自然言語処理入門
「貯め込んだテキスト」

東京大学　情報基盤センター
（情報理工学系研究科、情報学府　兼担）
中川裕志
nakagawa@r.dl.itc.u-tokyo.ac.jp
http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/

パソコン、ハードディスク、CD-ROM、インターネット

- ２０年くらい前にワープロが普及して電子テキストが一般的になった(パソコン、ハードディスク）)
- 電子テキストを共有、共用したくなる（CD-ROM、インターネット）
- 共通なフォーマットを持った電子テキスト
  - 平文（plain text)が使われ
  - タグつきの構造化文書へ SGML→XML
- 大量なテキストの蓄積、整備（コーパス）と利用が始まった。
- コーパスを利用するための言語資源の整備

言語情報資源とは

- コーパス
- 言語モデル
- 辞書
- 文法
- シソーラス

コーパス(Corpus, Corpora)

- コーパスとは電子化され大量の音声、テキストデータ
- 音声コーパス
  - １　一人　vs　多人数
  - ２　目的発話か自然発生的か
  - ３　書き起こしテキストの有無
  - ４　書き起こしテキストにタグがついているか
- テキストコーパス
  - １　生コーパス
  - ２　タグ付コーパス（形態素解析され品詞タグがついている）
  - ３　括弧付コーパス
  - ４　解析済みコーパス
- 単言語コーパス　vs　多言語コーパス（対訳コーパス）
- 分野、ジャンル

  新聞、雑誌、インターネット、専門的な学問分野

コーパスに基づく確率、統計を基礎にした言語処理の長短

- accurate
- broad-coverage
- robust
- data-driven　vs　rationalism
- コーパスの収録範囲を越えた分野への適用可能性に問題あり
- sparseness　あるいは　ゼロ頻度問題

二言語コーパス(bilingual corpus)

- 二言語コーパスとは二つの言語の文からなるコーパス
- comparable corpus
  - 内容が comparable なコーパス
- 対訳コーパス
  - ０ 単に対訳関係にあるということだけが分かっている二言語コーパス
  - １ 対訳文の対応付け (alignment) がされたコーパス:
  - ２ 単語の対応付けもされたコーパス
- alignment をつけること自体が90年代におけるコーパス言語処理の主要な研究テーマだった。

辞書

- 当然、電子化辞書
- 冊子体の辞書の代用（単言語、対訳）
- ワープロの仮名漢字変換用辞書（単言語）
- ワープロのスペルチェック用（単言語）
- 機械翻訳用の対訳辞書（意味の記述も場合によっては必要）
- 自然言語理解用（意味の記述された辞書）
- 情報検索用（キーワードと文書の対応の情報あり。）
- クロスリンガル検索用（対訳辞書）

辞書の記載項目

- 見出し
- 発音、読み
- 表記
- 品詞
- 文法情報（活用、文型）
  - １　活用は屈折型言語における　性、数、格、時制、アスペクト
  - ２　膠着型言語における　時制、アスペクト、接続
- 連接情報、共起情報
- 意味（どうやって記述するかが問題）、概念構造
- 用例
- 関連語（同義語、反義語）
- 訳語
- 選択制限など

辞書記述の例（IPAL辞書　部分）

見出し、読み	あげる（1）　アゲル
文法情報	下1　語幹は age
表記	上げる、揚げる
文法情報	他動詞<<統語>>
文型	N1がN2を（N3から）（N4に）（N5へ）
用例	彼は本を棚に上げた
意味	何かを上方に移す
関連語	上位語＝上、反義語＝下ろす

辞書の記載項目　意味や文法に係わる詳細　その1

- 動詞の主語や目的語にとなる名詞の性質
  - １　人称　　例）〇「私は暑い」、×「彼は暑い」
  - ２　有情（生物）か無情（非生物）　例）〇「部屋は暑い」
- 動詞の可能な態　（受動、能動、使役、相互など）
  - 〇「殴る　vs　殴られる」
  - ×「生まれる　vs　生まれられる」
  - 〇「生まれる　　vs　生ませる」
  - ×「病む　　vs　　病ませる」
  - ×「生まれる　vs　生まれあう」
- 動詞自体の意味素性
  - １　自動、他動、などの文法的性質
  - ２　移動、状態変化、着脱、模様替え、などなど
- 名詞の意味素性
  - １　場所、時間、生物、具体物、抽象物、組織、……
- 動詞の目的語や主語（格）に可能な名詞の意味素性
  - 「Nから来る」 Nは場所　など。
- 選択制限も意味素性で記述できる。

辞書の記載項目　意味や文法に係わる詳細　その2

- 　〇「早く走る」、×「早い走る」
  これを組織的にまとめると句構造文法になる。

- 接尾辞の連接
- 　動詞に後接する接尾辞の接続可能性も重要
  - 　〇「死んである」 vs　×「殺してある」
- 日本語の接尾辞
  - 　ている、てある、ておく、てしまう、てくる、ていく、がる、ない
  - 　てくれる、てやる、てあげる、てもらう、

シソーラス

□ 単語をその単語が表す概念と他の単語の概念の間の関係として記述した体系をシソーラスという。

□分類語彙表（国立国語研究所で開発された日本語のシソーラス）の例

１．１　抽象的関係
　　１．１００　こそあど

１．２　人間活動の主体
　　１．２００　われ、かれ

１．３　人間活動―精神および行為
　　１．３００　心

１．４　生産物および用具
　　１．４００　物品　－　金品、異物、現品、安物、名物、

　　１．４７１　道路・橋　－　国道、地下道、線路、掛け橋

１．５　自然物および自然現象
　　１．５００　刺激
　　１．５０１　光　―　光明、光輝、光彩

概念間の関係

- 上位下位関係:
  - モノ＞生物＞動物＞脊椎動物＞哺乳類＞人間
- 人工知能では is-a kind-of 関係などという。
- 兄弟の関係:
  - 特急と急行と鈍行と各駅停車（緩行）
- 下位語 hyponym
  - OS → Linux
- 反意後 antonym （反意語は多くの場合、兄弟間の関係となる。）
  - 出発←→ 到着, 北←→南
- 同義語 synonym
  - 日銀=日本銀行、 首相＝？総理大臣
- 部分 meronym vs 全体 holonym → part of
  - ボンネット、タイヤ ← 車

WordNet

□WordNetはプリンストン大学のGeorge A. Miller教授が中心となって開発が進められた英語のシソーラスである.

□　EuroWordNetは，WordNetと同じ方式でヨーロッパ言語のシソーラスを作成するプロジェクトである．各ヨーロッパ言語のSynset(意味クラス、同義語の集合)は，対応する英語のWordNetのSynsetへのリンクを持ち，これにより任意の言語対について同義語を検索することを可能にした.
また，シソーラスの上位の構造は完全に共有されている.
対象言語はオランダ語，イタリア語，スペイン語，ドイツ語，フランス語，など

WordNet 1.6 overview for "human"

The **noun** "human" has 2 senses in WordNet.

1. person, individual, someone, somebody, mortal, **human**, soul -- (a human being; "there was too much for one person to do")
2. homo, man, human being, **human** -- (any living or extinct member of the family Hominidae)

Search for of senses
Show glosses
Show contextual help

The **adjective** "human" has 3 senses in WordNet.

1. **human** -- (characteristic of humanity; "human nature")
2. **human** -- (relating to a person; "the experiment was conducted on 6 monkeys and 2 human subjects")
3. human (vs. nonhuman) -- (having human form or attributes as opposed to those of animals or divine beings; "human beings"; "the human body"; "human kindness"; "human frailty")

Search for of senses
Show glosses
Show contextual help

Choose a different search word

WORDNETの例1

The **noun** "human" has 2 senses in WordNet.

1. person, individual, someone, somebody, mortal, **human**, soul -- (a human being; "there was too much for one person to do")
2. homo, man, human being, **human** -- (any living or extinct member of the family Hominidae)

WORDNETの例2

- The **adjective** "human" has 3 senses in WordNet.

  1. **human** -- (characteristic of humanity; "human nature")
  2. **human** -- (relating to a person; "the experiment was conducted on 6 monkeys and 2 human subjects")
  3. human (vs. nonhuman) -- (having human form or attributes as opposed to those of animals or divine beings; "human beings"; "the human body"; "human kindness"; "human frailty")