自然言語処理論 I

11. テキスト処理 (コーパスの処理)

コーパス(corpus)とは?

- 例文集
  - 実際に使用されている例文を大量に集めたもの
- テキストコーパス ⇦
  - 文章を集めたもの
  - 新聞記事、雑誌、小説、辞書など
- 音声コーパス
  - 音声データを集めたもの
  - 対話、インタビュー、講演など

量から質へ

- 昔
  - 機械可読データを集めること自体が困難
  - コーパスの量が重視された
- 今
  - 電子化文書の普及
    - ✦ 新聞記事、レポート、ウェブ文書、blog
  - コーパスにどのような付加的な情報をつけるかということに重点が置かれている
  - 注釈付きコーパス(annotated corpus)

注釈付きコーパス

- コーパスに情報を付加したコーパス
- 主なもの
  - (平文コーパス)
  - 品詞タグ付きコーパス
    - ✦ 単語の品詞
    - ✦ 単語境界
  - 構文構造付きコーパス
    - ✦ 文の構文木
    - ✦ 文節の係り受け関係

注釈付きコーパス

- 主なもの
  - 語義タグ付きコーパス
    - ✦ 単語の語義
  - パラレルコーパス
    - ✦ 複数言語の対訳テキストデータ

      メアリーは朝食を食べて、学校へ行った。
      Mary ate breakfast and left for school.

    - ✦ 文と文の対応関係
    - ✦ 句、単語の対応関係

コーパスの例

- 品詞タグ付きコーパス
  - Brown Corpus
    - ✦ 100万語、英語
  - British National Corpus (BNC)
    - ✦ 1億語、英語

コーパスの例

- 構文構造付きコーパス
  - Penn Treebank
    - ✦ 110万語、英語
    - ✦ 品詞
    - ✦ 構文木
  - 京大コーパス
    - ✦ 100万語、日本語
    - ✦ 品詞
    - ✦ 文節の係り受け解析結果

コーパスの例

- 語義タグ付きコーパス
  - SEMCORコーパス
    - ✦ 英語
    - ✦ WordNet の意味クラス
  - RWCコーパス
    - ✦ 90万語、新聞記事3000個、日本語
    - ✦ 岩波国語辞典の語義
    - ✦ 品詞タグの付与
  - EDRコーパス
    - ✦ 500万語、20万文、日本語
    - ✦ EDR概念体系の語義
    - ✦ 品詞タグ、構文木も付与

注釈付きコーパスの作成

- 自動的に作る
  - エラーが多く含まれる
  - 現時点での自然言語処理技術では良い品質の注釈付きコーパスが得られない
- 人手で作る
  - 全ての情報を人間が与える ×
  - 自動解析(形態素解析、構文解析、意味解析)した結果を見て、誤りを修正する

注釈付きコーパスの作成

- どのような点が困難か?
  - 大量の文に情報を付加することが要求される
    - 人間の負担が大きくなる
    - 時間・費用がかかる
  - どのような情報を付加するかが難しい場合がある
    - 例外的な事象に対する情報の付与
  - 一貫性の問題
    - 複数の作業者による共同作業
    - 作業者による揺れの問題

コーパスの利用

- パラメータの計算
- コーパスの直接利用
  - 用例に基づく自然言語処理
- コーパスからの学習
- 自然言語処理システムの評価

パラメータの計算

- 解析における解の優先規則(確率モデル)のパラメータを学習する
- 形態素解析の場合
  - コスト最小法におけるコストの学習
    - 品詞間のコスト
    - 単語のコスト
  - 隠れマルコフモデルのパラメータの推定
    - $\prod_i P(C_i|C_{i-1}) \prod_i P(w_i|C_i)$

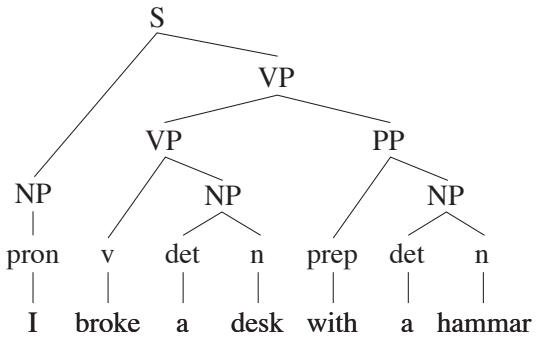
パラメータの計算

- 構文解析の場合
  - 確率文脈自由文法
    - ✦ Probabilistic Context Free Grammar (PCFG)
  - 規則 A → α_i の適用確率: $P(\alpha_i|A)$
    但し、$\sum_i P(\alpha_i|A) = 1$
  - 具体例

VP → v	0.3
VP → v NP	0.5
VP → VP PP	0.2

PCFG

- 構文木の生成確率
  - 構文木中に使われている規則の確率の積



S → NP VP	1
NP → pron	0.2
VP → VP PP	0.2
VP → v NP	0.5
NP → det n	0.8
PP → prep NP	1
NP → det n	0.8

(0.0128)

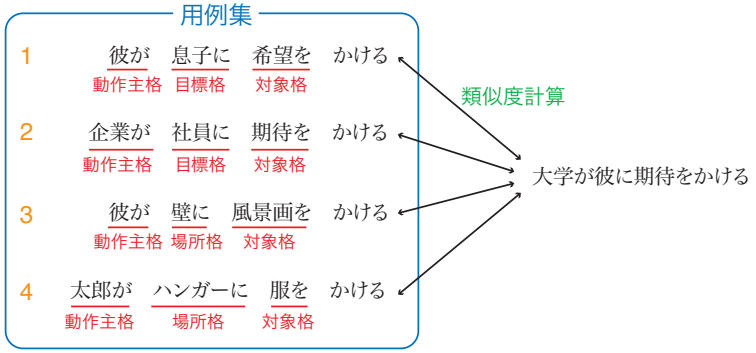
PCFG

- パラメータ(規則の確率)の推定
  - 構文木付きコーパスから推定可能
  - $P(\alpha_i|A) = \dfrac{O(A \to \alpha_i)}{\sum_i O(A \to \alpha_i)}$

✦ $O(A \to \alpha_i)$: コーパスにおける規則の出現頻度

コーパスの直接利用

- 用例に基づく格解析



- 用例に基づく機械翻訳
  - パラレルコーパスの利用

コーパスからの学習

- 様々な自然言語処理用知識の自動獲得
  - 文法
  - 格フレーム辞書
  - シソーラス
- 人間が作った知識と比べると
  - コストははるかに安い
  - 品質は悪い
    - ✦ 誤りも多く含まれる
  - 大規模な知識を容易に作ることができる

自然言語処理システムの評価

- 形態素解析システムや構文解析システムの正解率を調べる
  - コーパスに付加された情報を除いて解析
  - 解析結果とコーパスの情報がどれだけ一致しているかを調べる
  - システムの改良の結果がすぐ分かる
  - 同じコーパスで評価→システムの公平な評価
- よく使われるコーパス
  - Penn Treebank
  - 京大コーパス

まとめ

- 注釈付きコーパス
  - 形態素情報、構文情報、意味情報
- 人手で作成するのは多大な労力が必要だが、それだけ価値も高い
- 用途
  - パラメータの計算
  - 直接利用
  - 知識の自動獲得
  - 自然言語処理システムの評価