形態論的制約を用いた未知語の自動獲得

2008年3月20日

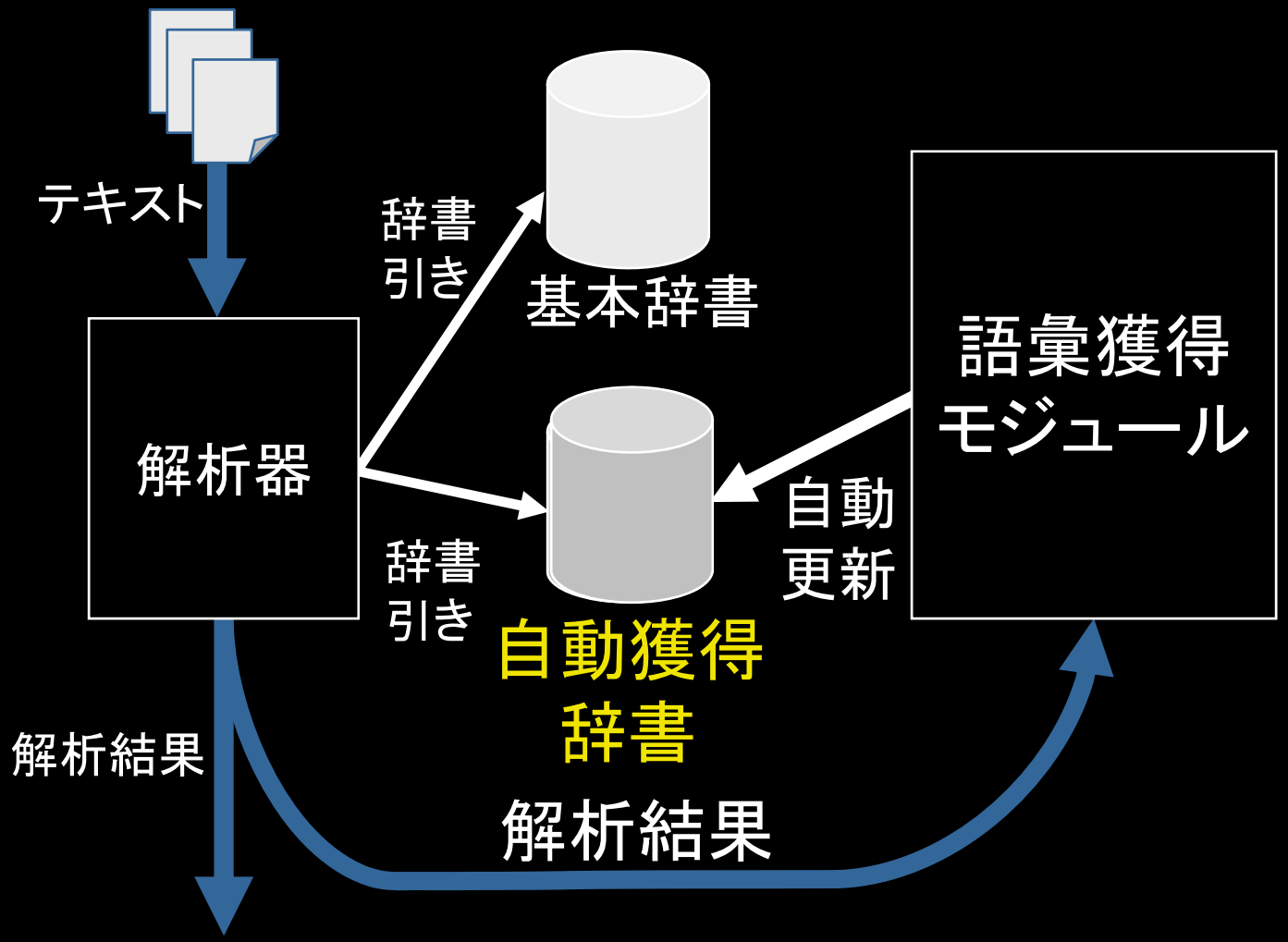
京都大学大学院情報学研究科

村脇 有吾　黒橋 禎夫

概要

- 自己を書き換えるシステム
  - 解析用辞書をシステムが動的に書き換える
- テキストから逐次的に未知語を獲得
  - その場で辞書を更新し解析にフィードバック
- 語彙獲得に形態論的制約を利用
  - 日本語に特化

システム構成



背景

- 言語の解析において未知語が障害
  - 基本語彙以外の名詞や動詞、形容詞
    - 難しい漢語
      - 完熟
    - 新語
      - ググる、准教授
    - 俗語
      - 殺る (やる)
  - 固有名詞
    - 壇ノ浦

語彙獲得タスク (1/2)

- 対象となる未知語は自立語
  - 普通名詞
  - サ変名詞
  - イ形容詞
  - ナ形容詞
  - 動詞
    - 母音動詞、子音動詞カ行など、活用型別に

語彙獲得タスク (2/2)

未知語について以下を同定

1. 形態素境界
   - 甘酸っぱい
2. 品詞
   - グ - って: 動詞? 名詞? ⇔ グ - る
   - 完熟 - の: 普通名詞? サ変名詞?
     ⇔ 完熟 - する

複数の用例を調べないと
確実には判断できない

従来手法 (1/2)

-

低頻度の形態素の

獲得精度に問題

より少量のデータから

獲得したい

存してください。
マンゴーって確か
しい限り♪今回は
類のフルーツ達が
、樹の上でじっと

完熟
完熟
完熟
完熟
完熟

のため、なるべく
のもののははじめて
の甘さを１ｋｇ分
を迎えます。スモ
を待ってから収穫

従来手法 (2/2)

- 入力から最適な出力候補を推定

$$\arg\max_{y}$$

- 1用例から
  2004]
  - 確実には
- 単純に用
  - 複数用例
    2006]
  - 多義性の

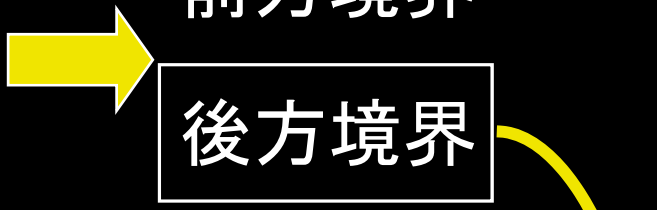
どこまで用例を増やせば
獲得に十分かを主体的に
判断すべき

比較的少数の

用例からの逐次獲得

提案手法

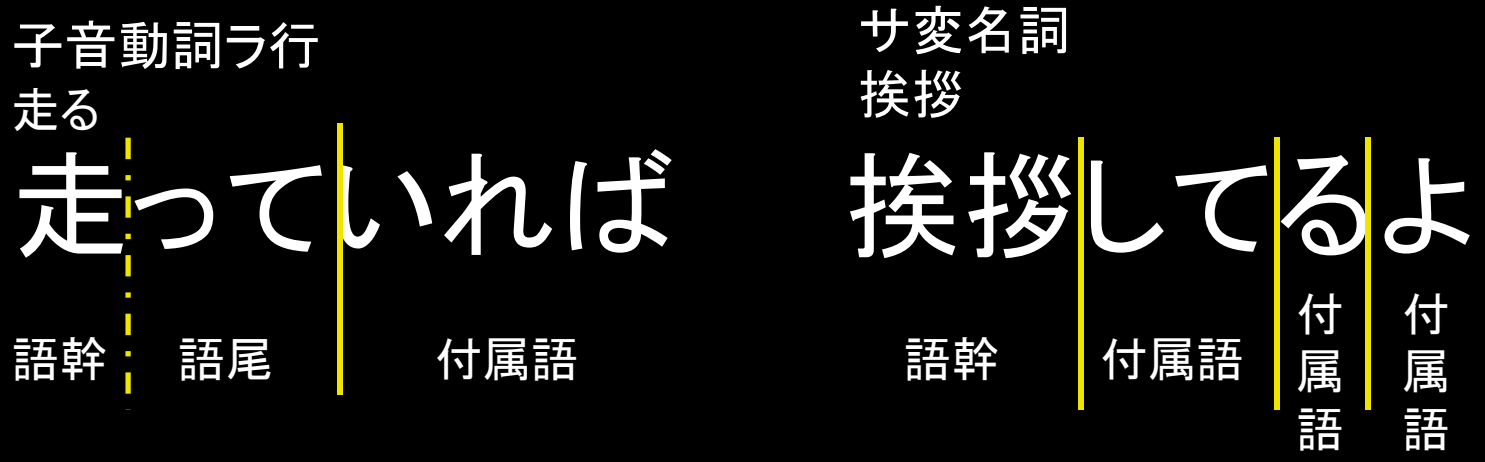
- 未知語について以下を同定
  - 形態素境界
    - 甘酸っぱ - い
  - 品詞
    - ググ - る: 子音動詞ラ行
    - 完熟 : サ変名詞



形態論的制約を利用

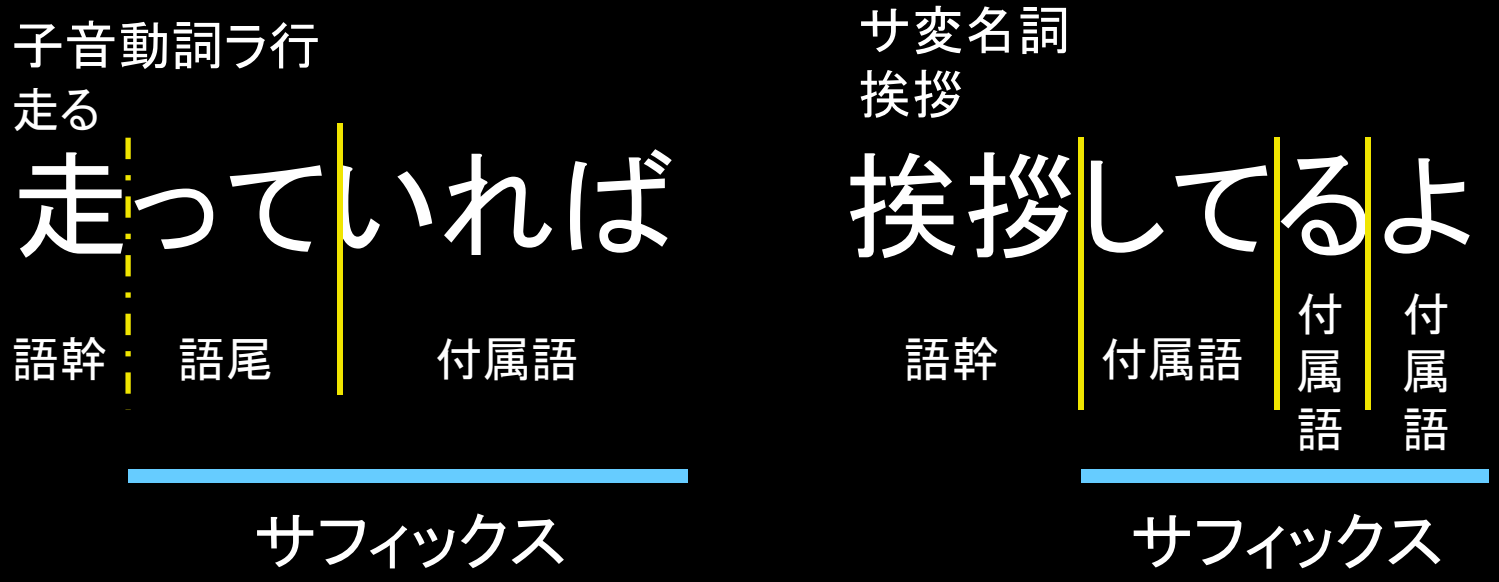
形態論的制約

- 日本語は膠着語
- 自立語にはいくつかの付属語が後続



形態論的制約とサフィックス

- サフィックスの接続は、品詞ごとに強い制約がある



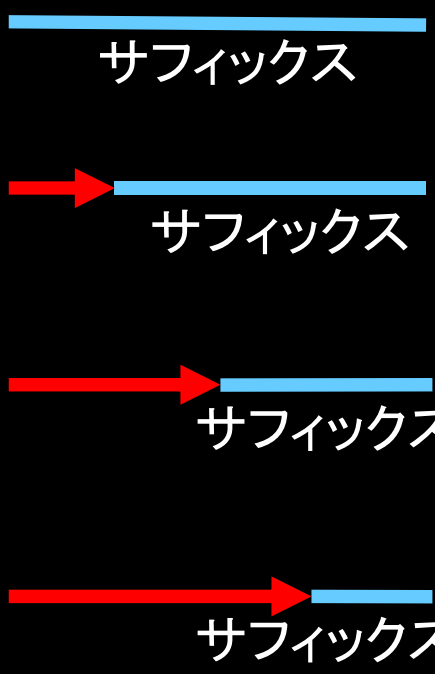
(準備) サフィックスの獲得

- コーパス: ウェブ文書1億ページの解析結果
- 既知の形態素に後続するサフィックス
- 自立語の品詞ごとにまとめる
- 獲得結果
  - 総出現回数　　約33億
  - 異なり数　　約53万
  - 品詞別異なり数　　約70万

(後方境界) 候補の列挙

ピボット

完熟してるよ



品詞	活用形
サ変名詞	する
子音動詞サ行	タ系連用テ形
母音動詞	基本連用形
母音動詞	タ系連用テ形
サ変名詞	る
子音動詞ラ行	基本形
普通名詞	る
母音動詞	基本形
サ変名詞	よ
ナ形容詞	語幹
母音動詞	省略意志形
普通名詞	よ

候補の選択

- 複数の用例を比較
- 多くの用例を説明する候補を選択

完熟してるよ		完熟の	
完熟	サ変名詞 子音動詞サ行 母音動詞	完熟	サ変名詞 ナ形容詞 母音動詞 イ形容詞 普通名詞

完熟し	母音動詞
⋮	⋮

獲得条件

- 候補を1つに絞り込む
- 活用形の異なり数3以上に獲得
  - 名詞の場合は後続する形態素の原形の異なり数

用例の集約 (1/3)

- 同じ未知語の複数の用例を集めたい
- 形態素境界が分からない
- どうやって集約するか?

用例の集約 (2/3)

- ピボット
  - 語幹の一部となる文字列
  - 字種に基づく簡単なルールにより選択

「ググる
またググったら

同じ未知語から同じピボットを選択

完熟してるよ
もう完熟の
追熟を楽しむ

別の未知語から同じピボットが選択されることもある

用例の集約 (3/3)

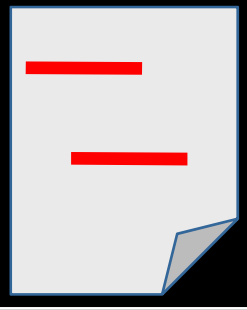
1. ピボットをキーに大雑把に用例を集約
2. 集まった用例を比較して形態素境界を確定させる
    1. 形態素の前方境界
        - 現在は、文頭や句読点などの明らかな境界のみで判定
    2. 形態素の後方境界

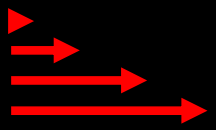
語彙獲得の手順

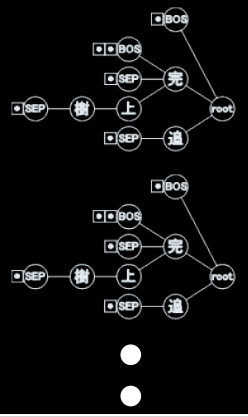
解析結果

用例

完熟してるよ







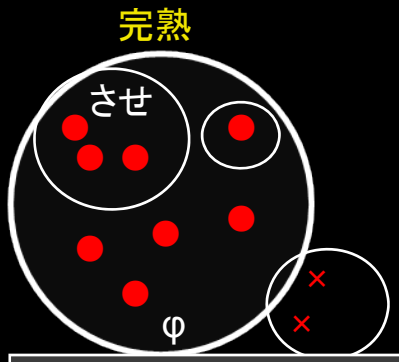
未知語を含む文字列(用例)を検出

ピボットの選択

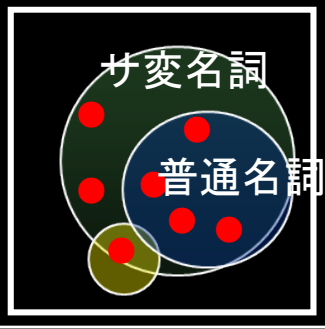
後方境界候補の列挙

ピボットごとに用例を集約

前方境界を決定



後方境界の決定



品詞の識別

完熟:
サ変名詞

辞書の更新

実験

- ウェブコーパス約1億ページから逐次獲得
- 調査内容
  - 獲得数
  - 獲得に要する用例の数 (中央値)
  - 獲得語彙の妥当性

実験結果 (抜粋)

品詞	獲得数	獲得に要した用例数 (中央値)	獲得例
母音動詞	1,854	18	棄る (すてる)
子音動詞カ行	355	9	ムカつく
子音動詞ラ行	893	6	ググる
イ形容詞	814	18	甘酸っぱい
ナ形容詞	936	12	イマイチだ
普通名詞	37,945	52	防具
サ変名詞	6,854	9	クリック

獲得語彙の評価

- ランダムに選んだ100個を評価

種別	数	例
正解	77	オヤヂ、瑞々しい
分割可能な複合語	9	在阪企業
名詞と⼗形容詞の誤り	5	ヤバン:普通名詞
その他	9	乗分布、ダレカ:普通名詞

今後の課題 (1/2)

- 未知語検出
  - 特にひらがなの未知語 (e.g. うざい)
- 提案手法の使い方
  - 獲得語彙が解析に副作用を起こす危険性
    - e.g. 短い形態素による過分割
  - 獲得語彙の使用範囲の決定

今後の課題 (2/2)

- 副詞、感動詞の識別
  - 識別に構文レベルの情報が必要
- 固有名詞の品詞再割り当て
  - 現在は普通名詞として獲得
- より頑健な語彙獲得
  - 獲得語彙の追跡調査・修正
- 頑健な形態素解析
  - 誤字、強調表現

提案手法の利点

- 計算が単純
- コーパス全体を見る必要がない
  - クロールしたばかりの最新の文書で辞書を更新
- 教師なし学習
  - コストがかからない
- パイプライン処理に向いている
  - 固有名詞、カテゴリ [原島 2007]、ドメイン [橋本 2008]

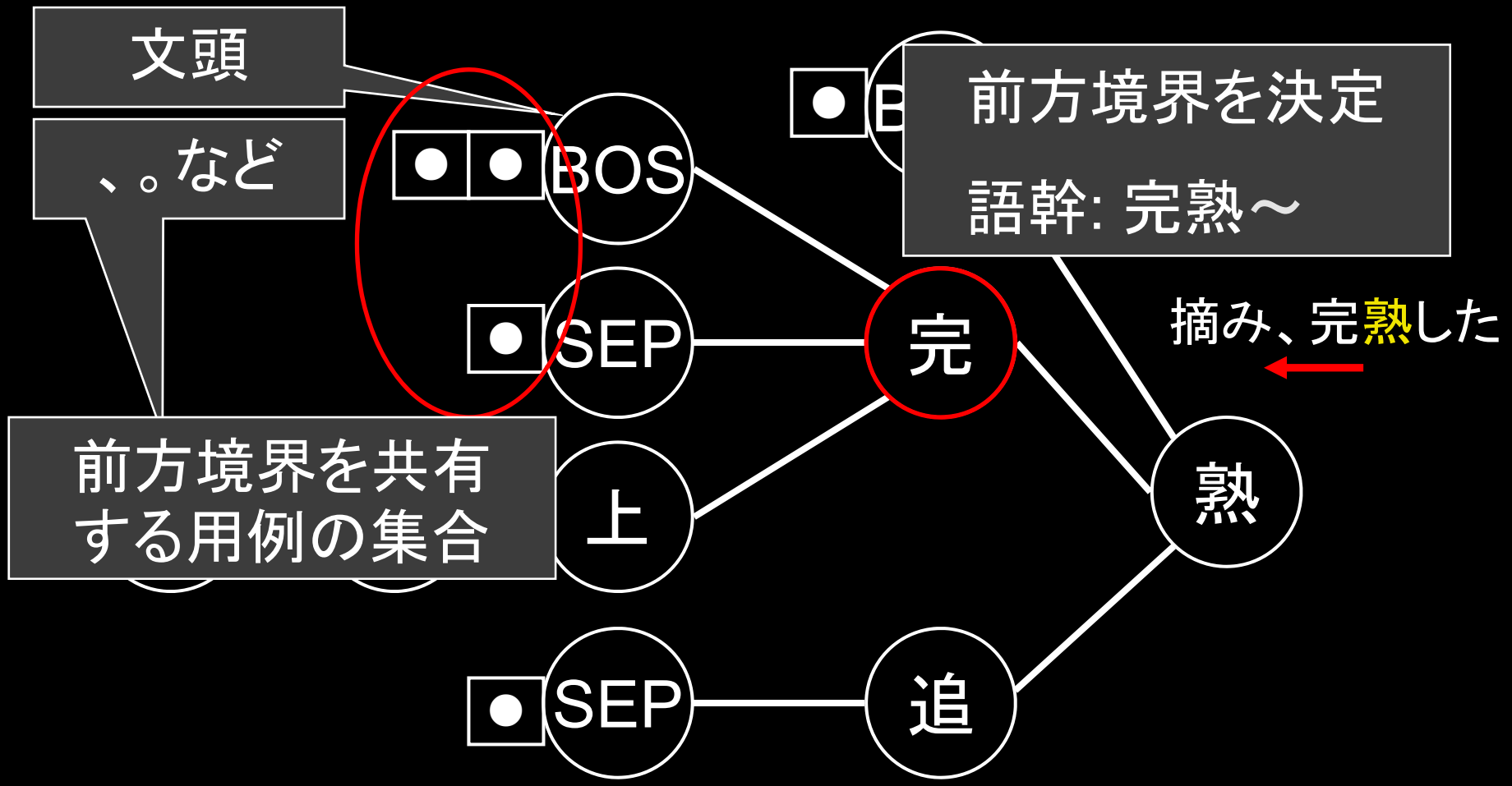
従来手法 (1/2)

- 解析時にその場で同定
  - 字種に基づくヒューリスティクス (JUMAN, Chasen, Mecab)
    - e.g. カタカナ連続を未知語候補とする
  - 統計的推定 [Nagata 1999][Uchimoto 2001][Asahara 2004]
- 問題
  - 語彙が一般に持つ手がかりは弱い
    1. 字種
    2. 形態素長

用例の集約の従来手法

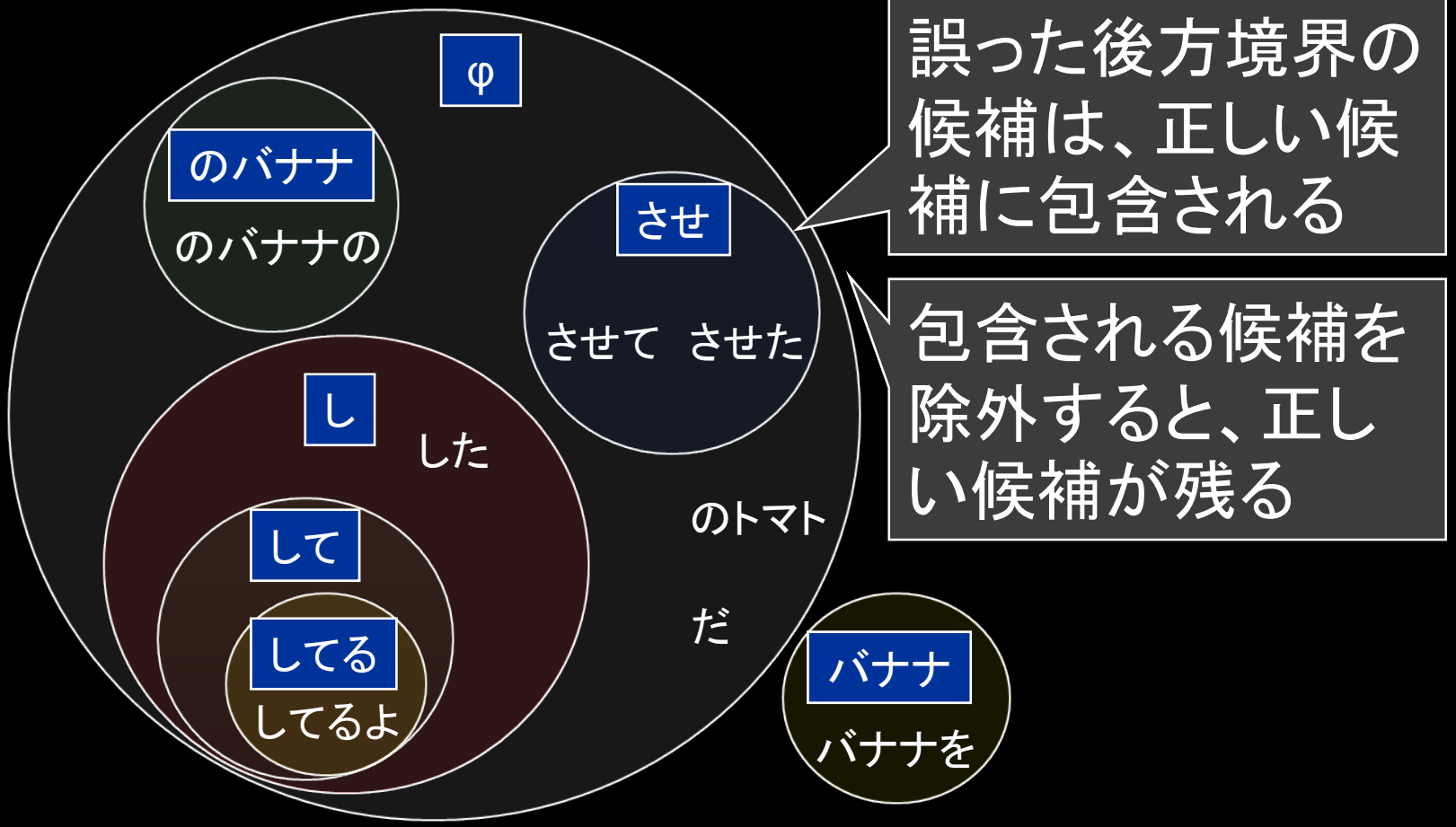
- 同じ未知語の複数の用例をどうやって集約するか?
  - 文字n-gram [森 1998]
    - コーパス全体から計算しないといけない
  - 局所的な情報だけでなんとかしたい

前方境界の決定



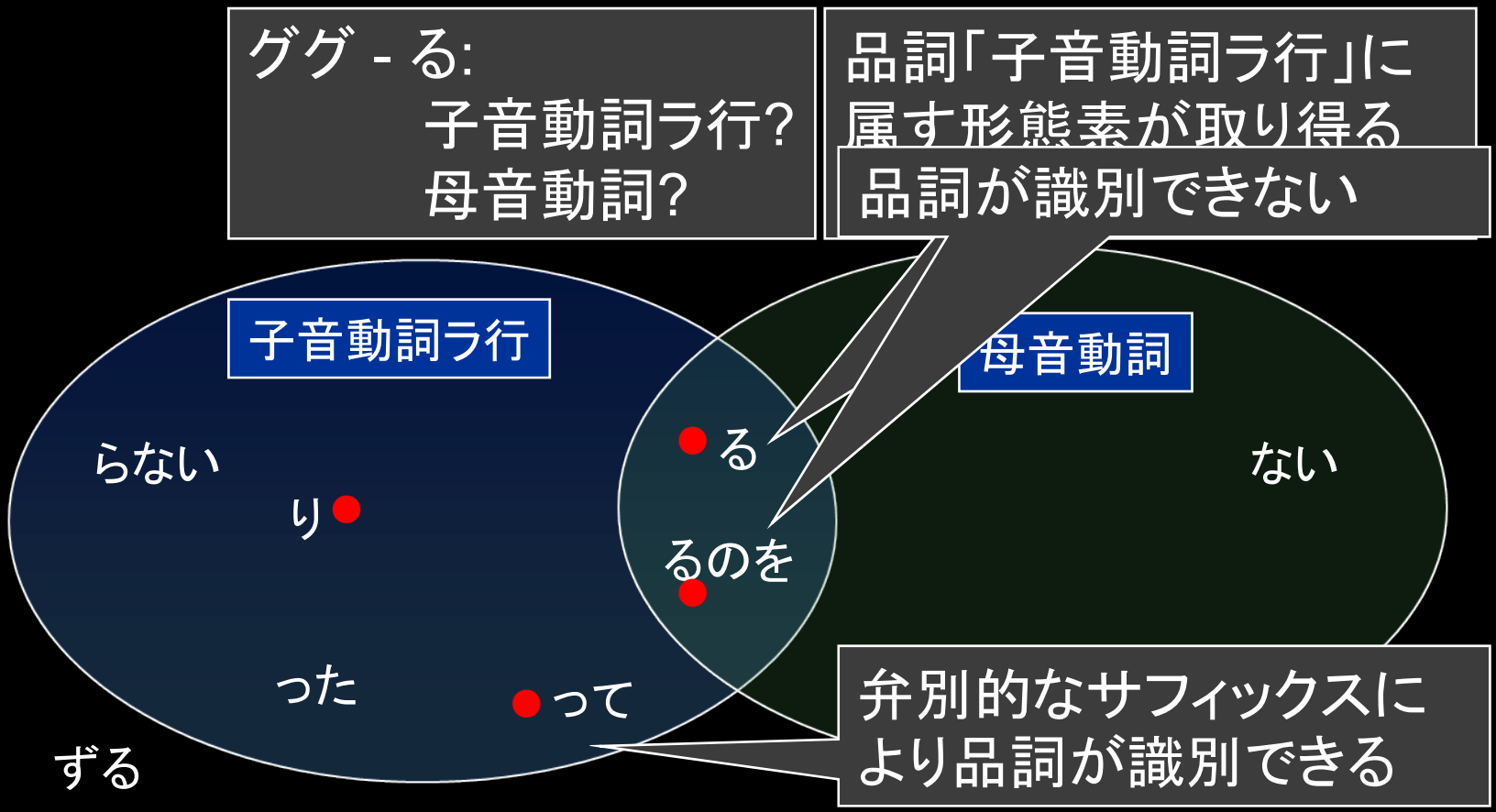
ピボット「熟」のトライ

後方境界の決定



「完熟～」の後方境界の候補

品詞識別 (1/2)

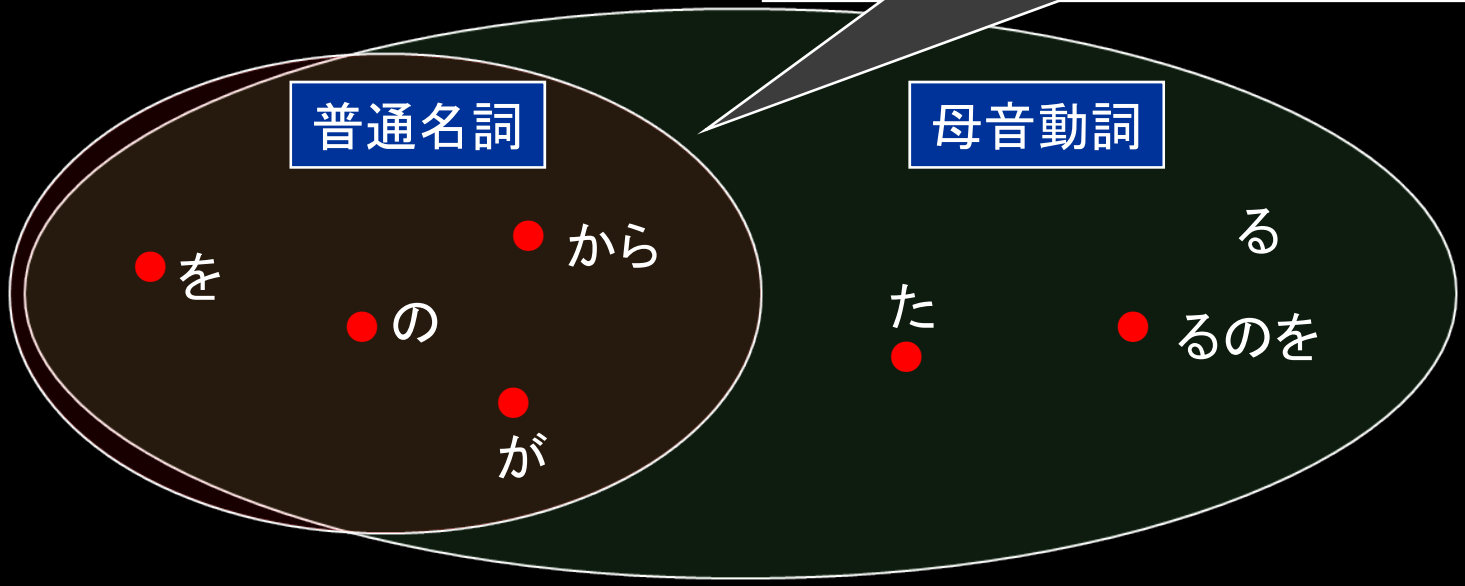

ググ - る:
子音動詞ラ行?
母音動詞?
品詞「子音動詞ラ行」に属す形態素が取り得る品詞が識別できない
子音動詞ラ行
母音動詞
らない
り
る
るのを
ない
った
って
ずる
弁別的なサフィックスにより品詞が識別できる

品詞識別 (2/2)

准教授:
母音動詞?
普通名詞?

憧れ

母音動詞の弁別的なサフィックスが現れないことに気付く必要



誤り分析

- 名詞とナ形容詞の区別は難しい
  - 誤ってもあまり問題ない
- 感動詞や副詞
  - プツッ
- 語尾までカタカナ
  - ウザイ
- カタカナの指示詞など
  - コチラ

結論

- テキストから逐次的に語彙を自動獲得
  - その場で獲得した語彙を解析器にフィードバック
- 1用例ごとに候補を列挙
- 同じ未知語の用例を集約
- 複数の用例の比較により決定
  - 形態素境界
  - 品詞