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

神崎享子[1]　戸室宣子[2]　井佐原均[1]

[1.] 独立行政法人情報通信研究機構

[2.] DeRaul University

１．　はじめに

本稿では、シソーラスで見られる構造のような、形容詞をカテゴライズする概念のタクソノミーを、コーパスから抽出する。具体的には、自己組織型神経回路網モデル（ＳＯＭ）(Kohonen 1997)によって、形容詞をカテゴライズする概念の類義関係と上位下位関係を求め、構造化された関係を2次元平面に可視化する。

タクソノミーの自動的な可視化よって、分類で得られた大きなクラスタや階層の塊どうしの、全体の構造の中での相対的関係をとらえることができる。それと同時に、形容詞の様相も、全体的に把握することができる。そして、現在利用されている人手で構築されたシソーラスやオントロジーなどを、実データからの自動分類の結果に基づいて、構造的に比較、再検討したりすることができる。

２．　データ

コーパスから形容詞をカテゴライズするような抽象的な名詞を取り出すために、根本(1969), 高橋(1975)などが着眼した、形容詞をカテゴライズする名詞の意味関係を、コーパスから探し、データ収集を行った。方法は、ＸがＹを範疇化するパタン（益岡 1994）「ＸトイウＹ」という文型を手がかりにＸが形容詞、Ｙが名詞というパタンをコーパスからとりだし、その後、Ｘの範疇化と考えられるＹを軸に、それを修飾する形容詞をコーパスから抽出し、Ｙと形容詞のデータを取り出し、ある程度、人手でデータを取捨選択した。「形容詞の概念名」として取り出された抽象名詞は、９４、９５年の毎日新聞２年分から取り出した。抽象名詞と共起する形容詞、形容動詞は、毎日新聞１１年分、日本経済新聞１０年分、産業金融流通新聞７年分、読売新聞１４年分、新潮文庫１００選、新書版１００冊の中から用例を調べた。抽出された抽象名詞は 365 語、形容詞の異なり語が10525 語、のべ語数は 35173 語であった。最大共起語数は、「こと」に対する 1594 語である。このリストの中で、頻度５以上出現した 2374 語の形容詞と 361 語の抽象名詞を対象とした。

本稿では形容詞をカテゴライズする意味関係をもつ名詞を形容詞概念として定義する。たとえば、「楽しい、嬉しい、悲しい」をカテゴライズする名詞「気持ち」は、形容詞概念としてここでは考える。

３．　TOP ノードを導入した SOM

入力データは、コーパスから抽出した、形容詞が属する概念（形容詞をカテゴライズするラベルとなる抽象名詞）のリストである。このリストの特徴は、共起形容詞が多くバラエティーに富んでいるほど、概念の抽象度が高くなることである。上位語から下位語へ（抽象から具体へ）の方向性あるこのデータを自己組織型神経回路網モデルＳＯＭによって分類することで、類義関係と上位下位関係を同時にとらえることを試みた。

Kohonen の自己組織型マップ(1997)は、教師なし学習で、入力データを n 次元空間に並んだノードに投射するものである。通常は入力データが多次元で、マップは 2 次元である。入力データは、第２節で述べたデータで、抽象名詞 361 語に対して、それぞれ形容詞 2374 のベクトルがある。

我々は、SOM マップを作るために、SOM_PAK (Kohonen et al, 1996)を利用した。我々は、全てのフィーチャーに１の値をもつ、最も抽象度の高い、「ＴＯＰノード」を設定した。
抽象・具体は、共起形容詞の数によって、かなりよい精度で決定されるので（Caraballo and Charniak, 1999)、全ての形容詞が修飾する名詞は、最も抽象度の高い名詞と考えられる。
学習の間、TOP ノードはマップの中心に割り当てられている（SOM_PAK のオプションで利用できる)。
結果として、マップ中心にある TOP 付近には、「こと」「状態」印象」など、共起する形容詞が多種多様で数も多く抽象度の高い名詞が分布し、そこから周辺へは、「家柄」「仲」「光沢」など、共起形容詞が特化した抽象名詞が分布していく、というマップが得られた（図１)。

さらに、SOM がどのような意味ある単語分布を出力しているかを探るために、SOM の学習によって得られる類義関係と、TOP を設定したことで得られる抽象から具体への単語分布の方向性（上位下位関係）をとらえ、形容詞をカテゴライズする抽象概念（形容詞が表現する属性概念）の構造的な分布について考察する。

４．タイトクラスター　―SOM マップから得られる抽象名詞の類義関係―

マップ上の名詞は、ノードが同じか近いノード上に分布していたら、類義語であると考えられるが、時に類義語が異なったノードに分布すること

がある。アルゴリズムの自己組織化は、設定するパラミーターに影響を受けやすいからである。

我々はこの問題点を解決するために、参照ベクトルがとても近いマップノードからタイトクラスターを抽出した。Average cosine coefficient (Salton and McGill 1983)は、閾値はノーマライズした値で 0.96 以上を対象にした。

$$\cos(v1, v2) = \frac{\sum_i v1_i \times v2_i}{\sqrt{\sum_i v1_i^2 \times \sum_i v2_i^2}}$$

vj は v の j 番目の特徴を表す。

図１では、タイトクラスターを丸で囲んでいる。

次に、我々の方法によって得られるタイトクラスタを分類語彙表の類義語と比較してみた。最初に、同じクラスタに分類された抽象名詞が、分類語彙表でも同じ分類項目になるかどうかを調べた。

タイトクラスは、８８つくられ、８１語はクラスタをつくらなかった。クラスタの例としては、以下のようになる。

- 時刻, 時間
- 気質, 気性, 気風,人柄
- 効果, 影響
- 関係, つながり
- やり方, 感覚, 言い方, 態度, そぶり
- 道のり, 形勢, 情勢, 状況, 環境
- 形状, かたち, 形態

……

これらが、分類語彙表でも、同じ分類項目になるかどうかを調べた。分類語彙表の番号が一致すれば、同じカテゴリということを表す。その結果、52% (46/88)のクラスタに、分類語彙表と一致する類義語が含まれ、47%(42/88)のクラスタが分類語彙表の類義語とは一致しなかった。半分以上が、言語学者の作成した「分類語彙表」の類義語と一致している。

分類語彙表と一致しなかった単語グループの中には、分類語彙表には類義語として登録されていないが、検討してもよいと思われる単語セットもみられた。

たとえば、「心, 明るさ」というセットでは、「明るさ」には物理的な光の「明るさ」しかなく、心の明るさは登録されていない。また、これ以外にも「勢い,速度」「迫力、剣幕」「身分、地位」「境遇、身の上」など、類義語、あるいは上位語・下位語として再検討してもよい思われるセットが、SOM から得られた。

次に、我々は、クラスターを作った名詞が、分類語彙表の特定のカテゴリに集中しているか否かを探った。生成されたタイトクラスタは 88 あり、このタイトクラスタに属する抽象名詞は、分類語彙表では 114 カテゴリに属していた。名詞が属しているカテゴリの範囲が、分類語彙表と 77%重なっていることがわかった。

以上の結果から、タイトクラスタの分類が、precision が 52%、recall が 77%であり、精度も高く、またいくつかのカテゴリに偏った分布になっていないことがわかる。

クラスタのマップ上での分布傾向をみると（図１）、「TOP」を中心に、「感じ、感情、気持ち、様子」などが一つの大きなクラスタとなる。その周りを「状態、情勢」などのクラスタ、「見方、言い方、態度、性格」などのクラスタ、「イメージ、姿」、「雰囲気、気配、持ち味」などのクラスタが囲んでいる。そしてまたその周りを、関係の密なクラスタが星のように散らばっている。

５．ＳOMノードから得られる階層関係

第三節でも述べたように、本マップでは、最上位にくる TOP を決め、そこから抽象名詞を分布させたことで、上位レベルから下位レベルへと名詞が分布している。既に，第４節では類義語をグループ化した。本節では、抽象名詞の階層を生成することで、抽象名詞や類義語のクラスタが、どのように他の名詞やクラスタと関係しているか、マップ上の分布の方向性を探る。

そして類義関係と上位下位関係の両面から、SOM によるマップ上の単語分布を捉える。

上位下位関係は、共起形容詞に基づいて Cosine と Entropy によって求めた。Entorupy による階層構築は Sharon Caraballo & Eugene Charniak によってその有用性が述べられている。

Cosine によって二単語間の類似度を求め、Entropy によって、どちらの単語が上位かを決めた。また、SOM の結果を反映させるために、SOM のノードから階層の構築を行った。

Fixed Center 5 million のマップで、Cosine を閾値 0.8 にして Entropy で上位下位を求めると、77 の階層が生成された。この 77 の生成された階層のうち、妥当そうな階層だけを選択する方法として、形容詞の継承性を一つの手がかりとした。

本研究では、抽象名詞が形容詞カテゴリーを示す概念と仮定している。形容詞は、そのカテゴリーの一事例（インスタンス）になっており、このインスタンスは、最上位概念か最下位概念まで継承されるものである。たとえば、「寿司」は、最上位概念の「もの」の一事例であり、「食べ物」概念の一事例であり、「料理」概念の一事例であり、最下位概念「日本料理」の一事例である。「寿司」は最上位概念から最下位概念までずっと成員となっている。この特徴を考慮し、生成された 77 階層のうち、形容詞が最上位概念から最下位概念まで連続して出現している階層を良さそうな階層と考え、SOM のマップ上での階層の分布をみた。

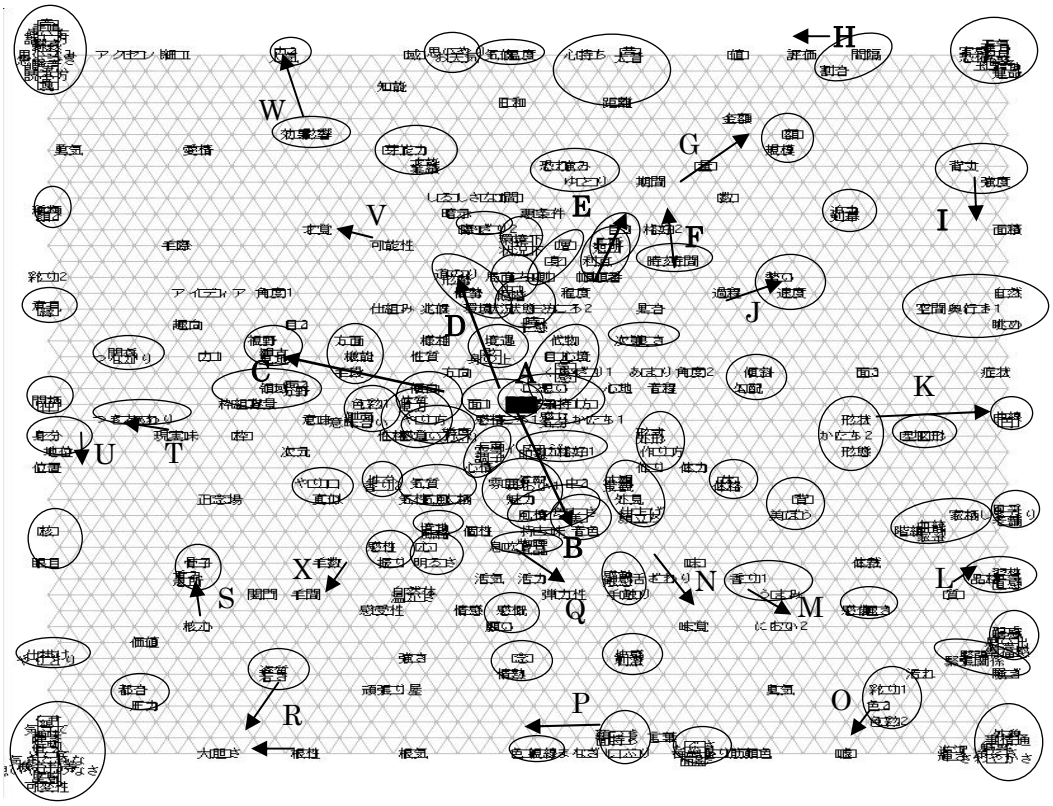
この考え方には問題点もあり、最下位ノードにある抽象名詞が一つの形容詞しか共起しない場

合、形容詞が共通していれば下位概念となるため、一見関係のない単語と結びつく可能性もある。その場合、多義語の形容詞の場合は、異なるノードをもつ可能性がある。たとえば、この方法だと「背丈」は「高い」という形容詞が一つだけ共起しており、「高い、安い」などの共起語をもつ「金額」が、たとえば、日本語シソーラス「分類語彙表」でその類似性をみると、それほど離れた関係では選択された階層は、77 階層中 37 階層あり、たとえば以下のようなものがある。[　]は、類義語のクラスタを示す。
の下位語になる。制約をさらに取り入れるべきだ
[面持ち, 顔つき, 口ぶり] → まなざし　→　[色 1, 視線] : いぶかしげな
過程 → [勢い, 速度] : めまぐるしい
[形状, 形態, かたち]　→ [型, 図形] : シャープな, 立体的な, 細長い、幾何学的な, 四角い

６．ＳOMマップに見られる類義関係と上位下位関係

直接的にマップ上の単語や類義語クラスタの関係を観察する方法としては、類義語クラスタや生成された階層を、マップ上にプロットさせることである。そこで、類義関係をもつ名詞のクラスタを丸で囲み、階層関係の方向をとらえるために、名詞や類義語クラスタを、矢印でつなげた。その結果が以下の図である。

　形容詞が継承しているかどうかをフィルターにし、77 階層のうち 37 の階層を得、それぞれの階層は、形容詞を一事例としてもつ属性概念の階層ということになる。図の黒い四角が我々が挿入した TOP である。



　この図から、名詞や類義語クラスタが TOP を最上位点として下位へ向かって放射状に分布していることがわかる。
　以下、階層と類義語クラスタの分布にあわせて、階層の分布している領域ごとに、どのような抽象名詞が分布しているかをいくつか述べる。

　1）中心の A 周辺 : TOP と同じクラスタには「こと」「感じ」「思い」「様子」などが分類されている。感情に関する抽象名詞形容詞が多く分布しておりクラスタも重複している。

　2）中心 A から B 方向、C 方向、D 方向、E,F,G,J,方向、K 方向、MN 方向、X、R、S 方向へと単語が放射状に分布している。特に B、C、D 方向は、類義語クラスタが多く一つの大きな塊になり、中心からわずかに四方へ移動している。抽象度が高くて、共起形容詞の重複が多いため、名詞を弁別的に分類することは難しいが、ある程度は共起形容詞に、傾向が見られることを表す。

階層例：

[ところ１　　顔１　　感情　　面 1]->
方向 -> 傾向 -> 性質 -> 様　相
-> 兆候 -> [情勢 道のり　形勢]

この領域に属する形容詞：
不幸な,乏しい,危険な,困難な,少ない,難しい、有利な、不利な

6.考察

本マップによって、新聞 2 年分の抽象名詞の範囲であるが、日本語形容詞の概念体系を、捉えた。形容詞概念として、「TOP,こと、感じ、様子など」にはほとんどの形容詞が共起しているということは、ほとんどの形容詞がこの概念を具体的に表現できるということである。また、そこを起点に分布しているのは、状態、程度、外観・外形、感覚（五感)、印象・雰囲気、人やものごとの特徴・人やものごとに対するかかわり、見方・観点・分野、能力・才能、効果・影響　などの抽象概念であることがわかった。それより外側のマップの端の方に位置するものは、抽象名詞がさらに具体性があり、特徴的な形容詞が共起している。たとえば、「評価など」や「階級、家柄など」「配慮、思いやり、心づかいなど」「顔つき、面持ち、くちぶり、まなざしなど」「身分、地位」「間柄、仲」などが分布している。

本マップは、近傍ノードとの類似関係の信頼性が向上したことで、かなり精度の良い類義語のクラスタが得られた（50%以上は日本語シソーラスと完全に一致し、残りの 50%には、単語が登録されていないものや、自動生成の類義語をフィードバックして再検討すべきものも含まれている)。そしてこの類義語のクラスタを踏まえ、ＳＯＭから上位下位関係の構造をとらえた。上位下位関係を求めることで、単語相互の関係を、横と縦の関係を含めて構造的に捉えることができた。たとえば、「雰囲気、気配、魅力」などは、「印象、イメージ、すがたなど」の下位概念であり、「イメージ、印象、すがた」などはＴＯＰの「感じ・様子」に近いので、感情形容詞によって表現されることも多い。

ＳＯＭによって単語を分類する利点の一つに、Visualization があげられる。本マップで抽象名詞が類義語のクラスタにならなくても、あるクラスタや名詞の近くにあれば、近い意味であることがわかる。既存のシソーラスに登録されていない重要な抽象名詞も、近いところに位置する抽象名詞をキーにして捉えることができた。たとえば、「臭気」は「香り・うまみ→におい」の階層の下位方向に位置しており、また、「角度」は「傾斜・勾配」の上位方向に位置しているので、それぞれ、類似性がありさらに、上位下位の関係であることもわかる。このように、未知語やうまくクラスタを作らなかったものなども、どの語と関係が近いかがわかる。

また、ＳＯＭのＭＡＰでは全体の関係も捉えられる。たとえば、「状態」と「程度」はとても近く、「形や図形」は、「程度」より、むしろ、「外形・外観」と近いこと、また、人の気性やものごとの特徴は、「印象・すがた・雰囲気」などを表現する形容詞などとも関係が深いことがわかる。

今後は、形容詞がどのような場合に、ある抽象概念から別の抽象概念へスイッチするのか、その条件も考え、取り入れる必要があろう。

７．終わりに

本研究の特徴は、形容詞をカテゴライズする抽象名詞を、ＳＯＭによるマップで類義関係と階層関係を同時に捉えたことである。そして、形容詞が表現する抽象概念体系を、明確な特徴をもって構造的に捉える。

＜参考文献＞

Caraballo S. A. and Charniak E. 1999. Determining the Specificity of Nouns from Text., *In Proceedings the joint SIGDAT conference on empirical methods in natural language processing (EMNLP) and very large corpora (VLC)*, 63-70.

Kohonen T., Hynninen J., Kangas J., and Laaksonen J. 1996. *SOM_PAK: The Self-Organizing Map Program Package. Technical Report A31,* Helsinki University of Technology, Laboratory of Computer and Information Science. http://www.cis.hut.fi/research/som_lvq_pak.shtml

Kohonen.T. 1997. *Self-organizing maps, 2nd Edition*. Springer

Salton G. and McGill M. J.　1983. Introduction to Modern Information Retrieval.　McGraw Hill.

益岡隆志. 1994.「名詞修飾節の接続形式―内容節を中心に―」田窪行則編「日本語の名詞修飾表現」くろしお出版.

根本今朝男. 1969.「が格」の名詞と形容詞とのくみあわせ. 電子計算機のための国語研究Ⅱ, 国立国語研究所, pp. 63-73.

高橋太郎. 1975. 文中にあらわれる所属関係の種々相. 国語学103 国語学会 *pp.1-16*.