarizer system, standard test corpus and evaluation tool for Afaan Oromo language were needed but there was no such a standard test corpus and evaluation tool for Afaan Oromo text news. Lack of standard test corpus and evaluation tool are the limitation of this work.

1.6 Significance of the research

Automatic text summarization is one of the two technologies used for finding out the information in a short period of time successfully while search engine is the other one [10]. A technology that helps to retrieve an initial set of relevant text for filtering information is called search engine. Whereas a technology that is used the user to locate the final set of desired documents is called text summarizers and both are used to save user time [11]. Also, text summarizer generates summary of document that enables users quickly identify and examine the content of the document to control its final set [39].

Hence, AAOTS used for the following profits.

- It helps students to collect important information and ideas easily that support them.
- It assists students to focus on key words and phrases of a given text that is worth noting and remembering.
- Create short report of television, radio, and entertainment programs.

1.7 Methodology

This section discussed the methodology that the researcher applied to accomplish the specified objectives of this thesis work.

1.7.1 Literature review

The researcher reviewed literatures to know automatic text summarization techniques that are based on extractive method and also Afaan Oromo related literatures to know the subject well. Reviewed materials are such as journals, literatures, articles, books and other scholarly published materials for the purpose of understanding the subject area of the study of this language.

1.7.2 Research design

Based on the concept got from reviewed materials and identified research questions, the researcher designed the summarization algorithm. In this work the researcher used an experimental method that evaluates the implementation of Afaan Oromo News Text summarization. The researcher designed the Afaan Oromo news text summarization system to extract sentences based on sentence scoring approach.

1.7.3 Data Gathering

The researcher gathered 30 Afaan Oromo news articles with length of an article is in an interval of 8 and 83 number of sentences. Among these articles the researcher used 10 articles for experimentally test the system. All these important data the researcher used for the experiment were collected from newspaper articles such as Bariisa, Kallacha Oromia and news sites and media organizations such as news agencies like Afaan Oromo BBC, OBN. Bariisaa is a weekly produced newspaper, whereas Kallacha Oromia is produced once with in two weeks.

Transmissions in Afaan Oromo that are available in our country are such as Ethiopian Radio, Fana Radio, OBN, OBS, ETV Afaan Orormo, different spiritual channels and from foreign BBC, OMN and VOA. For evaluating the system, the researcher used news with 8 sentences and greater than this number of sentences [15].

1.7.4 Tools and Techniques

The researcher developed text summarization system called AAOTS to achieve this research work. AAOTS system is developed by using natural language processing packages for python NLTK. "NLTK is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP) which contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning." The method the researcher used is extractive method based on sentence score. The system reads a text and decides sentences importance by calculating the score of each sentence and create a short summary of the main ideas in the text.

The techniques of generating a summary of today's most research still depend on extraction of necessary sentences from the original text to form a summary [11]. Extractive based text summarization selects useful sentences from the original text to form summary [44]. Challenges in extractive text summarization is how to determine the sentences that is significantly to be included in the summary, but sentence scoring works based on features of sentences [10].

So first, it assigns a score to each sentence based on feature, then selects the largest scores that are most likely to be included in final summary [31]. So in this way, most important sentences extracted from the document are displayed to the reader.

1.7.5 Evaluation

Evaluating summaries and automatic text summarization system is not a simple thing to simply test the system or summary which is the result of the system [24]; therefore, to evaluate the summary the researcher used an evaluation mechanism which tests the summary by itself and that mostly focuses on the coherence and in-formativeness of summaries based on human summary is called an intrinsic evaluation.

The evaluation is carried out subjectively by focusing on the structure of the summary measuring referential integrity, non-redundancy, coherence and in-formativeness of summary. Whereas evaluating objectively is carried out by measuring precision, recall and F-measure metrics. The researcher distributed the collected data to evaluators for evaluating the system summaries using different criteria.

1.8 Thesis Outline

The research work is organized in 6 various interconnected chapters. The first chapter starts by discussing about background of text summarization and Afaan Oromo language followed by statement of the problem, objective of the study, scope and limitation, research methodology and significance of the study. Consequently, the second chapter is all about Literature review on automatic text summarization and related works. In this chapter different types of summarizations, various approaches and techniques to automatic text summarization were reviewed. And also reviews of related works in automatic text summarization of different languages for Afaan Oromo Text summarization is reviewed regard to their objectives, statement of problem, limitations, contributions and justification in terms of the need for this work in that domain. Thirdly, it discusses overview of Afaan Oromo and newspaper. Then, the fourth chapter presents design and implementation. And the fifth Chapter presents the empirical results of the proposed system along with their interpretations. Finally, the last Chapter concludes the thesis with the research findings and recommend future works.

Chapter Two

Literature Review

2.1 Overview

In this chapter the main topics discussed are reviewed literatures and related works. Under these two main topics there are sub-topics such as history of text summarization, text summarization and its types, applications, stages, features and evaluation mechanisms. Consequently, the evaluation mechanism intrinsic evaluation which evaluates the generated summary is also discussed. Lastly, the related work which mainly discuss about text summarization researches discussed for the foreign and Ethiopian languages followed by challenges of text summarization.

2.2 History of Automated Text Summarization

The first interest on research work on automatic text summarization was started as early as fifties [19]. The first library catalogues was in 1674 and later generating abstracts for research articles in 1898 were the first applications [53]. The emphasis was on generating summaries that would help to choose the best articles for deeper reading rather than trying to generate summaries that would replace the original text [13]. Following this, the first summarization system was built on the first commercial computer, based on bag of words technique and counting word frequencies [13].

A paper produced in 1958 is an important paper of these days also [19]. According to this paper frequency of words used to suggest weight of sentences of a document. Later in 1969, in addition to frequency depending weights, also used Cue Method, Title Method and Location Method for determining the sentence weights [20]. Moreover in 1995, sentence extracting task which is based on a number of weighting heuristics was performed by Trainable Document Summarizer [54]. To parse sentences into tokens and stem words into their base forms machine learning was introduced between 1990 and 2000 in NLP [13]. In 2001 Conroy and O’Leary constructed a model with feature-based and non-sequential approaches that extract a sentence from a document using Markov model [71].

As Neelima Bhatia and Arunima Jaiswal [72] discussed in their work: using clustering method in 2008 Judith D. Schlesinger made summarization by joining clustering, linguistics and statistics. Nitin Agarwal also discussed based on clustering method concept and query oriented approach with that of unsupervised for summarization in 2011 [72]. Nitin Agarwal, Kiran Gvr, Ravi Shankar Reddy and Carolyn Penstein Ros[1] discussed in their article in 2010 template based generation of summaries possessing similar hierarchical topic structure.

During 2014-2015 in NLP community among some events that seem worth remembering the proposal of seq2seq learning framework based on LSTM and attention mechanism are the ground-breaking networks applied to machine learning [73]. And during 2015-2016 the development of the summarization filed seems to have begun to plateau. 2016-2017: is said the "eve" of the outbreak of abstract field research[2]. Specifically established Training methods, the introduction of CNN/DM datasets, model structures. 2017-2018: the development of summarization field has entered a golden period. During this time around 20 papers have been accepted. 2019 is based on the current acceptance of NAACL and ACL [73].

2.3 Text Summarization

Text summarization is the reduced version of the original text and it contains the essential element of the text [41]. And it creates a summary from article(s) by extracting the most significant fragments of the document and then listing them in a consecutive order [42].

Automatic text summarization is about creating a smaller version of text(s) by machines [41]. Various applications may be used to compress a text [42]. Automatic text summarization shows that the constant developments from 2001 up to 2007 through Document Understanding Conferences (DUC) organized by the National Institute of standards and Technology (NIST) among the first advances made in automatic extractive text summarization, has been considered in Luhn and Edmunson as the pioneers of Automatic Text Summarization [43].

Moreover, a summary is constructing smaller text with half or less than half of an existing document size [10]. In the time being, automation is a procedure that can be applied using expertise without any help from human beings [8].

Text summarization system can be used in situations such as a summarizer in a search engine to give a shortened information of each page to a user, in finding the most important contents

[1] https://www.aclweb.org/anthology/W11-0502.pdf

[2] http://pfliu.com/Historiography/summarization/summ-eng.html

of the text in a short time from large data, and for summarizing the letters and other document in offices [45].

2.4 Types of summaries and their properties

Text summarization used to extract needed information in a required period of time and depend on the nature of the input, purpose and output summary we can classify in to several categories [11]. These categories are such as informative vs indicative, abstractive vs. extractive, query-based vs generic, multi-document vs single document and domain-specific vs. general. Alongside the class of category, each of these classes can be combined to define a summarization task and any summary can be characterized by three major classes of characteristics [16]; input, output and purpose of the summarization.

1. Input: Characteristics of the source text(s)

This is based on the size of the text or number of documents to be summarized [17]. Single document if the summarization is performed for one text document, multi-document if the summarizer is performed for two or more than two text documents. Multi document is more complex and difficult than single document [11][18]. Even if summarizing single document is difficult summarizing multi-document is more difficult [15].

In addition to this, another class of text summarization is based on the domain of the text to be summarized which deals with domain [14]. Domain means the topic of the text whereas domain-specific means the model uses domain-specific knowledge along with the input text and it used in producing a more accurate summary.

A generic summarization is to generate summaries containing main topics of documents. In general, to perform extractive methods we do have three step procedures [23]. First creates a representation of the document. Some preprocessing such as tokenization, stop word removal, noise removal, stemming, sentence splitting, frequency computation etc. is applied here. In the second step sentence scoring are performed. The word scoring is assigning scores to the most important words, Sentence scoring is verifying sentences features such as its position in the document, similarity to the title, etc. and graph scoring is analyzing the relationship between sentences. And the last step is extracting high score sentences by using specific sorting order and generating the final summary if it is the single document summarization.

2. Output: Characteristics of the summary as a text:

There are two ways in which text summarization can be achieved by either extractive text summarization or abstractive text summarization based on the output [14]. In extractive text summarization important and relevant sentences are selected from the original source document and included in summery by concatenating sentences depend on their weight. Whereas abstractive text summarization is a summary that consists of ideas or concepts taken from the original document but are expressed in other words rather than direct words from the documents as extractive summarization. It needs extensive natural language processing and much more complex than extractive summarization [24]. Abstractive summarization produces an abstract summary by generating new phrases or sentences in order to offer a more coherent summary to the user and totally focuses on generating phrases and/or sentences from scratch in order to maintain the key concept alive in summary [14].

3. Purpose: Characteristics of the summary usage

Summarization is divided in to generic and query based, based on the purpose [10]. In query based the user gives to text summarization tool the query, and then it retrieves information related to that query [14]. Whereas generic is general in application and it does not make any assumption regarding the content of the text and treats all the inputs as equal, also it do not target any particular group and does not focus on a special user need [14][22]. Query based summarization mainly focuses on a user information rather than the author document and used to extract personalized summary based on user needs. In query-focused summarization, the relevant of the sentence to user question and sentence importance in the context determine importance of each sentence [22].

Depending on their purpose, the use or task for which they are intended, summaries are classified as indicative and informative [21]. Indicative is to provide an idea of what the document is about to the reader without giving any content [14]. The length of indicative summarization is around 5% of the given text [54]. Informative gives brief information of the main text [54]. And it is if the reader aim to substitute the original text by incorporating all the new or relevant information and that do provide a short version of the content [14]. The purpose of informative summaries is to deliver as much information as possible to the user and to serve as a substitute for the full document. The lengths of informative summaries is around 20% of the complete text [15].

2.5 The Stages of Automatic Text Summarization

The process of automatic text summarization has topic identification, interpretation and generation stages [18] [45]. Topic identification is used to identify the most important topics of the texts or it involves the identification of central or important topics from the document to be summarized [9][18]. Since it can automatically sort a set of documents into classes or categories or topics from a predefined set it is known as topic spotting or topic detection and tracking using techniques like positional importance, cue phrases, and word counting [25]. When the approach for text summarization is extraction, it does the topic identification stage only [26] and, most systems today use this stage only [21].

Interpretation is the analysis of the topics to understand how they are related and is performed only by summarization systems which are based on abstraction that oppose the extraction summary because extraction summary does not modify the original text [21]. At this stage the topics identified in the first stage are joined, represented in new terms or words which are not in the original text and expressed using new concepts to form an abstraction.

Summary generation is after sentence scoring is completed weather in extractive or abstractive summarization, and its purpose is to reconstruct the extracted and joined material into a coherent, densely phrased, new text [15].

2.5.1 Features for extractive text summarization

A number of researchers on text summarization [9] [17] [27] found some features that often increase the application of a sentence for inclusion in summary. These features are: sentence length, sentence position, cue phrases, sentence location, occurrence of proper names, word similarity with title, keyword occurrence, word similarity among paragraphs, word similarity among sentences, numerical data, first sentence, pronoun, weekdays, times, months and term frequency [9].

Sentence length feature is to give the attention for long sentences and provides less weight to short sentences as short sentences are relatively less important than the longer sentences in the text [27]. And this sentence weight is measured by calculating the number of words in the sentence divided by the number of words in the longest sentence of the document [27], then exclude sentences that are very short or very long from the summary [18].

Sentence position: the first sentence in the news is called the “leading sentence”, which usually indicates the main idea of the news and clearly describes the content of the news [70]. Thus sentence position is used to work out the position of a sentence in a document in terms of the normalized percentile score in the range between 0 and 1 [27]. According to Melese Tamiru [21], the significance score of a sentence is calculated based on three features: cosine similarity to the topic vector, the title vector, and sentence position in the document. In this work the researcher used sentence position in the document from these three features to calculate score of sentences in the document.

According to [66], Position of the sentence in the document, decides its importance. Sentences occurring first in the paragraph have highest score [65] and also according to Raja, Naol and Suresh the highest score is given to the first sentence of a document and the lowest score is given to the last sentence [69]. Thus, the score of each sentence in a paragraph can be calculated as follows: if we have a paragraph with n sentences then the score of every sentence is:

$$F_n(S_1) = \frac{n}{n}; \quad F_{n-1}(S_2) = \frac{n-1}{n}, \quad F_{n-2}(S_3) = \frac{n-2}{n}, \quad \text{and so on} \qquad 2.1$$

Cue phrases are summarizers which consider sentences containing cue phrases like “above all”, “actually”, “after that”, “also”, “as a result”, “at that time”, “conversely”, “in conclusion”, “in summary”, “to sum up”, “the point is”, “this letter”, “this report”, “argue”, “purpose”, “develop”, “attempt” etc. Sentences containing cue-phrases are given a higher score compared to other sentences for summary than other sentences in text summarization [28][46]. This feature score is calculated using the following equation.

$$Cue\ Phrase\ Score = \frac{\#\ Cue\ Phrases\ in\ the\ sentences}{\#\ Cue\ Phrases\ in\ the\ document} \qquad 2.2$$

Thematic words are the most frequent words in the given document [19] and words with maximum possible relativeness is determined by the number of thematic words [21]. In other words, sentences containing most frequent words have more chances to be considered as summary sentences since it is probably related to topic [61]. The top n frequent content words are considered as thematic words and the score for this feature is calculated by the following formula:

$$Score(Si) = \frac{No.of\ thematic\ words\ in\ Si}{Length\ (Si)} \qquad 2.3$$

Similarity of title words with all words in the document is another important thing to consider. Titles and headings are considered as short summaries of the texts, and the words that exist in the title or their synonyms are important and have high scores which gives important clues about the subjects of documents [57]. A sentence is with higher score if it contains the words that appear in the title and these sentence contains a words that occur in title may give what the document is intended to express [30]. This can be determined by counting the number of matches between the words in the title and the content words in a sentence [61]. Words in the title or heading of a document that reappear in sentences are directly related to summarization [31] and they are considered for summarization as they have some extra weight.

Hence, text summarization score of a sentence S can be computed by dividing the number of words in sentence that occur in the title over the number of words in the title [61].

$$Score\ (Si) = \frac{No.\ of\ title\ words\ in\ Si}{No.\ of\ title\ words\ in\ title} \qquad 2.4$$

Keyword-occurrence is choosing sentences that has keywords that most often used in the text represent theme of the document [10]. Keywords (Content words) are linguistic terms which refer to words such as nouns, most verbs, adjectives, and adverbs that refer to some object, action, or characteristic [17]. These content words are used as a criteria to select sentences to be included in the summary after they are identified.

A sentence is given a high score when the words in it appear in more number of other sentences in the document [31]. This can be identified by segmenting each sentence into words. The segmented words are searched in the other sentences of the given document to find out occurrence count of the word. Where occurrence count of the word is the number of other sentences in which a given word has occurred. Then Sentence Occurrence Count (SOC) is computed by summing up the occurrence count of all the individual words in the sentence [31]. The score for the feature, word similarity among sentences is computed as the ratio of the sentence occurrence count of the given sentences to the maximum sentence occurrence count in the document.

The numerical data in the document generally brings about some important status of the core idea of the document and thus the sentence with numerical data can reflect the intention of the document and may be selected for the summary. The score of this feature is calculated by dividing the number of numerical data that occur in sentence over the sentence length:

$$NU\_f(s) = \frac{Length\ of\ numerical\ data\ in\ the\ sentence}{Sentence\ length} \qquad 2.5$$

First sentences score given to first sentence of each paragraph, reflecting text order, normalized between 0 and 1. When all first sentences used, remaining sentences are scored in text order.

2.6 Approaches to Text Summarization

Extractive and abstractive approaches are the two techniques proposed by a number of researchers for automatic text summarization [18][19]. The abstraction approach requires fusing and phrasing concepts from original texts which is generally complicated due to the complexity of natural language. And it tries to understand the main ideas in a text and then express it clearly to reader. Whereas extraction summaries are expressed by taking out key text sentences or passages from the original document, by using statistical analysis of individual features or combination of two or more than two features.

2.6.1 Sentence Scoring to Extract

To extract sentence from original document, this method uses features such as term frequency, sentence position, cue phrase, sentence length mechanism features, name of numbers, name of times, days and months, decimal numbers, thematic words, Similarity with the title words (words in a header) [10]. The score of each feature is computed and combined together to give the weight of one sentence. Even though, there is difficulty how to combine these different scores, but there are several different literatures that described several approaches. So the most common point of these approach is that several weights are allocated by coefficients to the individual scores, those are combined together [9]. But due to lack of standard allocating weight, simple combination which is linear combination in which the constraints identified manually by conducting experimentation.

The score is computed by:

$$\text{Sentence score} = \sum_{j}^{n} P_j C_j \qquad 2.6$$

In this equation **C_j**: is the jth constraint coefficient, **P_j**: is the jth constraint, n: is the number of parameter.

2.6.2 Concept Counting by Knowledge-Based

Concept counting knowledge-based is used to identify the main concepts in the document, based on concept counting knowledge-based paradigm [9]. According to Lin to generalize the concepts, we use concept generalization taxonomy (WordNet) and its hierarchy helps to find the generalization of relating words with in a sentence to their category depend on type [10]. For the main topic for sentence “Tola bought shirt, jacket, and trouser.” These items may be cloths. But it is impossible to put any conclusion about the title of items by word counting and any other summarization approach since it needs deeper level of semantics and this can be solved by hierarchy [10]. Hierarchy describes a system that organizes or ranks things, often according to its importance. In using hierarchy, finding appropriate way how to generalize in taxonomy hierarchy is one big question. To answer this question let us discus ratio.

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

$$R = \frac{\boldsymbol{Max}(W)}{\boldsymbol{Sum}(W)} \qquad 2.7$$

In this equation w is the weight of children of a concept.

Parent concept weight is defined by the frequency of occurrence of a concept and its sub concepts in a document.

The branch ratio threshold (Rt) is defined and serves as a cut off point for the interestingness to determine the degree of generalization [10]. So the comparison of these concept’s ratio R and Rt is used for generalization. That is if R < Rt, it is an interesting concept. For instance in the following figure if Rt = 0.3, then Rt is less than R, so Spain is the main topic.

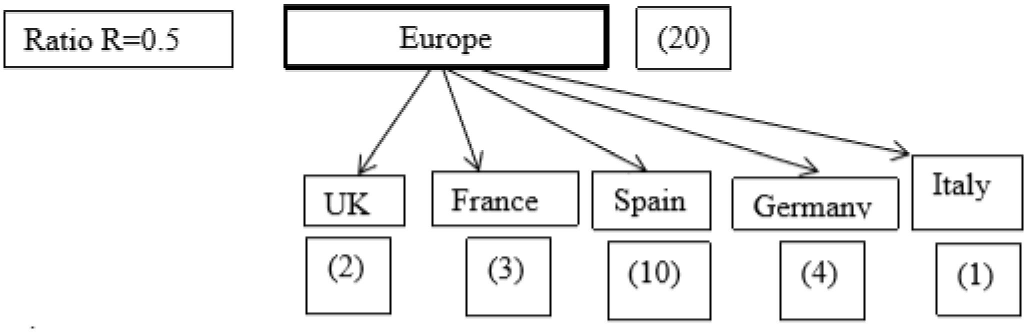


Figure 2.1 A sample hierarchy for world countries

2.6.3 Surface Level Approach

As Guesh and Melese stated in their paper most of the early works of automatic text summarization follow the surface level approaches for deciding which parts of the text are important [15][21]. This approach inclines to use surface features which then selectively mixed together to produce a salience function used for the extraction of the important information. For instance thematic feature is presence of statistically salient terms and based on term frequencies statistics, location is position in text, position in paragraph, section depth, and particular sections, occurrence of title or headings terms in the background, cue words and phrases are words or phrases that signal whether a sentence is important or not.

The oldest known automatic text summarization is that of Luhn [21][29]. According to Luhn sentence relevance is measured by term frequencies. That is the highest frequent words in a document mostly represent document content, so sentences containing them are more relevant. However, words in a document are not equally important, those words with less frequency and beyond high and words in a stop word list are out of the consideration, but the remaining words in text are arranged alphabetically, then similar words to be found by comparing pairs of words letter by letter (e.g., consider, considered, consideration).

The frequency of words is directly proportional to the relevance of that word to the whole document [26], and for text summarization after pre-processing is counted the frequency of words and putting the words in descending order and sentences will be given score based on the occurrence of those relevant words. Then highest scoring sentences included in the summary.

2.6.4 Lexical Chain Methods

A lexical chain is a groups or sequences of semantically related words or terms [47]. It occurs between pairs of words and over sequences of related words. Strong chains can be created using lexical relations which is determined from lexical database and it help to identify the important sentences [22]. And relations between words are stored in WorldNet database and words that appear as nouns in the WorldNet entry are chosen for text summarization. One of the efficient methods to compute the lexical cohesion is to use the lexical chain. There are three steps to be followed to construct lexical chains [9]. The first one is by making a list select a set of candidate words, then the second one is find a suitable chain for each candidate word and thirdly when the second condition fulfilled, then insert the applicant word and create a new chain or update it accordingly.

The strongest chains among those produced by the algorithm has to be identified in order to obtain the summary for a given input text.

2.6.5 Latent Semantic Analysis (LSA)

Latent Semantic Analysis is a statistical technique for extracting and representing words from large corpus of text by contextual-usage meaning [33]. LSA is used to find the correlation between the topics and sentences, terms and sentences of a given set of documents to find the effective summary of the document and dimension reduction also [33]. LSA uses information of word usage in context to extract words by meaning and similarity of sentences [10].

LSA is a statistical model that compares semantic similarity between fragments of textual information for word usage [21], and used for improving the efficiency for methods of information retrieval to analyze the relationships between the pair documents and their terms. And also the problem of synonym, in which a different word or term can be used to explain the same semantic concept, can be solved by latent semantic analysis and it tries to overcome the problem of literal term mismatch by allowing retrieval to be based on concepts rather than on terms.

There are three main steps in Latent Semantic Analysis [48]. The first is creation of input matrix: the input document is represented in a matrix form to perform the calculations and a matrix is created which represents the input text. In this the raw of matrix represents words in a sentence and column of the matrix represents sentences of the document. The cells of matrix is the intersection of the columns and rows which represents the importance of words in sentences.

The next is singular value decomposition which is an algebraic matrix factorization technique for matrix reduction. This step of latent semantic analysis decomposes a rectangular, giant, and sparse matrix and produces a smaller matrix with a small rank [34]. This step is a statistical model that shows relationship between words and sentences. Different algorithms use different approaches to select important sentences from the document for summarization using the results of singular value decomposition [35].

The last step is sentence selection process which is usually based on linguistic, mathematical and statistical techniques. This step performed to generate output summary using the result of SVD as different algorithms use different approaches to select important sentences from the document for summarization using this result [35].

2.6.6 Machine Learning Techniques

Machine learning was described in NLP from 1990 up to 2000 to parse sentences into tokens and stemming words into their base form [12]. And as Teferi stated in his paper machine learning is the ability of a machine to identify the repeatedly occurred pattern and improve itself from past practice by concentrating on induction algorithm and another algorithms which can be said to learn. Its techniques involves statistical learning that the probability of sentences to recognize as sentences can be included in a summary or not, by training document and their extractive summary using machine learning rule [36]. This approach used by training a program to identify summary sentences using an existing text or summary corpus. The training corpus is made by different techniques such as join two sentences from two different articles, for those abstract sentences exact matching of sentences from different articles, paragraph features related to sentence position and summary of sentences using word frequency [26].

2.6.7 Graph-based Approaches

Graph is a group of points that are connected by a line whereas dots are called nodes (or vertices) and the lines are called edges [17]. According to Melese Tamiru graph-based ranking algorithms are used in text summarization based on the assumption sentences that have more relationships with other sentences in a given text are more important because they can directly relate to more other sentences [21]. The similarity of sentences is measured by this approach is by using surface level statistical method to determine salient sentences using word overlap.

To construct a graph, vertex found in each sentences in the document are added together and edges between vertices are created by connecting sentence together which is defined as similarity relation. But similarity relation is content overlap between two sentences which can be determined by the number of common tokens between lexical representations of sentences [36].

2.7 Evaluations

Evaluation is crucial to assess the achievements of different approaches in a research, the key indicator of the applicability of the results and the way of measuring the performance of one thing [17][21]. According to Martin Hassel [37] summary evaluation is not straight forward. But while evaluating summary and summarization system there are at least two properties to be measured. These properties are:

Compression Ratio (CR): is used to determine how much (in what persentage) shorter the summary is than the original document.

$$CR = \frac{\text{length of Summary}}{\text{length of full text}} \qquad 2.8$$

Retention Ratio (RR): is used to determine how much information of the summary is contained the idea of the original document.

$$RR = \frac{\text{information in Summary}}{\text{information in full text}} \qquad 2.9$$

The first system automatic summary evaluation was started in the 1960s [38]. This evaluation of text summarization is used to determine how well the summary can be used to check as the summary achieved the goal of summarizing that document or is it helpful relate to the original document. Even though there are different approaches different researchers use to evaluate text summarization since there is no standard method of evaluation, generally it is classified in to intrinsic and extrinsic [26] [22] [15].

A. Intrinsic evaluation

In this evaluation technique judgement can be given by person on the generated summary conserning the quality of the summarization. And it is also known as normative evaluation, because it involves assessing the quality of a summary by human judges by comparing it to a manual summary, based on text quality. Where text quality refers to some aspect of the text such as grammatically, non-redundancy, reference clarity and coherence [15] [22]. Even if there are different intrinsic evaluation methods, this thesis mainly assess the coherence and in-formativeness of summaries:

Summary Coherence focus on, grammaticality, overall text coherence, and organization. Coherence is understood as connectivity of the summary text [38]. This method is based on extraction that use cut and paste operations on phrases, sentences or paragraph. That might result in a summary part extracted out of context which results in coherence problem which can be measured by having humans rank summary sentences for coherence and compare the ranks with the scores for reference summaries or source sentences.

Summary In-formativeness: This is to measure the content of summary as all information in the original document is contained in the summary part [9]. Also according to Addis Ashagre [22], it is by comparing the system generated summary with the input text being summarized analyzing how much of information from the input is represented in the summary, informativeness of the summary can be measured.

From the different methods of intrinsic evaluation co-selection measure is one another method. This method is used to evaluate summaries using the metrics of precision, recall and F-score. The computation of these metrics is based on the number of sentences that can be found from the system summary and the ideal summary.

Precision (P) is by dividing the number of sentences find in both system summary and manual summaries to the number of sentences in the system summary.

$$P = \frac{\text{correct}}{\text{correct} + \text{wrong}} \qquad 2.10$$

Where:

Recall (R) is dividing the number of sentences find in both system summary and manual summaries to the number of sentences in the ideal summary.

$$R\ (Recall) = \frac{\text{correct}}{\text{correct} + \text{missed}} \qquad 2.11$$

F-score is used as input precision and recall to compute f-score measure of summary. These measures are compute using the following equations

$$F\ (F - score) = \frac{2*\mathrm{P}*\mathrm{R}}{\mathrm{P}+\mathrm{R}} \qquad 2.12$$

Correct is the number of sentences find in both the reference summary and system summary. Wrong is the number of sentences not in the reference summary but in the sysstem summary. Missed is the number of sentences not in the summarizer’s summary but in the reference summary.

Advantage of using co-selection measure is that once human judge defined the ideal summary, it can be used to evaluate automatic summaries with a simple comparison. And its disadvantage is in terms of defining gold-standard summary. Recall measure of a summary may range from 25 % up to 50 % based on which of two accessible human extracts are used for evaluation. Therefore, using co-selection measure creates the possibility to have that two equally good extracts are judged very differently [22].

The other approach of intrinsic approach is using language experts to judge whether the summary is good or not; and it will assess the produced summary and decide on it. But this approach of evaluation is expensive since it needs human judges and the subjectivity of judging with human being may be biased. The problem of intrinsic approach evaluation is the difficulty of constructing the ideal summary for a given text. Even if we can construct the ideal summary, since it is constructed by human beings, different human beings will have different ideal summaries. So, selecting the ideal summary is a difficult task.

2.8 Related Works

2.8.1 Text Summarization on foreign language

As [46], conducted research on automatic text summarization for Punjabi language focusing on single document multi-news summarizer using extractive method. In this paper they applied preprocessing and processing with statistical features such as keywords identification, numbered data and sentence length.

They developed Punjabi summarizer lexical resources for Punjabi language by applying an algorithm which has seventeen steps and developed a system.

Then the system has been tested by over fifty Punjabi multi news documents (Data set containing 6185 sentences and 72689 words) from Punjabi news corpus. Then they applied four intrinsic and two extrinsic measures of summary evaluation. The intrinsic measures are F-Score, Cosine Similarity, Jaccard Coefficient and Euclidean distance and the extrinsic measures are Question Answering Task and Keywords Association Task for Punjabi multi news documents at different compression ratios. The results of intrinsic evaluation of Punjabi news documents at compression ratio 10% F-score is 97.87%, Cosine similarity is 0.98, Jaccard coefficient is 0.97 and Euclidean distance is 0.12, at 30% F-score is 95.32%, Cosine similarity is 0.96, Jaccard coefficient is 0.95 and Euclidean distance 0.32, and at 50% F-score is 94.63%, Cosine similarity is 0.95, Jaccard coefficient is 0.94 and Euclidean distance is 0.36. And the result of extrinsic evaluation accuracy of question answering task at 10%, 30% and 50% is 78.95%, 81.38% and 88.75% respectively and that of accuracy of keywords association task at 10%, 30% and 50% is 80.13%, 92.37% and 96.32%.

Kutlu [74], conducted a research for Turkish, applying extractive method using sentence scoring approach in order to extract sentences from original document which form a summary with the main content using surface-level document features such as term frequency, key phrase, centrality, title similarity and sentence position. He combined these features by a scoring function in which each feature has a different weight. He used two datasets, the first one is collection of 120 newspaper articles. And the second data set is a collection of 100 Turkish articles. His aim was to introduce the usage of key phrase as a surface level feature in text summarization and to analyze the effectiveness of the features in Turkish text summarization.

He evaluated his system and one of the evaluations is done to determine the effectiveness of each feature when it is used alone in text summarization and the other one is to measure the overall effectiveness of the system when all features are used in the summarization process. And he put intrinsic evaluation results of each feature tested individually for the newspaper data set for F-score of each features as follows: the performance of Term Frequency (TF) is 0.292, the performance of Title Similarity (TS) is 0.286, and performance for Key Phrase (KP) is 0.248, for Sentence Position (SP) performance is 0.314 and that of centrality (C) is 0.214. And he obtained 0.561 and 0.368 ROUGE-L scores for the newspaper and journal.

2.8.2 Amharic Text Summarization

2.8.2.1 Automatic Summarization for Amharic Text Using Open Text Summarizer

Addis conducted a research to investigate the applicability of the open text summarizer for single document Amharic news text summarization using methods such as Literature Review, Amharic Language Lexicon Gathering, Corpus Preparation and Customization of the OTS by using extractive approach. And tools he used are open text summarizer written in C#, visual studio 2016 is installed, for the table and graph preparations MS-excel and for the report writing MS-word were used. Using statistical approach Addis developed Amharic summarizer customizing OTS. He used features as frequency of terms to determine the importance of the sentence. In the first experiment he adapted the porter stemmer to fit his purpose and in second experiment he adopted a stemmer from the work of Tesema, 2007. He used 30 news texts which were collected from different sources to test the performance of his system. And he used 90 manual summaries prepared with two evaluators. And he applied three extraction rates 16%, 20%, and 30%. F-measure score for the three experiments (1^{st}, 2^{nd} and 3^{rd}) at 30% are 75.53%, 72.83% and 72.37% respectively. The performance evaluation of his work showed the second experiment (E2) outperformed the first experiment (E1) but, in terms of efficiency E1 which used Porter stemmer outperformed E2. And he recommended for future work text corpus and ideal summaries should be prepared, and a complete list of Amharic lexicons should be prepared and made available [22].

2.8.2.2 Automatic Amharic Text Summarization Using Latent Semantic Analysis

Melese did his thesis to investigate the application of Latent Semantic Analysis for automatic summarization of Amharic news texts. The methods he used to investigate is Literature Review, Data Corpus and tools such as Java (JAMA, a free Java library package) by applying sentence extraction approach. The system Architecture he used is Pre-processing, Semantic model analysis and Sentence ranking. He used dataset to evaluate summarization system that contains 50 Amharic news with 17 up to 44 sentence length. And he evaluated performance of the summarizer system using the most common co-selection measures recall (R), precision (P), and F-Score (F). He proposed two methods of sentence ranking which are refered as TopicLSA and LSAGraph. And performance evaluation of the summarizer conducted using these two approaches separately. And this thesis tried to fulfill gap identified how TopicLSA and LSAGraph methods can be used for summarizing Amharic news text [21]. And as the result of his evaluation shown the summarization system he proposed performed significantly better than previous one based on LSA and graph-based ranking algorithms.

2.8.2.3 Automatic Text Summarizer for Tigrinya Language

Guesh [15] conducted research to explore and design an automatic text summarizer for Tigrinya language that process texts to extract the most important information from a source (or sources) to produce an abridged version for a particular users using extraction method, Literature review and Data Corpus. And also Python programming language is used to build the summarization system as a tool. He used dataset of 30 Tigrigna News articles whose lengths are 16-42 sentences for evaluating the summarizer. His system is purely extractive type of summary; which means the process of selecting an important sentence based on the frequency of individual words and title words. The **Term Frequency:** is text representation techniques based on the extraction of terms of a text or documents which consist in choosing terms that are frequently occurring and then selecting sentences contains these terms to make the summary. **Scoring**: this module used to compute the score of each term and the sentence containing that term. So depend on the score of the term in each sentences it give score for each sentences in the document. **Sentence Ranking**: sentences that contain the most frequent words ranked first of the whole document. **Sentence generation:** The summarizer extracts sentences ranked based on the most frequent words which are intersection of the sentences.

His experimental result shows that the system registered for recall, precision and F-score 46%, 46% and 46% respectively for term frequency and 46%, 50% and 48% respectively for the feature title words.

How ever this work has some research gaps such as lack of scientific justification during weighting of sentence based on the two features stated in his work, lack of scoring adjustment mechanism, lack of cue phrases, thematic words, numeric datas, capital letters and number handling mechanism.

2.8.3 Afaan Oromo Text Summarization

2.8.3.1 Afaan Oromo news text summarizer

Girma [11] conducted a research on “Afan Oromo news text summarizer” in 2012 to explore appropriate statistical approaches for developing and implementing an automatic news text summarizer for Afaan Oromo. He generated summary by extract method, using features term frequency and sentence position together. To achieve this he used methods such as Corpus Preparation, Summary Generation, Experimentation methods and Literature review. He build the system from three basic subsystems with one XML based lexicon.

Preprocessing: this step is tokenizing, removal of stop-word, stemming and parsing. This is splitting input text into sentences. To remove stop word, he used the stop-word collected from different Afaan Oromo literatures and the stop-word list prepared by Debela. And stemmed the words.

Sentence Ranking: Sentences are ranked by their term frequency after the term frequency is computed. Term frequency is frequency of keyword occurrence in an article. Girma computed positional value of a sentence s, if the first sentence of a document allocated with the highest score and the last sentence is allocated with the lowest score. .

Summary Generation: produced by picking n-top ranked sentences, where n is number of sentences user need to pick. So value of n is given by the user.

The researcher used a corpus of 8 news articles averagely 11 sentences and 277 words. The researcher evaluated performance of system using three methods M1, M2 and M3. Method M1 used term frequency and position of sentence without stemmer & lexicon, Method M2 used stemmer and lexicon and the third method M3 improved all term frequency, position method, stemmer and lexicon. The performance of the first methods, the second methods and the third methods scored f-measure values of 34%, 47% and 81% respectively. That is the third method outperformed the two remaining methods. The subjective evaluation result showed that the three summarizers' M1, M2 and M3 performances regarding in-formativeness is (34.37 %, 37%, and 62.5%), linguistic quality is (59.37%, 60% and 53%) and coherence structure is (21.87%, 28.12% and 75%) respectively as it is compared the developed system with human evaluators.

2.8.3.2 Afaan Oromo Automatic News Text Summarizer Based on Sentence Selection Function

In his thesis Fiseha [10] developed a summarizer for Afaan Oromo based on Sentence Selection Function using extractive method to improve the work of Girma [11]. In his work he tried to justify how to give weight to sentences, scoring or weighting adjustment mechanism, computation of summary compression ratio, handling mechanism of different features. He used the methodologies like literature review, data gathering and experimental.

The researcher collected for validation and testing 33 different newspaper topics, of these, he used 20 of them for validation while the rest 13 for testing purpose. He developed the summarizer by using the architecture that has three parts. These are preprocessing, processing and Summarizer. In each of these parts different algorithms and how they work has been discussed. He computed sentence weighting based on sentence position, cue phrase, key word, event, number.

His work is to fulfill the research gaps identified like 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.

And also he has discussed how Sentence Compression module and summary sentence generator module designed and developed. Finally as a solution he was proposed improvement of the previous work using the same corpus and similar evaluation method his summarizer outperformed by 26.96% Fm than the previous work on this title, that is his system performs 87.47% [10]. How ever his work has some research gaps such as how to use title words, thematic words, capital letters and numeric data to summarize text documents.

2.8.3.3 Critics

After reviewing the work, the researcher identified the following drawbacks/gaps:

Always when a reporter write newspaper report conserning measurement or value of any thing he/she has to use number or digits. For example; 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.

The title is the first, and sometimes only, part of your article that potential readers will see, so it's important to grab their attention and entice them to read your article. An effective title, then, is key to getting your article noticed and read by identifying the first step toward making sure your work has an impact. It’s also important to accurately portray your work so readers aren’t misled by a catchy but not-quite-accurate title. So a sentence containing title words describe the title more. But all the reviewed articles not used title words.

A theme captures a common, recurring pattern across a dataset, clustered around a central organising concept. A theme tends to describe the different facets of that singular idea, demonstrating the theme’s patterning in the dataset. A central organising concept captures the essence of a theme. It is an idea or concept that captures and summarises the core point of a coherent and meaningful pattern in the data. If you can identify the central organising concept of a theme, you can capture the core of what your theme is about. If you cannot do this, your theme may lack coherence. So in all the reviewed articles they did not include the thematic words.

2.9 Challenges in Text Summarization

Challenges of text summarization discussed in different papers in different form are:

- Lack of ability to select the important features of a document summarization system which helps to extracts the main concept from source.
- Difficulty of how to identify ambiguous sentences in the original documents.
- Lack standard evaluation mechanism of text summarization system summary generated.

Chapter Three

Afaan Oromo Language

3.1 Introduction

This Chapter discuses an overview of Afaan Oromo, Afaan Oromo alphabets and writing system, punctuation marks in Afaan Oromo, Afaan Oromo Morphology, Word and Sentence Boundaries.

3.2 Overview of Afaan Oromo

Afro-Asiatic family is the largest family in Ethiopia and Cushitic is sub-family of this largest family in Ethiopia, and Afaan Oromo is sub-family of the Lowland East Cushitic family having the largest number of speakers among the Cushitic language family [55]. The language is the major African languages that is widely spoken and used in most parts of Ethiopia and also some parts of other neighbor countries like Kenya, Somalia and Egypt [7][29][5]. The Oromo peoples who speak this language are the largest ethnic group in Ethiopia, whose population around 34.5% [7] [11][29]. As a second language speakers, a number of members of other ethnicities who are in contact with the Oromos speak, for example, the Nilo-Saharan-speaking Kwama and Omotic speaking Bambassi in north western Oromia [29].

The Cushitic Language families Oromo is found in different parts of Ethiopia except the northern all areas retaining its homogeneity [55]. These geographical location are such as Southern, Eastern, Central and Western. Dialects of all these areas is almost similar while communication. The five major dialects of Afaan Oromo has are: Boorana (Southern), Tuulama (Central), Harar (Eastern), and Mecha (Western), Rayya (Northern) [55]. For instance, the Mecha dialect predominantly uses **koo** to mean ‘my, mine’, whereas other dialects use **kiyya**.

Nowadays, This language is an official language of the largest Regional State among the current federal states in Ethiopia called Oromia state [29][11] [10][56] and also Oromia Zone of the Amhara Region called Kemise [55]. Furthermore, beginning from primary school up to higher education it has been used as medium of instruction [10]. In addition to this Afaan Oromo language is now a language of research, administration, political and social interaction [56]. And also, literature works, newspapers, news, online education, educational resources, magazines, journals, books, videos, pictures, official credentials, religious documents, entertainment medias are increasingly published and available in this language [10][29][56].

3.3 Writing System

Afaan Oromo is a phonetic language, which means it is spoken in the way it is written [11]. Afaan Oromo writing system, Latin-based alphabet “Qubee” has been accepted and become the official script of Afaan Oromo since 1991 [29]. The writing system of this language is modified from Latin writing system and the language shares a lot of features of Latin writing system with some modification [1][11].

In Afaan Oromo language all letters in English language are also there, but the way they are written is different and there are no skipped or unpronounced sounds/alphabets unlike English or other Latin based languages [1]. In Afaan Oromo language, the sounds are more stressed when consonant is doubled in a word and the sounds are stretched or lengthened when the vowels are doubled [11].

In this language writing system there are 33 characters [29]; among this 33 characters 26 are the same to English letters whereas 7 letters are known as “Qubee Dachaa” which are formed by joining two consonant letters [56]. These are CH ch, DH dh, SH sh, NY ny, TS ts, PH ph and ZY zy. Qubee Afaan Orormo has five vowels like as in English with the same letters. These are a, e, i, o, and u. Vowels have two natures known as short and long in the language and they can result in different meaning.

P, V and Z are not found in the basic alphabet of Afaan Oromo [57] since there are no native words that are formed from these characters in Afaan Oromo. But, in this language, they are used to indicate borrowed words from another languages. For instance televiijinii to mean “Television”, poolisii to mean “Police”, Zeekkara to mean “Opera”. Afaan Oromo alphabets[3] are listed as follows with both upper case and lower case.

A	B	C	CH	D	DH	E	F	G	H	I	J	K
a	b	c	ch	d	dh	e	f	g	h	i	j	k
L	M	N	NY	O	P	PH	Q	R	S	SH	T TS	U
L	m	n	ny	o	p	ph	q	r	s	sh	t ts	u
V	W	X	Y	Z	ZY							
v	w	x	y	z	zy							

Figure 3.1 Afaan Oromo Alphabet/Qubee Afaan Oromoo

[3] Afaan Oromo alphabets

3.4 Punctuation Marks in Afaan Oromo

Afaan Oromo punctuation marks are used in text punctuation pattern as in other languages that follow latin writing system to make meaning clear and reading easier [1][11]. These punctuations are discussed as follows [62]. The punctuation mark that is used in abbreviations and at end of the sentence is called tuqaa (Full stop) (.). And the punctuation mark that is used at the end of question sentences or interrogative is called mallattoo gaaffii (question mark) (?). A punctuation that is used at the end of the exclamatory or command sentences is called raajeffannoo (exclamation mark) (!). The other punctuation mark used separate separated elements is called qoodduu (comma). Both languages Afaan Oromo and English use these punctuation marks are used in the same way except apostrophe is used in English to show possession. Whereas in Afaan Oromo this punctuation mark is used to represent a glitch ("hudhaa") sound and considered as part of a word [62]. For example, ka'e "stand up", bu'e "fall down" used to when two vowels appeared together.

3.5 Afaan Oromo Morphology

Studying the meaning of individual units or language morphemes, word formation and internal structure of words is called morphology[4] [10][29]. Words are the basic building blocks of a language whereas, morphemes are the basic building blocks in morphology [29]. Words are a freestanding unit of meaning whereas, morphemes the minimal unit of grammatical analysis and it may or may not stand alone [11].

Morphology is divided into inflectional and derivational branches [10][56]. Inflectional morphology is the study of processes, is the process of adding some meaning to the existing words rather than not as the creation of new words [56]. In inflectional morphology, a different form of the same word are added in words [10].

Derivational morphology is the process of creating a new word from a root word by adding a bound morpheme to a stem and also changing classes of words [7][29]. Derivational morphology changes the lexical meaning of the stem word from the other derived words by changing the class of the word. Suffixes such as -achuu, eenyaa, -ina, umsa and –ummaa are used for word formation in derivational morphology in tis language [63].

[4] https://www.merriam-webster.com/dictionary/morphology

Word can be formed in different ways in Afaan Oromo [7][10][29]. These different ways can be as the following categories are: nouns, verbs, adjectives, adverbs, prepositions, and conjunctions.

3.5.1 Nouns

This categories of in a document have person, gender, number and possession markers that are concatenated and affixed to a stem or singular noun form [10], and also true for pronouns and determinants have number, gender, adjectives, and quantifier markers in this language [55][56].

i. Gender

In Afaan Oromo there are two categories of gender [10][57], masculine and feminine like in another Afro Asiatic language. In Afaan Oromo some nouns use –essa, -aa, -isa for masculine and –ttii, -ttuu, -tee for and feminine [7][56].

For instance

Masculine	**Meaning**	**Femenine**	**Meaning**
Korbeessa	Male gaot	Goromtii	Female goat
Obboleessa	Brother	Obboleettii	Sister.
Barataa	Male Student	Barattuu	Female Student

In other way names of astronomical bodies and geographical places like cities and countries such as moon ‘ji’a’, sun ‘aduu’, star ‘urjii’ are feminine [62]. For example ‘Jiini baate’ (to mean the moon rises) and ‘Biyyi Itiyoophiyaa bara kam hundooftee?’ (To mean when Ethiopia country was established?), the suffix ‘tee’ indicates femenine gender in both sentences. Also ‘isa’ and ‘ishee’ in Afaan Oromo shows masculine and feminine respectively like that of English language third person singular pronouns he and she [56].

ii. Number

A noun is changed to plural form by using suffixes such –lee, -wwan, -een, -aan, -olii, -olee b [11][62], but use of suffixes different in different dialects [7][10].

Singular noun	**to mean**	**plural noun**	**to mean**
Barataa	Student	Baratt**oota**	Students
Barsiisaa	Teacher	Baresiis**ota**	Teachers
Farda	Horse	Fard**een**	Horses

Gaaffii	Question	Gaaffii**wwan**	Questions
Saawwa	Cow	Saawwa**n**	Cows
Gaangee	Mule	Gaang**olii**	Mules

Like the other world languages, Afaan Oromo uses the Arabic number as in any other languages, numbers can occur in this language too in the form of cardinal, ordinal and nominal numbers [10].

1. Cardinal number is a number that says how many of something there are, and also answers the question "How Many?" See table 3.1 below.

Table 3.1 Cardinal Numbers

Afaan Oromo	**English**
Zeroo	Zero
tokko	one
kudhan	ten
diddama	twenty

2. In Afaan Oromo an Ordinal Number is used when we use a list to tell position of something and they are formed from cardinal numbers by suffix 'ffaa' or '-affaa' [10]. See table 3.2 below.

Table 3.2 Ordinal numbrs

Afaan Oromo	**English**
1ffaa /tokkoffaa	1st /first
2ffaa/lammaffaa	2nd /second
3ffaa/ sadaffaa	3rd /third

3. A nominal number in Afaan Oromo is used to identify something and uses decimal digits 0 up to 9 to represent nominal numbers [29]. For example
    - a postal code ("1247")
    - a model number ("548")

In Afaan Oromo numbers can also be represented in decimal number [58]. Decimal number is a number whose whole number part and the fractional part is separated by a decimal point[5], and it has two parts. Whole number part and decimal part. The digits lying to the left of the decimal point form the whole number part. The places begin with ones, then tens, then hundreds, then thousands and so on. The decimal point together with the digits lying on the right of decimal point form the decimal part. The places begin with tenths, then hundredths, then thousandths

[5] https://www.splashlearn.com/math-vocabulary/decimals/decimal

and so on. For example: The decimal number 211.35 can be read as dhibba lamaa fi kudha tokko tuqaa sadii shan to men ‘two hundred eleven point three five‘; the whole number part is 211 and the decimal part is .35**.**

Table 3.3 Decimal numbers

English	**Afaan Oromo**
Zero point one	Zeeroo tuqaa tokko
Zero point two	Zeeroo tuqaa lama
One point one	Tokko tuqaa tokko
Two point three	Lama tuqaa sadii

Also in Afaan Oromo numbers can be represented by percentages and percentages are often used to describe changes in quantities or prices [59]. For example: '30% extra free' (dhibbantaa soddoma kaffaltii dabalataa ykn harka soddoma kaffaltii dabalataa), '10% discount' (dhibbantaa kudhan hirdhisa ykn harka kudhan hirdhisa) and 'add 15% VAT' (dhibbantaa kudha shan VAT dabali ykn harka kudha shan VAT dabali).

iii. Definiteness

Definiteness is used to distinguish noun phrases rendering to weather their reference in a given context is continued to uniquely identifiable [11].

There is no indefinite article such as a, an, and some as they exist in English in Afaan Oromo [57]. The definiteness article ‘the’ in English is with suffixes on the noun **-icha** and **-ittii** for masculine nouns and for feminine nouns respectively in Afaan Oromo [55]. In the language of Afaan Oromo the definite suffixes usage is less than that of the definite article the in English [56] and they seem not to co-occur with the plural suffixes [55].

iv. Case

This is to handle different cases by changing the whole documents case to similar case. This cases can be such as upper case or lower cases can be UPPER CASE, or lower case or Mixed Cases [57]. Or it can be grammatical category case to show relationship between noun and verb in sentences.

Case is the process of handling problem related with variation cases that is UPPER CASE, or lower case or Mixed Cases by converting the whole document in to similar case [57]. Or in other way case shows the nature of nouns verb relationship whose grammatical category is nouns [29]. Nouns in Afaan Oromo varied for nominative, ablative, instrumental and locative cases because number of cases varies from language to language.

Nominative Case is used in Afaan Oromo for nouns that are the subject of clauses and in this language ‘**-n**’ and ‘**-i**’ are nominative case indicators [55]. For instance ‘Duulaan buddeena nyaate’ (to mean Dula ate enjera). In this ‘Duulla’ shows the name of the man and ‘Duulaan’ is used as nominative and subject of the phrase ‘buddeena nyaate’. Nouns ending in short vowels with a proceeding single consonant drop a final vowel and add ‘ni’ to form nominative [56]. For instance: ‘Sangoota’ (Oxes), ‘Sangoonni’ (nominative). In this example the word ‘Snagootaa' shows the plural noun and the word ‘Sangoonni’ indicate nominative.

Instrumental Case in a language is the noun used to indicate the means by which something is occurred, the agent which makes something to be occurred, the reason why something is happened or the time of an event when it is occurred [56]. In Afaan Oromo suffixes like –n,-aan,-tiin,-iin,-dhaan,-aan are attached to the nouns based on the number of vowels and consonants the nouns are ending with to represent instrumental cases [29].

Table 3.4 some of the rules that enable to represent instrumental case

Suffix	Explanation	Examples
-n	Added to nouns end with short vowel or long vowel	Ija ‘eye’ ijaan ‘by eye’
-iin	Added to nouns end with consonants.	Halkan ‘night’,halkaniin ‘at night’ ‘Halkaniin dhufe’ (to mean he came in the night)
Tiin	Added to nouns end with long vowel or short vowel.	‘Afaan Engiliffaa’ (English language),’Afaan Engiliffaatiin’(in English language)
dhaan	Added to the nouns end with long vowel	Yeroo ‘time’,yeroodhaan,’on time’ ‘Inni yeroondhaan xumure’ (to mean He has finished on time)

Locative Case in Afaan Oromo is nouns that represent general locations of events or states and used to indicate more specific locations, prepositions or postposition (like booda, hanga, gara, etc) are used in this language [56]. The suffix ‘tti’ is added to the noun to form locative. For instance Jimmatti (in Jimma), Erga geessee booda bilbili (call after you arrived).

Ablative in Afaan Oromo is used to indicate the source of an event [29]. In this language the post position 'irraa' (to mean from in English) can be used as an alternative to the ablative by dropping its initial vowel letter. To form this ablative we follow the following rules:

Tbale 3.5 rules and examples for ablatives.

Rules	Examples
Vowel is lengthened when the words end in a short vowel	Biyya 'country' Biyyaa 'from country'
Suffix 'dhaa' is added when the words end in a long vowel	Mana barumsaa 'school' mana barumsaa**dhaa** Mana barumsaa**dhaa** dhufaa jira 'to mean he is coming from school'
Suffix 'ii' is added to the word when it ends in a consonant letter	Hararii 'to mean from Harar'

3.5.2 Verbs

Verbs are words that tell us action to be taken [29][55][62]. For example taa'i means sit, kottu means come on, etc. Stem and suffixes are used as follows in verb. Stem is used to represent the lexical meaning of the verb whereas suffix used to represent subject agreement in Afaan Oromo [56].

Depending on their stem endings verbs in Afaan Oromo are categorized in to four [56]. Regular Verbs[6]: are formed by attaching suffix to the stem word without changing the stem word and verbs with stem end with single consonant such as ch, a vowel, y, w or z [29]. For example for stem word 'dhug' we can derive other verbs like 'dhuge, dhugan, dhuguufi, dhuguu, dhugi, dhuguudhaafi, dhugnu, dhuganii, dhugna etc without changing the stem word dhug by simply attaching the affixes like '-e, -an, -uufi, -uu, -i, -uudhaafi, -nu, -anii, -na' etc. See the following example in table 3.6 below.

Table 3.6 Examples of regular verbs

	Deemuu – 'to go'		Dhuguu-'to dink'	
First person singular	ani /I	deem**a** 'I go'	ani /I	Dhuga 'I drink'
	ati/ you	deem**ta** 'you go'	ati/ you	Dhugda 'you drink'
third person singular	Inni/he	deem**a** 'he go'	Inni/he	Dhuga 'he drink'
	Isheen/she	deem**ti** 'she go'	Isheen/she	Dhugdi she drink'

[6] https://en.wikibooks.org/wiki/Afaan_Oromo/Chapter_04

First person plural	nuti, nuyi/we	deem**na** ‘we go’	nuti, nuyi/we	Dhugna ‘we drink’
second person plural	Isin/you	deem**tu** ‘you go	Isin/you	Dhugdu ‘you drink’
third person plural	Isaan/they	deem**u** ‘they go’	Isaan/they	Dhugu ‘they drink’

Double-Consonant Ending Stems: are formed when the stem word ends with double consonant, a slight modification is performed on the regular suffix to form other verbs because in Afaan Oromo it does not allowed three consecutive consonants to occur in a word [7]. This is different from regular verbs because it is not possible to form the verbs by using suffixes as we have used in regular verbs ‘a, ta, ti, na, tu, u. See the following example in table below. For example if we write ‘a**rgt**u’ this is wrong because three consecutive consonants are not allowed.

Table 3.7 Examples of double consonant ending stems.

Arguu – 'to see'		
First person singular	ani /I	arg**a** ‘I see’
	ati/ you	arg**ita** ‘you see’
third person singular	Inni/he	arg**a** ‘he see’
	Isheen/she	arg**a** ‘she see’
First person plural	nuti, nuyi/we	Arg**ina** ‘we see’
second person plural	Isin/you	arg**itu** ‘you see’
third person plural	Isaan/they	arg**u** ‘they see’

3.5.3 Adjectives

Adjectives are words that describe or modify nouns and pronouns [11][55]. In this language, adjectives used to describe or modify noun [57]. For instance, in **Guddataan kophee gurraacha godhate** “Gudeta use black shoose” the adjective gurraacha comes after the noun kophee.

In Afaan Oromo adjectives are categorized depend on their functions like in English such as those express color, quality, size, shape and taste [29].

Halluu (colour)	**hamma (size)**	**boca (shape)**	**Dhamdhama (taste)**
Gurraacha (black)	dheeraa (tall)	sirrii (straight)	mi’aa (sweet)
Cuquliisa (blue)	gabaabaa (short)	geengoo (circle)	soogidaawaa (salty)

Adii (white)
Daalacha (gray)
Diimaa (red)
Magariisa (green)

dhiphaa (narrow)
yabbuu (thick)
xinnaa (small)
gadifagoo (deep)

rogarfee (rectangle)
rogsadee (triangle)

oganaawaa (sour)
geengoo (circle)
asheeta (fresh)

3.5.4 Adverbs

In this language, adverbs are used to modify all parts speech [29] and also used to describe or modify the manner how activities are done or something is happened [56]. Afaan Oromo adverbs are classified into four categories [57]. They are adverb of time, place, manner and frequency. Lists of adverbs are given in table 3.9.

Table 3.9 Adverbs of Afaan Oromo

	Afaan Oromo	English	Afaan Oromo	English
Adverb of time	Kaleessa	Yesterday	gaafas	Then
	har'a	Today	eger	Later
	Bor	Tomorrow	amma isa ammaa	right now
	amma	Now	hatattamaa	immediately
Adverb of place	as	as -here	mana	home
	Achi	There	fagoo	Away
	gara sana	over there	dhihoo	Near
	iddoo hunda	every where	ala	Out
	eessayyuu	Nowhere		
Adverb of manner	Baayyee	Very	Qalbiidhaan	Carefully
	Baayyee	Quite	Walii wajjin	Together
	Dhugumaan	Really	Qofaa	Alone
	Dafee	Fast	Cimaa	Hard
	Suuta	slowly		
Adverb of frequency	yeroo hunda	always	darbee darbee	Seldom
	gaaf gaafii	some times	yoomiyyuu	Never

3.5.5 Conjunctions

Phrases, clauses, sentences are combined or joined by unchanged called conjunctions [11][29]. Conjunctions can be classified into coordinating and subordinating [29]. Coordinating is used to connect two independent clauses [62]. Some of Afaan Oromo conjunctions are: moo ‘or’, garuu ‘but’, haata’u male ‘however/so’, kanaafuu ‘therefore’, ta’ullee ‘even though‟ etc. For example: **Caaltuu carraa barumsaa Ameerikaatti bilisaan argatte garuu eeyyema deemuu**

himuu hin qabdu. Which means “Caltu has got scholarship at Amerika but she ha no visa”. ‘**garuu**’ in this sentence is coordinating conjunction which is used to join the two sentences **Caaltuu carraa barumsaa Ameerikaatti bilisaan argatte** and **eeyyema deemuu himuu hin qabdu** which are independent.

Subordinating conjunctions: connect main clause subordinate clause. hamma (until), wayta/yeenna (when),yoo (if), akka waan (as if), erga (after), dursa (before) etc are some of Afaan Oromo subordinating conjunctions . Example: <u>Wanta</u> ynaachuuf jedhu fakkaata. Akka wanta is a subordinating conjunction and it joins one subordinating clause that is ynaachuuf jedhu ‘as it is to eat‟ and fakkaata ‘it seems‟.

3.6 Word and Sentence Boundaries

Afaan Oromo uses white spaces to separate a word from another words like any another Latin languages [29]. In addition to this, word boundaries can bee seen by brackets, parenthesis and quotes [11]. Moreover, Afaan Oromo language one sentence can be separate from another sentence by a question mark (?), a period (.), or an exclamation point (!) [11] like that of in English language.

Chapter Four

Design and Implementation

4.1 Introduction

In this chapter the researcher discussed the design and implementation of Automatic Afaan Oromo news text summarizer (AAOTS) using sentence scoring. The proposed summarizer design and implementation consists of the following parts. The first part is the design part, in which corpus, validation data preparation and analysis are discussed. Next, the implementation part in which the required techniques from the input text to the summarized data are discussed deeply. Thirdly, the prototype which is used to display to user the summarized data present.

4.2 Design

Design in research is the arrangement of situations for gathering data and arranging or analysis of data based on the research objective. Research design shows all the procedures the researcher work in the paper. These works are collection, measurement and analysis of data. For this work the researcher used experimental type of research design. This method is good to test the system that why it is selected. So, the proposed system developed using NLTK and Python. And the design has three phases of text summarization: pre-processing, sentence scoring and summary generation.

This architecture is named Automatic Afaan Oromo Text Summarizer (AAOTS). This architecture has two parts named Database based part and sentence scoring part. The first part contains the information about Afaan Oromo language such as abbreviation, stop word, affixes, synonym, cue phrase and is called data base. Whereas the second part is the part that access data base to give for a sentence score and generate summary which is called sentence score. Sentence score part contains features such as thematic words, title words, cue phrases, sentence position, name of numbers, term weight, capital letters and event such as name of weeks, days and time. Then sentence scoring calculate score of sentences depend on these features. Then depend on the computed score the system select and include n top sentences in the summary. The basic architecture of Automatic Afaan Oromo Text Summarizer system is shown in Figure 4.1. The input for the preprocessing is the original text, then after the input is taken to preprocessing then tokenization, stop word removal and stemming is discussed under 4.3.1. And this is an input for sentence scoring. All the boxes represent a module of each function.

That means for example box which contain sentence score based on sentence position represents module which calculate sentence score based on the sentence position and description of all of the diagram discussed under 4.3.1 and 4.3.2. s

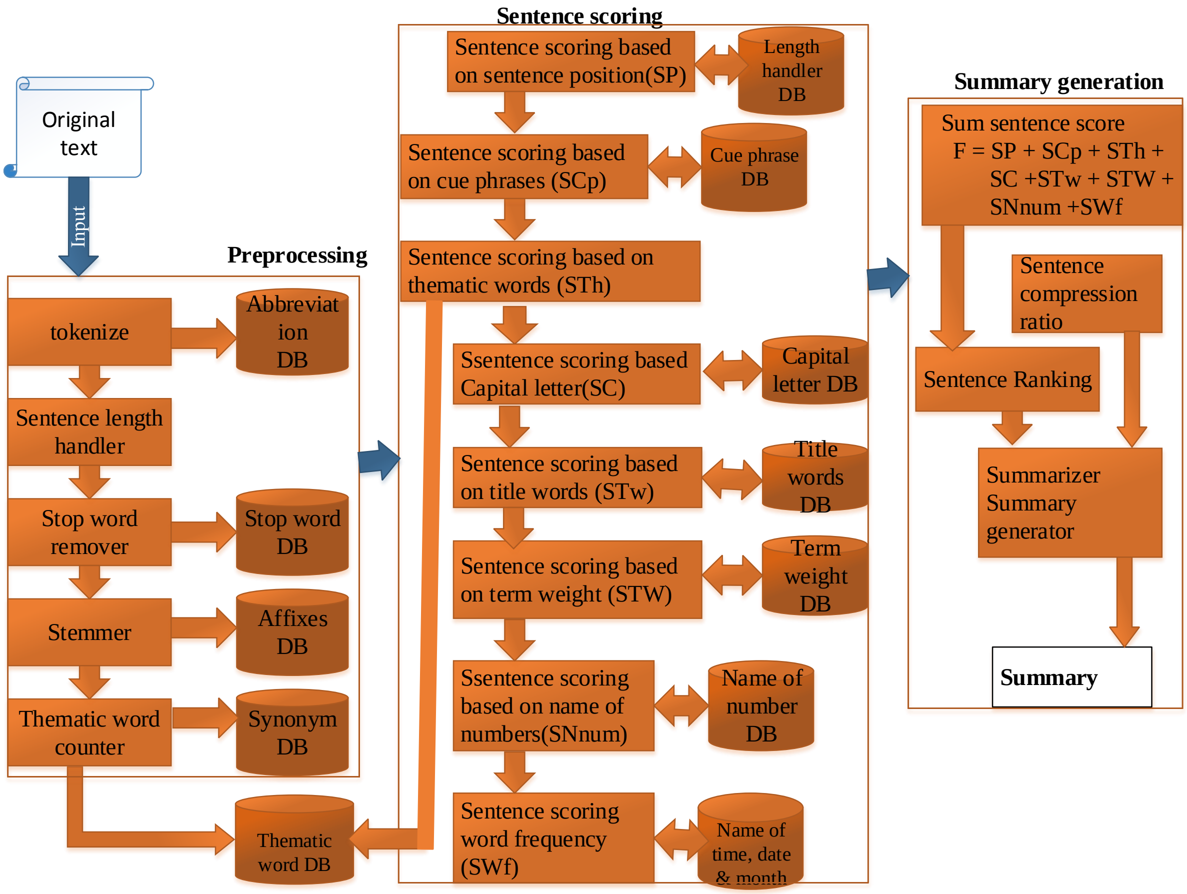

Original text
Input
Preprocessing
tokenize
Abbreviation DB
Sentence length handler
Stop word remover
Stop word DB
Stemmer
Affixes DB
Thematic word counter
Synonym DB
Thematic word DB
Sentence scoring
Sentence scoring based on sentence position(SP)
Length handler DB
Sentence scoring based on cue phrases (SCp)
Cue phrase DB
Sentence scoring based on thematic words (STh)
Ssentence scoring based Capital letter(SC)
Capital letter DB
Sentence scoring based on title words (STw)
Title words DB
Sentence scoring based on term weight (STW)
Term weight DB
Ssentence scoring based on name of numbers(SNnum)
Name of number DB
Sentence scoring word frequency (SWf)
Name of time, date & month
Summary generation
Sum sentence score
F = SP + SCp + STh + SC +STw + STW + SNnum +SWf
Sentence compression ratio
Sentence Ranking
Summarizer Summary generator
Summary

Figure 4.1 AAOTS Architecture

4.2.1 Data Base Preparation

Here the following sub sections discussed data deposited in each data base as a corpus and how to prepare this corpus.

4.2.1.1 Afaan Oromo cue phrase Corpus

In a given sentence words or phrases used to determine as that sentence is important or not is called cue phrase [29]. As [28] gives some example of English cue phrases are: “as a summary, as a result, the most important, precisely” can be a good indicator of the importance of the content.

In this work 772 cue phrases are used. 728 of these cue phrases are taken from Asefa’s and Fiseha’s paper and the remaining 44 cue phrases are prepared by the researcher. The researcher used three peoples to translate these 44 English cue phrases to Afaan Oromo who are a domain expert from higher education. The 44 English cue phrases to 44 Afaan Oromo cue phrases and Appendix A contains the cue phrase list. Table 4.1 shows sample on Afaan Oromo cue phrases. Cue phrases are stored in database after it were collected in to Afaan Oromo to use as knowledge based.

Table 4.1: Sample Afaan Oromo cue phrase

English cue phrases	Its translation to Afaan Oromo
attempt	Yaalii
because of that	Sababa sanaaf
conclusion	Xumura
consequence	Bu’aa
develop	Guddisuu

4.2.1.2 Afaan Oromo stop-words Corpus

Stop words are words that cannot descriptive text document, because they cannot change meaning of the sentence [11][22]. Stop words removed from the document or sentences at the preprocessing stage that means before processing [10].

In this work the researcher used 148 stop words, 21 from Girma’s paper, 99 stop words from Debela’s paper and 28 from Asefa’s paper [10]. Appendix B: contains list of stop words. And table 4.2 below shows sample of Afaan Oromo stop words.

Table 4.2: Sample Afaan Oromo stop words

Afaan Oromo	**English**
Booddee	Behind
Hanga	Until
Kana	This

4.2.1.3 Afaan Oromo synonyms Corpus

Word or phrase with in the same language having the similar meaning is called synonyms and they are used to represent the concept [11][22]. As Addis Ashagre [22] and other scholars stated the purpose of using the synonym words is to detect words that have similar or identical meaning and used interchangeable. Example dhara ‘false’ and soba ‘false’ refer to same thing.

Since extractive method extract sentence at once not word by word, synonym words are a challenge for extractive [10]. Due to list of synonym words is important to use. List of synonym that are used in this work is found on appendix D. And sample list of them is shown in table 4.4.

Table 4.4 Samples of Synonym words

Word	Synonym	Transit
geddaruu	Diddiiruu	Changing
herreguu	Yaaduu	Thinking
warra	Maatii	Family
Wayyaa/huccuu	Uffata	Cloth
mi'a	meeshaa	Material

4.2.1.4 Name of Events in Afaan Oromo Corpus

In Afaan Oromo language there are names for time, dates and months like in any other languages [10]. So corpus of time, dates and months is important in order to extract sentences containing these events. Fiseha collected list of these events from different Afaan Oromo books, news portal and websites [10]. And he integrated in name of time, date and month’s data base (Fig. 4.1). In this work the researcher used all names of events prepared by Fiseha. Here time, date and month listed on appendix E and sample of them are listed in the following table 4.6.

Table 4.6: Afaan Oromo time, date and month sample

Afaan Oromo	**English translated**
Waaree	Afternoon
Guyyaa	Day
Facaasaa, Kibxata, Balloo, lammaffoo, Salaasa	Tuesday
Arbii, Roobii	Wednesday
Bitootessa,	March
Caamsaa,	April

4.2.1.5 Afaan Oromo numbers Corpus

Corpus prepared under this title is to give score for sentences having number. As it is explained in chapter 3 numbers occur in different forms. Afaan Oromo numbers can exist in cardinal, ordinal, nominal, decimal number and decimal point. Decimal number and decimal points are separated by point (.). So decimal numbers are numbers located before point. Whereas decimal points are numbers come next to point. Hence, list of names of decimal numbers are collected and prepared by Fiseha [10], whereas list of names of decimal points are collected from textbooks, journals and news portal by researcher; And stored data base (Fig. 4.1). List of numbers are found on appendix F. And sample of them is shown in table 4.7.

Table 4.7: Sample list of Afaan Oromo decimal number and decimal point

Afaan Oromo numbers		
Sadii	Soddoma	Saadeettama
Afur	Sodomii	Sadetamii
Zeeroo tuqaa afur/afur kurnaffaa	Digdama tuqaa	
Zeeroo qutaa shan/shan kurnaffaa	Soddoma tuqaa	
Zeeroo tuqaa ja’a/ja’a kurnaffaa	Afurtama tuqaa	

4.2.1.6 Corpus of Afaan Oromo Capital Letters

Capital letters are letters with upper-case letters found at the beginning of words. And they are used in various conditions as in English in Afaan Oromo also [67]. These situations are a name given to company, name given days in the week, name of holidays, name of a person, name of a country, name state, name of city, name title of a book, and name of title of a song. This feature assigns higher scores to words that contain one or more capital letters [68] which can be a proper name or important word. So high scores are given to sentences if they contain high number of words that start with capital letters. And this sentence score can be computed by computing total of first letter capital words in sentence using the number of first capital words in sentence over number of words in sentence. i.e.

$$TCW(Si) = \frac{NCW(S)}{NW(S)} \qquad 4.1$$

NCW(s) is number of first letter capital words in sentence. NW(s) is number of words in sentence

And score of sentence is computed by the following equation:

$$Score\ f(Si) = \frac{TCW(Si)}{Max(TCW(Si))} \qquad 4.2$$

Where TCW (Si) is total of first letter capital words in sentence.

4.2.1.7 Afaan Oromo Term Weight Corpus

Technique giving an index to a term to determine the importance of word in a sentence is called Term weighting [29], by assigning a weight for each words in a sentence to represent the importance of a term. Thus, the number of occurrence of terms within a document has been used for computing the importance of sentence which can be calculated by first calculating weight of each individual terms using the following equation.

$$wt = \frac{frequency\ of\ the\ term}{total\ no.of\ terms\ in\ the\ document} \qquad 4.3$$

Score of a sentence can be computed as follows:

$$\mathrm{wt}_s = \sum_{i=1}^{n}(\mathrm{wt}i)/n \qquad 4.4$$

wt_s is total weights of sentences, wt_i weight of each sentence and n is number of sentences

4.2.2 Validation data preparation and analysis

For extractive type of text summarizer data is necessary in order to adjust a parameter validation [10]. The researcher used 30 news (articles) from different sources. These 30 articles are given to 40 peoples in the form of questioner. Then the articles are analyzed and organized as shown in the table 4.8. The compression ratio is analyzed by using equation 2.10. i.e. by dividing the length of summary by length of full text. Calculation was done for all the three extraction rates as indicated in the table 4.8 below.

Table 4.8 Statistics of the training corpus

Text ID	News size in words	News size in sentences	No of paragraph	Compression rate (% of summary tested) at different rate		
				20%	30%	40%
Topic 1	528	17	8	26.13%	35.79%	43.93%
Topic 2	221	8	4	41.17%	41.17%	53.39%
Topic 3	257	15	5	64.59%	77.43%	90.27%
Topic 4	217	12	3	22.58%	39.63%	48.38%
Topic 5	277	15	6	28.54%	38.54%	47.89%
Topic 6	299	24	7	77.59%	90.03%	99.33%
Topic 7	412	29	8	25.35%	34.12%	45.54%
Topic 8	474	41	8	24.75%	39.07%	52.18%
Topic 9	394	24	6	86.04%	95.93%	100%
Topic 10	304	14	5	46.71%	69.07%	91.11%
Topic 11	937	78	13	27.96%	37.99%	45.78%
Topic 12	638	37	8	29.45%	42%	60.45%
Topic 13	720	47	9	33.07%	52.50%	67.86%
Topic 14	280	19	4	45.71%	51.42%	56.07%
Topic 15	339	13	5	57.22%	72.56%	78.46%
Topic 16	292	12	4	22.60%	40.41%	44.52%
Topic 17	287	13	5	33.10%	42.16%	47.73%

Topic 18	268	12	4	32.40%	41.25%	49.54%
Topic 19	418	29	5	52.61%	67.16%	86.56%
Topic 20	354	15	6	12.52%	32.20%	44.06%
Topic 21	290	11	5	21.03%	31.03%	39.53%
Topic 22	291	12	5	8.59%	28.17%	34.70%
Topic 23	429	19	7	27.73%	33.33%	43.82%
Topic 24	239	9	5	26.77%	36.40%	52.71%
Topic 25	519	26	9	20.61%	36.99%	44.31%
Topic 26	544	19	8	38.97%	50.91%	59.74%
Topic 27	1054	83	19	23.71%	32.73%	40.32%
Topic 28	591	30	11	24.57%	34.25%	52.24%
Topic 29	189	8	4	40.74%	40.74%	60.31%
Topic 30	863	35	11	33.32%	43.20%	60.00%
Average	**430.83**	**24.2**	**6.9**	**35.20**	**46.94%**	**58.03%**

Table 4.8 shows the analysis of the 30 topics and it is used to make flexible the experiment of computing precision and recall.

4.3 Implementation

To conduct the experiment the researcher used NLTK using python 3.7.4 to develop the summarizer system first installed Python 3.7.4 and then NLTK on python 3.7.4 version. Different libraries are imported like beautiful soup, nltk, re, and so on. Researcher used this tool to develop the system and do experiments experiment with it. The workflow used to complete this work is discussed as follows:

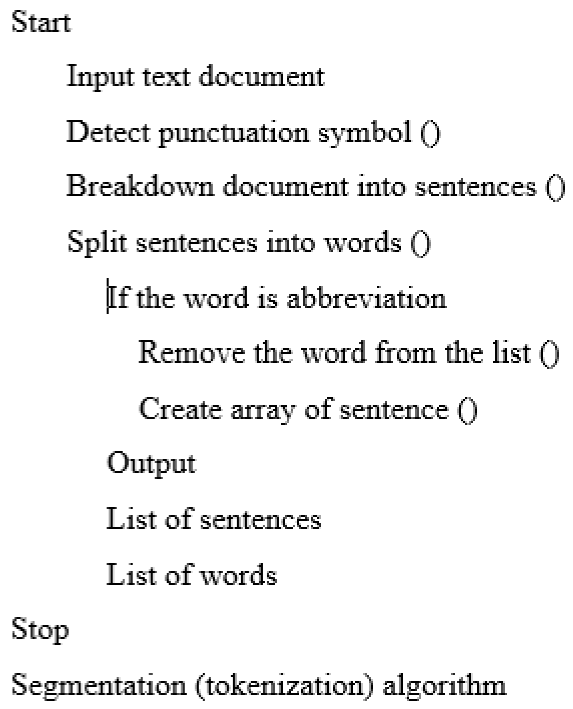
4.3.1 Phase I: Pre-processing

Preprocessing technique used by most of the researcher of text summarization [10]. And this phase is preparing data for processing. Chunking words into words by tokenization, removing stop words, stemming words in to verbs and suffixes, sentence length handling and etc happen in this phase.

4.3.1.1 Sentence Segmentation and Tokenization

Summary generation depend on the calculated value of each sentence, where calculated value a sentence depends on value of computed each words generate that sentence [29]. Hence, Sentence segmentation is the process of breaking down text documents into sentences and identifying boundary of sentences [10][29]. Thus, in this system text documents are breakdown into sentences by identifying the boundary of sentence which ends with period symbol (.), question mark (?), exclamatory mark (!).

Tokenization is the process of separating the input document into individual words [55]. And this is done by identifying the spaces ( ), comma (,) and special symbols between the words. In this process frequency of each word is computed and stored for further processing. One of the difficulties concerning the sentence boundaries is periods used as full stop and also in abbreviation [29]. For example w.k.f. waan kana fakkaatan (such a like), k.k.f. Kan kana fakkaatan (such like this), I.A.I.G.M.B itti aanaa itti gaafatamaa mana barumsaa (vice president of the school) (etc). To handle this problem the researcher used list of abbreviation prepared by previous researchers. And they are given in Appendix C. Sentence segmentation task given by algorithm below



```
Start
    Input text document
    Detect punctuation symbol ()
    Breakdown document into sentences ()
    Split sentences into words ()
       If the word is abbreviation
           Remove the word from the list ()
           Create array of sentence ()
       Output
       List of sentences
       List of words
Stop
Segmentation (tokenization) algorithm
```

Figure 4.2 Algorithm for Sentence Segmentation and Tokenization

4.3.1.2 Sentence Length Handler

When we summarize text documents too short sentences or too long sentences removed because they decrease the quality of the summary [28]. Hence, this section control this challenge by using array of tokenized sentences.

So, to build a module which handle this problem, what the researcher did in this study is, taking 20 topics that are prepared for validation purpose conduct an experiment to identify too long sentence and too short sentence. So in order to identify too long or too short sentence, length of sentences was grouped into equal range of clusters as shown in table 4.9. Whereas range is a sentence contains one to five words. The range is in five interval of number of words. This data has minimum sentence length 2 words and maximum sentence length 44 words; and nine different ranges.

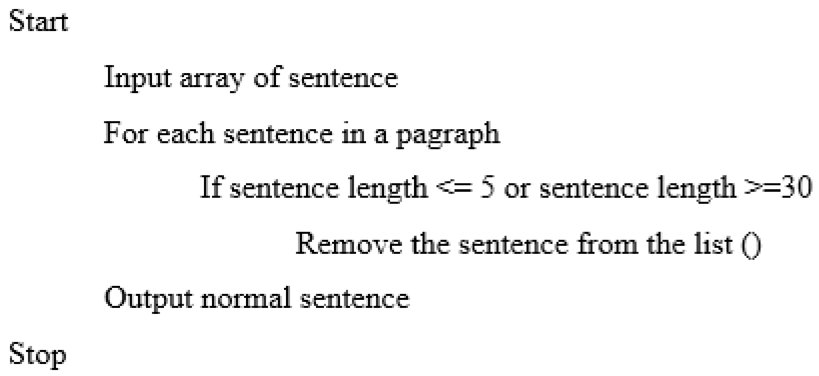
In table 4.9 below range shows the number of words. That is for instance [1, 5] means sentences that contain one up to five words. (5, 10] means sentences that contains greater than five words up to ten words and so on. And range contains sentence amount in number shows the total number of sentences that is with similar length in the given range. For instance in range (1,5] there are 35 sentences with the same sentence length in the interval one up to five and probability represents , the percentage of sentence that exist in a given range from the total 544 sentences and it is calculated using equation 4.5.

$$\text{Probability} = \frac{\text{range contains sentence amount in number}}{\text{total number of sentence}} * 100\% \qquad 4.5$$

Table 4.9 Probability of sentences with in each range by percent.

Range	Range contains sentence amount in number	Probability
[1,5]	35	6.43%
(5,10]	152	27.94%
(10,15]	118	21.69%
(15,20]	104	19.11%
(20,25]	72	13.23%
(25,30]	38	6.98%
(30,35]	18	3.30%
(35,40]	5	0.91%
(40,45]	2	0.36%
Total	544	100%

Generally, we observe from this table (table 4.9) sentences with less than five words are short sentences and sentences with greater than 30 words are moreover long sentences. So based on this result the following algorithm is designed for sentence handling.



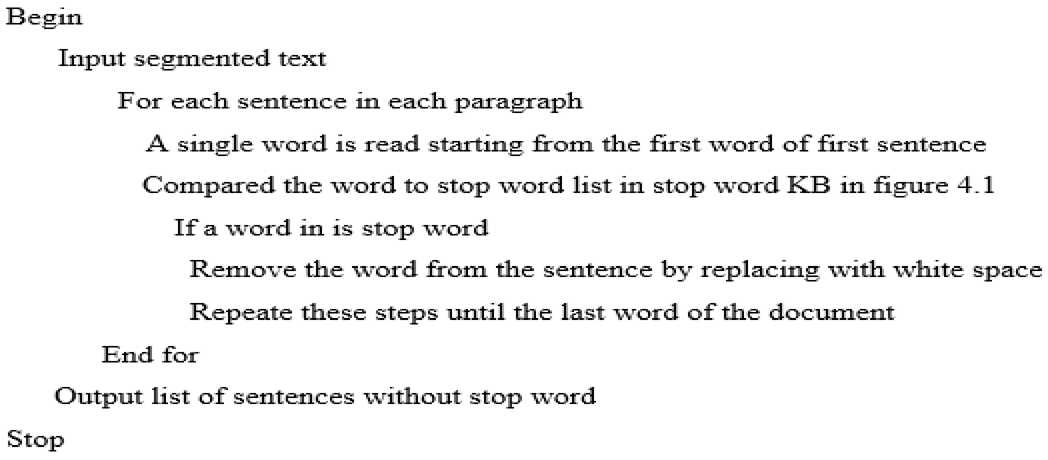
```
Start
    Input array of sentence
    For each sentence in a pagraph
        If sentence length <= 5 or sentence length >=30
            Remove the sentence from the list ()
    Output normal sentence
Stop
```

Figure 4.3 Algorithm for Sentence Length Handler

4.3.1.3 Stop Word Remover

Stop-word remover is removing words that has no significant contribution to the overall idea of the document but the most frequent terms that are common to every document [29]. Stop words may include words such as articles, conjunctions, prepositions, and pronouns, or are words appear in many sentences [28]. Some of Afaan Oromo stop words are “kun” means ‘this’, “fi” means ‘and’ and “koo” means ‘mine or my’ that cannot change the meaning of the document if they removed from the document.

The stop word remover removes stop words from the input document by finding the match with in the list of stop word in data base (Figure 4.1) after it is tokenized. List of stop words for this work is find in appendix A and its defined algorithm for stop words handling in this thesis work is given below.



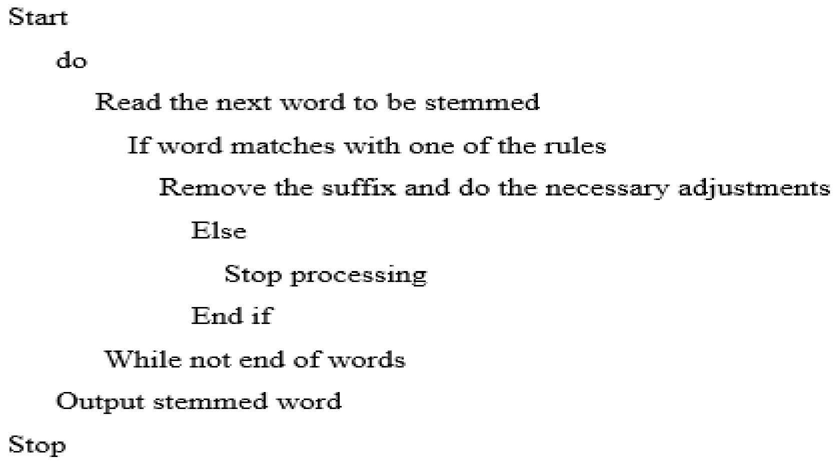
```
Begin
    Input segmented text
        For each sentence in each paragraph
            A single word is read starting from the first word of first sentence
            Compared the word to stop word list in stop word KB in figure 4.1
                If a word in is stop word
                    Remove the word from the sentence by replacing with white space
                    Repeate these steps until the last word of the document
      End for
    Output list of sentences without stop word
Stop
```

Figure 4.4 Algorithm of stop word remover adjusted from Fiseha

4.3.1.4 Stemmer

Stemmer come after stop words removed from document. This step of preprocessing is removing the affixes from the inflected forms of a word to get stem which is root form [55].

For example, walks, walking and walked can be stemmed to the same root word walk. In the same to English, Afaan Oromo words like ‘seeraan’,’seeraaf’,’seerota’ are stemmed to the stem word ‘seer’. The word ‘seer’ is no more stemmed to another word. So stemming is useful to search one word from different word forms to obtain information that contain another word in the same stem group [55]. Here as this example shows a single word has different morphological variants which miss guides the key counter during counting frequency of words in sentences [10]. So stemmer is important to split words into stem and affix.

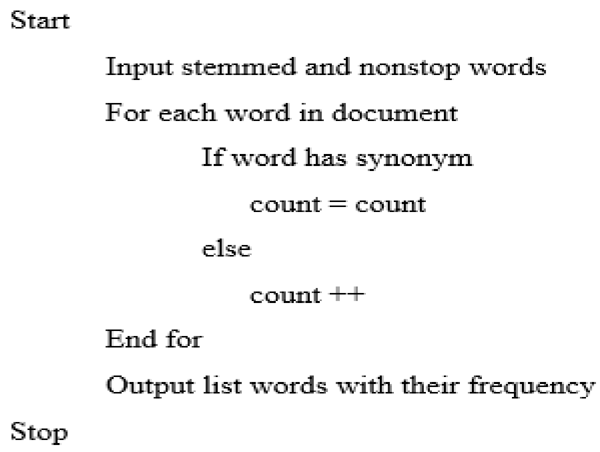


```
Start
  do
    Read the next word to be stemmed
      If word matches with one of the rules
        Remove the suffix and do the necessary adjustments
          Else
            Stop processing
          End if
     While not end of words
  Output stemmed word
Stop
```

Figure 4.5 Algorithm of Stemming taken from Debela

4.3.1.5 Thematic Words Counter

After the words are stemmed in to root and affixes then the next step is counting the frequency of words. Which is helpful for ranking words according to their frequencies. So in this section adjective, verb, noun and adverb which are out of stop words but found in every sentence are counted and the highest frequent one is taken as a thematic word. Thematic words are probably describe topic more, because the words that repeatedly occur in text are probably topic related word [61]. Under this section the inputs are the stemmed words and synonym words in data base (Figure 4.1). In this study it count the frequency of each words of a document. Finally, after the frequency of each word in document is counted, then generates data base known as thematic word data base, which contains list of thematic words with their frequency in descending order. The thematic word algorithm used in this work is given below in figure 4.6.



```
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
```

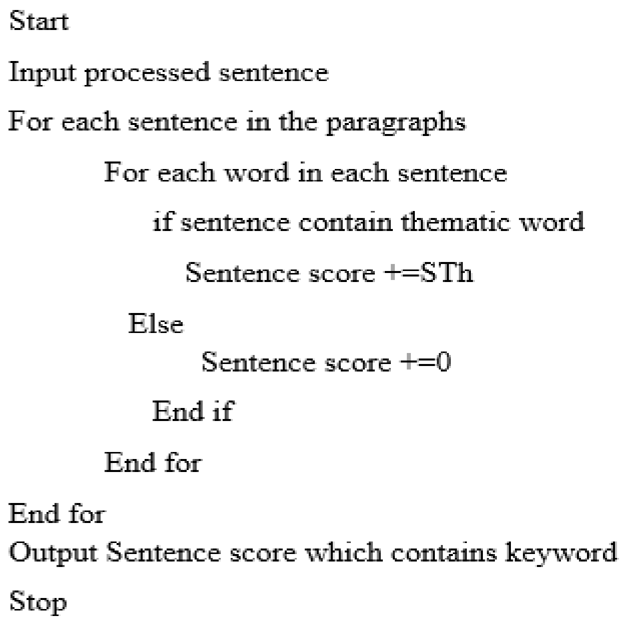
Figure 4.6 Thematic word counter algorithm

4.3.2 Phase II: Sentence scoring

Sentence scoring is one of the most used processes in the area of NLP while working on textual data. After the completion of phase 1 the input document for phase 2 is formatted and segmented into collection of words in which each word has its individual frequency. In this phase the score of the sentences computed based on the features: title words, thematic words, name of number, cue phrase, capital letter, term weight, sentence position, sentence length, and event such as name of weeks, days and time for summarization. Let’s see one by one the technique of these features score computation in the following sections.

4.3.2.1 Sentence Scoring Based on Thematic words

According to [46], thematic words are keywords which are helpful in deciding sentence importance. Thematic words are most frequent content words in the documents and their number shows that the words that mostly have relativity to topic words as discussed in section 4.3.1.5. A small number of thematic words are selected and each sentence which contain these words are scored as a function of frequency. So in this work top n keywords are used for consideration as thematic and a sentence that has maximum number of thematic words is selected for summary. And its score is computed by dividing number of thematic words that occurs in a sentence to the maximum number of thematic words in a sentence. Use the following algorithm for thematic word.



```
Start
Input processed sentence
For each sentence in the paragraphs
        For each word in each sentence
            if sentence contain thematic word
                Sentence score +=STh
          Else
                Sentence score +=0
            End if
        End for
End for
Output Sentence score which contains keyword
Stop
```

Figure 4.7 Algorithm for sentence scoring using thematic words

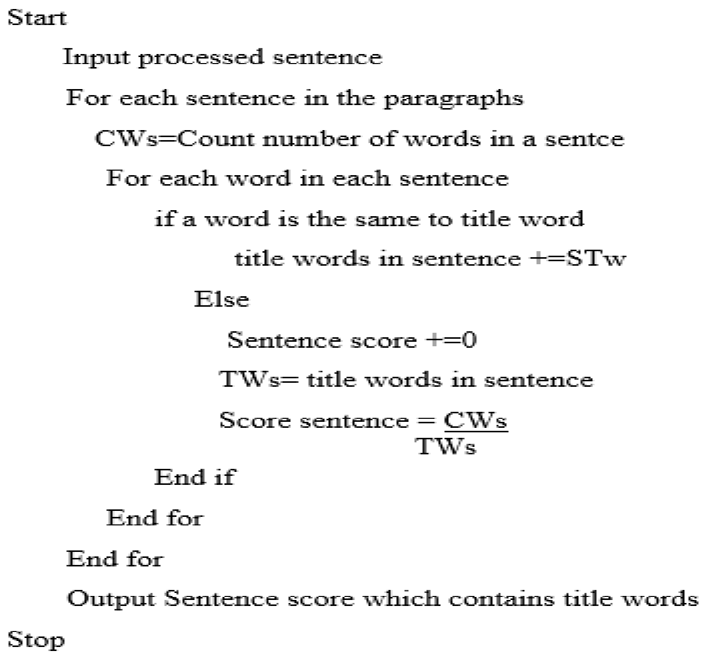
4.3.2.2 Sentence Scoring Based on title words

Title words could be considered as the essential part of the document because of the authors always used contents related to the title for filtering the article [15]. The amount of title word in a sentence, and also words in the text sentence also appear in title gives high score [61], So a sentence is considered as important if it contains maximum number of words occurring in the title since it is an indicative of the theme of the document. Not only this one, a high score is given to the sentence if it contains words occurring in the title as the main content of the document is expressed via the title word. Therefore, these sentences have greater chances for including in summary. The score of title words are calculated by counting the number of words common to both title words and sentence words and then divide by the total number of words in the title [61]. The following table shown as an example. And then rank in descending order. Then top 10 sentences are taken as summary of the document. Table 4.10 below shows how title words score computed. For example, Dubartii ulfa abbaa warraashii qurxummii shaarkiin nyaatamuurra hambiste.

Table 4.10 How title words computed

Sentence	No of title words in a sentence	Total No. of title words	$S(s) = \frac{\text{No of title words in sentence s}}{\text{total No. of words in sentences}}$
1	6	8	0.75
2	1	8	0.125

3	2	83	0.25
4	1	8	0.125
5	2	8	0.25
6	4	8	0.5
7	1	8	0.125
8	1	8	0.125
9	2	8	0.25
10	3	8	0.375

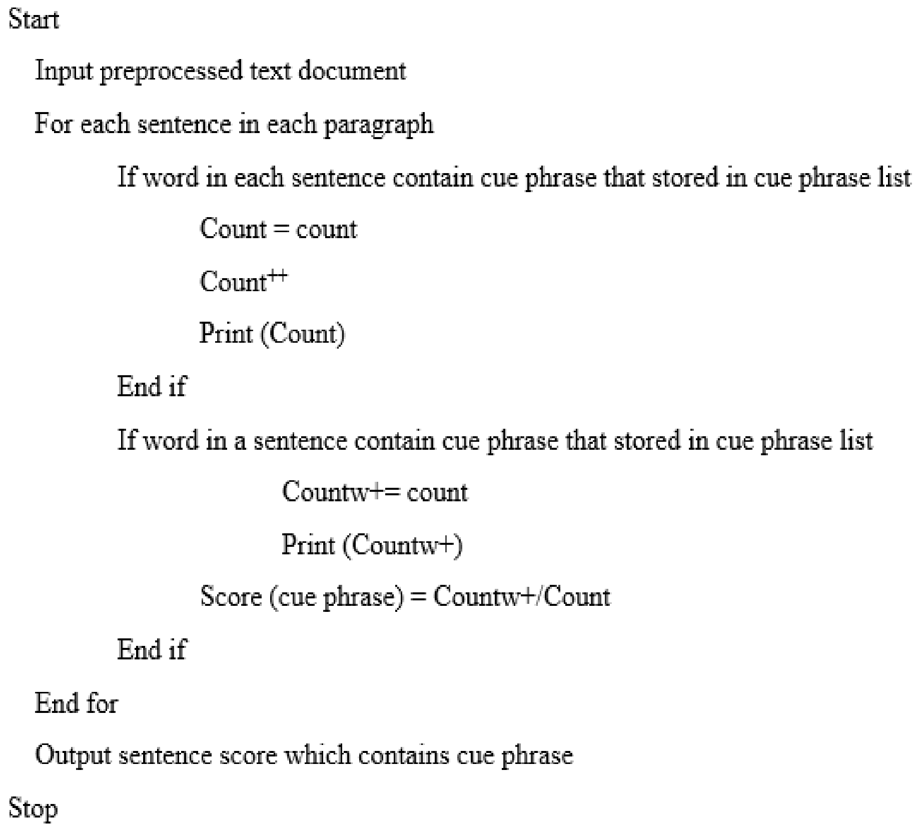


```
Start
    Input processed sentence
    For each sentence in the paragraphs
      CWs=Count number of words in a sentce
       For each word in each sentence
          if a word is the same to title word
               title words in sentence +=STw
            Else
              Sentence score +=0
             TWs= title words in sentence
             Score sentence = $\frac{CWs}{TWs}$
          End if
       End for
    End for
    Output Sentence score which contains title words
Stop
```

Figure 4.8 Algorithm for Sentence Scoring Based on title words

4.3.2.3 Sentence Scoring Based on Cue phrases

The existence of some phrases in the sentence can be a good indicator of the importance of the content [19][28]. This system compare cue phrases in a sentence with a list of cue phrases in cue phrase data base (Figure4.1). So, sentences that contain cue phrases are given a higher score compared to other sentences. The score of this feature is computed by dividing the number of cue phrases in a sentence to the number of cue phrases in the document.

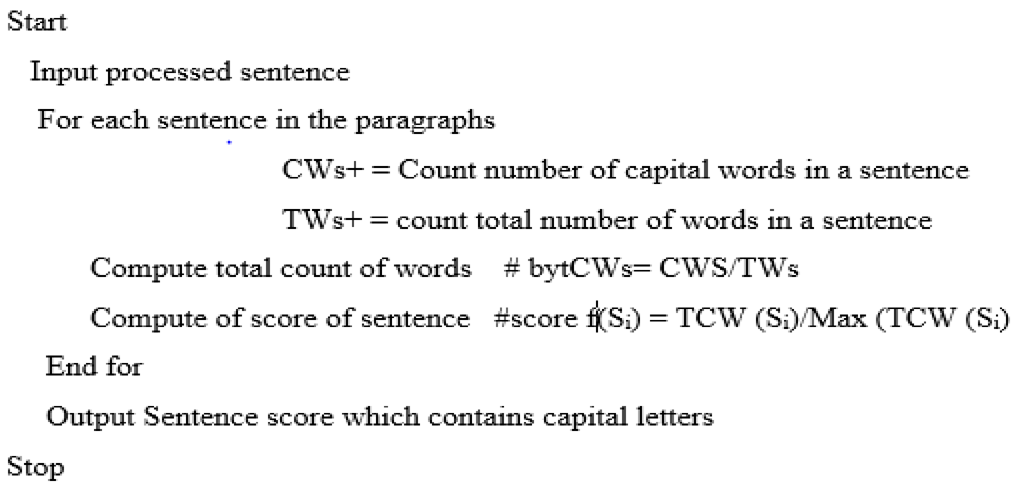


```
Start
  Input preprocessed text document
  For each sentence in each paragraph
        If word in each sentence contain cue phrase that stored in cue phrase list
              Count = count
              Count$^{++}$
              Print (Count)
        End if
        If word in a sentence contain cue phrase that stored in cue phrase list
                    Countw+= count
                    Print (Countw+)
              Score (cue phrase) = Countw+/Count
        End if
  End for
  Output sentence score which contains cue phrase
Stop
```

Figure 4.9 Algorithm for Sentence Scoring Based on cue phrases

4.3.2.4 Sentence Scoring Based on Capital Letters

Sentences that have upper case letters score higher [64]. This feature can be helpful for generating summary of news articles and short stories which contain plenty of proper nouns. As it is discussed in section 4.2.1.8 as in English capital letters are also used in Afaan Oromo letters in situations such as for name of company, name of days in the week, name of holidays, name of persons, name of country, name of state, name of city and name of the title book. So, this feature assigns higher scores to words that contain one or more capital letters of one of the above situations. Thus, high scores are given to sentences if they contain high number of words that start with capital letters. And sentence score by using this feature is computed by equation 4.2.



```
Start
  Input processed sentence
   For each sentence in the paragraphs
                    CWs+ = Count number of capital words in a sentence
                    TWs+ = count total number of words in a sentence
       Compute total count of words    # bytCWs= CWS/TWs
       Compute of score of sentence   #score f(S_i) = TCW (S_i)/Max (TCW (S_i)
   End for
   Output Sentence score which contains capital letters
Stop
```

Figure 4.10 Algorithm for Sentence Scoring Based on Capital Letters

4.3.2.5 Sentence Scoring based on Term Weight

Technique giving an index to a term to determine the importance of word in a sentence is called Term weighting [29], by assigning a weight for each words in a sentence to represent the importance of a term. So using term frequency, Inverse sentence frequency we calculate term weight value for words. Hence, taking the preprocessed document the system compute weight value for all words using equation 4.3. After weight value of each term computed then the system compute weight of sentences by adding weight of terms which is sentence score. Then put in descending order depend on their weight value. Finally summary extract taking top n sentences. Thus the algorithm of this is as follows:

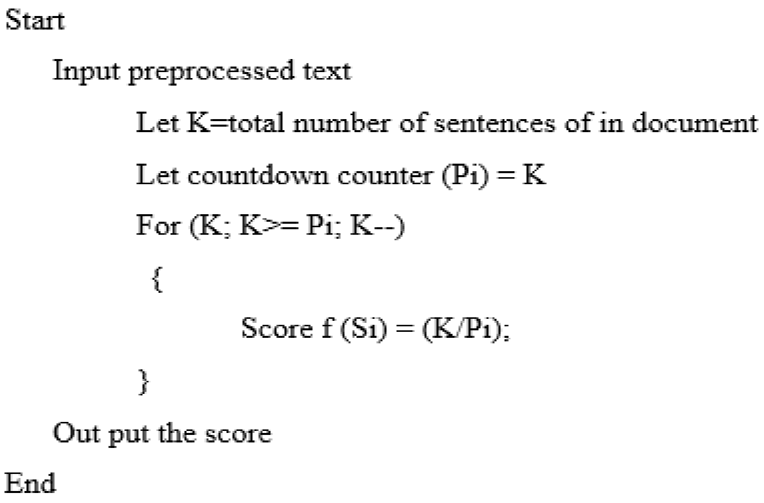
a. Read preprocessed document
b. Assign a weight score to each terms.
c. Compute sentence score or assign weight value to sentences, which can be computed as follows:

$$wt_s = \sum_{i=1}^{n} (wt_i) /n \qquad 4.6$$

d. Extract summary

4.3.2.6 Sentence Scoring Based on Sentence position

Sentence position is used to work out the position of a sentence in a document in terms of the normalized percentile score in the range between 0 and 1 as discussed in section 2.6. So the significance of score of a sentence in this work is calculated based on sentence position in the document from the three features to calculate score of sentence in the document. Since position of the sentences determine the importance of the sentences in the document, the first sentence in the paragraph score high whereas the last sentence scoreless [65][69]. In this work the researcher used the same original equation 2.1 for the sentences to compute their score. This calculation is held as follows: first total number of the sentences (K) is the same to the number of counter down counter (Pi), then for the second sentence the counter down counter decrements by one, and it continues until the last sentence of the paragraph. So score of the first sentence is 1 and the score of the last sentence is 0.1. In general computed score of sentences is between 0 and 1. The algorithm of scenario is:



```
Start
    Input preprocessed text
        Let K=total number of sentences of in document
        Let countdown counter (Pi) = K
        For (K; K>= Pi; K--)
         {
                Score f (Si) = (K/Pi);
        }
    Out put the score
End
```

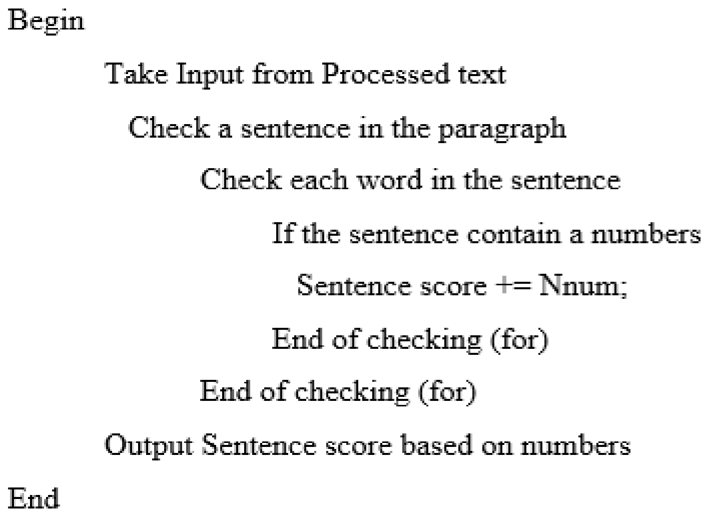
Figure 4.11 Algorithm of sentence position Adopted from Ali Al Dahoud

4.3.2.7 Sentence Scoring Based on numerical data and name of numbers

In this subsection sentences that contain ordinal, nominal, cardinal, percentage or decimal numbers are with higher score than sentences has no numbers. Data that contains numerical information important, so sentence that contains this important data most probably incorporated in the text summary [61]. Score of numerical data can be computed by dividing the number of numerical data or name of numbers in sentence by the sentence length.

$$\text{Score(Si)} = \frac{\text{No.of Numerical data or name of numbers in Si}}{\text{Length(Si)}} \qquad 4.7$$

An algorithm used to compute score Sentence Scoring Based on numerical data and name of numbers.



```
Begin
        Take Input from Processed text
          Check a sentence in the paragraph
                Check each word in the sentence
                      If the sentence contain a numbers
                        Sentence score += Nnum;
                      End of checking (for)
                End of checking (for)
        Output Sentence score based on numbers
End
```

Figure 4.12 Algorithm for Sentence Scoring Based on capital letters

4.3.2.8 Total sentence score

Sentence scoring is one of the most used processes in the area of Natural Language Processing (NLP) while working on textual data. So in this section the values of all the features like*: SP, SCp, STh, SC, STw, STW, SNnum, SWf* separately is joined by using linear combination function [10]. So for this work total score is computed by equation 4.9.

$$\sum_{i=1}^{8} Fi = F1 + F2 + F3 + F4 + F5 + F6 + F7 + F8 \qquad 4.9$$

Where, Fi is STh= F1, STw = F2, SCp= F3, SC = F4, STW = F5, SP = F6, SNnum = F7, SWf = F8

To adjust score for each eight features based on the above mentioned equation, the researcher did an experiment using validation data corpus stated in table 4.8.

The researcher conduct the experiment on ten topics by selecting randomly from the validation data corpus listed on table 4.8 to choose best equation precision, Recall and F-measure of each topic was computed using equation 2.12, 2.13 and 2.14. The experiments of each the eight features are done in chapter five.

4.3.3 Phase III: Summary Generation

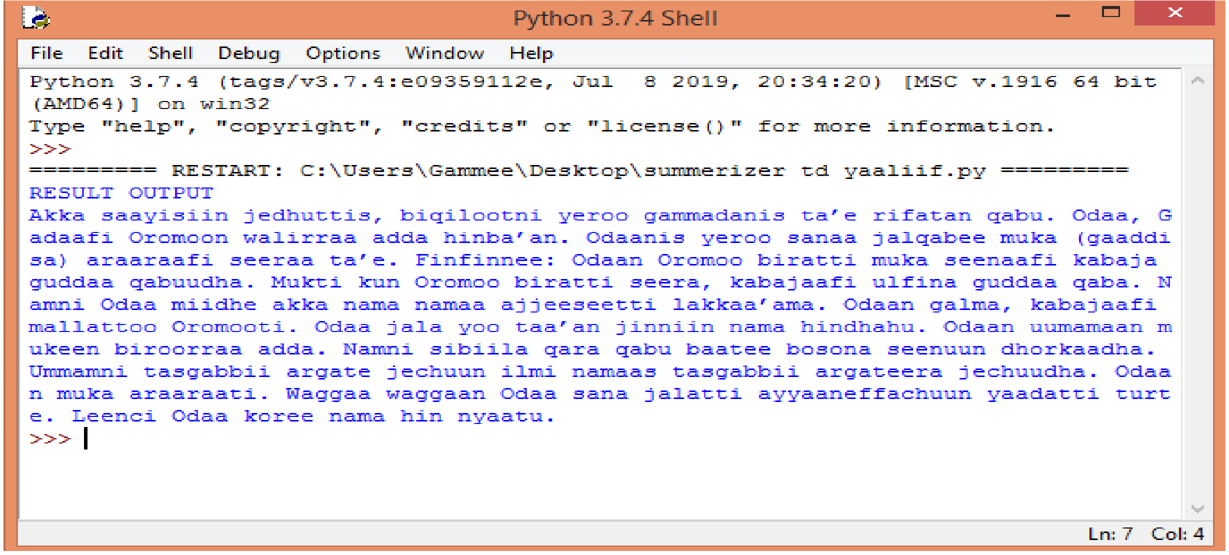
Extractive method of summarization is to identify the best applicant sentences for generating summary, which is the core of the summarization system [29]. So after each sentence is completely scored, then the sentence whose score value is highest comes first in top position and included in the summary whereas the smallest score value ignored from the summary [31].

So after all the score is computed as it is discussed under topic 4.2.3.9 then the ranking is implemented by using the function key heapq.nlargest ().

All activity concerning summary generation by taking the highest scored sentence is controlled by the function key heapq.nlargest () in this research work[7]. This function works on the first step it input an integer list in descending. Then next to first step it first traverse the list up to n times. Lastly on third step in each traverse find the largest value and store it in a new list.

Then heapq.nlargest () function extract the n top sentences depend on their score value. So after the heapq.nlargest () function extracts the n top sentences then resulting sentences put together by the sentence concatenation component to produce the document summary. According [28], extractive text summarization is used to extract important text segments (e.g. sentences) from the original document based on a set of important features from different levels (e.g. tokens, sentence, paragraph, and document). In this research work score-based summarization methods have been used to evaluate the effectiveness of the proposed features. And important sentences are extracted based on the total scores that are assigned to each sentences.
After the value of each sentence is computed then next phase is summary generation based on sentence score using heapq function.

4.4 The Prototype



```
Python 3.7.4 (tags/v3.7.4:e09359112e, Jul  8 2019, 20:34:20) [MSC v.1916 64 bit
(AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
========= RESTART: C:\Users\Gammee\Desktop\summerizer td yaaliif.py =========
RESULT OUTPUT
Akka saayisiin jedhuttis, biqilootni yeroo gammadanis ta'e rifatan qabu. Odaa, G
adaafi Oromoon walirraa adda hinba'an. Odaanis yeroo sanaa jalqabee muka (gaaddi
sa) araaraafi seeraa ta'e. Finfinnee: Odaan Oromoo biratti muka seenaafi kabaja
guddaa qabuudha. Mukti kun Oromoo biratti seera, kabajaafi ulfina guddaa qaba. N
amni Odaa miidhe akka nama namaa ajjeeseetti lakkaa'ama. Odaan galma, kabajaafi
mallattoo Oromooti. Odaa jala yoo taa'an jinniin nama hindhahu. Odaan uumamaan m
ukeen biroorraa adda. Namni sibiila qara qabu baatee bosona seenuun dhorkaadha.
Ummamni tasgabbii argate jechuun ilmi namaas tasgabbii argateera jechuudha. Odaa
n muka araaraati. Waggaa waggaan Odaa sana jalatti ayyaaneffachuun yaadatti turt
e. Leenci Odaa koree nama hin nyaatu.
>>> 
```

Figure 4.2 Prototype

[7] The nlargest () function of the Python module heapq returns the specified number of largest elements from a Python iterable list. The method heapq.nlargest () will return a list with the nlargest elements from the dataset defined by iterable. This function use the following steps to find N largest elements.

Chapter Five

Experimentation, Discussion and Evaluation

5.1 Introduction

In this Chapter, the researcher first discussed how to prepare and analysis test data, next to this tools used for implementation is described, then how to prepare experimental summary is discussed and lastly present how to evaluate developed system and discussed result.

5.2 Tools Used for Implementation

To discuss again, the development tool used to conduct the experiment is the NLTK using python version 3.7.4 have been implemented for Automatic Afaan Oromo Text Summarization. NLTK is an important platform for building Python programs to work with human language data for applying in statistical natural language processing (NLP). “This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response. And along with a group of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.”

5.3 Experimental Settings

The researcher gathered 30 Afaan Oromo articles and used for experiments and these articles are sourced from Afaan Oromo News present by different Medias as it is in section 1.6.3. The lengths of these articles ranges between 8 to 83 sentences, containing maximum number of words and sentences per documents 1054 and 83 respectively and lowest number of words and sentences per documents 189 and 8 respectively. Average length of collected news items is approximately 428.6 words and 24.43 sentences. In addition, table 5.1 shows number of sentences at each compression rate of the summary of the selected for experimentation. Using these data’s the researcher set 9 experiments as follows. The first experiment is when all the features are used in the experiment. That is when all the features used then the researcher computed for Precision (P), Recall (R) and then F-score for a single topic at three different extraction rates. For example: If we take Topic 6: The researcher took number common sentences for both manual summary and system summary which is correct =13, sentences which are only in the system summary which is missed = 3 and the sentences which are only manual summary called wrong = 3 in number at 20% extraction rate. And then, the researcher computed for R = $\frac{correct}{correct + missed}$ = $\frac{13}{13+3}$ = 0.813, P = $\frac{correct}{correct + wrong}$

= $\frac{13}{13+3}$ = 0.813 and F-score = $\frac{2*R*P}{R+P}$ = $\frac{2 * 0.813 * 0.813}{0.813+0.813}$ = 0.813. The researcher multiplied by 100% to change to out of a hundred. Again to compute for the three metrics at 30% the researcher followed the same procedure. That is taking the number of common sentences for both manual summary and system summary which is called correct = 13, sentences which are only in the system summary are called missed =4 and sentences which are only in manual summary called wrong = 5. Then

$$R = \frac{correct}{correct + missed} = \frac{13}{13+ 4} = 0.765, \; P = \frac{correct}{correct + wrong} = \frac{13}{13+5} =0.722 \text{ and}$$

$$\text{F-Score} = \frac{2*R*P}{R+P} = \frac{2*0.765*0.722}{0.765+0.722} = 0.743.$$

To compute for 40% extraction rate the researcher did the same work. That is taking the number of common sentences for both manual summary and system summary which is called correct = 15, sentences which are only in the system summary are called missed =4 and sentences which are only in manual summary called wrong = 1. And the three metrics computed as follows:

$$R = \frac{correct}{correct + missed} = \frac{15}{15+ 4} = 0.7895, \; P = \frac{correct}{correct + wrong} = \frac{15}{15+1} = 0.7375 \text{ and}$$

$$\text{F-Score} = \frac{2*R*P}{R+P} = \frac{2 * 0.7895 * 0.7375}{0.7895+0.7375} = 0.8572.$$

This is only when all the features are used for the three extraction rates, and when one of the features either thematic words or title words or capital letters or sentence position or numerical data's or term weight or cue phrases is not used in the generating summary the computation of all the metrics follows in the same way as did in above. So, for single topic the researcher did nine times summary generation and for all the summary computed the three metrics at three extraction rate.

Table 5.1 Afaan Oromo single document news summarization result by size using different sentence extraction rate.

Text ID	News size in words	News size in sentences	Numbers of sentence at extraction rate of		
			20%	30%	40%
Topic 6	299	24	5	7	10
Topic 8	474	41	8	12	16
Topic 10	304	14	3	4	6
Topic 11	937	78	16	23	31
Topic 14	280	19	4	6	8
Topic 19	418	29	6	9	12
Topic 20	354	15	3	5	6
Topic 21	290	11	2	3	4
Topic 22	291	12	2	4	5
Topic 27	1054	83	17	25	33

5.3.1 Experimental Summary Preparation

1. Manual Summary Test Data

This summary test data is prepared by three peoples who were first degree in Afaan Oromo graduated. They prepared manual summary of 10 news articles contained in the data corpus using the guideline indicated in appendix H. An evaluation guideline which is generic based was prepared and given to the three peoples. This guideline helps them to select the sentences mostly related to the features and important to be included in the manual summary. And 20%, 30% and 40% extraction rates are selected because of the summarizer is to summarize up to maximum half of the document and minimum to two line and above. So since among the selected topics, some of them are long and some of they are short documents as it is indicated in table 5.1 these three extraction rates are selected.

Manual summary of all articles in this work is prepared by take out the top ranking sentences based on the average rank of the sentences. The system of taking out the top documents of a given document which is called extraction rate is computed based on the number of sentences available in the document. That is by taking the product of total number of sentences in a document with the corresponding extraction rate and then rounding off the result to the nearest integer by using the rule of rounding digit to its nearest integer. For instance, a document with 23 number of sentences extract at 20%, can be calculated by taking the product of 23 and 20%. Which gives 5, and 5 is number of sentences in the generated summary.

In general, these three peoples prepared summary of 10 articles individually. So, they prepared 30 manual summary for Afaan Oromo single document news based on the three extraction rates given, and using this the reference summary is generated by taking the top n number of sentences with maximum score included in the summary.

2. System Summary Test Data

Summary generated by proposed system (i.e. AAOTS system) is called system summary. And it can be evaluated by peoples selected to evaluate the system (evaluators). Evaluators give judgement using the rule given in appendix I.

In this study, the researcher did 270 experiments on the selected 10 topics for the three extraction rate with all the 9 features. Which means using one topic the researcher did experiments when all features are included in the system, when one feature is absent from the system for all features one by one for all the three extraction rate. And did this for all ten selected topics. All these works are to find the effect of each features when absent from the system comparing to reference summary which is prepared manually with the system summary generated.

5.4 Evaluation and Result Discussion

In this section the performance of Automatic Afaan Oromo Text Summarizer was evaluated. As it is discussed under 2.8, evaluating the summary is difficult because there is no standard rule. But by taking number of sentences common for both manual summary and system summary the performance evaluation metrics (recall, precision and F-measure) computed, which means evaluating system summaries quality. To evaluate objectively and subjectively the linguistic quality of summary generated by developed system the researcher used intrinsic evaluation method as it is discussed under section 1.6.5.

2.9.2.1 Subjective Evaluation

Subjective evaluation is judgments given by human not by system depend on referential clarity, to check as there is any redundancy or not, in-formativeness, grammatical correctness and coherence [10]. To do this it is mandatory to read for evaluator the original document. After they read the original document, they used questioners prepared having the criteria of check point's description to make clear measurement, which is shown on appendix I. The check points were have five different values, valued 1 up to 5. This means 1 is to mean very poor, 2 is to mean poor, 3 is to mean not bad, 4 is to mean good and 5 is to mean very good.

Generally, the researcher used five check points to measure the performance of the summarizer by means of subjective evaluation mechanism are described below. For each experiment conducted the measurement of the summary is computed by the summation of the score given by evaluators for each topic divided by the sum of maximum score [10] and totally this evaluation is from 100%.

$$\sum_{i}^{5} K_i * 100\% \qquad 5.1$$

In this equation, K_i is result scored by evaluators, See table 5.2a, 5.2b and 5.2c below. For example: To make clear the concept of equation 5.1 using the result of Topic 6 of the experiment 1 (when all features are used) at 20% extraction rate shown as:

$(4 + 4 + 3 + 4 + 2) * 100\% = 17/25 * 100\% = 68\%$, so all the results in the tables (Table 5.2a, 5.2b and 5.2c) below computed in this way.

Table 5.2a: Subjective evaluation result at 20%

Text ID	Evaluation grade						
	Grammatical correctness	Non redundancy	Referential Clarity	Informativeness	Coherence & structure	Total score from 25	Total score in %
20%							
Topic 6	4	4	3	4	2	17	68%
Topic 8	5	4	2	3	3	17	68%
Topic 10	5	4	4	5	3	21	84%
Topic 11	4	3	3	3	4	17	68%
Topic 14	5	2	3	4	2	16	64%
Topic 19	4	4	5	3	4	20	80%
Topic 20	4	5	4	3	3	19	76%
Topic 21	5	4	3	3	3	18	72%
Topic 22	4	3	4	5	2	18	72%
Topic 27	5	3	2	4	5	19	76%
Average	90%	72%	66%	74%	62%		72.8%

Table 5.2b: Subjective evaluation result at 30%

Text ID	Evaluation grade						
	Grammatical correctness	Non redundancy	Referencial Clarity	Informativeness	Coherence & structure	Total score from 25	Total scorein %
30%							
Topic 6	4	4	5	4	2	19	76%
Topic 8	5	4	3	4	4	20	80%
Topic 10	4	5	4	5	3	21	84%
Topic 11	5	4	3	4	4	20	80%
Topic 14	5	4	3	4	3	19	76%
Topic 19	4	4	5	3	3	19	76%
Topic 20	4	5	4	3	4	20	80%
Topic 21	5	4	3	3	4	19	76%
Topic 22	4	3	4	5	3	19	76%
Topic 27	5	4	3	4	3	19	76%
Average	90%	82%	74%	78%	66%		78%

Table 5.2c: Subjective evaluation result at 40%

Text ID	Evaluation grade						
	Grammatical correctness	Non redundancy	Referencial Clarity	Informativeness	Coherence & structure	Total SCORE From 25	Total scorein %
40%							
Topic 6	4	5	5	4	3	21	84%
Topic 8	5	4	4	4	3	20	80%
Topic 10	4	4	4	5	3	20	80%
Topic 11	5	4	4	5	4	22	88%
Topic 14	4	3	4	4	4	19	76%
Topic 19	5	4	5	4	4	22	88%
Topic 20	4	5	4	5	4	22	88%
Topic 21	5	4	4	3	4	20	80%
Topic 22	5	4	5	5	4	23	92%
Topic 27	5	5	4	4	3	21	84%
Average	92%	84%	86%	86%	72%		84%%

Generally, the five check points were discussed as follows.

i. Is the generated summary grammatically correct?

This is to check as the generated summary is grammatically correct or as there is any error like fragments of sentence during the process of summarization and it is used to measure the linguistic quality used to evaluate the generated summary. Maximum of the results achieved are correct in terms of grammatical correctness as the techniques the researcher used was extractive text summarization technique. The summary is collected extracted sentences in their rank order.

The performance of the automatic summary in terms of grammatical correctness was checked at three level extraction rate. The first is the average performance of the system at 20% extraction rate concerning grammatical correctness is 90%, at second rate the performance of the system at 30% extraction rate concerning grammatical correctness is 90% and also lastly the performance of the system at 40% extraction rate regarding grammatical correctness is 92%. So, as we see from all the three extraction rates the result is promising concerning grammatically correctness.

ii. Is there redundancy in the summary?

Using this check point the evaluators checked for any repetition of ideas in sentences. It is assumed if possible the summary should contain one sentence to express one idea. No need of representing single idea with two or more than two sentences to precise the same conception with different words. Having summary without redundancy is more readable and interesting for readers.

The evaluators evaluated the summary by checking the redundancy of the sentences with in a summary. So, the result of subjective evaluation shows at 20% extraction rate performance of the summarizer system is 72%, at 30% extraction rate the performance of the summarizer is 82% and at 40% extraction rate the performance of the summarizer is 84%. This result shows that as the extraction rate increases then the performance of the system to reduce the redundance of the summary increases.

iii. Is referential clarity maintained in the summary?

This check point is used to measure the easiness of finding to whom or what the referential pronouns or words used refer to and if the reference is right in sequential sentences. Here the evaluators check that if summary has answer easily for questions like “who? or what?” for the referential words used.

So the performance automatic summarization of subjective evaluation result at 20% extraction rate performs 66%, at 30% extraction rate performs 74% and at 40% extraction rate 86% in concerning referential clarity. Concerning the referential clarity as we see from the result at the highest extraction rate the performance of the system is also high and at the lowest extraction rate the performance of the system is lower. This implies that as extraction rate increases the performance of the system also increases.

iv. Which summary is more coherent and structure?

This check point checks as there is a smooth transition of sentences in summary generated. So, in this part the respondents evaluated as transition of sentence for generated summary is smooth or not. When reading the sentences the summary it must contain a clear information about the topic. Hence, the coherent and structure of the generated summary evaluated by subjective evaluation performed 62% automatic summary at 20% extraction rate, 66% performed at 30% extraction rate and 72% performed at 40% extraction rate. As these results indicate the more coherent and structured sentences are at the highest extraction rate. Whereas at lower extraction rate the sentences are less coherent and structured.

v. To what level is the summary informative?

This check point is used for the evaluator to understand the amount of information present in both the reference summary and the generated summary as common, and to identify the summary that include the most significant information of the text. This measures the information content of the summary as it included the main information that relates to the topic of the article rather than sentences that are written about none related details.

The evaluators evaluated the summary by checking as it contains the whole content of the original document is in a summary. So, the result shows that at 20% extraction rate the performance of automatic summary was 74%, at 30% extraction rate the performance of automatic summary was 78% and at 40% extraction rate the performance of automatic summary was 86%. So these results show that the more informative summary is at the higher extraction rate. Totally the following table shows the performance of the extraction rate at three different rates for the 10 topics the researcher did an experiment on concerning subjective evaluation.

Table 5.2 subjective evaluation result of the summary at three extraction rate

No.	Text ID.	Performance of the summary based on subjective evaluation at 20%, 30% and 40% extraction rate		
		20%	30%	40%
1	Topic 6	68%	76%	84%
2	Topic 8	68%	80%	80%
3	Topic 10	84%	84%	80%
4	Topic 11	68%	80%	88%
5	Topic 14	64%	76%	76%
6	Topic 19	80%	76%	84%
7	Topic 20	76%	80%	88%
8	Topic 21	72%	76%	80%
9	Topic 22	72%	76%	88%
10	Topic 27	76%	76%	84%
Average		72.8%	78%	83.2%

Generally as the above discussions and table shows the subjective evaluation was made using three different extraction rate has promising. The subjective evaluation shows good performance measure at different extraction rate which is a promising of the generated summary. As we see from table 5.2 the average performance of the system at 20% extraction rate is 72.8%, at 30% extraction rate is 78% and at 40% extraction rate is 83.2%. And this shows as the extraction rate increases the performance of the system increases. So the subjective evaluation at last extraction rate which is at 40% shows a promising result.

2.9.2.2 Objective Evaluation

This is used to measure the performance, quality and effectiveness of the summarizer or the summarization system. Intrinsic evaluation method is used in this work to measure system summary quality based on the communication between manually generated summary and system generated summary which are measured by metrics of precision, recall and F-score [29]. So these three functions were computed using equations discussed under chapter two (equation 2.12, 2.13 and 2.14) of this work. In order to see the effect of all the 8 features on the system the researcher did the following experiments and also to answer the research question. After a lot of effort the result for all experiments discussed as follows:

- **Experiment 1**: Examining the performance of AAOTS when all features are used.

Experiment 1 is to talk over the performance of AAOTS system using all the features to extract sentence. So, the experiment did using all the features when all the features are used in the experiment, the average f-measure result of the system at 20% is 73%, at 30% is 80.4% and at 40% is 85.6% as it is shown in table 5.3. These results indicate that as the extraction rate increases the performance of the system increases.

Table 5.3: Experimental result when all features are used

Topic ID	P			R			F-measure		
	20%	30%	40%	20%	30%	40%	20%	30%	40%
Topic 6	81.3%	72.2%	82%	81.3%	76.5%	83%	81.3%	74.3%	82.5%
Topic 8	60%	85.7%	93.8%	75%	85.7%	79%	66.7%	85.7%	85.7%
Topic 10	83.3%	77.8%	83.8%	83.3%	77.8%	86.3%	83.3%	77.8%	90.3%
Topic 11	55.6%	75%	83.9%	55.6%	69.2%	74.3%	55.7%	72%	78.8%
Topic 14	83.3%	85.7%	88.9%	83.3%	85.7%	88.9%	83.3%	85.7%	88.9%
Topic 19	57.1%	54.6%	73.3%	44.4%	50%	64.7%	58%	52.2%	68.8%
Topic 20	33.3%	66.7%	80%	33.3%	80%	66.7%	33.8%	72.7%	72.7%
Topic 21	100%	100%	100%	100%	100%	100%	100%	100%	100%

Topic 22	100%	100%	100%	100%	100%	100%	100%	100%	100%
Topic 27	68.8%	84%	88.3%	68.8%	84%	88.3%	68.8%	84%	88.2%
Average	72.3%	80.2%	87.4%	72.5%	80.9%	83.1%	73.0%	80.4%	85.6%

- **Experiment 2:** Examining the performance of AAOTS excluding word frequency:

Experiment 2 is to talk over the effect of not using word frequency in the system. The performance of AAOTS system without using word frequency is described by f-measure out performance at 20% extraction rate is 71.7%, at 30% extraction rate is 81.3% and at 40% extraction rate is 86.1%. These results show that absence of the word frequency does not affect the system.

Table 5.4: Experimental result when word frequency feature is not used

Topic ID	P			R			F-measure		
	20%	30%	40%	20%	30%	40%	20%	30%	40%
Topic 6	78.6%	72.2%	93.8%	78.6%	76.5%	75%	78.6%	74.3%	83.3%
Topic 8	60%	85.7	93.8%	54.6%	85.7%	79%	53.5%	85.7%	85.7%
Topic 10	83.3%	88.9%	89.3%	83.3%	88.9%	89.3%	83.3%	88.9%	89.3%
Topic 11	55.6%	75%	83.9%	52.6%	69.2%	74.3%	54.1%	72%	78.8%
Topic 14	83.3%	85.7%	88.9%	83.3%	85.7%	88.9%	83.3%	85.7%	88.9%
Topic 19	50%	54.6%	84.6%	50%	50%	68.8%	50%	52.2%	75.9%
Topic 20	33.3%	66.7%	66.7%	33.3%	80%	80%	33.3%	72.7%	72.73%
Topic 21	100%	100%	100%	100%	100%	100%	100%	100%	100%
Topic 22	100%	100%	100%	100%	100%	100%	100%	100%	100%
Topic 27	68.6%	83%	87.5%	68.8%	80%	84.9%	68.8%	81.6%	86.6%
Average	71.3%	93.0%	88.8%	70.5%	81.6%	84%	71.7%	81.3%	86.1%

- **Experiment 3:** Examining the performance of AAOTS without using thematic words in the system

Experiment 3 is to talk over the effect of absence of using thematic words from the AAOTS system. So, the system performs without using thematic words is measured by f-measure and f-measure at 20% extraction rate is 68.2%, at 30% extraction rate is 77.5% and at 40% extraction rate is 75.6%. These results show that the absence of thematic words feature affect the system by the amount 10% when we compare with when all features are used.

Table 5.5: Experimental result when thematic words feature is not used

Topic ID	P			R			F-measure		
	20%	30%	40%	20%	30%	40%	20%	30%	40%
Topic 6	70%	86.7%	88.2%	63.6%	76.6%	75%	66.7%	81.3%	81.1%
Topic 8	60%	71.4%	81.3%	75%	76.9%	76.5%	66.7%	74.1%	78.8%
Topic 10	71.4%	77.8%	75.4%	71.4%	77.8%	75.4%	71.4%	77.8%	75.4%
Topic 11	61.1%	82.6%	87.1%	61.1%	79.2%	77.1%	61.1%	80.9%	81.8%
Topic 14	66.7%	71.4%	77.8%	66.7%	71.4%	77.8%	66.7%	71.4%	77.8%
Topic 19	57.1%	36.4%	75%	44.4%	33.3%	5%	50%	34.8%	59.7%
Topic 20	66.7%	80%	66.7%	66.7%	80%	66.7%	66.7%	80%	66.7%
Topic 21	50%	100%	75%	50%	100%	75%	50%	100%	75%
Topic 22	100%	100%	80%	100%	100%	80%	100%	100%	80%
Topic 27	82.4%	75%	78.8%	82.4%	75%	81.3%	82.4%	75%	80%
Average	68.5%	78.1%	78.5%	68.1%	77.0%	69%	68.2%	77.5%	75.6%

- **Experiment** 4: Examining the performance of AAOTS Without title words.

Experiment 4 is to talk over the role of the title words found in the document has on the system. So, the system without using the title words for extracting the sentence is measured f-measure at 20% extraction rate is 72.2%, at 30% extraction rate is 83.3% and at 40% extraction rate is 84.2%. The result of absence of title words also show that as the performance of the system is affected by the amount of 1.4%.

Table 5.6: Experimental result when title words feature is not used

Topic ID	P			R			F-measure		
	20%	30%	40%	20%	30%	40%	20%	30%	40%
Topic 6	81.3%	92.9%	93.8%	76.5%	76.5%	75%	78.8%	83.9%	83.3%
Topic 8	60%	85.7%	87.5%	66.7%	85.7%	77.8%	63.2%	85.7%	82.4%
Topic 10	66.7%	77.8%	84.5%	66.7%	77.8%	84.5%	66.7%	77.8%	84.5%
Topic 11	66.7%	82.6%	86.2%	63.2%	79.2%	78.1%	64.9%	80.9%	82%
Topic 14	83.3%	85.7%	77.8%	83.3%	85.7%	77.8%	83.3%	85.7%	77.8%
Topic 19	50%	81.8%	87.5%	50%	75%	93.3%	50%	78.3%	90.3%
Topic 20	100%	80%	83.3%	100%	80%	83.3%	100%	80%	83.3%
Topic 21	50%	100%	100%	50%	100%	100%	50%	100%	100%
Topic 22	100%	75%	80%	100%	75%	80%	100%	75%	80%
Topic 27	64.7%	84%	78.1%	64.7%	87.5%	78.1%	64.7%	85.7%	78.1%
Average	72.3%	84.6%	85.9%	72.1%	82.2%	82.8%	72.2%	83.3%	84.2%

- **Experiment 5**: Examining the performance of AAOTS without using capital letter

Experiment 5 is to talk over whether presence or absence of capital letters within the system has any effect or not. Hence, absence of capital letter from the system or without using the capital letters the performance of the system is measured by f-measure at 20% extraction rate is 66.9%, at 30% extraction rate is 80%% and at 40% extraction rate is 85.7% as shown in table 5.7. Absence of capital letter feature from the system doesn't affect the performance of the system.

Table 5.7: Experimental result when capital letter feature is not used

Topic ID	P			R			F-measure		
	20%	30%	40%	20%	30%	40%	20%	30%	40%
Topic 6	81.3%	92.9%	93.8%	72.2%	76.5%	75%	76.5%	83.9%	83.3%
Topic 8	60%	85.7%	93.8%	60%	85.7%	79%	60%	85.7%	85.7%
Topic 10	66.7%	77.8%	82.6%	66.7%	77.8%	82.6%	66.7%	77.8%	82.6%
Topic 11	55.6%	72%	81%	52.6%	66.7%	77.1%	54.1%	69.2%	79.0%
Topic 14	83.3%	71.4%	88.9%	83.3%	71.4%	88.9%	83.3%	71.4%	88.9%
Topic 19	50%	45.5%	85.7%	22.2%	50%	80%	30.8%	47.6%	82.8%
Topic 20	33.3%	80%	66.7%	33.3%	80%	66.7%	33.3%	80%	66.7%
Topic 21	100%	100%	100%	100%	100%	100%	100%	100%	100%
Topic 22	100%	100%	100%	100%	100%	100%	100%	100%	100%
Topic 27	64.7%	84%	87.9%	64.7%	84%	87.9%	64.7%	84%	87.9%
Average	69.5%	80.9%	88.1%	53.5%	79.2%	83.7%	66.9%	80%	85.7%

- **Experiment 6**: Examining the performance of AAOTS Without sentence position.

Experiment 6 is to talk over the effect of absence of the sentence position from the system. So, the performance of the system in absence of sentence position measured by f-measure at 20% extraction rate is 44.3%, at 30% extraction rate 64.1% and at 40% extraction rate 64.7% as shown in table 5.8. Absence of sentence position from the system highly affected the performance of the system by the amount 20.9%.

Table 5.8: Experimental result when sentence position feature is not used

Topic ID	P			R			F-measure		
	20%	30%	40%	20%	30%	40%	20%	30%	40%
Topic 6	75%	93.3%	93.8%	70.6%	77.8%	75%	72.7%	84.9%	83.3%
Topic 8	40%	42.9%	43.8%	40%	40%	41.2%	40%	41.4%	42.4%
Topic 10	50%	66.7%	58.3%	50%	66.7%	58.33%	50%	66.7%	58.3%
Topic 11	33.3%	75%	54.2%	31.8%	69.2%	50.5%	32.6%	72%	52.4%
Topic 14	50%	57.1%	66.7%	50%	57.1%	66.7%	50%	57.1%	66.7%
Topic 19	20%	40%	50%	16.7%	30.8%	50%	18.2%	34.8%	50%
Topic 20	33.3%	60%	50%	33.3%	60%	50%	33.3%	60%	50%
Topic 21	50%	100%	75%	50%	100%	75%	50%	100%	75%
Topic 22	50%	75%	100%	50%	75%	100%	50%	75%	100%

Topic 27	43.8%	46.4%	70%	50%	52.3%	67.7%	46.7%	49.3%	68.9%
Average	44.5%	53.6%	66.2%	44.2%	58.1%	63.4%	44.3%	64.1%	64.7%

- **Experiment 7**: Examining the performance of AAOTS Without using numbers

Experiment 7 is to talk over the presence or absence of the module of name of numbers and numerical data found in document from the system. So, the performance of the system in absence of number measured by f-measure at 20% extraction rate is 68.6%, at 30% extraction rate 80.7% and at 40% extraction rate 84.2% as shown in table 5.9. So this result shows that absence of the numerical data feature affect th performance of the system.

Table 5.9: Experimental result when number feature is not used

Topic ID	P			R			F-measure		
	20%	30%	40%	20%	30%	40%	20%	30%	40%
Topic 6	77.8%	93.3%	90%	87.5%	75%	82.3%	82.3%	83.%	86%
Topic 8	60%	85.7%	87.5%	60%	85.7%	73.7%	60%	85.7%	80%
Topic 10	83.3%	77.8%	80.5%	83.3%	77.8%	80.5%	83.3%	77.8%	80.5%
Topic 11	47.4%	75%	81%	50%	69.2%	77.1%	48.5%	72%	79.0%
Topic 14	66.7%	57.1%	77.8%	66.7%	57.1%	77.8%	66.7%	57.1%	77.8%
Topic 19	50%	72.7%	85.7%	44.4%	63.2%	80%	47.1%	67.6%	82.8%
Topic 20	33.3%	80%	66.7%	33.3%	80%	66.7%	33.3%	80%	66.7%
Topic 21	100%	100%	100%	100%	100%	100%	100%	100%	100%
Topic 22	100%	100%	100%	100%	100%	100%	100%	100%	100%
Topic 27	64.7%	84%	90.9%	64.7%	84%	88.2%	64.7%	84%	89.6%
Average	68.3%	82.6%	86.0%	69.0%	79.2%	82.6%	68.6%	80.7%	84.2%

- **Experiment 8**: Investigate the performance of AAOTS Without using Term Weight

Experiment 8 is to talk over the presence or absence of the module of term weight from the system. So, the performance of the system in absence of term weight measured by f-measure at 20% extraction rate is 68.1%, at 30% extraction rate is 77.8% and at 40% extraction rate is 86.0% as shown in table 5.10.

Table 5.10: Experimental result when term weight feature is not used

Topic ID	P			R			F-measure		
	20%	30%	40%	20%	30%	40%	20%	30%	40%
Topic 6	77.8%	92.3%	87.4%	87.5%	75%	83.3%	82.4%	82.3%	85.3%
Topic 8	60%	85.7%	93.8%	60%	85.7%	79%	60%	85.7%	85.7%
Topic 10	66.7%	57.1%	88.9%	66.7%	57.1%	88.9%	66.7%	57.1%	88.9%
Topic 11	52.6%	75%	81%	52.6%	69.2%	77.1%	52.6%	72%	79.0%
Topic 14	66.7%	57.1%	88.9%	66.7%	57.1%	88.9%	66.7%	57.1%	88.9%
Topic 19	44.4%	63.6%	80%	44.4%	58.3%	75%	44.5%	60.9%	77.4%
Topic 20	33.3%	80%	66.7%	33.3%	80%	66.7%	33.3%	80%	66.7%

Topic 21	100%	100%	100%	100%	100%	100%	100%	100%	100%
Topic 22	100%	100%	100%	100%	100%	100%	100%	100%	100%
Topic 27	75%	84%	87.9%	75%	80.8%	87.9%	75%	82.4%	88.4%
Average	67.5%	79.5%	87.4%	68.6%	76.3%	84.7%	68.1%	77.8%	86.0%

- **Experiment** 9: Examining the performance of AAOTS without using cue phrase:

Experiment 9 is to talk over the presence or absence of the module of term weight from the system. So, the performance of the system in absence of cue phrase measured by f-measure at 20% extraction rate is 66.9%, at 30% extraction rate is 76.4% and at 40% extraction rate is 85.3% as shown in table 5.11.

Table 5.11: Experimental result when term cue phrase is not used

Topic ID	P			R			F-measure		
	20%	30%	40%	20%	30%	40%	20%	30%	40%
Topic 6	73.3%	92.3%	83%	84.6%	75%	89%	78.6%	82.3%	86%
Topic 8	55.6%	85.7%	93.8%	55.6%	85.7%	79%	55.6%	85.7%	85.7%
Topic 10	66.7%	82.4%	85%	66.7%	82.4%	85%	66.7%	82.4%	85%
Topic 11	52.6%	70.8%	81%	52.6%	68%	77.1%	52.6%	69.3%	79.0%
Topic 14	66.7%	57.1%	88.9%	66.7%	57.1%	88.9%	66.7%	57.1%	88.9%
Topic 19	44.4%	50%	80%	44.4%	45.5%	75%	44.4%	47.6%	77.4%
Topic 20	33.3%	80%	88.3%	33.3%	80%	88.3%	33.3%	80%	88.3%
Topic 21	100%	100%	100%	100%	100%	100%	100%	100%	100%
Topic 22	100%	75%	75%	100%	75%	75%	100%	75%	75%
Topic 27	70.6%	84%	87.9%	70.6%	84%	87.9%	70.6%	84%	87.9%
Average	66.3%	77.7%	86.3%	67.5%	75.3%	84.5%	66.9%	76.4%	85.3%

Objective Evaluation Result and Analysis

As we can see from all tables above (table 5.3 up to table 5.11) of the experiment part the summarizer summaries were evaluated by comparing the manual summary and system summary using f-measure. As we see from the experiment results on the tables all the features has their own impact on the system even if it is different. Here as we can see absence of sentence position feature has more impact than other features. Because in absence of sentence position the system performance is significantly lower than in absence of other features from the system. This indicates sentence position has more impact than other features on the system. In other way, we can see that at the rate of expression increases the performance of the system increases.

Chapter Six

Conclusion, Contribution and Recommendation

6.1 Conclusion

This thesis had investigated the capability of the system developed for Afaan Oromo news text summarizer by NLTK using python for single document. So, the researcher tried to develop a prototype of extractive summarization system for text written based on sentence score. For this purpose the researcher developed the system called AAOTS to compute the significant sentence score and extract the top ranked sentences in the document. The researcher used 8 features like: words frequency, thematic words, title words, capital letters, sentence position, name of numbers, term weights and cue phrases. The researcher did experiments for all these features to see the impact of these features on the developed system. And put the conclusion as following.

- Performance of AATOS is without using the feature sentence position is low.
- Performance of the system without using the thematic words feature has no as such impact on the system.
- As the number of extraction rate increases the performance of the AATOS system increases.

And on the other hand the research questions are answered as follows:

- To what extent the features added in this study, affect the performance and quality of the summarizer?

For most of the features the performance of the system is almost the same, but for some the features the performance of the system is different for instance when sentence position is not used in the system the f-measure of the system at 20% extraction rate is 44.30%, at 30% extraction rate is 64.10% and at 40% extraction rate is 64.70%. But when all features are used the f-measure at 20% extraction rate is 70.03%, at 30% extraction rate is 80.44% and at 40% extraction rate is 85.59%. The difference between the performances of the system when all features used and when the feature sentence position is not used in the system at 20% extraction rate is 25.73% (70.03% - 44.30%), at 30% extraction rate 16.34% (80.44% - 64.1%) and at 40% extraction rate is 20.89% (85.59% -64.7%). And when we consider the added features in this work, when the thematic words is not used in the system the performance of the system decreases. That is performance of the system at 20% extraction rate is 68.2%, at 30% extraction rate is 77.5% and at 40% extraction rate is 75.6%. When we see its difference with when all features are used at 20% it is 4.8%, at 30% it is 2.9% and at 40% it is 10%. So the absence of thematic words affect the system this much. And when we see the other added feature in this work i.e absence of title words from the system, the system performance at 20% is 72.2%, at 30% is 83.3% and at 40% is 84.2%. Its difference with when all the features used is at 20% extraction rate 0.8%, at 30% extraction rate is 1.1% and at 40% extraction rate is 1.4%. This implies sentence position has great impact on the system. The new features added thematic words and title words are the second and the third ranked features that affect the system. Whereas term weight and capital letter features didn't affect the system as we see from table 5.12. Based on the result of the experiment when sentence position, thematic words and title words are incorporated to the system the performance of the system increases.

- Which feature is more important and less important, to contribute to the performance of the summarizer?

Using the experiment results sentence position, thematic words, title words and numerical data are the most important features because absence of these features affect the performance of the system, whereas cue phrases, capital letter, term weight and word frequencies are less important features because absence of these feature does not decrease the performance of the system.

6.2 Contribution

This study has contributed a paper contribution in addition to extraction based text summarization approach has been attempted to be used for summarizing Afaan Oromo single document news text.

6.3 Recommendation

The researcher has did Afaan Oromo single document summarization using extraction method. However, additional work or investigation is needed. Applying additional features to make complete the work more. Whatever the result of the study is interesting the researcher encountered with some like unable to get a well prepared Afaan Oromo news for the experimentation and tools for this languages are also not investigated.

To make more functional and suitable for the user the summarizer with better performance than this work additional features and future research directions are listed below:

- For scoring and extracting important sentence using features like: term weight, word frequency, capital letter, Thematic words, name of numbers, cue phrase and sentence position methods have been used in this study but by adding addition statistical features the summarizer's performance can be upgraded.
- It is important to make a convincing performance assessment for further studies, to develop or create standardized and well prepared Afaan Oromo text corpus.
- During summary evaluation objectively, there is a difficulty to compute the three metrics i.e precision, recall and f-measure. Thus, a tool that can automatically evaluate the summary is recommended.
- Since there is no complete stop-word list, synonyms, and abbreviations the researcher recommend that as they are very useful to enhance the extractive summarization method.

References

[1] Sisay A. M., “Information Extraction Model from Afaan Oromo News Texts.” M.Sc. Thesis, Bahir Dar University, Ethiopia, 2017.

[2] L. Douglas Bakerti, Andrew Kachites Mc.” Distributional Clustering of Words for Text Classification.” Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrievalAugust, 1998, Pages 96–103.

[3] Mekonen H. “Lexical standardization in Oromoo.” M.Sc. Thesis, Addis Ababa University, Ethiopia, 2002.

[4] Etnologue. Show Language (Online edition ed.). Available: http://www.ethnologue.com/web.asp, accessed on 15 Jan. 2020.

[5] Getachew Mamo, “Part-Of-Speech Tagging for Afaan Oromo Language.” M.Sc. Thesis, Addis Ababa University, Ethiopia, 2009.

[6] R. Kumar and K. Raghuveer. “Legal Document Summarization using Latent Dirichlet Allocation.” *International Journal of Computer Science and Telecommunications,* Volume 3, Issue 7, July 2012.

[7] Debela T.,” Designing a Stemmer for Afaan Oromo Text: A Hybrid Approach.” M.S. thesis, Addis Ababa University, Ethiopia, June, 2010.

[8] H. Han, “Automating TL; DR: An Approach to Text Summarization,“ May 16, 2019, volum 1. Available: www.haomiaohan.com

[9] M.A. Fattah and F. Ren, “Automatic Text Summarization.” *In World Academy of Science Engineering and Technology*, vol. 27, pp.192-195, 2008.

[10] Fiseha B., “Afaan Oromoo Automatic News Text Summarizer Based on Sentence Selection Function.” M.Sc. thesis, Addis Ababa University, Ethiopia, 2013.

[11] Girma D., "Afaan Oromoo news text summarizer." M.Sc. thesis, Addis Ababa University, Ethiopia, 2012.

[12] K. Kaikhah. “Automatic Text Summarization with Neural Networks,” Second International IEEE Conference, Volume: 1, 2004.

[13] Pauliina A. “Automatic Text Summarization.” M.Sc. Thesis, University of Turku, Pauliina Anttila, 2018.

[14] Vishwa P. and Nasseh T. “Automatic Text Summarization: A Systematic Review.” East Carolina University, Greenville NC 27858, USA, 2019.

[15] Guesh A. B.”Automatic Text Summarizer for Tigrinya Language.” M.Sc. Thesis, Addis Ababa University, Ethiopia, 2017.

[16] Eduard Hovy and Chin-Yew Lin. Automated Text Summarization and the Summarist System. Southern California: Association for Computational Linguistics, 1998, pp. 197–214.

[17] Mattias Gessesse A.”Efficient Language Independent Text Summarization Using Graph Based Approach.” M.Sc. Thesis, Addis Ababa University, Ethiopia, 2015.

[18] Eyob Delele Y. “Topic-based Amharic Text Summarization.” M.Sc. Thesis, Addis Ababa University, Ethiopia, 2011.

[19] V. Gupta and G. S. Lehal. “A Survey of Text Summarization Extractive Techniques”. *Journal of Emerging Technologies In Web Intelligence*, Vol. 2, No. 3, pp. 258-268, August 2016.

[20] H. P. Edmundson.,” New methods in automatic extracting”, *Journal of the ACM*, volume 16(2), pp.264-285, April 1969.

[21] Melese Tamiru. “Automatic Amharic Text Summarization Using Latent Semantic Analysis.” M.Sc. Thesis, Addis Ababa University, Ethiopia, 2009.

[22] Addis Ashagre T., “Automatic Summarization for Amharic Text Using Open Text Summarizer,” M.Sc. Thesis, Addis Ababa University, Ethiopia, 2013.

[23] Sumya Akter, Aysa Siddika Asa, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy and Masud Ibn Afjal “An Extractive Text Summarization Technique for Bengali Document(s) using K-means Clustering Algorithm,” 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Dhaka, Bangladesh, January 2017.

[24] M. Gambhir and V. Gupta, “Recent automatic text summarization techniques: a survey,” Artif Intell Rev, 2016, pp. 1-66, DOI: 10.1007/s10462-016-9475-9.

[25] Chin-Yew Lin, “Assembly of Topic Extraction Modules in SUMMARIST,” *Information Science Institute/USC4676 Admiralty Way*, Marina del Rey, CA 90292, USA, 1998.

[26] Mohammed A. H.”Modeling an Automatic Amharic Text Summarizer: Abstractive Approach.” M.Sc. Thesis, Addis Ababa University, Ethiopia, 2016.

[27] A. Jain, D. Bhatia, M. K Thakur, “Extractive Text Summarization using Word Vector Embedding,” *In Proc. IEEE International Conference on Machine learning and Data Science*, 2017.

[28] Qaroush, Farha, Ghanem, Washaha, Maali "An efficient single document Arabic text summarization using a combination of statistical and semantic features," *Journal of King Saud University – Computer and Information Sciences*, 2019.

[29] Asefa Bayisa K. "Query-based Automatic Summarizer for Afaan Oromo Text." M.Sc. Thesis, Addis Ababa University, Ethiopia, 2015.

[30] Visser, Wieling" Sentence-based Summarization of Scientific Documents," [Available at] "http://martijnwieling.nl/files/wielingvisser05automaticsummarization.pdf." 2017.

[31] T. Sri Rama Raju, Bhargav Allarpu," Text Summarization using Sentence Scoring Method," International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 04 Apr -2017.

[32] J. S. Kallimani, K. G. Srinivasa, Eswara R. B." Summarizing News Paper Articles: Experiments with Ontology-Based, Customized, Extractive Text Summary and Word Scoring," *Cybernetics and Information Technologies*, Volume 12, No 2, 2015.

[33] Mudasir Mohd, Rafiya Jan, Muzaffar Shah," Text Document Summarization using Word Embedding," Journal Pre-proof, 2019.

[34] J. Guadalupe Ramos, Isela Navarro-Alatorre, Georgina Flores Becerra, Omar Flores-Sánchez," A Formal Technique for Text Summarization from Web Pages by using Latent Semantic Analysis," 2019.

[35] Soe Soe Lwin, Khin Thandar Nwet," Extractive Summarization for Myanmar Language," University of Computer Studies, Yangon Yangon, Myanmar, 2018.

[36] Tsehay Diges D., "Automatic Amharic Multi document News Text Summarization using Open Text Summarizer", M.Sc. Thesis, University Of Gondar, Ethiopia, 2018.

[37] Martin Hassel, Evaluation of Automatic Text summarization, 2004.

[38] V. A. Yatsko and T. N. Vishnyakov, "A Method for Evaluating Modern Systems of Automatic Text Summarization," *SSN 0005-1535, Automatic Documentation and Mathematical Linguistics*, Vol. 41, No. 3, pp. 93–163, 2007.

[39] Yihong Gong," Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis", https://doi.org/16.1145/383952.383955.

[40] Daniel S. Alemu," Instructional Language Policy in Ethiopia: Motivated by Politics or the Educational needs of Children?," Vol. 37, No. 3&4, 2006, pp. 151–168.

[41] C Slamet, A R Atmadja, D S Maylawati, R S Lestari, W Darmalaksana and M A Ramdhani" Automated Text Summarization for Indonesian Article Using Vector Space

Model," *Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sunan Gunung Djati, Bandung, C Slamet et* al IOP Conf. Ser.: Mater. Sci. Eng. 288 012037, 2018.

[42] Hamzah Noori Fejer, Nazlia Omar" Automatic Arabic Text Summarization Using Clustering and Keyphrase Extraction," *International Conference on Information Technology and Multimedia (ICIMU),* Putrajaya, Malaysia, November 18 – 20, 2014.

[43] J. Rojas Sim´on, Yulia Ledeneva and Ren´e Arnulfo Garc´ıa-Hern´andez" Calculating the significance of automatic extractive text summarization using a genetic algorithm," *Journal of Intelligent & Fuzzy Systems* xx (2018), [available at DOI: 16.3233/JIFS-169588].

[44] Narendra Andhale, L.A. Bewoor" An Overview of Text Summarization Techniques," *Department of Computer Engineering, Vishwakarma Institute of Information TechnologyPune, India, Conference Paper,* · August 2016. [Available at DOI: 16.1169/ICCUBEA.2016.7860024]

[45] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh,"A Comprehensive Survey on Text Summarization Systems", Islamic Azad University of Mashhad, 2009.

[46] V. Gupta, G. S. Lehal "Automatic Text Summarization System for Punjabi Language", *Journal of Emerging Technologies in Web Intelligence,* Vol. 5, No. 3, August 2013.

[47] H. Gregory Silber, Kathleen F. McCoy" Efficient Text Summarization Using Lexical Chains", Computer and Information Sciences University of Delaware, International Journal of Research in Engineering and …, 2013.

[48] Makbule Gulcin Ozsoy, Ilyas Cicekli, Ferda Nur Alpaslan" Text Summarization of Turkish Texts using Latent Semantic Analysis," *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2016),* pages 869–876,Beijing, August 2016.

[49] https://qz.com/africa/1812085/ethiopia-adds-afan-oromo-somali-afar-tigrigna-languages-to-amharic/

[50] Sitti Rabiah" Language as a Tool for Communication and Cultural reality Discloser", Universitas Muslim Indonesia, Makassar, 2018.

[51] https://www.ethnologue.com/ethnoblog/gary-simons/welcome-23rd-edition

[52] https://ethiopiangazette.com/most-spoken-languages-in-ethiopia-2019

[53] Gary Miner, John Elder IV, and Thomas Hill. Practical text mining and statistical analysis for non-structured text data applications. Academic Press, 2012.

[54] S.A.Babara, Pallavi D.P.," Improving Performance of Text Summarization", *International Conference on Information and Communication Technologies (ICICT 2014), Procedia Computer Science 46 (2015) 354 – 363*. [Available online at www.sciencedirect.com.]

[55] Dursa Abduselam M." Constructing Sentiment Mining Model for Opinionated Afaan Oromo Texts on Ethiopian Politics." Department of Computer Science, M.Sc. Thesis, Gonder University, Ethiopia, June, 2017.

[56] Dejene Hundessa." Definition Question Answering System for Afaan Oromo Language." M.Sc. Thesis, Addis Ababa University, Ethiopia, October, 2015.

[57] G. Gutema, "Afaan Oromo Text Retrieval System." M.Sc. Addis Ababa University, Ethiopia, 2012.

[58] https://blogs.sas.com/content/iml/2020/02/10/fractional-part-of-a-number-sas.html

[59] https://www.cimt.org.uk/projects/mepres/allgcse/bkb11.pdf

[60] https://www.scribd.com/document/437350093/4H28000.pdf

[61] L Suanrnali, N Salim, M S Binwahlan" Automatic Text Summarization Using Feature Based Fuzzy Extraction", January 2008. [Available at: https://www.researchgate.net/publication/44161221]

[62] Chaltu Fita E." Afaan Oromo List, Definition and Description Question Answering System," M.Sc. Thesis, Addis Ababa University, April 14, 2016.

[63] http://ceur-ws.org/Vol-1173/CLEF2007wn-adhoc-KekebaTuneEt2007.pdf

[64] R Ferreira, L S Cabral, R D Lins, G F Silva, F Freitas, G D.C. Cavalcanti, R. Lima, S J. Simske, L Favaro" Assessing sentence scoring techniques for extractive text summarization ", *Expert Systems with Applications*, 10 May 2013.

[65] Ms.Pallavi D.Patil, Prof.N.J.Kulkarni" Text Summarization Using Fuzzy Logic," *International Journal of Innovative Research in Advanced Engineering (IJIRAE)* Volume 1 Issue 3 (May 2014) SPECIAL ISSUE.

[66] Rucha S. Dixit, Prof. Dr.S.S.Apte" Improvement of Text Summarization using Fuzzy Logic Based Method," *IOSR Journal of Computer Engineering (IOSRJCE) ISSN*: 2278-0661, ISBN:

[67] https://www.angelfire.com/art/danaesl/letter.html

[68] Ali Al Dahoud" Enhanced Feature-Based Automatic Text Summarization System Using Supervised Technique", *International Journal of Computers & Technology* April 2016.

[69] N.Kannaiya Raja, Naol Bakala, S. Suresh"NLP: Text Summarization by Frequency and Sentence Position Methods," *International Journal of Recent Technology & Engineering (IJRTE) ISSN: 2277-3878*, Volume-8 Issue-3, September 2019.

[70] Yan Du and Hua Huo" News Text Summarization Based On Multi-Feature and Fuzzy Logic", Received May 29, 2020, accepted June 10, 2020, date of publication July 14, 2020, date of current version August 11, 2020.

[71] Dipanjan DasAndr´e F.T. Martins" A Survey on Automatic Text Summarization", December 2007.

[72] Neelima Bhatia and Arunima Jaiswal" Literature Review on Automatic Text Summarization: Single and Multiple Summarizations", nternational Journal of Computer Applications (0975 –8887) Volume 117 –No. 6, May 2015.

[73] http://pfliu.com/Historiography/summarization/summ-eng.html

[74] M Kutlu, C Cığır and I Cicekli "Generic Text Summarization for Turkish", Department of Computer Engineering, Bilkent University, Ankara, Turkey 2009.

[75] Ily Amalina Ahmad Sabri, Mustafa Man" Improving Performance of DOM in Semi-structured Data Extraction Using WEIDJ Model", Indonesian Journal of Electrical Engineering and Computer Science, Vol.9, No.3, March2018, pp. 752~763. https://uom.lk/sites/default/files/staff_blog/dilum.bandara/files/Thesis-Nalinda-Herath.pdf.

List of Appendixes

Appendix A: The corpus of Afaan Oromo cue phrases

English	Afaan Oromo meaning
above all	hunda caalaa, hundaarra, hunda dura,hundaan olitti
accordingly	haaluma kanaan,kanaafuu haaluma sanaan, sana waliin
actually	dhugumaan, qabatamaan
admittedly	shakkii malee haqaaf, walirraa fuudhuu, amanuu
after	duuba,maayii, booda
after all	kanaan booda, sirumaa, hunda caalaa,hundaa ol, hunda booda, dhumarratti
after that	achihiin booda, sana booda
afterwards	isa booda, itti aansee, boodarra
again	Ammas, ammoo, itti dabalees, lammeessa,
all in all	dhibbaa dhibbatti, guutummaatti, guutuun, guutummaan guutuutti, dimshaashummatti
all the same	hata’u malee, yeroo hunda, unduu walfakkaata, gama hundaanuu
also	dabalataan, cinaan, innis, sunis, kunis
alternatively	akka filanootti, karaa biraa, gama biraatiin, carraa biroo, filannoof, yookaan
Although	ta'uu illee, ta’us
always assuming that	yeroo hundaa yoo akkasitti yaadame, yeroo mara yaaduun,
and	Fi
and/or	fi, ykn, yookiin, kana ykn sana,
another time	yeroo kan biraa, booda, yeroo biroo, maa'essa
anyway	hata’u malee, karaa kamuu, yaa ta’u, ta'us, sanas ta'e kana
apart from that	sana malee, isaa as, kana malees, sanas, kanaan alatti
as	haala, akka, sababa kanaaf
as a consequence	akka itti aanutti,kanaafuu, bu'aa, kana irraa kan ka’e
as a corollary	bu’aa, dhiibbaa
as a result	kana irraa kan ka'e, kanaafuu
as it happened	akka tasaa,akka ta’etti, osoo hin yaadin
as it is	akka jirutti
as it turned out	osoo hin yaadin
as long as	sun yoo ta’e, yoo
as luck would have it	akka tasaa, akka carraa
as obviously	Ifatti
as soon as	amma, booda, bakkumatti, akkuma sanaan, yeroo
as well	dabalataan, wajjin, dabalatamaan, waliin, akkasuma
at any rate	daddafina kamiinu, karaa kamuu, haala kamiinuu
at first	Dursa, jalqabaaf,yeroo duraa, jalqabarratti, tokkoffaarratti
at first sight	ilaalcha duraarratti,irra keessa, mil'uu duraan, osoo gadi hin fagaatin,
at first view	Mildhannaa, ilaalcha jalqabaarratti, millannaa dura,
at last	dhumarratti
at least	yoo xiqqaatee xiqqaate, yoo xiqqaate
at once	si'a tokkotti, al tokkotti, yeroo tokkoon

at that time	yeroo sanatti
at the moment	yeroo kanatti, amma , yeroos,yammus
at the outset	Jalqabarratti
at the same time	yeroo tokkotti, walfaana
at which point	bakka kamitti, sababa kamiif
back	teella, duuba
because	Sababa
before	dura, dursee
before then	achiin dura, san duraa
before long	osoo hin turin,dhiheenyaan, yeroo xiqqoo booda
before now	amman dura, kana dura ykn yeroo darbe
before ever	kamiyyuu dura, hunda dura
besides	dabalatamaan, bira,cinaa, itti dabalees
but	garuu
but then	yaggus garuu, yeroos garuu
by all means	dhugumatti, sirriidhumatti, yaalii huundaan, karaalee hundaa, gama kamiinu
by and by	osoo hin turin, yerootti, ammas ammas, yeroodhaa yerootti
by comparison wal	dorgomsiisuun, wal cinaa qabuudhan, walbira qabuun
by contrast	wal faallessuun, karaa biraa, garaagarummaa
by the same token	karaa wal fakkaatuun
by the time	yeroo sanatti, yerootti, yommus,yeroos,akka sanaan
by the way	osoo dubbannuu, osoo jennu,gidduutti, ani kanan jedhu
certainly	shakkii malee, dhugummaatti, haqumman, qabatamaan
clearly	dhoksaa malee, ifatti, mullinatti, ifa qabeessaan
come to think of it	mee itti yaadii, yaadatti seenuu
consequently	waan ta'eef, kanaafuu, kana irraa kan ka'e
considering that	ilaalcha keessa galchuun, kanaafuu, sababa kanaan, yaada keessa galchuudhan
conversely	karaa faallaa ta’een, karaa biraa
correspondingly	Walqabatee, bifa walfakkaatuun
despite this	ta'e iyyuu, yoo kana , ta'ullee
despite the fact that	hata’u malee, dhugaan akkasiis ta’u
each time	yeroo mara, yeroo hunda
earlier	subii,abboroo, durarratti
either	kana yookiin sana
else	kana malee
equally	Walqixhummaa
especially because	keessattu sababniisaa, hunda caalaa sababnisaa, sababni addumaan
especially if	keessattu yoo, hunda caalaa yoo, yoo addumaan
especially when	hunda caalaa yeroo, keessatuu yeroo
essentially then	yoosan, barbaachisaa, yeroos faayid-qabeessumman
essentially	bu’uuraan,dhugumaan,barbaachisummaadhaan
even	hanga-,ammayyuu, iyyuu, isii, isaa,sana iyyuu
even after	isa boodayyuu
even before	isa durayyuu
even if	ta'u malee, hata’u malee ,yoo ta'e iyyuu,ta'us

even so	ammayyuu, hata’u malee
even though	yaa ta’uu malee, hata’u malee
even when	yeroo kamuu
eventually	Dhumarratti
ever since	Yoonaatti, hanga ammaatti
every time	yoomiyyuu, yeroo kamuu, yeroo hunda
everywhere	eessayyuu/bakka hunda
except	iboo, sanamalee
except after	booda malee
except before	dura malee
except when	yeroo sana irra kan hafe, yeros malee
failing that	sana osoo hin darbin, bakka hin jirrettti
finally	Xumurarratti, dhumarratti
first	dursee, tokkoffaa
first of all	hunadura,hunda dura
firstly	jalqabaaf, dursee, tokkoffaarratti
following this	kanatti aanee, itti aansee, ittiaansuun
for	dhaaf,fi,akka, sababa kanaaf
for a start	jalqabaaf,eegaluuf
for another thing	gama biraan, kan biraaf, wanta biraatiif
for example	Fakkeenyaaf
for fear that	isa sodaaf,sodaa sanaa, yoo, kana sodaachuun
for instance	Fakkeenyaaf
for one thing	gama tokkoon, tokkoffaa, sababni tokko,
for one	tokkoof, jalqabuuf
for that matter	sanaaf,sababa kanaaf
for the simple reason	sababni xiqqaan, sababa salphaa sanaaf
for this reason	waan kana ta'eef, sababa kanaaf
fortunately	akka carraa, carraa ta’ee
from then on	sanatti aansee, sana booda, yerosii jalqabee, yerosii kaasee, yeroo sanii kaasee
further	garas, dabalataaf, fagoo
furthermore	gadi fageenyaan , dabalataan, kanbiraa,irra guddessaan,dabalataan
given that	isa kennuun, yoo ta’e, yoo akkas ta’ee
having said that	akka sana jechuun jedhee, , kana akkas jedhee ergan fixee ,hagana erga jennee
hence	yeroos kanaafuu,
however	garuu, ta’uyyuu,haa ta'u malee
i mean	kanan jedhu,jechuun, kan ani jedhu,kanaan jechuu fedhe
if	Yoo
if ever	Yoomiyyuu
if not	yoo ta’uu baate,yoo hin ta’u ta’e, ta'uu baatu
if only	yoo ta’e malee, yoo ta'e qofa
if so	yoo ta'e malee, erga ta’ee
in a deferent vein/way	karaa addaa, karaa biraa
in actual fact	akka dhugaatti,dhugaaf, haala qabatamaan

in addition	akka dabalaatti, dabalataan, itti dabaluun
in any case	karaa kamuu, haaluma huundaanuu, sababa kamiinuu
in case	yoo, tarii,sababa kanaan, kana,yoo sana, akkas ta'ee
in conclusion	xumerrarratti,dhumarratti, akka xumuraatti,walumaagalatti
in contrast	karaa biraa, akka faallaatti
in doing this	kana hojjachuun, kana gochuudhaan
in fact	dhugaatti,dhugaaf, dhugumman
in other respects	haala biraatiin,gama biraan,ilaalcha biraan
in other words	karaa biraa,jecha biraan, kana jechuun
in particular	qofaatti,addumaan
in short	Gabaabummati
in so doing	haala kanaan, akkasitti, akkas gochuun
in spite of that	ta’uyyuu,sanuma cinaatti,garuu/hanga ammaatti, ta'ullee, sana ta’us
in sum	ida'amaan,waliigala, dimshaashumatti
in that	sababni, achi kessatti,sana keessatti
in that case	karaa kana, sababa sana keessatti, kanaafuu, sababa sanaan
in that respect	haala sana keessatti
in the beginning	dursa, jalqabarratti
in the case of	sababa kanaan, karaa kanaa
in the end	xumurarratti, dhumarratti
in the event	osoo hin yaadin, akka jirutti, ta'umsa keessa,ta'insicha kessatti
in the first place	hunda caalaa, hunda dura, tokkoffarratti
in the hope that	abdiidhaan , abdiin, sana abdachuu keessatti
in the meantime	yeroo sana,wayitii sana, gidduu sanatti,yerooma sana, gidduutti
in this way	karaa kana, haaluma kanaan
in turn	tokko tokkoon, dabareen, wal duraa duubaan
in which case	karaa kamiinuu, sababa kamiinuu,sababa sana kessatti
in as much as	baayinuma kanaan, hanga, hammanatti
incidentally	utuu hin irraanfatin/akka tasaa, tasumatti,akka carra
indeed	shakkii malee,hojiin,isa dhugaa, dhugumaan
initially	akka ka'umsaatti, jalqabarratti, dursa irratti
in so far as	hamma, sanaan olitti,hanga, hanga kanatti,sanatti
instantly	battalumatti, yerosuma
instead	iddoosaa, bakkasaa, irra, sanaa, kana irra
it follows that	itti aanee, sanitti fufuun,iti fufee,sana hordofuun, kana irra kan maddu
it is because	sababiibsaa,sababa kanaaf
it is only because	sababniisa,sababii qofaa, wan kana qofa ta'eef
it might appear that	akka jedhametti/fakkaachuu,sana ta'uu danda'a
it might seem that	tarii fakkaachuu
just	reefuu, amma, amma kana
just then	yeroo sanatti, yerosuma,akkuma sanaan
largely because	caalmaatti sababni, bal’inaan sababni, sababni guddaan
last	dhuma,hafaa, maayii
lastly	Dhumarratti
later	gulana boodaan,
lest	ta'uu baannaan

let us assume	akkas jenne yaa yaadnu,hajennu,jenne haayaannu, akkasitti haa fudhannu
like wise	yoos,yeroo akkasii, akkasumas
luckily	carraa ta’ee, akka carraa
mainly because	sababni ijoon,hunda caalaa sababni, sababni guddaan
meanwhile	yeroo sanatti,akkuma ta'een,akkuma sanaan
merely because	sababa kanaan xiqqo,salphaatti sababni, sababa kana qofaaf
mind you	hubadhu, qalbeeffadhu
more	dabalee,caalaa
moreover	sanaa ol, dabalataan, kana caalaa,irra caalaan
most	irra guddeessaan,baayyinaan,caalmaan, hunda caaalaatti
much as	hanga danda'ame
much later	boodarra,duubarra
much sooner	ariitiin akkumasanaan,bay’ee dafee, hatattamaan,ariitiin
naturally	uumamaan,akka eegametti
neither is it the case	sababa ta'u dhiisuu, innnis sababa miti
nevertheless	karaa biraa,garuu, 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
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	tolmaama,haa jennuu, yaada sanarratti
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	irra deebi'ee,ammalle, 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	asliirraa,ka'umsarra, uumamumaan
otherwise	gama biraatiin,kana ta'uu baannaan
overall	jimlaa,dimshaasha
particularly when	yeroo addumaan
plainly	haala ifaa ta'een
presently	dhiyootti,amma, 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	Lama
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 daguu,akka tasa, 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	kan jechuu barbaade,akkas jechuun, 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	itti aanuun,sana boodaa, achumaan
then again	saniin booda,itti aansee,achiin booda
thereby	achi,achumarraan
therefore	Kanaafuu
third	Sadii
thirdly	Sadaffaa
this means	kana jechuun
this time	yeroo kana
though	ta'ullee,yaa ta'u malee
thus	kanaafuu,achirraan
to be precise	ifa taasisuuf
to be sure	mirkaneeffachuuf,dhugoomsuuf
to begin with	itti eegaluuf,itti calqabuuf,jalqabarratti
to conclude	Goolabuuf
to make matters worse	hammeessuuf,wantoota hammaataa taasisuuf
to start with	ittiin eegaluuf, ittiin jalqabuuf,dursa
to sum up	walitti qabuuf,
to summaries	Cuunfuuf
to take an example	fakkeenya dhiyeessuuf,fakkeenya kaa'uuf, akka fakkeenyaatti
to the degree that	hanga sanatti
too	baayyee,akkasuma,walfakkaataa
true	dhugaa,haqa
ultimately	kan xumuraa,kan dhumaa, olaanaa,daraan olaanaa
undoubtedly	shakkii malee
unfortunately	akka carraa
unless	ta'uu baannaan
until	Hamma
until then	hamma sanatti
we might say	tarii kan jechuun dandeenyu
well	gaarii,tole
what is more	caalaan,kana caalaa, irra guddeessaan
when	Yoom
whenever	Yoomiyyuu, yeroo kamuu
where	Eessa
whereas	gama biraatiin
where in	sana keessatti,achi keessatti
whereupon	Achirratti
wherever	Eessattuu
whether or not	ta'us ta'uu baatuus
which is why	sababnisaas,kun sababnisaa
which means	kana jechuun
which reminds me	yaada kana sammuutti qabachuun,kun kan nayaadachiisu

while	yeroo sana
whilst nearly same with that	gama sanaan,sana waliin
yet	hanga ammaattuu,hanga yoonaatti,ammayyuu
you know	beektee,bartee
alright	Gaarii dha.
argue	Falmuu
argument	Falmii
as a conclusion	Akka walii galaatti
as a consequence of	Kana irraa kan ka'e
as a logical	Akka namaaf tolutti
as a matter of fact	Dhugaadhumatti
as a summary	Walitti qabaattii
attempt	Yaalii
because of that	Sababa sanaaf
because of this	Sababa kanaaf
because our investigation	Sababa qorannaa keenyaatiif
cause	Sababa
claim	Gaafatachuu
conclusion	Xumura
consequence	Bu'aa
develop	Guddisuu
elaboration	Ibsuu
end	Dhumaa
enumeration	Lakkoofsa
first, second	Tokkoffaa, lammaffaa
for this reason	Sababa kanaaf
hardly	Kan hin fakkaanne
impossible	Kan hin taane
in consequence	Kana irraa kan ka'e
in summary	Walitti qabaatti
in this paper we show	Waraqaa kana keessatti wanti argisiifnu
it can be concluded that	Kana irraa kaanee kan jechuu dandeenyu
last of all	Hundumaa booda
okay	Haa ta'u
precisely	Gabaabaatti
purpose	Kaayyoo
reason	Sababa
result	Bu'aa
significantly	Mul'inaan
summarize	Walitti cuunfi
summary	Cuunfaa
that reminds me	Sun kan inni na yaadachiisu
the most important	Hundumaa irra barbaachisaan
the paper describes	Waraqaan kun kan inni ibsu

thereby	Achumaan
this letter	Xalyaan kun
this report	Gabaasni kun
well	Tole

Appendix B: Stop words

Aanee	bira	gidduu	ishii	keenya	otuma
agarsiisoo	Booda	gubbaa	ishiif	keessan	otumallee
akka	Booddee	hanga	ishiirraa	keessatti	otuu
akkam	dabalatees	hennaa	isii	keeti	otuullee
akkasumas	dhaan	Hoggaa	isiin	keetii	sana
akkum	dudduuba	hogguu	isin	kiyya	saniif
akkuma	dugda	ille	itti	koo	siin
ala	Dura	Illee	ittumalle	kun	silaa
alatti	duuba	immoo	ituu	malee	simmoo
alla	Eega	inaa	ituullee	moo	sitti
Amma	eegana	ini	jala	Na	sun
amma	eegasii	Innaa	jara	naaf	ta’ullee
ammo	ennaa	inni	jechaan	narraa	tahullee
ammoo	erga	irra	jechuu	natti	tanaaf
An	ergii	isaa	jechuun	Nu	tanaafi
an	F	isaaf	ka	nurraa	tanaafuu
Ana	faallaa	Isaan	kan	Nuti	tawullee
ana	fagaatee	isaanirraa	kana	nuyi	teenya
Ani	fi	isatti	kanaaf	odoo	un
ani	Fuullee	isee	kanaafi	ofii	utuu
Ati	gajjallaa	Iseen	kanaafuu	oggaa	waan
ati	Gama	Ishee	kanaf	Oo	waggaa
Bira	garuu	isheen	kee	Osoo	warra
woo	yammuu	Yemmuu	yeroo	yokiin	yoo
yookaan	yookiin	yookiinimmoo	yoom		

Appendix D: Afan Oromo synonyms

Haadhachiisuu/ aaddachiisuu	digdama / diddama	nafa / qaama/ dhaqna/ jismii
haaduu /aaduu	dirredawaa / dirreedhawaa	funyoo / haada
Jaldeessa/ aankoo	billaa / halbee	lukkuu / handaaqqoo
Aanaa/ aantii	bisaan / bishaan	waaqa / rabbi
Haarii/ aarii	biyyoo / biyyee	marga / citaa/ mayra
aayyaa /aayyoo	bokkaa / rooba	gadda / boo'a/ taziyaa
habaabayyuu/ abaabayyuu	boollo / boolla	callaa / qofaa/ qofaa
habaaboo/ abaaboo/ daraaraa	bukkee / maddii/ cinaa	dhibamuu / dhukkubsachuu
siruma/ abadan	foonaa / mooraa	geeddaruu / diddiruu
habashaa/ abashaa	fooyuu / foowuu	herreguu / yaaduu
abbaagadaa / luba	gaadduu / keettoo	rifachuu/ naasuu

abbala / hawwa	gaachana / gaalee	xiinxaluu / yaaduu
ajjaa / omborii	eega / eegee	warra / maatii
ilillii / habaaboo	eebba / heebba	jijjiiruu / diddiruu
abboomama / adabamaa	gaddii / milkii	horii / loon/ beeylada
abboomuu / adabuu	geedala / sardiida	wayyaa / uffata/ huccuu
abishii / sunqoo	waangoo / sardiida	kafana / uffata
ablee / hablee	habbayyii / abbayyii	mi'a / meeshaa
alalee / halalee	ja'a / jaha	miya / meeshaa
alamii / addunyaa	kaawoo / surraa	qodaa / meeshaa
ankarsaa / dhulaandhula	keenya / keenna	baallii / angoo/ tayitaa/ muudama
anqaaquu / hanqaaquu /buphaa/killee	kofla / kolfa	cimoo / cimaa
arcumee / harcumee	geedala / sardiida/ waangoo	coxee / catee
asimii / asmaa	habbayyii / abbayyii	cufantaa / cufaa
kudhaama / asmaa	ja'a / jaha	dubrummaa / qarree
axawuu / haruu	kaawoo / surraa	da'a / daha/ da'umasa
qulqulleessuu / haruu	keenya / keenna	daaktuu / daattuu
atamtama / harifannaa/ sardama/ muddama/ jarjara	micuree / mar'imaan	dabarsaa / dabaree
awaalama / hacuucama	harcummee / shaxxee	yeelloo / qaanii
cunqursaa / hacuucama	dhiluu / foowuu	yeella’aa/ qaana’aa
gidiraa / hacuucama	nahuu / rifachuu	makoodii / handarii
baaduu / areera/ hareera	obboroo / subii	gugee / handarii
baallama / beelama	qoonqoo / beela	jalqabuu / eegaluu
baasaa / riqicha	raajjuu / raagduu	dhaanuu / reebuu/ tumuu
baashee / beela	reettii / re'ee	beekuma / barumsa
hoongee / beela	nasuu / rifachuu	beenyaa / gumaa
bantii / qarree	siddisa / hamaaqixa	caatii/ jimaa
bara / beela	sooressa / dureessa	calii/ fo’aa jirbii
barchaa / ganboo	taa'aa / hudduu	ciciwwii / cuucii
bareeda / miidhaga	xiqqoo / bicuu	cilee / cilaattii
barraaqa / barii	yemmuu / yeroo	

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**
ixata,Dafinoo,Hojjaduree,Isinina	Monday	Amajjii	January
Kibxata ,Facaasaa, lammaffo ,Balloo, Salaasa	Tuesday	Guraandhala	February
Roobii ,Arbii	Wednesday	Bitootessa	March
Kamisa Jimaata Friday Ebla,Ebl. May	Thursday	Caamsaa	April
Sambata, Sambata duraa, Sambat'xinaa,Sabtii	Saturday	Waxabajjii	June

Aalaada, Sambata guddaa	Sunday	Adoolessa	July
		Hagayya	August
		Fulbaana, Birraa	September
		Onkoloolessa	October
		Sadaasa	November
		Mude, Arfaasaa, Afraasaa	December
Maqaa Yeroo/ Names of Time			
Ganama		Saafaa	
Waaree		Darbamtuu	
Guyyaa		Galgala	
Iyyandaanqoo		Waarii	
Obboroo		Halkanii	
Barraqa		Subii	

Appendix F: Afaan Oromo numbers corpus

0	Zeroo	Kudhaa	Jatamii	Biliyoona
1	Tokko	Digdama	Torbaatama	Tiriiliyoona
2	Lama	Digdamii	Torbatamii	Ku’aatiriiliyoona
3	Sadii	Soddoma	Saadeettama	1ffaa
4	Afur	Sodomii	Sadetamii	2ffaa
5	Shan	Afurtama	Sagaltama	3ffaa
6	Ja'a	Afurtamii	Sagaltamii	4ffaa
7	Torba	Shantam	Dhibaa	5ffaa
8	Saddet	Shantamii	Kuma	6ffaa
9	Sagal	Jahaatama	Miliyoona	7ffaa
	8ffaa	9ffaa	10ffaa	100ffaa
	1000ffaa			
0.	Zeeroo tuqaa	1.	Tokko tuqaa	
0.1	Zeeroo tuqaa tokko/ tokko kurnaffaa	1.1	Tokko tuqaa tokko/kudha tokko kurnaffaa	
0.2	Zeeroo tuqaa lama/lama kurnaffaa	10.	Kudhan tuqaa	
0.3	Zeeroo tuqaa sadii/ sadii kurnaffaa	10.1	Kudhan tuqaa tokko/ dhibbaa fi tokko kurnaffaa	
0.4	Zeeroo tuqaa afur/afur kurnaffaa	20.	Digdama tuqaa	
0.5	Zeeroo qutaa shan/ shan kurnaffaa	30.	Soddoma tuqaa	

0.6	Zeeroo tuqaa ja’a/ja’a kurnaffaa	40.	Afurtama tuqaa	
0.7	Zeeroo tuqaa torba/torba kurnaffaa	50.	Shantama tuqaa	
0.8	Zeeroo tuqaa saddeet/saddet kurnaffaa	60.	Jaatama tuqaa	
0.9	Zeeroo tuqaa sagal/sagal kurnaffaa	70.	Torbaatama tuqaa	
80.	Saddeettama tuqaa	90.	Sagaltama tuqaa	
		100.	Dhibba tuqaa	
		100.25	Dhibba tuqaa lama shan	

Appendix G: List compound words

Mana barumsaa	Caffeen Oromiyaa	Dhaabbilee Barnootaa ol’aanaa
Mana kitaabaa	waajjira Qonna Oromiyaa	muummee Saayinsii
godiina Shawwaa Kibba Lixaa	Qotee Bulaa	Muummee Afaan Oromo
aanaa Goorootti	nyaata horii	Mata duree
Manni Murtii	bona dabree	Jal-bultii
Iluu Abbaa Booraa	biqiltuu bunaa	Mana nyaataa
mana mootummaa	Bara dhufu	Waraqaa eenyummaa

St. Mary's University
School Of Graduate Studies
Department of Computer Science

Appendix H: Guide line for preparing manual summary

This guideline is to guide human summarizer to produce an extract summary by ranking sentences. The produced summary will be used as a reference summary to evaluate summary generated by the developed system. Dear evaluator please read the original news article until you understand well the content of the article. After you understand the article rank the sentences depending on their importance. Rank the sentences based on the number of the senteces in the article that is if the aticle has N number of sentences then the most important sentence ranked **N/N** or comes first. The second ranked sentence is **N-1/N,** third ranked sentence is **N-2/N** and so on. When you rank sentences by extaction method youe decision based on the following criteria.

i. **Informativeness**: are sentences that have the most important ideas of the original document. To be selected as a high rank, a sentence should be more informative and have the key idea of the article.

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

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

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

Thank you for your collaboration!

St Mary University
College of Natural Science
Department of Computer Science

Appendix I: System summary evaluation guide line

This questioner is aimed to measure the performance of the extractive summarization system developed for Afaan Oromo news single document. The summary the system generates is extractive type of summary. An extractive summary is a summary produced by selecting a certain number most important sentences from the original document. So, read the summary then using the following questions evaluate it. Tick in the box infront of the one you choise.

i. Grammatical correctness of the summary

1. Very Poor □ 2. Poor □ 3. Not Bad □ 4. Good □ 5. Very good □

ii. Non-redundancy of the summary

1. Very Poor □ 2. Poor □ 3. Not Bad □ 4. Good □ 5. Very good □

iii. The summary Referential Clarity.

1. Very Poor □ 2. Poor □ 3. Not Bad □ 4. Good □ 5. Very good □

iv. The summary informativeness:

1. Very Poor 2. Poor □ 3. Not Bad □ 4. Good □ 5. Very good □

v. Coherence and structure of the summary

1. Very Poor □ 2. Poor □ 3. Not Bad □ 4. Good □ 5. Very good □

Note:

- **Informativeness**: are sentences that have the most important ideas of the original document. To be selected as a high rank, a sentence should be more informative and have the key idea of the article.
- **Non-redundancy**: a summary should not contain unnecessary repetition of whole sentences or ideas or facts.
- **Referential integrity**: when you read the sentences in generated summary, it should be easy to identify who or what the pronouns and nouns phrases in each sentence are refereeing to.
- **Coherence**: is the smooth flow of information about a topic in consecutive sentences. When the sentences are read together, the idea should have a flow rather than having unrelated sentences.

Thank you for your collaboration!