テキストから自動獲得した名詞の分類

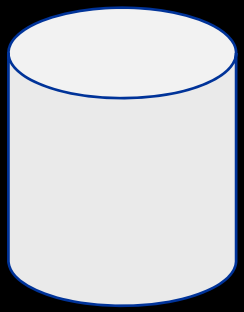
2010年3月11日

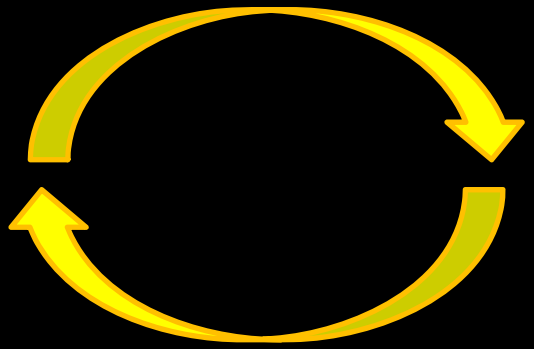
京都大学大学院情報学研究科

村脇 有吾　　黒橋 禎夫

従来: 形態素辞書の整備 (90年代-)

語彙増加による
解析精度向上

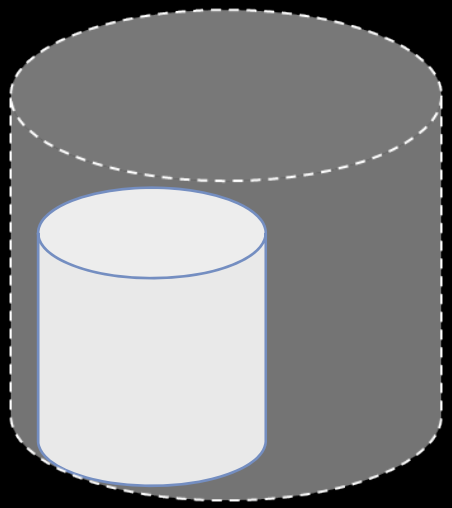




解析用の辞書

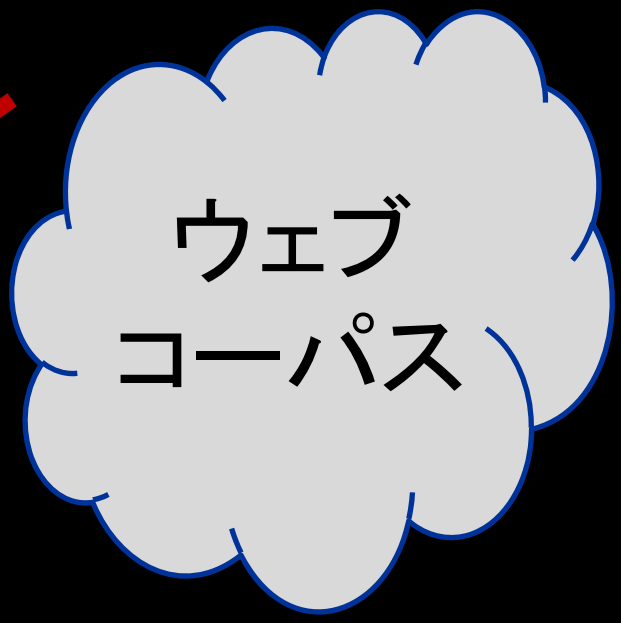
人手による**未知語**登録

現在: 未知語問題



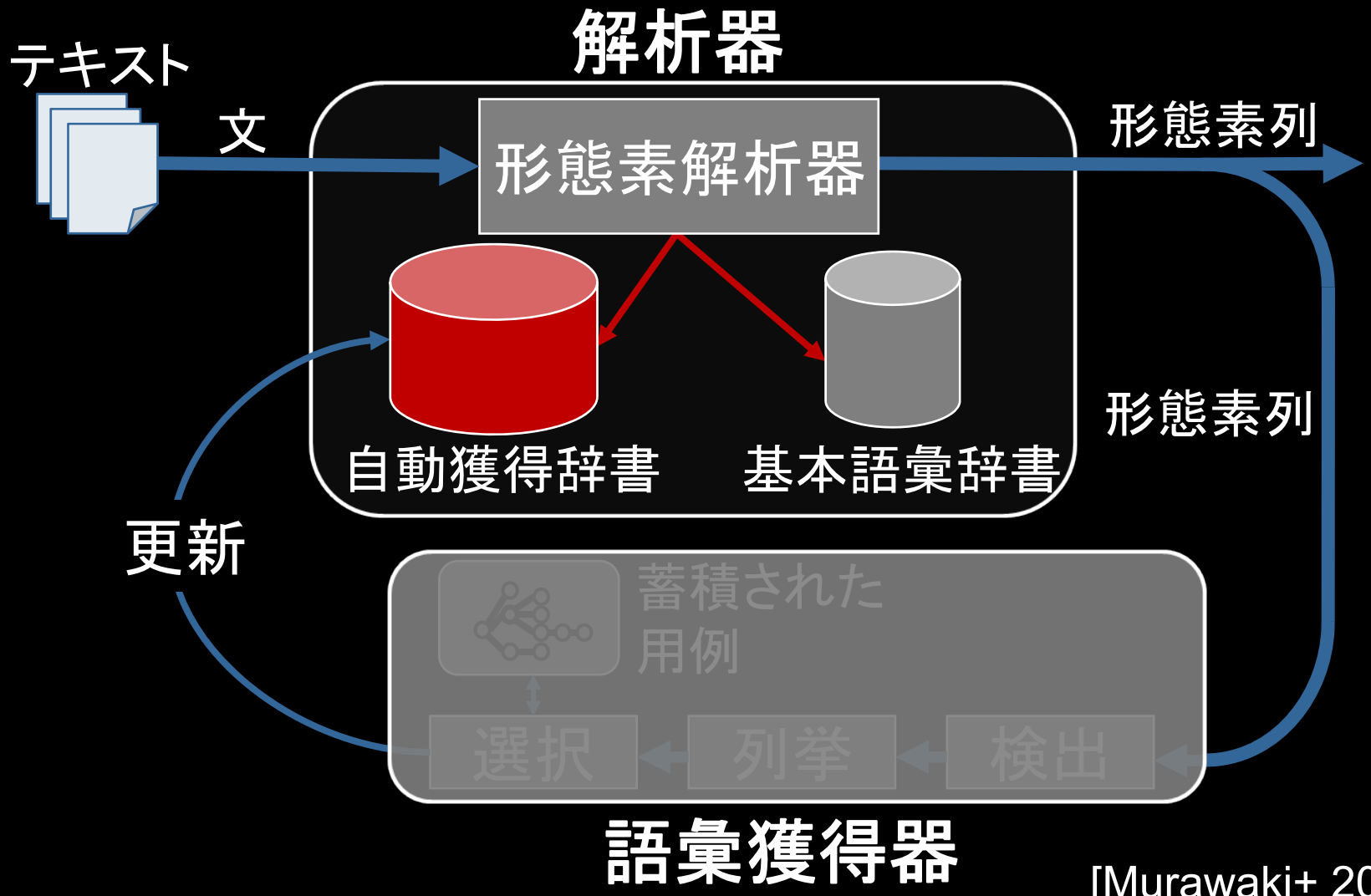
解析用の辞書

未知語による
解析誤り



大規模語彙の
人手登録は非現実的

テキストからの語彙の自動獲得



[Murawaki+ 2008]

知識の段階的獲得

形態レベルの分類
[Murawaki+ 2008]

品詞	品詞細分類	活用型	形態素の例 (基本語, 自動獲得語)
動詞		母音 (一段)	まぜる, 寝る, ボケる, ぶっちゃける
		ラ行	憧れる, 作る, 練る, ぐぐる, メタボる
		…	
形容詞		イ形容詞	暑い, 軽い, 濃い, かわゆい, うざい
		ナ形容詞	きれいだ, 巧みだ, ガチだ, シュールだ
名詞	普通名詞		被告, 部屋, 政府, 犬, ニート, 着メロ
	サ変名詞		主張, 離脱, スリップ, 試聴, コスプレ
	人名		裕介, 愛子, ジョージ, 佐祐理, キョン
	地名		京都, 銀座, ドイツ, アキバ, ウブド
	組織名		日銀, トヨタ, NHK, エリクソン, マツダ
	固有名詞		ガンダム, スラブ, 平成, ジプシー

意味レベルの分類
[今回の発表]

Outline

- 背景
- タスク設定
- 多値分類による解法
- 実験

Outline

- 背景
- タスク設定
- 多値分類による解法
- 実験

固有名詞にあわせて
普通名詞も分類

名詞の分類タスク

分類ラベル	普通/固有	形態素の例 (基本語, 自動獲得語)
人	普通名詞	被告, 先生, スタッフ, メル友, ニート
場所		職場, 部屋, カフェ, 囲炉裏, 圃場
組織		政府, 仲間, チーム, 興信所, 弊所
動物		犬, ムカデ, 顔, チワワ, マンタ
植物		花, ハーブ, 葉っぱ, ケナフ, クレマチス
普通名詞その他		主張, 花火, ビジネス, 甚平, 着メロ
人名	固有名詞	松井, 愛子, ジョージ, 佐祐理, キョン
地名		京都, 銀座, ドイツ, アキバ, ウブド
組織名		日銀, トヨタ, NHK, エリクソン, マツダ
固有名詞その他		ガンダム, スラブ, 平成, ジプシー

比較: 分類のための手がかり

レベル	分類対象	分類の手がかり	手がかりの強さ
形態	• 名詞 • 動詞-ラ行 • イ形容詞 • ナ形容詞 • etc	文法 (後続要素の形態論的振る舞い) • 「る」は動詞-ラ行の語幹に後続; 名詞には後続しない • 「を」は名詞に後続; 動詞-ラ行の語幹は後続しない	• 強い • 制約として利用可能
意味	• 人 • 人名 • 場所 • 地名 • etc	語彙的選好 • 「2人のX」のXは人? • 「Xを通る」のXは場所 or 地名? • 「Xが多い」のXは普通名詞?	• 弱い • 複数を組み合わせて信頼性を高める

タスク設定

- 形態素を分類
  - 複合語は固有表現認識におまかせ
- 普通名詞と固有名詞を識別
  - 英語の場合は綴りにより明らか
- 構文解析の情報 (e.g. Xが → 通行) が使える
  - 形態レベルでの語彙獲得により、自動獲得語の形態素解析が正しく行えるから

Outline

- 背景
- タスク設定
- 多値分類による解法
- 実験

多値分類問題

- 入力: x (d次元の素性列)
- 出力: y ($y \in Y$ の分類ラベル)

v_y: d次元の重みベクトル

素性テンプレート

テンプレート	素性の抽出例
「XというY」、「XといったY」など	Xという国 ⇒ 国 Xといった人 ⇒ 人
係り受け関係にある用言と格要素	Xを通る ⇒ 通る:ヲ格 Xが多い ⇒ 多い:ガ格
連体修飾する指示詞	このX ⇒ この どんなX ⇒ どんな
ノ格による係り先の要素	Xの部屋 ⇒ 部屋 Xの大自然 ⇒ 大自然
ノ格による係り元の要素	すべてのX ⇒ すべて 二人のX ⇒ <数量>人
後続接尾辞・末尾要素	Xさん ⇒ さん X大統領 ⇒ 大統領

訓練とテスト

1. テキストを形態素解析 + 構文解析
2. 訓練: 解析結果の基本語部分
   - ラベルの曖昧な形態素を訓練データから除く
     - e.g. 森:普通名詞その他 or 人名 or 地名
   - 平均化パーセプトロンで訓練
3. テスト: 解析結果の自動獲得語部分

事例の生成

- 入力 x をどうやって作るか
- 多義性の問題
  - タイガー: 動物 or 人名
- 仮定: 同一文書内では同一の語義 (分類ラベル) で用いられる
- 文書内で抽出された素性をまとめて1事例とする

Outline

- 背景
- タスク設定
- 多値分類による解法
- 実験

実験設定

- 対象テキスト: ウェブコーパス 25Mページ
- 素性: 126,653個

- 訓練用の基本語彙: 20,726個
  - 事例数: 6,163,366個
- テスト用の自動獲得語彙: 10,437個
  - 事例数: 302,303個
    - うち500事例に人手で正解を付与

実験結果

- 精度: 66.0% (330 / 500)
- 普通/固有の誤りを許容: 79.4% (397 / 500)
  - e.g.「人名」を「人」と誤判定しても正解とみなす

- cf 基本語彙の分類
  - Closed: 94.2%
  - Open: 78.6%

分類結果の混同行列

システムの出力 ＼ 正解ラベル	人	場所	組織	動物	植物	普通他	人名	地名	組織名	固有他
人	32		1	4		8	59	1	1	2
場所		4		1		4		3	1	
組織							1		1	
動物	3			20		7				
植物						4	1			
普通他	10	7		17	2	226	11	1	9	2
人名	3			2			38		1	
地名		1				1	1	2	3	
組織名							1		2	
固有他										2

議論:「人」(普通名詞)と「人名」(固有名詞)

- 「人」と「人名」をその他のラベルから識別:
  - Xが言う, Xに質問, Xと暮らす
- 「人」を「人名」から識別: 手がかりは多い
  - 数量表現: 多くのX, <数量>人のX, Xが増える
  - 属性表現: Xになる, Xを兼ねる, Xを募集
- 「人名」を「人」から識別: 手がかりが少ない
  - 若干の後続接尾辞・末尾要素
    - 「X氏」は例外が多い (「記者氏」、「管理人氏」)
    - 「X両氏」ならほぼ人名だが、頻度が「X氏」の 1/100
    - cf 「Xさん」は「人」の重みの方が大きい (「患者さん」)
  - Xという男/女性

今後の課題

- とにかく精度をあげる
  - 事例あたりの素性数を増やす
    - 文脈に基づく素性列のマージ
  - 語構成情報 + 組み合わせ素性の利用
  - 基本語と自動獲得語のラベル分布の違いの考慮
- 分類結果を高次の処理で利用
  - 固有表現認識、省略・照応解析
- 分類ラベル設計の見直し
- 基本語の曖昧性解消