言語コーパスからの語の共起性の推定

冨浦洋一[†] 日高達[††]

語の共起性は自然言語処理における基本的な知識の１つであり，これを利用して，自然言語文の統語的曖昧さや多義語の語義の曖昧さを解消することができる．本論文では，構文解析済みの言語コーパスから得られる共起データを基にして，語の共起性を推定する手法を提案する．係る語を実ベクトル（ワードベクトル）に対応させ，これを説明変量とする重回帰モデルにより語の共起性を推定する．通常の重回帰分析と異なり，説明変量であるワードベクトルも同時に学習することが本手法の特徴である．

Estimation of Words' Cooccurrency from Corpus

Yoichi Tomiura[†] and Toru Hitaka[††]

Words' Cooccurrency is one of the basic knowledge in Natural Language Processing, and it is used for syntactic disambiguation and word sense disambiguation. This paper proposes a new method for estimating words' cooccurrency with a syntactically analized corpus based on the multiple regression model. Independent variables of this model correspond to a satellite word (an independent word). Unlike the ordinary multiple regression analysis, the independent variables are also parameters of this model.

1. はじめに

自然言語文の解析では，様々な曖昧さが存在し，その曖昧さを解消する方法が多数提案されている．そのうち，統語構造の曖昧さ解消法および多義語の語義の曖昧さ解消法として，共起性[☆]を利用する手法がある．たとえば,「コスモスが咲いた丘に登る」という文において，文法的には,『コスモス』は『が』が規定する関係で,『咲く』にも『登る』にも係りうる．もし，共起性として，

- 〈『コスモス』,『が』,『咲く』〉 は共起性がある（『コスモス』は『が』が規定する関係で『咲く』に係りうる），
- 〈『コスモス』,『が』,『登る』〉 は共起性がない（『コスモス』は『が』が規定する関係で『登る』に係りえない），

という知識があるならば，これを利用して統語的曖昧さを解消することができる．

これまでは，意味素性と呼ばれる粗い意味分類やシソーラス上の概念による語の分類を用いて表現した共起性が広く利用されていた．しかし，曖昧さ解消の精度を上げようとすると，分類そのものを相当に細かなものにする必要があり，そのような細かな分類を用いて共起性を網羅的に人間が判断・記述するのは，かなりの労力を必要とする．一方，構文解析済み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，共起性がある語の組のごく一部しか収集されず，何らかの方法で，観測されていない語の組の共起性を推定する必要がある．

本論文では，観測された共起データを基にして，共起性を推定する手法を提案する．係る語を実ベクトル（ワードベクトル）に対応させ，これを説明変量とした重回帰モデルで共起性を表現する．モデルの学習において，重回帰モデル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本手法の特徴である．4.2 節で述べるように，重回帰モデルという単純なモ

† 九州大学大学院システム情報科学研究院
Graduate School of Information Science and Electrical Engineering, Kyushu University
†† 九州大学大学院システム情報科学研究院（平成 15 年 3 月退官）
Graduate School of Information Science and Electrical Engineering, Kyushu University (retired March 2003)

☆ 本論文では，語 w_1 が関係（あるいはそれを規定する前置詞や助詞などの機能語）f で語 w_2 に係りうるとき，$\langle w_1, f, w_2 \rangle$ に共起性があるといい，その係りやすさの程度を $\langle w_1, f, w_2 \rangle$ の共起性という．$\langle w_1, f, w_2 \rangle$ は $\langle w_1, \langle f, w_2 \rangle \rangle$ と記すこともある．以降，語の共起性を単に共起性と呼ぶ．

デルであるにもかかわらず，共起データ（学習データ）には含まれないが共起しうる語の組の共起性を上手く選択的に高い値として推定できる．

2. 共起性推定手法

2.1 モデル

本節では，記述の簡単のために，係りの種類（あるいはこれを規定する機能語）f と係られる語 w' の組 $\langle f, w' \rangle$ に通し番号を付与し，$\langle f, w' \rangle$ の全体集合 G を，$\{g_1, g_2, \cdots, g_N\}$ とする．また，係りの語 w の全体集合 W を $\{w_1, w_2, \cdots, w_M\}$ とする．ある適切な次元 K に対して，w_i と g_j の共起性 $C_{i,j}$ が重回帰モデルで，

$$C_{i,j} = \sum_{k=1}^{K} x_{i,k}\, a_{k,j} \qquad (1)$$

と表現できると仮定する．$[a_{1,j}, a_{2,j}, \cdots, a_{K,j}]$ は g_j に対応した重みベクトルで，$[x_{i,1}, x_{i,2}, \cdots, x_{i,K-1}]$ は w_i に対応したベクトル（ワードベクトル）である．$x_{i,K}$ 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 $x_{i,K} = 1$ である（つまり $a_{K,j}$ はバイアス項である）．

2.2 学習

当然のことながら，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よいというわけではない．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，式 (1) のような単純なモデルで精度良く共起性を表現することはおそらくできない．共起性記述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ータである $(K \times N)$ 行列 A（A の (k, j) 要素は $a_{k,j}$）だけでなく，$(M \times (K-1))$ 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，共起性の推定値にそのような傾向が表れるように，目的関数を

$$\sum_{\langle i,j \rangle \in S} (1 - C_{i,j})^2 + \sum_{\langle i,j \rangle \notin S} \alpha_{i,j} (0 - C_{i,j})^2 \quad (2)$$

と設定し，これを最小にするように X および A を求める（ただし，実際には式 (2) を極小にする X，A を求めることになる）．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）．S 中では観測されなかった $\langle i, j \rangle$ は重み付きの疑似的な負例（共起性 $C_{i,j} = 0$）と見なす．$\alpha_{i,j}$ はその重み，つまり，その負例に対する疑似的な頻度である．$\alpha_{i,j}$ をどのように設定するかについては，2.4 節で述べる．

目的関数 $F(X, A)$ は

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

と表現できる．ただし，

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

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

である．

目的関数を最小にするように学習することで，得られるワードベクトルおよび重みベクトルは次のような性質を持つと考えられる．

- S において，共起に関する類似した性質を持つ w_i と $w_{i'}$（どのような g が w_i と共起するかの傾向が，どのような g が $w_{i'}$ と共起するかの傾向と類似している）に対応するワードベクトルは類似したものとなる．
- S において，共起に関する類似した性質を持つ g_j と $g_{j'}$（どのような w が g_j と共起するかの傾向が，どのような w が $g_{j'}$ と共起するかの傾向と類似している）に対応する重みベクトルは類似したものとなる．

2.3 目的関数を極小にする X，A の求め方

X を任意に固定した場合に，$F(X, A)$ を最小にする A を求めてみよう．これは通常の重回帰分析と同様に，以下のようにして求めることができる．K 次正方行列 $D_j(X)$，K 次列ベクトル $\boldsymbol{b}_j(X)$ を

$$[D_j(X)]_{k,\ell} = \sum_{i=1}^{M} \beta_{i,j}\, x_{i,k}\, x_{i,\ell}\,,$$

$$[\boldsymbol{b}_j(X)]_k = \sum_{i=1}^{M} \beta_{i,j}\, x_{i,k}\, T_{i,j}$$

とおくと，$\partial F / \partial a_{k,j} = 0$ $(k = 1, 2, \cdots, K)$ 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．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 (4) を解いて得られる解 A は，$F(X, A)$ を最小にする．

次に，A を任意に固定した場合に，$F(X, A)$ を最

小にする X を求めてみよう．これも同様に，以下のようにして求めることができる．$K-1$ 次正方行列 $D_i(A)$，$K-1$ 次列ベクトル $\boldsymbol{b}_i(A)$ を

$$[D_i(A)]_{k,\ell} = \sum_{j=1}^{N} \beta_{i,j}\, a_{k,j}\, a_{\ell,j}\ ,$$

$$[\boldsymbol{b}_i(A)]_k \quad = \sum_{j=1}^{N} \beta_{i,j}\, a_{k,j}\, (T_{i,j} - a_{K,j})$$

とおくと，$\partial F/\partial x_{i,k} = 0$ $(k = 1, 2, \cdots, K-1)$ 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，$\boldsymbol{x}(i) = [x_{i,1}, x_{i,2}, \cdots, x_{i,K-1}]$ である（${}^t\boldsymbol{x}$ は $\boldsymbol{x}$ の転置を示す）．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 (5) 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 (4) を解いて得られる解を A_{m+1} とし，次に，$A = A_{m+1}$ 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，適当な初期 X から出発して，繰返し計算により，$F(X, A)$ を極小にする (X, A) を求めることができる．

2.4 負例の重みの設定法

疑似的に作成する負例の重み（頻度）$\alpha_{i,j}$ の設定法について述べる．ただし，ここで述べる設定法は考え方の例であり，改良の余地があるものである．

まず，正例列（つまり S）のサイズと $\alpha_{i,j}$ の関係について述べる．S のサイズ $|S|$ と独立に $\alpha_{i,j}$ が定まるとすると，$|S|$ が大きくなるに従い，推定される共起性の平均値は高くなってしまう．共起性の平均値は $|S|$ の大きさと独立であるべきであるから，以下のように負例の総量と正例列のサイズ $|S|$ の割合は一定でなければならない．

$$\sum_{\langle i,j\rangle \notin S} \alpha_{i,j} = \lambda \cdot |S|. \qquad (6)$$

つまり，

$$\alpha_{i,j} = \mu \cdot \alpha^*_{i,j} \quad ; \quad \mu = \frac{\lambda \cdot |S|}{\sum_{\langle i,j\rangle \notin S} \alpha^*_{i,j}} \qquad (7)$$

である．$\alpha^*_{i,j}$ は i，j にのみ依存するパラメータで，これを，式 (6) を満たすように正規化したものが $\alpha_{i,j}$ である．λ は，たとえば，S における w（$\in W$）および g（$\in G$）の発生確率に基づいて，$\langle w, g\rangle$ をランダムに発生させ，これを人間の内省に基づいて，共起性があるもの（正例）と共起性がないもの（負例）に振り分け，その比（負例数/正例数）と設定する．

$\alpha^*_{i,j}$ を i，j によらず一定とするのも 1 つの方法である．また，w_i も g_j も S 中で多く発生しているにもかかわらず，共起 $\langle w_i, g_j\rangle$ が S 中で観測されないならば，それだけ負例としての信頼度が高いと考え，$\alpha^*_{i,j}$ を

$$\frac{f(w_i; S)}{M} \cdot \frac{f(g_j; S)}{N} \quad (= P(w_i)P(g_j))$$

の増加関数とするのも 1 つの方法である（ただし，$f(w; S)$，$f(g; S)$ はそれぞれ，S における w，g の頻度である）．たとえば，$\langle i, j\rangle \notin S$ に対し，

$$\alpha^*_{i,j} = \begin{cases} -I(w_i, g_j) & ; \quad I(w_i, g_j) < 0 \\ 0 & ; \quad I(w_i, g_j) \geq 0 \end{cases} \qquad (8)$$

とすることが考えられる．ここで $I(w_i, g_j)$ は，共起 $\langle w_i, g_j\rangle$ に対する相互情報量[3]，

$$I(w_i, g_j) = \log \frac{P(w_i, g_j)}{P(w_i)P(g_j)}$$

である．$\langle w_i, g_j\rangle$ の共起性が低いほど $I(w_i, g_j)$ は絶対値の大きな負の数となる．上記の定義は，これを利用して，負例の信頼度が $-I(w_i, g_j)$ に比例すると考えたものである．ただし，負例は S 中には発生していないため，$P(w_i, g_j)$ を単純に最尤推定で求めると 0 となってしまう．そこで，Good-Turing 推定法[6]に基づいて，$P(w_i, g_j)$ を

$$\frac{[S \text{ で頻度 } 1 \text{ の共起の総数}]}{[S \text{ で頻度 } 0 \text{ の共起の総数}] \cdot |S|}$$

と見積もる．このように設定した $\alpha^*_{i,j}$ は $P(w_i)P(g_j)$ の増加関数になっていることが分かる．

また，$|S|$ に見合うだけの大きさの，確実な負例集合 SN をなんらかの方法で獲得できる（たとえば，人間の内省に基づき，共起性がきわめて低いものを集める）ならば，式 (8) に代えて，

$$\alpha^*_{i,j} = \begin{cases} 1 & ; \quad \langle w_i, g_j\rangle \in SN \\ -I(w_i, g_j) & ; \quad \langle w_i, g_j\rangle \notin SN \text{ かつ } I(w_i, g_j) < 0 \\ 0 & ; \quad \text{その他} \end{cases}$$

とすることも考えられる．

3. 関 連 研 究

提案手法は，観測された共起から未観測の共起（語の組）に対する共起性を推定するという，一種のスムージング手法ととらえることもでき，共起確率 $P(w, g)$ あるいは条件付き確率 $P(w|g)$ の推定におけるスムージングに関する先行研究と関連が深い．

単語 n グラムモデルにおけるスムージングである線形補間法やバックオフスムージング[6]を利用して，未観測の共起 $\langle w, g\rangle$ に対する $P(w|g)$ を求めることができる．しかし，これらのスムージングは，語の共起性に関する類似性を考慮したものではない．一方提案手法では，学習データにおいて，共起性に関する類似した性質を持つ w と w^* に対応するワードベクトルは類似したものになり，また，共起に関する類似した性質を持つ g と g^* に対応する重みベクトルは類似したものになる．その結果，$\langle w, g\rangle$ の共起性と $\langle w^*, g^*\rangle$ の共起性は類似する傾向にある．つまり，提案手法は，共起データ（学習データ）における語の共起性に関する類似性を考慮した推定法である．

語の類似性を反映したスムージングとして，語の意味クラスを利用することが考えられる．文献 10) では，名詞の意味クラスとして名詞シソーラスを用いたスムージングにより，w_i が名詞，g_j が助詞・動詞の場合の $P(w_i|g_j)$ を推定している．確かに，シソーラスは汎用的な意味分類であり，同じ分類（クラス）に属する語は共起性に関してある程度類似した性質を持つと考えられる．しかし，シソーラスはある一面的な分類であり，しかも，

- どのような概念を設定し，
- 概念をどのように配置（階層付け）するか，

はシソーラス作成者の主観に基づいている．したがって，シソーラスで得られる意味クラスはスムージングに利用するものとして最適とはいえない．文献 9) では，共起データから “soft clustring” と呼ばれる手法で語の意味クラスを求めている．

語の意味クラスを用いるのではなく，直接語の類似度を利用して，語の類似性を反映した共起性のスムージングを行う手法も考えられる．共起データ（学習データ）S 中の共起のうち，g との共起を重複を許して列挙したものを

$$\langle w_{t_1}, g\rangle, \langle w_{t_2}, g\rangle, \cdots, \langle w_{t_u}, g\rangle$$

とし，上記の列に対応して，w との類似度の列

$$sim(w, w_{t_1}), sim(w, w_{t_2}), \cdots, sim(w, w_{t_u})$$

を求める（$sim(w, w^*)$ は名詞 w と w^* の類似度）と，この列中の値のうち k 番目に大きな値と $\langle w, g\rangle$ の共起性の高さには強い正の相関があると考えられる（k は共起データの規模に依存する定数である）．そこで，この k 番目に大きな類似度を $\langle w, g\rangle$ の共起性と定義する．1 番目ではなく，k 番目とするのは，w と非常に類似する語 w^* と g との共起が 1 つだけ S 中で観測され，g との共起が観測されている他の語はどれも w との類似度が低いという場合に $\langle w, g\rangle$ の共起性が高い値として推定されることを防ぐためである．4.3 節では，シソーラス上の名詞間の類似度を用いて，上述の手法により共起性を推定する手法（本論文では，これをシソーラスを用いた用例主導の推定法と呼ぶ）を比較手法の 1 つとして取り上げる．文献 11) では，シソーラス上の単語間の距離を用いて，英語前置詞句の係り先の曖昧さ解消を用例主導の手法で行っているが，基本的には上記の考えに基づいて推定される共起性の大小で係り先を判定しているものととらえることができる（ただし，文献 11) では $k = 1$）．単語間の類似度（あるいは距離）として，シソーラスを用いる場合には，先の研究事例と同様の問題がある☆．観測された共起情報を用いて類似度を求め，これを反映するように共起性を推定する研究もされている．文献 2) では，w と w^* の類似性の尺度として，confusion probability $P_C(w|w^*)$ を用い☆☆，$\langle w, f, w'\rangle$ の共起頻度を $P_C(w|w^*)$ を重みとした $\langle w^*, f, w'\rangle$ の共起頻度の平均として求めている．また，文献 1) では，confusion probability のほか，共起情報から得られるいくつかの（非）類似度を用いた $P(w|g)$ のスムージングの比較検討を行っている．共起情報から得られる語の類似性尺度を用いている点は，提案手法におけるワードベクトルの考え方と通じるところもあるが，観測/未観測の語の組の共起性をより良く表現するようにモデルのパラメータを求めるという提案手法とは異なる手法である．なお，共起情報に基づいた単語間の類似度として，ほかにも，w，w^* と共通に共起する語との共起の相互情報量に基づいた類似度[3]，w，w^* と共通に共起する語との共起確率の和に基づいた類似度[7]など数多くの研究がある☆☆☆．

提案手法と関連の深い手法に PLSI[4]がある．PLSI は，情報検索の分野において，文書 $d(\in D)$ に語 $w(\in W)$ が発生する確率 $P(d, w)$ を推定する手法であ

☆ 提案手法は，未知の共起についてもその共起性を推定するものであるが，もちろん，語そのものが未知の場合（つまり学習データに現れない場合）には対処できない．一方，シソーラスを用いて共起性のスムージングを行う手法では，g さえ学習データに含まれていれば，シソーラスに登録されている学習データ中にない語 w に対しても，$\langle w, g\rangle$ の共起性が推定可能である．この点ではシソーラスを用いたスムージング手法の方が優位である．

☆☆ 実際には，$\sum_w P_C(w|w^*) = 1$ となるように正規化したものを使用．

☆☆☆ 文献 7) では，本論文で扱っている係り受け関係にある共起だけでなく，係り受け関係にはない単なる共起も利用し，前者の共起の影響を重み付けにより大きくすることで，より質の高い類似度を定義している．また，共起情報に基づく類似度の方が，シソーラスに基づく類似度よりも，類義語対/非類義語対の分離精度が高いことを実験により示している．この結果は本論文の主張とも一致しており興味深い．

る．これは，隠れクラス変数 $z \in Z = \{z_1, z_2, \cdots, z_K\}$ を考え（$|Z| \ll |D|, |W|$），文書 d における語 w の発生は，

$$P(d, w) = \sum_{z \in Z} P(d|z)P(w|z)P(z)$$

に従って起こると考えるものである．$P(d, w)$ を与えるためのパラメータ $P(d|z)$，$P(w|z)$，$P(z)$ は，次の対数尤度 L を目的関数とし，これを最大（実際には極大）にするように，適当な初期パラメータ値から始めて，EM アルゴリズムによる繰返し計算で推定する．

$$L = \sum_{d \in D} \sum_{w \in W} f(d, w) \log P(d, w) \qquad (9)$$

$f(d, w)$ は文書 d における語 w の発生頻度である．

w を g に，d を w に代え，PLSI のモデルを語の共起確率 $P(w, g)$ の推定に用いることを考える．つまり，

$$P(w, g) = \sum_{z \in Z} P(w|z)P(g|z)P(z) \qquad (10)$$

とする．このとき，以下のように対応をとると，

$C_{i,j} \Longleftrightarrow P(w_i, g_j)$,
$x_{i,k} \Longleftrightarrow P(w_i|z_k)$,
$a_{k,j} \Longleftrightarrow P(g_j|z_k)P(z_k)$,

PLSI のモデルと提案モデルが非常に良く対応付けられることが分かる．相違は，パラメータ空間とパラメータ推定のための目的関数である．PLSI では各パラメータは確率値であるため，たとえば，

$$\sum_{i=1}^{M} P(w_i|z_k) = 1,$$
$$P(w_i|z_k) \geq 0 \quad ; \quad i = 1, 2, \cdots, M$$

という制約がある．一方，提案モデルではそのような制約はない．このため，提案モデルでは負例を考えないと，つまり，$\alpha_{i,j} = 0$ とすると，$a_{1,j} = a_{2,j} = \cdots = a_{K-1,j} = 0$，$a_{K,j} = 1$ $(j = 1, 2, \cdots, N)$ という目的関数を 0 にする自明な最小解が存在する．これは，すべての語の組 $\langle w, g \rangle$ の共起性を 1 と推定するものであり，元々の目的に反する．そこで，提案モデルでは負例を考慮した目的関数となっている．しかし，負例を積極的に利用することにより，PLSI のモデルより精度良く共起性を推定できる可能性がある．実際，4.3 節で述べる比較実験では，提案手法により求めた共起性を用いた場合の方が，PLSI モデルにより求めた共起確率を用いた場合より，曖昧さ解消の精度が高かった．これは，

- 負例を積極的に用いた効果，
- PLSI モデルを語の共起確率の推定に用いるには，実験で用いた学習データが小さすぎる，

などの理由が考えられる．

タグ付けされていない生のテキストコーパスから格フレームを自動構築する研究が行われている[5]．これは，

(1) 生のテキストコーパスを構文解析し，解析結果から信頼度の高い係り受けのみを用例として収集する，
(2) 用例を，用言とその直前の格要素の組を単位としてまとめ，用例パターンを作成する，
(3) シソーラスを用いて用例パターンをクラスタリングし，（用例）格フレームを作成する，

というものである．得られた（用例）格フレームを用いた構文解析では，用例格フレームを用例とする用例主導の手法で曖昧さ解消を行う．このとき，用例との類似度計算にシソーラスを用いる．格フレームは（2 項の）共起性より，曖昧さ解消のための知識としては強力である．しかし，格フレームを構築する際と解析に用いる際にシソーラスを用いており，先のシソーラスを用いた研究事例と同じく，このことから生じる問題を含むと考えられる．ただし，文献 5) で述べられている『用言の直前の格要素が用言の用法を強く規定している』という考えは興味深い．この考えを取り入れて，共起性推定のモデル，式 (1) を格フレーム「$n_1\ f_1\ n_2\ f_2\ \cdots\ n_t\ f_t\ v$」（$n_i$ は名詞，f_i は格を規定する助詞などの機能語，v は動詞）の妥当性を推定する手法へ拡張することが考えられる．

4. 実　　験

本章では，共起 ⟨ 名詞，助詞，動詞 ⟩ を対象として行った共起性推定実験とその評価実験について述べる．

4.1 共起性推定実験とその諸条件

EDR 電子化辞書の日本語コーパス（JCO-V020E）から，4.3 節で述べる係り先判定実験の評価データの元となる文を除き，残った文に出現する名詞と助詞・動詞との共起を抽出，共起の列 S_0 を作成した．これから，

$$S = \{\langle n, \langle c, v \rangle \rangle \in S_0 \mid f(n\,;\,S) \geq 2,\ f(\langle c, v \rangle\,;\,S) \geq 2\}$$

を満足し，かつ，要素数最大の S を求め，これを共起性推定のための学習データとした☆．ただし，$f(n; S)$,

☆ S_0 では，名詞の異なり数 M，助詞・動詞の異なり数 N が大きくなりすぎ，利用した計算機での計算が困難になるため，なるべく多くの共起データが使えるように，上記のようにして規模を小さくした．

は S における名詞 n の出現頻度，$f(\langle c,v\rangle ; S)$ は S における助詞・動詞 $\langle c,v\rangle$ の出現頻度である．S の共起の総数は 213,663，異なり数は 162,589 であり，S 中の名詞の異なり数は 16,543，助詞・動詞の組の異なり数は 14,474 である．

S を学習データとして，2 章の手法により〈名詞，助詞，動詞〉の共起性の推定実験を行った．式 (7) の λ は，S における各名詞，助詞・動詞の発生確率（相対頻度）に基づいて，$\langle n,\langle c,v\rangle\rangle$ をランダムに 500 組発生させ，これを人間の内省に基づいて，正例と負例に振り分け，

$$\lambda = \frac{\text{負例数}}{\text{正例数}} = \frac{342}{158} \simeq 2.16$$

とした．また $\alpha^*_{i,j}$ としては式 (8) を用いた．ワードベクトルおよび重みベクトルの次元 K としては，6〜10 を試した．

2.3 節で述べたパラメータの推定法は，適当な X から出発して，目的関数 $F(X,A)$ を極小にする (X,A) を求める手法である．したがって，X の初期値が必要であり，また，求まる (X,A) は X の初期値に依存する．そこで，各次元 K ごとに，

- X の初期値として $[-1,1]$ 上の一様乱数に基づく 10 通りの初期値を用意，
- 各初期値から始めて十分に目的関数の値が収束するまで，A と X の値の更新を繰り返すことにより (X,A) を求め，
- 上記で十分に収束点に達した 10 通りの (X,A) のうち，$F(X,A)$ が最小となる $(\widetilde{X},\widetilde{A})$ を求め，

$(\widetilde{X},\widetilde{A})$ をその次元における学習結果とした．繰返し回数は，A を更新し，その後 X を更新するのを 1 ステップと数えて，400 回とした☆．

4.2 共起性の推定値の分布による直接評価実験

共起性の推定値を人間の内省に基づいて評価することは，非常に難しく，また，客観性に欠ける．そこで，以下の傾向がどのくらい顕著に表れているかで，得られた共起性の推定値の評価を行った．

(1) 学習データ S 中の語の組に対する共起性の推定値は比較的大きい．

表 1 共起性の推定値の分布
Table 1 Distribution of cooccurrency.

対象	総数	異なり数	平均	分散
S	213,663	162,589	0.769	0.056
S^C	約 24 千万	約 24 千万	0.076	0.032
$S^C \cap S'$	143,453	85,676	0.610	0.105

(2) S^C 中の語の組（学習データ S に含まれない語の組）☆☆に対する共起性の推定値の平均値は比較的小さい．

(3) 学習データ S には含まれないが，他のコーパスから作成した共起データ S' に含まれる語の組に対する共起性の推定値は比較的大きい．

他のコーパスとして RWC テキストデータベース第 2 版を用いて共起を抽出し，S' を作成した☆☆☆．次節の係り先の判定実験で最も正解率の高かった $K=7$ の場合の共起性の推定値の分布（平均，分散）を**表 1** に示す．表 1 に示した平均と分散から分かるように，S 中の語の組に対する共起性の推定値は比較的高く（S 全体のうち，共起性の推定値が 0.20 以上のものは 99.2%，0.40 以上のものは 91.8%，0.60 以上のものは 74.7%であった），S^C 中の語の組に対する共起性の推定値の平均値はきわめて低い．また，$S^C \cap S'$ 中の語の組に対する共起性の推定値は，平均的には S の場合よりも低いものの，比較的高い値である（$S^C \cap S'$ 全体のうち，共起性の推定値が 0.20 以上のものは 87.3%，0.40 以上のものは 73.7%，0.60 以上のものは 55.2%であった）．これから前述の傾向 (1)，(2)，(3) が顕著に表れていることが分かる．また，S^C からランダムに 143453 個取り出した語の組の共起性の推定値 $C(1), C(2), \cdots, C(143453)$ の（標本）平均 $\overline{C}$ が偶然 0.610 以上となる確率は，チェビシェフの不等式を用いて見積もると，

$$\begin{aligned} &P\left\{\overline{C} \geq 0.610\right\} \\ &\quad \leq P\left\{\left|\overline{C} - E\left[\overline{C}\right]\right| \geq 0.610 - E\left[\overline{C}\right]\right\} \\ &\quad \leq \frac{Var\left[\overline{C}\right]}{\left(0.610 - E\left[\overline{C}\right]\right)^2} \simeq 7.8 \times 10^{-7} \end{aligned}$$

と，きわめて小さな確率になる．ただし，

$$E\left[\overline{C}\right] \simeq 0.076, \quad Var\left[\overline{C}\right] \simeq \frac{0.032}{143453}$$

☆ 400 回目の目的関数の値は，$K=6$ で 52000 程度，$K=10$ で 42000 程度であった．どの次元においても，400 回以降では，試した 10 通りのうち 1 回の更新で最も変化が大きいものでも 0.5 未満であり（$K=10$ 以外は 10 通りのうち半数は小数点第 6 位まで変化がなかった），ほぼ収束していると考えた．得られた 10 通りの (X,A) がどの程度ばらついているかであるが，たとえば，$K=7$ の場合で，目的関数値で 17，また，4.3 節と同様の曖昧さ解消実験のクローズドテストでの type1 の文の正解率で，2.5%のばらつきがあった．

☆☆ S に現れる，名詞の全体集合を $NOUN$，助詞・動詞の全体集合を CV とすると，$S^C = NOUN \times CV - S$，つまり，すべての名詞と助詞・動詞の組の集合から S を除いたものである．

☆☆☆ RWC テキストデータベースには，構文構造は示されていないため，形態素情報だけで曖昧さなく係り先が特定できる〈名詞，助詞，動詞〉を取り出した．

と近似した．このことから，学習データに含まれないにもかかわらず，共起しうる語の組の共起性を上手く選択的に高い値として推定していることが分かる．

4.3 係り先判定による共起性の推定値の間接評価実験

4.1 節の実験で得られた〈名詞，助詞，動詞〉の共起性を評価する目的で，得られた共起性を利用して，

$$n \quad cs \quad \gamma \quad v_1 \quad \delta \quad v_2 \qquad (11)$$

という形態の文における「n c」の係り先の判定実験を行った．ただし，n は名詞，cs は助詞（格助詞および係助詞）の列，v_1，v_2 は動詞であり，γ，δ は単語列である．c は cs 中の助詞のうち，n の係りの種類を規定する助詞で，n を主辞とする後置詞句に格助詞と係助詞がともに含まれる場合は格助詞を c とする．また,「n c」は v_1 または v_2 に係り，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，

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

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いろいろと報告されている[12]．たとえば，

(a) n を主辞とする後置詞句に『は』が含まれるか（cs 中に『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，4.3.1 項の評価用データの内訳にもよく表れている．ところが，評価用データの内訳からも分かるように，このような表層的な手がかりがない（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）文の場合,「n c」の係り先を精度良く判定することは難しい．そこで，このような表層的な手がかりがない文について，$\langle n, c, v_1\rangle$ と $\langle n, c, v_2\rangle$ の共起性を利用した「n c」の係り先の判定実験を行った．共起性として，次の 3 つ，

- 提案手法で推定した共起性,
- 3 章で述べた，シソーラスを用いた用例主導の推定法により求めた共起性,
- PLSI モデルを用いた共起性,

を試し，提案手法で推定した共起性を評価した．

表 2 評価用データ
Table 2 Data for evaluation.

type	h	d	総数	係り先	
				v_1	v_2
1	0	0	764	474 (62.0%)	290 (38.0%)
2	0	1	1,341	1,285 (95.8%)	56 (4.2%)
3	1	0	342	23 (6.7%)	319 (93.3%)
4	1	1	41	5 (12.2%)	36 (87.8%)
全体			2,488	1,787 (71.8%)	701 (28.2%)

4.3.1 評価用データ

EDR 電子化辞書の日本語コーパス（JCO-V020E）から式 (11) の形態の文のうち，n，$\langle c, v_1\rangle$，$\langle c, v_2\rangle$ がすべて 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，d の値は真偽値で，h は 4.3 節冒頭で述べた (a) に関する情報，d は (b) に関する情報である．また，ans は「n c」の係り先である．

抽出したデータの詳細を**表 2** に示す．表の type1 の文が実験対象の文である．

4.3.2 共起性を用いた係り先判定法

4.3.1 項の評価用データの内訳から分かるように，表層的な手がかりがない場合,「n c」は v_1 に係る傾向が比較的強い．このことを考慮し，共起性を利用して，式 (11) の形態の文に対する「n c」の係り先を

$$\begin{array}{lll} C(n,c,v_1) < \theta \cdot C(n,c,v_2) & \Longrightarrow & v_2 \\ \text{その他} & \Longrightarrow & v_1 \end{array} \qquad (12)$$

と判定する．ただし，$C(n,c,v)$ は $\langle n, c, v\rangle$ の共起性，θ はスレッシュホールド（$0 < \theta < 1$）である．

提案手法による共起性を利用する場合，式 (12) の判定法でのパラメータは，スレッシュホールド θ と共起性推定のモデルの次元 K である．提案手法で推定した共起性は，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．

シソーラスを用いた用例主導の推定法による共起性を利用する場合，式 (12) の判定法でのパラメータは，θ と共起性推定モデルのパラメータ k である．シソーラスとしては，EDR 電子化辞書の概念辞書 CPD-V020.1 および日本語単語辞書 JWD-V020 を用いた．名詞間の類似度 sim としては，以下の 2 つの類似度 sim_1 と sim_2 を試した．sim_2 は基本的には文献 8)

で紹介されている類似度である．

$$sim_1(n_1, n_2) = \max_{c \in CM(n_1,n_2)} \frac{1}{2}\left(\frac{D(c)}{D(n_1\,;\,c)} + \frac{D(c)}{D(n_2\,;\,c)}\right),$$

$$sim_2(n_1, n_2) = \max_{c \in CM(n_1,n_2)} \frac{2D(c)}{D(n_1\,;\,c) + D(n_2\,;\,c)}$$

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）ノードの集合，$D(a)$ は a の深さ，すなわち，シソーラスのルートノードからノード a への最短パス長，$D(a\,;\,b)$ は b を経由する a の深さ，すなわち，シソーラスのルートノードからノード b を経由してノード a へ至るパスの最短パス長である．

PLSI モデルを用いた共起性を利用する場合，式 (12) の判定法でのパラメータは，スレッシュホールド θ と PLSI モデルの次元 K である．式 (10) で求められる $P(n, \langle c, v\rangle)$ を用いて，式 (12) の $C(n, c, v)$ を

$$C(n, c, v) = P(n|\langle c, v\rangle) = \frac{P(n, \langle c, v\rangle)}{P(\langle c, v\rangle)}$$

と定義する☆．なお，PLSI モデルのパラメータ学習では，提案モデルの場合と同様，各次元に対して，10 通りのパラメータ初期値を試し，このうち最適な（つまり対数尤度式 (9) 最大の）結果を，その次元の推定結果とした．試した次元は $K = 5$〜11 である．

係り先判定実験では，表 2 の type1 の文に対する正解率を，10 分割のクロスバリデーション（学習データ 9 ブロック，テストデータ 1 ブロック）により求めた．すなわち，各手法のパラメータ，(θ, K)，(θ, k)，(θ, K) は 9 ブロック分の学習データにおける最適値として求める．なお，θ は 0.025 刻で試した．

4.3.3 実験結果

表 3 に結果を示す．参考までに，type 2 の場合 v_1 に係り，type 3 および 4 の場合 v_2 に係ると判定し，type 1 の場合のみそれぞれの手法で判定した場合の全体の正解率を最後の行に示している．表中の「シソーラス法」はシソーラスを用いた用例主導の推定法を表す．

提案手法の場合，どのブロックを学習データとした場合でも，$K = 7$ が選ばれ，PLSI の場合は，どのブロックを学習データとした場合でも，$K = 10$ が選ばれた．また，シソーラスを用いた用例主導の推定法による共起性の場合は $k = 2$ が選ばれた．

表 3 曖昧さ解消の正解率
Table 3 Accuracy rate of disambiguation.

対象	正解率 (%)			
	提案手法	シソーラス法		PLSI
		sim_1	sim_2	
type1	70.3	67.3	66.9	66.4
全体	87.5	86.6	86.5	86.3

表 3 に示すように，提案手法で推定した共起性を用いた場合が，他の手法による共起性を用いた場合に比べ，若干正解率が高い．

シソーラスを用いた用例主導の推定法による共起性は，今回用いたシソーラス以外のシソーラスを用いると異なった結果になると予想され，また，シソーラスに基づく名詞間の類似度の定義自体にも改善の余地がある．したがって，この実験結果から単純に，提案手法がシソーラスを用いた用例主導の推定法よりも名詞間の類似性を上手く取り入れた手法であるとは断言できない．しかし，用例主導の推定法の場合，蓄積している用例（共起データ）の増加にともない，共起性推定に要する記憶容量も計算時間も増す．一方，提案した重回帰モデルによる共起性の場合は，モデルのパラメータ X，A の学習には一定の時間を要するが，共起性自体の計算にはほとんど時間を要しない．また，単語の異なり数が変わらなければ，共起データが増加しても必要な記憶容量は一定である．この点では，提案手法による共起性の推定は用例主導の推定法よりも明らかに有効といえる．

提案手法と同じくパラメトリックな推定を行う PLSI モデルによる共起確率の推定法は，明確なセマンティックを持つ優れた手法である．しかし，3 章で述べたように，実験で用いた規模の共起データでは十分に精度良く共起確率を推定することが困難であったと思われる．一方，提案手法はこのような規模のデータでも，疑似的ではあるが負例を用いることにより，比較的精度良く共起性が推定できたものと思われる．

5. おわりに

観測された共起データを基にして，重回帰モデルにより共起性を推定する手法を提案し，EDR 日本語コーパスから抽出した共起データを用いて，〈名詞，助詞，動詞〉の共起性の推定実験を行った．得られた共起性に対して，

- 共起性の推定値の分布による直接評価，

☆ 確率的な文の生成モデルを考えた場合，$C(n, c, v)$ として $P(n, \langle c, v\rangle)$ を用いるよりも，上記の条件付確率とする方が，妥当と思われる．ちなみに，PLSI モデルで求まる共起確率を共起性とした，つまり，$C(n, c, v) = P(n, \langle c, v\rangle)$ とした実験も行ったが，後述で示す結果より若干低い正解率であった．

- 共起性を用いた名詞句の係り先の判定実験による間接評価,

を行い，提案する共起性推定法の有効性を示した．提案手法は，重回帰モデルという単純なモデルであること，説明変量であるワードベクトルも学習すること，正例列から得られる疑似的な負例を積極的に利用していることに特徴がある．

今回の実験では，〈名詞，助詞，動詞〉の共起データから，〈名詞，助詞，動詞〉の共起性を推定したが，他の品詞の語に関する共起も学習データに含めるとどのようになるであろうか．たとえば，$\langle n,$『が』,『おいしい』$\rangle$ と $\langle n,$『を』,『食べる』$\rangle$ の共起性には強い相関がある．このことから，動詞との共起だけでなく，形容詞や形容動詞との共起も学習データに取り入れるならば，〈名詞，助詞，動詞〉の共起性自身もより精度良く推定されることが期待できる．

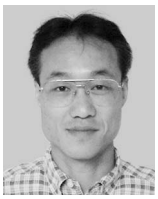
また，単語を語義で細分化し，単語と語義番号の組を対象にして，語義 i の語 w が f で語義 j の語 w' に係るという共起 $\langle (w,i), f, (w',j)\rangle$ を考えることもできる．このような共起性は，多義語の語義の曖昧さの解消にも使える．

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った．

参 考 文 献

1) Dagan, I., Lee, L. and Pereira, F.: Similarity-Based Models of Word Cooccurrence Probabilities, *Machine Learning*, Vol.34, No.1–3, pp.43–69 (1999).
2) Grishman, R. and Sterling, J.: Generalizing Automatically Generated Selectional Patterns, *Proc. COLING'94*, pp.742–747 (1994).
3) Hindle, D.: Noun Classification from Predicate-Argument Structures, *Proc. 28th Annual Meeting of ACL*, pp.268–275 (1990).
4) Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proc. SIGIR'99*, pp.50–57 (1999).
5) 河原大輔，黒橋禎夫：用言と直前の格要素の組を単位とする格フレームの自動構築，自然言語処理，Vol.9, No.1, pp.3–19 (2002).
6) 北　研二：言語と計算 4 確率的言語モデル，東京大学出版会 (1999).
7) 松永健司，田中英彦：コーパスから抽出した係り受け共起情報に基づく類似度と文書検索における評価，情報処理学会自然言語処理研究会，Vol.96, No.114 (96-NL-116), pp.73–78 (1996).
8) 長尾　真：岩波講座ソフトウェア科学 15 自然言語処理，p.601, 岩波書店 (1996).
9) Pereira, F., Tishby, N. and Lee, L.: Distributional clustering of English Words, *Proc. 31st Annual Meeting of ACL*, pp.183–190 (1993).
10) 白井清昭，乾健太郎，徳永健伸，田中穂積：統計的構文解析における構文的統計情報と語彙的統計情報の統合について，自然言語処理，Vol.5, No.3, pp.85–106 (1998).
11) 隅田英一郎，古瀬　蔵，飯田　仁：英語前置詞句係り先の用例主導あいまい性解消，信学論，Vol.J77-D-II, No.3, pp.557–565 (1994).
12) 内元清貴，関根　聡，井佐原均：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, pp.3397–3407 (1999).

(平成 15 年 4 月 23 日受付)
(平成 15 年 10 月 16 日採録)



冨浦　洋一（正会員）

昭和 59 年九州大学工学部電子工学科卒業，平成元年九州大学大学院工学研究科電子工学専攻博士課程単位取得退学．同年九州大学工学部助手，平成 7 年同助教授，現在，同大学院システム情報科学研究院助教授，工学博士．自然言語処理，計算言語学，人工知能に関する研究に従事．人工知能学会，言語処理学会各会員．

日高　　達（正会員）

昭和 40 年九州大学工学部電子工学科卒業，昭和 44 年九州大学大学院工学研究科電子工学専攻博士課程中退．同年九州大学工学部助手，昭和 48 年同講師，昭和 55 年同助教授，昭和 63 年同教授，平成 8 年同大学院システム情報科学研究科教授，平成 15 年退官．工学博士．