Segmenting Complex Scripts with Machine Learning

Line and word breaks

Word breaks

Dictionary based segmentation

where it fall short?

- size is too large
- new or specialized words are not easily recognized (xx-ing)
- longest match can fail by missing correct shorter words

2 Board cases needed difference solutions

- south east asian SEA
- East Asian CJK

CJK:

Adaboost: many tiny rules each vote on whether a break is good, combined votes decide word boundaries

RadaBoost: Radicals are the components of Han characters. Certain radicals frequently appear together, provides useful cues for word segmentation.

BudoX/RAdaBoost

AdaBoost learners

ICU dic 2.0M

BudoX zh-hant 64kb. zh-hans 63kb, Radical (all zh variants) 60kb

Revision #2
Created 12 November 2025 22:52:43 by itsLittleKevin
Updated 18 November 2025 21:35:02 by itsLittleKevin