Network Working Group
Request for Comments: 1842
Category: Informational

Y. Wei
AsiaInfo Services Inc.
Y. Zhang
Harvard Univ.
J. Li
Rice Univ.
J. Ding
AsiaInfo Services Inc.
Y. Jiang
Univ. of Maryland

August 1995

ASCII Printable Characters-Based Chinese Character Encoding for Internet Messages

Status of this Memo

This memo provides information for the Internet community. This memo does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Abstract

This document describes the encoding used in electronic mail [RFC822] and network news [RFC1036] messages over the Internet. The 7-bit representation of GB 2312 Chinese text was specified by Fung Fung Lee of Stanford University [Lee89] and implemented in various software packages under different platforms (see appendix for a partial list of the available software packages that support this encoding method). It is further tested and used in the usenet newsgroups alt.chinese.text and chinese.* as well as various other network forums with considerable success. Future extensions of this encoding method can accommodate additional GB character sets and other east asian language character sets [Wei94].

The name given to this encoding is "HZ-GB-2312", which is intended to be used in the "charset" parameter field of MIME headers (see [MIME1] and [MIME2]).

Table of Contents

1. Introduction

Chinese (and other east Asia languages) characters are encoded with multiple bytes to guarantee sufficient coding space for the large number of glyphs these languages contain. With the prolification of internetwork traffic around the world, it becomes necessary to define ways to facilitate the transfer of text in multiple-byte character-set languages (hereafter as Chinese text) over internet.

There are two layers of concerns need to be addressed by any mechanism whose purpose is to transfer Chinese text over internet. The first is on application layer, in which concerned applications should be able to recognize the encoding of the text and/or discern different character sets which might be mixed in the text and handle it accordingly. The second layer is the actual transport of Chinese text between point A to point B over the Internet. Because the prevailing mail transport protocol used over internet, the Simple Mail Transport Protocol (aka. SMTP) was designed originally for ASCII character set only, many internet mail agents are not 8 bit clean and therefore introduce challenges for any attempt to actually implement a mechanism for the transport of Chinese text over internet.

Here we describe a mechanism for transmission of Chinese text over IP network. This described mechanism has being implemented by various software package dealing with multi-language support and has been tested on USENET newsgroups and other types of internet forums over the last two years. The test results shows that the HZ representation can pass through almost all existing mail delivery agents without being corrupted. The HZ representation currently handles GB2312-80 Chinese character set only. Further expansion to other Chinese encoding systems and to other East Asia Language is under consideration.

2. Description

For an arbitrary mixed text with both Chinese coded text strings and ASCII text strings, we designate to two distinguishable text modes, ASCII mode and HZ mode, as the only two states allowed in the text. At any given time, the text is in either one of these two modes or in the transition from one to the other. In the HZ mode, only printable ASCII characters (0x21-0x7E) are meanful with the size of basic text unit being two bytes long.

In the ASCII mode, the size of basic text unit is one (1) byte with the exception '~~', which is the special sequence representing the ASCII character '~'. In both ASCII mode and HZ mode, '~' leads an escape sequence. However, as HZ mode has basic size of text unit being 2 bytes long, only the '~' character which appears at the first byte of the the two-byte character frame are considered as the start of an escape sequence.

The default mode is ASCII mode. Each line of text starts with the default ASCII mode. Therefore, all Chinese character strings are to be enclosed with '~{' and '~}' pair in the same text line.

The escape sequences defined are as the following:

```
~{       ---- escape from ASCII mode to GB2312 HZ mode
~}       ---- escape from HZ mode to ASCII mode
~~       ---- ASCII character '~' in ASCII mode
~\n      ---- line continuation in ASCII mode
~[!-z|]  ---- reserved for future HZ mode character sets
```

A few examples of the 7 bit representation of Chinese GB coded test taken directly from [Lee89] are listed as the following:

Example 1: (Suppose there is no line size limit.)

```
This sentence is in ASCII.
The next sentence is in GB.~{<:Ky2;S{#,NpJ)l6HK!#~}Bye.
```

Example 2: (Suppose the maximum line size is 42.)

```
This sentence is in ASCII.
The next sentence is in GB.~{<:Ky2;S{#,~}~
~{NpJ)l6HK!#~}Bye.
```

Example 3: (Suppose a new line is started for every mode switch.)

```
This sentence is in ASCII.
The next sentence is in GB.~
~{<:Ky2;S{#,NpJ)l6HK!#~}~
Bye.
```

3. Formal Syntax

The notational conventions used here are identical to those used in RFC 822 [RFC822].

The * (asterisk) convention is as follows:

```
    l*m something
```

meaning at least l and at most m somethings, with l and m taking default values of 0 and infinity, respectively.

```
message             = headers 1*( CRLF *single-byte-char *segment
                      single-byte-seq *single-byte-char )
                                  ; see also [MIME1] "body-part"
                                  ; note: must end in ASCII

headers             = <see [RFC822] "fields" and [MIME1] "body-part">

segment             = single-byte-segment / double-byte-segment

single-byte-segment = 1*single-byte-char

double-byte-segment = double-byte-seq 1*( one-of-94 one-of-94 )

single-byte-seq     = "~}"

double-byte-seq     = "~{"

CRLF                = CR LF
                                                ; ( Octal, Decimal.)

CR                  = <ASCII CR, carriage return>; (    15,      13.)

LF                  = <ASCII LF, linefeed>      ; (    12,      10.)

one-of-94           = <any one of 94 values>    ; (41-176, 33.-126.)

single-byte-char    = <any 7BIT, including bare CR & bare LF, but NOT
                       including CRLF, not including > / "~~">;

7BIT                = <any 7-bit value>         ; ( 0-177,  0.-127.)
```

4. MIME Considerations

The name given to the HZ character encoding is "HZ-GB-2312". This name is intended to be used in MIME messages as follows:

```
Content-Type: text/plain; charset=HZ-GB-2312
```

The HZ-GB-2312 encoding is already in 7-bit form, so it is not necessary to use a Content-Transfer-Encoding header.

5. Background Information

A GB code is a two byte character withe the first byte is in the range of 0x21-0x77 and the second byte in the range 0x21-0x7E. As the printable ASCII subset of characters are single byte character in the range of 0x21--0x7E, two printable ASCII characters can represent a two byte GB coded Chinese character if proper escape sequence is used to indicate the proper text mode. This form the base of the above described HZ 7-bit representation methods. Further, with the use of a printable ASCII character, '~', as the leading byte of the escape sequence, the HZ representation eliminated the need of reserving any non-printable ASCII characters, which are commonly used by application programs (as well as system environment) for various control function or other special signaling. Therefore, the HZ representation method described here posses the least probability of interfering with the host and network environment. This is also a convenient for application for implementing the HZ coding method.

HZ representation method has been implemented in various Chinese software across computer hardware platforms. It has also being tested for more than two years over USENET newsgroups, alt.chinese.text and chinese.*, for the transmission of Chinese texts over the internet. The original points of those transferred Chinese texts are geographically scattered around the world and under the constraints of vast different system and network environments. Therefore, such a test group may well represent a rather complete sample of the real internet world. The successful test of the HZ representation method therefore builds up the confidence that it is well suited for transmitting multi-byte text messages over the internet.

Under HZ representation, ASCII text remain as 7-bit characters and therefore HZ representation together with the 7-bit ASCII character set can be viewed as forming a superset of characters.

6. References

[ASCII] American National Standards Institute, "Coded character set -- 7-bit American national standard code for information interchange", ANSI X3.4-1986.

[GB 2312] Technical Administrative Bureau of P.R.China, "Coding of Chinese Ideogram Set for Information Interchange Basic Set", GB 2312-80.

[Lee89] Lee, F., "HZ - A Data Format for Exchanging Files of Arbitrarily Mixed Chinese and ASCII characters", RFC 1843, Stanford University, August 1995.

[MIME1] Borenstein N., and N. Freed, "MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1521, Bellcore, Innosoft, September 1993.

[MIME2] Moore, K., "MIME (Multipurpose Internet Mail Extensions) Part Two: Message Header Extensions for Non-ASCII Text", RFC 1522, University of Tennessee, September 1993.

[RFC822] Crocker, D., "Standard for the Format of ARPA Internet Text Messages", STD 11, RFC 822, UDEL, August 1982.

[RFC1036] Horton M., and R. Adams, "Standard for Interchange of USENET Messages", RFC 1036, AT&T Bell Laboratories, Center for Seismic Studies, December 1987.

[Wei94] Wei, Yagui, "A Proposal for a Consolidated Collection of East Asian Language Coding Standards Using Solely ASCII Printable Characters", June 30, 1994.

7. Acknowledgements

Many people have involved the design and specification of the HZ 7-bit Chinese representation system at different stages. Most notable among them are Ed Lai, Chunqing Cheng, Fung Fung Lee, and Ricky Yeung. This document is merely a recollection of thoughts and efforts made collectively by this group of people whose devotion has led to the current success of the HZ Chinese representation over the Internet. Further, the authors wish to thank AsiaInfo Services Inc. for sponsoring the preparation of this document and for facilitate the communication need to refine this document.

8. Security Considerations

Security issues are not discussed in this memo.

9. Authors' Addresses

Ya-Gui Wei
AsiaInfo Services Inc.
One Galleria Tower
13355 Noel Rd. Suite 1340
Dallas, TX 75240

Phone: (214) 788-4141
Fax: (214) 788-0729
EMail: HZRFC@usai.asiainfo.com

Yun Fei Zhang
CfA
Harvard University
MS 66
60 Garden St.
Cambridge, MA 02138

Phone: (617)-860-9444
EMail: zhang@orion.harvard.edu

Jian Q. Li
Rice University
ONS - MS 119
P.O. Box 1892
Houston, Texas 77251-1892

Phone: (713)285-5328
EMail: jian@is.rice.edu

Jian Ding
ISTIC Bldg, Room 431
15 Fuxing Road,
Beijing, China 100038

Phone: 86 10 853-7120
Fax: 86 10 853-7123
EMail: ding@Beijing.AsiaInfo.com

Yuan Jiang
Electrical Engineering Department
University of Maryland
College Park, MD 200742

Phone: 301-405-3729
EMail: yjj@eng.umd.edu

10. Appendix: List of Software Implementing HZ Representation

In the following, we compiled a list on software packages support the HZ Chinese representation method. Though this list is far from complete, it is visible that support for HZ representation has be implemented for major hardware and software platforms. For more information on the listed software packages (and for other information pertain to Chinese computing), please refer to the internet site: ftp://ftp.ifcss.org/pub/software/ or its mirrors at the following sites:

at Beijing, China: ftp://info.bta.net.cn:/pub/software/;
at Shanghai, China: ftp://info.bta.net.cn:/pub/software/;
at Taiwan: ftp://nctuccca.edu.tw/pub/Chinese/ifcss/;
or ftp://ftp.edu.tw:/Chinese/ifcss/software/;
At Singapore: ftp://ftp.technet.sg:/pub/chinese/;
at California, U.S.A.: ftp://cnd.org/pub/software/.

The software in the next section are listed by its name and followed by the current version number, release date (in parenthesis) and the author(s) of the software. A brief description of the functionality of the software starts at the line immediately after the headline and lead by character string "--". Two consecutive packages are separated by a blank line.

zwdos (V2.2, March 5, 1993) by Wei Ya-Gui
-- MS-DOS kernal extension that gives DOS text mode programs the ability to enter, display, manipulate and print ’zW’ and HZ Chinese text. Small memory requirement. Supports EGA, VGA or Hercules Monographic displays.

HZ (V2.0, Feb. 7, 1995) by Fung F. Lee
-- Conversion from HZ to GB, GB to HZ, and zW to HZ respectively. Versions for PC, Mac and Unix exist.

XingXing (V4.2, Mar 29. 1995) by Wang Xiangdong
-- chinese word processor for PC.

NJStar (V3.00, Feb. 10, 1994 by Hongbo Ni)
-- GB Word Processor (Viewer, editor, printing, converter) Supports EGA/(mono)VGA/SuperVGA monitors, and various printers, Chinese<->English dictionary lookup, HanziInfo and glossary; Includes more than 20 Chinese input methods with Intelligent LianXiang and fuzzy Pinyin; Speed up with sentence based Pinyin; Reads and writes GB,Hz,zW & Big5 files; DOS Shell; Configurable.

QuickStar (V3.0, June 7, 1995) by Anthony Mai
-- Compact size Chinese edit software for PC. PinYin, CiZu, WuBi, GuoBiao, ASCII etc input method. Translate to/from GB, HZ and Big5 coded Chinese files.

cnprint (V2.6, Jan. 25, 95) by Yidao Cai
-- print GB/Hz/BIG5/JIS/KSC/UTF8 etc or convert to PostScript (conforms to EPSF-3.0). Both DOS and UNIX version available.

dm24 (V2.0, Sept. 1993) by Gongquan Chen)
-- Chinese GB/HZ printing program for EPSON 24pin printer

HXLASER (V2.6, Feb. 1994) by Chen, Gongquan
-- A GB/HZ/BIG5 file printing program for HP LaserJet plus and later model printers.

CNVIEW (V3.0, Jan. 1, 1995) by Jifang Lin
-- View GB/Hz/Big5 encoded Chinese text file on IBM-PC & compatibles

ZWLIST (V1.1, Nov. 24, 1993) by Gongquan Chen
-- Chinese HZ/GB/BIG5 File Browser for ZWDOS

zwTool (V1.0, Oct. 30,1993) by Gongquan Chen
-- a MSDOS TSR program for input of Chinese characters in text mode; Developed primarily for Chinese programmers using IDE (Integrated Development Environment, like Borland's Turbo languages); Supports GB/HZ; EGA/VGA required;

DateStar (V1.1) by Youzhen Cheng
-- Chinese Calendar Producer. Displays Chinese and western calendar in ASCII code, BIG-5 code, GuoBiao code (PRC Standard), and HZ code (Network)

MacViewHZ (V2.21 Dec. 93) by Xiaodong Chen
-- Display and print GB/HZ or BIG5 coded Chinese text files on Macintosh without Chinese OS system, with easy to use Mac user interface including multiple windows and simple editing features such as delete, copy, cut and paste.

MacHZTerm (V0.52) by Xin Xu
-- a communication program using CommToolBox, capable of displaying GB, HZ, Big5 texts on line. No Chinese OS required. System 7 recommended.

HanziTerm (V0.5) by Ricky Yeung
-- A terminal emulator for Mac Chinese OS 6.0.x or later. Support 8-bit character code, HZ, and zW.

Tex-Edit-HZ (V1.0, Dec. 18 1993) by Tom Bender and Tie Zeng.
-- A MAC WorldScript savvy Text editor with HZ<->GB conversion feature.

MacBlue Telnet (V2.6.6, Feb 16, 1995) by MacBlue
-- A Telnet program that can handle all Chinese encodings (such as HZ, GB, Big5, ET etc), EUC-JIS and EUC-KSC; based on NCSA Telnet with built-in hanzi input methods.

rnMac (V1.3b5) by Roy Wood
-- Offline Newsreader including GB <-> HZ conversion

Weiqi267 (V2.67) by Xiangbo Kang
-- record Weiqi games and transfer them through net.
GB, HZ 100 % compatible (but Russian char disabled).
There is a user guide in HZ coding.
* Now can also be used for Chinese Chess.

TwinBridge (V3.2, Nov. 16, 1994) by Twinbridge Software Corporation
-- an interface between Windows and applications, it allows Chinese character processing in Windows applications like Word for Windows, Ami Pro, Excel, etc.
You can edit Chinese characters like English characters in most of applications.

WinHZ (V1.1, April 13, 1995) by Tian Bogang
-- HZ extension for Chinese systems for Windows

HZcomm (V1.5, Nov. 14, 1993) by Nick Ke Ning.
-- HZ coding supported communication program under Chinese Windows System (GB internal coded). Good for reading/writing HZ coded E-mail and news(alt.chinese.text) on line in Windows 3.1 for PCs.

SimpTerm (V0.8.0) by Jianqing Hu
-- A Chinese communication program for MS-Windows 3.1 with build in support for BIG5, HZ and GB encoded text.

ChPad (V1.31) by Tian Bogang
-- GUO BIAO and HZ file browser for MS WINDOWS 3.1

SilkRoad (V1.0) by Antony C. Hu
-- GB/HZ Viewer for MS-Windows 3.1

gnus-chinese (V1.0, Apr. 26 1994) by Ning Mosberger-Tang
-- convert HZ articles to the code understandable by your terminal automatically in GNUS newsreader (for GNU EMACS).

requires conversion program (e.g. hz2gb and gb2hz) to do the actual conversion.

irchat (V2.4jp4cn0) by HIROSE Tutomu
-- irc client e-lisp program on Mule
patched to handle HZ and Big5
now we can read/write all JIS/HZ/Big5 simultaneously on irc

hztty (V2.0 Jan 29, 1994) by Yongguang Zhang
-- This program turns a tty session from one encoding to another. For example, running hztty on cxterm can allow you to read/write Chinese in HZ format.

BeTTY/CCF/B5Encode package (V1.534, 1995.03.22) by Jing-Shin Chang
-- a chinese code conversion package for codes widely used in Taiwan and the GB code widely used in Mainland, plus a 7-bit Big5 encoding method (B5Encode3/B5E3, an extension to HZ encoding for GB),
including off-line converters (CCF/Chinese Code Filters and B5E/B5Encode) and an on-line converter (BeTTY) which simulates your native chinese terminal to become aware of the coding systems widely used in Taiwan and GB, HZ encoding.

gb2jis & jis2gb (V1.5, 1995.5.11) by Koichi Yasuoka
-- convert GB (or HZ) to/from JIS with two-letter pinyin

gb2ps (V2.02) by Wei SUN
-- convert GB/HZ to postscript, supports simple page formatting (change chinese fonts and font size, cover page, page number, etc). Five chinese fonts are provided in this release, they are Song, Kai, Fang Song, Hei and FanTi
The HZ ENCODING is also supported.

ChiRK (V1.2a) by Bo Yang
-- GB/HZ/BIG5 text viewer on terminals (or emulations) capable of displaying Tektronics 401x graphics, such as GraphOn,DEC VT240/330, Xterm, Tektool on Sun, EM4105 on PC, VersaTerm-Pro on Mac, etc.

Multi-Localization Enhancement of NCSA Mosaic X 2.4 (V2.4.0)
by TAKADA, Toshihiro
-- a patch to make use of various nat'l character sets in NCSA Mosaic for X 2.4. You can switch between char-sets in one Mosaic. Support ISO 8859-X, KOI-8, GB, HZ, BIG5, KSC & JIS.