

# 分布類似度のための文脈素性選択

萩原 正人, 小川 泰弘, 外山 勝彦 (名古屋大学)

## 背景

### 分布類似度

- 語の文脈の共通性を利用した類似度
- 分布仮定に基づいて計算  
→ 「文脈の類似した語は意味も類似している」

### 分布類似度の計算

- 語の文脈をコーパスから抽出・利用
  - 依存関係, 近接語, 依存パス, etc.
- 大量の共起データ → **効率化が必要**

## アプローチ

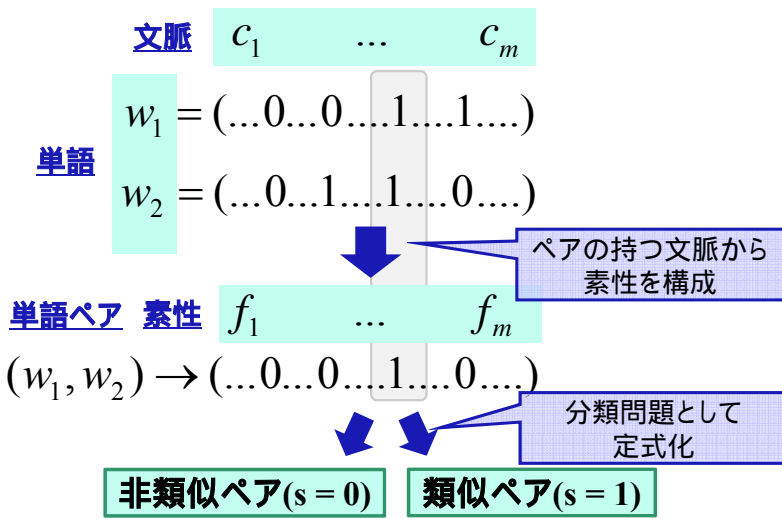
### 素性選択

- 文書分類の分野において広く提案・利用
- 分布類似度 ≠ 分類問題  
→ 既存の素性選択手法が適用できない

### 目的

**素性選択手法を  
分布類似度問題に適用し  
有効性を検証**

## 文脈選択手法



### 素性選択指標

- 文書頻度 (DF)  $df(c) = |\{w \mid N(w, c) > 0\}|$
- 索引語強度 (TS)  $s(c) = P(c \in C(w_2) \mid c \in C(w_1))$
- 相互情報量 (MI)  $I(f, s) = \log \frac{P(f, s)}{P(f)P(s)}$
- 情報利得 (IG)  $G(c_j) = \sum_{f_j} \sum_s P(f_j, s) \log \frac{P(f_j, s)}{P(f_j)P(s)}$
- $\chi^2$  統計量 (CHI2)  $\chi^2(c_j) = \frac{N(F_{11}F_{00} - F_{01}F_{10})}{(F_{11} + F_{01})(F_{10} + F_{00})(F_{11} + F_{10})(F_{01} + F_{00})}$

- 参照セットに対して各指標を計算
- 指標値の小さい文脈素性を除去

## 評価実験

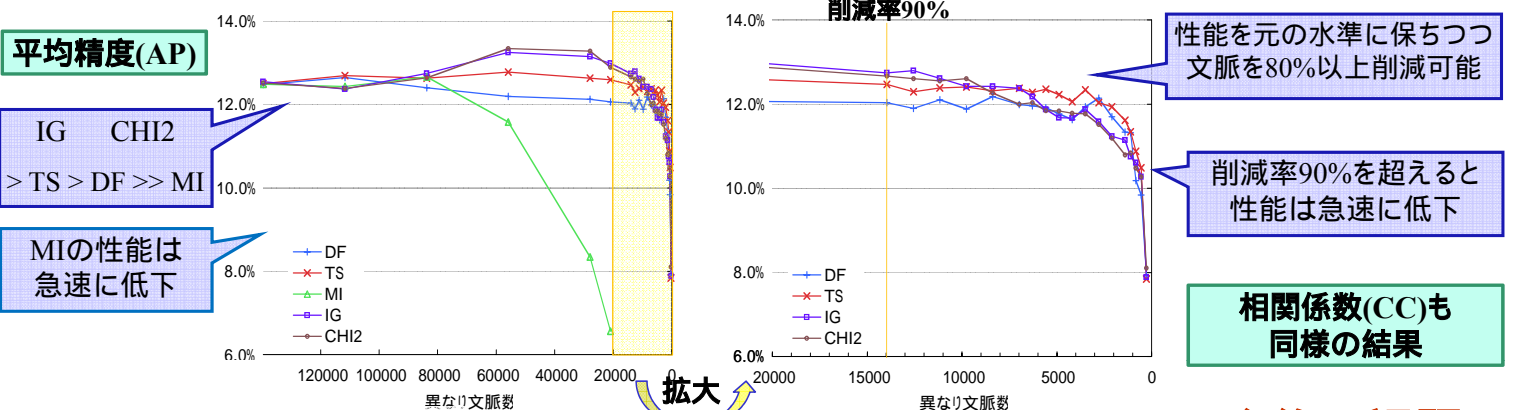
### 類義語自動獲得タスクによる性能評価

- 文脈: 依存関係 (パーサー-RASP2により抽出)
- 重み関数: 相互情報量, 類似度指標: Jaccard係数

### 評価指標

- 既存のシソーラスを正解とした平均精度 (AP)
- WordNetによる類似度との相関係数 (CC)

## 結果



### 文脈カテゴリへの適用

- IGをカテゴリ重要度へと拡張
- 性能と強い相関 → カテゴリ選択に対しても有効

### 今後の課題

#### 他の種類の文脈・タスクへの適用

- 近接語, 依存パス
- 語義曖昧性解消