JIMMA UNIVERSITY
INSTITUTE OF TECHNOLOGY
SCHOOL OF COMPUTING

Sentence prediction system for Afan Oromo text

Hunde Yegezu Guta

**A Thesis Submitted To The Department Of Information Technology
In Partial Fulfillment For The Degree Of Master
Of Science In Information Technology**

Jimma, Ethiopia
12/17/2020

JIMMA UNIVERSITY
INSTITUTE OF TECHNOLOGY
SCHOOL OF COMPUTING

Hunde Yegezu Guta
Advisor:Getachew Mamo(PHD)

This is to certify that the thesis prepared by Hunde Yegezu Guta, titled: Sentence prediction system for Afan Oromo text and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Information Technology complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by Examining Committee

Name	**Signature**	**Date**
____________ (Advisor)	____________	____________
____________ (Co-Advisor)	____________	____________
____________ (Program Chairman)	____________	____________
____________ (External Examiner)	____________	____________
____________ (Internal Examiner)	____________	____________

ABSTRACT

Data entry is a core aspect of human computer interaction. Sentence prediction is one of data entry systems to a computer and other hand held electronics device. It is a process of guessing the phrases or words which are likely to follow in a given text segment by displaying a list of the most probable sentences that could appear in that position. Sentence prediction assists physically disabled individuals who have typing difficulties, speed up typing speed by decreasing keystrokes, helps in spelling and error detection and it also helps in speech recognition and hand writing recognition. In this study, Sentence prediction model is designed and developed for Afan Oromo. We used RNN-LSTM for deep learning algorithm. Initially, the training corpus and user inputs are tokenized and analyzed. Subsequently, RNN-LSTM model is built on pre processed and cleaned data by using training corpus. After all the model is tested by input provided from the users. The developed model is evaluated based on Keystroke Saving (KSS) performance evaluation metrics.According the evaluation result 18.75% KSS is achieved based on the average of KSS of four testers.Therefore,Recursive neural network has good potential on sentence prediction for Afaan Oromo.

Keywords: *Sentence prediction,ANN, Recursive neural network, Keystroke Saving*

Acknowledgment

First of all, I would like to thank my Lord Christ Jesus for entitling me to this opportunity. I am deeply grateful to my advisor Dr. Getachew Mamo for his concern, constructive comments, supervision, and encouragements on my work. Finally, I thank all my families and friends for their continuous motivation and encouragement during my stay in the university.

CONTENTS

1. *Chapter one* . . . 1
   - 1.1 Introduction . . . 1
   - 1.2 Motivation . . . 4
   - 1.3 Statement of the problem . . . 4
   - 1.4 Objectives . . . 5
     - 1.4.1 General objective . . . 5
     - 1.4.2 Specific objectives . . . 5
   - 1.5 Methodology of study . . . 6
     - 1.5.1 Literature review . . . 6
     - 1.5.2 Data collection methodology . . . 6
     - 1.5.3 Design and experimentation . . . 6
     - 1.5.4 Testing and evaluation . . . 6
   - 1.6 Related work . . . 7
   - 1.7 Scope of the study . . . 8
   - 1.8 Application result . . . 8
   - 1.9 Organization of the thesis . . . 8

2. *CHAPTER TWO: Literature review* . . . 10
   - 2.1 Introduction . . . 10
   - 2.2 Overview of Afaan Oromo . . . 10
   - 2.3 Important linguistic characteristics of Afan Oromo . . . 12
     - 2.3.1 Afaan Oromo writings . . . 12
     - 2.3.2 Afaan Oromoo alphabets . . . 12
   - 2.4 Word class in Afaan Oromo . . . 13
     - 2.4.1 Afaan Oromoo noun class . . . 13
     - 2.4.2 Verb class in Afaan oromoo . . . 16

2.4.3 **Adjective class in Afaan oromoo** 17

2.4.4 **Adverb class in Afaan oromoo** 19

2.4.5 **Pre- and Post- Position class** 20

2.5 **Structure of Afaan oromoo phrases** 22

2.5.1 **Criteria for Afaan Oromoo phrase** 22

2.5.2 **Afaan Oromoo phrase classes** 23

2.6 **Sentences in Afaan oromoo** 27

2.6.1 **Types of sentences** 28

2.7 **Text predictive system** 31

2.8 **Sentence prediction** 33

2.8.1 **Approaches to text predictive system** 37

2.8.2 **Evaluation metrics for text predictive system** . . 38

2.9 **Summery** 39

*3. **CHAPTER THREE:Related works*** 41

3.1 **Introduction** 41

3.2 **Prediction system for local languages** 41

3.2.1 **Word sequence prediction for Afaan Oromoo language** 41

3.2.2 **Word sequence prediction for Amharic language** 42

3.2.3 **Enhancing the Text Production and Assisting Disable Users in Developing Word Prediction and Completion in Afan Oromo** 42

3.2.4 **Ethiopic Keyboard Mapping and Predictive Text Inputting Algorithm in a Wireless Environment** 43

3.2.5 **Implementation of Online Handwriting Recognition System for Ethiopic Character Set** 43

3.3 **Text prediction system for foreign languages** 45

3.3.1 **Bangla Word Prediction and Sentence Completion Using GRU** 45

3.3.2 **Dependency language models for sentence completion** 46

3.3.3 **Automated word prediction in Bangla language using stochastic language model** 46
3.3.4 **A Neural Language Model for query Auto-Completion** 47
3.4 **Summery** 47

4. ***CHAPTER FOUR:Afaan Oromo sentence prediction system*** 48
4.1 **Introduction** 48
4.2 **Architecture of Afaan Oromoo sentence auto completion** 48
4.3 **Language model** 50
4.3.1 **Afaan Oromoo corpus** 50
4.3.2 **Prepossessing module** 50
4.3.3 **RNN-LSTM training modules** 51
4.3.4 **Trained model** 52
4.3.5 **Summery** 52

5. ***CHAPTER FIVE: Experimentation, Prototype and Evaluation*** 53
5.1 **Experimentation** 53
5.1.1 **Data set** 53
5.1.2 **Tools and environmental setup** 53
5.2 **Result** 54
5.3 **Discussion** 55

6. ***CHAPTER SIX:Conclusion, Contribution and Future work*** 56
6.1 **Conclusion** 56
6.2 **Contribution** 57
6.3 **Future work** 57

LIST OF FIGURES

4.1 Afaan oromoo sentence prediction architecture. 49
4.2 How RNN works. 52

Acronyms

BERT.....Bidirectional Encoder Representations from Transformers

RNN.....Recursive Neural Network

LSTM.....Long Short Term Memory

KSS.....Key Stroke Saving

KUC.....Keystroke Until Complete

HR.....Hit Rate

IS.....Information Source

GPU.....Graphics Processing Units

GRU.....Gated Recurrent Unit

ANN.....Artificial Neural Network

1. CHAPTER ONE

1.1 Introduction

One of the most important things to transfer idea and understand what the writer or speakers intends to say are languages because the writer and reader or the speaker and listener share the same opinion. Each language offers a rich and unique insight into different ways of thoughts and lives as well as into the history of the myriad of cultures and peoples across the world [4].Most people who have access to computers in Africa, especially in Ethiopia tend to be educated in and socialized to some degree to use the official languages and thus less likely to actively seek to use their first languages in digital world even at regional level. But in Ethiopia specifically in Oromia region Afaan Oromoo is used as official language that means people who live in Oromia state uses their mother tongue as official language. According to the Ethiopian census of 2007, the first languages and the largest is the Afaan Oromo with about 24,930,424 of the population speaking the language which is equivalent to 33.8% and the Amharic with 21,634,396 users which is equivalent to 29.3% of the country's population[5].Ethiopia is one of a country with linguistically, ethnically and culturally diversified nation with more than 84 different languages spoken in the country. Out of 84 languages afaan oromoo is spoken by 34.4% percent of nation of the country [6].Afaan oromoo is spoken in Horn of Africa in country like Kenya and Somalia.

Natural language processing (NLP) is a sub field of Artificial Intelligence which deals with the ability of computer systems to analyze and synthesize spoken and written languages as human beings. It is (NLP) is concerned with the progress of computational models of human language processing. It is an interdisciplinary research area at the border between linguistics and artificial intelligence aiming at developing computer programs by using natural language text or speech according to the linguistics rules[1].NLP is making computers to perform useful tasks like enabling human-machine communication, improving human to human communication, or simply doing useful processing of text or speech [2]. NLP researchers aim at gathering knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks [3].The objectives Natural language processing is designing and building software/application that will analyze, understand, and generate human languages, so that eventually humans will be able to communicate with the machines using natural language like Afaan Oromo,Amharic ,English and etc. Some of applications that makes use of NLP to allow users interact with computer systems using natural languages are machine translation, information extraction and retrieval using natural language,text-to-speech synthesis, automatic written text recognition, grammar checking, named entity recognition, sentiment analysis, parsing, chunking , part-of-speech tagging (POS) and predictions like character prediction ,word prediction, word sequence prediction and sentence prediction. The researchers of NLP use natural language technologies to develop computer applications for natural language. Among this NLP investigation area or application, we a mainly focuses on sentence prediction or sentence auto completion.

Data entry is a core aspect of human computer interaction. Images, documents, music, and video data are entered to computers in order to get processed. Data entry can be through the use of keyboard, or other means. In most case, processing text using editor needs knowledge of the language to be used to speed up processing.So in this digital world many people face great difficulties in processing the documents while entering data to digital device like computer ,cellphone and so. But the computer has many possibilities to help people with these difficulties.Most difficulties comes from incorrect spelling, need of choosing between different words and need of grammatical mistakes related to that specific language. Therefore it is very necessary to think about the program with function of spell corrector, word predictor and grammar checker that help people to get rid of those difficulties. Research findings in Russia show that around 8% of the population has serious, specific reading and writing difficulties. Now we are in the era in which natural language processing plays a crucial role in processing natural language such predicting word, completing sentence or phrases when the writer try to produce the text document using one of data entry techniques. According to [7, 8] sentence auto completion is one parts of artificial intelligence system that predicts one or more words or phrases or sentences as alternative to user inputs when user input or write the first letters or letters of words. When the user starts to write fragment of the sentence, the system will provide list alternative words to complete that sentence [9]. Unless and otherwise the system gives the user no alternative words, the writer should input the next letter of the word or the next word of a phrase for the program to produce other set of words or phrases. Now a day smart phone and other device can complete partially written word to help the user input and speed up communication using digital device and to create document or text using mobile input interface. So many scholars agree that sentence prediction can play a crucial role in assisting the disabled just to improve the performance of typing and help them to get rid of misspelling [9]. For Afaan Oromoo Somali and Amharic, text predict entry system has designed [10, 11, 12].Beside this context-sensitive sentence auto completion has studied and designed for Amharic [13].But for Afaan Oromo it is limited to word prediction level using statistical approaches. Basically word prediction and sentences prediction tackles the same problem. But if appropriately designed and developed sentence prediction is the right choice.Even though sentence prediction is the most important tool to help the developer and the user, in domain of natural language, it is necessity remains unsound in our country for local languages.Thus why this work give consideration to this area.

1.2 Motivation

Afan Oromo uses Latin-based-script as writing system.People who uses Afan Oromo writes many letters compared to peoples who writes other languages that makes text entry process slower for Afan Oromo.Though sentence prediction is remained unsolved problem in domain of natural language processing most researchers is not paying attention to this area for local languages.As per our knowledge there is no attempted research work on sentence prediction for Afan Oromo which motivated as to makes this predictive system hot research area for local languages.

1.3 Statement of the problem

Afaan Oromo uses Latin character as symbol for data entry purpose.It is independent language with its own syntactical structure and semantic meaning.When writing Afan Oromo, many character need to be put to gather.Thus why typing Afan Oromo slow down document entry process in this digital world.For example the word **"kottu"** has five character in Afan Oromo but in Amharic it has only one character.Thus why typing in Afan Oromo is time consuming.Because of this it is difficult for many disabilities to process document in Afan Oromo using digital devices.Besides so,most people commits spelling error in Afan Oromo words which conveys completely different meanings if not spelled correctly For instance the word **"daboo"** and **"daabboo"** gives us different meaning because of their character arrangement.The first one means a group of peoples come to gather to help each other in doing something, in the second one , it is something to eat.Thus why committing spelling error is one of the difficulties in text entry process for Afan Oromo.

Researches is conducted on character prediction, word prediction and word sequence prediction for Afan Oromo to address the above stated problems.Researchers on[34] proposed and implemented word sequence prediction for Afan Oromo.They used statistical approaches like N-gram to develop the model.Researchers' on [36] also conducted research on Enhancing the Text Production and Assist-

ing Disable Users in Developing Word Prediction and Completion in Afan Oromo to address the above stated problems.This researcher also used statistical approaches Uni-gram, bi-gram and tri-gram with corpus collected from different sources.But those previously conducted researches are limited to only word sequence prediction.But the need for prediction in text entry process is not limited to word sequence.For example a disabilities need a system that can predict more than a word and help them in text entry process. Even normal people needs a system that can suggest them a sentences that they are intend to write.Thus , this work focus on sentence prediction for Afan Oromo.
After investigating the problem of sentence prediction,we will try to answer the following questions:-

1. How sentence prediction will be developed for Afan Oromo?
2. What is the solution to drawbacks of word sequence prediction?

1.4 Objectives

1.4.1 General objective

The general objective of this research is to design and develop sentence prediction model for Afan Oromo.

1.4.2 Specific objectives

To achieve the above mentioned general objective the following specific objectives will be carried out.

- literature review on structure of Afan Oromo,sentences prediction, approaches to sentence prediction will be conducted
- Review related works on word sentence prediction for other languages and supplementary researches like word sequence prediction,word prediction conducted on Afan Oromo just to spot out approaches for this study

- Collect representative corpus for training and testing the model.
- Collected corpus will be analyzed.
- Build the POS language models.
- Evaluate the performance of developed sentence prediction model

1.5 Methodology of study

In this section methodology applied to hit general objectives will be discussed in brief

1.5.1 Literature review

Researches and related works will be thoroughly reviewed to grasp a firm knowledge with the intention of developing appropriate sentence prediction model for Afan Oromo. Afan Oromo and morphology,Sentence prediction, Sentence prediction approaches, are some of the works that will be reviewed while conducting this research.

1.5.2 Data collection methodology

Because there is no ready-made corpus for Afaan Oromo; for purpose of development and training data will be collected from online data source of different classes like Sport, politics, health, education and etc

1.5.3 Design and experimentation

Different implementation tool will be used to tackle specified problem. For example python will be used as programming language and Google Co laboratory will be used as developmental environment.

1.5.4 Testing and evaluation

Testing will be conducted by inviting peoples from different classes. They may be allowed to enter whatever they want to write and tested according to their response.Evaluation metrics will be KSS.A Keystroke Saving (KSS) estimates saved effort percentage and is calculated through comparison of total number of keystrokes needed to type a text (KT) and effective number of keystrokes using word prediction.

1.6 Related work

Many works have been done on word prediction. Research has been conducted on word prediction for language like Hindi, English, and Bang lash and Hebrew, but till now research conducted on sentence completion is few in number. As per their necessity below is some of the work related to sentences auto completion.

In [8] Query auto completion is an important feature saving users many keystrokes from typing the entire query. In their work they studied the problem of query auto completion that tolerates errors in users' input using edit distance constraints. The work propose a novel neighborhood generation based algorithm, IncNGTrie, which can achieve up to two orders of magnitude speedup over existing methods for the error-tolerant query auto completion problem.

The other work [9] context-sensitive query auto completion is the work on the completion of user query even though the prefix enters is very small. Query auto completion is known to provide poor predictions of the user's query when her input prefix is very short (1 or 2 characters).They used Nearest Completion which outputs the completions of the user's input that are most similar to the context queries.

Word prediction [7] like Enhancing the Text Production and Assisting Disable Users in Developing Word Prediction and Completion in Afan Oromo also studied for Afan Oromo. The algorithm that used in this work was N-grams algorithms (Unigram, Bigram and Trigram) for auto completing a word by predicting a correct word in a sentence which saves time, reduces misspelling, keystrokes of typing and assisting disables. This work describes how we improve word entry information, through word prediction, as an assistive technology for people with motion impairment using the regular keyboard, to eliminate the overhead needed for the learning process [3].But this work is limited to word level prediction and they also didn't considered the context of the word.

1.7 Scope of the study

This study does not consider the semantic information of the language. This research will be undertaken with the aim to model sentence prediction for Afan Oromo based on trained RNN model.

1.8 Application result

The result of this work will be applicable to many days to day activity related to digital text processing. The result will be applied to the hand held device like mobile phone for speeding the text entry process and to tackle issues related to poor spelling and also problem related to incorrect grammar. In the same manner it's applicable for computer users when creating document using computer .As text processing is now the main activities of daily life, the result of this work will be there for disabled and normal personnel. Moreover, as development tool it can be used in speech recognition system, for recognizing unknown word using context of the sentences.

1.9 Organization of the thesis

The rest of this thesis is organized as follows. In Chapter 2, literature review briefly states fundamental concepts of structure Afan oromo languages and its grammatical rules, sentence prediction and methods of sentence prediction. grammatical rules. Chapter 3 presents researches conducted by different scholars on the topic of word sequence prediction, their approach, and findings. In Chapter 4, architecture of the proposed sentence prediction model, its approach, and related concepts are clearly explained. Experiment is presented in Chapter 5. Finally, conclusion and ,recommendation and future work are stated in Chapter 6.

2. CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

To conduct research one needs to take the advantages of the knowledge accumulated previously.Research cannot be undertaken in isolation of the work that has been done before on that problem directly or indirectly.According to [14] one of the important step in planning research study is careful review of research journal,books, dissertation,thesis and other sources of information on that problem.So as per statement the researcher reviewed is searched for other previous work on this area.Accordingly, this chapter primarily deals with the Afaan Oromoo grammer ,Afaan Oromoo sentence structure and state-of-the-art, relating to text auto-completion techniques and word predictions, which is the concern of this paper.

2.2 Overview of Afaan Oromo

Afaan Oromo is one of the most widely spoken languages in Africa, surpassed only by Arabic and Hausa [15]. The language is termed as, 'Afaan Oromoo' (Oromoo Language) because it is used by Oromo Society, the native ethnic group of Ethiopia that account for the largest population of the country. With regard to this, various scholars revealed that the Oromo People are the largest single ethno-nation in Eastern Africa [16], constituting at least 40% of the Ethiopian population [17]. According to Hussein [18], "The Oromo people speak Afaan Oromo (the language of Oromo), which belongs to the Eastern Cushitic family of Afro-Asiatic phylum." Studies reveal that Afaan Oromo is the most important language of Ethiopia where it is used not only as a national (official) language by the Oromo people but also as a lingua franca by several million speakers of other languages[19]. Outside Ethiopia, the language is spoken by thousands of other Oromo tribes in Kenya [18]. In line with this, Dejene [19] states as, "It is a language of a great people with national history going back at least to the 16th century that played a major political and cultural role in North-East Africa and whose cultural and social organization (e.g. the famous 'Gada' system) are among the most outstanding in Africa". Besides being the widely used language in Africa, Afaan Oromo has been included among the essential languages in the world. Justifying this, the report by the U.S Government and its Education Department (1985) has revealed that Afaan Oromo has been considered as one of the 169 critical languages of the world [20]. Based on the aforementioned historical and demographic issues of Afaan Oromo, researching the different aspects of the language is worth mentioning. Accordingly, the current research focuses on NLP for Afaan Oromo is among the basic issues that need to be studied in order to enhance the wide usage of the language during the current digital age when the use of technological resources is increasing dramatically. The study involves conceptualizing the writing system of the language including its alphabets and sound systems as well as describing its syllabification, morphological process and grammatical rules as these issues are the basics for studying the sentence prediction systems.Afaan Oromoo has many dialects because of geographical location.Even though there exist strong similarities between those dialects they also reflects some form of varieties.These varieties may be based on certain forms of pronoun,verb conjugation and colloquial lexicon.The major dialects of Afaan Oromoo are west dialects or Welega dialects mainly spoken in Ilubabor,Jimma and Welega ,Tulama dialects mainly spoken in north,west and east shewa,Wello mainly spoken in Walloo northern shewa and southern Amhara, Arsi dialects in Arsi and Bale zones,Borena dialects in southern most zone by the name.

2.3 Important linguistic characteristics of Afan Oromo

2.3.1 Afaan Oromo writings

Different scholars identified that the writing system Afaan Oromo relies on Latin Script; the alphabets and sounds of the language are modifications of Latin writing system. Thus, Afaan Oromo shares a lot of features with English writing system except some modifications, and the writing alphabet of the language is known as 'Qubee Afaan Oromoo' which is designed based on the Latin script. Thus, letters in English or Latin Alphabets are also found in Afaan Oromoo except the ways they are combined in phonetic alphabets and the styles in which they are uttered[2].

In order to get some insights into the writing system of Afaan Oromo, it is very important to lookthrough the alphabets and sound systems of the language; hence, the interplay between alphabets and sound systems of the language could be described below.

2.3.2 Afaan Oromoo alphabets

As it has been mentioned before, Afaan Oromo uses Latin character but with some modifications on sound of consonant and vowels. It has 28 letters called Qubee. However, later on a new letter ”Z” was included in the alphabet as there are words which require the letter. For example: ’Zelaya’ (gold), ’Zeeytuuna’ (guava), ’Azoole’ (river in Arsii), ’Zeekkara’ (Opera), ’Zalmaaxaya’ (mess), ’Waziiza’ (fire place or fire work) and Zawii (insanity) are Afaan Oromo words that can’t be written without letter ’Z’.Additionally ’P and V ’ are also added. ’P and V’ letters are not Afaan Oromo letters because there is no Oromo word written by use of either of them. But they are included by considering the fact of handling borrowed terms from other languages like English. For example: ”Police”, ”Piano”, ’Television’, ’video’ and etc. To sum up there are 31 letters of Afaan Oromo including ”Z ”, ”P ”, and ”V ”[21].

Vowels (Dubbachiiftuu)

There are five vowels in Afaan Oromo; these are ’a’, ’e’, ’o’, ’u’ and ’i’. They are similar to that of English, but they are uttered differently. Each vowel is pronounced in a similar way throughout its usage in every Afaan Oromo literature [21]. In other words, there is no rule which could violate their pronouncing style in difference contexts. There is no need to deal with phonetic transcription because the pronunciation is made just as the words are normally written in different texts.

Consonants(Dubbbifamoota)

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

2.4 Word class in Afaan Oromo

Word class in Afaan Oromo is some somewhat different from English word class. According to the [22, 23], Based on the context and form, in Afaan Oromo there are five types of main word classes. Those are noun, adjective, adverb, verb and pre and post position. The classifications of word types are almost common in most of the linguistic books[22, 24]. According to [25] and linguistic experts define, there is five word class that are head to phrases.

2.4.1 Afaan Oromoo noun class

Like English and in other languages, Afaan Oromo nouns are words used to name or identify any of a class of things, people, places, ideas or a particular one of these. For this study, the Afaan Oromo noun class is considered to consist of nouns and pronouns.
Example:**Tolaan** kaleessa dhufe. (Tola came yesterday)
In above sentence the underline word is noun. The reason word ”Tolaa” to be noun is position occupied. This gives a conclusion to some of subject word

in Afaan Oromo sentence is noun. Moreover, two numbers are recognized in Afaan Oromo nouns: singular and plural. Except a noun ended by -n in Afaan Oromo; A singular noun is marked by zero morpheme whereas a plural noun is marked by various forms. But those ended by -n is the same for singular and plural.

The plural forms are not used as often in Oromo as they are in English. Typically, the plural form is used to specify that one is talking about more than one object where no other indicators are given.For example, in conversation the plural is rarely used when the noun is modified by a number. One would say **Hoolaa lama** for **two sheep**, keeping hoolaa [sheep] in the singular, instead of **two sheeps**, where hoolota is the plural form of hoolaa. When a plural noun is modified by an adjective, only the adjective shows plurality (discussed in next). In Afaan oromoo written text, plural forms tend to be more common, and may occur with numbers, adjectives, and other indicators.

When the plural form is used, there are several forms it may take. Typically, the final vowel is dropped and the correct suffix attached: -oota, -toota, -lee, -een, -yyii, -wwan, -ootii, or -olii. Unfortunately, the correct suffix cannot be predicted from the noun, meaning plural forms must be learned individually. Plural forms also vary across dialects, and multiple forms may be correct for some words. The most common suffix is -oota.
For example:

Singular	Plular	English
Ilka	Ilkaan	Tooth
Guyyaa	Guyyoota	Day
Kitaaba	Kitaabolee	Book
Waggaa	Waggoota	Year
Wanta	Wantoota	Thing
Laga	Lageen	River
Gaara	Gaarota	Gaarreen
Muka	Mukoota	Tree

For nouns that may take either a masculine or feminine form, the feminine form is used as the stem to which the plural suffix is attached. For example, the plural of **barataa/ barattuu** [student] is barattoota. Many nouns have irregular plural forms (e.g. biraa [another] while [biroo] others)

Derived nouns in Afaan Oromoo

As expert define in [22, 24] noun can be derived from non-noun or other noun words by affixing nominal affixes. Those derived nouns are created in three different ways.

A.Noun drived from adjectives

Let we see the following example how nouns are derived from adjectives.

Base word	**Nominal Affixes**	**Drived noun**
Bal'aa	(-ina)	Bal'ina
Gabaabaa	(-ina)	Gabaabina
Adii	(-eenya)	Addenya
Jabaa	(-ina/-eenya)	Jabina/jabeenya

B.Noun drived from base noun in Afaan oromoo

In Afaan oromoo ,noun can be derived from base noun.see the following example.

Base word	**Nominal Affixes**	**Drived noun**
Guyyaa	(-saa)	Guyyaasaa/Guyyee
Galgala	(-oo)	Galgaloo
Waaree	(-iyoo/-tee)	Waariyoo/Waaritee

2.4.2 *Verb class in Afaan oromoo*

The main feature of the Afaan Oromo verbs is that any word that comes at the end of grammatically complete Afaan Oromo sentence are verbs[22]. They are independent words. As a a result of this feature a word at the end of complete sentence are expected to be tagged as a verb by an Afaan Oromo tagger.

For example in following sentences;

- Hundeen kaleessa **dhufe**

- Dammituun dammma **nyaatte**
- Dachaasaan kofoo **bite**

In above sentences the underlined words like dhufe, nyaatte, bite are verbs.All these word put the sentence to end because of their feature.

2.4.3 Adjective class in Afaan oromoo

In Afaan Oromo words that come after noun or pronoun to modify noun or pronoun is adjectives.
Example: Jabeessaan sangaa diimaa bite.
Jabesa bought red bull] In above sentence the word diimaa declares type of color of [bull] Jabeessa bought.As we notice in above sentences unlike English adjectives Afaan oromoo adjective do not come before the noun they modify instead they come after the word they modify.

Afaan Oromo adjective can be male, female, or neutral. Masculine adjective are used with masculine noun, feminine adjective modify feminine nouns, and neutral adjectives can be used with any noun. All non-neutral adjective can be made masculine or feminine by attaching the appropriate suffix. Masculine suffixes for adjectives are: -aa, -aawaa, -acha, and -eessa. Feminine suffixes are: -oo, -tuu, -ooftuu, and -eettii. Standard morphology rules apply when attaching suffixes. For example:

Masculine	**Feminine**	**English meaning**
Gurraacha	Gurraatti	Black
Boosacha	Boosettii	Messy
Xinnaa/xiqqaa	Xinnoo/Xiqqoo	Small
Mi'aawwaa	Minyooftuu	Sweet
Bareedaa	Bareedduu	Beautful
Hiyyeessa	Hiyyeettii	Poor
Godeessa	Godeetti	Skinny

Neutral adjectives (e.g. **adii** —"white") use the same form for both masculine and feminine nouns. When adjectives are used to modify a noun, typically the noun remains in the singular and number is shown by the adjective only. Plural adjective are formed by repeating the first syllable.
Example:

Singular	**Plural**	**English**
Gogaa	Gogoggaa	Dry
Adii	Adaadii	White
Bareedduu	Babbareedduu	Beautiful

Some masculine adjectives will change their ending to -oo when pluralized. Some of these do not repeat the first syllable as a plural marker.

Examples:

Singular	**Plural**	**English**
Guddaa	Guguddoo	Large
Olaanaa	Olaanoo	High
Beekaa	Beekoo	Knowledgeable
Cimaa	Ciccimoo	Strong

In written Oromo, the noun may be pluralized as well as the adjective, so that **'nama sosoressa lama'** and **'namoota sossooreyyii lama'** are correct ways to say [two rich people]. In conversational Oromo, the first method, keeping the noun in the singular, is more common.

2.4.4 ***Adverb class in Afaan oromoo***

In Afaan Oromo, adverbs are used to modify the coming verbs. Adverbs always come before the modified verb but it should be noted that any words come before verbs cannot be always considered as an adverb.
Example :Tolaan daddafee dhufe.
In above example a verb dhufe is preceded by an adverb(daddafee) that modifies it.Here everybody should note that not every word that comes before verb is necessarily an adverb.

In their nature, adverbs can be found either in their primitive form or compound form as grouping of preposition and other word categories. Adverbs indicate manner, time, place, cause, or degree and answers questions such as akkamitti [how], yoom [when],eessa [where]. The primitive adverbs are very few in number and these are: haamayyu [yet], daddafii [quickly] and etc.

2.4.5 Pre- and Post- Position class

Prepositions give meanings only if they combine with other words such as noun, adjective, verb, etc, unless they have no meaning. Pre- and post- positions link with nouns, pronouns and with other words in a sentence. The main properties of pre- and post- positions are: they never use affixes and they don‘t assist to form other words. According to , A preposition links a noun to an action (e.g.,”achirraa deemi”) or to another noun (’Qarshiin minjaalarraa jia’). For the purpose of clarity, this section will divide Oromo prepositions into two categories: prepositions and postpositions, with prepositions coming before the noun and postpositions coming after the noun they relate to.

Some common prepositions and post position in Afaan Oromoo

Post position	Preposition
’ala‘ — [out, outside]	’gara’ — [towards]
’bira’ — [beside, with, around]	’eega’, ’erga’ — [since, from, after]
’cinaa’ — [beside, near, next to]	’haga,’hanga’ — [until]
’dur, dura’ — [before]	’hamma’ — [up to, as much as]
’duuba’ — [behind, back of]	’akka’ — [like, as]
’itti’ — [to, at, in]	’waa’ee’ — [about, in regard to]
’booda’ — [after]	

Example:

boqonnaarra (boqonnaa irra) – [on vacation]

mana keessa – [in the house]

waaree booda – [afternoon]

irra deebi’i – repeat [lit. return on it]

mana nyaataa kanatti – [at this restaurant]

waa’ee fiilmii sun natti himi – [tell me about that film]

Chaaltuun akka Hawwiituu baratuu dha. – [Chaltu is a student like Hawitu.]

hanga torban dhufu – [until next week]

gammachuu wajjin – [with pleasure]

shaayee annan malee – [tea without milk]

Ani meetirii lama gadi. – [I am below (shorter than) 2 meters.]

Keeniyaan Itoophiyaarraa (gara) kibbatti argamti – [Kenya is located (to the) south of Ethiopia]

From the examples above, you may notice that the postpositions itti, irra, and irraa most often occur as suffixes, -tti, -rra, and -rraa, on the nouns they relate to. Often with place names, no preposition or postposition is used to be mean ”in”. Therefore, one can say ’Mattuu jiraatta’ for [you live in Mattuu], or ’hospitaalan ture for [I was in the hospital], using no preposition. Personal pronouns are not used with prepositions. Instead, possessive pronouns are used as personal pronouns.

Examples:

gara koo(tti) (it not similar ’gara na’] [toward me]

akka keenya — [like us]

akka isaatti — [according to him]

waa’ee kee — [about you]

Postpositions, on the other hand, take the accusative form of personal pronouns.

Examples:

sitti — [at you]
narraa —[from me]
isa jala —- under him

When an adjective modifies a noun, the postposition follows the adjective, as in 'Nama guddarraa' [from the big man].

2.5 Structure of Afaan oromoo phrases

A phrase is a structure in a language which is constructed from one or more words in the language. Phrases are composed of either only head word or other words or phrases with the head combination. The other words or phrases that are combined with the head in phrase construction can be specifies, modifiers and complements. As Afaan Oromo linguistic book [22]report, the phrase is part of sentence.

2.5.1 Criteria for Afaan Oromoo phrase

In Afaan Oromo; Single part of sentence or group words to be a phrase there is some criteria that identified by linguistic expert. A phrase is to be phrase one of the following three criteria must fulfilled [23].

Movement

Order of word in Afaan Oromo sentences is allowed by Afaan Oromo grammar. When move/interchange place those phrases are move with each. i.e. when it move in sentence moves as one part. See the following example;
Examples:
a. Bishaan [fayyaa namaatiif] gaariidha. [water is good for health of man]
b. [Fayyaa namaatiif] bishaan gaariidha.

In both sentence the part [Fayyaa namaatiif] [for health of man] is phrase. We cannot interchange the position if it does not give us complete meaning.

Replacement

To be phrase, those words that move as one word must be replaced by other word. One of this known replacement word is pronoun. Let we see following example;
a. Boontuun [meeshaa taphaa mucaa isheeti bitte.]
b. Kumeeni[s.]
c. Kumeeni[s akkasuma.]
d. Kumeeni[s sanuma mucaa ishiitii] bitte.

Interconnection

According to this criterion, any word can‘t enter between parts of sentence that construct single sentences.
Examples:
a. Sangaa diimaa guddaa kaleesssa argine sana
b. Sangaa **qalle** diimaa guddaa kaleesssa argine sana
c. Sangaa diimaa **qalle** guddaa kaleesssa argine sana
d. Sangaa diimaa guddaa **qalle** kaleesssa argine sana
e. Sangaa diimaa guddaa kaleesssa **qalle** argine sana
f. Sangaa diimaa guddaa kaleesssa argine**qalle** sana
g. Sangaa diimaa guddaa kaleesssa argine sana**qalle**.
In Afaan Oromo grammar only (a & g) is correct.a is phrase and g is sentence. Else others are grammatically incorrect.

2.5.2 *Afaan Oromoo phrase classes*

With above three criteria, In Afaan Oromo there are five phrase classes, which are noun phrases, verb phrases, adjectival phrases, adverbial phrases and prepositional phrases [22]. Those phrases are constructed from word classes.

Noun phrase(Gaalee maqaa)

Noun phrases are consists of a noun or pronoun and other related words that modify the noun or pronoun. It consists of noun or pronoun as head word and other words which come after or before the noun. The simplest NP consists of a single noun (e.g. Gammachuu) or pronoun such as **Inni** [he], **Isheen** [she], **Isaan** [they], etc. A complex NP can consists of a noun (called head) and other constituents (like complements, specifers, adverbial and adjectival modifiers) that modify the head from different aspects.
Example:
Tolosaan sangaa diimaa furdaa sana dheengadda bite
S
(NP (N Tolosaa) (dP -n))
(VP (NP (NP (N sangaa) (N diimaa)) (ADJ furdaa))
VP (AVP (AVB (dheengadda)) (VB bite))

Afaan oromoo verbal phrase (Gochibsii)

When we say verbal phrase we must see main components of sentences [Hima] in Afaan Oromo. In Afaan Oromo sentence structure verbs are found at the end of sentences. Example, In sentence Leensaan xalayaa barreessite.[Lensa wrote letter],barreessite[wrote] is head word and xalayaa [letter] is predicate. When categories sentence Leensaan xalayaa barreessite.in to noun phrase and verb phrase, Leensaan is noun phrase and xalayaa barreesite is verb phrase. The head word to phrase xalayaa barresssite is verb barressite that decided phrase is verb phrase.

Adjective phrase in Afaan oromoo

To take part of sentence is as adjective phrase or word or group of words are adjective phrase; head word of that phrase must be adjective. Example, Caalaan akkuma abbaa cimaadha. [Chala is strong as his father].akkuma abbaa cimaa is adjective phrase in verb phrase akkuma abbaa cimaadha‘.For this adjective phrase head word is a word cimaa. Because; concentration of adjective phrase akkuma abbaa cimaa is based on strengthens of chala not about chala like his father. That is why cimaa is head word to the phrase.

Adverbal phrases in Afaan oromoo

An adverb phrase is a phrase that is adverb is head word to phrases. In Afaan Oromo, adverbs are used to modify the coming verbs. Adverbs always come before the modified verb but it should be noted that any words come before verbs cannot be always considered as an adverb. In Afaan Oromo the main difference of adverb phrases from other phrase is adverb phrase is never come with other words [22]. Adverbs indicate manner, time, place, cause, or degree. Sometimes it is single word in sentences. In following example all the underline one is adverb phrases.
Example
Tolaan **kaleessa** deeme.(Tola went has gone yesterday)
In the above senescence the underlined word is adverb.

Prepositional phrase in Afaan oromoo

Afaan Oromo pre- and post- positional phrase (PPhr) is constructed from a pre- or post-position (PP) head and other constituents such as nouns, noun phrases, verbs, verb phrases, etc. that means there is no phrase that construct from only pre- or post- position. In Afaan Oromo prepositional phrases are not similar to other language like English. As in [26], there are not many possible forms for prepositional phrases in English, though adverbs can act as modifiers to prepositional phrases.

Examples: Post positional phrases with dependent post position

A)Tabba**tti**
B)Meeshaa**dhaan**
C)Abbid**aan**
D)Farda**an**
E)Muka**rra**

In above examples all are postpositional phrases. For these phrases head words are the bold once.Those are dependent one. But they have a right to guide structure that creates post positional phrases. In next example let will see prepositional and post positional phrase that were created from independent prepositional.

Example: prepositional and post positional phrase with independent pre- and post- positional.

A)Tabba **gubbaa**
B)**Gara** eessaa?
C)**Gara** manaa
D)Mana **jala**
E)**Waa'ee** mataaa
F)Garaa **keessa**

The bold words are head word to those phrases. Positions are either left or right of the structure. The difference of prepositional phrases from other phrases are head word can be either left or right of the structure. Also, dependent and independent prepositional can be taken as head in structure at the same time. However, it guides the structure by order or can't guide the structure at the same time. Let us take a look at the following example

Example: prepositional phrase with independent and dependent prepositional in the same structure.

Gara gaaraa**tti**
wa'ee koot**iif**

In above phrases a and b there is independent and dependent prepositional. But both can't guide the phrase. In both phrases the word **gara** and **waa'ee** is guide pre- and post- positional phrases **garaatti** and **kootiif**.

2.6 Sentences in Afaan oromoo

A sentence is group of words that are correctly structured to give one meaning. This definition is common in any language. But may be the structure is different based on the language. In English sentence must have subject, object, verb (S+O+V). But in Afaan Oromo, there is some difference. According to [22, 23], sentence is a complete thought or idea-subject + predicate. So, to be a sentence in Afaan Oromo it must have a subject(S) and a Verb (V). Structure of Afaan Oromo sentence look like this:-[23]

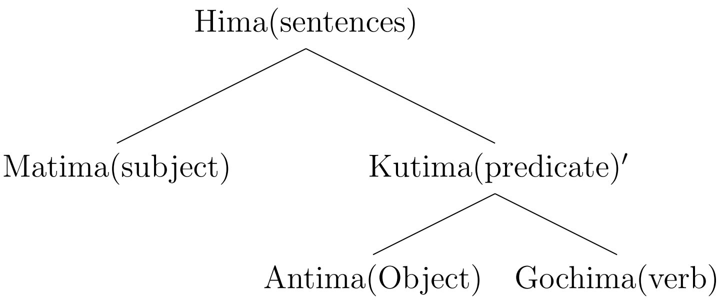

Figure 3. 1: Afaan Oromo sentence structure tree

Some sentences may have adjectives, adverbs and conjunctions. But that is not the case to define main structure (component) of the sentences. Let we see some of them in the following:

Example 3.1:
Sareen dutte.
Hundeen ogeessa IT ti.
Lagni Gibee guute.

A. Subject (Noun Phrase)

Subject (Noun Phrase) one of two main parts of a sentence containing the subject noun or a pronoun person, place or thing often accompanied by modifiers. Therefore, the noun or pronoun is who or what the sentence is about. If noun take as subject, it takes prefix like ’-ni‘, ’-n’, ’-ti’, ’-i’ and may be

zero morpheme.
Those morphemes are used in different ways.

- ‘-ni‘, ’ti‘, and ’i‘:- these are affixed on noun that end with single vowel letters. Among these three morphemes ’ni‘ is mostly used.
  Example: Mukni/Mukti/muki jige.

- ’-n‘ is affixed with noun end with double vowel letters to be noun as subject of the sentence.
  Example: Galataan arba ajjesse.

- Zero morphemes are noun end with ’-n‘ letters is no change when we take noun as subject of the sentence.
  Example: Bishaan dhuge.

B.Predicate
Predicate is one of two main parts of a sentence containing the verb, objects, or phrases governed by the verb. Basically sentence is made up of a noun and a verb. Verb is words that more declare a subject in sentence. In addition to subject and verb the sentence may have other words. Those words are used declare more subject and verb clearly. Verb is always at the end of the sentence in Afaan Oromo. In example 3.1; **’yuuse’**, **’Ogeessa fayyaati’** and **’guute’** are verbs. The object of a sentence is the noun or pronoun directly related to and affected by the subject‘s action (verb). The object is not who or what a sentence is mainly about; it‘s not the focus of the sentence.

2.6.1 *Types of sentences*

Basically Afaan Oromo sentences can be categorized in to different types based on structurally and functionally[27].

1.Structurally

Structurally, Afaan Oromo sentences can be categorized in to four types. Namely, Leexima[simple], Dachima[Compound], Xaxima[Complex] and Dachima xaxima[Compound Complex] sentences.

I. Simple Sentence

Simple sentence in Afaan Oromo contains subject (noun/noun phrase) and a predicate (Verb/verb phrase). It communicates one complete idea as an independent clause. It can be one independent clause.It‘s a complete sentence. To more explain let we see the following example.
i. Tolaan mana ijaare.
ii. Biliseen, intalli obbo Margaa, barsiistuu taate.
In example above both sentences are simple sentences because it has single verb. But subject of the sentence can be contains other words to more modify.

II. Compound Sentence

Compound sentence is constructed from the combination of two and more than two simple sentence or two and more than two independent clauses. In Afaan Oromo two or more simple sentences are combined to construct compound sentence, there are different techniques is there. Among those techniques some of them a using conjunction ’fi’ [and], using semicolon (;) and make affix -e double on verbs are techniques we can call.
Example:
I. Namni gaariidha.
II. Namni mana ijaare
. III. Namni horii horsiise.
The above three examples are simple sentences we can construct compound sentences from this three sentences, it create compound sentences ‘Namni mana ijaarefi namni horii horsiise gaariidha’.

III. Complex Sentence

A complex sentence includes a dependent clause linked to an independent clause by a subordinating conjunction of some kind to form a complete sentence. It contains a dependent clause and one or more independent clause. In complex sentence subordinating conjunction is affixed on dependent clause. Let we see in example.
I. Yoo dhufuu baattellee, xalayaa naaf barreessi.
II. Yoo finfinnee deemteef, meeshaa naa bitta.

IV. Compound Complex Sentence

Compound complex sentence is a sentence that contains one or more dependent clause and one or more independent clause.
Example;
I. Yommuu deemtuu fi yommuu deebitu naa dubbisii darbi.
II. sagal elmamus sagaltamni elmamus kan koo qiraaciidhuumaatti jette adurreen.

2. Functionally

Whereas sentence structure refers to the form of sentences in a language, sentence purpose refers to the function of sentences. Four types of sentence purposes exist in Afaan Oromo are similar to the other like English language: declarative sentences, interrogative sentences, imperative sentences, and exclamatory sentences.

A. Declarative Sentence

The first type of sentence in the Afaan Oromo is the declarative sentence. Declarative sentences, or declarations, convey information or make statements. They usually provide information, and are used to make statements.

For example:
Barattootni daree jiruu [Students are in class]

B. Imperative Sentence

The third type of sentence in the Afaan Oromo is the imperative sentence. Imperative sentences, or imperatives, make commands or requests.
For example:
Hojii manaa hojjedhu [Do homework]

C. Interrogative Sentence

The second type of sentence in the Afaan Oromo is the interrogative sentence. Interrogative sentences, or questions, request information or ask questions. It is a type of sentence that always has a question mark at the end.
For example;
Yoom biyya deemta? [When you go to your homeland]

D. Exclamatory Sentence

The fourth type of sentence in the Afaan Oromo is the exclamatory sentence. Exclamatory sentences, or exclamations, show emphasis. Unlike the other three sentences purpose, exclamatory sentences are not a distinct sentence type. Instead, declarative, interrogative, and imperative sentence become exclamatory through added emphasis.
For example;
Na gargaarii?(help me?)

2.7 Text predictive system

Text prediction, as its name indicates , is used to predict the word or a set of words as the user begins typing to help the user in reducing spelling error

,number of key strokes ,grammatical error and saving typing time.Predictive systems are not only for normal person but also for disable to help them in text entry process.

For text prediction there are many text entry system that can be used like keyboards ,speech,handwriting recognition,scanners ,gloved techniques ,microphone and digital cameras[26].Predictive text is one of the data entry techniques used in mobile or computer. In this technique, the system predicts what the user most likely intends to write based on some frequencies or other information.Scholars in [28], using a large dictionary of words for disambiguation, have developed a new text entry method for mobiles with a single key- press per letter on a standard phone pad. They have also discovered a system that predicts letters with much combination of key-presses. According to these scholars, there should be one or two words to match a given keystroke sequence (other combinations being non-sensible). This implies that, a predictive model of text entry method, if it makes use of a large dictionary of words, can suggest valid words to the user.

Other researchers on [4] is conducted research to address the problem of text entry on mobile phones. To solve the problem they have designed the text entry predictive methods that utilize the 12-key keypad and the model provides individual predictions for one-handed thumb and two handed index finger use. Further, the same researchers have discussed the three most common text entry approaches throughout the paper such as Multi -press, two-key press and T9 methods. However, they cannot provide word prediction.

The researchers in [10] have proposed a new text entry algorithm known as EasyET(ETWirelessKeyBoard) that predicts the next word while the user is typing a current word.However, they did not consider a new text entry algorithm for other devices like computer keyboard.

2.8 Sentence prediction

The predictive system work was first put forward in July 1995 [29]. Researchers in [30] have developed “a generic word prediction” model. A researcher in [31] argued that word prediction is predicting the most likely words tokens or words to follow a given segment of a text. This means that, a few keystrokes produce complete words or word sequences, and the number of keystrokes necessary to generate texts will be reduced.

Prediction can be a character, a word or a phrase. In the case of character prediction, the next character is to be predicted based on the previously inserted character(s), whereas in the case of word prediction the word is predicted based on some of the characters or words that were previously inserted in the editor [32]. On the other hand, phrase prediction system is guessing the number of appropriate phrases. Other researchers in [33] have studied the character prediction and its potentials for increasing recognition accuracy and provided a character predictor based on ngram with an optimal length of context for application to handwriting recognition.

Researchers in [4] believe that having more accurate predictions will provide a number of advantages like improving the quality as well as the quantity of message production for young people, for persons with language impairments, and for those who have learning disabilities and disambiguate sequences from ambiguous keypads and correct spelling errors. Another researcher in [66] has presented FASTY prediction system that includes several innovative features. A FASTY word prediction system have included prediction of compounds, prediction of proper inflectional form based on the use of parsing, dictionaries based on general language corpora and on users’ own texts and so on. Initially, a number of language were supported these are Dutch, French, German and Swedish.

The previous works, which were related to prediction and auto-completion systems, were to concern low-inflected languages such as English. The recent work, however, aims to consider highly inflected languages such as Amharic. According to some related researchers, reported in [66, 67, 68], morphological form variation is a problem in a prediction system. Morphological forms express mainly the number and, in some cases, the gender. When there are only fewer variations of a word, it is possible to store all of them in a dictionary. However, as highly inflected languages produce many forms, it may be difficult to store all of them. This is the main reason forthe search of new prediction methods in languages with a wide use of prefixes, infixes and suffixes.

To predict the next word, prediction algorithm may include different type of information about words. Some of them use only statistical information about the words in the sequence, such as unigram predictor, bigram predictor and trigram predictor. Others may also include syntactic information about the words using part-of-speech tags.

Furthermore, other applications such as word sense disambiguation, probabilistic parsing, part-ofspeech tagging, etc. [see in [31]] can be used in a word prediction system to develop it. A word prediction system facilitates the typing of text for users with physical or cognitive disabilities [18, 9]. As the user enters each letters of the required word, the system displays a list of the most likely probable words that could appear in that position. However, this study uses only part-of-speech tagging for finding the most frequent tag sequences.

In [19], said that current prediction function uses the word frequency lexicon, the word pair lexicon, and the subject lexicon. To train the system, they have used around 10,000 words with frequency information. The word pair lexicon consists of 3,000 reference words, each of which has an associated list of one to nine words that frequently succeed it. According to these researchers,the purpose of the subject lexicon is to allow the prediction system to adapt to the user's language by adapting the word frequency lexicon with those words of the user that are not in the lexicon or that have a rank higher than 1000. The user input is typically a prefix of a complete query q that the user intends to enter. The algorithm returns a list of k completions, which are suggestions for queries, from which the user can select.

The same researchers showed that when the input prefix is short (1 character) and the context is relevant to the user's intended query, then the weighted MRR of Nearest Completion is 48% higher than that of the standard Most Popular Completion algorithm. The Nearest Completion algorithm suggests the user's prefix input that is most similar to the recent queries s/he has just entered. However, when the context is irrelevant, Nearest Completion is useless. To solve this, these researchers have proposed hybrid completion, which is a convex combination of Nearest Completion and Most Popular Completion. Hybrid Completion is shown to be at least as good as Nearest Completion when the context is relevant and almost as good as Most Popular Completion when the context is irrelevant.

According to these researchers, Nearest Completion computes the similarity between queries as the cosine similarity between their rich representations. Nearest Completion is designed to work well when the user input has a non-empty context and this context is relevant to the query that the user is typing. They concluded that Nearest Completion relies either on no information or on false information and, thus, exhibits poor quality.

The Reactive Keyboard [15] works by attempting to predict what the user might want to select next on the basis of its preceding input. To predict the most likely next keystrokes, the system uses the sequence of the previous keystrokes. It uses an n-gram model for characters, created from text samples and from the user's input. The model is stored in a special tree structure that allows partial matches between context and model to be found economically. The idea is to use the n-1 previous characters to predict the n th one, where possible. If matches cannot be found, the context is shortened by one character, and the processes continue. Generally, keystrokes can only be predicted with limited accuracy.

For English language, VanDyke [38] has developed a word prediction system to provide the user with a list of grammatically appropriate words. The predictor works by trans versing the search space produced by constructing the parse tree of the input sentence. The parser holds all possible structures for the partial sentence entered so far, and thus at each point in the sentence, it knows what syntactic categories can be in the next position. This eliminates a number of words to choose from, resulting in predictions that are more appropriate. However, it requires a considerable amount of work to parse partially the input sentence every time the user completes a new word. There is no evidence of testing the system to see whether the use of grammar helps improve the prediction performance.

Researcher on [40] have incorporated additional information into the learning process of their word-prediction system in order to learn better language models in comparison to prediction systems that use n-gram models. In the proposed prediction system, the local context information along with the global sentence structure is considered. For this purpose, a very large set of features, characterizing the syntactic and semantic context in which the word tends to appear has used, is learned for each word in terms of the features, and a learning method that is capable of handling the large number of features is used. A language for introducing features in terms of the available information sources is also defined.

Each sentence is represented as a list of predicates called the information source (IS) of the sentence. Features are defined as relations over the information source, or aspects of the structure of the sentence. A few examples of features are the adjacency relations between words, word collocations, and the part-of-speech tag assigned to each word. There are also complex features such as the dependency relations between words and the role of each word or phrase inside the sentence, i.e., if it is a subject, object, etc.

A researcher in [39] has developed a new system whose scope was extended to include part-ofspeech tag trigrams and word bigrams at each prediction point. The prediction algorithm has been interacting with a first-order Markov Model for words and a second-order Markov Model for partof-speech tags. This considered conditional probability of a word by giving the probability estimation of the tag obtained by the tag Markov Model. The idea (i.e. assumption) of the prediction algorithm was first to obtain probability estimation for the tag of the next word using the tag Markov model. In the next step, probability estimation was found for the next word using the word Markov model. The tag probability estimation from the previous step was used to promote a rank of the words with the most likely tag. The tags unigram, bigram and trigram lexicons are created from the same corpus and used to build word unigram, bigram and trigram language models. Three texts of about 10, 000 words each have been used for evaluation. The system achieved a keystroke saving of about 43.2%, when given suggestions but no adaptations were used.

2.8.1 ***Approaches to text predictive system***

The methods for text prediction can be classified as statistical, knowledge based and heuristic (adaptive) modeling. Most of existing methods employ statistical language models using word n-grams and POS tags.

Statistical Prediction

In statistical modeling, the choice of text is based on probability that a text may appear in a corpus. The statistical information and its distribution could be used for predicting letters, words, and phrases. Statistical text prediction is made based on Markov assumption in which only last n-1 word of the history affects succeeding word and it is named n-gram Markov model. It is based on learning parameters from large corpora. However, one of the challenges in this method is when a language that is written with the help of text prediction system is of a different style than the training data.This statistical approaches is not only related to N-gram ,distance similarity and cosine similarity is categorized under these approaches.

Knowledge Based Prediction

Word prediction systems that merely use statistical modeling for prediction often present words that are syntactically, semantically, or pragmatically inappropriate and impose a heavy cognition load on users to choose the intended word in addition to decrease in writing rate. Syntactic, semantic and pragmatic linguistic knowledge can be used in prediction systems.

In this approach, Parts-of-Speech (POS) tags of all words are identified in a corpus and the system uses this knowledge for prediction. This approach requires a set of linguistic tools such as POS taggers and lemmatizes. However, these are not available in all languages. Statistical syntax and rule-based grammar are two general syntactic prediction methods, where statistical syntax uses the sequence of syntactic categories and POS tags for prediction. Therefore a probability would be assigned to each candidate word by estimating the probability of having this word with its tag in the current position and using most probable tags for previous one or more words. In rule-based grammar, syntactic prediction is made using grammatical rules of the language. A parser will parse current sentence according to grammar of the language to reach its categories

Heuristic Prediction

Heuristic (adaptation) method is used to make more appropriate predictions for a specific user and it is based on short term and long term learning. In short term learning, the system adapts to a user on current text that is going to be typed by an individual user. Recency promotion, topic guidance, trigger and target, and n-gram cache are the methods that a system could use to adapt to a user in a single text. However, in long-term prediction the previous texts that are produced by a user are considered

2.8.2 Evaluation metrics for text predictive system

In this section we will discuss some evaluation metrics for text predictive system.

Keystrokes Until Completion (KUC)

Keystrokes Until Completion (KUC) is another metrics to evaluate word prediction systems where, c1 ... cn being number of keystrokes for each of the n words before the desired suggestion appears in the prediction list [18]. It is the average number of keystrokes that a user enters for each word before it appears in the suggestion list [28]. Lower value of KUC shows better performance.

Hit Rate (HR)

Hit Rate (HR) is an additional word sequence prediction measuring metrics. It is the percentage of times that the intended word appears in the suggestion list and if its hit rate is high as the required number of selections decreases, the predictor is considered to have better performance [16, 28].

Keystroke Saving (KSS)

Keystroke Saving (KSS) is primarily used evaluation means in word prediction. The common trend in research is to simulate a ”perfect” user that will never make typing mistakes and will select a word from the predictions as soon as it appears [18, 22]. A Keystroke Saving (KSS) estimates saved effort percentage in keys pressed compared to letter-by-letter text entry and it is calculated using

Accuracy

Accuracy is the percentage of words successfully completed by a word prediction system before a user reaches the end of a word. It is the ratio of words correctly guessed to total words guessed. A system that completes words in early stages of typing is considered to have better performance [28].

2.9 Summery

This chapter covered overview of Afan Oromo like structure of Afan Oromo sentences and also covers predictive system,sentence prediction under this approaches to sentence prediction like statistical ,rule based and heuristic approaches and evaluation of predictive system like hit rate ,keystroke until complete ,accuracy and keystroke saving also discussed.

3. CHAPTER THREE:RELATED WORKS

3.1 Introduction

This chapter covers reviewed related work on word prediction and sentence prediction or sentence completion of local languages like Amharic and Afaan Oromoo and foreign language like English and Bangla languages.In this chapter work done, data set used, approaches and result obtained will be discussed.

3.2 Prediction system for local languages

In this section previous works done related to word completion and sentence completion will be reviewed for local language like Afaan Oromo and Amharic.Works like Word sequence prediction for Afaan Oromoo,word sequence prediction for Amharic languages, Enhancing the Text Production and Assisting Disable Users in Developing Word Prediction and Completion in Afan Oromo,Ethiopic Keyboard Mapping and Predictive Text Inputting Algorithm in a Wireless Environment and Implementation of Online Handwriting Recognition System for Ethiopic Character Set is reviewed

3.2.1 Word sequence prediction for Afaan Oromoo language

Researcher on [34] conducted research on word sequence prediction for Afaan Oromoo.In this work investigator designed word sequence prediction model for Afaan oromoo languages.For purpose of this work used a corpus collected from different sources like newspaper(Bariisaa,Bakkalcha Oromiyaa and Oromiyaa)journals ,criminal codes ,books,social media like Facebook with total number of 50 files consisting of 23,400 sentences and total word of 312,208 with 49,143 unique words.Just to show the acceptance of the designed word sequence prediction model for Afaan Oromoo prototype was developed using C# visual studio 2015.Data entry or text entry might be predictive or not-predictive.Because it is statistical based or measure frequency of word,the suggested word will be displayed according to their frequency.To test this model BBC news text was used with Key Stroke Saving(KSS) as standard performance metrics and the result was 22.4% KSS.

3.2.2 Word sequence prediction for Amharic language

Researcher [35]developed word sequence prediction system for Amharic language using N-gram language model.Prototype was developed to test the model using python programming language.For testing the model Amharic sentences from News with total of 107 sentences is used. KSS(Key Stroke Saving)is used as evaluation metrics.Bi-gram, Tri-gram and hybrid of both model is used and the three model produced independent result.Result obtained is 13.1%, 17.4% and 20% for bi-gram ,tri-gram and hybrid of them respectively.

3.2.3 Enhancing the Text Production and Assisting Disable Users in Developing Word Prediction and Completion in Afan Oromo

The researchers on[36]conducted research on word prediction and completion for disable.The idea behind this work is to assist the disables to interact with computer software and file editor by their mother tongue languages.For development purpose they used unsupervised machine learning.For implementation purpose this work is done using N-gram language model(unigram,bigram and trigram) for auto completing the word by predicting the correct word in a sentence which saves time ,reduce spelling error,keystrokes.For this work unsupervised machine learning is used because there is no officially annotated corpus for the languages .This work scored accuracy of 83.84%, 61.4%, 54.8% of unigram, bigram and trigram respectively with evaluation metrics of precision and recall 90% and 73.34% precision and recall respectively.The draw back of the model is that it waits until 3 or 4 character is input which might be in all full words.It is better if the evaluation metrics is KSS.

3.2.4 *Ethiopic Keyboard Mapping and Predictive Text Inputting Algorithm in a Wireless Environment*

In [10],researchers have proposed the system Ethiopic Keyboard Mapping and Predictive Text Inputting Algorithm in a Wireless Environment that predicts the text according to the user input and facilitates message text exchange with Ethiopic scripts on a wireless phone. They have used two methods, such as Multi-press and Three-key input methods as well to develop this system. In order to determine the frequency of words in the dictionary, they have used three rating attributes. Moreover, they have developed a wireless application for Ethiopic text messaging. The same researchers eliminated ‘superfluous’ characters, replaceable characters, Ethiopic numerals and labialized characters during text processing which reduce the number of characters in the font set. Accordingly, they have discussed two factors that have impacts on the prediction of the Amharic words. On the one hand, the structure of Amharic language and the statistical distribution of the order of characters, and the other was the habits of users to write words, phrases and sentences. Additionally, in this work, it was mentioned that to predict a word or phrase ‘apriori’ is difficult unless a dictionary is used in order to build a library of words, phrases and sentences commonly composed by the user. Eventually, they have achieved 40 % keystroke savings.

3.2.5 *Implementation of Online Handwriting Recognition System for Ethiopic Character Set*

Furthermore, another researcher in [37] has designed the word prediction system that takes recognized handwritten characters from an online Amharic handwriting character recognition system. She has designed the algorithm without considering the constraints of the handheld devices such as PDAs (Personal Digital Assistant). Another researcher in [38] has considered these constraints and developed his system for PDAs. However, the works of both researchers in [37] and [38] are just to recognize and display a single character not to write Amharic text. By observing the gap, other researchers in [39] have adopted it to give the recognized character to the word prediction system as an input, and developed an online handwriting word prediction system for Amharic language. The algorithms required for assigning word frequencies from the corpus to the dictionary (lexicon) and for predicting words have designed and implemented experimentally by using N-gram models. In order to determine the frequency of words in the dictionary, he has used word Smith tools. To predict the current or the next position of a word they only used word frequency information. In whatever way, the algorithm ignores the previous context and does not consider the syntactic structure. Finally, the researcher showed of 81.39% accuracy was achieved.In this research work, an approach based on the syntactic analysis and information retrieval of the sentence that tries to extend the previous statistical prediction methods has described above. The algorithm may suggest words, which are not grammatically appropriate.

In general, as mentioned before, the idea of developing prediction system for Amharic is not new and couples of researchers have presented their work. The first one is a Master's thesis focusing on word prediction online handwriting recognition for Amharic language [39] and the other is Ethiopic Keyboard Mapping and Predictive Text Inputting Algorithm in a Wireless Environment. As discussed in the above section the limitation of researchers work is related to the method in that they used frequency information only to predict the next word. In addition, another researcher in [40] design text predict entry system for Amharic text on mobile phone. This work has been used similar method in [39] but has used different training data set. The reliance on such statistical information only made the approach to lose the context of the previous word. According to recent works, researchers have manipulated word frequency effects in naming tasks are often considerably less than 100 msec. The problem of this fact is lack of context with word frequency interaction in the data. In this study, we include linguistic knowledge to solve such type of problem.

These researchers, particularly in [39, 40], found that written Amharic text has a high degree of redundancy. Based on these finding, it is natural to ask whether users can be supported in the process of writing text by systems that predict the intended next words, or sentences.

3.3 Text prediction system for foreign languages

In this section we will discuss word prediction and sentence prediction system of foreign languages like English and Bangla.Works like Bangla Word Prediction and Sentence Completion Using GRU,Dependency language models for sentence completion,Automated word prediction in Bangla language using stochastic language model,A Neural Language Model for query Auto-Completion is reviewed and presented.

3.3.1 Bangla Word Prediction and Sentence Completion Using GRU

Researchers on[41] has done word prediction and sentence completion using GRU(Gated recurrent unit) for Bangla language.In this work recurrent neural network on n-gram data-set is used.To do so data were collected from different sources like BBC News Bangla,The daily prothom alo news paper Bangla academic books.Data of 50,000 75,000 and 45,000 words were collected from BBC News Bangla ,The daily prothom Alo newspaper and Bangla academic book receptively and with 8239,9686 and 5075 unique words respectively.Firstly data set is divided into 5 data set for unigram ,bigram ,trigram, four gram and penta gram.Then RNN(GRU) was run on those prepared data.the trained Uni-gram model has an average accuracy of 32.17% and the average loss of 276.44% for our proposed approach, where the Bi gram model has an average accuracy of 78.15% and the average loss of 53.36%. Again, Tri-gram has 95.84% accuracy on average and 8.52% loss on average for the same dataset used for Uni-gram and Bi-gram. Uniformly 4-gram and 5-gram show an average accuracy of 99.24% and 99.70% where they have an average loss of 2.04% and 1.11%, which indicates that the accuracy and loss level is improved according to the number of n is increased.The problem with this model is that on the matter of sentence prediction it need to predict the whole sentences word by word.

3.3.2 Dependency language models for sentence completion

The researcher on [42] conducted research on dependency language model sentence completion for English languages.According to [42] Sentence completion is a challenging semantic modeling task in which models must choose the most appropriate word from a given set to complete a sentence. According to [42]variety of language models have been applied to the task sentence completion in previous work, none of the existing approaches incorporate syntactic information.In this paper work they propose to tackle this task using a pair of simple language models in which the probability of a sentence is estimated as the probability of the lexicalisation of a given syntactic dependency tree.In that work they used dependency language models rather than standard N-gram language models.According to this work N-gram is unable solve the problem of dependency between the word.But this model uses dependency tree in order to tackle this problem.As per[42]data sets from Sherlock Holmes novels by Arthur Conan Doyle which consists of 1040 sentences completion problem.Each problem has sentences in which single word is removed and replaced by blank space.For that blank space there are five candidate answers to be filled in the blank space.The total number of data set used for training purpose is 522 from project Gutemberg.The score is 8.7 better than standard n-gram languages.But these work is also limited to filling the blank space with only single word.

3.3.3 Automated word prediction in Bangla language using stochastic language model

The researchers on[43] proposed and designed automated word prediction for Bangla languages using stochastic language models.N-gram language model:-un-igram bi-gram tri-gram , deleted interpolation and back-off model is used .The data set were collected from news paper called daily Prothom Alo.The corpus consist of 0.25 million.It contains 14,872 word forms.Java program is used for experimentation purpose.The total corpus is set apart into two parts :-training and testing corpus.For splitting hold-out method is used.Average accuracy scored for each model is 21.4,45.84,63.04,63.50 and 62.86 for unigram,bi-gram,tri-gram ,backoff and Deleted Interpolation respectively.

3.3.4 A Neural Language Model for query Auto-Completion

A research[35] is done on query auto-completion using neural language model.As per researcher query auto-completion (QAC) systems suggest queries that complete a user's text as the user types each character.The work mainly focus on previously unseen queries.That means not like other model that suggest previously seen query based on their frequency,this mainly focus on the suggestion of query that has no seen history.Recurrent neural network is used to for modeling purpose.The result shows that the neural language model and wisdom of the crowd method (MPC) has the same performance for previously seen query.This modl improves the state of the art query aut completion by43%.

3.4 Summery

Generally around nine research work related to this works are reviewed in this chapter.In case for Afaan Oromoo there is no work done that directly related to this topic.All research conducted on Afan Oromo all about word sequence prediction and word completion.They all limited to word level and also limited to statistical approaches .But in today's digital world many users really want a system that can predict not only single word or next word but also next phrases or full sentences which is all about this paper

4. CHAPTER FOUR:AFAAN OROMO SENTENCE PREDICTION SYSTEM

4.1 Introduction

This chapter covers details of Afaan Oromo sentence prediction system architecture of proposed system.It contains the system architecture and it is detailed components and algorithm. In the architecture and algorithm language grammar is handled because it is context sensitive.

4.2 Architecture of Afaan Oromoo sentence auto completion

The following is architecture of Afaan Oromoo sentence prediction.

Figure **??** shows a boat. This Afaan Oromo sentence prediction system has three major components pre processing module,RNN-LSTM Training module and Trained model.Pre processing module.Trained model uses pattern developed during training process to propose the the list of most probable sentence. Finally, the prediction module produce the list 5 sentence.

The prediction take ways in the following order.The user inters three words and initiate the predictor by clicking enter key on the computer.Then the user input is accepted and analyzed using by prediction module .The prediction process is initiated after users enter three words. Accordingly, a user's input is accepted and analyzed using prediction module. At the end , the sentence generator give out in accordance with their probability order.

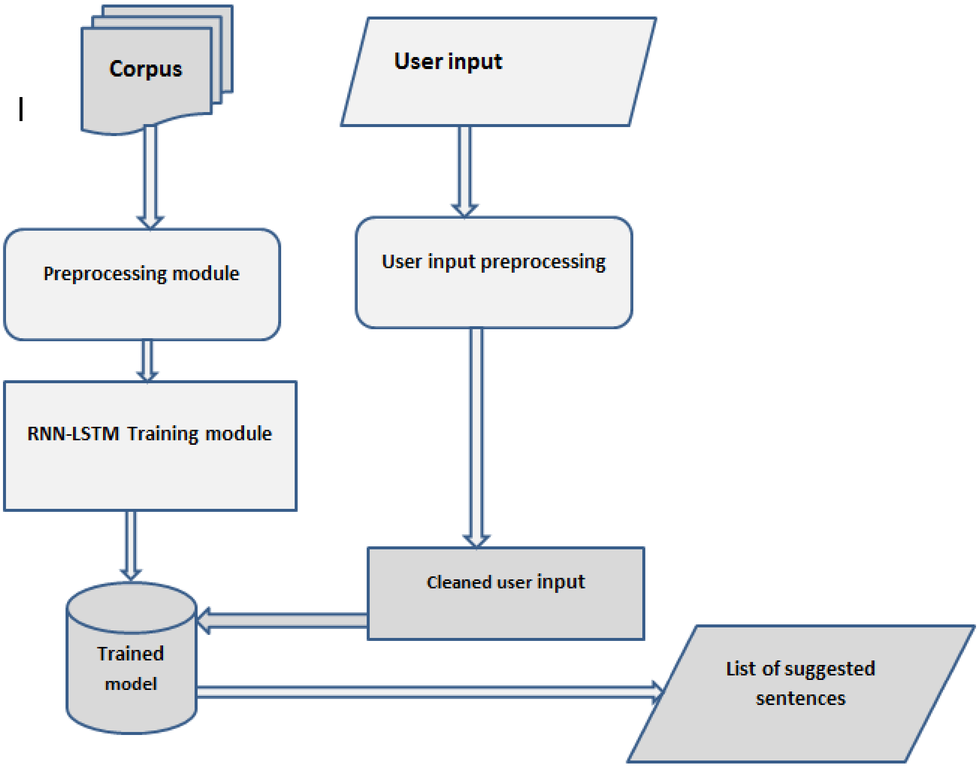

Fig. 4.1: Afaan oromoo sentence prediction architecture.

4.3 Language model

It is the language model that gives sentence predictor a competence to predict the next words phrases or sentence. The Language model gives the prediction engine by providing information collected from training corpus. Accordingly, to predict appropriate sequence of word or sentence for given user input RNN-LSTM is used.

4.3.1 Afaan Oromoo corpus

Sentence completion is not an easy task to design a model since the length of sentences differs, syntactic information of words and sentences are ambiguous, semantic information lacks ad hoc and so on. As mentioned above, there is no ready-made corpus for Afaan Oromo, the target language for this research work.Thus, to design the model and accomplish the below tasks of sentence prediction, the researcher has done his is best to train and test the model.For this research purpose Afaan Oromoo sentence from different media website are collected.The researcher collected document of different genre like health,sport,agriculture, crime from FBC Afaan Oromoo VoA Afaan Oromoo.

4.3.2 Prepossessing module

Data cleaning

There are many operations and products that are being developed using Natural Language processing. The text is the main input for any type of model like prediction, Sentiment analysis and many more.So consider the text contains different symbols and words which doesn't convey meaning to the model while training.So we will remove them before feeding to the model .This method is called Data Preprocessing. Alternately it is also called Text Cleaning.So as we stated in its definition data we gathered have much of dirty datum that need to be cleaned like removing URL ,removing stop words lemmatization,stemming and so on.

Removal of StopWords:

Stopwords are the words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For example, if you see the below example we can see the stopwords are removed. Nltk (natural language tool kit) offers functions like tokenize and stopwords.

Example:

Namni gabaa jira

Stemming and Lemmatization

Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.Stemming is a rule based approach, it strips inflected words based on common prefixes and suffixes.

Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma.It usually refers doing things properly with the use of a vocabulary and morphological analysis of words. It observes position and Parts of speech of a word before striping anything.

4.3.3 RNN-LSTM training modules

Under this components of the system reprocessed corpus is feed into RNN-LSTM model to train and develop model for Afan Oromo.RNNs are a powerful and robust type of neural network, and belong to the most promising algorithms in use because it is the only one with an internal memory.Because of their internal memory, RNN's can remember important things about the input they received, which allows them to be very precise in predicting what's coming next. This is why they're the preferred algorithm for sequential data like time series, speech, text, financial data, audio, video, weather and much more. Recurrent neural networks can form a much deeper understanding

Fig. 4.2: How RNN works.

of a sequence and its context compared to other algorithms.In a RNN the information cycles through a loop. When it makes a decision, it considers the current input and also what it has learned from the inputs it received previously.

Therefore, a RNN has two inputs: the present and the recent past. This is important because the sequence of data contains crucial information about what is coming next, which is why a RNN can do things other algorithms can't.Long short-term memory networks are an extension for recurrent neural networks, which basically extends the memory. Therefore it is well suited to learn from important experiences that have very long time lags in between. The units of an LSTM are used as building units for the layers of a RNN, often called an LSTM network. LSTMs enable RNNs to remember inputs over a long period of time. This is because LSTMs contain information in a memory, much like the memory of a computer. The LSTM can read, write and delete information from its memory.

4.3.4 ***Trained model***

This is RNN-LSTM deep learning model that is trained and learned the sequence of Afan Oromo sentences based the given training data set.This model is trained on Afan Oromo text that is cleaned and stemmed and lemmatized in prepossessing module.

4.3.5 ***Summery***

This chapter covered mostly the depicted architecture of Aafan Oromo sentence prediction system.This Afan Oromo sentence prediction architecture has three main components or modules like pre processing module for cleaning data set to be used ,RNN-LSTM training module for training purpose and RNN-LSTM trained module that learned Afan Oromo sentence sequence for predicting the sentence based on user input

5. CHAPTER FIVE: EXPERIMENTATION, PROTOTYPE AND EVALUATION

This chapter covers about data used for investigation ,excremental environmental setup under which the experiment is carried out,tools used ,prototype developed ,result obtained and evaluation of the results.

5.1 Experimentation

5.1.1 Data set

To accomplish a task of Afaan Oromo sentence prediction, it needs corpus from target language that means Afan Oromo.Since the Afaan Oromo has no ready made corpus, we are obliged to to collect our own corpus from various sources that include newspapers(Bariisaa, Bakkalcha Oromiyaa and Oromiyaa), journals, criminal code ,books, social media like Facebook, web pages, books written on different issues such as politics, religion, history and fiction. Around 50 different files, collected from various sources mentioned above, are provided to the tool and corpus consists of 23,400 sentences and a total of 312,208 words are generated filter of 49,143 unique words. All the 50 files are converted to txt format just to be feed to python tool. All the collected corpora are merged to one corpus named as: "SelectedText-dataset".The size of corpus affects the the result to be obtained and processing speed of the machine.

5.1.2 Tools and environmental setup

To experiment with design solution we used python as programming tools.Obviously very power full language for experimentation in areas of NLP and others.It is easy to write because it provides many built in libraries and enables the user to use them easily.As environment we used Google Co-laboratory.Google colab is used because RNN needs GPU for training purpose and Google co-lab provides free online GPU and RAM.Google drive is used for online data storage because Colab access data faster from the google drive than from local drives.

5.2 Result

Test result of the model is exhibited result obtained based on keystroke saving and accuracy performance evaluation metrics. To evaluate the performance of predictive system keystroke saving is one of the standard evaluation metrics.Keystroke Saving (KSS) estimates saved effort percentage which is calculated based on by comparing total number of keystrokes needed to type a text (KT) and effective number of keystrokes using word prediction (KE). Keystroke Saving (KSS) is referred to the percentage of keystrokes that the user saves by using the sentence prediction system.A higher value for keystroke saving implies a better performance.If the value of keystroke is high that means performance of the model is good.Test is done based on users inputs.Four testers are randomly selected and prompted to test the model.Based on each tester they produced different keystroke saving.They produced 15% KSS,22% KSS,18% KSS,20% KSS keystroke saving respectively.Then the average of those testers keystroke saving result is taken and recorded as a result in this work which is 18.75% KSS.

5.3 Discussion

The result of the experiment is discussed in above table.Four sample is taken to test the model and the model produced different keystroke saving with slight difference.Just to make is standard we take average of the four test sample result. Accordingly 18.75% KSS result is obtained from the experiment.We believe that the result in this work is promising and can be enhanced with addition of more linguistic resources in the language model.In this work, the testing result highly depends on the training data, and due to this the outcome can differ when tested on other training corpus. Rooms for improvement and extension of this work are presented in Section 6.2.

6. CHAPTER SIX:CONCLUSION, CONTRIBUTION AND FUTURE WORK

6.1 Conclusion

In this study, Afan Oromo sentence prediction model is developed using RNN-LSTM methods.Sentence prediction assists people in their text input means, and there have been a number of researches done on the topic for various languages as briefly stated in Chapter-3. Even though there are diverse linguistic researches in Afan Oromo, there is no work on the topic of sentence prediction that considers context information.This study is set out to suggest the next phrases to be typed by a user, based on previous history of input word. This is done using RNN-LSTM Deep Learning models which are developed using Afan Oromo news corpus, and syntactical rules of the language.

According to our evaluation better Keystroke saving (KSS) is achieved when using a RNN-LSTM models. In conclusion, the developed model has potential advantages since an effective sentence prediction can be carried out using very large corpus size, Deep learning based techniques, and linguistic rules. We believe that application of this technology is ample, and among them, it has capability to bring benefits of fast text typing to virtual keyboards, portable devices like Laptop and Desktop in assisting people with disabilities.

6.2 Contribution

1. Afan Oromo sentence prediction is investigated using deep learning model.

2. RNN-LSTM model performs better than statistical approaches like N-gram.

6.3 Future work

Sentence prediction system demands deep understanding of structural and semantic features of language under consideration. Hence, it seems that there is a plethora of gaps for improving and modifying Afaan oromo sentence prediction. To this end, we strongly believe that this kind of study can be further investigated in numerous ways to optimize the task of Afaan Oromoo sentence prediction. Accordingly, we suggest some direction for future works:

1. For future work, conducting this study with BERT is recommended since it might give a better result performance.

2. Because data set in this work is not subject to any linguistic science it is better to consider well prepared corpus.

3. Since this work does not consider about the semantic information, one can investigate this topic by taking into consideration about semantic information

BIBLIOGRAPHY

[1] P. Jackson and I. Moulinier, “natural language processing for online applications: text retrieval, extraction and categorization,” 1984.

[2] Jurafsky and Martin, “Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition,” 2007.

[3] K. S. Jones, “Natural language processing, a historical review,” *University of Cambrige, Cambrige,*, October 20.

[4] S. M. et al., *Predictive Text Entry Speed on Mobile Phones.* Proceedings of the ACM Conference on Human Factors in Computing Systems - CHI 2000, pp. 9-16, New York, 2000.

[5] “National census report-2007,” http://www.csa.gov.et/census-report/complete-report/census-2007, accessed: 2012-01-15.

[6] “ethiopian treasures language,” http://www.ethiopiantreasures.co.uk/pages/language.html, accessed: 2020-01-15.

[7] T. D. Tesema W, “Enhancing the Text Production and Assisting Disable Users in Developing Word Prediction and Completion in Afan Oromo.”

[8] Q. J. W. W. I. Y. T. K. S. K. Xiao, Chuan, “Efficient error-tolerant query autocompletion.” *Proceedings of the VLDB Endowment*, vol. 7, no. 200, 2013.

[9] K. N. Bar Yossef, Ziv, “Context-Sensitive Query AutoCompletion,” 2011.

[10] A. et al, “Ethiopic keyboard mapping and predictive text inputting algorithm in a wireless environment,” *ITEs, Addis Abeba*, 2004.

[11] M. Daud and V. Goyal, “Predictive text entry method for somali language on mobile,” *International Journal of Computer Science Trends and Technology (IJCST)*, no. 3, 2015.

[12] G.Tesema, “Design and implementation of predictive text entry method for afaan oromo on mobile phone,” *PhD diss., Addis Ababa University*, 2013.

[13] A. Solomon, “Unsupervised machine learning approach for word sense disambiguation to amharic words,” *M. Sc. thesis, University of Addis Ababa, Addis Ababa, Ethiopia*, 2011.

[14] L. Koul, “Methodology of educational research,” *Vikas Publishing House PVT LTD*, 1988.

[15] P. Isokoski, “Manual text input: Experiments, models, and systems.” *MSc thesis in Information science, University of Tampere*, 2004.

[16] N. S, “word prediction for amharic online handwriting recognition,” *MSc thesis in Computer science, Addis Ababa University*, 2008.

[17] P. Kristensson, “Five challenges for intelligent text entry methods,” *journal of Association for the Advancement of Artificial Intelligence*, 2009.

[18] Samuel and Johanna, “Effect of dynamic keyboard and word-prediction systems on text input speed in persons with functional tetraplegia,” *JRRD*, vol. 51, no. 3, 2014.

[19] B. A. M. Jacob O. Wobbrock, “Word completion for trackball text entry,” *nformation School, University of Washington Seattle, Washington USA*, 2012.

[20] A. S. Arif and W. Stuerzlinger, "Pseudo-pressure detection and its use in predictive text entry on touchscreens," *Information School, University of Washington Seattle, Washington USA*, 2014.

[21] Manning, "Foundation of statistical natural language processing," *MIT press*, 2014.

[22] R.Getachew, *FURTUU: Seerluga Afaan Oromoo(Oromo Grammar).*, 2009.

[23] B.Aduunyaa, *SEMMOO:Bu'uura Barnoota Afaaniifi Afoola Oromoo.*, 2014.

[24] B. Addunyaa, "Natoo: Yaadrimee caasluga afaan oromoo," *ADDIS ABABA*, 2012.

[25] M.Diriba, "Automatic sentence parser for oromo language using supervised learning technique," *ADDIS ABABA*, 2013.

[26] B. McCaul and A. Sutherland, "Predictive text entry in immersive environments," *Proceedings of the IEEE Virtual Reality*, vol. pp, no. 241, 2004.

[27] B. Addunyaa, "Semmoo: Bu'uura barnoota afaaniifi afoola oromoo." *Oromia*, vol. pp, no. 1–16, 2014.

[28] M. Dunlop and A. Crossan, "Predictive text entre methods for mobile," *SpringerVerlag London Ltd personal technology*, vol. pp, no. 134-143, 2000.

[29] A. C. et al, "Constructing a database for a new word prediction system,," *TMH-QPSR,*, vol. 32, no. 2, 1996.

[30] A. Bosch, "Scalable classification-based word prediction and confusible correction," *TAL*, vol. 46, 2006.

[31] M.Wester, "User evaluation of a word prediction system," *M. S. thesis, Uppsala University, Uppsala,*, 2003.

[32] "Wikipedia contributor language," http://http://www.wikepedia.com., accessed: 2020-05-15.

[33] S. Homayoon and M. Beigi, "Character prediction for on-line handwriting recognition," *Canadian Conf. on Electrical and Computer Eng., Toronto, Canada*, vol. 2, 1992.

[34] A. Bekele, "Word sequence prediction for afaan oromo," *AAU-ETD*, 2018.

[35] D. H. Park and R. Chiba, "A neural language model for query auto-completion." *In Proceedings of SIGIR '17, Shinjuku, Tokyo, Japan*, 2017.

[36] Workineh and Duresa, "Enhancing the text production and assisting disable users in developing word prediction and completion in afan oromo," *J Inform Tech Softw Eng 2017, 7:2*, 2017.

[37] A. Shimeles, "Online handwriting recognition for ethiopic characters," *M. S. thesis, Department of Computer Science, Addis Ababa University*, 2005.

[38] A. Abebaw, "Implementation of online handwriting recognition system for ethiopic character set," *Masters Project, Department of Computer Science, Addis Ababa University*, 2007.

[39] N. Suleiman, "Word prediction for amharic online handwriting recognition," *M. S. thesis, University of Addis Ababa, Addis Ababa*, 2008.

[40] A. Mulu and V. Goyal, "Amharic text predict system for mobile phone," *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 3, no. 4, 2015.

[41] Mulu and Goyal, "Bangla word prediction and sentence completion using gru," *International Conference on Sustainable Technologies for Industry 4.0*, vol. 3, no. 4, 2019.

[42] Gubbins and Vlachos, "Dependency language models for sentence completion," *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, vol. 3, no. 1405–1410, 2013.

[43] H. H. Rahman, "Automated word prediction in bangla language using stochastic language models," *International Journal in Foundations of Computer Science and Technology IJFCST*, vol. 5, no. 6, 2015.