RESEARCH ARTICLE OPEN ACCESS

Amharic Text Predict System for Mobile Phone

Alemebante Mulu [1], Vishal Goyal [2]
Research Scholar [1], Assistant Professor [2]
Department of Computer Science
Punjabi University Patiala
India

ABSTRACT

Amharic text prediction system describes the data entry techniques that are used to enter data into mobile devices, such as a smartphone. Data entry could be either predictive or non-predictive in which the first two characters is written and listed down all predicted word, based on the frequency of the word as well as going the alphabetical order if the frequency is the same. In this paper, we have designed text prediction model for Amharic language: a corpus of 1193719 Amharic words, 242383 Amharic lexicons and a list of names of persons and places with a total size of 20170 has been used. To show the validity of the word prediction model and the algorithm designed, a prototype is developed. The Experiment is tested by a database or lexicon of Alembante Mulu also conducted to measure the accuracy of the Amharic text prediction engine and finally the prediction accuracy is achieved 91.79%.

Keywords:- Amharic text prediction model, N-gram model for Amharic text prediction, Text prediction corpus, Text prediction lexicon, Natural Language Processing, SMS.

I. INTRODUCTION

A. Background

Amharic is the official language of Ethiopia that belongs from Geez and the language of Ethiopia that have been since 4th century AD. It is becoming to the second mostly vast semantic language next to Arabic in the world. Basically in Ethiopia they have more than 86 languages is speech over 90 million and above people. In the Ethiopian context, Amharic scripts used to industrial, commercial and political importance received the earliest, because of this they mentioned the above paragraph Amharic is the official language of Ethiopia.

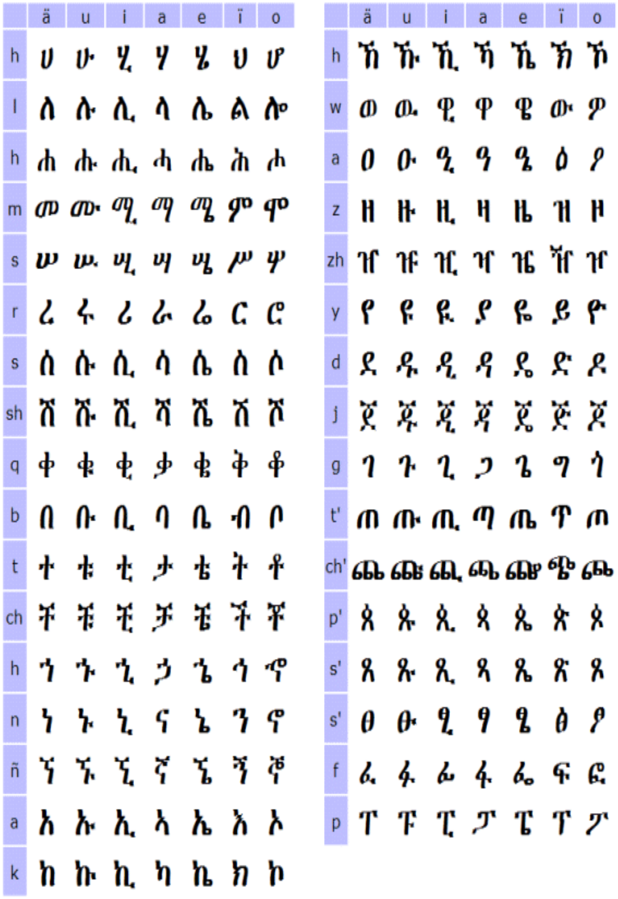
Current Ethiopic character is almost the same with recent Ethiopic character except some modification and they have 34 base characters with 6 tubular of column orders representing derived vocal sound basic character. Totally the script that used by Amharic have 265 characters including 27 labialized characters mostly represented two sound e.g.ቋ for ቅዋ .this all basic character is written in tabular format of columns and all the first character is the base character and other is derived characters. Amharic text prediction system in a mobile phone is that facilitate easily data entry techniques on a keyboard. In the Ethiopian context, People cannot type as fast as they think in Amharic character, especially when faced with the constraints of mobile devices.

B. Motivation

Amharic text prediction in a mobile phone has its own impact, regarding the social, economic and political development of a country. Since most computers work in English and other few languages, people who do not write and speak such languages are either forced to access computers as well as mobile phone in those languages or will not use them at all. Then it stands to do this thesis. In order to increase the usability of mobile phone and let people express their ideas using their local languages on mobile.

C. Problem definition

They have problems related to writing text and other some necessary information like save and search names of contact in The Amharic language. Basically, most of mobile user in Ethiopia is not well understand English language. According to this and a related problem it should have to handle this problem to develop an application Amharic text predict for mobile phone. So, why the Ethiopian people are not using own language? And why Amharic language is as a part of predictive language technology? In addition, Amharic language can give service alternative text prediction and entry method for SMS and any other social network like Facebook chatting and emailing of communication.



	ä	u	i	a	e	ï	o		ä	u	i	a	e	ï	o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ	h	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ	w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
h	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ	a	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ	z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ	zh	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ	y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ	d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
sh	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ	j	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	t'	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ	ch'	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
ch	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ	p'	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
h	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	s'	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ	s'	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
ñ	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ	f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
a	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ	p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ								

Figure 1.1 Overview of current Ethiopic character

D. *Objectives*

The objective of this research work to develop an application of Amharic text prediction system for a mobile phone.

E. *Research methodology*

The methodology of this research work was doing the following task.

Literature review: -It was been shown and understand related literature from different sources like observation, media, books, journals, Internet.

Data collection: - Data was been collected and stored this all filtrated data from the database.

Data analysis: -It was been created database, Based on collected data

Select implementation tools: -In this study, a number of tools are used like eclipse, android operating system, windows operating system SQLite database, a problem that is going to be addressed.

Design: -An overall design of the text input is implemented.

Implementation: - An overview coding, testing and maintenance of text input method is done.

Experiment: -Experiment measures the performance of the developed system.

II. DETAILED STUDY OF THE PROBLEM AND LITERATURE REVIEW SURVEY

Saied B. Nesbat [4], proposed the input methods can be sub dived into five text entry techniques in every mobile phone. Those are Multi press, two-key, three key, predictive and touch screen.

Multi press text entry technique

This is most common and easiest techniques text input method before a time that coming predictive and touchscreen text input method. This input technique sometimes call as multi-tap techniques and the user have to press each key at list one time and at most four times to enter one character.

Two key text entry techniques It is the second method which needs at list two key presses to write a single character. Where the first press indicates the desired key or group of character and the second press identifies the position on that key. For example, to write the character "C " then press the labelled 2 for selection of group and next press labelled 3 to select the character position. Therefore, a user should press 2 followed by 3 to enter the character "C".

Three key text entry techniques Gudisa [38], has proposed a greater understanding or improve main machine communication on three key text entry techniques. As the name implies, they need exactly three key to type a single character. The three key methods much related to two key input method and we have adopted for Ethiopic text input method.

Predictive text entry technique It is the fourth and most adaptive text entry method. This method is used in both key-based and pen-based paradigms. It is a dictionary based method (for example, T9 from Tegic Inc., iTap from Motorola, and eZiText from Zi Inc.) and it requires only one key press to enter each character. Predictive texts are classified in three ways.

Word Completion: - is the kind of word prediction system which is currently usable in the applications like Microsoft-Word and Microsoft-Excel. **Bigram/Trigram Prediction:-** which uses two/three-word patterns and their corresponding frequency for the sake of prediction. **Linguistic Word Prediction**: - in which the system knows the grammatical value of each word in the dictionaries (for example, Co-Writer) **Touch screen text entry technique** As **Scott macKenzie [17],** proposed touch screen keyboard is another kind of data entry technique which uses fingers to enter the data. This technique is more similar to the soft keyboard.

III. AMHARIC TEXT PREDICTION SYSTEM MODEL FOR MOBILE PHONE

This chapter, the developed Amharic text prediction model, the architecture of the text prediction Engine and the developed algorithms will be presented briefly.

A. Amharic text Prediction Model

To accomplish a task of Amharic text prediction, one has to have statistical information such as the frequency of occurrence of words. This can be achieved by using a corpus. Since the Amharic word corpses are not available easily, we prepare the corpses from various source that include Private and government newspapers, journals, social media like Facebook, Whats App, twitter, Instagram, Viber, Tango, Books which are written by different authors on different issues such as politics, religion, history, fiction and love. Around 2406 different files, collected from various sources mentioned above, are provided to the tool and a total of 1193719 words are generated filter 242383 unique words. All the 2406 files are converted to txt format in order to be used by the visual studio C# tool. All this collected word which we refer it as: "**Alembante Mulu Text Prediction Corpus**" that is used in this work is, thus, composed of 242383 unique word words. Table 3.1 show the lists of top 15 most frequently used distinct words, their frequencies, and the percentage of occurrence of that word compared to the total size of the corpus.

No	Words	Freq. of Occurre nce of Words	% of Words	Word Length
1	ነው	19917	1.66	2
2	ላይ	12152	1.01	2
3	ውስጥ	5472	0.45	3
4	ግን	5402	0.45	2
5	ወደ	5204	0.43	2
6	እና	4868	0.40	2
7	ነገር	4337	0.36	3
8	ጋር	4262	0.35	2
9	ነበር	4415	0.36	3
10	ጊዜ	3404	0.28	2
11	አንድ	3318	0.27	3
12	ደግሞ	3252	0.27	3
13	ብቻ	3009	0.25	2
14	ሰው	2986	0.24	2
15	ምን	2870	0.24	2

Table 3.1: List of Top 15 unique Words Extracted using the C#

B. Architecture of the System

The Word Prediction Engine has three components which are participating in the prediction Process; those are Start Engine, Word Selector and Word Ranker.

The **Start Engine** component, for the first time, gets two recognized characters one by one from the character recognition engine. After these two characters are received, this component waits for the time frame set up, which is 2000 milliseconds, to initiate the Word Selector component. To start predicting the intended word the prediction engine needs two or more characters. Once the **Word Selector** is initiated, it will start searching for words in the dictionary(Lexicon) of which their first two or more characters match with the recognized characters.

Word Ranker, by considering the frequency of each word, provides a rank to each word in the list of found words. Words with highest frequency will get highest rank and those with least frequency will get the least rank. In the case of two or more words having the same frequency; the word ranker will decide the rank of the words by considering their alphabetical order and predict based on Bigram/Trigram Prediction. As mobile devices have a result of these trials, it is found that displaying four words at a time will be much convenient.

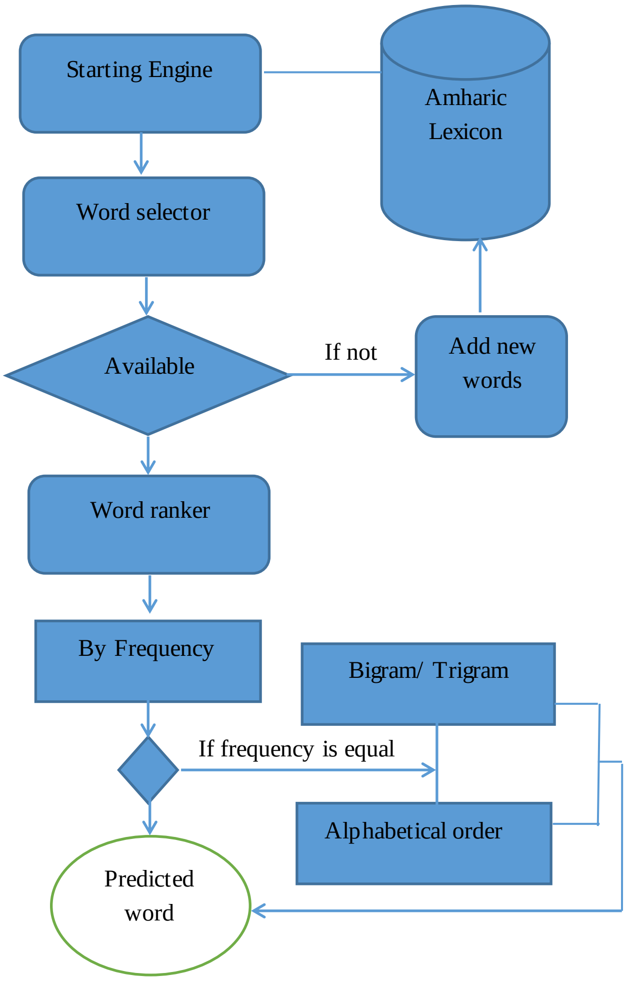


Figure 3.1 Architecture of the system

IV. IMPLEMENTATION AND EXPERIMENT

A. *Implementation*

Developing a prototype to demonstrate the validity and usability of the proposed text prediction system is one of the objectives of this work. In order to implement the algorithms and make the necessary experiment on the system, we have used different tools and development environments.

B. *Used Tools and Development Environment*

Eclipse is one of the android mobile and tablet application development platform. Eclipse SDK (eclipse Software Development Kit), which runs under Windows operating system, contains the eclipse SDK manager, class libraries and SQLite database. Eclipse SDK has three actives: Blank activity, full screen activity and Master/Detail flow.

Android Emulator is a virtual mobile device that runs on your computer. The emulator lets you develop and test Android applications without using a physical device.

SQLite Database has methods to create, delete and execute SQL commands in android, and perform other common database management tasks.

C. *Result of the Experiment*

The result, shown in Table 4.1, depicts the number of predicted and fully written words for the corresponding word-lengths. The column named as Top Ranked Predicted Word represents the number of highest ranked predicted words after two characters were written. These words are automatically displayed in the text box. The column, Predicted Words from Rank 2 to 4, shows the number of predicted words after two characters was written, but these words are found in the list of predicted words ranked 2 through 4.These words need to be selected from the list.be selected from the list. The last four columns indicate that the number of words either predicted after 3 or more characters or fully written words depending on the word length.

Word Length	Correct predicted Words	Incorrect word	Precision	Recall	F-measure
3-Leter	373	37	90.97%	40.54%	1.78%
4-Leter	300	20	93.75%	32.61%	1.82%
5-Leter	98	8	92.46%	10.65%	1.70%
6-Leter	78	6	92.86%	8.48%	1.67%
7-Leter	50	5	90.91%	5.43%	1.56%
8-Leter	14	1	94.11%	1.52%	1.16%
9-Leter	7	1	87.50%	0.76%	0.81
Accuracy	920	80	91.79%	14.28%	1.50%

Table 4.1 result of experiment

Precision: - that are % of selected item is correct on the system.in mobile prediction system is how to predict all words that are out of error.

$$\text{Precision} = \frac{\textbf{correct}}{\textbf{output length}}$$

Recall: - It is usually expressed as percentages that are percent of correct item are selected.

$$\text{Recall} = \frac{\textbf{correct}}{\textbf{reference length}}$$

F-measure: - a combined measures that access precision and recall trade off.

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

V. CONCLUSION AND FUTURE WORKS

Word prediction is a process of predicting a word that a user intends to write after the first few characters and based on a lexicon and statistical information obtained from the corpus. Text writing using an online handwriting recognition system follows a complex process that also requires training. When word prediction is used for handwriting recognition, the possible errors of character recognition can be minimized. To accomplish the task of word prediction process for Amharic language, a good collection of words which will be used as a corpus is a must. However, since there is no such corpus for Amharic language, collecting of words from different sources and preparing the corpus became the first task of this research. Analyses have been done on the corpus prepared. These analyses used to get information like the average word-length of Amharic language, the most frequently used Amharic word-length and the experiment conducted our word prediction system has shown a prediction accuracy of 91.79%. In this work, the searching efficiency of the system has not been taken into consideration.

REFERENCES

[1] Shimeles A., “Online Handwriting Recognition for Ethiopic Characters,” Department of Computer Science Addis Ababa University, pp. 8-19, 2005.

[2] Biadsy F, El-Sana J and Habash N., “Online Arabic handwriting Recognition using Hidden Markov Models”, In Proceedings of the 10th International Workshop on Frontiers of Handwriting and Recognition, pp. 105-109, 2006.

[3] Negussie D.,“Writer Independent Online Handwriting Recognition for Ethiopic Characters” ,

[4] Department of Computer Science, Addis Ababa University, pp. 29-36, 2006. Nesbat S., “A System for Fast, Full-Text Entry for Small Electronics Device”, Proceedings of the Fifth International Conference on Muultimodal Interface, ICMI 2003 (ACM-sponsored), Vancouver, pp. 55-56,2003.

[5] Abebe S, Seyoum T, Atnafu S, Kinde Kassegne S., “Ethiopic Keyboard Mapping and Predictive Text Inputting Algorithm in a Wirless Enviroment”, ITEs-2004, Addis Ababa, Ethiopia, pp. 39-41,2004.

[6] Vural E, Erdogan H, Oflazer K, Yanikoglu B. "An Online andwriting recognition system for Turkish", in Proceedings of SPIE Electronic Imaging Symposium, pp 3-5. 2005.