言語データのコーディング入門

コーパスに潜むどんな情報を,いかにして探り出すか

黒田　航

情報通信研究機構 けいはんな情報通信融合研究センター

08/08/2005

山梨研究室の院生を中心とした「統計勉強会」の下準備として

本日のお品書き

- コーパスの利用価値について
- 用法基盤の考え方について
- コーディングについて
- 観察の実践例

Co rp us Lin guis tics a nd B ey ond
Fr om Na ïv e, Int ui ti on-bas ed C ogn
it ive Li ngu istics to
Data-driven, Truly U sag e-b ase d Co
gn it ive Lin guis ti cs

コーパスの利用価値

コーパスを利用する意味

- 事例を網羅的，体系的に集めることで，観察レベルでのバイアスが回避可能
  - 作例中心の研究は観察のレベルでのバイアスを避けがたい
  - これは生成言語学で顕著だが，認知言語学でも是正されているわけではない
- ただしコーパスにはそれ自体の限界もある

コーパス利用の限界

- あらゆる可能性がコーパス事例で実現されているわけではない
  - 偶発的な実現バイアス，体系的な実現バイアス
  - 後者は使用域 (registers) の問題
- 実現されていない可能性を直観によって補うのは方法論的に誤りというワケでない
  - ただ統計解析に固執すると，この種の補完は禁じ手
  - 内観法を完全に排斥するのは本末転倒

二つのデータ定義の流儀

- 直観基盤 (intuition-based) の内観的 (introspective) 記述モデル=アプローチ
  - 作例が記述の対象; 実際の使用例は参考値
- 観察基盤 (observation-based) の外観的 (extrospective) 記述モデル=アプローチ
  - 実際の使用例が記述の対象; 作例は参考値
- 重要な点
  - これらのアプローチは本来は相補的なもので目的によって使いわける必要がある

いわゆる“コーパス言語学”の限界

- BNC コーパスを KWIC 検索して何千という実例が得られました．さて，どうする?
  - 実例を漠然と眺めているだけで何かがわかるワケではないので，何かしなければならない
  - コーパス言語学の常套手段は「とにかく数える」
- 問題
  - 頻度を調べるのはいいが，問題は何の頻度を，何のために数えるのか
  - これはコーパス言語学では真剣に考慮されていない

頻度情報の有用性

- 頻度は単に有用な指標でしかない
  - いわゆるコーパス言語学は頻度情報に依存しすぎで，頻度を調べられないものに関しては何も語れない
- 言語情報は頻度情報に還元できない
  - 数を数え(頻度を調べ)て意味があるのは，頻度に反映する情報のみ
- 頻度のような統計量の解析は重要だが，何の頻度を数えると何がわかるのかが問題

タイプとトークン

- 数えられるのは常に何らかのタイプ T のトークン t (= 実現値) の数
  - タイプが決まらないと数えることは無意味
- 例
  - 生コーパスを利用して数えられるのは文字列 (e.g., 襲 {い,う,え,っ}) をタイプとするトークンのみ
  - 品詞タグを利用して数えられるのは品詞列をタイプとするトークン (e.g., T = V; t = {襲い, 歌い, ...})

日本語の形態素解析の例

形態素	読み	見出し語	品詞-下位分類	付加情報1	付加情報2
日本語	ニホンゴ	日本語	名詞-一般		
の	ノ	の	助詞-連体化		
統語	トウゴ	統語	名詞-一般		
関係	カンケイ	関係	名詞-サ変接続		
解析	カイセキ	解析	名詞-サ変接続		
は	ハ	は	助詞-係助詞		
英語	エイゴ	英語	名詞-一般		
ほど	ホド	ほど	助詞-副助詞		
簡単	カンタン	簡単	名詞-形容動詞語幹		
じゃ	ジャ	じゃ	助詞-副助詞		
ない	ナイ	ない	助動詞	特殊・ナイ	基本形
.	.	.	記号-句点		
いや	イヤ	いや	接続詞		
,	,	,	記号-読点		
英語	エイゴ	英語	名詞-一般		
の	ノ	の	助詞-連体化		
解析	カイセキ	解析	名詞-サ変接続		
が	ガ	が	助詞-格助詞-一般		
簡単	カンタン	簡単	名詞-形容動詞語幹		
だ	ダ	だ	助動詞	特殊・ダ	基本形
って	ッテ	って	助詞-格助詞-連語		
意味	イミ	意味	名詞-サ変接続		
じゃ	ジャ	じゃ	助詞-副助詞		
ない	ナイ	ない	助動詞	特殊・ナイ	基本形
けど	ケド	けど	助詞-接続助詞		
.	.	.	記号-句点		
EOS					

- 形態素解析プログラム茶筅 (Chasen) による形態素解析 (morphological analysis) = トークン化 (tokenization) の結果
- 入力は
  - 日本語の統語関係解析は英語ほど簡単じゃない．いや，英語の解析が簡単だって意味じゃないけど．

品詞のタイプとトークン

- 名詞, 動詞, 助詞, 助動詞, 接続詞のような品詞はタイプで，それらのトークンは次のような語
  - 名詞: {日本語, 統語, 関係, 解析, ...}
  - 動詞: {関係する, 解析する, 疑う, ...}
  - 助詞: {の, は, が, ...}
  - 助動詞: {ない, だ, ...}

品詞タグつきコーパスは不可欠?

- 品詞タグつきコーパスがなくても，例えば次のコマンドで一文ごとに処理可能
  - % chasen sample-j.txt | nkf -s > sample-j.chasen-out.txt
  - chasen と nkf (Network Kanji Filter) の組み合わせ
    - Excel は SJIS コードしか読まないので，EUC-JP コード(Chasenの入力フォーマット) と SJIS (Excel の入力フォーマット) の仲介に nkf が必要
- ただし茶筅や寿満 (juman) の解析精度は100%ではない

コーパスの利用価値

- アノテーションなしの生コーパスはそれほど有用ではない
  - 十分な量のデータを収集することは必要だが，集めたデータを漠然と眺めるだけでは十分ではない
  - 特にコーパスに意味の生態を探りたい場合，語の頻度分析はそれほど有効ではない
- 重要な情報のみを，研究に使える形でうまく取り出す必要がある

意味の実態を探る際の困難

- 語の意味は見えない
  - コトバの意味の一部はイメージかも知れないが全部がそうだといワケではないし，イメージがコトバの意味の重要な部分だというワケでもない
  - これは認知言語学の主張の一つだが，その妥当性は鵜呑みされるべきでなく，実証的に示されるべき
- 語の意味は暗黙的
  - 語の意味は非言語的情報で与えられる

意味は(まだ)統計処理できない

- 意味のタイプが特定されていないコーパスを使っても意味のトークンは数えられない
  - 現時点で利用可能なほとんどのコーパスでは意味タイプはタグづけされていない
    - 例外: SemCor 1.6, 1.7, 1.7.1, 2.0 (http://www.cs.unt.edu/~rada/downloads.html)
  - 茶筅が品詞タグづけを実行してくれるように，意味タグづけを実行してくれるプログラムは存在しない
    - 自動意味タグづけは開発が進んでいるが，意味タグ体系 (WordNet, VerbNet, FrameNet) に依存

意味のタイプは見つけるもの

- 意味のタイプは研究者が自分で見つけ, 定義するもの
  - WordNet, VerbNet, FrameNet の意味のタイプは参考程度に使えばよい
  - これは品詞でも同じことだが，意味のタイプは品詞に較べて桁違いに多い
  - はじめは一般性を狙わず，自分の研究の目的にあったものを定義し，拡張してゆけばよい
- でも，どうやって見つけ, 定義するのか?

いかにして意味を認定するか

- 実はこれがなかなかの難題
  - 意味がどう認定されているかは，言語学の教科書では自明と見なされ，真剣に議論されることが少ない
  - 自明のことだと思われていることは，しばしばもっとも説明困難
- 私の提案する方法論
  - コーパスから得られた事例を網羅的にコーディングし，それを通じて帰納的に意味のタイプを発見しよう

まずは意味タイプの認定から

- タイプというのはトークンの一般化なので
- 意味のトークンからタイプへの一般化が必要
  - 品詞の体系も所与のものではなく，長年の研究を通じて発見的/帰納的に構築されたものである点に注意
  - 意味型の体系化はシソーラス (日本語語彙大系, WordNet) という形で与えられている
  - ただし意味役割の体系化は従来のシソーラスでは実現されていない

意味タイプの可視化が必要

- 「コーパスの海に意味の生態を探る」となったら，
  - 意味のトークンを見つけ，それらのタイプを定義する手法が必要
  - その作業を支援するために，意味のタイプとトークンの可視化に技法が必要

U sage-ba sed m o del of la ngu age
Where do int ui tions come from?
H ow to de sc ribe la nguage usage?
Aga inst the “P lato nist” vie w of lan guage
emp irici sm and
data
obs erva tio n

用法基盤の考え方

用法基盤のアプローチ

- 認知言語学では今までのところコーパス事例の分析に基づいた説明は見られない
  - 用法基盤主義は現在のところ完全にカケ声倒れ
- 例外
  - Gries, S. Th. 2003. Multifactorial Analysis in Corpus Linguistics. Continuum.
  - Stefanowitsch, A. & Gries, S. Th. 2003. “Collostructions.” International J. of Corpus Linguistics 8 (2), 209-243.
  - Barlow, M. and Kemmer, S, eds. 2000. Usage-based Models of Language. CSLI.

なぜ用法基盤主義が広まらないか

- 多くの言語学者は手抜き言語学が好き
- 用法基盤のモデルの根本的問題点
  - そもそも用法 (usages) の定義がないじゃん!!
  - usage event って図に書けばいいってもんじゃない!
  - 用法という概念に明示的な操作的定義を与える必要
- 私の提案
  - トークンの一般化の意味のタイプが用法の粗い近似と，ボトムアップに用法を定義したらどうか?
- 以下の話はこの案の具体化

W hy c ode d ata an d how?
Why us age-ba sed ap pro ach?
W hy Chom skian li nguists are b li nd to dat a?
Is ling uis tics an em piri cal scie nce, an d if so, how is it?

コーディングについて

コーディングは何のために?

- 生データはそのままでは何も語らない
  - データは解釈されて初めて意味をもつ
  - ただしデータの解釈は可能な限り研究者の恣意を排除した形でなされる必要がある
  - 妥当な解釈を保証するためには妥当な加工が必要
- 恣意性を避けながらデータの特徴を “浮き彫り” にする加工法がコーディング
  - コーディングはデータに内在する構造の可視化を支援

語の意味と用法との関係

- 作業仮説
  - 意味はトークンとしての語にではなく，タイプとしての語の用法に現われる
  - 用法をうまく特定できれば，それに基づいて意味のタイプを特定できる
- 基本はこれでよいとして，この案をどう実装する?

用法への生態学的アプローチ

- 有益なアナロジー
  - コーパスを生態系 (例えば海) に
  - 用法 = 意味を生物種に見立て
  - 用法 = 意味を発見, 観察, 記述する方法を考える
- 注意
  - これはアナロジーだが単なるアナロジーではない

生態系のアナロジーの帰結

- コーパス事例の検討 (外観) = 野外での観察
  - 対象の自然状態での自然な挙動が観察可能
  - 挙動の説明要因は統制困難
- 作例の検討 (内観) = 実験室での観察
  - 挙動の説明要因の統制が可能
  - 観察は自然な挙動を反映したものとは限らない
- 内観か外観かの二者択一ではなくて，両者の上手な組みあわせが必要

記述の前にまず観察

観察について

- 観察とは何をどうすることか
  - 言語学では観察が何であるかの説明がない
- 観察とはデータが値となる特徴の明示
- 特徴の例
  - 生物個体の {体長, 体重, 体色, 行動パターン, ...}
    - 体長-体重, 体色と雌雄の相関
    - 行動パターンと性別(雌雄)の相関

観察の内実

- 対象/現象 x の(科学的な意味での)観察とは x の属性/値マトリックスの値を可能な限り詳細に埋めること
- ある対象について思いつきをあれこれ述べることが観察ではない!
  - いわゆる「記述」言語学の記述は科学的な意味での記述のレベルには達していないことが多い

典型的な生態学的記述

- ハチの記述
  - http://www.insects.jp/konbunmakuhati2.htm
  - これはハチの種の記述で，感じとしては品詞分類に相当する記述
- 個体の特徴コーディングがこれに先行する
  - 種別は個体の特徴コーディングを通じて可能となる

架空の個体の特徴マーキング

個体 ID	種	雌雄	体長 (mm)	体重 (g)	体色
16	C	M	13	65	暗褐色
7	B	M	21	63	褐色
3	B	F	20	60	緑
12	B	M	20	60	褐色
4	A	M	19	57	緑
14	A	M	19	57	緑
2	A	M	18	54	緑
9	A	M	17	51	緑
5	C	F	10	50	暗褐色
15	C	F	10	50	暗褐色
6	C	F	9	45	暗褐色
11	C	F	9	45	暗褐色
10	C	F	8	40	暗褐色
1	A	F	13	39	褐色
8	A	F	12	36	褐色
13	A	F	11	33	褐色

これを見て，あなたは何を読み取れる?

架空の観察データ

- 架空の観察データから読み取れるもの=記述できるものは，例えば
  - A, C 種は普通種，B 種は大型の希少種
  - C 種は雌雄の別によらず体色は暗褐色
  - A 種の雄は褐色，雌は緑色
  - C 種では雌雄の大きさの差が大きい
  - どの種でも雄よりも雌の方が大きい

観察は技能である

- 観察はしばしば特殊な技術や装置が必要
  - 観察は訓練なしに体得できる技能ではない
- 観察の技法の例
  - 染色法 (特徴マーキングの一種)
    - 肉眼では困難な染色体の識別を可能にする
  - 認識札=タグづけや発信器による個体 (ID) マーキング
    - 識別困難な状態にある個体の認識
    - 行動範囲が広い生物 (e.g., クジラ) の追跡調査

コーディング = マーキング

- 極論すれば，観察対象 o のコーディングとは o の何かの特徴 c を一定の目的のためにマーキングすること
- 従って，言語データ d のコーディングとは，d の何かの特徴を一定の目的のためにマーキングすること
  - 特徴の選び方は任意だが，よい特徴を選べるかどうかにセンスの有無が現われる
  - それから，あまり欲張りすぎないこと!

コード可能な言語の特徴

- 要素の語彙クラスに関する(品詞)情報
- 要素の出現位置に関する情報
- 要素の担う意味の情報
  - 文法役割: Predicate, Subject, Object, ...
  - 主題役割: Agent, Patient, Theme, Instrument
  - 選択制限 (selectional restrictions)
- 統語情報は以上の特徴の混合
  - 音韻情報もコード可能だが音素変換なしでは難しい

Ob se rva tion is a s kill.
Obse rva tion is not f ree.
Good ob serv ation sk ills are aq uired throu gh experin ces.
G oo d in tuiti ons gro w out of goo d obs erva ti ons.

観察の実践例

コーディングの前処理

- 適当なコーパスを KWIC 検索
  - どんなコーパスが適当かは調査の性質による
  - ゴミがでないように検索パターンを工夫
    - Perl, Python, Ruby のような言語が検索に使えると最高だが，市販の検索ツールの利用も可
- 得られた検索結果の整形 (EXCEL が有用)
  - 不要データ (いわゆるゴミ) の濾過
  - 必要ならランダムサンプリングによってサイズ縮小
  - データの粗い分類

具体例

- 「襲う」のコーディングはいかに行われたか
  - 特にそれはどんな結果を見越して行われたか

Thank You for Your Observation