The best way to process Unicode input is to make somebody else do it

October 22, 2007

Raymond Chen

Andrew M asks via the Suggestion Box:

> I was hoping you could address how to properly code Unicode character input. It seems like a lot of applications don't support it correctly.

I'm not sure I understand the question, but the answer is pretty easy: Don't do it! Text input is hard. It should be left to the professionals. This means you should use controls such as the standard edit control and the rich edit control. Properly converting keystrokes to characters involves not just the shift state, but the management of various input method editors, some of which are quite complicated. For example, the IME Pad lets the user draw a Chinese character with the mouse (or if you're lucky, the stylus), and then it will take the result and try to figure out which character you were trying to write and generate the appropriate Unicode character. Other IMEs will generate provisional conversions of phonetic text into Unicode characters, and as more input is received, they can go back and revise their previous guesses based on subsequent input. You definitely don't want to get involved in this. Just leave it to the professionals.

Postscript: For those who have never used a phonetic IME, here's how a hypothetical English phonetic IME might work. Let's pretend there's an English phonetic keyboard with keys labeled with various phonemes. (Instead of IPA characters, I will use traditional American phonetics.)

You type	Result
ə	Uh
t	Ut
ĕ	A te
n	A 10

sh	A 10 sh
ə	A 10 sha
n	A tension
o	A tension o
l	Attention all

Notice how the IME keeps updating its guess as to what you're trying to type as better information becomes available. The text is underlined since it is all provisional. During the input, you can hit the left-arrow to go back to any part of the provisional text, hit the down-arrow, and see a list of alternatives, at which point you can override the guess with the correct answer. For example, if you really wanted to write "A tension all", you would arrow back to the word "Attention", hit the down-arrow, and select "A tension" from the menu. Eventually, you reach the end of a phrase or sentence, look over the provisional text, and after making any necessary corrections, you hit Enter, at which point the text is committed into the edit control and a new string of provisional text begins.

Raymond Chen

Follow