JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF INFORMATION SCIENCE

AUTOMATIC TEXT SUMMARIZATION FOR AFAN OROMO

BY:

GEMECHU KENA

JUNE, 2014

JIMMA, ETHIOPIA

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTEMNT OF INFORMATION SCIENCE

AUTOMATIC TEXT SUMMARIZATION FOR AFAN OROMO

A Thesis submitted to the school of graduate studies, in meeting the partial fulfillment for of the award of the degree of master of science in Information Science (Electronic and Digital Resource Management)

By:

GEMECHU KENA

ADVISOR: Debela Tesfaye (Assi. professor)

CO-ADVISOR: Dula Boru (MSc)

June, 2014

Jimma, Ethiopia

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTEMENT OF INFORMATION SCIENCE

AUTOMATIC TEXT SUMMARIZATION FOR AFAN OROMO

BY:

GEMECHU KENA

This thesis entitled "AUTOMATIC TEXT SUMMARIZATION FOR AFAN OROMO" has been read and approved as meeting the requirements of Department of Information Science in partial fulfillment for the award of the degree of Master of Science in Information Science (Electronic and Digital Resource Management), Jimma University, Jimma, Ethiopia.

Name	Title	Signature	Date
_____	Chairperson	_____	_____
_____	Advisor	_____	_____
_____	Co- Advisor	_____	_____
_____	External Examiner	_____	_____
_____	Internal Examiner	_____	_____

Dedicated

To

My beloved families

Declaration

This thesis is my original work and has not been submitted as a partial requirement for a degree in any University

_____

Gemechu Kena

June, 2014

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

_____

Debela Tesfaye (Assistance professor)

June, 2014

Acknowledgement

First and foremost, I would like to give a special gratitude to the Gracious God who provided me everything to finish my courses and this thesis. He has multiplied His mercy tremendously in my life. Had it been not His help, I would have not been complete my study.

My indebtedness goes to my advisors Mr. Debela Tesfaye for their commitment and patience reading for each and every section of the thesis, their valuable comments, encouragement and guidance from the initial to the final level of the research that enabled me to finish the thesis work.

My special thanks also go to my father Kena Terefa , my mother Shashitu Mohammed and my bothers Melkamu, Debela Chara and my sister Lensa for their moral support and encouragement during my study.

It is an honor for me to give my sincere gratitude to Mr. Kuusa who supports me to assign who voluntarily stood to be human annotator. This research is not possible without their expert knowledge on human summarizing.

Finally, I extend my heartfelt thanks and respect to my friends and all those people who were not mentioned here but their contributions have been inspiring for the completion of this work.

Abstract

With the rapid development of information technology, the world is flooded with information. Also information has become the most valuable and important resource of this fast growing information society. Today, with digitally stored information available in abundance, even for many minor languages this information must by some means be filtered and extracted in order to avoid drowning in it. Automatic text summarization is one such technique, where a computer summarizes a longer text to a shorter non-redundant form. This thesis thus sets the focus on automatic text summarization for Afan Oromo language by sentence extraction for the original source documents and the evaluation of the summaries using five human resources. The resources that are used for this study is nine corpus collected from different website. The field of automatic text summarization began with some classical approach by extracting sentence from original document attempted to identify the most salient sentences of the documents using some thematic features. This research was intended to develop extraction based automatic text summarization for Afan Oromo language by using two different feature namely, term frequency and title word features for achieving accurate summaries. The proposed method was evaluated by comparing the machine generated and human summaries. Results show that title word feature is the best individual feature for extracting most informative sentence from Afan Oromo text. According to the experimentation made the system registered recall 0.37(37%), 0.33(33%) precision and 0.35(35%) F-score for the method of Term frequency. Using the title word method 0.52(52%) recall, 0.39(39%) precision and 0.44(44%) F-score that shows the improvement of the summarizer with this method. In general, according the experiment result gives the best performance for the title word feature than term frequency in both subjective and objective evaluations.

Table of content

Contents page

List of Tables

List of Figures

List of Abbreviations

ANSI	American National standards institute
ARDA	Advanced Research and Development Activity
AS	Automatic summarization
ATS	Automatic text summarization
CAS	Chemical Abstracts Service
DUC	Document Understanding Conference
IDF	inverse document frequency
IR	Information retrieval
MMR	Maximal Marginal Relevance
NIST	National Institute of Standards and Technology
NLP	Natural Language process
OTS	Open Text Summarizer
SOV	Subject-Object-Verb
SVO	Subject-verb-object
Tf	term frequency
TS	Text summarization
TT	Test text

List of Appendices

CHAPTER ONE

INTRODUCTION

1.0 Background of the study

Nowadays, people widely use the internet to find information through information retrieval (IR) tools such as Google, Yahoo, Bing, and so on. However, with the sharp growth of information on the internet, information abstraction of retrieved results has become a necessity for users. A document summary keeps its main content and consequently helps users to understand and interpret large volumes of information available in the document. Summarization, as done by humans, involves reading and understanding an article, website, document, etc. The key points are then used to generate new sentences or selecting some sentences from the document which form the summary. The needs for automated summaries is becoming more and more apparent to automatically generate the summary and get the gist of long textual data.

In the current era of information overload, text summarization has become an important and timely tool for user to quickly understand the large volume of information. Text summary is a shorter text of the original document that still keeps the main content of information in the document. This task is performed by human after deep reading and understanding of the document content, selecting the most important information and paraphrasing them into a concise version. One of the natural questions to ask in summarization is “What are the text that should be represented or kept

in a summary?" The summary must be generated by selecting the important contents and conclusions in the original text.

Currently, the need for automatic text summarization has appeared in many areas such as email summary; news articles summary; short message news on mobile; information summary for businessman, government officials, and research, etc.; online search engines to receive the summaries of pages found and so on.

The earlier effort on automatic text summarization system that were developed in the late 1950s consisted of selecting significant sentences from a source document and concatenating them together (Luhn, 1950). Luhn uses term frequencies to measure sentence relevance. Sentences are included in the summary if the words in the sentence have high term frequencies. A number of techniques for automatic text summarization proposed in this area can be classified into different approaches. Some of these techniques are classified based on the input document used for the summary.

Single document summarization uses only one document to produce a single summary while multi-document summarization uses many documents that are related to the same topic to create a single summary. The summary methods can also be classified into two approaches: extraction and abstraction (Lin, 1997). Extraction summarization method refers to the selection of sentences or phrases from the source text and generating a new shorter text without changing the source text. Usually sentences are in the same order as in the source text. In contrast, abstraction summarization process consists of "understanding" the source text by using linguistic methods to interpret and examine the text.

The abstraction summarization finds new concepts in the original text and generates a summary that describes the most important concepts. This study considers extraction summarization for single documents. The technology of automatic document summarization is maturing and may provide a solution to the information overload problem.

Several efforts were made to develop automatic document summarization for many languages. Despite the necessity of such application, the progress in developing such systems for local languages is lagging behind. In this research we have presented an automatic document summarizer for Afan Oromo text using two features namely term frequency and title word

1.2. Statement of the problem

Automatic text summarization takes an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs(Mani, 2001). The goal of automatic text summarization is to present a document in a shorter text that still keeps the main content of information in the document; one important technique of the approaches applied in automatic text summarization is extraction methods. Automatic summarization (AS) is an increasingly important task in the current era of information overload, given the large volumes of available text documents in digital media. The first work in automatic text summarization was introduced by Luhn (1958). Users always seek simplicity by presenting less content which preserve the important information in long textual data. This issue makes it's a difficult to obtain the necessary information related to the need of a user. In order to solve this issue, text summarization systems can be used. Summarization task requires understanding the document and presenting the important parts. For humans, generating a summary is a straightforward process but it is time consuming. In contrast, for automatic summarizers, finding out important information becomes a truly challenging task. An automatic text summarization system works by selecting important sentences from a document and concatenating them together. The implementation of text summarization approaches has become more difficult due to the natural language complexity. In principle, text summarization is possible because of the natural occurring redundancy in text, and the important information is unevenly spread in textual documents. It is difficult to copy human

based text summarization since human can capture and relate deep meanings and themes of text documents.

Most of the AS is developed for only few major languages like English, French, Dutch, Spanish and Arabic. Such systems are highly language dependent in that the approach worked for one language cannot be directly utilized for another language and the approaches tried for the local languages are far from perfection. Recently some advancement is being made for the developing languages like Asian Languages and few of African Languages. The fact remains the same for Ethiopian Languages. Put differently, only few attempts were made to develop such system for Local languages including Afan Oromo.

Afan Oromo is one of the languages with large number of speakers under Cushitic family. Nowadays journal, magazines, newspapers, news, online education, books, entertainment Medias, videos, pictures, are available in electronic format both on the Internet and on offline sources. Huge amount of information being released with this language, since it is the language of education and research, language of administration and political welfares, language of service activities and social interaction in Oromiya region.

Despite the significant availability of electronic digital resources presented in the Language, the progress in developing automatic systems assisting users in accessing the documents is lagging behind and Automatic summarization being one of them. Put differently, the size Afan Oromo documents in electronic format is growing dramatically. As a result Afan Oromo text readers and writers are facing difficulty for text

summarization since it requires significant time to understand deep meanings of text document to make a summary or abstract when performed using human power. To save the reader's time it is required to produce summary important points by removing unwanted detail to solve the problems of Afan Oromo text readers.

Hence the aim of this study is to explore techniques and design an automatic text summarizer for Afan Oromo language that process texts to distill the most important information from a source (or sources) to produce an abridged version for a particular users using extraction method.

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

1. To what extent text extraction technique enables to design effective automatic text summarizer for Afan Oromo text?
2. What are the best features enables to design extraction based Afan Oromo text summarizer?

1.3. Objective of the study

The study has the following general and specific objectives.

1.3.1 General Objective

The main objective of this research is to explore techniques and design an automatic text summarizer for Afan Oromo language.

1.3.2. Specific Objectives

The specific objectives of the research are the following:

1) To review approaches and techniques in the area of automatic text summarization.
2) To investigate and select techniques for Afan Oromo language on automatic text summarizer.
3) To prepare and organize test documents for automatic text summarizer.
4) To evaluate the performance of the proposed model and recommend for future research direction.

1.4. Significance of the study

With the rapid increase of local language contents in electronic form and gradual improvement of Afan Oromo language resources for computational models, the possibility of developing some language processing applications for Afan Oromo has increased. An automatic summarizer was one such major application which, many people can benefit it through because it help to get the most important and relevant information in a shorter time.

- This study also helps the language to adapt to the technology rapidly in shorter time and less cost.
- It motivated the researchers to apply other summarizing techniques to Afan Oromo and find their applicability to languages for Afan Oromo text.
- Highly benefit in several information acquisition tasks such as to promote current awareness, save readers time, facilitate selection, and improve indexing efficiency.
- It helps Afan Oromo text reader and writer to have computer extracted summaries.

1.5. Scope and limitation of the study

The study involves extraction based text summarization to the selection of sentences or phrases from the source text and generating a new shorter text without changing the source text. The summarizer does not process document with varies formatting styles such as table, graphs, image and other data types and hence are out of focus of the research.

To prepare collected document for text summarization some preprocessing techniques were done such as tokenization, stop word removal and normalization are applied. As the information is published simultaneously on many media channels in different versions which include news paper, web news, radio news cast, and a spoken newspaper for the visually impaired.

As a result of time factor, limited corpus was used for evaluating the performance of the summarizer developed in the study. This is because it takes more time to prepare relevance judgment for both system summary and manual summary.

1.6. Organization of the Thesis

This thesis is organized in five chapters. The first chapter, presents out the background, statement of the problem, and the general and specific objectives of the study together with scope, limitations study and significance of the study are included.

Chapter two is literature review and it involves two main topics, related work and conceptual review. Conceptual review is review on basic norms of text summarization, concepts of automatic text summarization, process of automatic text summarization and related topics. Related work involves work done so far on the research topic.

Chapter three discusses about the methodology used in this research. Chapter four discusses of the work is Experimentation and result of the study. In this part corpus selections and preparations, implementations of the proposed work, experimentations, findings of the study, and issues in implementations are discussed in detail.

Finally in the chapter five major findings including faced challenges are written as a conclusion and works identified as future work and needs to get attention of other researchers are listed in recommendation section.

CHAPTER TWO

LITERATURE REVIEW

2.1 Basic Norms of Text Summarization

2.1.1 Type of a Summary

Summaries can be viewed in many dimensions. One angle would be the relationship between the summary and its input and the fundamental distinction between Extracts and Abstracts can be seen through it. Extracts contain the exact sentences appeared in its input while the abstracts are rewritten forms of the input. Extract need not consist of sentences but it may consist of a list of technical terms, proper nouns, noun phrases, truncated sentences among others. Abstracts contain at least some materials which are not present in its input. However, a short abstract may offer more information than a longer extract. Another way to look at summaries is in terms of the traditional distinction between Indicative and Informative summaries (Borko & Bernier, 1975). Indicative summaries provide a reference function for selecting documents for more in-depth reading while informative summaries are aimed at helping the user to decide whether to read the information source or not.

In the standard guidelines provided for abstractors by American National Standards Institute (ANSI) has specified that the indicative summaries are to be used for less-structured documents like editorials, essays, annual reports and others, whereas informative summaries are generally used for other documents.

Also, it has been mentioned that, in scientific investigation reports, an indicative summary should contain information about the article's purpose, scope and approach but not the results, conclusions and recommendations while an informative summary should cover all of these aspects (ANSI, 1997).

Another dimension of viewing summaries is the type of users that the summary is intended for. Two different summary types can be seen through it namely User-Focused summaries and Generic summaries. User-Focused summaries (also called topic-focused summaries or query-focused summaries) are for specific user or user groups and some users' interest will be taken into account when making summaries. User query and user background knowledge of the subject are most important factors for user-focused summaries. Generic summaries are aimed at a particular readership community and traditionally those are written by professional abstractors served as surrogates for full text.

However, user-focus summaries have increasing importance in computing environments since it is always able to capture user's requirements and the interest. The input document for the summarizer can be one (single-document) or a set of multiple similar documents (multi-document). Accordingly, the summaries can be categorized as single-document and multiple-document summaries.

Generally, a summary can be one or combination of types discussed above having different features. According to the above mentioned types of automatic text summarization, the summarization technique presented in this thesis can be called single document extraction based summarization.

2.1.2 Summarization Parameters

Automatic summarization is a highly interdisciplinary application, involving natural language processing, information retrieval, library science, statistics, cognitive psychology and artificial intelligence (Mani, Automatic Summarization, 2001). Therefore, many parameters from these paradigms are involved to fine-tune the summary against its input. There can be many lists for these parameters albeit most common parameters can be described as follows.

Compression Rate is the typical parameter for every summary, which is the ratio between the summary text length and the source text length. It allows user to determine how much information he needs from the source and usually it is set anywhere from 5% to 30% (Mani, 2001). Function allows user to select the types of summaries he needs. That can be just an indication of topics or informative as to content or evaluation of the content.

Audience is the parameter to set the user's type. It can be either user-focused summary or generic summary. Relation to the source is to select whether user needs extracted summary or abstracted summary. Summaries can be generated using either from a single document or from multiple documents. That can be set from the parameter called Span. Summaries can be monolingual (processing a single language and give the output in the same language) or multilingual (processing several languages and give the output in the same language as input) or cross-lingual (processing several languages and give the output in a different language from input) and language parameter can be set to get one of these values.

Summarizer will use different strategies for various types of text such as scientific or technical reports, news stories, email messages, editorials, books and others. Genre of a summarizer is to set such different varieties of the input. Summaries can take different media types such as text, audio, tables, pictures and diagrams and movies as the input and can produce the output in one of these different forms. Media can be set to indicate this feature for a summarizer. Importance of these parameters will vary according to the application. It is unlikely that any single summarizer will handle all of these parameters. However, the summarizers are built including only the relevant parameters to satisfy the purpose of the summarizer.

2.1.3 Summarization Machine

If the summarizer is considered as a machine, the typical architecture in figure 2.1. It references some parameters described above summarization.

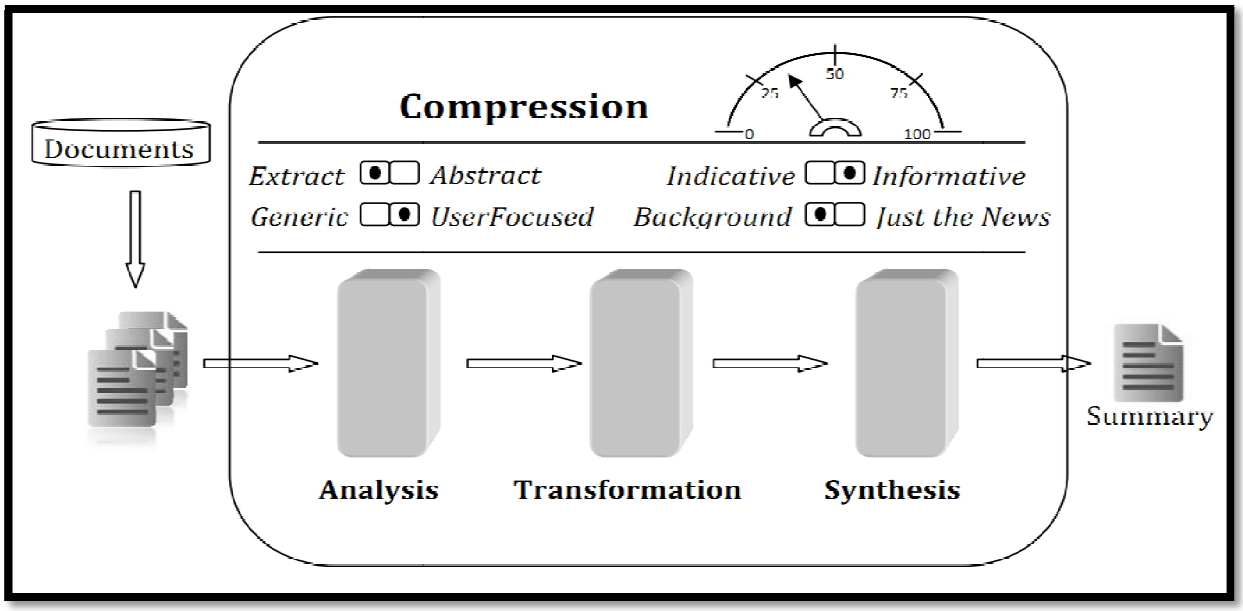

Figure 2.1: A high-level architecture of a summarizer (Source: Mani, 2001)

Researchers have identified three basic phases in text summarization, namely Analysis, Transformation and Synthesis. Summarizer analyzes the input and builds an internal representation of the input in the analysis phase. Transforming the internal representation into a representation of a summary is happening in the transformation phase. Summary representation is turned back into natural language in the synthesis phase. These two phases are mostly applicable to the systems which produce abstracts or which perform compaction or multi-document summarization. Systems which produce single-document extracts without compaction will be directly going from the analysis phase to the output.

Three basic condensation operations which summarizers carry out can be identified in any of the above phases. Selection is the operation for filtering of elements to obtain more salient information from the input. Aggregation is for merging the identified elements which were identified in the previous operation. Finally, the operation called Generalization is the substitution of elements with more general or abstract ones to make the summary. Other more complex operations such as paraphrasing or simplification can be described in terms of these three basic operations (Mani, Automatic Summarization, 2001).

2.1.4 Aspects of Summarization

A summary can be described mainly using three aspects, namely Input, Purpose and Output (Hovy & Marcu, , 1998). The domain of the source text, genre of the source (newspaper articles, editorials, letters, technical reports, emails etc.), form of the

source text (whether it is a regular text structure or a free- form) and the source text size (single document or multi documents) are the parameters for the aspect of input. These parameters can be set to define the input form and then the output will depend on it. Purpose of a summary can be described based on the situation, audience and usage of the summary.

Audience can be a focus group which has some background knowledge about the source or it can be a general audience. Output of the summary depends on its completeness, format and the style. Completeness is to indicate the level of user requirements while the style is to set the output form of the summary. It can be informative, indicative, aggregative or a critical summary. The format of the output will be a paragraph or indicative, aggregative or a critical summary. The format of the output will be a paragraph or a table or a chart.

2.2 Concepts of Automatic Text Summarization

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

The need to automatic summarization of document is increasing due to the fact that: it dramatically reduces the time required to produce a summary or abstract by experts; it enables a readers to quickly revise a content they have already seen and it enables one to

create certain standard or consistent summary format etc. Moreover, automatic text summarization systems can be applied in: summarizing news articles of newspapers and online news; can be embedded in large systems like search engines and in extracting key word and summaries of e-mail for SMS in mobile phones etc.

Though ATS is becoming a very interesting and useful task that serves the above mentioned purposes and gives support for many other tasks, it is still a challenging work (Lloret, 2008). Though early experiments in the field of automatic text summarization have showed the possibility and viability of creating text summary, it is not simple (Luhn, 1958 & Edmundson, 1969). In creating document summary automatically, one of the challenges is determining what information from the source text to be included in the summary.

The task of determining how important information to be included to the summary needs to consider several factors such as nature and genre (domain) of the source text, compression rate desired , the user's information need etc (Mani et al ,1998).

2.3. Process of Automatic Text Summarization

The process of text summarization can be decomposed in to three phases: analysis of source text, transformation, synthesis of output text. Analysis of the source text is to identify the essential content to build an internal representation. The techniques used for this task ranges from statistical methods that search for specific key content for extraction to complex techniques that employ natural language understanding. The statistical approaches in general concerned for identification of important topic terms and the extraction of contextual sentences that contain them.

On the other hand, other approaches for source analysis needs the complete understanding of the source text i.e. each sentence is processed into its propositions representing the meaning of the sentence. (Alguliev & Aliguliyev, 2009, Moens, 1997)

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

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

2.4 History of Automatic Summarization

Experiments on summarizing text using computers were begun in the late 1950's by characterizing surface level approaches. Luhn describes a simple, genre-specific approach that uses term frequencies for weighting sentences which are then extracted to make abstracts (Luhn, 1958). This work can be considered as the first computational paper on automated extraction (Mani & Maybury, 1999). Luhn was motivated by the need of dealing with information overload and it indicates that the problem of information overloading existed even before the 1950's. Rath, Resnick, and Savage have used five different word frequency and distribution based sentence selection approaches as Luhn did in his work (Rath, Resnick, & Savage, 1961).

Using these thematic features such as word frequency gave a positive start for the research in automatic summarization. In the early 1960's, researchers started to use entry level approaches based on syntactic analysis. Climenson, Hardwick and Jacobson's work has used such syntactic analysis for machine indexing and abstracting (Climenson, Hardwick, & Jacobson, 1961). Using the sentence location as a feature was introduced to the field in 1969 by Edmundson (Edmundson, 1969). He has used additional three features in addition to word frequencies, namely cue phrases, title and heading words and the sentence location. He has found that the combination of cue phases, title words and the sentence location was the best features.

He also has mentioned that the location being the best individual feature whiles the keywords alone the worst performing features. When early 1970's, there was a renewed interest in the field which led to develop first commercial application for automatic abstracting.

Pollock and Zamora have developed an automatic abstractor for the Chemical Abstracts Service (CAS) mainly using cue phrases specific to chemistry sub domain which they later used as a commercial product (Pollock & Zamora, 1975). More extensive entry level approaches have been used in the late 1970's. First discourse-based approaches based on story grammars were experimented in this time. Correira's work on computing story trees was one of early attempts for such approaches (Correira, 1980). Entry level approaches based on artificial intelligence such as use of scripts, logic and production rules, semantic networks as well as some hybrid approaches were experimented in the 1980's (Mani & Maybury,1999).

In the late 1990's the field of automatic summarization grew aggressively with all type of approaches being explored already due to the government and commercial interest for the applications.

Currently the research works have exclusively focused on extracts rather than abstracts along with a renewed interest in earlier surface-level approaches. However, more natural language generation works have been begun to focus on automatic summarization and the field is now exploring new areas such as multi-document summarization, multi lingual summarization and multimedia summarization rather than focusing on single document text summarization.

2.5 Single Document Summarization

2.5.1 Classical Approaches

The papers mentioned here are considered classics because they have created the foundations for modern applications and have also been a source of motivation for researchers. This work in those papers has been done using shallow linguistic analysis, primarily on the surface level of text. Luhn's paper on automatic abstracting provides a simple method for creating abstracts from specialized literature (Luhn, 1958). Here the author motivates the reader by expressing the advantages of automatic abstracting inexpensive and requires much less intellectual effort. Luhn (1958) used an algorithm which scans the source document for the most salient information. In this algorithm a weight is assigned to each sentence according to the term frequencies in the text. As the document is being scanned, pronouns, prepositions and other common words are filtered by using a stop list and then the remaining terms are sorted alphabetically. Next statistical analysis is applied to the list of sorted terms where pairs of succeeding words from the

input document are compared letter by letter. This allows for similar words to be found (e.g. differ, difference, different). All similar words are grouped together and are called significant words. The weight of a sentence or its significance factor is determined by the formula:

$$Sf=\frac{(\text{number of significant words})2}{\text{the total number of words}} \quad \text{..... (2.1)}$$

The sentences which have the highest weight value are extracted to produce the "auto–abstract". The system produced by Luhn was of reasonable quality given that, at that time, there were not many documents in electronic form. Luhn calls the final output an abstract, but it is essentially a summary produced by extraction. This should be made clear since we usually refer to abstracts as summaries "at least some of whose material is not present in the input" (Mani, 2001).

The work of Edmundson (1969) is one of the most influential in the area of automatic summarization. He created a framework for developing extraction based summarizers and also provided an evaluation for his system, the results of which were indeed very useful. The important innovation in Edmunson's work was the introduction of three new parameters for calculating the weights of sentences. Those were the sentence position in text, cue words and title and heading words.

The sentence position parameter indicates the place in the paragraph where a given sentence is located (e.g. the beginning or the end). This parameter was also used for assigning weights to sentences depending on their position in text. For example a positive score would be given to a sentence which was found in the first or last paragraphs of the document. This was a sensible ranking system as most informative sentences tend to be located either at the beginning ("Introduction") or at the end of the text ("Conclusion").

Examples of cue words are “significant,” “impossible,” and “hardly.” Cue words were stored in a cue dictionary which consisted of three smaller dictionaries: bonus words (positive relevance), stigma words (negative relevance) and null words (irrelevant) (Edmundson, 1969). Thus each word was ranked according to its relevance.

The final weight for a sentence equals the sum of the weights of the words in it. Keywords, on the other hand, are words that do not appear in the cue dictionary but are specific to a document. The title is very important part of a document because it can reveal the subject matter of that document. The assumption made by Edmunson was that authors usually use informative titles. While this is true for most cases, there can some exceptions. Edmunson used a large corpus of 200 documents of scientific papers which were chemistry related.

Apart from that he used another corpus of 200 documents which were related to other fields like physical science, life science and information science. This was used for purposes of statistical analysis (e.g. common words, sentence position etc.) and also determined the initial weights and parameters (Edmundson, 1969). The final weight, W (s), for a sentence s was calculated by a linear function containing the sum of the three parameters (C cue words, K keywords, L location, T title) above and the keyword parameter:

$$W(s) = \alpha C(s) + \beta K(s) + \gamma L(s) + \delta T(s) \quad \text{..... (2.2)}$$

Edmunson also experimented with the parameters in the above equation and, through evaluation, he found that keywords were not as good feature as the other three, and also that the combination of cue words, title and location gave the best performance.

He also found that location was the best and keywords were the worst individual features (Mani, 2001, p. 49). Pollock and Zamora (1975) focused on automatic extraction from specialized documents. They tried to demonstrate that using a genre specific algorithm for extraction yields better results that a more general approach. The aim of the paper was to develop a system which outputs a summary which conforms to the standards of the Chemical Abstracts Service (CAS).

Pollock and Zamora (1975) used an interesting algorithm which was used for sentence rejection rather than selection. The final output is an indicative summary, about 10–20% the size of the source. The idea behind the algorithm is that each word can be ranked using "semantic codes" which determine how suitable for extraction the word is. For that purpose Pollock and Zamora prepared a long list of words with more than 700 terms in it. Each term in this list was given a semantic code. Semantic codes were a sort of marking system and decided whether a word or a phrase is an indicator for informativeness.

For example the phrase "our results" is given the code "I" which means "very positive" which is also the highest mark. If a word is not suitable it is given a lower code like "B" (negative), or even "M" (super–negative, delete sentence). Once each word is marked accordingly, sentences are rejected of selected respectively, depending on their overall score. The first two papers discussed above proposed a similar solution to the problem by assigning weights to sentences. Luhn's system was very simple since it used only the term frequencies as a feature for extraction. Edmunson's system showed much better results in comparison to Luhn's. This was because Edmunson realized that features like sentence location and title were important features for extraction.

His paper also shows that using keywords as the only feature for extraction will give poor results (Edmundson, 1969). Then finally the third paper had a different approach to the problem of producing genre specific summaries (Pollock and Zamora, 1975). The algorithm here rejected sentences that were considered less informative. This technique proved to be quite effective for chemistry related documents.

2.5.2 Statistical (corpus–based) Approaches

In the previous section I talked about some of the early work that was done in the field. The approaches used there were simple and yet effective in most cases, but the analysis phase was done only on the single source document. A corpus is a collection of documents. Usually the documents in the collection are of different varieties. Corpus based approaches are different than other approaches in their analysis phase. This means that they analyses an entire corpus of documents instead of a single document (i.e. the source).

Machine learning techniques are often used in order to "learn" important information about the documents in the corpus. For example features like sentence location may have different values for different types of documents like newspapers and scientific papers (Mani, 2001).

So if a learning algorithm is applied to a corpus of newspapers articles, for example, it will learn that the L(s) term in equation (2.2) will have a higher value if sentence s is in the beginning of the article. Sometimes the term frequency alone is not a satisfactory measure of the importance of terms. That is why the tf.idf measure is used. It is also widely used for the purposes of information retrieval. The tf part stands for "term frequency" the number of times a term T occurs in a document.

The idf part stands for “inverse document frequency” and is calculated by the formula:

$$\text{idf} = \log_2 \frac{N}{n} \quad \text{.....(2.3)}$$

Where n is the number of document the term T appears at least once, and N is the number of documents in the collection. This is particularly useful because if a word appears a lot in one document but rarely in other documents then it is a relevant keyword for that document. The tf.idf value is usually only calculated for words that are not in the stop list. Kupiec et al. (1995) presented a more advanced extractive summarization system. It used two corpora a test corpus and a training corpus. The training corpus contained document/summary pairs. These summaries were abstracts created by professional abstractors. The algorithm was a Bayesian classifier which calculated the probability of a sentence being relevant. The formal definition for the classifier is derived from Bayes’ rule.

Here S is the summary to be produced and F1 to Fk are the features:

$$P(s \in S \mid F1, F2, \ldots Fk) = \frac{P(F1, F2, \ldots Fk \mid s \in S)(s \in S)}{P(F1, F2, \ldots Fk)} \quad \text{.....(2.4)}$$

This means that probability that a sentence s will be selected for extraction depends on the features F1 to Fk. Kupiec et al. (1995) used five main features sentence length, cue phrases, position of a sentence in paragraph (paragraph initial, paragraph final etc.), thematic words (most frequent words) and an uppercase word feature (proper names) In order to train their summarizer, Kupiec et al. (1995) used a sentence matching technique which would find correspondence between manual summary sentences and sentences in the original document (Kupiec et al., 1995).

Thus a sentence from the manual summary could be a direct match with another sentence in the source, a direct join, meaning two sentences were used from the source to produce one in the manual summary, or it could be unmatchable. The evaluation results in Kupiec et al. (1995) showed that the system produced good summaries with a high percentage of relevant sentences in them. The conclusion made at the end was that the best combination of features was paragraph cue phrase, sentence length, and the use of the keywords feature only decreased the overall performance. These observations were in agreement with Edmundson (1969)'s. Aone et al. (1999) used similar techniques to the ones presented in Kupiec et al. (1995) above. Their work goes beyond the typical frequency–based summarization systems and they used multi–word phrases as the basic text unit, instead of words.

A huge corpus of newspaper articles was pre-processed and tagged by Aone et al. (1999). A database was created from this corpus containing multi–word phrases and names. And since words were extracted along with their context (i.e. surrounding words) the database had different records for the same word or name.

For example the company "Ford" was different from President "Ford". The system also incorporated some knowledge of the corpus and was able to statistically derive collocation phrases (e.g. "computer chips", "potato chips"), find signature words by calculating idf values and recognize associated phrases (e.g. "Bayer" and "aspirin"). By gathering knowledge about the corpus, the system was able to adapt to different domains automatically. Another feature of the system was that it could be trained to better recognize signature words by using Bayesian statistics (see (Kupiec et al., 1995)).

The system explained by Aone et al. (1999) was implemented as a client–server application. The evaluation was carried out in two phases first without training and then with the trained system.

The results showed that the system was able to do better extraction when person names were removed from the text that was processed (but appeared in the summary). This was due to the fact that names have high idf values but do not indicate any relevant topics in the document. The conclusion was that names of people did not make good keywords and were rather misleading. The trained system gave better overall results in the tests and in particular it had a better precision and recall scores. They presented a summarization system called SUMMARIST which used topic identification, interpretation and generation operations to produce summaries.

The system combined statistical techniques, knowledge about the corpus and was designed to create both abstracts and extracts. Interpretation was the second step in the summarization process in Hovy and Lin (1999). Here two or more topics were “fused” into one concept. This process was considered to be the most difficult part of the summarization process because it requires knowledge about the world which is rarely included in the text explicitly. Generation was the final step in the process of summarization. The SUMMARIST system was able to produce extracts without generation by simply reproducing the sentences selected in the topic identification stage. Also it could also output topic lists with all keywords and fused concepts. And finally sentence generators together with a sentence scorer were be used to produce abstracts.

In this section complex summarization methods and techniques have emerged based on corpus statistics.

With the increasing amount of electronic publications and documents available, corpus based approaches become more popular and statistical analysis becomes a natural approach to the problem of automatic summarization. Nevertheless problems like coherence still exist when producing extracts and more advanced approaches are needed for natural language generation.

2.5.3Discourse Structure Based Approaches

Discourse structure approaches try to model the strategies that professional human abstractors use for producing abstracts. By studying the way humans create abstracts that can gain a better insight into the process of creating summaries. Then they can use this knowledge to create a better summarization system. Abstracts are usually very condensed summaries which try to follow the internal structure of the source document. Before I start discussing the advances in this area I let explain the difference between text coherence and text cohesion. Text coherence represents the relationships between sentences and clauses in the text. Text coherence as being related to the notion of a theme. And text cohesion "involves relations between words, word senses, or referring expressions, which determine how tightly connected the text is" (Mani, 2001). It also provides a way of finding the meaning of the text by examining linguistic relations such as anaphora, synonymy, hyponymy ("kind of") and metonymy ("part of"). Text cohesion represents the relationships between words in the text, as opposed to relationships between sentences (text cohesion). Boguraev and Kennedy (1997) explored phrasal analysis and the anaphoric relations in text.

The paper presented a system for summarization based on discourse structure. The system produced so called “capsule overviews” a set of key phrases and sentences from the original document. The system architecture contained several components: preprocessing, linguistic analysis, discourse segmentation, phrasal analysis, anaphora resolution, calculation of discourse salience and topic identification. These components worked together such that the output of one was the input of the next and so on. The capsule overview was in the form of a list of sentences, or parts of sentences, grouped together by topic.

Barzilay and Elhadad (1999) proposed a way of exploiting lexical chains. Barzilay and Elhadad created an algorithm which used several knowledge sources: the WordNet thesaurus, a part–of–speech tagger, a shallow parser and a segmentation component. In the process of summarization, text was first segmented (tokenized) and then lexical chains were produced. The procedure for chain construction was selecting a set of words and then for each one finding a related chain. A “relatedness” criterion was used for that purpose and if a related chain was found then the word was inserted into the chain.

The paper suggested that a lexical chain of low frequency words could carry the same salient information as the use of high frequency words. The evaluation showed that this method outperformed many commercial applications and could be used for building good quality summaries. The metric used for evaluation was the percent agreement. It measured the agreement among human judges. Results were better for a 10%summary with 96% average agreement, and for the 20% summary the agreement was 90%. Marcu (1995) used rhetorical structure theory to build trees. His rhetorical parsing algorithm used cue phrases to derive a rhetorical structure in the form of a tree.

The nodes were labeled with names of rhetorical relations (e.g. elaboration, concession) and the leaves of the tree contained elementary textual units. Each node in that tree was either a nucleus or a satellite. Nuclei nodes were assumed to hold more salient information than satellite nodes. Marcu (1995) presented a discourse–based summarizer which took rhetorical structure trees and used them to construct the final summary of a document. This was possible because the formalized structure of these trees allowed for the salience of clauses to be computed (Mani and Maybury, 1999). As part of the evaluation, Marcu (1995) compared his system with Microsoft's AutoSummarize which was part of the Office97 package.

The results expressed the percent agreement between human judges and the system with respect to the most important parts of the text. The tests showed that the discourse–based summarizer in Marcu (1995) created summaries with 60% precision and recall. Microsoft's commercial summarizer performed in the range of 40% precision and recall. And finally it was shown that the best performance was achieved by Marcu's system when manually constructed rhetorical trees were used 78% precision and 67% recall.

2.5.4 Knowledge Based Approaches

So far I have reviewed mostly domain–independent approaches to automatic summarization. The early works of Edmundson (1969) and Luhn (1958) presented more generic systems for text extraction. These systems had little knowledge about the type of document they were processing. Corpus based approaches use statistical analysis on text documents of various types in order to extracted common features from them.

The difference between all of the summarization systems discussed so far and knowledge rich systems is that, the latter are domain specific and incorporate a great deal of knowledge about a certain domain. This makes them very effective for creating summaries of documents in that domain.

The main disadvantage of knowledge rich systems is that they do not adapt easily to different types of documents, which is also a limitation. Some of the knowledge based systems discussed below output structured data and not ready–to–use summaries (Lehnert, 1981). In relation to that, others like McKeown et al. (1995) take structured data as input and generate natural language summaries from that data. This data represents the most salient information in a document and serves as a basis for creating the final summary. This decreases the complexity of the summarizer since sophisticated linguistic techniques like building rhetorical structure trees are not used (Mani and Maybury, 1999).

Lehnert (1981) talks about plot units as a way of representing the structure of narrative stories. The motivation behind this idea is that when humans read a narrative they create a mental representation of the story. And a lot of the information that learn about the story is actually inferred and is not explicitly present in the narrative. This means that by using classical summarization techniques they can only extract the information which is explicitly present in a document. In other words no inferred propositions will be recognized. Lehnert realized that the events in a story may have either a positive, negative or neutral effect on the reader. This is why she proposed affect states as building blocks of a plot unit. An affect state could be "+" (positive), "–" (negative) or "M" (neutral mental state).

State diagrams could be produced by simply connecting positive, neutral and negative states. A transition, or casual link, from a negative affect state to a neutral affect state was described as motivation, and a transition from a neutral state to a positive state was actualization (Lehnert, 1981). There were also two more casual links termination and equivalence. Primitive plot units could be constructed from the above casual links and affect states.

There were a number of different plot units like success, loss, resolution, problem etc.Thus sentences from a narrative could be labeled with a corresponding plot unit. For example the sentence "I fixed a flat tire today" is labeled as "success". More complex plot units were constructed from the primitive ones. For example the complex plot unit "giving up" consists of three primitive plot units — "failure" followed by "problem" followed by "change of mind". Plot units provided a means of "chunking" the information and then a summary could be produced by combining those "chunks" (Lehnert, 1981). Although Lehnert did not provide a full implementation of the system proposed, she created a framework for high–level analysis and summarization. The author suggested that plot units are good for generalization tasks and those they could also serve as a basis for natural language generation. McKeown et al. (1995) also used the knowledge based approach.

They described two summarization systems STREAK and PLANDOC. The first application was a summary generator which used structured data from basketball games, and the second generated summaries of telephone network plans. It took data file produced by the Bellcore PLAN software tool as input.

The goal that both systems were trying to achieve was to produce condensed summaries that contained as much data as possible. The STREAK system comprised three main components: sentence scorer, lexicalize and sentence reviser. The PLANDOC system had a different architecture. It used discourse planning and look–ahead operations such as conjunction and repetition deletion. There were several modules that carried out the plan processing in PLANDOC: fact generator, ontologizer, discourse planner, lexicalizer and sentence generator. The sentence scorer module in STREAK took a set of facts as input. These facts were produced by a fact generator which generated facts from a database of game scores. Based on these facts the sentence scorer created a semantic tree which was passed on to the lexicalizer. The lexicalizer processed that tree and mapped it onto a lexicalized skeletal syntactic tree (McKeown et al., 1995). The combined output of the sentence scorer and the lexicalizer modules was the first draft.

This draft and the facts served as the input of the sentence reviser module which produced the final draft. In the PLANDOC system a set of facts was passed to the ontologizer. The facts here were again generated by a fact generator. The role of the ontologizer was to enrich these facts with domain specific knowledge and then send them to the discourse planner. The discourse planner took the enriched facts and converted them to more complex facts. Finally the set of complex tasks was fed into the lexicalizer which did the same job as the one in STREAK (explained above). The final summary consisted of sentences automatically generated from syntactic trees. In comparison to STREAK the PLANDOC system used a simpler and more traditional approach to natural language generation.

Both systems implemented opportunistic methods for generating summaries. Although evaluation was not carried out in a formal manner, the two systems showed that the input of a summarization system was not limited to full text only. On the contrary, summarized or structured data could successfully be processed and turned into natural language by knowledge rich approaches (McKeown et al., 1995). As will be seen in the next section, knowledge based approaches are successfully used in multi–document summarization (McKeown and Radev, 1995).

2.6 Multi–document Summarization

This is a relatively new but very popular research area in automatic summarization. With the growing number of documents available electronically comes the need for some organization of information. Online news sources, for example, publish news articles every day and some of them have different versions of a story. Some may contradict; others may give exactly the same information. And since it is impossible for a user to read all the news on all the web sites, a summary of the news is desirable. Such a summary could give the users an overview of many news sources and inform them about uncertain facts by explicitly showing contradictions in the sources. Multi–document summarization could be used to solve this problem. As the name suggests, the number of documents used as source range from two to many. The different between multi–document approaches and corpus based approaches is that in the latter the corpus is usually composed of various types of documents, whereas here there should be at least two documents in the corpus that are on the same topic. One solution to the problem is to use a clustering algorithm to group similar documents together. This is done when there are many documents in a corpus on different topics.

Once the documents have been grouped, each cluster is processed and a summary is produced. There are several challenges in this area which were not an issue in previous single document approaches.

These are:

• Redundancy — eliminating redundancy is very important when processing many documents on the same topic.

• grouping — in order to group documents together by topic we need similarity measures for comparing them.

• Evaluation — human abstractors do not normally produce summaries from multiple documents so comparison between these and automatically generated ones can be problematic.

The following literature review focuses on the current state of research in this area. The papers herein are also discussed with respect to the challenges above. The first paper described here is by McKeown and Radev (1995). They present a system called SUMMONS which summarizes related news articles. The SUMMONS was a genre specific system which operated in the terrorist domain. They goal of the system was to generate fluent, variable length summaries. SUMMONS was based on traditional language generation architecture and had two main modules for doing content planning and linguistic operations. The content planner consisted of paragraph planner and combiner. The linguistic component was made up of a lexical chooser, ontologizer and a sentence generator. A similar architecture was seen in McKeown et al. (1995). The input of the system came in the form of MUC (Message Understanding Conference) templates which were directly fed to the combiner component.

These templates contained blank fields which had to be filled with some salient information from a single document. For example a template could contain the fields “victim” and “perpetrator”, which would be filled in by the system in the process of summarization. The output was a paragraph of automatically generated natural language text.

The role of the content planner was to determine the information which should be included in the summary. A set of planning operators was provided and it was used by the content planner. The operators were: change of perspective, contradiction, addition, refinement, agreement, superset, trend and no information.

Each one of these was essentially a manually written rule that linked two templates and as a result of that a third template was produced. For example if two news articles were contradicting each other’s then a contradiction operator will be used on their corresponding templates.

Thus a third template will be created which would contain the difference of the initial two. In general the whole linguistic component was reused from the PLANDOC system (McKeown et al., 1995). It contained grammar rules and constraints which were applied to words to produce natural language.

The lexical chooser managed the structure of each sentence by choosing appropriate words for each semantic role (McKeown and Radev, 1995). Finally the sentence generator produced natural language sentences by linearizing that syntactical structure. The algorithm outlined by McKeown and Radev (1995) had several steps: preprocessing, combination, discourse planning, format conversion. First templates were sorted in chronological order.

Then the templates were combined using any of the planning operators (contradiction, refinement, etc.) and the newly produced templates were sorted by priority. In the final step the sentence generator created the summary paragraph, which had variable length. As part of a testing stage in McKeown and Radev (1995), SUMMONS was given manually produced templates but no formal evaluation was carried out as a part of this project.

In the next paper, Radev et al. (2003) proposed a new approach to multi–document summarization. It was called *centroid based summarization* and was implemented in the MEAD system. This approach used clusters which were created by grouping similar documents together. A document was determined to be part of an existing cluster if its vector of highest tf.idf values was close to the vector of the *centroid* of that cluster.

As we have seen already in Aone et al. (1999), the tf.idf measure was used for single document summarization. Radev et al. (2003) proved that it can be used for multi–document summarization with the same success. It should be made clear that the term "clustering" is used in the sense of grouping random document together by topic.

Clusters are also used in the summarization process itself. This involves comparison between a sentence and the centroid of the cluster it is part of. A centroid is defined by Radev et al. (2003) as "set of words that are statistically important to a cluster of documents". If you think of clusters as being a circular area of space the centroid would be the centre of that area. And the border of that area is known as the threshold. It is important to note that a document D was only included in a cluster C if it was significantly similar to the documents in that cluster. Similarity between document D and

cluster C was calculated by the cosine similarity measure. Every cluster had a centroid which was represented by a list of tf.idf values. A centroid contained only values above a certain threshold. The overall value of a centroid was equal to the sum of these tf.idf values.

It was used along with two more parameters — positional value (P) and first–sentence overlap (F) — to give the final score for a sentence:

Score (si) = wc Ci + wp Pi + wf Fi …….. (2. 5)

By assigning a score to each sentence, the process of summarization was simplified to just picking the first n sentences with the highest score. Various other scoring functions were also used: position, centroid and overlap with first sentence. There results showed that combining these three features in a single scoring function produced the best summaries.

The formula for that function is shown below:

Score (si) = Ci + 2Pi + Fi……..…….. (2.6)

In order to cope with redundancy, Radev et al. (2003) used an algorithm which performed redundancy checks on sentences. If a sentence was considered redundant (i.e. contained overlapping words with another sentence), it was "penalized" by subtracting a redundancy penalty Rs from the overall score for that sentence.

This method was similar to Carbonell and Goldstein (1998)'s MMR (Maximal Marginal Relevance) but was modified to work with multiple documents. The evaluation framework used by Radev et al. (2003) was innovative and effective. Two new techniques, relative utility and information subsumption were introduced.

Human judges were also present, and their job was to mark each sentence in a cluster with an importance score from 0 to 10 (0 being “unimportant” and 10 — “very important”). In other words the score for a sentence is depended on the centroid (Ci), position (Pi) and the first sentence overlap (Fi) features.

In general the MEAD system performed well and the summaries produce contained informative sentences. The only drawback was that utility based evaluation required a significant amount of effort from the judges. Saggion and Gaizauskas (2004) also had a system based on clusters. Their system used extraction techniques to build personal profiles from clusters of documents. The system was presented in the Document Understanding Conference in 2004 and took part in the competition. There were two approaches covered by Saggion and Gaizauskas (2004) because the system was tested on two different tasks 2 and 5 from DUC (2004). Although both approaches were based around the idea of clustering similar documents together, the second task required more knowledge about the domain than the first. The features used in Saggion and Gaizauskas (2004) were sentence cluster similarity, sentence lead–document similarity and absolute document position. Documents were included in a cluster if the sentences in them were similar to the centroid of that cluster.

A cosine similarity measure was used. The goal of the system in Saggion and Gaizauskas (2004) was to create a profile of a person by extracting relevant information about that person from a set of documents. McKeown et al. (1995) populated blank templates of attribute/value pairs with data from the terrorist domain.

Here the goal was similar, but used a cluster of documents as the source. The personal profile was constructed from the information present in a list of facets (e.g. background, education, nationality etc.), which had to be populated by the system. This final summary was created on the basis of the personal profile, once any redundancy was removed. This was achieved by calculating an n–gram similarity between text fragments (Saggion and Gaizauskas, 2004).

The idea behind that was that if two fragments contained more than a certain number of identical text units (words), then one of those fragments was discarded as being redundant, and the other one was included in the summary. The evaluation stage in Saggion and Gaizauskas (2004), involved both human judges and the use of automatic summary evaluation tools such as SEE and ROUGE. The results from SEE with respect to task 2 were very promising but the text quality was average. Unfortunately the results for task 5 were not as high as expected.

2.7 Evaluation Methods

All summarization systems discussed thus far produced summaries of different kinds. And for the past 40 years people have worked hard to create better summarizer. But how do they decide which system produces the best summaries? Comparing summaries is not straightforward since there are many factors (coherence, relevance etc.) that have impact on the quality of the summary. Evaluation is not only used for comparison, but mainly for understanding the advantages and disadvantages of a system and learning from them (Mani, 2001). There is no universally established method for evaluation and research papers often use their own methods. Still we can distinguish between two types of evaluation methods: extrinsic and intrinsic (Mani, 2001).

An extrinsic evaluation is a test of the "usefulness" of a system. This could require some feedback from a third party like potential users of the system. Given the output of the system they are asked to perform a certain task (e.g. answer questions about a story) and then assessment is made on how good they perform that task. Thus the quality of the summary (the system output) can be inferred indirectly. An intrinsic evaluation is more of a "quality" test for a system. This process of evaluation usually involves human judges who analyses the performance of a system.

The evaluation itself could be a comparison between the summary and an ideal summary or between the system and another system (Mani and Maybury, 1999). The problems with the first approach come from the fact that it is hard to define the "perfect" summary, because there can be more than one way of summarizing a particular document. This is especially the case with abstracts. Although it is easier for judges to agree on the set of most important sentences in a text, the evaluation of extracts is very much related to their compression rates. As for second approach, comparison is usually made between the target system and a baseline system. A baseline system is a system which is set to perform a simple summarization and is considered to have the worst performance. For example a baseline system could be one which extracts random sentences from a document. The goal is to make a system that performs better than the baseline. Saggion and Gaizauskas (2004) compared the results of their system to the results of a baseline system. Another intrinsic approach was used by Marcu (1995), where the final summary was judged by humans. Radev et al. (2003) proposed the utility–based evaluation method. It was a more fine grained approach than the usual Boolean judgment (Mani, 2001).

In other words instead of judging whether a sentence should be in a summary, or not, assign to it a relative utility metric, which represents the degree of belief that a sentence should be included in the summary. Extrinsic evaluation was used by Mani and Bloedorn (2000) measured the usefulness of their system in the context of an information retrieval task.

2.8 Local Works on Automatic Text Summarization

This section describe the local works in the area of automatic text summarization and most of the attempts were made by student researchers

Amharic news summarization research is conducted by Kamil (2004). The extraction feature used by kemal in producing the summaries includes title words, head sentences, head sentences words, paragraph starting sentences, cue phrases and high frequency key words appearing in the texts. He recommended the development of good stemmer, preparing of standard Amharic corpus, organizing exhaustive lists of stop words, and the inclusion of more NLP, statistical and heuristic parameters. The Performance evaluation of his approach show 74.4% and 58% precision and recall respectively with 38.5% condensation rate.

Helen (2006) conducted a research towards Automatic Text Summarization for Amharic Legal documents used in Judgments. The study deal with the problem faced by legal experts in Ethiopia who spend their time on reading large volume documents to find relevant judgments in reaction to specific cases which results in delay of decision. As a result Helean (2006) proposed text summarization as solution to the above mention problem. The statistical extraction techniques were carrying out for the research.

Weight is assigned to each sentence based on its location and the cue words/phrases that it contains to extract the highest weighted sentences. Precision and recall measure is used for 20% and 10% compression rate. The system summary is compared against the human (ideal) summary. As a result, precision of the system summary is 33.9% and 39%; Precision of the random summary is 23% and 27%; recall of system summary is 57% and 50.5 %; recall of random summary is 46% is 38% for 20% and 10 % compression rate respectively.

The research conducted by Girma Debela (2012) on Afan Oromo news text summarizer based upon the Open Text Summarizer (OTS). His work done on customizing the OTS code. The summarizer basically uses the combinations of term frequency and sentence position method with language specific lexicons in order to identify the most important sentence for extractive summary. The result of objective evaluation shows that the three summarizers: M1(Method1), M2 (Method 2) and M3(Method3) registered f-measure values of 34%, 47% and 81% respectively i.e. M3 outperformed the two summarizers (M1 and M2) by 47% and 34 % . the subjective evaluation result shows that the three summarizers' (M1, M2 and M3) performances with informativeness, linguistic quality and coherence and structure are: (34.37 %, 37%, and 62.5%), (59.37%, 60% and 65%) and (21.87%, 28.12% and 75%) respectively as it is judged by human evaluators.

The research work by Teferi Andargie (2005) is on the same language, genre and similar problem as in the previous work by (Kamil, 2004). This study, however, employed machine learning technique (naïve Bayes). In this study, title, location, cue words and content words features are examined.

The results of the analysis shows that precision of 75.00%, recall 74.90 % and classification accuracy of 86.03% in predicting the summary sentences. The researcher recommends availability of standard Amharic corpus, analysis 26 of each single feature like cue words didn't help in the prediction of sentences for the summary and availability of standard stop-list.

CHAPTER THREE

METHODOLOGY

This chapter explains the methodology adopted to carry out the research on automatically summarizing Afan Oromo text. It is intended to generate the summary as extract rather than abstract due to certain limitations. Abstracts need some kinds of natural language generation techniques which require rich linguistic resources while extracts can be generated using sentence extraction approaches.

3.1 Research methodology

This research was conducted in order to figure out challenges of implementing an automatic text summarization for Afan Oromo language. Experimental research methodology are used for this study towards achieving the main objective, the following step by step procedures are followed.

3.1.1 Literature Review

To have conceptual understanding and identify the gap that is not covered by previous studies different materials, including journal articles, conference papers, books, and research paper have been reviewed. In this study the review is mainly concerned works that have direct relation with the automatic text summarization topic and the objective of the study.

3.1.2 Corpus preparation

The corpus is prepared from different sources namely, official website Oromiya National Regional state, International Bible Society Official website, Oromiya Radio and Television organization (ORTO) and Afan Oromo language department student research.

During selection Afan Oromo text considered to be on different topic such as sirna Gadaa 'Gada System', gumii Gaayoo, Hariroo wangeelaa fi adaa gidduu jiru 'relationship between bible and culture'. Those was selected and prepared from this websites. Two document from each site and three from Jimma University, Afan Oromo Department in different title namely, 'Sakatta'a Dogoggora Barreeffama', 'Rakkoolee Dandeettii Barreessuu' and 'mala qophii fi itti fayyadama qorichaa aadaa'

3.1.3 Text Pre-processing

The pre-processing step is perhaps the most important in the area of natural language, since the quality of the obtained summary depends on how efficient is the representation of a text. In this thesis, experiments were containing the pre-processing stage. When a pre-processing of a text is realized, an intermediate representation of it is obtained.

Stop words: When a pre-processing of a text is realized an intermediate representation of it is obtained on the pre-processing stage consists in eliminating stop word from text. For stop word removal Afan Oromo stop words are used. For example, 'aanee', 'akkam', 'ala', 'illee' 'irraa' etc for more detail see (appendix III). A preprocess of extraction of stop-word in the documents with the aim of reducing the content of the text to more specific expressions, containing only words that are useful and meaningful for the generation of automatic summaries.

Normalization and Tokenization: Text normalization is the issue tokens that matches occur despite superficial difference in the character sequences of the tokens, it also needs to normalize terms into the same forms. The text normalization in this study used is a rule-based component which removes the unimportant objects like table and handling of non standard words like web URL's and email.

Tokenization is done to spit the text into sentences a seemingly trivial task, but which can be complicated by the fact that punctuation marks also serve other purposes or a sentences ends with one or more points, exclamation marks and/or question mark. A sentence splitting also determine by a sentence delimiter are optionally follow by an ending quotation mark. The text is then divided into sentences for further processing.

3.1.4 Text Extraction

Text representation techniques based on the extraction of terms of a text or document which consist in choosing terms that extracted and then turned into summary. The research work carried out by sentence extraction module identifies the most important sentence in the input document. The input of the extraction module is the input document and the output is a list of key sentences that have been selected by extraction module. Title word and tem frequency feature were used for the identification of important sentence from Afan Oromo text.

3.1.5 Implementation tool

To develop automatic text summarization for Afan Oromo language, Python 3.3 programming language were used for window and Ubuntu, Fedora and other Linux environment. Python is dynamic programming language that is used in a wide variety of application domains.

It is simple, strong, involves natural expression of procedural code, modular, dynamic data types, and embeddable with in applications as a scripting interface.

3.1.6 Creating Reference summaries

The proposed methodology of this research in single document automatic summarizer for Afan Oromo text are evaluated with the reference summaries selected extent which is marked by the language experts. Corpus was selected from different website, books and student research papers (see section 3.1.2).

However, the number of document with human annotations was limited only nine documents due to the lack of human resources and time. Selected nine documents were randomly allocated for nine language experts who have professional experience with a guideline (see appendix I). They were asked to select the most important sentences of each document which could be described as the intention of the document. Additionally they were instructed to consider the summary's linguistic qualities, referential clarity, coherence, non-redundancy and informativeness of the summaries. The gold standard summaries are generated based on the three human summarizers' average rank of sentence. The average rank of a sentence is compute as the sum of the three divided by three. The prepared reference summaries for each test set is required to be compared with summaries generated by system for the purpose of performance evaluation.

3.1.7 Evaluation criteria

To evaluate the quality of system extracted summaries against the human or manually extracted summaries, the precision, recall and F-score were calculated for the system summaries. Calculating the precision and recall to measure the relevance of a set system summary with reference summaries.

$$\text{Precision} = \frac{\#of\ sentence\ in\ the\ automatic\ extract\ and\ also\ in\ the\ human\ extract\ sentence}{total\ \#of\ sentence\ in\ the\ automatic\ extract\ sentence}$$

$$\text{Recall} = \frac{\#of\ sentence\ in\ the\ automatic\ extract\ and\ aslo\ in\ the\ human\ extract\ sentence}{total\ \#of\ sentences\ in\ the\ human\ extract\ sentence}$$

$$\text{F-Score} = \frac{2\ Precision\ X\ Recall}{Precision + Recall}$$

3.2 The feature used for Afan Oromo text summarization.

3.2.1 Summarization features

This research attempted design a model using two different thematic features for assigning the weight of important sentence included in the summary. This research is carried out for the independent domain for data set. Also, it was assumed that the document or have a unique structure which is more suitable to be summarized. The research was conducted based on two thematic features.

1. **Identifying Term frequency**

Keywords of the document are primarily identified based on the term frequency. The main assumption of this paradigm is called "Thematic Term Assumption", that is relatively more frequent terms are more salient (Mani, 2001). A weight is assigned to each sentence according to the term frequencies in the text. As the document is preprocessed stop words and other common words are filtered by Afan Oromo stop word list. The sentences with hive the highest weight value are extracted to produce the summary.

$$\text{Tf} = \frac{\text{sentences salience}}{\text{total number sentences}} \quad \ldots\ldots 3.1$$

Where: tf is term frequency

2. Title words

The title words features it is assumed that authors always used contents related to the title for filling the article. Therefore the title can be considered as the essential part of the document. Edmundson (1969) has defined title words as a feature and that is used to assign a weight to the sentences based on the terms in it that are also present in the title.

Edmundson has used the title subtitle and heading to identify the title words and has manually assigned weight as it leads to the best performance. The selected corpus does not have any subtitle and heading by each document consists of an appropriate main title.

$$W(s) = \frac{\text{No of title words in the sentence s}}{\text{total number of words in the sentences s}} \quad \text{..... (3.2)}$$

Where, W(s) is the weight assigned for the sentence s based on title word.

Equation 3.2 which is defined to assign weight for the sentence s due to title words, always gives a value.

CHAPTER FOUR

IMPLEMENTATION, EXPERIMENTATION AND DISCUSSIONS

This chapter explains the implementation of Afan Oromo text summarization, experiments carried out using the designed methodology for the data-set which was explained in the previous chapter. Results of those experiments along with their evaluation are also described.

4.1 Implementation of Afan Oromo text summarization

4.1.1 Document preprocess

The test data consists of paragraphs written in Afan Oromo language. The corpus was containing different themes collected from website and books i.e it is a heterogeneous collection. The document contain are different paragraphs and contents (see section 3.1.2).

4.1.2. Summarization Process for Afan Oromo text.

This research carries out on Afan Oromo text summarization in single document summarization using sentence extraction from the original source to make a summary. The adopted summarization method is sentence extraction based. It has four major steps:

(1) Preprocessing, (2) Split the content into sentence (3) sentence ranking and (4) summary generation.

1- Preprocessing: - To create a summary of Afan Oromo text, first of all pre-processing is done on input Afan Oromo text before further processing. Since the quality of the obtained summary depends on how efficient is the representation of a text.

So, eliminating Afan Oromo stop words, tokenization is done to split the text in to sentence level and normalizations are applied as a rule-based component which removes the unimportant objects like table and handling of non standard words like web URL's and email are applied for the pre-processing Afan Oromo text.

2- Split the content into sentence: - After the preprocessing is done the content of the input document are split into sentence. The summarization system operations on sentence level computing feature value for each of them and generate a summary. Split the text are determined a sentence ends with one or more points, exclamation marks and/or question mark.

3- Sentence Ranking: - The sentence ranking is done by splitting the content into sentence and counting term frequency and comparing title sentence with the sentence containing title word in that document. Thus, sentence ranking is after an input document is formatted. The document is broken into sentence and sentence are ranked based the features identified for each sentence in the Afan Oromo text.

4- Summary Generation: - When all the sentences are assigns a final score the best sentences are selected to create a summary. The sentences with highest scores are considered as the best sentences and included in the summary.

4.1.4. Afan Oromo text summarization system Architecture

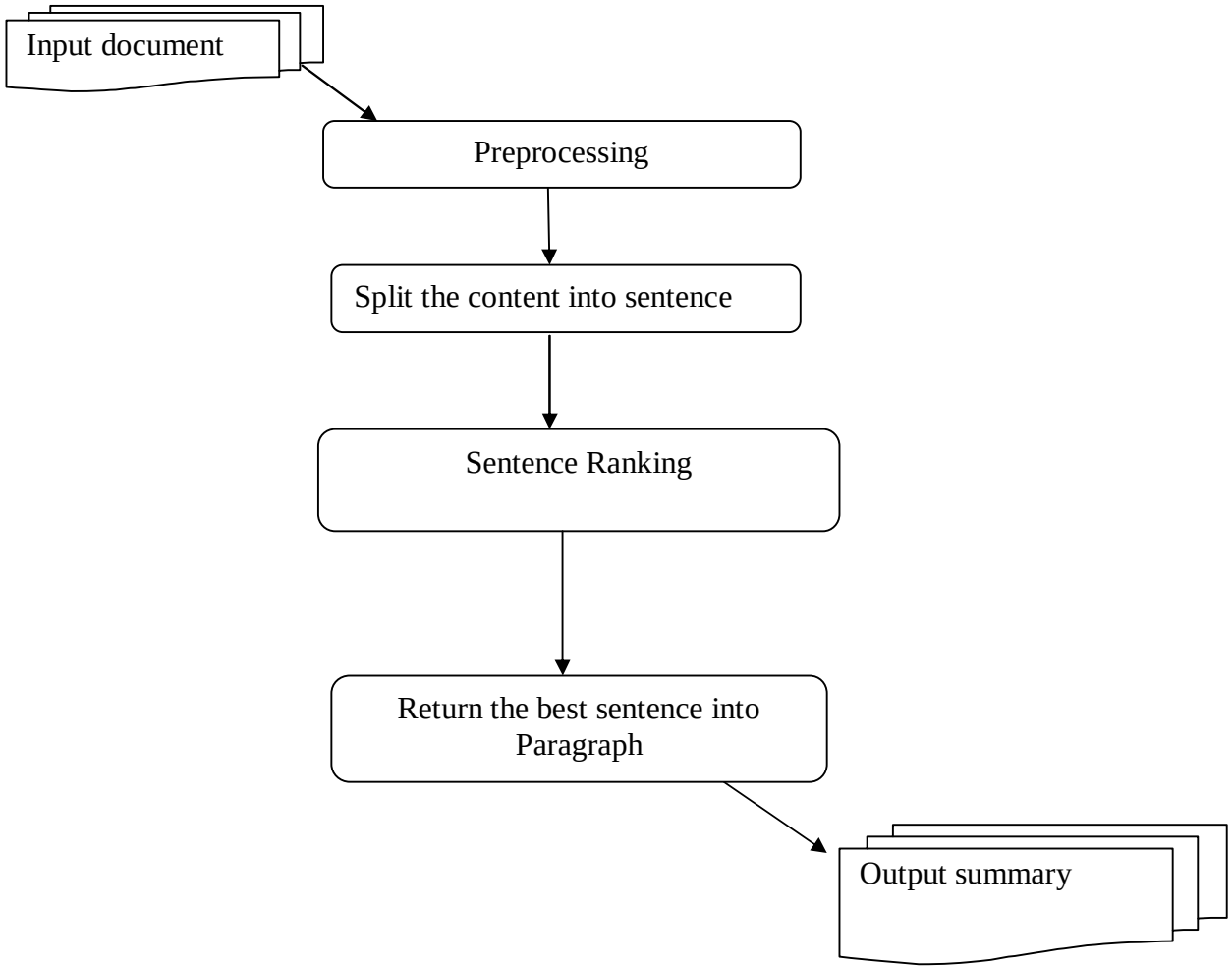

Figure 4.1 system Architecture

Figure 4.1 shows the system architecture. This is a domain independent single document summarization system for Afan Oromo text; therefore the input to the system is a single document for any domain and the output is a summary of the input document. The summarization process using extraction approach identifies the most important sentence in the input document.

4.2 Experimental setup

4.2.1 Data/corpus preparation

The test data consists of paragraphs written in Afan Oromo language. The corpus was containing different themes collected from website and books (see section 3.1.2)

Text _ID	Compression rate (%)	Document size in sentences
TT_001	5	76
TT_002	5	77
TT_003	5	51
TT_004	15	72
TT_005	15	121
TT_006	15	37
TT_007	30	162
TT_008	30	126
TT_009	30	274

Table 4.1 Basic statistics of Afan Oromo corpus.

4.3 Experimentation method.

For each Afan Oromo text, two experiments have been used with different compression rate. After the pre-processing phase weight is assigned to each sentence in the source text. Two features are applied in this research to identify the important sentences from the text. The final score of the sentence marks the importance of the sentence in Afan Oromo text are scored using term frequency and title word with 5%,15%,30% compression rate.

4.3.1 Identification of frequent Afan Oromo word.

The words which are frequently occurring in every sentence are considered to be the important words of the document. The sentences containing such words are considered to be the important and provide in the summary. The sentences containing the frequent words are assigned have high score. The sentences with hive the highest weight value are extracted to produce the summary.

4.3.2 Identification of Afan Oromo Title word.

The title words features it is assumed that authors always used contents related to the title for filling the article. Therefore the title can be considered as the essential part of the document. Edmundson (1969) has defined title words as a feature and that is used to assign a weight to the sentences based on the terms in it that are also present in the title.

Edmundson has used the title subtitle and heading to identify the title words and has manually assigned weight as it leads to the best performance. The selected corpuses do not have any subtitle and heading by each article consists of an appropriate main title. Sentence containing the title word were considered important for the summary.

4.4 Text summarization Evaluation Measures and discussion

A crucial phase of the development of any system, method or methodology is the evaluation and validation of said task. Natural Language Processing (NLP) systems are no exception. Rather, give the irregularities of (human) language it is simply an all too daunting task to logically prove this loosely defined body of possible utterances. Most automatic text summarization systems today are extraction-based systems. Summary evaluation methods attempt to determine how adequate and reliable or how useful a summary is relative to its source. Generally, there are two types of evaluation methods namely. The first is *intrinsic* evaluation in which users judge the quality of summarization by directly analyzing the summary. Users judge fluency, how well the summary covers stipulated key ideas, or how it compares to an ideal summary written by the author of the source text or a human abstractor. None of these measures are entirely satisfactory. The ideal summary, in particular is hard to construct and rarely unique. In most cases there is no only one correct ideal summary for a given document.

The second type of evaluation methods is *extrinsic*. For this study, the summarizers are evaluated using objective and subjective methods. For both subjective and objective evaluation methods used are intrinsic to the summary.

4.4.1 Subjective evolution

In order to establish criteria for evaluating automatic summary, nine automatic summaries were evaluated by five human subjects. The summaries were evaluated in terms of ease of understanding and appropriateness as summaries in five levels: 1-Very poor; 2-poor, 3-Fair, 4-good and 5-very good. The result of the subjective evaluation based evaluation point; results are available in (Appendix IV). The subjective evaluation results were converted into factor scores using factor analysis in order to normalize subjective differences. The evaluate check whether the summary has smooth transition of sentence, linguistic quality includes non-redundancy and referentially and check the best sentence are the contain the most important information of the topic sentence.

4.5 Result and Discussion of subjective evaluation of system summary

4.5.1 Content of summaries created by system

In this section I present result of the subjective evaluation based on evaluation criteria explain of Appendix II. For content of the summaries created for each text item is scaled out of 100 if the expected total 25 scales (1- V. poor, 2-poor, 3- fair, 4- good and 5-V. good) by five human subject evaluators. The results from the evaluator are turned into statistics based on the added score of the five results and compared on a scale out of 100. For example, if the summary test1 score 1 by evaluator 1, score2 by evaluator 2 and 3score by evaluator 3 score3 by evaluator4 and 4score by evaluator 5, the percentage of the overall grading for informativeness and content of sentence generated by machine

where as the average of the sum of the scores in percentage is i.e 1+2+3+3+4=13. The total score out of 25 that mean 13/25=0.52 (52%). For more detail of test document in terms of how much the machine summaries covers the important content of the original document and informativeness summaries measure best sentences that contain the most important information of the topic. For detail see table 4.2 and results are available in (Appendix IV).

Text_ID	Compression rate	System summary	
		Term Frequency	Title word
TT_001	5%	0.28	0.36
TT_002	5%	0.20	0.36
TT_003	5%	0.24	0.28
TT_004	15%	0.56	0.68
TT_005	15%	0.52	0.64
TT_006	15%	0.64	0.68
TT_007	30%	0.64	0.72
TT_008	30%	0.64	0.72
TT_009	30%	0.60	0.76
Average		**0.48**	**0.57**

Table: 4.2 content of system summaries result

As it is show in table 4.2 the obtained result by term frequency is 0.48(48%) and 0.57(57%) for title word feature. According to the result of title word feature is a better performance than term frequency.

4.5.2. Coherence

To the measure how the summary is structures the sentences and coherent. The results from the evaluator are turned into statistics based on the values what the evaluator give for coherence of the sentence are added the score of the five human subjects evaluator and compared the scale out of 100. For example, if the summary test1 score 4 by evaluator 1, score3 by evaluator2 and 2score by evaluator 3, score3 by evaluator4 and

2score by evaluator 5, the percentage of the overall grading for the coherence generated by machine where as the average of the sum of the scores in percentage is i.e 4+3+2+3+2=14 the total score out of 25 that means 14/25= 0.56(56%) For more detail see table 4.3 and results are available in (Appendix IV).

Text_ID	Compression rate	System summary	
		Term Frequency	Title word
TT_001	5%	0.40	0.44
TT_002	5%	0.40	0.44
TT_003	5%	0.40	0.48
TT_004	15%	0.64	0.72
TT_005	15%	0.60	0.64
TT_006	15%	0.56	0.68
TT_007	30%	0.68	0.76
TT_008	30%	0.68	0.72
TT_009	30%	0.60	0.68
Average		**0.55**	**0.61**

Table 4.3 coherence of system summaries result.

As it is show in table 4.3 the obtained result from five human subject evaluator average are 0.55 (55%) for term frequency feature and 0.61(61%) for title word to selected important sentence from a document. The system summary coherent of the title word features is better than performance of term frequency feature.

4.6 Objective Evaluation

4.6 .1 Precision, recall and F-Score

The evaluation of a summary quality is a very ambitious task. Serious questions remain concerning the appropriate methods and types of evaluation. There are a variety of possible bases for the comparison of summarization systems performance. Compares a system summary to the source text, to a human-generated summary.

To evaluate the quality of computer extracted summaries against the manually extracted summaries, the Precision and Recall were calculated for the computer extracted summaries.

Calculating the Precision and Recall to measure the relevance of a set of machine generated data with reference summary is a well established technique. Precision is defined as number of sentences occurring in both system and ideal summaries divided by the number of sentence in the system summary, while Recall is defined as the number of sentence occurring in both system and ideal summaries divided by the number of sentence in ideal summary . Equation 4.1 and 4.2 show the exact mathematical definitions of Precision and Recall respectively.

$$\text{Precision} = \frac{\#of\ sentence\ in\ the\ automatic\ extract\ and\ also\ in\ the\ human\ extract\ sentence}{total\ \#of\ sentence\ in\ the\ automatic\ extract\ sentence} \quad (4.1)$$

$$\text{Recall} = \frac{\#of\ sentence\ in\ the\ automatic\ extract\ and\ aslo\ in\ the\ human\ extract\ sentence}{total\ \#of\ sentences\ in\ the\ human\ extract\ sentence} \quad (4.2)$$

As it can be seen in equation 4.1 and 4.2, if it attempts to increase the Recall by retrieving more extract, it will cause to decrease the Precision and vice versa.

Therefore, to get the maximum values for both of these measures, the harmonic mean of the Precision and Recall, called F-Score is calculated. F-Score reaches its best value at 1 and worst score at 0. Even though there are some variations of the definition for the F-Score, the traditional definition which was used to evaluate these experiments is shown in equation 4.3

$$\text{F-Score} = \frac{2\ Precision\ X\ Recall}{Precision + Recall} \quad (4.3)$$

This F-Score measure was calculated for each computer generated and manually extracted summaries to evaluate the performance of the proposed methodologies.

Text ID	Compression rate	Term frequency			Title word		
		precision	recall	F-score	precision	recall	F-score
TT_001	5%	0.10	0.10	0.10	0.20	0.30	0.24
TT_002	5%	0.10	0.10	0.10	0.30	0.30	0.30
TT_003	5%	0.0	0.0	0.0	0.25	0.25	0.25
TT_004	15%	0.35	0.40	0.37	0.50	0.60	0.54
TT_005	15%	0.35	0.50	0.41	0.40	0.65	0.49
TT_006	15%	0.40	0.40	0.40	0.60	0.60	0.60
TT_007	30%	0.50	0.50	0.50	0.45	0.70	0.54
TT_008	30%	0.60	0.65	0.62	0.45	0.65	0.52
TT_009	30%	0.65	0.70	0.67	0.40	0.70	0.50
Average		**0.33**	**0.37**	**0.35**	**0.39**	**0.52**	**0.44**

Table 4.4 Objective evaluation result.

4.7Results of Objective evaluation and Discussion

The main thing to notice from the results above is that title word feature performs better than term frequency. The results of the experimentation have been compared with gold standard summary. I compute the standard recall, precision and F-score. As it has been discussed in section 4.6.1 Recall (R) is defined as the number of sentence occurring in both system and ideal summaries divided by the number of sentence in ideal summary. Precision (P) is defined as the number of sentences occurring in both system and ideal summaries divided by the number of sentence in the system summary. The F-score is composite measure that combines precision and recall. As it is shown in table 4.4 result of objective evaluation they obtained result is recall 0.37(37%), 0.33(33%) precision and 0.35(35%) F-score for the method of Term frequency. Using the title word method 0.52 (52%) recall, 0.39(39%) precision and 0.44(44%) F-score that shows the improvement of the summarizer with this method. According to table 4.4 test text ID TT_001-TT_003, F-score of both title word and term frequency are small percent by using compression rate 5% this is because system summary and reference summary have small number of

common sentences. But in both 15% and 30% compression rate the summarizer has better performance in both features.

4.8 Subjective Vs objective evaluation result.

For this study, the summarizers in both feature term frequency and title word are evaluated using subjective and objective method. The results of both subjective and objective show that title word feature have a better performance than term frequency. As it is shown in table 4.2 the obtained result by term frequency for the content and informativeness of created by system summaries is 48% and 57% using title word feature.

To measure how the system summary is structure the sentence and coherent. The average coherence and structure is 61% and 55% for title word and term frequency respectively. As it is shown in table 4.4 result of objective evaluation they obtained result is37% recall, 33% precision and 35 F-score by term frequency and (52%, 39% and 44%) are register using title word. In general, as the experiment show that title word feature has a best performance than term frequency in both subjective and objective evaluation.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

As amount of textual information available electronically grows rapidly, it becomes more difficult for a user to cope with all the text that is potentially of interest. Automatic document summarization methods are therefore becoming increasingly important. Document summarization is a problem of condensing a source document into shorter version preserving its information content. The extraction summarization method that extracts the most relevance sentences from the source document to form a summary.

As the title indicates, this research was carried out on extraction based automatic text summarization for Afan Oromo language. However, a vast amount of research has been carried out and many different approaches have been tried out over the last six decades to identify the best possible approaches to automatically summarize human languages. The experiments were carried out based on extraction approaches used in automatic text summarizer. Experimental results prove that some thematic features which researchers have identified for the languages such as English can be used for Afan Oromo language as well for the same objectives. This study basically, aimed to explore techniques and design an automatic text summarizer for Afan Oromo language. Two Different kinds of features were used for the weight of the sentence that included in the summary.

Key words of the document are primarily identified based on the term frequency. The main assumption of this paradigm is called “thematic Term Assumption” that is relatively more frequent terms are more salient. Title word is another feature used in this study by assumed that authors always used contents related to the title for filling the article. Therefore the title can be considered as the essential part of the document. Evaluation is an essential part of a practical discipline like automatic summarization. However, it is crucial to say one summary is better than another summary even though it can be easily said if it is a bad summary.

Researchers have been launched to find out most accurate ways to evaluate machine generated summaries. Since humans need to be involved to judge the machine outputs (summary) for giving a perfect evaluation of a summary. For this study, the summarizers are evaluated using objective and subjective methods. For both subjective and objective evaluation methods used are intrinsic to the summary. Automatically produced summaries of Afan Oromo text have been evaluated. The procedure involved five human subjects and included applying intrinsic measures to evaluate each summary.

Approach used for this study is extraction based single document summarization. According to the experimentation made the system registered 0.33(33%) precision, 0.37(37%) recall and 0.35(35%) F-score by term frequency feature and title word were registered 0.52(52%) recall, 0.39(39%) precision and 0.44(44%) F-score using three compressions rate which is promising to design text summarizer for Afan Oromo. So the term frequency has a low performance then title word feature in both objective and subjective evaluation.

5.2 Recommendations and future directions

This research was carried out based on the extraction approaches used in automatic text summarization for Afan Oromo Language.

1. More advanced method to implement in future is the practice of abstract summarization methods, synonym resolution and pronoun resolution which the resulting summary is an explanation of the text that the result will be much more coherent.
2. The study conduct using single document text summarization only it need expand work for the new areas on summarization such as multi-document summarization and other multi-media summarization.
3. To improve the performance of the summarizer a good stemmer, standard Afan Oromo corpus with the inclusion of more NLP are under consideration should be develop.
4. One of the most important features of improving performance of text summarization is cue word feature (Edmundson 1969). One can consider enhancing the model, so that model reformulates by adding cue word feature.
5. The size of the test collection used in this research is too small. However, one can increase the test collection and can evaluate the performance of the summarizer.
6. To improve the performance of the system an advanced summarization technique such as machine learning technique further studies should be conducted to design an effective text summarizer.

References

Abera N. (1988), “Long vowels in Afan Oromo: A generic approach” , Master’s thesis , School

Of graduate studies, Addis Ababa University, Ethiopia.

ANSI, A. N. (1997). Guidelines for Abstracts. Bethesda, Maryland: National Information

Standards Organization (NISO) Press.

Aone, C., Okurowski, M. E., Gorlinsky, J., & Larsen, B. (1999). A trainable Summarizer with

Knowledge Acquired from Robust NLP Techniques. In I. Mani, & M. T. Maybury, Advances in Automatic Text Summarization. Cambridge, Massachusetts: MIT Press.

Barzilay, R., & Elhadad, M. (1999). Using Lexical Chains for Text Summarization. In I.

Mani & M. T. Maybury, Advances in Automatic Text Summarization. Cambridge, Massachusetts: MIT Press.

Beeferman, D. and Berger, A. (2000). Agglomerative clustering of a search engine query log. In

Knowledge Discovery and Data Mining

Boguraev, B. and Kennedy, C. (1997). Salience-Based Content Characterization of Text Documents. In Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization, The MIT Press.

Borko, H., & Bernier, C. L. (1975). Abstracting Concepts and Methods. San Diego,

California: Academic Press.

Carbonell, J. G. and Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for

Reordering Documents and Producing Summaries. In Moffat, A. and Zobel, J., editors, SIGIR98, Melbourne, Australia.

Climenson, W. D., Hardwick, N. H., & Jacobson, S. N. (1961). Automatic syntax analysis in

Machine indexing and abstracting. American Documentation, 178-183.

Correira, A. (1980). Computing Story Trees. American Journal of Computational Linguistics.

Edmundson, H. P. (1969). New methods in automatic abstracting. Journal of the Association for

Computing Machinery.

Girma D, (2012), "Afan Oromo news text summarizer" Master's Thesis, Faculty of Informatics,

Addis Ababa University. Addis Ababa.

Grage G. & Kumsa T.(1982), "Oromo dictionary", African studies center. Michigan

State University.

Hahn, U., & Reimer, U. (1999). Knowledge-based text summarization: Salience and

Generalization operators for knowledge base abstraction. In I. Mani & M. T. Maybury, Advances in Automatic Text Summarization. Cambridge, Massachusetts: MIT Press.

Helen A. (2006), "Automatic Text Summarization for Amharic Legal Judgments",

Master's Thesis, Faculty of Informatics, Addis Ababa University. Addis Ababa.

Hoa, T.D. (2007) Overview of DUC (2007). In Proceedings of the Seventh Document

Understanding Conference (DUC). New York, USA

Hovy, E., & Lin, C.-Y. (1999). Automated Text Summarization in SUMMARIST. In I. Mani, &

M. T. Maybury, Advances in Automatic Text Summarization. Cambridge, Massachusetts: MIT Press.

Kamil N.(2005).," Automatic Amharic News Text Summarizer", Master's Thesis, Faculty of

Informatics, Addis Ababa University, Addis Ababa.

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In

Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'95), (pp. 68-73). Seattle, WA.

Kupiec, J., Pedersen, J. O., and Chen, F. (1995). A Trainable Document Summarizer.

Lehnert, W. G. (1981). Plot Units and Narrative Summarization. In Mani, I. and Maybury, M. T.,
Editors, Advances in Automatic Text Summarization. The MIT Press.

Lloret E.(2008) ,"Text summarization: an overview" ,Dept. Lenguajes y Sistemas

Inform_aticos Universidad de Alicante Alicante, Spain.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. IRE National Convention, New York.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. IBM Journal of Research Development.

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

Mani, I. (2001). Automatic Summarization. John Benjamins Publishing Company, Ams terdam/Philadephia.

Mani, I. and Bloedorn, E. (2000). Summarizing Similarities and Differences Among Related Documents. In Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization. MIT Press.

Mani, I. and Maybury, M. T., editors (1999). Advances in Automatic Text Summarization. MIT Press, Cambridge, MA.

Marcu, D. (1995). Discourse Trees Are Good Indicators of Importance in Text. In Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization, MIT Press, Cambridge, MA.

McKeown, K., Robin, J., & Kukich, K. (1995). Generating Concise Natural Language Summaries. Information Processing & Management.

McKeown, K. R. and Radev, D. R. (1995). Generating Summaries of Multiple News Articles. In Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization, MIT Press.

Myaeng, S. H., & Jang, D.-H. (1999). Development and Evaluation of a Statistically-Based Document Summarization System. In I. Mani, & M. T. Maybury, Advances in Automatic Text Summarization (pp. 61-70). Cambridge, Massachusetts: MIT Press.

Pollock, J. and Zamora, A. (1975). Automatic Abstracting Research at Chemical Abstracts Service. Journal of Chemical Information and Computer Sciences.

Radev, D. R., Jing, H., and Malgorzata Stys, D. T. (2003). Centroid-Based Summarization of Multiple Documents. Information Processing and Management.

Rath, G. J., Resnick, A., & Savage, T. R. (1961). The Formation of Abstracts By the Selection of

Sentences. American Documentation.

Spar k J ones , K .— Galliers , J . R (1995).: Evaluating Natural language processing Systems :

An Analysis and Review In lecture notes in Artificial Intelligence.

Teferi A.(2005), " The application of Machine learning Technique (NAÏVE BAYES) for

Automatic Text Summarization the Case of Amharic News Text", Master's Thesis. Faculty of Informatics, Addis Ababa University, Ethiopia.

Teufel, S., & Moens, M. (1999). Argumentative classification of extracted sentences as a first

step towards flexible abstracting. In I. Mani, & M. T. Maybury, Advances in Automatic Text Summarization. Cambridge, Massachusetts: MIT Press.

Tilahun G.(1993), " Qubee Afan Oromo : Reasons for choosing the Latin script for

developing an Afan Oromo Alphabet" Journal of Oromo studies.

Viatcheslav Y., Timur V. (2007), "Evaluating contemporary automatic text summarization

systems: an experiment", Computational Linguistics Laboratory (CLL).

Yatsko V. A. and Vishnyakov T. N. (2007), "A method for evaluating modern systems of

Automatic text summarization", Automatic Documentation and Mathematical Linguistics

Appendix I: Guideline for manual summarizers

The purpose of this guideline is to enable human summarizers to create a summary by extracting sentences from the document and ranking sentences according to the fellow of the context.

1. Evaluator must read the original articles deeply understand the document contains after deep understand the concepts, evaluator going to summarize a document into paragraph by ranking the sentence from original document.

2. Evaluator, it is reasonable to assume that the most common language error such as misspelled words, word separation errors and others are rare to appear in the document this help to get more accurate word frequencies, which significantly affect assigning the sentence for the summary.

3. All the evaluators have simple unique structure, which do not contain different formatting styles such as tables, graphs, image, and others. It contains only which is written fluently by separating paragraphs.

4. All the evaluator while ranking the sentence for the summary. A summary should not contain redundant idea or sentences, they should be a smooth transition of sentence and informative or the best sentence are the contain the most important information of the topic sentence.

5. 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.

Appendix-II Guideline for subjective evaluation

Dear evaluator, you are expected to read the document carefully. Then you are going to evaluate system summaries according to evaluation scale from 1-5. This evaluation scale is give based on text quality measure such as informativeness and content of the summaries and coherence of the sentence.

1. Is the summaries include best sentences are that contain the most important information of the topic sentence?

2. Does the summary have good structure and the sentences are coherence.

Appendix III Afan Oromo stop word

aanee	gararraa	ishiirraa	koo	siin
agarsiisoo	garas	ishiitti	kun	silaa
akka	garuu	ishiitti	lafa	silaa
akkam	giddu	isii	lama	simmoo
akkasumas	gidduu	isiin	malee	sinitti
akkum	gubbaa	isin	manna	siqee
akkuma	ha	isini	maqaa	sirraa
ala	hamma	isinii	moo	sitti
alatti	hanga	isiniif	na	sun
alla	henna	isiniin	naa	tahullee
amma	hoggaa	isinirraa	naaf	tana
ammo	hogguu	isinitti	naan	tanaaf
ammoo	hoo	ittaanee	naannoo	tanaafi
an	hoo	itti	narraa	tanaafuu
ana	illee	itumallee	natti	ta'ullee
ani	immoo	ituu	nu	ta'uyyu
ati	ini	ituullee	nu'i	ta'uyyuu
bira	innaa	jala	nurraa	tawullee
booda	inni	jara	nuti	teenya
booddee	irra	jechaan	nutti	teessan
dabalatees	irraa	jechoota	nuu	tiyya
dhaan	irraan	jechuu	nuuf	too
dudduuba	isa	jechuun	nuun	tti
dugda	isaa	kan	nuy	utuu
dura	isaaf	kana	odoo	waa'ee
duuba	isaan	kanaa	ofii	waan
eega	isaani	kanaaf	oggaa	waggaa
eegana	isaanii	kanaafi	oo	wajjin
eegasii	isaaniitiin	kanaafi	osoo	warra
ennaa	isaanirraa	kanaafuu	otoo	woo
erga	isaanitti	kanaan	otumallee	yammuu
ergii	isaatiin	kanaatti	otuu	yemmuu
f	isarraa	karaa	otuullee	yeroo
faallaa	isatti	kee	saaniif	yommii
fagaatee	isee	keenna	sadii	yommuu
fi	iseen	keenya	sana	yoo
fullee	ishee	keessa	saniif	yookaan
fuullee	ishii	keessan	si	yookiin
gajjallaa	ishiif	keessatti	sii	yoolinimoo
gama	ishiin	kiyya	siif	yoom

Appendix IV Subjective summary evaluation result

Content of system summaries result using term frequency feature						
Text ID	E1	E2	E3	E4	E5	Total
TT_001	1	1	2	1	2	7
TT_002	1	1	1	1	1	5
TT_003	1	1	1	2	1	6
TT_004	3	3	3	2	3	14
TT_005	3	3	3	2	2	13
TT_006	3	3	3	4	3	16
TT_007	3	4	3	3	3	16
TT_008	3	3	4	3	3	16
TT_009	4	3	3	2	3	15
Content of system summaries result using title word feature						
Text ID	E1	E2	E3	E4	E5	Total
TT_001	2	1	2	2	2	9
TT_002	2	2	2	1	2	9
TT_003	1	2	2	1	1	7
TT_004	4	3	3	4	3	17
TT_005	3	4	3	3	3	16
TT_006	3	3	4	4	3	17
TT_007	4	3	3	4	4	18
TT_008	3	4	4	4	3	18
TT_009	4	4	4	3	4	19

Coherence of system summaries result using term frequency feature						
Text ID	E1	E2	E3	E4	E5	Total
TT_001	2	2	2	2	2	10
TT_002	2	2	2	2	2	10
TT_003	2	2	1	3	2	10
TT_004	3	3	3	4	3	16
TT_005	3	2	4	3	3	15
TT_006	3	3	2	3	3	14
TT_007	4	3	4	3	3	17
TT_008	4	3	3	4	3	17
TT_009	3	3	3	3	3	15
Coherence of system summaries result using title word feature						
Text ID	E1	E2	E3	E4	E5	Total
TT_001	3	2	2	2	2	11
TT_002	2	3	2	2	2	11
TT_003	3	2	2	3	2	12
TT_004	4	3	4	4	3	18
TT_005	3	3	3	4	3	16
TT_006	4	3	3	3	4	17
TT_007	4	3	4	4	4	19
TT_008	3	4	4	3	4	18
TT_009	4	3	3	4	3	17

Appendix V Samples of Source Document, Machine Extracted Summaries.

Sample of source: TT_001

Barreessuun dandeettiiwwan gurguddoo afaanii arfan keessaa tokko yommuu ta'u barbaachisummaan isaas hammuma sadar kaafi beekuma keenya bal'achaaafi guddachaa deemu dabala. Akkumu beekamu namni hundisuu dhimmoota garaa garaatiif walquunnamuun barbaachisaadha. Yaada sammuu keessa jiru tokko baasanii walii dabarsuuf malli ittiin fayyadamnu inru guddaan barreeffamani. Walii galtee bifa barreeffamaan ergaa dabarfachuu kanatti kan dhimma bahu immoo nama barata qofadha. Kun immoo kan galma gahuu danda'u seera ittiin barreeffamu beekuun ergaa quutuufi ifa ta'e yoo dabarsedha. Barreessuuf yaada ofii karaa sirrii ta'e bakka buusuun kan danda'amu qubeetti gargaaramudhani.Barreessuu ilaalchisee, Nunan (1989:165) kitaaba "language teaching methodology" jedhamu keessatti akka armaan jedha. "Barreessuun dandeettiiwwan afanii gurguddoo jiran keessaa isa cimaa ykn ulfaataa akkasumas adeemsa qalxaxaa waan ta'eef kan akka bifa ykn unkaa, qabiyyee qubee sirrii, seerluga, qindoomina jechootaafi yaadaa eeganii barreeffamuu kan gaafatudha. Dabalataan Hayyuun afaanii Geetaachoo Robbirraa (2005:1) akka jedhutti "Qubeewwan yookiin mallattoolee uummatni irratti waliigaleetti fayyadamuudhaan waraqaa, dhagaa sibiilaa fi k.k.f. irratti barreessuudhaan tooftaan ergaa ofii ittiin dabarfatan keessaa inni beekamaan barreessuu jedhama" Walumaagalotti akka yaada beektota kanaatti barreessuu beekuun ergaa hiika qabeessa ta'e dabarsuuf, dogoggora yaadaa hambisuufi gahumsaan waliigaluuf bu'aa guddaa qaba. Ka'umsa Qorannichaa: Dandeettiin barreessuun uumamaan cimaa, walxaxaa, shakkalli irra deddeebii yeroo dheeraafi muuxannoo ximaa

kan barbaadudha. Barruun tokko ergaa guutuu dabarsuu kan danda'u immoo rakkoo dandeettii barreessuu irraa walaba yoo ta'e barreeffame qofaadha. Kunis jechoonni hima keessattii argaman akka sagalee isaaniitti sirnaan qindaanii yoo barreeffaman, tuqaaleen iddoo galuu qabanitti galan, jabeessuu laaffisuu, dheeressuu, gabaasuufi qubee guddeessi yoo haala barbaachisuun galaniifi yaadonni barraa'an walitti hidhiinsa haalaqabuun qindaa'anidha.Baayren (1988:4) akka jedhutti; "Barreessuun yaada ofii karaa seera qabeessa ta'een qindeessuun teessisuudhaafi dubbisuun dhimma itti ba'an akkasumas namoota biroof ittiin ergaa dabarsanidha." Kun kan nu hubachiisu yemmuu barreessinu daballi yookiin hir'inni qubee osoo hin jiraatiin jechicha akka jirutti qubeessuu danda'uudha. Akka yaada hayyootaatti dandeettiin qubeessuu dandeettii xiyyeeffannaa guddaan kennamuufii qabudha. Qo'attuunis dandeettii qubeessuu hojjettootaa qo'achuuf kan kaate kanarraa ka'uudhani.Gaaffileen bu'uuraa firii qorannoo kanaatiin deebii ni argatu jedhanii yaadamaniifi ka'umsa qorannoo kanaa ta'an keessaa muraasni akka armaan gadiitti dhiyaatu . Isaanis:

1. Barreessuun maaliif ulfaataa?
2. Wantoota barreeffaman keessatti irra caalaatti kan mul'atu dogoggora maaliiti?
3. Wantoonni gahumsa dandeettii Afaan Oromootiin barreessuu hojjettootaa irratti dhiibbaa uuman maal fa'ii?
4. Dogoggorri barreessuu maal irraa maddaa?

Kaayyoo Qorannichaa: Qorannoon kun kaayyolee lama qaba. Isaanis: Kaayyoo Gooroofi Gooree jedhamuun beekamu. Kaayyoo Gooroo: Kaayyoon guddaan qorannoo kanaa rakkoolee dandeettii barreessuu afaan Oromoo hojjettoonni waajjira bulchiinsa magaalaa Arjoo qaban adda baasuun yaada furmaataa kennuu ta'a.

Kaayyoo Gooree: Dandeettii barreessuu hojjettoota waajjira kanaa keessaa sakatta'uudha.

- ❖ Wantoota dandeettii barreessuu hojjettoota irratti dhiibbaa fidan ifa gochuudha.
- ❖ Dogoggora irra caalaatti barreeffama keessatti argaman addeessuu
- ❖ Barbaachisummaa barreessuun adeemsa ergaa walii dabarsuu keessatti qabu ifa gochuu.

Barbaachisummaa Qorannichaa : Qorannoo kanarraa qaamonni adda addaa fayyadamoo ta'uu ni danda'u fakkeenyaaf"

- ➢ Namoonni sirna qubeessuu irratti rakkoo qaban akka dogoggora isaanii sirreeffatan ni taasisa.
- ➢ Hojjettoonnii waajjirichaa qorannoo sana bu'uura godhachuun dogoggora barreessuu isanii ni fooyyessu
- ➢ Keessumaa barsiisaa afaaniifi barataa afaanii rakkoo qubeessuu qaban irraa maqisuuf ni oola
- ➢ Namoonni fuulduratti qorannoo haala kanaan walfakkaata ta'e gaggeessuu barbaadan akka ka'umsaatti itti ni gargaaramu.

Daangaa Qorannichaa: Matadureefi iddoon qorannoon kun irratti gaggeeffamu murtaa'uudha. Kunis Godina wallagga bahaa aanaa Jimmaa Arjoo; waajjira Bulchiinsa magaala Arjoo irratti kan gaggeeffamu yoo ta'u. Innis:-Sakatta'a dogongora dandeettii barreeffama Afaan Oromoo hojjettoonni waajjira kanaa qaban xiyyeeffannoo qorannoo kanaati xiyyeeffannoon isaas irra caalaa sakatta'a dogoggora qubguddeessa hudhaa fi k.k.f dha. Malleen Qorannichaa: Qorannoon tokko yommuu gaggeeffamu karaalee daatoon ittiin funaanamu murteessaadha. Karaaleen kunis mala qorannoo safarataa fi quleullee jedhamu. Isaan keessaa qorattuun kun kan qorannoo ittiin gaggeessitu safarata ta'a. Qorannicha keessatti tooftaaleen ittin adeemsifamus irraawwatama, Iddattoo qorannichaa murteessuufi mala odeeffannoon ittiin funaanamu kan akka afgaaffii bargaaffii fi galmee sakatta'uu ta'a. Irraawwatama Qorannichaa: Irraawwatamni qo'annichaa Hojjettoota waajjira bulchiinsa magadaa Arjoo yoo ta'u, Innis sakatta'a

dogoggora barreeffama Afaan oromoo hojjettoota waajjira kanaa irratti kan bu'uureffate ta'a. Iddattoo fi Iddatteessuu: Qorannoon kun kan inni irratti gaggeeffamee Godina Wallagga bahaa aanaa Jimmaa Arjoo yommuu ta'u hojjettoota waajjira bulchiinsa magaalaaa 32 jiran irratti adeemsifama. Kunis iddatteessuu carraa tasaafi sirnaawaatiin kan geggeeffamu ta'a. Tooftaalee Odeeffannoon Ittiin Funaanamu:Qo'annoo kana gaggeessuuf maddi odeeffannoo hojjettoota waajjira bulchiinsa magaalaa Arjoo irraa kan fudhatamu yoo ta'u malli odeeffannoon ittiin funaanamu afgaaffii bargaaffii fi galmee sakaatta'uu ta'a. Af-gaaffii: Karaa qo'attuun afaaniin ykn yaalaa fuulatti dubbii dhaan iddattoo irraa odeeffannoon ittiin funaanamudha. Haala kanaan qo'attuu fi iddattoon kallattiin waan walquunnamaniif odeeffannoo gahaa akka argattu gargaara.

Bargaaffii":Mala kana gargaaramuun kan barbaachisuuf qo'attuun yeroo gabaabaa keessattu iddattoo irraa bifa barreeffaman odeeffannoo fudhachuuf gargaara kunis bargaffii cufaafi banaa irratti hundaa'uun kan adeemsifamu t'a sababa qorattuun tooftaa kanatti dhimma baatus iddattoo mara irraa odeeffannoon waan fudhamuufidha. Galmee Sakatta'uu :Tooftaan kun immoo kan inni fayyadu barreeffamoota yeroo baay'ee tajaajila hawaasaa kennuuf waajjiricha keessaa bakka adda addaatti ergaman kallattiin sakataa'uu ta'a.

Kana jechuun xalayoota waajjiricha keessaa gara alootti barreeffaman irratti kan xiyyeeffatu ta'a. Kunis raga qabatamaa dogoggora qubeessuu irratti mul'atu mirkaneessa.

Sakatta'a barruun boqonnaa lammaffaa qaama qo'annooti namni tokko rakkoo tokko yeroo qo'atu, qo'annoo goggeessu sana irratti yaada isaa cimsachuuf yaada hayyootaa fi barruulee garaagaraa boqonnaa kana keessatti gargaarama. Kanarraa ka'uudhaan

qo'attuun kunis yeroo qo'annoo kana gaggeessitu yaada ishee haalaan cimsachuuf, yaada hayyootaafi barruulee adda addaa sakatta'uun kana keessatti kan dhiyeessitu ta'a. Maalummaa Barreessuu

Ronald, (1995:1-3) maalummaa barreessuu ilaachisee akkas jechuun lafa ka'a. Weitting is one of the most widely misunderstood of human activities. We can explain other compex tasks like driving and programming computer. But to most people the act of writing is mastery. Writing is exploring. It is a kind of competition b/n writer and reader. Ronald akka jedhutti barreessuun gochaa ilma namaa xiyyeeffannoo malee hubatamu keessaa isa tokko dha. Gochaawwan walxaxoo ta'an biroo kan akka konkolaachisuu, kompitera ajajuufaa ibsuu dandeenya. Garuu gochaa barreessuun abuurraadha. Innis gosa waldragommii barreessaafi dubbisaa gidduutti ta'udha. Yaada Ronald irraa wanti hubatamu barreessuun kana jedhanii lafa kaa'uun rakkisaa ta'uusaati. Ta'us wanta borreessaafi dubbisaa gidduutti ta'u ta'ee walumaagalatti barreessuun gochaa abuurraa ta'uu isaati. Akkasuma hayyuun Byrne (1991:1) akka jedhutti, "barreessuun yaada sammuu keenya keessa jiru waraqaarra kaa'uun namni biraa akka hubatu affeeruudha." Ammas hayyuun kun barreessuun dandeetti qubee walitti fiduun hima ergaa qabu tokko barreessuun ergaa ofii ittiin dabarsabidha jedha. Walumaagalatti barreessuun adeemsa yaada sammuu keenya keessa jiru mallattoo hiika kennuu danda'uun waraqaarra kaa'uudhaan namni biraan akka hubatu taasisuudha.

Kaayyoo Barreessuu Akka Tashoomaan (1998:216) jedhutti "Barreessuun sababaa fi kaayyoo mataa isaa qaba." Kaayyoowwan kanneenis akka dimshaashatti kan ka'an:

- Odeeffannoo dabarsuuf ykn waan tokko beeksisuuf
- Yaada yookiin waan haaraa tokko dabarsuufi amansiisuuf
- Amantaa, fedhiifi ilaalcha ofii ibsuufi kan kana fakkaatan isaan ijoodha.

Furmaata Rakkina Barreessuu

Akka barreessitoonni gumii qormaata Afaan Oromoo (1996:23) irratti jedhanitti “Afaan tokko Afaan Oromoo dabalatee jecha keessaa sagalee yookaan birsaga tokko jabeessuu, laaffisuu, dheeressuuniifi gabaabsuun hiika jechoota sanaa jijjiira jedhu.” Akka yaada barreessitoota kanaatti Afaan Oromoo keessatti jecha tokko yeroo barreessan sagalee ykn birsaga tokko tokko jabeessuun, looffisuun, dheeressuufi gabaabsuun jijjiirama hiika jecha sanaa waan fidhuuf xiyyeeffannoo argachuu akka qabu ibsu. Barbaachisummaa Barreessuu. Barreessuun faayidaan isaa inni guddaan yaada sammuu keessa jiru barreessoon mallattoolee waliigalteetti fayyadamuun nama biroof dabarsuudha. Barbaachisummaan barreessuus hanguma sadarkaan barnootaa dabalaa deemu tajaajilli isaas dabalaa ykn guddachaa kan adeemudha. Baddiluu, (1996:21) akkas jedha. . . . . namni dandeetti barreessuu gaarii qabu tokko tooftaa haala gooriin waayyooduu danda’uu gabbifata. Namni yeroo baay’ee waa barreessu qabiyyeewwan barreessuu sanaa ni xiinxala, ni qindeessa, adeemsa kana keessattis dandeettii waa yaaduu, xiinxaluu fi qindeessuu ni horata. Hiika jechoota adda addaas ni bora jechuun lafa kaa’a. Kanaafuu faaidaan barreessuu inni quddaan ergaa barbaachise tokko mallattoolee waliigalteetti gargaaramun nama biroof dabarsuufha. Kana malees, dandeettii yaaduu barreessaa kan cimsuufi dandeettii dubbisuu kana taana fooyyessachuun akka danda’amudha.

Summary with 5% compression rate

Barreessuun yaada ofii karaa seera qabeessa ta'een qindeessuun teessisuudhaafi dubbisuun dhimma itti ba'an akkasumas namoota biroof ittiin ergaa dabarsanidha. Kun kan nu hubachiisu yemmuu barreessinu daballi yookiin hir'inni qubee osoo hin jiraatiin jechicha akka jirutti qubeessuu danda'uudha. Irraawwatama qorannichaa: Irraawwatamni qo'annichaa Hojjettoota waajjira bulchiinsa magadaa Arjoo yoo ta'u, Innis sakatta'a dogoggora barreeffama Afaan oromoo hojjettoota waajjira kanaa irratti kan bu'uureffate ta'a. Sakatta'a barruun boqonnaa lammaffaa qaama qo'annooti namni tokko rakkoo tokko yeroo qo'atu, qo'annoo goggeessu sana irratti yaada isaa cimsachuuf yaada hayyootaa fi barruulee garaagaraa boqonnaa kana keessatti gargaarama. Kanarraa ka'uudhaan qo'attuun kunis yeroo qo'annoo kana gaggeessitu yaada ishee haalaan cimsachuuf, yaada hayyootaafi barruulee adda addaa sakatta'uun kana keessatti kan dhiyeessitu ta'a. Ta'us wanta borreessaafi dubbisaa gidduutti ta'u ta'ee walumaagalatti barreessuun gochaa abuurraa ta'uu isaati.

Summary with 15% compression rate

Barreessuun yaada ofii karaa seera qabeessa ta'een qindeessuun teessisuudhaafi dubbisuun dhimma itti ba'an akkasumas namoota biroof ittiin ergaa dabarsanidha. Kun kan nu hubachiisu yemmuu barreessinu daballi yookiin hir'inni qubee osoo hin jiraatiin jechicha akka jirutti qubeessuu danda'uudha. Qo'attuunis dandeettii qubeessuu hojjettootaa qo'achuuf kan kaate kanarraa ka'uudhani.Irraawwatamni qo'annichaa Hojjettoota waajjira bulchiinsa magadaa Arjoo yoo ta'u, Innis sakatta'a dogoggora barreeffama Afaan oromoo hojjettoota waajjira kanaa irratti kan bu'uureffate ta'a. Tooftaalee Odeeffannoon Ittiin Funaanamu:Qo'annoo kana gaggeessuuf maddi odeeffannoo hojjettoota waajjira bulchiinsa magaalaa Arjoo irraa kan fudhatamu yoo ta'u malli odeeffannoon ittiin funaanamu afgaaffii bargaaffii fi galmee sakaatta'uu ta'a. Galmee Sakatta'uu: Tooftaan kun immoo kan inni fayyadu barreeffamoota yeroo baay'ee tajaajila hawaasaa kennuuf waajjiricha keessaa bakka adda addaatti ergaman kallattiin sakataa'uu ta'a. Sakatta'a barruun boqonnaa lammaffaa qaama qo'annooti namni tokko rakkoo tokko yeroo qo'atu, qo'annoo goggeessu sana irratti yaada isaa cimsachuuf yaada hayyootaa fi barruulee garaagaraa boqonnaa kana keessatti gargaarama. Kanarraa ka'uudhaan qo'attuun kunis yeroo qo'annoo kana gaggeessitu yaada ishee haalaan cimsachuuf, yaada hayyootaafi barruulee adda addaa sakatta'uun kana keessatti kan dhiyeessitu ta'a. Ta'us wanta borreessaafi dubbisaa gidduutti ta'u ta'ee walumaagalatti barreessuun gochaa abuurraa ta'uu isaati. Furmaata Rakkina Barreessuu Akka barreessitoonni gumii qormaata Afaan Oromoo irratti jedhanitti Afaan tokko Afaan Oromoo dabalatee jecha keessaa sagalee yookaan birsaga tokko jabeessuu, laaffisuu, dheeressuuniifi gabaabsuun hiika jechoota sanaa jijjiira jedhu.Akka yaada barreessitoota kanaatti Afaan Oromoo keessatti jecha tokko yeroo barreessan sagalee ykn birsaga tokko tokko jabeessuun, looffisuun, dheeressuufi gabaabsuun jijjiirama hiika jecha sanaa waan fidhuuf xiyyeeffannoo argachuu akka qabu ibsu.

Summary with 30% compression rate

Barreessuuf yaada ofii karaa sirrii ta'e bakka buusuun kan danda'amu qubeetti gargaaramudhani. Barreessuun dandeettiiwwan afanii gurguddoo jiran keessaa isa cimaa ykn ulfaataa akkasumas adeemsa qalxaxaa waan ta'eef kan akka bifa ykn unkaa, qabiyyee qubee sirrii, seerluga, qindoomina jechootaafi yaadaa eeganii barreeffamuu kan gaafatudha. irratti barreessuudhaan tooftaan ergaa ofii ittiin dabarfatan keessaa inni beekamaan barreessuu jedhama. Walumaagalotti akka yaada beektota kanaatti barreessuu beekuun ergaa hiika qabeessa ta'e dabarsuuf, dogoggora yaadaa hambisuufi gahumsaan waliigaluuf bu'aa guddaa qaba. Kunis jechoonni hima keessattii argaman akka sagalee isaaniitti sirnaan qindaanii yoo barreeffaman, tuqaaleen iddoo galuu qabanitti galan, jabeessuu laaffisuu, dheeressuu, gabaasuufi qubee guddeessi yoo haala barbaachisuun galaniifi yaadonni barraa'an walitti hidhiinsa haalaqabuun qindaa'anidha. Baayren (1988:4) akka jedhutti; Barreessuun yaada ofii karaa seera qabeessa ta'een qindeessuun teessisuudhaafi dubbisuun dhimma itti ba'an akkasumas namoota biroof ittiin ergaa dabarsanidha. Kun kan nu hubachiisu yemmuu barreessinu daballi yookiin hir'inni qubee osoo hin jiraatiin jechicha akka jirutti qubeessuu danda'uudha. Qo'attuunis dandeettii qubeessuu hojjettootaa qo'achuuf kan kaate kanarraa ka'uudhani. Gaaffileen bu'uuraa firii qorannoo kanaatiin deebii ni argatu jedhanii yaadamaniifi ka'umsa qorannoo kanaa ta'an keessaa muraasni akka armaan gadiitti dhiyaat. Barbaachisummaa Qorannichaa: Qorannoo kanarraa qaamonni adda addaa fayyadamoo ta'uu ni danda'u fakkeenyaaf Namoonni sirna qubeessuu irratti rakkoo qaban akka dogoggora isaanii sirreeffatan ni taasisa. Hojjettoonnii waajjirichaa qorannoo sana bu'uura godhachuun dogoggora barreessuu isanii ni fooyyessu Keessumaa barsiisaa afaaniifi barataa afaanii rakkoo qubeessuu qaban irraa maqisuuf ni oola Namoonni fuulduratti qorannoo haala kanaan walfakkaata ta'e gaggeessuu barbaadan akka ka'umsaatti itti ni gargaaramu. Irraawwatamni qo'annichaa Hojjettoota waajjira bulchiinsa magadaa Arjoo yoo ta'u, Innis sakatta'a dogoggora barreeffama Afaan oromoo hojjettoota waajjira kanaa irratti Kan bu'ureffate ta'a. Tooftaalee Odeeffannoon Ittiin Funaanamu: Qo'annoo kana gaggeessuuf maddi odeeffannoo hojjettoota waajjira bulchiinsa magaalaa Arjoo irraa kan fudhatamu yoo ta'u malli odeeffannoon ittiin funaanamu afgaaffii bargaaffii fi galmee sakaatta'uu ta'a. Bargaaffii": Mala kana gargaaramuun kan barbaachisuuf qo'attuun yeroo gabaabaa keessattu iddattoo irraa bifa barreeffaman odeeffannoo fudhachuuf gargaara kunis bargaffii cufaafi banaa irratti hundaa'uun kan adeemsifamu t'a sababa qorattuun tooftaa kanatti dhimma baatus iddattoo mara irraa odeeffannoon waan fudhamuufidha. Galmee Sakatta'uu: Tooftaan kun immoo kan inni fayyadu barreeffamoota yeroo baay'ee tajaajila hawaasaa kennuuf waajjiricha keessaa bakka adda addaatti ergaman kallattiin sakataa'uu ta'a. Sakatta'a barruun boqonnaa lammaffaa qaama qo'annooti namni tokko rakkoo tokko yeroo qo'atu, qo'annoo goggeessu sana irratti yaada isaa cimsachuuf yaada hayyootaa fi barruulee garaagaraa boqonnaa kana keessatti gargaarama. Kanarraa ka'uudhaan qo'attuun kunis yeroo qo'annoo kana gaggeessitu yaada ishee haalaan cimsachuuf, yaada hayyootaafi barruulee adda addaa sakatta'uun kana

keessatti kan dhiyeessitu ta'a. Yaada Ronald irraa wanti hubatamu barreessuun kana jedhanii lafa kaa'uun rakkisaa ta'uusaati. Ta'us wanta borreessaafi dubbisaa gidduutti ta'u ta'ee walumaagalatti barreessuun gochaa abuurraa ta'uu isaati. Akkasuma hayyuun Byrne akka jedhutti, barreessuun yaada sammuu keenya keessa jiru waraqaarra kaa'uun namni biraa akka hubatu affeeruudha. Ammas hayyuun Kun barreessuun dandeetti qubee walitti fiduun hima ergaa qabu tokko barreessuun ergaa ofii ittiin dabarsabidha jedha. Kaayyoo Barreessuu Akka barreessitoonni gumii qormaata Afaan Oromoo (1996:23) irratti jedhanitti "Afaan tokko Afaan Oromoo dabalatee jecha keessaa sagalee yookaan birsaga tokko jabeessuu, laaffisuu, dheeressuuniifi gabaabsuun hiika jechoota sanaa jijjiira jedhu." Akka yaada barreessitoota kanaatti Afaan Oromoo keessatti jecha tokko yeroo barreessan sagalee ykn birsaga tokko tokko jabeessuun, looffisuun, dheeressuufi gabaabsuun jijjiirama hiika jecha sanaa waan fidhuuf xiyyeeffannoo argachuu akka qabu ibsu. Namni yeroo baay'ee waa barreessu qabiyyeewwan barreessuu sanaa ni xiinxala, ni qindeessa, adeemsa kana keessattis dandeettii waa yaaduu, xiinxaluu fi qindeessuu ni horata.