自 然 言 語 処 理 154−11
（２００３．３．６）

共起データに基づく名詞の n 次元空間への配置

冨浦洋一 *，田中省作 **，日高達 *

単語間の統語的・意味的な類似度（あるいは距離）は自然言語処理における基本的な知識の一つである．シソーラスに基づいて語の類似度や距離を求めるのが最も一般的であるが，言語コーパスから，多変量解析的な手法でこれを求めることも考えられる．本予稿では，『語 w と関係 f で共起する名詞は類似している』という考えに基づき，この類似性を反映するように，名詞を n 次元空間に配置する手法について報告する．

Placement of Nouns in n-Dimensional Space Based on Cooccurrency

Yoichi Tomiura*, Shosaku Tanaka**, Toru Hitaka*

The syntactic and semantic similarity (or distance) between words is one of the basic knowledge in Natural Language Processing. It's the most popular way to seek it through a thesaurus, but we can also get it from a corpus with Multivariate Analysis. This paper reports a way to place words in n-dimensional space based on the idea that "nouns which cooccur a word w with a relation f are similar", so that the placement reflects this similarity.

1 はじめに

単語間の統語的・意味的な類似度（あるいは距離）は自然言語処理における基本的な知識の一つである．たとえば，文の構文解析における曖昧さの解消では，文とその構文構造の対（用例）を多数用意し，入力文の構文構造を類似する用例に従って解析することができる [1][2]．また，多義語の意味の選択，機械翻訳における訳語の選択などでも類似用例に従った処理が可能である [3]．

類似用例に基づく手法では，入力文（句）と用例の間の類似度（あるいは距離）の計算をする必要があるが，その基本は単語間の類似度（距離）計算である．従来，単語間の上位下位関係を記述したシソーラスを用いて，単語間の類似度（あるいは距離）を求めていた．しかし，シソーラスは，人手で作成したものであるため作成者の主観の影響が大きく，また，単語の意味のある側面を捉えて単語を整理したものであり，単語間の距離の拠り所としては問題がある．

本稿では，共起情報（観測された共起性を持つ語の組の列）を基にして，単語を実数ベクトル（単語ベクトル）に対応させる手法を提案する．本手法で求められる単語ベクトルは共起情報という客観的な基準に基づいたものであるため，共起情報さえ十分に収集するならば信頼性の高いものとなり，求めた単語ベクトルを基に定義された単語間の類似度や距離も信頼性が高いと期待できる．

2 基本的な考え方

名詞 n が関係 f で語 w に係っている場合，n と $\langle f, w\rangle$ が共起すると呼ぶことにする．

* 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 情報基盤センター
Computing and Communications Center, Kyushu University

$\langle f, w \rangle$ の全体集合を S_G，名詞の全体集合を S_N（S_N の要素数を m）とし，各名詞を S_G との共起データを用いて，K 次元空間に配置する[*]（$K \ll m$）

$n \in S_N$ が K 次元ユークリッド空間上の $\boldsymbol{x}(n)$ に対応するとする．名詞ベクトルの平均ベクトルを $\boldsymbol{\mu}$，$g \in S_G$ と共起している名詞の平均ベクトルを $\boldsymbol{\mu}(g)$，

$$\begin{aligned} \boldsymbol{\mu} &= E[\boldsymbol{x}(N)] = \sum_{n \in S_N} \boldsymbol{x}(n) f_N(n) \\ \boldsymbol{\mu}(g) &= E[\boldsymbol{x}(N)\,;\, f_{N|G}(\cdot|g)] \\ &= \sum_{n \in S_N} \boldsymbol{x}(n) f_{N|G}(n|g) \end{aligned}$$

とおくと（$f_N(n)$ は名詞 n の発生確率，$f_{N|G}(n|g)$ は g と共起する場合の n の条件付発生確率である），名詞ベクトルの分散（厳密には共分散行列の対角成分の和）は，

$$\begin{aligned} & E[\,|\boldsymbol{x}(N) - \boldsymbol{\mu}|^2\,] \\ &= \sum_{n \in S_N} |\boldsymbol{x}(n) - \boldsymbol{\mu}|^2 f_N(n) \\ &= \sum_{n \in S_N} |\boldsymbol{x}(n) - \boldsymbol{\mu}|^2 \sum_{g \in S_G} f_G(g) f_{N|G}(n|g) \\ &= \sum_{g \in S_G} f_G(g) \sum_{n \in S_N} \\ &\qquad |\boldsymbol{x}(n) - \boldsymbol{\mu}(g) + \boldsymbol{\mu}(g) - \boldsymbol{\mu}|^2 f_{N|G}(n|g) \\ &= \sum_{g \in S_G} f_G(g) \sum_{n \in S_N} |\boldsymbol{x}(n) - \boldsymbol{\mu}(g)|^2 f_{N|G}(n|g) \\ &\qquad + \sum_{g \in S_G} f_G(g)\,|\boldsymbol{\mu}(g) - \boldsymbol{\mu}|^2 \qquad (1) \end{aligned}$$

である．

$g \in S_G$ と共起する名詞同士は類似していると考えられ，名詞ベクトルがこの類似性を反映しているならば，g と共起する名詞は，K 次元空間上でグループを成す．したがって，グループ内での名詞ベクトルの分散の S_G 全体での平均である (1) 式の第 1 項

$$\sum_{g \in S_G} f_G(g) \sum_{n \in S_N} |\boldsymbol{x}(n) - \boldsymbol{\mu}(g)|^2 f_{N|G}(n|g)$$

は小さく，一方，グループ間の分散である (1) 式の第 2 項

$$\sum_{g \in S_G} f_G(g)\,|\boldsymbol{\mu}(g) - \boldsymbol{\mu}|^2$$

は大きいと期待できる．そこで，(1) 式第 1 項を最小，第 2 項を最大にするように，各 $n \in S_N$ に対して $\boldsymbol{x}(n)$ を求める．

ここで，名詞のベクトルに対して，以下の制約を課す．

制約 1 $\boldsymbol{\mu} = \sum_{n \in S_N} \boldsymbol{x}(n) f_N(n) = \boldsymbol{0}$

制約 2 $E[(\boldsymbol{x}(N) - \boldsymbol{\mu})\ {}^t(\boldsymbol{x}(N) - \boldsymbol{\mu})] = I_K$
ただし，${}^tX$ は X の転置行列，I_K は K 次の単位行列．

名詞間の距離や類似度を求めるために名詞を K 次元空間に配置するのであるから，相対位置のみが重要であり，名詞ベクトルの原点はどこでも構わない．したがって，制約 1 は本質的な制約ではない．また，名詞ベクトルの各成分が互いに独立でその分散が等しいというのが制約 2 である（分散の大きさは本質的な制約ではない）．

制約 2 より，

$$E[\,|\boldsymbol{x}(N) - \boldsymbol{\mu}|^2\,] = K$$

であるから，(1) の第 1 項を最小，第 2 項を最大にする各名詞ベクトル $\boldsymbol{x}(n)(n \in S_N)$ は，制約 1，2 の下で，

$$F = \sum_{g \in S_G} f_G(g)\,|\boldsymbol{\mu}(g)|^2 \qquad (2)$$

を最大にすることが分かる．

3 解法

$S_N = \{n_1, n_2, \cdots, n_m\}$ とし，

$$X = \begin{bmatrix} {}^t\boldsymbol{x}(n_1) \\ {}^t\boldsymbol{x}(n_2) \\ \vdots \\ {}^t\boldsymbol{x}(n_m) \end{bmatrix} \quad \boldsymbol{f}(g) = \begin{bmatrix} f_{N|G}(n_1|g) \\ f_{N|G}(n_2|g) \\ \vdots \\ f_{N|G}(n_m|g) \end{bmatrix}$$

[*] 以降，ベクトルはすべて列ベクトルとする

とする．

$$\begin{aligned}\boldsymbol{\mu}(g) &= {}^tX\boldsymbol{f}(g)\\ |\boldsymbol{\mu}(g)|^2 &= \mathrm{trace}\,\boldsymbol{\mu}(g)\,{}^t\boldsymbol{\mu}(g)\\ &= \mathrm{trace}\ {}^tX\boldsymbol{f}(g)\,{}^t\boldsymbol{f}(g)X\end{aligned}$$

であるから，

$$\begin{aligned}F &= \sum_{g\in S_G} f_G(g)\,\mathrm{trace}\ {}^tX\boldsymbol{f}(g)\,{}^t\boldsymbol{f}(g)X\\ &= \mathrm{trace}\ {}^tX\left(\sum_{g\in S_G} f_G(g)\boldsymbol{f}(g)\,{}^t\boldsymbol{f}(g)\right)X\end{aligned}$$

と表せる．計算の見通しを良くするために，

$$\Delta=\begin{bmatrix} f_N(n_1) & & \boldsymbol{O}\\ & \ddots & \\ \boldsymbol{O} & & f_N(n_m)\end{bmatrix}$$

とおき，$Y=\Delta^{\frac{1}{2}}X$ なる変換を考える．制約 1 は，

$${}^tY\Delta^{\frac{1}{2}}\mathbf{1}=\mathbf{0} \tag{3}$$

となり，制約条件 2 は，

$${}^tYY=I_K \tag{4}$$

となる．また，

$$\Delta^{-\frac{1}{2}}\left(\sum_{g\in S_G} f_G(g)\boldsymbol{f}(g)\,{}^t\boldsymbol{f}(g)\right)\Delta^{-\frac{1}{2}}=A$$

とおくと，目的関数は，

$$F=\mathrm{trace}\ {}^tYAY \tag{5}$$

と表せる．

A は実対象行列ゆえ，直交行列 Φ，対角行列 Λ を用いて，

$${}^t\Phi A\Phi=\Lambda$$

と対角化できる．ただし，A の固有値を $\lambda_0,\lambda_1,\cdots,\lambda_{m-1}$，それぞれに属する大きさ 1 の固有ベクトルを $\boldsymbol{e}_0,\boldsymbol{e}_1,\cdots,\boldsymbol{e}_{m-1}$ とすると，

$$\begin{aligned}\Lambda &= \begin{bmatrix}\lambda_1 & & \boldsymbol{O}\\ & \lambda_2 & \\ & & \ddots\\ \boldsymbol{O} & & \lambda_0\end{bmatrix}\\ \Phi &= [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_0]\end{aligned}$$

である．ここで，

$${}^t\boldsymbol{f}(g)\mathbf{1} = 1$$

$$\sum_{g\in S_G} f_G(g)\boldsymbol{f}(g) = \begin{bmatrix} f_N(n_1)\\ f_N(n_2)\\ \vdots\\ f_N(n_m)\end{bmatrix} = \Delta\mathbf{1}$$

に注意すると，

$$\begin{aligned}&A(\Delta^{\frac{1}{2}}\mathbf{1})\\ &= \Delta^{-\frac{1}{2}}\left(\sum_{g\in S_G} f_G(g)\boldsymbol{f}(g)\,{}^t\boldsymbol{f}(g)\right)\mathbf{1}\\ &= \Delta^{-\frac{1}{2}}\left(\sum_{g\in S_G} f_G(g)\boldsymbol{f}(g)\right)\\ &= \Delta^{\frac{1}{2}}\mathbf{1}\end{aligned}$$

であるから，A は固有値として 1 を持ち，1 に属する固有ベクトルが $\Delta^{\frac{1}{2}}\mathbf{1}$ であることが分かる．$\lambda_0=1$，$\boldsymbol{e}_0=\Delta^{\frac{1}{2}}\mathbf{1}$ とし，$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_{m-1}$ とする．任意のベクトルは，$\boldsymbol{e}_0,\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_{m-1}$ の線形和で表現できるので，

$$Y=\Phi C \qquad (C\text{ は }m\times K\text{ 行列})$$

と表せる．これを目的関数に代入すると，

$$\begin{aligned}F &= \mathrm{trace}\ {}^tYA\,Y\\ &= \mathrm{trace}\ {}^tC\,{}^t\Phi A\Phi C\\ &= \mathrm{trace}\ {}^tC\Lambda C\\ &= \sum_{i=1}^{K}\left(\sum_{j=1}^{m-1}\lambda_j(C)_{ji}{}^2+\lambda_0(C)_{mi}{}^2\right)\\ &= \sum_{j=1}^{m-1}\lambda_j\sum_{i=1}^{K}(C)_{ji}{}^2\\ &\qquad +\lambda_0\sum_{i=1}^{K}(C)_{mi}{}^2\end{aligned} \tag{6}$$

となる．$\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_{m-1}$ は互いに直交するので，(3) 式より，

$$\mathbf{0} = {}^t(\Phi C)\boldsymbol{e}_0 = {}^tC \begin{bmatrix} {}^t\boldsymbol{e}_1 \\ {}^t\boldsymbol{e}_2 \\ \vdots \\ {}^t\boldsymbol{e}_{m-1} \\ {}^t\boldsymbol{e}_0 \end{bmatrix} \boldsymbol{e}_0 = {}^tC \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

したがって，

$$(C)_{mi} = 0 \quad ; \quad i = 1, 2, \cdots, K \qquad (7)$$

である．また，${}^t\Phi\Phi = I_K$ であるから，(4) 式より，

$$I_K = {}^tYY = {}^tC\,{}^t\Phi\Phi C = {}^tCC$$

したがって，

$$\sum_{i=1}^{K} {(C)_{ji}}^2 \leq 1 \qquad (8)$$

$$\sum_{j=1}^{m}\sum_{i=1}^{K} {(C)_{ji}}^2 = K \qquad (9)$$

である．

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m-1}$ および (6)(7)(8)(9) より，制約 1 および 2 を満たす X，つまり，(3)(4) 式を満たす Y では，

$$F \leq \sum_{j=1}^{K} \lambda_j \qquad (10)$$

である．等号成立条件は，各 $j(\in \{1, 2, \cdots, K\})$ で

$$\sum_{i=1}^{K} {(C)_{ji}}^2 = 1.$$

したがって，F の最大値 $\lambda_1 + \lambda_2 + \cdots + \lambda_K$ を実現する Y の一つ† は，

$$Y = \Phi \begin{bmatrix} I_K \\ \cdots \\ \boldsymbol{O} \end{bmatrix} = [\boldsymbol{e}_1\, \boldsymbol{e}_2\, \cdots\, \boldsymbol{e}_K] \qquad (11)$$

であり，X は Y より，$\Delta^{-\frac{1}{2}}Y$ で求まる．

実際には行列 A は，観測された共起データ D から

$$f_N(n) = \frac{D \text{ における } n \text{ の頻度}}{|D|}$$

$$f_G(g) = \frac{D \text{ における } g \text{ の頻度}}{|D|}$$

$$f_{N|G}(n|g) = \frac{D \text{ における} \langle n, g \rangle \text{の頻度}}{D \text{ における } g \text{ の頻度}}$$

として求める．

4 実験

EDR コーパス [4] から名詞と ⟨ 助詞, 動詞 ⟩ の共起データを抽出し，これを D とした．ただし，D 中の各 $\langle n, \langle c, v \rangle\rangle$（$n$ は名詞，$\langle c, v \rangle$ は助詞，動詞の組）の n および $\langle c, v \rangle$ の D 中での頻度がともに 10 以上となるように D を求めた．データの規模は，

$$|D| = 125,189$$
$$|S_N| = 3,144$$
$$|S_G| = 3,153$$

である．$K = 80$ として，本手法により各名詞のベクトルを求めた．

得られた名詞ベクトルを，類似用例に基づく手法に適用してその精度により評価することも考えられるが，ここでは，得られた名詞ベクトルにより名詞間の距離を求めて，人間の感覚との比較を行なった．

[評価法]

1. S_N からランダムに 100 個の名詞を選ぶ．これを T とする．

2. 各 $n\,(\in T)$ に対して，得られた単語ベクトルによる計算で距離が最小のものから 10 個の名詞を求め，これを $\Gamma(n)$ とする．

3. 各 $n\,(\in T)$ に対して，$\Gamma(n)$ 中の 10 個の名詞の内で n に最も類似している名詞を選ぶ（被験者の感覚による）．

† (3)(4) 式の下で F を最大にする Y には回転変換分の自由度が残る．

4. 上記で選んだ名詞が $\Gamma(n)$ 中で何番目に n との距離が小さかったかの平均値を求める．

ただし，上記 3 で，類似する名詞がない場合はその n を対象外とした．たとえば，

$$\Gamma(\text{傷}) = \{\text{一端}, \text{借金}, \text{任務}, \text{難}, \text{危険}, \text{債務}, \text{責任}, \text{動機}, \text{疲れ}, \text{税}\}$$

である．『傷』と Γ(傷) の各名詞との距離は 1.4 ～ 3.3 であった．T の名詞のうち評価対象は 79 で，最も類似しているとして選んだ名詞の距離の順位の平均は 2.4 であった．

付録に，全名詞間で最小距離のものから 50 個の名詞の組とその距離，および上記の評価に用いたデータの一部を示す．

5 関連研究

共起情報や辞書における参照関係を基に単語の特徴ベクトルを構成し，これから次元を圧縮することで単語ベクトルを求める方法がいくつか提案されている．たとえば，主成分分析により次元を圧縮する手法 [5]，シソーラスを利用する手法，シソーラスを利用して次元を圧縮した後特異値分解する手法 [6] などがある．これらの手法では，単語の特徴は基本的には元の特徴ベクトルが表しているが，以下の理由により次元を圧縮している．

- そのままでは次元が高く，利用に際して計算コストがかかるため，次元を圧縮する必要がある．
- 使用したデータのスパース性から来る情報の欠落の影響を減らすため，次元圧縮を通して細かな差異を切捨てる必要がある．

したがって，得られた単語ベクトルは与えられた次元で元の特徴ベクトルが持つ情報を極力保存するように求められている．

これに対して本手法は，最初に単語ベクトルの元となる特徴ベクトルを与えるのではなく，同一の $\langle f, w\rangle$ と共起する名詞のベクトルの分散が小さくなるように単語ベクトルを求めるもので，異なる視点からのアプローチと言える．

6 おわりに

共起データに基づき，名詞を n 次元空間へ配置する手法を提案し，本手法に基づいて求めた名詞ベクトルに対して，若干の評価（人間の感覚との比較）を行なった．

単語ベクトルを求めるために使用するデータの規模や，単語ベクトルを何に利用するかに，得られた単語ベクトルの性能は依存する．したがって，これまでに提案されて来た手法との優劣を議論することは一概にはできないが，得られた単語ベクトルを実際の処理に利用して，評価を進めて行く予定である．

参考文献

[1] 隅田，古瀬，飯田: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)．

[2] 冨浦，日高:k-NN 推定法に基づく統語的あいまいさの解消法，信学論, Vol.J80-D-II, No.9, pp.2475-2481 (1997)

[3] 古瀬，隅田，飯田：経験的知識を活用する変換主導型機械翻訳, 情報処理学会論文誌, Vol.35, No.3, pp.414-425 (1994).

[4] 日本電子化辞書研究所：EDR 電子化辞書日本語コーパス (JCO-V020E)

[5] 小嶋，伊藤：意味空間のスケール変換による動的シソーラスの実現，信学技報，NLC95-19, pp.1-8 (1995)

[6] 笠原，稲子，加藤:単語の属性空間の表現方法，人工知能学会誌 Vol.17, No.5, pp.539-547 (2002)

付録

距離	名詞 1	名詞 2
0.000	重軽傷 (12)	重傷 (16)
0.000	軽傷 (10)	重軽傷 (12)
0.000	軽傷 (10)	重傷 (16)
0.096	概念 (57)	関数 (41)
0.099	１１月 (50)	７月 (47)
0.133	需要 (157)	体重 (20)
0.137	４月 (90)	７月 (47)
0.139	形状 (38)	集合 (48)
0.145	皇太子 (10)	国王 (15)
0.146	８月 (45)	９月 (54)
0.151	確保 (24)	民主化 (36)
0.158	概念 (57)	形状 (38)
0.160	コンピューター (166)	パソコン (125)
0.161	県 (43)	中小企業 (21)
0.162	ジョブ (14)	ファイル (47)
0.163	威力 (20)	指導力 (13)
0.165	ＯＡ化 (17)	合理化 (28)
0.167	ＨＰ (11)	富士通 (80)
0.167	国際化 (30)	情報化 (15)
0.169	ヨーロッパ (36)	北海道 (23)
0.171	１１月 (50)	４月 (90)
0.171	建設 (88)	研究 (225)
0.172	緩和 (30)	設立 (31)
0.173	意思 (44)	考え (179)
0.173	ＮＴＴ (48)	日本ＩＢＭ (46)
0.175	解散 (21)	停止 (24)
0.175	出荷 (54)	販売 (402)
0.176	現地 (61)	東京 (171)
0.178	人員 (18)	数 (320)
0.179	プロセス (42)	方法 (332)
0.179	方式 (224)	方法 (332)
0.180	全域 (19)	都市 (55)
0.182	配列 (20)	論理式 (14)
0.182	見直し (61)	撤退 (22)
0.182	アメリカ (104)	欧米 (32)
0.184	関数 (41)	形状 (38)
0.184	国土庁 (12)	三井物産 (10)
0.185	西独 (58)	中小企業 (21)
0.185	活用 (24)	民営化 (18)
0.186	体内 (13)	町 (108)
0.187	外国 (65)	東独 (16)
0.192	ファイル (47)	符号 (37)
0.192	ところ (110)	体内 (13)
0.193	官僚 (10)	機関投資家 (14)
0.194	引き渡し (13)	停戦 (18)
0.194	段階 (122)	分野 (195)
0.194	自身 (75)	民社党 (12)
0.195	集合 (48)	文法 (14)
0.195	停戦 (18)	撤回 (21)
0.195	東京 (171)	名古屋 (10)

得られた単語ベクトルを用いた各名詞間の距離が小さい名詞の組，上位 50 組．名詞の横の括弧の中は共起データにおけるその名詞の頻度を示す．

名詞 1	距離	名詞 2
勢力 (41)	0.447	機械 (56)
	0.463	兵力 (10)
	0.511	要員 (13)
	0.519	人数 (16)
	0.544	収入 (66)
	0.551	貯蓄 (15)
	0.553	景気 (54)
	0.574	組織 (62)
	0.584	男子 (11)
	0.596	自動車 (28)
計算機 (99)	0.208	装置 (154)
	0.261	データベース (71)
	0.264	オフコン (15)
	0.293	変数 (27)
	0.301	入出力装置 (10)
	0.359	組合せ (19)
	0.361	ソフトウェア (89)
	0.362	回路 (50)
	0.363	アルゴリズム (35)
	0.371	ディスプレイ (11)
長男 (19)	0.689	子ども (43)
	0.711	弟 (51)
	0.725	娘 (41)
	0.748	孫 (22)
	0.778	子供 (106)
	0.812	一家 (25)
	0.880	父親 (48)
	0.882	人々 (143)
	0.883	息子 (38)
	0.896	母親 (70)
信頼性 (25)	0.399	性能 (87)
	0.558	互換性 (13)
	0.566	シェア (73)
	0.587	データ型 (17)
	0.588	特性 (59)
	0.589	価値 (46)
	0.591	率 (78)
	0.611	ノウハウ (16)
	0.612	質 (28)
	0.614	濃度 (25)
生徒 (41)	0.403	客 (111)
	0.466	子 (84)
	0.483	人 (494)
	0.499	読者 (25)
	0.517	報道陣 (13)
	0.519	犯人 (36)
	0.571	ファン (22)
	0.574	何度 (17)
	0.577	技術者 (22)
	0.599	学生 (89)
遠く (23)	0.369	店内 (13)
	0.402	目の前 (20)
	0.427	湖 (10)
	0.431	街 (56)
	0.457	ホーム (16)
	0.458	所 (49)
	0.468	庭 (33)
	0.476	校庭 (11)
	0.506	部屋 (99)
	0.510	ジャングル (11)

4 節の評価で用いた名詞間の距離の例．名詞の横の括弧の中は共起データにおけるその名詞の頻度を示す．