International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 10, Issue 5 (May 2014), PP.31-34

Romanized Language Identification and Transliteration System for Security with an Authentication System Using Persuasive Cued Click Points - RLITS

Sibi Jacob[1], Kochumol Abraham[2], Dr K Sridharan[3]
[1]*JKK Munirajah College of Technology, Anna University, Chennai.*
[2]*Marian College, M G University, Kuttikkanam P.O, Kerala.*
[3] *JKK Munirajah College of Technology, Anna University, Chennai.*

Abstract:- Romanized script is popular today for communication in every country, as the script is almost universally enabled in text processors. In countries like India which is a linguistic cauldron, it is very common to see English text in email messages and chat transcripts, with generous sprinkling of words from local languages in roman script. Dubbed as Manglish (Malayalam and English) etc., this roman transliteration of non-English languages contributes to a major noise in analyzing English text in these countries. Our system, RLITS introduces an automated language identification scheme from the Romanized message with SVM regression. It also provides a facility to transmit our message securely using symmetric key cryptography. The overall system is protected by an authentication scheme persuasive cued click point.

Keywords:- Romanization, Transliteration, character n-gram, Persuasive Cued Click Point.

I. INTRODUCTION

In the phonetic Indian languages, typing overhead is compared more to English and hence there is a natural inclination to use roman script to produce so called Hinglish (Hindi and English), Manglish (Malayalam and English), Benglish (Bengali and English) etc. This is very true of young generation non-resident Indians who may be familiar with spoken form of their mother tongue, but not with its script. When this kind of text is to be machine processed, word models of various languages are required to sieve out the underlined words. The task of identifying the language of text or utterances has a number of applications in natural language processing. If we engage human translators, it is very costly, time consuming and only a minimal set of languages can be handled by the human beings. In this paper, we consider the problem of language identification and present a system for the identification and Transliterisation of Romanized Malayalam and secure communication through cryptography. The overall system, RLITS has been protected by the authentication scheme Persuasive Cued Click Point.

II. RELATED WORK

A. Language Identification

Font Llitj´os and Black shows that language identification can improve the accuracy of letter-to-phoneme conversion[1]. Li etal. use language identification in a transliteration system to account for different semantic transliteration rules between languages when the target language is Chinese[2].Huang improves the accuracy of machine transliteration by clustering his training data according to the source language[3]. Konstantopoulos looks particularly at the task of identifying the language of proper nouns. He focuses on a data set of soccer player names coming from 13 possible national languages. He finds that using general n-gram language models yields an average F1 score of only 27%, but training the models specifically to these smaller data gives significantly better results: 50% average F1 score for last names [4]. McNamee presents sufficiently accurate language recognition methods for long texts consisting of tens of sentences [5]. Models based on frequencies of letter combinations are widely used to identify the language of a text. It was noted that, it is possible to use rank methods for language identification of a text, but they are not suitable for short texts. Also it is concluded in that, the language identification problem for short texts segments is still actual, and higher accuracy is achieved at the expense of larger model size and slower processing [6].

B. Transliterisation

Some focuses on back-transliteration [7][8][9].Transliterated words require special focus in Natural Language Processing. There is a growing body of research on automatic extraction of transliterated pairs [10] [11]. Sherif and Kondrak [11] use examples to jointly learn a bilingual string distance function and extract transliterated pairs. The task of identifying transliterated words has been less studied. Oh and Choi [12] studied identification of transliterated foreign words in Korean text, using an HMM on the word syllable structure. They used a corpus of about 1,900 documents in which each syllable was manually tagged as being either Korean or Foreign, and achieved impressive results. However, besides requiring a large amount of human labor, their results are not applicable to Hebrew (or Arabic) as syllable structures of these languages are not clearly marked in writing, and even the vowels are not available in most of the cases. Nwesri et al. [13] dealt with the identification of transliterated foreign words in Arabic text in the setting of an information retrieval system. They tried several approaches: using an Arabic lexicon (everything which is not in the lexicon is considered foreign), relying on the pattern system of Arabic morphology, and two statistical n-gram models, the better of which was based on Cavnar and Trenkle's rank order statistics [14], traditionally used for language identification. For the statistical methods, training was done on manually constructed lists of a few thousands Arabic and foreign words written in Arabic script. They also augmented each of the approaches with hand written heuristic rules. They achieved mediocre results on their training set, somewhat lower results for their test set, and concluded that the best approach for identification of transliterated foreign words is the lexicon-based method enhanced by hand written heuristics, by which they achieved a precision of 68.4% and recall of 71.1% on their training set, and precision of 47.4% and recall of 57.2% on their test set.

III. PROPOSED SYSTEM

A. Introduction

SMS provides a convenient means to people for communication with each other using text messages via mobile devices or Internet connected computers. But there are lots of security issues concerned with the use of SMS. A Short Message Service Centre (SMSC), usually owned and run by a telecommunication operator, is responsible for the routing and delivery of SMS. When an SMS message is delivered to the SMSC, a store-and-forward message mechanism is implemented, whereby the message is temporarily stored, then forwarded to the recipient's phone when the recipient device is available.

It is necessary to monitor the SMS communication for national security because most of the communication between the antisocial activists happens through this. The major bottleneck in sms monitoring is Romanization. Most of the people send messages in Romanized form. So we need an automated self-learning system for the identification and transliteration of Romanized messages. This paper deals with an SMS encryption for mobile communication. The transmission of an SMS in GSM network is not secure; therefore it is desirable to secure SMS by additional encryption. This paper focuses on this area of research and presents a system for the identification and transliteration of Romanized Malayalam and secure communication through cryptography. Authentication of the overall system is being provided using Persuasive Cued Click-Points. This system can be implemented in any SMSC centres.

B. Architecture of RLITS

RLITS can be explained in detail using Fig. 1. The whole system can be divided into two modes of operation.

a. Data collection, Training and Reporting.

b. Automatic Language Identification and detection of menacing words.

1) Data collection, Training and Reporting:

For implementing our system we create a dataset which consists of 500 romanized words and phrases of Malayalam and their English meaning. Training of the dataset is being performed using SVM approach. We choose SVMs because they can take a large number of features and learn to weigh them appropriately.

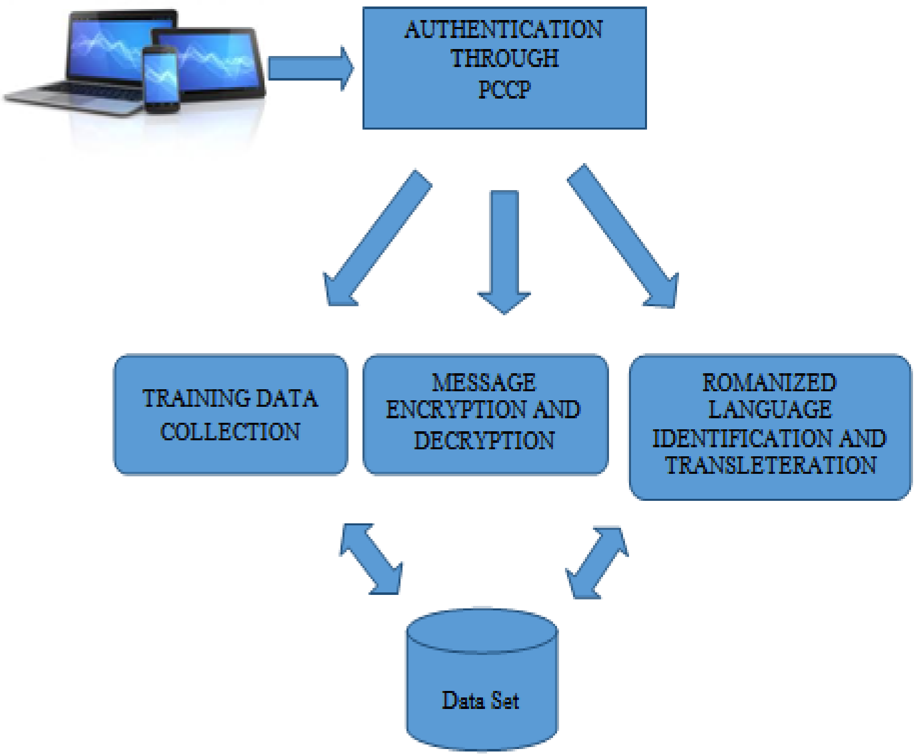
AUTHENTICATION THROUGH PCCP
TRAINING DATA COLLECTION
MESSAGE ENCRYPTION AND DECRYPTION
ROMANIZED LANGUAGE IDENTIFICATION AND TRANSLETERATION
Data Set

Fig. 2: Architecture of RLITS

2) Automatic Language Identification and detection of menacing word.

For communication, every SMS should pass through this system implemented in SMSC centres. Each SMS will be extracted and the language of the same will be identified. Language identification is done using n-gram approach. N-gram approaches have proven very popular for language identification in general. After this process, the meaning of words will be taken from the dataset and if any menacing words have been detected, it will be informed to the respective authority.

C. Authentication using PCCP

Graphical password system is a type of knowledge-based authentication that attempts to leverage the human memory for visual information. Here we have chosen a graphical password system, PCCP which support users in selecting passwords of higher security. Persuasion is being used to influence user choice in click-based graphical passwords which encourages users to select more random, and hence more difficult to guess, click-points [15].

IV. FUTURE ENHANCEMENTS

This paper focuses only on Malayalam. It can be extended to other languages too. Also, as this system has to be integrated with the SMSC centre, the challenges which can occur during implementation are to be identified.

V. CONCLUSION

We have proposed a new system, RLITS through which we can detect romanized menacing content from SMS. RLITS is being protected by an authentication scheme known as PCCP. If Malayalam is used by the sender, then it is very difficult to be identified as Malayalam is one of the toughest languages in the world. Most of the SMS contain Romanized text and hence it is very difficult to understand the meaning without a language identifier. Our system identifies romanized Malayalam. After identifying the language as Malayalam, the system transliterates it to the corresponding language font, detects the menacing content and reports it to the respective authority.

REFERENCES

[1]. Font Llitj´os and A. W. Black. 2001. Knowledge of language origin improves pronunciation accuracy of proper names. In Proc. of Euro speech, pages 1919–1922.

[2]. Li, K. C. Sim, J.-S. Kuo, and M. Dong. 2007. Semantic transliteration of personal names. In Proc. of ACL , pages 120–127.

[3]. F. Huang. 2005. Cluster-specific named entity transliteration. In Proc. of HLT-EMNLP, pages 435–442.

[4]. S. Konstantopoulos. 2007. What's in a name? In Proc. of RANLP Computational Phonology Workshop.

[5]. McNamee, B.P.: Language identification: a solved problem suitable for undergraduate in-structionl Journal of Computing Sciences in Colleges, 20(3), P. 94–101 (2005)

[6]. Cavnar, W. B., Trenkle, J. M.: N-gram-based text categorization. In.: Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, P. 161–175 (1994) .

[7]. Stalls and K. Night: Translating Names and Technical Terms in Arabic Text. In: Proc. of the COLING/ACL Workshop on Comp. Approaches to Semitic Languages. (1998)

[8]. Y. Al-Onaizan and K. Knight: Machine Transliteration of Names in Arabic Text. In: Proc. of ACL Workshop on Comp. Approaches to Semitic Languages. (2002)

[9]. Yoon, S.Y., Kim, K.Y., Sproat, R.: Multilingual transliteration using feature based phonetic method. In: Proc. of ACL. (2007).

[10]. Klementiev, A., Roth, D.: Named entity transliteration and discovery from multilingual comparable corpora. In: Proc. of NAACL. (2006)

[11]. Sherif, T., Kondrak, G.: Bootstrapping a stochastic transducer for arabic-english transliteration extraction. In: Proc. of ACL. (2007)

[12]. Oh, J., Choi, K.: A statistical model for automatic extraction of Korean transliterated foreign words. Int. J. of Computer Proc. of Oriental Languages 16 (2003)

[13]. Nwesri, A.F., Tahaghoghi, S., Scholer, F.: Capturing out-of-vocabulary words in arabic text. In: Proc. of EMNLP2006. (2006)

[14]. Cavnar, W.B., Trenkle, J.M.: N-gram-based text categorization. In: Proc. of SDAIR-94. (1994)

[15]. Sonia Chiasson, Member, IEEE, Elizabeth Stobert, Student Member, IEEE, Alain Forget, Robert Biddle, Member, IEEE, and Paul C. van Oorschot, Member, IEEE, Persuasive Cued Click-Points: Design, Implementation, and Evaluation of a Knowledge-Based Authentication Mechanism