言語解析論

講師　竹内孔一

本日の内容

- 構文解析
  - 確率的な構文解析
  - 学習
  - 評判情報抽出

確率的文脈自由文法

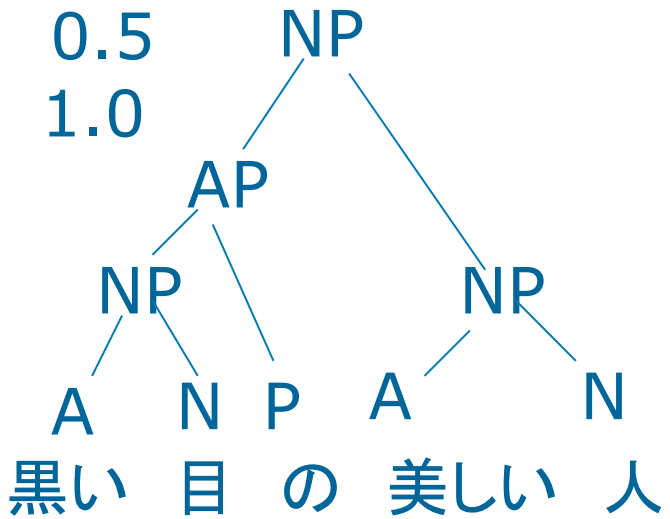
- 文脈自由文法
  - 係り関係を確率で結びつける

NP → A NP 0.2
NP → AP N 0.3
NP → AP NP 0.3
NP → A N 0.2
AP → N P 0.3
AP → AP A 0.4
AP → NP P 0.3

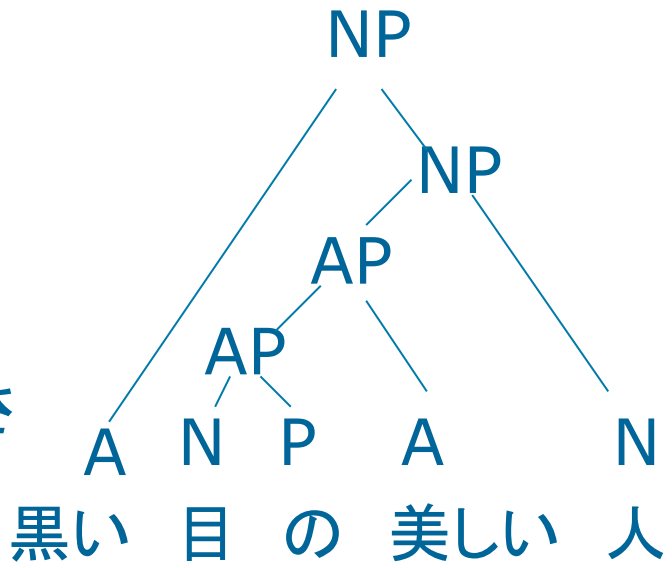
1

A → 黒い 0.5
A → 美しい 0.5
N → 目 0.5
N → 人 0.5
P → の 1.0

計算してみよう



練習15



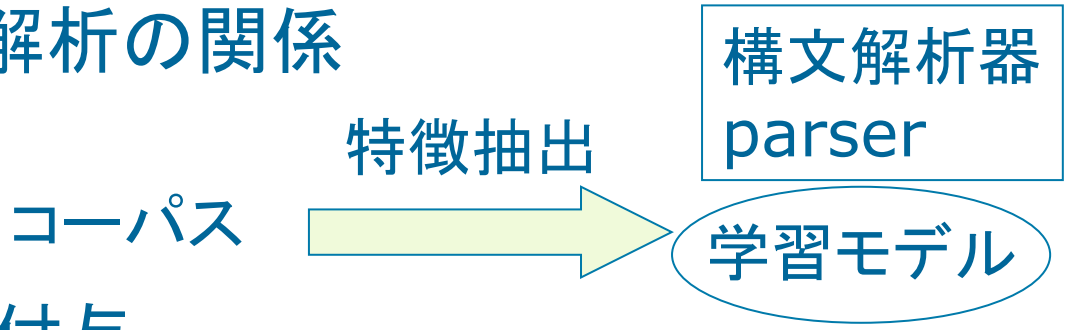
- 先ほどの確率的文法を用いて
  - 解析木の生成確率をもとめないさい
  - 確率値の大きいほうはどちらか

NP → A NP 0.2
NP → AP N 0.3
NP → AP NP 0.3
NP → A N 0.2
AP → N P 0.3
AP → AP A 0.4
AP → NP P 0.3

A → 黒い 0.5
A → 美しい 0.5
N → 目 0.5
N → 人 0.5
P → の 1.0

確率的文脈自由文法の学習

- 学習と解析の関係



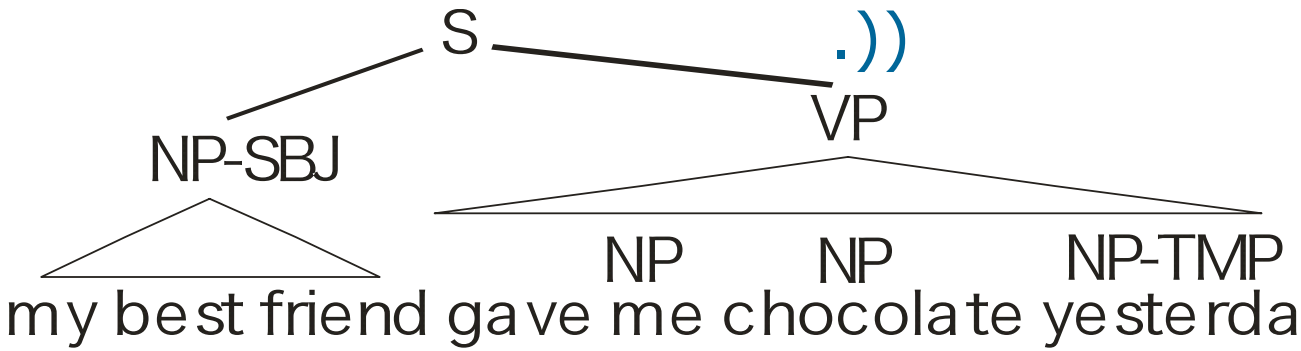
- 確率の付与
  - 統計的な学習による方法
    - 最尤推定
  - 人手で勝手に与える
- 前提
  - 文法セット　{NP->NP VP，，，}は既知
  - ChartParserなどで全組み合わせの構文木を作成

確率的文脈自由文法の学習

- コーパスからの推定
  - 正解つき学習　用意: 解析済みコーパス parsed corpus
  - 正解なし学習　用意: テキストコーパス
- Parsed corpus　例)Penn-tree bank

```
(TOP (S (NP-SBJ my best friend)
        (VP gave
            (NP me)
            (NP chocolate)
            (NP-TMP yesterday))
        .))
```

構文木を1行で書く
→ Lisp の様な形式



最尤推定

- 解析済みコーパスからの数え上げ

$$P(A \rightarrow B) = \frac{C(A \rightarrow B)}{C(A \rightarrow *)}$$

Cはコーパス中の頻度
＊は「すべて」

- 例) 以下のtree-bank で各規則の確率を求めよ

(TOP (S (NP 私　の　友達　が)
(VP くれた
(ADV 昨日)
(NP 私　に)
(NP チョコ　を))
。))

S → NP VP 。 1/1
NP → 私に 1/3
VP → くれた ADV NP NP 1/1
NP → 私の友人が 1/3
….

ここ注意

練習16

- 以下のコーパスがあるときに各規則の確率を求めよ。

(TOP (S (NP 私　の　友達　が)
(VP くれた
(ADV 昨日)
(NP 私　に)
(NP チョコ　を))
。))

(TOP (S (NP 彼　は)
(VP 買ってくれた
(ADV わざわざ)
(NP 私　に)
(NP チョコ　を))
。))

S→ NP VP 。

VP → ……. ??

練習16回答例

(TOP (S (NP 私　の　友達　が)
(VP くれた
(ADV 昨日)
(NP 私　に)
(NP チョコ　を))
。))
(TOP (S (NP 彼　は)
(VP 買ってくれた
(ADV わざわざ)
(NP 私　に)
(NP チョコ　を))
。))

P(S→ NP VP 。)=1/2
P(VP→ くれた ADV NP NP)=1/2
P(NP→ 私 の 友達 が) = 1/6
P(NP→ 彼 は) = 1/6
P(NP→ 私 に) = 2/6
P(NP→ チョコ を) = 2/6
P(ADV→ 昨日) = 1/2
P(ADV→ わざわざ) = 1/2

練習16-1

- 下記のS式を構文木で示せ

(TOP (S (NP 彼　が) (VP (NP-DAT 姉に) (NP-OBJ 本を)
買った) 。))

練習16-2

- 以下のコーパスがあるときに各規則の確率を求めよ。

(TOP (S (NP 彼　が) (VP (NP-DAT 姉に) (NP-OBJ 本を) 買った) 。))

(TOP (S (NP 太郎　が) (VP (NP-DAT 妹に) (NP-OBJ ゲームを) 買った) 。))

(TOP (S (NP 花子　が) (VP (NP-DAT 友人に) (NP-OBJ ゲームを) もらった) 。))

実用例

- 題: Webからある製品の評判情報を取り出す

  例） 対象: Web上のblog

  準備: 評価に関する述語を整理

  ～良い(+3), いい(+2), ちょっとね(-2), 微妙 (-1)

  解析: parser で構文木を作成

  - 主語(製品)と理由と述語(評価)をとるパターンを用意

  pattern: 主語[]a +のが[]b+述語[]c

  入力 **A社のBは〇〇が使えないのがちょっとね**

  →(**A社のBは) (((Cが使えない)の)が) (ちょっとね)**

  →抽出 B, -2, 理由:cが使えないから

練習17

- システムの構築について
  - 構文解析を使ってipod touch に関する評価をWebから集めるためのパターンを作成せよ

前提: 精度の高い構文解析器が利用できるとする

構文解析の現状

- 改善
  - 規則をより正確に記述 HPSG, LFG
- フリーソフト
  - KNP　京都先生
    http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html
  - CaboCha 奈良先端大
    http://chasen.org/~taku/software/cabocha/
- 精度
  - 80%以下
- 現在の問題点
  - 精度の向上が難しい
  - どう利用できるかという部分が発展途上
    → さきほどのパターンの構築の難しさ