Addis Ababa University
College of Natural sciences

TIGRIGNA QUESTION ANSWERING SYSTEM FOR FACTOID QUESTIONS

KIBROM HAFTU AMARE

A THESIS SUBMITTED TO THE DEPART MENT OF COMPUTER SCIENCE IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN COMPUTER SCIENCE

ADDIS ABABA, ETHIOPIA

17 JUNE 2016

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

KIBROM HAFTU AMARE

ADVISOR: YAREGAL ASSABIE (PhD)

This is to certify that the thesis prepared by **KIBROM HAFTU AMARE**, titled: ***TIGRIGNA QUESTION ANSWERING SYSTEM FOR FACTOID QUESTIONS*** and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining committee:

	Name	**Signature**	**Date**
Advisor:			
Examiner:			
Examiner:			

Abstract

Accessing relevant information is one of the major problems faced by Tigrigna language users for every domain of knowledge when dealing with huge amount of information especially in the Internet. Evidently, users are interested in obtaining a specific and precise answer to a specific question. However, obtaining a relevant and concise answer is a challenge to particular user question. For such situation, Tigrigna Question Answering system is a good solution.

The proposed QA system comprises of question analysis, document analysis and answer extraction modules. The main function of question analysis module is taking a Tigrigna Question as input and then generates a query, expands a query and determines its Question Particle and Question Type. A statistical language model approach is used to model the classification of Tigrigna questions to their category or type. The document analysis module performs the process of pre-processing of parallel corpora, which are documents that contain question sentences in one document and answer sentences in another one, and also ranking and extracting answer contents. Answer extraction also performs the detail analysis on the retrieved answer contents based on the question type, question particle and query using the techniques of language modeling called Answer Model. This statistical language model does the extraction process of exact and precise Tigrigna answer in probabilistic manner from sets candidate answers.

Generally, this system developed after reviewed literatures and related work, and selected the appropriate tools and data source such as Moses, GIZA++ and IRSTLM as tools and different Webs and Tigrigna newspapers and magazines as data sources. Our data sets are classified for training and testing activities of the system. Based on this, we collected around 1000 data sets for training and 200 data sets for testing. Performance evaluation conducted manually by comparing the system‟s answers with the answers exists in testing document, which is prepared for testing purpose. Finally the evaluation results of Tigrigna factoid QAS is expressed in terms of the average performance of a question type classifier which is 87%, and the average Precision, Recall and F – measure of the answer extraction, precision is 88.5%, recall is 85.9% and F – measure is 87.2%.

Keywords: Tigrigna question answering, Tigrigna Factoid questions, Language model based question classification, question analysis, Document Analysis, Answer Extraction.

Dedication

To my mother, she tries to make my life comfortable in different situation always. It is because of you that I am today;

Acknowledgements

First and foremost, I would like to thank God and St. Mary for their blessing and giving me the courage and wisdom to finish this thesis.

I especially want to thank my advisor Dr. Yaregal Assabie, whose support and guidance made my thesis work possible. He has been actively interested in my work and has always been available to advise me. I am very grateful for his patience; motivation, enthusiasm, and immense knowledge.

I would also like to thank who were involved in manually preparing training questions of Tigrigna Question answering parallel corpora for this thesis: Abrha G/kiros, Berhe Tareke, Mulatu Dagnachew, Feyisa Gemechu and Seid Hassen. Without their passionate participation and input, especially in translating previously done corpora of Amharic question answering systems by Seid Muhie and Desalegn Abebaw, could not have been successfully conducted.

Fana Broadcasting corporate, Mekalih Tigray, Dimtsi weyanie Tigray radio and Ethiopian Broadcasting corporate, thank you for your cooperation in supplying me the news articles used in the preparation of corpus.

Finally, I would like to thank my family for their continuous love and respect, support and encouragement to make my dream become real especially to my mother Tekeu Biru. Thank you. God bless you!

Table of Contents

List of Figures

List of Tables

Acronyms and Abbreviations

CLEF	Cross-Language Evaluation Forum
FAQ	Frequently Asked Questions
EM	Expectation Maximization
IE	Information Extraction
IR	Information Retrieval
NER	Named Entity Recognizer
NERC	Named Entity Recognizer Classification
NIST	National Institute of Standards
NLP	Natural Language
NLP	Natural Language Processing
LM	Language Model
QA	Question Answering
QAS	Question Answering System
SMT	Statistical Machine Translation
SVM	Support Vector Machine
TQA	Tigrigna Question Answering
TREC	Text REtrieval Conference
VSM	Vector Space Model
UC	UNIX Consultant

Chapter One: Introduction

1.1 Background

Automatic knowledge discovery and information retrieval are becoming more and more essential mainly in facts when we are dealing with a huge amount of information especially in the Internet. Accessing to the relevant information is one of the major problems faced by the user practically for every domain of knowledge. Particularly, the user lacks time to find a short and precise answer to a query among the variety of available documents. Therefore, precision in retrieving the accurate information is crucial and a challenging task [5].

Information Retrieval is the art and science of retrieving information from a collection of documents that are relevant to the user based on its query [2]. In the case of search engines, a user can ask natural language questions then the results that they return are usually sections of documents which may or may not contain the answer but which do have many words in common with the given question by the user. Evidently, the user is interested in obtaining a specific and precise answer to a specific question. But, there is a challenge capable of obtaining a relevant and concise answer. For such situation, Question Answering systems are a good solution in dealing with this challenge [2].

Question Answering (QA) system in Natural Language Processing (NLP) is a man machine communication system that provides correct responses to the user‟s questions in short and accurate manner. When we compare standard document retrieval systems and question answering system, standard document retrieval systems which just return relevant documents to a user query but a QAS has to respond with a specific answer to Natural Language (NL) query mean that traditionally IR concentrates on finding whole documents while QAS tries to provide only one or a small set of specific answers to an input question [2].This is due to question answering system requires a much deeper understanding and processing of text than most web search engines[3].

The historical development of question answering system passed different stages in different time. Two of the most famous QA systems are BASEBALL and LUNAR which were developed in the 1960s. BASEBALL answered questions about the US baseball league over a period of one year. LUNAR, also answered questions about the geological analysis of rocks returned by the

Apollo moon missions. Both those QA systems were very effective in their chosen domains. In fact, LUNAR was demonstrated at a lunar science convention in 1971 and it was able to answer 90% of the questions in its domain posed by people untrained on the system. The common feature of both these systems is that they had a core database or knowledge system that was hand-written by experts of the chosen domain [4].

The 1970s and 1980s saw the development of comprehensive theories in computational linguistics, which led to the development of ambitious projects in text comprehension and question answering. One example of such a system was the Unix Consultant (UC), a system that answered questions pertaining to the UNIX operating system [3]. The system had a comprehensive hand-crafted knowledge base of its domain, and it aimed at phrasing the answer to accommodate various types of users. Another project was LILOG, a text-understanding system that operated on the domain of tourism information in a German city. The systems developed in the UC and LILOG projects helped the development of theories on computational linguistics and reasoning [3].

In the late 1990s the annual Text Retrieval Conference (TREC) included a question-answering track which has been running until the present. Systems participating in this competition were expected to answer questions on any topic by searching a corpus of text that varied from year to year. This competition fostered research and development in open-domain text-based question answering. The best system of the 2004 competition achieved 77% correct fact-based questions [3].

In 2007, the annual TREC included a blog data corpus for question answering. The blog data corpus contained both "clean" English as well as noisy text that include badly-formed English. The introduction of noisy text moved the question answering to a more realistic setting. Real-life data is inherently noisy as people are less careful when writing in spontaneous media like blogs. In earlier years, the TREC data corpus consisted of only newswire data that was very clean [6].

An increasing number of systems include the World Wide Web as one more corpus of text. Currently, there is an increasing interest in the integration of question answering with web search. Ask.com is an early example of such a system, and Google and Microsoft have started to integrate question-answering facilities in their search engines [6].

In different question answering systems there are different question categories. As a result, it requires different strategies to answer for a particular question. Questions are normally asked using interrogatives in different forms such as [2, 7, 8, 9]:

- Factoids are those for which the answer is a single fact. Simple interrogative questions await an answer related to a named entity.
- List questions are factoid questions that require more than one answer.
- Definition type questions require word meaning, term definition, and description of term and so on.
- Descriptive questions require a more complex answer, usually constructed from multiple source documents.

1.2 Motivation

In today's Internet IR systems, users can submit a set of keywords, which represent their information needs. These queries are then processed by the system and a sorted list of relevant documents is returned. Such systems are also referred to as *Document Retrieval* systems [2]. For example, if a user wants to know **"ናይ መጀመሪያ ዩንቨርስቲ ኢትዮጵያ/the first university of Ethiopia",** he/she probably searches the Web using **"መጀመሪያ", "ዩንቨርስቲ"**, and**"ኢትዮጵያ"** as keyword query. Afterwards, the resulting documents have to find the desired answer for such queries by matching those words with words or phrases in the documents. However, if it was possible to directly ask the system by using "**ናይ መጀመሪያ ዩንቨርስቲ ኢትዮጵያ ማ ይብሃል? /What is the first university of Ethiopia?"**In this case, the answer of the system to this natural language question should not be a set of documents but the concise answer "**ኣዲስ ኣባባ ዩንቨርስቲ"** based on **"ማ ይብሃል"** pattern that refers the entity respect to the **organization** of "**Addis Ababa University"**.

The choice of factoid questions versus other types of questions is motivated by the following factors [12].

- A considerable percentage of the questions actually submitted to a search engine are factoid questions [12]. Current search engines are only able to return links to full-length documents rather than brief document fragments that answer the user's question.
- The frequent occurrence of factoid questions in daily usage.

1.3 Statement of the Problem

Tigrigna is an Afro-Asiatic language belonging to the Semitic family's branch, designated official language of Tigray Region in Ethiopia and the national language of Eritrea [11]. As a result, availability of Tigrigna language textual information is highly increasing from time to time in general. Literature work, a number of newspapers, magazine, educational resources, official credentials and religious documents have been written in the language, that information can be found electronically on different places online and offline as sources of information in the world [10].

Since the emergence of the idea of QAS, lots of researches have been done in different languages and showed good results. Few attempts have been made to develop a QA for local languages such as Amharic language [1, 13, 14] and Afaan Oromo [15]. Both Tigrigna and Amharic languages are Semitic languages and use the same Geez script called "Fidel (ፊደል)" however; they have different linguistic properties such as character representation and pronunciations, and sentence and question construction etc., such a difference shows that those system developed for Amharic language can not be used for Tigrigna language because they have language dependency . So, to the best knowledge of the researcher, nothing has been done for Tigrigna. In this research work, we would try to develop Tigrigna Question - Answering system for fact-based questions seeking answers related to different types of entities such as person, place, organization, time, quantity, etc.

1.4 Objectives

General Objective

The general objective of this research work is to develop **Tigrigna Question Answering System** for factoid type questions.

Specific Objectives

The specific objectives of this research are:

- Understand the grammatical structure of questions and statements of **Tigrigna** language related to factoid question types and question answering.
- Propose architezcture for **Tigrigna** Factoid Question answering system.
- Developing **Tigrigna** Question Answering corpus.
- Developing a language model for **Tigrigna** Factoid Question Answering system.
- Training the training data sets and performing analysis over Tigrigna questions and answers prepared for testing as testing data set.
- Develop a prototype.
- Evaluating the performance of the system.

1.5 Methods

Literature Review

Literature review has a vital role for identifying problems, finding gaps, identifying methodologies, etc. In addition, in order to understand the linguistic of Tigrigna language and state-of-the-art in question answering, books, articles, journals and other publications will be reviewed.

Tools

In order to achieve the research objectives, a number of tools and methods or approaches will needed such as Moses, GIZA++ and IRSTLM.

Data Sources

Question- Answer Tigrigna corpora will be collected from the Web and Tigrigna newspapers and magazines.

Testing

Performance evaluation will be conducted manually by comparing the system,s answers with their corresponding manual answers of the questions prepared as a document containing question and answer pairs for testing. That is, the evaluation metrics like precision and recall will be used for the evaluation.

1.6 Scope and Limitations

A fully developed QA system will require a number of natural language processing techniques and tools such as Sentences parser, Chunker, Part of Speech (POS) tagger, Stemmer, Named Entity Recognizer (NER) and so on. Even though some of the NLP tools have been developed by some researchers, they are not publicly available to use and integrating with our system. Having these limitations in mind, our scope would be:

- Answering only "መን (who)", "ኣበይ (where)", "መዓዝ (when)", "ክንደይ (how many)" types of Tigrigna Questions.
- Tigrigna News collected from different newspapers and magazines or free text used as a source of questions for training and testing purpose.

1.7 Application of Results

As QA is an extension to search engines that provides short and precise answer to a given natural language question, this system can be employed in retrieving short answers in Tigrigna language for factoid type of questions from collection of documents.

A fully developed Tigrigna Question Answering System would have a great contribution in different real-world applications such as automated customer services, suggests directions in driving system, Tigrigna E-learning by providing correct answers to students, reservation system and giving online help in the absence of information desk personnel by providing information about an organization automatically.

1.8 Thesis Organization

The remaining parts of this thesis are organized as follows. Chapter Two presents literature review. Chapter Three presents critical reviews related question answering works for factoid based questions of different languages and approaches they used. It presents the gaps in the reviewed researches. It also describes the proposed solution to bridge the gaps. Chapter Four explains a detailed description of architectural design of this proposed QA modules, components, subcomponents and then it describes architectural design along with its implementation. Chapter Five presents the evaluation mechanism or criteria, and the corresponding empirical results of the proposed system along with their interpretations. Finally, Chapter Six concludes the thesis with the research contributions, conclusions, and future works.

Chapter Two: Literature Review

2.1 Introduction

In this chapter, we concentrate on addressing Question Answering system development strategies. The first section presents the details of question answering system from Information Retrieval and Natural Language Processing perspective. The second section will cover the dimensions that determine question answering systems in terms of the level of understanding and reasoning, type of data they used and whether the system is domain specific or domain independent. The third section covers descriptions on general QA architectures. The fourth section explains the details on QA components. The remaining sections describe about Tigrigna Language, grammatical features of Tigrigna language, kinds of sentences in Tigrigna language and finally described about Tigrigna Factoid questions.

2.2 Question Answering System

Question Answering (QA) is commonly defined as either a type of Information Retrieval (IR) because it usually deals with large quantities of information or a subfield of Natural Language Processing (NLP) because it is mostly concerned with the interpretation of much smaller pieces of texts [7]. Answers are required to provide the correct information for questions that are requested by human being. Generally, Question Answering System (QAS) is an interaction of human and computer system that includes understanding a user information need, retrieving relevant documents from collection of documents, extracting and ranking available answers and finally providing short answers to the user‟s question [15].

2.3 Dimensions to Question Answering Systems

The problem of question answering can be described according to a number of different dimensions. Roughly, those dimensions can be divided into level of understanding and reasoning, type of data they use, and whether the system is domain specific or domain independent [9].

Level of understanding: systems can be distinguished by their overall purpose and level of information need and understanding [9]. Under this dimension, question types can be used to categorize a QAS, because they require different strategies to deal with them. From those,

Factoid Question Answering systems are the simplest and return their answers based on named entities. Three types of entities are distinguished to be recognized and categorized [12]:

ENAMEX: detection and classification of proper names and acronyms. The classes considered in this subtask are:

- Organization: named corporate, governmental or other organizational
- Person: named person or family
- Location: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)

TIMEX: Detection and classification of temporal expressions. The classes considered in this subtask are:

- Date: complete or partial date expression.
- Time: complete or partial expression of time of a day.

NUMEX: Detection and classification of numeric expressions monetary expressions and percentages. The classes considered in this subtask are:

- Money: monetary expression
- Percent: percentage

Based on the level of processing applied to the questions and documents, a system can be either shallow or deep system. Shallow QAS uses keyword based or question reformulation techniques to locate interesting passages and sentences from the retrieved documents. However, Deep QAS uses more sophisticated syntactic, semantic and contextual processing to extract or construct the answer of a question such as Named Entity Recognition and Classification (NERC) [19].

Type of data: question answering requires a data source which may be a single or a collection of documents. The type of data in those documents can be distinguished as structured data (e.g. databases), semi-structured data (e.g. comment fields in databases) and free text (e.g. news wire collections) [4].

Databases are the most popular answer sources that store structured data. Structured Query Language (SQL) is used to retrieve data from databases. LUNAR developed to answer Natural Language (NL) questions about the geological analysis of rocks returned by the Apollo Moon Missions is an example of such a database system. The performance of this system was excellent in terms of accuracy achieved [20].

Frequently Asked Questions (FAQ) are another answer resource in various commercial/business customer service and those only able to focus on processing input questions and matching them with FAQs. Like other systems they don"t require question analysis and answer generation stages. For an input question, if an appropriate FAQ is found, then using lookup table method corresponding answer is retrieved [21].

Question answering systems are nowadays almost exclusively concerned with free text data. TREC focuses on a fixed collection of news articles, however many systems participating in the contest are also capable of using the entire web as a knowledge source. Web QA uses search engines like Google, Yahoo and Alta-Vista etc. to retrieve web pages that contain answers to the questions. Some systems combine the web information with other answer resources to achieve better QA performance [3, 5]. AQUAINT corpus used in TREC QA Track consists of newswire text data drawn from three sources [22].

Generality: we can also make a distinction between systems that are either domain specific (closed -domain) or domain independent (open-domain).

Open domain Question Answering System is a research area of Natural Language Processing aimed at providing a convenient and natural interface for accessing information. It deals with questions about nearly everything [16]. TREC is concerned with open-domain systems, i.e., systems that attempt to answer whichever question a user wants to know the answer [5]. These systems usually have much more data available from which to extract the answer and then to answer unrestricted questions, world knowledge would be useful. ASKJEEVES is the most well-known open domain QA System [17].

Closed domain Question Answering Systems deal with questions under a specific domain (for example, medicine or automotive maintenance) and can be seen as an easier task because NLP systems can exploit domain specific knowledge frequently formalized [16]. Closed domain refers to a situation where only limited types of questions are accepted and correct answers to a question may often be found in only very few documents since the system does not have large retrieval set. BASEBALL system is a restricted domain QA System that only answers questions about the US baseball league over a period of one year [18].

2.4 General Architecture of QA system

The generic question answering system has four main components namely question analysis, document retrieval, document analysis, and answer selection. The question analysis component is used to analysis question structure and it‟s semantic. Depending on the class of the question, a query is formulated which is posed to the document retrieval component. Some information such as the question class and a syntactic analysis of the question, are also sent to the document analysis component [6].

The document retrieval component is generally a standard document retrieval system which identifies documents that contain terms from a given query and returns a set or ranked list of documents that are further analyzed by the document analysis component. The document analysis component takes documents as input that are likely to contain an answer to the original question, together with a specification of what types of phrases should count as correct answers. The document analysis component extracts a number of candidate answers which are sent to the answer selection component. The answer selection component selects the phrase that is most likely to be a correct answer from a number of phrases of the appropriate type. It returns finally the precise and concise answer to the user [8].

Figure 2.1, shows the main components of the generic architecture of QAS [6, 8, 15].

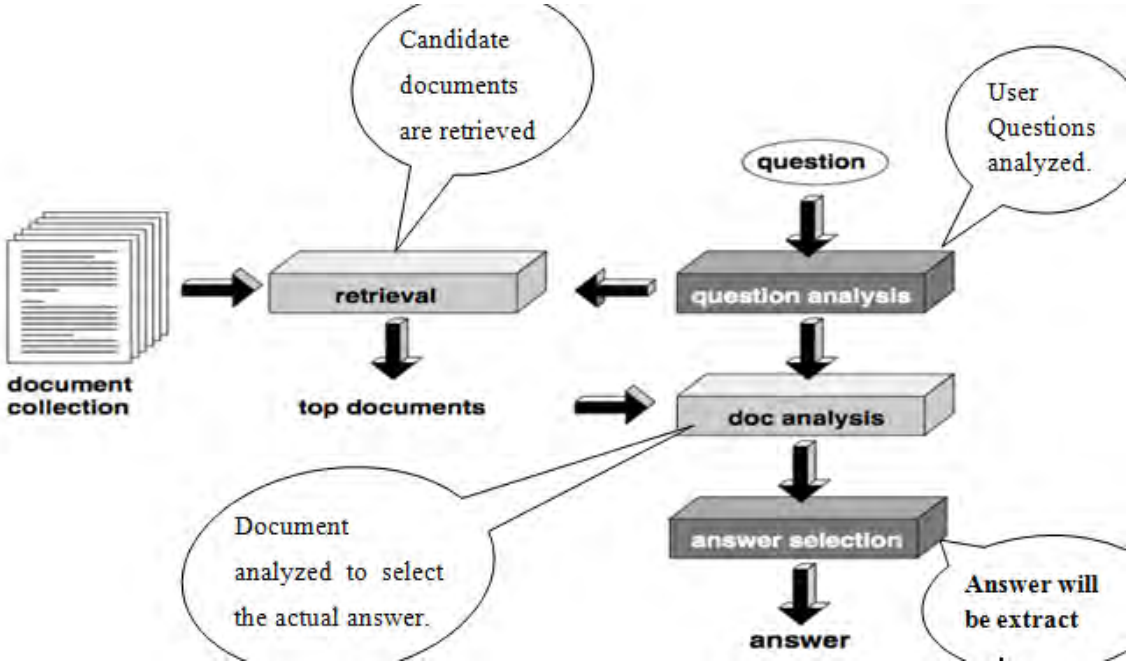


Figure 2 1 Summary for the main components of the generic architecture of QAS

2.5 Tasks in Question Answering System Components

Most QA systems perform different tasks inside their components. Those tasks can be grouped into tasks inside question processing, document processing and answer processing [4]. The tasks performed in question processing consists of tasks such as construction of question, derivation of expected answer type and keyword extraction. Additionally, question reformulation also performed [5, 9]. User question may have different forms. However, the need to understand the semantics of the question is essential and it termed as Question Processing [6].

Different tasks performed in Document processing typically include keyword expansion, document retrieval and passage identification. Keyword expansion involves taking the keywords extracted while question processing and looking them up in a thesaurus, or other resource and adding similar search terms in order to fetch as many relevant documents as possible[5].

Answer processing consists of components such as candidate answer identification, answer ranking and answer formulation tasks. Identifying candidate answers means taking the results from identified passages and then ranks according to a ranking algorithm or set of heuristics [4, 9]. Finally, it formulates specific answers.

2.6 Question Classification and Expected Answer Type

Question classification techniques help to answer a user question correctly by identifying user‟s need. Those techniques simplify the searching of an answer by giving clues such as predicts the type of the answer based on the provided user‟s query [14, 15]. The possible entity type for the factoid questions types are: Person, Location, Organization, Date and Quantity depend on question particles of the question. Table 2.1 Shows the Tigrigna factoid question particles and expected answer type.

Table 2 1 Question Particles and Expected Answer Type.

S.no	Question particle		Expected answer type
1	“...መን/ ንመን/ መነመን/ እንመን /መንንመን/ ናይመን/ ካብመን/ ብመን	Who	Person or organization
2	“...ኣበይ/ በይን/እስካብ ኣበይ/ካበይ / ናበይ/ ኣየነይቲ/ ኣየናይ/ ኣየኒኣም/ ኣበየኒኣም/ በበይ/ ካበየናይ/ በየናይ?”	Where	Place
3	“...መዓዝ/መኣዝ/እስካብ መዓዝ/ክሳብ መዓዝ/ እስካብ መኣዝ/ክሳብ መኣዝ/ካብ መኣዝ/ካብ መዓዝ/ንመኣዝ/ንመዓዝ/ ብመኣዝ/ ብመዓዝ /መዓዝ ግዜ / መኣዝ ግዜ / መኣዝ ጊዜ/ መዓዝ ጊዜ?”	When	Time and Date
4	“...ክንደይ/ ክንደይ ዝኣክል/ ንክንደይ/ ብክንደይ /ካብ ክንደይ/ ክሳብ ክንደይ / ኣብ ክንደይ / እስካብ ክንደይ / ብምንታይ ዝኣክል / ንመበል / ብመበል/ ክክንደይ?”	How many	Number or Quantity

Classification of a question can be categorized into coarse and fine for simplicity and low confusion at the moment of categorization. It will be done in two steps; in the first step the question will belong to one of the sixth coarse groups then with attention to coarse group, each coarse fine will belong to its sub-category. The coarse classes are as follows: abbreviation (ABBR), entity, description (DESC), person, location and numeric [27]. The abbreviation (ABBR) category consists of two subcategories. One subcategory concerns how acronyms

should be expanded (ABBR: expansion) and the other concerns how to abbreviate a given term (ABBR: abbreviation).

The entity category handles questions that ask for a specific object that fits a description, e.g., “ከንደይ ዝኣኽሉ ቋንቋታት ኣብ ኢትዮጵያ ይዝረቡ?” /“How many languages were spoken in Ethiopia?” This can belong to (Entity: language) or “ኢትዮጵያ ብዋናነት ናብ ኣውሮፓ ዝትልእኾ ምህርቲ ሕርሻ እንታይ እዩ?” /“What agricultural product does Ethiopia export to Europe mainly?” this also to (entity: product). There are 22 subcategories to this coarse class.

The DESC category is concerned with questions that ask for more elaborate answers. The category consists of the following subcategories: definition (DESC: definition), description (DESC: description), manner (DESC: manner) and reason (DESC: reason). Definitions basically refer to definitions of terms and concepts, for example “ዲሞክራሲ እንታይ ማለት እዩ? / What does the term democracy mean?” the description category covers questions like “ኮምፕዩተር ብከመይ ይሰርሕ?/ how does a computer work ?” that needs an elaborate factual description of a term or event or process. Manner refers to questions like “ግብሪ ምኽፋል ብኸመይ ግዴታ ይኸውን? /How does a bill become law?” that asks for a description of a process. Finally, the reason category covers all why-questions. These are perhaps the most difficult questions to answer. Question answering system for those question types has a goal to make inferences by itself from different sources of information and present to user the answer as well as the inference chain.

The person category covers questions relating to specific humans or human organizations. The individual (person: individual) covers questions that ask for a specific person that fits a given description such as “ቴሌፎን ወይድማ ስልኪ ዝመሃዘ መን እዩ? /who invented a telephone?” The group category (person: group) is concerned with questions where the answer is a group or organizations of people, such as a company. There is also a title category (Person: title) for questions that ask for a person‟s profession or title, example “ናይ ግብፂ ፕሬዚዳንት መን ይበሃሉ? / Who is the president of Egypt?” and there is a description category (person: description) are questions that request information about a person, such as ጥላሁን ገሰሰ መን እዩ? /Who is “Tlahun Gesese?”

The locations or place class covers geographic and virtual locations. Subcategories that cover geographic locations can be: city (location: city), country (location: country), mountain (location: mountain) and state (location: state). A fifth category (location: other) covers other geographic (e.g. planets and rivers) and also virtual locations (e.g. web addresses). For example ርእሰ ከተማ

ኢትዮጵያ መን ትብሃል?/ what is the capital city of Ethiopia?" "ካብ ዓለም እቲ ዝዓበየ ጎቦ ኣበይ ይርከብ?"/where do you found the largest mountain in the world?"

Finally, the numeric coarse class covers questions that request for some kind of numeric information such as dates, prices, ages and speed. There are 13 subcategories, 12 concerning specific numeric information and one category for those that do not fit the other 12 (number: other) example "ዘመናዊ ኦሊምፒክ መዓዝ ተጀመሩ? / When was a modern Olympic started? ". Table 2.2 shows lists of coarses and fines of classification of aquestion.

Table 2 2 List of coarses and fines of classification of a question

Coarse	**Fine**
ኣሕፅረተ-ቃል (ABBR)	ኣሕፅረተ-ቃል (Abbreviation), ትንትና (expansion)
መግለፂ (DESC)	ትረጉም (Definition), ገለፃ (description), ኣገባብ (manner), ምኽንያት (reason)
ኣካል (ENTITY)	እንስሳ (Animal) , ኣካል (body), ሕብሪ (color), ምህዞ (creation), ሸርፊ(currency), ሕማም/ሕክምና (disease/medical), ፍፃሜ (event), ምግቢ (food), መሳሪሒ (instrument), ቋንቋ (language), ፊደል (letter), ተኽሊ (plant), ምህርቲ (product), ሃይማኖት (religion), ስፖርት (sport), ነገረ (substance), ምልክት (symbol), ሜላ (technique), ፍታሕ(term), መጓዓዝያ (vehicle), ቃል (word), ካሊእ (other)
ወዲሰብ (PERSON)	ገለፃ (description), ጉጅለ (group), ሕድሕድ (individual), ማዕረግ (title)
ቦታ (LOCATION or Place)	ከተማ(city), ሃገረ(country), ጎቦ(mountain), ክልል(state) ክፍለክተማ (sub-city) ዞባ (zone), ወረዳ (wereda) ካሊእ (other)
ቑፅሪ (NUMBER)	ኮድ (code), ፈቀደ(count), ዕለት(date), ርሕቀት (distance), ገንዘብ (money), ደረጃ (order), ሚእታዊ (percent), እዋን (period), ፍጥነት (speed), መቐት (temperature), ዓቐን (size), ክብደት (weight) , ካሊእ (other)

2.7 Approaches to Question Answering System

The three main approaches to classify a question answering system can be: rule-based, machine learning based and language modeling based.

Rule based approach; hand-written grammar rules and a set of regular expressions are employed to parse a question and to determine the answer type [14]. This approach determines a question type based on the sentence pattern, which includes the interrogative word, certain sequences of words and some representative terms of particular question classes. Rules can be constructed manually to automatically classify questions and then does not require hand labeled training data. However, this approach has limitations; it is difficult to include all the possible patterns of the language in the rules, based on handcrafted rules that require linguistic knowledge, has human intervention and expensive to modify or maintain and extend [15].

Machine Learning Based, the type of the answer is predicted after the machine is trained with a training data set. However, it needs more training data in order to provide best results because it automatically learn and improve with experience. Here, learning means recognizing and understanding the input data and making wise decisions based on the supplied data [15]. Machine learning only focuses on predictive modeling. But, language model focuses on all aspect of data-analysis such as descriptive, exploratory, inferential, predictive and causal.

Language Model Based

A language model assumes human language generation as a random process and its aim is to represent that process by a statistical model to predict the next word by using context of the previous word. In another word, a language model is an estimate of words occurrence probabilities [2, 16]. This approach can be used in speech recognition, question answering and other natural language processing with lists of advantages such as novel way of looking problems of text retrieval based on probabilistic manner which is conceptually simple and explanatory, and formal mathematical model. In question answering task, a language model used to classify a question performs by developing a class to each questions based on the training data set. The probability of generating a question class is calculated for each class based on its language model and the highest probability determines the classification [15].

2.8 Language modeling for Question Answering

QA extends IR techniques to return a concrete answer to a question instead of references to full documents which are relevant to a query. QA attracted the attention of researchers for some years and several public evaluations have been recently carried in the TREC, CLEF, and NTCIR conferences using different approaches [30]. In a language model approach, to retrieve the correct answer, different techniques could be needed such as Statistical Machine Translation (SMT), which is a technique used to bridge the lexical gap between questions and answers of factoid QA. In such situation, the answer can be treated as a translation of the question using translation models by creating Question-to-Answer pair. Thus, the best translation for a given source sentence is the most probable one and the probability of each translation is given by the Bayes theorem [28].

In this case, the source sentence corresponds to the question Q and the target or translation is the sentence containing the answer A. With this correspondence, the fundamental equation of SMT can be written as:

$$A(Q) = argmax\, P\left(\frac{A}{Q}\right) \tag{2.7}$$

$$= argmax\, P\left(\frac{Q}{A}\right) P(A) \tag{2.8}$$

Where P (Q /A) is the translation model and P (A) is the language model and each of them can be understood as the sum of the probabilities for each of the segments or phrases that conform the sentence. The translation model quantifies the appropriateness of each segment of Q being answered by A. The language model is a measure of the fluency of the answer sentence and does not take into account which is the question. Since in identifying the concrete string that answers the question and not a full sentence, this probability is not as important as it is in the translation problem. The rationale of the process is that the SMT model can learn the context where answers appear depending on the structure of the question.

Generally, a Question Answering System in this approach can be performed firstly, the question is analyzed with several linguistic processors and then relevant documents are obtained from the document collection with straightforward IR techniques with a mathematical model and a list of

candidate answers is generated. Finally, these candidate answers are filtered and ranked to obtain the final list of proposed answers [30].

2.9 Lingustic Prperties of Tigrigna Language

2.9.1 The Tigrigna Language

Tigrigna (ትግርኛ) is an Afro-Asiatic language; belonging to the Semetic branch of language family with writing system is called Ge"ez script. Tigrigna has over 200 Characters that have different sound. The normal syllable of Tigrigna considered as a consonant followed by a vowel then those consonants are written in seven slightly different forms corresponding to the seven vowels. The ways that vowels are applied to all the consonants could be shown in Table 2.3 [11].

Table 2 3 Ge'ez script of Tigrigna Language

Name	ግዕዝ	ካዕብ	ሣልስ	ራብዕ	ሓምስ	ሳድስ	ሳብዕ
Order	1 st	2nd	3 rd	4 th	5 th	6 th	7 th
Transliteration	E	U	I	A	Ie		O
ሀ	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
	He	Hu	Hi	Ha	Hie	H	Ho
ለ	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
	Le	L**u**	Li	La	Lie	L	Lo
መ	መ	መ	ሚ	ማ	ሜ	ም	ሞ
	Me	Mu	Mi	Ma	Mie	M	Mo

Tigrigna also has word endings that vary according to the gender of the person you are speaking to. Thus, there are two grammatical genders called masculine and feminine, all nouns belong to either one or the other and then the distinguishing could be performed by their endings, for instance with the feminine signaled by *t* such as ከፈተ *kefete* 'open', ከፋቲ *kefati* 'opener (m.)', ከፋቲት *kefatit* 'opener (f.)' and ትግራዋይ *tigraway* 'Tigrean (m.)', ትግራወይቲ *tigraweyti* 'Tigrean (f.)'. Gender in Tigrigna normally agrees with biological gender for people and animals; thus nouns such as አቦ *"ʤbo* 'father', ወዲ *wedi* 'son, boy', and ብዕራይ *biəray* 'ox' are masculine, while nouns such as ኣደ *Ade* 'mother', ጓል *gual* 'daughter, girl', and ላም *lam* or ላሕሚ *lah.mi* 'cow' are feminine. However, most names for animals do not specify biological gender, and the words ተባዕታይ

teba,ɟay 'male' and ኣንስተይቲ *ansteyti* must be placed before the nouns if the gender is to be indicated. The gender of most inanimate nouns is not predictable from the form or the meaning, grammars sometimes disagree on the genders of particular nouns; for example, ጸሓይ *ṣeḥay* 'sun' is masculine.

Tigrigna has a singular and plural number. However, nouns that refer to multiple entities are not obligatorily plural. That is, if the context is clear, a formally singular noun may refer to multiple entities: ሓሙሽተ *ḥamushite* 'five', ሰብኣይ *seb'ay* 'man', ሓሙሽተ ሰብኣይ *ḥamushite sebut*, 'five men'. It is also possible for a formally singular noun to appear together with plural agreement on adjectives or verbs: ብዙሓት *bzuḥat* 'many (pl.)', ዓዲ *,,ɑdi* 'village'; ብዙሓት ዓዲ *bzuḥat ,ɑdi* 'many villages'. Noun plurals are formed both through the addition of suffixes to the singular form ("external" plural) and through the modification of the pattern of vowels within the consonants that make up the noun root ("internal" or "broken" plural).

Table 2 4 Numbers in Tigrigna Language

External plural		internal plural	
Singular	Plural	Singular	Plural
ዓራት *,ɑrat* 'bed'	ዓራታት(ዓራውቲ) *,ɑratat* 'beds	ፈረስ feres 'horse'	ኣፍራስ afras 'horses'
እምባ *əmba* 'mountain'	እምባታት *embatat* 'mountains'	ንህቢ *nhbi* 'bee',	*ኣናህብ anəhb* 'bees',
ጎይታ g oyita" master"	ጎይቶት goyita" masters"	ደርሆ *derho* 'chicken'	*ደርሁ derhu* 'chickens'
ሓረስታይ haresitay "farmer"	ሓረስቶት harestot "farmer"	መንበር *menber* 'chair',	*መናብር menabr* 'chair'
ገዛ geza "house"	ገዛውቲ gezawti "houses"	ቀላቢ *kelabi* 'feeder',	*ቀለብቲ kelebti* 'feeder's'

2.9.2 Grammatical Features of Tigrigna Language

In most languages, there are a small number of basic distinctions of person, number, and often gender that play a role within the grammar of the language. Tigrigna has its own basic grammatical features. Some of the grammatical features of Tigrigna language are described as follows [35, 36].

Noun

In Tigrigna, noun is treated as either masculine or feminine. However, most inanimate nouns do not have a fixed gender. It also has plural as well as singular forms, though the plural is not

obligatory when the linguistic or pragmatic context makes the number clear and the way forming through internal changes ("broken" plural) as well as through the addition of suffixes. For example, ፈረስ feres 'horse', ኣፍራሰ *ʾafras* 'horses'.

Adjectives

The adjectives precede the noun or pronoun, which it modifies and agrees with it in gender and number. The degree of modifying the noun or pronoun of those adjectives could be expressed with agreement of genders and numbers they modify. For instance, ጽቡቕ t*sbuqʾ* 'good (m.sg.)', ጽብቕቲ t*sbqti* 'good (f.sg.)' *ጾቡቓት Tsbuqat* 'good (pl.)'.

Verbs

Tigrigna has a verb system, which contains two "tenses" but several "aspects" (causative, intensive, reflexive etc) and the apparent derivation of related words from "roots" consisting of three consonants: {*sbr*} 'break', ሰበረ *sebere* 'he broke', ይሰብር *ysebr* 'he breaks', ምስባር *msbar* 'to break'. The inflections of these systems give the tense and specify the person and number of subjects and objects. Due to this, Tigrigna has a default word order in clauses as subject–object–verb and noun modifiers usually precede their head nouns. As a result, the simplest form of a verb is generally the third person masculine, singular, simple perfect tense eg ሰበረ *sebere* he broke.

Prepositions

Tigrigna has both simple and compound prepositions, simple preposition of one radical preposition which is prefixed to nouns, pronouns and adjectives and also the compound preposition which consists of a simple preposition plus another word. Some main simple prepositions are presented in the Table 2.5.

Table 2 5 Prepositions in Tigrigna Language

Simple preposition	Meaning	Simple preposition	Meaning
ኣብ *ab*	'on, in, at'	ናብ *nab*	'to, toward'
ብ *b*	'with' (instrument), 'by' (means, agent), 'in' (duration)	ከም *kem*	'like, as'
ን *n*	'for (the benefit of), to the detriment of'	ብዘይ *bzey*	'without'
ናይ *nay*	'of'	ምእንቲ *m'ənti*, ስለ *sle*	'for, because of, on the part of'
ምስ *ms*	'with' (accompaniment)	ድሕሪ *dḥri*	'after'
ካብ *kab*	'from'	ቅድሚ *qdmi*	'before'
ብዛዕባ *biza,ba*	'about' (concerning)	ክሳዕ *ksa,,* ክሳብ *kisab*, ስጋዕ *səga,,*	'until'

Pronouns

In Tigrigna, there are basic distinctions of person, number, and gender within the basic set of personal pronouns. Tigrigna also makes a two-way distinction between near ('this, these') and far ('that, those') using demonstrative pronouns and adjectives. Besides singular and plural, it also distinguishes masculine and feminine gender. Table 2.6 and Table 2.7 depict personal pronouns and demonstrative pronouns in Tigrigna.

Table 2 6 Personal Pronouns in Tigrigna Language

Personal Pronoun			
Tigrigna	English	Number	Gender
ኣነ	I	Singular	
ንስኻ	You	Singular	Masculine
ንስኺ		Singular	Feminine
ንሱ	He	Singular	Masculine
ንሳ	She	Singular	Feminine
ንሕና	We	Plural	
ንስኻትኩም	You	Plural	Masculine
ንስኻትክን		Plural	Feminine
ንሳቶም	They	Plural	Masculine
ንሳተን		Plural	Feminine

Table 2 7 Demonstrative Pronouns in Tigrigna

Number	Gender	Near	Far
Singular	Masculine	እዚ Ezi	እቲ Eti
	Feminine	እዚኣ Ezi'a	እቲኣ Etiá
Plural	Masculine	እዚኣም / እዚኣቶም Ezieom , ezi'atom	እቲኣም / እቲኣቶም Etieom , eti'atom
	Feminine	እዚኤን / እዚኣተን Ezi'en , ezi' aten	እቲኤን / እቲኣተን Eti'en , eti' aten

2.9.3 Tigrigna Sentence Construction

Tigrigna sentence (**ሙሉእ ሓሳብ**) **may** consist of one clause (independent clause (ዋና ሓረግ)) or more clauses (independent and dependent clauses (ተስሓቢ ሓረግ)). On the basis of the number of clauses and types of clauses present in a Tigrigna sentence, the sentences are divided into four kinds [36]:

- **ነፀላ ሙሉእ ሓሳብ (Simple Sentence):** a sentence consists of only one independent clause containing a subject (በዓል ቤት) and a verb (ግስ) and it expresses complete thought. There is no dependent clause.

Examples: ንሱ ስሒቁ ነይሩ።/ He laughed. ሳቶም ይድቅሱ ኣለዉ። / they are sleeping.

መፅሓፍ ገዚአ እየ። / I bought a book.

- **ድርብ ሙሉእ ሓሳብ (Compound Sentence):** a sentence consists of at least two independent clauses joined by coordinating conjunctions. There is no dependent clause in a compound sentence. The coordinating conjunctions used to join independent clauses are "ስለ */and,* እምበር/ *but,* ወይ ድማ /or, እኳ ድኣ /yet, እምበኣር /so*"*.

  For example, ስለ ዝሓገዝኩዎ ብጣዕሚ ሕጉሰ እዩ። / I helped him and he became happy**.**

- **ሕውስዋስ ሙሉእ ሓሳብ (Complex Sentence)**: a sentence consists of one independent clause and at least one dependent clause **j**oined by subordinating conjunction (ስለ /*because, ካብ / since, ምስ/ when, ተዘይ/ unless etc*) or relative pronoun (ከም/*that, መን/who, ኣየናይ/which etc*).

  Examples: ቅደም ዝረድኣኒ ዝነበረ ወዲ ረኺበዮ።/ I met the boy who had helped me.

  ንሳ ሚሒር ደስ ዝብል ካናቲራ ተኸዲና። / She is wearing a shirt which looks nice.

- **ድርብ -ሕውስዋስ ሙሉእ ሓሳብ (compound-complex Sentence): a** sentence consists of at least two independents and one or more dependent clauses**.**

  Examples: ንሱ ናብ ኮሌጅ ምስ ከደ አነ መፅሓፍ ናብ ዝሸወጠሉ ዕዳጋ ብምኻድ መፅሓፍ ገዚአ።

  He went to college and I went to a market where I bout a book.

  ሒሳብ ይፈትው እየ ኾይኑ ግና ሓወይ ዶክተር ክኸውን ስለ ዝደሊ ስነ-ህይወት ይፈትው እዩ**::**

  I like Mathematics but my brother likes Biology because he wants to be a doctor.

Based on their **purpose,** Tigrigna sentences are distinguished into the following four types of the sentences:

- **ገላፂ ሙሉእ ሓሳብ/ declarative sentence:** are **t**hose sentences that make a statement and end with ":: /ኣርባዕተ ነጥቢ".
  - For instance: ፍቅሪ ዝኾነ መዓልቲ ::/ The day was lovely.
- **ሓተታዊ ሙሉእ ሓሳብ / interrogative sentence:** are those sentences that could be asked a question for obtaining new information. It constructed with the help of interrogative pronouns or adverbs and finally at the end, it has an interrogative mark. Tables 2.8 depict the interrogative pronouns and adverbs in Tigrigna language.

Table 2 8 interrogative pronouns and adverbs in Tigrigna Language

Tigrigna	English	Tigrigna	English
እንታይ	What	መዓዝ	When
ኣበይ	Where	መን	Who
ካበይ	From where	ምስመን	With whom
ናበይ	To where	ናይ መን	Of whom , whose
ንመን	To whom	ኣየናይ	Which
ክንደይ	How many	ከመይ	How
ብምንታይ	How, by what means	ንምንታይ	Why , for what

- **ኢጋናዊ ሙሉእ ሓሳብ / exclamatory sentence:** are those sentences that could use a more emotional version of a common statement and end with an exclamation mark "!".
  - For example: ምሒር ፍቅሪ ዝኾነት ማዕልቲ! / What a lovely day!
- **ትእዛዛዊ ሙሉእ ሓሳብ/ imperative sentence:** are **t**hose sentences that could be usually used to express a demand, request or call for action:
  - For example: ረስዓዮ። / let's forget. ንክትርድኣኒ ሞክር። / Do try to understand me.

2.9.4 Tigrigna Factoid Question

Tigrigna factoid question is a question that requires an answer which presents only a single fact using simple interrogative question particles such as "መን (who)" that refers to a person, "ኣበይ (where)" that refers to places, "መዓዝ (when)" for specific time and months etc.., "ክንደይ (how many)" to indentify quantities or amounts [37]. Table 2.9 depicts lists of Tigrigna factoid questions and their type.

Table 2 9 Tigrigna factoid questions

Question	Type	Particle
ፍራንሷ ኦላንድ ናይ መን ሃገር ፕሬዚዳንት ኒይሮም?	Person	ናይ መን
ደራሲ ስብሃት ገብረ እግዚኣብሄር ከዚሀ ዓለም ብሞት ዝተፈለየ ብክንደይ ዓመቱ እዩ?	Quantity	ብክንደይ
ናይ ሞጆ ደረቅ ወደብ መዓዝ ተጀሚሩ?	Time	መዓዝ
ፀጋዬ ገብረ መድህን ኣበይ ተወለዱ?	Place	ኣበይ

Chapter Three: Related Work

3.1 Question Answering Systems for Amharic

Seid muhie [13] described the Amharic QAS called "TETEYEQ" which is developed for Amharic language factoid question types. It has five main modules namely document preprocessing, question processing, document retrieval, sentence/paragraph re-ranking and answer selection modules. The tasks performed on those modules would be described as follow; firstly, the document preprocessing module performs normalization of documents. Secondly, the question processing module accepts the user"s question and performs tasks such as question type determination, question focus identification, and expected answer type determination. Thirdly, in document retrieval component relevant documents are retrieved. Fourthly, the sentence/paragraph re-ranking module detected a possible answer particle in the returned document based on Named Entity and pattern techniques. Lastly, the answer selection module selected the best top 5(five) answers from the previously ranked documents. Finally, the researchers used a rule based approach for classifying questions and extracting answer, and then they tested the performance of their question answering system and got promising results. However, it is Amharic language dependent and has limitations of rule based approaches such as it is based on handcrafted rules requires linguistic knowledge, it is expensive to maintain and extend and it depends on human skill to define rules rather than the inter-structure and semantics of the Amharic languages data.

Desalegn Abebaw [14] designed a web based question answering tool for Amharic factoid questions called "LETEYEQ". This QAS had components and developmental stages which perform different tasks related to linguistics and computational tasks. Those components were classified into three main components namely question analysis, document retrieval with a search engine component and answer extraction components. The question analysis tool was designed using a machine learning approach. Thus, the major task in this research is how collecting/preparing huge amount of training set in the form of factoid questions and answers that trained the machine in the question analysis phase. Then, questions are given in a search engine like interface in natural language form and the question type and the expected answer type should be identified by the question analyzer component. Next, the web documents having a potential answer to the question that retrieved by the document retrieval component. Finally, the

answer extractor component performed the extraction of appropriate answers from the candidates returned by the document retrieval component. In this work [14], the system‟s performance is measured and compared against the performance of the rule based Amharic question answering system [13] and then by applying a machine learning approach to the question analysis component, the researchers improved precision of the question classification task in question analysis whose accuracy impacted the accuracy of the entire system. However, it is Amharic language dependent and has limitations of machine learning based approach. This is because of the system analyze a query or document based on geometric and detail use of document length and term in document or collection documents through the dependency on structural and syntactical features than similarity in semantical representation of query and documents .

3.2 Question Answering System for Afaan Oromo

Aberash Tesfaye [15], described the Afaan Oromo QAS which is developed for factoid question types of Afaan Oromo language. There are four main components namely: document pre-processing, question analysis, document retrieval and answer extraction. In the document pre-processing component, pre-processing document is performed. In question analysis component, question classification and query generation tasks are performed. The question classification sub component uses either machine learning or rule based models to classify the user‟s questions into one of the question classes (person, place, number and time). The query generation subcomponent task removes the question particles and pre-processes the user question. The document retrieval component retrieves relevant documents which have similar features with the user generated query. Finally, the answer extraction component extractes the actual answer of the given natural language by using the expected answer type from the candidate documents. The final output of the answer extraction component is a list of ranked Afaan Oromo answers. The system has a better performance due to evaluation of the machine learning and rule based question classification done using percentage of evaluation metrics.

3.3 Question Answering System for English

Andreas Merkel [31] presented a language model approach to parts of a complete QA system. It included the processing of the natural language query as well as the retrieval of relevant documents, passages and sentences. In the query processing module, the system tried to understand the user‟s natural language questions. This was necessary in decisions making in other parts of the system. The document retrieval module acted as a filter. It reduced the amount of documents that the other components had to handle. In the passage retrieval step, all relevant documents from the previous component were split up into text passages. Then, a re–ranking was done to find the best matching snippets. A special case of the passage retrieval module was selecting just one sentence as a text passage. The result in this QA system is shown in [31]. The performance this system would be equally well or even better than systems developed by other approaches. This happens due to the use of fast statistical language models. Generally, they explained the system is efficient, effective and very flexible.

Cristina España-Bonet and Pere Comas [32] presented a QAS using Statistical Machine Translation-based approach. The provided answer is a translation of the question obtained from the SMT system by using the n-best translations of a given question to find similar sentences in the document collection that contains the real answer. In this case, the source sentence corresponds to the question Q and the target or translation is the sentence containing the answer A. This QA system [32] had a pipeline of three modules. In the first one, the question is analyzed and annotated with several linguistic processors. In the second one, relevant documents are obtained from the document collection with straightforward IR techniques and a list of candidate answers is generated. Finally, these candidate answers are filtered and ranked to obtain a final list of proposed answers and returned.

Similarly to researchs [31,32], our system using language approaches that will be consider question answering as a translation problem which concern questions are translated into their answers. These generated answers are matched against all the candidates obtained with the retrieval module. The candidates are then ranked according to their similarity with the n- best list of translations. However, our system has additional features able to expand queries with lists of synonyms and perform deeply analysis on retrieved sets of candidate answers in particular sentences and word or phrase of retrieved sentence.

3.4 Question Answering System for Japanese

Seungwoo Lee and Gary Geunbae [33] described the Japanese QAS called "SiteQ/J" which is developed for factoid question types of Japanese language. First stage of this system is a passage selection module. Each passage is of size three consecutive sentences. After performing a preliminary analysis, a passage selection algorithm is used for ranking all passages in each document and selects top N passages for further processing. Then, each passage is scored using the count of query terms, their occurrence in the passages, and their inverse document frequency. One of the important requirements for a QA System is to predict what type of answer the question requires: a person name, location or organization. Once answer type of a question is determined entities belonging to the answer type within the passages selected in the previous step are extracted. After extracting answer candidates, some of them are filtered out and the remaining answers are scored using a specific expression. Finally, they evaluated the performance of their question answering system and they expressed it has good performance.

3.5 Question Answering System for Portuguese

Carlos Amaral et al. [34] described the Portuguese QAS which is developed for factoid question types of Portuguese language. Once the question is submitted, it is categorized according to question typology and through an internal query a set of potentially relevant documents is retrieved. Each document contains a list of sentences which were assigned the same category as the questions. Sentences are weighted according to their semantic relevance and similarity with the question. Next, through specific answer patterns these sentences are again examined and the parts containing possible answers are extracted and weighted. Finally, a single answer is chosen among all candidates and then, they got better performance results.

3.6 Summary

Those Amharic QA systems return answers to Amharic factoid type of questions by extracting concise and precise answers. Even though both Tigrigna and Amharic languages are belongs to the same Semitic category of language. However, those languages have their own linguistic properties in construction of sentences and questions such as they use their own grammatical features and representations, and question particles. Thus, those Amharic QASs aren"t able to answer Tigrigna factoid questions because those systems are language dependents.

The reviewed related work factoid Afaan Oromo, English, Japanese, and Portuguese QA systems are language dependents as those Amharic QASs, i.e., only answer their respective language factoid question types by depend on their own linguistic properties, the particular question particles and question focuses. Generally those QA systems can,t answer Tigrigna factoid types questions. Thus, in this research we aim to develop Tigrigna QA system that can answer Tigrigna factoid types of questions.

Chapter Four: Design of Tigrigna Question Answering System

4.1 Introduction

The objective of this research work is to develop a Tigrigna question answering system that provides answers for Tigrigna factoid questions by performing analysis on Tigrigna language questions and their answers within parallel corpora as a source of information. This chapter presents the architecture and descriptions of its components and modules of this system.

4.2 Architecture of Tigrigna QAS

The architecture of Tigrigna Factoid Question Answering System is depicted in Figure 4.1. The architecture has three major modules: question analysis, document analysis and answer extraction. The detailed description of each module and components within module is given in the subsequent sections.

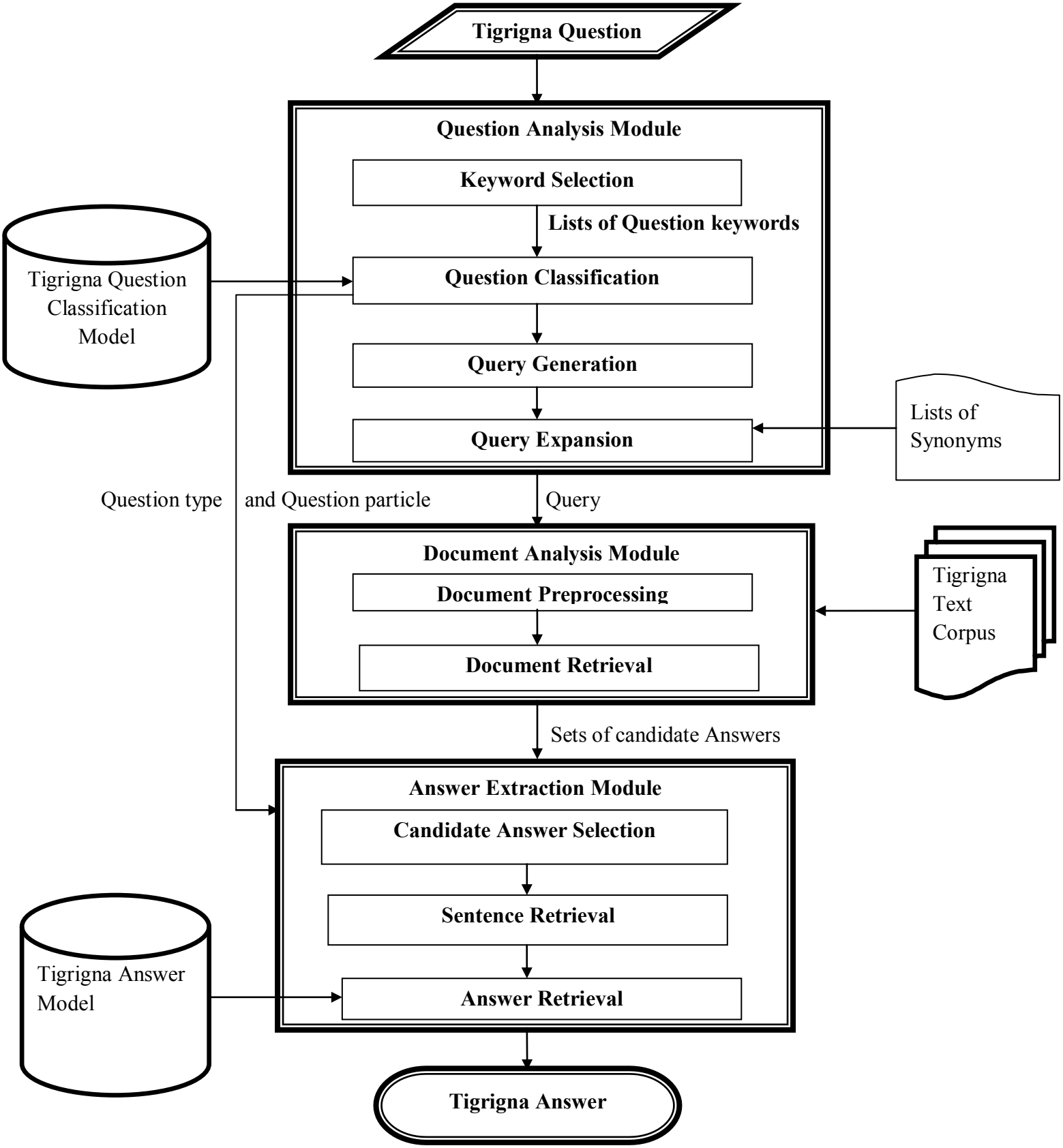


Figure 4 1 the Architectural Design of Tigrigna QA System

4.3 Question Analysis Module

Question Analysis is one of the fundamental tasks of Question Answering System and this is an initial step in the retrieval of relevant information. Before answering a Tigrigna question, we need to understand what type of information semantically it has because it is essential to determining question particles, question type and expected answer. This module includes components called Keyword Selection, Question Classification, Query Generation and Query Expansion and has a main function to take a Tigrigna Question as input and then convert it to a query via identifying and selecting keywords; next it generates and expands a query, determines question type and builds the semantic of the expected answer. The Keyword Selection component is used to receive and perform tasks such as identifying terms and selecting main terms for keywords that help the question classification component in determining the type of the question and question particle. This is an important step for extracting the actual answer through the query that query generator component generates from those keywords. Next, the generated query expanded by the query expander component and this module finally returns question particle, question type and Query that are used by the Answer Extraction Module to find out the answers relevant to the question.

Keyword Selector

This component analyses each and every word of the question as identified terms and then select keywords from those identified terms based on their probability through sequence of occurrences. Table 4.1 shows sample questions with the corresponding lists of question keywords.

Table 4 1 Sample Tigrigna question with the corresponding lists of question keywords

Question	Lists of Question keywords
ናይ ኢትዮጵያ ቀዳማይ ሚንስትር መን ይብሃል?/ who is the prime minister of Ethiopia?	ቀዳማይ ሚንስትር, ኢትዮጵያ/ prime minister: Ethiopia መን/who
ኣብ ጋና ናይ ኢትዮጵያ ኣምባሳደር መን እዮም?/ Who is the ambassador of Ethiopia in Ghana?	ኢትዮጵያ ኣምባሳደር, ጋና/ ambassador :Ethiopia, Ghana መን/who
ዋጋ ሓደ ኪሎ ስኳር ክንደይ እዩ? / How much does one kilo of sugar?	ስኳር/ sugar ክንደይ/ How much
ሓድሽ ዓመት ኢትዮጵያ መዓዝ ይኸበር? / When is the Ethiopian New Year holiday?	ሓድሽ ዓመት ኢትዮጵያ/ Ethiopian New Year መዓዝ/When

Question classification

The Question classification component is concerned with assigning questions to their category using question particles of the question and additionally it needs techniques to search similar or related question category by comparing the probabilities of keywords of the question with the corresponding question type it has. In this work, the possible question types of Tigrigna factoid questions are extracted based on the entity types they expresses such as Person, Location, Organization, Date and Quantity using the corresponding question particles of the given question as seen in chapter 2.

To make the process of question classification more dynamic and automatic, we use a language modeling, a statistical approach that had gained much attention recently in the IR and QA area. Thus, the model could be automatically constructed from the training set and its performance is competitive to other approaches. In our system, this is done by building language model for every class of questions based on the training data set to classify any question and then the probability would be generated. Finally, the highest probability determines the classification of the question to its category. The basic idea of language modeling behind the classification is that every piece of text or term can be viewed as being generated probability in the sentence leads to any category. Specifically, we built one language model for each category C of sample questions. When a new question Q comes, then by calculating the probability $P(Q|C)$ for each C and picks the one with the highest probability. The major advantage of language model over the other question classification techniques is its flexibility that can be automatically created and maintained. To summary the process, we used a statistical language modeling approach to model the classification of our questions to their category which is done automatically by training the system to understand the semantics of the training data sets and finally its performance is measured by preparing test data sets. Those training and test data sets are manually collected from the different Tigrigna news agency webs such as Mekalih Tigray, Dimtsi Weyane Tigray and Voice of America Tigrigna Program, some questions are also collected from Amharic news agency webs especially Fana broadcasting corporate and Ethiopian broadcasting corporate, then we used those articles by translating to Tigrigna sentences and the others are translated from the Amharic QA systems done before. For the training and testing tasks around 1000 Tigrigna factoid questions are prepared. Over training questions of each question types we developed a

language model and then every class C of the questions be presented by its question particles and question type via calculating the probability of a given question structure to the total given questions in training sets .

Generally, question classification is performed over training set by using the LM classifier that does its task as follow:

Step 1 *Collect Questions for training the system as training data sets.*

Step 2 *Aligned particular question to its question particle and type.*

Step 3 *Create a classifier language model to aligned questions.*

Step 4 *Train aligned particular questions and its question particle and type.*

Step 5 *Receive new Tigrigna question Q from the user.*

Step 6 *Calculate the probability of P (Q/CQ).*

Step 7 *Pick question type and question particle with the highest probability.*

Step 8 *Return the question type and question particle.*

Query Generation and Query Expansion

In this Question analysis task, extracting keywords helps us in classifying questions and understands the semantics of the question which is essential in query generation. The query generator acquires query after giving weight to those keywords of the question and then the highest weighted keywords are formed a query by merging each other. This generated query is later used as an input to query expansion component of this QAS. The goal of this query generation task is to create a query used in extraction of answers to a particular user question. For such task to identify the highest weighted keywords and formulating a query performed using language model.

The query expansion component gets query terms of query generated component as input and checks each term has a synonym word from the synonym word list, if so it appends the synonym

word to the query. We use a Statistical Machine Translation model for query expansion tasks and this is done by using a full-sentence paraphrase, which searches synonyms of context of the entire query or by translating each query terms into corresponding synonym term if it exists.

To this practically, firstly creating a translation model is essential for learning lexical correlations between words or phrases of the query with lists of synonym terms. After inserting space between words of the query to formulate a token, able to check whether each token has a synonym. The tokenization of query is essential in identifying keywords of the given query. Therefore, the process of query generation and expansion will be presented using the following steps.

Step 1 *Insert user Tigrigna question*

Step 2 *Tokenize each term and identify each terms.*

Step 3 *Get highest probable term or terms.*

Step 4 *Select query terms or keywords*

Step 5 *Check whether query terms or keywords have synonym*

If exist then check by full sentence paraphrasing or word/phrase level

Then return keywords with addition of the synonym (expansion query).

Else

Return the query terms or keywords that have generated as query

4.4 Document Analysis Module

Document Analysis module performs tasks such document pre-processing and retrieval of relevant contents from retrieved corpora which is done using document preprocessing and document analysis components.

Document Pre-processing

In this system, before retrieving documents as a process of QA activity a document pre-processing task is done over Tigrigna question answer corpora. Actually, the document pre-processing component included a task called tokenization that perform inside, helpful in

increasing the probability of getting similar terms within the query and the weighted answer sentences.

Tokenization breaks the stream of characters into raw terms or tokens. This process detects word boundaries of a written text. In Tigrigna, words can be separated using a white space, Tigrigna punctuation marks, or by new line. Therefore, by tokenization, we identify separate terms from parallel corpora. In this process, the parallel corpora pass the following steps:

- **Tokenization**: This means that spaces have to be inserted between (e.g.) words and punctuation.
- **Truecasing**: The initial words in each sentence are converted to their most probable casing. This helps reduce data sparsely.
- **Cleaning**: Long sentences and empty sentences are removed as they can cause problems and obviously misaligned question – answer pair sentences are removed. It performs by removing empty lines and redundant space characters.

Document Retrieval

The aim of document retrieval component is to retrieve relevant contents from Tigrigna corpora documents to a user‟s query. The relevancy of content is measured with the contents of corpora related to the query terms. Therefore, contents of those corpora which contain similar terms with a query terms are returned by the document retrieval part. Contents of corpora which do not bear any related particle is considered as non-relevant and is dropped. This means, the aim of question answering system is returning an exact answer to a user‟s question depends on the effectiveness of document retrieval. Generally, the document retrieval component does the ranking and extracting contents or candidate answer collections that exist in the corpora based on word overlap to the query using semantically interpretation.

Practically, the nature and content of documents preprocessing are stable after once preprocessed. However, before done those documents preprocessing tasks preparing parallel corpora is essential. It needs to train and test the system with the given Question sets and Answer sets by aligned at the sentence level as a source of information.

Preparing a Set of Question-Answer Pairs: our corpora contain around 1000 question- answer pairs for training as a training data sets as indicated the example of table 4.2.

Table 4 2 Sample question-answer pairs

No.	Question	Answer
1	ናይ ኢትዮጵያ አምባሳደር ኣብ አሜሪካ መን ይበሃሉ	አምባሳደር ግርማ Person
2	ናይ ዓለም ዋንጫ መዓዝ ተጀሚሩ	ብ 1959 Time
3	ፍራንሷ አላንድ ናይ መን ሃገር ፕሬዚዳንት ነይሮም	ፈረንሳይ Place
4	ከተማ ዓድዋ ካብ አዲስ አበባ ብክንደይ ኪሎ ሜትር ርሕቀት ትርከብ	1066 ኪሎ ሜትር Quantity
5	ሕቡራት መንግስታት ክንደይ አባል ሃገራት ኣለዉኡ	192 ሃገራት Quantity

After tokenization this parallel corpora, each Tigrigna words be separated using a white space. Thus, the tokenization process of the documents performed by inserting spaces between each Tigrigna words or terms and then those words or terms in each sentence are converted to their most probable casing. Then long sentences and empty sentences are removed. Such tokenization process performed with the help of Moses decoder to align a question – answer pairs in parallel line by line. The snapshoot seen below that taken after tokenization process of our corpora using Moses decoder.

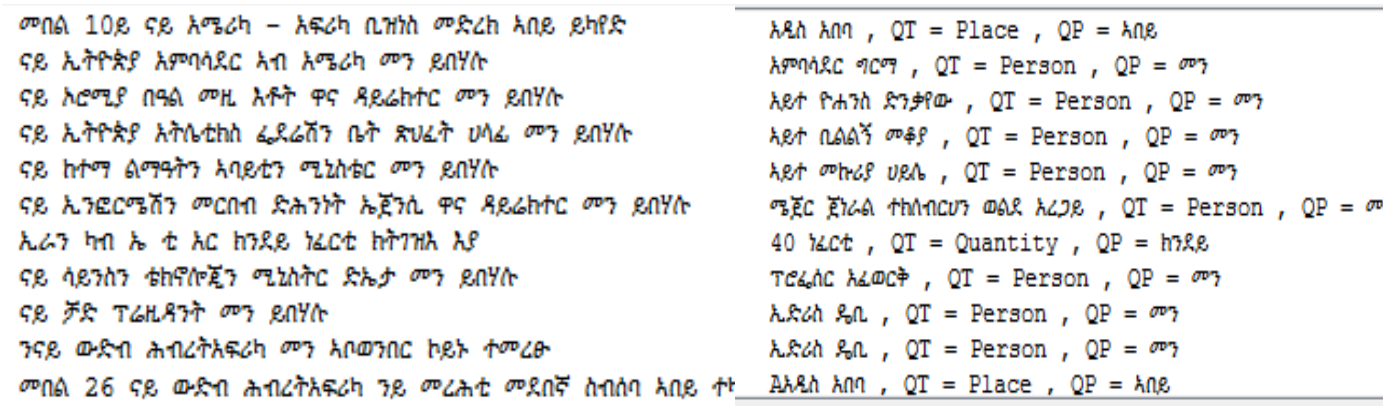


Figure 4 2 Corpus tokenization

Corpus truecasing model is used to model the probability of a given tokenized terms seen in query generated of a question. For example, the occurrence of sample words in the tokenized question expressed in numbers of occurrence as follow.

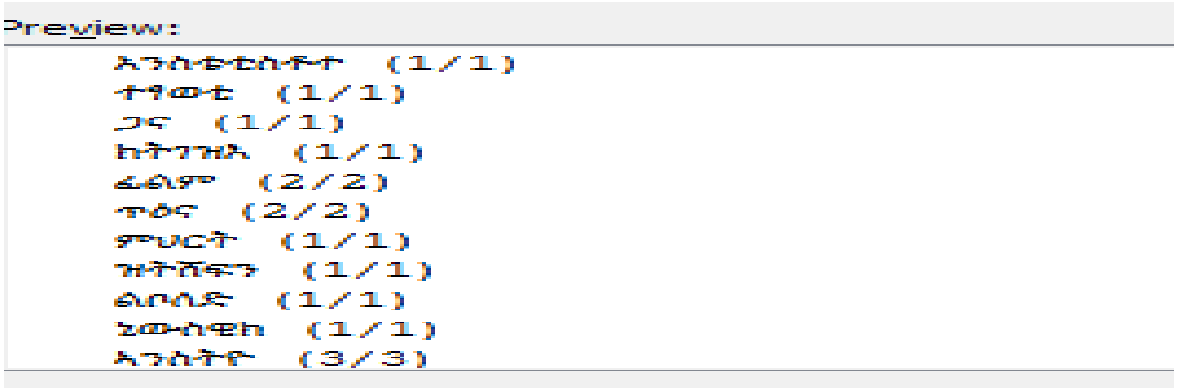


Figure 4 3 the occurrence of sample words in the tokenized question

The probability of a given tokenized answer as a target, it needs a model to estimate each word of an answer based on the numbers of occurrence in the corpora. For example, the occurrence of sample words in the tokenized answer.

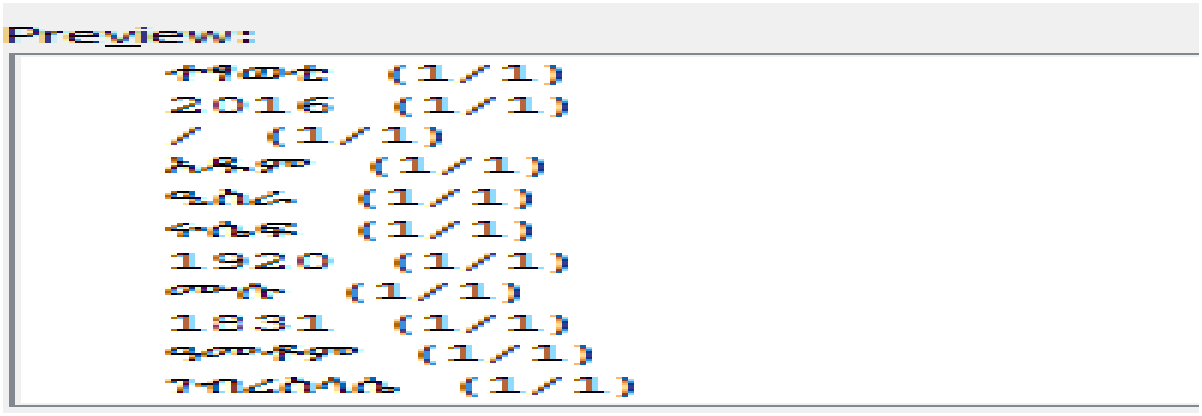


Figure 4 4 The occurrence of sample words in the tokenized answer.

Next, Moses decoder aligns the words exist in question and answer based on the probability calculated. Generally, the corresponding result is described the probability matching of question and answer terms as you seen below.

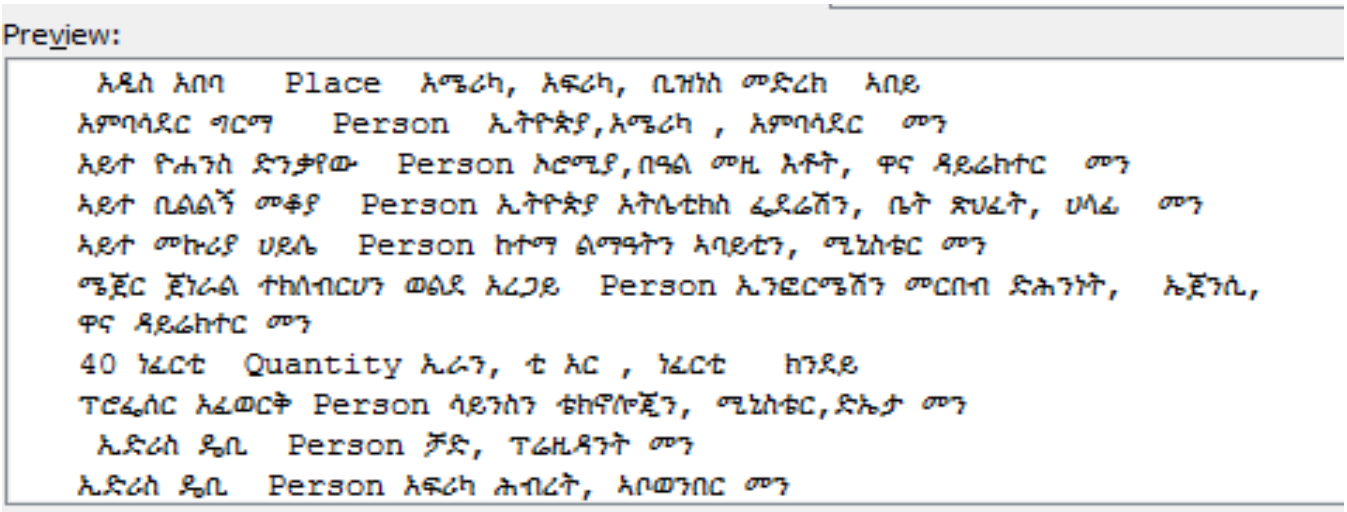


Figure 4 5 The lists of words exist in question and answer based on the probability.

Finally, this task returns the correctly aligned input sentences of question sentence and the corresponding output sentences of answer sentences as a result.

In our work, such sentence alignment and aligning word, phase in each sentence was created by Moses Decoder and GIZA++ program.

Table 4 3 Sample parallel corpora of Tigrigna questions and answers alignments.

Question	Answer
ናይ ውድብ ሑቡራት መንግስታት ዋና ፀሀፊ መን ይበሃሉ	ባን ኪ ሙን , QT=Person , QP= መን
መበል 20 ናይ ዓለም ዋንጫ ኣበይ ተኻይዱ	ብራዚል , QT=Place , QP= ኣበይ
ሳልቫ ኪር ናይ መን ሃገር ፕሬዚዳንት እዮም	ደቡብ ሱዳን , QT=Place , QP= ናይ መን
ዘመናዊ ኦሊምፒክ መዓዝ ተጀመሩ	ብ 1896 , QT=Time, QP= መዓዝ
ሕቡራት መንግስታት ክንደይ ኣባል ሃገራት ኣለዉኦ	192 ሃገራት , QT=Quantity , QP=ክንደይ

4.5 Answer Extraction

To identify relevant answer more accurately, the answer extraction module performs detailed analysis on the retrieved parallel Tigrigna corpora documents, the question type and question particle by extending the techniques of language modeling. Language models are essential in creating answer models in this part for ranking or selection of candidate answers that are retrieved from a given document by estimating the quantity, Q(S|P) where S is a candidate answer sentences and P, a given document. In this section, first we show how candidate answer sentences of a given

question type and question particle fit naturally using a probabilistic model in fetching relevant and precise sentence that contains the answer. Next, the ways to extract the Tigrigna answer from retrieved sentence. Generally, this is done via a simpler approach that extends the techniques of language modeling.

Candidate Answer Selection

Based on constructed query in Question Analysis module and information contents exist on Tigrigna parallel corpora retrieved from Document Analysis module, the answer extraction module extracts the answer. Ideally question type and question particle, lists of answer hypotheses are extracted. However, returning lists of answers is not explicitly performed.

Sentence Retrieval

In question answering systems, a concise and precise single answer sentence is essential than a collection of candidate answer sentences. This can be extracted by computing the highest probability over those given answers. So, our question answering system needs to have a component for the use of finer grained sentence retrieval.

Answer Retrieval

The extraction of possible answer from part of the sentence for factoid question answering is much more efficient to work on a smaller piece of texts than more than one sentence. This is because of factoid question needs short, exact and precise answering only. Hence, we need a narrower scope of data. The assumption here is that the relevant answer that we want to grab in a sentence is mostly stored in a single or few terms/ phrases. Due to this, our system has a component called Answer retrieval, to retrieve relevant answer from the retrieved sentence.

Practically, first we created an answer model using the help of IRSTLM with the Moses decoder and GIZA++ program integration. This answer model starts its work by focusing on training data sets. The second one is the answer model prepares the given targeted answers arrangement by specifying the starting and ending of the sentence using <s> answer sentence </s> then each answer sentence is added between the start and end symbol <s> </s> as seen in Figure 4.6.

Preview:

```
<s> አዲስ አበባ , QT = Place , QP = አበይ </s>
<s> አምባሳደር ግርማ , QT = Person , QP = ሙን </s>
<s> አይተ ዮሐንስ ድንቃየው , QT = Person , QP = ሙን </s>
<s> አይተ ቢልልኝ መቆያ , QT = Person , QP = ሙን </s>
<s> አይተ መኩሪያ ሀይሌ , QT = Person , QP = ሙን </s>
<s> ሜጀር ጀነራል ተከለጠርሀን ወልደ አረጋይ , QT = Person , QP = ሙን </s>
<s> 40 ኤርቲ , QT = Quantity , QP = ክንደይ </s>
<s> ፕሮፌሰር አፈወርቅ , QT = Person , QP = ሙን </s>
<s> ኢድሪስ ጄቢ , QT = Person , QP = ሙን </s>
<s> ኢድሪስ ጄቢ , QT = Person , QP = ሙን </s>
<s> ልአዲስ አበባ , QT = Place , QP = አበይ </s>
```

Figure 4 6 Creating answer model over sample answer sentence sentences.

After creating answer model in short and precise way, next we need to ensure the correct answer extraction by calculating the probability of answer sets of a given user question**.** The assumption on this task, creating answer model is used for rather than arranging a given question to the corresponding specific answer but to extract exact answer only which is done estimating probabilistic of answer words with the given Tigrigna question such as:

Question: መበል 10ይ ናይ አሜሪካ - አፍሪካ ቢዝነስ መድረክ አበይ ይካየድ

Answer alignment: <s> አዲስ አበባ , QT = Place , QP = አበይ </s>

Probabilistic value: each word given in a user question is compared with the given words as answer alignment then the corresponding result is calculated using the answer model as in Figure 4.7.

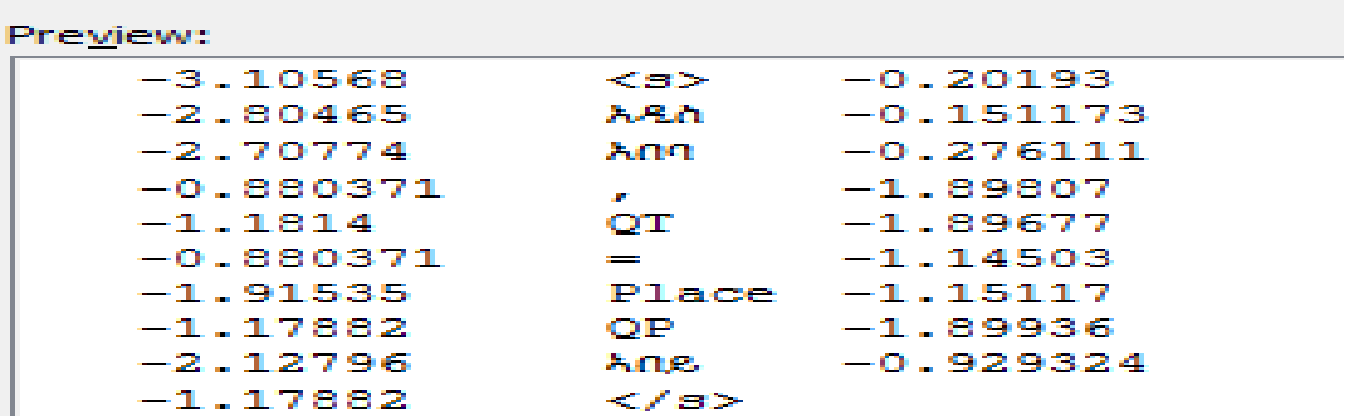


Figure 4 7 A Probabilistic value of words in question- answer alignment

Some steps performed in answer extraction process:

Step 1 *Get Tigrigna parallel corpora document from document analysis module and question particle, question type and query from question analysis module.*

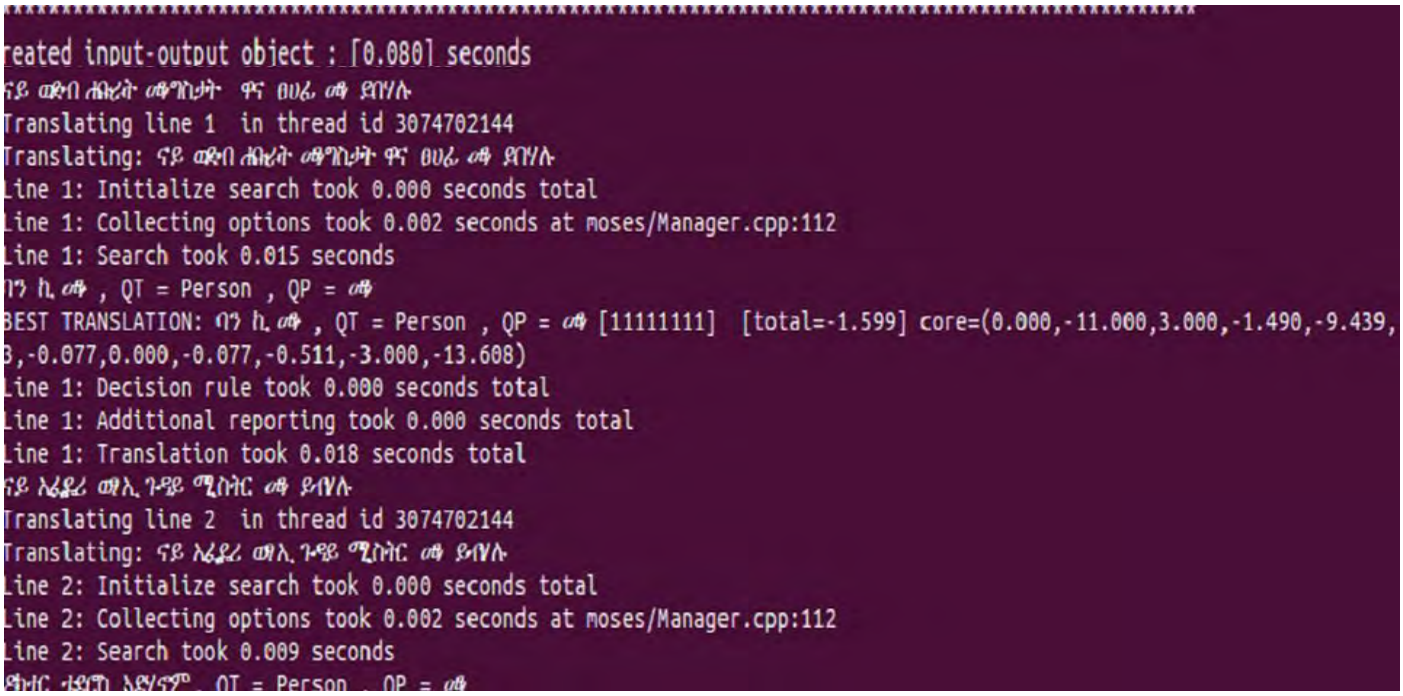
Step 2 *Perform calculating the probability of P(Q/CQ) to get candidate answers/*

Step 3 *Compute the score of answer words in each candidate answer sentences.*

Step 4 *Retrieve the top relevant answer sentence.*

Step 5 *Compute the most relevant term/phrase from the sentences.*

Step 6 *Return answer term /phrase.*



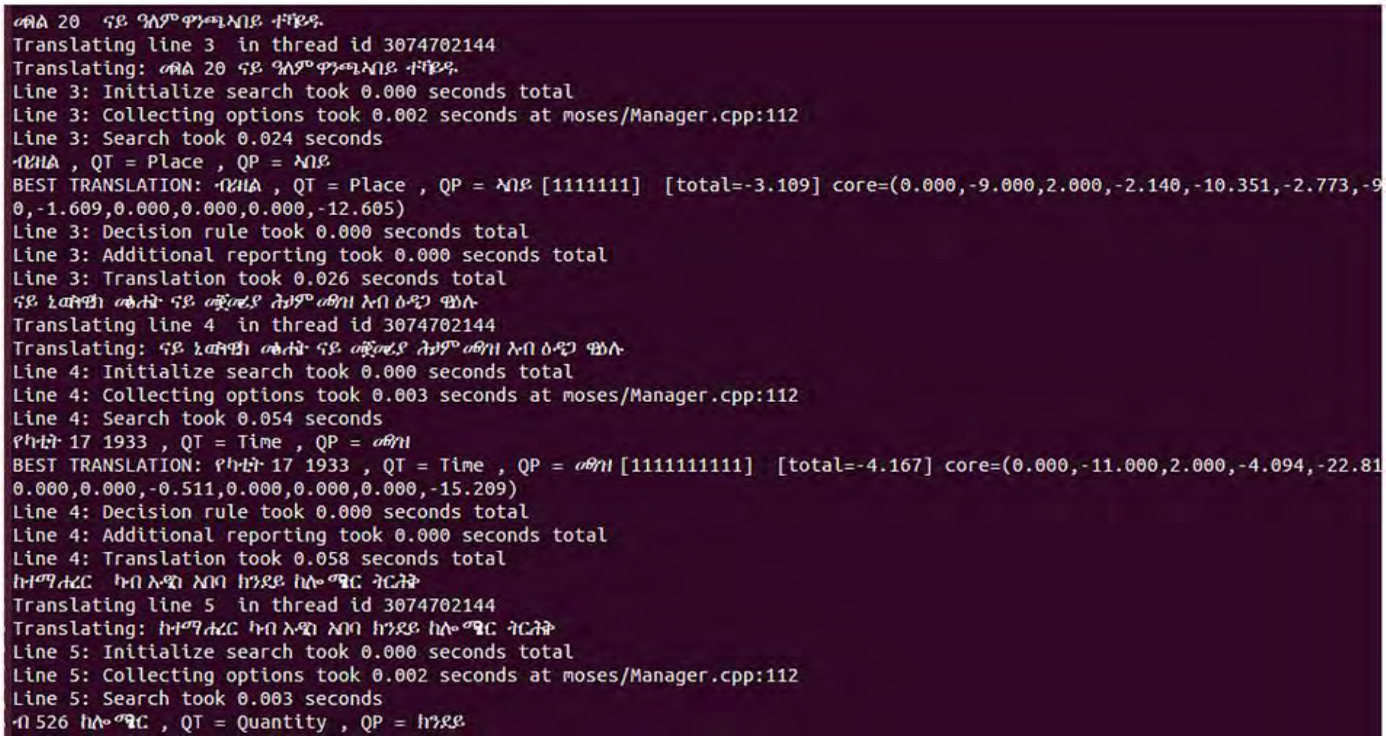


Figure 4 8 Sample extraction of answer for specific questions, question type and its question particle

Chapter Five: Experiments

5.1 Introduction

In this chapter we discuss the evaluation metrics to evaluate the performance of this question answering system. The experimental details and the ways of Tigrigna Question Answering System evaluations are also described and then finally the evaluation results of the performance of the Tigrigna factoid QAS are presented.

5.2 Developmental Environment

To achieve the research‟s objective, an essential number of developmental environments are needed. So, the system developed and evaluated on a computer with the following tools.

- **Moses**: we used to align question-answer parallel corpora, since the primary development platform for Moses is Linux, and this is the recommended platform. Thus, we installed this tool on Ubuntu 12.04.
- **GIZA++**: is a program for aligning words and sequences of words in aligned sentence in Question-Answer parallel corpora.
- **IRSTLM:** Language model estimation in extracting n-gram statistics of question particles, question type and answers to the given Tigrigna question. This provided by an efficient search algorithm that quickly finds the highest probability alignment of question sentences among the exponential number of question particles, question type and answer choices.

5.3 Evaluation Metrics

To accurately measure the performance of our QA System, we needed metrices that can provide good indication on how the system performs its goal. Due to this, there are a number of evaluation measures that can be used to evaluate the performance of this QAS. In such a situation each measure highlights a different aspect and use of several techniques‟ to describe the performance of the system.

Precision, Recall and F-Measure

These are the most commonly used indicators to measure IR retrieval and QA quality. Precision can be seen as a measure of exactness or quality whereas recall is a measure of completeness or quantity. High recall means that a system returned most of the relevant results. High precision means that the system returned more relevant results than irrelevant. Thus, the Precision and recall in our system can be described as follow. Consider a given question asked by the user called Q. Given that |RR| be the number of retrieved relevant answers, |RI| be the number of retrieved irrelevant answers and |RN| be the number of relevant answers not retrieved, then the precision and recall are given by the following equations.

$$\text{Precision} = \frac{|RR|}{|RR| + |RN|} \quad (5.1)$$

$$\text{Recall} = \frac{|RR|}{|RI| + |RN|} \quad (5.2)$$

F-Measure or F-Score can be defined as harmonic mean of precision and recall. It is a measure of system‟s accuracy. It considers both the precision and recall to compute the score. F-Measure reaches its best value at 1 and worst score at 0. It is 0 when no relevant answer has been retrieved and is 1 when exact answer is relevant.

$$\text{F} - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \quad (5.3)$$

Precision and recall can also be defined by a contingency table. We used typical contingency table to measure our system as shown below.

Table 5 1Contingency table of precision and recall

Total Answers	**Relevant Answers**	**Non Relevant**
Retrieved	Retrieved Relevant Answers (TP) ⇨ Return the Correct answer for the given question.	Retrieved Non- Relevant answers(FP) ⇨ Return Wrong answer for the given question.
Not Retrieved	Not Retrieved Relevant answers (FN) ⇨ Return Wrong answer for the given question.	Not retrieved Non –Relevant answers(TN) ⇨ There is no answer for the given question.

To measure our QAS performance the contingency table describes the above metrics as comparing the answers that return by our system with the answers given by human expert. Thus, TP answers are answers deemed relevant by both the human expert and this question answering system. FP answers are answers returned by this QA system, but irrelevant by the human expert. FN answers are relevant answers to the question which are not returned by the system. TN answers are answers not returned by this system and are considered irrelevant by the human expert. Then the accuracy, precision and recall of our question answering system expressed using the following metrics:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (5.4)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (5.5)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (5.6)$$

5.4 Performance Evaluation of Tigrigna Question Answer System

This section describes the evaluation of the Tigrigna QA system"s performance towards its correctness, completeness with recall, precision and F-measure computations on the factoid Tigrigna questions. Thus, to evaluate the system totally 200 test questions we used and those questions category according question type in four categories as seen the in the table below. The system also returns a single answer only for specific question.

Table 5 2 Amount of question to evaluate a question type.

Question type	No of answers present in the corpus	No of answers not present in the corpus	Total
Person	40	10	50
Time	40	10	50
Person	40	10	50
Quantity	40	10	50
Total			= 200

5.4.1 Question Classification Evaluation

In the question classification, the statistical language modeling automatically trained the training data sets and then its performance is measured on prepared test data sets contain 200 Tigrigna factoid questions with equal distribution question by the question type in five experiments. Then results of correct classification are shown in the Tables 5.3 and Table 5.4. Figure 5.1 depicts average performance of question classification component.

Table 5 3 the Performance of the language model question classifier on factoid type Tigrigna questions

Question type	Experiments	No .1	No.2	No.3	No.4	No.5
	Person	88	84	87	80	94
	Time	79	95	82	86	83
	Place	97	89	89	93	81
	Quantity	80	84	96	91	83

Table 5 4 average performance of question classification.

Question type	Average Exp result
Person	86.6
Time	85
Place	89.8
Quantity	86.8
Average	87

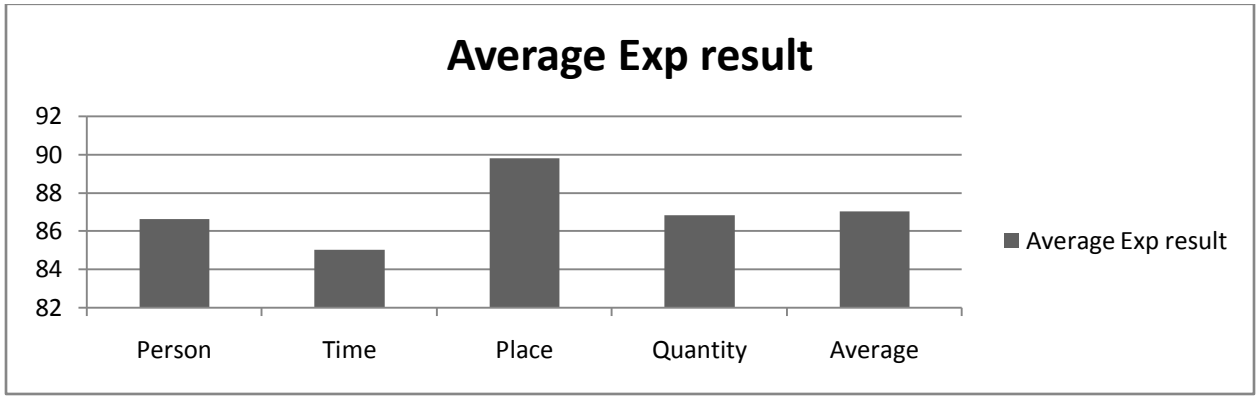


Figure 5 1 average performance of question classification component

Therefore, as shown in Table 5.4 and Figure 5.1, the average performance of statistical language model question type classifier is 87%. Sample questions and classification result are attached in Appendix B.

5.4.2 Answer Extraction Evaluation

Selection of evaluation criteria is a major task that is used in the completeness of performance analysis. To measure the performance of our system, we use two criteria"s" to group our testing data sets into different categories. The first one is based on the category of a question and then it divided into four categories. The second criterion is based on those questions need to get their answers from either present in the Tigrigna question answering corpora or not. Finally, these questions were fed to the system one by one and then the retrieved answers are analyzed. Generally the system is evaluated using the measures called precision, recall, and F-Measure. Table 5.5 shows the test results of a particular run.

Table 5 5 Contingency table showing Tigrigna question answering system test output

Total Answers	**Relevant Answers**	**Non Relevant answers**	**Total**
Retrieved	147	19	166
Not Retrieved	24	10	34
Total	171	29	200

Total number of questions asked = 200

Number of answers present in the corpus = 171 (147+24)

Number of questions correctly answered (TP) = 147

Number of questions wrongly answered (FP) = 19

Number of answers present in the corpus but not retrieved (FN) = 24

Answers which were not relevant (TN) = 10

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} = 147/\ (147+19) = 88.5\%$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad = 147/\ (147+24) = 85.9\%$$

$$\text{F} - \text{Measure} = \frac{2*\text{Precision}\ *\text{Recall}}{(\text{Precision}\ +\text{Recall}\ )} \quad = 2*88.5*85.9/\ (88.5+85.9) = 87.2\%$$

The performance of Tigrigna QAS is evaluated by measuring its ability to retrieve all and only relevant answer. It is strongly dependant on the size of the corpora because as the amount of aligned question answering pair of the corpora increases the ways of modeling a question is high. The system achieved an overall precision of 88.5% and 85.9% of recall. There is a tradeoff between precision (P) and recall(R). Higher the value of P lower will be the value of recall and vice versa. The F-Measure is the harmonic mean of P and R.

Referring to Table 5.5, it is clear that for 200 questions asked, 166 answers were retrieved out of which 147 answers were relevant to the query and 19 were non-relevant. Even though 24 other relevant answers existed in the corpus they are not extracted because some of the questions are not recognized properly or its semantics is not sufficient enough to identify the question sentences with corresponding answer sentences. In this work, retrieval scheme is purely based on language model. Hence, all the sentences with the candidate answers are retrieved. But all these answers might not be the answers to the question due to the result of probabilistically matching between the questions and answer.

Table 5 6 Performance of Tigrigna QAS

Questions	Relevant answers	Retrieved answers	Correct output	Recall	Precision	F-Measure
200	171	166	147	85.9	88.5	87.2

The overall performance of Tigrigna QAS is given in Table 5.6. Precision of 88.5% shows that the answers retrieved are correct answers and only very few non-relevant answers are retrieved. Percentage of recall is less than percentage of precision and it is 85.9%. This can certainly be improved by increasing the amount of training data sets. While raising the questions no restriction is kept to avoid any bias that may affect the system performance.

Table 5 7 describes the performance of Tigrigna QAS based on question type of factoid questions were considered in this work.

Table 5 7 Performance according to question types

Question type	No of questions	Retrieved answers	Correct answers	Wrong answers	Not retrieved	Precision	Recall	F–measure
Person	50	39	35	4	6	90.5	85.4	88
Time	50	41	34	7	4	83.5	90	87
Place	50	39	33	6	5	85.5	85	85.2
Quantity	50	47	44	3	9	94.5	83	88.4
Total	200	166	147	19	24			
					Average	88.5	85.9	87.2

Thus, the evaluation result of the answer extraction of Tigrigna QAS is shown in Table 5.7 and Figure 5.2 based on Recall, Precision and F-measure evaluation metrics.

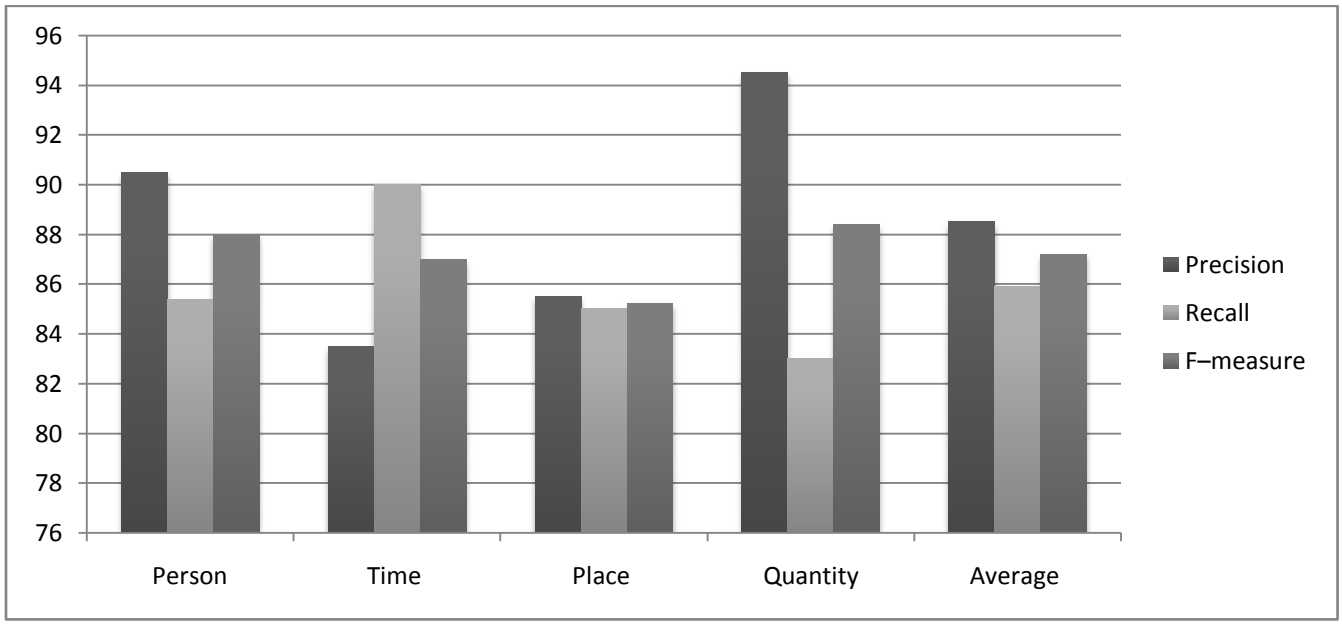


Figure 5 2 average performance of the answer extraction.

Chapter Six: Conclusion and Future Work

6.1 Introduction

Since a Question Answering system provides correct responses to the user"s questions in short and accurate manner by formulating Natural Language (NL) query to return only one or a small set of specific answers for a given question. Due to this, Tigrigna QAS is essential for retrieving short and precise answers to questions asked by Tigrigna language users through applying much deeper understanding and processing of text than most web search engines performed as a result of links of documents.

6.2 Conclusion

Because of the difference in grammatical construction and question particles on different natural languages, so many languages dependent QA systems are done. Thus, this thesis is done for Tigrigna fact-based questions using language model approach. This QA system consists of question analysis, document analysis and answer extraction modules. this system has 87% an average performance of the statistical language model question type classifier and also the average Precision, Recall and F – measure of the answer extraction as precision is 88.5%, recall is 85.9% and F – measure is 87.2%.

6.3 Contribution of the work

The most relevant contributions of this thesis work are listed as follows:

- The study has adopted other languages statistical language model based techniques of question answering systems to fact-based questions of Tigrigna language.
- The study has developed based on a statistical language modeling approach to model the classification of Tigrigna questions to their category.
- The study has understood the tasks of natural language processing that have impact on the performance of this QAS especially related to the parallel corpora processing.
- The study showed the language model techniques in developing this QA system

6.4 Recommendations and Future Works

To improve the performance of Tigrigna question answering system different Natural language processing activities needed to add to this QAS. Because of the tasks of fully developed QAS is a more complex, we required to specify additional features that can be added to increase the performance of the proposed QA system and future research works in this section. Those can be listed as follow:

- This research is done for fact-based questions. Thus, able to do for non-factoid question types and complex factoid question will be an open research area.
- Developing the QAS in the real-world situations in different platforms such Mobile application; can be considered as a future work and also such a system can be restricted to handle questions of closed domain for example in medical and other governmental, and non-governmental sectors.
- The performance of Tigrigna Question answering system depends on the size of the parallel corpora. Due to this, preparing of those corpora and using for training and testing activities needs standardization, which is not done currently. Thus developing standards of parallel corpora is open research for Tigrigna and other local languages.
- The main task of this question answering system is collecting and preparing corpora because it depends on size and semantically representation of contents rather than the Tigrigna language syntactical and structural representation. Due to this, we can develop a question answering system that supports two or more human languages.
- Adding Tigrigna WordNet to this system could be maximizing the performance of Expansion Query component tasks which helps in retrieval of answer from different representation.
- Adding Tigigna spelling checker to this system would help to avoid user input error when writing a Tigrigna question.

References

[1] Tilahun Abedissa, "Amharic Question Answering for Definitional, Biographical and Description Questions", Unpublished Master‟s Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa, 2013.

[2] Allen James, *"Natural Language Understanding",* The Benjamin Cummings Publishing Company, Redwood City,1995.

[3] Andrew Greenwood Mark, "Open-Domain Question Answering." Unpublished PhD Dissertation, Department of Computer Science, University of Sheffield,Sheffield ,2005.

[4] Das Biplab, "A Survey on Question Answering System" , Unpublished Master‟s Thesis,Department of Computer Science & Engineering, Indian Institute of Technology, Bombay, 2005.

[5] Gangwal Gaurav, "Question Answering System using Open Source Software", Unpublished Master's Projects, 2012, Paper 258.

[6] Haque Nafid, "A Prototype Framework for a Bangla Question Answering System using Translation based on Transliteration and Table Look-up as an Interface for the Medical Domain",M.Sc HLST, 2010.

[7] Kaisser Michael, "Acquiring Syntactic and Semantic Transformations in Question Answering" Unpublished PhD Dissertation, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh, Edinburgh, 2010.

[8] Monz Christof, "From Document Retrieval to Question Answering." Institute for Logic, Language and Computation, Amsterdam, Duitsland, 2003.

[9] Sundblad Håkan, "Question Classification in Question Answering Systems."Unpublished PhD Dissertation, Department of Computer and Information Science, Linköpings, 2007.

[10] Tesfaye Tewolde, "*A Modern Grammar of Tigrigna",*Unpublished PhD Dissertation, Savonarola Roma: Tipografia U. Detti – via G, 2002.

[11] *Tigrinya language* . http://en.wikipedia.org/wiki/Tigrinya_language (accessed Jan 20, 2015).

[12] Wissal Brini, Ellouze Mariem, and Trigui Omar, "Factoid and Definitional Arabic Question Answering System",*ANLP Research Group- MIRACL Laboratory*, 2009.

[13] Seid Muhie Yimam,“Teteyeq :Amharic Question Answering System for Factoid Questions",Unpublished Master‟s Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa, 2009.

[14] Desalegn Abebaw Zeleke, "LETEYEQ- A Web Based Amharic Question Answering System for Factoid Questions Using Machine Learning Approach ", Unpublished Master‟s Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa,2013.

[15] Aberash Tesfaye, "Afaan Oromo Question Answering System for Factoid Questions", Unpublished Master"s Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa,2014.

[16] TomekStrzalkowski, Sanda M. Harabagiu, "Advances in Open Domain Question Answering", Text, Speech and Language Technology Series, SpringerLink: Bucher, Vol. 32, Springer 2006.

[17] Ask Jeeves.http://www.ask.com (accessed May 20, 2015).

[18] Green W, Chomsky C, and Laugherty K., "BASEBALL: An automatic question answerer", Proceedings of the Western Joint Computer Conference, p.p. 219-224 San Francisco, CA, 1961.

[19] Chirstan Grant, Clint P.George, Joir-danGumbs,JosephN.Wilson, PeterJ.Dobbins, "Morpheus: A Deep Web Question Answering System", WAS2010, ACM, Paris, 2010.

[20] Woods, William A, "Semantics and Quantification in Natural Language Question Answering", Advances in Computers, Vol.17, Academic Press, 1978.

[21] Robin D. Burke, Kristian J. Hammond, Vladimir Kulyukin, Steven L. Lytinen, Noriko Tomuro, Scott Schoenberg , **"**Question Answering from Frequently Files", AI Magazine, Vol.18, No.2, AAAI,1997.

[22] Boris Katz, Gary Borchardt, Sue Felshin, "Natural Language Annotations for Question Answering", AQUAINT Phase II, AAAI, 2006.

[23] L. Hirschman and R. Gaizauskas, Natural language question answering: the view from here, *Natural Language Engineering,* vol. 7, no.4, pp. 275-300, 2001.

[24] SunitaSarawagi, Information Extraction, *Indian Institute of Technology,* India.,Vol. 1,No. 3, pp. 261–377, 2008.

[25] Hiemstra, Djoerd. *Using language models for information retrieval*. TaaluitgeverijNesliaPaniculata, 2001.

[26] Hobbs, Jerry R., and Ellen Riloff. "Information extraction." *Handbook of natural language processing,* 2010.

[27] Gharehchopogh, FarhadSoleimanian, and YaghoubLotfi. "Machine Learning based Question Classification Methods in the Question Answering Systems." *International Journal of Innovation and Applied Studies* 4.2, 264-273, 2013.

[28] Berger, A. and Lafferty, J. Information retrieval as statistical translation. In Proceedings of the 22nd Annual International ACM SIGIR Conference, pp.222-229,1999.

[29] Liu, Xiaoyong, and W. Bruce Croft. "Statistical language modeling for information retrieval." *Annual Review of Information Science and Technology* 39.1 pp. 1-31, 2005.

[30] Zhang, Dell, and Wee Sun Lee. "A Language Modeling Approach to Passage Question Answering." *TREC*. 2003.

[31] Merkel, Andreas. "Using language models in question answering." 2008.

[32] Espana-Bonet, Cristina, and Pere R. Comas. "Full machine translation for factoid question answering." Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra). Association for Computational Linguistics, 2012.

[33] Lee, Seungwoo, and Gary Geunbae Lee. "SiteQ/J: A question answering system for Japanese." *Proceedings of the third NTCIR Workshop*. 2003.

[34] Amaral, Carlos, et al. *Priberam's question answering system for Portuguese*. Springer Berlin Heidelberg, 2006.

[35] *Tigrinya language* . http://www.eritrea.be/old/eritrea-languages.htm (accessed Jan 20, 2015).

[36] *Tigrinya language* . http://www.omniglot.com/language/phrases/tigrinya.php (accessed Jan 20, 2015).

[37] *Tigrinya language* . http://www.omniglot.com/writing/tigrinya.htm (accessed Jan 20, 2015).

Appendix A: calendar and time notation, number system and punctuation of Tigrigna Language.

Months of the year			
January	ጥሪ	July	ሓምለ
February	ለካቲት	August	ነሓሰ
March	መጋቢት	September	መስከረም
April	ሚያዚያ	October	ጥቅምቲ
May	ግንቦት	November	ሕዳር
June	ሰነ	December	ታሕሳስ

Days of the week		
Sunday	ሰንበት	From Hebrew Sabbath
Monday	ሰኑይ	Semitic root for "two"
Tuesday	ሰሉስ	Three
Wednesday	ሮቡዕ	Four
Thursday	ሓሙስ	Five
Friday	ዓርቢ	Sunset
Saturday	ቀዳም	"first " or "prior " to Sunday

Time division of the Day	
Time Period Name	Meaning
ወጋሕታ	From first light to sunrise
ንጎሓ	From sunrise to mid morning
ፍርቂ መዓልቲ	Midday , noon
ድሕሪ ቀትሪ	From noon to early afternoon
ኣጋምሽት	Afternoon
ምሸት	From sunset to dark
ፍርቂ ለይቲ	Mid night
Number Systems	
Cardinal Numbers: some of Tigrigna written form of numbers derived from Geez language, Arabic and alphanumeric such as:	

Arabic	Ethiopic	Alphanumeric	Arabic	Ethiopic	Alphanumeric
1	፩	ሓደ	20	፳	ዒስራ
2	፪	ክልተ	30	፴	ሳላሳ
3	፫	ሰለስተ	40	፵	ኣርብዓ
4	፬	ኣርባዕተ	50	፶	ሓምሳ
5	፭	ሓሙሽተ	60	፷	ስልሳ
6	፮	ሽድሽተ	70	፸	ሰብዓ
7	፯	ሸውዓተ	80	፹	ሰማንያ
8	፰	ሸሞንተ	90	፺	ቴስዓ
9	፱	ትሽዓተ	100	፻	ሚእቲ
10	፲	ዓሰርተ	1000	፼	ሽሕ

Ordinal number: are numbers inflected for gender and agree with the noun they modify .masculine and the ordinal numbers eleventh and higher are made by using the word መበል mebel with cardinal number for instance መበል ዓሰርተው ሓደ mebel asertew hade which means eleventh .

Masculine	Feminine	
ቀዳማይ	ቀዳመየቲ	First
ካልኣይ	ካልኣየቲ	Second
ሳልሳይ	ሳልሰይቲ	Third
ራብዓይ	ራብዓይቲ	Fourth
ሓምሻይ	ሓምሸይቲ	Fifth
ሻድሻይ	ሻድሻይቲ	Sixth
ሻብዓይ	ሻብዓይቲ	Seventh
ሻምናይ	ሻምናይቲ	Eighth
ታሽዓይ	ታሽዓይቲ	Ninth
ዓስራይ	ዓስራይቲ	Tenth

Fractions: numbers

Tigrigna	English	Arabic

ፍርቂ	One –half	1/2
ሲሶ	One-third	1/3
ርብዒ	One-fourth	1/4
ሕምስት /ሓምሻይ ኣፍ ወይ ድማ ሓምሻይ ኢድ	One- fifth / A fifth portion	1/5
ስምኒት	One- eight	1/8
ዕስሪት	One- tenth	1/10

Tigrigna Punctuations

	Comma	Full stop / period	Colon	Semi colon	Question mark
Tigrigna	: or ፣	::	፡	፤	?
English	,	.	:	;	?

Appendix B: Sample test questions, their classification and answer retrieved results of the experiments.

Question	Answer	Question Classification Result	Answer retrived Result
መበል 10ይ ናይ አሜሪካ - አፍሪካ ቢዝነስ መድረክ ኣበይ ይካየድ	አዲስ አበባ	Correct	Correct
ናይ ኢትዮጵያ አምባሳደር ኣብ አሜሪካ መን ይበሃሉ	አምባሳደር ግርማ	Correct	Correct
ናይ ኦሮሚያ በዓል መዚ እቶት ዋና ዳይሬክተር መን ይበሃሉ	ኣይተ ዮሐንስ ድንቃየው	Correct	Correct
ናይ ኢትዮጵያ አትሌቲክስ ፌደሬሽን ቤት ጽህፈት ሀላፊ መን ይበሃሉ	ኣይተ ቢልልኝ መቆያ	Correct	Correct
ናይ ከተማ ልማዓትን ኣባይቲን ሚኒስቴር መን ይበሃሉ	ኣይተ መኩሪያ ሀይሌ	Correct	Correct
ናይ ኢንፎርሜሽን መርበብ ድሕንነት ኤጀንሲ ዋና ዳይሬክተር መን ይበሃሉ	ሜጀር ጀነራል ተክለብርሀን ወልደ አረጋይ	Wrong	Wrong
ኢራን ካብ ኤ ቲ አር ክንደይ ነፈርቲ ክትገዝእ እያ	40	Correct	Correct
ናይ ሳይንስን ቴክኖሎጂን ሚኒስትር ድኤታ መን ይበሃሉ	ፕሮፌሰር አፈወርቅ	Correct	Correct
ናይ ቻድ ፕሬዚዳንት መን ይበሃሉ	ኢድሪስ ዴቢ	Correct	Correct
ንናይ አፍሪካ ሕብረት መን ኣቦወንበር ኮይኑ ተመረፁ	ፕሬዚዳንት ኢድሪስ ዴቢ	Correct	Correct
መበል 26 ናይ አፍሪካ ሕብረት ንይ መረሕቲ መደበኛ ስብሰባ ኣበይ ተካይዱ	አዲስ አበባ	Correct	Correct
ናይ ኢትዮጵያ ሓይሊ ኤሌክትሪክ ህዝቢ ርክብ ዳይሬክተር መን ይበሃሉ	አቶ ምስክር ነጋሽ	Wrong	Correct
ናይ ፈረንሳይ ወፃኢ ጉዳይ ሚኒስትር መን ይበሃሉ	ሎረንት ፋብየስ	Correct	Correct
ናይ ሑቡራት መንግስታት ድርጅት ዋና ፀሀፊ መን ይበሃሉ	ባን ኪ ሙን	Correct	Correct
ናይ አፍሪካ ህብረት ኮሚሽን ኣዶመንበር መን ይበሃላ	ኒኮ ሳዛና ዲላሚኒ ዙማ	Correct	Correct
ናይ ኦሮሚያ ክልል ምክትል ርዕሰ ምምሕዳር መን ይበሃላ	ኣይተ እሼቱ ደሴ	Correct	Correct
መዓዝ እያ ሓንቲ ዝሓበረት አፍሪካ ሀገር ትምስረት	ብ2063	Correct	Correct
ና ኢትዮጵያ ክለባት ጥሎ ማለፍ ይጀመር	የካቲት 3 2008 ዓ.ም	Wrong	Wrong

ጃፓን ኣብ ኢትዮጵያ ብድርቅ ንዝተጎዱኡ ዜጋታት ዝውዕል ክንደይ ዝኣክል ገንዘብ ድጋፍ ገራ	21 ነጥብ 7 ሚሊየን የአሜሪካ ዶላር	Correct	Correct
ናይ ጣልያን ወፃኢ ጉዳይ ሚኒስትር መን ይብሃሉ	ፓውሎ ጀንተሎኒ	Correct	Correct
ናይ ቻይና ምክትል ወፃኢ ጉዳይ ሚኒስትር መን ይብሃሉ	ዣንግ ሚንግ	Correct	Correct
ናይ ሴኔጋል ፕሬዚዳንት መን ይብሃሉ	ማኪ ሳል	Correct	Wrong
ናይ ኢፌዴሪ ወፃኢ ጉዳይ ሚኒስትር መን ይብሃሉ	ዶክተር ቴድሮስ አድሃኖም	Correct	Wrong
ናይ ኢትዮጵያ ወፃኢ ጉዳይ ሚኒስትር መን ይብሃሉ	ዶክተር ቴድሮስ አድሃኖም	Correct	Correct
ናይ ኢፌዴሪ ምክትል ጠቅላይ ሚኒስትር መን ይብሃሉ	አይተ ደመቀ መኮንን	Correct	Correct
ናይ ኢትዮጵያ ምክትል ጠቅላይ ሚኒስትር መን ይብሃሉ	አይተ ደመቀ መኮንን	Correct	Correct
ናይ ብራዚል ፕሬዚዳንት መን ይብሃሉ	ዲልማ ሩሴፍ	Correct	Correct
ናይ ሃዋሳ ዩኒቨርሲቲ ፕሬዝዳንት ቤት ፅህፈት ሃላፊ መን ይብሃሉ	አይተ ካሳየ ጋዲሳ	Correct	Correct
ናይ ቀደም ሊቢያ መራሒ ሙኣሙር ጋዳፊ መዓዝ ብሓይሊ ስልጣኞም ለቀቁ	ከም አውሮፓውያን አቆፃፅራ 2011	Correct	Correct
ናይ ሊቢያ ዋና ከተማ መን ትብሃል	ትሪፖሊ	Correct	Correct
ናይ ኢፌዴሪ ወፃኢ ጉዳይ ሚኒስትር ዲኤታ መን ይብሃሉ	አምባሳደር ታዬ አጽቀስላሴ	Correct	Correct
ናይ ኢትዮጵያ ወፃኢ ጉዳይ ሚኒስትር ዲኤታ መን ይብሃሉ	አምባሳደር ታዬ አጽቀስላሴ	Correct	Correct
ናይ ኒውዝላንድ ወፃኢ ጉዳይ ሚኒስትር መን ይብሃሉ	ጀፍ ላንግሌ	Correct	Correct
ናይ ብረታ ብረት ኢንዱስትሪ ልማዓት ኢንስቲትዩት ዋና ዳሬክተር መን ይብሃሉ	አይተ ወርቅነህ ደለለኝ	Correct	Wrong
ናይ አፍሪካ ሕብረት ሓይሊ ሰብ አመራርሓ ዳይሬክተር መን ይብሃሉ	አሚን ኢድሪስ አዱም	Correct	Correct
ሳሞስ ደሴት ኣበይ ትርከብ	ኣብ ቱርክ አቅራቢያ	Correct	Correct
ኣብ መበል 26 ናይ አፍሪካ ሕብረት መረሕቲ ስብሰባ ክንደይ ዝኣክሉ መረሕቲ	ልዕሊ 40	Correct	Correct

ሃገራት ኣባል እቲ ሕብረት ይርከቡ			
ኣብ ናይ ጃንሜዳ ዓለም ለኸ ሃገር ምቁራፅ ውድድር ከንደይ ዝኣኽላ ሃገራት ይሳተፉ	ሓሙሽተ	Correct	Correct
ድሬቲዮብ ዝተብሃለ ድሕረ- ገፅ ብመን ተጀሚሩ	ቢኒያም	Correct	Correct
ሓድነት ንዲሞክራሲን ንፍትሕን ፓርቲ ምክትል ኣቦወንበርመን ይብሃሉ	ግርማ ሰይፉ	Correct	Correct
ኣብ ኢትዮጵያ ናይ ግብፅ ኣምባሳደር መን ይባሃሉ	መሀመድ ኢድሪስ	Correct	Correct
ኣብ ኢትዮጵያ ኮሪያ ዘመቲ ማህበር ፕሬዝደንት መን ይባሃሉ?	ኮሎኔል መለሰ ተሰማ	Wrong	Wrong
ናይ ጃፓኑ ናይ ኢኮኖሚ ሚኒስትር ኣኪራ ኣማሪ ከንደይ ዘኣክል ገነዝብ ጉቦ ተቀቢሎም	101 ሺሕ ናይ ኣሜሪካ ዶላር	Correct	Correct
ኣየርላንድ ኣብ ኢትዮጵያ ንዝተፈጠረ ድርቂ ከንደይ ዝኣክልገንዘብ ድጋፍ ገይራ	3 ነጥብ 8 ሚሊየን ናይ ኣሜሪካ ዶላር	Correct	Correct
ኣብ ኢትዮጵያ ብናይ ሩዝ ምህርቲ ዝግጅት ምኽንያት ዝትኻየደ ኣኼባ ኣበይ ነይሩ	ኣብ ባህር ዳር ከተማ	Correct	Correct
ናይ ኳታር እዝዳን ኩባንያ ኣቦመንበር መን ይብሃሉ	ሼህ ካሊድ ታኒ አልታኒ	Correct	Correct
ካልኣይ ናይ ኢትዮጵያ ህዝቢብሄራዊ መዝሙር ግጥም ደራሲ መን ይብሃሉ	ባህታ ሳህለ	Correct	Correct
ግብጻዊ ዉድብ ሑቡራት መንግስታት ድርጅት መራሒ ዝነበሩ መን ይብሃሉ	ሙራድ ዓሊ	Wrong	Correct
ናይ ሓለዋ ጥዕና ሚኒስቴር ሚኒስትር መን ይብሃሉ	ኣድማሱ ከሰተ	Correct	Correct
ናይ መጀመሪያ ፓኪስታናዊ ናይ ኖቤል ተሸላሚ መን ይብሃል	ፒካሚ	Correct	Correct
ናይ ደቡብ ክልል ርዕሰ ምምሕዳር መን ይብሃሉ	ደሴ ናደው	Correct	Wrong
ኣብ ሮም ናይ ማራቶን ውድድር ብባዶ እግሩ ጉያ ሀሸነፈ ኢትዮጵያዊ መን ይብሃል	ኣበበ በቃላ	Correct	Correct
ኢትዮጵያ ብዘመናዊነትን ኣብ ለውጥ ጉዕዞ ንክትጉዓዝ ዝገበሩ ናይ መጀመሪያ ሰብ መን እዮም	ሃይፀ ሚኒሊክ	Correct	Correct
ሕልሚ አለኒ ብዝብል ታሪካዊ ንግግር ዝገበሩ አሜሪካዊ መን ይብሃሉ	ኣብርሃም ሊንከን	Correct	Correct
ተማሪኹ ዝተወሰደ ሃይፀ ቴዎድሮስ ዉላድ መን ይብሃል	ያሬድ	Correct	Correct
ማራቶንን ንመጀመሪያ ጊዜ ዘሸነፈ ኣፍሪካዊ አትሌት መን ይብሃል	ኣበበ በቃላ	Correct	Correct
ናይ አይ ቢ ኤም ኩባንያ መስራቲ መን እዩ	ኪሊኑተን	Correct	Correct
ኢትዮጵያን ከብ 1889 ክሳብ 1913 ዘመሓደሩ ንጉስ መን ይብሃል	ልጅ እያሱ	Correct	Correct
ናይ ትግራይ ክልል ርዕሰ ምምሕዳር መን እዮም	ኣባይ ወልዱ	Correct	Correct

ከአድማስ ባሻገር ልቦለድ መፅሀፍ ደራሲ መን እዮም	በዓሉ ግርማ	Wrong	Correct
ናይ ቀዳማይ ሚኒስትር መለስ ንይ መጀመሪያ ስም መን ይብሃል	ለገሰ ዜናዊ	Correct	Correct
ናይ ደብረ ብርሃን ቅድስት ስላሴ ቤተክርስቲያን ብመን ዘመነ መንግስት ተሰሪሑ	ዘርዓ ያዕቆብ	Correct	Correct
ኣብ 1986 ናይ ኢትዮጵያ ፕሬዝዳንት መን ነይሩ	ኮኔሬል መንግስቱ ሃይለ ማርያም	Correct	Correct
ብኣሜሪካ ዝተቀተሉ ናይ ቀደም ናይ ኢራቅ ፕሬዝዳንት መን ይብሃሉ	ሳዳም ሑሰን	Correct	Correct
ሃይፀ ዳዊት ክብ መዓዝ ን ምዓዝ ነጊሶም ነይሮም	ካብ 1733 እስካብ 1804	Correct	Correct
ምኒሊክ 2ይ መዓዝ ተወሊዱ	ብ 1869	Correct	Correct
ኣብ ኢትዮጵያ ንይ መጀመሪያ ስልክ አገልግሎት መዓዝ ተጀሚሩ	ብ1917	Correct	Correct
ናይ ማይጨው ኩናት መኣዝ ተኻይዱ	ብ1928	Correct	Correct
ናይ ዓለም ዋንጫ መዓዝ ተጀሚሩ	ብ 1959	Wrong	Correct
ናይ አፍሪካ ሕብረት መዓዝተመስረቱ	ብ 1945	Correct	Correct
ክብ አዲስ አበባ ጂቡቲ ዝተዘርገሐ ኛይ ባቡር መስመርመዓዝ ስራሑ ጀሚሩ	ብ 1934	Correct	Correct
ናይ ለንደን ማራቶን መዓዝ ተኻይዱ	ብ 1960	Correct	Correct
ምስጢር ዝብሃል መፅሐፍ ዝደረሰ መን ይብሃል	መሀመድ ሰልማን	Correct	Correct
ሓድነት አፍሪካ ድርጅት እንትምሰረት ናይ መጀመሪያ ዋና ፀሐፊ መን ነይሮም	ኣይተ ክፍሌ ወዳጆ	Correct	Wrong
መስራቲ ናይ ኢትዮጵያ ፉትቦል ፌዴሬሽን መን ነይሮም	ይድነቃቸው ተሰማ	Correct	Correct
ኣብ ጋና ናይ ኢትዮጵያ አምባሳደር መን ይበሃሉ	ጊፍቲ አባሲያ	Correct	Correct
ናይ ሚሽን ፎር ኮሚዩኒቲ ዴቨሎፕመንት ፕሮግራም መስራቲትን ኤክስኪዮቲቭ ዳይሬክተር መን ይብሃላ?	ወይዘሮ ሙሉ ሀይለ	Correct	Correct
ኣብ ኢትዮጵያ እግር ኩዕሶ ፌደሬሽን ናይ መጀመሪያ ምክትል ፕሬዝዳንት መን ነይሮም	ተካ አስፋው	Wrong	Correct
ናይ ሸገር ሬዲዮ መስራቲት በዓልቲ ዋናን ሥራ መካየዲት መን ይብሃላ	መአዛ ብሩ	Correct	Correct
ናይ ግርማዊ ቀዳማዊ ሃይፀ ኃይለሥላሴ መዘከርታ ማህበር ናይ ቦርድ ኣቦመንበር መን ይባሃሉ	ናሁሰናይ አርአያ	Correct	Correct
ናይ ትምህርት ሚኒስቴር ሚኒስትር ዴኤታ መን ይበሃሉ	አቶ ፉአድ ኢብራሂም	Correct	Correct
ናይ መጀመርያ ዓብዩ ጉያ ኣብ ኢትዮጵያ መን ኣሸኒፉ ነይሩ	ሀይሌ ገብረስላሴ	Correct	Wrong

ናይ እንጀራ መጋገሪ ማሽን ዲዛይን ዝገበረ መን እዩ	ሙሉጌታ በጋሻው	Correct	Correct
ኣብ ናይ ለንደን ኦሊምፒያድ ኢትዮጵያን ብ800 ሜትር ጉያ ወከሉ ዝተወዳደረ መን እዩ	መሀመድ አማን	Correct	Correct
ናይ ኢትዮጵያ ደቂ ኣንስትዮ ሰናይ ኣድራጉት ማሕበራት ሕብረት ሥራ መካየዲ መን እየን	አዜብ ቀለመወርቅ	Correct	Correct
ናይ ግብፅ ፕሬዚዳንት ዝነበሩ መን ይበሃሉ	መሀመድ ሞርሲ	Correct	Correct
ናይ መጀመርታ ናይ ኢትዮጵያ እግሪ ኩዕሶ ክለብ መን ይበሃል	ቅዱስ ጊዮርጊስ	Correct	Correct
ናየ ኢትዮጵያ ብሔራዊ ቡድን ኣሰልጣኒ መን ይበሃሉ	ሰውነት ቢሻው	Correct	Correct
ናየ መጀመሪያ ጓል ኣንእስተየቲ በዓሊቲ ቅኔ መን ይብሃላ	እማሆይ ገላነሽ	Correct	Correct
ናይ ኢትዮጵያ ምህርት ዕደጋ መስራቲትን ዋና ስራሕ መካየዲ መን ይብሃላ?	እሌኒ ገ/መድህን	Correct	Correct
ናይ ኢትዮጵያ ደቂ ኣንእስትዮ ደራስያን ማሕበር ፕሬዚዳንት መን ትብሃል	የምወድሽ በቀለ	Correct	Correct
ናይ መጀመሪያ ጓል ኣንእስተየቲ ናይ ፊልም ደራሲ መዳለዊትን ፕሮዲዩሰርን መን ትብሃል	ቅድስት ባየልኝ	Correct	Correct
ናይ መንግስቲ ኮሙዩኒኬሽን ጉዳያት ሚኒስትር መን እዮም	ኣይተ ጌታቸው ረዳ	Correct	Wrong
ጉማ ናይ ፊልም ሽልማት መዓዝ ተኻይዱ	ኣብ ታህሳስ ወርሒ	Correct	Correct
ናይ ሞጆ ደረቅ ወደብ መዓዝ ተጀሚሩ	ብ 2001	Correct	Correct
ልጅ እንዳልካቸው መኮንን መዓዝ ተወሊዶም	ጳጉሜ 3 ቀን 1920	Correct	Correct
ዲናሞ መዓዝ ተሰሪሑ	ብ 1831	Correct	Correct
ኔልሰን ማንዴላ ዓብዩ ዓለም ለኸ ናይ ሰላም ሽልማት መዓዝ ተቀቢሎም	ብ 1985	Correct	Correct
ኢኳቶሪያል ጊኒ ካብ ስፔን ኣገዛዝኣ ነፃነታ መዓዝ ረኺባ	ብ 1968	Correct	Correct
ናይ ኢትዮጵያ ደቂ ኣንስትዮ ሰናይ ኣድራጉት ማሕበራት ሕብረት መዓዝ ተመስረቱ	ብ 2010	Correct	Wrong
እስራኤል ምስ ግብጽ ብካምፕ ዴቪድ ናይ ሰላም ስምምዕነት መዓዝ ተኻይዱ	ብ 1979	Correct	Correct
ፀጋዬ ገብረ መድህን መዓዝ ተወለዱ	ነሓሴ 1928	Correct	Correct
ናይ ኒውስዊክ መፅሔት ናይ መጀመሪያ ሕታም መዓዝ እብ ዕዳጋ ዊዕሉ	የካቲት 17 1933	Correct	Correct
የፍታሄ ንጉሴ መዓዝ ተወለዱ	ብ 1885	Correct	Correct
ኣቡነ ጴጥሮስ መዓዝ ተወለዱ	ብ 1885	Correct	Correct
ናይ ቀዳማዊ ኃይለሥላሴ መዘከሪታ ማሕበር መዓዝ ተመስረቱ	ሓምለ 16 1987	Correct	Correct

ናይ ፅሕፈት መሐተሚ ማሽን ናብ ኢትዮጵያ መዓዝ ኣትዩ	ብ 1889	Correct	Correct
ኣብ ዓለምና መጠን ምብፃሕ ተንቀሳቓሲ ቴልፎን ሚኢቲ ካብ ሚኢቲ መዓዝ ይኸውን	ብ 2016	Wrong	Correct
ናይ አፍሪካ ሕብረት ኣዶመንበር ዶ/ር ዙማ መዓዝ ቃለ መሐላ ፈፂመን	ሓምለ 10 2004	Correct	Correct
ዘመናዊ ኦሊምፒክ መዓዝ ተጀመሩ	ብ 1896	Correct	Correct
ኢትዮጵያ ንመጀመርያ ጊዜ እብ ኦሊምፒክ ምስታፍ ዝጀመረትሉ መዓዝ እዩ	ብ 1956	Wrong	Correct
ናይ ኢትዮጵያ አትሌቲክስ ፌዴሬሽን ዓመታዊ ጠቅላላ ኣኼባኡ ዘካይዶ መዓዝ እዩ	ሕዳር 22 ን ሕዳር 23	Correct	Wrong
ኣብ ናይ ንጉሱ ዘመን ናይ ኢትዮጵያ ህዝቢ መዝሙር መዓዝ ተፃሒፉ	ብ 1919	Correct	Correct
ኣብ ዳግማዊ ምኒልክ ሆስፒታል ውሽጢ ዝርከብ ጥዕና ሳይንስ ኮሌጅ መዓዘ ተመስረቱ	ብ 1941	Wrong	Correct
ኢህአዴግ ናብ ስልጣን መዓዝ ወፅኡ	ግንቦት 20	Correct	Correct
ናይ ቀደም ሶቪየት ሕብረት አፍጋኒስታንን ዝወረረትሉ እዋን መዓዝ ነይሩ	1979	Correct	Correct
መሐመድ አማን መዓዝ ተወለዱ	ጥሪ 1 1986	Correct	Correct
ሱዳን ኣብ ናይ አፍሪካ ዋንጫ ሻምፒዮን ዝኾነትሉ መዓዝ ነይሩ	ብ 1970	Correct	Correct
ፍራንሲ ኦላንድ ናይ መን ሃገር ፕሬዚዳንት ነይሮም	ፈረንሳይ	Correct	Correct
ፀጋዬ ገብረ መድህን ኣበይ ተወለዱ	አምቦ ከተማ	Wrong	Correct
ሳልቫ ኪር ናይ መን ሃገር ፕሬዚዳንት እዮም	ደቡብ ሱዳን	Correct	Correct
ዘመናዊ ኦሊምፒክ ኣበይ ተጀመሩ	ኣብ ግሪክ አቴንስ	Correct	Correct
መበል 20 ናይ ዓለም ዋንጫ ኣበይ ተኻይዱ	ብራዚል	Correct	Correct
ኤልክላሲኮ ናይ መን ሃገር ክለባት ደርቢ ጭዋታ እዩ	ስፔን	Correct	Correct
ናይ ሞዛምቢክ ዋና ከተማ መን ትብሃል	ማፑቶ	Wrong	Correct
ናይ ሶሪያ ዋና ከተማ መን ትብሃል	ደማስቆ	Correct	Correct
ናይ ኢትዮጵያ አትሌቲክስ ፌዴሬሽን ዓመታዊ ጠቅላላ ኣኼባኡ ኣበይ ኣካይዱ	ባህር ዳር	Correct	Correct
መበል 6ይ ቢግ ብራዘርስ አፍሪካ-ሓቢሪካ ምንባር ውድድር ኣበይ ተካሄደ	ደቡብ አፍሪካ	Correct	Correct
ናይ ፖላንድ ዋና ከተማ መን ትብሃል	ዋርሶ	Correct	Correct
መበል ሳላሳን ሓደን ኦሊምፒክ ኣበየናይ ሃገር ይካየድ	ሪዮ ዲ ጃኔሮ ከተማ	Correct	Correct
ዮፍታሄ ንጉሴ ኣበይ ተወሊዱ	ጎጃም	Wrong	Wrong
ኣቡነ ጴጥሮስ ኣበይ ተወሊዱ	ፍቸ	Correct	Correct

ናይ ኡጋንዳ ዋና ከተማ መን ትብሃል	ካምፓላ	Correct	Correct
ቀዳማይ ናይ ምብራቅ አፍሪካ ናይ መናእሰይ ናይ ሙያ ሻምፒዮን ኣበይ ተካይዱ	አዲስ አበባ ከተማ	Correct	Correct
ናይ አርቲስት አስናቀች ወርቁ ስርዓተ ቀብር ኣበይ ተፈፂሙ	መንበረ ፀባኦት ቅድስት ስላሴ ደብሪ	Correct	Correct
ናይ እሥራኤል ናይ ዓይን ሐኪማት ነፃ ሕክምና ዝሃቡሉ ኣበይ ነይሩ	አዲስ አበባ	Correct	Correct
ገረብ ሱባ ኣበይ ይርከብ	ሆለታ	Correct	Correct
ናይ ምብራቅን ማእኸላይ አፍሪካ ዋንጫ ኣበይ ተኻይዱ	ኡጋንዳ	Correct	Correct
ናይ አማራ ብሔራዊ ክልላዊ መንግሥቲ ርዕሰ ከተማ መን እያ	ባህር ዳር	Correct	Correct
ናይ ቦትስዋና ዋና ከተማ መን እያ	ጋብሮኒ	Correct	Correct
ቤትሆቨን ኣበይ ተወለዱ	ጀርመን	Wrong	Wrong
ናይ ሑቡራት መንግስታት ድርጅት ብኣብዝሓ ወፃኢ ዝትሸፍን ሃገር መን እያ	አሜሪካ	Correct	Correct
ናይ አቡነ ጳውሎስ ናይ ቀብር ስነስርዓት ኣበይ ተፈፂሙ	ኣብ መንበረ ፀባኦት ቅድስት ስላሴ ካቴድራል	Correct	Correct
ድሬቲዮብ ናይ ቪዲዮ ድረገጽ ክንደይ ዝኣክል ቅርሺ ወፅኢዎ	ክልተ ሚሊዮን ብር	Correct	Correct
ናይ አንስቴቲስቶተ ማሕበር ክንደይ አባላት አለዉዎ	ካብ 300 ንላዕሊ	Correct	Correct
እብ አዲስ አበባ ልምዲ ሲጋራ ኣብ ክንደይ ደቂ ተባዕትዮ ይርኣይ	ሸሞንተ ካብ ሚኢቲ	Correct	Correct
ከተማ ዓድዋ ካብ አዲስ አበባ ብክንደይ ኪሎ ሜትር ርሕቀት ትርከብ	1066 ኪሎ ሜትር	Correct	Correct
ቀዳማይ ሚንስተር መለስ ናብ በረኻ ዘተፀመበሉ ብክንደይ ዓመቶም እዩ	ብ ዒስራ	Wrong	Correct
ናይ ዓለም ሎሬት ሜትር አርቲስት አፈወርቅ ተክሌ ብክንደይ ዓመቶም ሞይቶም	ብ 80 ዓመቶም	Correct	Correct
ሕቡራት መንግስታት ክንደይ አባል ሃገራት ኣለዉኦ	192 ሃገራት	Correct	Correct
ኣብ ኣድስ ኣባባ ክኘደይ ዝኣኽሉ ናይ ኤፍ ኤም ሬዲዮ ጣቢያታት ኣለዋ	ዓሰርተ	Correct	Correct
ናይ ኢንተርፖል አባል ሃገራት ብዝሓት ክንደይ እዩ	188 ሃገራት	Correct	Wrong
ደሌ ቢራን ሄኒከንን ንኢትዮጵያ ብሔራዊ ቡድን ክንደይ ቅርሺ ስፖንሰር ገይሮም	ከብ 24 ሚሊዮን ንላዕሊ	Correct	Correct
ኣብ ሎካርቢ አውሮፕላን ነቷጊ ፍንዳታ ክንደይ ዝኣኽሉ ሰባት ሞይቶም	270 ሰባት	Correct	Correct
ኣብ መበል 30 ናይ ለንደን ኦሎምፒያድ አፍሪካብምውካል ክንደይ ሃገራት	53 ሃገራት	Wrong	Correct

ተሳቲፈን			
ኣብ ዓመት ክንደይ ዝኣክላ እብ ገጠር ዝኑብራ ደቂ ኣንስትዮ ንገዛ ሰራተኝነት ናብ ኣድስ ኣባባ ይመፅኣ	10,000 ደቂ ኣንስትዮ	Correct	Correct
ከተማ ሐረር ካብ አዲስ አበባ ክንደይ ኪሎ ሜትር ትርሕቅ	ብ 526 ኪሎ ሜትር	Correct	Correct
ካብ ዓለም ክንደይ ዝኾኑ ሰባት በቢዓመቱ በስትሮክ ምኽንያት ይሞቱ	ሽድሸተ ሚሊዮን	Correct	Wrong
ናይ ቅዱስ ያሬድ ዜማ ምልክታት ብዛሓት ክንደይ ይኸውን	ዓሰርተ	Correct	Correct
ክንደይ ዝኾኑ ናይ ኤርትራ ተፃወቲ ዕቁባ ሓቲቶም	17 ተፃወቲ	Correct	Correct
አሰላ ናይ ቡቅሊ ፋብሪካ ክንደይ ኩንታል ናይ ቢራ ስገም ካብ ወፃኢ ገዚኡ	ካብ 132 ሽሕ ንላዕሊ	Correct	Correct
ካብ ሁምቦ አርባ ምንጭ ዘሎ መንገዲ ክንደይ ኪሎ ሜትር ይርሕቅ	106 ኪሎ ሜትር	Correct	Correct
ናይ አፍሪካ ሕብረት ህንፃ ክንደይ ዶላር ወፂኡሉ	200 ሚ .	Correct	Correct
ሐበሻ ሲሚንቶ ንደቡብ አፍሪካ ኩባንያታት ክንደይ ዝኣክል ዶላር አክሲዮን ሸይጡ	21 ሚሊዮን ዶላ ር	Wrong	Wrong
ሊፋን ሞተርስ ክንደይ ዝኣኽሉ ናይ ውሽጢ ዓድ ግብረሰናይ ድርጅታት ንምርዳእ ወሲኑ	ክልተ	Correct	Correct
ኣብ ጎንደር ከተማ ዝርከብ ደስታ ቤት ሲኒማ ክንደይ መቀመጢ ኣለዎ	250 ወንበር	Correct	Correct
ደራሲ ስብሃት ገብረ እግዚአብሄር ከዚህ ዓለም ብሞት ዝተፈለየ ብክንደይ ዓመቱ እዩ	ብ 76 ዓመቱ	Correct	Correct
አርቲስት አስናቀች ወርቁ ብክንደይ ዓመታ ከዚህ ዓለም ብሞት ተፈሊያ	ብ 78	Correct	Correct

Signed Declaration Sheet

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

Declared by:

Name: _____

Signature: _____

Date: _____

Confirmed by advisor:

Name: _____

Signature: _____

Date: _____