情報処理学会論文誌 Vol. 49 No. 12 3997–4006 (Dec. 2008)

単語・意味属性間共起に基づく コーパス概念ベースの生成方式

別所 克人[†1] 内山 俊郎[†1] 内山 匡[†1]
片岡 良治[†1] 奥 雅博[†1]

コーパス中の単語間の共起頻度に基づいて生成する概念ベースは，単語の意味表現としての概念ベクトルを提供するが，単語間共起には単語間の関連性やメモリ使用量の制約に起因する品質の問題がある．本論文では，メモリ使用量を増やすことなく，より高品質な概念ベースを生成する手法を提案する．具体的には，高頻度の単語と単語意味属性とのコーパス中における共起頻度に基づき概念ベースを生成し，次に低頻度語の概念ベクトルを，概念ベクトルの分布の分散最小性に基づき推定し，概念ベースを拡張することにより実現する．生成した概念ベクトルを用いた文書検索の評価実験により，従来の単語間共起に基づいて生成した概念ベースと比較して，検索精度が向上することを検証した．

A Method for Generating Corpus-concept-base Based on Co-occurrences between Words and Semantic Attributes

Katsuji Bessho,[†1] Toshio Uchiyama,[†1]
Tadasu Uchiyama,[†1] Ryoji Kataoka[†1]
and Masahiro Oku[†1]

A concept base generated based on co-occurrence frequencies between words in a corpus provides the concept vectors, which are the semantic representations of words. Co-occurrence between words has a problem of quality due to relationships between words and restriction by amount of memory use. This paper proposes a method generating a concept base of good quality without increasing the amount of memory use. In detail, this method generates a concept base for high frequent words based on co-occurrence frequencies between words and those semantic attributes in a corpus, and then it estimates the concept vectors for low frequent words based on minimum variance criterion of the distribution of concept vectors. The experimental results of document retrieval using the generated concept vectors showed that the proposed method can improve retrieval accuracy than the conventional method using the concept base based on co-occurrence frequencies between words.

1. はじめに

単語間の意味的類似性を定量化するための技術として概念ベース技術がある．概念ベースは，単語とその意味表現である概念ベクトルの対の集合を格納したものである．単語間の類似性は，対応する概念ベクトル間の類似性により定量化される．概念ベースの生成手法としては，文献 1) における単語間共起に基づく手法がある．この手法では，各行・各列がそれぞれ 1 単語に対応している単語間共起行列をとる．各行に対応する単語が概念ベースに登録する単語であり，行列の各成分は，対応する単語間のコーパスにおける共起頻度である．行に対応する単語の対に対し，それらの単語の意味が近ければ，それらの行ベクトルも近く，行ベクトルを，対応する単語の概念ととらえることができるというのが基本的な考えである．この共起行列に対し特異値分解を行い，列数の縮退した行列に変換する．変換後の行列の各行ベクトルが，対応する単語の概念ベクトルである．概念ベースは，情報検索[2)–4)]や，テキストセグメンテーション[5)]等に適用され，効果をもたらしてきた．

このような単語間共起に基づく概念ベースは有用ではあるものの，依然として以下の課題がある．

まず，単語間共起行列の列に対応する単語の中には意味の近いものがあり，そういった単語との共起を別々にカウントしているため，行ベクトルが，対応する単語の概念を十分的確には表現できないという問題がある．

また，概念ベース生成における特異値分解処理は，共起行列の行数と列数の積のオーダのメモリ使用量をとる．数百 GB のメモリであっても，行数，列数が 20 万ほどに限定される．比較的低価格で取得できるメモリの範囲内で計算可能とするためには，特異値分解の対象となる共起行列の行および列を，高頻度語に対応するものだけに制限する必要がある．制限した場合，生成される概念ベースの品質は低下する．

これらのことから，メモリ使用量を増やすことなく，より高品質な概念ベースを生成する技術が必要となる．

†1 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

文献 6) においては，単語・意味属性間共起の属性行列に対し特異値分解を行う手法が提案されている．ここでいう意味属性とは，単語の意味のカテゴリを表し，各単語には対応する意味属性が付随している．文献 6) では，単語間共起の特徴行列が与えられれば，意味属性が同じ単語の列をマージすることにより，単語・意味属性間共起の属性行列に変換する．

単語間共起の特徴行列としては，上述したコーパスから生成するコーパス共起行列のほかに，見出し語とその説明文から構成される国語辞典から生成する辞書共起行列がある[7)]．辞書共起行列では，各行が見出し語に相当し，各列が説明文に出現する特徴語に相当する．行列の成分は，対応する見出し語の説明文中における，対応する特徴語の出現回数である．この辞書共起行列にさらに各種精錬化処理を行う．辞書共起行列では，説明文中の特徴語間の共起頻度は考慮せず，コーパス共起行列とは異なるものである．文献 6) では，辞書共起行列をもとにした単語・意味属性間共起行列に対し評価が行われ，効果が確認されている．しかし，見出しや説明文といった構造を持たないコーパスから生成するコーパス共起行列をもとにした単語・意味属性間共起行列による効果を検証した研究はいまだないのが現状である．

国語辞典と比較してコーパスは，膨大な量を用意しやすく，また，分野・時期に応じたものを用意しやすいという利点がある．

本論文では，コーパス共起行列をもとにした単語・意味属性間共起行列に対し特異値分解を行うことにより生成されるコーパス概念ベースが，コーパス共起行列に対し特異値分解を行うことにより生成されるコーパス概念ベースの諸問題をいかに解決するかを考察し，その有効性を検証する．

この単語・意味属性間共起に基づく手法では，共起行列において，意味属性が同じ単語との共起頻度がマージされるため，行ベクトルが，対応する単語の概念を的確に表現できるようになる．また，特異値分解のメモリ使用量の制約により，単語間共起行列の列となる単語集合は限定されていたが，この列から漏れた単語との共起頻度も，該単語の意味属性との共起頻度に含まれるので，特異値分解の対象となる共起行列の列数を増やしていくことなく，共起行列の情報量を増やすことができる．

一方，特異値分解のメモリ使用量の制約上，特異値分解の対象となる共起行列の行は制限されたままであるため，共起行列中のある行に対応する単語で，概念ベースに含まれないものが存在する．このような単語を，未登録語と呼ぶことにする．本論文では，未登録語の概念ベクトルを，概念ベクトルの分布の分散最小性に基づいて推定する手法を提案する．未登録語とその推定概念ベクトルを概念ベースに追加し，概念ベースを拡張することにより，概念ベースの品質を向上させることができる．

以下，2 章でコーパス概念ベースの生成アルゴリズムについて述べ，3 章で未登録語の概念ベクトルの推定手法を述べる．4 章で評価実験の結果を述べ，5 章でまとめを述べる．

2. コーパス概念ベース生成アルゴリズム

本章では，まず従来手法である単語間共起に基づく手法について紹介した後，その問題点を考察し，それを解決する手法として単語・意味属性間共起に基づく手法を述べる．

2.1 従来手：単語間共起に基づく手法

本節では，文献 1) における単語間共起に基づく概念ベース生成アルゴリズムを紹介する．

まずコーパスを形態素解析し，名詞，用言等の内容語のみを残す．用言は終止形に統一する．残った異なり単語の集合を G，K（$G = K$）とする．G 中の単語を概念語，K 中の単語を共起語と呼ぶ．任意の概念語と共起語とが 1 文中に共起する頻度をカウントし，各行が概念語に対応し，各列が共起語に対応しているような共起行列を生成する．共起行列の各行ベクトルは，対応する概念語の共起パターンを表しており，この行ベクトルを共起ベクトルと呼ぶ．ある 2 単語に対応する共起ベクトルが近ければ，共起パターンが似ている（共通の隣人となる単語を持っている）ので，この 2 単語は意味的に近いということが推測される．ただし，限定されたテキストデータから抽出される単語間の共起頻度は，単語間の関連性を完全に正確に表現しているとはいえないため，このままでは共起ベクトル間の類似度の精度は低いと考えられる．また，一般に共起ベクトルの次元数は非常に大きなものとなるので，共起ベクトルを利用する言語処理の計算量も無視できないものとなる．このため共起行列を特異値分解により，次元数を縮退させた行列に変換する．次元圧縮により，共起ベクトル間の細かな差異を切り捨て，テキストデータ中の情報の欠落の影響を減らすことができる．また，ベクトルを用いた言語処理の計算量も減らすことができる．

G，K の要素数が多いと，特異値分解のメモリ使用量は多量になるので，低コストで実行することが不可能となる．そこで，G，K を，高頻度語の集合に限定したうえで特異値分解を実行する．

最初の単語間共起行列の生成は，G は限定せず，K を高頻度語の集合に限定したうえで，**図 1** のように行う．このようにして**図 2** のような共起行列が生成される．共起行列を生成した後，共起行列から零ベクトルである行ベクトルを削除する．高頻度語間の共起頻度は非常に大きな数で，高頻度語の共起ベクトルと低頻度語の共起ベクトルとの乖離が大きくなり，次元圧縮後の精度を低下させるので，共起頻度が過度に大きくならないように，共起行

コーパス中の任意の文に対し，以下の処理を行う．

1) 文 c における，以下の異なり共起語とその頻度の組の集合を導出する．

$\{(w_i, z_i)\}$　w_i : c 内の異なり共起語
z_i : w_i の c における頻度

2) c における任意の概念語 w（同一表記のものが複数ある場合，別物とする）に対応する行ベクトルに，上記の，異なり共起語とその頻度の組の集合の情報を加算する．加算の仕方として以下の，w 自身との共起はカウントしない自己非共起方式 A) と，カウントする自己共起方式 B) がある．

A) **自己非共起方式**
$\{(w_i, z_i)\}$ から，w の頻度を 1 だけ減じたものを，w のベクトルに加算する．

B) **自己共起方式**
$\{(w_i, z_i)\}$ を，w のベクトルに加算する．

図 1　単語間共起頻度算出処理
Fig. 1　Algorithm of calculating co-occurrence frequencies between words.

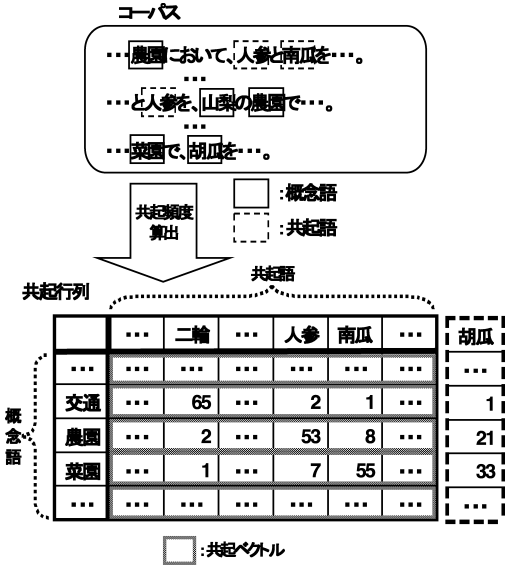

図 2　単語間共起行列
Fig. 2　A matrix of co-occurrences between words.

列中の各成分をその平方根に変換しておく．このようにすることにより次元圧縮後の精度が向上する．共起行列の G を高頻度語の集合に限定した部分行列 X に対し特異値分解を実行する．

X を $p \times q$ の行列としたとき，特異値分解により X は，以下のように分解できる．

$$\underset{p\times q}{X} = \underset{p\times r}{U}\ \underset{r\times r}{\Sigma}\ \underset{r\times q}{V^T} \tag{1}$$

ここで，添字 T は行列の転置を表す．$r = \mathrm{rank}\, X \leq \min(p, q)$，$U^T U = V^T V = I$（$I$：単位行列）であり，$\Sigma = (\delta_{ij})$ としたとき，$\delta_{ii} \geq \delta_{jj} > 0$ $(1 \leq {}^\forall i \leq {}^\forall j \leq r)$，$\delta_{ij} = 0$ $(i \neq j)$ である．δ_{ii}（$1 \leq i \leq r$）を X の特異値と呼ぶ．

ここで，$1 \leq {}^\forall r' \leq r$ に対し，U の最初の r' 列，V^T の最初の r' 行，Σ の最初の r' 行，r' 列をとり，

$$\underset{p\times q}{X'} = \underset{p\times r'}{U'}\ \underset{r'\times r'}{\Sigma'}\ \underset{r'\times q}{V'^T} \tag{2}$$

とする．U' の行ベクトルを長さ 1 に正規化したものを単語概念ベクトルと呼び，概念語とその概念ベクトルの対の集合を単語概念ベースと呼んでいる．

2.2　単語・意味属性間共起に基づく手法の考察

2.1 節で述べた単語間共起に基づく手法は，以下に述べるような精度上の問題があるため，文献 6) で提案されている単語・意味属性間共起に基づく手法が，これらの諸問題をいかに解決するかを考察する．

まず単語間共起に基づく手法では，共起行列の列となる単語の中に同一のカテゴリに属するものがあり，それらの単語との共起頻度が別々にカウントされるため，共起ベクトルが適切なものでなくなるという問題がある．たとえば，図 2 の “人参” と “南瓜” は同一のカテゴリ “野菜” に属するが，それらとの共起頻度が別々にカウントされるため，“農園” と “菜園” の共起ベクトルが適切なものでなくなり，“農園” と “菜園” は意味的に近いにもかかわらず，対応する共起ベクトルは遠くなる．

また，単語間共起に基づく手法では，共起語を高頻度語に限定するため，共起行列の列となる単語から漏れる単語が多数あり，そのような単語との共起頻度は考慮されないという問題がある．たとえば，図 2 の “胡瓜” は限定した共起語集合に含まれないため，“胡瓜” との

コーパス中の任意の文に対し，以下の処理を行う．

1) 文 c における，以下の概念語の意味属性とその頻度の組の集合を導出する．

$\{(s_i, z_i)\}$ $s_i : c$ 内の概念語の意味属性
$z_i : s_i$ の c における頻度

2) c における任意の概念語 w（同一表記のものが複数ある場合，別物とする）に対応する行ベクトルに，上記の，意味属性とその頻度の組の集合の情報を加算する．加算の仕方として以下の，w 自身の意味属性との共起はカウントしない自己非共起方式 A) と，カウントする自己共起方式 B) がある．

A) **自己非共起方式**
$\{(s_i, z_i)\}$ から，w の各意味属性の頻度を 1 だけ減じたものを，w のベクトルに加算する．

B) **自己共起方式**
$\{(s_i, z_i)\}$ を，w のベクトルに加算する．

図 3　単語・意味属性間共起頻度算出処理
Fig. 3　Algorithm of calculating co-occurrence frequencies between words and those semantic attributes.

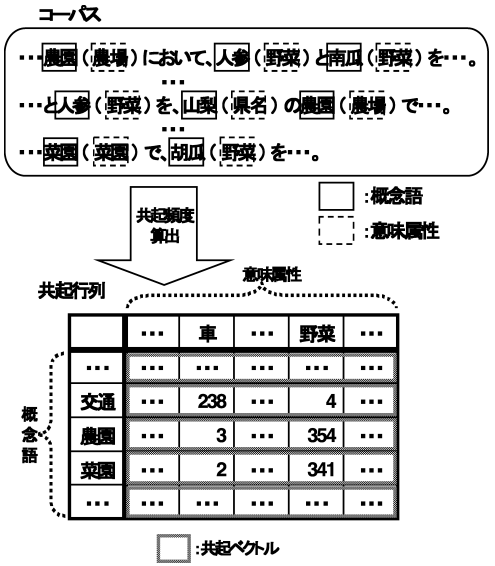

図 4　単語・意味属性間共起行列
Fig. 4　A matrix of co-occurrences between words and those semantic attributes.

共起頻度が考慮されない．このような情報の欠落により，共起ベクトルの品質が低下する．

この単語間共起に基づく手法の問題点を解決するためのものとして，単語・意味属性間共起に基づく手法では，コーパスにおける単語どうしの共起頻度ではなく，コーパスにおける単語と，単語に付随する意味属性との共起頻度をとる．この意味属性とは，日本語語彙大系[8] における一般名詞意味体系の意味属性を意味している．

日本語語彙大系における一般名詞意味体系は，名詞と用言の意味を体系立てたシソーラスであり，各ノードを意味属性と呼ぶ．このシソーラスは 12 階層であり，2,715 個のノードからなる．

本論文の手法では，形態素解析プログラムとして JTAG[9] を用いる．JTAG が参照する単語辞書では，各名詞と用言に意味属性が付与されている．1 つの単語に複数の意味属性が付与されていることもあるが，これらの意味属性は，使用される局面が高いと思われる順に順序付けられている．形態素解析結果において，各単語には，対応する意味属性の情報が付随している．

概念語の集合 G を限定することなく，任意の概念語と意味属性とが 1 文中に共起する頻度を**図 3** のようにしてカウントする．図 3 の処理において，1 単語に複数の意味属性がついている場合，使用される局面の高い最初の一定数の意味属性のみ考慮するというバリエーションも考えられる．このようにして，**図 4** のような，各行が概念語に対応し，各列が意味属性に対応しているような共起行列を生成する．

このように単語ではなく，意味属性との共起頻度をとることにより，同一の意味属性を持つ個々の単語との共起頻度は，該意味属性との共起頻度に含まれるため，共起ベクトルが，より適切なものとなる．たとえば，図 2 における “二輪” の意味属性は “車” で，“人参”，“南瓜” の意味属性は “野菜” であるため，“人参”，“南瓜” それぞれとの共起頻度は，“野菜” との共起頻度に含まれる．これによって意味的に近い “農園” と “菜園” の共起ベクトルは値が近くなる．単語・意味属性間共起行列は，意味属性という知識が混入されているため，単語間共起行列よりも，単語間の意味的類似性をより反映させることができる．

また，意味属性の数はたかだか 2,715 であるため，全意味属性を共起行列の列として採用

することができる．このため，単語間共起に基づく手法で，共起行列の列となる単語から漏れていた単語との共起頻度も，該単語の意味属性との共起頻度に含まれるため，共起ベクトルが，より豊富な情報を持つようになる．たとえば，図 2 における “胡瓜” の意味属性は “野菜” であるため，“胡瓜” との共起頻度が “野菜” との共起頻度に含まれる．単語間共起に基づく手法では，考慮されなかった “胡瓜” との共起頻度が，単語・意味属性間共起に基づく手法では考慮されるようになる．単語・意味属性間共起に基づく手法は，共起行列の列数を増やしていくことなく，列となりうる全単語を考慮した共起行列を生成できる．

共起行列を生成した後，共起行列から零ベクトルである行ベクトルを削除する．精度向上のため，共起行列中の各成分をその平方根に変換しておく．共起行列の G を高頻度語の集合に限定した部分行列 X に対し特異値分解を実行する．その結果得られる式 (2) における U' の行ベクトルを長さ 1 に正規化したものを単語・意味属性間共起に基づく手法の単語概念ベクトルとし，概念語とその概念ベクトルの対の集合を単語・意味属性間共起に基づく手法の単語概念ベースとする．

3. 未登録語の概念ベクトルの推定手法

2 章で述べた単語間共起に基づく手法も単語・意味属性間共起に基づく手法も，特異値分解のメモリ使用量の制約上，特異値分解の対象となる共起行列の行を制限する必要がある．共起行列中のある行に対応する単語で，概念ベースに含まれないものを未登録語と呼ぶ．概念ベースを利用する言語処理において，未登録語の概念はいっさい考慮されないため，精度の低下を招く．

未登録語の概念ベクトルを推定する手法として，従来手法である意味空間への射影による手法[10] と，提案手法である概念ベクトルの分散最小性に基づく手法を述べる．

3.1 従来手法：意味空間への射影による手法

文献 10) においては，フォルディング・イン（folding-in）と呼ばれる，特異値分解によって得られる意味空間へ射影する手法が述べられている．意味空間とは，式 (2) における $\underset{r' \times q}{V'^T}$ の r' 個の行ベクトルが張る空間である．式 (1) より，

$$\underset{p \times q}{X}\ \underset{q \times r}{V}\ \underset{r \times r}{\Sigma^{-1}} = \underset{p \times r}{U}$$

であるため，

$$\underset{p \times q}{X}\ \underset{q \times r'}{V'}\ \underset{r' \times r'}{\Sigma'^{-1}} = \underset{p \times r'}{U'}$$

となる．これにならい，任意の共起ベクトル $\underset{1 \times q}{h_w}$ に対し，

$$\underset{1 \times q}{h_w}\ \underset{q \times r'}{V'}\ \underset{r' \times r'}{\Sigma'^{-1}} \tag{3}$$

といたものは，各成分が，$\underset{r' \times q}{V'^T}$ の対応する行ベクトルと h_w の内積に，対応する特異値の逆数を乗じたものであるため，確かに，意味空間への h_w の射影となっている．任意の共起ベクトル h_w に対し，式 (3) で得られるベクトルを長さ 1 に正規化したものを推定概念ベクトルとする．

3.2 提案手法：分散最小性に基づく手法

分散最小性に基づく手法とは，概念ベクトルの分布が正規分布に従っているという仮定に基づく手法である．以下，背景を述べつつ，本手法を説明する．

1 つのトピック区間内の単語は意味的に近いので，その概念ベクトルも互いに近いと考えられる．そのため，トピック区間内の概念ベクトルは正規分布に従っているという仮定をおくことができる．

文献 5) においては，この仮定のもとにテキストセグメンテーションを行うことにより，一定の精度を出している．また，文献 11) においては，この仮定のもとに，ある動詞にある助詞で係る名詞の集合をクラスタとしたうえで，クラスタ間の分散が，一定の制約下で最大となるように名詞にベクトルを付与する手法が提案されている．

本仮定のもとではトピック区間内に未登録語が存在する場合でも，トピック区間内の未登録語を含めた単語の概念ベクトルは正規分布に従っている．このとき，テキスト中の各トピック区間内の概念ベクトル集合をクラスタとみれば，クラスタ内分散の平均は最小となる．このため，提案する分散最小性に基づく手法では，コーパス中の 1 文における単語集合をクラスタとし，登録語の概念ベクトルを固定したうえで，クラスタ内分散の平均が最小となるように，未登録語の概念ベクトルを付与する．ここで，クラスタの範囲を 1 文としたのは，通常，1 文がトピック区間として保証されている最大の範囲だからである．

ただし，文献 12) で述べているように，この手法では，未登録語の異なり数だけの変数を持つ 2 次式が最大となる解を，ある制約条件の下で解く必要があり，多量の計算量を要する．未登録語の異なり数が多いと，一般のコンピュータでは実行が不可能となる．

このため，文を 1 つずつとっていき，取得した文集合における未登録語の異なり数が，あ

る一定数を超えた時点で，取得した文集合における各未登録語の概念ベクトルを求め，概念ベースに追加する．この操作を，取得する文がなくなるまで繰り返すことにより，全未登録語の概念ベクトルを求める．

しかしながら，この手法では，一部の文集合における情報から，それに含まれる未登録語の概念ベクトルを求めるので，全文集合から求めるのと比べ，情報源の量が圧倒的に少ない．このため，未登録語の推定概念ベクトルの品質に問題がある．

このため，本論文では，着目している 1 つの異なり未登録語以外の異なり未登録語の概念ベクトルの存在は無視したうえで，各文内の概念ベクトルの分散の平均が最小となるように，該異なり未登録語の概念ベクトルを求める．以下，提案手法について説明する．

対象としている 1 個の異なり未登録語を含む，コーパス中の文の集合を $C = \{c_1, c_2, \cdots, c_g\}$ とする．

また，C 中に出現する，異なり登録語の集合を $\{w_1, w_2, \cdots, w_x\}$ とし，対象としている異なり未登録語を w_{x+1} とする．

文 c_j 内の異なり単語 w_i の出現回数を $z(w_i|c_j)$ とする．また，c_j でののべ単語数を，

$$z(c_j) := \sum_{1 \le i \le x+1} z(w_i|c_j)$$

と定義する．

概念ベクトルは f 次元ベクトルとし，単語 w_i の概念ベクトルの m 番目の成分を $v^m(w_i)$ とする．

c_j での m 番目の成分の平均を

$$\mu^m(c_j) := \frac{\sum_{1 \le i \le x+1} z(w_i|c_j) \bullet v^m(w_i)}{z(c_j)}$$

と定義する．

各文 c_j における概念ベクトルの集合をクラスタとしたとき，クラスタ内分散は，概念ベクトルの平均と，各概念ベクトルとの距離の自乗の和を，概念ベクトル数で割った式 (4) で表される．

$$\frac{\sum_{1 \le i \le x+1} z(w_i|c_j) \sum_{1 \le m \le f} \left(v^m(w_i) - \mu^m(c_j)\right)^2}{z(c_j)} \tag{4}$$

クラスタ内分散の平均は，各クラスタ内の概念ベクトル数で重み付けした平均である式 (5) で表される．

$$\frac{\sum_{1 \le j \le g} z(c_j) \dfrac{\sum_{1 \le i \le x+1} z(w_i|c_j) \sum_{1 \le m \le f} \left(v^m(w_i) - \mu^m(c_j)\right)^2}{z(c_j)}}{\sum_{1 \le j \le g} z(c_j)} \tag{5}$$

式 (5) の分母は定数であるため，分子である以下の式 (6) を考えればよい．

$$\begin{aligned} &\sum_{1 \le j \le g} \sum_{1 \le i \le x+1} z(w_i|c_j) \sum_{1 \le m \le f} \left(v^m(w_i) - \mu^m(c_j)\right)^2 \\ &= \sum_{1 \le m \le f} \sum_{1 \le j \le g} \sum_{1 \le i \le x+1} z(w_i|c_j) \left(v^m(w_i) - \mu^m(c_j)\right)^2 \end{aligned} \tag{6}$$

式 (6) を最小にする $v^m(w_{x+1})$（$1 \le m \le f$）を求めるには，任意の成分 m（$1 \le m \le f$）に対し，

$$\sum_{1 \le j \le g} \sum_{1 \le i \le x+1} z(w_i|c_j) \left(v^m(w_i) - \mu^m(c_j)\right)^2 \tag{7}$$

を最小にする $v^m(w_{x+1})$ を求めればよい．

式 (7) は以下のように変形される．

$$\begin{aligned} &\sum_{1 \le j \le g} \sum_{1 \le i \le x+1} z(w_i|c_j) \left(v^m(w_i) - \mu^m(c_j)\right)^2 \\ &= \sum_{1 \le j \le g} \left[\begin{array}{l} \displaystyle\sum_{1 \le i \le x} \left[z(w_i|c_j) \bullet \left[\begin{array}{l} v^m(w_i) - \dfrac{\sum_{1 \le p \le x} z(w_p|c_j) \bullet v^m(w_p)}{z(c_j)} \\ -\dfrac{z(w_{x+1}|c_j)}{z(c_j)} \bullet v^m(w_{x+1}) \end{array} \right]^2 \right] \\ +z(w_{x+1}|c_j) \bullet \left[\begin{array}{l} -\dfrac{\sum_{1 \le p \le x} z(w_p|c_j) \bullet v^m(w_p)}{z(c_j)} \\ +\left(1 - \dfrac{z(w_{x+1}|c_j)}{z(c_j)}\right) \bullet v^m(w_{x+1}) \end{array} \right]^2 \end{array} \right] \end{aligned}$$

$$= \sum_{1\le j\le g} \left(a_j \bullet v^m (w_{x+1})^2 + b_j \bullet v^m (w_{x+1}) + d_j \right) \quad (8)$$

ここで，

$$z(c_j) = \sum_{1\le i\le x+1} z(w_i|c_j)$$
$$= \sum_{1\le i\le x} z(w_i|c_j) + z(w_{x+1}|c_j) = z_k(c_j) + z(w_{x+1}|c_j)$$

とおくと，a_j，b_j は，以下のように表される．

$$a_j = \sum_{1\le i\le x} z(w_i|c_j) \left(\frac{z(w_{x+1}|c_j)}{z(c_j)} \right)^2 + z(w_{x+1}|c_j) \left(1 - \frac{z(w_{x+1}|c_j)}{z(c_j)} \right)^2$$
$$= z_k(c_j) \left(\frac{z(w_{x+1}|c_j)}{z(c_j)} \right)^2 + z(w_{x+1}|c_j) \left(\frac{z_k(c_j)}{z(c_j)} \right)^2$$
$$= z_k(c_j)\, z(w_{x+1}|c_j) \left(\frac{1}{z(c_j)} \right)^2 \left(z_k(c_j) + z(w_{x+1}|c_j) \right)$$
$$= \frac{z_k(c_j)\, z(w_{x+1}|c_j)}{z(c_j)}$$

$$b_j = -2\frac{z(w_{x+1}|c_j)}{z(c_j)} \sum_{1\le i\le x} \left[z(w_i|c_j) \bullet \left[v^m(w_i) - \frac{\sum_{1\le p\le x} z(w_p|c_j) \bullet v^m(w_p)}{z(c_j)} \right] \right]$$
$$- 2z(w_{x+1}|c_j) \frac{\sum_{1\le p\le x} z(w_p|c_j) \bullet v^m(w_p)}{z(c_j)} \bullet \left(1 - \frac{z(w_{x+1}|c_j)}{z(c_j)} \right)$$
$$= -2\frac{z(w_{x+1}|c_j)}{z(c_j)^2} \left[\begin{array}{l} z(c_j) \sum_{1\le i\le x} z(w_i|c_j) \bullet v^m(w_i) - \sum_{1\le i\le x} z(w_i|c_j) \\ \bullet \sum_{1\le p\le x} z(w_p|c_j) \bullet v^m(w_p) \end{array} \right]$$
$$- 2\frac{z(w_{x+1}|c_j)}{z(c_j)^2} \left(z(c_j) - z(w_{x+1}|c_j) \right) \bullet \sum_{1\le p\le x} z(w_p|c_j) \bullet v^m(w_p)$$
$$= -2\frac{z(w_{x+1}|c_j)}{z(c_j)} \sum_{1\le i\le x} z(w_i|c_j) \bullet v^m(w_i)$$

式 (8) は，2 次式

$$a \bullet v^m (w_{x+1})^2 + b \bullet v^m (w_{x+1}) + d \quad (9)$$

と表される．

C 中に異なり登録語が存在する場合，$a > 0$ であり，式 (9) を最小にする $v^m(w_{x+1})$ は，

$$v^m (w_{x+1}) = -\frac{b}{2a}$$

となる．

a_j は成分 m に依存しないので，a も成分 m に依存しない．

m 番目の成分が $v^m(w_i)$ である，未登録語 w_i のベクトルを $v(w_i)$ とする．$v(w_i)$ を長さ 1 に正規化したものを，未登録語 w_i の推定概念ベクトルとする．

提案手法では，1 変数の 2 次式が最小となる解を求めるので，計算量的に問題がない．このため，着目している 1 つの異なり未登録語を含むすべての文の集合から，該異なり未登録語の概念ベクトルを推定することができる．他の異なり未登録語の概念ベクトルを考慮したうえでは分散最小とはならないものの，情報源の量は文献 12) の手法と比べ圧倒的に多いため，結果として，推定概念ベクトルの品質は高くなる．

4. 評価実験

単語間共起に基づく手法と，単語・意味属性間共起に基づく手法の両方に対し，2 つの推定手法を適用し，生成した概念ベースを用いた文書検索の精度を比較した．

単語概念ベース生成用コーパスとしては，101,302 個の Q&A 文書と，それを包含する 1,695,818 個の Q&A 文書の 2 つを用い，コーパス量ごとの各手法の精度を調べることとした．

コーパス中の名詞，用言等の異なり単語すべてを，共起行列の行となる単語として，共起行列を生成した．共起行列の列数は，単語間共起に基づく手法と単語・意味属性間共起に基づく手法とで条件が揃うように 2,715 とした．このため，単語間共起に基づく手法では，列となる単語は 2,715 個の高頻度語とした．

単語・意味属性間共起に基づく手法で共起頻度をとる際は，形態素解析結果中の 1 単語の意味属性が複数ある場合は，それらの中で最も使用される局面が高い意味属性のみ考慮した．

共起行列から零ベクトルである行ベクトルを削除し，共起行列中の各成分をその平方根に変換した．

特異値分解の対象となる部分行列の行となる単語は，26,900 個の高頻度語とした．この

表 1　概念ベースの単語数
Table 1　Number of words in the concept bases.

共起：推定手法	コーパス文書数	
	101,302	1,695,818
全異なり単語数	86,263	155,430
単語間：射影	85,121	153,872
単語間：分散	85,878	154,997
単語・意味属性間：射影	86,003	155,165
単語・意味属性間：分散	85,878	154,997

部分行列から特異値分解により，200 次元の概念ベクトルを生成した．特異値分解処理のメモリ使用量は 1 GB 強であった．

射影による手法では，未登録語の，共起行列における共起ベクトルを h_w として，推定概念ベクトルを導出した．

いずれのケースも，自己共起よりも自己非共起の方が精度が高かったため，自己非共起の結果を報告する．

各ケースごとの，拡張後の概念ベースの単語数は，**表 1** のとおりである．表 1 において，「射影」とは射影による手法を，「分散」とは分散最小性に基づく手法を意味する．

検索アルゴリズムは，以下のとおりである．各検索対象文書を形態素解析し，名詞，用言等の内容語のみ残す．用言は終止形に統一する．各検索対象文書において，残った単語（同一表記のものが複数ある場合，別物とする）の概念ベクトルの和を長さ 1 に正規化したものを，該検索対象文書の概念ベクトルとする．各検索対象文書の概念ベクトルは，あらかじめ生成しておきインデックスに格納しておく．検索キーとなる入力文書に対しても同様の手順で，その概念ベクトルを生成し，入力文書概念ベクトルと各検索対象文書概念ベクトルとの距離の近い順に，検索対象文書集合をランキングして検索結果とする．

検索対象文書集合として，単語概念ベース生成用コーパスとは共通部分を持たない 99,404 個の Q&A 文書の質問文部分を用いた．あらかじめ 1 つの検索対象文書と質問意図が同じで，可能な限り該文書中の内容語を含まない入力文書を作成した．一例として，「不景気の原因」という検索対象文書からは，「不況の理由」という入力文書を作成する（実際の検索対象文書は，質問文としてもっと長いものである）．入力文書を検索キーとして検索を実行し，得られた検索結果における，該入力文書に対応する検索対象文書の順位を n としたとき，$1/n$ の平均値（平均逆順位と呼ぶ）を精度の指標とする．入力文書を作成する際は，入力文書中の内容語で，元の文書に含まれるものすべてを含む検索対象文書が 10 件以下となるような入力文書は，全文検索でも容易に対応する検索対象文書が見つけられるため，除外することとした．このようにして，6,068 個の入力文書を作成した．

表 2　検索精度（平均逆順位）
Table 2　Retrieval accuracy (Mean reciprocal rank).

共起：推定手法	コーパス文書数	
	101,302	1,695,818
単語間：拡張前	0.3276	0.3557
単語間：射影	0.3494	0.3865
単語間：分散	0.3518	0.3830
単語・意味属性間：拡張前	0.3339	0.3565
単語・意味属性間：射影	0.3572	0.3887
単語・意味属性間：分散	0.3589	0.3861

各ケースごとの平均逆順位は**表 2** のようになった．1 つの共起：推定手法の出力結果における逆順位の分布について，それが正規分布であるという仮説は，いずれの分布に対しても，正規性検定により p 値がほぼ 0%であり，右片側有意水準 1%で棄却された．そこで，任意の 2 つの分布の平均に差があるかを，右片側有意水準 1%でのウィルコクソンの符号付順位和検定で検証することにする．この検定では，比較する 2 つの手法間で違いが出たデータ群について，有意な差があるかを検証できる．

他の条件が同じ場合，単語・意味属性間共起に基づく手法は，単語間共起に基づく手法より，つねに高精度となった．だが，コーパスが多量（1,695,818 文書）のときは p 値が 0.24%〜0.87%であり有意差が認められたが，コーパスが少量（101,302 文書）のときは p 値が 2.58%〜25.19%であり有意差が認められなかった．この原因として，コーパスが少量のときは，単語間共起行列の列から漏れる単語が比較的少ないため，意味属性導入による情報量拡大の効果がそれほど大きくなかったことが考えられる．コーパスが多量のときは，単語・意味属性間共起をとることにより，特異値分解のメモリ使用量を増やすことなく，概念ベースの品質を向上させることができる．

いずれの推定手法を用いても，概念ベースを拡張することにより，検索精度が向上し，p 値がほぼ 0%で有意差が認められた．

コーパスが少量（101,302 文書）のときは，単語間共起であっても単語・意味属性間共起であっても，分散最小性に基づく手法の方が，射影による手法よりも，高精度であり p 値がほぼ 0%で有意差が認められた．一方，コーパスが多量（1,695,818 文書）のときは，分散最小性に基づく手法は，射影による手法より精度が低いという結果となっている．

表 3　検索精度（平均逆順位）
Table 3　Retrieval accuracy (Mean reciprocal rank).

共起：推定手法	コーパス文書数 1,695,818
単語間：射影＋分散	0.3869
単語・意味属性間：射影＋分散	0.3891

この原因として，コーパスが少量のときは，未登録語の共起ベクトルの成分値はスパースなものであり，高頻度の概念語から生成された意味空間と，未登録語の共起ベクトルとの乖離が大きく，射影しても信頼性が低いことが考えられる．意味空間上にある登録語の概念ベクトルから，未登録語の概念ベクトルを分散最小性によって求める方が，推定の確度が高い．

逆に，コーパスが多量のときは，未登録語の共起ベクトルの成分値はスパースではなく，意味空間が，未登録語の共起ベクトルの分布も，ある程度反映しているため，射影によって得られる推定概念ベクトルの品質が高いと考えられる．

このことから，コーパスが多量のときでも，高頻度の未登録語に対しては射影により概念ベクトルを推定し，低頻度の未登録語に対しては分散最小性により推定するというバリエーションも有効だと考えられる．コーパスが少量（101,302 文書）のとき，共起行列において零ベクトルでない未登録語の出現頻度の最大値は 31 個であった．そこで，コーパスが多量（1,695,818 文書）のとき，共起行列において零ベクトルでない未登録語で出現頻度が 32 個以上のものの概念ベクトルを射影による手法で推定し，残りの未登録語の概念ベクトルを，初期の概念ベースをもとに，分散最小性に基づく手法により推定した．単語間共起も単語・意味属性間共起もともに，拡張後の概念ベースの単語数は 154,997 となり，128,097 個の未登録語のうち，射影による手法の対象となった単語は 35.7%を占め，分散最小性に基づく手法の対象となった単語は 64.3%を占めた．

各ケースごとの平均逆順位は**表 3** のようになった．これらの結果における逆順位の分布も，正規分布であるという仮説は，正規性検定により p 値がほぼ 0%であり，右片側有意水準 1%で棄却された．そこで，これらの分布を含めた任意の 2 つの分布の平均に差があるかを，右片側有意水準 1%でのウィルコクソンの符号付順位和検定で検証した．単語間共起も，単語・意味属性間共起も，「射影」のみあるいは「分散」のみによる拡張よりも，「射影＋分散」による拡張の方が，高精度であり p 値がほぼ 0%で有意差が認められた．また，表 3 においても，単語間共起より単語・意味属性間共起の方が，高精度であり p 値が 0.43%で有意差が認められた．

これらのことから，コーパスが多量のときは，共起としては，単語・意味属性間共起をとり，未登録語の推定においては，高頻度語に対しては射影による手法で推定し，低頻度語に対しては分散最小性に基づく手法で推定するハイブリッド方式をとることにより，最も高い精度が得られるといえる．

5. ま と め

単語・意味属性間共起に基づく概念ベース生成手法と，概念ベクトルの分散最小性に基づく未登録語の概念ベクトル推定手法により，メモリ使用量を増大させることなく，高品質の概念ベースを生成できることを検証した．今後は，生成した概念ベースを利用した言語処理アプリケーションのさらなる研究を進めていきたい．

参 考 文 献

1) Schütze, H.: Automatic Word Sense Discrimination, *Computational Linguistics*, Vol.24, No.1, pp.97–123 (1998).
2) Schütze, H. and Pedersen, J.: A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval, *Proc. RIAO'94*, pp.266–274 (1994).
3) Kato, T., Shimada, S., Kumamoto, M. and Matsuzawa, K.: Idea-Deriving Information Retrieval System, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.187–193 (1999).
4) 熊本　睦，島田茂夫，加藤恒昭：概念ベースの情報検索への適用—概念ベースを用いた検索の特性評価，情報処理学会研究報告，Vol.SIG-ICS 115, pp.9–16 (1999).
5) 別所克人：クラスタ内変動最小基準に基づくテキストセグメンテーション，情報処理学会論文誌，Vol.47, No.3, pp.957–967 (2006).
6) 笠原　要，稲子希望，加藤恒昭：単語の属性空間の表現方法，人工知能学会誌（JSAI），Vol.17, No.5, pp.539–547 (2002).
7) 笠原　要，松澤和光，石川　勉：国語辞書を利用した日常語の類似性判別，情報処理学会論文誌，Vol.38, No.7, pp.1272–1283 (1997).
8) 池原　悟，宮崎正弘，白井　諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦：日本語語彙大系，岩波書店 (1997).
9) Fuchi, T. and Takagi, S.: Japanese Morphological Analyzer using Word Co-occurence-JTAG, *COLING-ACL*, pp.409–413 (1998).
10) Berry, M.W., Dumais, S.T. and O'Brien, G.W.: Using linear algebra for intelligent information retrieval, *SIAM Review*, Vol.37, No.4, pp.573–595 (1995).
11) 冨浦洋一，田中省作，日高　達：共起データに基づく名詞の n 次元空間への配置，情報処理学会研究報告，Vol.SIG-NL 154, pp.71–76 (1999).

12) 別所克人，奥　雅博：未知語の概念ベクトル推定手法，情報処理学会研究報告，Vol.SIG-NL 164, pp.59–64 (2004).

(平成 19 年 11 月 1 日受付)
(平成 20 年 9 月 10 日採録)

別所　克人（正会員）

1992 年大阪大学理学部数学科卒業．1994 年同大学大学院修士課程修了．同年日本電信電話（株）入社．現在，NTT サイバーソリューション研究所所属．自然言語処理の研究に従事．電子情報通信学会，言語処理学会各会員．

内山　俊郎（正会員）

1987 年東京工業大学工学部電気電子工学科卒業．1989 年同大学大学院修士課程修了．同年（株）NTT データ入社．1991〜1993 年南カリフォルニア大学客員研究員．1999〜2005 年通信・放送機構研究員，特別研究員．2006 年より NTT サイバーソリューション研究所所属．Web データマイニング，分光色再現の研究に従事．博士（工学）．

内山　　匡（正会員）

1985 年名古屋大学理学部物理学科卒業．1987 年同大学大学院修士課程修了．同年日本電信電話（株）入社．1998〜2001 年 NTT コミュニケーションズ，2004〜2006 年 NTT レゾナントにてポータルサービスの開発等に従事．2007 年より NTT サイバーソリューション研究所所属．ポータルサービスシステムの研究開発に従事．電子情報通信学会，日本応用数理学会各会員．

片岡　良治（正会員）

1987 年千葉大学大学院電子工学専攻修士課程修了．同年日本電信電話（株）入社．以来，トランザクションの並行処理制御方式の研究，マルチメディア情報システムの研究，ポータルサービスシステムの研究開発に従事．現在，NTT サイバーソリューション研究所所属．

奥　　雅博（正会員）

1982 年大阪府立大学工学部電子工学科卒業．1984 年同大学大学院修士課程修了．同年日本電信電話公社（現 NTT）に入社し，日英機械翻訳システム等における自然言語処理技術，特に日本語処理技術の研究開発に従事．現在，NTT サイバーソリューション研究所において，検索をはじめとするブロードバンドインターネットサービスに関する研究開発に従事．博士（工学）．電子情報通信学会，言語処理学会各会員．