コーパスからの形容詞概念階層の自動構築と EDR 概念体系との比較評価

神崎享子† 山本英子† 馬青†‡ 井佐原均†
†独立行政法人情報通信研究機構／‡龍谷大学

１．はじめに

本研究は、コーパスからの形容詞概念階層の自動構築と、得られた概念のＥＤＲ概念体系との比較評価法について述べたものである。

上位下位関係の研究は、Hearst (1992)、Caraballo (1999) 以降多く見られるが、対象にしている品詞が名詞や動詞であり、その手法は形容詞に対してそのままでは使えない。また、上位下位関係も二単語間の関係が主流であって、最上位のノードから最下位のノードまで通して体系的に構造化する手法についての研究はまだ発展途上である。

本研究では、形容詞の概念階層の自動構築を複数の方法で行い、生成された階層の特徴を、階層の作られ方の観点から比較し、妥当そうな階層を選択した。階層の構造の良否や、形容詞の階層としての適否をこのように数値的あるいは構造的に把握するだけでは、外見的な評価になりかねないので、人間の内省による階層の質の評価もあわせて行うこととし、得られた階層を人手で人間の直感に基づいて作られたＥＤＲ概念体系と比較した。

２．データ

コーパスから形容詞を範疇化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、形容詞を範疇化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(神崎他 2003)。方法は、X が Y を範疇化するパタン（益岡 1994）である「X トイウ Y」という文型を手がかりに X が形容詞、Y が抽象名詞というパタンをコーパスからとりだした。続いて、X の範疇化と考えられる抽象名詞Ｙを修飾する形容詞をコーパスから抽出し、ある程度、人手でデータを取捨選択した。

「形容詞の概念名」として用いる抽象名詞 Y 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。

３．三種類の方法による階層の自動構築

神崎他(2003), Kanzaki et al.(2004)では、劣化印刷文字の認識のために澤木・萩田（1995a, 1995b）によって開発され、山本・梅村(2002)が自然言語処理に応用した、包含関係を表す尺度である CSM を採用して、二単語間の上位下位関係を推定した。本研究でも、２節で構築した形容詞グループの類似性に基づいて、まず頻度を考慮せずに、CSM を用いて抽象名詞の階層を構築した。そして、さらに、Manning and Shütze (1999)で、包含関係を推定する尺度として紹介されている Overlap Coefficient(以下 Ovlp と記述する)を用いた頻度情報を考慮しない階層と、コーパス内の形容詞と名詞の共起頻度を考慮した CSM(以下 Freq)によって構築された階層(Yamamoto et al. 2005)を、比較の対象とした。

階層構築方法は以下のようになる。

(1) 包含関係を示す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) まず、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。次に、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、上位下位関係は必ず保存する。上位下位関係が壊れる場合は、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、二つの階層が一単語だけ異なる場合は、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、それに沿って結合した。

(4) 最後に各階層の最上位に「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、最も抽象的な概念と考えることができる。計算時間の便宜上、「こと」は最後に各階層の最上位に結合させることとした。こうして、最終的に抽象名詞によって構成される、「こと」を最上位概念とした階層が得られる。

４．複数の手法を使って自動生成された階層間の比較　―妥当な手法と閾値の判定のために―

CSM、Ovlp、FreqCSM によって自動構築された階層を、その作られ方の観点から比較検討し、妥当そうな階層を特定する。

CSM, Ovlp はそれぞれ閾値を、正規化した数値 0.3 と 0.2 の２つに設定し(CSM0.3, CSM0.2, Ovlp0.3, Ovlp0.2 と表示する)、頻度を使った CSM の閾値は、0.2 と 0.1 に設定した(Freq0.2, Freq0.1)。これらの閾値は実験の結果得られたものであり、これ以上閾値をあげると、極端に階層を構築する名詞が少なくなり、これ以上閾値を下げると、冗長でしかも上位下位関係が、見た目にもバラバラな階層ができる。ここで用いる閾値によって、適度な数の名詞が階層化される。

自動生成されたこれらの階層の中から、最も妥当そうな階層を見つけるため、まず、自動生成された階層のうち、形容詞の階層として生成された階層数を調べる。これによって、自動生成された階層のうちの、外見的に妥当そうな階層を判定する。「形容詞の階層」とは、具体的には、最上位の抽象名詞から最下位の抽象名詞まで共通して共起する形容詞が存在する、ということである。これは、言い換えると、最下位のカテゴリーにある事例は、最上位のカテゴリーまで、共通の事例（ここでは形容詞に相当）を持っている、という階層構造の一つの特徴を表す。たとえば、スズメは鳥のカテゴリーにも動物のカテゴリーにも生物のカテゴリーにも事例として属する。

次に、自動生成された階層が、対象とした形容詞のうちのどれくらいの割合をカバーできたか、そして、対象とした抽象名詞のうちのどれくらいの割合をカバーできたか、を調べる。

これらに加え、自動生成された階層の特徴をみるために、階層のノードの深さごとの階層数を調べる。その結果を図１に示す。階層の長さを調べるのは、単に階層の長短によって妥当性を判断するためではなく、階層の特徴を見るためである。

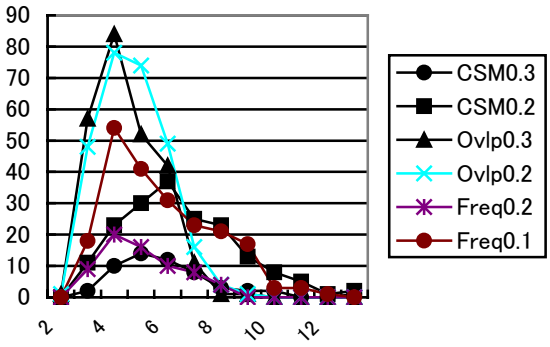

図１　手法別階層の深さと階層数

	階層の深さ(最多ノード)	*形容詞の階層としてできた階層の割合*	*階層が生成された形容詞数(語)*	*階層を構成する名詞の割合*
CSM0.3	13*(事実上 10)	90 %	175 語	× 24 %
CSM0.2	13	58 %	101 語	90 %
Ovlp0.3	9	63 %	99 語	86 %
Ovlp0.2	9	52 %	53 語	92 %
Freq0.2	8	53 %	70 語	× 28 %
Freq0.1	12	× 36 %	67 語	82 %

CSM0.3 の 13* については深さ 10 でほとんどの階層が作られており、深さ 11 や 12 の階層数は 0 なのだが、深さ 13 の階層が 1 つだけ作られた。

表１　階層の作られ方からみた手法別の階層の特徴

図１から、Ovlp0.3(▲)と Ovlp0.2(×)が階層の深さに関して同じような特徴を持つ、すなわち、浅い階層が多くできていることがわかる。また、CSM0.2(■)と Freq0.1(◆)では比較的深い階層ができていることがわかる。

次に、1） 自動生成された階層のうち、形容詞の階層として生成された階層の割合、2）対象にした形容詞のうちで階層に含まれた形容詞の数、3）対象にした 365 語の名詞のうちで階層に使われた名詞の割合を表１にまとめる。

表１では、数値がよいものベスト３には○、極端によい場合は太線の○をつけた。また、数値が極端に低いものには×を付与した。上記の３つの特徴から総合的にみると CSM0.2 と Ovlp0.3 が、形容詞の階層としてできた階層の割合も、階層が生成された形容詞数や 365 語の名詞のうちで階層を構成する名詞の割合もよいとわかる。CSM0.3 は、表１に示す６種類の階層の中で形容詞の階層を最も多く作っているが､階層を構成している名詞の数が最も低い(つまり precision は高くて recall が低いとも言える)。これは同じ階層をもつ形容詞のグループが、未分化である可能性がある。また、Ovlp0.2 の階層は、形容詞の階層はあまり作られていないが、対象にしている 365 語の名詞をほとんど使って、階層を作っている(つまり、precision は低いが recall が高いとも言える)。これは、冗長に階層を作っている可能性がある。程度の差こそあれ同様の傾向がみられるのは Freq0.1 である。得られた形容詞の階層は少ないが、階層を構成している名詞が多いことがわかる。Freq0.2 は、形容詞の階層の数は少なめであるが、それより顕著な特徴は、階層を構成する名詞の種類が少ないことである。程度の差こそあれ、その点では、CSM0.3 に似た傾向がみられる。

５．自動生成された概念階層とＥＤＲの概念階層との比較実験

ＥＤＲは計算機用大規模辞書である。このＥＤＲの概念階層は人手作業によって構築されているので、人間の直感が反映されたものである。そこで、自動生成された階層と人手による階層で、どちらがより良い形容詞の階層ができているかを、人間が評価する実験を行うことにした。

ＥＤＲの概念階層の各ノードには、概念ＩＤが付与されており、そのそれぞれに概念記述が付与されている。概念記述は単語で定義されている場合もあれば､文で説明していることもある。ＥＤＲでは「肯定的な」という形容詞の概念階層は複数作られているが、その一例をあげると次のようなものがある。

ＥＤＲ階層：「肯定的な」

概念(3aa966)→事象(30f7e4)→移動(30f801)→情報の移動(30f832)→情報の受信(3f96e7)→知る(30f876)→認知主体と認知対象との認知的距離減少(3f972c)→(意見などに)同意しているさま(0f0ae2)

一方、自動生成された階層でも概念階層が複数作られているが、各ノードの概念は、抽象名詞で表現されている。その一例に次のようなものがある。

自動生成の階層：「肯定的な」

こと→言い方→言葉→評価

一見してわかるように、概念の定義の仕方が、ＥＤＲと自動生成の階層では異なっている。ＥＤＲの概念記述文の表現は、人間のための説明であり、概念階層を構築するときにそれほど厳密なものではなく、むしろ上位から下位へのノードの繋がりが適当かどうかが比較のポイントになる。

形容詞や形容動詞に妥当な階層が作られているかどうかを評価するために、形容詞や形容動詞の階層として自動生成された階層をＥＤＲの階層と比較した。

比較実験で対象にした手法と閾値は、4 節の結果から CSM で閾値 0.2、Ovlp で閾値 0.3 そして、頻度を導入した CSM で閾値 0.2、の３種類とした。また、被験者は 20 人で、辞書編纂経験者、言語学や自然言語処理を背景知識としてもつ人から構成されている。

実験は各手法が作った階層とＥＤＲの階層を対象に一対比較法で行った。被験者がそれぞれの階層とＥＤＲの階層と比べて、ある形容詞に対して、直感的にどちらが妥当な階層として選択するかを求めた。

6. まとめ

本研究では階層の構築を行い、手法と閾値別に比較を行った。その結果、CSM で閾値 0.2 と Ovlp で閾値 0.3 が、階層の作られ方の観点でよいことがわかった。また、ＥＤＲと自動生成の階層について人手による評価実験を行った。

今後、形容詞や形容動詞の上位下位関係について精度を高める方法を探り、また現在構築している自己組織化マップ(Kanzaki 2004)に、最も妥当な上位下位関係の階層を反映させたいと考える。

参考文献

EDR Electronic Dictionary. 1995. http://www2.nict.go.jp/kk/e416/EDR/index.html

Sharon A Caraballo. 1999. Automatic Acquisition of Hyponym – Labeled Noun Hierarchy from Text. *In proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*: pp.120-126.

Marti. A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. *In proceedings of 14th International Conference on Computational Linguistics*: pp.539-545.

神崎享子, 馬青, 山本英子, 村田真樹, 井佐原均. 2003. 形容詞が内包する抽象的意味の抽出と自動分類の試み, 言語処理学会第9回年次大会発表論文集.

Kyoko Kanzaki, Qing Ma, Eiko Yamamoto and Hitoshi Isahara. 2004. Construction of an objective Hierarchy of Abstract Concepts via Directional Similarity, Proceedings of the 20th International Conference on Computational Linguistics (COLING2004), pp.1147-1153.

Christopher D. Manning and Hinrich Shütze. 1999. Foundations of Statistical Natural Language Processing, The MIT Press.

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.

澤木美奈子, 萩田紀博. 1995a. 補完類似度による劣化印刷文字認識, 電子情報通信学会 信学技法, PRU95-14, pp.101-108.

澤木美奈子, 萩田紀博. 1995b. 補完類似度に基づく新聞見出し文字の領域抽出と認識. 信学技法, PRU95-106, pp.19-24.

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.

Eiko.Yamamoto, Kyoko.Kanzaki, and Hiroahi. Isahara. 2005. Extraction of hierarchies based on inclusion of co-occurring words with frequency information, the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005), pp.1166-1172.

山本英子, 梅村恭司. 2002. 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.