論 文

Technical Papers

共起データに基づく名詞の多次元空間への配置

Placement of Nouns in a Multi-Dimensional Space Based on Words' Cooccurrency

冨浦 洋一
Yoichi Tomiura
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
tom@is.kyushu-u.ac.jp

田中 省作
Shosaku Tanaka
九州大学 情報基盤センター
Computing and Communications Center, Kyushu University
sho@cc.kyushu-u.ac.jp

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

keywords: word vector, multivariate analysis, semantic similarity

Summary

The semantic similarity (or distance) between words is one of the basic knowledge in Natural Language Processing. There have been several previous studies on measuring the similarity (or distance) based on word vectors in a multi-dimensional space. In those studies, high dimensional feature vectors of words are made from words' cooccurrence in a corpus or from reference relation in a dictionary, and then the word vectors are calculated from the feature vectors through the method like principal component analysis. This paper proposes a new placement method of nouns into a multi-dimensional space based on words' cooccurrence in a corpus. The proposed method doesn't use the high dimensional feature vectors of words, but is based on the idea that "vectors corresponding to nouns which cooccur with a word w in a relation f constitute a group in the multi-dimensional space". Although the whole meaning of nouns isn't reflected in the word vectors obtained by the proposed method , the semantic similarity (or distance) between nouns defined with the word vectors is proper for an example-based disambiguation method.

1. はじめに

自然言語文の統語解析では，文 (用例) とその統語構造の対を多数用意し，入力文の統語構造を類似する用例の統語構造に従って解析することにより，統語的な曖昧さを絞り込むことができる [隅田 94, 冨浦 97] . また，多義語の意味の選択 [Niwa 94, 藤井 95]，機械翻訳における訳語の選択 [古瀬 94] などでも類似用例に従った処理が可能である . これらの用例主導の曖昧さ解消法では，入力文 (句) と用例の間の類似度や距離を計算をする必要があるが，その基本は単語間の類似度や距離計算である . このように，単語間の意味的な類似度や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 .

従来，最も一般的には，単語間の上位下位関係を記述したシソーラスを用いて，単語間の類似度や距離を求めていた . しかし，シソーラスは，単語や概念を上位-下位 (一般-特殊) の関係で階層付けしたもので，本来，単語間の類似度や距離を求めるためのものではない . しかも，基本的には人手で作成されるため，その階層付けも作成者の主観の影響が大きい . したがって，シソーラスに基づく単語間の類似度は，単語間の意味的な類似度の重要な候補の一つであるが，特定の問題に必要となる単語間の意味的な類似度として，最適とは限らない .

名詞 n が，どのような関係 f でどのような語 w に係り得るか (共起性) は，名詞 n の意味と深く関係している . 共起性だけで名詞の意味を完全に捉えられるわけではないが，少なくとも，類似した共起性を持つ名詞は意味的にも類似していることが期待できる . 本論文では，共起性で捉えられる名詞間の意味的な類似性を反映するように，名詞を多次元空間に配置する新たな手法を提案する . 用例主導の各種の曖昧さ解消法で必要となる名詞間の意味的な類似度や距離の多くは，共起性で捉えられる名詞間の意味的な類似度や距離が適していると考えられる . 提案手法で得られる名詞ベクトルは主に用例主導の各種の曖昧さ解消法で必要となる類似度や距離の計算に用いることを想定している .

本論文では，まず，提案する名詞ベクトルの獲得法について述べた後，名詞ベクトルを求める他の手法との関

係について述べ，最後に，名詞ベクトル獲得実験とその評価実験について報告する．

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．$\langle f,w\rangle$ の全体集合を S_G (S_G の要素数を ℓ)，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[*1] ($K \ll m$)．K 次元空間に名詞 n を配置したベクトルを名詞ベクトルと呼び $\boldsymbol{x}(n)$ で表す．

ここで，求める名詞ベクトルに対して，以下の制約を課しておく．

制約 **1** $\quad \boldsymbol{\mu} = E[\boldsymbol{x}(N)] = \mathbf{0}$，
制約 **2** $\quad E[(\boldsymbol{x}(N)-\boldsymbol{\mu})\,{}^t(\boldsymbol{x}(N)-\boldsymbol{\mu})] = I_K$

ただし，${}^tX$ は X の転置行列，I_K は K 次の単位行列である．また，$\boldsymbol{\mu}$ は名詞ベクトルの平均ベクトルで，

$$\boldsymbol{\mu} = E[\boldsymbol{x}(N)] = \sum_{n\in S_N} \boldsymbol{x}(n) f_N(n)$$

である ($f_N(n)$ は名詞 n の発生確率)．名詞間の類似度や距離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，どの方向の分散も等しいというのが制約 2 である．分散の大きさが 1 というのは本質的な制約ではなく，0 でない有限の値であれば構わない．

関係 f で語 w に係り得る，つまり，g ($=\langle f,w\rangle$) と共起し得る名詞は，ある共通の意味 (概念) を持っていると考えられる．したがって，名詞ベクトルが，共起性で捉えられる名詞間の意味的な類似性を反映するものであるならば，『g と共起する名詞は，多次元空間中で一つのグループを成す』と考えられる．グループを成すのであれば，g と共起する全名詞に対する名詞ベクトルの分散 (つまり，グループのまとまり具合) は比較的小さいと言える．ただし，g と共起する全名詞に対する名詞ベクトルの分散は，g と共起するという単独の性質では決まらない．なぜなら，一般に一つの名詞が複数の g と共起し得るため，g_1 と共起する全名詞に対する名詞ベクトルの分散を小さくすることと，g_2 と共起する全名詞の名詞ベクトルの分散を小さくすることとが，上記の制約 2 の下では相反する要請になる場合があるからである．そこで，グループ内分散 (g と共起し得る全名詞の名詞ベクトルの分散の各 g に対する平均)

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

を求めた名詞ベクトルの良さの尺度と考え，上記の制約条件の下でこれが最小になるように各名詞ベクトルを求める．ただし，$f_G(g)$ は g 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，$\boldsymbol{\mu}(g)$ は g と共起している全名詞に対する名詞ベクトルの平均ベクトル

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

である．

全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，

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

と，グループ内分散とグループ間分散 (上記最終式第 2 項) の和として表すことができる．これは，多変量解析や統計的パタン認識の分野で良く知られている関係である．

制約 2 より，

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

であるから，(1) 式より，制約 2 の下で，グループ内分散を最小にする名詞ベクトルは，グループ間分散を最大にすることがわかる．したがって，制約 1，2 の下で

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

を最大にする各名詞に対する名詞ベクトルが求める名詞ベクトルである．

3. 解　　法

$S_N = \{n_1, n_2, \cdots, n_m\}$ とし，

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

とする．

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

[*1] 以降，特に断わらない限りベクトルはすべて列ベクトルとする

であるから，

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

と表せる．計算の見通しを良くするために，

$$\Delta = \begin{bmatrix} f_N(n_1) & & \boldsymbol{O} \\ & f_N(n_2) & \\ & & \ddots \\ \boldsymbol{O} & & f_N(n_m) \end{bmatrix}$$

とおき，$Y = \Delta^{1/2}X$ 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（$\mathbf{1} = {}^t[1\,1\,\cdots\,1]$），制約条件 2 は，

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

となる．また，

$$A = \Delta^{-\frac{1}{2}}\left(\sum_{g\in S_G} f_G(g)\boldsymbol{f}(g)\,{}^t\boldsymbol{f}(g)\right)\Delta^{-\frac{1}{2}} \tag{5}$$

とおくと，目的関数は，以下のように表せる．

$$F = \mathrm{trace}\ {}^tYAY. \tag{6}$$

A は m 次実対称行列であるから，重根も含めると m 個の実数の固有値を持ち，直交行列 Φ，対角行列 Λ を用いて，$A = \Phi\Lambda\,{}^t\Phi$ と表現できる．ただし，$\lambda_1, \lambda_2, \cdots, \lambda_m$ を A の固有値，$\boldsymbol{e}_i$ を λ_i に属する大きさ 1 の固有ベクトルとすると，

$$\Lambda = \begin{bmatrix} \lambda_1 & & \boldsymbol{O} \\ & \lambda_2 & \\ & & \ddots \\ \boldsymbol{O} & & \lambda_m \end{bmatrix},$$
$$\Phi = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_m]$$

である．(i,k) 成分が

$$\sqrt{f_{N|G}(n_i|g_k)f_{G|N}(g_k|n_i)}$$

である $(m\times\ell)$ 行列 B（2 章冒頭で述べたように，ℓ は S_G の要素数，m は S_N の要素数）を用いて，A は，

$$A = B\,{}^tB$$

と表せるので，

$$\lambda_i \geq 0 \qquad (i = 1,2,\cdots,m) \tag{7}$$

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，(5) 式より，

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

であるから，A は固有値として 1 を持ち，1 に属する大きさ 1 の固有ベクトルが $\Delta^{1/2}\mathbf{1}$ であることが分かる．$\lambda_m = 1$，$\boldsymbol{e}_m = \Delta^{1/2}\mathbf{1}$ とし，

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m-1} \tag{8}$$

とする．任意の m 次元ベクトルは，$\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_m$ の線形和で表現できるので，

$$Y = \Phi\Gamma \tag{9}$$

と表せる．ただし，Γ は $(m\times K)$ 行列で，(i,j) 要素を c_{ij} と表記する．${}^t\Phi\Phi = I_m$ および $\boldsymbol{e}_m = \Delta^{1/2}\mathbf{1}$ であるから，(9) を (3) 式に代入して，

$$c_{mj} = 0 \qquad (j = 1,2,\cdots,K) \tag{10}$$

を得る．また，(9) を (4) 式に代入すると，${}^t\Gamma\Gamma = I_K$ であるから，

$$\sum_{j=1}^{K} {c_{ij}}^2 \leq 1 \qquad (i = 1,2,\cdots,m), \tag{11}$$
$$\sum_{i=1}^{m}\sum_{j=1}^{K} {c_{ij}}^2 = K \tag{12}$$

を得る．

したがって，(9) を (6) 式に代入すると，

$$F = \mathrm{trace}\ {}^tYA\,Y$$
$$= \mathrm{trace}\ {}^t\Gamma\,{}^t\Phi A\Phi\Gamma$$
$$= \mathrm{trace}\ {}^t\Gamma\Lambda\Gamma$$
$$= \sum_{j=1}^{K}\sum_{i=1}^{m}\lambda_i {c_{ij}}^2$$
$$= \sum_{i=1}^{m-1}\lambda_i\sum_{j=1}^{K} {c_{ij}}^2 \qquad (\because\ (10)\,)$$
$$\leq \sum_{i=1}^{K}\lambda_i \qquad (\because\ (8)(11)(12)\,) \tag{13}$$

となる．等号成立条件は，

$$\sum_{j=1}^{K} {c_{ij}}^2 = 1 \qquad (i = 1,2,\cdots,K)$$

である．

したがって，制約 1 および 2 を満たす X，つまり，(3)(4) 式を満たす Y で，F の最大値 $\lambda_1+\lambda_2+\cdots+\lambda_K$ を実現する Y の一つ[*2]は，

$$Y=\Phi\begin{bmatrix} I_K \\ \cdots \\ \boldsymbol{O} \end{bmatrix}=[\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_K] \qquad (14)$$

であり，X は Y より，

$$X=\Delta^{-\frac{1}{2}}Y \qquad (15)$$

として求まる[*3]．

なお，実際には，観測された共起データ D から，f_N，f_G，$f_{N|G}$ を最尤推定で求め，これから上記にしたがって単語ベクトルを求める．

4. 関 連 研 究

共起データや辞書の定義文における参照関係を元に，単語をその意味的特徴に応じた実ベクトル (単語ベクトル[*4]) に割り当てる手法がいくつか提案されている．たとえば，以下のものがある．

- 自己組織化神経回路網モデルを利用して 2 次元に圧縮するもの [馬 01]，
- 主成分分析を利用して次元を圧縮するもの [小嶋 95]，
- 特異値分解を利用して次元を圧縮するもの [笠原 02, 佐々木 03]．

上記 3 つの手法に共通するのは，何らかの情報に基づいて，単語の特徴ベクトルを構成し，この次元を圧縮することで単語ベクトルを求める点である．これに対して，提案手法は，最初に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．

[馬 01] の手法では，自己組織化神経回路網モデル SOM を用いて，特徴ベクトルを 2 次元空間に配置する．固有値問題に帰着される提案手法と異なり，自己組織化に基づく手法であり，しかも，2 次元空間上への配置という可視的な表現に重点をおいたもので，本研究とは目的が異なる．

主成分分析により次元を圧縮する [小嶋 95] の手法は，提案手法と同じく固有値問題に帰着される．他の手法が共起頻度や辞書の定義文における参照頻度の関数として，特徴ベクトルを構成しているのに対し，[小嶋 95] では，辞書の定義文における参照関係を意味ネットワークと捉え，活性伝搬の結果を用いて特徴ベクトルを構成している．このため提案手法との直接的な対応関係はない．

特異値分解により次元を圧縮する手法も，提案手法と同じく固有値問題に帰着される．$m\times\ell$ 行列 W は，

$$W=U\Sigma\,{}^tV$$

と特異値分解できる．ここで，Σ は $W\,{}^tW$ の正の固有値 $\lambda_1,\lambda_2,\cdots,\lambda_r$ ($r=\mathrm{rank}\ W$) の平方根 $\sigma_1,\sigma_2,\cdots,\sigma_r$ (W の特異値と呼ばれる) を成分とする対角行列で，U の第 i 列は λ_i に属する大きさ 1 の $W\,{}^tW$ の固有ベクトルである．また，${}^tWW$ の正の固有値も $\lambda_1,\lambda_2,\cdots,\lambda_r$ であり，V の第 i 列は λ_i に属する大きさ 1 の ${}^tWW$ の固有ベクトルである．$\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_r$ とし，Σ_K を $\sigma_1,\sigma_2,\cdots,\sigma_K$ ($K\leq r$) を成分とする K 次対角行列，U_K を U の第 1 列から第 K 列までから成る行列，V_K を V の第 1 列から第 K 列までから成る行列とすると，

$$Z=U_K\Sigma_K\,{}^tV_K$$

は階数 (rank) が K 以下の $(m\times\ell)$ 行列の内で最も良い W の近似であることが知られている．今，W が単語の特徴行列 (第 i 行を単語 w_i の特徴ベクトルとする行列) であるとする．単語間の類似度として単語の特徴ベクトルの余弦を用いたとすると，$W\,{}^tW$ が分かっていれば単語間の類似度が求まる．W を Z で近似すると，

$$\begin{aligned} W\,{}^tW &\simeq (U_K\Sigma_K\,{}^tV_K)\,{}^t(U_K\Sigma_K\,{}^tV_K) \\ &= (U_K\Sigma_K)\,{}^t(U_K\Sigma_K) \end{aligned}$$

である．このことを利用し，単語の属性行列 (第 i 行を単語 w_i の単語ベクトルとする行列) を $U_K\Sigma_K$ に取れば，単語ベクトルは K 次元に圧縮できることになる[*5]．

根拠は示されていないが，[笠原 02] では，求める属性行列の各列ベクトルが正規直交系となることを重視し，$U_K\Sigma_K$ ではなく，U_K を単語の属性行列としている．この場合，以下に示すように提案手法とかなり関係が深い．提案手法では，まず，(5) 式で定義される行列 A の固有ベクトルから，(14) 式の Y を求める．3 章で示したように，(i,k) 成分が

$$\sqrt{f_{N|G}(n_i|g_k)f_{G|N}(g_k|n_i)}$$

である $(m\times\ell)$ 行列 B を用いて，A は

$$A=B\,{}^tB$$

と表せる．B を上記の単語の特徴行列 W と考えると[*6]，Y は [笠原 02] の U_K に相当する．ただし，提案手法では，

- Y には，A の固有値 1 に属する固有ベクトル $\Delta^{1/2}\mathbf{1}$ は含まれない．
- Y から，さらに (15) 式の変換により求めたものが属性行列 X である．

---

*2 (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．

*3 $r=\mathrm{rank}\ A$ とすると，$\lambda_r=\lambda_{r+1}=\lambda_{m-1}=0$ であるから，$K\leq r-1$ でなければ意味がない．

*4 研究者によっては属性ベクトルとも呼ぶ．

*5 [佐々木 03] では，W の行と列の取り方が逆であるため，得られる属性行列を $\Sigma_K\,{}^tV_K$ としている．

*6 $\sqrt{f_{N|G}(n_i|g_k)f_{G|N}(g_k|n_i)}$ が高い g_k は n_i を良く特徴付けると考えられる．

という点が異なる．なお，[笠原 02] では，シソーラスモデルによる次元の圧縮と特異値分解による圧縮を併用することが主眼にある．

単語ベクトルを求める手法ではないが，共起データから名詞間の類似度を求めるものとして，[Hindle 90] がある．5･2･2 節で述べるように，この手法では共起の相互情報量に基づいて直接名詞間の類似度を定義している．したがって，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，このベクトルを元に距離や類似度を求める提案手法は，[Hindle 90] とは異なるアプローチである．

これらの手法および提案手法とも，単語の意味的特徴を表す元となる情報として，共起データや辞書の定義文における参照関係を用いている．語の共起はその語の意味的性質を良く表していると考えられる．しかし，低頻度の語に関しては，当然のことながら十分な情報となり得ない．これは，本論文の 5･2･1 節の実験にも表れている．一方，辞書における参照関係を利用する場合は，網羅的に語の情報が得られる．しかし，定義文に出現する参照語の（定義文における）役割などを考慮していないため，その点では，粗い情報である．

単語を実ベクトルに対応させる方法は色々考えられ，その元となる情報も様々である．得られた単語ベクトルの有効性は，元となる情報の種類と規模，得られた単語ベクトルの利用法に依存するため，一概に手法の優劣は判断できない．

5. 実　　験

5･1 名詞ベクトル獲得実験

EDR 電子化辞書の日本語コーパス（JCO-V020E）[*7] から，5･2･2 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〉の共起を抽出，共起の列 D_0 を作成した．これから，

$$D = \{\langle n,\langle c,v\rangle\rangle \in D_0 \mid freq(n\,;D)\geq 2,\ freq(\langle c,v\rangle\,;D)\geq 2\}$$

を満足し，かつ，要素数最大の D を求め，これを f_N，f_G，$f_{N|G}$ を推定するための共起データとした[*8]．ただし，$freq(n\,;D)$，$freq(\langle c,v\rangle\,;D)$ は，それぞれ，D における n の出現頻度，$\langle c,v\rangle$ の出現頻度である．D の共起の総数は 213663，異なり数は 162589 であり，D 中の名詞の異なり数は 16543，助詞・動詞の組の異なり数は 14474 である．

D により，推定される f_N，f_G，$f_{N|G}$ を用いて，3 章で述べた手法により名詞ベクトルを獲得した．名詞ベクトル獲得実験は，当大学情報基盤センターの研究用計算機システム，富士通 VPP 5000/64 上で行ない，固有値・固有ベクトルは，数値計算サブルーチンライブラリ SSL II を用いて，(5) 式の行列 A を平衡化し，さらにハウスホルダー法でヘッセンベルグ行列に変換したものに 2 段 QR 法を適用して求めた．なおこの手法は，固有値が重根あるいは近接根となる場合においても，比較的精度良く固有値・固有ベクトルを求めることができる．

D 中における頻度が 10 以上の名詞（3908 個）間の距離[*9]の小さいもの上位 20 組，D 中における頻度が 40 以上の名詞（1042 個）間の距離の小さいもの上位 20 組を表 1 に挙げる．括弧の中の数値はその名詞の D 中での頻度である．

5･2 評　価　実　験

名詞 n と n' が同義語の場合，5･1 節で得られた名詞ベクトル $\boldsymbol{x}(n)$ と $\boldsymbol{x}(n')$ の距離は小さいことが望ましい．そこで，同義語間の距離の分布を調査する実験を行なった．対象とする名詞ベクトルは，後述の 5･2･2 節での実験結果が最も良かった $K=120$ の名詞ベクトルとした．

上記の実験では，近くに配置されるべき名詞同士が近くに配置されているかを調べることはできるが，遠くに配置されるべき名詞同士がそのように配置されているかは調べることができない．また，さらに一般的に，

$$|\boldsymbol{x}(n)-\boldsymbol{x}(n_1)| < |\boldsymbol{x}(n)-\boldsymbol{x}(n_2)|$$

のとき，名詞 n_1 と n_2 を比較すると，n_1 の方が n に類似しているのかを人間の内省に基づいて調査することは非常に困難である．そこで，提案手法により得られた名詞ベクトルに基づく類似度を用例主導の曖昧さ解消に用いた実験を行ない，提案手法の有効性を評価した．

§1 同義語間の距離の分布の調査

対象とする名詞の全体集合 S_N に対して，共起データ D 中での頻度 ζ（$\zeta=2,10,20,30,40$）で制限した名詞対集合 $P(\zeta)$

$$P(\zeta)=\{(n,n')\in S_N^2 \mid freq(n\,;D)\geq\zeta,\ freq(n'\,;D)\geq\zeta,\ n\neq n'\}$$

を作成した．さらに，EDR 電子化辞書の日本語単語辞書 JWD-V020[*10] を用いて，以下の完全同義語対集合 $P_a(\zeta)$，部分的同義語対集合 $P_p(\zeta)$ を作成した．

$$P_a(\zeta)=\{(n,n')\in P(\zeta)\mid \mathcal{M}(n)=\mathcal{M}(n')\},$$
$$P_p(\zeta)=\{(n,n')\in P(\zeta)\mid \mathcal{M}(n)\cap\mathcal{M}(n')\neq\emptyset\}$$

---

*7 新聞，雑誌，書籍，教科書から収集された約 20 万文からなるテキストデータベースで，各文に，形態素情報，構文情報，意味構造（依存構造）が付与されている．共起データはこれらの情報を利用して自動抽出した．

*8 名詞 n が特定の一つの $\langle c,v\rangle$ とのみ共起し得る名詞だからではなく，たまたま，n の D_0 中での頻度が 1 であることから，D_0 中では「n と共起しているのは $\langle c,v\rangle$ のみである」となることを避けるために，頻度 2 以上に制限した．

*9 求まった名詞ベクトルにより，名詞 n_1，n_2 が K 次元空間上で $\boldsymbol{x}(n_1)$，$\boldsymbol{x}(n_2)$ に配置されるとする．本論文では，$\boldsymbol{x}(n_1)$ と $\boldsymbol{x}(n_2)$ とのユークリッド距離 $|\boldsymbol{x}(n_1)-\boldsymbol{x}(n_2)|$ を単に名詞 n_1 と名詞 n_2 との距離と呼ぶ．

*10 日本語単語辞書は，日本語の各単語に対して，表記などの見出し情報，文法情報，意味情報，用法などの補助情報を記述したものである．意味情報は見出し語が持つ概念（単語概念）の識別子でその語の直接の上位概念である．

表 1 名詞間の距離

頻度 10 以上

距離	名詞 1	名詞 2
0.156	開発 (378)	整備 (84)
0.164	引き渡し (13)	停戦 (19)
0.175	出荷 (54)	販売 (419)
0.182	引き渡し (13)	撤回 (21)
0.187	見直し (67)	撤退 (27)
0.192	強化 (124)	撤退 (27)
0.195	引き渡し (13)	提出 (20)
0.202	停戦 (19)	撤回 (21)
0.204	読出し (10)	入出力 (18)
0.207	開発 (378)	実用化 (26)
0.209	引き渡し (13)	設立 (34)
0.211	交渉 (190)	折衝 (30)
0.214	構造 (194)	性質 (140)
0.214	づくり (45)	開発 (378)
0.217	削減 (76)	転換 (51)
0.220	テスト (42)	査察 (19)
0.221	コンピューター化 (37)	拡充 (24)
0.222	ネットワーク化 (19)	実用化 (26)
0.223	現地生産 (12)	発売 (29)
0.223	増強 (19)	普及 (52)

頻度 40 以上

距離	名詞 1	名詞 2
0.156	開発 (378)	整備 (84)
0.175	出荷 (54)	販売 (419)
0.214	構造 (194)	性質 (140)
0.214	づくり (45)	開発 (378)
0.217	削減 (76)	転換 (51)
0.225	自由化 (72)	普及 (52)
0.225	改善 (89)	自由化 (72)
0.225	あり方 (52)	政策 (215)
0.228	活動 (255)	研究 (254)
0.234	改善 (89)	普及 (52)
0.235	西独 (80)	日本 (1142)
0.235	形状 (43)	集合 (56)
0.236	方式 (253)	方法 (387)
0.240	研究 (254)	作業 (237)
0.242	開発 (378)	構築 (59)
0.243	傾向 (161)	動き (382)
0.244	強化 (124)	見直し (67)
0.246	改革 (159)	活動 (255)
0.246	構築 (59)	整備 (84)
0.252	意 (41)	意向 (94)

表 2 同義語間の距離の分布

	$P(\zeta)$		$P_a(\zeta)$				$P_p(\zeta)$			
ζ	平均	分散	総数	平均	分散	95%点	総数	平均	分散	95%点
2	6.55	978.28	3842	7.61	603.31	14.08	9280	10.53	1730.25	15.54
10	4.70	67.48	146	1.67	2.46	3.12	1109	2.82	25.62	6.06
20	3.82	32.02	46	1.63	6.52	2.27	612	2.37	3.94	5.99
30	3.46	20.08	28	1.13	0.25	1.88	399	2.27	3.54	6.14
40	3.13	10.27	25	1.13	0.26	1.88	293	2.15	2.13	5.99

ただし，$\mathcal{M}(n)$ は上記日本語単語辞書で記述されている n が持つ概念 (つまり，直接の上位概念) の集合である．

提案手法により獲得される名詞ベクトルの内，D 上の出現頻度の低い名詞に対するベクトルは，たまたま出現した共起の影響が色濃く表れ，質が良くないと予想される．D 上の出現頻度ごとに完全同義語対集合および部分的同義語対集合を考えるのは，低頻度語の名詞ベクトルの質を調査するためである．

$\zeta(= 2, 10, 20, 30, 40$) ごとに，$P(\zeta)$，$P_a(\zeta)$ および $P_p(\zeta)$ の名詞対の距離の平均，分散を調べた．さらに，$P_a(\zeta)$，$P_p(\zeta)$ に関しては，それぞれの集合の名詞対を距離の昇順で並べたときの 95%番目の名詞対の距離 (95%点) も求めた．結果を表 2 に示す．

$P_a(2)$ と $P_p(2)$ 中の名詞対の距離の平均は，$P(2)$ 中の名詞対の距離の平均より大きく，予想通り，共起データ D において低頻度の名詞に対しては，意味的な類似性を表すような名詞ベクトルは獲得されていないことが読み取れる．たとえば『美女』と『美人』はともに D における頻度は 2 であった．この二つの名詞は完全同義語対であるが，5･1 節の実験で獲得された名詞ベクトルを用いた距離は，154.98 と非常に大きい．一方，頻度が 10 以上の場合は，シソーラス上で同義語である名詞対の距離は，比較的小さく，意味的に類似する名詞は近くに配置されていることがわかる．

§ 2 用例主導の統語的曖昧さ解消実験

名詞間の類似度 (あるいは距離) と共起データ D に基づいて，名詞 n が助詞 c で動詞 v に係る係り易さの程度 ($\langle n, \langle c, v\rangle\rangle$ の共起性と呼び，$C(n, c, v)$ で表す) を推定し，これを利用して，後置詞句の係り先の判定を行なうことを考える．これは，[隅田 94] で提案されている手法と基本的には同じ手法である．

対象とする文 (統語的曖昧さを持つ文) は，

$$n \quad cs \quad \gamma \quad v_1 \quad \delta \quad v_2 \tag{16}$$

という形態の文である．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．cs 中の助詞の内，n の係りの種類を規定する助詞を c とする (cs 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．

EDR 電子化辞書の日本語コーパス (JCO-V020E) から (16) の形態の文のうち，n，$\langle c, v_1\rangle$，$\langle c, v_2\rangle$ が全て D

表 3　抽出したデータの内訳

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	1341	1285 (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%)
合計			2488	1787 (71.8%)	701 (28.2%)

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 は $1/0$ の 2 値で , h は以下の (a) に関する情報 , d は以下の (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 3 に示す .

係り先を決める要因は色々と報告されている [内元 99] . 本実験では , 共起性以外に ,

(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 ,

という要因を考えた . cs に『は』が含まれる場合 ,「$n\ c$」は v_2 に係る傾向が非常に強く , また , n の次の自立語が v_1 である場合 ,「$n\ c$」は v_1 に係る傾向が非常に強い . このことは , 表 3 の内訳にも顕著に表れている . このような表層的な手がかりがある場合は , 共起性よりもむしろ表層的な手がかりに基づいて一意に係り先を判定する方が精度が良い . 今回の実験では , このような表層的な手がかりがない文 (表 3 の type1 の文) を対象として , $\langle n, \langle c, v_1\rangle\rangle$ と $\langle n, \langle c, v_2\rangle\rangle$ の共起性を利用した「$n\ c$」の係り先の判定実験を行なった . なお ,「$n\ c$」の正しい係り先を v , 他方を v' とすると , type1 の文 764 のうち , $\langle n, \langle c, v\rangle\rangle \in D$ なる文は 194 文 (25.4%) , $\langle n, \langle c, v'\rangle\rangle \in D$ なる文は 70 文 (9.2%) , $\langle n, \langle c, v\rangle\rangle \in D$ かつ $\langle n, \langle c, v'\rangle\rangle \in D$ なる文は 29 文 (3.8%) であった .

基本的には「$n\ c$」は係り易い方 , つまり , 共起性が高い動詞の方に係ると考える . しかし , 表 3 から分かるように , 表層的手がかりがない type 1 の文でも若干 v_1 に係る傾向が強い . そこで , この傾向を捉えるために , 適当なスレッシュホールド θ を設定し ($0 < \theta < 1$) ,「$n\ c$」の係り先を , 共起性に基づいて以下のように判定する .

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

共起性 $C(n,c,v)$ を名詞間の類似度と共起データ D を用いて定義することを考えよう . 一般に ,

> 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 .

と期待できる . D 中の共起の内 , $\langle c, v\rangle$ との共起を列挙したものを

$$\langle n_{t_1}, \langle c, v\rangle\rangle, \langle n_{t_2}, \langle c, v\rangle\rangle, \cdots, \langle n_{t_u}, \langle c, v\rangle\rangle$$

とする . ただし , 重複を許す (つまり , $t_i = t_j$ $(i \neq j)$ もあり得る) . 次に , 上記の列に対応して , 類似度の列

$$sim(n, n_{t_1}), sim(n, n_{t_2}), \cdots, sim(n, n_{t_u})$$

を求める ($sim(n, n')$ は名詞 n と n' の類似度) . この列中の値の内 , k 番目に大きな値と $\langle n, \langle c, v\rangle\rangle$ 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 $C(n,c,v)$ と定義する[*11] .

名詞間の類似度 sim としては , 5･1 節で得られた名詞ベクトルに基づく類似度 sim_d , シソーラスを用いた類似度 sim_1 , sim_2 , および提案手法と同じく共起データを用いる他手法として , [Hindle 90] の類似度 sim_{H1} とその修正版 sim_{H2} を試し , 比較実験を行なった . シソーラスとしては , EDR 電子化辞書の日本語単語辞書 JWD-V020 および概念辞書 CPD-V020.1 (概念体系辞書) を用いた[*12] .

提案手法により得られた単語ベクトルに基づく類似度 sim_d は

$$sim_d(n_1, n_2) = e^{-|\boldsymbol{x}(n_1) - \boldsymbol{x}(n_2)|}$$

である . 名詞ベクトルを用いた類似度の定義を [笠原 02] のように , 二つのベクトルの余弦とするのも一般的である . しかし , 提案手法で獲得された名詞ベクトルでは , 共起の傾向が似ている名詞同士が近くに配置されるため , 類似度は距離の減少関数とすべきである . そこで , sim_d を上記のように定義した .

シソーラスを用いた類似度では , シソーラスを有向グラフと考え , グラフ上での単語 n_1 と n_2 の位置関係に

---

*11　1 番目ではなく , k 番目とするのは , n と非常に類似する名詞 n' と $\langle c, v\rangle$ との共起が 1 つだけ D 中で観測され , $\langle c, v\rangle$ との共起が観測されている他の名詞はどれも n との類似度が低いという場合に $C(n,c,v)$ が高くなることを防ぐためである . k 番目までの類似度の平均とするのも一つの考え方であり , また , より一般的に , k 番目の類似度 (あるいは k 番目までの類似度の平均) のある増加関数として $C(n,c,v)$ を定義することも考えらる . 本実験では最も単純な定義を選んだ .

*12　概念体系辞書は , 概念 (JWD-V020 に記載の単語概念も含む) 間の上位下位関係を記述したものであり , 多重継承 (1 つの概念の直接の上位概念が複数存在すること) を許している . 日本語単語辞書 JWD-V020 および概念辞書 CPD-V020.1 (概念体系辞書) から , 5･1 節の実験対象の名詞 16543 個に関する部分シソーラスを作成した . このシソーラスでは ,

単語概念数 (異なり数)	33755
全概念数 (異なり数)	38104
単語-単語概念の関係数	43038
下位概念-上位概念の関係数	40048
最大深さ	17
平均の深さ	9

であった .

表 4 曖昧さ解消の正解率

用いた類似度		正解率 (%)					
		type1			全体		
		$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
sim_d	(次元 $K=10$)	62.4	62.8	63.1	85.1	85.2	85.3
〃	(〃 $K=20$)	65.8	65.2	63.6	86.1	85.9	85.5
〃	(〃 $K=40$)	68.1	66.4	66.0	86.8	86.3	86.2
〃	(〃 $K=60$)	68.7	68.5	66.9	87.0	86.9	86.5
〃	(〃 $K=80$)	69.4	67.9	66.1	87.2	86.8	86.2
〃	(〃 $K=100$)	69.6	67.1	65.6	87.3	86.5	86.1
〃	(〃 $K=120$)	70.3	66.5	66.1	87.5	86.3	86.2
〃	(〃 $K=140$)	69.2	66.4	66.6	87.2	86.3	86.4
sim_1		66.4	67.9	65.4	86.3	86.8	86.0
sim_2		66.5	67.5	65.4	86.3	86.7	86.0
sim_{H1}		66.9	67.4	66.9	86.5	86.6	86.5
sim_{H2}		66.8	67.8	66.5	86.4	86.7	86.3

基づいて，以下のように類似度を定義する．

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

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

ここで，$CM(n_1,n_2)$ はシソーラス上での名詞 n_1 と n_2 の共通の上位概念のノード (直接の上位概念以外も含む) 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．sim_2 は基本的には [長尾 96] で紹介されている類似度である．今回用いたシソーラスでは語の多義性や概念の多重継承のために，シソーラス上の一つのノードの直接の上位ノードが複数存在する場合がある．このことを考慮し，sim_1，sim_2 では max を取っている．

[Hindle 90] に基づく類似度は，以下のように定義される[*13]．

$$sim_{H1} = \sum_c \sum_v SIM_c(v,n_1,n_2).$$

ここで，$SIM_c(v,n_1,n_2)$ は，助詞 c で v に係るという共起性で見た場合の名詞 n_1 と n_2 の類似度で，以下のように定義される．

$$SIM_c(v,n_1,n_2) = \begin{cases} min(|S_c(n_1,v)|,|S_c(n_2,v)|) \\ \quad ;\quad 0 < S_c(n_1,v)\cdot S_c(n_2,v) < \infty \\ 0 \quad ;\quad otherwise \end{cases},$$

$$S_c(n,v) = \log_2 \frac{f(n,v|c)}{f(n|c)\,f(v|c)}.$$

$f(n,v|c)$ は『c で係っている場合に n が v に係る条件付確率』，$f(n|c)$ は『c で係っている場合に，係っている名詞が n である条件付確率』，$f(v|c)$ は『c で係っている場合に，係られている動詞が v である条件付確率』である．これらは共起データ D より推定する．また，[Hindle 90] では共起の関係 subject と object を同等に扱って類似度を定義しているが，今回の共起データでは共起の関係 (助詞) の頻度分布には偏りがあるため，以下の修正版の類似度 sim_{H2} も試した．

$$sim_{H2} = \sum_c f(c) \sum_v SIM_c(v,n_1,n_2).$$

ただし，$f(c)$ は助詞 c の発生確率であり，これも共起データより推定する．

係り先判定システムのパラメタは θ 一つであり，実験対象数は 764 と十分に大きいため，オープンテストでの結果とクローズドテストでの結果は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 θ を 0.025 刻みで最適に調節した場合の評価用データに対する正解率による評価) とした．実験結果を表 4 に示す．表において，K は名詞ベクトルの次元，k は類似度を用いた共起性の定義における k 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 に係ると判定し，type1 のみ前述の手法で判定した場合の全体の正解率を「全体」として示している．

獲得した名詞ベクトルに基づく類似度 sim_d では，名詞ベクトルの次元 K が 120，共起性の定義に用いられる k が 1 のとき，最高の正解率で，type1 の文に対しては 70.3% (全体では 87.5%) であった．これに対して，他の類似度では，sim_1 が k が 2 のとき最高の正解率で，type1 の文に対しては 67.9% (全体では 86.8%) であった．このように，若干ではあるが，提案手法により獲得した名詞ベクトルに基づく類似度 sim_d の方が良い結果を示している．

*13 英語の名詞を対象とする [Hindle 90] では，$c \in \{\text{subject},\text{object}\}$ である．

また，同義語間の距離の分布の調査からわかるように，本手法により獲得された名詞ベクトルは，低頻度の名詞に対しては質が良くない．しかし，低頻度の語が実際の応用で処理の対象になることはまれであるため，本実験結果が示すように，その影響は小さいと考えられる．

今回の評価実験で用いた共起性の定義が唯一のものではない．また，シソーラスに基づく類似度の定義や，利用するシソーラスも他のものが考えられる．したがって，今回の評価実験により，提案手法の有効性を十分に検証できたとは考えていない．今後，大規模な共起データに基づいて名詞ベクトルを求め，様々な用例主導の曖昧さ解消に適用して，その有効性を検証する必要がある．

6. ま と め

共起データに基づき，名詞を多次元空間へ配置する手法を提案した．従来研究が，何らかの情報に基づいて，名詞の特徴ベクトルを構成し，この次元を圧縮することで単語ベクトルを求めているのに対して，提案手法は，最初に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あり，異なる視点からのアプローチと言える．

〈 名詞,〈 助詞,動詞 〉〉 の共起データに基づいて，提案手法による名詞ベクトルの獲得実験を行なった．さらに，得られた名詞ベクトルを用いた二種類の実験・調査

- 同義語間の距離の分布の調査，
- 用例主導の統語的曖昧さ解消実験，

により，提案手法の評価を行なった．

今回の名詞ベクトルの獲得実験では，動詞との共起データのみを用いて行なったが，提案手法は，元々名詞と様々な品詞の語との共起データを用いることができる．〈 名詞,〈 助詞,動詞 〉〉 以外に，〈 名詞,〈 が,形容詞 〉〉，〈 名詞,〈 が,形容動詞 〉〉，〈 名詞,〈 の,名詞 〉〉 などの共起を用いるならば，より質の良い名詞ベクトルが獲得できるものと考えられる．

今後，共起データを大規模にして，獲得される名詞ベクトルの質を上げるとともに，[笠原 02] のようにシソーラスとの併用も考慮に入れた手法を検討する必要もある．

◇ 参 考 文 献 ◇

[藤井 95] 藤井 敦, 秋山 典丈, 徳永 健伸, 田中 穂積：動詞の多義性解消における格の弁別能力と集中度の有効性について, 言語処理学会第 1 回年次大会, pp. 117–120 (1995)

[古瀬 94] 古瀬 蔵, 隅田 英一郎, 飯田 仁：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)

[Hindle 90] Hindle, D.: Noun Classification from Predicate-Argument Structures, in *Proceedings of the 28th Annual Meeting of ACL*, pp. 268–275 (1990)

[笠原 02] 笠原 要, 稲子 希望, 加藤 恒昭：単語の属性空間の表現方法, 人工知能学会論文誌, Vol. 17, No. 5, pp. 539–547 (2002)

[小嶋 95] 小嶋 秀樹, 伊藤 昭：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)

[馬 01] 馬 青, 神崎 享子, 村田 真樹, 内元 清貴, 井佐原 均：日本語名詞の意味マップの自己組織化, 情報処理学会論文誌, Vol. 42, No. 10, pp. 2379–2391 (2001)

[長尾 96] 長尾 真：岩波講座ソフトウェア科学 15 自然言語処理, 岩波書店 (1996)

[Niwa 94] Niwa, Y. and Nitta, Y.: Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries, in *Proceedings of the International Conference on Computational Linguistics(COLING-94)*, pp. 304–309 (1994)

[佐々木 03] 佐々木 稔, 新納 浩幸：単語クラスタリングの語義判別問題への応用, 情報処理学会研究報告 NL-154, pp. 145–152 (2003)

[隅田 94] 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)

[冨浦 97] 冨浦 洋一, 日高 達：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)

[内元 99] 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)

〔担当委員：堀 浩一〕

2003 年 5 月 9 日 受理

著 者 紹 介

冨浦 洋一（正会員）

1984 年九州大学工学部電子工学科卒業．1989 年同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，1995 年同助教授，1996 年同大学大学院システム情報科学研究科助教授，2000 年同大学大学院システム情報科学研究院助教授，現在に至る．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．情報処理学会，言語処理学会各会員．

田中 省作

1995 年岡山大学工学部情報工学科卒業．2000 年九州大学大学院システム情報科学研究科知能システム学専攻博士課程修了．同年九州大学大学院システム情報科学研究院助手，2001 年同大学情報基盤センター助手，現在に至る．工学博士．自然言語処理，自然言語処理技術の言語教育への応用に関する研究に従事．情報処理学会，言語処理学会，英語コーパス学会各会員．

日高 達

1965 年九州大学工学部電子工学科卒業．1969 年同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，1973 年同講師，1980 年同助教授，1988 年同教授，1996 年同大学大学院システム情報科学研究科教授，2000 年同大学大学院システム情報科学研究院教授，2003 年退官．工学博士．情報処理学会会員．