日本語意味マップの自己組織化

馬 青
独立行政法人 通信総合研究所

1 はじめに

インターネットの普及に伴い，大量の情報を手に入れることができるようになった半面，情報の洪水からどのようにして必要な情報を取り出すかが大きな問題となっている．情報を効率的に扱うために，これまで多数の情報検索システムが開発されてきている．しかし，これらのシステムに用いられる検索条件は基本的にキーワードという単語ベースのものであり，同じ意味であっても使う単語表現が違えば検索結果も違ってしまう．従って，よりよい検索結果が得られるようにするためには，検索を単語ベースの代わりに意味ベースで行う必要がある．そのために，何らかの方法で単語の意味表現を得る必要がある．

本章で紹介する日本語意味マップとは，日本語単語の意味空間へのマッピングを指す．ここで言う意味空間は 2 次元の可視的な表現である．この 2 次元表現においては，単語は，意味的類似性を距離として，意味的に類似性の高い単語どうしは近くに，意味的に類似性の低い単語どうしは遠くに配置される．このような意味マップを用いれば，単語の意味はそのマップ上の二次元座標で表現することができる．

本章は，このような日本語意味マップを，Kohonen が提案した自己組織化型神経回路網モデル（Self-Organizing Map，略して SOM[1]）を用いて，自動構築することについて述べる[1)]．以下，2 節では，SOM について簡単に紹介する．3 節では，

[1)] 本章で紹介した内容は神経回路網を用いた自然言語処理の一例である．神経回路網を用いた自然

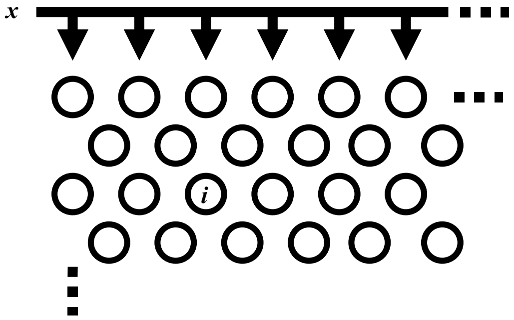

図 1: コホーネンの 2 次元 SOM

日本語意味マップを自己組織させるにはどのような学習データを使えばよいか，また，このようなデータはどのように SOM に適するようにコーディングすればよいかについて述べる．4 節では計算機実験の結果について述べる．この中で，まず，自己組織化によって得られた意味マップが妥当かどうかの検討を行う．そして，従来のクラスタリング手法との比較を行うことによって，SOM の分類能力を検証するとともに意味マップの特徴を明らかにする．更に，多くの変量が持っている情報を少数個の主成分あるいは因子に縮約するという意味で，SOM と同様に可視化能力を有す多変量解析手法との比較も行い，SOM を用いる有効性を示す．5 節ではまとめと今後の展開について述べる．

2　自己組織化神経回路網モデル SOM

自己組織化神経回路網モデル SOM（図 1）は高次元入力を持つ 2 次元配列のノード（ニューロン）で構成され，以下に述べる自己組織化によって高次元データを 2 次元空間にその特徴を反映するように非線形的に射影（マッピング）することができる．

入力 x が n 次元のベクトル，すなわち，$x = [\xi_1, \xi_2, \cdots, \xi_n]^T \in \Re^n$，ならば，個々

言語処理についてさらに知りたい読者は文献 [2, 3, 4] を参照されたい．

のノード i はそれぞれ同じ n 次元のベクトル $m_i = [\mu_{i1}, \mu_{i2}, \cdots, \mu_{in}]^T \in \Re^n$ を持つものとする．このベクトルは参照ベクトルと呼ぶ．但し，参照ベクトルの要素 μ_{ij} はノード i と入力要素 ξ_j の間の重みであり，自己組織化過程において少しずつ修正される．具体的には，入力ベクトル x が与えられたとき，まず，その入力をすべてのノードの参照ベクトルと比較し，ユークリッド距離の一番小さいノードを活性化する．その活性化されたノードを勝者ノードと呼ぶ．即ち，勝者ノード c は以下のように選ばれる．

$$c = argmin_i\{||x - m_i||\} \tag{1}$$

そして，自己組織化が大局的に行われるように，勝者ノードの近傍のノードも活性化させ，リラックス処理を行う．即ち，活性化されたすべてのノードに対し，それらの参照ベクトルを入力ベクトルに近付くように修正を行う．

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \tag{2}$$

ここで，t は学習回数で，$h_{ci}(t)$ は例えば以下のように定義された近傍関数である．

$$h_{ci}(t) = \alpha(t) \cdot exp(-\frac{||r_c - r_i||^2}{2\sigma^2(t)}) \tag{3}$$

但し，$r_c \in \Re^2$ と $r_i \in \Re^2$ はそれぞれ勝者ノード c と近傍ノード i の位置ベクトルである．従って，項 $||r_c - r_i||$ は近傍ノード i が勝者ノード c から離れて行くにつれ，h_{ci} が小さくなり $m_i(t)$ の修正量が小さくなることを意味する．また，$\alpha(t)$ は学習率で，$\sigma(t)$ は近傍の大きさ（半径）である．これらは時間と共に単調に減少していく関数であり，例えば以下のようなものである．

$$\alpha(t) = \alpha(0)\frac{T-t}{T}, \tag{4}$$

$$\sigma(t+1) = 1 + (\sigma(t) - 1)\frac{T-t}{T} \tag{5}$$

但し，$\alpha(0)$ は初期値で，T は総学習回数である．

自己組織化が終った後のマッピング処理は，単に入力 x に対する勝者ノードを選び出す[式 (1)] ことである．従って，上に述べた自己組織化（即ち，入力データに応じた参照ベクトルの修正）は入力にマッチするノードの整列過程と見なすことができる．即ち，マッチするノードを入力に近付ける過程である．但し，この処理は大局的に行われるため，結果的には，似た入力への勝者ノードが近いところに配置され，そうでないノードは離れたところに配置されるようになる．

一般的に，学習は「整列」と「微調整」の二つのフェーズに分けて行われる．つまり，上に述べた自己組織化アルゴリズムは実際，二回使われることになる．但し，「整列」フェーズでは，$\alpha(t)$ と $\sigma(t)$ の初期値を共に大きく取り，時間と共に (4)(5) に従って減少して行く．ノードの配置の基本形はこのフェーズで形成される．一方，残りのフェーズでは $\alpha(t)$ と $\sigma(t)$ は小さい値のまま長時間かけて初期フェーズで形成された基本形を微調整する．

3 意味マップの自己組織化

本章で紹介する意味マップの取り扱う対象は日本語名詞に限定する．本節では，日本語名詞の意味マップを自己組織させるために，どのような学習データを用いればよいか，また，このようなデータをどのようにコーディングすればよいかについて述べる．

3.1 学習データ

名詞の意味マップを自己組織化させるためには，名詞間の意味的な関係を反映できるような学習データが必要である．一般的に，名詞間の意味的な関係は名詞と共起する動詞や形容詞，さらには係り先の単語などを通じて結びつけることができる．ここでは，一番簡単に名詞間の関係を共起する連体修飾要素を通じて結びつけることとした．連体修飾要素とは，名詞句において，形容詞，形容動詞（＋な），名詞（＋の）といったように，後ろの主名詞を修飾するものを指す．実際，ある意味

では，意味的に近い名詞どうしは多くの共通する連体修飾要素を持ち，意味的に遠い名詞どうしはあまり共通する連体修飾要素を持たないと考えることができる．ここで「思い」,「気持」，そして「観点」という三つの名詞を例として考えよう．「思い」と「気持」は意味的に近い名詞どうしのため，共通する連体修飾要素「うれしい」や「あこがれの」を持ち得る．しかし一方,「思い」や「気持」と意味的に遠い関係にある名詞「観点」は，同様な連体修飾要素を持つことができない．従って，ここでは，個々の名詞をそれと共起する連体修飾要素のセットを用いて定義し，そのセットを学習データとする．

名詞句を構成する連体修飾要素と被修飾名詞の間には種々の関係が存在する．たとえば,「うれしい思い」の例では，連体修飾要素「うれしい」は名詞「思い」の具体的な内容を表す．一方,「太郎の思い」の例では,「太郎の」は「思い」の所有者を表す．連体修飾要素は名詞の属性として働く場合もある．例えば,「赤いリンゴ」の例では,「赤い」は「リンゴ」の属性である．これらの関係のうち，連体修飾要素が被修飾名詞の具体的な内容を表すものは特に自己組織化にとって有効であると考えられる．なぜなら，被修飾名詞の具体的な内容を表す連体修飾要素は被修飾名詞と共通する意味要素を持っているからである．すなわち，この種の名詞句の場合，被修飾名詞は，連体修飾要素の表す事柄を抽象化した意味を持つと考えることができる．例えば,「思い」は「うれしい」を抽象化した意味を持ち,「うれしい」は「思い」を具体化した表現で，互いに,「思い」という意味要素を共有している（詳細は文献[5]を参照）．

そのため，名詞の意味マップの構築に，まずこのタイプのデータを抽出して用いることにした．表1は新聞から人手で収集した，連体修飾要素が名詞の具体的な内容を表すようなデータを示す．この表の1列目は被修飾名詞であり，2列目は共起する連体修飾要素である．この表からも分かるように，連体修飾要素として，形容詞の他，形容動詞，そして「名詞＋の」も用いられている．但し，このタイプのデータは自動抽出が困難で言語学の専門家に頼らなければならないため，データの収集にコストがかかる．そのため，さらに，連体修飾要素と被修飾名詞の関係を考慮せ

表 1: 収集されたデータの例

名詞	連体修飾要素
思い	うれしい，あこがれの，気の毒な，いとしい，$\cdots$．
気持	うれしい，あこがれの， いい加減な，うらやましい，$\cdots$．
観点	アカデミックな，教育的な，歴史の，$\cdots$．
$\vdots$	$\vdots$

ずに形容詞と形容動詞を共起語とする名詞句を新聞から自動抽出して用いた名詞意味マップの構築も試みた．

3.2 データコーディング

前の節で述べたように，個々の名詞は共起する連体修飾要素のセットで定義する．例えば，表 1から名詞「思い」を

$$思い = \{ うれしい，あこがれの，気の毒な, \cdots \}$$

のように定義する．ここで，一般に ω 種類の名詞 $w_i (i = 1, \cdots, \omega)$ が存在し，それらの意味マップを構築すると仮定する．このような場合，名詞 w_i は以下のように連体修飾要素のセットで定義される．

$$w_i = \{a_1^{(i)}, a_2^{(i)}, \cdots, a_{\alpha_i}^{(i)}\} \tag{6}$$

但し，$a_j^{(i)}$ は w_i と共起する j 番目の連体修飾要素で，α_i は w_i と共起する連体修飾要素の数である．このように定義された名詞を SOM が取り扱えるようにするためにはまずそれらをコーディング（ベクトルで表現）しなければならない．コーディングの方法としては，まず英語の意味マップの構築に用いられてきた「ランダムコーディング法」[6] が考えられる．しかし，この方法は日本語マップの構築においてあ

まり有効でないことが先行実験によって判明したため，相関コーディング法を新たに提案した．以下，この二つのコーディング法について述べる．

3.2.1　ランダムコーディング法

ランダムコーディング法においては，名詞 w_i をコーディングするために，まずそれを定義する連体修飾要素 $a_j^{(i)}$ $(j = 1, \cdots, \alpha_i)$ [式 (6)] をコーディングする．すべての異なる連体修飾要素にできるだけ互いに独立するコードを付与するために，個々の連体修飾要素を単位ノルムの n 次元ランダムベクトル

$$V(a_j^{(i)}) = [v_{j1}^{(i)}, v_{j2}^{(i)}, \cdots, v_{jn}^{(i)}]^T \tag{7}$$

でコーディングする．このようなベクトルは，個々の要素 $v_{jk}^{(i)}$ をそれぞれゼロ平均を持つ独立なガウス分布から生成し，それを大きさ 1 のベクトルに正規化することによって得られる[2)]．このように得られた $V(a_j^{(i)})$ や $V(a_k^{(i)})$ は近似的に直交していると考えてよく，少なくとも統計的に独立している [6]．従って，名詞 w_i は，式 (6) に示された個々の連体修飾要素を表すベクトル [式 (7)] の和でコーディングすることができる．すなわち，

$$V(w_i) = \sum_{j=1}^{j=a_i} V(a_j^{(i)}) = [\sum_{j=1}^{j=a_i} v_{j1}^{(i)}, \sum_{j=1}^{j=a_i} v_{j2}^{(i)}, \cdots, \sum_{j=1}^{j=a_i} v_{jn}^{(i)}]^T \tag{8}$$

あるいは，名詞 w_i の定義を式 (6) の代わりに，その名詞自身も用いて

$$w_i = \{w_i, a_1^{(i)}, a_2^{(i)}, \cdots, a_{\alpha_i}^{(i)}\} \tag{9}$$

のように定義すれば，名詞 w_i は以下のようにコーディングすることもできる：

$$V(w_i) = [\varepsilon V'(w_i), \frac{\sum_{j=1}^{j=a_i} v_{j1}^{(i)}}{a_i}, \frac{\sum_{j=1}^{j=a_i} v_{j2}^{(i)}}{a_i}, \cdots, \frac{\sum_{j=1}^{j=a_i} v_{jn}^{(i)}}{a_i}]^T \tag{10}$$

但し，ε は例えば 0.2 のような小さな定数で，$V'(w_i)$ は式 (7) に示す $V(a_j^{(i)})$ と同様，ランダム的に生成されるベクトルである．

---
2) $v_{jk}^{(i)}$ は Press らが書いたプログラム [7] を用いて生成することができる．

表 2: 名詞の相関行列

	w_1	w_2	$\cdots$	w_ω
w_1	d_{11}	d_{12}	$\cdots$	$d_{1\omega}$
w_2	d_{21}	d_{22}	$\cdots$	$d_{2\omega}$
$\vdots$			$\ddots$	
w_ω	d_{w1}	d_{w2}	$\cdots$	$d_{w\omega}$

3.2.2 相関コーディング法

相関コーディング法においては，まず，それぞれ二つの名詞間の意味的相関（あるいは意味的距離）を反映するものを求める．ここでは，名詞 w_i と名詞 w_j 間の意味的距離 d_{ij} を (6) の定義式を用いて以下のような計算式によって求めることとした．

$$d_{ij} = \begin{cases} \frac{(\alpha_i - c_{ij}) + (\alpha_j - c_{ij})}{\alpha_i + \alpha_j - c_{ij}} & \text{もし } i \neq j \\ 0 & \text{その他} \end{cases} \tag{11}$$

但し，α_i と α_j はそれぞれ w_i と w_j が共起する連体修飾要素の数で，c_{ij} は w_i と w_j に共通する連体修飾要素の数である．従って，意味的関係 d_{ij} は名詞 w_i と w_j の間にどれだけ共通の連体修飾要素を持つかという意味での両名詞間の正規化された距離である．即ち，d_{ij} が小さければ，両名詞は意味的に近く，d_{ij} が大きければ，両名詞は意味的に遠い．

個々の d_{ij} は名詞 w_i と名詞 w_j の間の局所的な関係のみを反映しているが，このような個々の局所意味的距離から表 2に示すような行列を作成すれば，各行はそれぞれ見出しとなる名詞とそれを除いた $\omega - 1$ 個の名詞との局所意味的距離から構成されていることが分かる．即ち，各行はある程度大局的にその見出しとなる名詞と他の名詞との意味的な関係を反映していると考えられる．

従って，相関コーディング法では，名詞 w_i をこの行列を用いて以下のような多次元ベクトルにコーディングする．

$$V(w_i) = [d_{i1}, d_{i2}, \cdots, d_{i\omega}]^T \tag{12}$$

ここで，$V(w_i) \in \Re^{\omega}$ は SOM の入力である（即ち，$x = V(w_i)$，$n = \omega$）．即ち，SOM の役割は，このような多次元ベクトルを自己組織化によってそれらの間に存在する意味関係を顕在化させて 2 次元的に表現することである．

4　計算機実験

4.1　実験に用いたデータ

計算機実験では，3.1 節で述べた二種類の学習データを用いた．具体的にはまず，毎日新聞から人手で収集した，連体修飾要素が名詞の具体的な内容を表すような名詞句を用いた（今後，このデータをデータ 1 と言う）．名詞の種類 ω は 66，連体修飾要素ののべ総数 $\sum_{i=1}^{i=\omega} \alpha_i$ は 7,852，連体修飾要素の異なり総数は 4,998 であった．このうち，名詞「状態」の連体修飾要素の数（$\alpha_{\text{状態}} = 391$）がもっとも多く，名詞「中 2」[3]の連体修飾要素の数（$\alpha_{\text{中 2}} = 4$）がもっとも少なかった．また，名詞「思い」と「気持」がもっとも多く共通の連体修飾要素を持ち（$c_{\text{思い, 気持}} = 79$），多数の名詞が互いに共通する連体修飾要素を持たなかった．

次に，8 年分の毎日新聞から共起する形容詞と形容動詞の頻度順から自動的に集めた形容詞 / 形容動詞と名詞で構成されるような名詞句を用いた（今後，このデータをデータ 2 と言う）．名詞の種類 ω は 100，形容詞 / 形容動詞ののべ総数 $\sum_{i=1}^{i=\omega} \alpha_i$ は 33,870，形容詞 / 形容動詞の異なり総数は 4,023 であった．このうち，名詞「もの」と共起する形容詞 / 形容動詞の数（$\alpha_{\text{もの}} = 1,870$）がもっとも多く，名詞「政治家」と共起する形容詞 / 形容動詞の数（$\alpha_{\text{政治家}} = 192$）がもっとも少なかった．また，名詞「もの」と「こと」に共通して共起する形容詞 / 形容動詞の数がもっとも多く（$c_{\text{もの，こと}} = 764$），名詞「活動」と「気分」に共起する形容詞 / 形容動詞がもっとも少なかった（$c_{\text{活動，気分}} = 6$）．

[3]添字は同じ単語の（連体修飾要素の観点からの）異なる意味を区別するために用いている．例えば，「くらやみの 中」と「新しい 中（から一つを選ぶ）」の場合では，連体修飾語群の性格が異なる（前者が状態，後者が範囲を表す）ので，「中 1」（状態），「中 2」（範囲）などとして区別しておいた．

4.2 SOM による実験結果

実験に用いたパラメータは，データ１とデータ２のどちらの場合でも，SOM への入力の次元 ω が異なった以外，すべて同じように設定した．具体的に，SOM は 13×13 の 2 次元配列のノードで構成し，近傍の形状は六角形にした．整列フェーズにおいては，学習総回数 T を 10,000 に，学習率の初期値 $\alpha(0)$ を 0.1 に，そして，近傍の初期半径 $\sigma(0)$ を 13（即ち，SOM の直径に等しい）に設定した．微調整フェーズにおいては，学習総回数 T を 100,000 に，学習率の初期値 $\alpha(0)$ を 0.02 に，そして，近傍の初期半径 $\sigma(0)$ を 13 に設定した．初期参照ベクトル $m_i(0)$ は 0.0 から 1.0 までの乱数値からなる．SOM への入力の次元はランダムコーディングの場合では（式 (8) の n を）10 から 500 まで変化させて設定した．一方，相関コーディングの場合では，入力ベクトルの次元は取り扱う名詞の数に等しいので，データ１とデータ２の場合で（式 (12) の ω が）それぞれ 60 と 100 であった．

まず，ランダムコーディング法を用いた計算機実験を行った．その結果，入力ベクトルの次元を小さく取っても大きく取っても，また，名詞の定義に式 (6) を用いても式 (9) を用いても，十分納得の行く日本語意味マップを得ることができなかった．そのため，日本語の意味マップの自己組織化に相関コーディング法を用いることにした．以下，相関コーディング法を用いて得られた２種類の意味マップを示すとともに簡単な説明を行う．より詳細なことを知りたい読者は文献 [8] を参照されたい．

図 2(a) にデータ１を用いて得られた名詞の意味マップを示し，図 2(b) にこれを人手で意味的に七つの領域（グループ）に分けた図を示す．例えば，左上から左下までの領域にマッピングされた名詞（思い，気，気持，気分，くらい，$\cdots$，どん底，色 2）は基本的に <感情・感覚> を表すものと考えてよい．このグループはより詳細には直接感情を表すもの（思い，気，気持，気分，$\cdots$，ため息，色 2）と感情の程度を表すもの（くらい，ほど，どん底）に分けられる．この領域には，左端中央に名詞「空気」が出現しているが，これは「気まずい<u>空気</u>」などのように「雰囲気」を表す名詞である．連体修飾要素からみると，「空気」は <感情・感覚> の名詞と共通

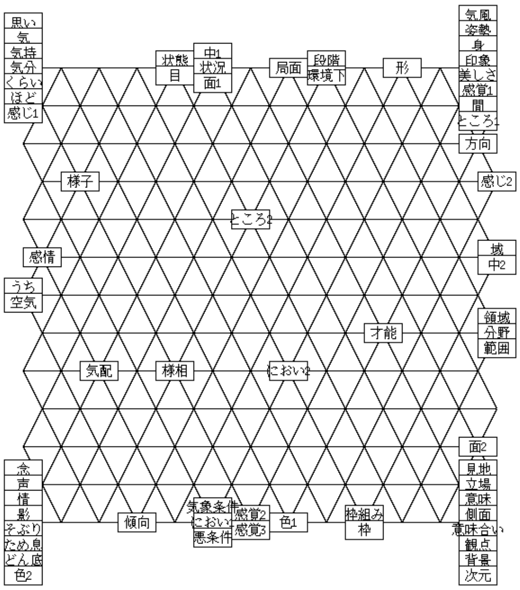

図 2(a) データ 1 を用いた意味マップ

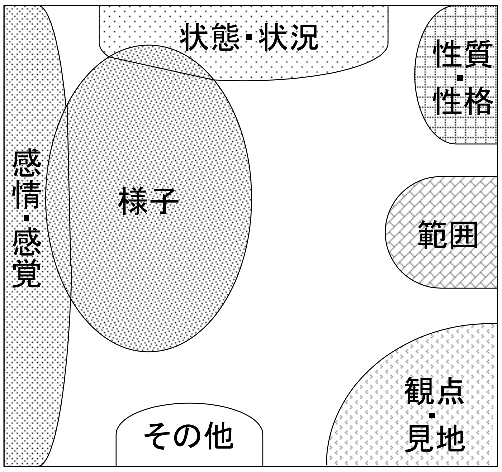

図 2(b) 意味マップに対応する意味グループ

の連体修飾要素を多く持つが，それだけではなく＜様子＞の領域の名詞（様子，空気，気配，様相）とも共通する連体修飾要素が多い（例えば「不穏な 空気」）．従って，図 2(b) の意味グループの図では＜様子＞の領域の一部を＜感情・感覚＞の領域に重ねて図示している．また，同様の例として，名詞「様子」は＜様子＞の領域だけでなく，＜感情・感覚＞の領域（例えば「感激の 様子」）や，＜状態・状況＞の領域（例えば「悪戦苦闘の 様子」）とも連体修飾要素を共有している．実際，図 2(a) をみると，名詞「様子」の位置は，＜感情・感覚＞や＜状態・状況＞の領域に寄っていることが分かる．従って，＜様子＞の領域を＜感情・感覚＞の領域とだけでなく＜状態・状況＞の領域とも一部分重ねて図示した．

右下の領域にマッピングされた名詞（{ 面 2，見地，$\cdots$，次元 } と { 枠組み，枠 }）は観点や見地などを表すものである．このグループにおいては，前者のセットと後者のセットの間の意味上の微妙な違いは異なる領域にマッピングされることによって表現されている．もちろん，分け方によってはこの二つの領域にマッピング

される名詞を二つのグループに分けることも可能である．しかし，この二つの領域が近接しているため，これらの中の名詞が意味的に近いという結果には変わりがない．似たようなことは<観点・見地>グループと<範囲>グループの間にも言える．この二つのグループの名詞は意味的に相似しているため，それぞれ隣同士の領域に位置するグループにマッピングされている．実際，<範囲>グループの名詞「分野」や「領域」は<観点・見地>グループに分類されてもおかしくない．今回は，これらを二つのグループに分けたが，いずれにせよこの二つのグループの中の名詞が意味的に近いという結果に変わりがない．このように，意味マップは意味の相似性を連続的に表現しているため，意味マップを用いれば，どの名詞がどのグループに分類できるかを知ることだけでなく，名詞同士が互いにどれぐらい意味的に近いかを知ることもできる．そして，名詞間の意味的相似を見る際にはそれらの名詞が同じグループに属するかどうかを考える必要はなく，また，その名詞がどのグループに分類すべきかが決まらない場合でも意味的な近さの比較が可能である．

以上に述べたように，自己組織化されたマップは全般的に我々の直観に合っていると言ってよい．うまく行かなかった例もあるが，本実験に限って言えばそれは基本的に訓練データのスパースネス問題に起因するものであった．その例として<その他>にマッピングされた名詞（傾向，気象条件，$\cdots$，色 1）がある．訓練データを調べた結果，これらの名詞は他の名詞との間に共通する連体修飾要素が極端に少ないことが分かった．例えば，名詞「気象条件」は他の名詞との間に最も多い場合でも三つの共通の連体修飾要素しか持っていない．

次に，データ 2 を用いた実験結果について述べる．図 3(a) にデータ 2 を用いて得られた名詞の意味マップを示し，図 3(b) にこれを人手で意味的に分類した図を示す．この意味マップを大まかに見て，まず，感性的活動<感情>を表す名詞（表情，態度，気持，感じなど）と理性的活動<思考活動>を表す名詞（意見，判断，考え，批判など）がそれぞれ左下と右上の対角領域に，そして，社会性<社会的トピック>を表す名詞（社会，環境，情報，政治，数字など）と文化性<文化活動>を表す名詞（作品，音楽，舞台，プレーなど）がそれぞれ上中央と下中央の領域にマッピング

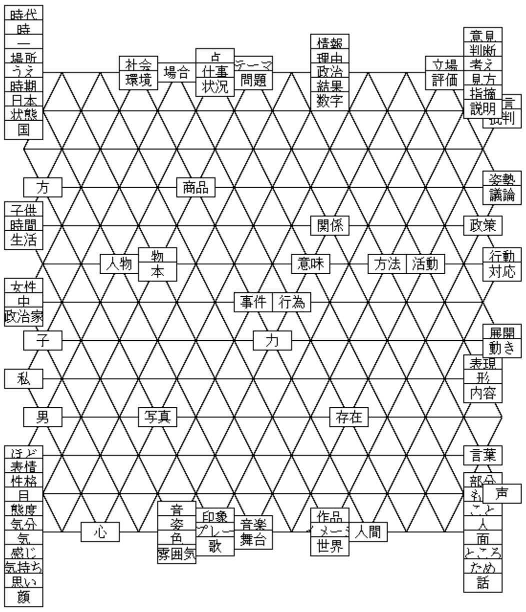

図 3(a) データ 2 を用いた意味マップ

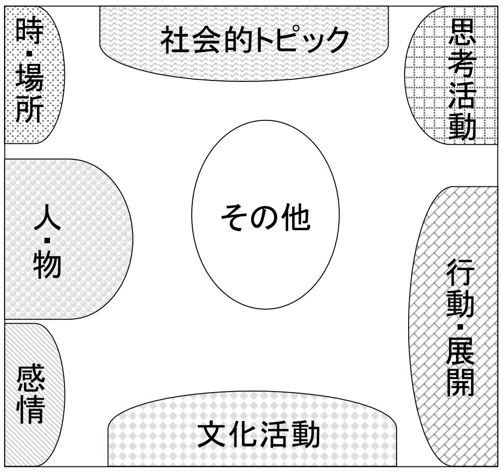

図 3(b) 意味マップに対応する意味グループ

されていることが分かる．そして，右の大半の領域を占める < 行動・展開 > グループは < 文化活動 > のグループとともに，感性と理性の表現領域の中間に配置されている．この配置は，その結果の必然性を別にして，文化活動を含む人間の種々の行動は理性と感性の総合作用によって発生するという人間の外的行為と内的精神活動との関係を反映している，というように解釈を与えることが可能であろう．そして，外的行為の中も，文化活動のほうがより感性に頼っているという意味で，< 文化活動 > が < 行動・展開 > よりも < 感情 > の近くに配置されていることも解釈可能となろう．残りの領域は，他の領域との相対的な関連性が薄く，それぞれ独立していると考える．具体的に，左上の領域にマッピングされた名詞（時代，時，$\cdots$, 国）は基本的に < 時・場所 > を表すものとしてとらえることができ，左中央の領域にマッピングされた名詞（方，子供，$\cdots$，商品，人物，物，本，女性，$\cdots$，男）は基本的に < 人・物 > を表すものと考えることができる．そして中央にマッピングされた名詞（関係，意味，$\cdots$，力）はどのグループにも分類されていないものとしてとらえ

ることができる．データ 2 のように，ただ単に共起語の頻度順で選んだ，我々人間にとってもその意味付けや分類が非常に困難と思われる名詞群も，SOM を用いることによってある程度我々人間の直観に合う意味マップを得ることができる．

4.3　他手法との比較

4.3.1　階層型クラスタリング手法

SOM が意味マップの構築に有効であることを確かめるためには，その分類能力を評価する必要がある．そのため，同じデータ 1 とデータ 2 を用いて階層型クラスタリング手法[4)]との比較実験を行った．クラスタリングにかけられたデータは SOM への入力ベクトルと同様，3.2.2 節に述べた相関コーディング法を用いて構成した．表 3と表 4はそれぞれデータ 1 とデータ 2 に対する分類結果（1 列目と 2 列目）を示す．但し，得られたクラスタリング結果は階層的になっているため，ここに示した分類結果は，SOM で得た意味マップをグループ分けする場合と同様，階層的にクラスタリングされた結果をできるだけ最適であるように人手でグルーピングすることによって得た．

まず，図 2 の意味マップと表 3のクラスタリング結果を比較してみよう．基本的にこの両者はよく対応が取れていると言える．分かりやすくするために，表の右側には，それぞれのクラスに対応すると思われる意味マップ上のグループ名を付け加えた．但し，両者間に細かいところに色々な違いもある．ここではまず，両者ともうまく分類ができていない例をみてみよう．例えば，図 2(a) の意味マップにおいては，学習データの不足のため我々人間も分類できない名詞「色 1」と「におい 2」はそれぞれ <その他> と他の名詞群から離れて孤立に，そして，「忙しい 間」や「留守の間」など状況の意味を持つ名詞「間」は隣接の <性質・性格> のグループに誤って配置されている．一方，クラスタリングの結果では，その三名詞はクラス 5 と 6（<

[4)] パッケージソフト SPSS for Windows 6.0J の階層型クラスタリングプログラムを用いた．但し，クラスターの結合方法としてはグループ間の平均連結法を，ベクトルやクラスタ間の距離計算には平方ユークリッド距離を採用した．

表 3: クラスタリング手法によるデータ 1 の分類結果

クラス	名詞	意味マップとの対応
1	ほど くらい 気 感じ 1 気分 思い 気持 感情 様子	＜感情・感覚＞
2	ところ 1 面 1 空気 姿勢 印象 感覚 1 気風	＜性質・性格＞
3	ところ 2 目 身 様相 気配 状態 状況 局面 段階 環境下 形 うち 中 1	＜状況・状態＞＜様子＞
4	そぶり ため息 どん底 影 念 情 色 2 声	＜感情・感覚＞
5	枠 枠組み におい 2 色 1 観点 見地 面 2 背景 側面 意味 立場 意味合い次元 領域 才能 分野 方向	＜範囲＞＜観点・見地＞
6	域 範囲 間	＜範囲＞
7	感じ 2 傾向 悪条件 気象条件 におい 1 中 2 感覚 3 感覚 2 美しさ	＜その他＞

範囲＞と＜観点・見地＞）に誤って分類されている．次に，意味マップの結果が正しく，クラスタリングの結果が間違っている例としては次のようなものがある．「清楚な 美しさ」など＜性質・性格＞に属する名詞「美しさ」は意味マップでは正しく配置されているが，クラスタリングにおいては＜その他＞に相当するクラス 7 に分類されている．同様に「一文なしの 身」など＜性質・性格＞に属する名詞「身」も意味マップでは正しく配置されているが，クラスタリングにおいては＜状態・状況＞と＜様子＞を一まとまりにしたものに相当するクラス 3 に誤って分類されている．＜感情・感覚＞と＜様子＞に属する名詞「空気」は意味マップでは正しく配置されているが，クラスタリングにおいては＜性質・性格＞に相当するクラス 2 に誤って分類されている．そして，「好人物の 感じ」のように＜性質・性格＞に属す場合もある「感じ 2 」は意味マップではこの領域に配置されているが，クラスタリングにおいては＜その他＞に相当するクラス 7 に分類されている．さらに，意味マップでは

表 4: クラスタリング手法によるデータ 2 の分類結果

クラス	名詞	意味マップとの対応
1	もの こと 人 ところ 言葉 表現 内容 形 話 声 面	＜行動・展開＞
2	物 本 商品 人物 政治家 子供 子方 女性 人間	＜人・物＞
3	日本 国 時 時期 時代 場所 一 状況 ため うえ 部分 点 生活 時間 中	＜時・場所＞
4	結果 数字 指摘 説明 理由 場合 発言 批判 意見 考え 判断 立場 見方 評価	＜思考活動＞
5	政策 活動 議論 行動 方法 問題 テーマ 仕事 情報 政治 社会 関係 環境	＜社会的トピック＞
6	作品 舞台 音楽 歌 音 写真 色 感じ 印象 イメージ 雰囲気 姿 世界 存在	＜文化活動＞
7	ほど 気 気持 思い 心 顔 表情 目 気分	＜感情＞
その他[1]	動き 展開 プレー 力 意味 行為 事件 対応 姿勢 態度 男 性格 私	＜その他＞

意味の微妙な違いを表現できているのに対し，クラスタリングではうまく反映できていない例として次のようなものがある．クラスタリングの結果においては，クラス 3 は状況と様子の名詞を一まとまりにし，クラス 5 も＜観点・見地＞だけでなく「領域」や「分野」といった＜範囲＞の名詞も含んでいる．それに対し，意味マップにおいてはそれらのグループ間に重なりはあるものの，離れた場所に配置することによりそれらの違いもはっきり表現している．一方，意味マップの結果が間違っていてクラスタリングの結果が正しい例は以下の一個しか見つけることができなかった．すなわち，「意地っぱりの 面」のように＜性質・性格＞に属する名詞「面 1 」は意味マップでは隣接の＜状態・状況＞に誤って配置されてしまっているが，クラスタリングにおいては正しく＜性質・性格＞に相当するクラス 2 に分類されている．

次に，図 3 の意味マップと表 4のクラスタリング結果の比較について簡単に述べる．この二つの結果の対応付けは表 3と図 2 の対応付けに比べてやや困難であるが，大雑把には表の右側に示しているような対応付けも可能であろう．そして，この両者を細かく比べていくと，ほぼ同程度によい分類結果が得られていることが分かる．その比較の詳細を知りたい読者は文献 [8] を参照されたい．

このように，SOM とクラスタリング手法の両者の比較について，厳密な評価を下すことはできないが，SOM を用いた手法は意味的分類を可視的連続的に表現できるという利点を持っているだけでなく，分類能力においても従来のクラスタリング手法よりやや優れていることが言えよう．

4.3.2 多変量解析手法

可視的な 2 次元意味マップの構築には SOM による自己組織化以外の方法も考えられる．その有力候補として主成分分析や因子分析といった多変量解析手法を挙げることができる．しかし，以下に述べるように，本タスクにはこのような多変量解析手法がうまく適用できない．

主成分分析も因子分析も，多くの変量が持っている情報を少数個の主成分あるいは因子に縮約するための解析手法である．両者の基本的な考え方には相違があるが，少数個の変量で多変量データを説明するという意味で共通性があり，同じ課題に利用されることが多い．そして，もとの変量の持っている情報の大部分を説明できるようにするために主成分あるいは因子を幾つまで必要とするかは，それらの寄与率（即ち，それぞれの分散が全体の中で占める割合）の累積値によって決まる．意味マップの構築に用いた 2 種類のデータに対し主成分分析[5]を行った結果，第 1 と第 2 主成分が合わせた累積寄与率はデータ 1 とデータ 2 に対してそれぞれ 15%と 20%にすぎないことが分かった．一般的には主成分の累積寄与率が 70 ないし 80%以上大き

[1]本手法は階層型クラスタリング手法であるため，下位の層で分類されていないものは「その他」にした．

[5]パッケージソフト SPSS for windows の因子分析プログラムを用いた．

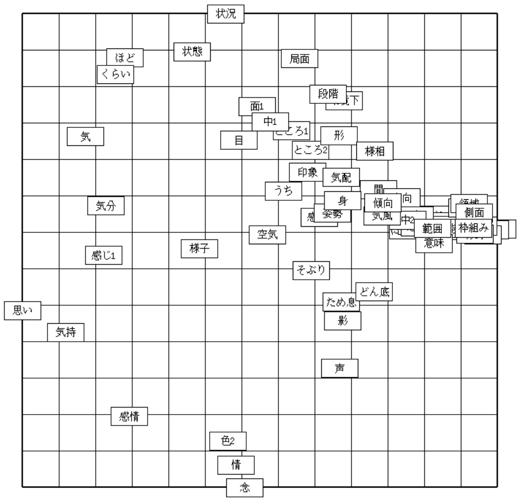

図 4 主成分分析結果のプロット図

くならないと，多変量データを正しく縮約できないとされている [9]．したがって，明らかにデータ 1 とデータ 2 を用いた意味マップの構築においては，多変量解析を用いることは困難である[6)]．実際，データ 1 を第 1 主成分と第 2 主成分を軸とする 2 次元平面にプロットする計算機実験も行ってみた．その結果，図 4 に示しているように，意味マップに相当する < 感情・感覚 > グループ（全エリアのほぼ半分を占める左下の領域）と < 状態・状況 > グループがある程度形成されたとしても，それ以外の約三分の二のデータが真中の小さなエリアに集中してうまく分類できなかった．

多変量解析がこのタスクにうまく適用できない主な理由は今回扱ったデータが強い非線形性（つまり線形分離不可性）を持っているためであると推測される．つまり，既存の主成分分析や因子分析手法は，非線形処理を行う SOM と違って，その主成分や因子はもとの変量の線形結合だけで求められているため，線形処理しか行っておらず，非線形性の強いデータ処理には対処できないと思われる．明らかに，このような結果は可視的な 2 次元意味マップの構築に SOM を用いる必要性を補強

6) 因子分析において座標軸への種々の回転操作を行っても累積寄与率をあげることができなかった．

していると言えよう．

5 おわりに

本稿は，自己組織化型神経回路網モデル SOM を用いた日本語意味マップの自動構築について述べた．自己組織化に用いる学習データは新聞から収集したもので，意味マップの自己組織化に有効と思われる，連体修飾要素が名詞の具体的な内容を表すような名詞句と，大規模意味マップの構築にはデータ収集の自動化が不可欠という観点から，新聞から名詞及びそれと共起する形容詞と形容動詞を共起頻度の高いものから自動的に取り出して構成される名詞句という二種類のものを用いた．データのコーディングには，これまでよく用いられて来たランダムコーディング法の代わりに相関コーディング法を新たに導入した．計算機実験で得られた意味マップは実際に用いた学習データを用いて検討し，意味マップ上の名詞は全般的に学習データが示唆する意味で配置されていることを示した．そして，本手法の有効性及び必要性を階層型クラスタリング手法と多変量との比較を通じて明らかにした．

本研究の今後の展開としては，まず，提案手法の改善及び拡張を行う．これは SOM のアーキテクチャ，コーディング，及び学習データの構成という三つの側面から考えられる．アーキテクチャにおいては，これまでサイズが固定の SOM を用いていたため，問題に応じた最適なサイズの設定が困難であるという問題がある．そのため，問題駆動のサイズ可変型 SOM を導入する必要がある．コーディングにおいては，二つの側面からその改善を考える必要がある．本稿で述べた相関コーディング法では，形成される入力ベクトルの次元が対象名詞の数に等しいため，対象名詞の数が増えればベクトルの次元も増えるという欠点がある．また，今のコーディング法には，修飾語の共起頻度を組み込むことができない．しかし，一般的に，ある名詞に高い頻度で共起し，他の名詞にあまり共起しないような修飾語はその名詞にとって重要であると考えられる．従って，コーディングの一改善法として，情報検索において重要なキーワードを決める TFIDF 法の適用が考えれる．実際，TFIDF

の適用はすでに中国語の意味マップの自己組織化に行われ，その効果が確認されている（詳細は文献 [10] を参照）．また，学習データの構成においては，これまで，名詞をそれと共起する連体修飾要素のみで定義していた．より一般的な意味マップの構築を目指すためには，動詞などを含むより広範囲な情報（例えば名詞の係り先のすべての単語）の利用も考慮する必要がある．以上のように提案手法の改善や拡張を行った後，より大規模で，実用指向の日本語意味マップの構築を行っていく予定である．そして，このように構築された意味マップは高性能情報検索システムの開発に役立つであろう．

参考文献

[1] T. Kohonen: Self-organizing maps, Springer, 2nd Edition, 1997.

[2] R. G. Reilly and N. E. Sharkey (Eds): Connectionist Approaches to Natural Language Processing, Lawrence Erlbaum Associates, Publishers, 1992.

[3] N. Sharkey (Ed): Connectionist Natural Language Processing, Kluwer Academic Publishers, 1992.

[4] R. Miikkulainen: Subsymbolic Natural Language Processing, The MIT Press, 1993.

[5] K. Kanzaki, Q. Ma, and H. Isahara: Similarities and differences among semantic behaviors of Japanese adnominal constituents, *ANLP/NAACL'2000 Workshop: Syntactic and Semantic Compexity in Natural Language Processing Systems*, Seattle, Washington, pp. 59-68 (2000).

[6] H. Ritter and T. Kohonen: Self-organizing semantic maps, *Biological Cybernetics*, 61, pp. 241-254 (1989).

[7] Press, W. H., Teukolsky, S. A., Vetterling W. T., and Flannery, B. P.: Numerical recipes in C, Cambridge University Press, 2nd Edition, 1994.

[8] 馬青，神崎 享子，村田 真樹，内元 清貴，井佐原 均：日本語名詞の意味マップの自己組織化，情報処理学会論文誌，Vol. 42, No. 10, pp. 2379-2391, 2001.

[9] 田中 豊，脇本 和昌：多変量統計解析法，現代数学社 (1983)．

[10] Q. Ma, M. Zhang, and M. Zhou: Self-Organization of Chinese Semantic Maps Using TFIDF Term Weighting, NLPNN2001, pp. 85- 92, 2001.