Advanced Lecture on Internet Applications

5. Text based Communication:

Character Code and Internationalization (2)

Masataka Ohta

mohta@necom830.hpcl.titech.ac.jp

ftp://ftp.hpcl.titech.ac.jp/appli5e.ppt

What is “Character”

- unit to represent language by graphic symbol
  - phonetic character
  - ideogram

What is Script（用字系）?

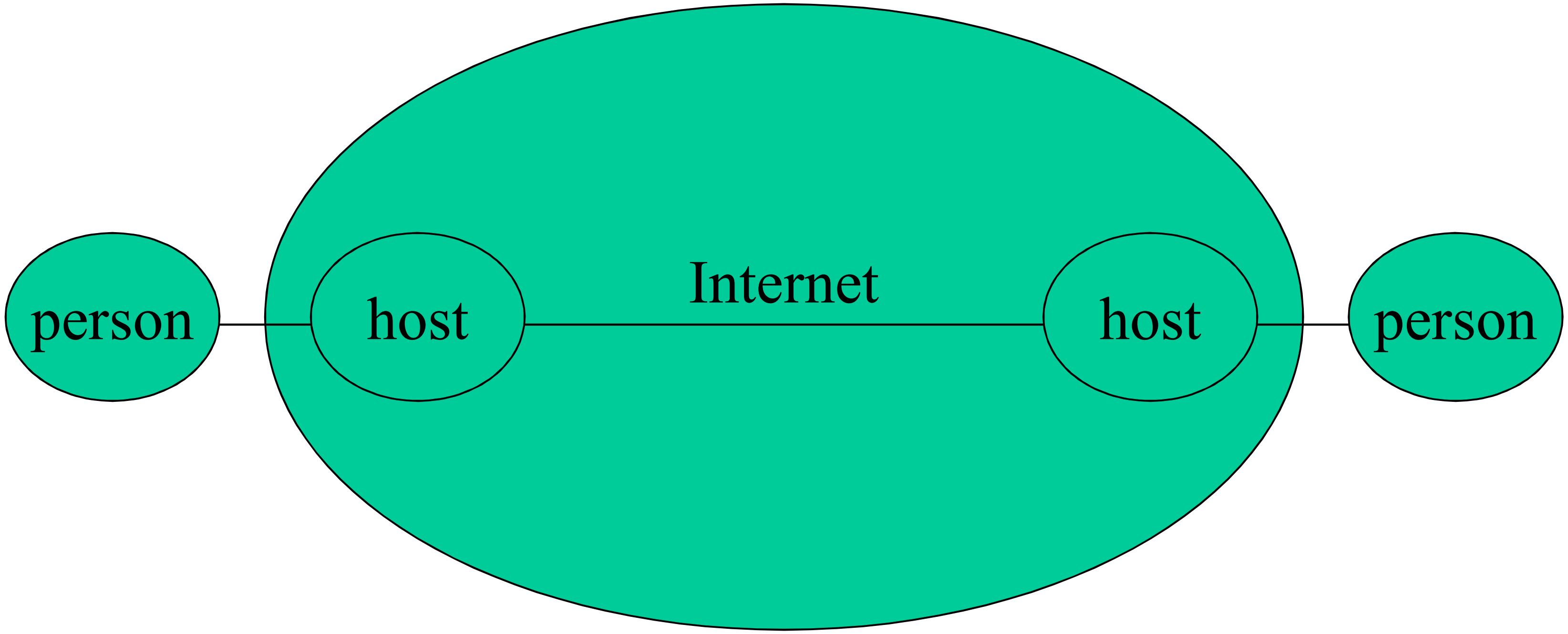
- system to represent language by characters
- never confuse language and script
  - –「でぃすいずあぺん」is English by Kana script
  - – “Koreha pen desu.” is Japanese by Roman (Latin) script

Scripts to Represent Japanese

- kana (hiragana, katakana, manyogana)
- mixed kanji kana
- romaji (Hepburn, Monbusho, etc.)
  - “masataka” in French should be “massataka”
- and phonetic representations in various local script systems such as Hangul

Internet and Internationalization (I18N)

- Internet
  - connects hosts around the world
- should all the hosts be internationalized?
  - maybe
- Internet
  - connects people around the world
- should all the people be internationalized?
  - maybe, but, ...



end to end principle beyond hosts

Internationalization

- internationalized person?
  - person to be able to use English?
  - person to be able to use major six languages?
  - person to be able to use all the languages in the world?
  - person to be able to use internationally used language in each field!
    - Japanese literature is discussed in Japanese internationally
      - manga and animation, too?
    - internet technology is discussed in English internationally

Internationalization of Hosts (1)

- internationalization of computer languages and text based protocols (ftp etc.)
  - artificial language with limited characters
  - has nothing to do with natural languages
- K&R C and Internationalized ISO C
  - K&R C use national variant characters of ISO 646
  - escape mechanism of trigraph is introduced for ISO C
    - mostly ignored even in Japan

Localization (L10N) of Hosts (1)

- localized computer language may exist
  - but no one use
  - kana COBOL
- localized protocol is meaningless
  - internet connects all the hosts in the world

Localization (L10N) of Hosts (2)

- output locally used characters
  - localized character code
  - bitmap display + localized font
    - was using character generator ROM
- input locally used characters
  - localized key board?
  - information of dictionary on characters necessary (e. g. for kana kanji conversion)

Localized Character Code

- national versions of ISO646
  - have freedom of 12 characters
    - too small even for western Europe
- the first multibyte character set developed by Japan
  - 2 byte kanji code JIS X 0208

Simplicity of ASCII (or Latin Script for English) (1)

- small number of characters
- horizontal only
- single (left to right) directional only
- ligature (variation of character shape by previous/next characters) is not necessary
- commonly shared recognition for character identifications and character shapes

Simplicity of ASCII (or Latin Script for English) (2)

- correspondence between small/capital characters is clear and regular
- no characters with diacritical marks
  - such as “ä”
- character width can be constant
- widely spread and usable everywhere

Complexity and Simplicity of JIS X 0208 (1)

- large number of characters
- horizontal and vertical
  - vertical was not supported so seriously
- single (left to right) directional only
- ligature (variation of character shape by previous/next characters) is not necessary
  - though circle mark for composition exists
    - not really used for composition

Complexity and Simplicity of JIS X 0208 (2)

- no commonly shared recognition for character identifications and character shapes
  - is the serious problem
- correspondence between hiragana/katakana characters is not so clear and regular「ブ」
- diacritical (?) marks「゛゜」are precombined
- character width can be constant
- widely spread and usable everywhere

Internationalization of Hosts (2)

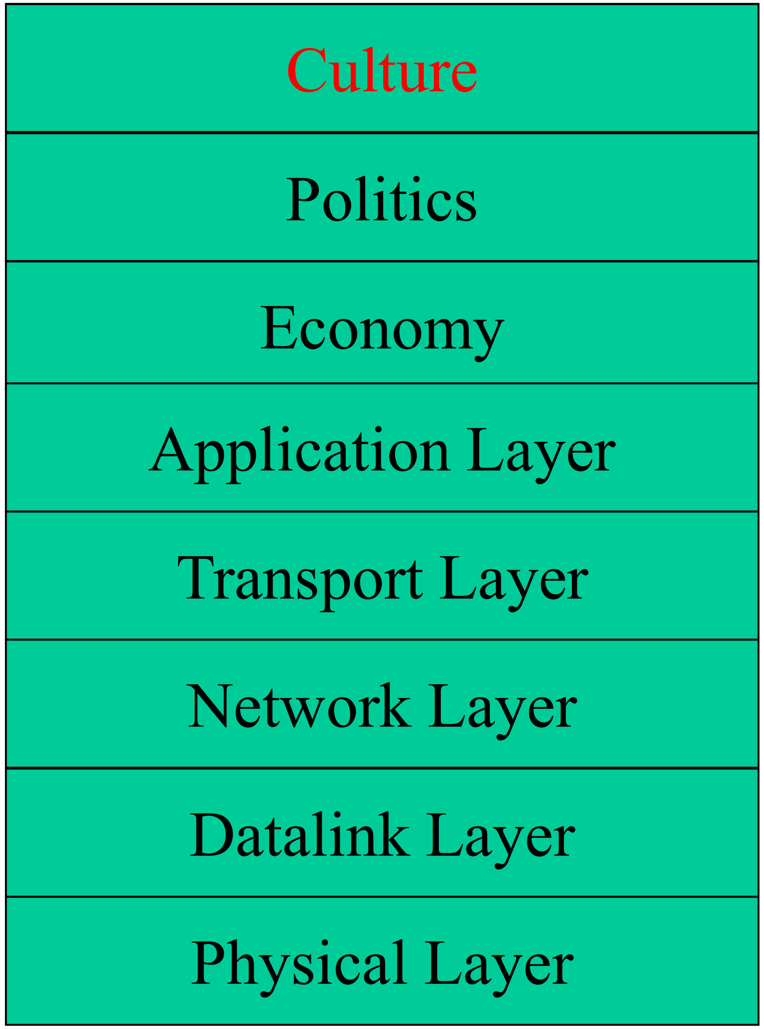
- output all the locally used characters in the world
  - need character code supporting all the local characters in the world
    - ISO 2022

Character Code

- an encoding rule for strings using characters of a character set
  - not merely assign code (number) to characters
    - the rule may be very complicated
- the number of characters of a character set matters
  - if large, many bits are necessary
  - if small, many characters can’t be represented
    - small differences between similar characters can’t be represented

Internationalized Character Code

- must contain large number of characters
- requirements for character code differs culture by culture
  - how can various requirements unifited?
    - in practice, must have separate character code for each country
    - least common framework desired
      - ISO 2022
      - finite stateness



Culture

Politics

Economy

Application Layer

Transport Layer

Network Layer

Datalink Layer

Physical Layer

Layering Structure over the Internet!

ISO 2022

- standard to switch national character sets
- can simultaneously handle four (G0, G1, G2, G3) 7 bit (multi)byte character sets
  - G0 as is, G1 with SI/SO or 8th bit set
  - G2/G3 after SS2/SS3
- assignment to Gn set can be switched by escape sequences
  - “ESC$B” assign G0 to ASCII

Shift Code

- escape sequences of ISO 2022 is necessary for internationalization
  - localization for a few sets can be simpler
- within (?) a localized hosts
  - 1~2 byte variable length code desired
- Shift JIS was developed in Japan (Fukase of “ASCII”), BIG5 and BGK followed
- is used also for communication

ISO 2022 and EUC

- EUC (Extended Unix Code)
  - ISO 2022 with fixed assingments to Gn
    - often only use G0 and G1
      - in Japan, ASCII and JIS X 0208
- caused a lot of confusion by incompatibility between Shift JIS and EUC
- EUC-GB was exebded ti shifted GBK

Structures of Text

- text has various structures
  - line changes
  - font changes
  - underlining
  - switch vertical/horizontal
  - left/right directionality
  - chapter, section, subsection, paragraph, ...
  - indent

Plain Text and Structured Text

- plain text does not have structure?
  - have line changes, at least
  - may have underlining on some text processing tools of UNIX
- structure text may have any structures
- where is the border between plain and structured text?
  - abstract cultural arguments can last forever

Why do We Need Charater Code?

- input characters
- output characters
  - so far, picture without coding is OK
- process (especially, search) characters
  - the primary reason to use character code

Character Code and Finite Stateness

- if a string is finite state
  - everything can be deterministic
  - search is efficient
- if a string is not finite state
  - needs prohibitively large processing time
    - e. g. search with LBA is O(L^3)
  - may processing is uncomputable
- character code should be finite state!

Structures supported by Plain Text

- can not support nesting (require PDA)
  - line changes
  - font changes
  - underlining (not many lines)
  - switch vertical/horizontal
  - left/right directionality (no nesting)
  - no chapter, section, subsection
    - paragraph, maybe
  - indent as simple as TAB

Unified Framework for International Character Code

- must support input/output of all the characters in the world
- must be finite state
  - though ISO 2022 support nested directionality
- differentiation of plain and structured texts is essentially important

Internationalization of Hosts (3)

- hosts supporting character input
  - must support internationalized character code
  - must support local input methods of characters locally used in various locations in the world?
    - must have standard input device
      - ASCII keyboard
      - smart phone
    - input methods of characters different by culture
      - download software?
        - binary is different CPU by CPU

Internet and Internationalization (I18N)

- Internet
  - connects hosts around the world
- should all the hosts be internationalized?
  - maybe
- Internet
  - connects people around the world
- should all the people be internationalized?
  - maybe, but, ...

To Internationalize Character Processing Environment

- character code should be internetionalized
  - character font must be accessible over the Internet
  - dictionary information on characters must be accessible over the Internet
  - input methods must be accessible over the Internet

Wrap Up

- internet makes internationalized character code necessary
- internationalize character code needs internetinalization
- never confuse I18N and L10N