NLP

Introduction to NLP

Preprocessing

Text Preprocessing

- Removing non-text (e.g., ads, javascript)
- Dealing with text encoding (e.g., Unicode)
- Sentence segmentation
- Normalization
  - labeled/labelled, extra-terrestrial/extraterrestrial, extra terrestrial
- Stemming
  - computer/computation
- Morphological analysis
  - car/cars
- Capitalization
  - Now/NOW, led/LED
- Named entity extraction
  - USA/usa

Text Preprocessing

- Types vs. Tokens
  - To be or not to be
- Tokenization:
  - ALS vs. A.L.S.
  - Paul's, Willow Dr., Dr. Willow, New York, ad hoc, can't
  - "The New York-Los Angeles flight" vs. "Minneapolis-St.Paul"
  - Numbers, e.g., (888) 555-1313, 1-888-555-1313
  - Dates, e.g., Jan-13-2012, 20120113, 13 January 2012, 01/13/12
  - URLs

Word Segmentation

- 金属製品製造の日立金属は19日、世界最大手の鉄鋳物メーカー「ワウパカ　ファウンドリー　ホールディングス」（米国・デラウェア州）を米投資ファンドから買収し、完全子会社にすると発表した。買収額は13億ドル（約1330億円）で、10月中にも手続きを終える。

Word Segmentation

- Arabic:

  كتاب

- Japanese:

  この本は重い。

  (kono hon ha omoi)

- German:

  Finanzdienstleistung = financial services

- Chinese:

  电视 (television)
  电 (diàn = electric) 视 (shì = to look at)

Text Preprocessing

ニューヨーク (New York) は、アメリカ合衆国ニューヨーク州にある都市

- Kanji, Katakana, Hiragana, Rōmaji, (numbers)
- Nyūyōku wa, Amerikagasshūkoku nyūyōku-shū ni aru toshi

Sentence Boundary Recognition

- Decision trees
- Features
  - punctuation
  - formatting
  - fonts
  - spacing
  - capitalization
  - case
  - use of abbreviations, e.g., Dr., a.m.
- Example
  - If there is no space after a period, don't assume that there is a sentence boundary

NLP