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

2006－ICS－144 (5)
2006／7／28

単語・意味属性間共起に基づく概念ベースの拡張方式

別所 克人[*1] 内山 俊郎[*1] 片岡 良治[*1]

[*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

コーパスから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。本稿では、未登録語の概念ベクトルを推定する方式、すなわち概念ベースを拡張する方式を提案する。概念ベクトルを用いた文書検索の評価実験により、拡張した概念ベースを用いることで、検索精度が向上することを検証した。また、単語・意味属性間共起に基づく概念ベースは、単語間共起に基づく概念ベースよりも高精度であるが、この拡張によっても優位性は変わらないことを確認した。

Expansion Method of Concept Base Based on Co-occurrences between Words and Semantic Attributes

Katsuji Bessho[*1] Toshio Uchiyama[*1] Ryoji Kataoka[*1]

[*1] NTT Cyber Solutions Laboratories, NTT Corporation

A concept base generated from a corpus provides the concept vectors, which are the semantic representations of words. We propose a method that estimates the concept vectors of unregistered words and that expands the concept base. The experimental results of document retrieval using the concept vectors showed that the expansion of concept base improves retrieval accuracy. The results also showed the advantage of the concept base based on co-occurrences between words and those semantic attributes against that based on co-occurrences between words.

1. はじめに

コーパスにおける単語同士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、テキスト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。本稿では、概念ベースに登録されていない単語の概念ベクトルを推定する手法について論じる。

単語間共起に基づく概念ベースの生成においては、特異値分解処理で一般に多量の計算量を要するため、共起頻度をとる単語の集合を制限する必要がある。このため、生成された概念ベクトルの質に問題があった。

[5]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方式が提案されており、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概念ベースをもとにしたものに対し評価が行われ、効果が確認されている。

[6]においては、[5]の考えに基づき、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記録した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に対し評価を行い、単語間共起に基づくコーパス概念ベースよりも高精度であることを検証した。

この単語・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における共起行列の各列に対応する任意の単語を考慮し、計算量を増やすことなく、概念ベクトルの質を向上させることができる。しかし、各行に対応する単語の集合は、特異値分解の計算量の制約上、制限されたままであるため、コーパス中の単語でその概念ベクトルを生成できないものが存在する。

本稿では、未登録語の概念ベクトルを推定し、未登録語とその推定概念ベクトルを概念ベースに追加することにより、概念ベースを拡張する方式を提案する。この推定手法は、概念ベクトルの分散最小性に基づくもので、[7]で提案されている同種の手法と比べ、より高速・高精度な性質を持つ。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、拡張方式を適用し、生成した概念ベクトルを用いた文書検索の精度を比較した結果を報告する。

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紹介し、3章で提案する推定手法を述べる。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

2.1 単語間共起に基づく手法

本章では、[8]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述べる。

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。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作成する。共起行列から零ベクトルである行ベクトルを削除する。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

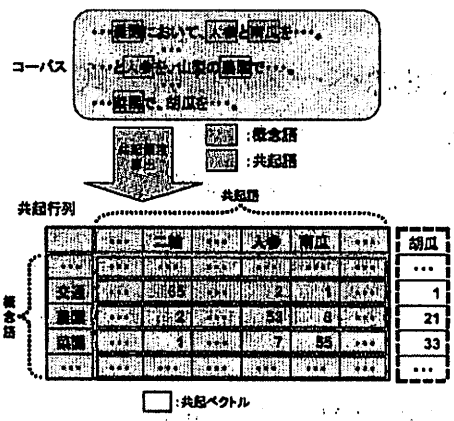

図 1: 単語間共起行列

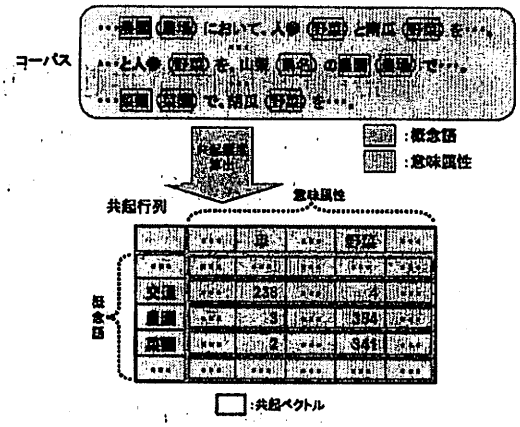

図 2: 単語・意味属性間共起行列

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(図 1)。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、共起ベクトルを利用した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。

G, K の要素数が多いと、特異値分解の計算量は多量になるため、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上で特異値分解を実行する。ここで、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得られる行列 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。

$$\underset{p\times q}{X} = \underset{p\times r}{U}\ \underset{r\times r}{\Sigma}\ \underset{r\times q}{V^t} \qquad (1)$$

ここで、添字 t は行列の転置を表す。

$r = rank\,X \le \min(p,q)$, $U^tU = V^tV = I$ (I :単位行列)であり、$\Sigma = (\delta_{ij})$ としたとき、$\delta_{ii} \ge \delta_{ij} > 0 \left(1 \le^{\forall} i \le^{\forall} j \le r\right)$, $\delta_{ij} = 0\,(i \ne j)$ である。$\delta_{ii}\ (1 \le i \le r)$ を X の特異値と呼ぶ。

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、

$$\underset{p\times q}{X'} = \underset{p\times r'}{U'}\ \underset{r'\times r'}{\Sigma'}\ \underset{r'\times q}{V'^t} \qquad (2)$$

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

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。例えば、図 1 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関わらず、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。例えば、図 1 の"胡瓜"との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、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同士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[9]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、2715 個のノードからなる。

本稿の手法では、形態素解析プログラムとして JTAG[10]を用いているが、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。一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。

任意の概念語と意味属性とが 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、各列が意味属性に対応しているような共起行列を作成する(図 2)。

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、同一の意味属性をも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。例えば、図 1 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。

また、意味属性の数は高々2715 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、より豊富な情報をもつようになる。例えば、図 1 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。

概念語・意味属性間共起行列中の各成分をその平方根に変換して得られる行列 X に対して特異値分解を実行する。その結果得られる(2)式における 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 章で述べた手法では、特異値分解による計算量の制約のため、共起行列の行数を制限する必要がある。共起行列の行に対応する単語の集合から漏れた単語については、概念ベクトルが付与されない。この結果、概念ベースを利用した言語処理において、概念ベースに登録されていない未登録語の概念は一切考慮されないため、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、意味空間への射影による手法[11]と、概念ベクトルの分散最小性に基づく手法について述べる。

3.1 意味空間への射影による手法

[11]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、(2)式における $\underset{r'\times q}{V'^{t}}$ の r' 個の行ベクトルが張る空間である。(1)式より、

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

であるため、

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

となる。これに倣い、任意の共起ベクトル $\underset{1\times q}{h_w}$ に対し、

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

とおいたものは、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、(3)式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 分散最小性に基づく手法

[7]で提案されている分散最小性に基づく手法は、コーパス中の単語について、概念ベースの登録語に対しては、その概念ベクトルを割り当て、未登録語に対しては、変数としての概念ベクトルを割り当てた上で、各文内の概念ベクトルの分散の和が最小となる未登録語の概念ベクトルを推定概念ベクトルとする。

この手法では、未登録語の異なり数だけの変数をもつ 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、一般的なコンピュータでは実行が不可能となる。

このため、文を一つずつとっていき、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質に問題があった。

本稿で提案する推定手法は、着目している一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上で、各文内の概念ベクトルの分散の和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というものである。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 $z(w_i \mid c_j)$ とする。また、c_j でののべ単語数を、

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

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 番目の座標を $v^m(w_i)$ とする。

c_j での m 番目の座標の平均を

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

と定義する。

各文 c_j における、概念ベクトルの平均と、各概念ベクトルとの距離の自乗の和を、文集合 C の全文にわたって加算した和は、以下の式(4)で表される。

$$\sum_{1\le j\le g}\ \sum_{1\le i\le x+1} z(w_i \mid c_j) \sum_{1\le m\le f} \left(v^m(w_i)-\mu^m(c_j)\right)^2$$
$$= \sum_{1\le m\le f}\ \sum_{1\le j\le g}\ \sum_{1\le i\le x+1} z(w_i \mid c_j)\left(v^m(w_i)-\mu^m(c_j)\right)^2 \qquad (4)$$

式(4)を最小にする $v^m(w_{x+1})$ $(1 \le m \le f)$ を求めるには、任意の座標 m $(1 \le m \le f)$ に対し、

$$\sum_{1 \le j \le g} \sum_{1 \le i \le x+1} z(w_i \mid c_j)\left(v^m(w_i) - \mu^m(c_j)\right)^2 \qquad (5)$$

を最小にする $v^m(w_{x+1})$ を求めればよい。

式(5)は以下のように変形される。

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

$$= \sum_{1 \le j \le g} \left[\sum_{1 \le i \le x} \left[z(w_i \mid c_j) \bullet \left(v^m(w_i) - \frac{\sum_{1 \le p \le x} z(w_p \mid c_j) \bullet v^m(w_p)}{z(c_j)} - \frac{z(w_{x+1} \mid c_j)}{z(c_j)} \bullet v^m(w_{x+1}) \right)^2 \right] + z(w_{x+1} \mid c_j) \bullet \left(- \frac{\sum_{1 \le p \le x} z(w_p \mid c_j) \bullet v^m(w_p)}{z(c_j)} + \left(1 - \frac{z(w_{x+1} \mid c_j)}{z(c_j)}\right) \bullet v^m(w_{x+1}) \right)^2 \right]$$

$$= \sum_{1 \le j \le g} \left(a_j \bullet v^m(w_{x+1})^2 + b_j \bullet v^m(w_{x+1}) + d_j \right) \qquad (6)$$

ここで、

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

とおくと、a_j, b_j は、以下のように表される。

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

$$= z_k(c_j) \left(\frac{z(w_{x+1} \mid c_j)}{z(c_j)} \right)^2 + z(w_{x+1} \mid c_j) \left(\frac{z_k(c_j)}{z(c_j)} \right)^2$$

$$= z_k(c_j) z(w_{x+1} \mid c_j) \left(\frac{1}{z(c_j)} \right)^2 \left(z_k(c_j) + z(w_{x+1} \mid c_j) \right)$$

$$= \frac{z_k(c_j) z(w_{x+1} \mid c_j)}{z(c_j)}$$

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

$$= -2 \frac{z(w_{x+1} \mid c_j)}{z(c_j)^2} \left(z(c_j) \sum_{1 \le i \le x} z(w_i \mid c_j) \bullet v^m(w_i) - \sum_{1 \le i \le x} z(w_i \mid c_j) \bullet \sum_{1 \le p \le x} z(w_p \mid c_j) \bullet v^m(w_p) \right) - 2 \frac{z(w_{x+1} \mid c_j)}{z(c_j)^2} \left(z(c_j) - z(w_{x+1} \mid c_j) \right) \bullet \sum_{1 \le p \le x} z(w_p \mid c_j) \bullet v^m(w_p)$$

$$= -2 \frac{z(w_{x+1} \mid c_j)}{z(c_j)} \sum_{1 \le i \le x} z(w_i \mid c_j) \bullet v^m(w_i)$$

式(6)は、2 次式

$$a \bullet v^m(w_{x+1})^2 + b \bullet v^m(w_{x+1}) + d \qquad (7)$$

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(7)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 は座標 m に依存しないので、a も座標 m に依存しない。

m 番目の座標が $v^m(w_i)$ 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、一変数の 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、着目している一つの異なり未登録語を含む全ての文の集合から、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上では分散最小とはならないものの、情報源の量は[7]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、生成した概念ベクトル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、110,000 個の Q&A 文書と、それを包含する 1,874,553 個の Q&A 文書の 2 つを用い、コーパス量ごとの各推定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全てを、共起行列の行となる単語として、共起行列を作成した。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 2715 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一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と思われる意味属性のみを使用することとした。

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。このようにとったのは、与えられたメモリ(8GB)内で特異値分解を実行できる行数の上限がこの値であったからである。

特異値分解により、200 次元の概念ベクトルを生成した。

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。

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。

表 1：概念ベースの単語数

共起：推定手法 ＼ コーパス文書数	110,000	1,874,553
全異なり単語数	85,687	154,195
単語間：射影	84,622	152,690
単語間：分散	85,319	153,774
単語・意味属性間：射影	85,687	154,195
単語・意味属性間：分散	85,320	153,774

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。各検索対象文書において、残った単語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。

あらかじめ一つの検索対象文書と文意が同じで異なる表現の入力文書を作成する。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。

検索対象文書集合としては、単語概念ベース生成用コーパスとは共通部分をもたない 110,000 個の Q&A 文書を用いた。これを用いて 6,068 個の入力文書を作成した。

表 2：検索精度(平均逆順位)

共起：推定手法 ＼ コーパス文書数	110,000	1,874,553
単語間：拡張前	0.3067	0.3256
単語間：射影	0.3280	0.3553
単語間：分散	0.3307	0.3515
単語・意味属性間：拡張前	0.3288	0.3383
単語・意味属性間：射影	0.3539	**0.3690**
単語・意味属性間：分散	**0.3544**	0.3645

各ケースごとの平均逆順位は表 2 のようになった。

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、検索精度が向上する。

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、常に高精度となっており、概念ベースを拡張しても、前者の後者に対する優位性は変わらない。

分散最小性に基づく手法は、射影による手法と比較して、コーパス量が少ない場合は、精度がより高く、コーパス量が多い場合は、精度がより低い結果となった。

この原因として、コーパス量が少ない場合は、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、コーパス量が多い場合は、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質が高いと考えられる。

単語概念ベース生成用コーパスは、アプリケーションの適用分野と同じものにした方が、精度上、望ましい。適用分野のコーパスの量をどれだけ用意できるかに応じて、推定手法を選択するのがよいと考えられる。

5. まとめ

未登録語の概念ベクトルを推定することにより概念ベースを拡張する方式を提案し、拡張により精度が向上することを検証した。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く予定である。

参考文献

[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.
[5] 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 5 月号(JSAI), Vol.17, No.5, pp.539-547, 2002.
[6] 別所克人, 古瀬蔵, 片岡良治: 単語と意味属性との共起に基づく概念ベクトル生成手法, 人工知能学会全国大会(第 20 回), 3C3-1, 2006.
[7] 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.
[8] H. Schutze, Automatic Word Sense Discrimination, Computational Linguistics, Vol.24, No.1, pp.97-123, 1998.
[9] 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.
[10] T. Fuchi, and S. Takagi, Japanese Morphological Analyzer using Word Co-occurence-JTAG, COLING-ACL, pp.409-413, 1998.
[11] M. W. Berry, S. T. Dumais, and G. W. O'Brien, Using linear algebra for intelligent information retrieval, SIAM Review, Vol.37, No.4, pp.573-595, 1995.