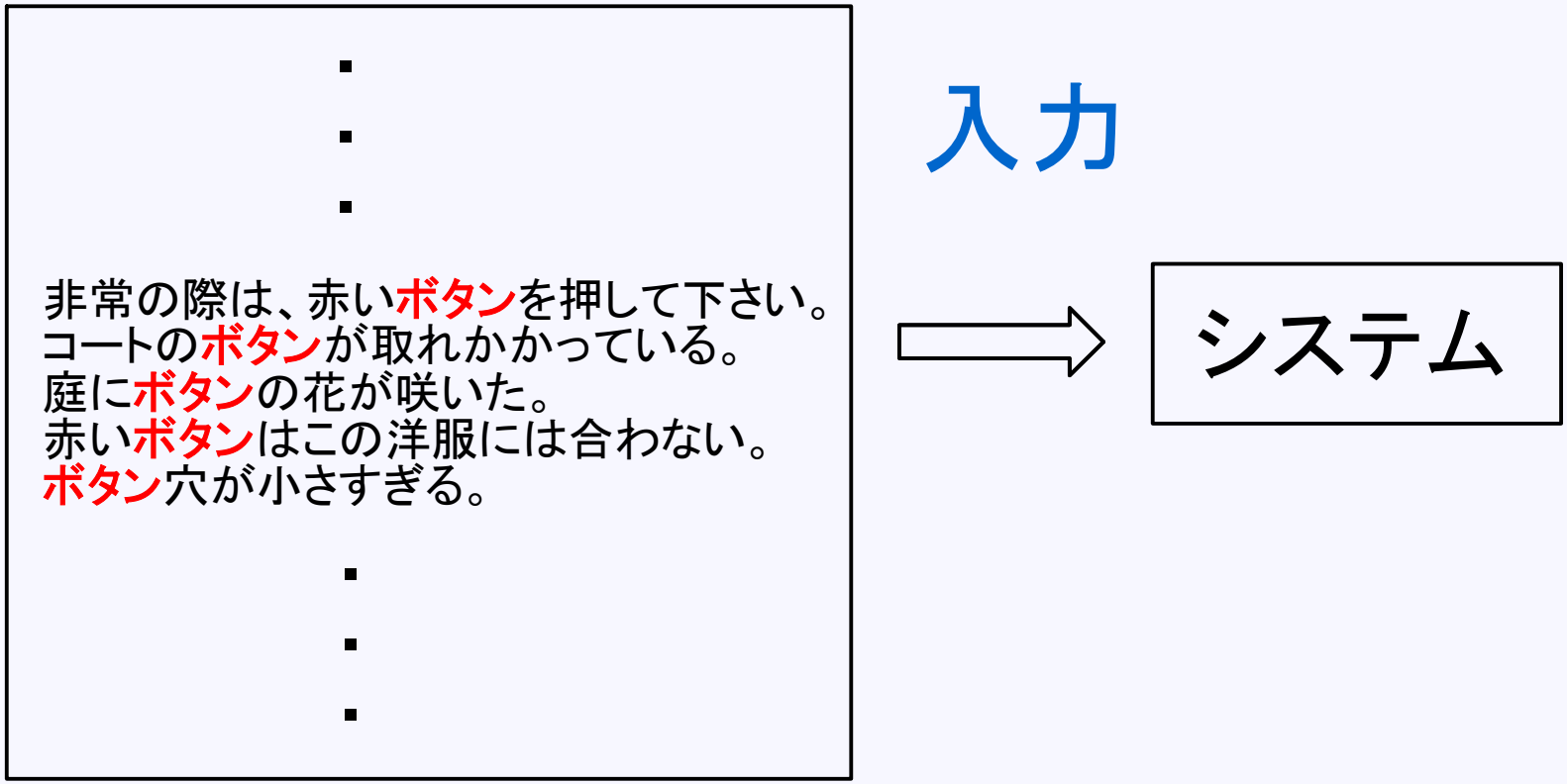
用例間類似度測定のための属性重みの推定

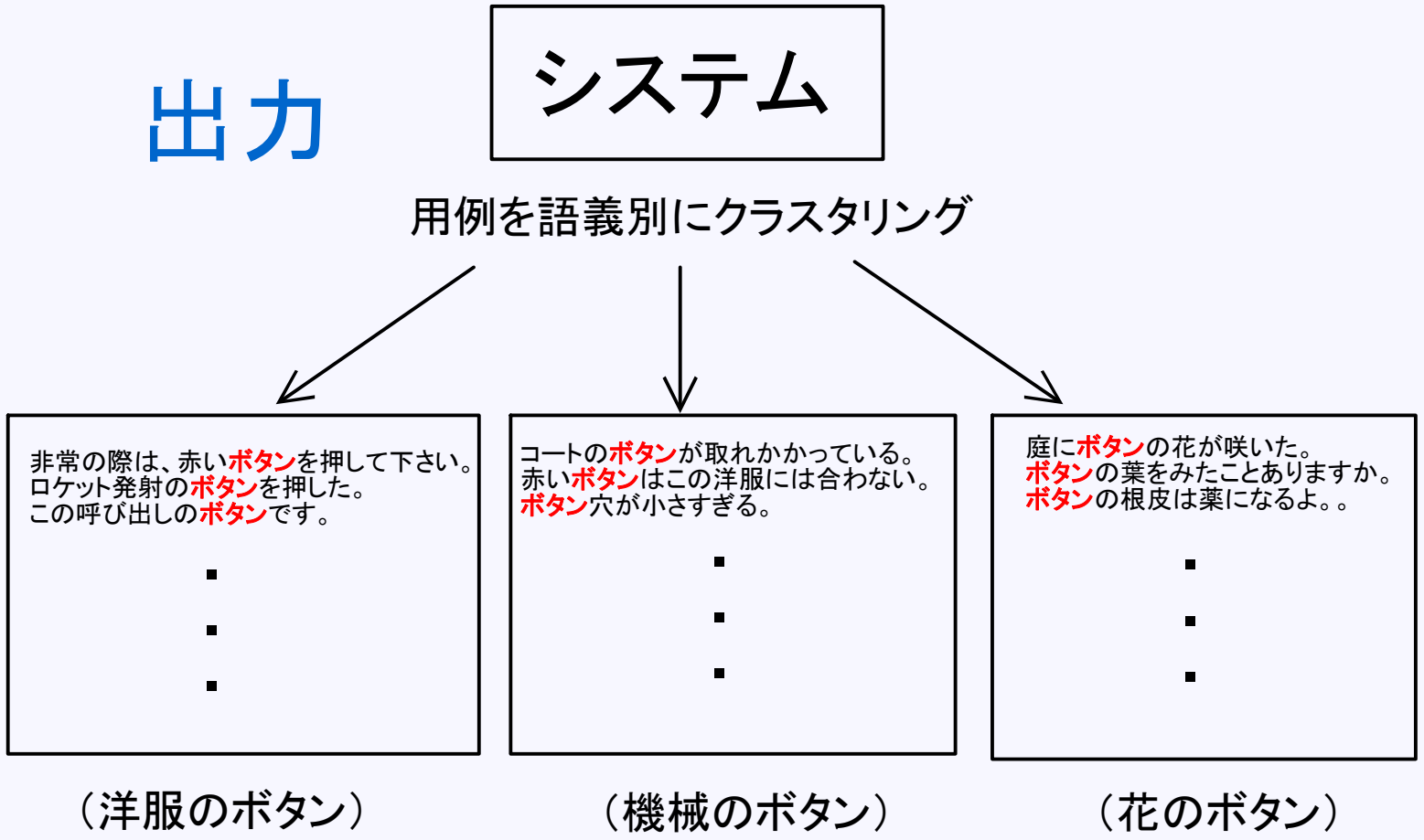
新納浩幸・佐々木稔（茨城大学）

用例のクラスタリング（1）

用例



用例のクラスタリング（2）



背景

目的　語義別用例の収集

語義識別規則の学習

動詞の格フレームの獲得

語義別のシソーラスの作成

辞書の編纂、語学学習

・・・・

利用可能

アプローチ

語義識別（教師有り学習）

⇐ 語義のセットが予め準備できない

クラスタリング（教師なし学習）

⇐ 語義の数の推定が困難

⇒ 半教師有りクラスタリング

問題点

本質的問題はクラスタリング手法ではなく、
用例間の類似度の設定

非常の際は、赤いボタンを押して下さい

この洋服に赤いボタンは似合わない

類似度はいくつ？

ここでやったこと

用例間の類似度の設定方法を提案

50個の名詞に対する用例のデータセットを
提案手法の類似度でクラスタリング

⇒ 提案手法は有効

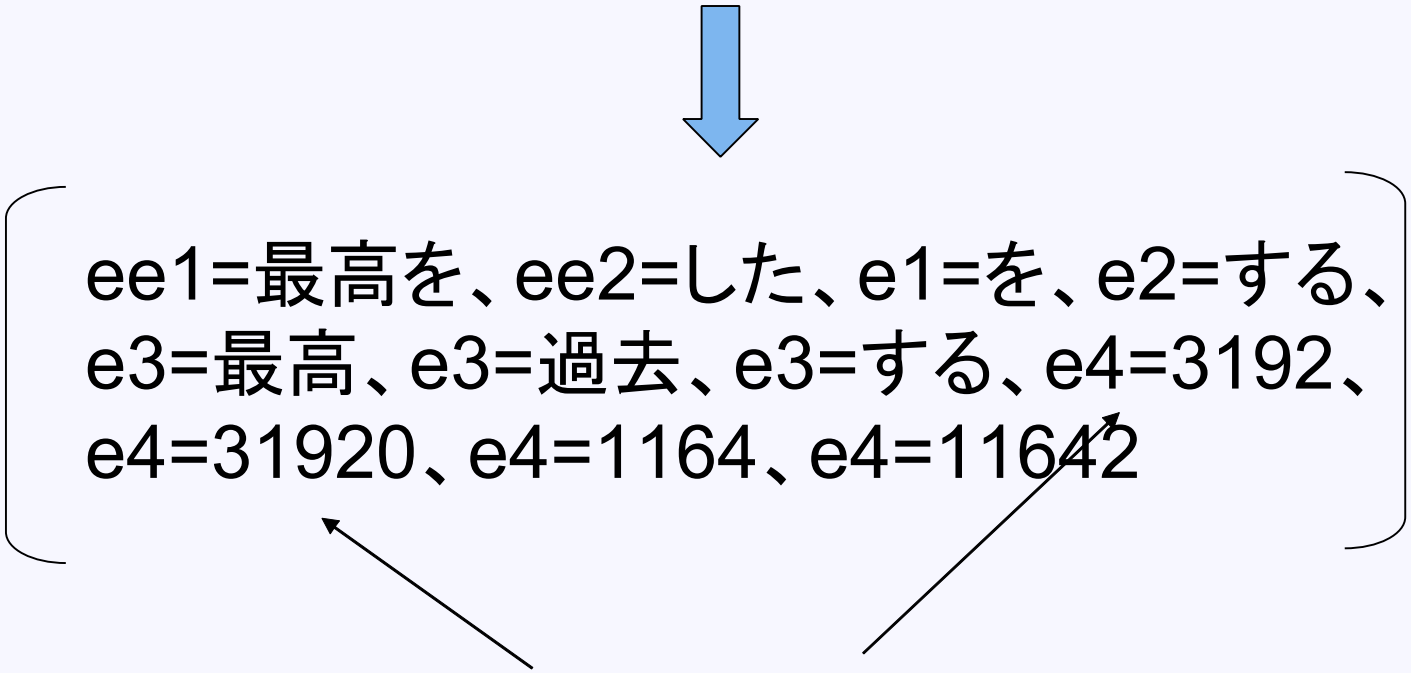
用例を素性リストへ（１）

用例を素性リストで表現
素性リスト間の類似度を定義

６つの素性
- ee1: 直前の2単語
- ee2: 直後の2単語
- e1: 直前の単語
- e2: 直後の単語
- e3: 前方と後方の内容語
  それぞれ2つまで
- e4: e3 の分類語彙表の番号

用例を素性リストへ（2）

過去 / 最高 / を / 記録 / する / た / 。



素性リスト間の類似度のモデル

線形モデル

$$sim(s_1, s_2) = a \cdot M(ee1) + b \cdot M(ee2) + c \cdot M(e1) + d \cdot M(e2) + e \cdot M(e3) + f \cdot M(e4)$$

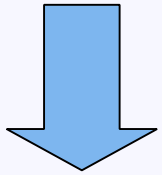
$M(X)$: 素性 X の一致数

$a = b = c = d = e = f = 1$ 一般の類似度

重みの学習

$$sim(s_1, s_2) = a \cdot M(ee1) + b \cdot M(ee2) + c \cdot M(e1) + d \cdot M(e2) + e \cdot M(e3) + f \cdot M(e4)$$

a, b, c, d, e, f の値（素性の重み）が問題

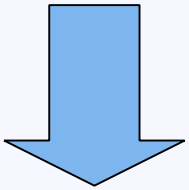


学習により設定

アプローチ

訓練データがあれば重みの推定は容易

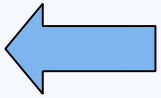
しかし本タスクでは訓練データの作成は不可能
(類似度の値は人間でもわからない)



重みを経験的に与え、擬似的な訓練データを
作成し、それを用いて重みを推定

WSDの訓練データの利用

WSDの訓練データ



同じ語義かどうかはわかる

用例のペア

異なる語義なら類似度は０、
同じ語義なら経験的重みと線形
モデルから類似度を与える

線形モデルの訓練データを構築

最小2乗法（1）

用例対と類似度 → 線形モデルの訓練データ

$$\begin{cases} 2a+b+c+3d+4e+f=8 \\ d+e+2f=0 \\ \cdots \\ a+3c+d+e+2f=5 \end{cases}$$

ａ～ｆ の推定は重回帰分析の手法が利用可能

最小2乗法

最小2乗法(2)

$$\sum_{i=1}^{n}\left(y^{(i)} - \sum_{j=1}^{6} a_j x_j^{(i)} \right)^2 \longrightarrow \text{最小化}$$

極値問題として解ける

$$\begin{bmatrix} a \\ b \\ \vdots \\ f \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_6 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{16} \\ S_{21} & S_{22} & \cdots & S_{26} \\ & & \ddots & \\ S_{61} & S_{62} & \cdots & S_{66} \end{bmatrix}^{-1} \begin{bmatrix} S_{1y} \\ S_{2y} \\ \vdots \\ S_{6y} \end{bmatrix}$$

実験（重み推定）

WSDの訓練データ:

SENSEVAL2辞書タスクの50個の名詞に対する訓練データ（846,045用例対）

経験的な重み:

a=b=10 , c=d=5, e=f=1

最小2乗法による推定値:

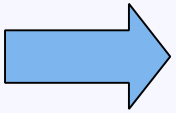
a = 8.1987　b = 8.0044　c = 3.3696

d = 3.8949　e = 1.0512　f = 0.4125

実験（クラスタリング）

テストデータセット：

SENSEVAL2辞書タスクの50個の名詞に対する訓練データ



50 個のテストデータセット

利用した類似度：

(1) 単純な類似度（a=b=c=d=e=f=1）

(2) 経験的重み（a=b=10 , c=d=5, e=f=1）

(3) 提案手法（推定した a～f）

実験結果（1）

エントロピーによる評価：
（値が小さい方が良いクラスタリング）

50個のテストデータセットに対するクラスタリング結果のエントロピーの平均値

(1) 単純な類似度： 0.4824
∨
(2) 経験的重み： 0.4786
∨
(3) 提案手法： 0.4767

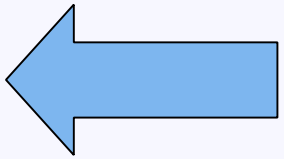
実験結果(2)

単語	用例数	語義数	類似度(1)	類似度(2)	類似度(3)
(1)間	266	9	0.368	0.384	0.383
(2)頭	169	6	0.524	0.505	0.511
(3)一般	267	5	0.635	0.652	0.649
…	…	…	…	…	…
(50)問題	636	4	0.114	0.115	0.117
平均	**278.4**	**3.76**	**0.4824**	**0.4786**	**0.4767**

考察（各データセットでみると・・・）

50個のデータセット中、最良のクラスタリング結果を出した個数

(1) 単純な類似度：　25個
(2) 経験的重み：　15個
(3) 提案手法：　19個



実は単純なものが安定して良い結果を出す

(3) が (2) よりも良いということが重要

考察（用例によるWSD）

用例によるWSDは用例と入力文との類似度を測る必要がある

動詞の場合、格フレームの選択問題に帰着

$$\sum w_{stype} \times \left(\text{入力文中の格要素と用例中の格要素のシソーラスによる類似度} \right)$$

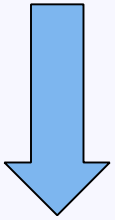
w_{stype}：表層格に応じて与えられる重み

→ 線形モデル

考察（経験的パラメータ値は必要？）

判別分析やSVMなど により
線形モデルのパラメータは推定可能

教師信号は同じクラスかどうかだけですむ
経験的パラメータ値は必要ない



線形モデルでは、うまくはいかない
多くの用例対の類似度は0
アンバランスな訓練データ

考察（距離学習）

距離学習

教師あり
教師なし

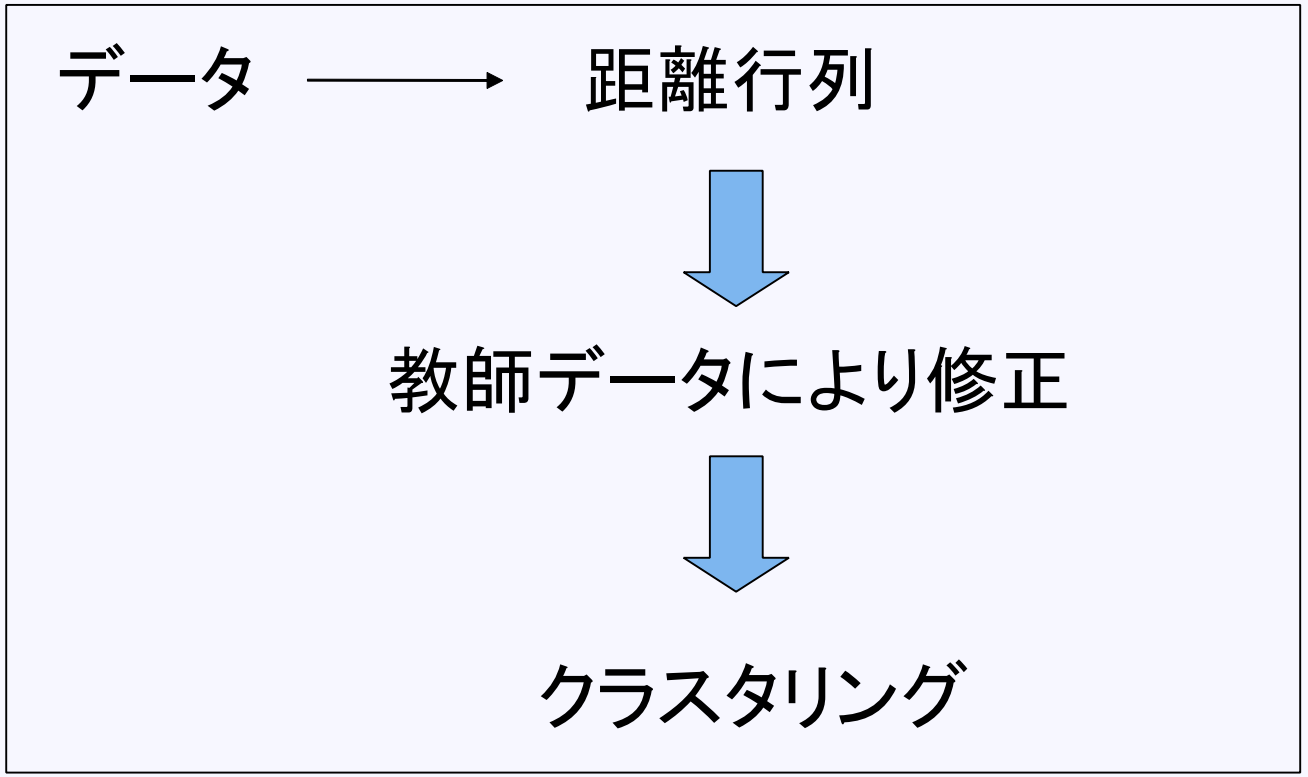
本タスク

次元縮約、結局、(ソフト)クラスタリング
解決にはならない（？）

本手法は教師なしに背景知識を利用した
という位置づけ

考察（半教師有りクラスタリング）

距離修正型
(CCL など)



本手法もこの形の一種

まとめ

- 単語用例のクラスタリングがタスク
- このタスクは用例間類似度の設定が鍵
- 線形モデルのパラメータを推定

  WSDの訓練データを利用
  経験的なパラメータ値から訓練データを作成
  最小2乗法からパラメータを推定

- 50データセットの実験から本手法の有効性を示した
- 単語間類似度表の作成が今後の課題

問題点（再）

本質的問題はクラスタリング手法ではなく、
用例間の類似度の設定

非常の際は、赤いボタンを押して下さい

この洋服に赤いボタンは似合わない

類似度はいくつ？