共起語の包含関係に基づく語彙の階層化への頻度情報の影響

山本英子　神崎享子　井佐原均

独立行政法人　情報通信研究機構

eiko@nict.go.jp　kanzaki@nict.go.jp　isahara@nict.go.jp

1. はじめに

語彙の階層関係は言語資料として有用である．これまでにさまざまな観点から階層関係を含むシソーラスの構築がなされ，公表されている．これらの資料は元となる資源と編集者に依存し，それぞれ独自の体系に基づき，手作業で構築されている．そのため，個々のユーザの思考と合致しない体系も存在する．また，既存のシソーラスにおいて，同義語や類義語が列挙されているが，これらの間にある意味的または統計的な階層関係が明記されていない場合がある．我々はこれまでに，語彙の客観的な階層構造を得るために，二値画像認識のために開発された補完類似度を用いて，コーパスから階層構造を自動抽出することを試みた[Yamamoto et al., 2004]．その文献では，語彙の階層関係を見つけるために，抽象名詞と形容詞・形容動詞との共起関係を二値ベクトルで表し，そのベクトル間の包含関係を補完類似度で測ることによって，二語間ごとの上位下位関係を推定し，その値を元に語彙を連結していくことで，語彙の階層化を行う手法を提案している．本稿では，それをさらに進めて共起関係を表すベクトルの要素として，共起頻度に基づく重みを用い，多値画像認識のために改良された補完類似度を適用し，語彙の階層化への頻度情報の影響を考察する．

2. 抽象名詞に関する言語コーパス

本稿では，提案する手法が階層関係を構築する問題に適用可能であるかを示すため，形容(動)詞の上位語として定義された抽象名詞[Kanzaki et al., 2003]の階層関係を自動構築することを試みる．本稿で対象とした語彙は，2 年分の毎日新聞に含まれた抽象名詞である．この抽象名詞に前接する形容(動)詞を 100 の小説，100 のエッセイ，11 年分の毎日新聞，10 年分の日本経済新聞，7 年分の産業金融流通新聞，14 年分の読売新聞に含まれ，KNP によって構文解析できた文から，収集した．このようにして収集した語彙は抽象名詞 354 種類，形容(動)詞 6407 種類である．

3. 補完類似度

本稿では，階層関係を構築するために，二語間の関係を統計的指標で推定し，それを元に階層関係を構築する．このため，補完類似度という文字認識の分野で開発された類似尺度を適用する．

3.1. 二値画像のための補完類似度

補完類似度は劣化印刷文字を認識するために提案された類似尺度である[Hagita and Sawaki, 1995]．二値画像のための補完類似度(Complementary Similarity Measure for binary images: CSM-b)はテンプレート文字と印刷文字を二値ベクトルで表し，ベクトル間の包含関係を測る尺度である．ベクトル $F=(f_1,\ldots,f_i,\ldots,f_n)$ と $T=(t_1,\ldots,t_i,\ldots,t_n)$ (f_i, $t_i=0$ または 1) における補完類似度は次のように定義される．

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

$$a=\sum_{i=1}^{n} f_i\cdot t_i, \qquad b=\sum_{i=1}^{n} f_i\cdot(1-t_i),$$

$$c=\sum_{i=1}^{n}(1-f_i)\cdot t_i, \quad d=\sum_{i=1}^{n}(1-f_i)\cdot(1-t_i),$$

$$n=a+b+c+d$$

この定義式において，n は次元数，a,b,c,d は要素間の関係を表すパラメータである．本研究では，n は形容(動)詞の種類数に相当し，各要素 f_i, t_i は抽象名詞が i 番目の形容(動)詞と共起するならば 1，共起しなければ 0 である．この要素で構成されるベクトル F, T は抽象名詞がどの形容(動)詞と共起出現し，どれと共起出現しないのかを表す出現パターンである．したがって，パラメータ a はベクトル F を持つ抽象名詞とベクトル T を持つ抽象名詞の双方と共起する形容(動)詞の数，b はベクトル F を持つ抽象名詞とは共起するが，他方とは共起しない形容(動)詞の数，c はベクトル T を持つ抽象名詞とは共起するが，他方とは共起しない形容(動)詞の数，d はどちらの抽象名詞とも共起しない形容(動)詞の数に相当する．

3.2. 多値画像のための補完類似度

多値画像のための補完類似度(Complementary Similarity Measure for gray-scale images:

CSM-g)は二値画像のための補完類似度 CSM-b を拡張した尺度である[Sawaki et al., 1997]. CSM-b はグラフィカルデザインなどの汚れに強いが，二値化状態やスキャンの条件が強く影響する．CSM-b は 2×2 分割表の特殊な例であり，その一般形として，CSM-g を定義した．この尺度は直接，グレースケールで表される多値画像を扱うため，二値化状態やスキャンの条件に影響されにくいという特徴を持つ．ベクトル $F_g = (f_{g1},\ldots,f_{gi},\ldots,f_{gn})$ と $T_g = (t_{g1},\ldots,t_{gi},\ldots,t_{gn})$ (f_{gi} , $t_{gi} = 0$ から 1) における補完類似度は次のように定義される．

$$CSM_g(F_g,T_g) = \frac{a_g d_g - b_g c_g}{\sqrt{nT_{g2} - T_g^2}}$$

$$a_g = \sum_{i=1}^{n} f_{gi}\cdot t_{gi}, \qquad b_g = \sum_{i=1}^{n} f_{gi}\cdot(1-t_{gi}),$$

$$c_g = \sum_{i=1}^{n}(1-f_{gi})\cdot t_{gi}, \quad d_g = \sum_{i=1}^{n}(1-f_{gi})\cdot(1-t_{gi}),$$

$$T_g = \sum_{i=1}^{n} t_{gi}, \quad T_{g2} = \sum_{i=1}^{n} t_{gi}^2$$

この定義式において，二値画像のための補完類似度と同じように，次元数は形容(動)詞の種類数に相当し，各要素 f_{gi}, t_{gi} は抽象名詞と i 番目の形容(動)詞とが共起するかどうかを表すが，0 または 1 ではなく，共起頻度に基づく重みである．実験では，下記のような重みを用いた． 式中の $Freq(noun, adj)$ は抽象名詞 $noun$ が形容(動)詞 adj と共起する頻度である．

$$Weight(noun, adj) = \frac{Freq(noun, adj)}{Freq(noun, adj) + 1}$$

4. 階層関係の構築方法

コーパスからの階層関係の構築工程を示す．

4.1.階層構築工程

1. 各尺度を用いて，単語対ごとに出現パターン間の類似度を測り，包含関係を推定する．この推定された包含関係を二語間の階層関係とする．単語対 X,Y の間で，X が上位語，Y が下位語と推定された場合，その単語対を(X,Y)と表す．
2. 類似度を正規化し，閾値 TH 未満の単語対を削除する．
3. 各語彙 C について，
   (ア) 上位語が C である，かつ最も高い類似度を持つ単語対(C,D)を階層の初期値とする．
   (イ) 階層の最後尾に位置する単語 D を上位語に持つ単語対の中で最も高い類似度を持つ対(D,E)を見つけ，E を階層の最後尾に連結する．ただし，下位語 E は現行の階層に含まれていないものに限る．
   (ウ) 工程(イ)に沿った単語対(E,F)が選択できる間，工程(イ)を繰り返す．
   (エ) 階層の先頭に位置する単語 C を下位語に持つ単語対の中で最も高い類似度を持つ対(B,C)を見つけ，B を階層の先頭に連結する．ただし，上位語 B は現行の階層に含まれていないものに限る．
   (オ) 工程(エ)に沿った単語対(A,B)が選択できる間，工程(エ)を繰り返す．
4. 構築した階層について，もし短い階層が順序の保たれた状態でより長い階層に包含されるなら，短い階層を階層の集合から削除する．

4.2.条件設定

実験において，二つの CSM を用いて構築した階層を比較するために，抽象名詞「こと」から始まる階層をできるだけ多く構築するように，工程 2 で用いる閾値を設定した．「こと」は意味的に広く使える抽象名詞であり，もっとも出現頻度が高い抽象名詞である．実験的に設定した閾値 TH はそれぞれ以下のような値である．

- CSM-b の場合，TH = 0.2
- CSM-g の場合，TH = 0.12

5. 比較評価

5.1.抽出された階層間の比較

まず初めに，CSM-b を用いて得られた階層と，CSM-g を用いて得られた階層を比較する．表 1 に抽出された階層の数を示す．

表 1：抽出された階層の数

	階層の種類	数
(A)	CSM-b によって得られた階層	189
(B)	CSM-g によって得られた階層	178
(C)	共通して得られた階層	28
(D)	CSM-g の階層を包含する CSM-b の階層	5
(E)	CSM-b の階層を包含する CSM-g の階層	38

実験において，CSM-b と CSM-g の双方で得られた共通の階層は 28 階層しかなく，そのほとんどは深さ 3 から 6 と短い階層であった．また，一方の補完類似度で得られた階層を他方の階層が包含する階層を比べると，CSM-b の階層を包含する CSM-g の階層は，CSM-g の階層を包含する CSM-b の階層より多い((D) < (E))．これは，CSM-g が CSM-b よりも長い階層を抽出できる特徴を持つことを示唆している．

CSM-b の階層を包含する CSM-g の階層の例を次に示す．ここで，下線を引いた抽象名詞は包含される CSM-b の階層に現れなかった名詞である．

- こと -- 時 -- 日和 -- 温度 -- 幸福感
- こと -- ところ -- しぐさ -- 面影 -- 可愛さ

実験において抽出された階層の深さを比較すると，CSM-b の階層は深さ 3 から 12 の範囲に，CSM-g の階層は深さ 3 から 15 の範囲に分布する．また，階層を構成する抽象名詞の異なり数を比較すると，全 354 種類中，CSM-b は 318，CSM-g は 314 であり，網羅性には差が見られなかった．表 2 に深さによる階層の分布を示す．これらの結果から，CSM-g は抽出できる階層の数は CSM-b より少ないけれども，より長い(深い)階層を抽出できることがわかる．

表 2：深さによる階層の分布

深さ	3	4	5	6	7	8	9
CSM-b	4	33	45	36	22	16	20
CSM-g	5	29	24	32	29	18	10

深さ	10	11	12	13	14	15	計
CSM-b	7	2	4				189
CSM-g	16	8	4	1	1	1	178

5.2.EDR 電子化辞書にある階層との一致度

次に，それぞれ得られた階層を人手によって構築された EDR 電子化辞書 [1995]にある形容(動)詞の概念階層と比較し，階層の一致度を測る．EDR 中の形容(動)詞に関する概念階層は 932 階層あり，深さ 3 から 14 の範囲に分布する．実験において得た階層はそれぞれ，深さ 3 から 12，3 から 15 の範囲に分布するため，この EDR が持つ階層との比較は得られた階層を評価することに適していると考えた．

しかし，EDR の概念階層を構成する各概念は単語ではなく，概念 ID と説明文で記述されているため，抽象名詞で構築した階層と比較するには，変換が必要であった．そこで，各概念記述について内容語である名詞，動詞を取り出し，さらに，それらの単語に類義語を付与し，その列で文を置き換えた．同様に，抽象名詞にも類義語を付与し，使用単語の違いを軽減した．用いた類義語は EDR 電子化辞書から抽出した語である．このように変換した階層について一致度を測る．ここで，一致度とは順序を保持した一致ノード数である．

具体的には，構築された階層の各ノードにある抽象名詞は，その抽象名詞とその類義語で，ノード(抽象名詞，類義語 1，類義語 2, . . .)と表す．一方，EDR から抽出した階層の各ノードにある概念は，その概念記述にある内容語とそれらの類義語で，ノード(内容語 1，類義語 1_1，類義語 1_2，内容語 2，類義語 2_1，類義語 2_2, . . .)と表す．このとき，構築した階層のノードにある単語が EDR の階層のあるノードの内容語や類義語と一致するのであれば，そのノードは EDR の階層のノードと一致すると考える．たとえば，実験において構築された階層が以下のように表され，

A(a, a', a") -- B(b, b', b") -- C(c, c', c") -- D(d, d', d")

この階層に対応する EDR の階層が以下のように表現される階層であるならば，

P(a, a', x, x') -- Q (b, b") -- R(r, r', r") --
S(s, s', d, f, f', f') -- T(t, t', g, g')

構築された階層は EDR の階層と一致するノードを三つ持っていると数え，その階層の一致度は 3 と定義する．下線で示す単語は二つの階層間で一致する単語とその単語を持つノードを示し，構築された階層のノード A，B，D はそれぞれ P，Q，S と一致することを表している．

このように，階層を比較する中で，構築された階層において上位下位関係にある単語が EDR 電子化辞書では類義語となる場合があることがわかった．たとえば，我々の手法では次のような階層を構築する．

- こと --- ところ --- イメージ ---
  雰囲気 --- 空気 ---
  感情 --- 心情 --- 心境 ---
  感慨 ---思い出

EDR 電子化辞書では，類義語とは「同じ概念にリンクされる単語」と定義されており，我々は EDR における類義語を集められる．実際に，上の階層において， EDR では「心情」と「心境」が「感情」の類義語であり，「空気」と「雰囲気」も類義語である．EDR には「感慨」は出現しない．

上の例で、EDR の階層との一致度を厳密に数えた場合，「こと --- ところ --- イメージ--- 雰囲気(または，空気) --- 感情(または，心情，心境) --- 思い出」が一致し，一致度は 6 となる．しかし，もし EDR では類義関係にある単語間の上位下位関係を許すならば，一致度は 9 となる．本稿では，後者の方法で，EDR の階層との一致度を測った．

表 3 に EDR の概念階層と CSM-b による階層との一致度の深さごとの分布を示す．同様に，表 4 に CSM-g による階層の分布を示す．たとえば，深さ 3 の CSM-b 階層は 4 つあり，そのうち EDR との一致度が 1 のものは 1 つ，一致度が 2 のものは 2 つ，一致度が深さと同じ 3，すなわち双方が完全に一致するものは 1 つある．この完全に一致する部分は表中で下線付きで表している．

「平均」は深さ 3 の階層の一致度の平均((1*1+2*2+1*3)/4=2.00)であり，「全体の平均」は全階層の一致度の総和を階層の総数で割った値である．

これらの表から，より深い階層はより高い一致度を持つ傾向にあることがわかる．表 4 においては，その傾向を表 3 よりはっきりと観ることができる．階層の深さごとに一致度の平均を比べると，深さ 8 と 9 以外の深さにおいて，CSM-g のほうが CSM-b より高い値を持つことがわかる(図 1)．これらのことから，CSM-g は全体的に CSM-b より EDR の概念階層に近い，つまり，共起頻度を考慮したほうが，人間の直感に近い階層を構築していると考えられる．

表 3：深さごとの CSM-b 階層の一致度の分布

階層の深さ	EDR の概念階層との一致度									平均
	1	2	3	4	5	6	7	8	9	
3	1	2	**1**							2.00
4		6	18	**9**						3.09
5		7	23	12	**3**					3.24
6		4	12	9	7	**4**				3.86
7		2	2	10	4	3	**1**			4.32
8			1	6	6	3				4.69
9			1	4	5	4	5	1		5.55
10			2		2	2			1	5.29
11			1			1				4.50
12			1	1					2	6.25
全体の平均										4.28

表 4：深さごとの CSM-g 階層の一致度の分布

階層の深さ	EDR の概念階層との一致度									平均
	1	2	3	4	5	6	7	8	9	
3	1	3	**1**							2.50
4		6	13	**10**						3.14
5		3	9	9	**3**					3.50
6		1	11	12	6	**2**				3.91
7		1	5	10	8	5				4.38
8			4	5	7	2				4.39
9				6	1	3				4.70
10					2	6	4	3	1	6.69
11			1	2	1	3	1			5.13
12								1	3	8.75
13							1			7.00
14								1		8.00
15									1	9.00
全体の平均										5.47

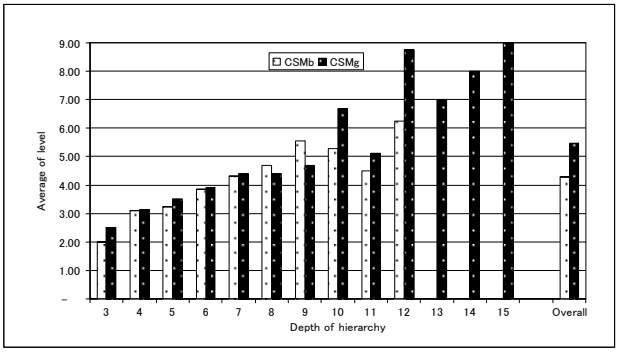

図 1：深さごとの一致度の平均による比較

また，EDR の概念階層と一致したノード(抽象名詞)を分析すると，多くの場合，最高位近くの上位概念と一致する傾向にあった．既存のシソーラスでは，語彙のカテゴリ化は人間の直感に基づきトップダウン方式で分類されている．したがって，我々は少なくとも上位概念の最高位の辺りにおいては，人間の直感にあった抽象名詞の階層関係を構築できると考察する．

6. まとめ

本稿では，共起関係を表すベクトルの要素として，共起頻度に基づく重みを用い，多値画像認識のために改良された補完類似度を適用した．そして，構築された階層を EDR の概念階層との一致度において比較した結果，共起頻度を考慮したほうがより人間の直感に近い階層を構築しうることを示した．

参考文献

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

[Hagita and Sawaki, 1995] Hagita, N. and Sawaki, M. Robust Recognition of Degraded Machine-Printed Characters using Complimentary Similarity Measure and Error-Correction Learning, In *Proceedings of the SPIE – The International Society for Optical Engineering*, 2442: pp. 236-244, 1995.

[Kanzaki et al., 2003] Kanzaki, K., Ma, Q., Yamamoto, E., Murata, M. and Isahara, H. Adjectives and their abstract concepts --- Toward an objective thesaurus from a semantic map. In *Proceedings of the Second International Workshop on Generative Approaches to the Lexicon*, pp. 177-184, 2003.

[Sawaki et al., 1997] Sawaki, M., Hagita, N. and Ishii, K. Robust Character Recognition of Gray-Scaled Images with Graphical Designs and Noise, In *Proceedings of the International Conference on Document Analysis and Recognition,* pp. 491-494, 1997.

[Yamamoto et al., 2004] Yamamoto, E., Kanzaki, K. and Isahara, H. Hierarchy Extraction based on Inclusion of Appearance, In *ACL04 Companion Volume to the Proceedings of the Conference*, pp. 149-152, 2004.