Sentence Prediction on SMS in Sinhala Language

M. S. Karunarathne*, L. D. J. F. Nanayakkara **, Kapila Ponnamperuma **

* Department of Computing & Information Systems, Sabaragamuwa University of Sri Lanka
** Department of Industrial Management, University of Kelaniya

Abstract- There is a rapid growth on Information Technology through e-Government concept and the usage of mobile phone is also increasing day by day. The majority of SMS writers in Sri Lanka transliterate messages because of language barriers, which create various communication problems and ambiguity of messages. Computing in Sinhala language is an emerging trend in Sri Lanka. This research is an attempt to predict Sinhala sentences in mobile short messages. This is a timely necessity in Sri Lanka.

The main advantage of the research is developing an effective algorithm for reducing the typing effort, saving time and avoiding language ambiguity. The approach is not based on dictionary but by identifying a user's writing patterns through several dimensions such as time, trend and social user groups. The mobile application predicts words through an identified language model and tests the system in a given testing corpus through the emulator. The main deliverables of this research are a language model for Sinhala Short Messages, an efficient and optimized algorithm to predict words in sentences, and a testable emulator. The algorithm is created based on Genetic Algorithm considering time series and user categories. The proposed algorithm has been tested with the existing system (without predicting feature) based on three measurements such as performance, accuracy and efficiency. This system is personalized; hence the capacity to applying algorithm to other users' data was validated by installing and testing application with two other personalized message suits.

Index Terms- Sinhala SMS, Genetic Algorithm, Sentence Prediction, Mobile Intelligence

I. Introduction

The mobile phones have been used in Sri Lanka for over two decades. Currently most mobile phones and applications in Sri Lanka work only in English although only 3-5% of Sri Lankan population literate in English while 91.4% literate in National language in 1999 according to ICT profile [8]. Throughout this period, the Sinhala language computing has also evolved gradually. However, compared with other Asian countries, the use of Sinhala language in ICT is at the basic level due to attitude and small Sinhalese market [4]. Nowadays people try to compute in mother tongue and most of the new mobile phones support Sinhala. As an initiative to computing in Sinhala, SLS 1134 Unicode standard and character set were developed by Information and Communication Technology Agency (ICTA) of Sri Lanka.

Text messaging or texting refers to the exchange of brief written messages between fixed line phone or mobile phone and fixed or portable devices over a network. Text messaging terminology is changing from region to region. It may simply be referred to as a text in European countries where it is called as SMS in Asian countries. The Sri Lanka mobile market, as a competitive growing market, 841 Sri Lankans own a mobile phone out of each 1000 Sri Lankans [2]. Unlike adults most young people (age between 19 -24 years) use to writing text messages. One limitation of text message is the maximum number of letters per message. As a result most SMS writers have developed their own abbreviations. Hence the language becomes unstructured and colloquial. The other limitation is higher consumption of battery power for texting. Hence an effective algorithm for predicting and suggesting words by reducing typing time and number of keystrokes per word is required.

This research is an attempt to develop an efficient algorithm for mobile phone applications to predict Sinhala sentences in SMS. Genetic Algorithm (GA) has been used because it is a natural algorithm and has fewer computations which overcome several constraints such as less memory, limited processing power, battery power and screen size in mobile phones.

II. RELATED WORK

Sinhala is an Indo-Aryan language, spoken in Sri Lanka by about 13 million people, and also known as Sinhalese or Singhalese [9]. The standard character set of Sinhala language includes 56 characters in a number of categories such as consonants, vowels and semi-consonants.

Independent Vowels [18]
අආඇඈඉඊඋඌඍඎඏඐඑඒඓඔඕඖ
Dependent Vowels [22]
◌ා ◌ැ ◌ෑ ◌් ◌ි ◌ී ◌ු ◌ූ ◌ෘ ◌ෲ ◌ෟ ◌ෳ
◌ෙ ◌ේ ◌ෛ ◌ො ◌ෝ ◌ෞ
Sinhala Consonants [47]
කඛගඝඞඟචඡජඣඤඥඦටඨඩඪණඬතථදධනඳපඵබභමඹ
යරලවශෂසහළෆ

Sinhala, as a language is different for speaking and writing. Most people who write Sinhala SMS words use colloquial sentences rather than grammatically correct sentences. The Sinhala Unicode encoding [9] is a major milestone of Sinhala language computing.

There are many literature reviews on natural language processing (NLP) for computers but there are less literature reviews on NLP for mobile phone domain. The literature reviews on NLP are divided into two categories as word prediction and sentence prediction. This research is mainly focused on sentence prediction rather than word prediction because meaning of sentences should generate by a series of individual words.

For the prediction of sentences there are several methods such as Neural Networking, Knowledge Base Systems (KBS), Genetic Algorithm (GA), Statistical Inference (SI) and Markov Transition Processes. It is important to identify a feasible and efficient algorithm to implement a solution for mobile phones.

A. T9 Technology [3]

The algorithm in the T9 technology is an optimized and compressed algorithm which compresses 1 byte per word. The main drawback of the above algorithm is that it over-generates words which are sometimes visible to the user as 'junk words' and the database size (30 -100 kb) is high.

A comparison of the potential algorithms for the research is shown in Table I:

Table I: Comparison of Available Methods

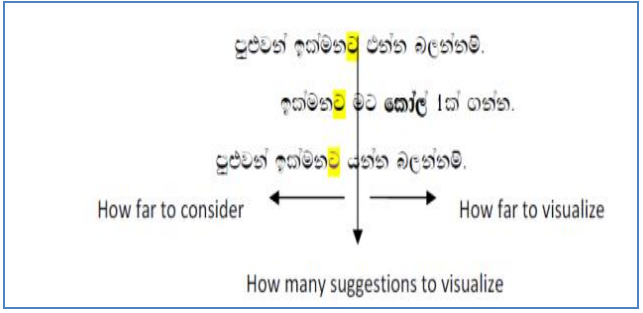
Method of prediction	GA[1]	Neural Network [5]	SI[7]	KBS
Memory Usage	Less	Less	Very High	Less
Ease of implementation	Easy, natural	Difficult, need to develop network	Easy	Time consuming
number of calculation	Less	High, complex formulae	Less	Less
Dependence on data	Low, predictive method	Low	Very high	Low

The Genetic Algorithm was selected for the research as the base algorithm considering low memory consumption and ease of implementation.

III. METHODOLOGY

Mobile phones are used to communicate with each other where the speed and convenience are valued. Since mobile phones are used by almost every person regardless of his or her language proficiency, it is important to develop the capability for text messaging in the local language. Sinhala is a well-structured and complex language. So assigning each character to the nine keys is difficult and writing messages using nine keys which are used for several letters is also cumbersome. A sentence prediction algorithm is to be developed for successfully overcoming these limitations and weaknesses.

Developing a strong language model is difficult because the SMS writers do not follow language rules. It is also difficult to develop a fixed vocabulary because the terms in SMS dynamically change from time to time and from person to person. The main challenges of predicting words can be demonstrated as in the following example:


පුළුවන් ඉක්මනට එන්න බලන්නම්.
ඉක්මනට මට කෝල් 1ක් ගන්න.
පුළුවන් ඉක්මනට යන්න බලන්නම්.
How far to consider
How far to visualize
How many suggestions to visualize

Based on literature review and the comparison of Table 1, a hybrid model (Figure 1) is using both statistical inference and natural language processing.

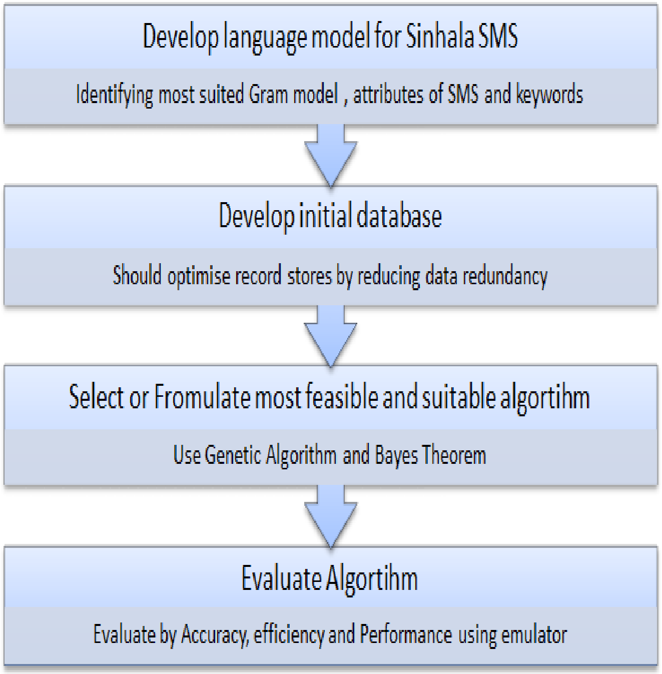

Develop language model for Sinhala SMS
Identifying most suited Gram model , attributes of SMS and keywords
Develop initial database
Should optimise record stores by reducing data redundancy
Select or Fromulate most feasible and suitable algortihm
Use Genetic Algorithm and Bayes Theorem
Evaluate Algortihm
Evaluate by Accuracy, efficiency and Performance using emulator

Figure 1: Methodology used in the Research

A. Language Model Observed

Initially, in a user study their SMS writing patterns, terminology and key attributes of SMS were identified. SMS messages were collected for a six month period from 30 Users in 5 categories. Two main attributes of SMS messages were found as Receiver Categories and Time Series. There are sub attributes such as Venue and Purpose but those attributes are interrelated with the former main attributes and they are reflected in main attributes themselves. Each message has a referring time, although some messages have particular referring time.

E.g. In April, Most SMS has New Year (Aurudu) Greetings.

Based on the importance of time period user can configure the minimum time unit as month, year, week, day etc. This research is based on periods of one month. Receiver category also should be configured based on user where a pattern can be seen in it. Set Diagram for user category and for time wise is developed.

B. Gram Model

An n-gram model is a type of probabilistic model for predicting the next item in such a sequence. N-gram models are used in various areas of statistical natural processing and genetic sequence analysis. The Markov assumption is applied in n-gram model as a base. The suitable n-gram selection is tested with training corpus for the sentence as Table 2.

Table II: Results of n-Gram model Test

Gram	Length of the list (excluding first letter)	Length of the list (including first letter)	Number of absences	Percentage
1	114	114	4	80%
2	6	15	14	84.375%
3	6	15	16	85.185%
4	6	15	32	71.42%

When we consider the four combinations in Table II, the following observations can be made.

1. Length of the list (excluding first letter in sentence) decrease and become constant at 6

2. First letter of a sentence prediction is unreliable in all combinations.

3. The number of absences of correct prediction is rapidly increasing when the number of letters is increasing.

4. The probability of predictions between first four suggestions in the list increases up to a certain point and then rapidly decreases.

It is important to identify the relationship between auto-generated suggestions and the target words of sentences. Hence the Regression analysis was done. Since Multiple R (co-efficient of correlation) is 0.667944, there is a positive relationship between absence of suggestions and appearance of target output in the four suggestion of the list. The point where absence of prediction is minimizing and the appearance of target word in the list is increasing is well suited with 3-Gram model. 3-Gram model is applied for the research.

C. Develop the Initial Database

The memory is one critical factor to develop the conceptual model for the mobile phone. Indexing and relational database mechanisms have been used to optimize the memory usage and reduce data redundancy. Record Management System (RMS) is both an implementation and API for persistent storage on Java ME devices. Data is stored and must be retrieved from the Record Store using a Byte array.

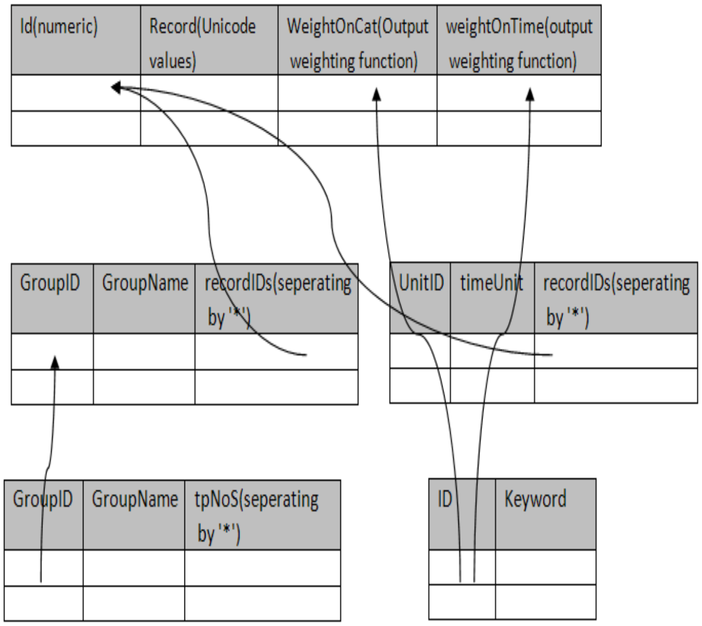

Id(numeric)
Record(Unicode values)
WeightOnCat(Output weighting function)
weightOnTime(output weighting function)
GroupID
GroupName
recordIDs(seperating by '*')
UnitID
timeUnit
recordIDs(seperating by '*')
GroupID
GroupName
tpNoS(seperating by '*')
ID
Keyword

Figure 2: Table Structure

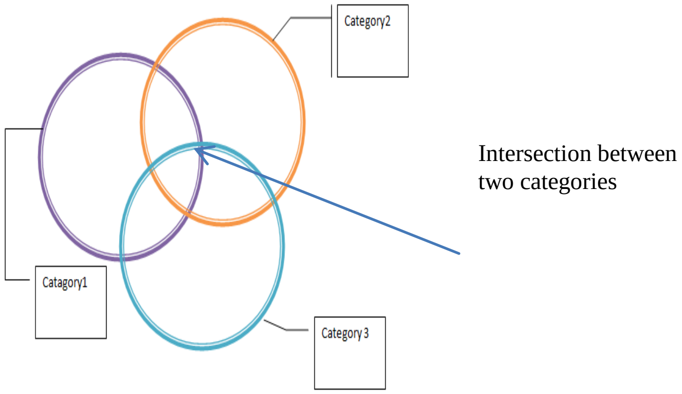
D. Feasibility of Genetic Algorithm [6]

Tournament selection is more suitable for mobile phone in efficiency. Boltzmann Selection is highly applied in dynamic environment but the word predictor is not dynamic. Steady State Selection is not applicable because it lost some parent genes. In a SMS predictor the same word can be repeated, so Steady State Selection is not applicable. Hence Roulette-wheel selection is applied.

E. Algorithm Development

1. Crossovering Mechanism

Venn diagrams which are based on two types of attributes in SMS writing were found. The intersection between the sets among categories, among time series and among both categories and time series can be seen. So it gets two- dimensional and the intersection between different sets will be the crossovering.



E.g.:- A = {Category 1 - Boarding friends}
B = {Category 2 - University friends}

$$A \cap B = \{\, friends\ in\ bording\ \&\ university \}$$

$A \cap B$ gets the mutual behavior between friends who are in boarding as well as university. In this research, we can find crossovering between three ways such as among receiver's categories, among time series and between receiver's categories and time series.

2. *Mutation Mechanism*

Mutation mechanism is mitigating the records by giving choice to delete as in normal process. However, system will automatically find the least keyword containing record and suggest removing the record to release memory.

3. *Weighting Mechanism*

The records are weighted according to the appearance of keyword on time series and categories. The basic algorithm in pseudo code is as follows:

```
If(User is in contact){
        Find category
        If(more than one category){
              Selection Set = union of categories
        }else if( only one category){
             Selection Set = intersection of category
        }
}else{
        Selection Set = All records
}

If(Selection set is not empty){
  Next letter/word = get Most Highest ranked word
}
```

IV. IMPELEMENTATION

The mobile application containing two parts; config and message writing is a predictive application by weighting the keywords and applying genetic Algorithm. In the config part of application, the language model and other configuration are configured necessary to run the application. Keywords can be modified according to time and category. The SMS writing part is basically for writing message in Sinhala and predicting next words. The application is developed by extending a GameCanvas class. So all the letters are images which is assigned unique Unicode value. Screen shots of word prediction are as Figure 3.

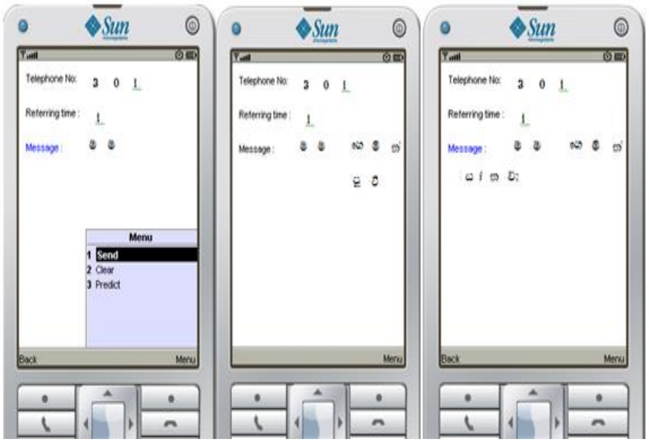


Figure 3: SMS Writer

V. EVALUATION

The new system will be compared with existing system using developed test corpus. Both of the application are developed using the same algorithm but for two different users message suites. The correctness and applicability of algorithm to different users is tested by conducting F Test as follows.

Where;
$$H_0 : \sigma_A^2 = \sigma_Y^2$$
$$H_1 : \sigma_A^2 \neq \sigma_Y^2$$

σ_A^2 is the variance of person A

σ_Y^2 is the variance of person B

The significance level is 0.05

Since P value (0.187743616) in time wise is higher than significance level, we do not reject null hypothesis under the 0.05 significance level. So time wise the algorithm has given the same output for both persons.

Since P value (0.084329096) in key stroke wise is higher than significance level, we do not reject null hypothesis under the 0.05 significance level. So key strokes wise the algorithm has given the same output for both persons.

VI. CONCLUSION AND FUTURE WORKS

In the evaluation phase, the following results are identified in proposed system.

Table III: Results of Proposed System

Measurement	**Person A**	**Person B**
The time necessary to generate each word	10.86 s	13.7 s
The number of keystrokes per a word	8.38	12.5
The percentage of tallying expected result and auto-generated corpus	62.33%	61.29%

The time necessary to generate each word in non-predictive existing system = 24.19 s
The number of keystrokes per word in non-predictive existing system = 15.45 s

We can conclude that the developed application has reduced typing time compared with the non-predictive existing system while the algorithm is worked for different individuals without significance difference in performance. Considering the accuracy and performance of the application, the users can experience reliable, accurate and convenient mechanism for typing SMS messages. The developed algorithm could be an effective algorithm to achieve the goals and objectives of the research.

Some of research areas as to extend to this research are interoperability of the algorithm in various types of mobile phones and identifying other attributes and terminology of Sinhala language model. Since the application development for mobile phones is emerging, it is expected that this research will be useful for future research and development activities on mobile application development and natural language.

REFERENCES

[1] Ali Reza Ebadat, “Using Genetic Algorithm for the Decoder in SMT”

[2] Central Bank of Sri Lanka, “Sri Lanka – Socio Economic Data 2012”, vol. XXXV, June 2012

[3] Dale L. Grover, Martin T. King, Clifford A. Kushler, “Reduced Keyboard Disambting Computer”, Patent No: 5,818, 437, 6.10.1998

[4] Gihan Dias, Aruni Goonetilleke, “Development of Standards for Sinhala Computing”, 1st Regional Conference on ICT and E-Paradigms, 24th – 26th June 2004, Colombo, Sri Lanka

[5] João Luís Garcia Rosa, Mestrado em Sistemas de Computação, Rodovia D. Pedro I, “A Biologically Motivated Connectionist System for Predicting the Next Word in Natural Language Sentences”

[6] Melanie Mitchell, An Introduction to Genetic Algorithms,ISBN 0−262−13316−4 (HB), 1996

[7] Nicola Carmignani, “Predicting Words and Sentences using Statistical Models”, Departement of Computer Science, University of Pisa, 2006

[8] Shahid Akhiguatar, Patricia B. Arinto,” Digital review of Asia Pacific by United Nations Development Programme”, 2009-2010

[9] Sri Lanka Standards Institute, Draft Sri Lanka Standard Sinhala Character Code for Information Interchange, available at http://www.fonts.lk/doc/sls1134.pdf , 2004.

AUTHORS

First Author – M. S. Karunarathne, BSc (Hons), Sabaragamuwa University of Sri Lanka and sajeewani@sab.ac.lk
Second Author – L. D. J. F. Nanayakkara, PhD, University of Kelaniya and Julian@kln.ac.lk
Third Author – Kapila Ponnamperuma, PhD, University of Kelaniya and kapila@kln.ac.lk.

Correspondence Author – M. S. Karunarathne, .sajeewani@sab.ac.lk, sajee_87@yahoo.com, +94772382036