



**Improving Afaan Oromo Question Answering System: Definition,
List and Description Question Types for Non-factoid Questions**

A Thesis Presented

By

Endale Daba

To

The Faculty of Informatics

Of

St. Mary's University

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

In

Computer Science

July, 2021

ACCEPTANCE

Improving Afaan Oromo Question Answering System: Definition, List and Description Question Types for Non-factoid Questions


By

Endale Daba


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

Thesis Examination Committee:

Internal Examiner

Full Name	Signature	Date
1. Dr. Michael Melese, Advisor	_____	_____
2. <u>Dr. Temtim Assefa</u>	<u>_____</u>  _____	<u>25/07/2021</u>
3. _____	_____	_____

External Examiner

Full Name	Signature	Date
1. <u>Dr. Temtim Assefa</u>	 _____	<u>25/07/2021</u>
2. _____	_____	_____
3. _____	_____	_____

Dean, Faculty of Informatics

Full Name	Signature	Date
1. Dr.Alemebante Mulu	_____	_____
2. _____	_____	_____
3. _____	_____	_____

July, 05, 2021

DECLARATION

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

Endale Daba

Signature

Addis Ababa

Ethiopia

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

Dr. Michael Melese, Advisor

Signature

Addis Ababa

Ethiopia

Exact Date of Defense

July 05, 2021

DEDICATED TO:

This thesis is dedicated to all people who help me into the person I am today,

Especially to my Family.

ACKNOWLEDGMENT

Most of all, I would like to thank God, who makes everything possible, for helping me pass all those hard times that I will never forget in my life.

I owe my deepest gratitude to my advisor Dr. Michael Melese for his time, patience and undeniably helping comments all the way through this study. He really was an inspiration for me to proceed whenever I face difficulties and he is easily approachable. This thesis would not have been possible without his constructive comments on every aspect of the study.

Finally, my earnest gratitude goes to the people who have supported me during my stay in the University and encouragement of my family, friends and relatives are worth being acknowledged.

Table of Content

List of Figures.....	iv
List of Tables.....	iv
List of Acronyms.....	vi
Abstract.....	vii
Chapter One – Introduction 1.....	1
1.1. Overview	1
1.2. Motivations.....	2
1.3. Statement of the Problem	3
1.4. Objective of the study	5
1.4.1. General objective	5
1.4.2. Specific objectives	5
1.5. Research justification.....	5
1.6. Scope and limitation of the study	6
1.7. Methodology.....	6
1.7.1. Research Design	6
1.7.2. Literature Review	7
1.7.3. Data collection	7
1.7.4. Implementation Tool	7
1.7.5. Evaluation.....	8
1.8. Application of the Study	9
1.9. Thesis Organization	9

Chapter Two - Literature Review.....	10
2.1. Overview	10
2.2. Information Retrieval.....	10
2.3. Information Extractor	13
2.4. Question Answering	13
2.4.1. The History of Question Answering	14
2.5. The General Architecture of QAS	16
2.5.1 Question Analysis	17
2.5.2. Document Retrieval	18
2.5.3. Document Analysis	18
2.5.4. Answer Selection	19
2.6. Related Works	20
2.6.1 Amharic QA for List Questions	20
2.6.2 Factoid Question Answering for Afaan Oromo	21
2.6.3 Factoid type QAS for Tigrigna Language	22
2.6.4 Factoid QAS for English Language	22
Chapter Three: Afaan Oromo Language	25
3.1. Basics of Afaan Oromo Language	25
3.2. Afaan Oromo Alphabets and Writing System.....	27
3.3. Afaan Oromo Punctuation Marks.....	29
3.4. Afaan Oromo Part of Speeches	29
3.5. Word Categories in Afaan Oromo	31

3.6. Abbreviations in Afaan Oromo	38
3.7. Afaan Oromo Questions.....	38
3.8. Afaan Oromo Morphology	38
3.8.1. Types of morphology in Afaan Oromo	38
3.9. Summary	41
Chapter Four: System Design and Implementation.....	42
4.1. Overview	42
4.2. System Architecture	42
4.3. Document Preprocessing	44
4.4. Question Analysis	47
4.4.1. Question Classification	48
4.4.2. Query Generation	50
4.5. Document Selection	51
4.5.1. Document Retrieval	51
4.5.2. Document Analysis	52
4.6. Answer Extraction	52
4.6.1. Definition Description Answer Extraction.....	52
4.6.2 List Answer Extraction.....	56
4.7 Summary	57
Chapter Five: System Evaluation and Results	58
5.1 Overview	58

5.2 Experimentation Environment.....	58
5.3 Evaluation Criteria.....	58
5.3.1. Question Classification Evaluation.....	59
5.3.2. Document Selection Evaluation.....	59
5.3.3. Answer Extraction Evaluation.....	61
5.4. Discussion.....	66
Chapter Six: Summary, Conclusion, Recommendations and Future Work.....	68
6.1. Summary.....	68
6.2. Conclusion.....	69
6.3. Contribution.....	70
6.4. Recommendations and Future Works.....	70
References.....	74
Appendices.....	78
Appendix 1: some of Afaan Oromo Short words and their Expansion.....	79
Appendix 2: List of place names.....	79
Appendix 3: Sample Test Questions and their Question Type.....	79
Annex.....	85

List of Figures

Figure 2.1: The General Architecture of QAS-----	16
Figure 4.1: The Architectural Design of Afaan Oromo List, Description and Definitional QA System-----	38
Figure 4.2: Architecture of Afaan Oromo Analyzer Component-----	39
Figure 5.1 Screenshot of Correct List Answer Example-----	58
Figure 5.2 Screenshot of Correct Description Answer Example-----	59
Figure 5.3 Screenshot of No answer Description Example-----	60
Figure 5.4 Screenshot of No answer Definition Example-----	61

List of Tables

Table 3 plural noun formation using suffix-----	29
Table 3.1 plural noun formation using numerals-----	29
Table 3.2 definiteness form of nouns-----	29
Table 3.3 derivational nouns formation-----	30
Table 3.4 personal pronouns that can represent subject-----	30
Table 3.5 personal pronouns that can replace object-----	31
Table 3.6 Afaan Oromo Possessive pronouns-----	31
Table 3.7 adjective inflection for number and gender-----	33
Table 3.8 Afaan Oromo inflectional morphology-----	36
Table 3.9 Afaan Oromo derivational morphology-----	36

List of algorithms

Algorithm 4.2: Rule Based Question Classification Algorithm-----	49
Algorithm 4.2: Query Generation-----	51

Acronyms

IAOQAS:	improvement of Afaan Oromo Question Answering System
QAS:	Question Answering System
IR:	Information Retrieval
NLP:	Natural Language Processing
AI:	Artificial Intelligence
RBQ:	Rule Based Question
AOQAS:	Afaan Oromo Question Answering System
POS:	Part of speech
NER:	Named entity recognizer
IE:	Information extraction
TREC:	Text Retrieval Conference
KE:	Keyword Extractor
DDAE:	Definition Description Answer Extraction
API:	Application Programming Interface
AVG:	Average
LAE:	List Answer Extraction
QA:	Question Answering
TREC:	Text Retrieval Conference
URL:	Uniform Resource Locator

Abstract

Question Answering (QA) can go beyond the retrieval of relevant documents, it is an option for efficient information access to such text data. The task of QA is to find the accurate and precise answer to a natural language question from a source text. The existing Afaan Oromo QA systems handle questions that usually take named entities as the answers.

A different type of Afaan Oromo Question answer such as list, definition and description. The goal of this study is to propose approaches that tackle important problems in Afaan Oromo non-factoid QA, specifically in list, definition and description questions. The proposed QA system comprises of document preprocessing, question analysis, document analysis, and answer extraction components.

Rule based techniques are used for the question classification. The approach in the document analysis component retrieves relevant documents and filters the retrieved documents using filtering patterns for list, definition and description questions a retrieved document is only retained if it contains all terms in the target in the same order as in the question. The answer extraction component works in type by type manner.

The extracted sentences are scored and ranked, and then the answer selection algorithm selects top 5 non-redundant sentences from the candidate answer set. Finally the sentences are ordered to keep their coherence.

The system is tested using evaluation metrics and used percentage ratio for evaluating question classification which classified 98.3% correctly. The document retrieval component is tested on two data sets that are analyzed by a stemmer and morphological analyzer. The F-score on the stemmed documents is 0.729 and on the other data it set is 0.764. Moreover, the average F-score of the answer extraction component is 0.592.

Keywords: Non-factoid Question-Answering, Afaan Oromo Question Answering System, Description Question types, Question Classification, Document Filtering, Sentence Extraction, Answer, Selection, RuleBased.

Chapter One

Introduction

1.1. Overview

The Oromo people have their own language which is called Afaan Oromo or Oromiffa [1]. It is a Cushitic language spoken in most parts of the Ethiopian Empire and northern Kenya. It is considered one of the five most widely spoken languages from among the approximately 1000 languages of Africa [2].

The Afaan Oromo language has very rich vocabulary and it is the third most widely spoken languages in Africa, surpassed only by Arabic. It is the most widely spoken language in the family of Cushitic branch, and the most populous language of Ethiopia. It is different from other languages in morphological properties, patterns of word synthesis and grammatical rules [3]. With the rapid growth of internet, non-factoid Question-Answering tasks are in high demand. The amount of information available on an Internet is growing at alarming rate from time to time [4]. Processing this large amount of information is difficult and time consuming using the traditional approach from the practical point of view as it requires the state-of-the-art techniques in various fields, such as Information Retrieval (IR), Natural Language Processing (NLP), and Artificial Intelligence (AI) [5].

NLP has many application areas such as: Information Retrieval, Information Extraction, Machine Translation, Text Summarization, and Question Answering [4]. Internet is a place where human beings seek for information, ask questions and accordingly a system provides answer for the given query by searching from different source using information retrieval [4].

Among the NLP application areas, Question Answering System (QAS) provides an answer for natural language questions rather than a linked list of documents [3]. Question answering is a challenging task in general. The main goal of a question answering system is to give a precise answer to user's queries in natural language. NLP also has links to research in cognitive science, psychology, philosophy and mathematics (especially logic).

Question answering systems categorized into factoid and non-factoid type [6]. Factoid type questions deals with fact finding which includes fact seeking questions i.e. factoid questions,

asking about (who, when, where, what and how much) relates to different types of entities like person, location, organization, time and quantity. There are different question types such as acronym, counterpart, definition, biography, description, famous, stand for, synonym, why, name-a, name-of, where, when, who, what/which, how, yes/no and true/false. Where, when, which, yes/no, true/false, and name of are kinds of factoid questions whereas definition, description, list, biography are non-factoid questions.

Another category of Question Answering Systems are mainly classified as Open and Closed Domain [7]. Open domain QA is responsible for handling large amounts of data and wide range of questions designed to answer questions about everything. Google, Wikipedia, etc. are examples of such systems. Unlike open domain closed domain systems are designed to handle questions and answering in a give specific domain.

Natural Language Processing is the core of any Question Answer system. Natural Language Processing is the core of any Question Answer system. Through a number of modules build an efficient Question Answering system which adds to the complexity of such a system [7, 3]. The aim of a Question Answer System is basically to allow a user to ask a question in everyday language and receive an answer in user comprehensible format.

1.2. Motivations

We are always in a quest of Information. However, there is difference in information and knowledge. Information Retrieval or web search is mature and we can get relevant information at our finger tips. Question Answering is a specialized form of information retrieval which seeks knowledge. We are not only interested in getting the relevant pages but we are also interested in getting specific answer to queries. Question Answering in itself is an intersection of Natural Language Processing, Information Retrieval, and Rule Based Representation.

Currently some natural language processing applications are being developed for Afaan Oromo language. As it becomes more and more difficult to find answers on the WWW using standard search engines question answering technology will become increasingly important. Natural Language Processing is a theoretically motivated range of computational Techniques for analyzing and representing naturally occurring finds answers to question types of:

Definition, List, and Description at one or more levels for analysis for the purpose of achieving language processing for a range of tasks [8].

Question answering system in its being is an art, at the same time it has science in its essence. Question Answering Systems are needed everywhere. It is necessity in every aspect where we need some assistance from computers. So, it is worth exploring the exiting field of question answering.

1.3. Statement of the Problem

Afaan Oromo is the official language of Oromia regional state of Ethiopia. The language Afaan Oromo is spoken by more than 40 million peoples and most of native speakers are people living in Ethiopia, Kenya, Somalia and Egypt [2]. It is third largest language in Africa following Kiswahili and Hausa; 4th largest language, if Arabic is counted as Africa language. Currently Afaan Oromo is a language that is spoken by large number of speakers (40% of the total population in Ethiopia) and some countries of East Africa [9]. It is a Latin based script called Qubee which has 37 basic characters [9]. Afaan Oromo is Cushitic language which is a family of Afro Asiatic languages.

Therefore, each question needs special consideration to return the correct answer according to their languages question answering techniques. In the idea of question answering system, a lot of researches have been done worldwide on QA in many languages. In the case of Ethiopia, the field needs to be exploited more as Ethiopia is the home for speakers of more than 80 languages [10]. For local Languages, there are researches related to this task in different domains and question types. Some of the local works related to Afaan Oromo question answering systems are: Afaan Oromo List, Definition and Description Question Answering System by Chaltu Fita [11], Amharic Question answering for Factoid Question by Seid Muhie [12], Web based Amharic Question Answering systems for Factoid Questions by Desalegn Abebaw [10], Amharic Question Answering for Definitional, Biographical and Description Questions by Tilahun Abdissa [13], Afaan Oromo Question Answering System for Factoid Questions by Aberash Tesfaye [1], Definition Question answering system for Afaan Oromo Language by Dejene Hundesa [34]. These works contribute their role for the Afaan Oromo Language to handle the described problems even if their accuracy and efficiency should be

further improved. Since Afaan Oromo is spoken and written by large number of people, the grammatical structure of the language is different from other languages and it has a very complex rich morphology the development of the language should be supported with a technology, therefore it is logical to study the Question Answering System of Afaan Oromo language. The problem that this research work tries to address is how to design and develop Afaan Oromo Question Answering System, Definition, List and Description Question Types. To fill the gap identified An effective definition extraction approach to extract answer for definition, questions from Afaan Oromo documents, factors affect the performance of definition question answering system, the language dependent components of Afaan Oromo definition question answering system are the gap filled.

In our country, Ethiopia, the numbers of electronically generated Afaan Oromo documents is increasing on an increasing rate as researches, historical documents, fictions, magazines and many newspaper publishers started providing their works electronically. On the other hand, asking questions is the very nature of human beings, so peoples need answer for their question that is found in the documents. This gap will raise need to have some system which can understand questions and then look for an answer in the knowledge base and give the direct answer for the question.

Thus, there is a need to develop question answering system for definition, list and description types for Afaan Oromo language. In attempt to develop a question answering system for Afaan Oromo language, the research questions have been formulated. The following research questions are to be answered to fill the gap identified above:

The following research questions are to be answered to fill the gap identified above:

RQ1. Which question answer technique is effective for Afaan Oromo QA system?

RQ2. To what degree the performance of Afaan Oromo QA for list, Description and definition system is achieved?

RQ3. What are the language dependent components of Afaan Oromo definition question answering system?

1.4. Objective of the study

To achieve the objective of the study the following general and specific objective are set for Question Answering system for Afaan Oromo language.

1.4.1. General objective

The general objective of this research is to design and develop Afaan Oromo Question Answering System, definition, list and description question types.

1.4.2. Specific objectives

- ✓ To collect and prepare a representative dataset for Afaan Oromo Question Answering System.
- ✓ To design algorithm suitable to identify adding question and answer from user natural language.
- ✓ To develop a general architecture of AOQAS for Question Answering System, Definition, List and Description Question Types,
- ✓ To develop a prototype for Afaan Oromo Question Answering System, Definition, List and Description Question Types and evaluate its performance,
- ✓ To evaluate the performance of Question answer system and,
- ✓ To report the finding and forward the recommendation

1.5. Research justification

In this very changing world new technologies have been emerged dramatically and the information need of the people. Therefore almost in every discipline people are using automated systems that generate information in electronic format in different natural languages. This has brought a problem of dramatic increase in information and it has become a major issue in the field of information management. However in order to facilitate accessing huge amount of information automated systems that use Information Retrieval (IR) technique is needed.

The existing Question Answering System, definition, list and description question types

developed for Afaan Oromo language are not capable in providing design and develop of Afaan Oromo Question Answering System, definition, list and description question types. The proposed method is intended to bring design and develop over the existing Afaan Oromo Question Answering System, definition list and description question types, so that a better information retrieval system can be built.

1.6. Scope and limitation of the study

Scope

In regarding the scope of the research, this question answering system have been designed for Afaan Oromo language since it is familiar for the researcher and it is the regional language of the Oromia state. The type of question to be addressed are Non-factoid type, i.e., specific to Afaan Oromo list, definition and description.

Limitation of the study

Lack of large size document corpus that are prepared using Afaan Oromo language is the major limitation of this research. Collecting possible answers for the prepared question are another problem because some question might be difficult to know the exact answer by the researcher.

1.7. Methodology

Methodology delivers an understanding of the way a research is conducted and studied. In order to accomplish the general and specific objectives of this study, different methodologies have been applied.

1.7.1. Research Design

This study is an experimental quantitative method research approaches employed to Afaan Oromo Definition Question Answering system. It is used surface pattern based design approach to extract concept description pairs from the Afaan Oromo language documents by adding question answer using the system.

1.7.2. Literature Review

To understand problem, gap and state-of-the-art techniques for developing QA systems, different literature books, journal article, Internet publications and other scholarly published materials reviewed for the purpose of understanding. The review helped to understand QA approaches and techniques, as the research focuses on Afaan Oromo language, different language specific features and properties have been studied in light with QA systems for definition, description and listing questions.

1.7.3. Data collection

The necessary data or documents for this research were collected from different Afaan Oromo newspapers on the web (Bariisaa, Bakkalcha and Oromiyaa) and other official websites to evaluate the document retrieval and answer extraction and report related data. The documents that are collected from difference source were selected and processed to be suitable for definition, description and list type questions. 2700 question-answer pair's datasets are prepared to evaluate our system.

1.7.4. Implementation Tool

In order to successfully achieve the research objectives different methods and tools engaged in the process. In order to design and develop Afaan Oromo Question Answering System for definition, list and description question types as a developmental tool Java programming language is the major developmental tool for the prototype whereas Apache lucene [14] used for indexing and relevant document retrieval task method implemented for the question classification task and stemmer for stemming [15] and morphological analyzer [16] for lemmatization used.

Java is one of the important programming languages and computing platform for many applications. It's released by Sun Microsystems in 1995. Many applications use Java specially web application high potential in the field of programming web, games, databases, and many other applications. It has many compilers and editors such as text pad, Eclipse platform, and NetBeans platform.

The reason to use Java is on one hand, the exposure of the researcher to the language and on the other hand, due to the fact that more Web-based tools that can be used for this research are based lucene, an open source based API, is utilized for its indexing and searching capability. Lucene is a high performance, scalable Information Retrieval (IR) library. It helps to add indexing and searching capabilities to applications. Lucene is a mature, free, open-source project implemented in Java. It is a member of the popular Apache Jakarta family of projects, licensed under the liberal Apache Software License [14]. Lucene can index and make searchable any data that can be converted to a textual format.

1.7.5. Evaluation

The performance of Afaan Oromo Question Answering System, definition, list and description question types system done by collecting list, definitional and description questions types and evaluate the systems performance avoid manual answers.

RECALL: It is one of the metrics used to evaluate the performance of the QAS. It is calculated as by dividing the number of relevant records retrieved to the total number of relevant records in the corpus or database depending on the structure of the data from which the records are retrieved.

PRECISION: Precision is also one of the standard metrics used to evaluate the performance of QAS. Its value is calculated by dividing the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved.

F-measure: is used to evaluate the performance of the system by considering both the precision and recall of the system (it is taken as the weighted average of precision and recall).

Although the criteria listed above valuable to evaluate the answer of QAS, in some cases optimizing along one criterion may reduce the goodness along another criteria. For instance if the answer is judged for answer justification criteria and fulfils it, it may reduce the answer conciseness when judged by the answer conciseness criteria. Therefore the intended use of QAS, the intended user and the interface of QAS should be considered in evaluating the answer of QAS. Manual evaluation of an answer of QAS is one of the main approaches used for judging the correctness of an answer to natural language question. In this approach a team

of assessors manually judge the correctness of the answer. TREC adopted manual evaluation of an answer to judge the correctness of the answers which has been accepted in advance by several QAS. In this approach a pair consisting of an answer and a supporting document is considered as a system's responses to a natural language question. The system's responses are judged by at least one human expert and one of the four labels: 'correct', 'Unsupported', 'inexact' or 'incorrect' are assigned to the answer. The answer is considered as correct when the answer to a NLQ contains only the relevant information along with the supporting document which enables the user to justify the answer. When the same answer string which was paired with a document that enable the user to justify the answer and taken as 'correct' again paired with other document which does not enables the user to justify the answer would be judged as 'unsupported'. An answer is judged as 'inexact' when a QAS returns an answer string with extraneous words. And finally, when the answer string provided by QAS is not related to the information requested in the question the response would be judged as 'incorrect'.

1.8. Application of the Study

As QA provides short and precise answer to a given natural language question, the Afaan Oromo Question Answering system can be applicable in finding short answers for Afaan Oromo definitional, list and description type of questions from collection of documents. In general, if it is transformed to full-fledged system, an Afaan Oromo Question Answering System can be applicable for different real world applications such as automated customer services, driving direction system, reservation system and giving online help in the absence of information desk personnel.

1.9. Thesis Organization

This thesis is organized into six chapters. The first chapter discusses basic concept of question answering and Afaan Oromo language and statement of the problem. It also presents general and specific objectives of the study, methodology of the study, scope and limitation of the research and significance of the study. Chapter two address literature reviews on question answering and essential elements of QA. Moreover, brief review techniques related to QA, general framework of QA, types of QA, related work global and local researches are

discussed. Chapter three addresses about Afaan Oromo language. Chapter four describes the Design and Implementation of the proposed of the system. The Experiment and Evaluation of the system is discussed in chapter five. Finally, conclusions drawn from the thesis result, the contributions of this research work and recommendations on possible future works related to this research are given in chapter six.

Chapter Two

Literature Review

2.1. Overview

This chapter starts by discussing about Information Retrieval (IR), Information Extractor , Question Answering, The History of Question Answering, general architectures and particularly on techniques and approaches in question analysis, document retrieval, document analysis and answers extraction components of a QA system, Morphological Analysis, and Afaan Oromo Language.

2.2. Information Retrieval

Information retrieval (IR) has most usually been constructed as the problem of selecting texts from a database in response to some more or less well specified query [4]. Information Retrieval (IR) is the discipline that deals with retrieval of unstructured data, especially textual documents, in response to a query, which may itself be unstructured like sentence or structured like Boolean expression [17]. Search depends on query users and full text search, Meta data, indexing, etc. The goal of IR is to retrieve accurate documents contain results and ranked depending on user's queries [4].

When the query passed to the IR system, the system search in one corpus or many corpora to find the relevant document(s), that matches the phrase or key words in the query then the IR system must return Ranked the documents depending on many criteria's such as: matching full phrase, matching key words, number of key words matches, etc. and filtering the documents to find the repeated documents to decrease the number of candidate documents. The main problems of dealing with IR systems stated by [18] in three main points as follows:

- ✓ The search engine retrieves the relevant documents but not the exact answer, this means you need to search inside the documents for answers.
- ✓ The quality of data because users of social media explains their opinions.

Indexing: Not all the pieces of an information item are equally significant for representing its meaning [4]. In written language, for example, some words carry more meaning than others. Therefore, it is usually considered worthwhile to pre-process the information items to select the elements to be used as index objects. Indices are data structures constructed to speed up search. It is worthwhile building and maintaining an index when the item collection is large and semi-static. The most common indexing structure for text retrieval is the inverted file. This structure is composed of two elements: the vocabulary and the term occurrences. The vocabulary is the set of all words in the text. For each word in the vocabulary a list of all the text positions where the word appears is stored. The set of all those lists is called occurrences.

Query processing: The user needs, the query, is parsed and compiled into an internal form. In the case of textual retrieval, query terms are generally pre-processed by the same algorithms used to select the index objects. Additional query processing (e.g., query expansion) requires the use of external resources such as thesauri or taxonomies.

Searching: user queries are matched against information items. As a result of this operation, a set of potential information items is returned in response to user needs. The way this is achieved may vary considerably depending on the format of information (text, audio, video, etc.), but in all cases, some form of simplification is done in the information model to make it tractable. For instance, text retrieval commonly builds on the assumption that the matching between information items (the documents) and user information needs (the query string) can be based on a set of index terms. This obviously involves a loss of semantic information when text is replaced by a set of words. A similar situation occurs in multimedia retrieval where matching is performed based on numeric signal features.

Ranking: The set of information items returned by the matching step generally constitutes an inexact, by nature approximate answer to the information need. Not all the items contain relevant information to the user. The ranking step aims to predict how relevant the items are comparatively to each other, thus returning them by decreasing the order of estimated relevance. Thus, in a way, ranking algorithms can be considered the core of IR systems, as they are keys to determine the performance of the system.

Information Retrieval Models:-

The ranking algorithm is one of the main characteristic components of an IR system. A ranking algorithm operates according to basic premises regarding the notion of document relevance. Distinct sets or premises yield different IR models [3]. This section tries to cover three of the most important classic text IR models, namely: Boolean, Vector, and Probabilistic. In the Boolean model documents and queries are represented as a set of index terms. In the Vector space model documents and queries are represented as vectors in a t -dimensional space. In the basic probabilistic model, documents and queries representations are based on probability.

Boolean Model:-

The Boolean Model is a simple retrieval model based on set theory and Boolean algebra. Documents are represented by index terms extracted from documents, and queries are Boolean expressions on terms. The Boolean model suffers from two major drawbacks. First its retrieval strategy is based on a binary criterion (i.e., a document is predicted to be either relevant or non-relevant), the index terms can only be given Boolean weights i.e. $\in \{0, 1\}$. This means that with too restrictive expressions no documents will qualify. On the other hand a very general expression will result in too many documents being returned. Therefore, it does not provide a proper basis for ranking the retrieved results, which May likely result in low precision levels when the retrieval space is too big. Second, it is not always easy for most users to translate an information need into a Boolean expression with logic operators [19].

Vector Space Model:-

The vector space model VSM recognizes that the use of binary weights is too limiting and proposes a framework in which partial matching is possible. This is accomplished by assigning non-binary weights to index terms in queries and documents. These term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query. By sorting the retrieved documents in decreasing order of this degree of similarity, the VSM takes into consideration documents which match the query terms only partially. The main resulting effect is that the ranked document answer set is considerably more precise than the answer set retrieved by a Boolean model [19].

2.3. Information Extractor

It is the task of automatically extracting structured information from unstructured and/or semi structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Recent activity in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as information extraction [20].

Information Extraction is the part of a greater puzzle which deals with the problem of devising automatic methods for text management, beyond its transmission, storage and display. The discipline of information retrieval (IR) has developed automatic methods, typically of a statistical flavor, for indexing large document collections and classifying documents. Another complementary approach is that of natural language processing (NLP) which has solved the problem of modelling human language processing with considerable success when taking into account the magnitude of the task. In terms of both difficulty and emphasis, IE deals with tasks in between both IR and NLP. In terms of input, IE assumes the existence of a set of documents in which each document follows a template, i.e. describes one or more entities or events in a manner that is similar to those in other documents but differing in the details.

2.4. Question Answering

The task of returning a particular piece of information in response to the users' query is called Question Answering [4]. The task is termed factoid question answering if the answer is a simple fact such as a location or a date or a person. However, most interesting questions are not factoid questions, and hence we concentrate on non-factoid QA. Focused Summarization is another term used for this task, where we try to summarize multiple documents for the purpose of generating an answer to a user's query. We use these two terms interchangeably as they are synonymous for the problem at hand. A question answering is a task that aims to automatically give answers to questions described in natural language [21]. It allows users to have exact answer rather than having list of potentially relevant documents. The traditional search engine focuses on retrieving related documents and returns list of related documents for the users and users must scan to get the necessary information. Whereas, QA system answers the question in the form of exact answer which is extracted from source documents. QA system needs more

complex natural language processing (NLP) tools for precisely understanding the user's intention as well as to extract correct answers. But, in the case of IR, a simple technique is sufficient to return content-rich documents. In recent time, the automatic question answering system has become an interesting research field and resulted in a significant improvement in its performance which has been largely driven by the TREC (Text Retrieval Conference) QA Track [21]. Question answering system is two form factoid and non-factoid questions. Factoid questions return answers in the form of a name of a person, name of a country, name of an organization, quantity of something and date or time on which something happens. On the other hand, non-factoid questions are questions that ask for definitions, reasons, biography, methods, and procedures. The answers for non-factoid type of questions are more complex than factoid questions.

2.4.1. The History of Question Answering

The interest in natural language was identified in 1665 by Simmons by his paper review which was entitled as "Answering English Questions by Computer" natural language question answering had got a great attention since in the beginning of the Question Answering track in the Text Retrieval Conferences in 1999.

In 1999 the first Question Answering task in TREC 8 (Text Retrieval Conference) revealed an increasing need for more sophisticated search engines able to retrieve the specific piece of information that could be considered as the best possible answer to the user question. Such systems must go beyond documents selection, by extracting relevant part of them. The problem intersects two domains Information Retrieval (IR) and Natural Language Processing (NLP). IR is improved by integrating NLP functionalities at a large scale.

Some of the early AI systems were question answering systems. Two of the most famous QA systems of that time are BASEBALL and LUNAR, both of which were developed in the 1960s. BASEBALL answered questions about the US baseball league over a period of one year [4]. LUNAR in turn answered questions about the geological analysis of rocks returned by the Apollo moon missions. Both QA systems were very effective in their chosen domains. In fact, LUNAR was demonstrated at a lunar science convention in 1971 and it was able to answer 90% of the questions in its domain posed by people untrained on the system. Further

restricted- domain QA systems were developed in the following years.

The common feature of all these systems is that they had a core database or knowledge system that was hand-written by experts of the chosen domain. Some of the early AI systems included question-answering abilities. Two of the most famous early systems are SHRDLU and ELIZA. SHRDLU simulated the operation of a robot in a toy world (the blocks world), and it offered the possibility to ask the robot questions about the state of the world.

The history of question answering system started in 1961 BASEBALL and in 1973 LUNAR QAS [22]. BASEBALL was a program for answering questions about baseball games played in the American league over only one season. Given a question such as who did the Red Sox lose to on July 5? Or how many games did the Yankees play in July? BASEBALL analyzed the question, using linguistic knowledge, into a canonical form which was then used to generate a query against the structured database containing the baseball data. The second QAS was LUNAR designed to enable a lunar geologist to conveniently access, compare and evaluate the chemical analysis data on lunar rock and soil composition that was accumulating as a result of the Apollo moon mission. LUNAR could answer questions such as what is the average concentration of aluminum in high alkali rocks. Or how many Brescia's contain Olivine? The system was able to answer 90% of the in-domain questions posed by working geologists, without prior instructions as to phrasing.

The best known early question answering system is a Program for answering questions about BASEBALL games (BASEBALL) played in the American league over one season. Green et al (1961) has developed a QA system BASEBALL which is the first of a series of programs designed as natural language front ends to databases. BASEBALL is restricted to questions about baseball facts and most QA system is used for a long time. It is restricted to structured database model. In this closed domain QA, users are allowed to input queries in natural language then the interface is used to analyze the syntaxes and meanings by using linguistic knowledge. The most well remembered other early work in this tradition is the Application to LUNAR geology (LUNAR) system. LUNAR is designed to enable a lunar geologist to conveniently access, compare and evaluate the chemical analysis data on lunar rock and soil composition that is accumulating as a result of the Apollo moon mission.

TRECs question answering track which is motivated in the field of question answering. The initial efforts in question answering is focused on fact-based and short answered questions [22]. It is aimed for comparing IR systems implementation by academic and commercial research groups. The TRECs systems are run by pre-selected queries and retrieved text documents. The results are evaluated manually. This type of TRECs system is used by shallow NLP technique based on pattern matching algorithm.

2.5. The General Architecture of QAS

Recently, a number of Question Answering systems were submitted in each evaluation task of QA, such as TREC, CLEF and NTCIR etc. In 2002, 2003 and 2004, 34, 16 and 28 research groups participated in the question answering track of the annual TREC respectively, each group having implemented their own system [4].

These systems cover a wide spectrum of different techniques and architectures which are impossible to capture all variations within a single architecture. Most of the time, question answering system has four basic components such as question analysis, document retrieval, passage retrieval and answer extraction. The prototypical system has four components: question analysis, document retrieval, document analysis, and answer selection.

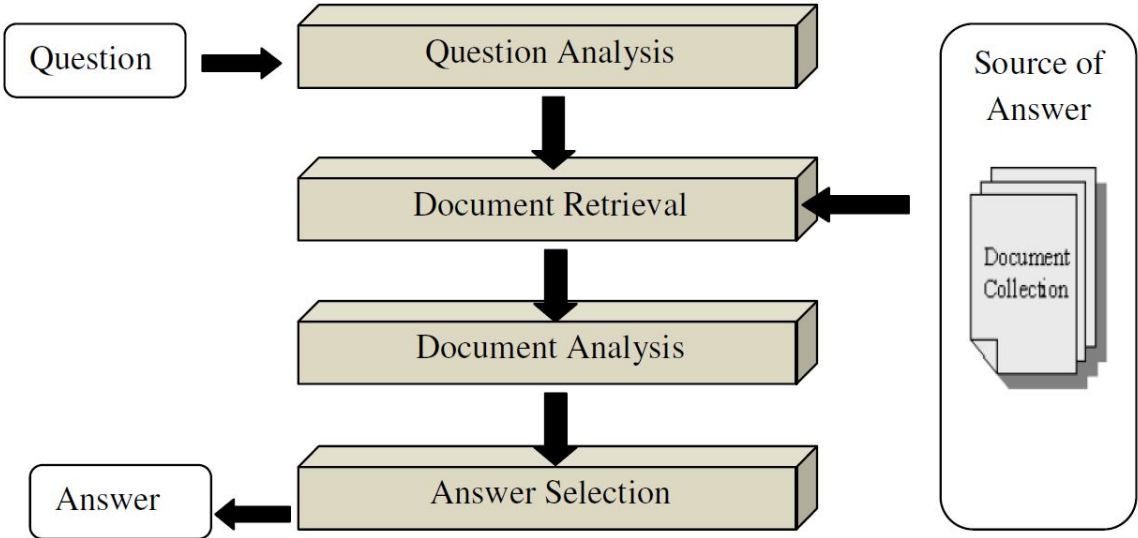


Figure 2.1: The General Architecture of QAS [4]

2.5.1 Question Analysis

Analyzing the natural language question provided as input to the system is the first step toward finding the answer. Question analysis is the process of constructing representation of questions, deriving of expected answer types, and extracting of keywords [23]. In the question analysis stage, the type of question will be analyzed. The question type further illustrates what will be the expected answer type. It is the question analysis stage that is also responsible for constructing proper query for the IR component of the QA system. Correctly identifying the expected answer type will help the later stage of answer extraction to correctly identify answers. Therefore, wrong question analysis means that the document retrieval component will retrieve wrong documents as well as the answer extraction component will extract wrong answer or no answer. The question analysis component has two sub-components: question classification and query generation.

Question Classification: is the process of putting the questions into several categories. The set of possible classes is predefined, and ranges from few basic sets only depending on looking at the key question word, such as: Enyuu dha? / Enyuu Isheen? / Enyuu isaan? Maal jechuu dha? Faayidaan Isaa maali? The accuracy of question classification is very important to the overall performance of the Question Answering system. Thus, most systems resort to more detailed analysis of the question which determines additional constraints on the answer entity.

Classification of questions can be implemented in a variety of ways. The simplest method is to use a set of rules that map patterns of questions into question types. The patterns are expressed by means of regular expressions on the surface form. The identification of the answer type is usually performed by analyzing the interrogative terms of the question wh-terms. For example, given the question: Qu'anoo fi Qoraanno maal jechuu dha? Maal jechuu dha? Indicates that the question is looking for a definition.

Query Generation: Question-analysis also extracts the information in the question that allows the generation of queries that could be used by an IR system for the extraction of the answer-bearing text from the entire document collection. These queries are commonly obtained using keyword selection and answer-pattern generation processes.

2.5.2. Document Retrieval

QA architecture is chosen, answering questions over a closed corpus or even the web almost always involves some kind of searching for and retrieval of documents as a first step to narrow the search space for the answer to the question[3].

The function of the document retrieval component is not to find actual answers to the question, but to identify documents that are likely to contain an answer [4]. Document retrieval aim to return relevant documents to a user's query, where the query is a set of keywords. A document is considered relevant if its content is related to the query [24]. The main purpose of the document retrieval component is to select an initial set of candidate answer-bearing documents from a large text collection prior to sending them to a downstream answer extraction module. In an effort to pinpoint relevant information more accurately, the documents are split into several passages, and the passages are treated as documents. Thus, many QA systems also have a passage retrieval stage, interposed between the document retrieval and answer extraction components, which can be thought of as a second, smaller scale IR module. Using a passage-based retrieval approach instead of a full-document retrieval approach has the additional advantage that it returns short text excerpts instead of full document which are easier to process by later components of the question answering system. Document retrieval has a long tradition and many frameworks have been developed over the years, resulting in sophisticated ways to compute the similarity between a document and a query. Depending on the retrieval engine that is actually used, the retrieval component returns either an unordered set of documents that are likely to contain an answer, or a ranked list of documents, where the documents are ranked with respect to their likelihood of containing an answer. Document retrieval effectiveness is critical to the overall performance of a question answering system. If the document retrieval component fails to return any document that contains an answer, even optimally functioning answer extraction and answer selection components will inevitably fail to return a correct answer to the user [4].

2.5.3. Document Analysis

Once candidate answer-bearing documents or document passages/segments have been selected, these text segments may then be further analyzed [25]. Sentence/Passage extraction

can be performed by segmenting each document into small sentence/passage and selects suitable sentence/passage related to keywords. In segmenting the set of relevant documents, in order to detect sentence in a document, punctuation marks can be used as separators. In detecting paragraphs of a document, empty lines can be used as separators. So, in this way from the set of candidate documents the set of candidate sentences/passages which are supposed to contain the candidate answers are retrieved. Once candidate answer-bearing documents or document passages/segments have been selected, these text segments may then be further analyzed. The document analysis component searches through the documents returned by the retrieval component to identify phrases that are of the appropriate type, as specified by the question analysis component.

The selected documents or document portions using at the very least a named entity identifier, which recognizes and classifies multiword strings as names of includes person, organization, dates, locations, temporal and spatial distances, etc. At this stage, there are a number of ways to further analyze documents. Such as sentence splitting, part-of-speech tagging, and chunk parsing. In order to establish an explicit link between a phrase of the appropriate type and the question, the syntactic structure, pattern matching, or lexical chaining, then linear proximity is often used.

2.5.4. Answer Selection

The final component of answer selection, the representation of the question and the representation of the candidate answer-bearing texts are matched against each other and a set of candidate answers is produced, ranked according to likelihood of correctness. Usually originate from different passages and are often extracted using different strategies. Moreover, these textual fragments may not always constitute full answers. The set of answer candidates obtained through answer extraction could include [26]: Incorrect: the answer string does not contain a correct answer. Not Supported: the answer string contains a correct answer but the document returned does not actually answer the question. Not Exact: the answer string contains a correct answer and the document supports that answer, but the string contains more than just the answer or is missing bits of the answer. Locally Correct: the answer string consists of exactly a correct answer that is supported by the document returned, but the document collection contains a contradictory answer that the assessor believes is better.

Globally correct: the answer string consists of exactly the correct answer, that answer is supported by the document returned, and the document collection does not contain a contradictory answer that the assessor believes is better.

Answering definitional, biographical, how, why and other complex questions require to put together partial answers from different documents, in which this task is handled by the answer generator. So, answer generation is about taking a candidate answer and produce correct and complete answers with corresponding confidence scores. This task involves combining evidence from several answer candidate components in order to generate meaningful correct answers with high confidence scores [26]. Some question answering systems use the whole document collection to count how often a candidate answer co-occurs with terms from the question. Other systems even go beyond the actual document collection and use the World Wide Web to get these frequencies.

2.6. Related Works

Many QAS have been developed by researchers in different language. The information contained in a fixed size of corpus and also the web is used for extracting the answer. Different researchers have attempted to develop QAS in foreign language like English, Chinese, Arabic, Spanish and the like and also some attempts have been done to develop QAS for local languages. We presented QAS for Local Languages and QAS for Foreign Languages

2.6.1 Amharic QA for List Questions

Another Amharic QAS was done on Amharic Question Answering for list questions: A case of Ethiopian tourism by Brook Eshetu Bete [27] which was focused on list questions in closed domain of Amharic question answering system. The aim of the research was to 55 extract a list of answers for the users' question.

The architecture of the system comprises of the modules like answer type retrieval module to identify question type, query interface, candidate answer extraction, co-occurrence of candidate answers and answer type, classification module that divides the answer that are related with the questions as relevant and at the end the answer module returns the answer to the users' question. The researcher had applied the hypothesis which states that answers to a

list questions shares identical semantic class, answers that occur together with in each sentences of the document have relationship to the target and the sentences in the document and the natural language question for which an answer is sought have the same context. The type of answer is identified by answer type identification module by analyzing the type of question posted whether it is list type question or other type. The domain area that was selected by the researcher was the Ethiopian tourism center.

The researcher used recall, precision and F-measure to evaluate the performance of the system. The evaluation was performed independently on the components such as document retrieval, candidate answer extraction, answer type recognition, co-occurrence information extraction and candidate answer selection. Accordingly the researcher reported 100% performance in answer type recognitions, 55% in candidate answer extraction, 54% performance in document retrieval, 61% in candidate answer selection and 100% in co-occurrence information extraction. The researcher has also pointed that the performance of the candidate answer extraction module is less because of the reduced performance of the document retrieval module. Since the candidate answer extraction module depends on the amount of the relevant document returned by the document retrieval module, if no relevant candidate documents are not returned by the document retrieval document the performance of answer extraction module will be poor.

2.6.2 Factoid Question Answering for Afaan Oromo

An attempt to design factoid question answering for Afaan Oromo was done on Factoid Question Answering for Afaan Oromo by K. Abdissa [28]. The objective of his study was to extract fact based answer to user from Afaan Oromo electronic documents collected from Oromia Radio and Television Agency, Fana broadcasting Afaan Oromo service, Online VOA, Magazines prepared in the language like Barisa, Kallacha and Oromia culture and tourism bureaus.

The architecture of the system comprises of the modules like question analysis, answer extraction and IR module. The question analysis module is used to identify answer type identification, IR module is used to extract candidate passages from documents and the answer extraction module is used to extract candidate answer. The researcher also used synonym for

query expansion. Rule based patterns were used for identifying the answer types. To retrieve the documents containing phrases, the researcher has used phrase based indexing for questions that contain phrase?

The researcher has reported that the pattern based answer type identification achieved 92.2%. And also the researcher has pointed out that the system has shown 0.83 recall, 0.71 precision and F-measure of 0.78 and as the researcher has reported the result was encouraging and usage of synonyms and phrase based indexing more improved the performance of the system.

2.6.3 Factoid type QAS for Tigrigna Language

Tigrigna Question Answering System for Factoid Questions by Kibrom Haftu[29] attempted to develop factoid type Question Answering systems for Tigrigna language for selected type of questions, i.e., who, where, when and how many types of question only, and the researcher used the common Question Answering System Architectures, i.e. question analysis, document analysis and answer extraction module. To identify the categories or types of question for the Tigrigna language question, the researcher uses a statistical language model approach. The document analysis module performs the process of preprocessing of parallel corpora, which are documents that contain question sentences in one document and answer sentences in another one, and also ranking and extracting answer contents. Answer extraction also performs the detail analysis on the retrieved answer contents based on the question type, question particle and query using the techniques of language modeling called Answer Model. This statistical language model does the extraction process of exact and precise Tigrigna answer in probabilistic manner from sets candidate answers.

Generally, the researcher uses Moses, GIZA++ and IRSTLM as tools and different Webs and Tigrigna newspapers and magazines as data sources. The researcher uses 1000 data sets for training and 200 data sets for testing. Performance evaluation conducted manually by comparing the system's answers with the answers exists in testing document, which is prepared for testing purpose. He gets 87% of the average performance of the question type classification, 88.5% of the average precision, 85.9% of average recall and 97.2% of the average F-Measure.

2.6.4 Factoid QAS for English Language

The dissertation in factoid question answering systems for English language [30] attempts to design factoid question answering systems for English language. As the researcher identifies, the key challenge in QA is the generation and recognition of inductive signals for answer patterns. So, the dissertation proposed the ideas of feature-driven QA, a machine learning framework that automatically produces rich features from linguistic annotations of answer fragments and encodes them in a compact log-linear models. These features are further enhanced by lightly coupling the question and answer snippets via monolingual alignment.

With the help of modern search engine, database and machine learning tools, the proposed method is able to effectively search billions of facts in the web space and optimize from millions of linguistic signals in the feature space. In this dissertation, the QA is modeled as a pipeline of the form: Question (input) ----->Information Retrieval (search) ----->Answer Extraction (from either text or knowledge base) ----->Answer (output) This dissertation demonstrates the feature-driven approach applied throughout the QA pipeline: the search front end with structured information retrieval, the answer extraction back end from both unstructured data source (free text) and structured data source (knowledge base). Error propagation in natural language processing (NLP) pipelines is contained and minimized. The final system achieves state-of-the art performance in several NLP tasks, including answer sentence ranking and answer extraction on one QA dataset, monolingual alignment on two annotated datasets, and is competitive with state-of-the-art in answering web queries using Freebase.

The work of Aunimo, [28], focuses on question typology and feature sets. In this work, the question typology is very important in mapping question types to answer types. Hence, 18 question classes have been identified which are used for question classification with the help of 700 question sets. The feature sets, that are terms that further help in classifying questions, have been also identified. The research shows techniques that will help automatically extract features from questions so that it can be matched with identified feature sets to successfully construct the question typology.

The paper in [29], attempts to design a simple Question Answering Systems written in Perl that uses the CMU link parser, the REX systems for XML parsing and the Managing Gigabyte search engine. We have discussed this paper in detail because it helps for this research work. In this work, documents are prepared off-line for the pure information retrieval tasks of identifying potentially relevant paragraphs.

Research gap: In our country, Ethiopia, the numbers of electronically generated Afaan Oromo documents is increasing on an increasing rate as researches, historical documents, fictions, magazines and many newspaper publishers started providing their works electronically. On the other hand, asking questions is the very nature of human beings, so peoples need answer for their question that is found in the documents. This gap will raise need to have some system which can understand questions and then look for an answer in the knowledge base and give the direct answer for the question.

Chapter Three:

Afaan Oromo Language

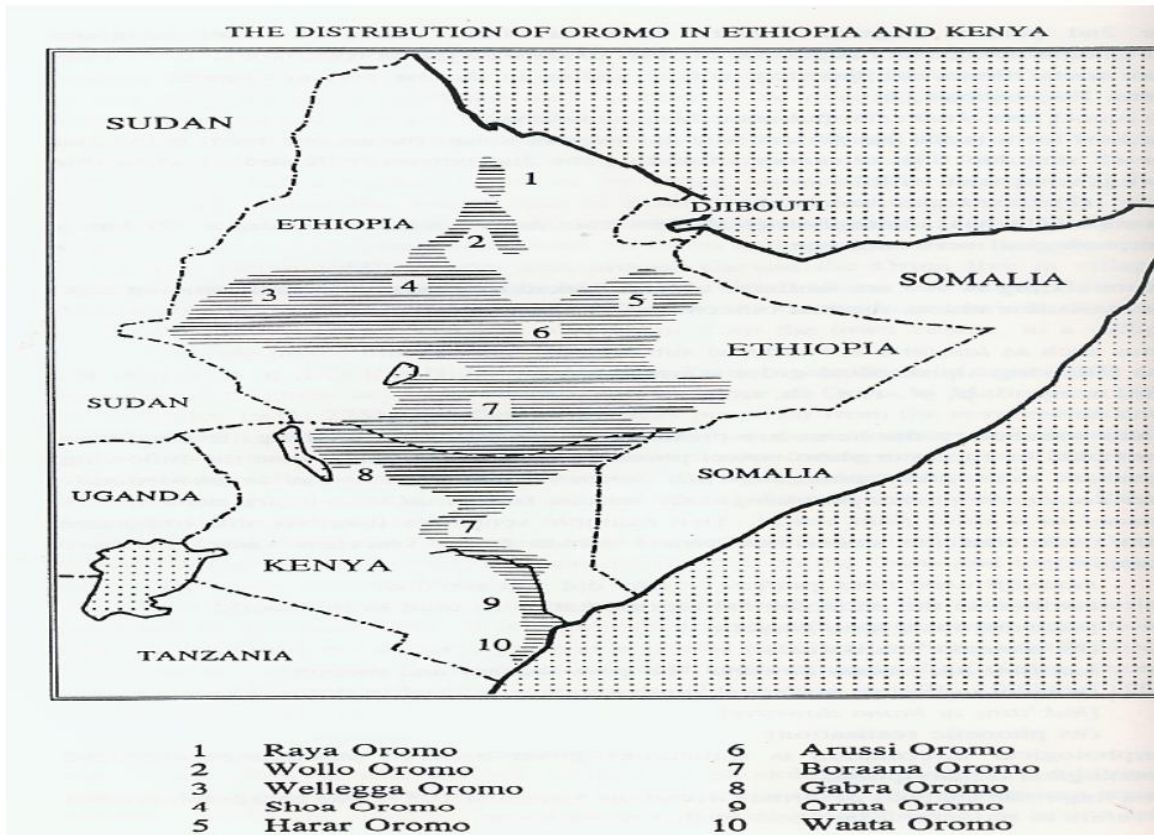
In this chapter the basic structure of Afaan Oromo is presented in order to understand the nature of the language which helps in designing the proposed prototype. The language's nature like to what extent the language is spoken, application area of the language (areas where the language is used like newspapers, in the different offices of Oromia Regional states, in different research publications, in higher educational institutes, etc.), how words are formed, morphological nature of the language and other important features of the language that are specifically important for this thesis are discussed under this chapter.

3.1. Basics of Afaan Oromo Language

Afaan Oromo is an Afro-Asiatic language, and the most widely spoken of the Cushitic subphylum. It is widely used as both written and spoken language in Ethiopia and some neighboring countries, including Kenya and Somalia. Besides being a working language of Regional Government of Oromia, it is the instructional medium for primary and junior secondary schools throughout the region. Moreover, a number of literatures, newspapers, magazines, educational resources, official documents and religious writings are written and published in Afaan Oromo [31]. There are also television and radio programs on which information in Afaan Oromo is being broadcasted in Ethiopia. These include Ethiopian Television (ETV), Oromia Television (TV Oromia), Ethiopian Radio, and Radio Fana. Before 1991, Ge'ez script was used for writing Afaan Oromo documents. A Latin-based alphabet called Qubee has been adopted and became the official script of Afaan Oromo since 1991. There are about twenty-six consonants and ten vowels (five short and five long) in the language [31].

The Oromo belong to the Cushitic group of people. They live in Ethiopia and Kenya [2]. In Ethiopia, the Oromo have an estimated population of 25,488,344, which accounts for 34.5% of the whole population of the country [2]. According to Tesema they are the largest ethnic group in the horn of Africa. The Oromo occupy an area stretching from the Western end of Ethiopia to the Eastern end, and from the Northern end Ethiopia to Southern Kenya. The Oromo speak Afaan Oromo (lit. Oromo Language) which belongs to the Lowland East Cushitic sub-family

of the Afro-Asiatic phylum. Apart from Ethiopia, Afaan Oromo is also spoken in Kenya and Somalia [2]. In Ethiopia, Afaan Oromo is the official language of Oromia Regional State. It is used as a medium of instruction in schools and in the region's Teachers Training Colleges.



Distribution of Oromo in Ethiopia and Kenya [2].

The language belongs to Cushitic language family such as Somali, Sidama, Afar and Geedo which are spoken in Ethiopia [9]. Afaan Oromo is a category of the Lowland East Cushitic group in the Cushitic family of the Afro Asiatic language. It is the most widely spoken language in the families of Cushitic branch (Kula & Varma, 2007).

3.2. Afaan Oromo Alphabets and Writing System

Afaan Oromo is a phonetic language, which means that it is spoken in the way it is written. The writing system of the language is straightforward which is designed based on the Latin script. Unlike English or other Latin based languages there are no skipped or unpronounced sounds/alphabets in the language. Every alphabet is to be pronounced in a clear short/quick or long /stretched sounds. In a word where consonant is doubled the sounds are more emphasized. Besides, in a word where the vowels are doubled the sounds are stretched or elongated. Like in English, Afaan Oromo has vowels and consonants. Afaan Oromo vowels are represented by the five basic letters such as a, e, i, o, u. Besides, it has the typical Eastern Cushitic set of five short and five long vowels by doubling the five vowel letters: „aa“, „ee“, „ii“, „oo“, „uu“. Consonants, on the other hand, do not differ greatly from English, but there are few special combinations such as “ch” and “sh” (same sound as English), “dh” in Afaan Oromo is like an English "d" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins. Another Afaan Oromo consonant is “ph” made when with a smack of the lips toward the outside “ny” closely resembles the English sound of “gn”. We commonly use these few special combination letters to form words. For instance, ch used in barbaachisaa ‘important’, sh used in shamarree ‘girl’, dh use in dhadhaa ‘butter’, ph used in buuphaa ‘egg’, and ny used in nyaata ‘food’. In general, Afaan Oromo has 36 letters (26 consonants and 10 vowels) called “Qubee”.

Qubee (Sagaleewwan) dubbii Afaan Oromo.

a,	b,	c,	d
e,	f,	g,	h
i,	j,	k,	l
m,	n,	o,	q
r,	s,	t,	u,
w,	x,	y,	ch
dh,		ny,	ph
sh, ykn ‘(hudhaa)			

Consonant and Vowel in Afaan Oromo

Like most other Ethiopian languages, whether Semitic, Cushitic, or omotic, Afaan Oromo has a set of ejective consonants, that is, voiceless stops or affricates that are accompanied by glottalization and an explosive burst of air. Afaan Oromo has another glottalized phone that is more unusual, an implosive retroflex stop, "dh" in Afaan Oromo orthography, a sound that is like an English "d" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins. Afaan Oromo has the typical Southern Cushitic set of five short (a, e, i, o, u) and five long vowels, indicated in the orthography by doubling the five vowel letters (aa, ee, ii, oo, and uu). The difference in length of vowels results in change of meaning.

Vowels-Dubbifttu

Afaan Oromo vowels are represented by the five letters, a, e, o, u and i.

All vowels are pronounced basically the same way throughout Oromia. These vowels when stressed may be opened: deemu (go), nyaadhu (eat) or closed: bada, rafi.

The Afaan Oromo vowels always are pronounced in sharp and clear fashion which means each and every word is pronounced strongly, for example:

- ✓ A: Ar’bba, Fardda, Haadha
- ✓ E: Gannale, Waabee, Noole, Roobale, colle
- ✓ I: Arsii, laali, Rafi, Lakki, Sirbbi
- ✓ O: Oromo, Cilaalo, Haro, caancco, Danbidoollo
- ✓ U: Ulfaadhu, Gudadhu, dubadhuu, arbba guugu, Ituu

Consonants-Sagaleewwan

Most Afaan Oromo constants do not differ greatly from Italian, but there are some exceptions and few special combinations.

- ✓ The consonant “g” has a hard sound. Gaari, gadi bayi, gargaari.

✓ The combinations NY and DH have a hard sound. E.g. Nyaadhu, Dhugi.

3.3. Afaan Oromo Punctuation Marks

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

Exclamation mark (Rajeffannoo) (!): It is used at the end of command and exclamatory sentences.

Comma (Qooduu) (,): It is used to separate listing in a sentence or to separate the elements in a series.

Full stop (Tuqaa) (.): it is used at the end of a sentence and in abbreviations.

Question mark (Mallattoo Gaafii) (?): It is used in interrogative or at the end of a direct question.

Colon (Tuqlamee) (:): It is used to separate and introduce lists, clauses, and quotations, along with several conventional uses, and etc.

3.4. Afaan Oromo Part of Speeches

Afaan Oromo language words can be categorized into nouns, verb, adverb, adjective, pronoun and prepositions.

Nouns: Noun is a word that helps to identify the categories of things, people, places and ideas. Nouns in Afaan Oromo are inflected for gender, definiteness and number. Most Afaan Oromo nouns are not morphologically marked for gender, except some sets of animate nouns which have the same root. These are respectively marked as feminine and masculine with [-ti: /-tti:] and [-SA /-ssa] as in the examples below. (Amanuel 2007)

Verbs: Verbs are words or compound of words that expresses action, a state of being and/or relationship between two things. In their normal position, they are found at the end of the sentence as shown below.

Example,

- ✓ Caalaan mana bite. : Chala bought a house.
- ✓ Ayyaantuun dhufte. : Ayantu has come.

Adverbs: Afaan Oromo adverbs are words that are used to modify verbs. Adverbs usually precede the verbs they modify or describe. They have the function to express different adverbial relations such as relations of time, place, and manner or measure.

Adjectives: An adjective is a word that describes or modifies a noun or pronoun. It specifies to what extent a thing is as distinct from something else. The masculine form terminates in one of the following suffixes – aa, -eessa, or -(a)acha, and the feminine form terminates in one of the following suffixes –oo, -tuu, -eettii, or –aattii.

Example,

- ✓ Dursaan gabaabaa (m) dha. :Dursa is short
- ✓ Hawwiin furdoo (f) dha. : Hawi is fat.

Pronoun: In Afaan Oromo, like in other languages, a pronoun is a word that is used instead of a noun or noun phrase. They are characterized based on number and gender. The Pronoun in Afaan Oromo can be independent or hidden with the verb based on their existence in a sentence. Independent pronouns are pronouns exist in a sentence as a separate word in the sentence. In the following example, “Inni” is an independent personal pronoun.

Example: Yohaannis yeroo dhufe homaa hin nyaatu homaas hin dhugu ture; Isaanis Inni dhukuba qaba jedhu.

Preposition: In Afaan Oromo prepositions are much less numerous than postpositions. The most common are: akka - according to, like, as. eega - since, from, after. eegasu - in that case, therefore gara - in the direction, towards, side. haga – until. Hamma - upto, until such that, as much as. Example,

- ✓ Akka isaa jabaan namu hin jiru. : There is no one as strong as he is.
- ✓ Inni gara manaa deema. : He goes (towards) home.

Gender: Afaan Oromo nouns gender variations are identified depending on the suffixes attached to the nouns. Some of them are “aa” is attached for masculine and “tuu” for feminine. For instance “Barataa” (indicate male student),”Baratuu” (female student), “Barsiisaa” (male teacher) and “Barsiistuu” (female student). When the suffixes “ssa” and “tti” are attached to nouns they also indicate masculine and feminine respectively. For example “Obbolleessa” (brother) and “Obbolleettii” (sister). In this language names of astronomical bodies and geographical places like cities and countries are feminine. For example “Aduun Baate” (to mean the sun rises) and “Magaalli Finfinnee barakam hundooftee?” (To mean when Addis Ababa City was established?), the suffix “tee” indicates feminine gender in two sentences. In Afaan Oromo the term “isa” and “ishee” shows masculine and feminine respectively like that of English language third person singular pronouns he and she.

3.5. Word Categories in Afaan Oromo

Words are the basic unit of a given language. The combination of these words on the bases of the language gives us phrases, clauses and sentences. The meanings of these sentences depend on each word of the sentence and the way they are arranged. Afaan Oromo words can be placed in to different categories. These categories are Noun, Verb, Adjective, Adverb, Adposition, Pronoun, Conjunction and Interjection and numerals.

Nouns:

Afaan Oromo nouns are words used to name or identify any of the categories of things, people, animal, places or ideas. However, sometimes lexical classes like noun can be defined functionally (morphologically and syntactically) so that some words for people, places, and things may not be nouns. In Afaan Oromo, nouns mainly occur at the beginning of a sentence. In the following examples, Afaan Oromo nouns are italicized and underlined.

- ✓ Dimaan fillannoo kooti : Red is my preference.
- ✓ Sareen manatti olixxe : A dog entered into a house.

Words that are categorized as nouns in a sentence can be a subject or object. Subject mostly comes at the beginning whereas an object mostly comes after subject and before verb in a sentence. Consider the following example.

- ✓ Warabessi harree nyaate. : A hyena eats a donkey
- ✓ Tolaan mana ijare. : Tola built a house.

In the above two sentences the underlined and italicized words “Warabessi” (A hyena) and “Tolaan” (name of person) are the subject of the sentences, whereas the two italicized words “harree” (donkey) and “mana” (house) are the objects of the two sentences respectively.

Afaan Oromo nouns are also inflected for number i.e. singular and plural. Singular nouns are nouns which are built without adding any affix. On the other hand, plural nouns are mainly formed by adding suffix or using numerals with single noun. The following are suffixes used to form plural nouns in Afaan Oromo. “-Oota”, “-ota”, “-wwan”, “-een”, “-lee”, “-yyi” etc.

Singular	Plural	Plural Marker
Mana :House	Manneen :Houses	-een
Sangaa :ox	Sangota :Oxen	-ota
Waraabeessa :Hyena	Waraabeyyii :Hyenas	-yyi
Hojii :Work	Hojiilee :Works	-lee
Barsiisaa :Teacher	Barsiisota :Teachers	-ota
Sa’a :Cattle	Saawwan :Cattle	-wwan

Table 3 plural noun formation using suffix

Additionally, plural nouns can be formed by using numerals. This can be done by writing singular noun followed by numeral. For example, the following are valid plural nouns in Afaan Oromo.

Singular	Plural
Mana :House	Mana sadi :Three house
Nama :Man/Woman	Nama Lama :Two man
Hoola :Sheep	Hoola kudhan :Ten sheep

Table 3.1 plural noun formation using numerals

Afaan Oromo analyses masculine and feminine genders and most of the nouns belong to either of the two. However, there are some nouns used for both masculine and feminine. For example, the noun “nama” (man or woman) can be used for masculine and feminine.

Afaan Oromo nouns are also inflected for definiteness, but not for indefiniteness. The definiteness (the English ‘the’) in Afaan Oromo can be different for masculine and feminine. For masculine “-icha” is mainly used while “-ittii” is used for feminine.

Noun	Definiteness
Nama :man	Namicha :The man
Harree :Donkey	Harricha :The donkey
Qaalluu :Priest	Qaallitti :The priest

Table 3.2 definiteness form of nouns

Afaan Oromo nouns are also inflected for case. Case is a grammatical category of nouns that indicates the nature of their relationship to the verb in sentences. Afaan Oromo nouns can be formed by adding derivational suffix to different categories of words such as noun, verb, adjectives etc.

Noun	Derived noun
Bilisa :Free	Bilisummaa :Freedom
Fira :Relative	Firummaa :Relationship
Nagaa :Peace	Nageenya :Peaceful
Jabaa :Strong	Jabeenya :Strength
Nama :Man	Namooma :Humanity

Table 3.3 derivational nouns formation

In the above Table 3.3, abstract nouns are formed from other nouns by adding “ummaa” or “eenya” or “ooma” suffix.

Pronouns:

Afaan Oromo pronouns can be used for replacing nouns and noun phrases. Like nouns, Afaan Oromo pronouns decline for number and gender. Consider the following pronouns

- ✓ Ishee: she, represents feminine noun and singular
- ✓ Isa : he, represents masculine noun and singular
- ✓ Isaan: they, represents plural nouns and either masculine or feminine.

There are different categories of pronoun in Afaan Oromo based on their functionality and meaning in a sentences. These are personal, possessive, demonstrative, reflective and reciprocal pronouns. Personal pronouns can be used to replace the subject and the object of a sentence. The following are Afaan Oromo personal pronouns used to represent subject of the sentence.

	First person	Second person	Third person
Singular	Ani: I	Ati :You	Inni :He
Plural	Nuyi :Nuti :We	Isiin :You	Isaan :They

Table 3.4 personal pronouns that can represent subject

Consider the following two sentences.

- ✓ Firomsaan leenca ajjeese. :Firomsa kills a lion
- ✓ Inni leenca ajjesse. :He kills a lion

The above two sentences have the same meaning. Even if in the second sentence the pronoun “Inni” (he) replaces the subject “Firomsaan” name of a person. A personal pronoun can also replace an object of a sentence in Afaan Oromo.

	First person	Second person	Third person
Singular	Ana :Me	Si :You	Isa :Him
Plural	Nu :Us	Isiin :You	Isaan :They

Table 3.5 personal pronouns that can replace object

Consider the following example:

- ✓ Leensaan barattota barsiiste. : Lensa teaches students.
- ✓ Leensaan isaan barsiiste. : Lensa teaches them.

In this example, “isaan” (them) replace the object “barattota” (students). Another category of Afaan Oromo pronoun is possessive pronouns which are used to indicate the ownership of something.

	First person	Second person	Third person
Singular	Koo :Kiyya :Mine	Kee :Yours	Isaa:His, Ishee:Her
Plural	Keenya :Ours	Keessan :Yours	Isaan :Their

Table 3.6 Afaan Oromo Possessive pronouns

Consider the following example:

Demonstrative pronouns are pronouns that are used to refer to a thing that was known previously or mentioned earlier. It can also be used to refer to the objects which are in the speaker's mind. Both proximal and distal demonstrative pronouns exist in Afaan Oromo. Proximal pronouns have masculine and feminine whereas distal do not have. However, plural and singular demonstratives are not distinguished. The proximal Afaan Oromo Pronoun are:

- ✓ Kana / kuni (this / these) (masculine)
- ✓ Tana / tuni (this / these) (feminine).

And the distal Afaan Oromo pronouns are:

- ✓ san (that)
- ✓ sun (those)

Verbs:

Verbs are words or compound of words that express action, state of being in or relationship between two things. In Afaan Oromo verbs mostly appear at the end of sentence.

Consider the following example:

- ✓ Guutaan *kaleessa deeme*. : Guta went yesterday.
- ✓ Caalaan kubbaa *dhiite*. : Chala kicked a ball.

In this example, the words written in italic and underlined ‘deeme’ (went) and ‘dhiite’ (kicked) are verbs of the sentence.

Afaan Oromo Verbs are inflected for number, gender and tense [35]. Additionally Afaan Oromo verbs can be categorized into main (transitive or intransitive) and auxiliary verbs. Intransitive verbs are main verbs which do not take object or complement in a sentence. The following examples illustrate intransitive verb in Afaan Oromo.

- ✓ Namichi fige. : The man runs.
- ✓ Isaan Kaleessa dhufan. : They came yesterday.

In the above example, the words written in italic and underlined are transitive verb. They do not transfer any message from subject to complement. Transitive verbs are main verbs which transfer message to complement (objects). Consider the following two sentences:

- ✓ Tolaan ulee cabse. : Tola broke a stick.
- ✓ Caalaan muka mure. : Chala cut a tree.

In the above two examples the verb “cabse” broke and “mure” cut are transitive verbs. They interrelate subject and object in the sentences. Auxiliary verbs support the main verbs used in a sentence. The following are Afaan Oromo auxiliary verbs “dha”, “ta’e”, “qabda”, “ture”, “jira”, etc. In the following two examples the auxiliary verbs are written in bold and italic.

- ✓ Isheen barattu cimtu dha. : She is a clever student.
- ✓ Hojii kana hojjechuu qabda. : You have to work this job.

In the above two sentences the word “dha” and “qabda” are auxiliary verbs.

Adjectives:

Adjectives in a sentence are used to modify nouns to show the quality of things. I.e. it specifies to what extent a thing is distinct from something else.

Consider the following examples:

- ✓ Leenseen bareeddu dha. : Lense is beautiful.
- ✓ Tolaan dheeraa dha. : Tola is tall.

In the above examples the word “bareeddu” beautiful and “dheeraa” tall are adjectives. Afaan Oromo adjectives can be formed from compound words. For instance, “humna dhabeesssa”

weak, “simbo qabeessa” handsome are some of adjectives constructed from compound words. Adjectives inflect for number and gender in Afaan Oromo language.

Singular	Plural	Masculine	Feminine
Guddaa	Gurguddaa	Guddaa	Guddoo
Jabaa	Jaboota	Jabaa	Jabduu
Ko'eessa	Ko'eeyyii	Ko'eessa	Ko'eettii
Cimaa	Ciccimoota	Cimaa	Cimtuu

Table 3.7 adjective inflection for number and gender

Adverbs:

Adverbs are words which are used to modify verbs. In Afaan Oromo adverbs come before the verb they modify. Afaan Oromo adverbs are categorized as adverbs of time, place and manner (condition). Adverbs of time show the time the action takes place. The following are the words that can be used as adverbs of time in Afaan Oromo language.

- ✓ Amma :now , Boru :Tomorrow ,Kaleessa :Yesterday , Yoom :When
- ✓ Har'a : Today, Galgala : Tonight etc.

Consider the following example.

- ✓ Boonsaan kalessa dhufe. : Bonsa came yesterday.
- ✓ Qananiisaan boru ni figa. : Kenenisa will run tomorrow.

In these examples the word “kaleessa” yesterday and “boru” tomorrow are adverbs of time. Mostly adverbs of time answer the question of when the action takes place. Adverbs of place show the place where the action takes place. The following are the words that can be used as adverb of place in Afaan Oromo. “as” here, “achi” there, “gadi” below, “gubbaa” above, ‘jidduu’ middle, ”irra” on etc.

Consider the following example,

- ✓ Tolaan mana jira. : Tola is at home.
- ✓ Inni konkolaataa irra jira. : He is on the car

Adverb of manner show how the action of the sentence is done. The following are Afaan Oromo words that can be used as adverb of manner “ariitin” quickly, “suuta” slowly, “akka gaarii” well etc.

Consider the following example,

- ✓ Inni ariitin figa. : He is running quickly.
- ✓ Caalaan baay’ee cimaa dha. : Chala is very clever

In the above sentences the word “ariitin” quickly and “baay’ee” very are adverbs of manner.

3.6. Abbreviations in Afaan Oromo

Abbreviations are mostly formed by taking initial letters of multiword sequences to make up a new word. Sometimes, they can be formed from initial and non-initial letters. Afaan Oromo, abbreviations are used to represent dates A.L.I Akka Lakkoofsa Itiyooophiyaa to mean in Ethiopian calendar, A.L.A Akka Lakkoofsa Awurooppaa to mean in Gregorian calendar, months and dates by short words. Moreover, personal titles can be abbreviated like that of English language. For examples: “Aadde” is abbreviated as “Aadd.” Mrs., Obbo is abbreviated as “Obb.” Mr. Organization’s names are also abbreviated.

3.7. Afaan Oromo Questions

Different languages have different ways in the use of the word order and question particles. However, question statements are constructed with the help of interrogative words and question marks to indicate the statement is a question, in every language. In the English language, interrogative articles such as who, what, where, when, why, how are used to construct a questions. In Afaan Oromo interrogative particles help to construct a question sentence. Interrogative particles are also known as interrogative pronouns. Some of the Afaan Oromo interrogative particles are: “eessaatti” where “maaliif” why, “yoom” when, “maali”

what “akkamitti” how and so on. These interrogative particles are used to construct the factoid and non-factoid questions.

3.8. Afaan Oromo Morphology

Morphology is a branch of linguistics that deals about the knowledge of the meaningful component of words [32]. Jurafsky and Martin [32] defined morphology as the study of the way words are built up from smaller meaningful units called morphemes. Word is the most basic unit of linguistic structure. Like other Ethiopian languages, Afaan Oromo has complex and rich morphology.

3.8.1. Types of morphology in Afaan Oromo

In Afaan Oromo language, we have two broad types of morphology namely derivational morphology and inflectional morphology [33].

Inflectional Morphology: Inflectional morphology is the processes when words are adapted to their proper functions within a given sentence without changing the meaning of base words. Alternatively, can be defined as the change the form of a word for grammatical usage. It occurs in the form of different word classes. Inflection of verbs, Inflection of Nouns and Inflection of Adjectives [33].

Inflection of Nouns: most of Afaan Oromo nouns end with a vowel except few which ends with consonants like n, l, t. [Inflection Morphology Oromo]. Inflectional categories under nouns exists mainly in the forms of marking number, definite and gender.

In Number, marking inflections distinguish plural and singular. Several types of suffixes are attached to nouns to make plural forms.

Example:

The plural – (o) ota attached on Nouns

Waggaa [Base form] Wagg-oota [Inflected form] Years

The plural –lee attached on Nouns

Baatii [Base form] Baatiilee [Inflected form] Months

In definite marker or singulative marker shows noun is marked for being used as single form.

Example: Nama [Base form] Namicha [Inflected form] The man

In gender marker, we use inflection to identify masculine and feminine through gender suffixes.

Example, boonaa [base form] boont-aa [inflected form -m] / proud boy boont-uu [inflected form -f]/proud girl

Inflection of Verbs – verbs are the most classes in which inflection are occurred. Mainly verb inflection happens in the form of inherent and agreement properties. Inherent properties is a verb inflection that triggers inflection on that word class includes aspect, mood, and voice. However, agreement properties indicate inflection of word class for properties out of its members include person, number, gender and case. In the Afaan Oromo language, the roots or stems of verbs, usually end in consonant, take inflectional morphemes showing distinction between aspects or mood or gender or number.

Example, Mur – [root form] Mur-te [Inflected form]

Inflection of Adjectives – are the same with that of nouns. Adjectives are inflected for number, gender and singulatives like nouns. If adjectives occur within sentences, number is marked on both of them. Inflection for number of adjectives occurred in the form lexical, reduplication and-(o) ota.

Example: Inflection for Numbers Lexical: Sooressa (s) Sooreeyyi (p)

Reduplication: Guddaa (s) gud-guddoo (p) - (o) ota: hamaa(s) ham-oota (p)

In Afaan Oromo, the base forms of adjectives are normal to be used as masculine, but inflection occurs when we make them for feminine.

Example: Hamaa (m) Hamtuu (f)

In Afaan Oromo, singulative markers are not used on both noun and adjective at the same time. Which means, when nouns is marked, adjective is not and vice versa.

Example: Muk-ni (n) dheer-icha (adj.). The long stick.

Derivational Morphology: Derivation morphology is the creation of new words from already existing words in the language. And is the alternate word form which changes the meaning and word class. Different derivational suffixes are attached to the root or stem of the word [33]. It occurs in the form of different word classes. Derivation of verbs, Derivation of Nouns and Derivation of Adjectives.

Derivation of verbs – is the creation of new verb word class from other given word class stem.

Example: arguu [base form] to see argachuu [derivated form] to get, find

Derivation of Nouns – is the creation of new noun word class from other given word class stem.

Example: bulchuu [base form] to administer bulchiinsa [derivated form] administration

Derivation of Adjectives – is the creation of new adjectives from nouns or from other adjectives or from verbs. Example: Jabaa [base form] strong jabaacuu [derivated form] to be strong

3.9. Summary

Afaan Oromo language is an Afro-asiatic language in family of Cushitic language spoken by people live in Ethiopia and neighboring country. Afaan Oromo word classifications are important in the writing system of the language, these includes noun, verb, adjectives etc. The morphology of Afaan Oromo language is same as other language in broad categorization, like derivational and inflectional where inflectional is adapting new forms of word but the same in meaning and derivational is the creation of new words from already existing words. Unlike other language like Philippines, Afaan Oromo does not have infix morpheme.

Chapter Four

System Design and Implementation

4.1. Overview

This Chapter gives the architectural design and implementation of the List, Description and definition Afaan Oromo Question Answering system. The document preprocessing section presents the architecture of AO analyzer and explains the techniques. The third section covers detailed strategies and algorithms implemented in analyzing and generating questions. The next section covers the specific methods used in retrieving and filtering documents retrieved from the corpora. The fifth section details about the techniques and algorithms used for selecting the best answers. Finally the summary section summarizes the chapter.

4.2. System Architecture

QA system consists of the document preprocessing component for document normalization and indexing, the question analysis component to identify question type, generate query, and expand the query, the document analysis component for retrieving relevant documents using the queries from the question analysis component and filtering the retrieved documents according to the question type, and the answer extraction component to produce answers from the filtered documents. The whole design of the proposed QA system and the interaction among the components is shown in Figure 4.2.

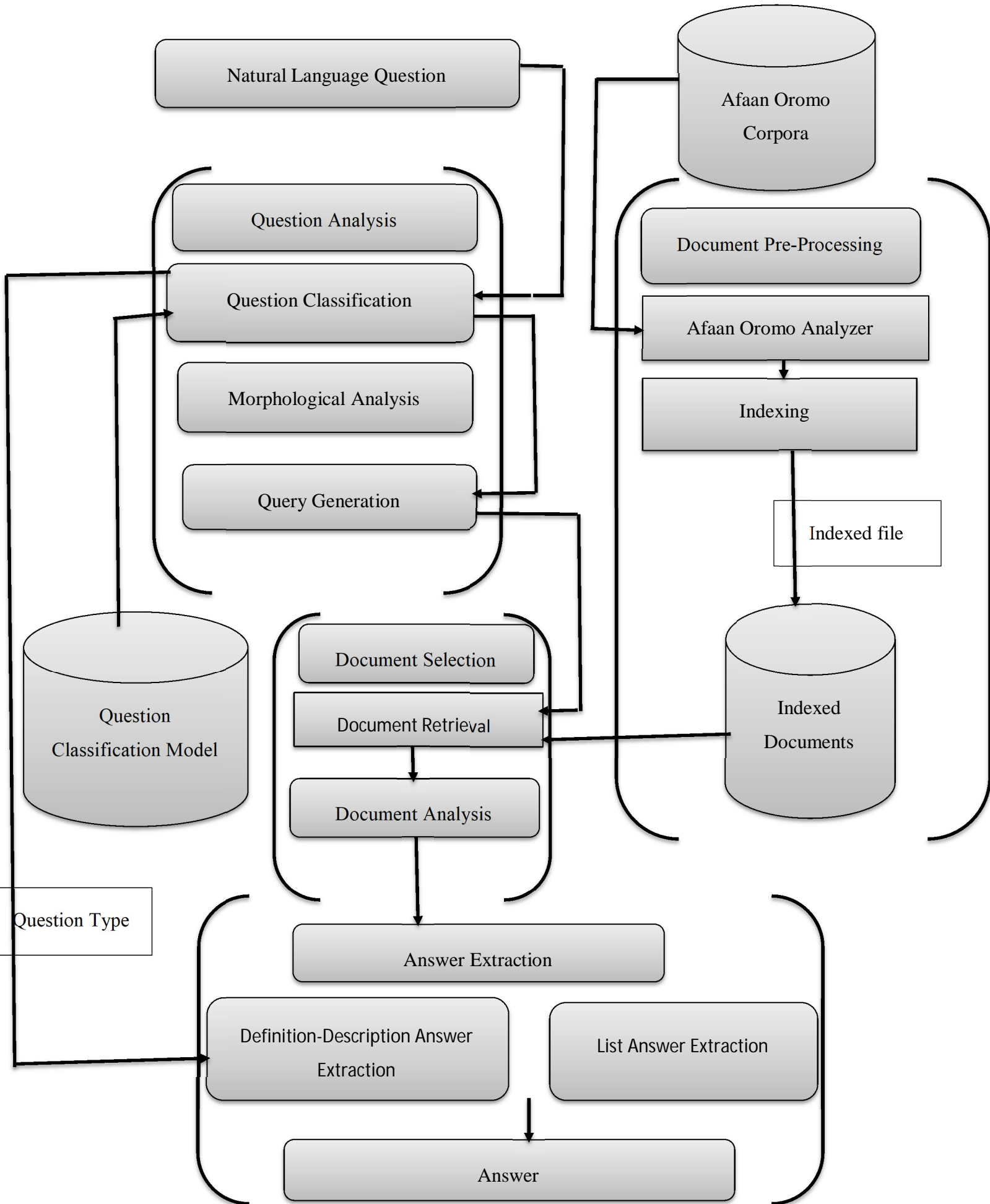


Figure 4.2: The Architectural Design of Afaan Oromo List, Description and Definitional QA System [4]

Type-by-type approach proposed for answer extraction. The design and implementation process of the proposed question answering system consists of document pre-processing, question analysis, document selection, and answer extraction. The document pre-processing module preprocesses documents. Question analysis determines question types, pre-process queries and constructs proper query for the document selection component.

4.3. Document Preprocessing

The documents work are that Afaan Oromo corpora collected from different web sites and books. In the process of question answering activity, before retrieving documents which contain an answer of a question from Afaan Oromo corpora, different text pre-processing tasks are involved. The main pre-processing techniques we have used are text tokenization, short word expansion, case normalization, stop word removal, stemming, morphological analysis and indexing that should be done in order to accomplish the question answering task.

Figure 4.3 shows Afaan Oromo Analyzer used for preprocessing questions and documents.

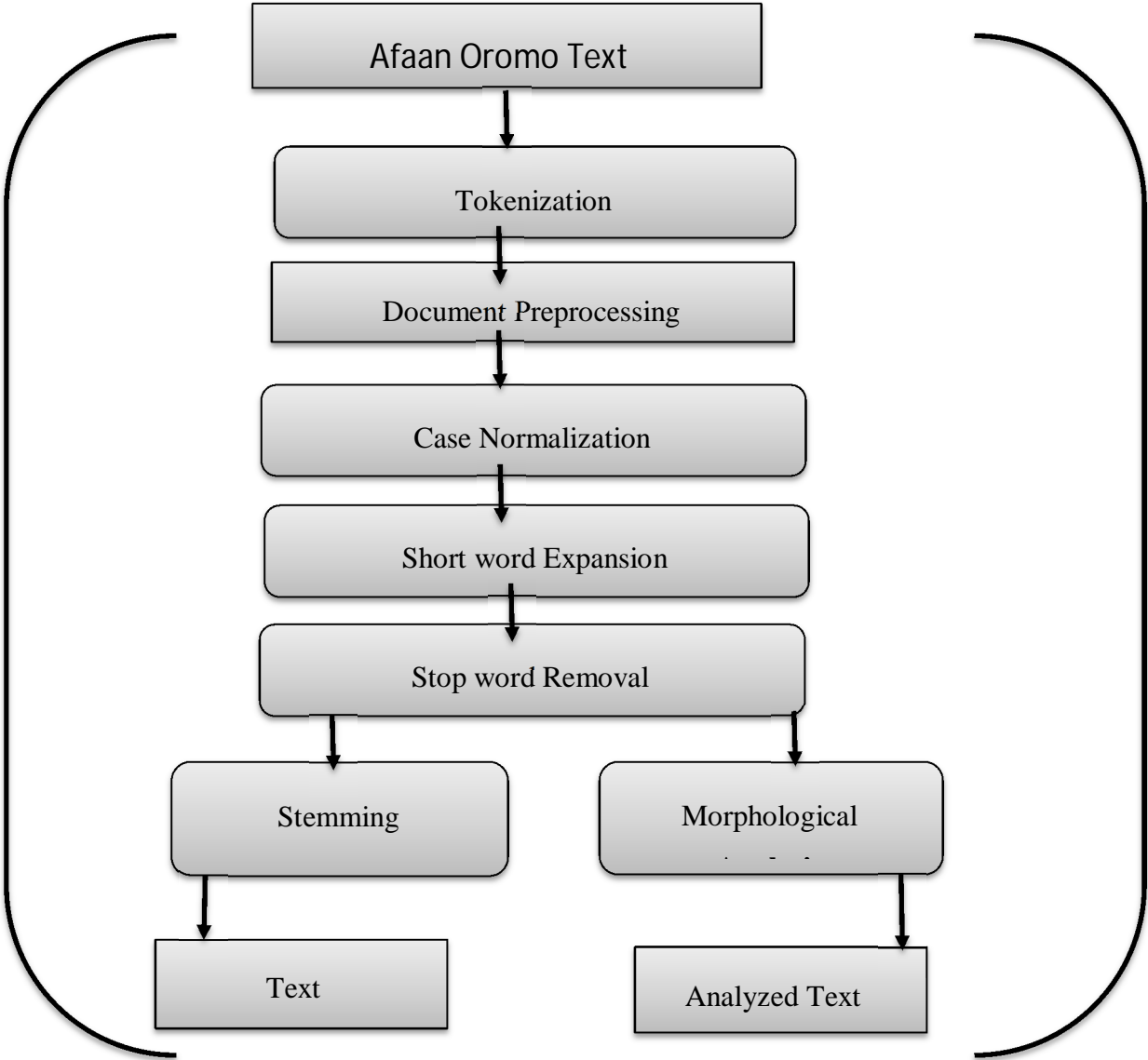


Figure 4.3: Architecture of Afaan Oromo Analyzer Component [11]

Tokenization: It is the process of dividing the sentences to words depending on the sentences separator which is vary from sentence to another and the boundary of the words of the sentence. It is defined as Tokenization is the task of cutting a string into identifiable linguistic units that constitute a piece of language data. Many separators can be used. Tokenization breaks the stream of characters into raw terms or tokens, detects word boundaries of a written text and at the same time it can be taken as the process of removing non alphanumerical characters.

The document retrieval subcomponent of question answering system fetches answers of natural language questions from Afaan Oromo indexed files. Tokenization is one of the text preprocessing tasks that should be done before the file indexing task. Tokenization helps in matching the tokens of the query with tokens in the document. In Afaan Oromo white spaces are used to separate the boundary of a token and punctuation marks such as commas, periods, question marks, exclamation marks and hyphens are important to demarcate the boundaries of tokens but, in Afaan Oromo language apostrophe mark are considered as a part of a word. For example, in the word ‘Sa’a’ (cow), the apostrophe is used to show that the vowels are produced independently. Thus, the word ‘Sa’a’ has to be treated as a single token in the tokenization process. White spaces and punctuation marks except “”, “/”, “.” “,” and “-” are used as a word delimiters.

Case Normalization: It is the process of transforming text into some other forms. It is the process of handling problems related with variation of cases of upper case, or lower case or mixed cases. So the good way to handle this problem is converting the whole document into similar case. In some languages like Amharic which does not have a distinction between upper and lower case, this might not be a big deal. However, it is very important for languages that use Latin characters for writing. In this research we will use lower case letters for questions and corpus.

Short Word Expansion: Short words are short form of words or phrases which can be formed from initial letters of important terms of a word or a phrase or from the combination of letters of a word or a phrase and other characters. Usually in Afaan Oromo, ‘and, /’ are used while writing words in short form.

Stop Word Removal: Stop-words are most frequent terms which are common to every document, and have no discriminating power.

Stemming: Is the process of removing some additional characters from the word to categorize a group of words into one main category to help in semantic analysis. It is defined by [8] as "is the task of correlating several words onto one base form". Stemming is an activity to find the stem of a word by removing affixes, i.e., it enables to merge morphological variants of a word under a single index entry or its common form. Thus, for this research work, we have used Debelas stemmer, which takes a word as an input and removes its affixes using a rule based algorithm [15].

Morphological Analysis: Is the process of dealing with relationship of the words with the same inflection (structure of the word depending on its position and its classification in the sentence). Morphological analysis is the segmentation of words into their component morphemes and the assignment of grammatical morphemes to grammatical categories and lexical morphemes to lexemes [16]. Thus, the morphological analyzer returns the root of a word and it enables to merge morphological variants of a word under a single index entry or its common form. Horn Morpho [16] is used for the morphological analysis task. Horn Morpho is a Python program that analyzes Amharic, Oromo, and Tigrinya words into their constituent morphemes (meaningful parts) and generates words, given a root or stem and a representation of the words grammatical structure [16].

Indexing: Is a process that converts documents in a repository into cross reference lookup. The index stores statistics about terms in order to make term based search more efficient. It is the last step on document preprocessing, i.e., before being indexed, it is necessary to perform the techniques discussed above.

4.4. Question Analysis

The main function of the question analysis component is to understand the kind of information the question is asking for. In addition, it is responsible to formulate proper queries for document retrieval. When the user poses a question to the system, the question analysis component takes in the user query and passes it to its sub components.

Question analysis is conducted with the two following objectives: the first is to extract some characteristics which will likely be used in the answer extraction module; the second is to extract the terms which will re-index the selected documents in order to retain only a subset of them and to supply further evidences during the final matching.

4.4.1. Question Classification

Question classification represents the main part of Question Answering Systems (QAS) regardless of various types of architectures. Question classification as different types of question (definition, factoid, list, yes/no etc.) require different processing for answer extraction in a large collection of documents and texts, at first the system should know what it looks for. In this case, questions should be classified regarding their types. The question classifier subcomponent determines the type of a question as list, definition or description.

Types of questions: non-Factoid questions:

✓ Definition, Description, List

Class	Interrogative term	Class indicative term
Definition	Maal Jechuu dha, Maali dha, Maali Isheen, Maali Isaan	Hiikni, Yoo Hikamu, Hiika, Yeroo Hikamuu
Description	Maali dha, Maali, Maali qaba/qabdi/qabu, Maali akka ta'e ibsi	Faayida, Gahee, Gayee, Faayidan.../Isaa/Isaan/Ishee/
List	Maali fa'i , Enyuu fa'a, kami fa'a, Tarreessi, Caaqasi	Sababa/ Sababoota/ Ulagalee/Goosa/Goosota

Table 4.1: Question classes, interrogative terms, and class indicative terms

The task of a question classifier is to assign one or more class labels, depending on classification strategy, to a given question written in natural language. Question classification is a subcomponent of question analysis which is concerned with assigning questions to semantic classes. This semantic classification can be used to reduce the search space of possible answers. In order to assign a question to semantic classes, a rule based question classification and machine learning approaches are used most of the time but, for this thesis we only used a rule based approach to determine the type of question. This is due to the study in [19] and [23] which show that the rule based question classification approach got good performance than the machine learning approach.

```
Input the question
If the question contains (one of the definition indicative
Terms) then
Return question type 'Definition'
Else if the question contains (one of the definition
Question particles and (one of the definition indicative
Terms) then
Return question type 'Definition'
Else if the question contains (one of the description
Indicative term and one of the descriptive question
Particle) then
Return question type 'Description'
Else if the question contains (one of the list indicative
Term and one of the list question particle) then
Return question type 'List'
Else Return question type 'Unknown' End If
```

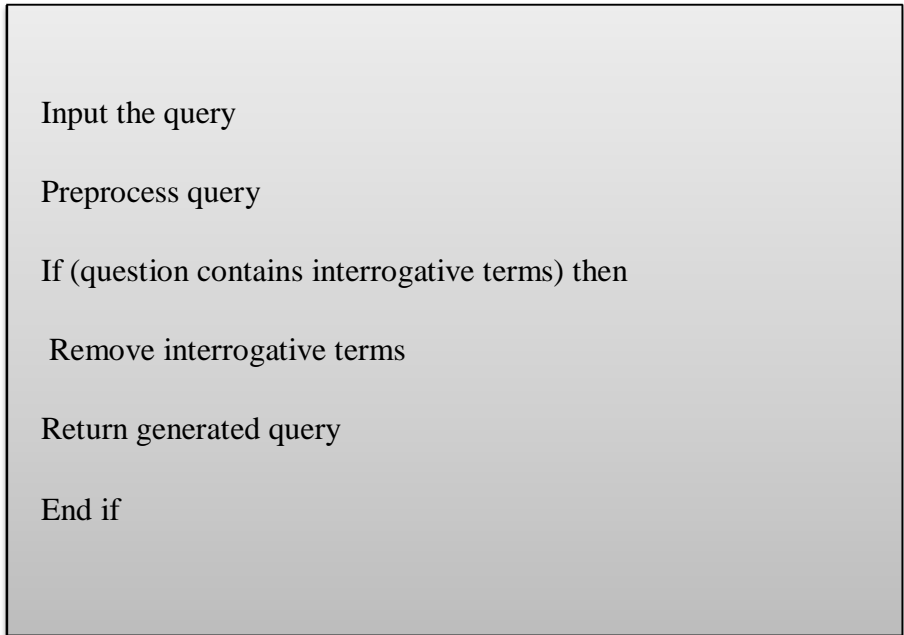
Figure 4.2: Rule Based Question Classification Algorithm

Thus, we have used a rule based approach to identify type of questions using Algorithm 4.2. The algorithm determines the question type by using the interrogative terms of the question and class indicative terms shown in Table 4.1 For example, given the question ‘Faayidaan man maali?’ (What is the use of house?) The terms ‘faayidaa’ use and ‘maali’ what indicate that the question is looking for a description. Another example for definition question, ‘Hiikni Aadaa maali?’ ‘What is the meaning of Culture?’ The terms ‘Hiikni’ meaning and ‘maali’ what indicate that the question is looking for definition question. Question asking for a thing ‘Diirqama mootumma caqasii?’ ‘List Governments duty?’ The terms ‘caqasii’ list and ‘diirqama’ duty indicates that the question is looking for list question. Question asking for a place ‘lagawaan umamaa Itiyooophiyaa tarreessi’ ‘name Ethiopian natural Rivers’ the terms ‘tarreessi’ name and ‘Lagawaan’ lakes indicates that the question is looking for list question. Algorithm 4.1 shows a rule based question classification method for classifying queries to their classes.

4.4.2. Query Generation

Question analysis also extracts the information in the question that allows the generation of queries that could be used by an IR system for the extraction of the answer-bearing text from the entire document collection. The query generator acquires queries by removing the interrogative terms from the question.

The query generation is performed in two steps. The first step is to generate an action sequence that imitates the overall user behavior of adding terms, dropping terms, and submitting queries. The second step is term selection, where the actual terms of the queries are chosen. Query generation is used to convert the user’s natural language questions into suitable form for document retrieval. First users query will be preprocessed using AO Analyzer which contains tokenization, case normalization, stop word removal, short word expansion, stemmer and morphological analyzer tasks. Then, the query generator removes interrogative terms from the preprocessed query and generates a query which is used by the document retrieval. The question particles like ‘maali’, ‘jechuun maal jechuu dha’, ‘faayidaan’, ulaagaa, tarreessi, caqasii etc., are removed from the question because it doesn’t worth for searching.



Algorithm 4.2: Query Generation

4.5. Document Selection

The document selection component consists of the document retrieval sub component which is responsible for retrieving documents which may contain information pertinent to the list, definition or description of a target and document analysis subcomponent responsible for filtering documents by identifying the relevant document from the irrelevant one.

4.5.1. Document Retrieval

In fact the document retrieval component can be quite complex involving numerous strategies for retrieving both documents and associated data as well as determining the amount and structure of the text to be passed to the final component in the system pipeline (Andrew, 2005). Document retrieval has a responsibility to fetch documents which are related to the generated query; it takes keywords produced by the query generator component. It starts with user's query and terminates with a list of documents ready to be processed by document analysis and later for answer extraction, also used as an intermediary between question analysis and answer extraction components. For this study, a Lucene package [14] was used for searching. It returns a ranked list of candidate documents by considering the number of keywords of the query in the documents from the Lucene index.

4.5.2. Document Analysis

The document analysis takes the most likely answer list with the question classification description that shows what answer should be. For document analysis, the system implements a rule-based method on relevant documents, and adopts calculation to extract the correct answer. Retrieved documents should be filtered before further analysis in order to identify relevant documents from the irrelevant. Thus, document analysis first locates the question keyword using keyword extractor [1] and based on the keyword it filters the documents. Keyword Extractor (KE) extracts the keyword target term(s) of the question [1]. Keyword target term is obtained by removing indicator terms from the query; the query term is used later in sentence extraction subcomponent. After the target is extracted, the documents will be tested by their respective rules (regular expressions). Then, if a text is extracted by one of the rules, the document will be kept; otherwise it will be removed.

4.6. Answer Extraction

Answer extraction is the act of extracting text strings constituting the exact answer or answers to a question from the candidate answer bearing documents retrieved by the document analysis module. To identify relevant information more accurately, the answer extraction module performs detailed analysis on the retrieved documents and pinpoints the answer to the question as specified by the question analysis component, i.e., according to the question type, the filtered documents will be processed by the list answer extraction or definition, description answer extraction subcomponents.

4.6.1. Definition Description Answer Extraction

Questions require a single fact or a set of facts for a list question to be returned to the user. So the two factoid Afaan Oromo QA works in [9, 10] extract answers using the named entity gazetteer and pattern based answer pinpointing algorithms. These algorithms only identify person or place name, dates, and numbers. But definition and description questions require a substantially more complex response a short paragraph which succinctly defines the target or state concepts the user wishes to know more about. That means the methods in [9, 10] cannot

be used for list, definition, and description questions. Thus finding snippets or piece of information about the current target, ranking, selecting, and ordering them is very important. To do so, the definition, description answer extraction subcomponent comprises snippet sentence extraction, sentence score computation, answer selection, and sentence ordering units.

Sentence Snippet Extraction

The main purpose of this component is to create candidate answer set. First every filtered document is tokenized to sentences by the sentence tokenizer. Then according to the question type the snippet/sentence extractor extracts sentences from the tokenized sentences using manually crafted indicative patterns that are listed in Table 4.5. About rules, regular expressions for definition and rules for description questions are crafted by inspecting different Afaan Oromo definition and description bearing documents. We observed that many documents, while defining a term they usually state the core concepts about the term at the beginning. Due to this reason, for definition questions, if the first sentence of a document is extracted by one of the rules, the next two sentences are incorporated to the candidate answer set.

Question Type	Sentence Extraction Patterns
Definition	<p>Rule 1: target + “jechuun “+ “. *”</p> <p>Rule 2: target + “(jechuun)? “+”. *”+” jechuu dha[;]”</p> <p>Rule 3: target +”. * Hiikan [Hiikni Hiiki [Isaa] (. * jechuu dha.)</p> <p>Rule 4: “. *” + target +”. *”+” Yookaan”+”. *”</p> <p>Rule 5: “. *” + “Yookaan “+ target +”. *”</p>

	<p>Rule 6: “.*”+ target +” .*”+”(Hiikamuf jechuu dha)”</p> <p>Rule 7: target + “.*”+” (jedhuu Yoo ta’u Yoo jedhamu) “+”.*”</p> <p>Rule 8: “.*” + target +” jedhama”</p> <p>Rule 9: target +”.*”+” (jedhamun jedhame jedhameme)?”+”(beekama hiikama Waamama)”</p> <p>Rule 10: target +” .*”+”(tati ta’u dha)”</p>
<p>Description</p>	<p>Rule 1: target +”.*”+” (faayidaan [isaa isaani] gaheen [Isaa isaan] tajajiila [isaa isaan] jechuu dha?”</p> <p>Rule 2: target +”.*”+” (faayidaa [isaa isaan olaan gahee qaban] jechuu dha?”</p>
<p>List</p>	<p>Rule 1: target+”.*”+ (isaanis kan dalagaan isaas kanneen keessaa muraasni)</p> <p>Rule 2: target+”.*”+ (bakka akka kannatti) + ”.*”+ qoodama qoodamu.</p> <p>Rule3:dalagaa(n wwan) + target</p> <p>Rule4:maddi dirqama mirga karaalee (kaayyoo dammee hariiroo (n wwan))</p>

	+ target Rule 5: hariiroo kaayyoo (wwan) + target + beekamoo ta'an Rule 6: ".*" + dalagaa (n wwan) + target Rule 7: dammee(n wwan) + target + muraasni
--	---

Table 4.5: Sentence/Snippet Extraction Patterns

Sentence Score Computation

An answer to a definition, description or list question should contain all the vital snippets or sentences. Thus, in order to select the appropriate sentences from the candidate answer set we formulate the sentence scoring function given in Equation 4.2, i.e., the score of a sentence S is calculated as the sum of the percentage of the query (Q) terms in the sentence, weight of the pattern that identifies the sentence, the reciprocal of the position of the sentence in the document.

$$score(S) = \frac{N_{S \cap Q}}{N_Q} + weight(S, P) + \frac{1}{pos(S)} + luceneScore(D, S)$$

Where $NS \cap Q$ is the number of terms that are found in both S and Q , NQ is the number terms in Q , $weight S, P$ is the weight of the pattern P that matches with S , $luceneScore(D, S)$ is the score of document D that contains S by Lucene, and $pos(S)$ is the position of S in the document that contains S .

Since the position of a sentence does not have any impact for description questions, score of sentence S is computed by the formula given in Equation 2.

$$score(S) = \frac{N_{S \cap Q}}{N_Q} + weight(S, P) + luceneScore(D, S)$$

Answer Selection

At this stage this subcomponent faced a set of scored text fragments which are possibly used to generate the correct answer. But it is likely that some of the snippets or sentences might be redundant and all are not equally important. That is, we have to guarantee that, given sentence A does another sentence B provide any new and relevant information for the purpose of defining or describing a given target? Therefore, we need an approach that would determine the semantic similarity of the extracted snippets/sentences and their importance to the target relative to each other. As the work in [13] suggested one way of determining the similarity of texts is to use word overlap. The more different text fragments share common non-stop words it indicates that they are highly similar [13]. A sentence profile is constructed for each sentence which contains the set of non-stop words, T , in the sentence. Then the similarity, sim , between sentences A and B is calculated by the formula given in Equation 3 which is adopted from [13]. This is the percentage of tokens in A which appear in B or the percentage of tokens in B which appear in A.

$$sim(A, B) = \frac{|T_A \cap T_B|}{\min(|T_A|, |T_B|)}$$

where $sim A, B$ is the similarity of the sentences A and B, T_A and T_B are the number of non-stop tokens in sentences A and B respectively, and $T_A \cap T_B$ is the number of common tokens in A and B.

4.7.2 List Answer Extraction

Answer extraction is selection of an answer for a given query from collection of text documents. Answering List questions is more difficult compared to answering factoid

questions because it requires a system to acquire the answer instances from different sources answer fusion. This component is used to answer two types of list question. Before extracting candidate answers, the filtered documents are tokenized. Depending on the question focus List Answer Extraction (LAE) used two methods for extracting candidate answer from the tokenized sentences. The first one is for answering about things in this case, the pattern matching method were used for extracting the tokenized sentences.

4.8 Summary

This chapter described the architectural design of the Afaan Oromo List, Definition, Description, Question Answering system and the implementation of its main components. The Afaan Oromo List, Definition, Description, Question Answering system implementation consists of four main modules. The document pre-processing component is used to normalize, remove stop words, expand short words, stem, lemmatize and index documents. Once documents are normalized and indexed, they will be ready for the succeeding components for further processing. Question pre-processing is used to manipulate the questions to create a proper query and is done in query generation. The question analysis component determines question type by using a rule based technique and pre-process and creates a proper query that will be submitted to the document selection component. The document retrieval component retrieves documents using the query from the question analysis component and filters the retrieved documents using filtering patterns. The answer extraction component has sentence tokenize, Definition-Description Answer Extraction and List Answer Extraction. Definition-Description Answer Extraction contains sentence extraction subcomponent which extracts sentences from the sentence splitter using manually crafted answer extraction patterns. The score of each sentence is computed by the sentence scoring subcomponent. Then, the answer selection algorithm selects top 5 non-redundant sentences from the candidate answer set. Finally, the sentences are generated by the sentence generator subcomponent and returned to the user. List Answer Extraction contains candidate answer selection, rules and the gazetteers are incorporated to extract answers. Questions asking about a thing are matched with the rules developed and place name based questions are matched with the gazetteers and answer selection selects the answer.

Chapter Five

System Evaluation and Results

5.1. Overview

In this Chapter we describe the experimentation environment, the data set, the evaluation metrics, and the results of the performance of the Afaan Oromo definition, description, and list QA system components. In addition discussion on the methods used, activities followed, and results obtained are presented in detail.

5.2. Experimentation Environment

The prototype was developed using Java programming language. Standard Java classes from `java.io` and `java.util` packages were used in collaboration with Lucene and external Java libraries to perform tasks such as accessing files, dealing with arrays and handling exceptions. The eclipse Java editor has been used to develop our system.

Java programming language

Java is one of the important programming languages and computing platform for many applications. It's released by Sun Microsystems in 1995. Many applications use Java specially web application high potential in the field of programming web, games, databases, and many other applications. It has many compilers and editors such as text pad, Eclipse platform, and NetBeans platform.

In addition, Java has the following characteristics:

- ✓ Secure.
- ✓ Fast.
- ✓ Reliable.
- ✓ Java works in many machines such as laptops, datacenters, game consoles, supercomputers, cell phones

There are a lot of applications and websites that will not work, unless you have Java installed, and more are created every day. From laptops to datacenters, game consoles to scientific supercomputers, cell phones to the Internet, Java is everywhere!

Lucene search engine

Apache Lucene is a text search engine with high performance, full search featured library written in Java, it is suitable for nearly any application that requires full-text search, especially cross platform which is an open source project available for free. It can be enhanced and modified depending on user requirements to be more efficient. The project used it to stem words and designing patterns to enhance the results of search.

The system is developed and tested on a computer with the following specifications:

- ✓ Windows 10 pro operating system,
- ✓ Installed memory RAM size 8GB,
- ✓ Hard disk size 550GB and
- ✓ Processor Intel(R) Core(TM) [i5-6440HQCPU@2.60GHZ](#) 2.59 GHZ.

5.3. Evaluation Criteria

For evaluation, three parameters are used: precision, recall, and F-measure. They are slightly the same across all benchmarks. Automatic QA evaluation has been studied from a number of different perspectives such as the use of test collections reading comprehension tests, and the application of automatic systems that evaluate the correctness of the answers returned by comparing them with human-proposed answers to the same set of questions.

Evaluation for QA system mainly focuses on the accuracy of the answers returned. The accuracy of question classification module is done by evaluating the percentage of correctly identified question types. The percentage is computed by taking the ratio of correctly identified questions to the total test questions. The document selection and answer extraction modules are evaluated by precision, recall and F-score. Generally, the performance of a question classifier can be measured by computing the accuracy of a particular classifier on a test set.

5.3.1. Question Classification Evaluation

A correct answer for a question would probably be extracted if its question type and the expected answer type are correctly identified. Since this task is performed by the question classifier, its performance should be evaluated. Question classification is one task of question analysis used to classify questions into its intended types and identify an expected answer types. The performance of question classification is crucial for answer extraction, i.e., wrongly classified questions will lead to return wrong or No answer as result. The question classification is evaluated by the percentage of correctly and wrongly classified questions. Where, the percentage is computed by taking the ratio of correctly identified questions to the total test questions. The experiment is conducted on 270 test questions that are chosen from list, definition and description question types and the system correctly classified 98%, 99% and 97% respectively.

5.3.2. Document Selection Evaluation

Generally, there is a range of tests required to measure the effectiveness and performance of IR systems. These tests consist of documents collection (indexed documents according to the system needs), questions in a natural language, and a set of relevant judgments of questions. In terms of systems evaluation, two types of cases can occur, namely: ranked and unranked retrieval. In the case of rank retrieval, the evaluation depends on the list of retrieved documents and sorts them for the best match with the question pattern and keyword. While the case of unranked retrieval, the QA system is evaluated by using the following formulas. The standard approach to information retrieval system revolves around the concept of relevant and non-relevant documents [1]. Document retrieval systems are evaluated with respect to the notion of relevance judgment by human that a document is relevant to a query based on the presence of correct answer particles on the retrieved documents. In information retrieval system, precision and recall are defined in terms of a set of retrieved and relevant documents as follows: Precision is the ratio of the number of relevant documents returned to the number of documents returned, used to assess the measure of how many of the documents returned for a given query are actually relevant.

For text-span questions whose answer is string(s), we need to compare the predicted string(s) with the ground truth answer string(s) (i.e., the correct answer). RCstyle QA task generally uses evaluation metrics Exact Match (EM) and F1 score (F1) proposed by Rajpurkar ET al. for text-span questions. EM assigns credit 1.0 to questions whose predicted answer is exactly the same as the ground truth answer and 0.0 otherwise, so the computation of EM is the same as the metric Accuracy but for different categories of RC-style QA. F1 measures the average word overlap between the predicted answer and the ground truth answer. These two answers are both considered as bag of words with lower cases and ignored the punctuation and articles 'a', 'an' and 'the'. For example, the answer 'The Question Answering System' is treated as a set of words question, answering, system. Therefore, F1 of each text-span question can be computed at word-level by Equation

Precision (P): which represents the percentage of retrieved documents related to the query.

$$\text{Precision (P)} = \frac{\text{Number of relevant documents retrieved}}{\text{Retrieved total number of documents}}$$

Recall (R), which represents the percentage of related documents retrieved. It is the ratio of the number of relevant documents returned to the total number of relevant documents in the collection.

$$\text{Recall (R)} = \frac{\text{Number of relevant retrieved documents}}{\text{Relevant total number of documents}}$$

F - Measure (F), which represents the percentage of combination for precision and recall. Synthetic measure from both precision and recall, the harmonic mean between the two (known as F1 or F score) is also used.

$$\text{F - Measure (F)} = \frac{2 * P * R}{P + R}$$

The document retrieval performance has been evaluated based on the presence of correct answer particles on the retrieved documents. The performance of our document selection

component is evaluated by 75 questions on 250 documents. According to the query search result, our system scored a recall of 0.87, 0.686 precision and 0.767 of F-score.

5.3.3. Answer Extraction Evaluation.

The QA system can be evaluated either the whole system and / or single module. Therefore, the above formulas are not valuable because they assess the system's effectiveness in retrieving relevant documents without highlighting the system's ability to deliver documents ranked as relevant. Question (input) information retrieval (“search”) - answer extraction (from either text or knowledge base) - answer (output). The answer extraction component is responsible to extract answer from the relevant documents which are retrieved by the document retrieval. It is evaluated using precision, recall and F-score by comparing the answers that our system returned with manually constructed answers using 50 test questions for each question types.

Precision: percentage of instances returned that are correct. It is the outcome of this formula is a number between 0 and 1 which indicates the probability of providing a correct answer by the QA system.

$$\text{Precision} = \frac{\text{Number of correct answers}}{\text{Returned total number of answers}}$$

Recall: percentage of the expected correct instances that are returned.

$$\text{Recall} = \frac{\text{Number of correct answers}}{\text{Expected total number of answers}}$$

Recall is a weighted harmonic mean of precision and recall, equation 6 were used to evaluate the quality of an answer.

Question Type	Precision	Recall	F-score
Definition	0.629	0.744	0.682
Description	0.562	0.720	0.64
List	0.7	0.707	0.649
Average	0.597	0.724	0.654

Table 5:1 the Answer Extraction component Recall, Precision, and F-score result.

From Table 5:1, we conclude that the answer extraction patterns of definition are better than description because the F-score of definition is greater than the F-score of description. Description terms are incorporated within their definition which leads the F-score result of the description to be less. The F-score of list question is also good (better than the description).

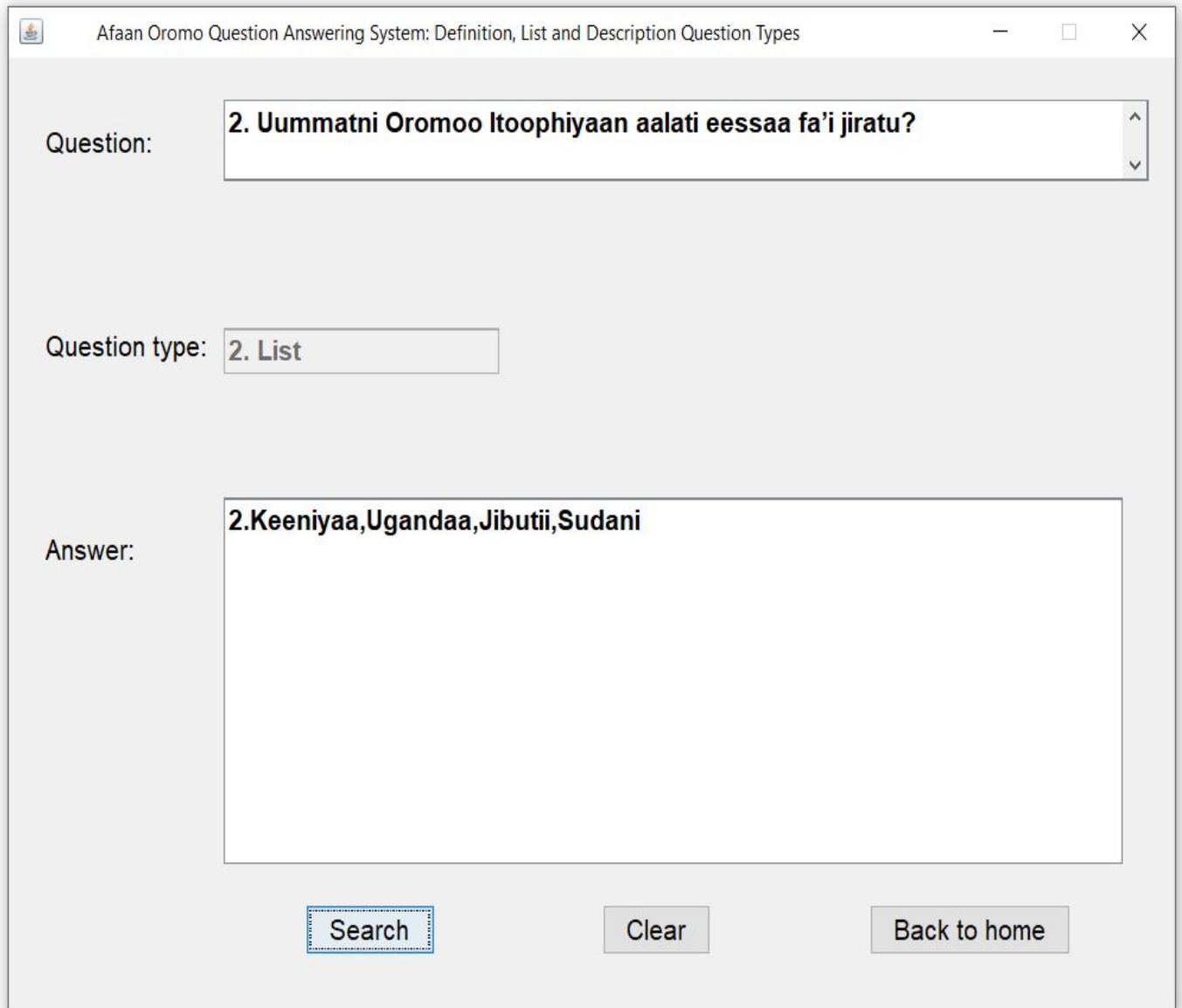


Figure 5.2 Screenshot of Correct List Answer Example

As Figure 5.2 shows, the answer for the question “Uummatni Oromoo Itoophiyaan aalati eessaa fa'i jiratu??” (“Lists where Oromo people are found out of Ethiopia?”) Return Answer Kenya, Uganda, jibout and Sudan exist in document corpus about Oromo people.

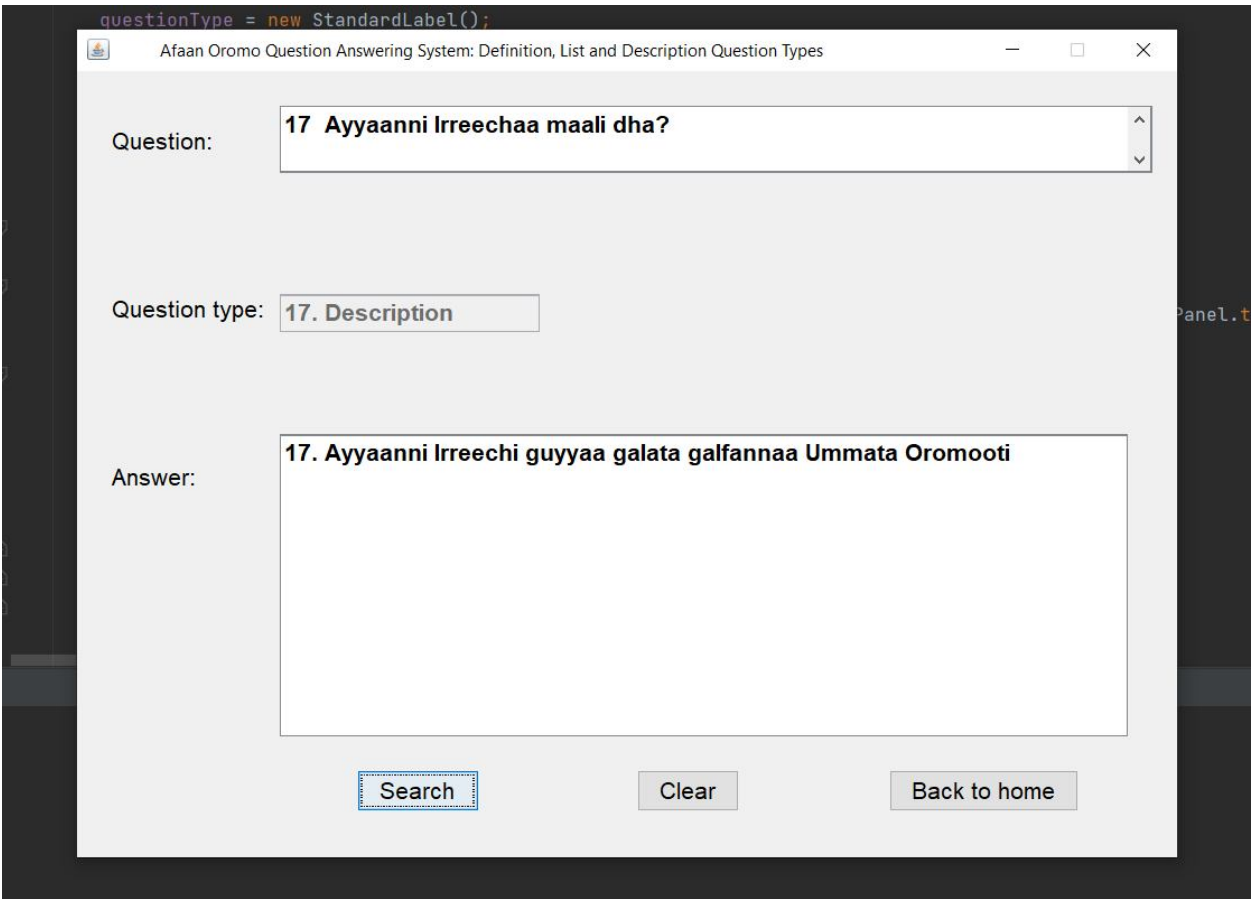


Figure 5.3 Screenshot of Correct Description Answer Example

As Figure 5.3 shows, the answer for the question “Ayyaanni irreechaa maali dha??” (“Describes what Irecha festival is?”) Return Answer ayyaanni irreechi guyyaa galata galfannaa ummata oromoti exist in document corpus about irecha which means irecha is Oromo people thanks day.

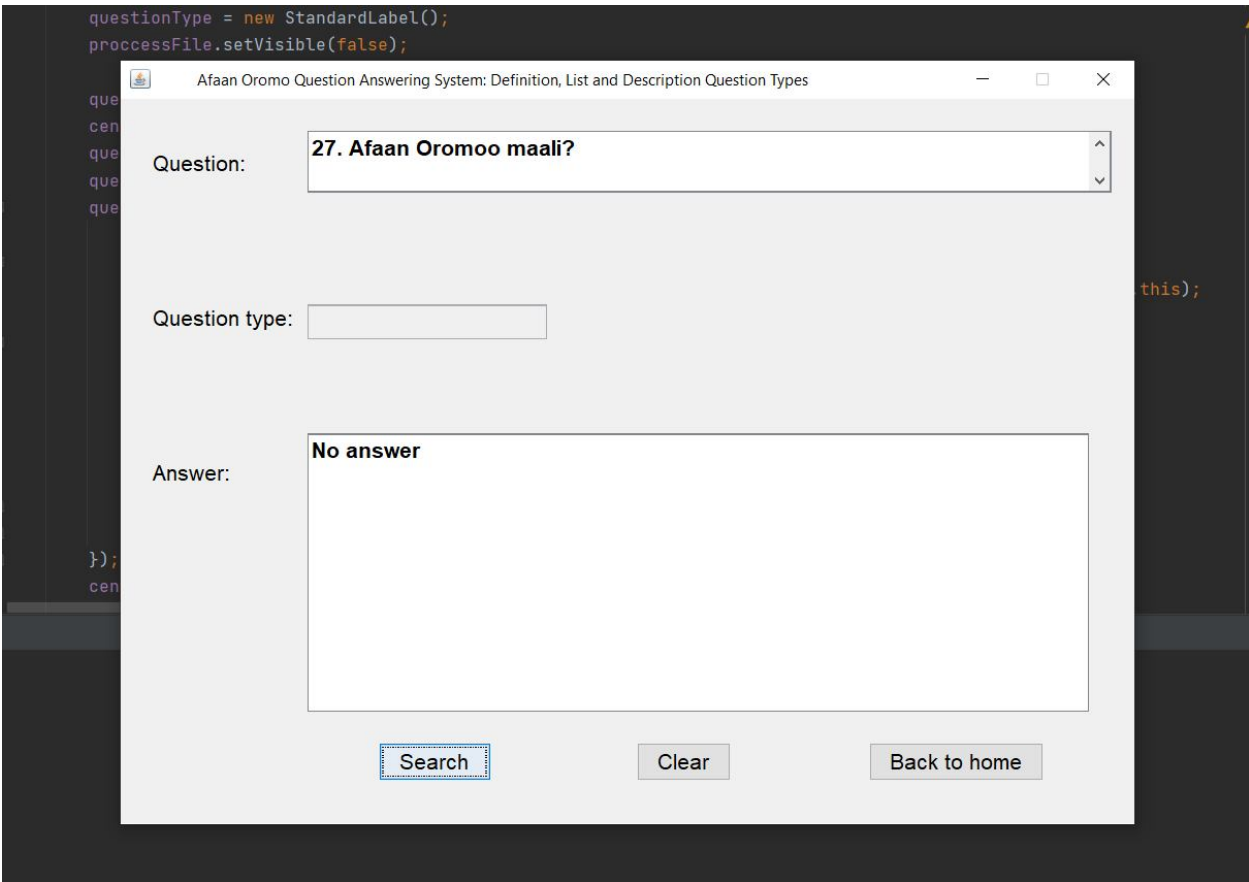


Figure 5.4 Screenshot of No answer Description Example

As Figure 5.4 shows, the answer for the question "Afaan Oromoo maali?" ("Definition about what is Afaan Oromo?") Return No Answer but, there is a document in the corpus about Afaan Oromo, the error occurred due to document retrieval.

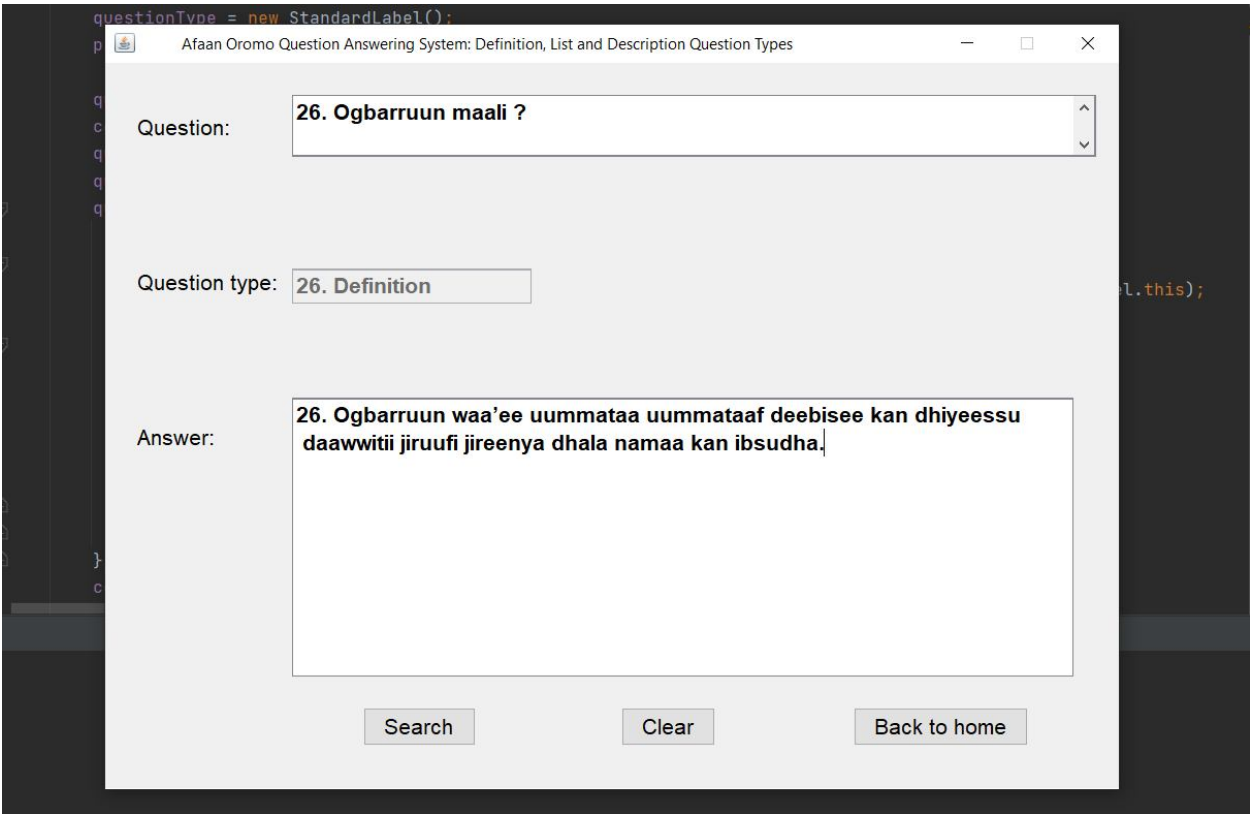


Figure 5.5 Screenshot of No answer Definition Example

As Figure 5.5 shows, the answer for the question "Ogbarruun maali?" ("Definition about what is literature?") Return Ogbarruun waa'ee uummataa uummataaf deebisee Kan dhiyeessu daawwitii jiruufi jireenya dhala Nama Kan ibsudhaAnswer which exist in document in the corpus.

5.4. Discussion

This part depicts the results of our system experiments that were conducted to measure the system performance in the context of QA. During the evaluation we used two evaluation criteria. The first evaluation criterion is used for question classification component, computes correctly classified question types. The other evaluation criterion was precision, recall and F-score used for evaluating document selection and answer extraction components. In doing so, we faced some issues which are listed below.

- ✓ Spelling errors in extracting correct answer, for example instead of writing the word "Laafaa" (culture) if the question term is written as "lafaa" it leads to return no answer.

- ✓ Even though the techniques we used in this thesis have performed well, there are questions which are not answered correctly and got answers that contain sentences unrelated to them. Improvements to the stemmer and specially the morphological analyzer probably result in improvement of performance.
- ✓ We observed that documents with more number of user's query terms have higher probability of correct answer matching.

Chapter Six

Summary, Conclusion, Recommendations and Future Works

This chapter focuses on summaries that indicate the whole picture of the study, conclusion based on the findings of the experiment and recommendations that the researcher has suggested as the future work.

6.1. Summary

QAS is important in retrieving relevant answers for user's natural question. Unlike that of common search engines like Google, Yahoo, etc. that return a ranked list of documents QAS returns an exact answer to users' natural language questions. This study has focused on List, definition and description question answering system for Afaan Oromo language. The objective of this research is to explore the possibility to design and develop Afaan Oromo Question Answering System, definition, list and description question types. Pattern based approach was employed to extract the words to be defined from the natural language questions and also to identify document corpus used for the study.

With the help of natural language processing, the information extraction is performed automatically and the user will be presented with answers believed to fulfill the user's request. Question answering systems could use preformatted corpora and provide concise answers in the form of paragraphs, sentences or phrases to natural language questions.

Opening the address could present the user with lots of pages of information and it is the user's duty to go through the information and extract the actual fact. This is how ordinary search engines help users in need of information. What they do is accept the users' query, search documents in their repository which contains any of the words in the query, rank the retrieved documents and present to the user with title of the page, a snippet, and address (URL) of the page included in the response.

6.2. Conclusion

In this thesis, the aim was to improve accuracy of retrieving answers to the user questions in a restricted QAS. The approach to answering list, definition and description questions. Also lists future works for the question answering system.

Question answering system is one of the applications of NLP that provides precise answers to human language questions. QA system for definition and description question could allow someone to know about a term. QA system for list question provides a list of answers for a question. We developed Afaan Oromo list, definitional and description QA system. Afaan Oromo is morphologically rich, so we tried to use a morphological analyzer for simplifying the complexity of words which allows as in generating root words.

We have used a preprocessing technique, in which the data sets were preprocessed using the tasks such as tokenization, case normalization, stop word removal, short word expansion, stemming, lemmatization and indexing allows us to have the same standard between query terms and index terms. Rule based question classification model were used for classifying users natural language question, which classify users query to their semantic types, queries were generated by removing the interrogative terms after the queries are pre-processed, which allows us to know the kind of information the question is asking for and also to retrieve relevant documents from indexed file. Retrieved documents need to be filtered in order to provide correct answer for the user and we showed how to filter relevant documents from the irrelevant one.

We have used two different methods in extracting answers for list, definition and description questions. The first is a pattern matching (regular expression) method for extracting answers for definition, description and list (where the focus is thing) questions. The other answer extraction method is a gazetteer (NER) for answering list question (where the question focus is place). The performance of the QA system is affected by the different components as mentioned in the experiment section.

In order to evaluate the performance of our system we used two criteria. The first criterion was percentage for evaluating question classification component which classified 98.3% correctly and the other criterion were precision, recall and F-score for evaluating document selection

and answer extraction components. The document selection component is tested and scored 0.767 of F-score. The answer extraction component is evaluated with an average precision, recall and F score, 0.596, 0.723 and 0.653 results are obtained respectively.

6.3. Contribution

The contributions of this thesis work are summarized as follows:

- ✓ Applying highlights the existing systems of Afaan Oromo QA that will help us compare and measure our contribution to other systems.
- ✓ Making use of rule based automatic question classification model for IAO list, definition and description questions types.
- ✓ Making use of information extraction technique to capture the embedded information in documents.
- ✓ Extending the existing AOLDDQAS of Afaan Oromo QAS for answering non-factoid question.

6.4. Recommendations and Future Works

In this thesis work, we present question answering systems that can be applied in specific sectors, Improving Afaan Oromo Question Answering System: Definition, List and Description Question Types for Non-factoid Questions answering requires deep analysis of the question as well as the corpus by using NLP tools. So, there are a number of rooms for improvement and modification for Afaan Oromo question answering system. Some of the recommendation we propose for future works are listed below.

- ✓ Applying all the patterns in one prototype to add more enhances to the question answering system for Afaan Oromo language.
- ✓ Applying prototypes in other domains such as fully support QA verify the effectiveness of the system using a larger-scale special domain database and transfer to other domains.
- ✓ Applying Query expansion for future researches to improve the performance of the system by exploring and integrating query expansion techniques.

- ✓ Doing speech based question answering is highly recommended future work for the Afaan Oromo language.
- ✓ Using Query expansion for future researches to improve the performance of the system by exploring and integrating query expansion techniques.
- ✓ Integrating Afaan Oromo spelling checkers with Amharic question answering systems to enhance the performance of the system.

References

- [1] Aberash Tesfaye, Afaan Oromo Question Answering System for Factoid Questions, Unpublished MSc Thesis, Department of Computer Science, Addis Ababa University, July 2014.
- [2] Amanuel Raga Yadate LINGUISTIC SEXISM IN GENDER ASSIGNMENT SYSTEMS OF AFAN OROMO, AMHARIC, AND GAMO, Doctor of Philosophy in Linguistics, 20 September 2019, Bologna.
- [3] Djoerd Hiemstra, Information Retrieval Models, University of Twente, November 2009.
- [4] Haiqing Hu A Study on Question Answering System Using Integrated Retrieval Method, Information Science and Systems Engineering the University of Tokushima, Tokushima, Japan February, 2006.
- [5] Ann Copestake, Natural Language Processing: part 1 of lecture notes, 2003, 8 Lectures, (aac@cl.cam.ac.uk), <http://www.cl.cam.ac.uk/users/aac/lectures.pdf>, Access date: Feb. 12, 2015.
- [6] Alvin Andhika Zulen and Ayu Purwarianti, Study and Implementation of Monolingual Approach on Indonesian Question Answering for Factoid and Non-Factoid Question, 25th Pacific Asia Conference on Language, Information and Computation, pp.622–631, 2011.
- [7] Patricia Nunes Gonçalves, António Branco "Open-Domain Web-Based List Question Answering with LX-ListQuestion" University of Lisbon, June 2014.
- [8] Saad Ahmad, Tutorial on Natural Language Processing, University of Northern Iowa, United States, 2007.
- [9] Meskerem Derese, "Afaan Oromo List Question Answering System", WOLLEGA UNIVERSITY, DEPARTEMENT OF INFORMATICS.

- [10] Desalegn Abebaw Zeleke, LETEYEQ: A Web Based Amharic Question Answering System for Factoid Questions Using Machine Learning Approach, Unpublished Master's Thesis, Computer Science Department, Addis Ababa University, 2013.
- [11] Chaltu Fita: Afaan Oromo List, Definition and Description Question Answering System, MSc Thesis, Department of Computer Science, Addis Ababa University, 2016.
- [12] Seid Muhe, TETEYEQ: Amharic Question Answering System for Factoid Questions, Unpublished MSc Thesis, Department of Computer Science, Addis Ababa University, 2009.
- [13] Tilahun Abedissa, Amharic Question Answering For Definitional, Biographical and Description Questions, Unpublished Master's Thesis, Computer Science Department, Addis Ababa University, Addis Ababa, Ethiopia, November 2013.
- [14] <http://lucene.apache.org/core/>, Last Accessed on May 27, 2015.
- [15] Debela Tesfaye and Ermias Abebe, Designing a Rule Based Stemmer for Afaan Oromo Text, International journal of Computational Linguistics, Addis Ababa, vol. 1, no. 2, pp.1-11.
- [16] Michael Gasser, HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya, Indiana University, July 2012.
- [17] G. Suresh kumar and G. Zayaraz, 'Concept relation extraction using Naïve 116 Bayes classifier for ontology-based question answering systems', Journal of King Saud University Computer and Information Sciences, May 2014.
- [18] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, Introduction to information retrieval, 2008.
- [19] Miriam Fernández Sánchez. —Semantically Enhanced Information Retrieval: An ontology based approach, Doctoral dissertation, Universidad de Autónoma, Madrid, unpublished 2009.
- [20] DEREJE MAMO BEYENE, AMHARIC FACTOID QUESTION ANSWERING SYSTEM, BAHIR DAR, OCTOBER, 2017.

- [21] Burger, J. D. MITRE's Qanda at TREC-12, the Twelfth Text REtrieval Conference, NIST Special Publication SP 500-255, 2004.
- [22] L. Hirschman and R. Gaizauskas, Natural language question answering: the view from here, Natural Language Engineering, vol. 7, no.4, pp. 275-300, 2001.
- [23] Wesley Hildebrandt, Boris Katz, and Jimmy Lin. Answering Definition Questions Using Multiple Knowledge Sources, In Proceedings of the 12th Text REtrieval Conference (TREC 2003).
- [24] Christof Monz Document Retrieval in the Context of Question Answering, University of Amsterdam, Netherlands, 2003.
- [25] Håkan Sundblad, Question Classification in Question Answering Systems, Dissertation of Department of Computer and Information Science, University of Linköping, Sweden, 2007.
- [26] Patricia Nunes Gonçalves, António Branco "Open-Domain Web-Based List Question Answering with LX-ListQuestion" University of Lisbon, June 2014.
- [27] B. Eshetu, Amharic Question Answering for list questions: A case of Ethiopian tourism, Addis Ababa University, Addis Ababa, Ethiopia, 2013.
- [28] K. Abdissa, "Factoid Question Answering for Afaan Oromo," Addis Ababa University, Addis Ababa, Ethiopia, June, 2014.
- [29] K. Haftu, Tigrigna Question Answering System for Factoid Questions, Addis Ababa University, Addis Ababa, Ethiopia, 17 JUNE, 2016.
- [30] Dissertation in factoid question answering systems for English language.
- [31] Tesfaye Guta Debela, AFAAN OROMO SEARCH ENGINE, MSc Thesis, Department of Computer Science, Addis Ababa University, November 2010.
- [32] Meiws C.G., "A grammatical sketch of Written Oromo", ISBN 3- 89645- 039-5, 2001.
- [33] Tesfaye Guta Debela, AFAAN OROMO SEARCH ENGINE, MSc Thesis, Department of Computer Science, Addis Ababa University, November 2010.

[34] DEJENE HUNDESSA: DEFINITION QUESTION ANSWERING SYSTEM FOR AFAN OROMO LANGUAGE, MSc Thesis, SCHOOL OF INFORMATION SCIENCES, Addis Ababa University, 2015.

Appendices

Appendix 1: some of Afaan Oromo Short words and their Expansion

Abbreviation/Acronym	Afaan Oromo word(s)	Gloss
A.L.A.	Akka Lakkofsa Awurophaa	European calendar (Gregorian calendar)
A.L.I. (also ALH)	Akka Lakkofsa Itophiya (Habashaa)	Ethiopian calendar
Dh.K.B.	Dhaloota Kirstoos Booda	A.D. (anno Domini; in the year of our Lord).
Dh.K.D.	Dhaloota Kirstoos Dura	B.C.
FK.	fakeenyaafi	For example
Kkf	Kan kana fakkaatu	This and the like [and so on]
L.Bil.	Lakkoofsa bilbilaa	Telephone number
WB	Waaree booda	After noon (pm)
WD	Waaree dura	Morning (am)

Appendix 2: List of place names

Country Names	City Names	Capital Names
Itoophiyaa	Adaamaa	Finfinnee
Ameerikaa	Alamata	Landan
Jarmanii	Amboo	Waashingitan
Inglizii	Arbaa Minchi	Abu Dhaabii
Jaappan	Asoosaa	Abujaa
Chaayinaa	Asallaa	Akraa
Meeksikoo	Asasaa	Aljersi
Kanaadaa	Asaayitaa	Amaan

Appendix 3: Sample Test Questions and their Question Type

NO	Question	Question Types			
		Definition	List	Description	Unknown
1	Yuunivarsiitoota naannoo Oromiyaatti argaman kami fa'i?		List		

2	Uummatni Oromo Itoophiyaan aalati eessaa fa'i jiratu?		List		
3	Fiigicha gabaabaa fi dheeraan Itoophiyaaf badhaasa warqee hedduu argamsiisan eenyuu fa'i?		List		
4	Aadaan Oromo Maali fa'i dha?		List		
5	Naannoo Oromiyaa keessaa Nama du'eef goodini ililchu kami fa'i?		List		
6	Oromoo Tuulamaa Godiina Shawaa Bahaa Aanaa Dugdaafi Booraa, Lixa Shawaafi Giddu Gala Shawaa keessatti argaman kami fa'i dha?		List		
7	Goosota shanan Gadaa Oromoo maali fa'i?		List		
8	Shanan Oromoo eessatti argamu?		List		
9	Maqaan Sadarkaa Gadaa Oromoo maali fa'i?		List		
10	Oromiyaa keessa odaawwan jiran maali fa'i?		List		

11	Dandeettiwwan afaanii arfan maali fa'i?		List		
12	Rakkoo baay'achuun uumataa fidu tarreessi?		List		
13	Gosoota albuudoota qama namaf fayyadan tarreessii?		List		
14	Haariiroowwan hawaasumaa beekamoo ta'an kami fa'i dha?		List		
15	Naannolleen Itoophiyaa keessaa horsiisee bulluun beekaman eenyu fa'i dha?		List		
16	Naannollee Itoophiyaa keessaa warqii oomishuun beekaman eenyuu fa'i?		List		
17	Pirootiiniin maddi isaa maal fa'i?		List		
18	Akaakuwwan Isiporti barreessi?		List		
19	Dalaggaawwan maashaa qaama keenyaa maal fa'aa dha?		List		

20	Faayidaaleen siinqeen qabdu tarreessi?		List		
21	Odaan Oromoo eenyuu fa'i dha?		List		
22	Goosonni Ayyaana ateetee maali fa'i dha?		List		
23	Ulaagaalee biiyyaa gurguddoo ta'an keessaa sadii tarreessaa?		List		
1	Aadaan Oromoo maali dha?	Definition			
2	Gosni Oromoo Warjii jedhamu Gosa Oromoo kamiitii?	Definition			
3	Pirofeesar Asmaroom eenyu?	Definition			
4	Lagni Burqaa maali?	Definition			
5	Sikkoo Mandoon maali?	Definition			
6	Odaan maali?	Definition			
7	Irreessi maali?	Definition			

8	Afaan maali?	Definition			
9	Ogbarruun maali?	Definition			
10	Afaan Oromoo maali?	Definition			
11	Odaan Maali?	Definition			
12	Maayikirooorganizimiin maali dha?	Definition			
13	Buddeeni maali?	Definition			
15	Callaan maali?	Definition			
16	Gadaan maali?	Definition			
17	Tullun maali dha?	Definition			
18	Dimookiraasii jechuun maal jechuu dha?	Definition			

19	Kuurunbaan maali?	Definition			
20	Barruu jechuun maal jechuu dha?	Definition			
1	Afaan Oromoo afaan hojii mootummaa federaalaaf maalif barbachise?			Description	
2	Inistitiyuutiin Aartii Oromiyaa maalif fayyaada?			Description	
3	Booranni fayyida maaliif Abbaa Gadaa ijoollummaatti filata?			Description	
4	Faayidaa Sirna Barnootaa maali?			Description	
5	Faayidaan dhadhaa maali?			Description	
6	Faayidaa barumsaa Afaani maali?			Description	
7	Ameerikan guudina aduunyaatiif gahee maali qabdi?			Description	
8	Gaheen Ilkani maali dha?			Description	

9	Qonnii faayidaa akkami kenna?			Description	
10	Fardi faayidaa akkami kenna?			Description	
11	Nyanni maaliif fayyada?			Description	
12	Roobin faayidaa maali keena?			Description	
13	Qootiyyoon maaliif fayyada?			Description	

Annex

```
package home;

import addNewQuestionAndAnswer.AddQuestionAndAnswerFrame;
import addNewQuestionAndAnswer.AddQuestionAndAnswerMainPanel;
import lucene.FilePathService;
import lucene.MainLucene;
import questionAndAnswers.QuestionAndAnswersFrame;
import standards.StandardButton;

import javax.swing.*.*;
import java.awt.*.*;
import java.awt.event.ActionEvent;
import java.awt.event.ActionListener;

public class HomePanel extends JPanel {
    private SpringLayout layout;
    private JLabel informationLabel;
    private StandardButton addNewQuestion, startQuestionAndAnswer, exitButton;
    private MainLucene mainLucene;
    private FilePathService filePathService;
    public HomePanel(){
        layout = new SpringLayout();
        setLayout(layout);

        filePathService = new FilePathService();
        mainLucene = new MainLucene(filePathService.getPath());
        informationLabel = new JLabel("Welcome to Afaan Oromo Question
Answering System",JLabel.CENTER);
        informationLabel.setFont(new Font("arial",Font.BOLD,15));
        add(informationLabel);

        layout.putConstraint(SpringLayout.WEST,informationLabel,50, SpringLayout.WEST,
this);

        layout.putConstraint(SpringLayout.NORTH,informationLabel,100, SpringLayout.NOR
TH,this);

        addNewQuestion = new StandardButton();
        addNewQuestion.setText("Add new question and answer");
        add(addNewQuestion);

        layout.putConstraint(SpringLayout.WEST,addNewQuestion,100, SpringLayout.WEST,t
his);

        layout.putConstraint(SpringLayout.NORTH,addNewQuestion,100, SpringLayout.NORTH
,informationLabel);
        addNewQuestion.addActionListener(new ActionListener() {
            @Override
            public void actionPerformed(ActionEvent e) {
                AddQuestionAndAnswerFrame addQuestionAndAnswerFrame = new
AddQuestionAndAnswerFrame();
                addQuestionAndAnswerFrame.setSize(new Dimension(500,600));
                addQuestionAndAnswerFrame.setVisible(true);
            }
        });
    }
}
```

```

        SwingUtilities.getWindowAncestor(HomePanel.this).dispose();
    }
});
startQuestionAndAnswer = new StandardButton();
startQuestionAndAnswer.setText("Just start question and answer");
add(startQuestionAndAnswer);

layout.putConstraint(SpringLayout.WEST, startQuestionAndAnswer, 100, SpringLayout.WEST, this);

layout.putConstraint(SpringLayout.NORTH, startQuestionAndAnswer, 75, SpringLayout.NORTH, addNewQuestion);

startQuestionAndAnswer.addActionListener(new ActionListener() {
    @Override
    public void actionPerformed(ActionEvent e) {
        QuestionAndAnswersFrame addQuestionAndAnswerFrame = new
QuestionAndAnswersFrame();
        addQuestionAndAnswerFrame.setVisible(true);
        SwingUtilities.getWindowAncestor(HomePanel.this).dispose();
    }
});

exitButton = new StandardButton();
exitButton.setText("Exit");
add(exitButton);

layout.putConstraint(SpringLayout.WEST, exitButton, 350, SpringLayout.WEST, this);
;

layout.putConstraint(SpringLayout.NORTH, exitButton, 100, SpringLayout.NORTH, startQuestionAndAnswer);
exitButton.addActionListener(new ActionListener() {
    @Override
    public void actionPerformed(ActionEvent e) {
        SwingUtilities.getWindowAncestor(HomePanel.this).dispose();
    }
});
});

}

public void showHome(){
    MainFrame mainFrame=new MainFrame();
    mainFrame.setVisible(true);
}
}

```