



# 統計的特徴を利用した機能語の自動認定

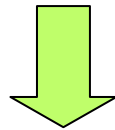
---

NHK放送技術研究所 人間・情報

木下 明德 後藤功雄 熊野正 加藤 直人 田中英輝

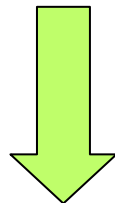
# NHKの国際放送

- 現在、20程度の言語で放送が行われている



翻訳作業が必要

- 多言語用例提示システムによる支援



- 言語によっては、十分な辞書などがないのが現状
  - 過去記事を有効活用したい
  - 言語によらない普遍的特徴でシステムを改良したい
- 内容語と機能語をうまく分類して検索精度を向上したい

例) インドネシア語

内容語

機能語

Dia datang dengan ibunya.

(彼女は、彼女の母親と来ました。)



# 内容語と機能語の分類

---

- 統計的特徴を利用して、機能語を認定する
  - 単語の出現頻度
  - 前後に隣接する単語の異なり数
  - エントロピー
  - エントロピー再計算
- 今回の実験の対象とした言語 (NHKニュースコーパス)
  - 日本語 (mecab)、英語 (penn treebank tagging set)
  - フランス語、スペイン語、イタリア語、ロシア語 (treetagger)
  - インドネシア語、マレー語 (GSK辞書)

# 出現頻度の例(日本語上位50単語＝機能語)

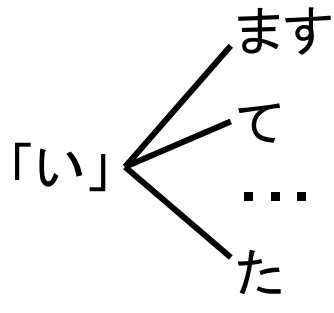
機能語とする

内容語

順位	単語 $w_i$	頻度 $F(w_i)$	順位	単語 $w_i$	頻度 $F(w_i)$	順位	単語 $w_i$	頻度 $F(w_i)$
1	の	1071301	11	い	213690			
2	を	586049	...			...		
3	に	580025	22	人	84726	73	選手	22748
4	た	536030	...			...		
5	が	509366				272	ながら	6399
...			48	日本	32887	...		
9	し	310266	49	で	32315			
10	と	270596	50	これ	31835	...		

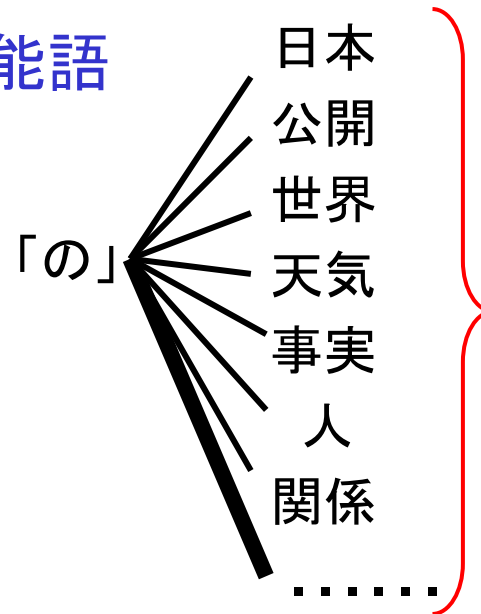
# 前後に隣接する単語の異なり数

内容語

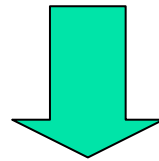


異なり数  
少ない

機能語



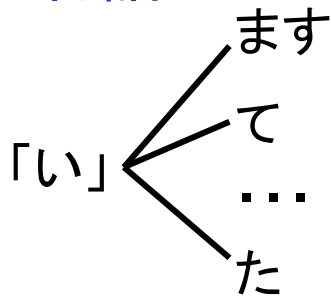
異なり数  
多い



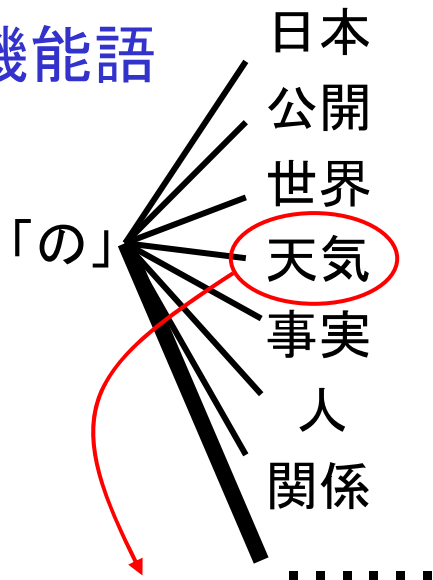
異なり数の多い単語を機能語とする

# エントロピー $H(w_i)$

内容語

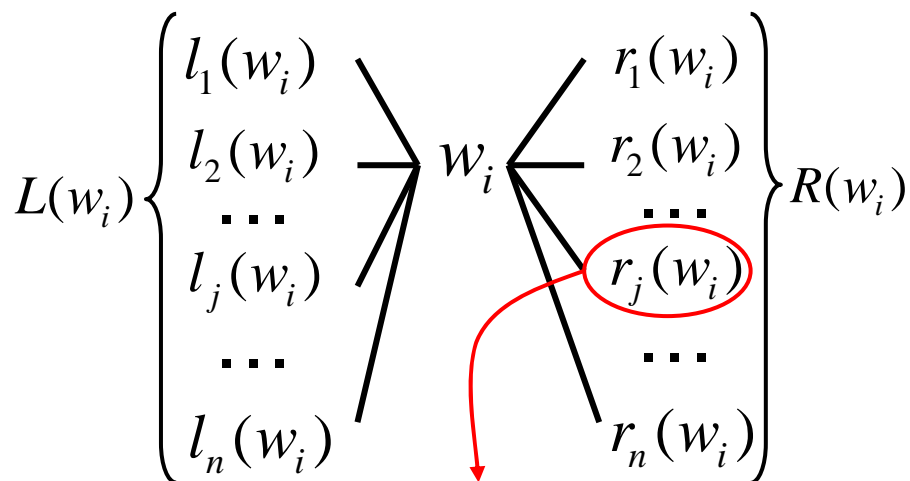


機能語



$\frac{\text{「の」の直後の「天気」の頻度}}{\text{「の」の総出現頻度}}$

単語

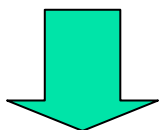


$\frac{r_j(w_i) \text{ の頻度 } f_{r_j}(w_i)}{w_i \text{ の総出現頻度 } F(w_i)}$

$$H(w_i) = - \sum_{l_j(w_i) \in L(w_i)} \left( \frac{f_{l_j}(w_i)}{F(w_i)} \log_2 \frac{f_{l_j}(w_i)}{F(w_i)} \right) - \sum_{r_j(w_i) \in R(w_i)} \left( \frac{f_{r_j}(w_i)}{F(w_i)} \log_2 \frac{f_{r_j}(w_i)}{F(w_i)} \right)$$

# エントロピー再計算 (基本的なアイデア)

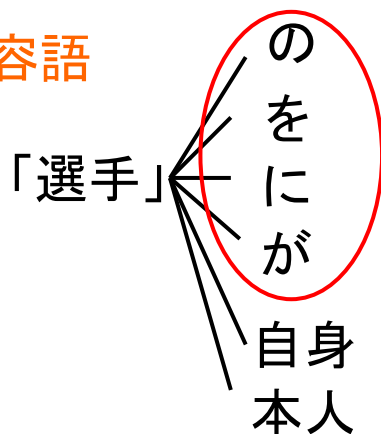
- 機能語は連続しにくい
- 内容語は隣に機能語を持ちやすい



エントロピーの計算結果で機能語としたもの (= 仮機能語) をひとつにまとめる

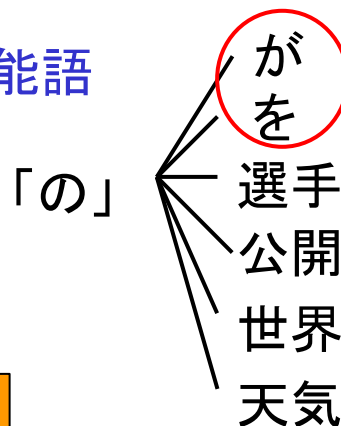
機能語と認定					
順位	単語	順位	単語	順位	単語
1	の	11		103	し
				...	
				43	日本
9		49	選手	...	
10	と	50			い

内容語



6単語  
↓  
3単語

機能語



6単語  
↓  
5単語



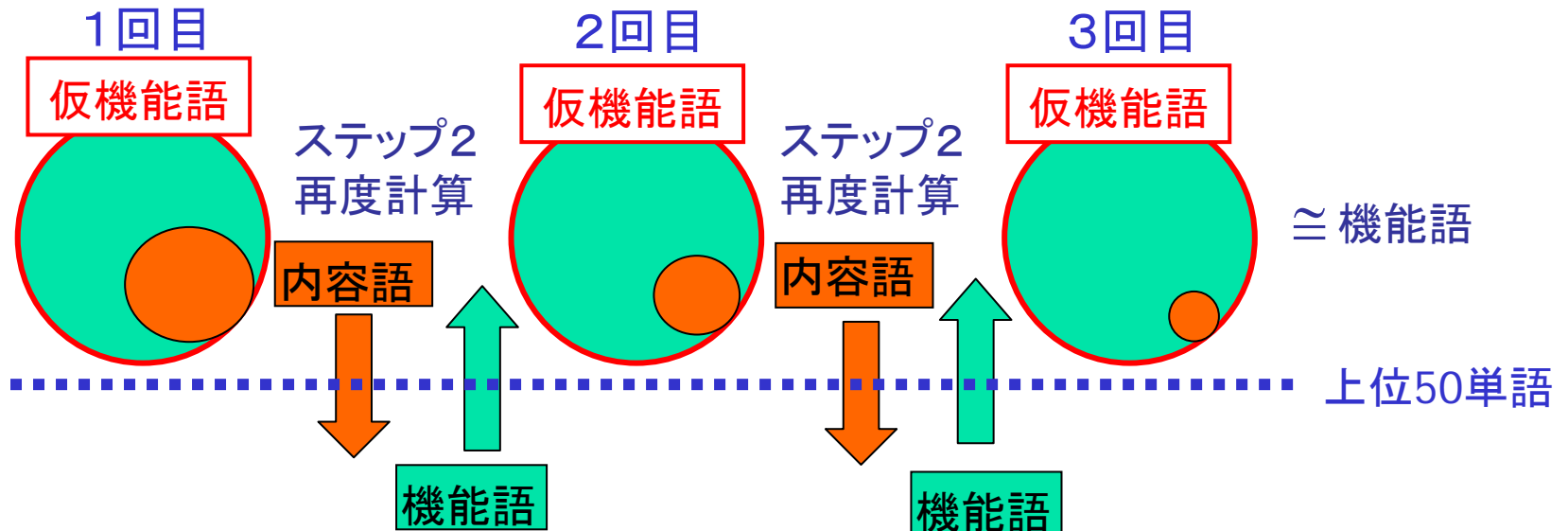
まとまった分だけエントロピーの値が低くなる

# エントロピー再計算(アルゴリズム)

- ステップ1: エントロピーの高い上位50語を求める
- ステップ2: 上位50語を仮機能語としてひとつにまとめてエントロピーの計算を再度行う
- ステップ3: ステップ1と2を繰り返す

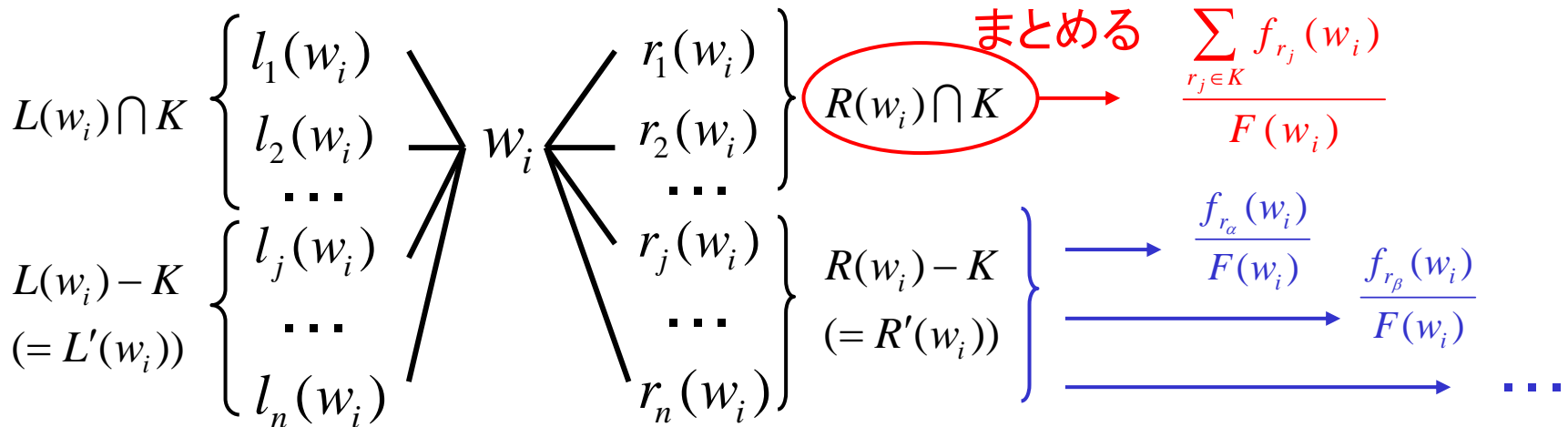
機能語

内容語



# エントロピー再計算 $H'(w_i)$

仮想機能語  $K = \{k_1, k_2, \dots, k_i, \dots, k_m\}$



$$\begin{aligned}
 H'(w_i) = & - \sum_{l_j(w_i) \in L'(w_i)} \left( \frac{f_{l_j}(w_i)}{F(w_i)} \log_2 \frac{f_{l_j}(w_i)}{F(w_i)} \right) - \frac{\sum_{l_j \in K} f_{l_j}(w_i)}{F(w_i)} \log_2 \frac{\sum_{l_j \in K} f_{l_j}(w_i)}{F(w_i)} \\
 & - \sum_{r_j(w_i) \in R'(w_i)} \left( \frac{f_{r_j}(w_i)}{F(w_i)} \log_2 \frac{f_{r_j}(w_i)}{F(w_i)} \right) - \frac{\sum_{r_j \in K} f_{r_j}(w_i)}{F(w_i)} \log_2 \frac{\sum_{r_j \in K} f_{r_j}(w_i)}{F(w_i)}
 \end{aligned}$$



# 実験結果A

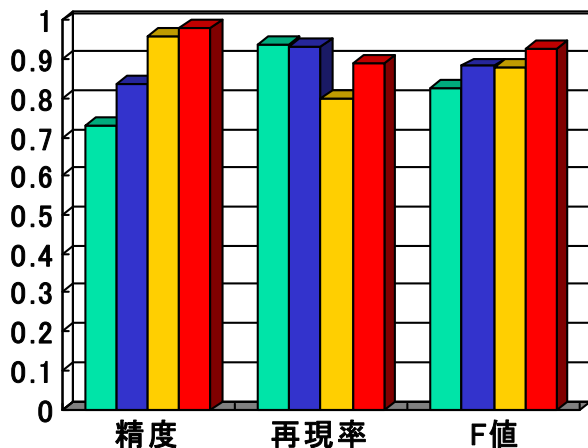
赤文字語数: 機能語と判定された語の数  
 青文字%: 文章全体に占める機能語の割合

言語	手法	上位50	上位200	言語	上位50	上位200
ロシア語  397語 21%	出現頻度	23	39	イタリア語  164語 36%	29	60
	異なり数	25	49		38	80
	エントロピー	22	55		42	78
	エントロピーの再計算	25	65		43	82
インドネシア語  [152語] [11%]	出現頻度	18	42	マレー語  [114語] [18%]	14	24
	異なり数	25	49		18	27
	エントロピー	29	58		18	30
	エントロピーの再計算	31	67		20	34

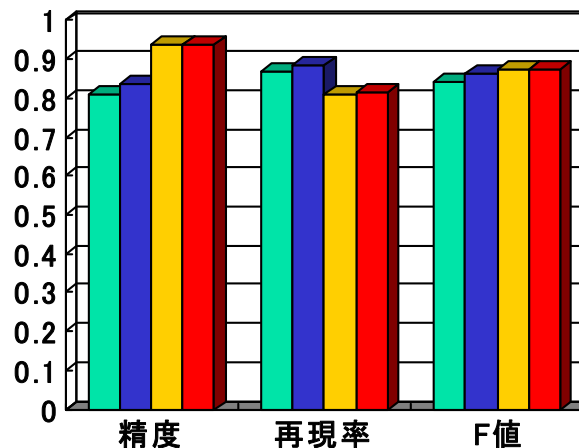
# 評価実験B(上位50単語=機能語)

- 出現頻度
- 異なり数
- エントロピー
- エントロピー再計算

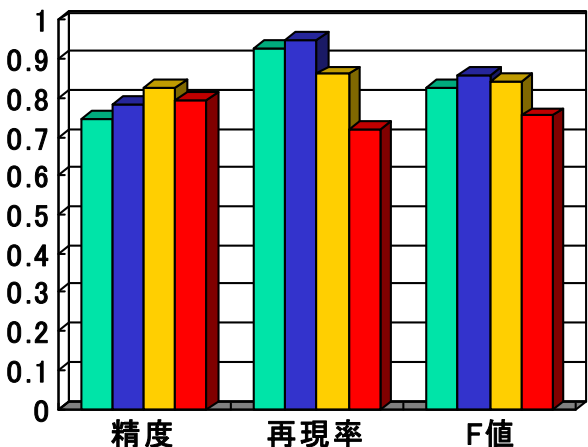
日本語



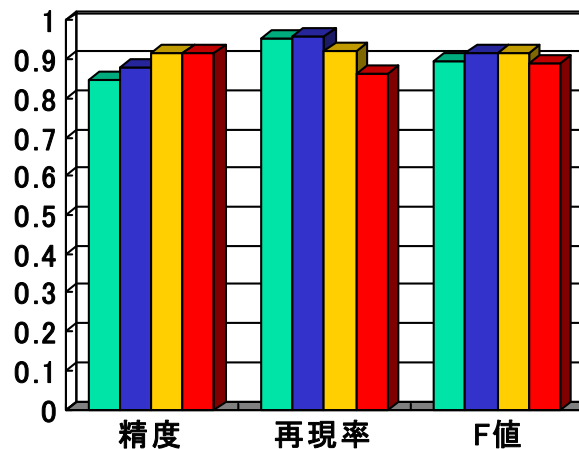
英語



フランス語



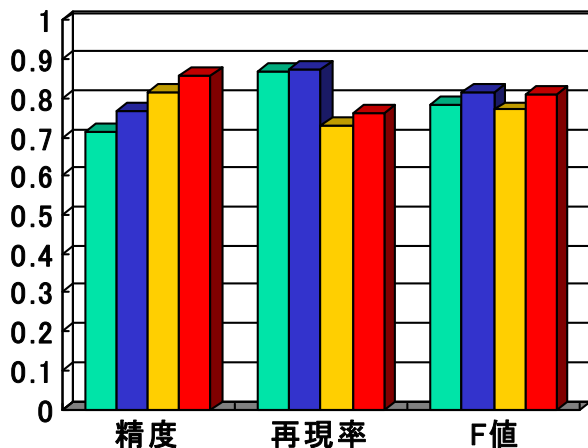
スペイン語



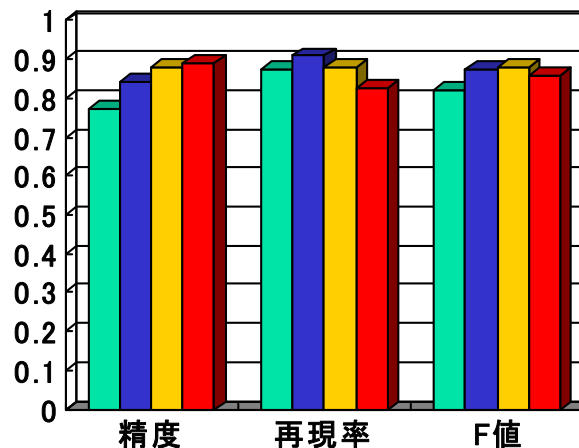
# 評価実験B(上位50単語=機能語)

- 出現頻度
- 異なり数
- エントロピー
- エントロピー再計算

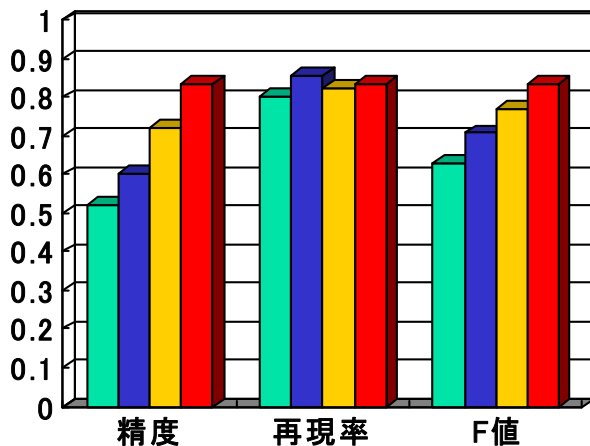
ロシア語



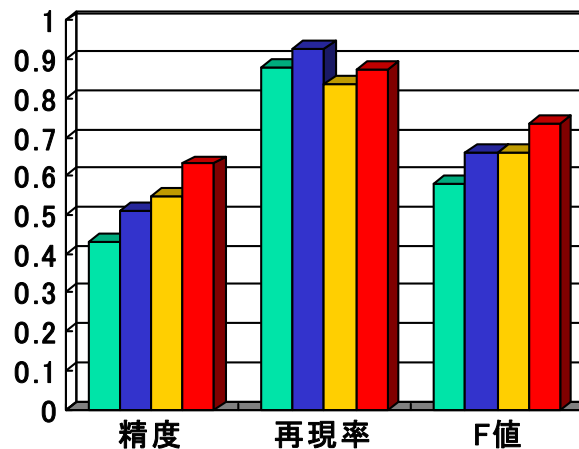
イタリア語



インドネシア語



マレー語





## まとめ

---

- 言語が違ってても手法による傾向は似ている
- 一定の性能では、機能語と内容語を分類が可能

## 今後の課題

- 機能語の割合の差、閾値の考え方
- 基本的手法からの発展、後処理
- 分かち書きされていない語は??