Developing Efficient Text Entry Methods for the Sinhalese Language

Shyam Reyal
University of St Andrews
St Andrews, UK
smr20@st-andrews.ac.uk

Keith Vertanen
Montana Tech
Butte, Montana, USA
kvertanen@mtech.edu

Per Ola Kristensson
University of St Andrews
St Andrews, UK
pok@st-andrews.ac.uk

CHI'13, April 27 – May 2, 2013, Paris, France.
ACM 978-1-XXXX-XXXX-X/XX/XX.

Abstract
Western languages and alphabets dominate the text entry field. This position paper outlines the background, related work and the challenges of implementing an efficient text entry method for the Sinhalese language. Sinhalese falls into the Indo-Aryan family of languages, along with Hindi, Urdu, Sanskrit and other languages.

Author Keywords
Text entry; Sinhalese; Language modeling

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction
Indo-Aryan languages are known for their complex alphabets, syllables, and grammars. For example, the Sinhalese language is spoken, read and written by over twenty-two million people worldwide including almost all the citizens of Sri Lanka. Sinhalese has sixty base characters in the alphabet, with each character having thirteen variations when combined with vowels (Figure 1). The language itself features a greater diversity of rich and complex contextual phrases and idioms than many Western languages. Sinhalese sentences can also be very long: a single sentence can occupy a full paragraph with no punctuation, and without violating any grammar rules.

Figure 1. The Sinhalese characters in the Unicode character tables [1].

Sinhalese was designed to be written with pen or pencil (originally quills or chalk). The language was not designed or optimised for typewriting, word-processing, or to be entered via a computer. The designs of keyboards, numeric keypads, and touchscreens were mostly innovations of the West, invented by the people who had written and communicated using Western languages. Accordingly, text entry methods were designed to target western languages and alphabets (e.g. Morse code was made for the English alphabet). Indo-Aryan languages are based in the Eastern world, mostly in third world countries where technology arrives three, four, or even ten years after it has matured in the West. Therefore, Indo-Aryan languages' text entry methods were not invented to fit the languages, but rather copied from the Western world and naïvely applied with little focus on creating an optimal text entry experience for users.

Related Work

Today, suboptimal methods of text entry are commonly used in countries such as Sri Lanka. One such keyboard-based input method for the Sinhalese language is the Wijesekara keyboard [2] (Figure 2), which simply assigns the Sinhalese base characters and vowel variants onto a standard US keyboard using a different keyboard mapping (i.e. the letter "K" in the keyboard is mapped to the Sinhalese equivalent character of "N"). This causes problems for users who are already familiar with typing in English. Those users expect the Sinhalese "K" to be displayed when the "K" key on the US keyboard is pressed. It is plausible that this mapping inconsistency causes an increase in error rate for users familiar with typing in English.

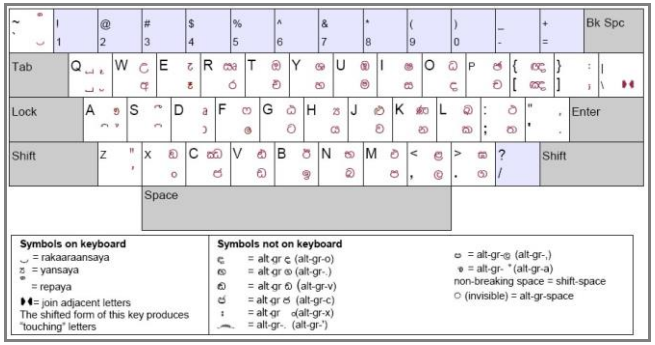

Figure 2. The Wijesekara keyboard layout.

A later approach was the Realtime Singlish Unicode Converter [3]. Singlish, as the name suggests, means Sinhalese through English. In Singlish, a key on the English keyboard is mapped to a Sinhalese character in Unicode by direct transliteration. This is a better way of achieving faster Sinhalese typing on a keyboard for users who are familiar with typing in English, but this too has its limitations, such no word prediction and no error correction. The latter is particularly important, as the transliteration process is highly sensitive to errors. A single incorrect key press can result in a completely different transliteration result.

Work addressing Sinhalese text entry for mobile devices is very minimal. Since there are sixty base characters and thirteen vowels (plus vowel variants), dividing all these keys among a numerical keypad would be difficult. Furthermore, due to the high number of possible characters per key and the high number of possible words for a given key sequence, popular mobile text entry methods, such as disambiguating keypads (e.g. T9), would provide too many ambiguous words to make it effective for Sinhalese text entry.

Milestones in Sinhalese Text Entry

1985 *CINTEC establishes a committee for the use of Sinhala & Tamil in Computer Technology. [4]*

1989 *"WadanTharuwa" (means WordStar in Sinhala) developed by the University of Colombo. [5]*

1994/1996 *Samanala Transliteration Scheme developed by Prasad Dharmasena. [5]*

1997 *Helewadana for Windows developed by Microimage (Pvt) Ltd and Harsha Punasinghe. [5]*

1998 *SLS1134/Unicode standards released by CINTEC for the first time. [5]*

2002 *Siyabasa sinhala typing software is released. [5]*

2002 *Madura English-Sinhala Dictionary is released. [5]*

2004 *Sinhala Text Box developed by Dasith Wijesiriwardena. [5]*

Even on modern large touchscreen devices it would be difficult to fit the entire set of Sinhalese characters on the screen.

Advances in word prediction and language modelling have resulted in optimised word prediction software for many languages, including Western languages, Korean, Chinese and Japanese. However, currently there are no successful predictive text entry methods available for the Sinhalese typist, whether the input method is a keyboard, a numeric keypad or a touchscreen. One reason for the lack of a successful word prediction system for Sinhalese is that very little language modelling work has been done in the area. The absence of in-domain training data for models and test sets means building a word prediction system from scratch is difficult. Furthermore, creating a word prediction system designed for mobile text entry is even more problematic as there are no representative mobile training texts available in Sinhalese. Thus we have a bootstrap problem: we need mobile Sinhalese text to build a good mobile text entry method, but we need a good mobile text entry method to obtain representative mobile training data.

Nonetheless, there does exist a small Sinhalese blogging community. Furthermore, there are news journalists and Wikimedia-publishers that use the suboptimal methods of Sinhalese text entry available to them. The texts that these writers generate can be collected via a web crawler and used as training data for long-span statistical language models. However, a challenge is that due to the non-standard nature and differences in the input methods used by these online authors, creating a high-quality corpora can be difficult as writers tend to type the same words using different and often incorrect grammar. This results in writers injecting noise in the form of grammatical errors into the training corpora.

Research Questions

Given these challenges, the research questions that need to be resolved in order to create efficient text entry methods for Sinhalese include:

- Empirically understanding the limitations and implications of existing methods for Sinhalese text entry.
- Identifying the classes of text entry methods that are suitable to be adapted to support Sinhalese.
- Identifying factors for determining whether the Sinhalese user community would accept a new text entry method, and why. Here, it is important to consider not only Sinhalese text entry users who are already fluent in English text entry systems, but also users who do not know English at all. An example of a potential factor is social acceptability.
- Identifying which text entry methods for Western languages can be directly applied to non-Western languages, and which require changes, and what those changes are.
- Identifying which text entry methods would suit different use-scenarios (e.g. text entry while travelling, text entry while walking etc.)
- Investigating the major differences between text entry methods designed for Western languages and text entry methods designed for Indo-Aryan languages.
- Identifying the most important parameters when devising efficient text entry methods for Indo-Aryan

Milestones in Sinhalese Text Entry (Continued)

2004 *The Iskole potha Unicode fonts released by Microsoft [5]*

2004 *Tusitha Randunuge and Nianjan Meegammana at Kaputa.com released Kaputa Unicode Fonts and Keyboard drivers [5]*

2006 *Sinhala SP for Windows developed by Native Innovation (Pvt) Ltd is a more complete software solution to its predecessor Sinhala Text Box. [5]*

2009 *A novel word-based predictive text input system named SriShell Primo. [6]*

2011 *Sinhala T9 text entry system, Dewapura, M.H. [7]*

2012 *The Language Technology Research Laboratory attempts to build a 10 million word Sinhalese corpus [8]*

languages (e.g. error rate, speed, number of key-presses, keystroke savings, language model characteristics, user feedback mechanisms, error correction interfaces, etc.)

Current Progress and Future Work

This position paper has outlined the research questions we are currently exploring in our effort to devise an efficient touchscreen-based text entry method for Sinhalese. Such a text entry method would benefit a majority of Sri Lankans and the Sinhalese-literate community of the world. We aim to develop a text entry method that would run on existing hardware and input devices, ensuring compatibility and easy adaptability with existing user interfaces on the dominant operating systems, such as Microsoft Windows, Google Android, and Apple iOS.

Current progress includes the design of a web crawler that collects Sinhalese text from online forums, blogs and news websites. We have also downloaded the Sinhalese Wikipedia database. Using this data we have generated a Sinhalese vocabulary and created long-span character and word-based language models. We are currently investigating which domains provide the best performance for mobile-like Sinhalese text. We hope to create statistical language models representing different scenarios for text input (e.g. writing blog entries, writing news articles, writing wiki articles, or simply texting), thereby creating more personalised and efficient text entry methods for users.

Once the language modelling phase has been completed, we will be investigating a series of approaches to devise efficient predictive Sinhalese text entry methods for touchscreen-based mobile devices.

Conclusions

Western languages and alphabets dominate the text entry field. In this position paper we have outlined the background, related work and the challenges of implementing an efficient text entry method for the Sinhalese language.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (grant number EP/H027408/1) and the Scottish Informatics and Computer Science Alliance.

References

[1] *Sinhala Character Code Chart*. http://www.unicode.org/charts/PDF/U0D80.pdf

[2] *Wijesekara Keyboard Layout*. http://www.ucsc.lk/ltrl/services/layout/

[3] *Real Time Unicode Converter*. http://www.ucsc.cmb.ac.lk/ltrl/services/feconverter/

[4] *UNESCO META Survey on the Use of Technologies in Education*. http://www.unescobkk.org/fileadmin/user_upload/ict/Metasurvey/SRILANKA.PDF

[5] *History of Sinhala Software*. http://en.wikipedia.org/wiki/History_of_Sinhala_software

[6] Goonetilleke, S. Hayashi, Y. Itoh, Y. Kishino, F. 2008. SriShell Primo: A Predictive Sinhala Text Input System. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 43–50.

[7] Devapura, M. H. 2011. *Sinhala T9 Text Entry System*. MSc Thesis, University of Moratuwa, Sri Lanka.

[8] *Corpus and Corpus Analysis Tool*. http://www.ucsc.cmb.ac.lk/ltrl/?page=panl10n_p1&lang=en#corpus