E-003

回帰分析による上位語推定の検討

別所 克人　　牧野 俊朗　　松尾義博
日本電信電信電話株式会社　メディアインテリジェンス研究所

1．はじめに

用語の上位下位関係を規定した知識は、情報検索や要約、含意認識、談話解析といった自然言語処理を高精度に行うために重要なものとなっている。本稿では、WordNet[1]に代表される上位下位関係を規定したシソーラスを拡張する手法を提案する。具体的には、大量の文書から獲得した単語の概念ベクトルの集合である概念ベース[2]を用いて、シソーラス内の下位語の概念ベクトルと、上位語の概念ベクトルとの対応関係を回帰分析により学習する。この結果得られた最小自乗行列を、任意の語の概念ベクトルに適用することにより、上位概念相当の概念ベクトルが得られるので、この上位概念相当ベクトルに近い概念ベクトルをもつ単語を上位語と推定する。

2．回帰分析による上位語推定

本稿で対象とするシソーラスは、(りんご, 果物), (みかん, 果物), (大根, 野菜), (白菜, 野菜)のような下位語と上位語の対からなるレコードの集合である。

文献[2]ではテキストコーパスから、単語とその意味属性間の共起頻度行列を生成し、この行列に対し SVD を施して列数を縮退させて得られる行列の各単語に対応する行ベクトルを、概念ベクトルと呼び、単語とその概念ベクトルの対の集合を概念ベースと呼ぶ。各単語の概念ベクトルは k 次元空間内にあり、長さ 1 に正規化され、意味的に近い単語の概念ベクトルは近くに配置されている。

シソーラス中のレコード集合を $\{(l_t,u_t)|1\le t\le n\}$ とする。ここで、レコード数が n であり、(l_t,u_t) が t 番目のレコードで、l_t が下位語、u_t が上位語である。

下位語 l_t の構成単語の概念ベクトルの和を長さ 1 に正規化したものを、l_t の概念ベクトル $(x_{1t},x_{2t},\cdots,x_{kt})$ とする。また、上位語 u_t の構成単語の概念ベクトルの和を長さ 1 に正規化したものを、u_t の概念ベクトル $(y_{1t},y_{2t},\cdots,y_{kt})$ とする。

下位語概念ベクトルを並べた行列 X、上位語概念ベクトルを並べた行列 Y に対し、数式(1)を満たす行列 B、E を考える。行列 X の各下位語概念ベクトル L に右から行列 B を乗じて得られるベクトル V に、行列 E の対応する行ベクトルを加算したものが、行列 Y の対応する上位語概念ベクトル U となる。

文献[2]での共起頻度行列生成では、各単語に対し、各成分が意味属性に対応するベクトルで、各成分値が、該

数式(1)

$$Y=\begin{pmatrix} y_{11} & y_{21} & \cdots & y_{k1} \\ y_{12} & y_{22} & \cdots & y_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1n} & y_{2n} & \cdots & y_{kn} \end{pmatrix}$$
$$=\begin{pmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}\left(\begin{pmatrix} \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1k} \end{pmatrix}\begin{pmatrix} \beta_{21} \\ \beta_{22} \\ \vdots \\ \beta_{2k} \end{pmatrix}\cdots\begin{pmatrix} \beta_{k1} \\ \beta_{k2} \\ \vdots \\ \beta_{kk} \end{pmatrix}\right)$$
$$+\begin{pmatrix} \varepsilon_{11} & \varepsilon_{21} & \cdots & \varepsilon_{k1} \\ \varepsilon_{12} & \varepsilon_{22} & \cdots & \varepsilon_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{1n} & \varepsilon_{2n} & \cdots & \varepsilon_{kn} \end{pmatrix}=XB+E$$

単語と、該成分に対応する意味属性との共起頻度であるようなベクトル（共起ベクトルと呼ぶ）を生成する。一般に上位語は、下位語が共起する意味属性と共起し、下位語が共起しない意味属性ともわずかに共起する傾向があると推測される。そのため、図 1 は 1 共起ベクトルの成分値の分布（全成分値の和が 1 となるように正規化したもの）をヒストグラムで表現したものであるが、上位語の共起ベクトルの成分値の分布（図 1 の破線部分）は、下位語の共起ベクトルの成分値の分布（図 1 の実線部分）と近く、下位語のそれと比べて、より一様となる。下位語の共起ベクトルで、値が比較的大きい成分は、上位語の共起ベクトルでは、やや小さくなるものの他の成分と比べ大きいままであり、反対に、下位語の共起ベクトルで、値が比較的小さい成分は、上位語の共起ベクトルでは、やや大きくなるものの他の成分と比べ小さいままである。次元を縮小した概念ベクトルでも、同様な傾向がある。図 2 は、概念ベクトルの次元数を 2 とした場合の上位語・下位語の概念ベクトルの関係を示した図である。概念ベクトルは、半径 1 の球面上にある。下位語概念ベクトルは、一般に、特定の成分群の値が大きく、該成分群の軸がなす平面に近い位置にある。上位語概念ベクトルは、下位語概念ベクトルに近く、かつ、成分値の分布がより一様になる方向（原点を通る破線の直線に近い方向）に位置する傾向にある。

上位語概念ベクトル U と、V との差分である行列 E の対応する行ベクトルのノルムができるだけ小さくなるように行列 B を設定する。こうすると、下位語概念ベクトル L と上位語概念ベクトル U は上述のように近いので、U の第 i 成分値を導出するのは行列 B の第 i 列だが、L の第 i 成分値がなるべく反映されるように、該行列 B の第 i 列においては、第 i 成分値が大きく、他の成分値は小さくなる傾向がある。図 2 では U の第 i 成分値を導出するのは、行列 B の第 i 列を係数とする平面で、任意の L

Estimation of Hypernyms based on Regression Analysis
Katsuji Bessho, Toshiro Makino, and Yoshihiro Matsuo
NTT Media Intelligence Laboratories, NTT Corporation.
1-1 Hikarinooka, Yokosuka-Shi, Kanagawa, Japan

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

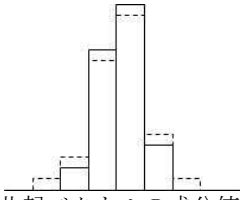

図 1　共起ベクトルの成分値の分布

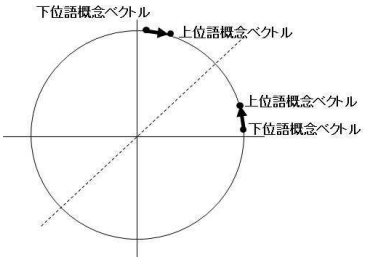

図 2　上位語・下位語の概念ベクトルの関係

を L の第 i 成分値に近い値に写像するような平面である。このような行列 B を、半径 1 の球面上の任意の概念ベクトルに右から乗じて得られるベクトル $\hat{V}$ は、上位概念に相当する概念ベクトルとなる。

これより、行列 XB と行列 Y との自乗距離を、$\sum_{1\le i\le k}\sum_{1\le l\le n}\varepsilon_{il}^{\,2}$ と定義し、該自乗距離が最小となる行列 $\hat{B}$ を求める。該自乗距離が最小となる行列 $\hat{B}$ は、以下の式により算出できる。この $\hat{B}$ は、シソーラス内の下位語の概念ベクトルと、上位語の概念ベクトルとの対応関係を表しており、最小自乗行列と呼ぶ。

$$\hat{B}=\left(X'X\right)^{-1}X'Y$$

任意の語 A に対し、語 A の構成単語の概念ベクトルの和を長さ 1 に正規化したものを、語 A の概念ベクトル H とする。H と最小自乗行列 $\hat{B}$ とを乗じて得られるベクトル $H\hat{B}$ は、語 A の上位概念相当の概念ベクトルである。

概念ベース内の単語 w の概念ベクトルと $H\hat{B}$ との類似度を余弦測度として算出したとき、類似度の高い単語 w は、語 A の上位概念に相当すると考えられるので、単語 w を、語 A の上位語として出力する。

3．評価実験

評価実験では、シソーラスとして日本語 Wordnet[1]を使用した。Wordnet から、synset と、該 synset とのリンク情報が Hypernym（上位語）、Domain Category（被包含領域）、Instances（例）のいずれかである synset を、それぞれ、下位語、上位語として、下位語・上位語対を抽出した。この結果、579,480 個の対が得られた。この対集合の中から、上位語が概念ベース内にある対を 1,000 個、ランダムに（但し、全ての語が異なるように）選び、テストセットとした。テストセットに含まれない対で、上位語がテストセット中の上位語として出現せず、かつ、下位語がテストセット中の下位語として出現しないような対を全部とり、学習セットとした。学習セットは 459,398 個の対からなる。このように構成したのは、テスト用対と上位語（下位語）が同じ学習用対があると、下位語（上位語）の概念ベクトルの類似度が高いだけで、上位推定が有利にできてしまうため、学習による汎化の効果を適切に測れないからである。

学習セットに対して回帰分析を行い最小自乗行列を求めた。各テスト用対に対し、下位語の概念ベクトルと最小自乗行列とを乗じて上位概念相当ベクトルを求め、概念ベース内の単語を、その概念ベクトルと上位概念相当ベクトルとの類似度の降順にランキングし、上位語の順位を求めた。上位語が高順位であるほど、上位概念相当ベクトルは適切であるといえる。比較手法として、各テスト用対に対し、概念ベース内の単語を、その概念ベクトルと下位語の概念ベクトルとの類似度の降順にランキングし、上位語の順位を求めた。各手法の、上位語の順位の逆数の平均（平均逆順位）は、表 1 のようになり、ウィルコクソンの符号付順位和検定により有意水準 1%で逆順位の母平均に有意差が認められた。

表 1　評価結果

	提案手法	比較手法
平均逆順位	0.015650	0.011611

表 2 は、テスト対（下位語:西洋南瓜、上位語:野菜）に対し、提案手法と比較手法それぞれでの、下位語「西洋南瓜」に対する 20 位までの単語ランキング結果である。上位語「野菜」は、提案手法で 3 位、比較手法で 15 位であり、提案手法の方が高順位である。比較手法では「人参」や「牛蒡」等、「西洋南瓜」の兄弟概念にあたる野菜の種類名が多いが、提案手法では「植物」や「作物」等、「西洋南瓜」の上位概念にあたる単語が多い。この例に見られるように高順位の単語群には、比較手法は兄弟概念にあたる単語がくる傾向があるのに対し、提案手法では上位概念にあたる単語がくる傾向がある。

表 2　単語ランキング結果

順位	提案手法	比較手法
1	西洋	西洋
2	栽培	南瓜
3	野菜	カボチャ
4	南瓜	薩摩芋
5	一年草	ジャガいも
6	品種	人参
7	マメ科	裏漉し
8	草本	菠薐草
9	植物	里芋
10	収穫	東洋
11	カボチャ	ポタージュ
12	作物	ブロッコリー
13	草	ビシソワーズ
14	種子	牛蒡
15	は種	野菜
16	園芸	大根
17	植える	インゲン
18	蔓草	パンプキン
19	多年草	カリフラワー
20	雑草	ポトフ

4．おわりに

回帰分析により語の上位語を推定する手法を提案した。だが、提案手法による単語ランキング結果における高順位の単語群には、上位語でないノイズも多い。今後は、アルゴリズムの改良による精錬化を図っていきたい。

文　献

[1] F. Bond et al.,“Japanese SemCor:A Sense-tagged Corpus of Japanese,” Proc. 6th Int. Conf. on GWC-2012.

[2] 別所克人 他,“単語・意味属性間共起に基づくコーパス概念ベースの生成方式,”情処論, vol.49, no.12, pp.3997-4006, Dec.2008.

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.