言語コーパスからの語の共起性の推定と統語的曖昧さ解消実験による評価

冨浦 洋一，　日高 達

九州大学大学院 システム情報科学研究院

E-mail : {tom,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．この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，あるいは，その係り易さの程度）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，著者らは観測された共起情報を基にして，重回帰モデルにより共起性を推定する手法を提案している [1]．重回帰モデルに適用するためには，語をユークリッド空間上の点 (ワードベクトル) に対応させる必要があるが，[1] の手法は，モデルのパラメタ (重回帰式の重み係数) だけでなく，ワードベクトルも同時に学習するのが特徴である．この推定手法を評価する目的で，[1] に基づいて得られた名詞と〈助詞，動詞〉との共起性を用いて，名詞句の係り先の曖昧さ解消実験を行なった．本発表では，提案手法の概説，評価実験とその結果について報告する．

2　共起性推定手法

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．

記述の簡単のために，$\langle f, w' \rangle$ に通し番号を付与し，$\langle f, w' \rangle$ の全体集合を，$\{g_1, g_2, \cdots, g_N\}$ とする．また，w の全体集合を $\{w_1, w_2, \cdots, w_M\}$ とする．w_i と g_j の共起性 $C_{i,j}$ が重回帰モデルで，

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

と推定できると仮定する．ただし，$a_{k,j}(k = 1, 2, \cdots, n)$ は g_j に依存した重みで，$[x_{i,1}, x_{i,2}, \cdots, x_{i,n-1}]$ は w_i に依存したベクトル（ワードベクトル）である．$x_{i,n}$ 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（つまり $a_{n,j}$ はバイアス項）．

2.2　学習

当然のことながら，推定式 (1) で用いられるワードベクトルはどのようなものでも良いと言うわけではない．たとえば，シソーラス上での単語間の類似度を反映するようにして求めたワードベクトルを適用しても，(1) 式のような単純なモデルで共起性を推定することはおそらく出来ない．共起性推定という問題に特化したワードベクトルを用いる必要がある．つまり，重回帰モデルのパラメタである $(n \times N)$ 行列 A（A の (k,j) 要素は $a_{k,j}$）だけでなく，$(M \times (n-1))$ 行列 X（X の (i,k) 要素は $x_{i,k}$）も推定する必要がある．

共起が観測された $\langle w, g \rangle$ は共起性が高く，観測されなかった $\langle w, g \rangle$ は共起性が低いと考えられる．そこで，共起が観測された組の共起性は 1，観測されなかった組の共起性は 0 と割りきり，目的関数

$$\sum_{\langle i,j \rangle \in S} (1 - \widehat{C}_{i,j})^2 + \alpha \sum_{\langle i,j \rangle \notin S} (0 - \widehat{C}_{i,j})^2 \qquad (2)$$

を最小にするように 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 $\langle w, g \rangle$ の w の通し番号と g 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）．α は学習データのサイズ

に依存する 1 以下の正の定数で，

$$\text{学習データサイズ} \to \infty \quad \text{のとき} \quad \alpha \to 1$$

となるように定める．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（α 倍）軽く見ようというものである．

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

とすると，目的関数 $F(X,A)$ は

$$F(X,A) = \sum_{i=1}^{M} \sum_{j=1}^{N} \beta_{i,j} (C_{i,j} - \widehat{C}_{i,j})^2 \qquad (3)$$

と表現できる．ただし，

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

である．

2.3 解法

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

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

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

$$D_j(X) \, \boldsymbol{a}(j) \ = \ \boldsymbol{b}_j(X) \qquad (4)$$

が得られる．ただし，$\boldsymbol{a}(j)$ は，A の j 列である．また，$n-1$ 次正方行列 $D_i(A)$，$n-1$ 次列ベクトル $\boldsymbol{b}_i(A)$ を

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

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

$$D_i(A) \ {}^t\boldsymbol{x}(i) \ = \ \boldsymbol{b}_i(A) \qquad (5)$$

が得られる．ただし，$\boldsymbol{x}(i) = [x_{i,1}, x_{i,2}, \cdots, x_{i,n-1}]$ である（${}^t\boldsymbol{x}$ は $\boldsymbol{x}$ の転置を示す）．

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 (4) 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 (5) 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 X の初期値 X_0 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．

3 評価実験

前節の手法により推定した名詞と〈助詞, 動詞〉との共起性を評価する目的で，得られた共起性を利用して，

$$n \quad c \quad \gamma \quad v_1 \quad \delta \quad v_2 \qquad (6)$$

という形態の文に対する「$n\,c$」の係り先の判定実験（v_1 に係るか v_2 に係るか）を行なった．ただし，n は名詞，c は助詞，v_1，v_2 は動詞であり，γ，δ は単語列である．n を主辞とする後置詞句に格助詞と副助詞がともに含まれる場合は格助詞を c とする．また，v_1 は δ 中の名詞を修飾し，γ 中の単語の係り先は v_1 より後方にはないものとする．たとえば，

$$\underset{n}{\underline{\text{メーカー}}} \ \underset{c}{\underline{\text{が}}} \ \text{プラスチック製の危険物を}$$
$$\underset{v_1}{\underline{\text{探知する}}} \ \text{X線を} \ \underset{v_2}{\underline{\text{売り出す}}}$$

のような文である．この形態の文の場合，「$n\,c$」は文法的には v_1 にも v_2 にも係る可能性がある．

係り先を決める要因は色々と報告されている [2]．たとえば，

(a) n を主辞とする後置詞句に「は」が含まれるか，

(b) n の次の自立語が v_1 か，

という要因を考えることができる．「は」を含む後置詞句は v_2 に係る傾向が非常に強い．また，n の次の自立語が v_1 である場合，「$n\,c$」が v_2 に係るとすると，δ 中の名詞を修飾する連体修飾文が v_1 単独となり，連体修飾文としては意味的に不適切な場合が多いため，v_1 に係る傾向が非常に強い．たとえば，「熊が出没する山に入る」という文で，「熊が」が「入る」に係るとすると，「出没する山」では意味的に不自然である（「出没する」

は「山」の連体修飾文としては,「山」を十分に修飾限定していない)．このような傾向は，3.5節の評価用データの内訳にもよく表れている．そこで，今回の評価では，n を主辞とする後置詞句に「は」が含まれず，n の次の自立語が v_1 でない場合について，本手法で推定した共起性が係り受けの曖昧さ解消にどの程度有効かで，得られた共起性の評価を行なった．

3.1 推定した共起性を用いた係り先判定法

本手法で推定された共起性を利用して，(6) の形態の文に対する「$n\,c$」の係り先を以下のように判定する．

$$\begin{array}{ll} \theta \cdot C(n,c,v_1) < C(n,c,v_2) & \Longrightarrow \quad v_2\text{に係る} \\ \text{その他} & \Longrightarrow \quad v_1\text{に係る} \end{array}$$

$C(n,c,v)$ は推定された $\langle n,c,v\rangle$ の共起性である．ただし，今回の手法では，推定された共起性が $[0,1]$ 上の値とは限らないので，最大値が1，最小値が0となるようにスケール変換したものを用いた．

3.2 シソーラスを用いた手法（比較手法1）

多くの研究者が用いている手法として，シソーラス上の名詞間の類似度を利用した類似用例に基づく手法がある．これを比較手法として上げる．

名詞 n と n' の意味が類似していて，n' と $\langle c,v\rangle$ の共起が観測されているならば，n と $\langle c,v\rangle$ の共起性も比較的高いと考えられる．そこで，3.1節での $C(n,c,v)$ の代わりに，共起性として以下を用いる．

> S 中の $\langle n',\langle c,v\rangle\rangle$ で（重複も許す），n と
> の類似度 $sim(n,n')$ の k 番目に大きな値．

k は適当な定数（k-Nearest Neighbor 法の k）である．

名詞間の類似度 sim は，シソーラスとしてEDR概念体系辞書と日本語単語辞書 [3] を用いて，以下のようにして求めた．

$$sim(n_1,n_2) = \max_{c\in CM(n_1,n_2)} \frac{D_r(c\,;\,n_1)+D_r(c\,;\,n_2)}{2}$$
$$\left(D_r(c\,;\,n) = \frac{L(R,c)}{L(R,c)+L(c,n)}\right)$$

と定義した．ここで，$CM(n_1,n_2)$ は名詞 n_1 と n_2 の共通の上位概念（直接の上位概念以外も含む）の集合，$L(a,b)$ は a と b の間の最短パス長，R はシソーラスのルートノードである．つまり，$sim(n_1,n_2)$ は二つの名詞の共通の上位概念 c の n_1 に対する相対的な深さ $D_r(c\,;\,n_1)$ と n_2 に対する相対的な深さ $D_r(c\,;\,n_2)$ の平均の最大値である．

3.3 人間の内省による共起性の強弱に基づく手法（比較手法2）

共起性に基づく係り先の判定手法がどのくらい有効であるかの目安として，人間の内省により共起性の強弱を求め，これを利用して係り先の判定を行なう実験も行なった．

$\langle n,c,v_1\rangle$ と $\langle n,c,v_2\rangle$ を順不同で提示して，4人の被験者に

$$\begin{cases} C(n,c,v_1) \gg C(n,c,v_2) \\ C(n,c,v_1) \ll C(n,c,v_2) \\ \text{あまりかわらない} \end{cases}$$

の3つに判別させた．これを利用して以下のように判定する．

$$\begin{array}{ll} C(n,c,v_1) \ll C(n,c,v_2) & \Longrightarrow \quad v_2\text{に係る} \\ \text{その他} & \Longrightarrow \quad v_1\text{に係る} \end{array}$$

3.4 共起性の学習データ

EDR 日本語コーパス [3] から，これに出現する共起 $\langle n,\langle c,v\rangle\rangle$（名詞 n が助詞 c で動詞 v に係る）を抽出し，S 中の n および $\langle c,v\rangle$ がともに頻度2以上になるように低頻度の共起を削除して，共起の列 S を作成し（同一の $\langle n,\langle c,v\rangle\rangle$ が複数含まれることも許す），名詞と〈助詞,動詞〉との共起性の学習データとした．ただし，実際には S は，名詞，〈助詞,動詞〉それぞれに付与した通し番号の組の列である．データの規模は，$|S| = 213,663$，名詞異なり数 16,543，助詞・動詞異なり数 14,474 である．

3.5 評価用データ

EDR 日本語コーパスから (6) の形態の文のうち，n, $\langle c,v_1\rangle$, $\langle c,v_2\rangle$ が全て S に含まれているものに対して

$$\langle n,c,v_1,v_2,h,d,ans\rangle$$

を抽出した（2488組）．h は3節冒頭で述べた (a) に関する情報，h は (b) に関する情報である．また，ans は「$n\,c$」の係り先である．

抽出したデータの内訳を下記に示す．このうち，上記の type 1 について曖昧さ解消実験を行なった．

type	h	d	総数	係り先	
				v_1	v_2
1	no	no	764	474 (62.0%)	290 (38.0%)
2	no	yes	1341	1285 (95.8%)	56 (4.2%)
3	yes	no	342	23 (6.7%)	319 (93.3%)
4	yes	yes	41	5 (12.2%)	36 (87.8%)
合計			2488	1787 (71.8%)	701 (28.2)

3.6 実験結果

(1) 式の n として，6, 7, 8, 9, 10, 11，(2) 式の α として，0.006, 0.007, 0.008, 0.009, 0.010, 0.020 の合計 36 通りで，共起性の推定実験を行なった．ただし，2.3 節で述べた手法は目的関数を極小にする X，A を求める手法であり，これは X の初期値に依存するため，X の初期値としてランダムに 10 通りを試し，同一の n，α で目的関数値が最小となる X，A を学習結果とした．

また，3.1 節および 3.2 節の手法での曖昧さ解消実験では，本来はオープンテストとすべきであるが，パラメタ（スレッシュホールド）1 つに対して数百のデータがあるため，クローズドテストでの結果とほぼ同じと期待できる．そこで，クローズドテストによる評価（つまり，スレッシュホールドを最適に調節した場合の評価用データに対する正解率）とした．

以下に結果を示す．

	正解率 (%)					
	提案手法	比較手法 1	比較手法 2			
			1	2	3	4
type1	69.6	67.9	68.5	65.6	66.5	66.4
全体	87.3	86.8	86.9	86.1	86.3	86.3

提案手法は，$n = 10$，$\alpha = 0.007$ として共起性を推定したの場合の結果であり，これが試した n，α の中では最も良かった．また，比較手法 1 は最も結果の良かった $k = 2$ の場合を示している．比較手法 2 の 1〜4 は被験者番号である．参考までに，type2 の場合 v_1 に係り，type3 および 4 の場合 v_2 に係るとして，type1 のみそれぞれの手法で判定した場合の全体の正解率を，上記表の最後の行に上げた．

4 おわりに

観測された語の共起情報を基にして，重回帰モデルにより共起性を推定し，得られた共起性を用いて，名詞句の係り先の曖昧さ解消実験を行なった．シソーラスから得られる語の類似性を利用した共起性に基づく手法（比較手法 1），人の内省による共起性の強弱に基づく手法（比較手法 2）と比較してもほぼ同性能であった．若干提案手法によって推定された共起性に基づく手法の方が高いが有意な差ではない．しかし，比較手法 1，つまり，類似用例に基づく手法では，蓄積している用例（共起データ）の増加に伴い，記憶容量も計算コストも増す．一方，提案した重回帰モデルによる共起性の推定では，共起データが増加しても記憶容量，計算コストは一定であり，この意味では有効と言える．

今回の推定では，共起が観測されたものの推定値が 1 より大きくても誤差としている．同様に，共起が観測されなかったものの推定値が 0 より小さくても誤差としている．本来は，共起が観測されたものは共起性が高く，観測されなかったものは低いとすべきである．計算の簡便さから今回のようにしたが，たとえば，$x_{i,n}$ も推定すべきパラメタとして $\boldsymbol{x}(i)$ の成分に含め，

$$\begin{cases} |\boldsymbol{x}(i)| \leq 1 & ; \quad i = 1, 2, \cdots, M \\ |\boldsymbol{a}(j)| \leq 1 & ; \quad i = 1, 2, \cdots, N \end{cases}$$

なる制限下（推定される共起性は最大で 1，最小で -1 となる）で (3) 式（ただし，$\langle i, j \rangle \notin S$ の場合 $C_{i,j} = -1$）を最小にするように X，A を求めたり，あるいは，(1) 式に変えて，推定モデルを

$$\widehat{C}_{i,j} = \left(1 + \exp\left\{-\sum_{k=1}^{n} x_{i,k}\, a_{k,j}\right\}\right)^{-1}$$

として，(2) 式を最小にするように X，A を求めたりすることも考えられる．

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] 冨浦，田中，日高：言語コーパスからの語の共起性の推定，言語処理学会第 8 回年次大会，p.631-634 (2002)

[2] 内元，関根，井佐原：最大エントロピー法に基づくモデルを用いた日本語係り受け解析，情報処理学会論文誌，Vol.40, No.9, p.3397-3407 (1999)

[3] EDR 電子化辞書 日本語コーパス (JCO-V020E)，概念辞書 (CPD-V020.1)，日本語単語辞書 (JWD-V020)