B3ゼミ 2014/1/30（第4週）
コーパスについて

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

- 「ベーシックコーパス言語学」の本にそって勉強資料を作成
  - 石川慎一郎　著
  - 応用言語学

- 2回のゼミで、終了予定

- 自分が自然言語処理をする側ということを忘れない

はじめに

- 「コーパス」って自然言語を扱う分野でよく聞くけど・・・・
- 定義は?どうやって決めてるの?本質は?
- 扱う上で分かっておかなければいけない

このスライドはコーパスを理解して扱うための学習の補助

第2章　コーパスとはなにか

- コーパス･･･corpus(複数形corpora)
  - ざっくりと「言語のデータベース」を指す
  - 自然な言語を代表する標本
  - 言語の特性を反映している
    - 基準にしたがって作られる
    - 実例を集めている
    - 電子化されている（用意に検索可）
    - ・・・・
- コーパスとはどのようなものか考える

2.1　コーパスの5つの観点

コーパスを考える上での観点

1.書き言葉や話し言葉などの現実の言語を

2.大規模に

3.基準に沿って網羅的･代表的に収集し

4.コンピュータ上で処理できるデータとして

5.言語研究に使用するもの

これらを踏まえてコーパスの様子からコーパスを知る

2.2　コーパスとは?

- 対象
  - 原則　「自然言語」

- 実例と作例
  - 実例:現実のコミュニケーションで使われた文例
  - 作例:辞書の用例など人為的に用意された文例
  - コーパスでは「実例」を利用　→　「本物の言語」

2.2　コーパスとは？

- 書き言葉と話し言葉
  - どちらもコーパスの対象
  - 書き言葉
    - 書籍、新聞　おもに公刊物からデータを収集
  - 話し言葉
    - 映画の台本、自発音声　話しで使用される言葉を収集

日本語書き言葉コーパスの紹介

2.2　データの規模

- コーパスにおけるデータの規模
- 「大規模」
  - 少なくとも100万語
    - 1億語程度？
- なぜ規模が必要？
  - ある語が使われていることの実在証明
    - ある語が使われないという不在証明にはならない
      - （使用されないのか、たまたま未出現なのか分からない）

1964年　Brown Corpus　100万語

1994年　British National Corpus 1億語

2.2 データの規模

- これからは・・・
  - 1億語を遥かに超える「超大規模」コーパス
  - ウェブ上のデータを利用して、コーパスを自動構築
- 「大きければ大きいほど良い」という立場が主流
- 統計学より
  - コーパスサイズが100倍になれば頻度の精度は10倍向上

2.2 データの規模

- でもコーパスがあまりに大きくなると？
  - 例　enTenTenでlargeについて検索
    - 80万件近い用例が得られるが、すべてを細かく調査不可
- 研究目的ごとに最適なコーパスサイズがあるとする立場がある
- 大は小を兼ねる??

2.2 データの規模

- コーパスの規模を論じる際の留意点

- 規模について
  - 例　日本人中学生の英作文コーパス
    - 10万語でも大規模と言えるのではないか？
- 語数について
  - 語認定（tokenization）
    - 英　語　例　isn't 1語、is n't 2語
    - 日本語　例　私は 2語、私は 1語

2.2.3 データの収集方法

- コーパスの原則
  - 網羅的で代表的なものである
- コーパスデータ収集の方法論
  - 悉皆（しっかい）的収集法
    - 収集対象となるデータの全体（母集団）を漏れ無く収集
  - 均衡的収集法
    - 母集団から一定の基準に基づく取捨選択を行い、母集団のさまざまな側面や特性が均衡的に再現されるようデータを収集
  - 大規模収集法
    - 入手しやすいデータをできる限り広範にかつ大規模に収集し巨大コーパスをつくる

2.2.3 データの収集方法

- 利点で見比べると?
  - 悉皆的収集法
    - より大きな母集団を扱いやすく、構築コスト小
    - サイズが一定範囲に収まるため、容易に検索可

      均衡的手法と比べて
    - 英語や日本語といった個別言語全体を母集団とする場合悉皆的収集法を採用できない。

2.2.3 データの収集方法

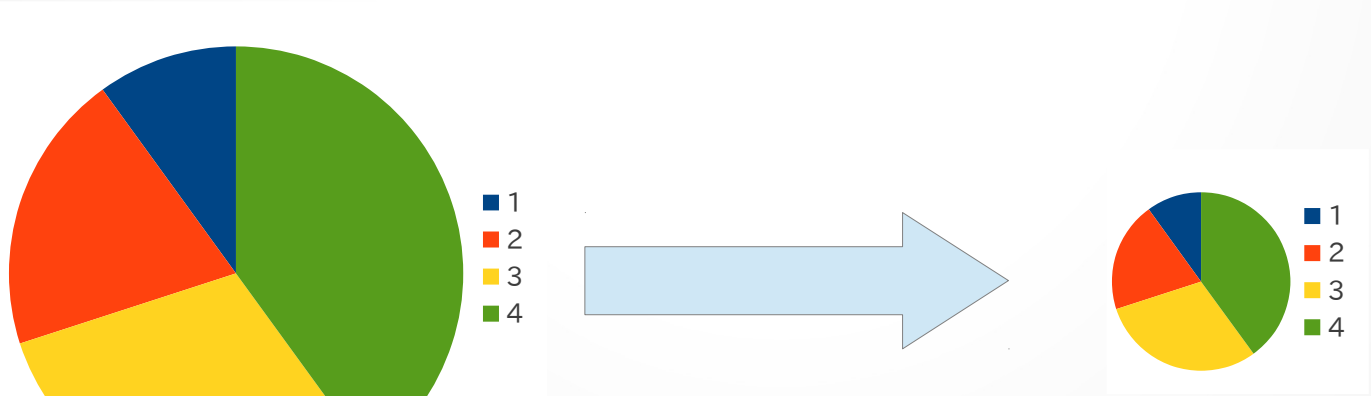
- 利点で見比べると?
  - 大規模収集法
    - 母集団との連関性が強い、コーパスで見つけた知見を母集団に還元して解釈しやすい
    - 無計画にデータを収集したコーパスに含まれる事実は偶然かもしれない
      - 例） コーパスデータの大半が特定の雑誌からなら、得られた結果が言語全般の特徴なのか、当該雑誌の特徴なのか不明
    - 均衡的収集法に基づいて構築されたコーパスでは、観察結果の信頼度が高まる

2.2.3 データの収集方法

- 均衡的収集法の実際
  - 層化抽出と無作為抽出を組み合わせて使用
  - 層化抽出
    - 母集団を複数の層に分け、層ごとにデータを収集すること
  - 無作為抽出
    - 作為や意図が全く働かない条件のもとでデータを選んで抜き出す

2.2.3 データの収集方法

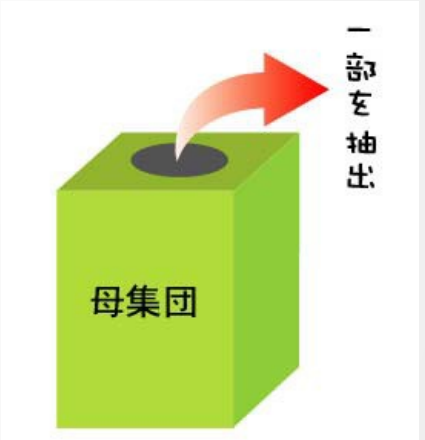
- 層化抽出
  - 母集団を複数の層に分け、層ごとにデータを収集
  - 例　母集団：日本語の書き言葉
    - 第1層：新聞・雑誌・書籍・論文・書簡などのジャンル分け
    - 第2層：新聞ー（全国紙・地域紙・地方紙、朝刊・夕刊・・・）



層化抽出法の標本抽出イメージ
全体のサイズが圧縮され、要素の比率が保持

2.2.3 データの収集方法

- 無作為抽出
  - 作為や意図が全く働かない条件のもとでデータを選んで抜き出す
    - 例　袋に入っている玉の中から目隠し状態で1つ取り出す
  - 書籍から標本をとる場合
    - 出版目録などすべての書籍に通し番号をふる
    - 乱数表などを使用しデータを取る書籍を選ぶ
    - どのページからデータを取るか、ページ内のどの場所から収集を開始するか乱数表から決定する



2.2.3 データの収集方法

- 均衡的収集法の限界
  - 母集団についての問いの唯一の解はない

    総語数はいくつ？　書き言葉と話し言葉の比率は？

    - 現代アメリカ英語　Brown Corpus

      ほんとに現代アメリカ英語？議会図書館の収録目録？

  - 母集団としての言語の不定性のなかにある、本質的な制約を意識し、方法論をさまざまに工夫しなければいけない

おわり

- 今回はこれで終わりです

おつかれさまでした