RESEARCH ARTICLE OPEN ACCESS

Predictive Text Entry Method for Somali Language on Mobile Phone

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

ABSTRACT

To enter data into mobile phones, there are different data entry techniques that can be used like keyboards, stylus, soft keys, speech etc. It can be used these techniques to enter data into mobile phones which may be either in predictive format or non-predictive format. Text prediction is a technique which predicts what the user wants to write. The user writes some characters and the system predicts the remained ones. Text prediction reduces the redundancy and makes the system more efficient because of time saving. In this research work, it was prepared text prediction model for Somali language. To extract the statistical information like the frequency of each word, 119923 Somali words was collected and analysed. This information was used to determine the n-gram model which is after how many characters the system starts prediction that was decided to be two, and to know priority of the words and the most frequently used one has the highest priority to predict. To check the accuracy of the text prediction system, an experiment was conducted and 92.83 accuracies were achieved.

Keywords:- Text prediction, Somali Text Prediction Model, Text prediction corpus, N-gram model for Somali Text Prediction

I. INTRODUCTION

A. Background of the study

Somali language is originated from different languages like Arabic, English Italian etc. Somali language is spoken by different countries like Somalia, Somaliland, Ethiopia, Djibouti and Kenya, and also the Somali people who live in abroad. The Somali language is one of the Cushitic Afro-Asiatic families of languages. The Cushitic contains around 40 different languages which are mainly spoken in Ethiopia, Somalia, Somaliland, Djibouti, Eritrea and Kenya. These languages can be subdivided into East, South, Central and North Cushitic languages [1]. Somali language is one of the east Cushitic languages and it has some relation with some other languages in the Cushitic family like Afan Oromo and Afar languages. Somali language is one of the most popular and widely used languages among Cushitic languages because more than five countries use as primary language [2]. Before the colonial power came in Africa Somali language was using Arabic script, but the British, Italian and France came and colonize the area that the Somali language was spoken, they made some influence and the language is converted form Arabic script to Latin script [3]. Today, Somali language is mainly used all the 26 letters of the English-language alphabet except for P, V and Z.

To communicate with a machine it needs a process from data entry to output, to get the output, a data should be inputted by using one of the data entry techniques, the system processes and produces output. To improve the efficiency of the system better data entry technique should be chosen. One of the best data entry techniques is text prediction. This technique reduces the redundancy and increases the speed of the system. Text prediction is the technique which the user writes part of the intended word and the system completes the rest by suggesting the suitable words according the preceding characters [4]. To make suggestions the system compares the frequencies of the words and lists by decreasing order; means the word with the highest frequency will be the top of the selected list. The development of the technology increases and there are many applications which the users are mostly using on their mobile phones like SMS, viber, facebook, whatsapp and all the social networks. All these applications need an easy, fast and efficient data entry technique. This gives moral and more energy for the researchers to develop new systems which are suitable for all the applications and gives the users suitable interface. The most important area that all the researchers are focusing on is text prediction, because it is needed a system that saves time for users and improves the correctness of their text. Moreover, text prediction gives suitable and preferable interface for elderly and disabled people. So to get an efficient and suitable system text prediction should be used as data entry technique [4], [5].

B. Motivation/Significance

The development of the technology is increasing rapidly and reaches where every country and even individuals takes the advantage. The development of a country is dependent on how much it uses and utilizes the technology. To improve the utilization of the technology it should be localized. The people in a given country should have the technology which supports their socio economic live, like their language, if this happen and the people able to get technology which followed by their culture, then they can participate the development of the technology. Most of the computers use specific language like English and view others, this makes problem for the usability of the technology because those who don't know these languages faced a problem which is how to use these devices, there is only two options: they have to learn these languages or will not use those devices. This is a big obstacle for the development of the technology, to prevent this obstacle and increase the usability of handheld devices it is needed to localize and let the people to express their ideas in their own languages. After observed these problems which jeopardize the development of the technology this research became necessary to alleviate these problems.

One of the most needed research area which demands the researcher's attention is text prediction in a particular language. Text prediction is important for text writing because it reduces writing time and improves the correctness, this improves the quality of the system.

All the above mentioned reasons are motivated us to do a research for text prediction technique on mobile phones for Somali language to utilize the Somali people and play their crucial role to the technology advancement.

II. LITERATURE SURVEY

`SachinAgarwal and ShilpaArora(2007)[6], mentioned that the technology is increasing and the usage of the mobile phones is also increasing in parallel because of the wireless technology. The usage of the mobile brought new applications like short message services (SMS).these applications need a fast and efficient data entry method. So according to the authors this method is predictive text entry because this method increases speed of the system and reduces time. Predictive method is one of the easiest and most efficient data entry methods.

Barry McCaul and Alistair Sutherland(2004)[7] mentioned that whenever data entry is required in virtual environment means soft keyboard, the user is presented with a graphical representation of a keyboard, by using finger of stylus pen. After a sequence of finger flexes the user is presented with the predicted word. Users may go through to rotate alternative matching words to indicate the desired word, if the initial prediction is incorrect. The user gets an ordered list of words which are ranked according their probability and the most suitable word will be selected for the input.

GudisaTesema(2013)[4] , proposed system that reduces the ambiguity that the was present in previous systems which were not dictionary based , so as the researcher said to reduce this ambiguity a dictionary based disambiguation should be used by adding dictionary to the system. For example, T9 from Tegic Communication., iTap from Motorola, and eZiText from Zi Corporation and it requires only one key press to enter each character. In this method, when the key is pressed, the system compares the key sequence with the word possibilities in a linguistic database to guess the required word. When two or more words match the given key sequence ambiguity is possible. Under such condition, the word with the highest frequency of occurrence is chosen. Moreover, a down-up arrow or a special "next" key is used to choose an alternative word if the intended word is different from the displayed one.

Lee Butts, Dr. Andy Cockburn (2001)**[8]** said, if the keyboard has less than 26 keys, more than one character should be mapped to the same key because there is 26 character and the keys are less than 26 so the available keys should be shared. For example in the standard 12 key, there is only 12 keys and its needed to map the 26 characters to these keys. Since there are unequal number of keys and number of character, multi_ press is needed to input data. In mobile phones, the multi_ press is commonly used, although nowadays some new keyboards are increasing in use. The multi_ press can be categorized in to two different techniques which are multi_ press with timeout and multi_ press with next button. Nowadays typically a combination of both concepts is used; the selected character is either approved after a timeout of some seconds or by clicking the next button. Our implemented multi-tap input concepts differ in the number of keys and in the arrangement of the characters on these key.

HedyKober. (2001)[9], explained the main reason to explore and examine the efficacy of mobile keypad text entry is the growth of Short Message Services (SMS) messages on mobile phones, especially in Europe. As the researchers mentioned, in earlier times the majority of SMS users were using multi-press as their method of text entry. Always there are some errors when the users typing on mobile keyboards with multi_ press keys. These errors are mainly comes from

hitting keys adjacent to the desired key. Mobile phones, generally, are not suited to text input. Earlier mobile phones were using an interface which contains 12-15 keys. These keys deal with the 26 letters of English alphabets, punctuation marks and numerical characters. Since the number keys are less than the number of characters that they are dealing they should be overloaded, from key number 2 to key number six and key number eight contains three letters each while keys number 7 and 9 contains four letters. For example, key number 5 is mapped to J, K, L while key number 9 is mapped W, X, Y and Z. Each of these keys may also be assigned additional special characters and punctuation marks. So that, since each key contains three or more characters. We need a method that specifies the character that we want, when we press a specific key. For example if we press key number 5, the mobile phone does not know which letter do we need because there is three possibilities which are J,K and L. To solve this problem most of the mobile applications apply the Multi_ press with time out, Multi_ press with next button and two key with predictive text inputs. These three input methods may access a database for checking spelling error after the word is entered, before these three methods there was another type of text input that uses a keyboard which inserts a single character by applying multi_ press mechanism using the next bottom, this method was very simple because it does not need to access database Multi_ press with time out, Multi_ press with next button and Two key are commonly used for many different languages spoken in the world [4],[5],[7],[12].

Amal Sirisena(2002)[10], explained that there are different types of input techniques which can be used mobile phones. Key based text method is from standard keyboard which are more ambiguous since each key contains three or more letters to the keyboards which are less ambiguous since each key contains one character like fictitious alphabetic keyboard which have different keys for upper and lower cases [5],[11]. mobile keypad based text entry method is improving time after the time, parallel with the improvement of the technology, the usage of mobile phones are increasing with different purposes like text messaging service, this initiates the key based text entry method.

III. WORD PREDICTION MODEL FOR SOMALI LANGUAGE

A. The Word Prediction Model

In this research work two main statistical methodologies are used. These methodologies are frequency of usage or word counting and recently used. The first methodology is based on the generality of the language means it is not consider the usage of the individuals. This methodology selects the words which are frequently used when a specific word is inputted and suitable word should be select for next. The second methodology is recently used, this methodology uses the concept that if the word is used recently has more priority than the others and more likely to use again.

1) Statistical Prediction: Most of the prediction systems are based on the statistical analysis to predict what the user wants to write. When the prediction process is going, the probabilities of the words are used. This probability may be fixed or dynamic means adaptive.

Fixed Lexicon is one of the methods that use the statistical prediction and it is easiest method, the words of this method has fixed frequency that is used to predict that word throughout the general language. This method uses two techniques to make prediction, the first technique, all words are arranged in order based on their frequencies, and some few words are at the top of list. When the user wants to use this technique, first it is checked whether the desired word is among those words which are at the top. If it is so the user can simply select and put the sentence. If not the user enters the first character of the word that is needed to be part of the sentence and the system arranges the words and puts at the top few of the words that start this character, the user again checks whether the needed word is among these words if it is true then it should be selected if not the user again enters the next character. The process is going entering and checking until the words is entirely entered by the user character after character or it is short listed as a prediction list. The second technique of the fixed lexicon is more advanced than the first one because this technique considers how the words are following each other. When a word is entered the system predicts the words which are mostly followed this word and adds the prediction list. This technique is better than the first technique because this is basically depend on the current situation and always result correct prediction, that is why it became more familiar than preceding technique and most of the prediction designers use this technique [4],[13],[14].

An adaptive lexicon is another method for statistical prediction; this method changes the frequency of the words that the dictionary or lexicon contains when the user builds sentences. Rather than the fixed lexicon, adaptive lexicon considers the recently used words, when a word is used the priority of that word increases, and gives high priority to use it again [14]. The adaptive lexicon is somehow similar to the fixed lexicon, because both of them are frequency based and the frequency of the words are independently, means there is

no link between the preceding and coming words, it predicts the intended word only with its frequency, the main difference between fixed lexicon and adaptive lexicon is that the adaptive lexicon is dynamic and it dependent on the usage of the words, if word is used the frequency of that word increases and it will get high priority than the others, so it should be updated the lexicon, the other difference is if a new word is used which was not in the lexicon or dictionary it adds the dictionary with frequency of one.

To complete the task of word prediction we need to get some statistical information such as the words and their frequency. To get this information "**Word Prediction Corpus of Mahamed"** was prepared. Having in mind that a good collection of words results to design better model, a data was collected to prepare this corpus from different sources like newspapers which may be private or governmental, books which are written by different authors with different issues like political, cultural, education, love, religion, teenagers, women affairs, and so on. This corpus was used to get statistical information like frequency of the words, the average word length of Somali language. It is also used to decide after how many characters the system starts prediction which is N. The statistical information and the value of N were used to design the word prediction system and the algorithm that was implemented. This corpus were collected from 1030 files from different sources and contains total of 688830 words out of which 119923 of them are unique words. Table 1 shows the total number of words in the corpus, unique words among the total words, the average word length and the description of these unique words based on the word length. It is also mentioned in this table the percentage of the words based on the word length. To complete the preparation of this corpus we have also used different Somali dictionaries with different authors. Figure 1 show us the word length distribution based on five randomly selected Somali dictionaries. There are some words which have more than 16 characters in length and they are not mentioned here, since they are insignificant number. To develop the Somali text prediction system which predicts the word that user wants to write, we have decided after how many characters the prediction will start which is 2, based the information of the table 1 we have got that the most frequently used word length of Somali language is 8 characters long and the average word length of Somali Language is 7.2.

Table I Description of the Word Prediction Corpus of Mahamed

Data Item	**Description of the corpus**	
Total Words	688830	
Unique words	119923	
Average word length	7	**Percentage of words based on length**
1-character	123	0.1%
2-characters	315	0.26%
3-characters	1060	0.88
4-characters	2830	2.36%
5-characters	7602	6.34%
6-characters	12529	10.45%
7-characters	15361	12.81%
8-characters	18701	15.59%
9-characters	17215	14.36%
10-characters	14726	12.28%
11-characters	10998	9.17%
12-characters	7320	6.1%
13-characters	4744	3.96%
14-characters	2828	2.36%
15-characters	1686	1.41%
16-characters	975	0.81%
17-characters	558	0.47%
18-characters	206	0.17%
19-characters	92	0.08%
20-characters	45	0.04%

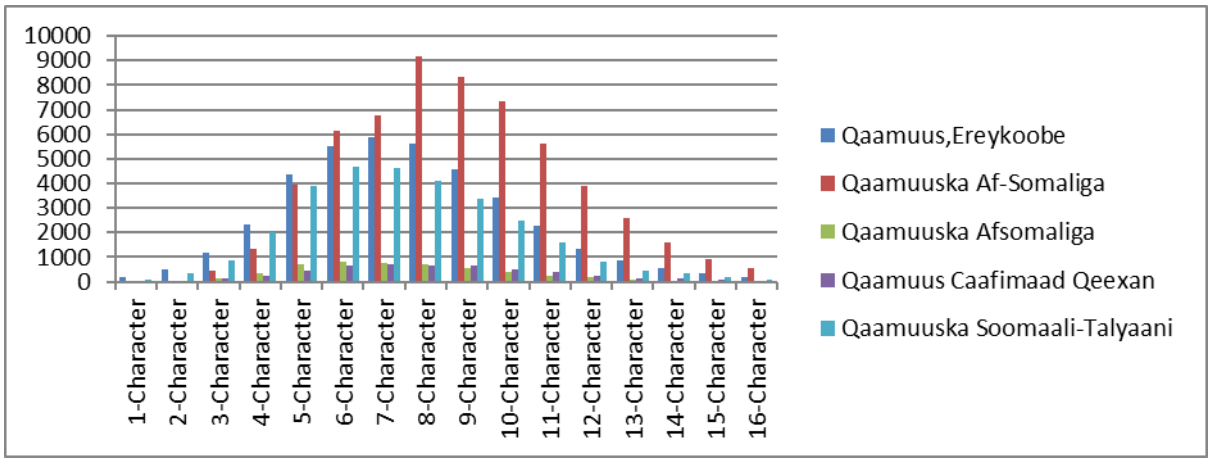

Figure 1 Word-Length Distributions in the Reference Dictionary

IV. EXPERIMENT

To evaluate the accuracy of the developed system which is Somali text prediction, an experiment was performed this experiment clarified the accuracy of the system by using sample data which was collected from different sources. These sources were mostly collected from private newspapers, social media like Facebook, twitter, and some other social medial where the people meet. The sources that the test data was collected were around 30 different sources, each of them are randomly collected from 50 words which makes the total words of this test data 1500 words. So these words were tested to the system to check the accuracy of the system. The experiment was conducted on Android emulator.

A. Result of the Experiment

After the experiment of the system, the result of the experiment is mentioned in this section, and it is explained how the result of the experiment was interpreted. The result is discussed in tubular form as shown in the table 2; this table contains 16 rows and 4 columns. This table explains the test data based on their word length. Table 2 classifies this data in to two parts the first part is predicted words while the rest is non-predicted words which the user must write. It also mentioned in this table how many words do the system predicts among this test data and how the system predicts based on the word length. The column name collected words represent the total test data based on the word length, the column named predicted words represents total predicted words among the collected test data based on the word length while the last column which is named as non-predicted words indicates the number of words which the system does not predict or fully written words by the writer for the corresponding word-lengths. In addition to this, table 3 shows the percentage of the predicted words based on the word length.

The total percentage of the predicted words among the test data is 91.83%. Based on this experiment we can conclude the total prediction accuracy of our system which Somali text prediction is 92.83. Furthermore this accuracy can be improved during the usage of the system. Because we used the adaptive technique where there is an option which can be added the non-predicted words when the user writes at the first time and it will be predicted if it used again.

Table II Number of Predicted and Written Words

Word length	collected words for test	predicted words	non predicted words
3-Characters	62	61	1
4-Characters	150	149	1
5-Characters	263	258	5
6-Characters	240	231	9
7-Characters	209	202	7
8-Characters	215	202	13
9-Characters	127	115	12
10-Characters	100	89	11
11-Characters	68	58	10
12-Characters	40	35	5
13-Characters	13	11	2
14-Characters	10	8	2
15-Characters	2	2	0
16-Characters	1	1	0
Total	**1500**	**1422**	**78**

Table III Prediction Accuracy of Each Word Length

Word length	Predicted words	Prediction accuracy
3-characters	61	98.29%
4-characters	149	99.33%
5-characters	258	98.1%
6-characters	231	96.25%
7-characters	202	96.65%
8-characters	202	93.95
9-characters	115	90.55%
10-characters	89	89%
11-characters	58	85.29%
12-characters	35	87.5%
13-characters	11	84.62%
14-characters	8	80%
15-characters	2	100%
16-characters	1	100%
Average Prediction Accuracy		**92.83**

V. CONCLUSION

This paper is mainly focused on how the Somali language users get an easy system that support them whenever they are using their mobile phones like when they writing messages. This system allows them to write some characters and the system predicts the remained characters of the word that they want to write. This is more helpful for elderly and disabled people to write their messages easily without spelling errors.

Somali language is a language which has the Latin script with long word length. The average word length of Somali language is 7.2. This makes difficult to write the whole word for the users without prediction, fortunately, our system solves this problem. This system contains 119923 unique words that were used as a lexicon or dictionary and it is updatable during the usage of the system. After experiment it was concluded that the accuracy of the system is 92.83%.

ACKNOWLEDGEMENTS

All praises and thanks are due to Allah (SWT) The creator, the exalted, sustainer and the most merciful, who gave us the courage, endurance, willingness and ability to complete this paper successfully. We would like to acknowledge Khadar Dahir Abdi and Farah Mohamed Hassen for giving us variable ideas and guidance from the very beginning.

REFERENCES

[1] Martin Orwin. (1995). *Complete Language Course*, Retrieved from: http://www.somalicsc.org/wp-content/uploads/2013/07/Somali-Language-Learning-Resource-List-2013-14.pdf

[2] Abraham, Major R.C. *Somali-English Dictionary*, London: University of London Press. 1962.

[3] Armstrong, Lilias E. *The Phonetic Structure of Somali*, Westmead: Gregg International Publishers. 1964.

[4] Gudisa tesema. *Design and implementation of predictive text entry method for afan Oromo on mobile phone*, Addis Ababa Institute of Technology (aait), Electrical and Computer Engineering Department.2013.

[5] MacKenzie, I. S., &Soukoreff, R. W. *Text Entry for Mobile Computing Models and Methods,Theory and Practice*, Human-Computer Interaction, 17, 147-198.2002.

[6] Sachin Agarwal and Shilpa Arora. *Context Based Word Prediction for Texting Language*, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213.2007.

[7] Barry McCaul and Alistair Sutherland. *Predictive Text Entry in Immersive Environments*, Proceedings of the IEEE Virtual Reality 2004 (VR'04), P: 241. 2004.

[8] Lee Butts and Andy Cockburn. *An Evaluation of Mobile Phone Text Input Methods.* Human-Computer Interaction Lab, Department of Computer Science, University of Canterbury,Christchurch, New Zealand. 2001.

[9] HedyKober, Eugene Skepner, Terry Jones, Howard Gutowitz, and Scott MacKenzie.*Linguistically Optimized Text Entry on a Mobile Phone*.2001.

[10] Amal Sirisena. *Mobile Text Entry, Department of Computer Science*,University of Canterbury, Christchurch, New Zealand. , 2002.

[11] Jacob O, Wobbrock, Brad A. Myers, and John A. Kembel. *A Stylus-Based Text Entry Method Designed for High Accuracy and Stability of Motion*.2003.

[12] Kumiko Tanaka-ishii. *Word-Based Predictive Text Entry Using Adaptive Language Models*, Natural Language Engineering 13 (1): 51–74, Cambridge University Press. 2006.

[13] Vladimir N. *Vapnik.The Nature of Statistical Learning Theory*, Springer, New York.1995

[14] Y. Yang, *An evaluation of statistical approaches to text categorization*, Technical Report CMU-CS-7--27, Carnegie Mellon University. 1997.