オンライン語彙獲得を用いた リアルタイムウェブの言語処理

京都大学大学院 情報学研究科

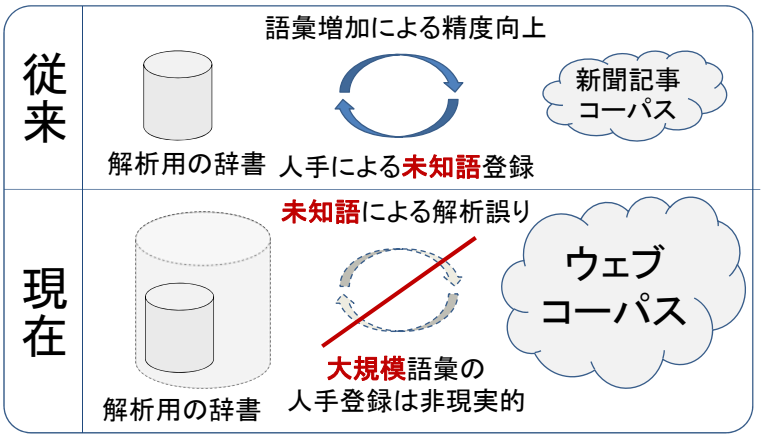
村脇 有吾 murawaki@nlp.kuee.kyoto-u.ac.jp

黒橋 禎夫 kuro@i.kyoto-u.ac.jp

オンライン語彙獲得 + リアルタイムウェブ = リアルタイムに言葉を覚えるボット

1. 背景

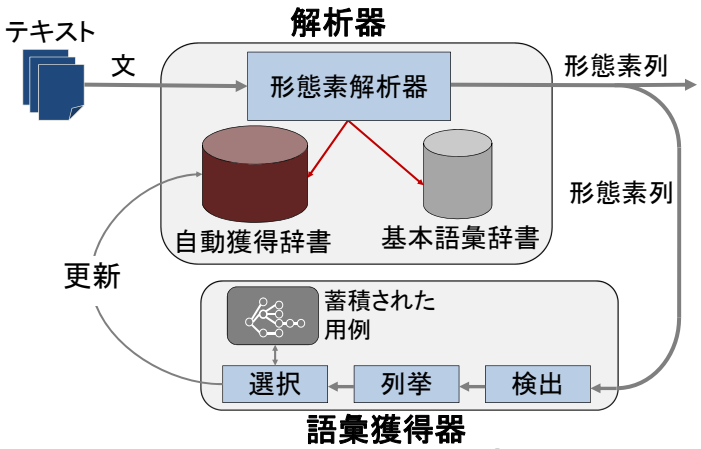
- テキストからの情報抽出に**形態素解析**が必要
  - 日本語は分かち書きされていないため、形態素 (単語) が抽出できない
- 形態素解析には**辞書**が重要



- 未知語による形態素解析誤り
  - ググる ⇒ ググ + る
  - ついったー ⇒ つ + いった（言った）+ ー

2. テキストからのオンライン語彙獲得

- 基本語彙 (約12万) は人手で整備済み
- 足りない語彙を**テキスト**から**オンライン**で自動獲得
  - **ググ-る**:**動詞-ラ行**
  - **ついったー**:**名詞**
- 人手による獲得語彙の修正は原則なし



[Murawaki+ 2008]

3. ツイッター上で言葉を覚えるボット

- **ツイッター** http://twitter.com
  - ツイート: 利用者の投稿 (最大140文字)
  - タイムライン: ツイートの時系列順の表示
  - フォロー: 他の利用者のツイートを自分のタイムラインに登録
- 言葉を覚えるボット
  - タイムラインのツイートを語彙獲得システムに入力
  - 獲得されたら「覚えた」とツイート

ジャーゴンの獲得と、獲得の決め手となったツイート

特徴と今後の課題

- 誤字に対してもある程度頑健
  - 複数の使われ方を見比べて候補を絞り込むから
- 同一文の解析結果が変化し得る
  - 辞書を刻々と更新するため
- 超大規模語彙の扱いがオープンな問題
  - 従来: 高頻度語のみ人手で登録; 低頻度語は無視
  - 従来: 対象分野特有の語彙を整備 (分野適応)
  - 今後: 明確な分野境界が存在しないテキストで低頻度語も正確に扱う