Input Methods for Chinese/Japanese/Korean Word Processing

Presented by Yue Zhu

July 17th 2018
CJK Characters

• CJK is a collective term for the Chinese, Japanese, and Korean languages, all of which include Chinese characters and derivatives in their writing systems.

• Chinese (Han) Characters are called
  • Hanzi in Chinese
  • Kanji in Japanese
  • Hanja in Korean

• CJK characters include hànzì in Chinese, kanji, kana in Japanese, hanja, hangul in Korean.
Issues for CJK Characters input
Issues for CJK Characters input

• Think of typing English text...

• What about CJK Characters?
  • Chinese: 106,330 Chinese characters are recorded in variant character table (异体字表) in which 3,500 are commonly used.
  • Japanese: 50,305 Chinese characters are recorded in The Great Chinese–Japanese Dictionary (大漢和辞典) in which 2,136 are commonly used.
  • Korean: 1,800 Chinese characters are recorded in Basic Hanja for Educational Use (대한민국 중고등학교 기초한자 목록)
Issues for CJK Characters input

• Input each character directly using a specifically designed keyboard?
  • Large keyboards that hold thousands of individual keys exist, but...
    • Special training is required
    • Difficult to use
Solutions: input methods
Outline

• General categories for input methods
• Input stages
• Input techniques
  • Input by reading
  • Input by structure
  • Input by multiple criteria
  • Input by encoding
• User interface
## General categories for input methods

<table>
<thead>
<tr>
<th>Category</th>
<th>What are they</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct methods</td>
<td>Employ a unique value for the target character</td>
<td>Unambiguous, only one candidate if the type of the encoding is fixed</td>
<td>Hard to use, Not intuitive</td>
</tr>
<tr>
<td>Indirect methods</td>
<td>Obtain the encoded value of the target character or characters usually by typing out the reading or shape on a standard or specialized keyboard</td>
<td>Easy to use, Intuitive, Take advantage of the readings or the structure of the characters</td>
<td>Might have long candidate lists which lower down the input efficiency</td>
</tr>
</tbody>
</table>
Input stages

• Stage 1:
  • The user types raw keyboard input (Transliteration, Native-script input)
  • The computer interprets the raw keyboard input by using the input method and the conversion dictionary to display a list of candidate characters

• Stage 2:
  • The user selects one choice from the list of candidate characters, or requests more choices.
Input techniques

- Input by reading
- Input by structure
- Input by multiple criteria
- Input by encoding
Input techniques – by reading

• By reading (pronunciation)
• Most intuitive and frequently used
• Easiest to learn
• May have long candidate lists
• Three basic units by which input readings can be converted into ideographs:
  • Single ideograph
  • Ideograph Compound
  • Ideograph phrase
### Input techniques – by reading

- **Example:** Input by reading -- three basic units of input
- **Text:** “漢字は” in Japanese which means “The ideograph” in English

<table>
<thead>
<tr>
<th>Units of input</th>
<th>Characters</th>
<th>Latin input</th>
<th>Kana input</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single ideograph</td>
<td>漢</td>
<td>kan</td>
<td>かん</td>
<td>漢感間菅缶艦管館甘完勘巻官環関神鑑韓冠康観棺燗寛...</td>
</tr>
<tr>
<td></td>
<td>字</td>
<td>ji</td>
<td>じ</td>
<td>字寺時自二痔路慈持辞事寰耳磁地姬師辭爾尔児次...</td>
</tr>
<tr>
<td></td>
<td>は</td>
<td>ha</td>
<td>は</td>
<td>n/a</td>
</tr>
<tr>
<td>Ideograph compound</td>
<td>漢字</td>
<td>kanji</td>
<td>かんじ</td>
<td>漢字 感じ 幹事 監事 完司</td>
</tr>
<tr>
<td></td>
<td>は</td>
<td>ha</td>
<td>は</td>
<td>n/a</td>
</tr>
<tr>
<td>Ideograph phase</td>
<td>漢字は</td>
<td>kanjiha</td>
<td>かんじは</td>
<td>漢字は 感じは 幹事は 監事は 完司は</td>
</tr>
</tbody>
</table>

- **It is much more efficient to input ideographs as they appear in compounds or even phrases.**
Input techniques – by reading

• Example: Input by reading -- three basic units of input
• Text: “汉字” in Chinese Pinyin simplified

<table>
<thead>
<tr>
<th>Units of input</th>
<th>Characters</th>
<th>Input</th>
<th>Candidates</th>
</tr>
</thead>
</table>
| Single ideograph        | 汉          | han   | 韩喊含汖寒涵翰横旱悍憨撼摴毹gboolean碣寔椃...
|                         | 字          | zi    | 字字自紫姿滋姿籽梓渍渍孜姿滋滓...
| Ideograph compound      | 汉字        | hanz   | 汉字 汉子 汗渍 憨子 荨子 賢子
Input techniques – by reading

- Example: Input by reading – indication of tones
- Chinese is the only language among CJK that has an indication of tones.
- Text: “汉” in Chinese Pinyin simplified

<table>
<thead>
<tr>
<th>Tone</th>
<th>Reading</th>
<th>Total # of Candidates</th>
<th>Candidates</th>
</tr>
</thead>
</table>
| No tone       | han     | 169                   | 韩喊含汗寒涵翰汉函旱憾憨晗罕瀚焊悍撼鼾...
| First tone    | hān     | 14                    | 愁頃軒魃蚶倶甜窳翕倏...
| Second tone   | hán     | 53                    | 寒軒邗邯函椮.HandlerFunc敲...
| Third tone    | hǎn     | 23                    | 喊罕浮DateString...
| Forth tone    | hàn     | 83                    | 汉閬扌軒斥早軒閑軒悍捍...

Input techniques – by structure

• Why it is needed?
  • Input by reading fails to locate the desired ideographs
  • Not fast enough

• By structure
  • By indexing radical
  • By number of strokes
  • By stroke shapes
  • By corner
  • By other structures
Input techniques – by structure

• Ideographs are composed of radicals or radical-like elements
Input techniques by structure, by indexing radical

• Historically, the ideographs arranged in a character set by radical (e.g. GB 2312-80 Level2 ordered by radical followed by the number of residual strokes)

• Currently, allow users to input by indexing radical even though the characters are ordered differently.
Input techniques by structure, by indexing radical

- Example: Chinese (mainland) input by indexing radical
- Text: “汉字” in simplified Chinese

<table>
<thead>
<tr>
<th>Characters</th>
<th>Indexing radical</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>汉</td>
<td>氵</td>
<td>汀江汝污</td>
</tr>
<tr>
<td></td>
<td>汀江汝污</td>
<td>法治</td>
</tr>
<tr>
<td></td>
<td>汀江汝污</td>
<td>洪润</td>
</tr>
<tr>
<td>字</td>
<td>子</td>
<td>字字字</td>
</tr>
<tr>
<td></td>
<td>子</td>
<td>其他字</td>
</tr>
</tbody>
</table>
Input techniques by structure, by indexing radical

• Example: Chinese (mainland) input by indexing radical
• Keyboard array (the first stroke + the number of strokes of the radical)

<table>
<thead>
<tr>
<th>Q</th>
<th>W</th>
<th>E</th>
<th>R</th>
<th>T</th>
<th>Y</th>
<th>U</th>
<th>I</th>
<th>O</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\ >= 5 | 
\ | 4 | 
\ | 3 | 
\ | 2 | 
\ | 1 | 
\ | 1 | 
\ | 2 | 
\ | 3 | 
\ | 4 | 
\ | >=5 |
| 穴立广衣方火门  | 
\ | 4 | 
\ | 3 | 
\ | 2 | 
\ | 1 | 
\ | 1 | 
\ | 2 | 
\ | 3 | 
\ | 3 | 
\ | 4 | 
\ | >=5 |
| 穴立广衣方火门  | 
\ | 4 | 
\ | 3 | 
\ | 2 | 
\ | 1 | 
\ | 1 | 
\ | 2 | 
\ | 3 | 
\ | 3 | 
\ | 4 | 
\ | >=5 |

Pinyin input  Page up  Page down
Input techniques by structure, by number of strokes

- Most of the characters have unique number of strokes
- Simple and straight forward
- Prior knowledge of readings and radical structures are not required
- Good for input characters which have large / small number of strokes
  - Stroke number = 1 -> 3 characters;
  - Stroke number = 31 -> 3 characters;
- Popular stroke numbers may result in extremely long candidate lists
  - Stroke number = 6 -> 217 characters;
Input techniques
by structure, by stroke shapes

• The “Five-stroke Input Method (Wubi Method)”, developed by Yongmin Wang in 1983
  • Mainly used in simplified Chinese.
  • In the past, convenient and commonly used
  • Short candidate list
  • Special training is required
Input techniques
by structure, by stroke shapes
Input techniques by structure, by stroke shapes

• Case1: Main root
  • Wubi code: keystroke * 4

• Case2: Associated root
  • Wubi code: keystroke + first stroke + second stroke + last stroke

• Case3: Other cases
  • Contain more than 3 roots
    • Wubi code: first root + second root + third root + last root
  • Contain less than or equal to 3 roots
    • Wubi code: all roots + id code (last stroke[region #] + character shape[index #])
Input techniques by structure, by corner

• The four corner code
  • Used for locating characters in ideograph dictionaries.

• The three corner code
  • Developed by liren Hu, Yuanwei Zhang, Kedong Huang
  • Derived from the four corner code
  • Designed specifically for character input
  • Used for Chinese input in early days
Input techniques by structure, by corner

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>头 (Head)</td>
<td>Lid</td>
</tr>
<tr>
<td>1</td>
<td>横 (Horizontal)</td>
<td>Horizontal bar</td>
</tr>
<tr>
<td>2</td>
<td>垂 (Vertical)</td>
<td>Vertical bar</td>
</tr>
<tr>
<td>3</td>
<td>点 (Dot)</td>
<td>Dot</td>
</tr>
<tr>
<td>4</td>
<td>叉 (Cross)</td>
<td>Cross</td>
</tr>
<tr>
<td>5</td>
<td>插 (Skewer)</td>
<td>Skewer</td>
</tr>
<tr>
<td>6</td>
<td>方 (Box)</td>
<td>Box</td>
</tr>
<tr>
<td>7</td>
<td>角 (Angle)</td>
<td>Angle</td>
</tr>
<tr>
<td>8</td>
<td>八 (Eight)</td>
<td>Eight</td>
</tr>
<tr>
<td>9</td>
<td>小 (Small)</td>
<td>Small</td>
</tr>
</tbody>
</table>

Example: Four corner code
Order follows “Z” shape:
- upper-left
- upper-right
- lower-left
- lower-right
(extra above bottom right)

Example: Three corner code
Order from left to right, up to down

Example:
Input techniques by structure, by other structures

• Cangjie Method
  • Developed by Bangfu Zhu in 1976
  • Commonly used in Hong Kong and Taiwan
  • Short candidate lists
  • Complex rules

• Zheng Code Method
  • Developed by Yili Zheng, Long Zheng in 1989
  • Similar to Wubi Method but has different stroke coding
Input techniques by structure, by other structures

- Text: “漢字” in traditional Chinese by Cangjie input method

<table>
<thead>
<tr>
<th>Characters</th>
<th>Cangjie Code</th>
<th>Cangjie Code --- QWERTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>漢</td>
<td>水+廿+中+人</td>
<td>ETLO</td>
</tr>
<tr>
<td>字</td>
<td>十+弓+木</td>
<td>JND</td>
</tr>
</tbody>
</table>

![Cangjie input method diagram](image)
Input techniques by structure, by other structures

• Text: “汉字” in simplified Chinese by Zheng Code input method
Input techniques – by multiple criteria

• Useful for users to narrow down a search

• Input by structure and reading
  • Reduce the number of collisions by combining reading and structure attributes
  • Tze-loi Method, developed by Tze-loi Yeung

<table>
<thead>
<tr>
<th>Character</th>
<th>Tze-loi Code</th>
<th>Reading</th>
<th>Tze-loi Code (QWERTY)</th>
</tr>
</thead>
<tbody>
<tr>
<td>法</td>
<td>丶+△</td>
<td>FA</td>
<td>6ZF</td>
</tr>
</tbody>
</table>

• Renzhi Code Method, developed by Kekang He in 1994
Input techniques – by encoding

- Direct input
- Unambiguous input is possible
- Make use of the encoded value of the target character
- When to use?
  - Past: Telex Code for hanzi interchange
  - Present: Input uncommon words
- Example: Korean input by encoding

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>漢</td>
<td>89-51</td>
<td>F9D3</td>
<td>6F22</td>
<td>7953</td>
<td>F7D3</td>
</tr>
<tr>
<td>字</td>
<td>77-14</td>
<td>EDAE</td>
<td>5B57</td>
<td>6D2E</td>
<td>F1AE</td>
</tr>
</tbody>
</table>
User interface

• Consider the following scenario...
  • The input takes place in a separate window
  • After the input completes, the text is then sent to the application
  • Finally the text is inserted into the current cursor position

• Obviously this is not satisfactory...
  • The user’s eyes must travel back and forth between the current cursor position and the input method’s own UI
User interface

• Hence, inline conversion is desired.

• Input methods are separate processes from the application which accepts the inserted text.

• How can we provide inline conversion at run time?
User interface

• The facilities for inline conversion
  • provided by the OS (Mac OSX, Windows)
  • used by input methods and applications.

• Inline conversion support is very common these days.
Questions
Thanks for listening!