自然言語処理論I

9. 辞書と概念階層

辞書

- 単語に関する様々な情報を記載した知識データベース
- 様々な知識=自然言語処理に必要な知識
  - 形態素解析
  - 構文解析
  - 意味解析
- 今までの例
  - 単語辞書
  - 格フレーム辞書

辞書に記載される情報

- 形態素情報(英語の場合)
  - 発音
  - 品詞
  - 語形変化
    - ✦ 名詞、代名詞の数・格による変化
    - ✦ 動詞の数・人称・時制による変化

辞書に記載される情報

- 形態素情報(日本語の場合)
  - 読み
  - 品詞
  - 左右の接続可能性
  - 語形変化
    - ✦ 活用型(ア行五段活用、一段活用...)
    - ✦ 活用形(未然形、連用形...)
  - 表記のゆれ

表記のゆれ

- 異表記
  - 「言い換える」の場合
    - 言い換える、言い替える、言い変える、言いかえる、言換える、いい換える...
  - 個別に記述すると効率が悪い
- 異表記に関する情報を辞書に記載する

言	い	換	える
い(読み)	(省略可)	か(読み)	
		替(別表記)	
		変(別表記)	

辞書に記載すべき情報

- 構文情報
  - 表層格フレーム(日本語)
    - 動詞、形容詞、形容動詞
    - 取り得る表層格(ガ格、ヲ格など)
  - 下位範疇化フレーム(英語)
    - subcategorization frame, SUBCAT
    - 動詞は支配する主語、目的語、前置詞句など

辞書に記載すべき情報

- 意味情報
  - 格フレーム辞書
    - 表層格と深層格の対応
    - 選択制限
      (格として現われる名詞に関する意味的な制約)
  - 意味素
    - 特に名詞
    - human, concrete, abstract, ...

辞書のシステム

- 辞書を使うときの流れ
  - 単語を検索する
  - 情報を取り出す
- 表形式の辞書モデル

表形式の辞書モデル

- key=単語
- 内容=品詞、読み、格フレームなど
- 単語を入力とし、keyが同じレコードを検索

レコード$_1$	key$_1$	内容$_1$
レコード$_2$	key$_2$	内容$_2$
	｜	｜
レコード$_n$	key$_n$	内容$_n$

辞書探索アルゴリズム

- 実用的な辞書は登録単語数も多い
  - 数千〜数十万語
  - 効率よく検索することが重要
- 辞書探索アルゴリズム
  - 線形探索
  - 二分探索
  - ハッシュ法
  - トライによる辞書検索

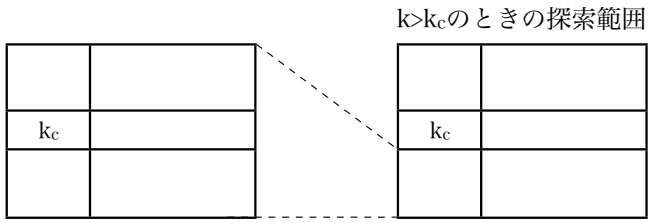
線形探索

- linear search
  - 表の一番上から順に調べる
  - 計算量 O(n)
    - ✦ nはレコードの総数
  - 実用的ではない

二分探索

- binary search
  - レコードをある順序(ABC順など)で並べる
  - 探索範囲の真中にあるキー k_c を調べる
  - 見つからないときは探索範囲を狭める
    - ✦ $k < k_c$ なら上半分、$k > k_c$ なら下半分 (kは検索キー)
  - 計算量 $O(\log_2 n)$

繰り返す



ハッシュ法 (hash)

- ハッシュ関数 H(q)
  - 検索単語qからインデックスを計算する関数
  - ex. (qの文字コードの総和) mod N
- ハッシュ表 (表形式の辞書)
  - インデックスがH(q)のところに単語qの情報を置く
- 問題点: 衝突
  - 複数のキーに対してH(q)が同じ値を返すこと
  - チェイン法
    - ハッシュ表のレコードを線形リストにする
    - 同じインデックスに複数の単語を登録

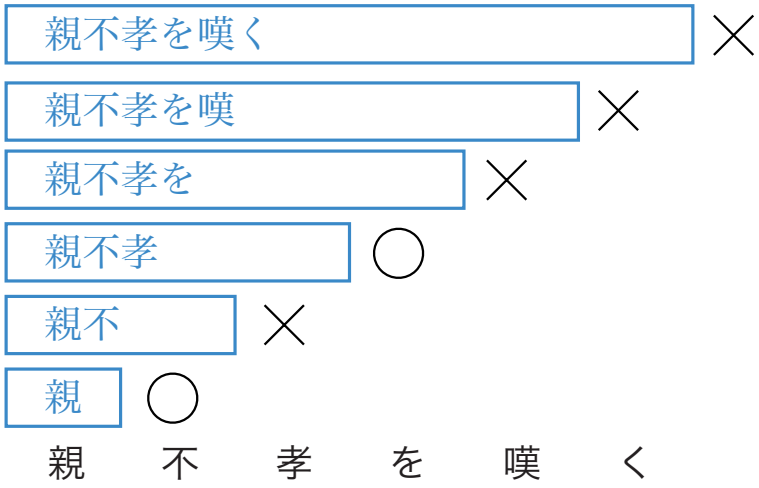
ハッシュ法 (hash)

- 計算量 O(1)
  - 衝突が発生しない理想的な状態のとき
- 衝突を避けるには
  - ハッシュ表のサイズを大きくする
    - 一般に、辞書(ハッシュ表)のサイズは登録単語数よりはるかに大きくなる
  - 良いハッシュ関数H(q)を使う
    - qからただちに計算できる
    - 衝突をなるべく起こさない

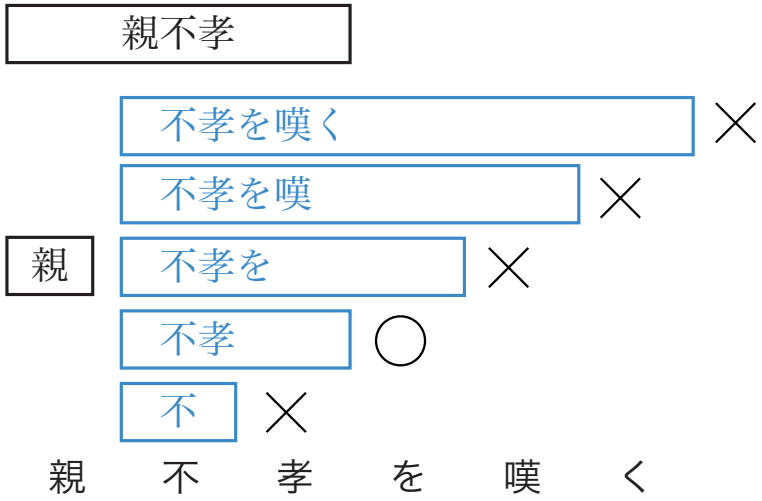
トライによる辞書検索

- トライ(trie)
- 日本語の形態素解析(復習)
  - 分かち書き(単語分割)も同時に行う
  - ありとあらゆる部分文字列に対して辞書引きが必要
    - 文字数がnのとき、$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ 回の辞書引きが必要

辞書引きの過程

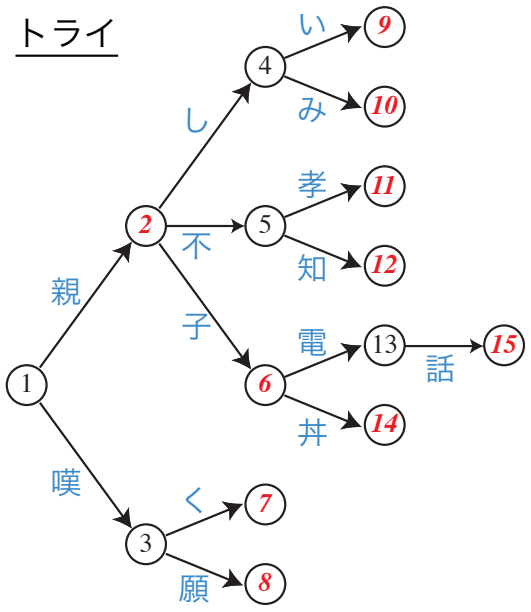


辞書引きの過程



トライ

- 単語の共通接頭辞を併合した木構造
- 日本語の形態素解析に最適
  - 「親不孝を嘆く」の場合
    - トライを1→2→5→11とたどる
    - ノード2と11から、「親」と「親不孝」を単語ラティスに加える
  - 探索効率が良い
    - 文字数nの文の場合、トライをn回たどれば全ての単語候補を取り出すことができる
    - 矢印がなければ、その先はたどる必要がない



	登録単語
2	親
6	親子
7	嘆く
8	嘆願
9	親しい
10	親しみ
11	親不孝
12	親不知
14	親子丼
15	親子電話

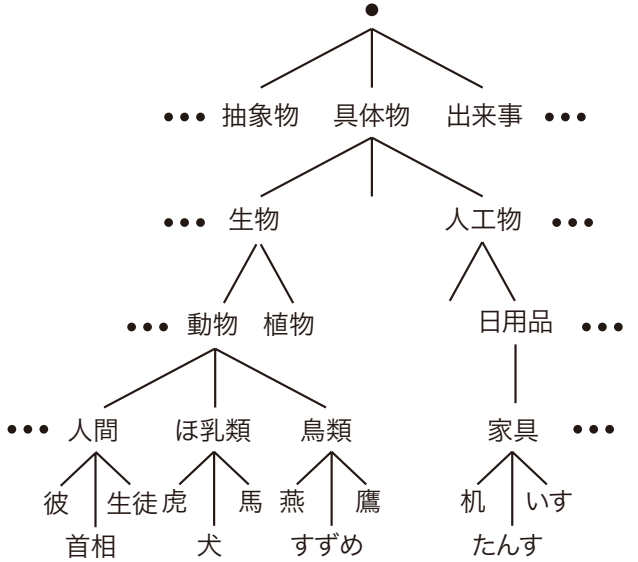
辞書の例

- EDR日本語単語辞書
  - 40万語
  - 品詞、読み、品詞間の接続可能性、意味分類
- IPAL辞書
  - 格フレーム辞書

概念階層

- シソーラス (thesaurus)
  - 語と語の間の意味的関係を記述したデータベース
  - 意味的関係の例
    - ✦ 上位語、下位語
      - ○ 生物—動物，果物—りんご
    - ✦ 同義語
    - ✦ 反義語
- 木構造で表現されることが多い
  - 親子関係が上位-下位関係を表わす
  - 意味の似ている単語ほど近い位置に配置される

シソーラスの例

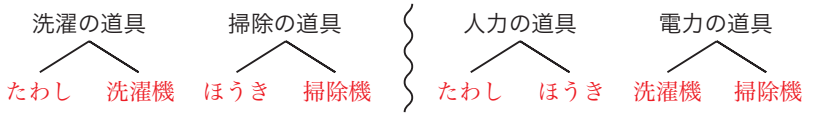


シソーラス

- 用語
  - 意味クラス
    - ✦ シソーラスの内部ノードが表わす単語の集合
    - ✦ 自然言語または記号で表現される
  - 深さ
    - ✦ 根ノードからのパスの長さ
- 2種類のシソーラス
  - 分類シソーラス
    - ✦ 単語はシソーラスの葉のみに存在
  - 上位下位シソーラス
    - ✦ 単語はシソーラスの内部ノードにも存在

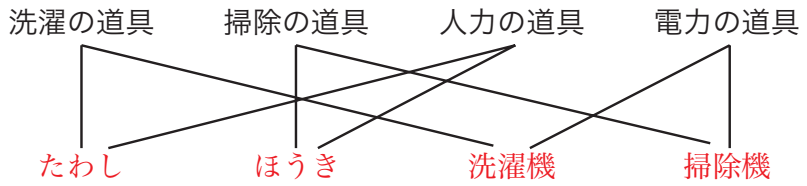
シソーラスの構造

- 一般には木構造だが、グラフ構造のときもある
- 視点
  - 単語の持つ意味の1つの側面
  - 「学校」は建物でもあり、組織でもある
  - 視点が異なれば上位語も異なる
    - ✦ 単語の上位語は一般に複数存在



シソーラスの構造

- 視点を考慮した場合、シソーラスの構造はグラフ構造になる



シソーラスの利用

- シソーラスはどのように使われるか?
- 意味素の辞書として使う
  - 意味クラスは意味素とみなせる
  - 意味素の細かさを柔軟に調整できる
    - ✦ シソーラスが階層構造を持っているため
- 二単語の類似度の計算
  - シソーラス上の距離
  - 共通上位ノードの深さ
  - $\frac{d_c \times 2}{d_i + d_j}$ (d_i, d_j, d_cは、単語i, 単語j, 共通上位ノードcの深さ)

シソーラスの例

- 英語
  - ロジェのシソーラス
  - WordNet
- 日本語
  - 分類語彙表
  - 角川類語辞典
  - 日本語語彙体系
  - EDR概念体系辞書