言語コーパスからの語の共起性の推定

冨浦 洋一*，　田中 省作**，　日高 達*

* 九州大学大学院 システム情報科学研究院 知能システム学部門

** 九州大学情報基盤センター

E-mail : {tom@is,sho@cc,hitaka@is}.kyushu-u.ac.jp

1　はじめに

自然言語文の構文解析では，一般に，一つの入力文に対して文法的には正しい複数の統語構造が得られ，しかもその多くの統語構造は意味的に不自然である．どのようにして，意味的に不自然な統語構造を排除して，可能な統語構造を絞り込むかが構文解析の大きな問題の一つである．意味的に不自然な統語構造を排除する代表的な解決法として，語の共起性（たとえば日本語では，名詞 n が格助詞 c を伴って動詞 v に係り得るか否か）を利用した方法がある．

ところが，共起性を持つ語の組は膨大である．それらを，人間が列挙するのも困難であるし，また，構文解析済の言語コーパスから，『共起が観測された語の組は共起性がある』として，自動的に抽出したとしても，共起性を持つ語の組の極一部しか収集されない．

そこで，本稿では，観測された共起性を持つ語の組を基にして，観測されていない語の組の共起性を判定する手法を提案する．本手法は，係る語をユークリッド空間上の点(ワードベクトル)に対応させ，これを説明変量，係の種類(あるいはこれを規定する格助詞などの機能語)と係られる語の組を目的変量とする重回帰モデルに基づくものであり，回帰式に現れる重みだけではなく，係る語のワードベクトルも同時に学習する手法である．

2　提案手法

2.1　推定モデル

本稿では，語 w が係りの種類（あるいはこれを規定する格助詞などの機能語）f で，語 w' に係り得るとき，$\langle w, f, w' \rangle$ に共起性がある（w と $\langle f, w' \rangle$ に共起性がある）と言うことにする．

係る語を $n-1$ 次元ユークリッド空間上の点（ワードベクトル）に対応させる．もしこの対応が良いものであるならば，$\langle f, w' \rangle$ との共起性は，これを目的変量，ワードベクトルを説明変量とする重回帰モデルによってある程度推定可能であることが期待できる．重回帰モデルが語の共起性推定のモデルとして適切かどうか，議論の余地のあるところであるが，まずは，このモデルにより推定を試みた．

語 w_i のワードベクトルを，$[x_{i,1}\, x_{i,2} \cdots, x_{i,n-1}]$ で表す．係りの種類と受けの語の組に通し番号を付与し，j 番目の組 $\langle$ 係りの種類, 受けの語 $\rangle$ との共起性を目的変量 Y_j で表す．語 w_i の Y_j に対する推定値 $\widehat{y}_{i,j}$ を，

$$\widehat{y}_{i,j} = \sum_{k=1}^{n} x_{i,k}\, a_{k,j} \qquad (1)$$

なる重回帰モデルで推定する．ただし，$a_{k,j}(k = 1, 2, \cdots, n)$ は Y_j に依存した重みで，$x_{i,n}$ は，記述を簡潔にするために導入したものであり，任意の i に対して $x_{i,n} = 1$ である．以降，$\langle w, f, w' \rangle$ の共起性推定において，w の総数を M，$\langle f, w' \rangle$ の組の総数を N とし，X は (i,k) 要素が $x_{i,k}$ である $(M \times n)$ 行列，A は (k,j) 要素が $a_{k,j}$ である $(n \times N)$ 行列であるとする．

2.2 学習

通常の回帰分析では，各 entity の説明変量および目的変量の値が既知の学習データを基に，重回帰モデルの重みを学習する．

しかし，本稿で提案するモデルでは，

- entity（語）の説明変量の値（ワードベクトル）は未知である．

さらに，構文解析済みの言語コーパス上で観測された組〈係る語, 係りの種類, 受けの語〉を取り出して学習データとすると，

- 「共起性がない」という情報は学習データ中に含まれない．つまり，正例のみからの学習である．

そこで，目的関数を，

$$\alpha \sum_{\langle i,j\rangle \notin S} (y_{i,j} - \widehat{y}_{i,j})^2 + \sum_{\langle i,j\rangle \in S} (y_{i,j} - \widehat{y}_{i,j})^2$$

とし，これを最小（極小）にするように X および A を求める．ここで，S は学習データで，共起が観測された $\langle w, f, w'\rangle$ の w の通し番号と $\langle f, w'\rangle$ の通し番号の組の列である（$\langle i,j\rangle$ が S 中に重複して現れることも許す）．$y_{i,j}$ は Y_j の観測値で，

$$y_{i,j} = \begin{cases} 1 & ; \ \langle i,j\rangle \in S \\ 0 & ; \ \langle i,j\rangle \notin S \end{cases}$$

である．α は学習データのサイズに依存する 1 より小さな定数である（学習データサイズ $\to \infty$ のとき，$\alpha \to 0$ である）．有限の学習データに，共起が観測されなかったとしても，それは偶然かも知れず，共起しないとは断定できない．上記の目的関数は，共起が観測されなかった組に対する共起性の推定誤差を共起が観測された組に対する共起性の推定誤差より軽く見よう（α 倍）というものである．

$$\beta_{i,j} = \begin{cases} \langle i,j\rangle \text{ の } S \text{ 中での頻度} & ; \ \langle i,j\rangle \in S \\ \alpha & ; \ \langle i,j\rangle \notin S \end{cases}$$

とすると，目的関数 $F(X,A)$ は

$$F(X,A) = \sum_{i=1}^{M}\sum_{j=1}^{N} \beta_{i,j}(y_{i,j} - \widehat{y}_{i,j})^2$$

と表現できる．

2.3 目的関数を極小にする X, A の求め方

$$\frac{\partial F}{\partial a_{k,j}} = -2\sum_{i=1}^{M} \beta_{i,j} \left\{ y_{i,j} - \sum_{\ell=1}^{n} x_{i,\ell}\, a_{\ell,j} \right\} x_{i,k}$$

ゆえ，$\partial F/\partial a_{k,j} = 0$ より，

$$\sum_{\ell=1}^{n} \left\{ \sum_{i=1}^{M} \beta_{i,j}\, x_{i,k}\, x_{i,\ell} \right\} a_{\ell,j} = \sum_{i=1}^{M} \beta_{i,j}\, y_{i,j}\, x_{i,k}\,.$$

ここで，n 次正方行列 D_j^X，n 次列ベクトル $\boldsymbol{b}_j^X$ を

$$\begin{aligned} \left[D_j^X\right]_{k,\ell} &= \sum_{i=1}^{M} \beta_{i,j}\, x_{i,k}\, x_{i,\ell}\,, \\ \left[\boldsymbol{b}_j^X\right]_k &= \sum_{i=1}^{M} \beta_{i,j}\, y_{i,j}\, x_{i,k} \end{aligned}$$

とおくと，$\partial F/\partial a_{k,j} = 0 \ \ (k = 1, 2, \cdots, n)$ より，連立方程式

$$D_j^X\, \boldsymbol{a}(j) \ = \ \boldsymbol{b}_j^X \qquad (2)$$

が得られる．ただし，$\boldsymbol{a}(j)$ は，A の j 列である．

同様に，$\partial F/\partial x_{i,k} = 0 \ \ (k = 1, 2, \cdots, n-1)$ より，連立方程式

$$D_i^A\ {}^t\boldsymbol{x}(i) \ = \ \boldsymbol{b}_i^A \qquad (3)$$

が得られる．ただし，D_i^A，$\boldsymbol{b}_i^A$ はそれぞれ，$n-1$ 次正方行列，$n-1$ 次列ベクトルで，

$$\begin{aligned} \left[D_i^A\right]_{k,\ell} &= \sum_{j=1}^{N} \beta_{i,j}\, a_{k,j}\, a_{\ell,j}\,, \\ \left[\boldsymbol{b}_i^A\right]_k &= \sum_{i=1}^{M} \beta_{i,j}\, y_{i,j}\, a_{k,j} \end{aligned}$$

であり，$\boldsymbol{x}(i)$ は $[x_{i,1}\, x_{i,2} \cdots x_{i,n-1}]$ なる $n-1$ 次の行ベクトルである（^{t} は転置行列を示す）．

X を任意に固定した場合，$j = 1, 2, \cdots, N$ に対して連立方程式 (2) を解いて得られる解 A は，$F(X,A)$ を最小にする．一方，A を任意に固定した場合，$i = 1, 2, \cdots, M$ に対して連立方程式 (3) を解いて得られる解 X は，$F(X,A)$ を最小にする．したがって，今，$X = X_m$，$A = A_m$ である

とき，X を固定して，A に関して，連立方程式(2)を解いて得られる解を A_{m+1} とし，$A = A_{m+1}$ と固定して，X に関して，連立方程式(3)を解いて得られる解を X_{m+1} とすると，

$$F(X_m, A_m) \geq F(X_{m+1}, A_{m+1})$$

が成立する．このことを利用して，適当な初期 X から出発して，繰り返し計算により，$F(X, A)$ を極小にする X，A を求めることができる．

2.4 共起性の判定

Y_j に対応する $\langle f, w' \rangle$ と w_i との共起性は，(1)式で推定されるが，これは連続値である．共起するか否かという2値の判定を行う場合は，適当な閾値 SH を設定して，

$$\widehat{y}_{i,j} < SH \implies \langle w, f, w' \rangle \text{ は共起性無し}$$
$$\widehat{y}_{i,j} \geq SH \implies \langle w, f, w' \rangle \text{ は共起性有り}$$

と判定する．

3 評価実験

〈名詞,助詞,動詞〉の共起性推定を対象として，提案手法の評価実験を行った．

3.1 学習データ

EDR 日本語コーパス [1] から，これに出現する共起 $\langle n, c, v \rangle$（名詞 n が助詞 c で動詞 v に係る）を抽出し，この3つ組の列 S_0 を作成した（同一の $\langle n, c, v \rangle$ が複数含まれることも許す）．

余りにも出現頻度の低い単語に関しては，本手法で精度良く共起性を推定することはできない．そこで，S_0 の中の低頻度（今回は，5回以下とした）の n, $\langle c, v \rangle$ を含む組を削除して，S_1 を作成した．S_1 中の名詞の異なり数は 4865，助詞・動詞の組の異なり数は 4947 である．

共起性の推定結果を構文解析で用いる場合，共起性を判定しようとする語の組の分布に偏りがあるため，学習データはその偏りが反映できるよう，重複を許すべきである．しかし，今回の評価実験は，ランダムに w と $\langle f, w' \rangle$ を与えたときの共起性判定の評価であるため，学習データには，重複を許すべきではない．そこで，S_1 から重複を取り除いて，学習データ S を作成した．S のサイズ（共起が観測された組の異なり数）は，102901 である．

3.2 評価用データ（正解データ）

S 中の名詞，〈助詞,動詞〉をそれぞれ，ランダムに 70 個選び，これらの全ての組み合わせ 4900 組のうち，S に現れない 4881 組に対して，二人の被験者に共起性を判定してもらった．そのうち，答えが一致するもの 3407 組を評価対象とした．このうち，被験者により共起性があると判定された組は，44.2% であった．

3.3 比較実験

名詞 n_1 と n_2 の意味が類似しているならば，n_1 の $\langle c, v \rangle$ に対する類似性と n_2 の $\langle c, v \rangle$ に対する類似性も類似していると考えられる．そこで，ある適当な μ を設定しておき，共起性が未知の組 $\langle n_1, c, v \rangle$ に対して，

$$\exists n_2 \text{ 類似度 } (n_1, n_2) \geq \mu \ \&\ \langle n_2, c, v \rangle \in S$$
$$\implies \langle n, c, v \rangle \text{ は共起する}$$

と推定することが考えられる．

名詞間の類似度は，EDR 概念体系辞書と日本語単語辞書 [1] から求めた．概念 c_1 と c_2 の類似度は，c_1 と c_2 に共通の最下位の上位概念の深さ (ルートノードからのパス長) が深ければ高く，また，c_1 と c_2 に共通の最下位の上位概念が c_1，c_2 に近い場合にも高いと考えられる．基本的にはこの考えに基づいて名詞間の類似度を定義するのであるが，一般に名詞は複数の意味を持つ (名詞の直接の上位概念は複数) ため，今回の実験では，名詞 n_1 と n_2 の類似度を

$$\max_{\substack{c_1 \in C(n_1) \\ c_2 \in C(n_2)}} \frac{1}{2}\left(\frac{D(c_{12})}{L(c_{12}, c_1)+1} + \frac{D(c_{12})}{L(c_{12}, c_2)+1}\right)$$

と定義した．ここで，$C(n)$ は名詞 n の直接の上位概念の集合，c_{12} は概念 c_1 と c_2 に共通の最下

位の上位概念，$D(c)$ は概念 c のルートノードからの最大の深さ（ルートから c への最長のパス長），$L(c,c')$ は概念 c と c' との最短のパス長である．

3.4 実験結果

$n=7\sim10$，$\alpha=0.008, 0.01$ の各場合に対して，目的関数 F_1 でモデルを学習したときの評価用データに対する正解率（正しく判定した組数/3407），共起性がある組の再現率，適合率を表1に示す（スレッシュホールド SH は正解率が最大になるように設定）．

また，比較実験として行った3.3節で述べた手法での，評価用データに対する正解率は，$\mu=0.57$ のとき，最高で，62.0 % であった．

4 考察

2節で述べた提案手法での正解率は，比較実験として行った3.3節で述べた手法での正解率より高かったが，67 % 前後という正解率は，それほど高いというわけではない．

その最大の原因は，学習データの規模が小さいことにある．学習データの元となる言語コーパス中で出現頻度が低い w や $\langle f,w'\rangle$ に対して，本手法で，$\langle w,f,w'\rangle$ の共起性を推定する場合，勿論高精度は期待できない．逆に，言語コーパス中での出現頻度が高い，w や $\langle f,w'\rangle$ に関しては，高い正解率が得られることが期待できる．そこで，評価用データの $\langle n,c,v\rangle$ のうちで，S_1 中での出現頻度が γ（$\gamma=10,20,30,40,50$）以上の組についてのみ，$n=8,\alpha=0.01$ の場合の正解率，再現率，適合率を求めてみた（表2）．これから分かるように，高頻度の $n,\langle c,v\rangle$ に対する共起性の判定精度はかなり高く，大規模な学習データさえ作成することができれば，提案手法はかなり有効な手法だと言える．

また，別の原因として，X の初期値に対する工夫がなされていないこともあるかも知れない．今回の実験では X の初期値をランダムに設定したが，シソーラス上の類似度を反映するように初期ワードベクトルを設定することで，よりよい X，A を推定できるかも知れない．これは，今後の検討課題である．

n	α	正解率	再現率	適合率
7	0.008	67.3	55.5	65.3
7	0.010	67.2	52.3	66.3
8	0.008	66.7	57.5	63.5
8	0.010	66.8	49.8	66.6
9	0.008	66.6	57.3	63.5
9	0.010	66.5	57.1	63.3

表1 F_1 により学習したモデルでの共起性推定結果

γ	データ数	共起割合	正解率	再現率	適合率
10	2209	44.0	68.6	56.3	67.1
20	618	46.9	69.3	69.0	66.7
30	223	53.8	71.3	75.8	72.2
40	90	55.6	80.0	92.0	76.7
50	67	56.7	82.1	92.1	79.5

表2 出現頻度で制限した評価用データでの共起性推定結果
データ数：評価データの数
共起割合：評価データの内の共起性を有する組の割合

5 おわりに

観測された共起性を持つ語の組を基にして，観測されていない語の組の共起性を重回帰モデルに基づき推定する手法を提案した．また，共起性が未知の3407組の〈名詞,助詞,動詞〉に対して行った評価実験では，67 % 前後の正解率であったが，評価用データを高頻度の名詞，〈助詞,動詞〉に限った場合の正解率は82% であった．このことから，観測された共起の組を蓄積してゆくことにより，提案手法は共起性推定のための現実的な手法となり得ると考えている．

なお，本研究の一部は，科研費基盤研究(C)，大川情報通信基金研究助成により行なった．

参考文献

[1] 日本電子化辞書研究所，EDR 電子化辞書仕様説明書，1995