ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCES

DEFINITION QUESTION ANSWERING SYSTEM FOR AFAN OROMO LANGUAGE

BY

DEJENE HUNDESSA

OCTOBER, 2015

ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCES

DEFINITION QUESTION ANSWERING SYSTEM FOR AFAN OROMO LANGUAGE

BY

DEJENE HUNDESSA

A THESIS SUBMITED TO THE SCHOOL OF GRADUTE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTERS OF SCINECE IN INFORMATION SCIENCE

OCTOBER, 2015

ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCES

DEFINITION QUESTION ANSWERING SYSTEM FOR AFAN OROMO LANGUAGE

BY

DEJENE HUNDESSA

ADVISOR: MARTHA YIFIRU (phD)

Name and Signature of the Board of Examiners for Approval

	Name	**Signature**
1.	_____	_____
2.	_____	_____
3.	_____	_____
4.	_____	_____

Dedication

This thesis is dedicated to my uncle Efa Weyessa.

Acknowledgment

First of all I would like to thank **GOD** who guides me with his care in my life. I would like to gratefully thank my advisor Dr.Martha Yifiru for her support, understanding, guidance, encouragement and patience throughout the entire process of this thesis. Her guidance helped me in all the time of writing this thesis. Besides my Advisor my great gratitude goes to Dr.Solomon Tefera for his politeness and help to get the resource which was valuable for this thesis. My sincere thanks also goes to Dr.Million Meshesha for insightful comments.

I would like to thank Mr.Milki Mekuria , who had provided me legal documents that I have used for my evaluation corpus. I would like to thank my friends Biruk Retta and Birhanu Herano for their motivation and encouragement.

Finally I sincerely thank my family for their support and encouragement. Many thanks to my little sister Chaltu Efa for her love and good wishing for me.

Abstract

The amounts of electronic documents on the web are increasing explosively from time to time. Users were relying on IR system to explore their information of interest.IR cannot satisfy today's users' information need because it only returns a ranked list of documents rather than a precise and exact answers for users' needs. Question answering system is a solution to provide exact answers for users' information need with the help of information extraction techniques. The demand of users for retrieving precise information from the electronic Afan Oromo documents is increasing. An attempt had been done to develop factoid question answering for this language. In this study we have designed Afan Oromo Definition Question Answering system. The system architecture had been specified and its major components are question analysis, document processing and answer extraction. Pattern based approach had been used to extract target word from users' natural language question and definition from the corpus. We have found that the question analysis and answer extraction components are the major language dependent component of the system. And also we have noticed that the usage of stemming and synonyms list have a great effect on the performance of the system. Using the surface pattern approach, the definition extraction component correctly extracts 93.3% of the definiendum from users' question. Only 6.7% have been wrongly identified. The standard performance measures metrics recall, precision and F-measure was used for evaluation. We have performed the experiment with and without applying the stemming and synonyms. Accordingly without the usage of stemming and synonyms the system performance was precision 85%, recall 58.72% and F- measure 69.45%.And with the application of stemming and synonym the performance of the system was recall 78%, precision 86.4% and F-measure 81.9%. Usage of surface pattern approach for answer extraction from Afan Oromo documents had shown an encouraging result. Preparing patterns in detail for compound words more improve the performance of the system.

Keyword: Question answering, Afan Oromo Question Answering, Definition question

Table of Contents

List of Tables

List of Figures

List of Acronyms

EAT: Expected Answer Type

IE: Information Extraction

IR: Information Retrieval

MRR: Mean Reciprocal Rank

NE: Named Entity

NER: Named Entity Recognition

NLP: Natural Language Processing

QA: Question Answering

QAS: Question Answering System

TREC: Text Retrieval Conference

NLQ: Natural Language Question

MUC: Message Understanding Conference

DBOQAS: Data base -oriented question answering system

QARAB: A Question Answering System to Support the Arabic Language

START: Syntactic Analysis using Reversible Transformations

AQUASYS: An Arabic Question answering system based on extensive question analysis and answer relevance scoring

FHS: First Hit Success

FARR: First Answer Reciprocal Rank.

PREC: precision

List of Algorithms

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

Nowadays there is a huge volume of data available on the web. This huge volume of data on the web can satisfy most of the information need. But without the appropriate search facilities it is difficult to get the required information from the web documents. Search engines like Google, Yahoo, etc, help the users to get new concepts from different documents on the web. This is to mean that such kind of search engines return a ranked list of documents that contain the concepts that the user requested. Then the users by themselves go through the returned list of documents to filter the concepts that satisfy their needs (Zhang et al., 2004).

Because information retrieval systems do not have the capability to fully understand the users' questions, they return a list of documents that are not relevant to the queries as most of the search engines do. To get what they need the users provide the query to the search engine, then the search engine returns the documents related to the query terms by ranking them. The users themselves read and select the information of their interest from the returned document which is time consuming and the information returned by the search engines are not specific to the query.

Although information retrieval to some extent satisfies the user interest related to information need, these days' users need a better tool which can satisfy their interest more. First, most of the time the relevance of the search result returned by the search engines depends on the query formulated by the user. Thus, to get relevant documents from search engines, it demands on users to formulate queries that should maximize document matching. But users need a tool which enables them to reduce their time and effort to construct effective queries. Secondly, rather than a list of ranked documents users want to get real answers (Guo et al., 2007).

Therefore, without an appropriate search tools it is difficult to get the intended specific information from a large collection of data on the web. Question answering researches aim

to resolve these problems by providing the exact answers to the users' query(Denicia-Carral et al., 2006).

A question answering system (QAS) is an information retrieval application whose purpose is to help inexperienced users to access the information they need by enabling them to write a query in natural language and obtaining a specific answer to their query rather than a set of documents that contain the answer (Denicia-Carral et al., 2006).Complicated search tool is required because of the increased number of different types of information that are available on the web. Different successes in the area of question answering have been reported in the question answering evaluations which was started in 1999 as a part of Text retrieval Conference (TREC) in which two third of factual question types have been correctly answered in the series of these evaluations.

These successes in QAS and also the users' interest to get a short and concise description rather than a list of ranked document have motivated international interest towards question answering (Hirschman & Gaizauskas, 2001). Question answering system is a very challenging task because natural language is a powerful tool and questions can also be formulated in different formats. And also returning a short precise answer is more difficult than returning a list of related documents.

There are different types of question answering system like factoid, list, definition, cause/consequence, Evaluative/comparative, questions with examples, questions about opinion and etc. Factoid question answering system answers questions that are started with interrogative pronouns like when, who, how many, where, etc. It has exactly one correct answer which can be extracted from short text segment. For example, questions like "Who is the prime minister of Ethiopia in 2004?" and "When did Nelson Mandela die?" are some of the examples of factoid question types. List questions types require multiple facts to be returned in answer to a question. For example, list the universities in Ethiopia and list the famous athletes in Ethiopia are some of the list question types. Definition question types require a description for the target word. For example, in the question 'who is Nelson Mandela?' a description is required for the target word 'Mandela', thus this kind

of questions are definition question types. Definition question types are generally requested about organization or things. Even though dictionaries and encyclopedias can be a source for definition question types, they do not contain the latest information about a specific organization or things due to lack of immediate update (Trigui et al, 2008).

QAS that can automatically answer definition questions have been widely researched in the context of the question answering track of the (TREC) (Figueroa, 2010).A definition question answering system helps to extract the definition nuggets that contain the most descriptive information about the question target from a collection of documents. A nugget is a piece of relevant or factual information about the target term. And also it is a short sentence which shows the information about the question topic (Figueroa, 2010).

1.2 STATEMENT OF THE PROBLEM

The number of legal related electronic documents written in Afan Oromo (Latin script) is increasing from time to time. The information released on this kind of documents helps legal experts to perform their daily activities and also other users who want to have legal related information. Different attempts have been done to help the information seekers to access Afan Oromo electronic documents. For instance, Tesfaye (2010) has developed Afan Oromo search engine which aimed to enable users to retrieve Afan Oromo electronic documents. And also an attempt was done by (Gezehagn & Gutema, 2012) to design Afan Oromo text retrieval system. The aim of these researchers was to provide a list of ranked documents to the users. However, their works have limitations to provide specific and concise information for the users that in turn demand the users' more effort and time to read the returned documents to extract what they want(Guo, et al., 2007).

But currently users need tools that enable them to get precise and specific information from a large collection of online documents, because to use their time effectively and efficiently they do not prefer to go through the ranked list of documents. And also they want solutions that demand the users less time and effort to formulate queries for accessing information. Therefore, to return precise and specific result to users' question, a question answering system that extracts exact answers to user queries are required. Currently different QAS like list, factoid and definition have been researched by different

researchers in different countries for different languages. And also some attempts like Amharic Question answering for factoid Questions, Amharic definition question, Amharic question answering for list questions and Factoid question answering for Afan Oromo have been done to design QAS for local language.

To the knowledge of the researcher, no research attempt has been done on Afan Oromo definition question answering system. Because of this, the researcher motivated to conduct this research. Besides, definition question answering system that are designed for other foreign and local languages cannot be directly applied for Afan Oromo language because of the fact that the Afan Oromo language's syntactic, morphological and grammatical features are different from other languages. Moreover, Kasahun (2014) and Wondwossen(2013) have recommended that definition question types should be done for other local languages and this question types would be useful for many applications where a lot of information can be extracted.

The focus of this study is, therefore, exploring the possibility of developing definition question answering system for Afan Oromo language. The users of the system can provide their questions in the form of natural language and their questions are compared with the required documents to extract the exact definition.

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

- What is an effective definition extraction approach to extract answer for definition questions from Afan Oromo documents?
- What are the factors that may affect the performance of definition question answering system?
- What are the language dependent components of Afan Oromo definition question answering system?

1.3 OBJECTIVES OF THE STUDY

1.3.1 GENERAL OBJECTIVE OF THE STUDY

- ✓ The general objective of this research is to explore the possibility of developing Afan Oromo Definition question answering system.

1.3 2 SPECIFIC OBJECTIVES

In order to achieve the above general objective, the following specific objectives are listed:

- To review different literatures related to question answering systems to have a conceptual understanding on different approaches and techniques used.
- To study the structure of Afan Oromo language specific to definition question types.
- To prepare Afan Oromo legal related document corpus and questions those are used to evaluate the performance of the prototype system.
- To identify factors that can affect the performance of Afan Oromo definition answering.
- To identify the components of Afan Oromo Definition question answering that mainly affected by the language structure.
- To check the effectiveness of pattern based approach for Afan Oromo definition question answering.
- To develop a prototype for Afan Oromo definition question answering and evaluate its performance.

1.4 SCOPE AND LIMITATION OF THE PROPOSED STUDY

This study focused on a closed domain legal area. Closed domain QAS deals with questions from a specific domain like Law, Medicine, Geography and etc. The researcher has selected legal related area because the electronic documents are easily available and as the study focuses on definition question type answering, documents from this area contains more definition sentences which are suitable for this study. Because of the time issues the researcher didn't consider other question types like factoid and list to develop fully fledged

question answering system. Thus, this study focused on only definition question types. Stemming algorithm is one of the sub module used in this research, but it was not readily available for easy access. The researcher has spent much time in contacting the developer of Afan Oromo stemmer and customizing it. Corpus preparation for this study has taken more time. Because this work is the first of its kind for Afan Oromo language and no standard test corpus was available so far for this kind of question, it demanded the researcher to prepare the corpus which was time consuming.

1.5 METHODOLOGY OF THE STUDY

According to (Kothari, 2004) research methodology is a technique of systematically answering a research question. And also it is a way of studying how research is accomplished scientifically. The methodology part is crucial as the performance of the system is affected by the approaches and techniques followed by the researcher during the course of the study (Kasahun, 2014).The methodology of this study focuses on the approaches, techniques and the source of data that the researcher has been used during the study. Therefore, the methodology the researcher has been followed is discussed below:

1.5.1 RESEARCH DESIGN

This study is an experimental quantitative method that evaluates the implementation of Afan Oromo Definition Question Answering system. The researcher has used surface pattern based design approach to extract concept –description pairs from the Afan Oromo language documents.

1.5.2 LITERATURE REVIEW

Different literature, books and other scholarly published materials for the purpose of understanding the subject area of the study, to avoid duplication of research and also to go through the different techniques and algorithms that are applied by different researchers to design question answering system for different languages specifically giving attention to definition question answering system have been reviewed. Additionally local researches

that have been done on question answering system have been reviewed to get more understanding about the approaches and techniques followed by the researchers. And also since this study has been focused on designing Definition question answering system for Afan Oromo documents, characteristics and challenges of the Afan Oromo language have been studied in detail.

1.5.3 DATA COLLECTION

Afan Oromo corpora of legal documents that include proclamations, procedures, training materials and research papers have been collected from Oromia justice sector professionals Training and Legal research institute for the purpose of evaluating the performance of the proposed system. We have compiled 25 documents each containing ten pages as a total of 250 pages as a corpus.

The 25 documents were compiled from different legal proclamation, research papers, procedures and other legal related documents.35 definition questions have been prepared from the compiled documents by five individuals that two of them are legal professionals and the rest are non legal professionals for evaluating the performance of the system. The legal professionals have first degree in Law and the others have first degree in civil engineering, earth science and accounting. All of them are fluent speakers of Afan Oromo language. From a total of 35 questions 15 questions have been used for evaluating the performance of question analysis module and the rest 25 questions have been used for evaluating the performance of the system as a whole.

1.5.4 IMPLEMENTATION TOOL

For the development of the proposed prototype system, python programming language was used. This programming language has been selected because of the following:

- The researcher is familiar with the language.
- The language has many built in functions to process texts.
- It is simple to develop patterns as required.
- It is an interoperable and easily available tool.

- It has rich implementation of the widely used NLP algorithms.

1.5.5 TESTING AND MEASUREMENT PROCEDURES

For the purpose of evaluating the correctness of the answer returned by the prototype system, the information nuggets prepared by the legal domain expert were used.
As stated by (Figueroa, 2010) the performance of a question answering system is affected by each of the individual components in the system and thus, we have evaluated the question analysis component separately as it has a great effect on the performance of the system. And also we have evaluated the overall performance of the system with and without applying the synonym list and stemming.IR effectiveness measures mainly recall, precision and F-measure have been used to measure the performance of the system.

1.6 SIGNIFICANCE OF THE STUDY

The major significance of this study is to provide an input for designing and developing a full fledged Afan Oromo question answering system because this study provides prototype to definition questions which is a one step forward in producing a complete question answering system. And also it is a good start for helping users who want to get precise and exact answers for definition question types from Afan Oromo documents.

1.7 ORGANIZATION OF THE THESIS

This thesis has been organized into six chapters. The first chapter of the thesis talks about the background of the study, statement of the problem, general and specific objectives of the study, the methodology that has been used, scope and limitation of the study and the significance of the study. In chapter two different global and local literatures related to question answering have been reviewed.

Chapter three addresses the structure of Afan Oromo language related to question answering. Chapter four discusses the proposed Afan Oromo Definition Question

Answering System Architecture, the design approaches used to extract answers and each of the components of the proposed system. In chapter five dataset preparations, result analysis and experimental evaluation of the prototype is discussed, And also the core function and implementation of the major components of the proposed system are presented.

Finally depending on the findings of the research the Summary, conclusion and recommendations are presented in chapter six.

CHAPTER TWO

LITERATURE REVIEW

Users of the web in their daily live seeking information about things, persons, political, social and others which are required in their daily activities. Different search engines enable to retrieve information from the huge collection of resources on the web. But the content of the returned information by the current search engines are not as the user's needs. Question answering system is a solution which enables the users to get the exact answers to the query term rather than a ranked list of documents related to the query word. In this chapter the researcher had discussed overview of question answering system, approaches to question answering, question answering system architecture, related research works and others related concept to definition question answering system.

2.1 OVERVIEW OF QUESTION ANSWERING SYSTEM

Now a day because of the internet explosion there is a huge volume of data on the web. These huge volumes of data to some extent satisfy the need of users. But in order to make the retrieval task economical, it is necessary to use the appropriate search facilities and it is required to have a system which can allow a user to ask a question in everyday language and get the answer quickly with enough contexts to validate the returned answers.

Current search engines do not return exact answer to the users' question rather it returns a ranked list of documents that have some relation to the question key word term. This situation is a cause for the emergence of QAS.QAS is an area of natural language processing study intended to provide human users with a natural and appropriate interface for accessing information.QAS is taken as a combination of two related information access activities called information extraction (IE) and information retrieval (IR).But unlike them, the goal of QA is to provide exact and precise answers to users' questions posted in natural language.

Current successes in question answering (QA) have been reported in question answering evaluation which was started in 1999 as part of the text retrieval Conference (TREC).Two third of factual questions have been answered in series of this evaluation with best QAS.

The international interest and activities in question answering have been motivated by both an effective result in QAS and users' demands towards QAS(Hirschman& Gaizauskas,2001).

Question answering system is an information retrieval system application which enables an inexperienced user to formulate their queries in natural language easily in order to get the exact answer rather than a collection of documents that contain the answer (Denicia-Carral et al., 2006).According to (Denicia-Carral et al., 2006) QAS designed by different researchers have shown an encouraging result that in turn have motivated international interest and activity in QA. This issue begins from an invitation to the research community to discuss the performance, requirements, uses and challenges of QAS. In question answering system in order to answer a question a system should analyze the question, should find one or more answers from online resource, databases or text corpus and should provide the answer to the user in an appropriate form (Hirschman& Gaizauskas, 2001).

The difference between question answering and search engines can be seen in two aspects: In search engines a string of keywords are used as a search term but in the case of question answering a natural language question is used to search the required answer. And search engine returns a list of related documents or URL rather than returning an exact answer at phrase level or sentence level (Gupta & Vishal Gupta, 2012).

2.1.1 HISTORY OF QUESTION ANSWERING SYSTEM

Even though 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.

Baseball was the first question answering system which was designed for answering questions about base ball games played in the American season during a single season.

Even though Base ball was relatively sophisticated in its semantic structure and syntax of questions it was restricted to a single domain rather than dealing with a huge collection of text data (Hirschman & Gaizauskas, 2001).

LUNAR was the other early question answering system which was designed for the purpose of enabling a lunar geologist to access, compare and evaluate the chemical analysis data on the composition of soil and lunar rock that was accumulating as a result of the Apollo moon mission (Hirschman & Gaizauskas, 2001).Lunar was able to answer 90% of the question provided by working geologists at a lunar science convection in 1971.As it was stated in the paper the basic limitation of the Baseball and Lunar was both system have limited to answering question from a closed domain like structured data not from open ended unstructured collection of texts.

The questions provided to Baseball and Lunar were usually analyzed using linguistic knowledge to produce a canonical form, which was then used to formulate a standard database query. For example, for the question:"List the authors who have written books about business", to access information for this query an SQL query like :'select first Name, second Name, last Name from authors where title=""(Bill Woods,1973).These early QAS depend on having the knowledge needed for answering a question available in structured form, not as completely unstructured text. Although it shows that little research into QA as an independent task was being performed, most of the early research projects were concerned with related tasks that would form the basis for future QA research (J.Grosz et al., 1986).

2.1.2 APPLICATION OF QUESTION ANSWERING SYSTEM

QAS can be applied for different areas. According to (Hirschman & Gaizauskas, 2001) QAS's application can be categorized depending on the source of the answer. According to the authors these classifications include structured data, semi-structured data or free text. And also others include searching over a fixed set of documents, searching from the web, searching from book or collection of books like encyclopedia, domain specific QAS, domain independent QAS, annotated images and speed data.QAS can be applied in educational institution to help students to get short and precise answer for their question, used in financial institution to enable users in order get short information about the company like

year of establishment, their services, etc and QAS can be used by users in their day to day activities to get precise and short answer for their enquires.

2.1.3 USERS OF QUESTION ANSWERING SYSTEM

The users of QAS include those who use QAS for one time (demand-oriented) and use it in their day to day activities repeatedly. These classes of users require different interfaces in which the questions should be presented, answers to a question. To design an effective QAS it is important to consider the users who will use the system (Hirschman & Gaizauskas, 2001). AS it was stated by the researcher the interface should be simple and clear which users can use without any complexity when they use the system.

2.1.4 CLASSIFICATION OF QUESTION ANSWERING SYSTEM

As stated in (Dom`enech, 2007) QAS can be classified as open domain and closed domain depending on the domain in which the system is focused. Open domain question answering system deal with different types of question from different sources. It uses the World Wide Web or large amount text corpus as a data source to extract the answer.

Closed domain QAS deals with question from a specific domain like Law, Medicine, Geography and etc. And also QAS can be divided into database oriented and text based depending on the structure of the data used to the answer. Database oriented QAS uses the structured data stored in the database for the purpose of extracting the answer. In this QAS the main difficulty is to formulate database queries from natural language queries to extract the answer (Dom`enech, 2007).

Text based QAS use unstructured information as its data source for extracting answer. This unstructured information can be news papers, encyclopedia, manuals and books. In text based QAS the higher the amount of data in the corpus the more the probability of getting the answer in the text. Even though using huge data collection in text based approach helps to get more answer it increases the computation costs of searching the answer (Dom`enech, 2007).Question answering system can be classified into list, factoid and definition depending on question types.

FACTOID QAS: Factoid question types seek short fact based answers like entities, organizations, persons, dates and the like. For example:

- What is the capital city of Ethiopia?
- Who is the president of America in 2014?
- Which is the color of the sky?

The answer to a factoid question can be a noun (example: Which is the color of the sky? blue), a noun phrase (which is the color of the sky? slightly blue) or a named entity (who was the first black president of United States of America? Barack Obama).

LIST QAS: List question types require multiple facts to be returned as an answer to a question. For example:

- List 5 cities in Ethiopia.
- List 4 companies that manufacture computer.

DEFINITION QAS: Definition question types require textual answers that cover essential as well as non essential descriptions of the definiendum. Definiedum is a term which needs to be defined. For example:

- Who is Mandela? Mandela was the president of South Africa.
- What is aspirin? Aspirin is a drug.

Not all types of questions are answered by using only a single exact answer. For instance questions like 'what is information?' 'Who is Obama?' and 'Define law' do not answer only with a single answer. Such kind of question requires information that can describe the target term of the question. Answers to a definition questions are similar to the definition of terms in encyclopedia or biographical dictionaries. For instance, if the question asks for organization the answer may include information about date on which the company was established, what the company produces, the owner of the company and other interesting information about the company. Thus, definition question can be defined as a question

which requires textual answers that cover essential descriptions of the definiendum(Greenwood, 2005).

Definition questions can be natural language questions and most of them have the form 'Who/What is /was X?' In these question forms X is the definiendum which is referred as 'target term' of the question and it can be a person, thing, organization, event or object which the users want to have information about them from the system.

According to(Greenwood&Saggion,2004)a different strategy is required for answering definition questions to that used for answering factoid and list questions. This is because definition questions provide very little information which can be used to help find relevant definition-bearing passages in the text corpus apart from the definiendum. Embedding the target term in a question seems artificial. Users of electronic encyclopedias usually enter just the name of the thing they were interested in but not they have to enter a full question to find an article.

Even though question is more natural to the user it actually complicates the task of designing definitional QAS for researchers. With full questions it is easier to distinguish if the target term is a person or some other entity enabling target definition for a person to be constructed differently to those for an organization or generic name such as aspirin. By taking only the target as input there is no clear indication of what type of thing the target is (i.e. no words like who) and as such all targets are probably to be treated as the same. Test sets that are used in the TREC evaluation (from 2003 onwards) are currently the only accepted for definitional question answering and the researchers assume the same scenario for producing definition for a question.

According to (Greenwood, 2005) the currently accepted evaluation methodology for definitional questions mainly focuses on the inclusion in the definition of given nuggets of information. Nugget can be defined as a piece of information about the target which should be defined. For instance, when a system is asked to define 'Nelson Mandela', according to the TREC provided answer key the following facts like South Africa President, anti-apartheid revolutionary',' politician' and 'philanthropist' may be included in the definition. But this information does not fully describe who the person is. For example it does not contain information about when he was selected for the president, where he was born, where he was thought his higher education and the like. From this example it is essential to

consider that the TREC view point may not exactly represent that of the real world users (information which is contained in the encyclopedia or biographical dictionary entries).Even though systems designed to answer factoid question is similar to that of definitional question answering system, question analysis component of definition question is not equivalent to that of factoid question because of the little information contained in the factoid question (Greenwood,2005).In definitional QAS whatever the method to find relevant sentences the activities that should be performed is to cluster, rank and simplify the sentences to present a short precise definition. Different systems use indicative patterns in order to select highly relevant sentences or to extract short phrases of information (Ravichandran & Hovy, 2002).

There are also other question types like cause/consequence (example: what are the consequences of climate change?), Evaluative/comparative (example: What are the difference between hypothesis A and B?), Questions with examples (example: I am looking for a hotel in Addis Ababa near to airport.) and questions about opinion (example: What do people think about Mandela?).

2.1.5 EVALUATION OF QUESTION ANSWERING SYSTEM

As stated by (Hirschman & Gaizauskas, 2001) answers in QAS have to contain sufficient context in order to help the users to find the correct answers when multiple answers are presented by the system and also when the returned answer is not ranked at the top. According to the author it is easier to provide longer segments that contain an embedded answer than a short answer segments. In evaluation of QAS selecting the basic criteria for judging an answer is important to select the answer which can satisfy the users' questions. According to (Burger et al., 2001) the criteria that should be used to evaluate the answer of QAS include:

RELEVANCE: - It shows the returned answer should be equivalent to the user expectations.

CORRECTNESS: Depending on the question types the answer should be factually correct. For example question types like factual and list, the answer for these kinds of questions should be factually correct and approvable. The correctness of the answer returned by QAS like definition question should be judged by domain expert to validate their correctness.

CONCISENESS: The answer returned by the QAS should not contain irrelevant information and it should be brief and clear.

COMPLETENESS: The answer should contain all the appropriate points that should be contained in the answer.

COHERENCE: The answer should be consistence and simple

JUSTIFICATION: The returned answer should enable the user to find out why the answer was chosen as an answer to the question.

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 'in correct'.

2.1.6 PRESENTATION

Users interact with the QAS with the interface. Designing a good interface which enables the users to interact with the system like dialogue interaction increases usability of the system and users' satisfactions. The way in which the information can be presented to the user may be affected by the volume of the information that should be presented to the user, the context to be provided, returning short or long answers (Hirschman & Gaizauskas, 2001).As it is stated in the above section interface design can affect the usage of the system.

2.1.7 DISCIPLINES RELATED TO QUESTION ANSWERING

QAS has relation with scientific fields such as information retrieval, natural language processing and human computer interaction. Information retrieval focuses on extracting documents that best matches the query term.IR has components like query formulation, document analysis and relevance feedback that enable to retrieve documents.NLP enables to understand and generate natural language text in question answering process. Human computer interaction deals with how users interface with computers. Other scientific disciplines like recommender technology to search the required answer, knowledge representation and reasoning for question and answer analysis and multimedia information processing which helps to extract answers from audio or video sources and

also information visualization to display results play a crucial role in question answering system(Moldovan& Surdeanu,2003).

2.1.7.1 INFORMATION RETRIEVAL

The need to access and store information electronically was started after the invention of computers. Several research works were emerged in the mid 1950s related to information retrieval from a large collection of knowledge sources. Information retrieval is defined as the process of finding documents that are unstructured in their nature from a collection of documents that can satisfy the need of the users (Moldovan & Surdeanu, 2003).The main purpose of information retrieval is to extract documents that match a user's query. Indexing and searching are the main tasks in IR.

Indexing helps to quickly search the required documents from a collection of document sources by representing the documents by their key features when queries are issued to the IR system (D. Manning et al., 2009).Indexing techniques like term weighting gives a degree of importance to a word in a description and word proximity used to capture the linguistic structure between words (Moldovan & Surdeanu, 2003).Similarity algorithms that are based on term weighting are used to search documents that are relevant to the query term and for ordering the returned documents in terms of their relevance.

In addition term weighting helps to controls the IR precision (Moldovan & Surdeanu, 2003).Term weighting can be initial and relevance weights. In initial weight is used during the searching process while relevance weight is used after the required documents are retrieved for the purpose of ordering the returned list of documents (Moldovan & Surdeanu, 2003).Inverse document frequency (idf) is one of the type of term weight assignment which assigns weights to the keyword terms such that the weights are assigned inversely proportional to the frequency of occurrences of the keyword term in documents (Moldovan &Surdeanu,2003).Tf.idf (Term frequency inverse document frequency) is more improved weighting technique which multiplies the term frequency by the collection frequency.

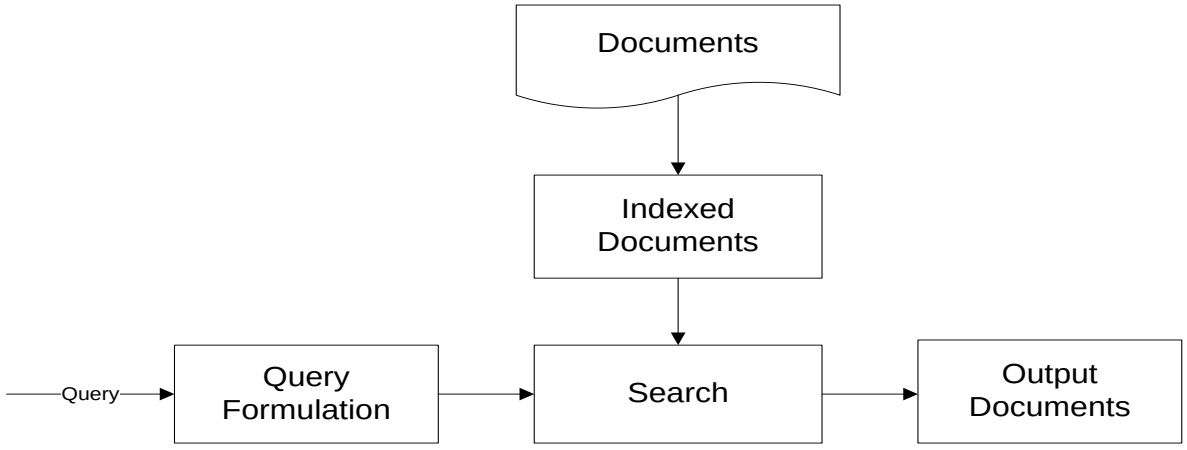

Fig 2.1.Generic IR System Architecture

IR is relevant to question answering in that IR techniques can be used to locate relevant documents that are used to extract the answer.IR is useful to QAS because of two reasons. The first reason is that IR techniques not only enable to return relevant documents but also relevant passages with in documents. Since the size of the returned passage can be reduced and taken as answer to a question, QA can be taken as passage retrieval by specifying some limit. Second, the recent QA evaluation was developed from systematic methodology evaluation which in turn developed by the IR community which is the most known today's selective of which are the annual Text Retrieval Conference (TREC) performed by the US National Institute of Standards and Technology. This QA evaluation methodology has initiated much of today's interest in QA (Hirschman & Gaizauskas, 2001).

2.1.7.2 INFORMATION EXTRACTION

Information extractions systems try to extract information from a collection of documents by populating some pre defined template (Moldovan & Surdeanu, 2003). It is defined as the task of filling predefined formats from natural language texts, where the formats are prepared to capture information about key role players in stereotypical events (Hirschman & Gaizauskas, 2001).

The information that is extracted by IE includes entities and the relation between them. The concept of IE was advanced with the Message Understanding Conference (MUC) in the late 1980's and early 1990's.IE focuses on a specific domain at a time and it is required

from the users to specify their need in some predefined formats. IE is one of the research areas that has included into the current TREC question answering track. IE template can be taken as expressing a question and a filled template as having an answer. Therefore, IE can be taken as a limited form of question answering in which the data from which the questions to be answered are a large dynamic collection of text that selected randomly. The message understanding conferences which was run between 1987 and 1998 was formulated by the IE community for the purpose of evaluating IES.

2.2 APPROACHES TO QUESTION ANSWERING

A realistic question answering system accepts a natural language question and accesses a source from which an answer should be extracted and returns an answer. All of the stage of the process performed for extracting the answer is executed automatically without any human intervention except inserting the input question. In this section three types of question answering systems are specified and their basic features are discussed. These QAS include Data base-oriented, text based and inference base.

2.2.1 DATA BASE ORIENTED SYSTEM

Data base -oriented question answering system(DBOQAS) use a traditional data base to store facts about things, places, people, events and the like which can be retrieved. These facts are used when system is processing users' query and extracting answers.

Natural language query will be translated into a database language query like SQL to retrieve the facts stored in the database. Because of these kinds of systems do not deal with the problem of answer extraction; they are regularly referred to as front -end system. In this approach standard data base techniques handle the problems related to answer extraction (Dom`enech, 2007). Some of the data base-oriented questions answering systems are LUNAR, START, BASEBALL and FREEBASE (Dom`enech, 2007).

2.2.2 TEXT CORPUS-BASED QAS

The data which is used for extracting an answer in text corpus based question answering system is not pre formatted (structured) unlike that of database- oriented QAS. So it is required to analyze both the question and data in order to extract an appropriate answer in the corpus.

Textual question answering systems are information systems that accept a natural language question as input, retrieve an answer from a huge volume of database of unstructured text and returns an exact answer which is a text string (Dom`enech, 2007).For instance, the large database of unstructured text may consist of newspaper. Textual QAS usually combines techniques from the fields of information retrieval and natural language processing. Sometimes textual QA is taken as corpus-based QA. Textual QAS can be closed or open domain system. Open domain QAS can take its input from all kind of question while closed domain QAS takes its inputs from only a restricted domain like law, medicinal or company's product.

According to (Dom`enech, 2007) in real situation almost there is no purely textual QAS as it is acceptable to store the already retrieved answers to frequently asked questions and to evaluate new questions for similarity with the extracted answers and also to use structured data in corresponding with unstructured data. Some QAS extract answer for natural language question only by comparing the new next question to previously asked questions. This approach of answer extraction is good for cases where questions with the same semantic contents likely asked repeatedly. Some of text based QAS are oracle which produces a syntactic analysis of both the question and text corpus which may contain an answer. This analysis converts the question and a text corpus into a canonical form, identifying the subject, object, verb and time and place indicator. Even though the approaches works for simple sentences it completely fails if the sentences are complex enough like if sentences have more than two objects.

2.2.3 INFERENCE BASED SYSTEM

Like data based –oriented QAS most inference-based systems need the data to be pre-prepared. Even though pre- formatting data is not required as crucial, it simplifies the

process of inference drawing. Inference –based QAS mainly focused on to figure out relationships that are not explicitly identified between entries in the knowledge base on the one hand and the question and the knowledge base on the other hand (Dom`enech, 2007).As it was stated by the researcher in inference based QAS to infer means to conclude or decide starting from something recognized or assumed. In other case to reason is to think rational and logically to form inferences or conclusion starting from the identified facts. Thus the reasoning process involves the understanding of inferences, starting from distinguished facts. To perform inferences means to draw out new facts starting from facts which are taken as true.

2.3 GENERAL QUESTION ANSWERING SYSTEM ARCHITECTURE

QAS can have different architecture. In the following section the most common architecture of QAS is presented. The main components of QAS that are used most of the time include Question Analysis, Document retrieval, Document analysis and answer extraction (Aunimo, 2007).

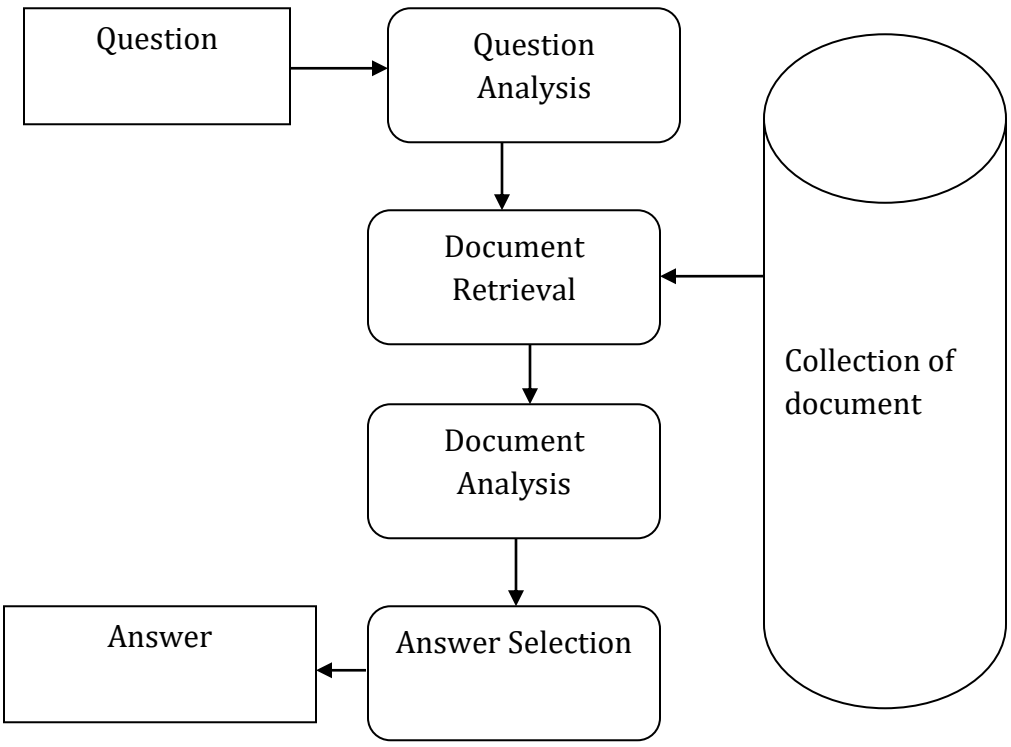

Fig 2.2 Generic Question Answering System Architecture

(Adopted from "Methods for Answer Extraction in Textual Question Answering", Aunimo, L. 2007).

2.3.1 DOCUMENT RETRIEVAL

According to (Greenwood, 2005) Question analysis component enables to construct an appropriate IR query. In addition to retrieving both documents and data the document retrieval component helps to determine the amount and structure of the text to be passed to the answer extraction component in the system.

Even though full syntactic and semantic parsing takes much time many QAS depend on extraction of answer for a natural language question. In real time QAS the number of time needed to process full syntactic and semantic representation affects their usage. In QA over closed domain such kind of problem may be resolved by pre parsing the whole collection, storing the resulting parse trees and semantic representation. In fact there is no fundamental reason why all question independent processing (Tokenization, POS tagging, named entity detection and the like) cannot be carried out in advance, as long as the time to retrieve the data is not longer than that taken to generate it as this would defeat the main use of pre-processing – faster question answering.

The document retrieval component plays a basic role in retrieving a subset of the entire collection of documents which will be processed in advance at question extraction stage. Specifying the IR model and the volume and structure of text to be retrieved at this stage is crucial. Even though many researchers have proposed that Boolean model is more appropriate for QA system for instance (Moldovan & Surdeanu, 1999; A. Greenwood& Saggion, 2004) many QA systems use ranked IR engines like (Robertson &Walker, 1999).

Even though query formulation can be done automatically using a ranked IR engine but it is more complex because of due care and consideration has to be given to how to rank or limit an unordered list of relevant documents.

As stated in (Greenwood, 2005) the document retrieval component should return a large number of documents per each question in order to guarantee that at least one relevant document is retrieved especially the report of (Hovey et al.,2000) show that still when considering top 1000 text segments there is no relevant documents are retrieved for 8% of the questions. According to (Gaizauskas et al., 2003) increasing the amount of text certainly increase the coverage of answer and it can decrease the accuracy of the answer extraction

components. As it was stated in (Gaizauskas et al., 2003) as the IR ranking with a coverage value of 69.2% when 200 pages are considered the coverage of retrieved documents continuously increases.

Therefore the document retrieval component is responsible for balancing the retrieved document coverage and answer extraction accuracy. Obviously the total amount of text that should be passed to the answer extraction module can be structured depending on the performance of the answer extraction component. For instance if the answer extraction component works better with a small amount of text then it shows that passing a few full documents will provide a good performance than passing a larger number of short relevant forming up the same volume of text.

A number of researchers reported many methods of segmenting full documents to give passage selection in an attempt to rise up the coverage of while keeping the volume of text to a minimum. (Monz,2004) suggest an approach to passage selection based not around fixed passage sizes but relatively round the smallest logical text unit to contain the question terms.

2.3.2 QUESTION ANALYSIS

As it is depicted in the above diagram question analysis is the first step in the process of QAS. A natural language question is submitted to the question analysis stage. This input may be presented in different ways like it is required from the user to use a subset of natural language 'controlled language ', which is constrained in terms of vocabulary and syntax. For instance most of front ends to database based QAS uses the above method. And also natural language question can be submitted to the system by filling pre prepared formats which can simplify the system's task of interpreting the question.

In addition to identify the input question explicitly, it is also possible to specify the natural language question implicitly. To identify the question implicitly the QAS should support an on-going dialog like some abbreviations in the question which requires access to dialog context to be described. The system's knowledge of the goal of the users can also be taken as implicit input (Hirschman & Gaizauskas, 2001).

To simplify the process of identifying answer in the next steps of QAS a more detailed analysis of the question should be performed. The first one is specifying the semantic type of the entity required by the question (for instance a date, person, company.ect). In this step it is required to look at the key question word (for instance when seeks a date time, where for a location and who for persons).

The other is specifying additional constraints on the answer entity (like determining keywords in the questions which will be used in matching candidate sentences which may be candidate answer. And specifying syntactic or semantic relations between a candidate answer entity and other entities or events listed in the question).And also at question analysis stage it is important to focus on the context. Context is considered when a question is a part of a longer sequence of questions or when it is a question scenario. To answer such kind of question it is required to consider the answer or scope of previous questions (Greenwood, 2005).

2.3.3 DOCUMENT ANALYSIS

After the candidate answer-bearing documents or candidate passages/segments have been retrieved at the document retrieval stage, these text segments should more analyzed to maximize the chance of getting correct answer for the questions. At the document analysis stage processes like syntactic analysis and semantic analysis are performed in order to improve the extraction of answer for a natural language questions. The documents are analyzed syntactically by using NLP techniques like part of speech tagging and named entity recognition in order to identify phrasal sentences that match with phrasal chunks identified at the question analysis stage to extract relevant sentences. At the semantic analysis stage semantic phrases are identified by performing shallow parsing in order to extract sentences that semantically similar to that of question (Athira et al., 2013).

2.3.4 ANSWER SELECTION

Even though at the question analysis, document retrieval and document analysis question processing is done in order to get relevant document or passages is tough, most of the task needed to answer a question performed at the answer extraction stage. Most answer

extraction modules depend on the relevant documents being subjected to different type of standard text processing ways to provide a better representation than just as the wards appear in the documents. This text processing techniques include tokenization, sentence splitting, POS tagging and named entity recognition. Depending on the approach selected to perform answer extraction, other techniques will be applied using the output from the above simple techniques.

For instance surface matching text patterns (Soubbotin & Soubbotin, 2001; Ravichandran & Hovy, 2002; Greenwood & Saggion, 2004) usually need no additional processing documents. These approaches by depending on a comparatively wide list of surface patterns to extract answers from the surface structure of the retrieved documents.

For example, according to (Ravichandran & Hovy, 2002) questions like its answer require a birth date can be answered by extracting answers by applying patterns such as:

<NAME>(<ANSWER>-)

<NAME> was born on <ANSWER>,

<NAME> was born <ANSWER>

Even though assembling broad lists of such patterns can be time consuming and difficult once they are assembled they can be exceptionally perfect for such a simple approach. According to (Soubbotin,2001) which used this approach as its main technique of answering questions was in fact the best performing system in the TREC2001 QA evaluation, correctly answering 69.1% of the questions.

As stated in (Greenwood,2005) surface matching text patterns work effectively using little in the way of NLP techniques other than essential named entity recognition. Semantic type extraction systems require more detailed processing of the documents which is used for answer extraction.

According to (Greenwood, 2004), answer extraction is simply done by extracting the most often happening entity of the expected answer type. In this approach identifying each entity of the expected answer type is required in free text resulting in more complex entity identification than is required by the surface matching text patterns. In order to get associations between questions and possible relevant documents QAS like FALCON (Harabagiu et al., 2000) and (Greenwood et al., 2002) use deep linguistic processing which require syntactic and semantic parsing. The representation of the question and candidate

answer-bearing texts are compared against each other and a collection of candidate answers are produced and ranked depending on their correctness (Hirschman & Gaizauskas, 2001).

According to (Nico, 2005) pattern matching approach is used to extract an answer from a text collection. The answer extracted is short text snippets, typically named entities or numeric or temporal expressions. As stated in (Hirschman & Gaizauskas, 2001).in addition to analyzing the question into an expected answer type some collection of additional constraints have analyzed the candidate documents as far as some explanation with semantic type extracted from the set of answer types.

Thus, the matching process may necessitate first that a text unit from a candidate answer text (possibly a sentence, if sentence splitting has been carried out) contain a string whose semantic type matches that of the expected answer. Matching here can be far type substitution (perhaps constructed as hyponymy in a lexical resource such as Word Net) and need not be restricted to identity. After a text component containing an expected answer type has been identified, other constraints may be applied to the identified text. Those text units that fail satisfy the constraints do not taken as candidate answers. There are difference among systems like the types of constraints used, how constraint fulfillment is performed and how weights are assigned to constraints.

2.4 QUESTION ANSWERING USING TEXT MINING APPROACH

The figure below shows that how definition patterns are discovered and answer is extracted from the web (text collection).The pattern discovery module uses a small set of concept description pair to collect from the web an extended set of definition instances. Then, it applies a text mining method on the collected instances to discover a set of definition surface patterns. The answer extraction module applies a definition catalog consisting of a set of potential concept-description pairs created by applying the discovered patterns over a target document collection in the answer extraction module. When an answer is required for a given question; it extracts from the catalog the set of associated descriptions to the requested concept. Finally, it retrieves the selected descriptions to find the more adequate answer to the given question.

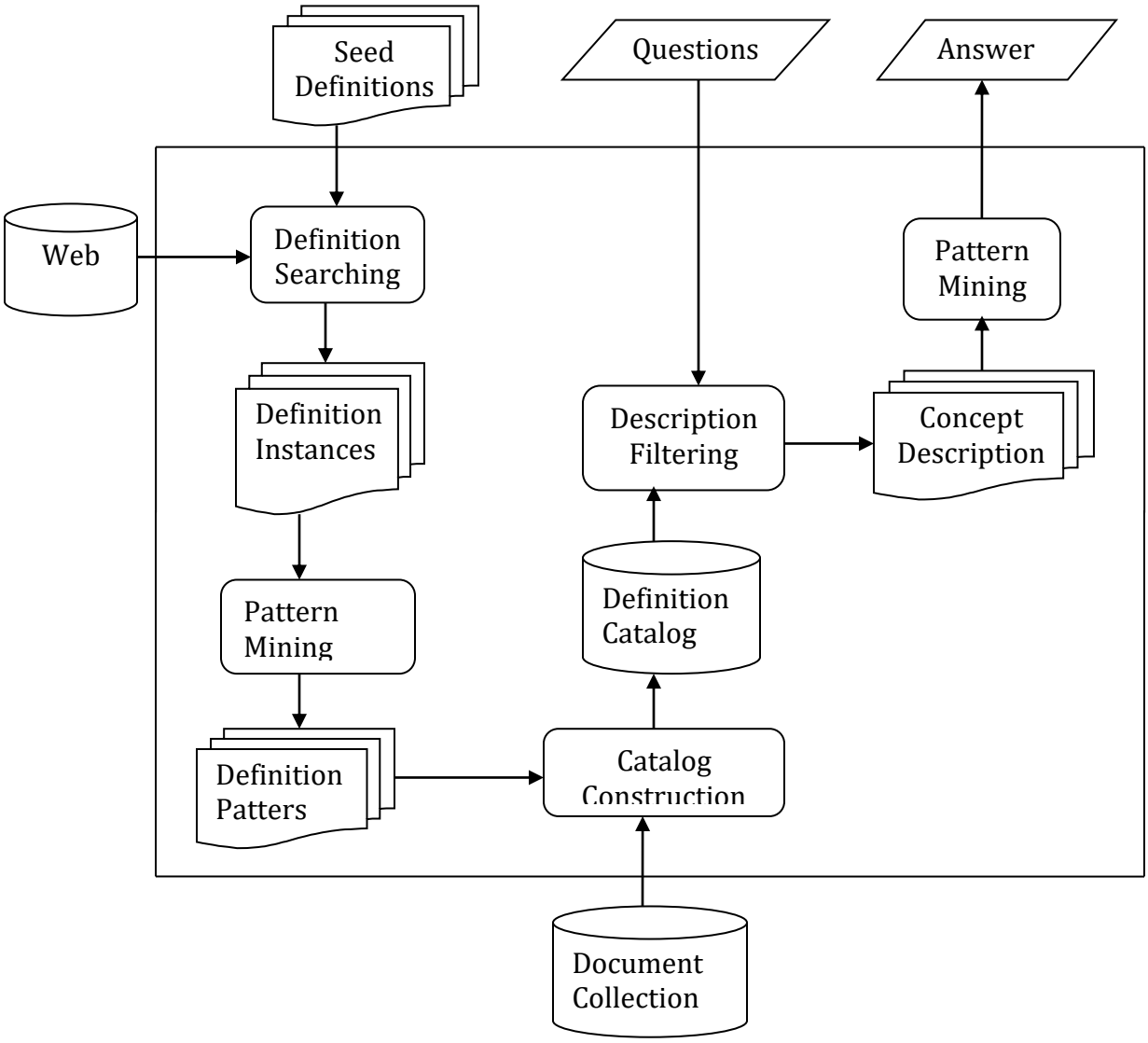

Fig 2.3.General Diagram of Question Answering Using Text Mining Approach
(*Adopted from "A Text Mining Approach for Definition Question Answering",* Denicia-Carral et al., 2005*)*

2.4.1 PATTERN DISCOVERY

The pattern discovery module applies a small set of concept-description pairs to collect from the web a comprehensive collection of definition instances. After the definition instances are collected, it applies a text mining technique on the collected definition

instances to determine a collection of definition surface patterns. Different QAS approach uses certain stylistic convections by means of a set of lexical patterns. Since in natural language concepts can be described in many ways it is difficult to come up with a complete set of linguistics patterns. And also these patterns described depending on the text domain, writing style and the language used (Denicia-Carral et al., 2005). According to the authors (Denicia-Carral et al., 2005) a general method which has two main sub tasks for pattern discovery was used.

2.4.1.1 DEFINITION SEARCHING

According to the author (Denicia-Carral et al., 2005) the definition searching process is initiated by a small set of empirically defined concept-description. These concept-description pairs help to search a number of sentences (phrases) that can be taken as a definition instances. To be useful in retrieving from the web the definition instance should contain the concept and its description in a single phrase.

2.4.1.2 PATTERN MINING

Pattern mining is the other process which is performed under the pattern discovery module. Data preparation, data mining and pattern filtering are the steps to be followed in pattern mining. Data normalization is performed at the data preparation phase. In this phase the input data is normalized by transforming all definition instances into identical format by using special tags for the concepts and their description. In the data mining step all maximal frequent sequences of words and punctuation marks from the set of definition instances are obtained by applying a sequence mining algorithm. The pattern-filtering phase helps to select the more discriminative patterns that satisfy a pre defined patterns.

2.4.2 ANSWER EXTRACTION

The answer extraction module creates a definition catalog consisting of a collocation of potential concept-description pairs by applying the discovered patterns at pattern discovery module over a target document collection. The main objective of this module is to find the more appropriate description for a question from a definition catalog based on text

mining approach. Since the construction of the definition catalog can be guided by definition patterns, it includes an enormous diversity of information, including information that is incomplete and incorrect. Even though the definition catalog may contain incorrect and incomplete, it is assumed that the incorrect information more scarce than the correct one. This assumption helps the idea of using a text mining method to differentiate between the adequate and improbable answers to a given questions. The following sub tasks are performed under the answer extraction module:

- **DEFINITION CATALOG CONSTRUCTION:** At this phase the definition patterns that specified at the pattern discovery module are applied on the collection of documents from which the answer to be extracted. A set of matched segments that most likely contain a concept and its description is the output of the definition catalog construction phase.
- **DESCRIPTION FILTERING:** In this phase for the question posted from the user, all descriptions corresponding to the requested concept is extracted from the definition catalog.
- **ANSWER MINING:** In this phase from a collection of extracted concept-description pair a single answer is selected for a given question. Under this phase there are three other phases such as data preparation, data mining and answer ranking. The focus of the data preparation phase is to normalize the descriptions related to the requested question. In the data mining phase, all maximal frequent word sequences from the collection of concept-description pairs is selected by applying a sequence mining algorithm. In this case each sequence shows a candidate answer to the requested concept. At the end in the answer ranking phase, each candidate answer is evaluated depending on the frequency of occurrences of its consequent. The logic behind is that a candidate answer constructed from frequent subsequences has more probability of being the correct answer than the one formed by the rare ones. Thus, the sequence with the greatest ranking score is selected as the correct answer.

2.5 RELATED RESEARCH 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. In this section we present some of the pertinent global as well as local research works in the area of question answering.

2.5.1 GLOBAL RESEARCH WORKS

Under this sub section different global researches that are related to QAS are presented. There was an attempt to develop QAS by different researchers for different language like English, Arabic, Chinese and Indian language.

2.5.1.1 QUESTION ANSWERING FOR ARABIC LANGUAGE

According to (Hammo et al., 2004) because of the fact that Arabic language is more inflectional and derivational language, there is sparseness of terms in text which leads to inadequacy in many statistical IR and NLP techniques. And also because of its inflectional and derivational variation it makes the process of question analysis and answer extraction difficult. When it is compared to other languages like English, Arabic language has complex morphology which highly complicates morphological analysis of the language, it does not have capital letters which adds difficulties to differentiate between named entities and other words, since alphabets are written from right to left and numerals are written from left to right editing Arabic text is more difficult when both are written on the same line and also Arabic language lacks corpora, lexicon and electronic dictionaries.

As stated by (Hammo et al., 2004) lack of more computerized tools and resources in Arabic language affects negatively on the growth of QAS in Arabic language. Some of the QAS developed in Arabic language include:

QARAB (A QUESTION ANSWERING SYSTEM TO SUPPORT ARABIC LANGUAGE)

The main purpose of the QARAB system was to recognize a text passage that answers a natural language question. The main steps that are performed in QARAB system for extracting the answer are processing of the input question, candidate document retrieval; processing each of the retrieved documents. The user question is also processed for extracting the answer. It is a stand –alone (non-web-based) QAS that uses traditional IR and NLP approaches to extract answers from a collection of Arabic newspaper texts. The authors didn't specify the size of the corpus used for the experiments (Hammo et al., 2004**)**. To extract the answer the author had adopted a keyword matching strategy in addition to matching simple structures extracted from both the question and the candidate documents selected by the IR system. In order to identify the proper names and other important lexical items the author has used the existing tagger. It provides answers to factoid questions but it does not support other types of questions like how or why questions. The authors of the QARAB system had reported a precision and recall of 97.3 %(Rosso et al., 2004).The evaluation was done by four Arabic native speakers who presented 113 questions to the system and have judged the correctness of the answer.

JAWEB

It is an Arabic web-based QA system that focuses on factoid questions types. It takes questions associated to any named entity including person, location, organization, time and the like. After the question is analyzed, important information are extracted from the Arabic corpus to return relevant answer. The system has user interface, question analyzer, passage retrieval and answer extractor as its component. The domain of the corpus was an open domain. The authors have used an Arabic corpus containing 39,660 words with of 457 kb as the information collection to retrieve an answer for the users' questions. The system was evaluated with recall, precision and response time. It had 100% recall, its average precision was 80% considering the limited size of the corpus and its average response time was 108.2 nanoseconds which depend on the size of the corpus and the performance of the server machine(Kurdi et al,2014).

AQUASYS: AN ARABIC QUESTION ANSWERING SYSTEM BASED ON EXTENSIVE QUESTION ANALYSIS AND ANSWER RELEVANCE SCORING

The AQUASYS was designed to answer a users' question formulated in Arabic language related to a named entity that could be of any types like person, organization, time, location, quantity and etc. It extensively utilizes NLP techniques to analyses questions and retrieves answers from an Arabic corpus that had been developed by the authors. Retrieved answers are scored and presented based on their relevance. The modules of the system are question analysis, sentence filtering, candidate answer finding and candidate answers scoring and ranking. The performance of the system was evaluated over a variety of question types posted by native Arabic speakers. The system had scored a recall of 97.5% and precision 66.25% score (Bekhti et al., 2011).

2.6.1.2 QUESTION ANSWERING FOR ENGLISH LANGUAGE

Many question answering system are developed with English language and publically available on the web. Some of them are the following:

START (SYNTACTIC ANALYSIS USING REVERSIBLE TRANSFORMATIONS)

START was the first web based QAS which was developed at the MIT (Massachusetts Institute of Technology) Artificial intelligence Lab. It is a natural language multimedia information access system in response to the user question and contains two modules which use the same grammar. The module named understanding module analyzes English text and provide a knowledge base which contains information found in the text. Provided an appropriate segment of the knowledge base, the generating module produces English sentences. The information stored in the knowledge base is retrieved by the query which is formulated in English and the system provides the response in English (Katz, 1997).

NSIR

NSIR (pronounced 'Answer') is a web-based question answering system under development at the University of Michigan. It uses the existing web search engines to research related documents on the web. As soon as the system gets list of returned documents from the search engine, it extracts the answers by extracting the top ranked

returned documents. The proximity algorithm and probabilistic phrase ranking techniques are used to rank possible answer before they are returned to the users (R. Radev et al., 2002). The proximity algorithm is based on the closeness in the text between the question words and the neighbors of each phrasal answers. A possible answer that is close to the question words gets higher score than the one which is far from it.

The probabilistic phrase ranking considers expected answer type to rank the answer. A probability score which indicates the extent to which the phrase matches the expected answer type with respect to the part of speech tag sequence is assigned to each phrase. The system returns the answer with its context to the user which in turn helps the user to justify the returned answer.

The authors evaluate the performance of the system by considering user effort in getting the correct the answer. They have developed the metrics named FHS(First Hit Success) which shows if the first answer returned by the system is the correct answer for the user question then the FHS is 1.when the first answer returned by the system is not the correct answer 0 is assigned to FHS. If the first answer to each question on a set of questions is considered and it is assumed as the web contains answer to all the questions, then the average of FHS represents the recall ration of a system (R. Radev et al., 2002).

The other metric is used to evaluate the system was FARR (First Answer Reciprocal Rank).This shows the effort of the user to get the correct answer from the returned answer. For example according to the authors if from the returned answer the third answer is the highest ranked correct answer the FARR is 1/3 and if no answer is returned by the system the FARR value is 0.And also a user can read a supporting documents to get the correct answer. In such cases, the order of the answer returned by the system affects the users' effort required to filter the answer.

FARWR (First Answer Reciprocal Word Rank) shows the number of words a user has to read before getting the correct answer. For example for question 'In which city is the office of Supreme court of Ethiopia is Located?', if the first answer is like 'capital city of Ethiopia Addis Ababa', the correct answer is starts from the fourth word, therefore the FARWR is ¼.FARWR metrics represents the users' time to reach at the correct answer. The other metrics they have used was PREC (precision).This indicates the percentage of useful content in the list of answer returned by the system. It is calculated as the total character

length of all correct answers divided by the total character length of all answers provided by the system. The authors have performed correlation analysis for some metrics, they have found that precision is a poor performance measure for web based QAS (R. Radev et al., 2002)

BASEBALL

According to (F. Green, 1961), it is a computer program which extracts for users' questions which is requested in ordinary English language. It has two parts that are the Linguistic and processor part. The linguistic part is responsible for reading the question from the punched card, analysis it synthetically and determines what information is given about the data being requested .The second part which is the processor retrieves through the data for the proper information, processes the results of the search and prints the answer.

EPHYRA

According to (Nico et al., 2006) Ephyra is a modular and extensible that enables to integrate different approaches to question answering system in a single system. Their framework can be adaptable to other language other than English language by replacing language specific components. The framework helps to combine several techniques for question analysis and answer extraction and to include different knowledge bases in order to accommodate user requirements. The Ephyra QAS uses pattern learning and matching, answer type analysis and redundancy elimination as its techniques for extracting an answer (Nico, 2005).The author combine different approaches for question analysis and answer extraction to more fulfill the requirement of TREC 2006.The system automatically learns text patterns that are applied for extraction and it is trained on question –answer pairs and utilizes convectional web search engines to extract text snippets that are appropriate for pattern extraction (Nico et al., 2006).

LCC (LANGUAGE COMPUTER CORPORATION)

It combines different techniques from information extraction with the huge amount of knowledge representation techniques which is derived from Word Net to justify answers

extracted from text corpus. For improving the system performance sophisticated tools like Syntactic pattern, logic form Transformer, Named Entity Recognition, Word sense disambiguation, Logic prover, Lexical Chenier and the like was applied. It was participated at TREC 2002, 2003 and 2004 and shows progressive performance for factoid, definition and list type questions (Harabaigu et al., 2004).

ASKMSR

According (Banko et al., 2002) the system was developed by taking advantage of the large volume of online text available through the worldwide web rather depending on sophisticated linguistic analysis like part-of-speech tagging, parsing, named entity extraction, semantic relations, dictionaries and Word Net of either questions or candidate answers. As stated in (Brill et al., 2002) the architecture of the system includes Query Reformulation, N-Gram Mining, N-Gram Filtering and N-Gram Tiling. At query reformulation stage given a question, the system produces a number of weighted rewrite strings which are likely substrings of declarative answers to the question.

For instance 'When was Mandela selected as a president?' rephrased as 'Mandela was selected as a president'. These patterns are applied to a collection of documents that contain such patterns in the process of extracting an answer for a question. After the query reformulation has been generated, the rephrased strings are taken as a query and sent to a search engine from which different page summaries are collected and analyzed. After that n-grams are collected as a possible answer from the returned collection of page summaries. After the n-grams are collected they are filtered and reweighted based on how much they are matched with the expected answer type. Finally, an answer tiling algorithm is applied to select an answer from Candidate answer that have a higher score in order to take as a final answer for a question.

2.5.1.3 QUESTION ANSWERING FOR THE BENGALI LANGUAGE

Even if the Bengali language is the most spoken language many attempts was not done to develop question answering system. According to the author presence of many interrogatives, interrogative position in the text and language processing tools scarcity makes development of QAS for the English language (Banerjee et al., 2014).The Bengali

QAS named BFQA has three modules namely question analysis, sentence extraction and answer extraction.

At question analysis stage natural language question in Bengali language is accepted and activities like Question Type extraction, expected answer type identification, Named Entity Identification, named entity recognition and question topical target extraction are done. At the sentence extraction stage relevant sentences that may contain an expected answer are extracted and after the retrieved sentences are ranked depending on answer score, the answer to the natural language question is specified by the Named Entity identification which is recommended by expected answer type identification module.

The author proposed a sentence ranking strategy for the system. According to the author (Banerjee et al., 2014) the accuracy of the system was not equivalent with those of factoid QAS developed with European language. And the basic reason for the low performance of the system was because of the low accuracies of the shallow parser and the Name Entity recognizer system as the performance of the name entity recognizer and parser component affects the performance of the system.

2.5.2 LOCAL RESEARCH WORK

2.5.2.1 AMHARIC QUESTION ANSWERING FOR FACTOID QUESTIONS

An attempt to develop QAS in local language is also done by local researchers. Seid (2009) tried to develop the first QAS in Amharic language. He attempted to identify the main language specific features like character normalization, number normalization and document normalization for Amharic language. As stated by the researcher since characters can be written in different symbols and used interchangeably in the language they would normalized to a specific character to have a better recall. Because of sentence delimiters can be formed in different forms in the Amharic document in order to identify the end of sentences and paragraphs, sentence and paragraph demarcation was integrated.

The documents for the thesis work were Amharic news corpuses collected from different electronic news papers.

He has performed document pre processing like stemming, normalization and stop word removal which makes the document ready for the next module. His question analysis module used to clearly identify what the users' question is, identify question focus and the expected answer type. The document retrieval component retrieves the documents that are expected to contain the expected answer type identified at the question analysis phase. The answer extraction module which is the last module will extracts the candidate answer from the retrieved documents by the document retrieval module. Sentence indexing was performed on the documents by using the Amharic full stop. And also paragraphs were indexed separately based on the special paragraph markers ($$).

The Author has developed and used Gazetteer and pattern based answer selection algorithm for improving the activity of selecting the proper answer for natural language question. According to the author as it is the initial attempt for Amharic language QAS it has a good performance. The document retrieval component shows a more coverage of searching a relevant document 97%, the sentence based retrieval component scored 93%, the gazetteer based answer selection techniques answer 72% of the question correctly and the file based answer selection technique shows better recall 91% which indicate that most relevant documents are returned which are considered to be as correct answer are returned. The pattern based answer selection technique scored a better accuracy for person names using paragraph based answer selection techniques whereas the sentence based answer selection techniques performs more for numeric and date question type.

2.5.2.2 DESIGNING AMHARIC DEFINITIVE QUESTION ANSWERING (DEFAMHARICQA)

The first attempt to design Amharic definition question answering was performed by (Wondowossen, 2013) .It was aimed to solve the problem of extracting answers for a natural language question from a collection of documents. The designed system was classified into two components which are the indexing and definition searching components. The indexing components have two subcomponents that are preprocessing and the extraction of possible definition from the legal document corpus by using patterns.

And the definition searching module divided into question analysis and definition extraction subcomponents.

The question analysis module extracts the definiendum from the users' natural language question. Lexical patterns are manually constructed and used to extract candidate definition from the document. The researcher had performed preprocessing on legal document processing such as paragraph demarcation, character normalization, stemming and synonym indexing. He has developed an algorithm used to extract definiendum from user's question. To return answer for the user question the researcher opened the dictionary catalog and compares the definiendum with each of the concept-description pairs. If the definiendum existed in the dictionary catalog it displays the description of the concept as answer to the question otherwise it displays an error.

The researcher had used some known Amharic definition keywords that always appear after the definiendum to identify the key word to be defined. He had used such kind of Amharic key words to identify concept-description pairs and the keyword before the definition keyword are taken as definiendum and the rest sentence after the definition keyword is considered as the definition. In the query generation sub module he has included features like stemming, character normalization in the question words and synonym expansion to improve the performance of the system.

The researcher had used the recall and precision to compute the effectiveness of the system for performance measurement. He had also evaluated the query processing module to find out to what extent it identifies the definiendum from the users' natural language question. He had used legal documents containing 20 Ethiopian proclamations, five law directives and four training manuals a total of 125 pages for his experiment. According to the author the system returns correct answer for 17 questions from a total of 20 questions which score 85%. And the system score nugget recall of 73%, nugget precision of 85.6% and F-measure of 78.8%.

2.5.2.3 AMHARIC QA FOR LIST QUESTIONS

Another Amharic QAS was done by (Brook, 2013) which was focused on list questions in closed domain of Amharic question answering system. The aim of the research was to

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 had identified semantic entity classes such as hotel, museums, birds, celebrations, languages cultures, nationalities, lakes, tour travels, monuments, tourist sites and wild animals and associates each of the question with one of the entity class to develop a lexical patterns for each of the entity class and to identify the question terms that delimits question to a list question type.

The IR modules identifies the relevant documents and based on the question focus that retrieves for a specific entity class like birds, hotels, lakes and etc the answer extraction module is applied. The terms that match to the answer type specified by the answer type recognition module are extracted from the source document collections.

The researcher had computed the similarity between the lists of initial potential answers to analysis the frequency of the co- occurrence of answer instances in the document which enables the candidate answer selection module to select terms that co-occur more frequently. The researcher had used a classification method to classify the candidate answer extracted into relevant and irrelevant in the candidate answer selection module and then the candidate answer are returned in the class of relevant. The candidate terms that co-occur more frequently with other candidate answers, therefore the answer is the one which has a more similarity value. Thus, the researcher had used the sum of the similarities each term with the other terms as a pointer of how frequently each term co-occurs with the other terms.

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.

2.5.2.4 FACTOID QUESTION ANSWERING FOR AFAN OROMO

An attempt to design factoid question answering for Afan Oromo was done by (Kasahun, 2014).The objective of his study was to extract fact based answer to user from Afan Oromo electronic documents collected from Oromia Radio and Television Agency, Fana broadcasting Afan 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 phrases.

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.

CHAPTER THREE

AFAN OROMO LANGUAGE

In this chapter the basic structure of Afan 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 news papers, 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 OVERVIEW OF AFAN OROMO

Afan Oromo is one of the major African languages which is extensively spoken and used in most parts of Ethiopia and also some parts of other neighbor countries like Kenya, Somalia and Egypt. Afan 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).The native language speakers are mostly found in Shoa, Illubabour, Jimma, Wollega,Arsi, Bale, Hararghe, Wollo, Borana, and the southwestern part of Gojjam. The language speakers in Ethiopia is more than 30 million and also the language is spoken other than Ethiopia such as Kenya, Somalia and Egypt(Gezehegn,2012).The Oromo people are the largest ethnic group in Ethiopia and about 40% of the populations are Oromos.

The language is used as an official language of the Oromia regional state which is one of the largest states in Ethiopia. Qube(a Latin-based alphabets) has been adopted as a writing system and taken as the official script of Afan Oromo since 1991(Kula & Varma, 2007).Currently Afan Oromo language is a language of research, administration, political and social interaction. The language is also academic language in universities in Ethiopia like Jimma University, Addis Ababa University, Ambo University, etc and Oromia regional sate in primary schools and also as a single subject in high school. Currently news papers,

news, online education, magazines, journals, books, videos, pictures and entertainment Medias are increasingly published in this language.

The Afan Oromo language uses 26 Latin symbols from A to Z which are both lower and upper case letters. There are two categories of symbols like English language namely vowel and consonant. From 26 symbols used in the language a, e, I, o and u (A, E, I, O, U) are called vowel letters and the remaining are consonant letters. There are also other symbols that are specific to Afan Oromo language. These symbols are called 'Qubee dachaa' which include CH, DH, NY, SH and PH and categorized in consonant symbols. Including the symbols called 'Qubee dachaa' the language has 31 letters (26 consonants and 5 vowels).Even if the writing styles of English and Afan Oromo languages are different, all the letters of English language also found in Afan Oromo language.

The same rule is followed to form words in Afan Oromo language with that of English which is the combination of (consonant + vowel).Although Afan Oromo language follows similar rule in word formation with that of English language there are many language specific in Afan Oromo language. Some of the Afan Oromo language specific features are having one or two vowels in between consonants convey different meanings which are called as 'Jecha dheeraa' and 'jecha gabaabaa' depending on the number of vowel letters used. In the language if there are more than two vowels next to each other a glottal stop consonant or which is called 'udhaa' is used to make the word to be meaningful otherwise it is not allowed to have more than two vowels next to each other. And also in the word formation it is not allowed to have more than two similar consonant letters next to each other.

For instance: L + a + f + a= Lafa to mean 'Ground or Earth'

L + aa + f + aa=Laafaa to mean 'Soft'

B + u + n + a =Buna to mean 'Coffee'

B + uu +n +aa= Buunaa to mean 'To go'

In Afan Oromo language a word which has two similar consonant letters has different meaning with that of one consonant letter. A word which has two similar consonant letters is named as 'Jecha jabaa' and the one which has a single consonant is called 'Jecha laafaa'.

For instance: s + a + m + uu=samuu to mean 'something with bad odor or smell'

S + a + mm + uu=sammuu to mean 'brain'

Q + o + r +e =Qore to mean 'examine'

Q + o + rr + e=Qorre to mean 'Cold'

Forming a word by putting an apostrophe between vowels which called in the language 'hudha' which mean glottal is also another specific feature to Afan Oromo word formation.

For instance:

S.No	Afan Oromo words	Their meaning in English
1	Sa'a	Cow
2	Baay'ee	Many
3	Re'ee	Goat
4	Mul'dhata	Vision
5	Dhangala'aa	Liquid
6	Boba'aa	Fuel
7	Boba'a	It is burning

Table 3.1some glottal words in Afan Oromo

3.2 WORD MORPHOLOGY IN AFAN OROMO

Morphology as a sub-discipline of linguistics was named for the first time in 1859 by the German linguist August Schleicher who used the term for the study of the form of words. The word morphology has different meaning in different field of study. For example in biology the word morphology indicates the study of the forms and structure of organisms, in geology it refers to the study of the configuration and evolution of land forms and in linguistics the term morphology indicates that the scientific study of words, their internal structure and how they are formed in a language. The term morphology was named as a

sub-discipline of linguistics for the first time in 1859 by the German linguist August Schleicher who was applied the term for the study of the form of words (Tom Ritchey, 2013). In this paper the linguistic context of the term morphology is considered.

As it is stated in different literature morphology in linguistic is classified in to inflectional and derivational morphology (Abebe , 2010).Inflectional morphology describes the word variants that can be formed from the same stem word. It is the process by which affixes are combined with the root word to indicate basic grammatical classification like tense or plurality. Inflectional morphology also viewed as the process of adding some meaning to the existing words not as the creation of new words. In inflectional morphology different word forms are formed from one root word to indicate person, numbers, and gender, tense or case.

For instance in the words 'car-s' and 'finish-ed' the characters 's' and 'ed' are inflectional suffixes that are used to form the words 'cars' and 'finished' . Even though the words 'cars' and 'finished' indicate different number and tense respectively , they have the same word class with their root word. In Afan Oromo also different affixes are added to the stem word to form different words. For example the affix 'oota' is added to the stem word 'seer' to from the plural word 'seeroota' which is to mean laws. Derivational morphology is the process of creating a new word from a root word. Derivational morphology changes the lexical meaning of the stem word from the other derived words by changing the class of the word. For example the derivational suffix 'ly' changes adjectives to adverb like slow to slowly.

Afan Oromo has a very rich morphology like others African and Ethiopian language (kula et al., 2007).It has the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative language like Afan Oromo, Amharic and Zulu most of the grammatical information is conveyed through affixes (prefixes, infixes and suffices) attached to the roots or stems. In comparison to English plural markers (-s or -es) there are more than 12 major and very common markers in Afan Oromo nouns. For instance (oota, ooli, -wwan, -lee, -an, een, -oo, etc). In Afan Oromo language nouns are inflected to show numbers and verbs are inflected for gender, number, tense, voice, aspect

and mood (kula et al.,2007).In Afan Oromo language every word contains one or more morphemes. These morphemes are categorized as free and bound morphemes. Free morphemes can stand as a word on its own and they can constitute words by themselves but a bound morpheme does not occur as a word on its own. They are parts of words and occur combined to free morphemes. For example in English word 'cats' the word 'cat' is free morpheme because it can stand alone and the character '-S' is a bound morpheme because it does not stand alone like that of the unbound (free) morpheme.

In Afan Oromo root words are considered as bound morpheme because of the reason that they cannot stand alone and it is required to add affixes to them in order to make them free morpheme and more meaningful. For example to make the words 'Nyaat**-'**(eat) and 'deem**-**'(walk) pronounceable the appropriate affixes should be added to them like 'Isheen deemte' (to mean She walked) or 'Isaan deemuuf'(to mean They are to walk)(Debela & Ermias,2010).An affix is also cannot stand alone unlike that of free morphemes and are considered as bound morpheme. Affixes are attached to the root word and they are classified as prefix, suffix and infix. Prefixes are found at the beginning of the root word where as suffixes are located at the end of the root word. And infixes occur in between characters of the word. For examplein the word nyaadhuu'eat',-dhuu is a suffix and nyaa- is a stem word.

In Afan Oromo language suffixes can be categorized into three basic types. They are derivational, inflectional and attached suffixes. Inflectional suffixes are the mostly used in Afan Oromo language and some of them are –n,-lee,-een,-ichaa,-tu,-oo,-ootaand–wwan. For instance in the word 'baratoota' (students) 'barat' is the stem term and 'oota' is the suffix. Derivational suffixes in Afan Oromo are used to create a new word by attaching them to the root word.

These derivational suffixes are '–achuu','-lee','-eenyaa','-ina','-ummaa' and etc. For instance in the word 'nammummaa' (humanity) 'namm' is the root word and 'ummaa' is the derivational suffixes (Abebe,2010).In Afan Oromo language affixes such as 'arra','bira','irra','tti' and 'dha' are considered as attached suffixes and most of the time they are located next to the word they are attached with them(Abebe,2010).

In Afan Oromo language words can be formed in different ways. Morphological analysis of this language can be classified into nouns, pronouns and determinants, case and relational concepts, function words, verbs and adverbs. Almost all Afan Oromo nouns in a given text have person, number, gender and possession markers which are concatenated and affixed to a stem or singular noun form. Similar to Afan Oromo noun determinants have number, gender, adjectives and quantifier markers.

Afan Oromo verbs are also highly inflected for gender, person, number, tenses, voice and transitivity. In Afan oromo adverbs can be categorized as adverb of time, adverb of place and adverb of manner in which some of them are affixed (Debela& Ermias,2010).In Afan Oromo prepositions, postpositions and article markers are indicated through affixes and conjunctions in Afan oromo can be used to separate words and some of them are affixed. For example 'Gammachuu fi Jiituun mana barumsaa deemaa jiru'(to mean Gemechu and Jitu are going to school),in this sentence the word 'fi' is used as a conjunction. As it is stated above the morphological analysis of Afan Oromo language such as nouns, pronouns and determinants, case and relational concepts, function words, verbs and adverbs are discussed below:

3.2.1 NOUN MORPHOLOGY

Nouns are words that enable to name places, people, things or ideas (Gezehagn, 2012).

3.2.1.1 NOUN INFLECTION

In Afan Oromo different grammatical functions like number, gender, definiteness and case are achieved by inflecting nouns. A word of identical class as the original stem can be formed by adding suffixes to stem (root) word.

PLURALIZATION: There are different ways of expressing plural numbers in Afan Oromo language. This is mostly performed by adding suffixes at the end of nouns. In this language different suffixes are added to different nouns to form plural. As it was stated in (Abebe, 2010) in Afan Oromo nouns there are more than 12 basic and common plural markers unlike that of English language which is s or (es). Some of these plural markers in Afan Oromo nouns are shown below:

S.No	Suffixes	Afan Oromo Nouns		Equivalent word in English
		Singular	Plural	
1	-oota	Nama	Namoota	Men
		Barataa	Barattoota	Students
		Hojataa	Hojjatoota	Workers
2	-lee	Lammii	Lammilee	Relatives
		Kitaaba	Kitaabolee	Books
3	-(w)wan	Gaara	gaarrewwan	Mountains
		Hojii	Hojiiwwan	Jobs
		Sa'a	Saawwan	Cows
		Tuluu	Tulluwwan	Mountains
4	-olii	Jaarsa	Jaarsolii	Elders
		Gaangee	Gaangolii	Mules
5	-een	Lagaa	Laggeen	Rivers
		Farda	Fardeen	Hourses
6	-aan	Ilma	Ilmaan	Sons

3.2 Some pluralized nouns with common suffixes

GENDER: In Afan 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 Afan Oromo the term 'isa' and 'ishee' shows masculine and feminine respectively like that of English language third person singular pronouns he and she.

DEFINITENESS: In Afan Oromo there are no indefinite articles corresponding to that of the English a and an. The definite article corresponding to the English the can be indicated by adding suffixes such as 'icha' and 'ittii' to nouns after dropping vowels if the nouns are vowel ended in Afan Oromo .The suffixes 'icha' and 'ittii' indicates masculine and feminine nouns respectively. In Afan Oromo the usage of the definite suffixes is less than that of the definite article the in English. For example 'namicha'(the man).For animate nouns that can take either gender like 'qaalluu'(to mean priest), 'qallicha'(the priest shows masculine) and 'qallittii'(the priest shows feminine) the definite suffix may indicate the intended gender(Abebe,2010).

CASES: In Afan oromo noun has a base form that is used when the noun is the object of a verb, preposition or postposition. For example 'konkolaataa binne'(We bought a car)- in this sentence the noun 'konkolaataa'(which is to mean car in English) is used as an object of the verb 'binne'. Case is a grammatical category of nouns that shows the nature of their relationship to the verb in a sentence. It can be indicated by adding suffixes or lengthening of the noun's final vowel. In this language nouns are inflected for nominative, ablative, instrumental and locative cases (Abebe, 2010).

I. NOMINATIVE CASE

In Afan Oromo nominative is used for nouns that are the subject of clauses. For example: 'Beekaan mana bite'(to mean Beka has bought a house).In this example 'Beekaa' shows the name of the man and 'Beekaan' is used as nominative and subject of the phrase 'mana bite'. To form nominative most nouns ending in short vowels with a proceeding single consonant drop a final vowel and add 'ni'. For example: 'Baratoota' (students), 'baratoonni' (nominative).In this example the word 'Bratoota' shows the plural noun and the word 'baratoonni' indicate nominative. There are rules that should be followed to suffix the base noun to change to nominative. Some of them are the following (Abebe, 2010):

- (-i) is suffixed when a final short vowel is preceded by two consonants or a geminated consonant. For instance: 'Namicha' (the man) it becomes 'Namichi

dubataa jira,(the man is speaking (nominative)). In this example the letter a is dropped and i is added and make the noun as the subject of the clause.)

- If the noun ends in a long vowel, 'n' is suffixed to the noun to form nominative. For example 'Maqaa' (name),'maqaan'(name(nominative)).In the sentence 'Maqaan kee eenyu?' the word 'maqaan' indicative nominative.
- If the noun ends in n, the nominative is identical to the base form. For example: 'Afan'(mouth, language (base form or nom.))
- Some feminine nouns ending in a short vowel add 'ti' to indicate nominative. For example:

  - mana 'house',manni.
  - Haadhaa 'mother',haati.

In the above two examples the words 'manni' and 'haati' are indicative nominative.

II. INSTRUMENTAL CASE

The instrument in language is the nouns that are used to indicate the means by which something is occurred, the agent which makes something to be occurred, the reason why something is happened or the time of an event when it is occurred.

To represent instrumental cases in Afan Oromo suffixes like *–n,-aan,-tiin,-iin,-dhaan,-aan* are attached to the nouns based on the number of vowels and consonants the nouns are ending with. Some of the rules that enable to represent instrumental case are:

- -n is added to nouns end with short vowel or long vowel and 'iin' is added to nouns end with consonants.

  For example:

  - Ija 'eye' ijaan 'by eye'.
  - Halkan 'night',halkaniin 'at night'
  - 'Halkaniin dhufe' (to mean he came in the night)
- The suffix 'tiin' is added to nouns end with long vowel or short vowel.

  For example:

  - 'Afan Engiliffaa' (English language),'Afan Engiliffaatiin'(in English)

- ‘Afan Engiliffaatiin dubata’(to mean He is speaking in English)

➢ The suffix ‘dhaan’ is added to the nouns end with long vowel.

For example:

- Yeroo ‘time’,yeroodhaan,’on time’
- ‘Inni yeroondhaan xumure’(to mean He has finished on time)

III. LOCATIVE CASE

In Afan Oromo the locative is used for nouns that represent general locations of events or states. To indicate more specific locations, prepositions or postposition (like booda,hanga,gara,etc) are used in this language. The suffix ‘tti’ is added to the noun to form locative.

For example:

- Finfinneetti ‘in Addis Ababa’.
- ‘Manni murttii walii gala federaalaa Itiihoophiaa finfinneetti argama’(to mean Federal supreme court of Ethiopia is found in Addis Ababa).

IV. ABLATIVE

The ablative is used to indicate the source of an event. In Afan Oromo the ablative is equivalent to that of English from. Some of the rules that are followed to form ablative are:

➢ The vowel is lengthened when the words end in a short vowel.

For example:

Biyya ‘country’,biyyaa ‘from country’

➢ The suffix ‘dhaa’ is added when the words end in a long vowel.

- Mana barumsa ‘school’,mana barumsaadhaa
- ‘Mana barumsaadhaa dhufaa jira’(to mean he is coming from school)

➢ The suffix ‘ii’ is added to the word when it ends in a consonant letter.

- For example: ‘Hararii’(to mean from Harar).

The post position ‘irraa’ (to mean from in English) can be used as an alternative to the ablative by dropping its initial vowel letter.

For example: Biyya ‘country’,biyyarraa(to mean from country).

V. GENITIVE CASE

Genitive is used to indicate Possession or belonging. It is equivalent to the English possession markers(‘s). In Afan Oromo language the genitive is formed by lengthening a final short vowel, by adding the suffix ‘ii’ to the final consonant when the word ends in a consonant and by leaving a final long vowel as it is. In Afan Oromo genitive phrase the possessor noun follows the possessed noun.

For example:

- Konkolaataa’car’,Beekaa(name of a person),’konkolaataa beekaa’(Beka’s car)
- Obboleetti ‘sister’,Beekaa(name of a person),’obboleettii Beekaa’(Beka’s sister).

PRONOUNS: The most common types of pronoun in Afan Oromo are:

- **DEMONSTRATIVE PRONOUN:** This pronoun type is similar to that of English like ‘sana’(that), whose(kaneenyuu), kun/kana(this),who(eenyu) and kam(which).

For example:

- ‘Konkolaatan kun kan keenya’(to mean this car is ours).
- ‘Konkolaata kana biti’(to mean buy this car).
- ‘Barsiisas ana waami’(to mean call that teacher).
- ‘Barumsa kamtu huulfaata?’(to mean which field of study is difficult?)
- ‘Mana keessaa eenyutu dubata?’(to mean who is speaking in the house?)

- **PERSONAL PRONOUN:** some the examples of personal pronouns in Afan Oromo are:

- You are walking slowly. ‘Ati suuta jeteed eemta.’
- He has a car. ‘Inni konkolaataa qaba.’
- They are eating lunch. ‘Isaan laaqana nyaachaa jiru.’

In the above example the words ‘ati,inni and isaan’ are personal pronouns in Afan Oromo that are corresponding to the English pronoun you ,he and they respectively.

- **POSSESSIVE PRONOUNS:** Some examples of possessive pronoun in Afan Oromo are:
    - My ->'ko/kiyya/tiyya.'
    - This is my car. 'Konkolaatan kun kan kiyya'(in this the word 'kiyya' indicate possession which is equivalent to the pronoun my in English.)
    - Our ->'keenya/teenya.'
    - Our school is profitable. 'Manni barumsa keenya bu'aqabeessa.' (The word 'keenya' indicate possession which equivalent to the pronoun our in English in the sentence.)

DERIVED NOUNS :A noun can be derived from another noun by attaching derivational suffixes like '-ummaa,-annoo, -eenya, -ina , -ii, -ee, -a, -iinsa, -aa,-i(tii), -umsa, -oota, -aata, and–ooma' in Afan Oromo. The derived noun may be different from the root word by their grammatical category (Abebe, 2010). Some of the examples are shown in the table below:

S.No	Noun	Derived Noun
1	Hiryyaa	Hiryyummaa
2	Lammii	Lammummaa
3	Saba	Sabummaa
4	Gaarii	Gaarummaa
5	Hollaa	Hollummaa
6	Biyya	Biyyummaa

Table 3.3 List of nouns derived from another noun

In Afan Oromo derivation of nouns from verbs are done by attaching the appropriate suffixes like '-baa,-aa, -eenya, -taa , -ii, -ee, -a, -iinsa, -aa,-i(tii), -umsa, -oota, -aata, and–ooma' to verbs.

Some of the examples are shown below:

S.No	Verb	Nouns
1	Kijibe	Kijibaa,Kijibduu
2	Ariifate	Ariiftuu,Ariifataa
3	Deeme	Deemaa,Deemtuu
4	Dadhabe	Dadhabaa,Dadhabduu
5	Barate	Barataa,Baratuu

Table 3.4 List of nouns derived from verbs

In Afan Oromo language there is also compound nouns like other language. These compound nouns are formed from two independent nouns by combining them and they have a single meaning even if they are two separate words.

For example:

- 'Abbaa seeraa'(to mean lawyer)
- 'Abbaa manaa'(to mean husband)
- 'Haadha mana'(to mean wife)
- 'Abbaa gadaa'(to mean traditional Gada System Oromo president)

3.2.2 VERB

In Afan Oromo a verb consists of simply of a stem which represent the lexical meaning of the verb and suffix which represents tense or aspect and subject agreement. For example in the word 'fixne'(to mean we finished), the stem is 'fix' and 'ne' indicates the tense which is past and even if the subject of the verb 'fixne' is not explicitly specified it is first person plural(nuyi,nuti(to mean 'we' which is the subject)).

Most dictionaries in Afan Oromo lists verbs in their infinitive form and all infinitives are end in 'uu'. For example 'deemuu' to mean to go. The verb stem is formed from the infinitives by dropping the suffix 'uu'. For example the stem of the word 'deemuu' is 'deem' which is formed by dropping the suffix 'uu'. Different words are formed from the stem word deem by attaching different affixes to it. For example affixes like 'aan' and 'uuf' are

attached to the stem word ‘deem’ to form the words ‘deemaan’ and ‘deemuuf’ respectively. Afan Oromo verbs are classified into four groups depending on their stem endings.

- **REGULAR VERBS:** These are verbs that are formed by attaching suffix to the stem word without changing the stem word. They are verbs with stem that do not end in a double consonant, ch, a vowel,y, w or Z. For example: From the stem word ‘deem’ other verbs like ‘deema ,deemna, deemta, deemna, deemtu, deemu’ etc are formed without changing the stem word deem by simply attaching the affixes like ‘–a,-na,-ta,-na,-tu,-u’ etc.
- **DOUBLE-CONSONANT ENDING STEMS:** when the stem word ends with double consonant, a slight modification is performed on the regular suffix to form other verbs. For ‘nuti, ati, isin and isheen’ form the character ‘I’ is added to the regular suffix during formulation of other verbs. For example: ‘nuti argina’(to mean We see),’ati argita’(to mean You see),’isheen argiti’(to mean she will see).

  In these example like the regular verbs shown above other verbs are not formed by simply adding regular suffix such as ‘n,tu,na,ta’ etc.Because in Afan Oromo language occurrence of three consonant letter is not allowed in a row. To formulate other verbs from stem word which ends in double consonant some modification is done on the regular affix and attached to the stem word. From the above example the affixes ‘ina,ita, iti’are modified from the regular affix ‘na,ta and ti’ respectively. By attaching the modified affixes ‘ina,ita and iti’ to the stem word ‘arg’ the verbs ‘argina, argita and argiti’ are formulated.
- **VERBS INFINITIVES ENDS WITH -CHUU**: For verbs that are ends with the affix ‘chuu’ the ‘ch’ changes to ‘dh’ in the ‘ani’ form and changes to t for other forms. For example from the word ‘nyaachuu’, other forms like ‘nyaadhaa, nyaatta, nyaata, nyaatti, nyaanna, nyaattu, nyaatu’ etc can be formulated by attaching the suffixes ‘ta, tti, nna, ttu, tu’ at the end of each of word.
- **IRREGULAR VERBS (VOWEL –ENDING STEMS):** Infinitive verbs that end with ‘-a'uu , -o'uu, -u'uu, -e'uu, and -i'uu’ considered as regular verbs. For example from the word ‘du’uu’(which is to mean to die) other forms like ‘du’a,duuta,duuna,duutu,du’u,duuti’ etc are formed

3.2.3 PREPOSITION

A preposition is used to link a noun to an action or to another noun. In Afan Oromo preposition can be pre preposition which comes before the noun (like 'hamma'(to mean up to, as much as),'gara'(to mean towards),'hanga or haga'(to mean until),etc. And post preposition Comes after the noun like 'bira'(to mean beside, with, around),'booda'(to mean after),'keessa'(to mean inside, in),etc(Gezahegn,2012).

For example:

- 'Gammachuun kutaa keessa jira'(to mean Gemechu is in the class).In this sentence the word 'keessa' indicate post position and it came after the noun 'kuta'.
- 'Gammachuun gara Awaasaa deemaa jira'. (To mean Gemechu is going to Awassa.).In this sentence the word 'gara' indicates preposition and it came before the noun Awaasaa.

3.2.4 ADJECTIVES

Like other language in Afan Oromo adjectives are words that stand for describing the characteristics of the noun. For example:

- 'Qormaatni kun salphaa dha'. (To mean this exam is easy).In this sentence the word 'salphaa' is an adjective which describes the noun 'Qormaatni'.
- 'Kitaabni kun mi'aa dha'. (To mean this book is expensive).The word 'mi'aa' describes the noun 'kitaabni' and stands as an adjective.

3.2.5 ADVERB

Adverbs are used to describe or modify the manner how activities are done or something is happened. Adverbs can describe the manner of an action. For example: 'Inni suuta jedhee nyaata'(to mean he eats slowly.),shows the time at which the action is occurred(For example: 'Isheen boru dhufti'(to mean she will come tomorrow.),indicate location(For example: 'Bakki hojii koo mana koo irraa fagoo dha'.(to mean My work place is far from my

home) and indicate degree of something.(For instance:'Biyya koo baay'een jaala dha'.(to mean I like my country a lot)))(Gezehagn,2012).

3.3 INTERROGATIVE SENTENCE IN AFAN OROMO

Questions in Afan Oromo are constructed with words like 'eenyu,eessatti,yoom,bakkakamitti,hangam, ibsi,maal jechuu dha,maal,maali, tarressi', and etc to request for information. For example:

- 'Maqaan kee eenyu?'(What is your name?)
- 'Eessa jiraata?'(where are you living?)
- 'Ibsi seeraa maal?'(what is the definition of law?)
- 'Seeraa jechuun maal jechuu dha?'(What does law mean?)
- 'Mandeelan eessatti dhalate?'(Where was Mandela born?)
- 'Magaalota Itiyoopi'aa keessa jiran tareessi'.(List the cities in Ethiopia.)

The above example indicates how different question types are formulated in Afan Oromo. For example: The questions 'Abbaa seeraa jechuun maal?','Maandeelan eessati dhalate?' and 'Magalota Itiyoopi'aa keessatti argaman tarreesi' indicates definition ,factoid and list question type respectively. The keywords like 'jechuun,eessatti,tarreessi' in the above example also shows the question type definition, factoid and list respectively. More over keywords like 'ibsi,maal jechuudha,maal,hiika(hiikni)' and the like are used in forming definition question in Afan oromo. For example:

- 'Jecha Abba Seeraajedhu ibsi'. (Describe the word lawyer.)
- 'Abbaa seeraa jechuun maal?'(What is lawyer means?)

CHAPTER FOUR

DESIGN OF AFAN OROMO DEFINITION QUESTION ANSWERING SYSTEM

Surface Pattern based approach is used in the study to identify the required concept from the document collection. The patterns that are applied to extract the concept with its description are manually constructed for the purpose of this study. The manually constructed patterns were also used to extract the definiendum from the users' natural language questions.

The details of how the patterns are constructed and used in the process of answer extraction are discussed under this chapter. This chapter mainly deals with the design of the proposed system and an overview about the main components of Afan Oromo Definition Question Answering system. The components of the system are identified, the activities that are accomplished by each component are described; the interactions among the components are discussed and the outputs that are generated by each of them are specified.

4.1 ARCHITECTURE OF AFAN OROMO DEFINITION QUESTION ANSWERING SYSTEM

The architecture of the proposed system mainly includes question analysis, document processing, answer extraction and answer ranking components. The processes under the online components are performed when the questions are issued to the system. The processes under the offline components are performed before the user questions are submitted to the system. The sub components under document processing are sentence tokenization, stemming of definiendum in the corpus and definition catalog construction. A natural language question from the user is taken as an input to the system. The question analysis and description filtering (matching) are considered as a process in the architecture. And finally the extracted answer is considered as the output from the system. Figure 4.1 below shows the architecture of the system:

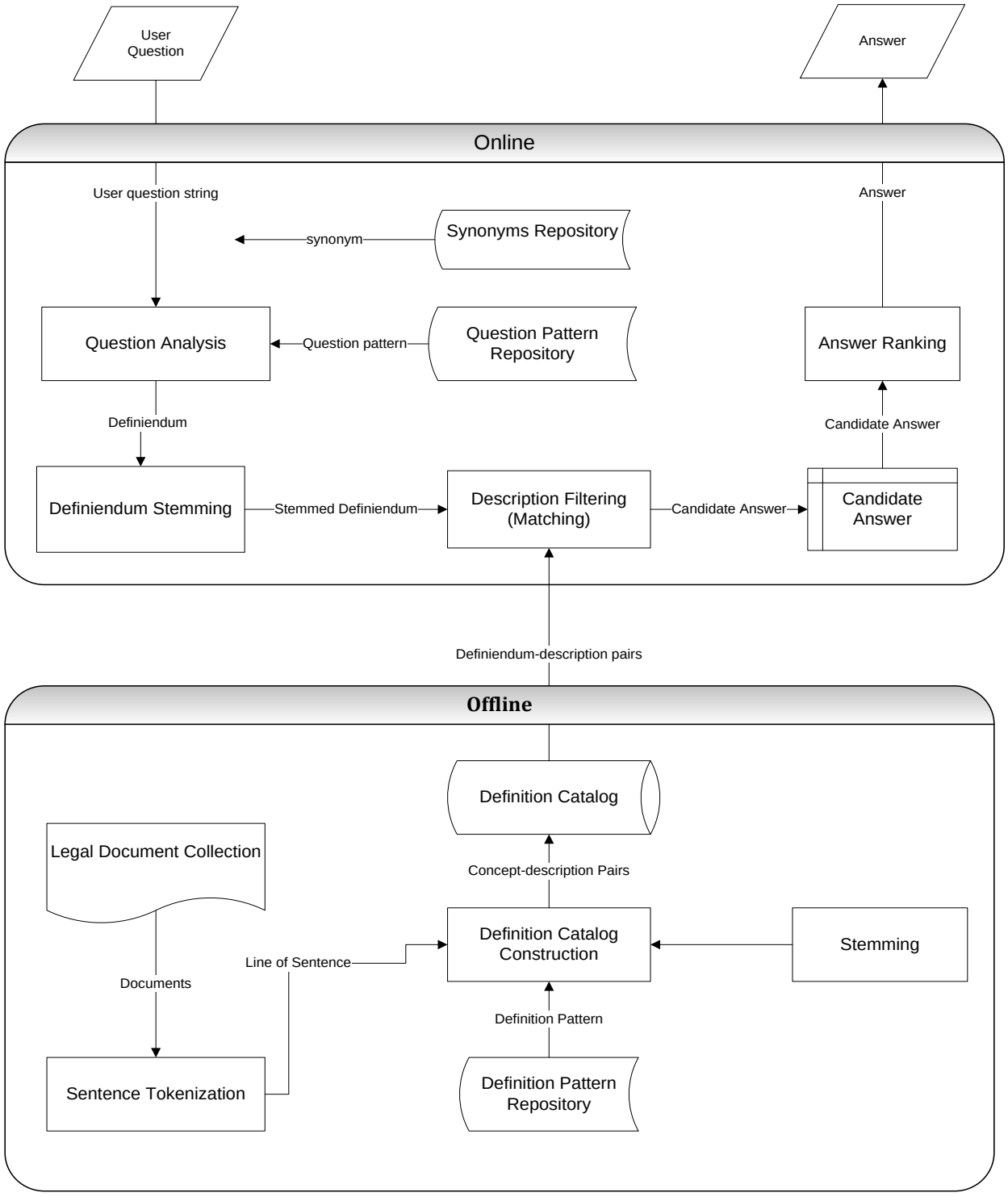

Fig 4.1 Architecture of the system

The main components of the Afan Oromo definition question answering system are Question analysis, Document processing and Definition answer Extraction. Different activities are performed under each component to extract the right definition for the natural language question posed by the user. The main components of the Afan Oromo definition question answering are discussed below:

4.1 QUESTION ANALYSIS

As it is depicted in figure 4.1 above question analysis is one of the component in the system architecture. The main goal of QAS is to extract and return a short and exact description for the natural language question posted by the user. In order to extract the required answer from the collection source (from where the answer should be extracted) the QAS should first understand the natural language question posted by the user. This is to mean that the system should identify information like what the questions mean (their semantic), what type of answer is expected by the question (expected answer type) and what is the main focus of the question (the question focus).Therefore if such kind of information are clearly identified from the question posted by the user, the chance of extracting the correct answer for the posted question by the system is high.

The first stage of any question answering system is the question analysis phase and the input to this phase is natural language question posed by the users (Greenwood, 2006).So the question analysis module of the proposed system is responsible for analyzing the natural language question posed by the user. The basic function of this component is to extract the definiendum (to mean the key word or the word required to be defined by the system) from the natural language question posed by users.

The definiendum extraction is performed by the system based on the manually constructed definition question patterns. Some of the manually constructed definition question patterns are indicated in Table 4.1. We have manually constructed different definition question patterns by analyzing the different ways in which definition questions are formulated in Afan Oromo language. This is done by reviewing in detail the legal document

corpus collected for the purpose of the system performance evaluation how definition sentences are constructed. And also Afan Oromo language expert has been consulted on the ways in which definition sentences and questions are constructed in Afan Oromo language.

Although definitional question answering is challenging because of the fact that the natural language question posed by the users does not contain extra clue which may simplify the process of extracting the correct definition for the question, in our case the question patterns that are formulated manually are applied in extracting the right definiendum from the question.

Question analysis component is a key for extracting the correct definition for the natural language question because of the fact that if the definiendum is wrongly identified at this phase the chance of extracting the correct answer by the rest of the components of the system is very less. Therefore, it is essential to carefully analysis the user question in the process of identification of definiendum from the users' questions at this phase.

4.1.1 DEFINIENDUM EXTRACTION FROM USERS' NATURAL LANGUAGE QUESTIONS

Identification of the definiendum is the basic task which should be performed prior to the answer extraction. As we have said above, identification of the correct question focus has a great effect on the answer extracted by the answer extraction module. To clarify how the definiendum identification from the users' natural language question is performed it is necessary to see some of the different ways how interrogative sentences are constructed in Afan Oromo language. We have clearly identified how different interrogative sentences like factoid, list and definition are formulated in Afan Oromo language in Chapter Three under interrogative question subtopic. Since the focus of this thesis is on definition question type we have focused in detail on how the definition questions are formulated in the language.

QUESTION PARTICLES: users use different distinctive structures to indicate definition type question in Afan Oromo. The question particles are used for different types of question formulation. Table 4.1 shows some of question particles that are used in natural language question formulation in Afan Oromo language.

S.No	Question Particles	Question Types
1	<definiendum> jechuun maal jechuu dha? <definiendum> jechuun maalii dha? What does <definiendum> mean?	Term/expression
2	<definiendum> ibsi. Describe <definiendum>	Term/expression
3	Ibsi <definiendum> maal jechuu dha? What is the description of < definiendum >?	Term/expression
4	<definiendum> eenyuu dha? <definiendum> eenyuu? Who is < definiendum >?	About person ,organization, places, things, etc
5	Jechi < definiendum > jedhu maal? What does < definiendum > mean?	Term/expression
6	Hiikni < definiendum > maal? What is the meaning of < definiendum >?	Term/expression
7	< definiendum > hiiki Define <definiendum>	Term/expression
8	<definiendum> maali? Maali < definiendum >? What is < definiendum >?	Term/expression
9	Jecha < definiendum > jedhu ibsi. Describe the word < definiendum >.	Term/expression

Table 4.1 Afan Oromo Definition question particles.

As indicated in Table 4.1 definition questions can be formulated in different ways. Identification of definiendum from poorly formulated definition question is difficult. To improve the process of definiendum identification from the definition question posted by

the user the researcher has analyzed in advance how the definition questions are formulated in Afan Oromo. As shown in the Table 4.1 above the keyword such as 'jechuun,ibsi,jechi,hiiki,hiikni,hiikti,eenyu,maali,maal' etc are used as a clue in identifying the definiendum from the question.

For example definition question of the word 'seera' can be constructed like

- ✓ 'Seera jechuun maal?'(to mean in English what does law means?)
- ✓ 'Jecha Seera jedhu ibsi'(to mean Describe the word law.)
- ✓ 'Hiikni Seera maali?'(to mean what is the meaning of law?)
- ✓ 'Seera ibsi'(to mean describe law.)

Thus, the above example indicates that in Afan Oromo language one definition question can be constructed in different ways in which all of them seek the same information for the definiendum 'seera'. By analyzing in detail the ways in which the definition questions are constructed, we have manually constructed question patterns in order to identify the definiendum when the user has requested for some information.

In Afan Oromo language like other languages there are compound words that include two or more words that are taken as one word. We have also applied the patterns we have constructed manually on the natural language question to extract the definiendum which is compound word from the user question. For example in the sentence 'Abbaa seeraa jechuun nama wa'ee seeraa irra hojatu jechuu dha.' In this sentence the phrase 'Abbaa seeraa' is identified as a definiendum by the question analysis module and the keyword 'jechuun' is considered to extract the definiendum. To take the phrase as one word we have used the hyphen (-) to combine the words in the phrase. In the above example the phrase 'Abbaa seeraa' after identified as a definiendum,the two words are combined like 'Abbaa-seeraa' and transferred to the next module as a definiendum.

STEMMING: It is one of the techniques used to handle word variants by grouping words that share the same morphological root. Afan Oromo is one of the language which has extensive inflectional and derivational features. These extensive inflectional and derivational features of the language are presenting various challenges in text processing

and information retrieval tasks (Debela,2010).Since one root word can have different word variants, stemming has a great effect on information retrieval process. Root word is part of a word that cannot be changed whether or not the word is morphologically inflected.
Many modern search engines associate words with prefixes and suffixes to their root word to make the search result to have a greater number of relevant matches. Afan Oromo is morphologically rich language where many words can be conflated into a single stem (Debela, 2010).A morphological variant of a word can differ in tense, case, plurality, etc. or can be different in meaning or class.

Just like information retrieval system stemming has also a great effect on QAS. For example the word 'seer' is the stem for 'seera, seeraaf,seeraan, seerotaa' , etc. If stemming is not done on the word 'seera' and all of the words that are morphologically variant with the word 'seera' are considered differently during the answer extraction process. In this paper stemming is applied on the definiendum word identified from the question of the user and also from each of the document in the corpus.

During the search process when the question is posed by the user it is analyzed by the question analysis module and the definiendum is identified. Each of the definiendum identified by the question analysis module is stemmed to their respective stem word. For the purpose of this thesis Afan Oromo stemmer which was developed by (Debela, 2010) has been adopted and used. Debela(2010) has developed the stemming by java programming language. For the purpose of this study the researcher has implemented the stemming by python programming language. Words that are different because of morphological changes like tenses, number, case and gender are stemmed to the same root by the stemming algorithm.

SYNONYM IDENTIFICATION: For the purpose of this study**,** a list of synonyms from the legal documents were identified and compiled manually to use during answer extraction. Each of the definiendum is matched against each of the synonym word in the synonym file and for all the words in the synonym files that match with the definiendum, the answer extraction is also performed for them.

4.2 DOCUMENT PROCESSING COMPONENT

Under this module pre-processing like stemming, sentence demarcation and definition catalog construction are performed. Preprocessing of documents can improve the process of extracting the correct answer from the corpus. The legal document corpus is analyzed and the candidate definition sentences are identified from the corpus. These candidate definition sentences are identified by applying the definition patterns (definition patterns are the way definitions are constructed in Afan Oromo language) we have prepared manually.

4.2.1 SENTENCE TOKENIZATION

In Afan Oromo language sentences are end with full stop (.) like that of English. And they are also separated by Question marks (?) for interrogative sentence and exclamation marks (!) for commands. Because of the fact that like English language most of definition sentence types are delimited by full stops, during document preprocessing we didn't consider the exclamation marks (!) and question marks (?).Thus we have removed all the sentences that end with exclamation mark or question mark during our corpus pre-processing.

During corpus preparation we have marked the end of a sentence with full stops in order to simplify sentence tokenization. Full stop can also be used in abbreviation like that of English language. Some time it creates ambiguity when it is used as sentence delimiter and to solve this ambiguity we have also excluded all words that are written in abbreviation from our corpus.

We have exclude abbreviation because of the fact that in Afan Oromo language the usage of abbreviation is very less. Most of the time instead of abbreviation slash mark (/) is used in the language. For example, the phrase 'Mana Murtii Ol'aanaa' can be written as M/M/O (which is to mean Supreme Court).And also we didn't consider this type of cases in this study and we have excluded phrases or words that are written in this format. Because to study this ways of writing in detail it requires more time and also corpus from different domain. For the purpose of tokenizing the sentences in our experimental corpus we have used full stop as a sentence delimiter.

4.2.2 DEFINIENDUM EXTRACTION FROM THE DOCUMENT CORPUS

Definiendum extraction from the legal document corpus is done under this document processing module. Definiendum extraction from the document corpus is required because of the fact that during answer extraction each of the definiendum from the users' natural language question is compared with the definiendum in the corpus. So specifying each definiendum in the document corpus simplifies the task of extracting the definition from the corpus.

As the researcher has discussed in the literature review part there are different approaches to extract answers for a definition type questions. Among them pattern based matching is used to identify definiendum with their description in each of the document (Wondowossen, 2013). In our case in order to identify the definiendum from the corpus we have applied definition patterns. We have manually constructed definition patterns after analyzing different ways of constructing definition sentences in Afan Oromo.

S.NO	Patterns
1	‹definiendum› jechuun
2	‹definiendum› yemmuu ibsamu
3	‹definiendum› yeroo ibsamu
4	Ibsi ‹definiendum›
5	Hiikti ‹definiendum›
6	‹definiendum› yeroo hiikamu
7	Hiikni ‹definiendum›
8	Jechi ‹definiendum› jedhu yeroo ibsamu
9	Jechi ‹definiendum› jedhu yammuu hiikamu/hibsamu
10	‹definiendum›

Table 4.2 Definition patterns used to identify definition-description pairs in the corpus.

As indicated in the table 4.2 definition sentences in Afan Oromo can be formulated in different ways.

For example:

- Seera jechuun sirna ittin biyyi tokko buluu dha.
- Ibsi seera sirna ittin biyyi tokko buluu dha.
- Jechi seera jedhu yeroo ibsamu sirna ittin biyyi tokko buluu dha.
- Hiikni seera sirna ittin biyyi tokko bulu jechuu dha.
- Abbaa seera jechuun ogeessa seeraa jechuu dha.

The above example indicates that one definiendum term which is 'seera' can be defined in different ways. We have manually constructed the definition pattern by considering such kinds of definition format in the language. In the above example the keywords 'jechuun,Ibsi,Hiikni',etc are used as a clue in definiendum extraction process from the documents of our corpus.

In the above example in the last sentence the definiendum term which is 'Abbaa seeraa' is compound words. The phrase has two keywords that are 'Abbaa and seeraa'.To solve such kinds of complexity during the identification of the definedum we have used hyphen (-) in order to consider the phrase as one word. Thus, in the sentence 'Abbaa seeraa jechuun ogeessa seeraa jechuu dha',after the phrase 'Abbaa seeraa' is identified by applying the definition patterns we have used the hyphen(-) to make the phrase like 'Abbaa-seeraa' after that the phrase is considered as a single word. This is required because each of the definiendum is considered as a single word and it simplifies the process of definition catalog construction at definition catalog construction module.

4.2.3 DEFINIENDUM PREPROCESSING

Some pre-processing tasks are performed on the definiendum that is extracted from the document at the definiendum extraction stage. Performing preprocessing tasks on the definiendum at this stage will improves the answer extraction process later at the definition extraction component. Special character like ("@#%$^&*()/,~;}]") and punctuation marks (. ? !) are removed from the definiendum at this stage. When the identified definiendum from the document corpus is a compound word, after preprocessing

is applied on it the words are interconnected with the symbol hyphen (-) to consider it as a single word. Stemming is only performed on the words that are not compound words. As stated by different local (wondowossen, 2013) and (kasahun, 2014) and different global researchers stemming has an effect on the QAS.

4.2.4 DEFINITION CATALOG CONSTRUCTION

Definition catalog is the one which is used by the definition extraction component during answer extraction process. At this stage every definition in the document corpus are identified and indexed and saved as concept- description pairs (concept indicates definiendum) in a separate file.

We have also applied the definition patterns that we have constructed manually to identify each definition from the list of sentences tokenized at the sentence tokenization stage. The candidate definitions are those lists of sentences that have the same format with the definition patterns we have constructed. After all of the potential definitions are extracted from the document we have applied the definition pattern on the candidate definitions to separate into definiendum and its description. We have used the special character (#) to separate the definiendum from its description in the definition catalog (like concept#description).

Then the concept-description pairs are indexed and saved in a separate file as a dictionary which is used by the answer extraction component later to extract the required answer. Separating the definiendum from its respective description in the definition catalog simplifies the answer extraction process because at the answer extraction stage the candidate definitions in the definition catalog are tokenized by using the special character (#) to store each definiendum and its equivalent definition separately.

4.3 DEFINITION ANSWER EXTRACTION COMPONENT

At this stage the actual matching is performed to extract the candidate answer**.** At the question analysis stage the definiendum from users' question is identified and preprocessing like special character removal, punctuation mark removal and stemming are applied on each of them as discussed above under the question analysis component.

And also as we have indicated under the document processing component all the candidate definitions are extracted and stored in a separate file as concept-description pairs.

Under this module the comparison between the definiendum indentified at question analysis stage and with each of the definiendum in the definition catalog that was identified at the document processing stage is performed. During the comparison when a match exists between the definiendum which was identified at the question analysis stage and the definiendum in the definition catalog, a description of a definiendum in the definition catalog is taken as a candidate answer and stored in a temporary variable. After the comparison is completed if candidate answers are specified, they are forwarded to the Answer ranking module. But if nothing is identified as a candidate answer during matching, the message 'There is no answer' is returned for the user at this module.

4.4 ANSWER RANKING

Under this module the candidate answers identified at the answer extraction stage are ranked. We have considered the frequency of each of the candidate answer in the documents of the corpus for candidate answer ranking purpose. The frequency of each of the candidate answer is to mean that the frequency of the answer in the document.

The ranking are performed as follows: Each of the candidate answer is taken and compared with all the sentences in each of the document of the legal corpus.

The frequency of the candidate answer per each of the document is assigned to each of the candidate answer as a weight. After all the frequency of the candidate answer in the documents are specified, the candidate answer with their frequency are ranked by their frequency in descending order and stored in as a dictionary. The candidate answers with greater frequency are ranked at the top in the dictionary. When the answers are displayed for the user, they are displayed as organized in the temporary dictionary variable which means the answer with a greater weight is displayed at the top. For example: if the candidate answers for the user question 'seera jechuun maali' are 'seera jechuun sirna ittiin bulmaatati', and 'seera jechuun sirana biyyi tokko ittin oogannamuu dha', and if the first candidate answer occurs in two documents (doc1 and doc2) and the second candidate

answer occurs in three documents (doc1, doc2 and doc3) then the second candidate answer is ranked at the top and when the answers are displayed to the user it is located first. That means the answers are displayed as:

seera jechuun siran biyyi tokko ittin oogannamuu dha.

seera jechuun sirna ittiin bulmaatati.

CHAPTER FIVE

IMPLEMENTATION AND EXPERIMENTAL EVALUATION

In this chapter the performance of prototype system is evaluated with the legal corpus which the researcher has prepared. And the implementation of each of the module of the prototype system is discussed. Although all the modules in the prototype system are important in the process of answer extraction, the question analysis module is a key in extracting the correct answer to the question.

At the question analysis stage if the definiendum is wrongly identified from the user question the chance of getting the correct answer is very less. Since the question analysis module has a great effect to extract the correct answer its performance is separately evaluated. Then the performance of the prototype system as a whole is evaluated. As we have specified in chapter four of this thesis the prototype system has four general components. They are question analysis, document processing and answer extraction. Under each of the core components there are some sub modules. The implementation of each of the components with their respective sub modules are discussed in this chapter.

5.1 QUESTION ANALYSIS MODULE

This module mainly focuses on how the process of definiendum extraction from natural language is implemented. At this stage provided the natural language question, the questions are analyzed and represented in a way suitable to extract the definiendum. Most question answering systems start from the question analysis phase that attempts to determine the semantic and syntactic elements of the question that are important for the implementation of other modules that depend on the output of the question analysis module. The activities implemented in this module are accepting the natural language question, tokenizing the question phrase if the word in the requested natural language question is more than one and analyzing the position of each of the token in the question phrase with the manually constructed Afan Oromo definition question type patterns. As we have indicated in detail about the definition question particles in chapter four, Afan Oromo definition question particles are essential in this module to analyze the question that

improves the accuracy of this module in extraction of definiendum from the users' question.

The Afan Oromo definition question particles like 'jechuun,hiikni,ibsi,jechi,hiiki',etc. are keys in the implementation of definition extraction module. These definition keywords indicate the question phrase is looking for a certain description. If any of the definition question keywords do not exist in the natural language question, the definiendum extraction task is complicated and nothing is returned as a definiendum .When such kinds of situations are occurred the system returns 'There is no answer'. After the definiendum from the natural language question are identified tasks like stemming, special character removal and punctuations mark removal are performed on each of the definiendum before the definiendum is forwarded to the next module. And also query expansion is performed by using synonyms.

```
For a definiendum D:
if there is a match exists in Synonyms list  L then
take all the tokens equivalent with the definiendum d
else
return null
      End If
End For
```

Algorithm 5.1 pseudo code for query expansion

```
For Question Q:
tokenize the question phrase in to tokens
if number of token in the question is one then
take the word as a definiendum
else if the number of token are two
if the tokens contain[ibsi,hiiki,maali,maal,eenyu] then
take the first token as a definiendum
else
take the two tokens as a definiendum
        End if
else if the number of tokens are more than two then
if the tokens contain[jechuun,ibsi,maal,maali,hiiki] && the number of tokens before
these keywords are two or one then
take the token as a definiendum
else if the tokens contains[jecha,ibsi,jechi] && these keywords come at the beginning of
the question then
take the second token as a definiendum
else if the tokens contains none of [jecha,jechuun,ibsi,maal,maali,hiikni,hiiki,jechi...] then
take null as a definiendum
End IF
    End If
End For
```

Algorithm 5.2 pseudo code for definiendum extraction from users' natural language question

5.1.1 SPECIAL CHARACTER AND PUNCTUATION MARK REMOVAL

Before the definiendum is forwarded to the definition extraction module, special character and punctuation marks removal are done on the definiendum . Removal of special character and punctuation mark are necessary at this stage because of that if the definiendum contains these kinds of character the definition extraction component wrongly understands the words and the chance of extracting the correct answer is minimized. For example, if the user question is like 'Sirna jechuun maal?',the word 'sirn&a' is identified as a definiendum by the definiendum extraction module. Then before the word is forwarded to the next module, the character '&' is removed and the word 'sirna' is forwarded to the definition extraction module.

```
For a definiendum D:
for each  character in D
if a character is in [?.!] || a  character in [@#%$^&~] then
remove the character from d
      End If
  End For
End For
```

Algorithm 5.3 pseudo code for Special character and punctuation mark removal

5.1.2 STEMMING SUB MODULE

Stem term is part of the word that never changes even when morphologically inflected. For example, in English the word finish is the stem for words like finishes, finished, finishing, etc. and if the word finish is taken it is no more stemmed to any other word. Like that of English in Afan Oromo language the words 'seeraan','seeraaf','seerota' are stemmed to the stem word 'seer'. Here the word 'seer' is no more stemmed to other word. To analyze the effect of stemming on the performance of the system in extracting the answer we have

performed the experiment with and without applying stemming on a definiendum extracted from the users' question and from the corpus. In addition to both the definiendum from the users' question and the corpus we have applied stemming on synonyms of definiendum extracted from users' question. For the purpose of this study we have adopted Afan Oromo stemming developed by Debela(2010).He has developed with java programming language. The researcher has implemented the stemmer by using python programming language which python was selected as an implementation tool for this study. The stemmer coverts each definiendum from user natural language question and definiendum identified in the corpus in to their respective stem word. The stemmer removes different suffixes from the definiendum to change into their respective root word.

5.2 DOCUMENT PROCESSING MODULE

Under this module the experimental corpus is analyzed and the definiendum in the document is identified. As we have discussed in chapter four part of this thesis paper to identify the definiendum from the document corpus we have applied the definition patterns that we have constructed manually. In order to more simplify the process of answer extraction at the definition extraction module, from the experimental legal corpus the maximum potential concept-description pair is extracted in this module. All the identified definiendums from the document corpus are stemmed into their stem word because we have performed our experiment before and after stemming the definiendum words. After all the possible concept-description pairs are identified, they are saved in a separate file as a definition catalog that used later by definition extraction module to extract the answer.

5.3 DEFINITION EXTRACTION MODULE

The candidate definitions are extracted at this module. Definition question types require full word meaning, term definition, description of a word, etc. In this module we have extracted full sentences as an answer for a term. All the candidate answers are extracted from the definition catalog prepared at the document processing component. The output of the question analysis module which is the definiendum is accepted by this module and matched with each of the definiendum in the definition catalog file. Then when the exact

match occurred with the definiendum in the definition catalog, the description of the definiendum in the definition catalog is taken as a definition and stored as a temporary variable.

5.4 ANSWER RANKING MODULE

This module is the last sub module which sorts the candidate answer before returned to the user. To rank the candidate answer we have used the frequency of each of the candidate answer in each of the document in the corpus. Each of the answer in the candidate list are taken one by one and compared with the sentence in each of the document in the corpus and the frequency of the answer sentence in each of the document is taken as a weight for the answer candidate. Then the candidate answers are ranked depending on their weight and stored in a temporary dictionary variable. After the ranking is done, the candidate answers are returned to the requested user by their descending order of their weight.

5.5 PERFORMANCE MEASURE

The effectiveness of the system was evaluated in terms of how the correct answer is returned. This was tested by calculating recall and precision in terms of how the answers of a given question are correct, complete and exact.

The question analysis module is a key in the system because at this stage if the definiendum from the user's question is not correctly identified the chance of returning the correct answer by other module is very less. Because of this in addition to the performance of the system as a whole we have evaluated question analysis module in terms of the extent it identifies the definiendum from the user's natural language question. Stemming also has an effect in extracting the answer. We have evaluated the system by applying and without applying stemming. Applying synonyms is one of the query expansion methods in question answering system which helps to use the semantically related terms with the definiendum to increase the coverage of the retrieval phase(Kasahun,2015).And also we have evaluated the performance of the system with and without applying synonyms. The total performance of the prototype system as a whole was evaluated by the degree to which the

definition is correctly extracted from the corpus. Thus; we have evaluated the performance of the proposed system as a whole with precision, recall and f-measure.

5.5.1 DATASET PREPARATION

Since there is no corpus prepared so far for Afan Oromo definition question answering by other researchers it requires the researcher to compile and prepare a corpus for the experiment purpose. Thus, the documents we have used for the experiment were collected from Oromia legal research and training institute. The document contains different proclamations, training manuals, research papers and other law related directives. From each of these categories we have compiled 250 pages document. From this 250 pages of documents we have prepared twenty five documents with each of them has ten pages for experimentation purpose. The legal area is selected because of the fact that the documents from such area contain different definition types that are appropriate for evaluating the performance of definition question answering system, the documents are easily available and most of the time legal experts seek definition for different legal related terms in their daily activities. Rather than searching from documents using the question answering system helps them to perform their activities efficiently.

5.5.2 QUESTION PREPARATION

For the purpose of evaluating the performance of the system we have provided the documents to five individuals. Two of the selected individuals are legal professionals and the rest are selected from other areas. Each of the individuals has formulated five questions from the document. Each of the individuals was informed to prepare definition question types.

A total of twenty five questions have been formulated by the selected individuals for evaluating the performance of the proposed system. Later all the 25 questions are provided to the assessor for the purposing of evaluation .The selected assessor was a legal expert and has first degree in law. And the assessor is a fluent speaker of Afan Oromo language.

The number of questions has been limited to 25 because the researcher thinks that these questions can show the ways in which definition question types are formulated in Afan Oromo language.

5.5.3 ANSWER JUDGMENT

Users have different opinion in evaluating the information retrieved from information retrieval system. Question answering system is designed to satisfy real world information by providing exact answer rather than a ranked list of document. Relevance of the result returned by question answering system is one of the criteria used to evaluate definition question answering system.

Users can agree or disagree on the answer returned by question answering system, because different users have different needs for the answer returned by the system. The mismatch of opinions related to the answer returned by the system among different users affect the conclusion which is drawn from the evaluation result. To solve such kind of variation, we have selected the domain expert, to be the assessor of the prototype system and he has identified the correct and incorrect answers returned by the system.

One assessor was selected because of the fact that there is no standardized body which is responsible for evaluating the correctness of the answer returned by question answering system in case of our country and the selected assessor has a good experience on legal related activities.

Depending on the information nugget prepared by the assessor we have evaluated the performance of the system. There are cases where a system can return more than one answers, in this case if one or more the returned answer are judged as vital the system will not be penalized. The system is penalized only when it does not retrieve a correct answer (this is to mean that when the answer returned by the system is judged by the human assessor and the answer is taken as wrong, the system response is not taken as vital).

5.5.4 EXPERIMENTATION

5.5.4 .1 EXPERIMENT ONE: DEFINIENDUM IDENTIFICATION FROM QUESTIONS

Definiendum identification is a key in definition question answering system. Because when it is wrongly identified the overall performance of the system is affected. The question analysis module has been evaluated in terms of its performance in identifying the

definiendum from users' natural language question. If the definiendum is identified incorrectly at this stage the chance of getting the correct answer is very low.

This module has a great effect on the answers that returned by the system. In order to identify the definiendum from the users' natural language question we have manually constructed patterns by studying in detail the ways in which definition questions are constructed in Afan Oromo language. In addition to the performance testing questions, we have prepared fifteen definition questions other than 25 questions from the corpus separately to evaluate the definiendum identification capability of the question analysis module. From the prepared question three of them contain compound words as their definiendum.

From the 15 questions the question analysis module identifies the definiendum correctly from 14 of them. The module only identifies a definiendum wrongly from one question. The question from which the module returns a wrong definiendum was that 'Jechi itti gaafatamaa jedhuu maali?', from this question the module returns empty value. This is wrongly identified, because we didn't consider in detail about compound words patterns in Afan Oromo language in this study like that of none compound words.

For example, the question 'seera jechuun maal jechuu dha?' In this question the key word 'jechuun' is considered to identify the definiendum . The system considers every words that are located before the keyword 'jechuun' and when the index of the words are less than or equal to two. If the identified definiendum is a compound word, the systems concatenate the phrase with hyphen (-) to consider it as a one word.

For example: Abbaa seeraa jechuun maali?.From this question the system takes the phrase 'Abbaa seeraa' as a definiendum . In order to consider the phrase as one word the system concatenates the words like 'Abbaa-seeraa'. Then the system returns 'Abbaa-seeraa' as a definiendum to the next module in the system. From the experiment we have seen that the module correctly identifies 93.3% ofthe definiendum from the users' natural language questions that are prepared for evaluation purpose and only identifies 6.7% of the definiendum wrongly.

5.5.4.2 EXPERIMENT TWO: EFFECT OF STEMMING ON PERFORMANCE

In order to identify the effect of stemming on the performance of the system we have performed evaluation before and after stemming. Stemming has an effect on the answer which is returned by the system. If the words 'seera,seeraan,seeraaf,seerota'are not stemmed to their root word the system retrieves their answer separately.

For example: From the questions 'Seera jechuun maal jechuu dha?' and 'seerota jechuun maal jechuu dha?',the definiendum extraction module has identified the words 'seera and seerota' as a definiendum .If the two words are not stemmed to their stem word 'seer',at answer extraction stage the system extracts an answer for only one of them or it returns 'There is no answer'. For example, if the document corpus contains definition for the word 'seera', when the user request definition for the word 'seerota' the system returns 'here is no answer'. This is because when both the definiendum in the users' natural language question and in the definiendum are not stemmed to their stem word, the system considers the two words differently.

Thus, evaluating the system with stemming is crucial to see such scenarios. Accordingly the evaluation result with stemming shows recall 76%, precision 85.5 and F-measure 81 % and without applying stemming the evaluation result indicates 85% precision, 58% recall and 70.13% F-measure. The experimental result shows that the stemming has an effect on the performance of the system.

Effect of Stemming on Performance

Before Stemming			**After Stemming**		
Precision	Recall	F-measure	Precision	Recall	F-measure
85%	58.72%	69.45%	85.5%	76%	80.4%

Table 5.1 Effect of Stemming on Performance

5.5.4.3 EXPERIMENT TWO: EFFECT OF SYNONYMS ON PERFORMANCE

We have manually constructed list of synonyms and used for query expansion. After the question that has been posed by the user has been analyzed and the definiendum is identified from them, the query generation module compares the definiendum with each of the term in the synonym file in order to increase the coverage of the queries. At answer

extraction module, in addition to the definiendum the answer extraction is performed for words in the synonym file that are semantically equivalent to the definiendum from the users' question.

For example:'seera jechuun maal jechuu dha?' After the definiendum which is 'seera' is identified from this question, it is compared with each of the manually constructed synonym. If there are words in the synonym file that are semantically equivalent with the definiendum 'seera' in addition to the definiendum 'seera' the answer extraction process is performed for the synonym words. Even though the definition for the synonym word exists in the corpus unless the query expansion is performed with the synonym the answer is not extracted for the synonym term.

For example in the synonym file there is a word 'haqa' which has equivalent meaning to the word 'dhugaa'. The user requested as 'ibsi dhugaa maali?'.The system returns 'There is no answer' to the user, because the definiendum is not compared with the synonyms list in the synonym file. But when the semantically equivalent word with the definiendum 'dhugaa' in the synonym file is found, the system searches definition for the definiendum and its synonyms.

The above example shows that when the answer exists with the synonym word in the definiendum but not with the definiendum unless the query expansion is done with the synonym word the correct answer is not returned. Therefore, to see the effects of synonym it is required to evaluate the performance of the prototype system with and without applying synonym on definiendum. Accordingly, the result shows that applying synonyms on definiendum from natural language question more improves the answer coverage than without using synonym.

5.5.4.4 GENERAL EVALUATION OF AFAN OROMO DEFINITION QUESTION ANSWERING SYSTEM

Evaluating definition question answering system is much more difficult than evaluating QA system that answer factoid question because it is not acceptable to judge a system response as simply right or wrong (M. Voorhees, 2003).As the researcher (M. Voorhees, 2003) states some mechanism should be used to assign a particular weight to the system response in order to match the concepts in the desired response to that of the concepts in the system's

response. Before the testing is started all the prepared questions by the selected individuals and the prepared corpus are presented to the selected assessor. Then he has created a list of information nugget about the definiendum depending on the list of the questions, corpus and also his opinion. Since the assessor is a domain area expert he has used his opinion in developing the information nugget.

Information nuggets are facts that help the assessor to decide whether the system response is valid or not. The prepared information nuggets are per each question and are assumed as vital (vital in this case to mean nuggets that must exist in a definition for that definition to be acceptable).Then after all the information nuggets are prepared, the assessor has started to evaluate the system response. Depending on his judgment and the information nugget, he has assigned a number for vital and non vital nuggets in the retrieved system response.

During the evaluation, the assessor has focused on the content (concept) of the system response. And he has ignored wording difference and considered conceptual match rather than syntactic matches. Depending on the assessor judgment, the recall has been calculated as the number of correctly retrieved nuggets divided by the number of nuggets on the assessor's list. But by considering only the number of nuggets correctly retrieved and the total number of nuggets retrieved for calculating the precision is difficult, because the correct value of the denominator which is the total number of nuggets retrieved is unknown. To solve such kind of problem we have used a length as a crude approximation to precision. In this case the length is to mean that the number of characters in the answer returned by the system.

The length based measure shows the concept that users would desire more the shorter of two definitions that have the same concepts (Harman &Over, 2002).The length based measure that was used in the TREC for evaluation purpose has assigned allowance of 100 non white space characters for each of the correct nuggets retrieved. The precision score has been assigned to one if the system response is less than or equal to the allowance. But when the system response is greater than the allowance the precision score is decreased. In the formula length shows that the number of non white space characters in the retrieved vital nugget. The system has penalized for not retrieving vital nuggets. The nugget recall is calculated as the vital nugget retrieved over the total vital nuggets in the assessor's list.

Non vital nugget retrieved is not considered at nugget recall calculation but it is considered in precision calculation (because allowance is calculated by considering the number of non vital nuggets retrieved).The final score for the definition from the system was computed by using F-measure. We have taken the assumption that both recall and precision are equally important (during F-measure calculation we considered β=1).The recall, precision and F-measure has been calculated manually.

Given the number of vital nuggets retrieved from the system (V) and the total number of vital nugget in the assessor list (A) ,recall (R) is calculated as:

R= V/A……..equation 5.1

Given the number of vital nuggets retrieved from the system (V) and the total number of non vital nugget retrieved from the system (N), Allowance is calculated as:

Allowance=100*(V+N)……equation 5.2

Given the total number of the non white space characters in the retrieved nuggets from the system (L) and Allowance, precision (P) is calculated as:

P = 1 if L is less than allowance else ……equation 5.3

P = 1- (L-Allowance)/L

Given recall and precision-measure is calculated as:

F-measure=2PR/P+R……equation 5.4

Therefore, according to the procedure and the formulas we have discussed above, the evaluation metrics shows that precision is 86.4%, recall is 78% and f-measure is 81.9%.Out of 25 questions the system correctly extract answer for 21 of them. And the system returns wrong answer for four question.

From the result we have realized that the application of stemming component and also usage of synonym for definition question answering system improves the performance of the system.

Precision	**Recall**	**F-measure**
86.4%	78%	81.9%

Table 5.2 performance of Afan Oromo Definition Question Answering system as a whole

According to the experiments the performance of Afan Oromo definition question answering system is affected by the usage of stemming and also synonyms list. And also the findings shows that the question analysis and answer extraction components mainly the language dependent of the components. The pattern based approach for the process of extracting answer from the document corpus is effective and as indicated in the table 5.2 the performance of the system as a whole is encouraging.

CHAPTER SIX

SUMMARY, CONCLUSION AND RECOMMENDATION

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 users' 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 definition question answering system for Afan Oromo language. The objective of the study is to explore the possibility of developing Afan Oromo definition question answering system. Pattern based approach was employed to extract the words to be defined from the natural language questions and also to identify the definiendum from the document corpus used for the study.

The domain of the study was a closed domain area which is specific to legal related documents collected from Oromoia Legal research and training institution. Python programming language was selected for implementing the system. In the study the feasible architecture for Afan Oromo definition question answering system was explored and the main components of the architecture are question analysis, document processing, and definition answer extraction and answer ranking. At the question analysis stage the natural language question is accepted and analysis like tokenization, stemming and synonyms applications are done and the definiendum is identified. At the document processing stage sentence tokenization from the document corpus and stemming of the definiendum words are performed.

Definition catalog construction which contains each definiendum with its description was performed under the document processing stage. The answer extraction was performed by matching the definiendum identified at question analysis stage and the definiendum in the definition catalog. The description of the definiendum in the definition catalog is taken as an answer when there is an exact match between the two definiendum. The answer

ranking was performed depending on the frequency of the answer in each of the documents used for the experiments. The answer with the greatest frequency has been ranked at the top when it is displayed to the user. The performance of the system has been evaluated with the standard evaluation metrics precision, recall and f-measure. The result of the evaluation is a recall of 78%, precision 86.4% and F-measure of 81.9%.

6.2 CONCLUSION

According to the findings of the study applications of stemming and synonyms have an encouraging effect on system's performance than without usage of stemming and synonyms list. As each of the module of the system has an effect on the performance of the system, we have evaluated the performance of the question analysis module and the experiment result shows that from 15 questions the question analysis module identifies the definiendum correctly from 14 of them which 93.3 % and only from 1 question wrongly identifies the definiendum . We have also realized that the question analysis and definition construction module basically depends on the language structure. This is because the construction of patterns basically depends on the ways in which the question phrase and definitions sentences are constructed in Afan Oromo language. Dealing with all compound words and classifying the type of answer after extraction are the challenges we have faced in this study.

6.3 RECOMMENDATION

Even though, the result of the experiment is a promising result, based on the findings of the experiment we recommend the following points as a future research area to improve the effectiveness and efficiency of the system:

- We recommend that other researcher perform researches to explore the algorithm used to extract an answer for compound words in detail.
- We recommend that other researcher conducts researches on how to classify the answers of the question answering system to their categories.
- This study is performed on closed domain area. We suggest that other researcher can conducted Afan Oromo Definition QAS on Open domain.

- We recommend that other researcher can use more number of questions to evaluate the performance of the system.
- We have manually constructing patterns to identify definiendum from natural language question and document corpus. We recommend that other researcher can explore the ways of developing patterns automatically.

REFERENCE

Abebe Abeshu, "Automatic Morphological synthesizer for Afaan Oromo," Department of Computer Science, Master's thesis, Addis Ababa University, Addis Ababa, 2010.

Aunimo, L., "Methods for Answer Extraction in Textual Question Answering," University of Helsinki, Computer Science, Helsinki, 2007.

Banerjee et al.," BFQA: A Bengali Factoid Question Answering System,"
Department of Computer Science and Engineering, Jadavpur University, India, 2014.

Banko et al., "AskMSR: Question Answering Using the Worldwide Web,"*In Proceedings of 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases* , Palo Alto, California, March 2002.

Bekhti and al-harbi,"AQUASYS: An Arabic question-answering system based on extensive question analysis and answer relevance scoring *,"International journal of academic research, Department of computer science*, College of computer and information sciences, al-imam muhammad ibn saud university, vol. 3. No. 4, July, 2011.

Biruk Eshetu, " Amharic Question Answering for List Questions: A case of Ethiopian Tourism", Master's thesis, School of graduate studies, Department of Information Science, Addis Ababa University,2013.

Brill et al.," An Analysis of the AskMSR Question-Answering System," *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 257-264, July 2002.

Burger et al.,"Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A),"2001.

Cordell green &Raphael," Research on intelligent question-answering system,"
Stanford research institute Menlo Park, California, May 7967.

C.R.Kothari," Research Methodology: Methods & Techniques," New Age
International Publishers, ISBN (13): 978-81-224-2488-1, India, 2004.

D. Manning et al.,"An Introduction to Information Retrieval," Online edition Cambridge UP, 2009.

Debela Tesfaye, "Designing a Rule Based Stemmer for Afaan Oromo Text,"

Master's thesis, Department of Information Science, Addis Ababa University, 2010.

Denicia-Carral et al.,"A Text Mining Approach for Definition Question Answering," Language Technologies Group, Computer Science Department, National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico, 2005.

Figueroa, "Finding Answers to Definition Questions across the Spanish Web," German Centre for Artificial Intelligence DFKI Stuhlsatzenhausweg 3, D 66123, Saarbr ¨ucken, Germany,2010.

F. Green et al.," BASEBALL: An automatic question answer," Lincoln Laboratory, Massachusetts Institute of Technology, pp. 219-224, 1961.

Girma Debele and Martha Yifiru," Afan Oromo News Text Summarizer," *International Journal of Computer Applications (0975 – 8887),* Vol.103, No.4, October 2014.

Greenwood & Saggion," A Pattern Based Approach to Answering Factoid, List and Definition Questions," Department of Computer Science University of Sheffield Regent Court, Portobello Road Sheffield S1 4DP UK, 2004.

Greenwood, "Open-Domain Question Answering, " Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Department of Computer Science, University of Sheffield, UK, September 2005.

H. P. EDMUNDSON," New Methods in Automatic Extracting,"*Journal of the Association for Computing Machinery,* Vol. 16, No. 2, pp. 264-285,University of Maryland, College Park, Maryland, April 19691.

Hammo et al.," QARAB: A Question Answering System to Support the Arabic Language," DePaul University School of Computer Science, Telecommunications and Information Systems, 2004.

Jianan," An Intelligent FAQ Answering System Using a Combination of Statistic and Semantic IR Techniques," master's Thesis ,Department of Computer Science and Electronics at Mälardalen University, Nov.2006.

John J. McCarthy," Formal Problems in Semitic Phonology and Morphology," Linguistics Department Faculty Publication Series, University of Massachusetts – Amherst, 1985.

Kasahun Abdissa," Factoid Question Answering for Afan Oromo," Master's thesis, School of graduate studies, Department of Information Science, Addis Ababa University, 2014.

Katz," Annotating the World Wide Web using Natural Language," *In Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*,1997.

Linand Demner-Fushman," Automatically Evaluating Answers to Definition Questions," *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP),* pp. 931–938, October 2005.

L. HIRSCHMAN & R. GAIZAUSKAS," Natural language question answering: the view from here," Natural Language Engineering **7,** pp.275-300, Department of Computer Science, University of Sheffield, and Sheffield, UK 2001.

Malhotra and Dixit," An Effective Approach for News Article Summarization*," International Journal of Computer Applications (0975 – 8887),* Vol.75, No.17, Department of Computer Engineering YMCA University of Science &Technology Faridabad, India, Aug.2013.

Moldovan et al.,"LCC Tools for Question Answering," *Proceedings of the 11th Text Retrieval Conference*, 2000.

Moldovanand Surdeanu," On the Role of Information Retrieval and Information Extraction in Question Answering Systems," Human Language Technology Research Institute, University of Texas at Dallas, 2003.

Monz," Minimal Span Weighting Retrieval for Question Answering," Institute for Advanced Computer Studies (UMIACS),University of Maryland College Park, MD 20742, USA,2004.

M. M. Soubbotin," Patterns of Potential Answer Expressions as Clues to the Right Answers,"2001.

Poonam Gupta and Vishal Gupta,"A Survey of Text Question Answering Techniques," *International Journal of Computer Applications (0975 – 8887),* Vol.53, No.4, Sep. 2012.

Ravichandran and Hovy," Learning Surface Text Patterns for a Question Answering System, "Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*,* July 2002.

Ritchey, "General Morphological Analysis A general method for non-quantified modeling," *Downloaded from the Swedish Morphological Society* (*www.swemorph.com*), 2013.

R.Mervin and Dr.A.Jaya," Knowledge Based Question Answering System Using Ontology," *International Journal of Engineering Sciences & Research Technology*, Computer Science and Engineering, B.S.Abdur Rahman University, India, October, 2014.

Rosso et al.," Towards an Arabic Question Answering System," Natural Language Engineering Lab., RFIA Dept. of SIC, Polytechnic University of Valencia, Spain Group, 2004.

R. Radev et al.," Evaluating Web-based Question Answering Systems," School of Information, Department of EECS _Business School, University of Michigan Ann Arbor, MI 48109,2002.

R. Radev et al." Introduction to the Special Issue on Summarization,"©*Association for Computational Linguistics*, Vol.26, No.4, pp. 400-408, 2002.

Schlaefer1et al., "A Pattern Learning Approach to Question Answering
Within the Ephyra Framework," pp. 687–694, Springer-Verlag Berlin Heidelberg, 2006.

Seid, Yimam, "Amharic Question Answering (AQA)," Master's thesis, School of Graduate Studies, Department of Computer Science, Addis Ababa University, 2009.

Trigui et al.," DefArabicQA: Arabic Definition Question Answering System," ANLP Research Group- MIRACL Laboratory, University of Sfax, Tunisia, 2008.

Wondwossen Teshome," Designing Amharic Definitive Question Answering," Master's thesis, School of graduate studies, Department of Information Science, Addis Ababa University, 2013.

Wu et al.," CLVQ: Cross-Language Video Question/Answering system," Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering (ISMSE'04), 2004.

Zhang et al.," Answering Definition Questions Using Web Knowledge Bases," Department of Computer Science and Engineering, Fudan University, Shanghai, China, 2004.

APPENDIXES

APPENDIX 1: AFAN OROMO DEFINITION QUESTION ANSWERING SYSTEM PYTHON CODE.

```
import string
import re
import os
import operator
def openfiles(fi):#This function  used to open the file
try:
fp=open(fi,'r')
except IOError:
print("Fa'ila banuu hin dandeenye")
else:
return fp
def splitString(fs):
words=[]
words=fs.split()
   #print(len(words))
def removePunct(word):#This function is used to remove punction
   fWord=""
for cr in word:
if cr in "?.!":
        cr=""
fWord+=cr
return fWord
def removeSpecialCr(fWord):#This function removes the special characters from the definiendum
   npWord=""
for cr in fWord:
```

```
if cr in "@#%$^&*()/~;}]":
        cr=""
npWord+=cr
return npWord
def definiendum(f):#This function identifies defiendum in the corpus
   Pd=[]
phr=''
   Def=[]
global Disc
   Disc=[]
   D=[]
   L=[]
words=[]
   #print(f)
for line in f:
     L=line.split(None)
for i in range(len(L)):
        #print(len(L))
if L[i]=='jechuun':
if int(L.index("jechuun"))==2:
phr=L[i-2] + '-' + L[i-1]
D.append(phr)
Disc.append(line)
else:
              f=int(L.index(L[i]))
st=stem(L[f-1])
D.append(st)
Disc.append(line)
if L[i]=='yeroo-ibsamu':
            f=int(L.index(L[i]))
```

```
st=stem(L[f-1])
D.append(st)
Disc.append(line)
if L[i]=='yeroo-hiikamu':
        f=int(L.index(L[i]))
st=stem(L[f-1])
D.append(st)
Disc.append(line)
if L[i]=='yemmuu-hiikamu':
        f=int(L.index(L[i]))
st=stem(L[f-1])
D.append(st)
Disc.append(line)
if L[i]=='yemmuu-ibsamu':
        f=int(L.index(L[i]))
st=stem(L[f-1])
D.append(st)
Disc.append(line)
else:
continue
return D
def getDefinitionWord(q):#This function extracts defiendum from the users' question
compWord=''
qr=''
if q[-1:]=='?':
qr=q[:-1]
else:
qr=q
words=qr.split()
for word in words:
```

```
if 'jecha' in words:
      f=int(words.index('jecha'))
      d=stem(words[f+1])
elif 'jechuun' in words:
if int(words.index("jechuun"))==2:
compWord=words[0] + '-' + words[1]
         d=compWord
         #print (compWord)
else:
         d=stem(words[0])
elif 'ibsi' in words:
if int(words.index('ibsi'))== 0:
         d=stem(words[1])
else:
       f=int(words.index('ibsi'))
       d=stem(words[f-1])
elif 'maal' in words:
      f=int(words.index('maal'))
       d=stem(words[f-1])
elif 'maali' in words:
       f=int(words.index('maali'))
       d=(words[f-1])
elif 'hiikni' in words:
       f=int(words.index('hiikni'))
       d=stem(words[f-1])
elif 'hiiki' in words:
       f=int(words.index('hiiki'))
      d=stem(words[f-1])
elif 'jechi' in words:
      f=int(words.index('jechi'))
```

```
      d=stem(words[1])
else:
word=q.split()
if len(word)== 2:
        d= word[0] + '-'+ word[1]
else:
      d=stem(q)
return d
#declaration of affixes
cluster1=['ittii','ttii','tii','irra','rra','dha']
cluster2=['olii','oolii','ota','oota','oolee','olee','icha','ichi','fis','siis','siif','fam','ata','ooma','oma']
cluster3=['ti','t','tee','te','tuu','tu','nu','nna','na','nne','ne','nnu','dhaaf',
       'dhaa','chaaf','tiif','ach','adh','chuu','at','att','ch','tanuu','tanu','tani','tan']
cluster4=['du','di','dan','lee','wwan','een','an']
cluster5=['aa','uu','ee','a','e','u','s','suu','sii','sa','se','si','ssi','sse','ssa','nye','nya']
cluster6=['eenya','ina','annoo','umsa','ummaa','insa','am','ni','affaa','offaa']
cluster7=['a','e','i','o','u','f','n']
def measure(unsteemd_word):#This function counts the number of CV(Consonant Vowel) Sequence
   voul_con=0
   Cons_con=0
   Cons="c"
   Seq_vc=""
   temp1=""
   temp2=""
   temp3=""
measure=0
vowel=['a','e','i','o','u']
for s in unsteemd_word:
if s not in vowel:
        Seq_vc=Seq_vc+"c"
```

```
else:
        Seq_vc=Seq_vc+"v"
   temp1=Seq_vc[:1]
for i in Seq_vc:
if i=="c" and temp1[-1:]=="c" and temp3[-1:]!="c":
        temp3= temp3 + "c"
elif i=="c" and temp3[-1:]=="c":
        temp3=temp3[:-1]+ "c"
elif i=="c" and temp3[-1:]=="v":
        temp3=temp3 + "c"
elif i=="v" and temp3[-1:]=="c":
        temp3=temp3 + "v"
elif i=="v" and temp3[-1:]=="v":
        temp3=temp3[:-1]+ "v"
elif i=="v" and temp1[-1:]=="v":
        temp3=temp3 + "v"
for s in temp3:
if s=="c":
        Cons_con=Cons_con+1
else:
        voul_con=voul_con+1
if (voul_con==1 and Cons_con==1) or (voul_con==2 and
     Cons_con==1)or (voul_con==1 and Cons_con==0) or(voul_con==0 and Cons_con==0):
measure=0
     #print measure
return measure
if Cons_con==2 and voul_con==1:
measure=1
     #print (measure)
return measure
```

```
if Cons_con==2 and voul_con==2:
measure=2
    #print (measure)
return measure
else:
measure=3
    #print (measure)
return measure

def stem(token):#stemming of words begin here
for i in range(4):
for i in cluster1:
measure(token)
if measure(token)>=1:
if token.endswith(i):
stem=token.rstrip(i)
measure(stem)
if measure(stem)>=1:
token=stem
else:
token=token
break
else:
token=token
for i in cluster2:
measure(token)
if measure(token)>=1:
if token.endswith(i):
stem=token.rstrip(i)
measure(stem)
```

```
if measure(stem)>=1:
token=stem
else:
token=token
break
else:
token=token
for i in cluster3:
measure(token)
if measure(token)>=1:
if token.endswith(i):
stem=token.rstrip(i)
measure(stem)
if measure(stem)>=1:
token=stem
else:
token=token
break
else:
token=token
for i in cluster4:
measure(token)
if measure(token)>=1:
if token.endswith(i):
stem=token.rstrip(i)
measure(stem)
if measure(stem)>=1:
token=stem
else:
token=token
```

```
break
else:
token=token
for i in cluster5:
measure(token)
if measure(token)>=1:
if token.endswith(i):
stem=token.rstrip(i)
measure(stem)
if measure(stem)>=1:
token=stem
else:
token=token
break
else:
token=token
for i in cluster6:
measure(token)
if measure(token)>=1:
if token.endswith(i):
stem=token.rstrip(i)
measure(stem)
if measure(stem)>=1:
token=stem
else:
token=token
break
else:
token=token
for i in cluster7:
```

```
measure(token)
if measure(token)>=1:
if token.endswith(i):
stem=token.rstrip(i)
measure(stem)
if measure(stem)>=1:
token=stem
else:
token=token
break
else:
token=token
        #token=stem
   #print (token)
return token
if os.path.isfile('D:\corpus4\DefIndex.txt'):
os.remove('D:\corpus4\DefIndex.txt')
defn=open('D:\corpus4\DefIndex.txt','a')
docname=[]
docname=os.listdir("D:\corpus4")
corp="D:\corpus4\ "
cop=corp.strip(None)
for i in range(len(docname)):
filename=cop+docname[i]
   f=openfiles(filename)
dff=definiendum(f)
for i in range(len(dff)):
defn.write(dff[i].lower())
defn.write('#')
defn.write(Disc[i].lower())
```

```
defn.close()
Disc=[]
f=openfiles('D:\corpus4\DefIndex.txt')#Calling function to open the file
fr=f.readlines()
unq=[]
for line in fr:
   #sy2.append(stem(sy[1].strip('\n').lower()))
line.strip('\n')
if line not in unq:
unq.append(line)
def ExtractAnswer(df):#This function extract the answer from the corpus.
sentence=[]
disc=[]
dff =[]
ans=[]
   #sy=[]
   sy1=[]
   sy2=[]
op=[]
   s3=[]
gl=[]
syn=openfiles('D:\corpus3\synonyms.txt')#This  function manipulates synonyms
   #print(syn)
sys=syn.readlines()
   #sys.lower()
for line in sys:
sy=line.split('=')
      #print(sy)
if len(sy)>1:
sy1.append(stem(sy[0].lower()))
```

```
sy2.append(stem(sy[1].strip('\n').lower()))
if len(sy1)>=1:
for i in range(len(sy1)):
if df==sy1[i]:
op.append(sy2[i])
elif df==sy2[i]:
op.append(sy1[i])
   #os.remove('D:\corpus2\DefIndex.txt')
   f=openfiles('D:\corpus4\DefIndex.txt')#Calling function to open the file
fr=f.readlines()
ddf=[]
for line in fr:
line.lower()
      D=line.split('#')
dff.append(D[0])
disc.append(D[1])
for i in range(len(dff)):
if df==dff[i]:
if disc[i] not in ans:

ans.append(disc[i])
else:
continue
for s in range(len(op)):
for i in range(len(dff)):
if op[s]==dff[i]:
if disc[i] not in ans:
ans.append(disc[i])
else:
continue
```

```
return ans
qr=input("Gaaffii Keessan Galcha:")
qr.lower()
dfw=getDefinitionWord(qr)
pfw=removePunct(dfw)
spcf=removeSpecialCr(pfw)
#stemedWord=stem(spcf)
w=[]
d=[]
count=0
docname=[]
l=[]
docname=os.listdir("D:\corpus4")
corp="D:\corpus4\ "
cop=corp.strip(None)
rans = {}
ar={}
ans=ExtractAnswer(spcf)
for a in range(len(ans)):
count=0
for d in range(len(docname)):
filename=cop+docname[d]
    f=openfiles(filename)
fr=f.readlines()
for line in fr:
if ans[a]== line:
count=count+1
else:
continue
ar[ans[a]]=count
```

```
w.append(count)
if len(ans)==0:
print("Deebii hin qabu")
else:
rans=sorted(ar.items(),key=operator.itemgetter(1),reverse=True)
for a in rans:
print(a[0])
```

APPENDIX 2: LIST OF PREPARED QUESTIONS FROM THE LEGAL DOCUMENT CORPUS USED FOR EVALUATION

1. Malaamaltummaa jechuun maali?

2.Ibsi malaamaltummaa maal?

3.Jecha malaamaltummaa jedhu ibsi.

4. Ibsi seera maali?

5.Seera jechuun maal jechuu dha?

6.Faayidaa jechuun akkamitti ibsama?

7.Hojimaata dhaddachaa jechuun maal?

8. Jecha madaallii raawwii jedhu ibsi.

9.Hiikni labsii maal?

10.Labsii hiiki.

11. Jechi caffee jedhu maal agarsiisa?

12.Jecha koree jedhu hiiki.

13.Waajira hiiki.

14. Jechi iyyannoo jedhu akkamitti ibsama?

15. Jechi garee jedhu maal agarsiisa?

16.Mana mare jechuun maa ibsa?

17.Dawoo seeraa jechuun maal agarsiisa?

18.Jechi mummichaa jedhu akkamitti ibsama?

19. Jecha maatii jedhu hiiki.

20.Walkaadhimmachuu yoo jedhamu maal ibsa?

21.Jecha mindaa jedhu akkamitti ibsama?

22.Ragaa yaalaa jechuun maal ibsa?

23.Ibsi Ministeera maal?

24.Hiikni Magaalaa maal?

25.Bulchiinsa magaalaa yeroo jedhamu maal ibsa?

26.Qabeenya xifaafiree jechuun maal?

27.Komishiinara jechuun maali?

28.Jecha kantiibaa jedhu ibsi.

29. Kantiibaa maal agarsiisa?

30. Jecha raggaasuu jedhu hiiki.

31. Jechi baasii jedhu maal ibsa?

32.Kaffaltii jechuun akkamitti ibsama?

33. Jallisii aadaa jechuun maal?

34 Faayidaa ummataa hiiki.

35.Seera ibsi.

36.Jechi jiraataa jechu maal ibsa?

37. Hiikni Jiraataa maali?

38. Jechi Labsii jedhu maal ibsa?

39.lafa manaa jechuun maali?

40. Hooggansa dhaddachaa jechuun maal?

APPENDIX 3: EVALUATION RESULTS

3.1. RESULT WITHOUT STEMMING

S.No	V	N	A	AW	L	L-AW	Precision	Recall	F-measure
1	1	0	1	100	139	39	0.72	1	0.837
2	2	1	3	300	97	-203	1	0.67	0.802
3	1	0	1	100	83	-17	1	1	1
4	1	0	2	100	89	-11	1	0.5	0.667
5	1	0	2	100	94	-6	1	0.5	0.667
6	0	2	1	200	103	-97	1	0	0
7	1	0	2	100	46	-54	1	0.5	0.667
8	0	0	2	0	0	0	0	0	0
9	1	2	2	200	160	-40	1	0.5	0.667
10	1	0	2	100	72	-28	1	1	1
11	2	0	3	200	184	-16	1	0.67	0.802
12	0	0	1	0	0	0	0	0	0
13	1	1	2	200	167	-33	1	0.5	0.667
14	1	0	2	100	70	-30	1	0.5	0.667
15	1	0	1	100	55	-44	1	1	1
16	1	2	1	200	211	11	0.95	0.5	0.327
17	2	0	3	200	116	-84	1	0.67	0.802
18	1	0	1	100	90	-10	1	1	1
19	2	0	2	200	80	-120	1	1	1
20	1	1	2	200	120	-80	1	0.5	0.667
21	0	0	2	0	0	0	0	0	0
22	2	0	2	200	141	-59	1	0.67	0.802
23	1	0	1	100	241	-41	0.58	1	0.734
24	1	1	2	200	76	-124	1	0.5	0.667
25	1	1	2	200	132	-68	1	0.5	0.667
Average Weight							**0.85**	**0.5872**	**0.6945**

3.2. RESULT WITH STEMMING

S.No	V	N	A	AW	L	L-AW	Precision	Recall	F-measure
1	1	0	1	100	139	39	0.72	1	0.837
2	3	1	3	300	97	-203	1	1	1
3	1	0	1	100	83	-17	1	1	1
4	2	0	2	200	89	-111	1	1	1
5	2	0	2	200	226	26	0.11	1	0.206
6	1	1	1	200	103	-97	1	1	1
7	1	0	2	100	46	-54	1	0.5	0.667
8	2	0	2	100	146	5	1	1	1
9	1	2	2	200	160	-40	1	0.5	0.667
10	2	0	2	100	72	-28	1	1	1
11	3	0	3	200	184	-16	1	1	1
12	0	0	1	0	0	0	0	0	0
13	1	1	2	300	167	-33	1	0.5	0.667
14	1	0	2	200	70	-30	1	0.5	0.667
15	1	0	1	100	55	-44	1	1	1
16	1	2	1	200	211	11	0.95	1	0.974
17	3	0	3	200	116	-84	1	1	1
18	1	0	1	100	90	-10	1	1	1
19	2	0	2	200	80	-120	1	1	1
20	1	1	2	200	120	-80	1	0.5	0.667
21	0	0	2	0	0	0	0	0	0
22	1	1	2	200	141	-59	1	0.5	0.667
23	1	0	1	100	241	-41	0.58	1	0.734
24	1	1	2	200	76	-124	1	0.5	0.667
25	1	1	2	200	132	-68	1	0.5	0.667
Average Weight							**0.855**	**0.76**	**0.804**

3.3. RESULT WITH STEMMING AND SYNONYMS APPLICATION

S.No	V	N	A	AW	L	L-AW	Precision	Recall	F-measure
1	1	0	1	100	96	39	1	1	1
2	3	1	3	300	97	-203	1	1	1
3	1	0	1	100	83	-17	1	1	1
4	2	0	2	200	89	-111	1	1	1
5	0	0	2	0	0	0	0	0	0
6	1	1	1	200	103	-97	1	1	1
7	1	0	2	100	46	-54	1	0.5	0.667
8	2	0	2	100	146	5	1	1	1
9	1	2	2	200	160	-40	1	0.5	0.667
10	2	0	2	100	72	-28	1	1	1
11	3	0	3	200	184	-16	1	1	1
12	0	0	1	0	0	0	0	0	0
13	1	1	1	300	167	-33	1	1	1
14	1	0	1	200	70	-30	1	1	1
15	1	0	1	100	55	-44	1	1	1
16	1	2	1	200	89	11	1	1	1
17	3	0	3	200	116	-84	1	1	1
18	1	0	1	100	90	-10	1	1	1
19	2	0	2	200	80	-120	1	1	1
20	1	1	2	200	120	-80	1	0.5	0.667
21	0	0	2	0	0	0	0	0	0
22	1	1	2	200	141	-59	1	0.5	0.667
23	0	0	1	0	0	0	0	0	0
24	1	1	1	200	76	-124	1	1	1
25	1	1	1	200	132	-68	1	1	1
Average Weight							**0.864**	**0.78**	**0.819**

Where

V=Vital nuggets retrieved from the system

N=Non vital nuggets ret retrieved from the system

A=Concepts in the assessors list

AW=Allowance

L=Length of characters in the retrieved vital nuggets without white characters