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.

<table>
<thead>
<tr>
<th></th>
<th>English mode without shift</th>
<th>English mode with shift</th>
<th>Thai mode without shift</th>
<th>Thai mode with shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘a’ (lowercase a)</td>
<td>‘A’ (uppercase a)</td>
<td>‘ฟ’ (for-fun)</td>
<td>‘ฎ’ (lor-lur)</td>
<td></td>
</tr>
</tbody>
</table>
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**

\[
Tprob = \prod_{i=1}^{m-1} \rho_T(K_iK_{i+1})
\]

\[
Eprob = \prod_{i=1}^{m-1} \rho_E(K_iK_{i+1})
\]

\(\rho_E\) is the normalized probability of the bigram key buttons considered in English.

\(\rho_T\) is the normalized probability of the bigram 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.

<table>
<thead>
<tr>
<th>Case</th>
<th>Language to be identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Eprob&gt;Tprob)</td>
<td>English</td>
</tr>
<tr>
<td>(Tprob&gt;Eprob)</td>
<td>Thai</td>
</tr>
</tbody>
</table>
Thai-key prediction employing trigram

\[ \tau = \max \arg \prod_{i=1}^{n} p(c_{i} \mid c_{i-1}, c_{i-2}) \cdot p(K_{i} \mid c_{i}) \]

\( \tau \) 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.
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 “asdkf” 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 $Rm(xyz)$ is less than 1.2, the pattern $xyz$ is reduced to $xy$.
  - If $Lm(xyz)$ is less than 1.2, the pattern $xyz$ is reduced to $yz$.
  - The 2 rules above are applied recursively until no pattern can be shortened.

- $Lm(.)$ and $Rm(.)$ are defined as:

$$Lm(xyz) = \frac{p(xyz)}{p(x)p(yz)}, \quad Rm(xyz) = \frac{p(xyz)}{p(xy)p(z)}$$

where $xyz$ is the pattern being considered,
$x$ is the leftmost character of $xyz$,
$y$ is the middle substring of $xyz$,
$z$ is the rightmost character of $xyz$,
$p(\cdot)$ is the probability function.
## Experimental results

<table>
<thead>
<tr>
<th>Thai-English language identification accuracy</th>
<th>99.10 %</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Thai-key prediction accuracy</th>
<th>Training corpus</th>
<th>Test corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigram</td>
<td>93.11 %</td>
<td>92.21 %</td>
</tr>
<tr>
<td>Trigram + error correction</td>
<td>99.53 %</td>
<td>99.42 %</td>
</tr>
</tbody>
</table>
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.