形容詞に関係する概念体系の自動構築の試み

― 階層の自動構築とその評価 ―

神崎享子[1]　山本英子[1]　馬青[1,2]　井佐原均[1]

[1] 独立行政法人情報通信研究機構　　[2] 龍谷大学

1．はじめに

本研究は、実データから、辞書などを用いずに形容詞に関係する概念体系を自動構築することをめざすものである。そのためには以下の点を考慮する必要がある。

１）実データから、どのように形容詞の概念を抽出するか。
２）どのように概念体系を自動構築するか。
３）構築した概念体系を、人間の直感を用いずにどのように評価するか。

従来の研究で、辞書を用いずに自動的な単語分類をしたものには、二つのタイプがある。

一つめのアプローチしとしては、統語パタンなどの表層構造を利用して単語分類を行う方法がある（Hindle 1990, Hatzivassiloglou and Mckeown 1993, Tokunaga et al. 1995)。二つめのアプローチとしては、言語学的知見を利用して単語分類を行う方法がある(Walde and Brew 2002, Boleda et al. 2004)。

我々は、全ての単語を対象にして表層構造を利用するのではなく、特定の関係の形容詞と抽象名詞の共起関係を利用する。抽象名詞は、カテゴリ名や上位概念を定義するときに用いられるので、形容詞の上位概念を抽出するためのデータとして利用することにした。したがって、本稿では、「コーパスから抽出したある特定の抽象名詞」が「概念」相当と仮定する。また、単語の概念体系を構築するためには、類義関係や上位下位関係からなる全体の骨組みと、実際の運用で単語と単語が結びつくための規則が必要であるが、われわれは、現在、単語の類義関係や上位下位関係からなる全体の骨組みに焦点をあてている。本稿では、そのうち、上位下位関係の自動構築について焦点をあてる。

具体的には、共起する形容詞の類似性によって Kohonen (1997)の自己組織型マップ（SOM）を用いて抽象名詞を分類し、単語の上位下位関係を求めるために、補完類似度（CSM）を用いた。また、最終的には、他の手法と比較実験を行い、評価を試みた。

2．データ

抽象的な名詞の統語的役割に着目した先行研究には、根本（1969), 高橋（1975)などがあげられる。例えば高橋（1975)においては、

①やぎは性質がおとなしい
②ぞうは鼻が長い

の二例を比較し、①の「性質」を側面語、②の「鼻」を部分語と仮に呼び、文中の役割が異なることを述べている。側面語になる単語は主語の示すものや人の側面を表すとともに、述語の示す属性の類概念（上位概念）を表す単語である。また、根本（1969)においても「色が白い」「速さがはやい」「年が若い」「背が高い」などは、「顔が赤い」などのような名詞が状態の持ち主を表す場合と違って同義反復的な性格が強いと述べている。このように、我々の言語活動の中にも、形容詞の上位概念を示すような用法がみられる。このようなパタンは、形容詞の抽象的意味をコーパスから探るのに重要な手がかりになるのではないかと考える。

対象とした抽象名詞は、９４、９５年の毎日新聞二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、毎日新聞十一年分、日本経済新聞十年分、産業金融流通新聞七年分、読売新聞十四年分、新潮文庫百選、新書版百冊の中から用例を調べた[1]。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。最初に作成され

[1] 用例を検索するにあたっては情報通信研究機構で開発したツール Tea を利用した。

るのは、以下のような抽象名詞とそれを修飾する形容詞の表である。

思い：うれしい　楽しい　悲しい ……
気持ち：楽しい　嬉しい　幸せな ……
観点：医学的な　歴史的な　学術的な ……

3．SOM を用いた抽象名詞の自動分類

第二節の単語のリストを SOM の入力とするには、符号化する必要がある(Ma et al. 2000)。

ここで、一般に ω種類の名詞 w_i (i = 1, … , ω)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。

思い＝ {悲しい、楽しい、幸せな、…}

$$w_i = \{ a_1^{(i)}, a_2^{(i)}, \ldots, a_{\alpha i}^{(i)} \}$$

ただし、$a_j^{(i)}$は、w_iと共起する j番目の連体修飾要素である。この後、二語間の類似度計算をする。それを元に、Ma et al. (2000)の「相関コーディング法」を用いる。ここで提案する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符号化する。

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

V（w_i）は h への入力であり、この多次元ベクトルを、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、二次元空間に表現する。

本研究では、類似尺度に補完類似度を用いる。山本・梅村(2002)の補完類似度は、包含関係を取り出す類似尺度である。意味マップの一つの問題点として、マップ上での語どうしの関係がわかりづらいということがある。補完類似度を用いることで意味マップ上に分布している名詞どうしの上位下位関係がわかる。山本・梅村(2002)によれば、補完類似度は以下のような式になる。

今、 $\vec{F} = (f1, f2, \ldots, fi, \ldots, fn)(fi = 0\text{または } 1)$

$$\vec{T} = (t1, t2, \ldots, ti, \ldots tn)(ti = 0\text{または}1)$$

とする。

$$Sc(\vec{F}, \vec{T}) = \frac{ad - bc}{\sqrt{(a+c)(b+d)}}$$

ここで、a は、二つのラベルが同時に現れるデータの数、b は *label 1* が現れ、*label 2* は現れないデータの数、c は、*label 2* が現れ、*label 1* は現れないデータの数、d は、二つのラベルがどちらとも現れないデータの数である。本データに対してこの尺度を用いる際には、*label* にあたるのが抽象名詞となり、a は、ある形容詞が双方の抽象名詞と共起しているパタン、b と c は、ある形容詞が一方の抽象名詞とだけ共起しているパタン、d は形容詞がいずれの抽象名詞とも共起していないパタンの数ということになる。そして最後に補完類似度の数値を正規化し、相関コーディング法によって多次元ベクトルに符号化して入力データにする。CSM を導入したマップは以下のようになる。上位下位関係が推定された二単語間に線を引いた図である。

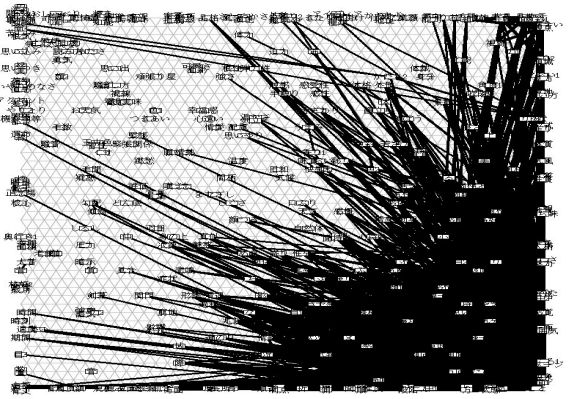

図１．CSM に基づいた抽象名詞のマップ

右下角を基点にして、放射状に単語が分布していることがわかる。右下角には、「こと」という抽象名詞があり、そこから離れるに従って、たとえば、「時間」などのように比較的具体的な単語になっていく。

しかし CSM は基本的に二単語間の上位下位関係を推定するものである。全体がどのような上位下位関係をもっているのかを知るために、CSM の数値を利用して単語の階層化を図った。

4．CSM による階層化

以下に CSM を用いて階層化構築の手順を示す。

(1) 包含関係を示す補完類似度の値の高い順に単語 A, B をつなげる。ここでは、仮に単語 A が上位語、単語 B が下位語という

関係とする。

(2) 単語Aを下位語として、最高値で上位語となる単語Xを探してAの前に連結するというように、A-B を基点にして上位（前）へ向かって連結を繰り返す。一方、単語Bを上位語として、最高値で下位語となる単語Yを探しBの後ろに連結するというように、A-B を基点として下位（後ろ）に向かって連結を繰り返す。上位下位関係は必ず保存する。もし上位下位関係が壊れたら、その関係は連結しない。こうして一本の階層を作る。

(3) 長い階層に完全に含まれる短い階層はマージし、二つの階層のうち一単語ずつ異なる場合は、差異となる二単語の補完類似度が上位下位関係を示せば結合した。

(4) 最後に各階層のトップに「こと」を結合する。「こと」は全ての形容詞と共起することができ、意味マップ上でも、一番右下に配置されている。計算時間の便宜上、「こと」は最後に各階層のトップに結合させることとした。

5．階層の比較

CSM で構築した階層を頻度つきの CSM (Tf.CSM)と Overlap Coefficient (Ovlp)によって同様に階層を構築し比較を行った。Overlap Coefficient も、単語の上位下位関係を推定する類似尺度である(Manning and Shütze 1999)。以下の表は、各手法によってできた階層の深さごとの階層数を表している。

Depth	***3***	***4***	***5***	***6***	***7***	***8***	***9***
CSM	0	3	16	27	32	23	23
Tf.CSM	1	5	10	18	13	25	11
Ovlp	32	56	61	57	21	7	2
depth	***10***	***11***	***12***	***13***	***14***	***15***	**計**
CSM	19	7	3	4	3	1	161
Tf.CSM	24	13	14	14	7	2	158
Ovlp	2	0	0	0	0	0	240

表１　三手法による深さ別階層数

表から、最も深い階層を作る手法が Tf.CSM で、最も浅い階層を作る手法が Ovlp であることがわかる。また階層の合計を比較すると、Ovlp が 240 階層と、最もたくさん階層を作っている。CSM はどちらかというと、Tf.CSM に近い。

次に、階層のつながり方を共起形容詞から検討する。最下位を n 番目とすると、n 番目と n-１番目の抽象名詞について、それぞれの手法で共起形容詞の差が最大になった場合について表にした。

	$n\text{-}1^{st}$	n^{th}	Differences
CSM	10 adjs	2 adjs	8 adjs
Tf.CSM	6 adjs	1 adjs	5 adjs
Ovlp	1594 adjs	1 adjs	1593 adjs

表２　n^{th} と $n\text{-}1^{st}$ の共起形容詞の数

表２から、Ovlp による階層では、n^{th} と $n\text{-}1^{st}$ の階層の共起形容詞の数の差が最も大きい。これは、中間的な抽象名詞を経由せずに、上位の抽象名詞と下位の抽象名詞が直接的に結合していることを意味する。

次に共起形容詞を利用して、各手法による階層で、どれくらい形容詞が上位から下位まで共通して出現しているかを調べる。たとえば、「ふくろう」は、「鳥」で「動物」で「生物」であり、これら全ての概念をもっている。したがって、最下位の抽象名詞と共起している形容詞がどれくらい上位に位置する抽象名詞と共通して共起しているかを調べることで、階層のつながりのよさを評価した。そして、以下のような結果を得た。

CSM　11/161 (0.068)
Tf.CSM　9/158 (0.056)
Ovlp　88/240 (0.366)

この結果から、最下位まで共通形容詞が連続して出現している階層が少ないことがわかる。そこで、階層の構築方法について以下の２点を再検討した。

１）長い階層を短い階層にマージしたこと
２）CSM と Yates を組み合わせた手法の妥当性

そこで、CSM のみを使い、また長い階層と短い階層をマージしない方法で、共通形容詞の連続性という点から評価を行った。すると以下のような結果になった。ここで記号の後ろの数字はＣＳＭの正規化した値がこの数字以上である範囲で階層を作ったことを示す。

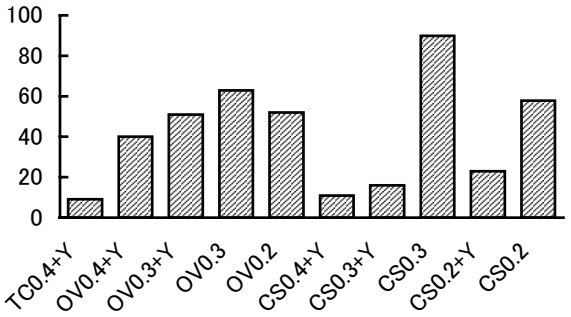

図２　各手法の共通形容詞を持つ階層の数の比較

CSM0.3 については９０％の階層で共通形容詞が最上位から最下位の抽象名詞まで連続して出現していたが、階層の数自体が５４階層しかできず、また、階層を構成している名詞も少数であるので対象外とする。これは、CSM の数値が高いほど信頼度は上がるが、上位下位関係を推定された単語対の種類が少ないことによる。すると、Ovlp0.3 と CSM0.2 がそれぞれ６３％と５８％の階層で共通形容詞が最上位から最下位へ連続して出現することがわかった。そこで、共通形容詞の出現の連続性という観点では、Yates を用いず、なるべくその手法一つで階層を用いたほうがよいことがわかる。

ただし、全ての数値で階層を作ると、上位下位関係が崩れランダムで冗長な単語の並びとなることが経験的にもわかっている。

そこで、Ovlp0.3 と CSM0.2 の階層の比較であるが、Ovlp の場合は、短い階層がたくさんでき、最上位と最下位が直接結合する傾向にあることは、第五節の最初に述べた傾向と同様である。したがって、共通形容詞の出現の連続性という点では、なるべく Yates を使わず、また長短の階層をマージしない CSM0.2 の手法が、もっともらしい階層だという結果になった。

６．まとめ

形容詞が一つの具体事例となる抽象名詞を概念ととらえ、SOM と CSM を用いて、抽象名詞を分類し階層化することを試みた。そして、包含関係を推定するほかの尺度とも比較を行い、また、最下位の抽象名詞の共起形容詞が最上位まで連続して出現しているかという点で階層を評価した。すると、CSM の正規化した数値 0.2 までで上位下位関係を推定した単語対を用いることが、もっともらしい階層を作ることがわかった。今後は、SOM 上での単語の類義関係を推定するとともに、共起形容詞の側も詳細に分析する。

参考文献

[Boleda et al., 2004] G. Boldea, T. Badia and E. Batlle. Acquisition of Semantic Classes for Adjectives from Distributional Evidence. *In Proceedings of the 20th International Conference on Computational Linguistics* (COLING), VolumeⅡ, pp. 1119-1125, 2004.

[Hatzivassiloglou and McKeown, 1993] V. Hatzivassiloglou and R.K. McKeown. Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 172-182, 1993.

[Hindle, 1990] D. Hindle. Noun Classification From Predicate-Argument Structures. *In Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 268-275, 1990

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

[Ma et al, 2000] Q. Ma, K. Kanzaki, M. Murata, K. Uchimoto and H. Isahara. Self-Organization Semantic Maps of Japanese Nouns in Terms of Adnominal Constituents, *In Proceedings of IJCNN'2000*, Como, Italy, Vol. 6., pp. 91-96, 2000.

[Manning and Shütze, 1999] C.D. Manning and H. Shütze. Foundations of Statistical Natural Language Processing, The MIT Press, 1999.

[根本, 1969] 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.

[Tokunag et al., 1995] T. Tokunaga, M. Iwayama and H. Tanaka. Automatic Thesaurus Construction Based On Grammatical Relations. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence* (IJCAI), pp. 1308-1313, 1995.

[高橋, 1975] 高橋太郎, 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, 1975.

[Walde and Brew, 2002] S. S. Walde and C. Brew. Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. *In Proceedings of the 40th Annual meeting of the Association for Computational Linguistics* (ACL), pp. 223-230, 2002.

[山本・梅村, 2002] 山本英子, 梅村恭司. コーパス中の一対多関係を推定する問題における類似尺度. 自然言語処理, vol.9, No.2, pp.46-75, 2002.