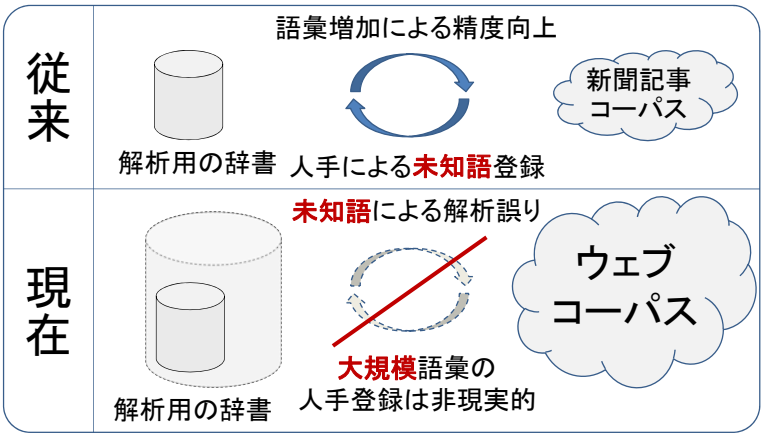
自然言語の解析のための テキストからの語彙の自動獲得

知能情報学専攻 黒橋研究室 D2　村脇 有吾　murawaki@nlp.kuee.kyoto-u.ac.jp

背景

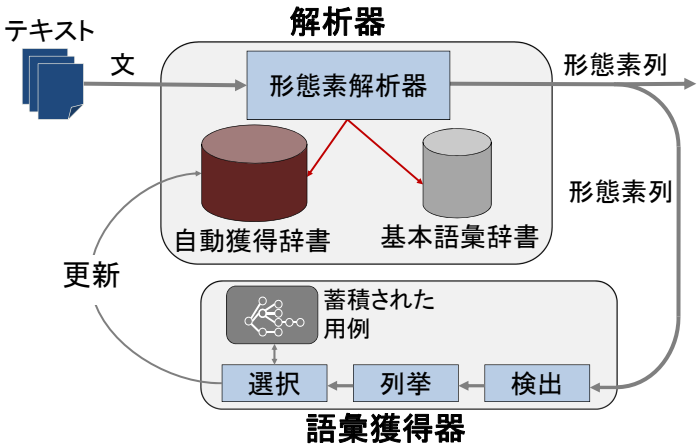
- テキストからの情報抽出に**形態素解析**が必要
  - 日本語は分かち書きされていないため、形態素 (単語) が抽出できない
- 形態素解析には辞書が重要



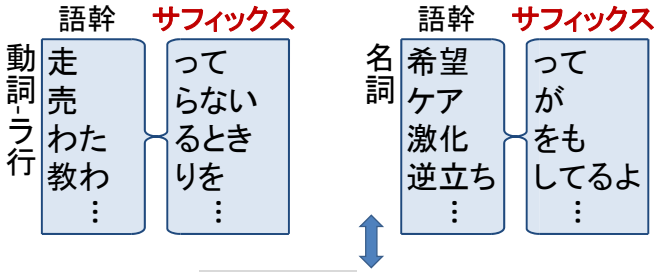
- 未知語による形態素解析誤り例
  - ググる　⇒　ググ + る
  - ついったー　⇒　つ + いった（言った）+ ー

テキストからの語彙の自動獲得

- 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自動獲得
  - e.g. **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



形態論的制約による識別



基本語彙から構築した知識を未知語に適用

ググって
ググらない
ググるとき

ついったーって
ついったーが
ついったーをも

実験

1. 設定
   - ウェブの検索結果上位1,000ページから語彙獲得
     - 84,498—271,898文 (5文書セット)
2. 結果
   - 獲得語彙数: **74--460**
   - 精度: **97.3—98.5%**
   - **4-7**個の用例を見た時点で獲得 (中央値)

文書セット	獲得形態素の例
捕鯨問題	モラトリアム, ツチ鯨, 混獲, 解-る:動詞-ラ行
赤ちゃんポスト	ダンナ, 棄て-る:動詞-母音, 訊-く:動詞-カ行
ジャスラック	シャ乱Q, ぱく-る:動詞-ラ行, スゴい:イ形容詞
ツンデレ	アキバ, 腐女子, しまんぬ, モテ-る:動詞-ラ行
アガリクス	サプリ, アロマ, 食効, αリポ酸, すぐりむん

獲得語彙を用いた再解析による変化

文書セット	分割				分割 + 品詞				計
	誤→正	正→正	誤→誤	正→誤	誤→正	正→正	誤→誤	正→誤	
捕鯨問題	19	**36**	0	3	19	**36**	0	3	58
赤ちゃんポスト	**33**	19	0	1	**33**	19	0	1	53
ジャスラック	18	**32**	2	0	12	**32**	8	0	52
ツンデレ	4	**53**	0	1	4	**53**	0	1	58
アガリクス	**30**	29	0	0	**30**	29	0	0	59

応用

1. ツイッターでリアルタイムに言葉を覚えるボット

2. 構文レベルの知識の利用
   - タスク: 自動獲得した名詞をさらに分類
     - e.g. 倖田來未:名詞-**人名**, けいはんな:名詞-**地名**
   - 語彙獲得により形態素解析が正しく行えると、構文解析結果が利用できるようになる
     - Xを通行する　⇒　Xは場所（固有 or 普通)?
     - Xを遣わす　⇒　Xは人（固有 or 普通)?
     - Xが多い　⇒　Xは普通名詞?
     - 2人のX　⇒　Xは人の普通名詞?