

CJK UNICODE LANGUAGE IDENTIFICATION METHOD

BACKGROUND OF THE INVENTION

5 1. Field of the Invention

The present invention is related to detecting possible language ("Simplified Chinese, Traditional Chinese, Japanese and Korean") based on statistics of Chinese character range and other language
10 specific range in CJK Unicode range.

2. Description of the Related Background

Unicode is a standard method for representing written language in
15 computers and designed to provide unique number for each character on any platform, and program. It includes almost all language characters on the earth that people are using. The standard character encodings, an enumeration of character properties such as upper and lower case, a set of reference data computer files, and a number of related items, such as character properties, rules for
20 normalization, decomposition, collation, rendering and bidirectional display order (for the correct display of text containing both right-to-left scripts, such as Arabic or Hebrew, and left-to-right scripts).

Unicode uses 16-bit word to encode characters, allowing up to 65,536
25 characters to be encoded without re-sorting to more complicated schemes involving multiple machine words per character.

Now, Unicode characters are quite common in many OS(operating system) and facilities. A lot has been done for the Unicode language detection.
30 Regarding this, "Identifying language and character set of data representing text" US 6157905, Microsoft Corporation, 12/11/1997 may be referred to. Unicode Detection is towards all countries languages Detection. Depending on the Unicode character range, we can recognize the language which the characters belong to. With Unicode character value, most languages can be recognized
35 because each language occupies a certain range in the Unicode table. There is one exception that is CJK languages. As Chinese character range doesn't indicate in

Unicode character data, for the Chinese character, it is unable to correctly render out the language-dependent character as shown in Figure 3.

SUMMARY OF THE INVENTION

5

CJK language Unicode is introduced from the above description. CJK is the acronym for Chinese, Japanese, and Korean and Chinese has two formats: Simplified Chinese and Traditional Chinese. So CJK include four different written language styles. These language styles origin and inherit from the same language, which is Ancient Chinese. Now, these language codes are using a lot of cross characters in writing glyphs. In Unicode table, even though there is very little difference in writing style for the same character among CJK languages, Unicode still uses the same U+16 value for them. Thus, it is hard to distinguish the language only by these characters U+16 value.

15

As shown in Figure 2, a Unicode string is displayed with its U+16 values and corresponding languages. All characters are available in GB, BIG5 and SJIS and most characters can be shown in Korean. In this case, the writing glyphs of these characters in different language fonts are displayed with a little difference as in Figure 3.

20

Then, this detection method deals with the matter when the unknown language characters are all CJK Unicode and can not be distinguished the language belongings from their nature value under the above situation.

At first, the following CJK language character-occupation table under Unicode is studied. As can be seen from the table, CJK Unicode set is divided into several segments. Each segment is described as one set of CJK languages.

25

CJK language	Range		
Hangul Jamo	0x 1100	--	0x 11FF
Kangxi Radicals	0x 2F00	--	0x 2FDF
CJK Radicals Supplement	0x 2E80	--	0x 2EFF
Hiragana	0x 3040	--	0x 309F
Katakana	0x 30A0	--	0x 30FF
Hangul Compatibility	0x 3130	--	0x 318F
CJK Stokes	0x 31C0	--	0x 31EF

Katakana Phonetic Extension	0x 31F0	--	0x 31FF
CJK Unified Ideographs Extension A	0x 3400	--	0x 4DBF
CJK Unified Ideographs	0x 4E00	--	0x 9FCF
Hangul Syllables	0x AC00	--	0x D7AF
CJK Compatibility Ideographs	0x F900	--	0x FAFB
Halfwidth and Fullwidth	0x FF00	--	0x FFEF

The first several segments, such as Hangul, Hiragana, and Katakana, can be easily recognized due to the separate range without mixing with other languages. But in CJK Unified Ideographs segment, language characters are mixed together.

5 In this case, some references for the most often used characters for each language were found. Microsoft Code Pages (CP) gives the normal Native-Code characters used by each language(CP932-GB, CP936-SJIS, CP949-KEUC and CP950-BIG5). From the Code Pages, each language's Native-Unicode conversion table on U+16 values can be achieved. With these tables, recognizing the Unicode
10 range in each language is possible. Then CJK language Unicode character table is separated and each table covers part of the CJK region, which can be seen from Figure 5.

It can be clearly seen from the Figure 5 that there are total 12 zones in CJK Unicode languages. Each language has their own zone and also has cross
15 zones between any 2 languages. Each zone is marked with special language record. E.g., Zone 4 is the cross zone language for GB, Japanese and Korean. The character belonging zone is clearly defined as in Figure 5. This structure of the record data table consists as shown in Figure 7.

According to that above ideal design, a 1-1 mapping table with each
20 Unicode CJK character and its belonging zone U+16 value is made as shown in Figure 6. If a rendering character's attribute needs to be known, the table can be looked through with this exact character value. With this 1-1 table, the CJK language detection issue can be solved. E.g., a character '載' (0x8F09) can be used for GB, BIG5, SJIS and KEUC, then it is in Zone 2, its corresponding Bit
25 Value is '1111', and U+16 value is 'F'. While character '甄' (0x750E) can be only used for GB and SJIS, it is in Zone 5, its corresponding Bit Value is '1010', and UTF16 value is 'A'. For any string, the language can be estimated by the calculation of possibility of CJK languages.

In the detection mode, a parameter with detection priority is adopted. For

each language with first priority, the detection order is different and the first priority language will be chosen when this language and other language have same largest possibility.

5 Additionally, GB has a large number of characters and occupies most areas of the table. Thus, we found a table that shows traditional Chinese to Simplified Chinese transfer made by China National language character committee on 10 Oct,1986. Regarding this, homepage whose URL is “<http://www.china-language.gov.cn/wenziguifan/managed/002.htm>” may be referred to. Because GB has the “1 to 1” replacement of these
10 characters in its own range, we can reduce the GB possibility for them. From the zone shown in Figure 5, a small zone 0 is emptied for GB only as the Chinese simplified to traditional characters conversion zone. For compensation, priority of GB detection is raised. Also, Korean language mostly use Hangul characters and seldom uses only CJK characters in a whole string, thus Korean priority is always
15 set to the lowest one. Any language given the first priority will be raised to the top selection order. Then, to detect a CJK string’s language, we check each character belonging zone and value the string language possibility. The language with the largest possibility value is the detected one.

20 This method allows the user to select the priority language on demand. Each time, the detection program also depends on the user-selected language. It does calculation and selects the most possible language as the result. E.g., products selling in Japan; people use Korean as high priority language(In the meantime, Korean priority will be the highest one due to the user-defined priority).
25 Then, the characters in any zone will be calculated possibility and compared with the possibility of other languages, the language font with the largest corresponding value and the highest priority will be used. Of course, in this case, if all characters are in Korean zones, the string will be displayed as Korean. It will be displayed as other languages only when the other language’s possibility is
30 larger than Korean’s.

BOM (Byte Order Mark) is the information header marker for text format data. UTF-16 follows the standard with BOM Big Endian “0xFE, 0xFF” or Small Endian “0xFF, 0xFE”. In normal data source, Native Code does not have BOM and Unicode must have BOM. Non-BOM string can only be Unicode
35 without BOM in Unicode Data Device input. Then the string will be gone through for UTF8 style checking.

From each character's hexadecimal value in binary sequence, the character number can be detected and checked. If it is other languages such as, Unicode Arabic, Russian, Latin5/ASCII and Thai, the Unicode value series can be groups of 0x06, 0x0E, 0x00, 0x04 and etc. If it fulfills these conditions and is
5 larger than 90% of total string length, it is taken as non-BOM U+16 format. If the series follows BE format, then it is taken as Big Endian Unicode. If the series follows SE format, then it is taken as Small Endian Unicode.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows the overall process of the algorithm used in the present invention for the CJK Unicode language detection.

5 Figure 2 outlines the characters exist in different languages.

Figure 3 shows the character display glyphs in different language fonts.

Figure 4 shows the CJK language detection algorithm with comparisons for both value and priority selection.

Figure 5 is the CJK zone in Unicode distribution for each languages.

10 Figure 6 is a clear corresponding value table for each languages.

Figure 7 shows mapping table structure for each code.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention of CJK Unicode Language detection deals with the language distinction of characters in Unicode range. This invention applies on any digital document that has an unknown language Unicode string (a series of Unicode characters). The method is suitable for any operating system in which language is unknown.

The CJK Unicode Language detection uses statistical method to analyze the possibility of being any kind of CJK language. In the detection, some mapping value is used and a preferable value must be set to optimize the result.

The facility selects part of the sample digital document as source data. In step 201, the source data is examined with regard to mapping table. In step 201, the facility loops through each Unicode character. In step 202, each Unicode character is checked in the mapping table. If the Unicode is found, a record value will be made in step 203. If it can not be found, then it goes to step 204.

For all CJK language Unicode characters, a corresponding zone value is given as shown in Figure 5. The facility checks the source data from the corresponding zone value. If the data is found in the zone table, a value will be given for all CJK languages possibility.

Finally, all language possibilities for the source data will be presented. The preference of the user will affect the priority of the language selection with possibilities as shown in Figure 4 .

CLAIMS:

1. A CJK Unicode Language Identification method which identifies CJK language in an electronic system built with a processor, a memory and an input data stream, the method comprising:
 - 5 (a) accessing an input text data for CJK language detection, the input text data being Unicode; and
 - (b) analyzing the input text data by using statistical analysis.
2. The method according to claim 1, wherein, in the step (a), the CJK language detection divides CJK Unicode range into different zones.
- 10 3. The method according to claim 2, wherein each character in CJK Unicode range is assigned to a zone based on language usage.
4. The method according to claim 2, wherein CJK languages using same Unicode characters are considered as one zone.
5. The method according to claim 2, wherein each defined zone represents one or more languages used in the CJK Unicode range.
- 15 6. The method according to claim 2, wherein a specified zone, "zone α " represents a less frequently used GB character zone in the zone where only GB and BIG5 Unicode characters both are used.
7. The method according to claim 1, wherein, in the step (a), the CJK language detection analyzes each character, determines the character in the defined CJK language zone and increases the possibility value only for languages in that zone.
- 20 8. The method according to claim 6, wherein, in the step (a), the CJK language detection analyzes CJK Unicode data, and does not consider the GB possibility value increase in the "zone α ".
- 25 9. The method according to claim 1, further comprising:
 - (c) determining a language with highest possibility value.
10. The method according to claim 1, further comprising:
 - 30 (c) if two or more languages possibility values are same, determining a language according to pre-defined language priority order sequence.
11. A Unicode Language Identification method which identifies language in an electronic system built with a processor, a memory and an input data stream, the method comprising:
 - 35 (a) accessing an input text data for language detection, the input text data being Unicode; and
 - (b) analyzing the input text data by using statistical analysis,

wherein the input text data is SE, BE Unicode data of languages among Thai, Arabic, Russian or ASCII without BOM,

12. The method according to claim 11, wherein, in the step (a), the language detection checks for non-BOM data stream U+16 value sequence,
5 both SE and BE.
13. The method according to claim 11, wherein, in the step (a), the language detection analyzes the Unicode data, determines the character sequence, and increases the possible value for language following the language U+16 structure.
- 10 14. The method according to claim 11, wherein, in the step (a), the language detection determines a language with the highest possible value.