Word segmentation

Word definition

Maximum matching

- Linguistics
  - Object: natural languages
- Lexicography
  - Sub domain of linguistics
  - Object: vocabulary

- History of Vietnamese language
- Chinese
- Han-Nom
- Quoc-ngu

- Categorization
  - Content words: noun, verb, adjective, adverb
  - Function words: pronouns, preposition, conjunction

- Noun: cơm, gạo…
- Verb: chạy
- Adjective: (vấn đề) khó
- Adverb: (chạy) nhanh

- Containing only one syllable
- Most of monosyllable words are thuan-Viet

- Contains multiple syllables
- Từ ghép đẳng lập: to lớn, khỏe mạnh
- Từ ghép chính phụ: bằng khen, phòng ăn

- Chinese orgin
- Composing of meaningful Chinese characters
- E.g.: học viên, nhân viên, công trường, đại học, phổ thông

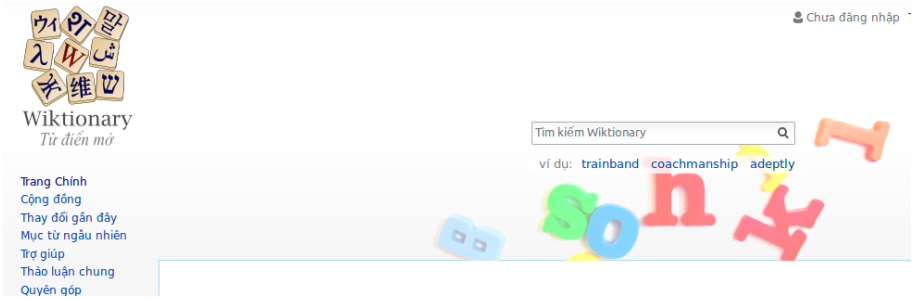
Word segmentation
Word definition
Foreign words

- Borrowed from other languages like English, French, Russia, Japanese, Korean.
- Domain terminologies
- E.g: cây (borrowed from ‘case’ in English), web, data.

Word segmentation
Word definition
Dictionaries

- http://vlsp.org.vn/
- Containing 40,000 common Vietnamese words

Word segmentation
Word definition
Dictionaries



Word segmentation
Word definition
Dictionaries

- Persons: Hồ Chí Minh, Ngô Bảo Châu
- Locations: Hà Nội, Nhật

Độ dài	# từ	%
1	6,303	15.69
2	28,416	70.72
3	2,259	5.62
4	2,784	6.93
5	419	1.04
Tổng	40,181	100

"*Tách_từ là bài_toán nhận_diện từ trong văn_bản tiếng_Việt .*"

- WS is a problem in many Asian languages including Chinese, Japanese, Vietnamese, Thai, and Bummese.

- Input:
  - Dictionary
  - Unsegmented texts
- Algorithm:
  - Greedy algorithm
  - Move from left to right; Get the longest word; Repeat until finish

- **START** — *initialize*
- (1) Input sequence [*w_0 w_1 … w_n-1]*
- (2) *words* ← *[]*
- (3) *s* ← *0*
- (4) *e* ← *n* — *iteration*
- (5) When [*w_s … w_e]* **has not been a word yet**: *e* ← *e – 1*
- (6) *words* ← *words* + *[w_s … w_e]*
- (7) *s* ← *e* + *1*
- (8) If *e* < *n:* Return to (4)
- (9) Get the segmented *words* — *finish*
- **END**

- Example 1:
  “thời khóa biểu đang được cập nhật”
- Example 2:
  “môn học xử lý ngôn ngữ tự nhiên”
- Example 3:
  “con ngựa đá con ngựa đá”

- Example 1:
  “thời khóa biểu đang được cập nhật”
  → *“thời_khóa_biểu đang được cập_nhật”*
- Example 2:
  “môn học xử lý ngôn ngữ tự nhiên”
  → *“môn_học xử_lý ngôn_ngữ tự_nhiên”*
- Example 3:
  “con ngựa đá con ngựa đá”
  → *“con_ngựa đá con_ngựa đá”*

“*học sinh học sinh học*”

- Left to right
  “*học_sinh học_sinh học*”
- Right to left
  “*học sinh_học sinh_học*”

Q&A

- Advantages:
  - Simple implementation
  - O(V.n)
- Shortcomings:
  - Require dictionaries
  - Segmentation ambiguities

hieunk@soict.hust.edu.vn