Some notes on Japanese TEXt Processing

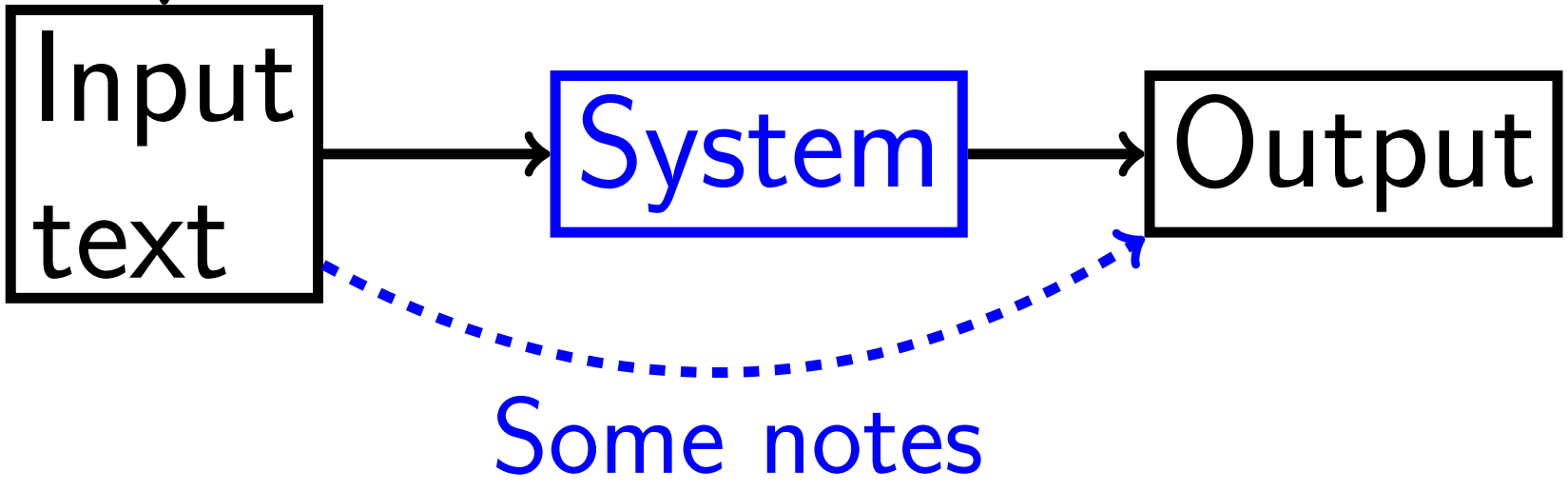
KUROKI Yusuke

kuroky(at)users.sourceforge.jp

October 24, 2013

Overview

– IME: input method editor



IME: input method editor

- There are several ways to input Japanese into computer. Usually,
  1. input *kana* first (directly, by romanization, by pocket bell style, by flick input[1], etc.), then
  2. change them to *kanji-kana-majiri* correctly by human
- The software, IME, helps both operations above
- Users freely to choose where they change *kana*s to *kanji-kana-majiri*.
- Users often turn on IME to input Japanese & off to Latin. In writing TEX source, we change the modes frequently.

[1] With help of Moe Masuko

TEX-related systems to operate Japanese

- De facto standard in Japan:
  pTEX (engine extention) + jsclasses class files
- New age: LuaTEX-ja (macros of TEX & Lua for LuaTEX)
- Experimental stage?: ConTEXt MkIV

- upTEX
  (change the internal operations of pTEX into Unicode)
- ConTEXt MkII + pTEX
- CJK package + Takayuki YATO's package
- X∃TEX+ Takayuki YATO's package

Note for line-breaks

- Roughly speaking, Japanese words could be split anywhere due to line-ending
- Input (e.g., in case of 5 em line-breaking):

これは僕が飼っている犬です。	v.s.	This is the dog which I keep.

- Output:

No Good **これは僕が 飼っている 犬です。**
Good **これは僕が飼っている犬です。**
v.s. This is the dog which I keep.

- Sometimes, we need a little space as the author indicates, e.g., pTEX **は中野 賢さんほかにより作られた。**

Note for Unicode input

When we use JIS X 0208 character set, we could sort out which areas are for Japanese and which for Latin easily.

- multi-byte area should be for Japanese
- ASCII area should be for Latin

§ § (input \S before Unicode age)
“ “ (‘ ‘)
” ” (’ ’)

In Unicode age, since some signs and marks are combined, we will need indicate which area is in which language.