言語解析論

講師　竹内孔一

本日の内容

- 形態素解析とは
- 形態素解析器のモデル化
  - 隠れマルコフモデル(hidden Markov model)

言語研究の相関

形態素解析

- 目的
  - 入力文字列を最小の意味単位に分割
    → 構成素に分解
  - 時系列データの基本単位を取り出す
    ➔最小のchunking (まとまりを見つけること)
- 例
  - 私の友人は吹奏楽団に所属している
    私/名詞 の/格助詞 友人/名詞 は/格助詞 吹奏楽/名詞
    団/接尾辞 に/格助詞 所属/サ変 して/動詞 いる/助動詞

モデル化

- 2つの重要な点
  定義: 形態素という基準をどう決めるか
  - 最小単位にはいろいろ考えられる
    例) 吹奏|楽団, 吹奏楽|団, 現(会長と副(会長))
    男の子, 火の見|やぐら, (新しい町)作り
    →切るレベルを決め辞書を作成しておく

  モデル: 形態素列のマッチング
  → 入力記号と辞書とのマッチ (音声認識)
  DTW dynamic time warping
  隠れマルコフモデル hidden Markov model (HMM)

単語まわりの考察

- 句, 複合語, 名詞?, 形容詞?
  - どこが単語の区切り?
    例) 訪米, 訪中, 訪韓, , ,　厚さ, 分厚さ
    [国立ないし公立]大学,　[不精な中年男性]用
    現[会長と副会長], 電子メール(ご)使用の際は
    飲み始める/そうし始める　駆け込む/*そうし込む

  - 品詞は?
    例)　積極, 消極, 国際　カチカチに, さらさらに/する

練習8

次の言葉を意味ある最小単位に区切り，その係り関係を示せ

a. 火の見やぐら，がまの油売り
b. 値上げ，時間切れ，肩こり
c. 元同社カメラマン，カーター元大統領
d. テープの頭出し，モーツァルトの墓参り
e. 光男の子

パス

モデル化(一般)

- モデル化とは
  - 対象となる問題を規定
    - 入力，出力はどんな形になるか
  - 必要とする情報源を整理
    - その問題を解くために必要な情報
  - 特徴を選択して数式・手続きを作る　どの特徴に注目したか
    - 統計的手法による学習機能付きのモデル
      - 隠れマルコフモデル(HMM)，SVM
    - 決定的なモデル
      - 人手による制約に基づくモデルなど　DTW
  - テストを行い修正する

形態素解析のモデル化

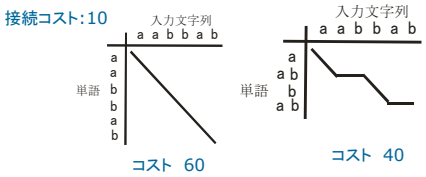
- 問題の設定
  - 入力：文字列，出力：形態素と品詞
- 必要とする情報源
  - 形態素の辞書，品詞，活用形
- とりあげる特徴
  - となりどうしの接続規則
    例）統計(名詞)的(接尾辞)手法(名詞)
    -扱わないこと：入れ子関係，意味的な関係　例 火/の/見/やぐら
- 数式・手続き化
  - dynamic time warping (DTW)
    隠れマルコフモデル hidden Markov model (HMM)

DTW

- 文字列マッチングの基礎
  - 入力文字列に対する単語列をすべて生成してその接続コスト最小のものを選択する

準備　接続コストのルール，辞書，辞書引きシステム

- 例) 入力文字列　aabbab，辞書 {a, ab, b}



DTWのまとめ

- 機能
  - 文字列に対して単語列を提示
- 利点
  - 接続コストにより部分的な情報から計算可能
- 欠点
  - 接続コストを修正する理論が無い
    - 人手で修正するのはかなり困難

→数学的な学習モデルの利用
→マルコフモデル

マルコフモデル

- マルコフモデル

仮定
- ある時系列はその前の時系列の出方によって決まる

式　$P(W) = P(w_1, w_2, \ldots, w_n)$　w_i は単語

$$= P(w_1)P(w_2 \mid w_1) \cdot P(w_3 \mid w_1, w_2) \cdot$$
$$\cdot P(w_n \mid w_1, .., w_{n-1})$$
$$= \prod_{i=1}^{n} P(w_i \mid w_1, .., w_{i-1})$$

小練習：上の式が展開して元に戻るか試してみよう

マルコフモデル

- マルコフ仮定(Markov assumption)
  - $P(w_1, w_2, \ldots, w_n)$ はほとんど確率が0

1重マルコフモデル

1つ前だけを見るモデル

$$P(W) = \prod_{i=1}^{n} P(w_i \mid w_{i-1})$$

2重マルコフモデル

2つ前を見るモデル

$$P(W) = \prod_{i=1}^{n} P(w_i \mid w_{i-2}, w_{i-1})$$

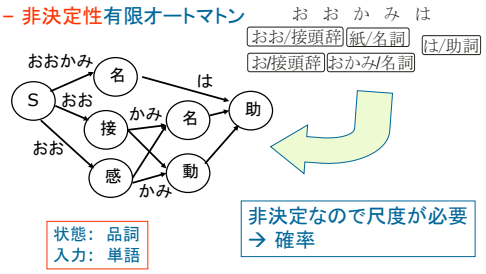
w0,w1,はdummyを利用

音声認識，株価予測など時系列予測に使われる

マルコフ仮定を非決定性有限オートマトンに適用してみる

形態素解析への拡張

- 形態素解析のモデル化
  - 非決定性有限オートマトン

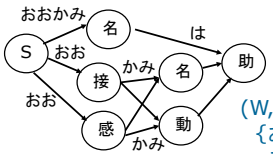


状態：品詞
入力：単語

非決定なので尺度が必要
→ 確率

問題の定式化

- 形態素解析



入力:文字列
出力: 単語W+品詞列S
の組み合わせ
入力: おおかみは
候補:

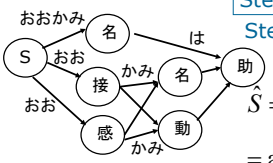
(W, S) =
{おおかみ/名詞 は/助詞,
おお/接頭辞 かみ/名詞 は/助詞,
おお/接頭辞 かみ/動詞 は/助詞,
おお/感動詞 かみ/名詞 は/助詞,
おお/感動詞 かみ/動詞 は/助詞}

この中で最適
パス(組み合わせ列)
を求める

隠れマルコフモデル

- 形態素解析

Step1 単語列を固定して考える
Step2 単語列の組み合わせを考慮



Step1
入力: 単語列W, 状態列:
品詞列S とし 最適なSを求める

$$\hat{S} = \arg\max P(S \mid W)$$

$$= \arg\max \frac{P(S,W) \cdot P(S)}{P(W) \cdot P(S)}$$

Bayesの式

Wは省略できる

$$= \arg\max \frac{P(W \mid S)P(S)}{P(W)}$$

隠れマルコフモデル
hidden Markov model
(HMM) →

$$\overset{def}{=} \arg\max P(W \mid S)P(S)$$

HMMによる日本語形態素解析器

- 形態素解析の定式化

Step2: Wも入力文字列に対して組み合わせを許す

$$P(W,S \mid L) = \frac{P(W,S,L)}{P(L)} \propto \underset{s.t. W \subset L, S \subset W}{P(W,S)}$$

$$= \frac{P(W,S)P(S)}{P(S)} = P(W \mid S)P(S) \qquad st. W \subset L, S \subset W$$

$$(\hat{W}, \hat{S}) = \underset{W \subset L, S \subset W}{\arg\max} P(W \mid S)P(S)$$

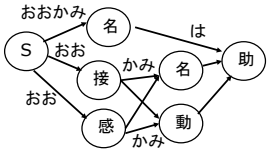
入力文字列L, 単語列W, 品詞列S

HMMを少し応用した形態素解析のモデル

竹内孔一，松本裕治：隠れマルコフモデルによる日本語形態素解析のパラメータ推定，情報処理学会論文誌，Vol.38, No.3, pp. 500–509, 1997.

練習9

- 入力「おおかみは」に対して下図のような曖昧性がある。HMMのL, W, Sを答えよ



$$(\hat{W}, \hat{S}) = \underset{W \subset L, S \subset W}{\arg\max} P(W \mid S)P(S)$$

形態素解析器のモデル化

- HMMによる定式化

$$(\hat{W},\hat{S}) = \arg\max_{W \subset L.S \subset W} P(W \mid S)P(S)$$

品詞列 $S = s_1, s_2, \ldots, s_n$　全列を見るのは無理!!

単語列 $W = w_1, w_2, \ldots, w_n$

マルコフ仮定 (Markov assumption)
1重マルコフモデル(1つ前だけ見よう)

$$P(S) = \prod_{i=1}^{n} P(s_i \mid s_{i-1}) \quad P(W \mid S) = \prod_{i=1}^{n} P(w_i \mid s_i)$$

よって

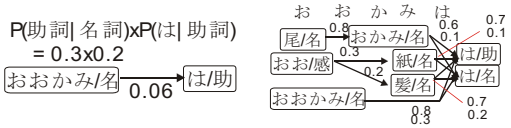
$$(\hat{W},\hat{S}) = \arg\max_{W \subset L.S \subset W} \prod_{i=1}^{n} P(w_i \mid s_i)P(s_i \mid s_{i-1})$$

形態素解析

- 結局どういうモデル化をしたのか

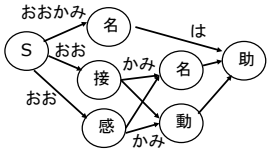
$$(\hat{W},\hat{S}) = \arg\max_{W \subset L.S \subset W} \prod_{i=1}^{n} P(w_i \mid s_i)P(s_i \mid s_{i-1})$$

– 入力文字列に対して
  - 辞書引きしたあらゆる組み合わせのパスを求める
  - その中で確率値が最大のものを選択する



練習10

- 1重のHMMで形態素解析の結果を求めよ
  入力:「おおかみは」



状態遷移確率:
S->名: 0.3
S->接: 0.2
S->感: 0.1
名->助: 0.5
接->名:0.3
接->動: 0.1
感->名: 0.3
感->動: 0.1
動->助: 0.2

単語の確率:
名-おおかみ: 0.1
名-かみ: 0.2
接-おお: 0.3
感-おお: 0.3
助-は: 0.5
動-かみ: 0.1