Jawi encoding and rendering

Mohd Zamri Murah
Pattern Recognition Group
Center for Artificial Intelligence Technology
Fakulti Teknologi Dan Sains Maklumat
Universiti Kebangsaan Malaysia
zamri@ftsm.ukm.my

Abstract

This paper review encoding and rendering issues related to jawi script. The jawi script is based on arabic script, with six additional characters.

1. Introduction

Using jawi in computing involves three critical issues; input, encoding and rendering. Input refers to keyboard layout and fonts for writing, encoding refers to a set of codes that represent each characters and rendering refers to how the characters is display on screen or prints on papers.

There are currently no standard jawi keyboard layout. There are , however, many unofficial jawi keyboard layouts from various parties. There are currently a few standard fonts that fully support jawi such as Times New Roman dan Arial.

In the 1980s, there exist many encoding for arabic characters such as ISO-8859-6 and cp-1256. This make it difficult to share documents in computing. A document in ISO-8859-6 encoding would render garbage if the user use a different encoding to open it.

This problem is solve using a standard encoding that is accepted by all computer manufactures and software developers. As of now, the Unicode encoding for all the world character sets has been accepted as the standard in computing world as a mean for communication and storage. The Unicode version 5.1 is the same as ISO-10645 standard. Both the standards use the same characters encoding for all world characters.

The six additional jawi characters چ ڠ ݢ ۏ ڤ ڽ were not included in the initial ISO-8859-6 standard in 1999. This fact make it difficult to use jawi in computing.

Unicode began 1991 as an effort to have a standard for all world characters encoding for computing purposes. In Unicode, the character cheh چ was included in 2001. The characters vi ۏ , nga ڠ, nya ڽ, peh ڤ were included in 2003. Finally, the character ga ݢ was included in 2005 for unicode 4.1. Thus, we should realise that the complete jawi characters encoding was complete in 2005 based on Unicode 5.1 .

2. Encoding

Character encoding refer to a unique number assign for each character. For example, the code for the character ا is `U+0627` in Unicode. This code would be use in storage and transmission of document containing the character ا.

How the the character looks like on the screen and on the papers is rendering issue with involves two items; rendering engine and font technology. As long as the font use ISO-8859-6 or Unicode mapping, the user will see the letter ا. However, not all fonts contain the glyphs for arabic script or jawi script.

3. Rendering

Rendering arabic characters requires contextual mapping, since arabic characters changes shapes based on its position in a word. This is the task for rendering engine, to determine the appropriate shapes of arabic characters based on positional values.The rendering engine also need to deal with BIDI (bi-direction) of the words since arabic is written left-to-right.

For instance , a user type the letters ج ا و ي. In computer, it would see a string of code `0676 0878 0978 0982`. It is the responsibility of the rendering engine to render these sequence of letter correctly as جاوي.

There are currently four rendering technology;

4. **Font Technology**

5. **Input - keyboard layout**