F-004

出現状況の包含関係を利用した語彙の階層関係の自動構築

山本 英子† 神崎 享子† 井佐原 均†
Eiko Yamamoto Kyoko Kanzaki Hitoshi Isahara

1. まえがき

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．この背景に伴い，我々はユーザが扱う情報に基づき階層関係を自動的に抽出することを考える．本稿では，コーパス中の出現状況の包含関係を利用し、階層関係の構築を試みる。そこで、包含関係を測る統計的指標として，補完類似度とオーバーラップ相関係数を検討する．実験では，それらの指標の適用性を示すために，抽象名詞を取り上げ，階層関係の構築を試みる．そして，EDR 電子化辞書と比較し，考察する．

2. 階層関係構築の概要

2.1.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おいて適用可能であるかを示すため，形容詞・形容動詞の上位語として定義された抽象名詞(Kanzaki 2003)の階層関係を自動構築することを試みる．使用するコーパスは形容(動)詞を分類するための足がかりとなるべく上位語である抽象名詞を分類することを目的として，多種類の新聞データなどから各形容(動)詞と隣接する抽象名詞の集合を収集したものである．

2.2.統計的指標による二語間の関係推定

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によって推定し，それを基に階層関係を構築する．二語間の関係を推定する問題において，補完類似度とオーバーラップ相関係数を検討した．用いる情報はコーパスにおいて抽象名詞が各形容詞・形容動詞と共起するかどうかを表す出現状況である．

2.2.1. 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 1995)．この尺度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．これまでに，この尺度の特徴を生かし，各語彙の出現状況をベクトル化し，コーパス中の一対多関係を推定する問題に適用されている (山本 2002)．本稿では，上位語である語彙は広義語であるため，下位語である狭義語よりも頻繁に用いられる傾向にある．その語彙対を出現状況で比較すると，重なる（包含する）状況が観察できる．この包含関係を測定する．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 or 1) における補完類似度は次のように定義される．

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

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

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

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

パラメータ a,b,c,d はそれぞれ次の数を表す．

a : 二語がどちらとも共起する形容(動)詞の数
b,c : 一方が共起し，他方が共起しない形容(動)詞の数
d : 二語がどちらとも共起しない形容(動)詞の数

2.2.2. オーバーラップ相関係数

本稿では、比較対象として、補完類似度と同様に、二つのベクトル間の包含関係を測る特徴を持っているオーバーラップ相関係数を比較対象とした(Manning 1999).

$$OVLP(F,T) = \frac{|F \cap T|}{MIN(|F|,|T|)} = \frac{a}{MIN(a+b,a+c)}$$

2.3.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．本稿では，二語間の関係推定結果から，閾値を 0.2 とした．閾値を低く設定すれば，深い階層関係を得られるが，低すぎる値を持つ関係は信用しがたい．そのため，閾値を設け，0.2 未満の単語対は考慮しないこととした．

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．
2. 類似度を正規化し，ソートする．
3. 各語彙 B について，
   (ア) B が上位語であり，かつ最も高い類似度を持つ単語対 (B,C)を選択する．この単語対を階層の初期値とする．
   (イ) 現行の階層の最下位語 C を上位語とし，下位語 D が現行の階層に含まれておらず，かつそのような単語対の中で最も高い類似度を持つ単語対(C,D)を選択する．
   (ウ) 下位語 D を階層の最後尾に連結する．
   (エ) そのような単語対が選択できる間，(イ)と(ウ)を繰り返す．
   (オ) 現行の階層の最上位語 B を下位語とし，上位語 A が現行の階層に含まれておらず，かつそのような単語対の中で最も高い類似度を持つ単語対(A,B)を選択する．
   (カ) 上位語 A を階層の先頭に連結する．

† 独立行政法人 情報通信研究機構,
National Institute of Information and Communications Technology

(キ) そのような単語対が選択できる間，(オ)と(カ)を繰り返す.

4. 構築した階層について，
   (ア) もし短い階層が語彙の順序が保持さ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.
   (イ) もし二つの階層の違いが小さい（数個の語彙についての関係が異なる）なら，二つの階層を統合する.

3. 実験結果

実験において，すべての階層の最上語は「こと(事)」であった．これは意味的に広く使える抽象名詞であり，出現頻度が高い語彙である．実験において得られた階層をいくつか示す.

こと -- 面 -- イメージ -- 印象 -- 顔立ち -- 品格 -- 血筋
こと -- 面 -- イメージ -- 印象 -- 感じ -- 気分 -- 気持ち -- 感情 -- 心情 -- 心境 -- 感慨 -- 思い出
こと -- 面 -- イメージ -- 印象 -- 風格 -- 家柄 -- 血統 -- 家系 -- 血筋
こと -- 面 -- イメージ -- 性格 -- 印象 -- 一面 -- 態度 -- 人柄 -- 気質 -- 気風 -- 気性
こと -- 面 -- イメージ -- 美しさ -- 若さ -- 大胆さ
こと -- 面 -- イメージ -- 体 -- 体格 -- 背
こと -- 面 -- 側面 -- 意味 -- 方向 -- 観点 -- 目 -- 視野 -- 角度 -- アイディア
こと -- 状態 -- 関係 -- かかわり -- つきあい
こと -- 状態 -- 状況 -- 兆候
こと -- 状態 -- 傾斜 -- 勾配
こと -- 時 -- 温度 -- 幸福感
こと -- 規模 -- 数 -- 量

4. EDR 電子化辞書との比較評価

本稿では，EDR 電子化辞書(1995)から形容(動)詞の上位概念を抽出し，評価に用いることとした．しかし，EDR の上位概念は語彙ではなく，文で表されているため，構築した階層と比較するためには，整形が必要であった．そこで，上位概念について名詞，動詞を取り出し，文と置き換えた．さらに，階層に現れる各語彙に「分類語彙表」(1964)にある同義語を付与し，使用語彙の違いを軽減した．このように整形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語彙数である．たとえば，「A - B - C - D – E」と「A - B - D - F – G」では，「A - B - D」と順に一致するので，一致度は 3 となる． 図 1 と 2 に各手法の階層の一致度(”level”)を示す.

まず、深さについて考察する。EDR の階層は深さ 3 から 14 である．図 1 から補完類似度の階層は深さ 3 から 14 であり，図 2 からオーバーラップ相関係数の階層は深さ 2 から 9 である．これにより，補完類似度のほうがオーバーラップ相関係数よりも深い階層を得られることがわかる．オーバーラップ相関係数では，深さが 4 から 6 に集中しており，その深さの階層を観ると，関連ある名詞が別々の階層に分かれていた．これに対し，補完類似度では，関連ある名詞が一つの階層内に組み込まれていた.

次に，一致度については、深さ 5 や 6 の階層を比較すると，補完類似度のほうが高い一致度を持つことがわかる．このことから，本実験では補完類似度のほうが EDR の階層関係と一致する階層を抽出できたと考える．どちらの手法とも level2 から 4 である階層が多いことがわかる.

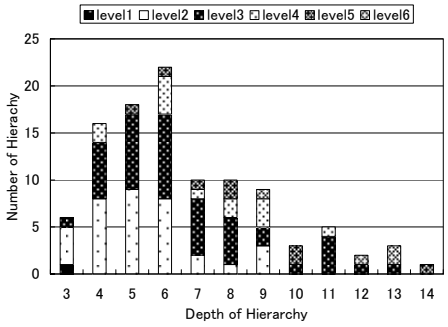

図1 補完類似度に関する一致度の分布

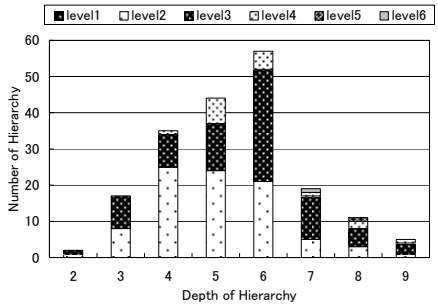

図2 オーバーラップ相関係数に関する一致度の分布

一致した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．現在のシソーラスでは，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たと考察する.

5. まとめ

本稿では，コーパスから出現パターンの包含関係に基づき，自動的に階層関係を抽出する方法を提案した．そして，提案手法の適用可能性を示すために，形容(動)詞の上位語として定義される抽象名詞の階層関係を構築することを試みた．今後の課題は、他の階層化手法との比較や EDR との相違を分析することである.

参考文献

EDR 電子化辞書, 1995. http://www2.nict.go.jp/kk/e416/EDR/

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

Kanzaki, K., Ma, Q., Yamamoto, E., Murata, M., and Isahara, H. 2003. Adjectives and their Abstract concepts --- Toward an objective thesaurus from Semantic Map. In Proceedings of the Second International Workshop on Generative Approaches to the Lexicon, pp.177-184.

国立国語学研究所,「分類語彙表」, 秀英出版, 1964.

Manning, C. D. and Schutze, H. 1999. Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge MA.

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.45-75.