B4ゼミ2014/2/20(第1週)

コーパスからのコロケーション情報抽出

--分析手法の検討とコロケーション辞典項目の試作--
(論文紹介)

長岡技術科学大学　電気系
自然言語処理研究室　高橋寛治

はじめに

- 論文を読んできて、紹介
- コロケーションについての考え方を中心に紹介

- 言語学の論文ということに注意

- 論文
  - 田野村忠温(2009)「コーパスからのコロケーション情報抽出：分析手法の検討とコロケーション辞典項目の試作」『阪大日本語研究』,21, 21-41.

コロケーションとは？

コロケーション

- **よく使われる組み合わせ、自然な語のつながり**
  - 例　「辞書」
    - 自　然:「辞書を引く」、「辞書で調べる」、「分厚い辞書」
    - 不自然:「辞書を読む」、「太い辞書」
- **言語表現全般（語、語の連続、句、節など）のあいだに観察される習慣的な共起関係**

コロケーションの用途

- **語義の精密な分析・記述のための考察材料**
  - 類義的な表現の意味の差を考えるとき、どのような文脈でよく生起するか観察する
- **辞書の編集、外国語の教育など実用的な分野**
  - 辞書編集　語義の記述、例文の選択に役立つ

コロケーションに関する情報は有用？

- **文法の知識から**
  - 分かること
    - 名詞には格助詞、動詞には助動詞や接続助詞が後接する
  - 分からないこと
    - ある動詞はほどんど常に否定の文脈で使われる、受け身でよく使われる
- **語義の知識から**
  - 分かること
    - 「バイオリン」「フルート」→「弾く」「吹く」
  - 分からないこと
    - 「シンバル」→「弾く」「吹く」??
    - 万全ではない

コロケーションに情報は有用？

- **何を表現する機会が多いかという事実の反映にしか過ぎない情報も価値が乏しい**
  - 「飲む」
    - 「飲む」の目的語にどんな飲み物が多く現れるかが明らかになってもしかたない
    - 実際の場面では、発話意図に即した表現を選ぶ必要がある
  - 「飲む」
    - 「飲む」の対象は、「スープを飲む」「薬を飲む」「息を飲む」

コロケーション情報をいかに獲得？

コロケーション情報をいかに獲得？

- **正確なコロケーション情報を得るには？**
  - 大量の用例（大規模コーパス）を収集して分析
  - 内省は正確なコロケーション情報獲得手段として絶望的に非力
  - 紙媒体の言語資料の手作業による分析は時間と労力の点で非現実的

- **答えはない**
  - 日本語コーパスに対し、どのような処理を施すか？

コロケーション情報をいかに獲得？

- **筆者の認識**
  - コーパスから獲得できる情報は、我々が真に得たいコロケーション情報のごく粗い近似
    - コーパスの調査は表面的な文字レベルの分析
  - コーパスを入力として与えれば完成した形の分析結果を出力してくれるソフトウェアを作ることは不可能
    - 有用な情報をより分けるには「人手」に頼る
  - コロケーション情報の評価では、統計的な厳密さを重視する意味は乏しい
    - コロケーションは理論的な価値の乏しい概念
    - ひと手による情報の取捨を前提としている

コロケーション情報をいかに獲得？

- **コーパスから得られるコロケーション情報抽出が資料ごとに異なることを示す単純な例**
- **「西瓜」**
  - Webコーパス
    - 食べる 10315、買う 2152、ある 1528、切る 1095、なる 1092
  - Yahoo!知恵袋ベータ版データ
    - 食べる 275、出る 80、取りこぼす 71、出す 65、引く 57
  - 朝日新聞記事11年分
    - 食べる 32、作る 10、切る 8、売る 8、割る 6
  - BCCWJ2008の書籍データ
    - 食べる 8、盗む 7、切る 5、持つ 4、できる 3

使用するコーパス

使用するコーパス

- **大規模なものが必要　→　Webコーパスを主**
  - 約150GB（約750億文字）
  - その一部約10GB（約50億文字分）を使用
- **なぜ一部を利用？**
  - すべての用例数が3桁以上を望む（筆者の感覚）
  - ハードウェアの性能??

コーパスからの
コロケーション情報の抽出

共起語分析

- **分析の基本方針**
  - コーパスにおいて所与の語句の近傍にどのような語ないし形態素がよく現れるか
- **分析手順の検討**
  - 調査対象とする語句とその共起語の関係を限定
    - 例　所与の名詞にどのような動詞が後続するか

共起語分析
ー所与の名詞に後続する動詞ー

- **分析事項と考察**
  - 内省だけでは容易に知り得ない情報を獲得
    - 才能に恵まれる、能力を高める
  - 名詞と動詞の意味・表現上の関係が動詞ごとに異なる
    - 「熱意**が**伝わる」
    - 「熱意**を**感じる」
    - 「熱意**に**負ける」
    - 「熱意**に**あふれる」

共起語分析
ー所与の動詞に先行する名詞ー

- **動詞「任せる」に先行する名詞**
  - 「身」「仕事」「流れ」「他人」

- **助詞でみると**
  - 「身**を**任せる」　「仕事**を**任せる」
  - 「流れ**に**任せる」「他人**に**任せる」

- **助詞との組み合わせを考慮に入れることが必要**

評価

- コロケーションを単純な2単語間の関係として調査･分析する方法には限界がある
  - 「健康 に 気 を 付ける」「健康 に 害 を 及ぼす」
  - 「健康 に 気 を 及ぼす」「健康 に 害 を 付ける」

- これらの問題を解決するために、共起分析に代わる分析方法の可能性を検討する必要がある

別の手法の分析

- **共起語連鎖分析**
  - 所与の語句とよく共起する語連鎖への着目
  - 語連鎖、語レベルのN-gramを使用

- **共起文字連鎖分析**
  - 所与の語句とよく共起する文字連鎖への着目
  - 文字連鎖、文字レベルのN-gramを使用

日本語コロケーション辞典

日本語コロケーション辞典

- **制限　せいげん**
  - 1名詞
    - 「制限」＋動詞
      - ーを設ける、ーを緩和する、ーを解除する
        - 多くの国は外国人の長期滞在に関していろいろと制限を設けている
  - 2 複合サ変動詞語幹
    - 自由をーする、権利をーする
      - この法案は個人の自由・権利を不当に制限するものだ
    - 理由をーする、行動をーする、活動をーする
      - 混雑時はご入場を制限することがあります／輸入を制限することによって自国の産業を保護している国は多い

おわりに

おわりに

- 精密化や効率化を図る余地
- まったく異なる着想に基づく手法の可能性
- 大規模なコーパスは言語の仕様に関わる統計的事実を圧倒的な説得力を持って我々の前に提示してくれる

高橋の所感

言語学の分野の人が
かゆいと思っているところに
手が届くよう、自然言語処理を
していかなければいけない