自己組織化マップを用いた自由記述形式アンケートの内容分析のための単語クラスタリングに関する検討

Consideration on Word Clustering for Content Analysis of Essay Questionnaire Using Self-Organizing Map

1.はじめに

アンケート調査は多数の人間の意見をまとめる方法として実施されている。

日常的に行われているアンケート例

公共事業に関するアンケート

積極的に公共事業に住民の意見を取り入れる。

授業評価アンケート

授業の改善に役立てる。

企業に関するアンケート

商品の開発・改善の参考にする。

1.はじめに

記述形式アンケート集計の問題点

・アンケートに全て目を通し分析を行うため、作業時間がかかる。
・分析者によって分析結果が異なる。

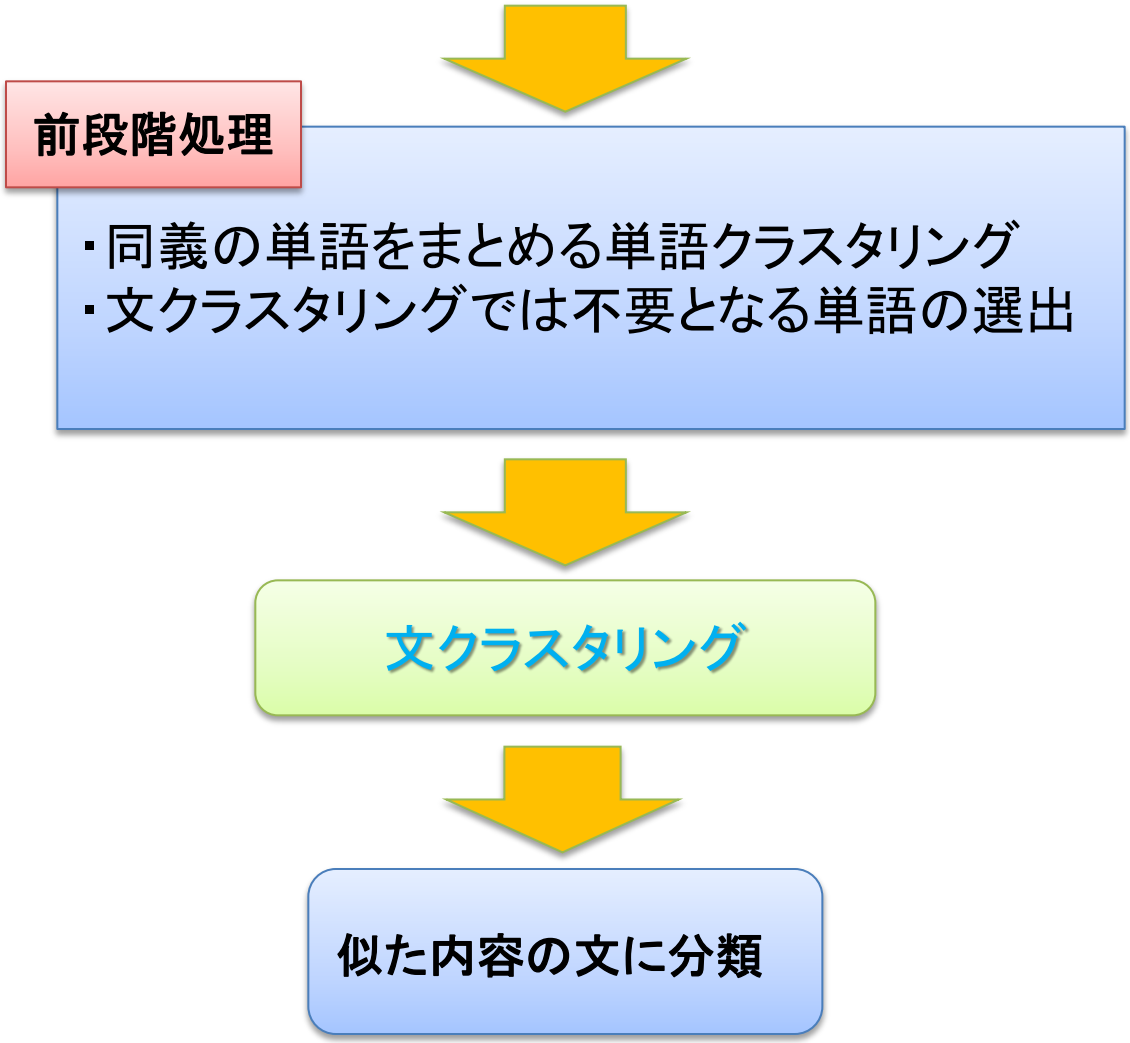
計算機によるアンケートの集計作業の自動化が求められる。

目標

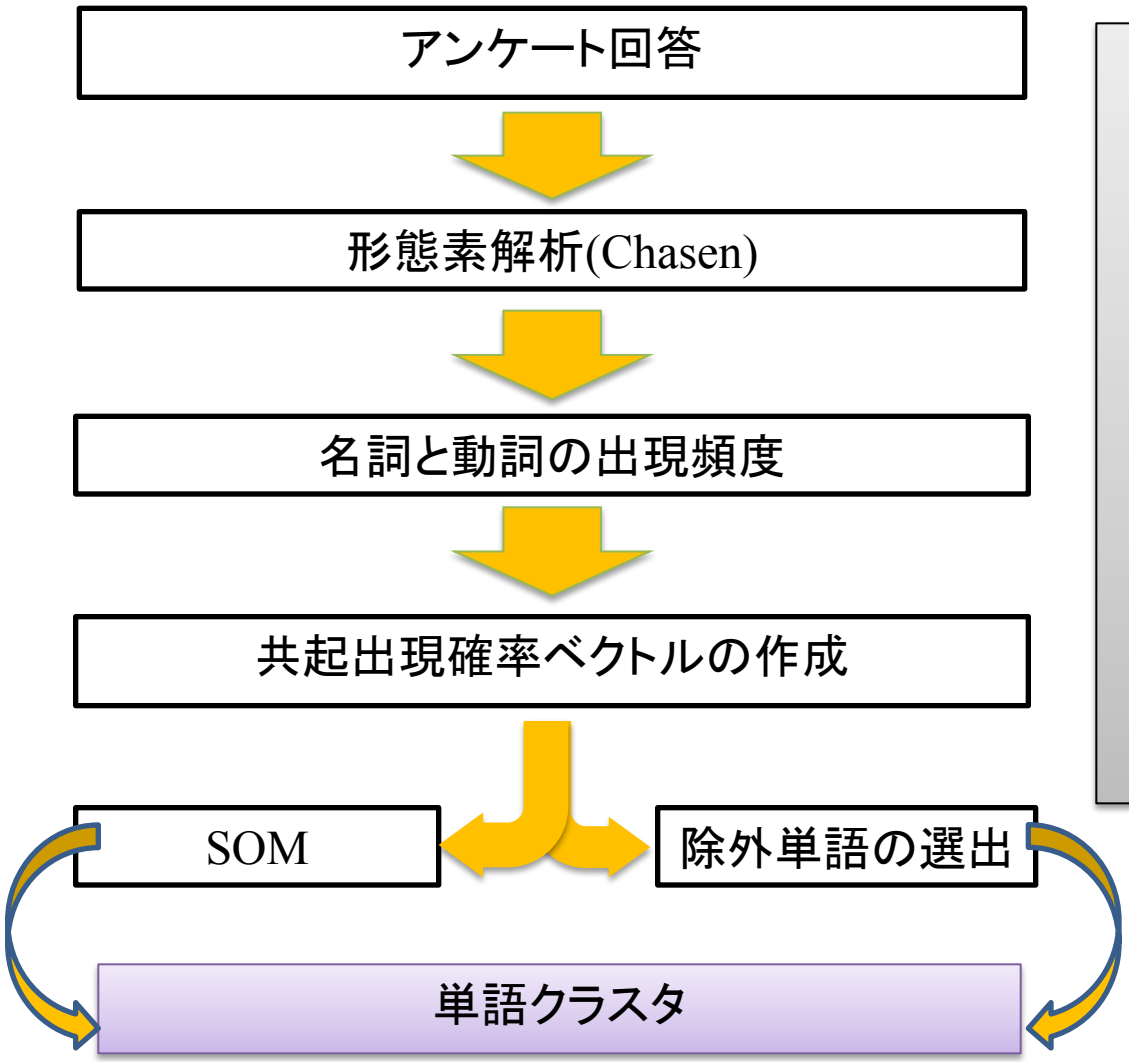
似た内容の文に分類することで、アンケートの内容分析の補助を行う。

2.自由記述形式アンケートの内容分析補助

似た内容の文には**似たような単語**が用いられる可能性が高いと仮定



3.単語クラスタリング手法



形態素解析ソフトChasen
による解析例

例文：文を単語ごとに分割する

文	名詞-一般
を	助詞-格助詞-一般
単語	名詞-一般
ごと	名詞-接尾-一般
に	助詞-格助詞-一般
分割	名詞-サ変接続
する	動詞-自立

図1　形態素解析例

3.単語クラスタリング手法

3.1共起出現確率ベクトルの作成

共起出現確率の作成

単語の共起する範囲が「文」の場合

1:アンケート回答内の単語w_iを含む「文」を抽出
2:文中の単語w_jを含む「文」を抽出
3:単語w_iを含む文A_i 個　単語w_jを含む文B_j個の共起出現確率を算出

$$P_{i,j}^{sente} = \frac{B_j}{A_i}$$

4:同様に単語w_iと共起する単語全ての共起出現確率を算出
5:全ての単語N個分のベクトルを作成

$$\mathbf{P}_{i,}^{sente} = \left[P_{i,1}^{sente}, P_{i,2}^{sente}, \cdots, P_{i,N}^{sente} \right]^T$$

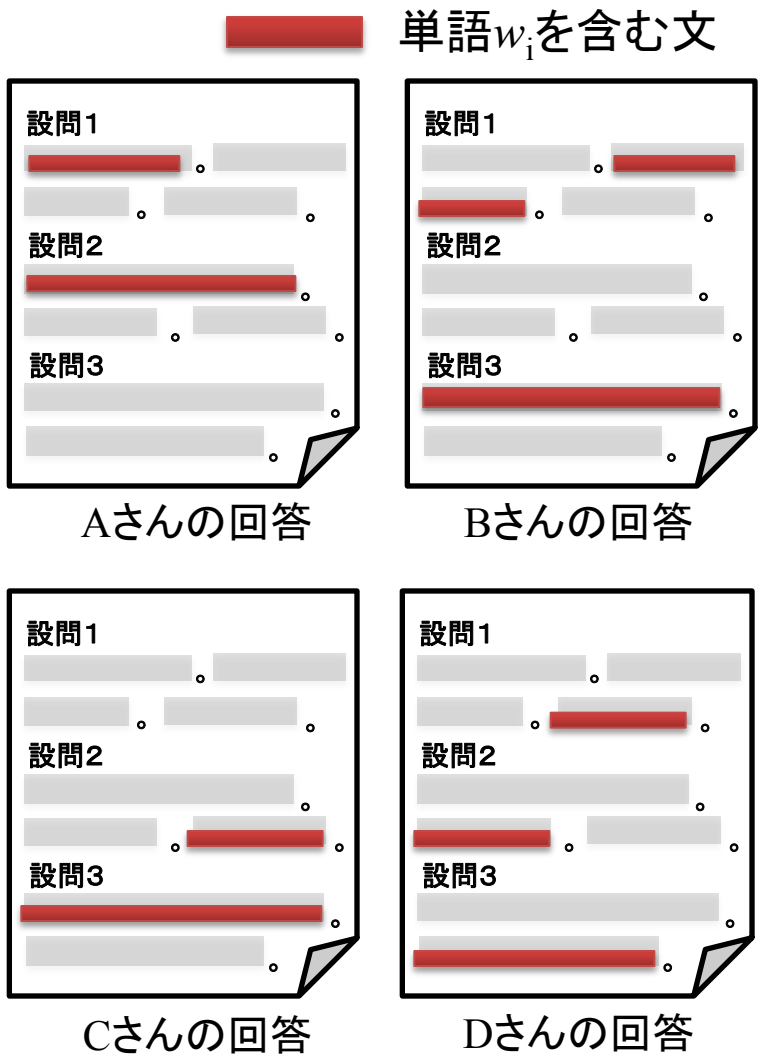


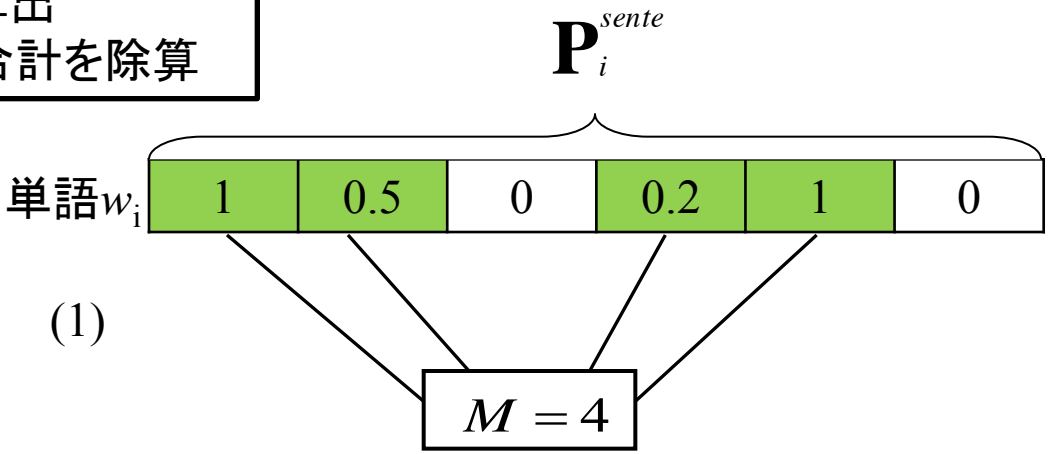
図2　アンケート回答

3.単語クラスタリング手法

3.2単語クラスタリング対象単語の選出

共起出現確率ベクトルの部分平均

1：単語w_iの共起出現確率の合計を算出
2：単語w_iと共起している単語数Mで合計を除算



$$D_{i,} = \frac{P_{i,1}^{sente} + P_{i,2}^{sente} + \cdots + P_{i,N}^{sente}}{M} \quad (1)$$

$$0 < D \leq 1$$

共起範囲が狭く、出現頻度も少ない

重要性が低い

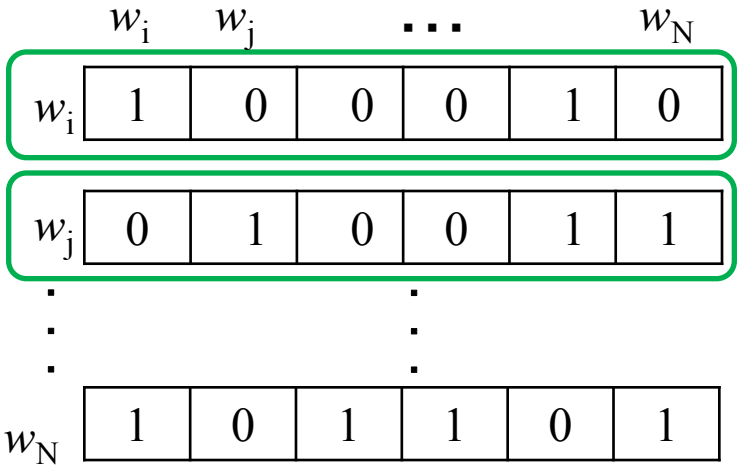
除外

3.単語クラスタリング手法

3.2単語クラスタリング対象単語の選出

1：単語w_iの共起出現確率ベクトルの相関を算出

$$C_{ij} = \frac{\sum_{a=1}^{N} P_{ia}^{sente} P_{ja}^{sente}}{\sqrt{\sum_{a=1}^{N}\left(P_{ia}^{sente} - \overline{P_i^{sente}}\right)^2}\sqrt{\sum_{a=1}^{N}\left(P_{ja}^{sente} - \overline{P_j^{sente}}\right)^2}} \qquad (2)$$

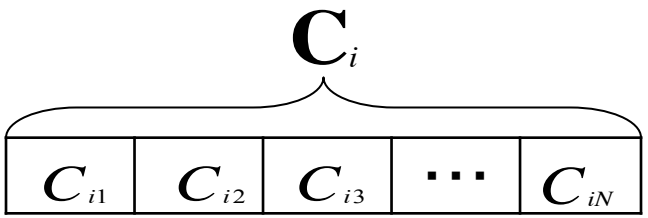


単語間の関係性

2：各単語の相関ベクトルの合計を算出

$$S_i = \sum_{j=1}^{N} C_{ij} \qquad (3)$$

汎用性



単語に意味はなく、共起している単語によって意味が変わる

除外

3.単語クラスタリング手法

SOMとは多次元の入力データを二次元に写像するクラスタリングアルゴリズム

1:参照ベクトルをランダムな値に初期化
2:入力ベクトルと最も近い参照ベクトル(勝者ユニットc)を見つけ出す

$$c = \arg\min_i \| X(t) - m_i(t) \| \qquad (4)$$

3:勝者ユニットと周りのユニットを次式に従って更新する

$$m_i(t+1) = m_i(t) + \alpha(t)[X_j(t) - m_i(t)] \qquad (5)$$

$$\alpha(t) = \alpha(0)(1 - t/T) \qquad (6)$$

$\alpha(t)$:学習係数　　T :学習回数

m :参照ベクトル

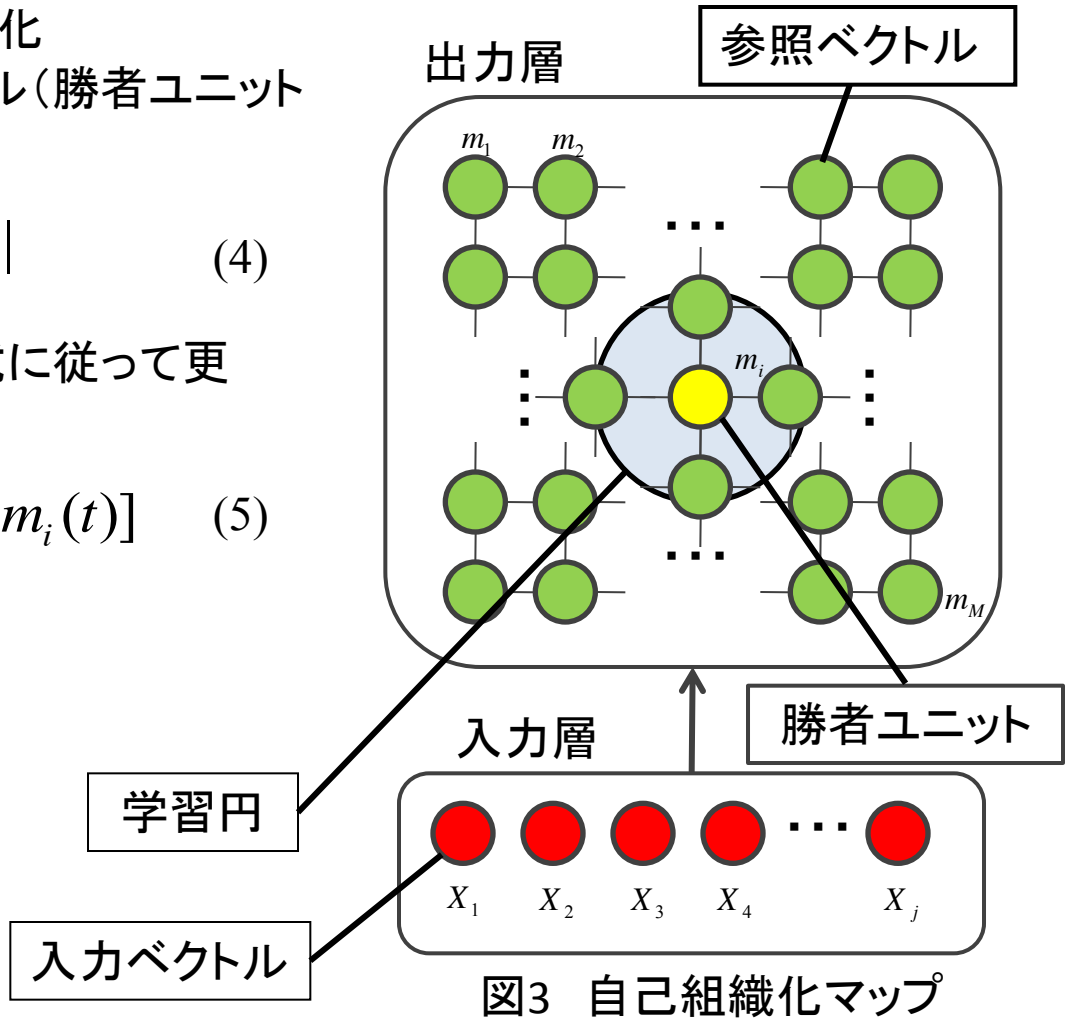


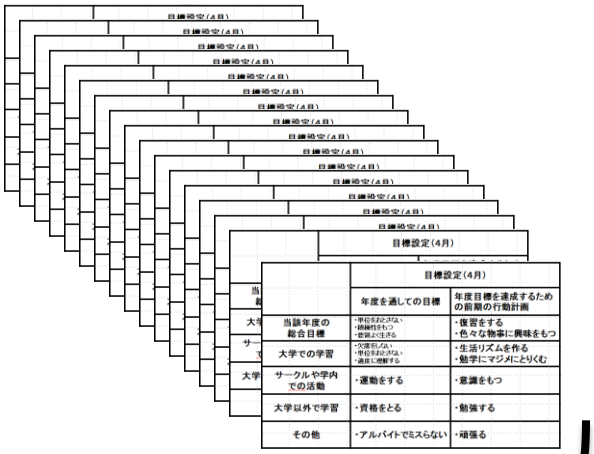
図3　自己組織化マップ

学習結果をWard法によりクラスタを作成する

4.評価実験

実験対象

文章数540の記述形式アンケート



アンケート回答

形態素解析(Chasen)

単語単位に分割

単語の集計

共起出現確率ベクトルの作成

全ての単語

全ての単語	1	0.5	…	0
	0.4	·		·
	·		·	·
	·			·
	·			·
	0.3	…	…	0

名詞

動詞	1	0.5	…	0
	0.4	·		·
	·		·	·
	·			·
	·			·
	0.3	…	…	0

全ての単語

名詞	1	0.5	…	0
	0.4	·		·
	·		·	·
	·			·
	·			·
	0.3	…	…	0

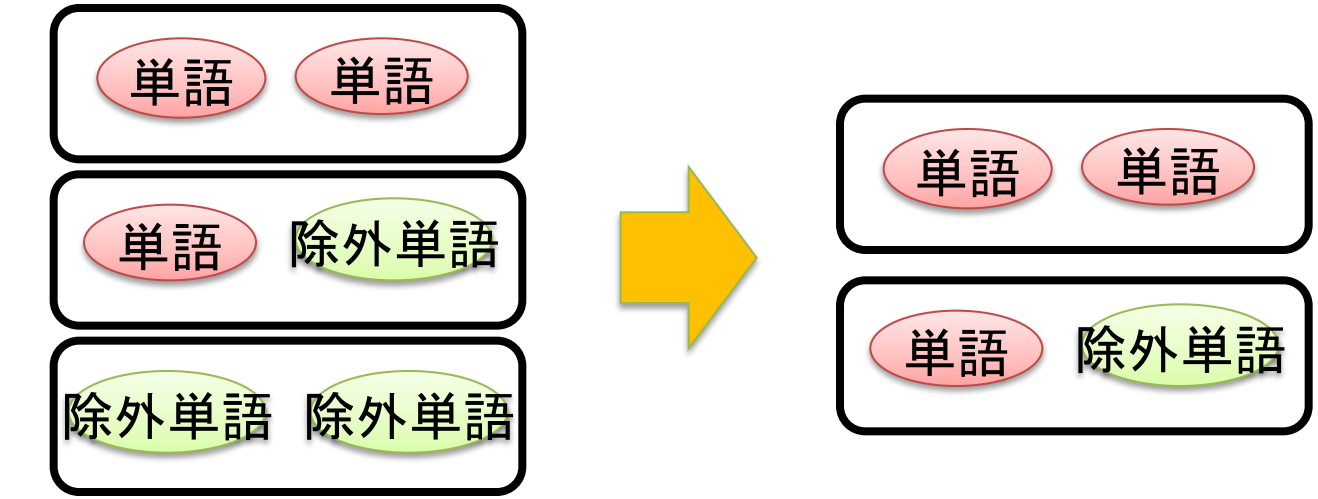
4.評価実験

共起出現確率ベクトル

球面SOMを用いたデータ可視化ツール「blossom」

クラスタ作成

相応しくない単語の除外



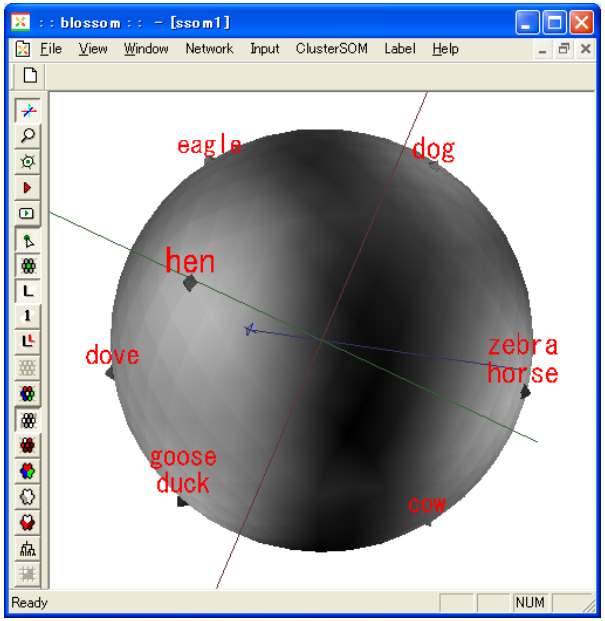


図4　blossom

5. 評価結果

SOMにより作成された**クラスタ数**

全ての単語のクラスタ数	440		227
動詞と名詞のクラスタ数	159	クラスタを除外	105
名詞と全ての単語のクラスタ数	255		124

良い例

- 「引っ掛かる」「躓く」
- 「降りる」「出る」

悪い例

- 「ボタン」「押す」
- 「転倒」「怪我」
- 「取り残す」「はじめる」

5. 評価結果

除外クラスタ

「アクセサリー」「鞄」

「オークラ」

「連動」「スライド」

アクセサリー付き鞄
ホテルオークラ
連動スライド

文を修飾語である単語や特殊な使われ方している単語

・似た意味をもつ単語クラスタが作成された。

・各クラスタ内の単語は同義とは判断できないものがあった。

・相応しくない単語クラスタが除外できた。

6.おわりに

まとめ

・文を分類する前段階の処理として単語クラスタリング手法を提案した
　共起出現確率ベクトルを作成し、単語クラスタを作成した
　文クラスタでは不要となる単語の除外をおこなった

・単語クラスタリング手法をアンケートに対して適用した
　同義の単語クラスタが作成された
　作成したクラスタから相応しくないクラスタを除外した

今後の課題

・単語クラスタリングの精度の向上
・単語クラスタの作成に共起出現確率以外の手法を提案
・クラスタリング対象単語の選出方法の検討
・プログラム処理時間の短縮