クラスタリング手法を利用したコーパスからの容器状物体の形状獲得

黒澤　義明　　竹澤　寿幸
広島市立大学大学院　情報科学研究科
{kurosawa, takezawa}@ls.info.hiroshima-cu.ac.jp

1.　はじめに

コーパス資源の蓄積による自然言語処理研究の対象拡大により，文の解析（形態素解析，構文解析）だけではなく，比喩等の言語現象を扱う研究も増えている．しかし，まだ十分とは言えない．

例えば，以下に挙げた(1)の表現の解釈としては，通常文字通りの解釈は成立しない．鍋は食用可能な物体ではないからである．したがって，適切な解釈のためには『容器－中身』という関係性が必要となり，その結果，“鍋（の中身）を食べる”という解釈が可能となる．このような表現を換喩と言う．

(1) 鍋を食べる

そこで，いわゆる容器を『容器－中身』の関係性を持つ物体と考えれば，上記の解釈は可能と考えるかもしれない．しかし，実際には不十分である．山梨(1988)も指摘しているように，以下の (2)の適切解釈のためには，知識～容器ではない物体も『容器－中身』の関係性を持つという知識～が必要だからである．

(2) 押し入れをかき回す

すなわち，文字通りの容器でなくとも，何かを入れるという機能を持つ物体は『容器－中身』の関係性を満たすと言える．ここに広い意味での「容器状物体」という新しい分類が必要となる．では，こうした容器状物体の定義を如何にコンピュータに与えればよいか？

黒澤ら(2008)では，『容器－中身』の関係性記述の試みとして，“A の奥”，“A の底”等，物体 A を表す名詞とともに共起する表現（彼らは「見立て詞」と呼んでいる）に着目した．例えば，鍋には深さはあるが奥行きはそれほどない．このため“鍋の底”とは言えても“鍋の奥”という表現はしづらい．また，瓶は細長いため“瓶の先”と言える．しかし，鍋の場合には“鍋の先”という表現は難しい．つまり，物体 A と共起する表現には，人間の言語直感が含まれていると考え，コーパスから物体の形状獲得を試みたわけである．“鍋の先”と言えるかについては，コーパスからの頻度抽出により可能である．しかし，「言えない」というためには長さ情報等の新たな基準が必要である．つまり，このような形状獲得により，一種の言語直感のコンピュータへの構築を目指す．

本稿ではその考え方を再検討し，物体の形状分類に際し，Hoffman(1999)による pLSA(probabilistic Latent Semantic Analysis)を用いた次元の縮約・整理を行った上で，Kohonen(2001)の自己組織化（Self-Organizing Map : SOM）を行う．新たな手続きの追加により，物体の形状を有効に反映したマップが作成され，換喩の解釈に必要となる関係性が得られる筈である．

2.　自己組織化マップ SOM による容器状物体の形状分類

本研究では，Kohonen(2001)による自己組織化マップ(Self-Organizing Map, SOM)を使用する．SOM は，多次元ベクトルにより表されたデータを，その特徴を残し，他のデータとの相互関係を保ったまま，2 次元マップに写像することが出来る．すなわち，多次元のデータの関係を 2 次元平面上の距離として表し，視覚的に理解し易いと言う特徴を持つ．

2.1.　自己組織化マップのアルゴリズム

SOM は二層からなる神経回路網モデルである．教師なし学習～入力層への入力により，競合層の特定の領域が反応するような～を行う．

入力層への入力ベクトル $\boldsymbol{x}$ は n 次元のベクトルであり，$\boldsymbol{x}=\{x_1, x_2, \cdots, x_n\}$ と表現する．また，競合層にはノードと呼ばれるユニットがあり，全ノードから，入力層との間に参照ベクトル $\boldsymbol{m}$ と呼ばれるリンクが行われる（図 1）．

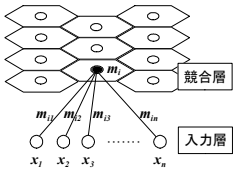

図 1　SOM の基本概念

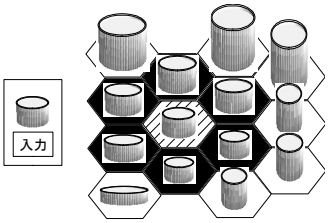

図 2　勝者ノード，近傍

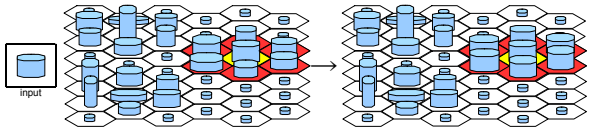

図 3　参照ベクトルの更新

ここで，次式を満たす勝者ノード c の発見を試みる．次式は入力ベクトルに最も類似した参照ベクトルを持つノードを見つける操作と考えられる．

$$\forall i,\quad \|x-m_c\| \le \|x-m_i\|$$

上記の勝者ノードの発見に続いて，近傍内の複数の参照ベクトルを入力ベクトルに近づける操作を行う．つまり，時間が経つにつれ，近隣のノードの類似性が増し，隣接ノード間のベクトル距離が近づくこととなる（図 3 の右中央部の変化）．以下に，時間軸 t を用いた式を示す．

$\forall i \in N_c(t)$ を満たすとき，

$$m_i(t+1) = m_i(t) + h_{ci}(t)\left(x(t) - m_i(t)\right)$$

それ以外のとき，　$m_i(t+1) = m_i(t)$

以上の勝者ノード発見，近傍更新を繰り返すことにより，学習を行うこととなる．なお，$h_{ci}(t)$ は作用範囲を示す近傍と呼ばれる領域であり，本研究ではガウス関数を用い，時間とともに縮小させた．

本節に紹介した SOM は自然言語処理に適用され，オノマトペ等，様々な対象の分類に有効性が確認されている(cf. Kurosawa et al. 2010; 神崎ら 2007; 金 2003; 馬ら 2001)．

3.　実験と考察

3.1.　言語データ

本研究で用いる言語データは，基本的に黒澤ら(2008)で収集されたデータと同一である．その収集・加工手続きを示す．

① 容器状名詞の収集

野口(2005)の現代仮名遣い作品から，10 回以上登場する「A の中」という名詞句 265 個を収集した．この手続きにより，典型的な容器だけでなく，比喩的に解釈可能な名詞 A が収集される（ex. 頭の中）．なお，この手続きだけでは筒状の物体が少なくなるため，単語を追加し，283 語とした[1]．

② 見立て詞毎出現率算出

池原ら(1997)の「2610 場」の下位分類から，8 語（端，角，口，奥，先，席，底，隅）を選び，容器状名詞との共起頻度を計算し，出現率にデータ変換を行った．

$$m_i\text{の出現率} = m_i \Big/ \sum_{i=1}^{n} m_i$$

③ pLSA による次元整理

Hoffman(1999)による pLSA(probabilistic Latent Semantic Analysis)を用い，次元の整理・縮約を行う．②に挙げた 8 個の間に相関があるかもしれないため，もしあるなら，容器状名詞との共起を確率的に整理・縮約することにより，よりよい学習結果が得られるはずである．この追加手続きは，黒澤ら(2008)の論文とは異なる[2]．

工藤による pLSA の実装を用いて，上記手続きを実行した．この実装は温度パラメータ β の変化により，確率値のスムージングが可能である．そこで，β=1.0（厳密な EM の実行）と，β=0.8 のときの比較を行い，容器状名詞表現の有効性について考察を加える．

3.2.　実験手続き

2 章で説明した手続きにより，som_pak を使用した分類を行った．なお，2 段階の学習を行った．予備実験により決定されたパラメータを以下に示す．

マップサイズ：64 ノード×48 ノード

1st: 学習回数 100,000，初期学習率係数 0.05

2nd: 学習回数 1,000,000，初期学習率係数 0.01

1 黒澤ら(2008)は 283 語からさらに選別し，209 語を実験に用いた．しかし，本研究ではこの選別を行わず，283 語を使用した．

2 今回は 8 次元表記のまま縮約を行わず，結果の検討を行う．

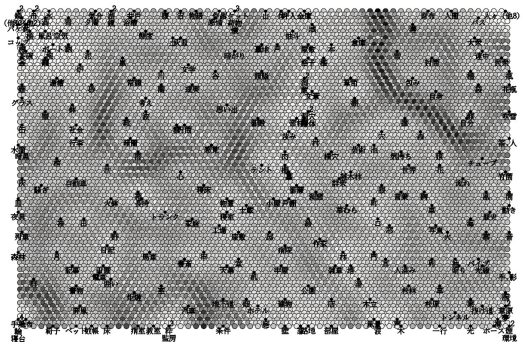

図 4　SOM による分類結果（β=1.0）

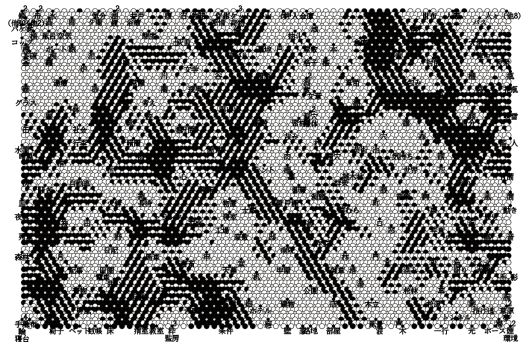

図 5　結果（図 4）のノード間距離を強調

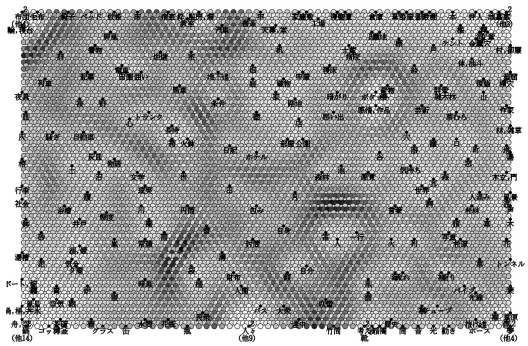

図 6　SOM による分類結果（β=0.8）

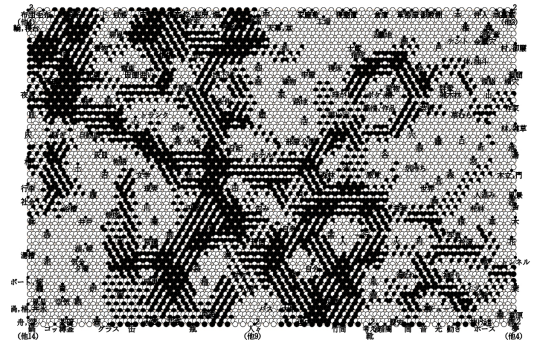

図 7　結果（図 6）のノード間距離を強調

3.3.　実験結果

pLSA の温度パラメータが β=1.0 のときの実験結果を図 4 に示す．隣接ノード間距離の最大値と最小値を元に，距離が 0-1 になるよう変換し，明度で表現した．図中，より黒く見えるところは隣接ノード間の距離が遠いことを，より白く見えるところは距離が近いことを示す．したがって，黒い線分に囲まれているように見える領域では，その領域の内側と外側の特徴が異なっていることになる．例えば図 4 では，右上に大きな領域があり，クラスタが生じていることがわかる．

ただし，領域がはっきりしないところもある．そこで，この領域を明確にするため，図を強調する（図 5）．0.25 以上の距離を持つノードを黒，それ以外を白で着色した図である．

同様に，図 6，図 7 は温度パラメータ β=0.8 のときの結果とその強調図である．β=1.0 のときの方が黒いところが多く，細かく分類されているように見える．この点については後述する．

4.　考察

4.1.　一般的な傾向について

黒澤ら(2008)での結論同様，大まかに形状分類がなされていると言える．例えば，図 5 の左上を拡大する（図 8）．「鍋」，「胃」，「湖」等，基本的に水に関連した容器状の名詞が多く含まれる．また，左下にかけて，「水筒」など，高さを持った物体が配置されているように見える．同じ配置は図 7 の左下にも見られる．同様に，「布団」等の平たい（高さが低い）

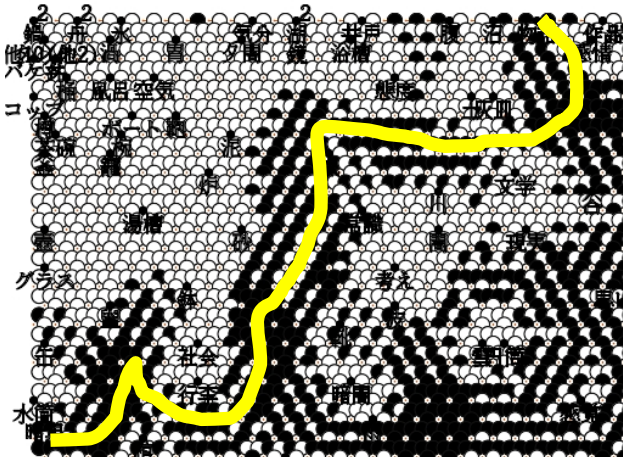

図 8　図 5 の左上拡大図

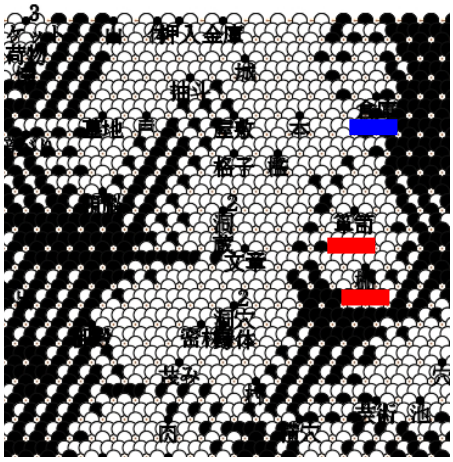

図 9　図 5 の中央上部拡大図（β=1.0）

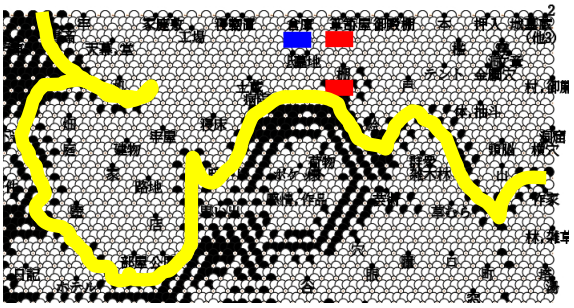

図 10　図 7 の中央上部拡大図（β=0.8）

グループ（図 5 左下，図 7 左上）等も確認される．

4.2.　黒澤ら(2008)と比べ悪化した点

「バス」等の乗り物を示すクラスタ検出に失敗した．pLSA による次元整理の結果，温度パラメータ β に関わらず，名詞との共起組合せ数が小の見立て詞では，効果消失の可能性があることが確認された．

4.3.　温度パラメータ比較

前述のように温度パラメータが 1.0 の場合に，より小さなクラスタ検出の可能性が示唆された．例えば図 9 では「倉庫（青線）」，「箪笥，棚（赤線）」が別クラスタに分類されているように見える．しかし，図 10 で同単語は別クラスタではない．逆に，中央から右にまたがる大規模なクラスタに分類され，家具・部屋・建物などが混在している．一方，β=1.0 の方には部屋や建物に関連したクラスタが見られる（図 4，図 5 の中央下側）．したがって，このクラスタについては，特に 1.0 の場合により有効なクラスタリング結果につながっていると考えられる．

5.　おわりに

本研究は，コーパスから容器状物質の形状抽出を試み，SOM を使用した実験を行った．実験の結果，物体の高さ情報をマップ上に表現する等，本手法の有効性を確認した．また，温度パラメータの設定により，クラスタの大きさが変更されることを確認した．以上の結果により，言語直感の実現が可能となると考えられる．

今後の課題としては，コップと浴槽のように大きさが異なる物体の別クラスタへの分類が挙げられる．動詞に着目する等，新たな分析を行った上で，こうした課題を解決したい．

謝辞

この研究の一部は，平成 22 年度広島市立大学特定研究費（一般研究）の補助を得ている．関係各位に感謝申し上げる．

参考文献

Hofmann, T. (1999). ” Probabilistic Latent Semantic Indexing. ” in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval”, pp.50-57.

池原悟・宮崎正弘・白井諭・横尾昭男・中岩浩巳・小倉健太郎・大山芳史・林良彦(1997)『日本語語彙大系』岩波書店.

金明哲 (2003).“自己組織化マップと助詞. 分布を用いた書き手の同定及びその特徴分析.”計量国語学，pp.369-386，計量国語学会.

神崎享子，戸室宣子，井佐原均 (2007).“自己組織化マップによる形容詞抽象概念の階層関係・類義関係の自動抽出.”言語処理学会年次大会，pp.986-989.

Kohonen，T.(2001).“Self-Organizing Map, 3rd Edition.”徳高平蔵，岸田悟，藤村喜久郎訳 (2005).“自己組織化マップ.” シュプリンガー・ジャパン.

工藤拓. “PLSI”, http://chasen.org/~taku/software/plsi/

Kurosawa, Y., Mera, K., and Takezawa, T. (2010). “Psychomime Classification and Visualization Using a Self-Organizing Map for Implementing Emotional Spoken Dialog System.” In Spoken Dialogue Systems Technology and Design, Wolfgang Minker, W., Lee, G. G., Nakamura, S., and Mariani, J. (eds), pp.107-134, Springer.

黒澤義明，原章，市村匠 (2008).　“換喩検出を目的とした自己組織化マップ SOM による物体の形状マップ生成.”言葉と認知のメカニズム，pp.353-374，ひつじ書房.

馬青，神崎享子，村田真樹，内元清貴，井佐原均 (2001).　“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，pp.2379-2391，情報処理学会.

野口英司(2005)『インターネット図書館 青空文庫』はる書房.

som_pak, “som_pak.” http://www.cis.hut.fi/research/som_pak/

山梨正明(1988)『比喩と理解』東京大学出版会.