Towards an Intelligent Multilingual Keyboard System

Tanapong Potipiti, Virach Sornlertlamvanich, Kanokwut Thanadkran
Information Research and Development Division,
National Electronics and Computer Technology Center(NECTEC), Thailand

Introduction

- Two annoyances for Thais to type Thai-English bilingual texts.
  - To switch between languages by using a switching key.
  - To employ the shift key to type half of Thai characters.
    (Because there are more than 100 characters in Thai, to input about half of all Thai characters, a user has to use combinations of 2 keys —the shift key and a character key— to input them.)
- Other multilingual users face the same problems.
- We have proposed a practical solution to solve these problems. Through our system, a user can type Thai-English bilingual texts without using the shift and switching keys.
- Our approach is general and applicable to other multilingual keyboard systems.

The Thai-English keyboard system

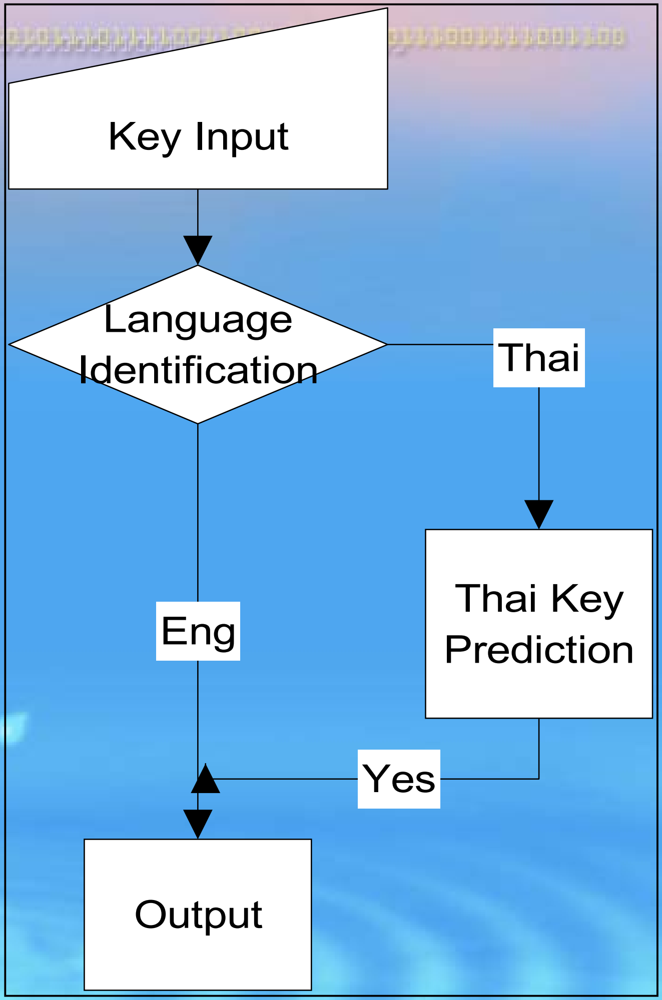
- Thai-English keyboards employ the language switching key and shift key to help typing. For example, in the Thai-English keyboard the ‘a’-key button can represent 4 different characters in different modes as shown below.

English mode without shift	English mode with shift	Thai mode without shift	Thai mode with shift
‘a’ (lowercase a)	‘A’ (uppercase a)	‘ฟ’ (for-fun)	‘ฤ’ (lor-lur)

Overview

There are two main processes in our system.

1) Automatic language identification
2) Key prediction without using the shift key in Thai



Thai-English language identification

$$Tprob = \prod_{i=1}^{m-1} p_T(K_i K_{i+1})$$

$$Eprob = \prod_{i=1}^{m-1} p_E(K_i K_{i+1})$$

p_E is the normalized probability of the bi-gram key buttons considered in English.

p_T is the normalized probability of the bi-gram key buttons considered in Thai.
K is the key button considered.

Tprob is the probability of the considered key-button sequence to be Thai.
Eprob is the probability of the considered key-button sequence to be English.

Case	Language to be identified
Eprob>Tprob	English
Tprob>Eprob	Thai

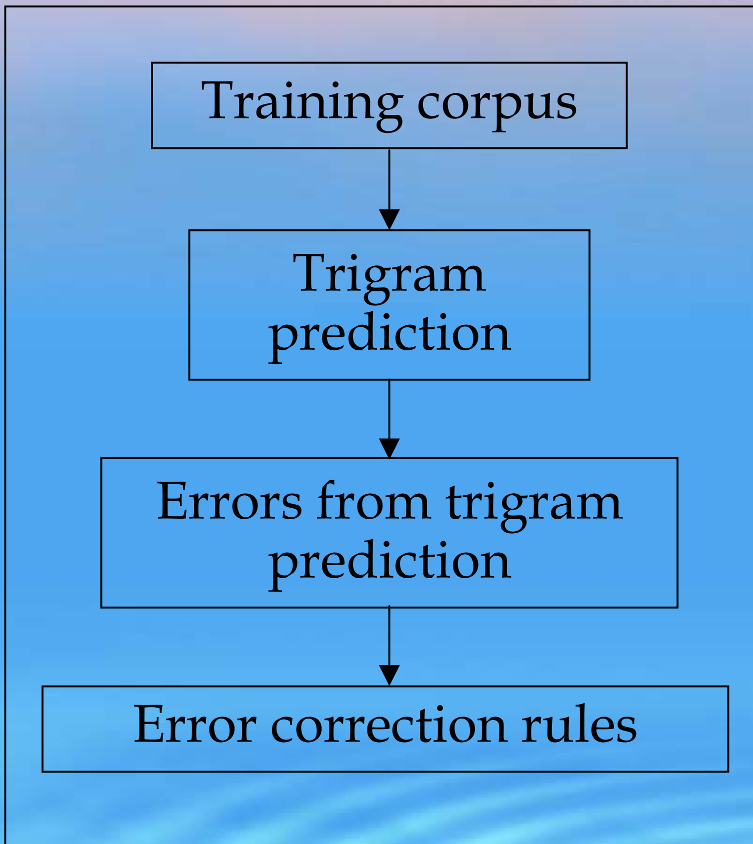
Thai-key prediction employing trigram

$$\tau = \max_{c_1, c_2.., c_n} \arg \prod_{i=1}^{n} p(c_i \mid c_{i-1}, c_{i-2}).p(K_i \mid c_i)$$

τ is the sequence of characters that maximizes the character string sequence probability,
c is the input characters possible for the key button ***K***,
K is the key button consider.

The character sequence that maximizes the probability will be selected.

Error-correction rules



After the trigram model was applied, there are some character sequences that often generate errors. These patterns are collected for error-correction process.

For example, if the key sequence pattern "*asdfk*" always generates wrong prediction, this pattern and the correct predicted pattern corresponding with it will be collected for error correction.

To reduce and collect patterns with most appropriate length, mutual information is employed.

Pattern shortening

- To collect the correction patterns with optimal lengths the following rules are applied.
  - -Initially, patterns are collected 7-character length.
  - -If $Rm(x\mathbf{y}z)$ is less than 1.2, the pattern $x\mathbf{y}z$ is reduced to $x\mathbf{y}$.
  - -If $Lm(x\mathbf{y}z)$ is less than 1.2, the pattern $x\mathbf{y}z$ is reduced to $\mathbf{y}z$.
  - -The 2 rules above are applied recursively until no pattern can be shortened.
- $Lm(.)$ and $Rm(.)$ are defined as:

$$Lm(x\mathbf{y}z) = \frac{p(x\mathbf{y}z)}{p(x)p(\mathbf{y}z)}, \qquad Rm(x\mathbf{y}z) = \frac{p(x\mathbf{y}z)}{p(x\mathbf{y})p(z)}$$

where $x\mathbf{y}z$ is the pattern being considered,
x is the leftmost character of $x\mathbf{y}z$,
$\mathbf{y}$ is the middle substring of $x\mathbf{y}z$,
z is the rightmost character of $x\mathbf{y}z$,
$p(\,)$ is the probability function.

Experimental results

Thai-English language identification accuracy
99.10 %

	Thai-key prediction accuracy	
	Training corpus	Test corpus
Trigram	93.11 %	92.21 %
Trigram + error correction	99.53 %	99.42 %

Conclusion

- We have applied the trigram model and error-correction rules for intelligent Thai key prediction and English-Thai language identification.
- The experiment reports 99 percent in accuracy, which is very impressive.
- Hopefully, this technique is applicable to other Asian languages and multilingual systems.
- Our future work is to apply the algorithm to mobile phones, handheld devices and multilingual input systems.