自己組織化マップSOMによる心情を表すオノマトペの意味分類と可視化

著者：中村沙織 (広島市立大学 情報科学部)
黒澤義明 (広島市立大学大学院 情報科学研究科)
竹澤寿幸 (広島市立大学大学院 情報科学研究科)

発表者：三上健太

はじめに

- オノマトペ
  - 擬音語・擬態語の総称
  - 感覚的な表現(語義が曖昧)
  - 種類が多い
  - 絶えず新語が作られる(辞書に載ってない)
- 理解困難さを解消したい
- 新語への対応を容易にしたい
- そこで・・・

  オノマトペを自動分類するシステム(Webより用例抽出)

  先行研究(k-means法でクラスタリングする手法)と比較

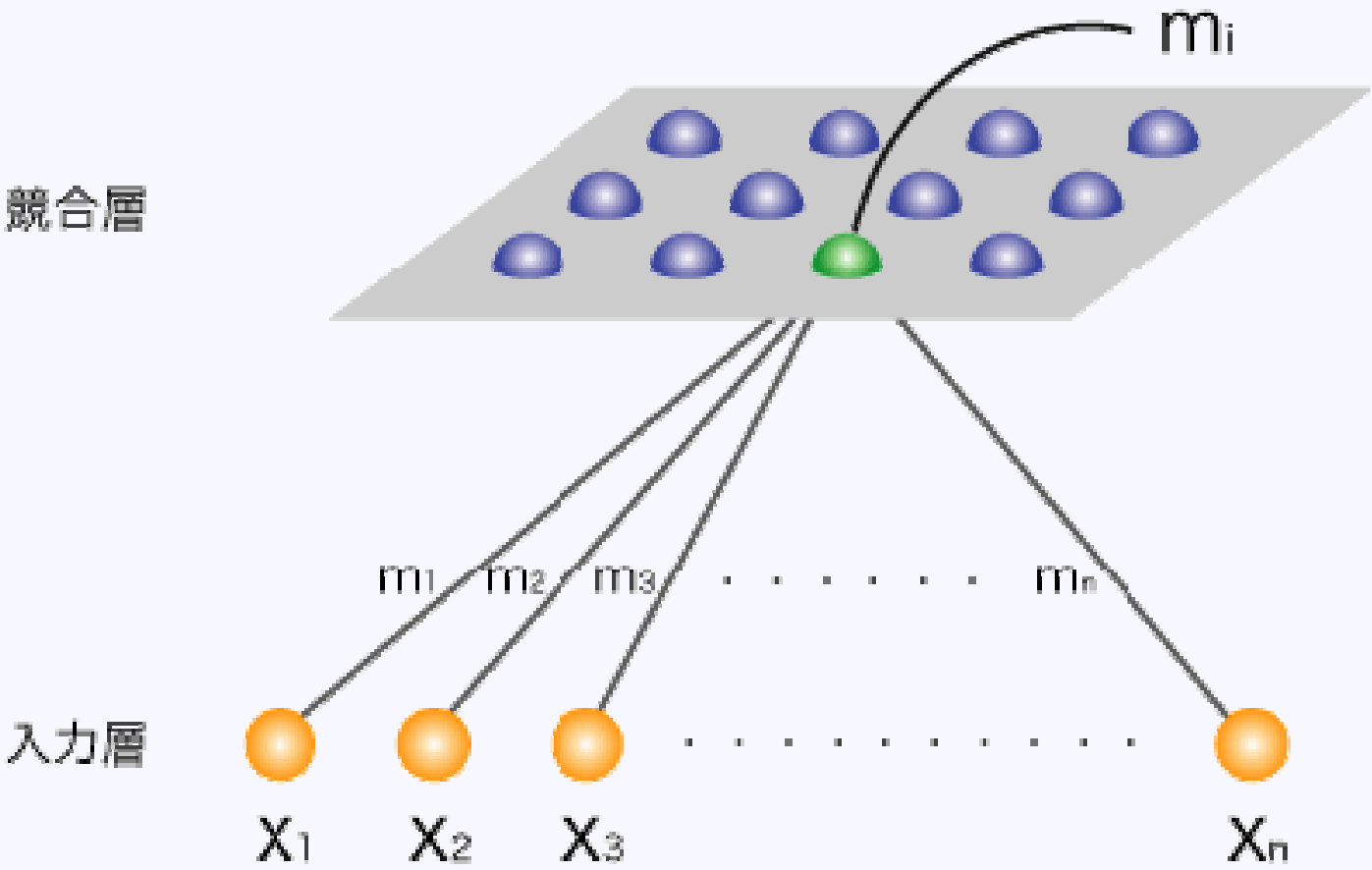
オノマトペの可視化

- オノマトペは係っている動詞によって意味が大きく異なる
  → ・動詞項構造シソーラスを用いる
    ・対象となるオノマトペと共起性が高い動詞を意味概念によって分類
    ・動詞の概念ごとに可視化
- 可視化手法に自己組織化マップ(SOM)を用いる
- 自己組織化マップ(Self-Organizing Map:SOM)
  - 多次元ベクトルで表されたデータを2次元マップ上に写像
  - 多次元データの特徴,データ間の相互関係を保っている
  - 2次元マップでの表現 → 視覚的に理解しやすい

自己組織化マップ

- 階層型ニューラルネットワークの一種,2層のネットワーク
- 第1層はn次元の入力層x(t),第2層は競合層(一般的に2次元配列となっている)
- 競合層のベクトルは参照ベクトル$m_i(t)$で表現され,n個の要素をもつ
- 入力層への特定の入力により,競合層の特定の領域が反応するような学習が行われる
- 教師なし学習
- 学習にユークリッド距離を用いる

自己組織化マップ-概念図



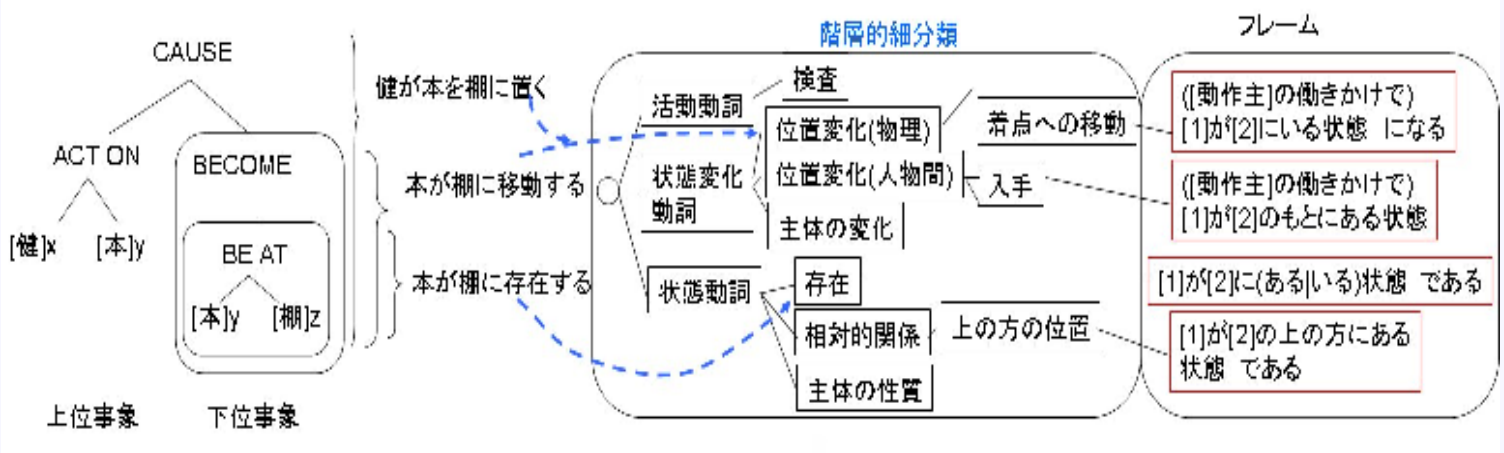
研究の流れ

- SOMによる分類を行うための用例としてWebコーパスを使用
- コーパス上のオノマトペとその後方一語に出現する動詞を取得
  → 動詞項構造シソーラスを用いて動詞を概念ごとに分類し
  出現頻度を求める
- データ変換して自己組織化を行う

- 使用コーパス：Web日本語Nグラム
- 対象語：心情を表すオノマトペ228語
  (片仮名・平仮名表記，対象語の後方に「っ」，「ッ」，「っと」，
  「ット」を含む語を実験に用いる)

研究の流れ(2)

- 動詞項構造シソーラスのうち，フレームに着目して分類を行う
  フレーム - 動詞が何を表す(どんな状態なのか)を表すもの



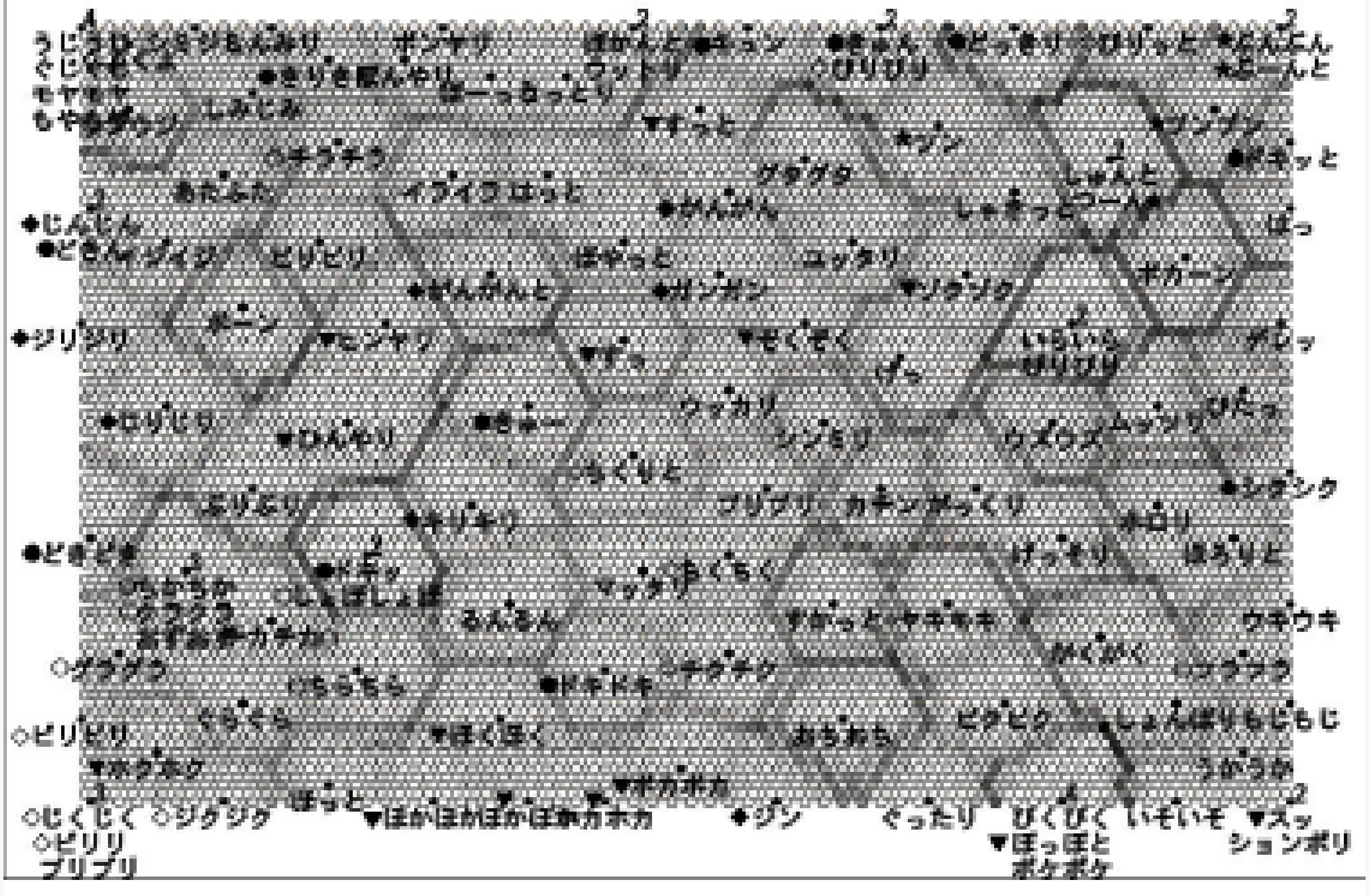
- データの変換
  - 動詞フレームの出現回数をオノマトペごとに出現率で示し，
    ベクトルとした。動詞フレーム軸を $v_1, v_2, v_3, \cdots, v_n$ とする

$$v_i \text{ の出現率} = \frac{v_i}{\sum_{i=1}^{n} v_i}$$

実験

- データはオノマトペ120語，フレーム軸数165となった
- マップサイズ：64×48
- 学習は2段階
  - 実験1：学習回数 100000，　初期学習率係数 0.05
  - 実験2：学習回数 1000000，初期学習率係数 0.01

実験結果

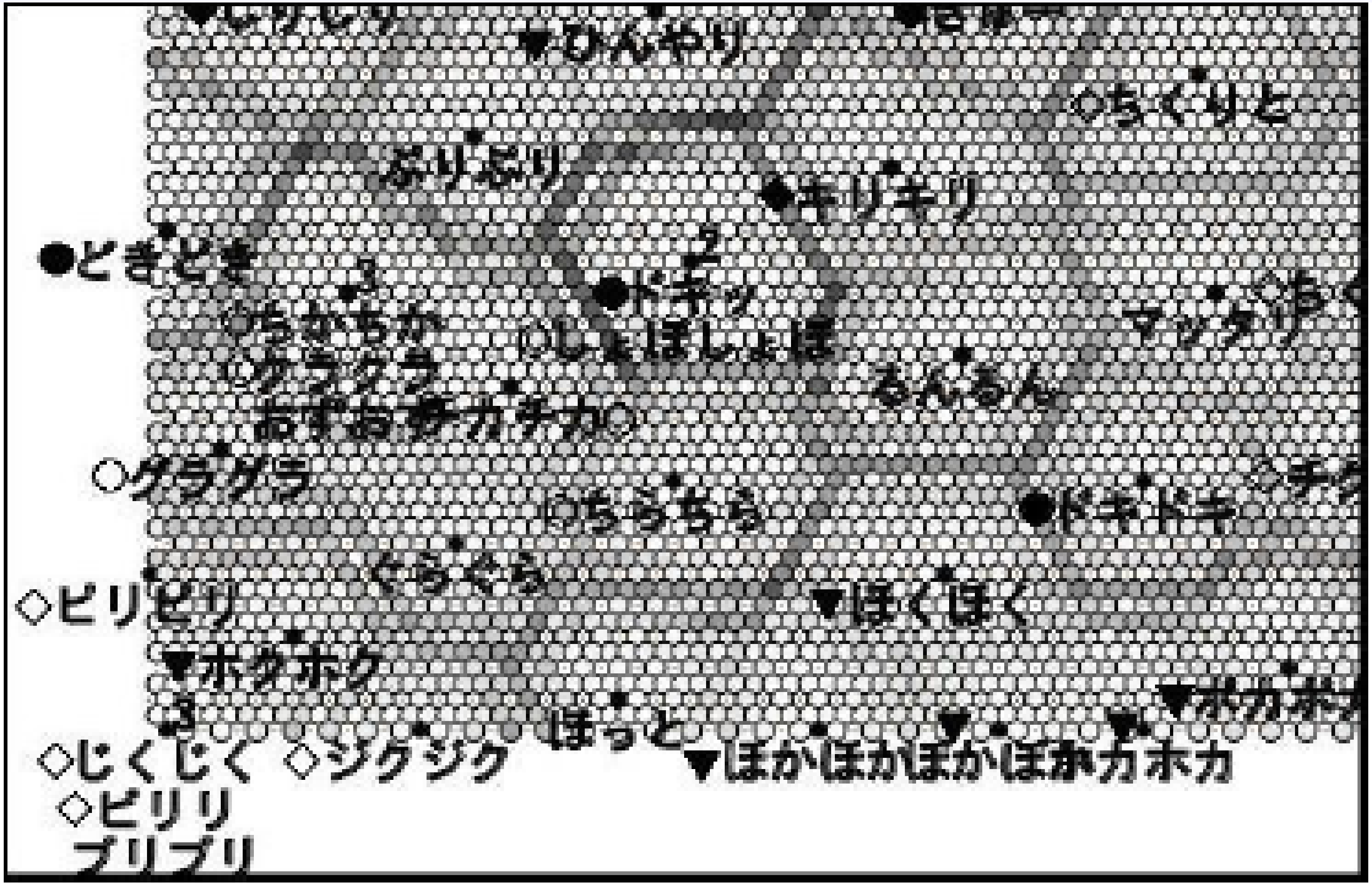


実験結果(2)

- 考察のため，対象語を分類した
- 対象語には以下のように英語で意味が示されている
  - dokidoki 'feeling one's heart throbbing'
  - hinyari 'feeling pleasantly cool'
- 英語の説明文中の特徴を示す単語に着目

特徴	記号	個数	着目した単語
目眩	○	9	eye, dizzy
鼓動	●	9	throb, heart jump
痛み	◆	10	sore, pain, pang,griping
臭い	★	5	smell
温度	▼	14	cool, chilly, cold, hot, warm, glowing
刺激	◇	9	pungent, skin
他		64	

実験結果(3)



ひんやり
ちくりと
ぶりぶり
キリキリ
どきどき
しょぼしょぼ
るんるん
グラグラ
ちらちら
ドキドキ
ぐらぐら
ピリピリ
ほくほく
ホクホク
ほっと
ほかほかぽかぽかホカホカ
じくじく
ジクジク
ピリリ
ブリブリ

考察

- 前述のグループに対して,グループに属するオノマトペごとの軸のうち最もグループ全体に影響している軸をそれぞれ3つ選び,精度・再現率を検証
- 形態素解析誤りによるデータの誤差を除外するため,閾値を設定する(1/4,1/3,1/2)
- 閾値以下となる値の小さいベクトルの語彙を削除

考察(2)

	目眩	鼓動	痛み	臭い	温度	刺激	目眩	鼓動	痛み	臭い	温度	刺激
	閾値なし						閾値1/4					
精度 %	52.9	23.1	19.2	36.4	66.7	20.7	77.8	36.4	20.0	57.1	70.0	42.9
	9 / 17	6 / 26	5 / 26	4 / 11	8 / 12	6 / 29	7 / 9	4 / 11	4 / 20	4 / 7	7 / 10	6 / 14
再現率 %	100.0	66.7	50.0	80.0	57.1	66.7	77.8	44.4	40.0	80.0	50.0	66.7
	9 / 9	6 / 9	5 / 10	4 / 5	8 / 14	6 / 9	7 / 9	4 / 9	4 / 10	4 / 5	7 / 14	6 / 9
	閾値1/3						閾値1/2					
精度 %	87.5	44.4	21.1	57.1	70.0	46.2	85.7	66.7	25.0	80.0	70.0	60.0
	7 / 8	4 / 9	4 / 19	4 / 7	7 / 10	6 / 13	6 / 7	4 / 6	4 / 16	4 / 5	7 / 10	6 / 10
再現率 %	77.8	44.4	40.0	80.0	50.0	66.7	66.7	44.4	40.0	80.0	50.0	66.7
	7 / 9	4 / 9	4 / 10	4 / 5	7 / 14	6 / 9	6 / 9	4 / 9	4 / 10	4 / 5	7 / 14	6 / 9

- 除外する対象を多くする→精度増,再現率低
  → 目的に応じて閾値の検討が必要
- ex. 第二言語学習者のようなオノマトペについて知識がない場合,精度が高いことが優先

SOM分類の有効性

- 階層的クラスタリング(ユークリッド距離を用いた最遠隣法)
  - クラスタ数:10個
  - 結果がかたよった
- 非階層型クラスタリング(k-means法)
  - クラスタ数:7個(前述のグループ6つ＋その他)
  - 最遠隣法よりはばらつくが，正しく分類されないものもある

  ex. 温度のグループ:寒が正しく分類されなかった

クラスタ	1	2	3	4	5
個数	1	111	1	1	1
語彙	ウッカリ		ぞくぞく	うっとり	すっと
クラスタ	6	7	8	9	10
個数	1	1	1	1	1
語彙	がんがん	ガンガン	マッタリ	じりじり	しみじみ

最遠隣法

クラスタ	1	2	3	4	5	6	7
個数	7	4	2	15	83	4	5
語彙	ほくほく ぽかぽか ジン	ぴりっと ぷんぷん ぷーんと	しゅんと つーん	うじうじ シミジミ ぽかんと	どっきり しみじみ チラチラ	キュン ぴりぴり ツン	あたふた どきん イジイジ

k-means法

動詞項構造シソーラスの有効性

- 対象語の後方の動詞をそのまま軸としてSOMで分類

	動詞項構造シソーラス 使用						動詞項構造シソーラス 不使用					
	目眩	鼓動	痛み	臭い	温度	刺激	目眩	鼓動	痛み	臭い	温度	刺激
精度 %	33.3	23.1	36.8	19.2	66.7	23.3	26.3	50.0	36.4	71.4	42.9	36.8
	8 / 24	6 / 26	7 / 19	5 / 26	8 / 12	7 / 30	5 / 19	4 / 8	8 / 22	5 / 7	9 / 21	7 / 19
再現率 %	88.9	66.7	70.0	100.0	57.1	77.8	55.6	44.4	80.0	100.0	64.3	77.8
	8 / 9	6 / 9	7 / 10	5 / 5	8 / 14	7 / 9	5 / 9	4 / 9	8 / 10	5 / 5	9 / 14	7 / 9

- 動詞項構造シソーラスを用いると・・・
  - 目眩のグループでは精度・再現率が上昇
  - 臭いのグループでは精度が大幅に低下

  → グループにより有効性が異なる

おわりに

- 本手法でのSOMによる有効性が確認された

- 今後の課題
  - 動詞項構造シソーラスによる分類で，1つの動詞に対し複数のフレームが存在する多義語の分類を可能にする
    → グループごとで有効性に差が生じる問題の解決へ