A Sinhala Predictive Text Input System for Android Devices

Randil Pushpananda
Language Technology Research Laboratory
University of Colombo School of Computing
No 35, Reid Avenue, Colombo 07.
rpn@ucsc.lk

Introduction

Nowadays, mobile phone is very important and valuable device for everyone to communicate with others and it is the most popular communication device in Sri Lanka. Short Message Service is now used more and more by most of the people in Sri Lanka and still there is no proper input system to send SMS in Sinhala language.

Objectives

- Develop an effective, user friendly and Unicode support Sinhala predictive text input system for Android devices.
- Design a standard Sinhala onscreen touch keyboard layout for Android devices to represent Unicode Sinhala characters.

Methodology

Predictive input system was developed using an open-source built-in input system named *ScandinavianIME* [1] because it has a special feature of adding external dictionaries to the system.

Application Development

The basic architecture of the predictive input system is shown in figure 1.

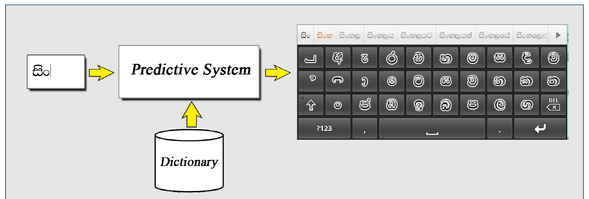

Figure 1: Architecture of the predictive input system

The following code shows the special method which was added to handle the Sinhala keyboard layers.

```
void toggleShiftSinhala() {
   if (Current Keyboard ID = SinhalaShifted KB ID)){
       Enable Non-Shifted Keyboard   }
   else {
       Enable Shifted Keyboard   }
}
```

Edit distance algorithm was used in this system to suggest words correctly when the typing error occurs. The following figure 2 shows how the prediction works when typing the word "ආයුබෝවන්" (*ayubo̱wan*). Prediction starts after the second keystroke.

Figure 2: How the words are predicting while typing

Keyboard Design

Two direct input keyboard layouts were created.

1. By analyzing characters sorted in descending order using the distinct words extracted from *UCSC 10M Sinhala Corpus* (figure 3).

Figure 3: Corpus based keyboard layout with non-shifted and shifted

2. Based on the Wijesekara keyboard layout (figure 4).

Figure 4: Wijesekara based keyboard layout with non-shifted and shifted

These two layouts were created to increase the user friendliness of the system by displaying Sinhala characters instead of English characters.

Dictionary Creation

Since the predictive input method is totally based on the dictionary, distinct word list with frequencies was created UCSC 10M Sinhala Corpus. One letter and two letter words are removed from the list because this prediction algorithm works after entering two characters. Most of the words with frequency less than five are removed from the list. A total 97731 words were remained after cleaning. Word list is sorted in two ways to optimize the word prediction.

- According to the alphabetical order
- According to the frequency in descending order.

Generated words and their frequencies added in to a standard XML format as shown in figure 5.

```
<?xml version="1.0" encoding="UTF-8" ?>
    <wordlist>
        <w f="2303">අංක</w>
        <w f="5">අංකන</w>
        <w f="5">අංකනය</w>
        <w f="401">අංකය</w>
        <w f="31">අංකයක්</w>
        <w f="89">කැමක්</w>
        <w f="26">කැමකට</w>
        <w f="7">කැමකි</w>
        <w f="543">කැමට</w>
        <w f="10">කැමටත්</w>
        <w f="5">කැමටය</w>
        <w f="40">කැමත්</w>
        <w f="7">කැමද</w>
    </wordlist>
```

Figure 5: word list with frequencies

XML is converted in to binary code which is the standard way to access the predictive input method.

Evaluation

Two components of the input system were evaluated separately as described below.

1. Evaluation the effectiveness of the input system

It was calculated by analyzing number of key strokes used to type a given sentence with the dictionary support and without the dictionary support.

Sample of 50 unique Sinhala sentences were selected from the editorial articles of daily newspapers. Table 1 shows sample sentences used.

Sentence	Keystrokes without Dictionary	Keystrokes with Dictionary	No of predicted words
මෙබඳු තත්ත්වයකින් පසුව උදාවන්නේ ආර්ථිකයට වෙහෙවින් හිතකර තත්ත්වයෙකි.	73	46	7/8
නිරිඟු පිටි මෙරටට ආනයනය කරන්නේ සහල්වලට ආදේශකයක් ලෙසිනි.	61	42	8/8
යුනෙස්කෝ අධිපති කෝච්යෝ මන්ත්‍රීරා මහතාට අපගේ ආචාරය හිමිවේ.	70	55	6/8

Table 1: Sample editorial sentences

The following equation used to calculate the improved efficiency.

$$improved\_efficiency = \frac{\sum_{1}^{50}(KWOD - KWD)}{\sum_{1}^{50} KWOD} \times 100\%$$

KWD – Total number of keystrokes used to type sentence with the help of dictionary
KWOD - Total number of keystrokes used to type sentence without the help of dictionary

Improved efficiency of the predictive input system for given sentences is 37.19%.

2. Evaluation of coverage of the keyboard

The main objective of this evaluation is to identify the designed keyboard covers all the characters, modifiers and non-vocalic strokes in Sinhala language. Only ඎ, ඏ, ඐ, ඃ are not possible to type using the created keyboard layouts.

Conclusion

Since this is the first attempt in developing a Sinhala predictive system for Android devices, this can be used as a baseline for further developments.

Reference

[1] (2011, Aug) Scandinavian Keyboard [Online] Available - http://tinyurl.com/3gcjozj