# 形態素解析へのフィードバックのための未知語の自動獲得

知能情報学専攻
黒橋研究室
村脇 有吾

2008年12月4日 情報学若手コロキアム

# 概要

未知語の獲得手法を提案

e.g. ググる, ようつべ (YouTube)

- 形態素解析器の出力から未知語を獲得し、解析器の辞書を直接更新してフィードバック
- 解析器の出力ごとに未知語の用例を検出
- 記憶に用例を蓄積し、複数用例を比較して、最適な辞書項目の候補を選択

# 形態素解析のタスク

入力

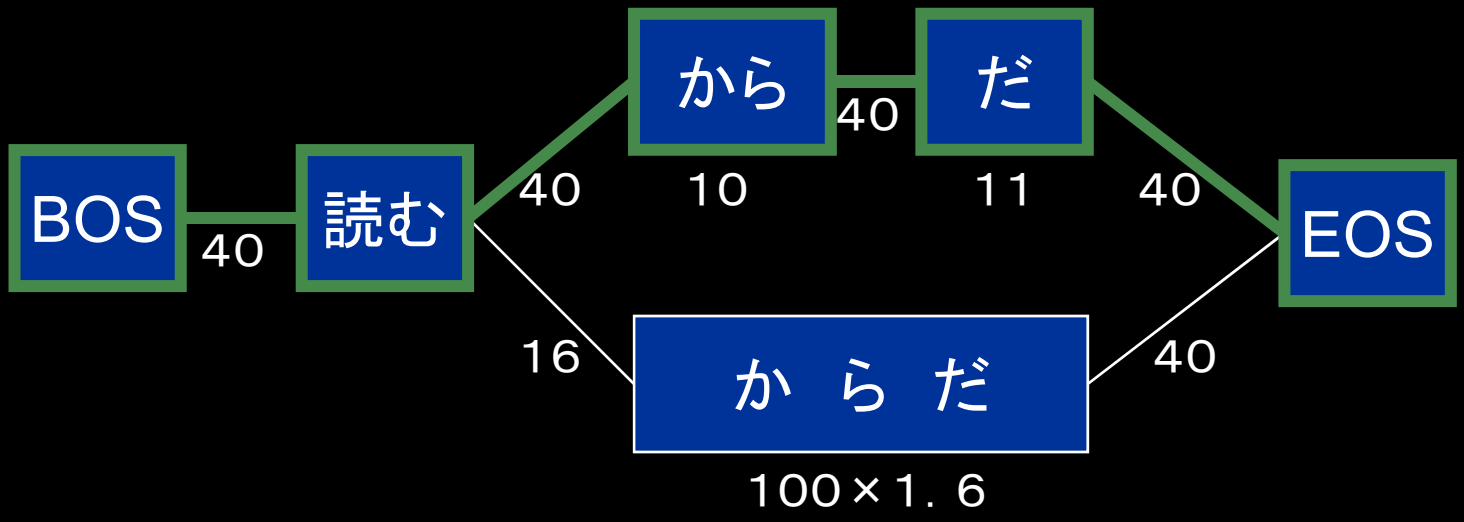
| |
|---|
| 読むからだ |

出力

| | | |
|---|---|---|
| 読む | 子音動詞マ行 | 基本形 |
| から | 接続助詞 | * |
| だ | 判定詞 | 基本形 |

- 文を形態素に分解
- 各形態素に品詞を割り当て

# 辞書引きによる候補列挙




1. 予め定義された辞書を引いて候補を列挙
2. コストにより最適なパスを選択

[Kurohashi et al. 1994] [Asahara and Matsumoto 2000] [Kudo et al. 2004]

# 形態素解析は既に高精度

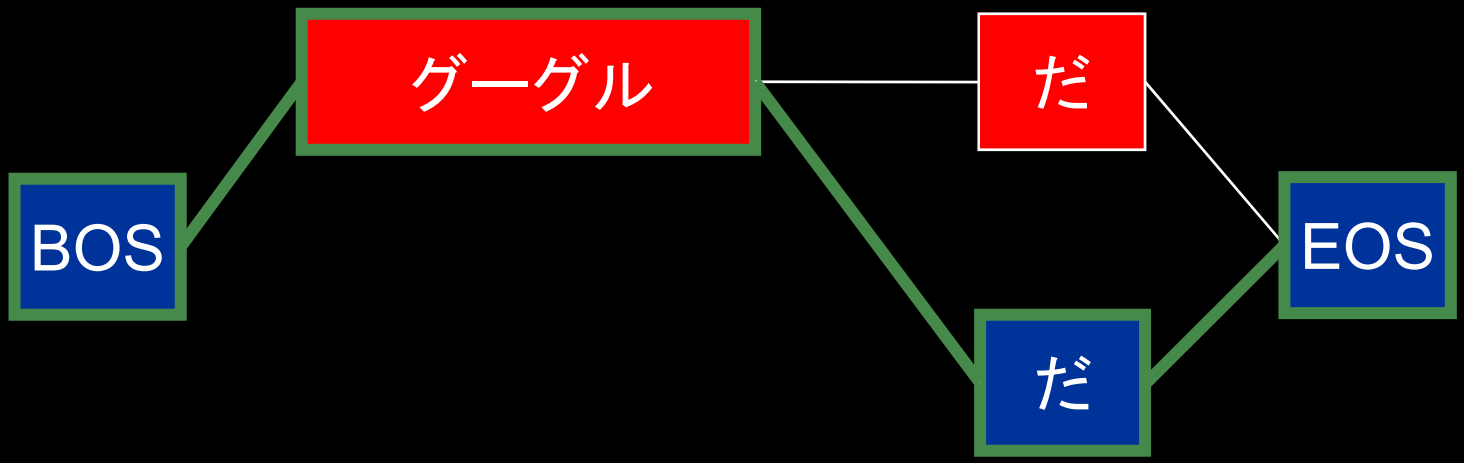
Table 3: Results of KC, ($F_{\beta=1}$ (precision/recall))

| system | seg | top | all |
|---|---|---|---|
| L2-CRFs ($C=1.2$) | **98.96 (99.04/98.88)** | **98.31 (98.39/98.22)** | **96.75 (96.83/96.67)** |
| L1-CRFs ($C=3.0$) | 98.80 (98.84/98.77) | 98.14 (98.18/98.11) | 96.55 (96.58/96.51) |
| MEMMs (Uchimoto 01) | 96.44 (95.78/97.10) | 95.81 (95.15/96.47) | 94.27 (93.62/94.92) |
| JUMAN (rule-based) | 98.70 (98.88/98.51) | 98.09 (98.27/97.91) | 93.73 (93.91/93.56) |
| HMMs-bigram (baseline) | 96.22 (96.16/96.28) | 94.96 (94.90/95.02) | 91.85 (91.79/91.90) |

[Kudo et al. 2004]

- ただし、形態素が辞書に登録済みなら

# 未知語の処理




- 辞書にない形態素 (未知語) の扱い
  - 字種に基づくヒューリスティクス
  - 動詞は? ググる
  - ひらがな語は? ようつべ (YouTube)

# 未知語による誤りの連鎖

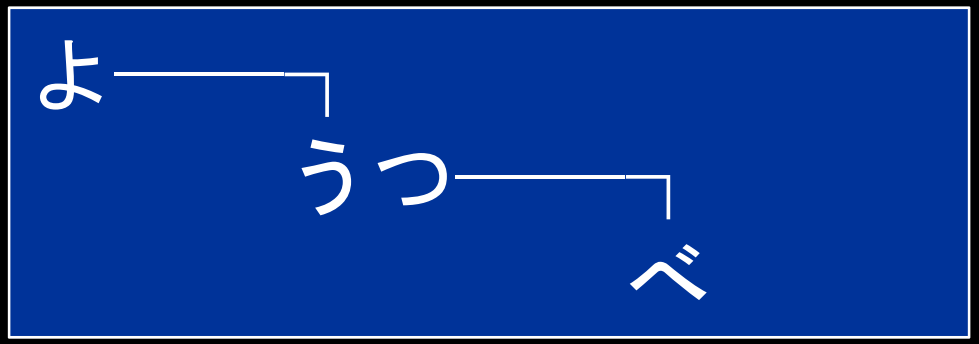
入力文

ようつべ

形態素解析

よ（夜）+うつ（打つ）+べ

構文解析



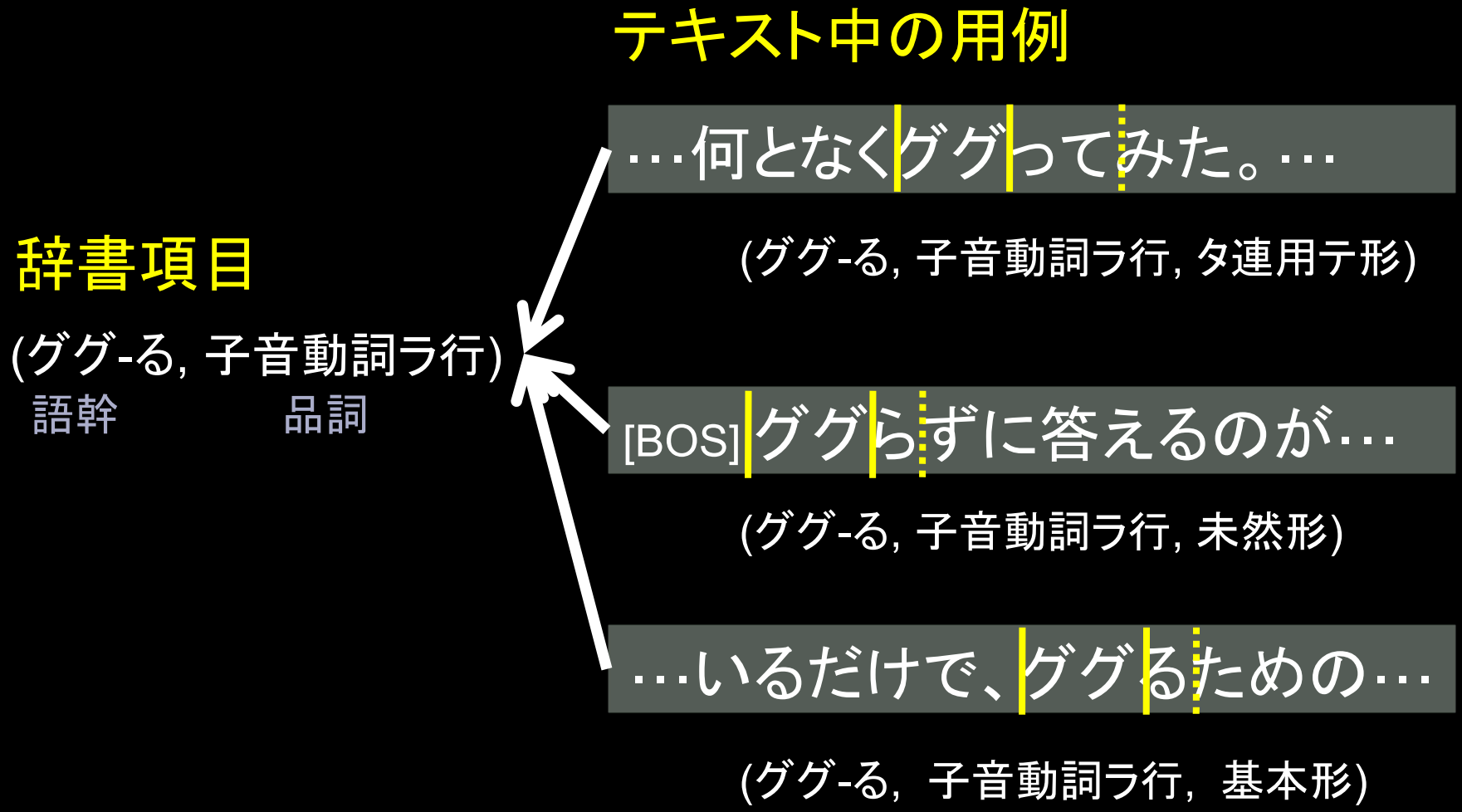

格解析

打つ（べ:ガ，よ:時間）

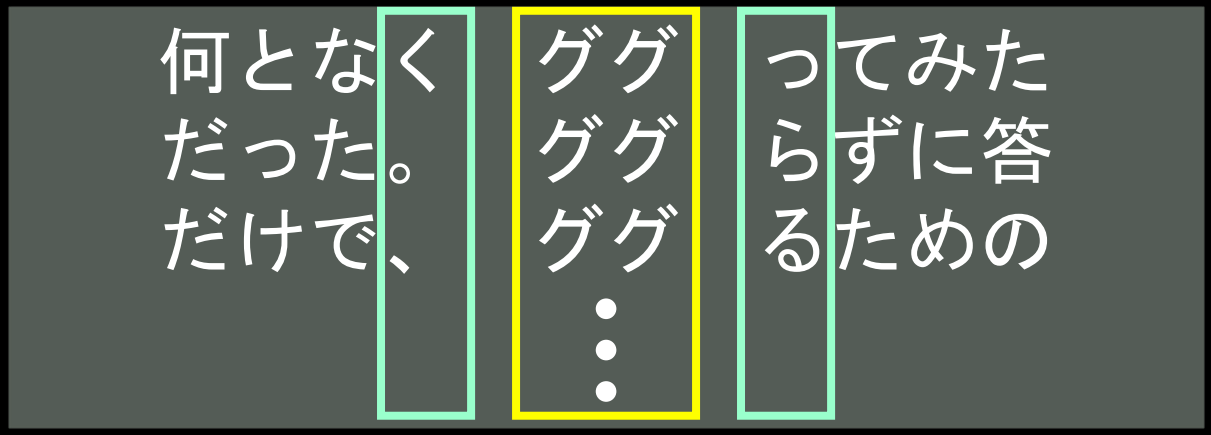
# 未知語問題の解決方法

1. もっと賢い未知語モデルを作る [Nagata 1999] [Uchimoto et al. 2001]
   - 文字N-gram、字種、形態素長、etc
2. 語彙獲得により未知語を辞書登録 [Mori and Nagao 1996] [福島ほか 2007]

# 語彙獲得のタスク


テキスト中の用例
…何となくググってみた。…
(ググ-る, 子音動詞ラ行, タ連用テ形)
辞書項目
(ググ-る, 子音動詞ラ行)
語幹
品詞
[BOS]ググらずに答えるのが…
(ググ-る, 子音動詞ラ行, 未然形)
…いるだけで、ググるための…
(ググ-る，子音動詞ラ行，基本形)

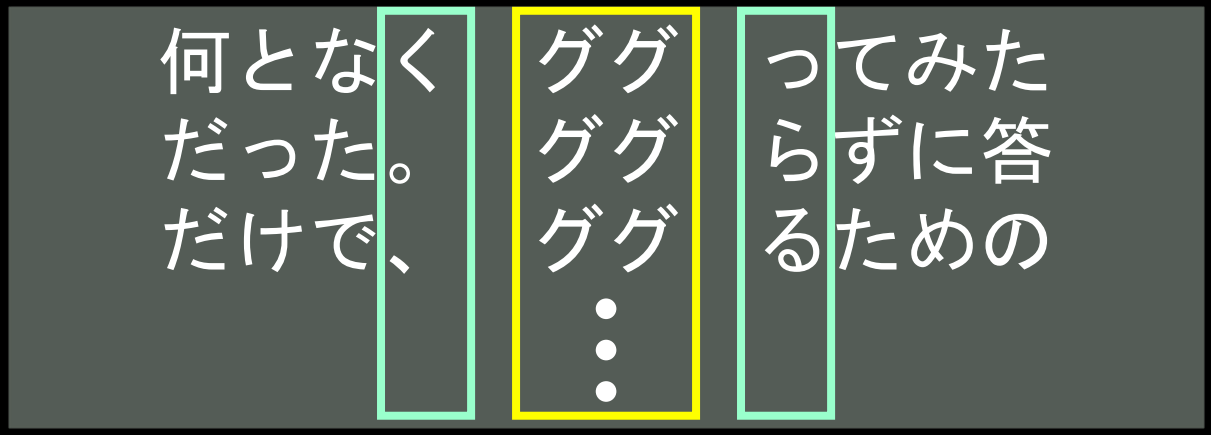
# 語彙獲得の従来手法




1. コーパス全体をソート
2. コーパスから任意の文字列 (語幹候補) を抽出
3. 前後の環境 (文脈) ベクトルにより品詞を決定

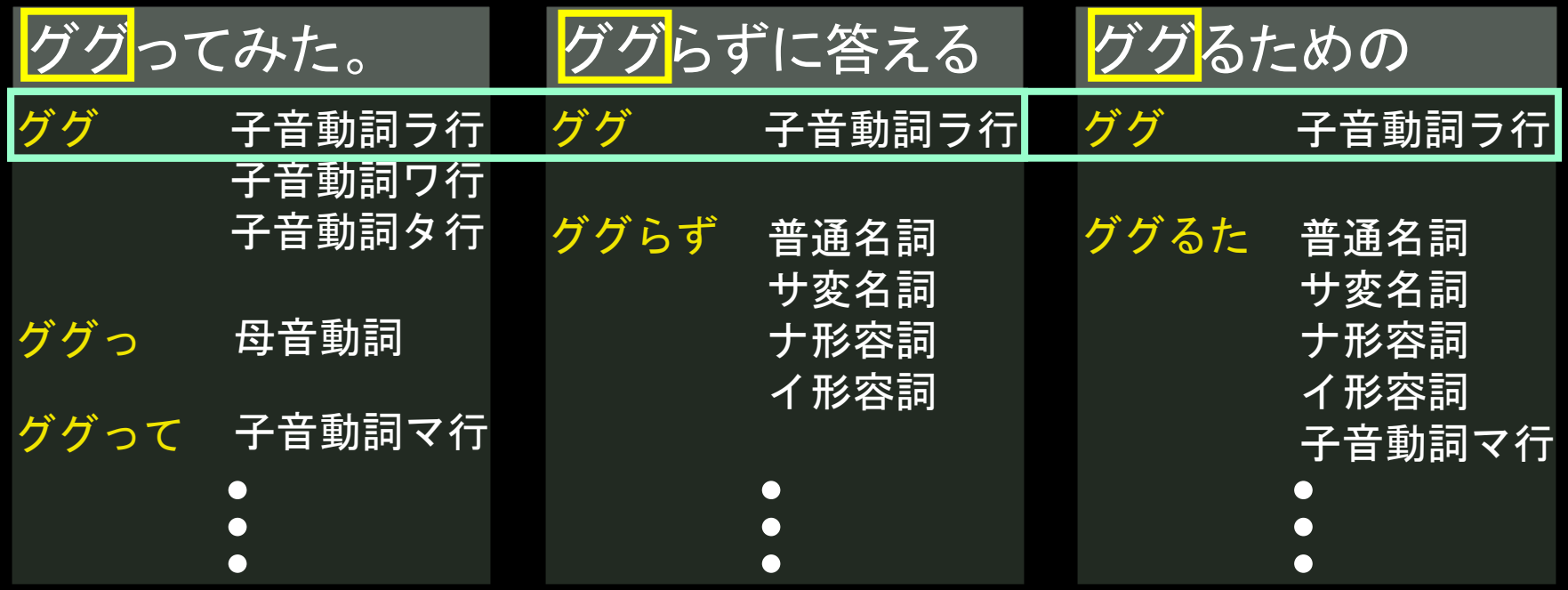
[Mori and Nagao 1996]

# 語彙獲得の従来手法



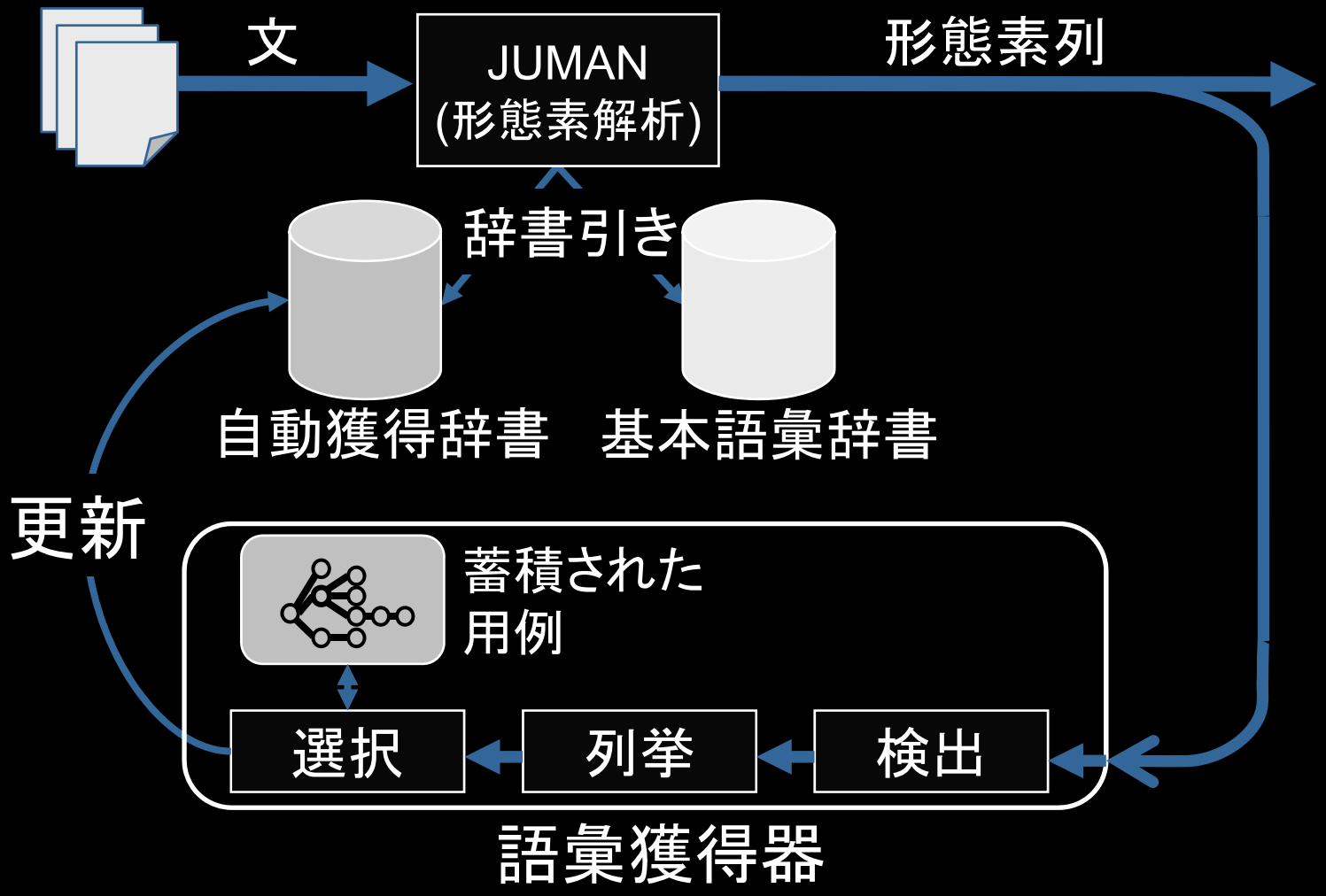

- コーパス全体を見ないでも獲得できるはず
  - ウェブコーパスは全部見るには巨大すぎる
- 環境 (文脈) はノイズが多い
  - 頻度10未満の候補を無視しているが、10回も出現すれば十分わかるはず

# 提案手法のアイデア




- バッチではなくオンライン
  - 各文から未知語用例を検出、辞書項目候補を列挙
  - 複数用例の比較により曖昧性を解消

# システム構成


文
JUMAN
(形態素解析)
形態素列
辞書引き
自動獲得辞書
基本語彙辞書
更新
蓄積された
用例
選択
列挙
検出
語彙獲得器

# 未知語の用例の検出

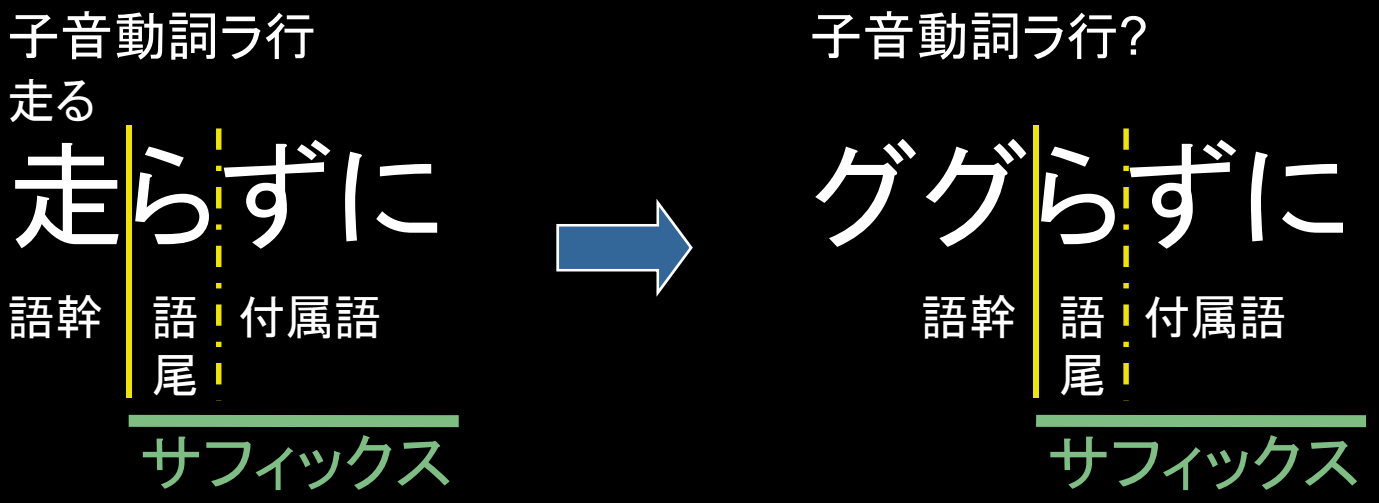
- 形態素列を走査して、未知語の用例が含まれていそうな箇所を見つける
- タスクの性質
  - 検出されなければ獲得しようがない
  - 多少間違えても問題ない

ググ：未定義語+って+みた

# 辞書項目候補の列挙

- タスク: 語幹と品詞の組を列挙
  - 語幹は前方境界と後方境界の組
- 利用できる手掛かり
  1. 形態素解析器の出力した形態素境界
     - 間違っているかもしれない
  2. 文頭、句読点などの明示的な境界
  3. 語尾や付属語の連接制約 (形態論的制約)

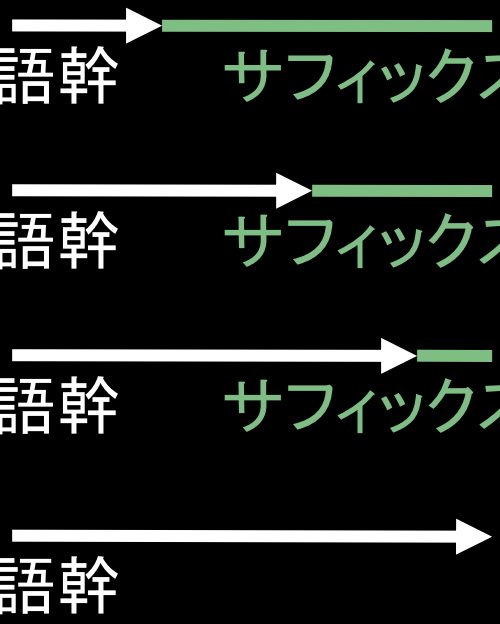
# サフィックスと形態論的制約




- 語幹に後続し得る文字列に着目 [福島ほか 2007]
  - あらかじめ登録済み形態素の用例から収集
  - 未知語の用例に適用
    - 後方境界と品詞の組の候補

# 文字列マッチングによる列挙

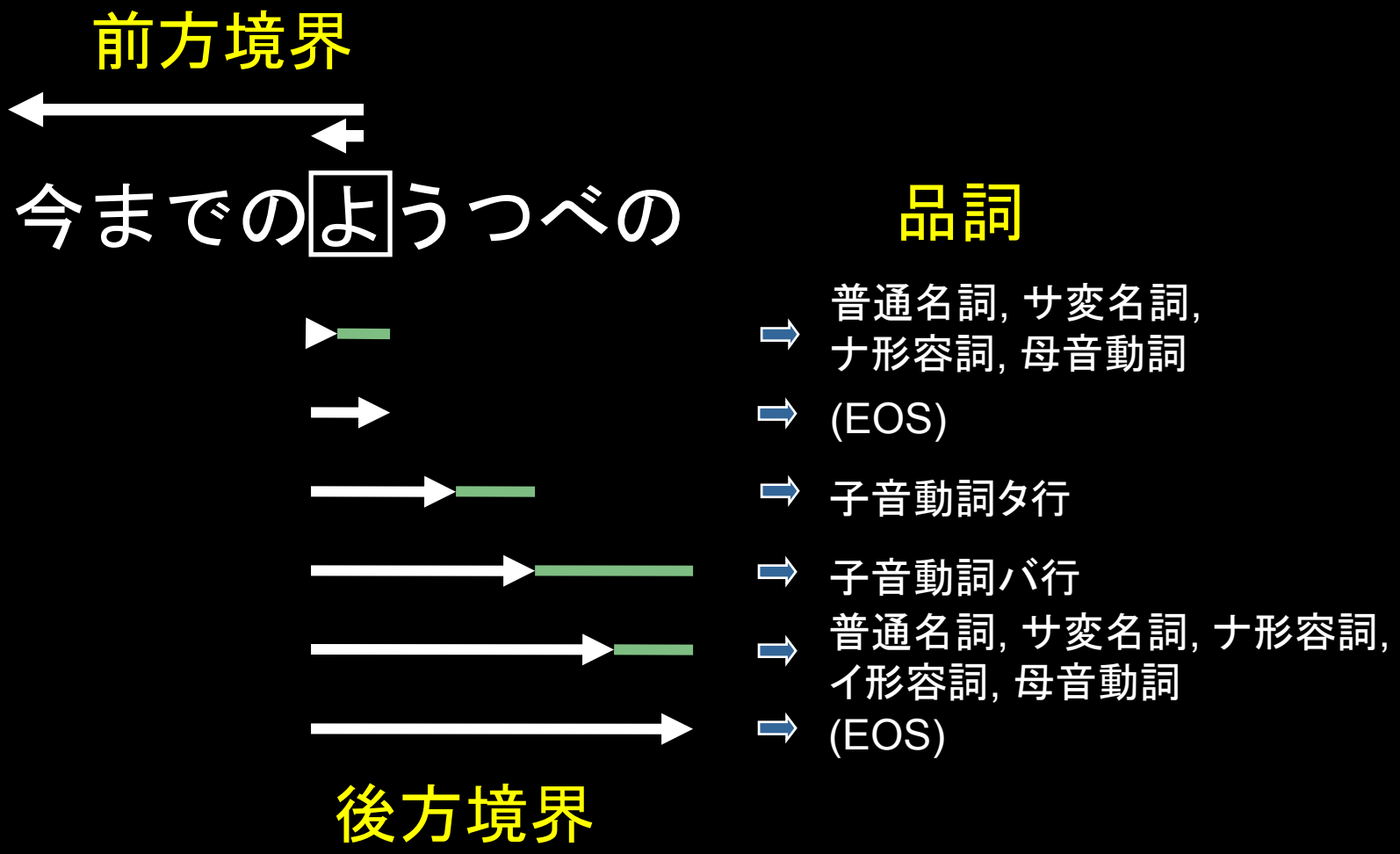

何となく[ググ]ってみた。




品詞

- ⇒ 子音動詞ラ行
- ⇒ 子音動詞ワ行
- ⇒ 子音動詞タ行
- ⇒ 子音動詞マ行
- ⇒ 母音動詞
- ⇒ 子音動詞タ行
- ⇒ (EOS)

# 文字列マッチングによる列挙

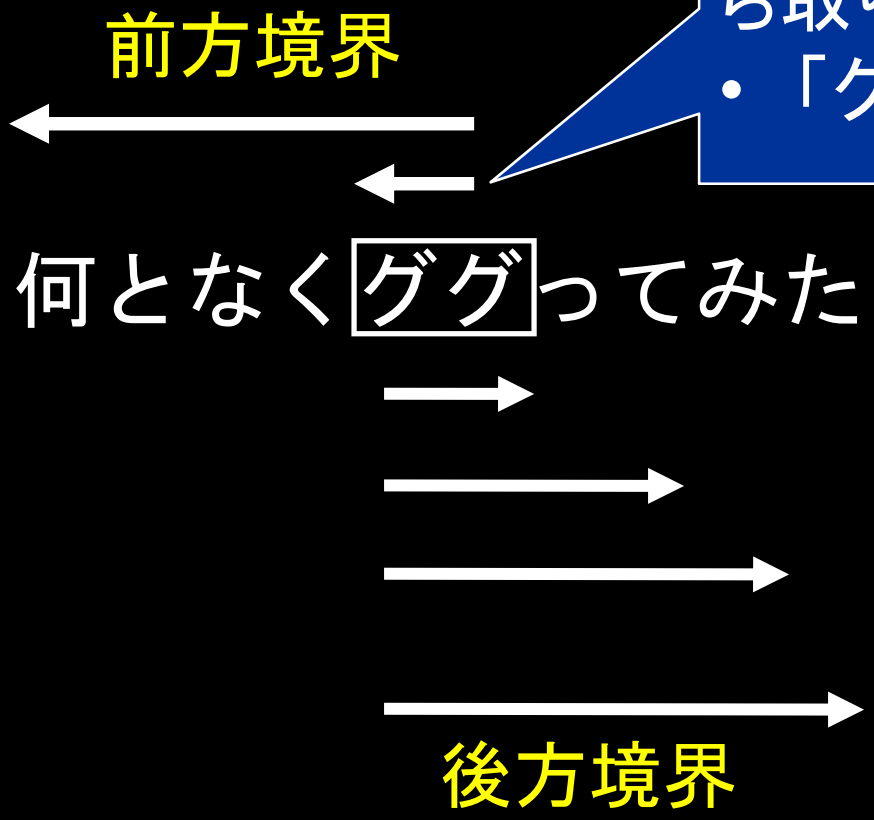


# 最適な辞書項目候補の選択

- 用例を記憶に蓄積
- 複数の用例を比較して、うまく説明できる辞書項目の候補を選択
- 判断を保留というオプション

# 記憶からの用例の取り出し

前方境界を共有している可能性のある用例を記憶から取り出す
- 「ググ」から始まる用例




品詞

⇒ 子音動詞ラ行, 子音動詞ワ行
子音動詞タ行

⇒ 子音動詞マ行

⇒ 母音動詞, 子音動詞タ行

⇒ 普通名詞, サ変名詞,
ナ形容詞, 母音動詞

# 最適な辞書項目候補の選択

| ググってみた。 | |
|---|---|
| ググ | 子音動詞ラ行 |
| | 子音動詞ワ行 |
| | 子音動詞タ行 |
| ググっ | 母音動詞 |
| ググって | 子音動詞マ行 |
| | ● |
| | ● |
| | ● |

| ググらずに答える | |
|---|---|
| ググ | 子音動詞ラ行 |
| ググらず | 普通名詞 |
| | サ変名詞 |
| | ナ形容詞 |
| | イ形容詞 |
| | ● |
| | ● |
| | ● |

| ググるための | |
|---|---|
| ググ | 子音動詞ラ行 |
| ググるた | 普通名詞 |
| | サ変名詞 |
| | ナ形容詞 |
| | イ形容詞 |
| | 子音動詞マ行 |
| | ● |
| | ● |
| | ● |

- 終了条件
  - 3種類以上の異なる活用パターンが出現

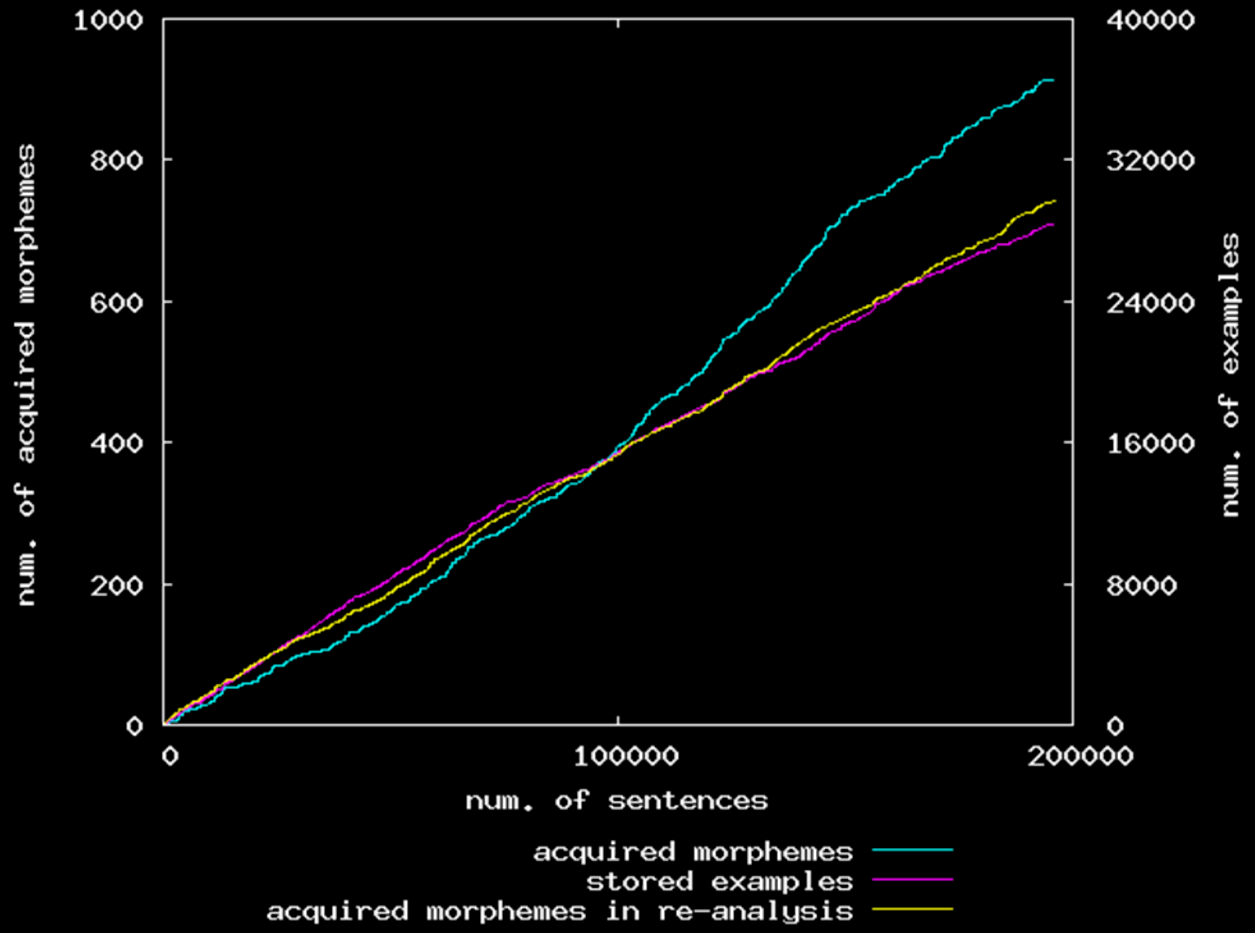
# 実験

- 入力データ: 検索エンジンが返す文書上位1,000ページ
  - クエリ: “捕鯨問題”, “赤ちゃんポスト”, “JASRAC”, ”ツンデレ”, “アガリクス”
- 評価
  1. 獲得される形態素の精度
  2. 獲得語彙による形態素解析の改善

# 獲得された形態素

- 総数: 107—903個
- 精度: 93.9—99.3%
- 獲得に使った用例数: 4—9
  - cf [Mori and Nagao 1996] では10回未満の候補は無視
- 検出された未知語の用例のうち、頻度にしておよそ半分をカバー

# 語彙獲得の経過


1000
800
600
400
200
0
num. of acquired morphemes
40000
32000
24000
16000
8000
0
num. of examples
0
100000
200000
num. of sentences
acquired morphemes
stored examples
acquired morphemes in re-analysis

# 獲得例

| クエリ | 獲得例 |
| --- | --- |
| 捕鯨問題 | モラトリアム, ツチクジラ,混獲 |
| 赤ちゃんポスト | ダンナ, 助産師, 棄てる, 訊く |
| JASRAC | ソフ倫, シャ乱Q, ヲタ |
| ツンデレ | アキバ, 腐女子, モテる |
| アガリクス | サプリ, アロマ, 食効 |

# 形態素解析の改善

- 2つの辞書の解析結果を比較し、差分を評価
  - 文単位では1.04—15.4%が変化
  - 50文を人手で評価

| クエリ | 分割/品詞 | | | | 合計 |
|---|---|---|---|---|---|
| | E → C | C → C | E → E | C → E | |
| 捕鯨問題 | 11 | **45** | 0 | 2 | 58 |
| 赤ちゃんポスト | **37** | 12 | 9 | 3 | 52 |
| JASRAC | 16 | **23** | 1 | 12 | 52 |
| ツンデレ | 17 | **39** | 0 | 1 | 57 |
| アガリクス | 22 | **31** | 0 | 0 | 53 |

# まとめ

- オンラインの語彙獲得手法を提案
  - 比較的少数の用例から獲得可能
  - 高精度
- 獲得により形態素解析が改善される