社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－NL－164（10）
2004／11／4

未知語の概念ベクトル推定手法

別所 克人† 奥 雅博†

†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

〒239-0847 神奈川県横須賀市光の丘 1-1

E-mail: †{bessho.katsuji, oku.masahiro}@lab.ntt.co.jp

あらまし　他の単語との共起パターンから導出される単語の意味表現である概念ベクトルとその集合である概念ベースは、テキストの話題構造化や検索に非常に有用なものであるが、概念ベースに含まれない単語としての未知語には概念ベクトルが付与されず、概念ベースを利用した処理において考慮されないという問題がある。本稿では、テキストにおける共起情報から、未知語の概念ベクトルを推定する手法を提案する。新聞記事を用いたトピックセグメンテーションの評価実験の結果、推定ベクトルを用いる手法は、用いない手法よりも精度が向上することを検証した。

キーワード　未知語，概念ベクトル，トピックセグメンテーション

Estimate of Conceptual Vectors for Unregistered Words

Katsuji BESSHO‡ and Masahiro OKU‡

†NTT Cyber-Solutions Laboratories, NTT Corporation 1-1 Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847 Japan

E-mail: †{bessho.katsuji, oku.masahiro}@lab.ntt.co.jp

Abstract　A conceptual base consisting of conceptual vectors, which are semantic representations of each word generated from co-occurrence patterns with other words, is useful for topic structure extraction from text or information retrieval. However, unregistered words from the conceptual base are not assigned conceptual vectors, and cannot be applied to the processing using conceptual base. We propose a method of estimating the conceptual vectors for unregistered words based on co-occurrence information in the texts. The experimental results of topic segmentation using several articles in the newspapers show that the proposed method, using estimated vectors, can improve segmentation accuracy.

Keyword　Unregistered Word, Conceptual Vector, Topic Segmentation

1. はじめに

コーパスにおける単語間の共起頻度を記録した共起行列に対し特異値分解を行い、単語を次元数の縮退したベクトルで表現したものを概念ベクトルと呼び、単語とその概念ベクトルの対の集合を概念ベースと呼ぶ。概念ベースは、単語の意味的類似性を定量化できるため、情報検索[1,2,3]や、トピックセグメンテーション[4]等に適用され、効果をもたらしてきた。

しかしながら、処理対象テキストにおいて、単語辞書には含まれるものの、概念ベースには含まれない単語（以下、未知語と呼ぶ）には、概念ベクトルを付与することができない。

テキストの話題構造化等の言語処理で概念ベースを利用する際、話題に特徴的な単語で概念ベースにないようなものがあると、精度の低下を招く。

また、概念ベースを利用した検索処理において、検索語が検索対象テキスト中に存在しても未知語であったならば、概念ベクトルが付与されないために、全く考慮されなくなる。

コンピュータのメモリの制約から、概念ベース生成時に格納できる単語の数は限られている。また、メモリの問題を度外視しても、単語辞書中の全単語を格納する概念ベースを生成する時間は極めて長くなる。単語辞書には逐次新しいエントリが追加されることを考え合わせると、実際の運用は不可能といえる。現実的な時間で未知語にベクトルを付与する技術が必要となる。

未知語対策として、処理対象テキストから概念ベースを生成し、処理対象テキスト中の全ての単語に概念ベクトルを割り当てるという方法があるが、処理対象テキストが少量ならば情報量が少ないため、生成される概念ベースの質は低い。

また、概念ベース生成用コーパスと処理対象テキストをマージしたテキストから、処理対象テキスト中の単語は全て概念ベースに登録されるように、概念ベー

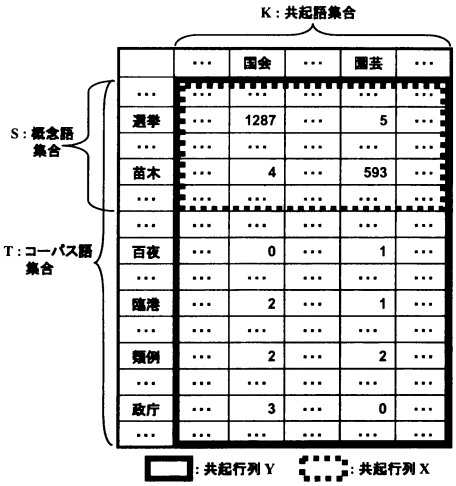

図 1 共起行列

スを生成するという方法もある。しかしながら、一般に特異値分解処理は時間がかかり、時々刻々と新しい処理対象テキストが出現するような環境で、処理対象テキストが出現するたびに概念ベースを生成するのは非効率である。また、この方法では、既に概念ベクトルを割り当てられた単語のベクトル値も変化する。既存の処理対象テキストのベクトル値は温存しておきたい場合など、既に概念ベクトルを割り当てられた単語のベクトル値は固定したままにしておきたいタスクには都合が悪い。

したがって、概念ベースを生成し直すことなく、未知語の概念ベクトルを推定する技術が必要となる。本稿では、未知語の概念ベクトル推定手法として、概念空間法と分散最小法の 2 つの手法を提案する。

以下、2 節において概念ベースの生成についての説明を行う。3 節において提案手法の説明を行い、4 節で評価実験結果を述べ、5 節でまとめを述べる。

2. 概念ベース

本節では、[5]における潜在的意味解析の手法に基づいて、概念ベースの説明をする。概念ベースの生成においては、まずコーパスを形態素解析した後、内容語以外を除去する。残った異なり単語の集合を T とする。T 中の高頻度語の集合 K をとる。T 中の単語をコーパス語、K 中の単語を共起語と呼ぶ。任意のコーパス語と任意の共起語との間の 1 文中に共起する頻度をカウントし、各行がコーパス語に対応し、各列が共起語に対応しているような共起行列 Y を作成する（図 1 参照）。共起行列の各行ベクトルは、対応するコーパス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。

但し、このままではデータのスパースネス性があることを始めとして、テキストデータから抽出される単語の情報には常に欠落があると予想されるため、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ため、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。

T の要素数は非常に多いため、コンピュータのメモリの制約から、一般に特異値分解の実行が不可能となる。そこで共起行列の行を、T 中の高頻度語の集合 S に限定して得られる共起行列 X に対して特異値分解を実行する。S 中の単語を概念語と呼ぶ。

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、$r = rank\,X \le \min(p,q)$，$U^tU = V^tV = I$（$I$：単位行列）であり、$\Sigma = (\delta_{ij})$ としたとき、

$\delta_{ii} \ge \delta_{jj} > 0 \left(1 \le^{\forall} i \le^{\forall} j \le r\right)$，$\delta_{ij} = 0\,(i \ne j)$ である。$\delta_{ii}\,(1 \le i \le r)$ を X の特異値と呼ぶ。

V^t の r 個の行ベクトルは、q 次元空間中の正規直交基底であり、X の第 i 番目の q 次元行ベクトルは、この正規直交基底の張る r 次元部分空間において、$U\Sigma$ の第 i 番目の r 次元行ベクトルで表される。

ここで、$1 \le^{\forall} r' \le r$ 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。$U'\Sigma'$ の第 i 番目の行ベクトルは、$U\Sigma$ の第 i 番目の行ベクトルの 1 番目から r' 番目までの座標をとったものであり、$U\Sigma$ の第 i 番目の行ベクトルを、V'^t の行ベクトルの張る r' 次元部分空間に射影して得られるものである。X の第 i 番目の q 次元行ベクトルは、$U'\Sigma'$ の第 i 番目の r' 次元行ベクトルに射影される。V'^t の行ベクトルが張る r' 次元部分空間は、X の各行ベクトルとその射影した点との距離の自乗和が最小となる r' 次元部分空間であり、その意味で X の行ベクトルの分布に最もあてはまりのよい r' 次元部分空間である。この

V'^tの行ベクトルが張るr'次元部分空間を概念空間と呼ぶ。$U'\Sigma'$の各行ベクトルは、U'の対応する行ベクトルを、各座標ごとに対応する特異値の割合で伸縮したものである。U'の行ベクトルをその長さで割って単位ベクトルに正規化したものを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概念ベースと呼ぶ。

3. 提案手法

未知語の概念ベクトル推定手法として、概念空間法と分散最小法の2つの手法を説明する。概念空間法は概念ベースの生成過程を利用する手法であり、分散最小法はトピック区間内の概念ベクトルの分布の性質を利用する手法である。

3.1. 概念空間法

概念空間法の基本的な考え方は、未知語に対しても共起ベクトルを生成し、その共起ベクトルを概念空間へ射影したものを概念ベクトルとしようというものである。特異値分解の対象となる点の集合が異なれば、概念空間も異なり、その結果、対応する概念ベクトルも異なってくる。しかし、点集合の分布の差異がそれ程大きくなければ、Xから得られた概念空間をそのまま使っても、概念空間におおきなずれはなく、したがって射影した点は、真の概念ベクトルと大きくずれることはなく、精度よく真の概念ベクトルが推定できる。

$XV=U\Sigma$より$XV'=U'\Sigma'$となる。これに倣い、q次元ベクトルh_wに対しh_wV'としたものは、各成分がV'^tの対応する行ベクトルとh_wとの内積であるため、確かに概念空間へのh_wの射影となっている。従って、$h_wV'\Sigma'^{-1}$の長さを1に正規化したものが、h_wから得られる概念ベクトルに相当する。

実際には、処理対象テキストが与えられたとき、処理対象テキスト中の異なり単語（概念語も含む）の推定概念ベクトルを、以下の手順で求める。

処理対象テキスト中の異なり単語の集合Wをとる。W中の単語を対象語と呼ぶ。任意の対象語と、任意の共起語（概念ベース生成用コーパスから導出されるK中の語のことである）との間の、処理対象テキストにおける1文中に共起する頻度をカウントし、各行が対象語に対応し、各列が共起語に対応しているような共起行列Zを作成する。

対象語wのZにおける共起ベクトルをh_{wZ}とする。wのYにおける共起ベクトルをh_{wY}（wがTに存在しない場合、$h_{wY}=0$とする）としたとき、$(h_{wZ}+h_{wY})V'\Sigma'^{-1}$の長さを1に正規化したものを、$w$の推定概念ベクトルとする。但し、$h_{wZ}+h_{wY}=0$である$w$には、概念ベクトルは付与されないものとする。

これは、概念ベース生成用コーパスと処理対象テキストとをマージして得られるテキストから導出される共起ベクトルを、Xから得られた概念空間へ射影したものを推定概念ベクトルとすることを意味する。従って概念語については、共起ベクトルが、Yにおける共起ベクトルと異なれば（即ち、$h_{wZ}\neq 0$であれば）、推定概念ベクトルは、概念ベース中の概念ベクトルとは違ったものとなる。もちろん概念空間法においては、概念語についてはベクトル推定をせず、概念ベース中の対応する概念ベクトルをそのまま使用することも可能である。

3.2. 分散最小法

[4]においては、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いており、この仮定のもとにトピック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。従って、トピック区間内の未知語を含めた単語の概念ベクトルは近似的に正規分布に従っていると推察される。トピック区間内の概念ベクトルが正規分布に従うとき、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、クラスタ間分散が最大となる。

[6]においては、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上で、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方法が提案されている。本稿で提案する分散最小法は、処理対象テキスト中の1文における単語集合をクラスタとし、概念語の概念ベクトルを固定した上で、クラスタ間分散が最大となるように、未知語の概念ベクトルを付与するものである。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障されている最大の範囲だからである。

分散最小法の手順を説明する前に、いくつかの定義を述べる。

処理対象テキスト中の異なり単語である対象語の集合を、$W=\{w_1, w_2, \cdots, w_m\}$と表す。また、クラスタ集合である、処理対象テキスト中の文集合を$C=\{c_1, c_2, \cdots, c_n\}$とする。文$c_j$内の対象語$w_i$の出現回数を$z(w_i \mid c_j)$とする。また、以下の値を定義する。

・ w_iの出現回数

$$z(w_i) := \sum_{1\le j\le n} z(w_i \mid c_j) \qquad (3)$$

・c_j内ののべ単語数　$z(c_j) := \sum_{1 \le i \le m} z(w_i \mid c_j)$　(4)

・のべ単語数　$z(A) := \sum_{1 \le i \le m} z(w_i) = \sum_{1 \le j \le n} z(c_j)$　(5)

対象語w_iの概念ベクトルを$v(w_i)$とし、以下の値を定義する。

・平均　$$\mu := \frac{\sum_{1 \le i \le m} z(w_i) \bullet v(w_i)}{z(A)} \quad (6)$$

・c_jの平均　$$\mu(c_j) := \frac{\sum_{1 \le i \le m} z(w_i \mid c_j) \bullet v(w_i)}{z(c_j)} \quad (7)$$

以下の関係式が成り立つ。

$$\sum_{1 \le i \le m} z(w_i) \bullet \|v(w_i) - \mu\|^2 = \sum_{1 \le j \le n} \sum_{1 \le i \le m} z(w_i \mid c_j) \bullet \|v(w_i) - \mu(c_j)\|^2 + \sum_{1 \le j \le n} z(c_j) \bullet \|\mu(c_j) - \mu\|^2 \quad (8)$$

上式の左辺を全変動、右辺の第一項をクラスタ内変動、第二項をクラスタ間変動と呼ぶ。全変動を固定した場合、クラスタ内変動が最小となるとき（即ちクラスタ内分散の全クラスタにわたる平均が最小となるとき）、文集合Cはクラスタ群として最適となる。このとき、クラスタ間変動及びクラスタ間分散は最大となる。

分散最小法では、処理対象テキストが与えられたとき、処理対象テキスト中の未知語の推定概念ベクトルを、以下の手順で求める。

処理対象テキスト中の概念語w_iに、Xから生成した概念ベース中の対応する概念ベクトルを割り当てる。ここで、概念語のベクトルの平均が零ベクトル、各成分の分散が1となるように変数変換をし、変換後のベクトルを$v(w_i)$とする。

処理対象テキスト中の未知語w_iに、未知ベクトル$v(w_i)$を割り当てる。未知語のベクトルの平均が零ベクトル、各成分の分散が1であるという制約条件の下で、クラスタ間変動

$$\sum_{1 \le j \le n} z(c_j) \bullet \|\mu(c_j)\|^2 \quad (9)$$

を最大にする未知ベクトル$v(w_i)$を求める。この変数変換と制約条件は、対象語の平均を零ベクトル、各成分の分散を1とするためである。

処理対象テキスト中の概念語については、概念ベース中の対応する概念ベクトルを、未知語については、推定したベクトル$v(w_i)$に対し、上述の変数変換の逆変換を行い、長さ1に正規化したものを割り当てる。

4. 評価実験

推定ベクトルの妥当性を検証するため、処理対象テキスト中の単語の概念ベクトルの分布を利用したトピックセグメンテーションの評価実験を行った。未知語の概念ベクトルが的確に推定されるならば、未知語に対しても推定概念ベクトルを付与することにより、正解トピック区間内の概念ベクトル集合が、より鮮明にクラスタの形状をなし、セグメンテーション精度が上昇すると期待される。

4.1. 評価データ

概念ベース生成用コーパスは、毎日新聞2000年版1年分の106614記事のヘッドラインと本文を用いた。コーパス語は109953個であった。共起語として出現頻度数が上位51番目から1500個をとり、コーパス語と共起語から共起行列Yを生成した。共起語として上位50個を排除したのは、高頻度語であるが故に多数のコーパス語と共起し、コーパス語の共起パターンを差異化する力が弱いと考えられたからである。また、概念語として出現頻度数が上位の30000個をとり、これと共起語から生成される共起行列Xに対し、縮退後の次元数が200となるように特異値分解を実行した。

処理対象テキストは、毎日新聞2001年版の社会面の記事100個の本文の部分を連接したテキストと、国際面の記事100個の本文の部分を連接したテキストとした。ともに、1記事を1トピックと仮定している。

各処理対象テキストに関する情報を表1に示す。$h_{wZ} + h_{wY} > 0$の未知語とは、共起語と共起する未知語であり、概念空間法により概念ベクトルを付与される未知語である。さらにその中の$h_{wY} > 0$である未知語は、概念ベース生成用コーパスにおける共起ベクトルh_{wY}も利用できる未知語である。

表1 処理対象テキストの情報
（未知語の括弧内は全単語に対する割合）

対象テキスト		社会面	国際面
文数		838	1079
トピック数		100	100
単語	のべ数	7771	13116
	異なり数	3384	3600
未知語	のべ数	418(5.4%)	292(2.2%)
	異なり数	310(9.2%)	197(5.5%)
$h_{wZ} + h_{wY} > 0$, $h_{wY} > 0$の未知語	のべ数	381	279
	異なり数	276	190
$h_{wZ} + h_{wY} > 0$, $h_{wY} = 0$の未知語	のべ数	36	13
	異なり数	33	7
$h_{wZ} + h_{wY} = 0$の未知語	のべ数	1	0
	異なり数	1	0

4.2. 提案手法の評価実験

各処理対象テキストに対し、クラスタ内変動最小アルゴリズム[4]により、セグメンテーションを行った。クラスタ内変動最小アルゴリズムでは、まず処理対象テキスト中の単語を概念ベース中の対応する概念ベク

トルに変換する。次に、文の系列を分割するクラスタ列で、任意の分割数に対し、クラスタ内変動が最小となるものを動的計画法による求める。得られたクラスタ列をトピックセグメント列とする。

概念ベースのみを用いる手法では、処理対象テキスト中に、概念ベクトルを割り当てられない未知語が存在する。概念ベースのみを用いる手法の他に、概念空間法、分散最小法により処理対象テキスト中の単語に概念ベクトルを付与した上でセグメンテーションする手法を実行した。また、処理対象テキスト中の単語頻度情報に基づくHearst法[7,8]によるセグメンテーションも行った。評価実験においては、出力トピック数を正解トピック数と同じにした上で、各手法間の精度差をみることとした。

処理対象テキスト中の任意の概念語 w に対し、$h_{wZ}+h_{wY}\neq 0$であった。概念空間法の処理では、概念語に対してもベクトル推定を行い、推定したベクトルの方を用いた。

分散最小法におけるクラスタ間変動を最大にする未知ベクトルを求める処理は、数値計算ソフトであるMATHEMATICA 5[9]を用いて行った。計算量を抑えるために、処理対象テキストの文を順次取り込み、未知語の異なり数がはじめて100以上になった時点で、それまでに取り込んだテキストに対して推定処理を行い、未知語とその推定概念ベクトルを概念ベースに追加するという操作を繰り返した。また、制約条件が成分間で独立であり、クラスタ間変動の式(9)は各成分ごとの値の和なので、各成分ごとに最大となる未知数を求めるようにした。

各手法の精度を表2に示す。精度は正解トピック境界集合と出力トピック境界集合との間の再現率と適合率から導出されるF値である。正解範囲は、正解とみなす正解トピック境界からの範囲を意味する。

表2 セグメンテーション精度
(括弧内は概念ベースのみの手法との差)

対象テキスト	社会面		国際面	
正解範囲	±0	±1	±0	±1
概念空間法	81.8% (+2.0%)	88.4% (+4.6%)	89.9% (+2.0%)	97.5% (+1.0%)
分散最小法	81.8% (+2.0%)	87.3% (+3.5%)	88.9% (+1.0%)	97.0% (+0.5%)
概念ベースのみ	79.8%	83.8%	87.9%	96.5%
Hearst法	46.5%	59.2%	42.4%	59.9%

Hearst法よりも、概念ベースのみを用いるクラスタ内変動最小アルゴリズムの方が精度は高い。概念空間法、分散最小法によりベクトル推定を行うことにより、概念ベースのみを用いる手法よりさらに精度が向上している。表1における各処理対象テキスト中の未知語のべ数の、単語のべ数に占める割合に、概念ベースのみの精度を乗じただけの精度上昇がほぼ認められる。

4.3. 考察

概念空間法の長所として、処理対象テキストにおける共起ベクトル h_{wZ} のみならず、概念ベース生成用コーパスにおける共起ベクトル h_{wY} も利用できることが挙げられる。$h_{wY}>0$の未知語ののべ数は、全未知語ののべ数の91.1%ないし95.5%を占めているため、この事実が概念空間法の方が分散最小法よりも、精度がやや高い要因の一つと考えられる。

また概念空間法は、概念空間が求まれば、それに射影すればよいだけなので、推定処理時間が短い。

一方、概念空間法は、$h_{wZ}+h_{wY}=0$である単語には、概念ベクトルを付与することができない。

これに対し、分散最小法は、$h_{wZ}+h_{wY}=0$である未知語も含め、全未知語に概念ベクトルを付与することができるという原理的な長所がある。このことは、特に検索処理において重要である。

しかしながら、分散最小法は未知語の数が多いと推定処理時間が長くなる。

表1のように、$h_{wZ}+h_{wY}=0$である未知語は、$h_{wZ}+h_{wY}\neq 0$である未知語と比べ数が少ない。そこで、概念空間法により、共起語と共起する単語にベクトルを付与してから、残りの共起語と共起しない単語に分散最小法によりベクトルを付与するハイブリッド方式が好ましいといえる。ハイブリッド方式の精度は、いずれの処理対象テキストでも、概念空間法の精度と同じであった。

5. まとめ

本稿では、概念ベースに含まれない未知語の概念ベクトルを、テキスト中の共起情報から推定する手法として、概念空間法と分散最小法を提案した。推定ベクトルを用いた新聞記事に対するトピックセグメンテーションの評価実験の結果、用いない手法よりも精度が向上することを確認した。また、両手法を組み合わせたハイブリッド方式が、精度や速度の点で好ましいとの結論に達した。

今後は、Web上の掲示板やブログ等、より未知語が頻出するデータに適用し、提案手法の有効性と問題点を検証していく予定である。

文　献

[1] H. Schutze, and J.O. Pedersen, A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval, Proc. RIAO '94, pp.266-274, 1994.

[2] T. Kato, S. Shimada, M. Kumamoto and K.

Matsuzawa, Idea-Deriving Information Retrieval System, Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.187-193, 1999.

[3] 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.

[4] 別所克人, "クラスター内変動最小アルゴリズムに基づくトピックセグメンテーション, "情報処理学会研究報告, Vol.SIG-NL 154, pp.9-16, 1999.

[5] H. Schutze, Dimensions of Meaning, Proc. Supercomputing '92, pp.787-796, 1992.

[6] 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.

[7] M. A. Hearst, Multi-Paragraph Segmentation of Expository Text, 32nd Annual Meeting of the Association for Computational Linguistics, pp.9-16, 1994.

[8] M. A. Hearst, TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, Computational Linguistics, Vol.23, No.1, pp.33-64, 1997.

[9] MATHEMATICA 5, http://www.wolfram.com/