自 然 言 語 処 理 154－21
（2003．3．7）

単語クラスタリングの語義判別問題への応用

佐々木 稔† 新納 浩幸‡

†茨城大学 工学部 情報工学科 ‡茨城大学 工学部 システム工学科

††〒316-8511 茨城県日立市中成沢町 4-12-1

† sasaki@cis.ibaraki.ac.jp ‡ shinnou@dse.ibaraki.ac.jp

概要

本稿では，文書要約の支援を目的としたシソーラスの自動構築を行うために，大規模な単語集合に対するクラスタリング手法の提案を行う．これまでの単語クラスタリングに関する研究は，索引語・文書行列を利用してさまざまな要素間類似度やアルゴリズムを用いてクラスタリングが行われている．この索引語・文書行列を利用した場合，索引語の分布はどのような文書内容で出現するかを統計的に示したもので，文書内における語と語の間にある意味的なつながりはそれほど強くない．そのため，結果として出力されるクラスタにはある話題に共通する単語が集まりやすくなると考えられる．意味的につながりを持つクラスタを構築するために，共起関係を持つ単語の組を抽出し，ある単語に対して意味的につながりやすい単語を統計的に表現し，それをクラスタリングすることで意味的な共通性を持つクラスタの自動構築を目指す．

Word Clustering for Word Sense Disambiguation

Minoru Sasaki† Hiroyuki Shinnou‡

†Department of Computer and Information Sciences, Faculty of Engineering, Ibaraki University

‡Department of Systems Engineering, Faculty of Engineering, Ibaraki University

4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan

† sasaki@cis.ibaraki.ac.jp ‡ shinnou@dse.ibaraki.ac.jp

Abstract

In this paper, we propose a new clustering algorithm for large scale document size to construct the thesaurus automatically in aid of summarization. The existing word-clustering systems use various similarity and clustering algorithm based on the context of the information retrieval. In case of the clustering using term-document matrix, the distribution of the index word represents the frequency of the word appearance in a certain contents of a document. Therefore, semantic relation between these words in the document is not so strong. As a result, the words which appear frequently in the contents tend to be gathered for one cluster. To construct a cluster set in which semantic relation between these words is contained, we show a word clustering using a pair of words with cooccurrence relation automatically. We further show that our clustering is effective for word sense disambiguation in comparison with using term-document matrix.

1 はじめに

ユーザが自分の検索要求を表現するためによく使われるのが自然言語であり，Lycos や Goo のようなインターネット上にある WWW サイトの検索エンジンなどで検索を行う場合，ユーザは自分の検索要求を少ない数の索引語からなる検索質問として表現している．しかし，その検索要求をユーザが正確に索引語として表現できる場合もあるが，時としてユーザの意図している索引語が見つからずに，ユーザが検索したい意味内容を持つ単語を表現できない場合もある．また，情報検索システムでは，検索質問と文書中の索引語が一致することにより検索が行われ，言い換え表現などのような概念に対して表現の多様性を考えることなしに，字面での検索が行われてしまうという問題が生じる．

また，文書要約では情報検索と同様に，基本的に文中で使われている単語の頻度や出現位置などをヒントにして要約が行われている．特に，何度も繰り返し出現する単語が存在すると，その単語に含まれる概念がその文において非常に重要な要素を持っていると考えられる．この場合，同じ単語が連続して出現すれば重要語がどの単語になるのか分かりやすいが，一般的にひとつの単語には複数の語義を持っている場合や同じ概念を表現するときに言い換えなどによって使われている単語が異なっている場合もある．

情報検索システムにおける効率的な検索作業や文書の高精度な要約結果を出力するために，シソーラスを利用することが考えられる．シソーラスは一般的に同義語や類義語を概念関係により分類・整理した辞書で，ユーザの情報要求により近くなるように検索質問へキーワードを追加するなどといった検索支援に使われている [4]．このようなシソーラスを利用するためには，計算機で扱えるような形式で構築する必要があり，この作業を人手によって行うにはデータの数が非常に多く，手間がかかるという問題が生じる．また，情報検索や文書要約の支援に対してシソーラスを用いる場合，分類語彙表で定義されているほどの階層構造はあまり必要ではなく，簡単なクラスタ構造でも十分に支援することが可能ではないかと考えられる．そのため，これらの目的に対応させた専用のシソーラスの構築を自動的に行うことが必要である．

本稿では，語義判別問題の解決を目的としたシソーラスの自動構築を行うために，大規模な単語集合に対するクラスタリング手法の提案を行う．これまでの単語クラスタリングに関する研究は，索引語・文書行列を利用してさまざまな要素間類似度やアルゴリズムを用いてクラスタリングが行われている [8]．この索引語・文書行列を利用した場合，索引語の分布はどのような文書内容で出現するかを統計的に示したもので，文書内における語と語の間にある意味的なつながりはそれほど強くない．そのため，結果として出力されるクラスタにはある話題に共通する単語が集まりやすくなると考えられる．本研究では，意味的につながりを持つクラスタを構築するために，共起関係を持つ単語の組を抽出し，ある単語に対して意味的につながりやすい単語を統計的に表現し，それをクラスタリングすることで意味的な共通性を持つクラスタの自動構築を目指す．

2 単語クラスタリング

2.1 共起データの抽出と共起関係行列の作成

単語のクラスタリングを行うために，本研究では大規模な文書データから共起関係を持つ名詞単語の組を抽出し，それを特徴要素として利用する．抽出する共起関係としては，「〜の〜」といった表現となる「名詞」＋「格助詞」＋「名詞」，複合名詞など名詞が組み合わさってまとまりをなした「名詞」＋「名詞」の 2 つのパターンを持つ名詞単語の組とする．

この共起関係を抽出する際，いくつかの例外を定義してより正確な共起関係が抽出できるようにしている．まず，名詞間に副詞や形容詞が挿入されている場合は、これらの単語は無視して上記の形に従っていれば単語の組を抽出する．また，今回の実験では名詞単語のみを考慮しているため，代名詞は共起データの候補として残さず，簡単化のため「１月」、「一月」のような月日を表す名詞、ひらがな、カタカ

ナ、アルファベット語は無視して今後対応することとした．さらに，名詞単語の中で「名詞-数」や「名詞-接尾-助数詞」と判定される名詞単語については，「1本の鉛筆を買った」や「鉛筆を1本買った」というように「1本」という語が使われる位置が前後するために共起データの候補としては考慮しなかった．

検索対象のデータに対して，茶筌を用いて形態素解析を行い，上述のパターンに従い共起データを抽出し，このデータをもとに統計的な重みを与えて共起関係行列を作成する．このとき，名詞単語の組でどちらの名詞単語を対象としてクラスタリングを行うかが問題となる．例えば「A の B」というパターンを考えた場合，名詞単語 A と B のどちらを用いるかということであるが，修飾される単語の方が重要性が高くなると考える．そのため，本研究では名詞単語 B を対象にクラスタリングを行い，以降これを対象単語と呼ぶこととする．この対象単語を数値的に表現するために，修飾単語 A の部分に出現する単語を要素とする単語ベクトルとしてベクトル空間内の点とみなす．

これらの単語を要素とする文書ベクトルを作成するとき，単語の頻度に重みを加えた数値をベクトルの要素とする．数多く提案されている重みづけ手法で，今回の実験では一般的に文書検索で用いられる対数エントロピー重み [2] を用いて重みづけを行った．この重みづけは，局所的重み L_{ij} を第 j 番目の対象単語に対する，i 番目の修飾単語への重み，大域的重み G_i を全共起データに出現する i 番目の修飾単語への重みとしてそれぞれ以下のように表す．

$$L_{ij} = \begin{cases} 1 + \log f_{ij} & (f_{ij} > 0) \\ 0 & (f_{ij} = 0) \end{cases} \tag{1}$$

$$G_i = 1 + \sum_{j=1}^{n} \frac{\frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i}}{\log n} \tag{2}$$

ここで，n は対象単語数，f_{ij} は j 番目の対象単語に共起する i 番目の単語の頻度，F_i は全体の共起データにおける i 番目の修飾単語の頻度を表す．これより，j 番目の対象単語を表す単語ベクトルにおける i 番目の修飾単語を表現した，共起関係行列 A の要素 d_{ij} は，

$$d_{ij} = L_{ij} \times G_i \tag{3}$$

となる．

2.2 共起関係行列の潜在的意味解析

前節における処理を行うことにより，n 個の対象単語と m 個の修飾単語から，$m \times n$ である共起関係行列 A が得られたとする．このとき，行列 A の階数が r であるとすると，A の特異値分解は次のように定義される．

$$A = U\Sigma V^T \tag{4}$$

ここで，$U = (u_1, \cdots, u_m)$ と $V = (v_1, \cdots, v_n)$ は $U^TU = V^TV = I_n$ を満たすユニタリ行列，Σ は $\Sigma = diag(\sigma_1, \sigma_2, \cdots, \sigma_n)$，$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq \sigma_{r+1} = \cdots = \sigma_n = 0$ を満たす対角行列で，この σ_i $(i = 1 \cdots r)$ は A の特異値と呼ばれる．これらの特異値の値 σ_i により，左特異ベクトル u_i と右特異ベクトル v_i が導き出され，A の i 番目の3つ組 $\{u_i, \sigma_i, v_i\}$ が定義される．この3つ組を用いることで，行列 A は次のように表すことができる．

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \tag{5}$$

共起関係行列 A の特異値分解 $A = U\Sigma V^T$ が得られたとすると，行列 U に含まれる左特異ベクトル $u_1, \cdots, u_m$ が修飾単語がなす空間の正規直交基底となる．また，行列 V に含まれる右特異ベクトル $v_1, \cdots, v_n$ が対象単語がなす空間の正規直交基底となる．LSI は行列 Σ に含まれる特異値の小さいものを除くことにより行列 A の次元が削減され，元のランク r よりも低いランク k である近似行列 A_k が得られる．

このように共起関係行列 A から特異値分解により近似された行列 A_k が得られたとき，A_k では対象単語がなす空間において修飾単語間の共起情報が計算されており，結果として意味的な関連付けが自動的に行われている．たとえば，「経営」と「経済」という2つの修飾単語を考えたとき，「環境」などのようなさまざまな対象単語に対してこれらの修飾単語が頻繁に共起している場合，これらの修飾単語の間には非常に深い関係が存在すると考えられる．そのため，修飾単語「経済」を含む単語ベクトルにお

いて修飾単語「経営」が存在していなくても，次元削減を行い近似行列を計算すると「経済」に引っ張られる形で意味的に関係の深い修飾単語「経営」に対しても比較的高い重みが与えられる．

2.3 単語クラスタリング

全節までの処理において得られた $m \times n$ の共起関係行列 A から単語の意味的な関係をとらえるために，同様の意味を持つ単語をいくつかのクラスタに自動分類を行う．本研究では，クラスタリングアルゴリズムとして，非階層で内積を類似度として分類を行う球面 k 平均アルゴリズム [3] を利用する．この内積計算において特異値分解の性質を利用することにより，計算コストが大幅に削減されるとともに，大規模な単語集合に対してもクラスタリングが容易に行うことが可能となる．

球面 k 平均アルゴリズムは，ユークリッド空間内における重心ベクトルとベクトル間の内積を類似度として，多次元空間の単位円を分割することによりクラスタリングを行うものである．このアルゴリズムは，ベクトル x_i の属する各クラスタ $\pi_j (1 \leq j \leq s)$ の密度を

$$\sum_{\mathbf{x}_i \in \pi_j} \mathbf{x}_i^T \mathbf{c}_j \tag{6}$$

とし，クラスタの結合密度の和を目的関数とし，以下に示す目的関数 Q が局所的に最大となるまで，ベクトル x_i のクラスタリングが繰り返される．

$$Q = \sum_{j=1}^{s} \sum_{\mathbf{x}_i \in \pi_j} \mathbf{x}_i^T \mathbf{c}_j \tag{7}$$

このとき，各クラスタにおける重心ベクトルの初期値は，対象となるベクトルの集合を任意に指定した数の部分集合に分割してできるクラスタの重心を計算することで得られる．このようにしてクラスタリングを行うアルゴリズムは，ベクトルのスパースさを逆に利用して高速に収束する利点を持ち，そのときに参照する重心ベクトルは特異値分解でできる特異値ベクトルに近いことが分かっている．

このアルゴリズムの類似度となる内積計算を行う際，単語ベクトル v_i と v_j の類似度 d_{ij} は，

$$d_{ij} = v_i \cdot v_j \tag{8}$$

となる．このとき共起関係行列を A とすると，類似度を表す行列 D は $D = A^T A$ と表現することができ，行列 A を特異値分解により得られる式 $A = U\Sigma V^T$ を用いて D を表現すると，

$$\begin{aligned} D &= (U\Sigma V^T)^T (U\Sigma V^T) \\ &= (V\Sigma U^T)(U\Sigma V^T) \\ &= (\Sigma V)^T (\Sigma V) \end{aligned} \tag{9}$$

となる [1][5]．この式は，類似度 d_{ij} の計算が特異値行列と右特異値ベクトルとの積 ΣV における第 i 列と第 j 列の内積に等しくなることを表している．また，特異値行列 Σ にある特異値の数をより少ない k 個とすると，k 次元に近似されたベクトルが得られ，LSI と同様にノイズを取り除く効果が期待されるだけでなく，類似度の計算時間が少なくなるため大規模な単語集合においても 1 回のクラスタリングで分類が可能となる．

ここで，k 次元に近似された特異値行列と右特異値ベクトルとの積ベクトル $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ を s 個のクラスタ $\pi_1^\star, \pi_2^\star, \cdots, \pi_s^\star$ に分割するためのアルゴリズムを以下に示す．

1. すべての単語ベクトルを s 個のクラスタに任意に分割する．これらの部分集合を $\{\pi_j^{(0)}\}_{j=1}^s$ とし，これより求められた重心ベクトルの初期集合を $\{\mathbf{c}_j^{(0)}\}_{j=1}^s$ とする．また，t を繰り返しの回数とし，初期値は $t = 0$ である．

2. 各単語ベクトル $\mathbf{x}_i (1 \leq i \leq N)$ に対し，余弦が最も大きい，最も単語ベクトルに近い重心ベクトルを見つける．このとき，すべての重心ベクトルは正規化されているので，余弦は単語ベクトル $\mathbf{x}_i$ と 重心ベクトル $\mathbf{c}_j^{(t)}$ の内積を求めることと同値である．これにより，前回の繰り返しで求めた重心ベクトル $\{\mathbf{c}_j^{(t)}\}_{j=1}^s$ から，単語ベクトルが新たな部分集合 $\{\pi_j^{(t+1)}\}_{j=1}^s$ に分割される．

$$\begin{aligned} \pi_j^{(t+1)} = \{\mathbf{x}_i : \mathbf{x}_i^T \mathbf{c}_j^{(t)} \geq \mathbf{x}_i^T \mathbf{c}_l^{(t)}\} \\ (1 \leq l \leq N,\ 1 \leq j \leq s) \end{aligned} \tag{10}$$

ここで，$\pi_j^{(t+1)}$ は重心ベクトル $\mathbf{c}_j^{(t)}$ に近いすべての単語ベクトルの集合とする．

3. 新たに導かれた重心ベクトルの長さを正規化する．

$$\mathbf{c}_j^{(t+1)} = \frac{\mathbf{m}_j^{(t+1)}}{||\mathbf{m}_j^{(t+1)}||}, \quad (1 \leq j \leq s) \qquad (11)$$

ここで，$\mathbf{m}_j^{(t+1)}$ はクラスタ $\pi_j^{(t+1)}$ の単語ベクトルの重心を表す．

4. 目的関数 $Q^{(t+1)}$ の値を求め．前回の繰り返しにおける目的関数の値 $Q^{(t)}$ との差を計算する．このとき，

$$\|Q^{(t)} - Q^{(t+1)}\| \leq 1 \qquad (12)$$

を満たす場合，$\pi_j^\star = \pi_j^{(t+1)}$，$\mathbf{c}_j^\star = \mathbf{c}_j^{(t+1)}$ $(1 \leq j \leq s)$ とし，アルゴリズムを終了する．停止基準を超えていない場合は，t に 1 を加え，ステップ 2 に戻る．ここで，停止基準における目的関数の差は，繰り返しでの 1 以下の差は無視できるとして便宜的に 1 という値を設定した．

3 語義判別実験

本節では，前節において述べた単語クラスタリングを用いることにより，得られる単語のクラスタが語義判別問題に対して有効であるかどうか評価を行う．語義判別問題には，SENSEVAL2 の日本語辞書タスクにおいて課題として出題された動詞 50 単語を判別実験の対象とする．判別実験では，訓練データの学習に Naive Bayes 手法を用いて語義判別の手がかりとなる素性から確率値を推定し，この確率モデルをもとにテストデータに対する判別精度を求める．

3.1 判別のための素性

語義判別を行うための素性として，[6] で設定されたものを用いている．しかし，本研究では分類語意表を使わず単語クラスタを利用しているため，クラスタ番号を素性として利用している．抽出する素性の定義は以下のようになっている．

e1: 直前の単語

e2: 直後の単語

e3: 前方の内容語 2 語まで

e4: 後方の内容語 2 語まで

e5: e3 に含まれる語のクラスタ番号

e6: e4 に含まれる語のクラスタ番号

3.2 実験

本実験において用いたデータには，毎日新聞 1994 年の新聞記事 1 年分を利用した．これらの新聞記事データに対して，タグを全て取り除き茶筌により形態素解析を行い，前節において示したように名詞単語の組を抽出する．得られた名詞単語の組に対して共起関係行列を求めるために，クラスタリングを行う対象単語が，修飾される単語の頻度をもとに統計的な重みを計算する．このとき，クラスタリングを行う対象単語は 55,597 語存在し，これらの単語を一度のクラスタリング処理によりクラスタを求めている．このようにして得られた共起関係行列に対して，特異値分解を行い，3 つの行列に分解する．このとき 93 個の特異値が得られたが，今回の実験では特異値の数を削減し次元数を減少させることはしなかった．

最後に球面 k 平均アルゴリズムを用いて単語クラスタリングを行い，得られたクラスタリング結果から語義判別実験により評価を行う．クラスタ数には，1000，3000，5000 の各クラスタ数において実験をし，単語のクラスタを自動的に構築する．日本語辞書タスクで用いられる訓練データに対して，得られた単語クラスタを素性に利用して Naive Bayes 手法にて学習を行い，テストデータに対する判別精度を求める．

この実験での判別精度を評価するために，索引語・文書行列により同様に単語のクラスタリングを行い，得られたクラスタリング結果から動詞の語義判別実験を行う．この場合についても，クラスタ数には 1000，3000，5000 の各クラスタ数についてクラスタリングを行っている．また，クラスタの番号を利用せずに単語のみを素性とした場合と素性のクラス

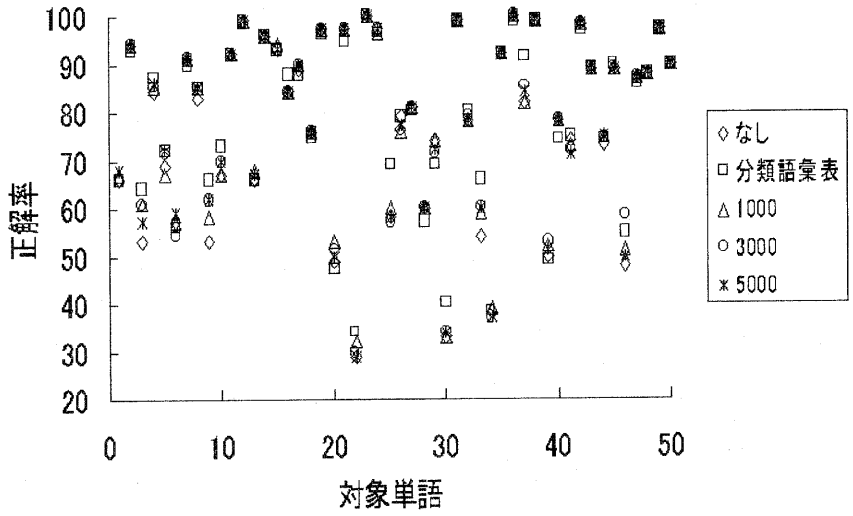

図 1: 判別対象単語に対する実験結果の比較

タ番号に分類語彙表を用いて語義判別を行った場合も同様に行い，共起関係行列を用いた場合との相対的な精度の違いを調査する．

3.3 実験結果および考察

このようにして，共起関係行列，索引語文書行列，および分類語彙表を用いて語義判別実験を行った結果，図 1 のグラフに示す正解率が得られた．正解率の算出方法には解答された結果に部分点を与える mixed-gained-scoring 方式を用いている [7]．この図において，横軸は判別対象となる動詞単語の番号を表し，縦軸はその単語における正解率を表している．この図より，クラスタ情報を何も使っていない場合でも正解率が 90% 以上となる単語に対しては，クラスタ情報を利用しても正解率に大きな変化は見られなかった．しかし，それ以下の正解率になる単語に対しては分類語彙表や単語クラスタを用いることで若干の変化が生じている．また，単語クラスタの正解率が分類語彙表を用いた場合と比較して，正解率が上昇・減少した単語数を表 1 に示す．この表を見ると，個々の判別単語に注目した場合，分類語彙表を用いた場合に比べて正解率の上昇する単語の数が若干多く，単語クラスタを用いることによる効果が分類語彙表よりも高いのではないかと考えられる．

次に，全体的な結果を見るためにそれぞれの場合における平均正解率を，表 2 と表 3 に示す．次元数 3000 において 78.2%，次元数 1000 と 5000 において 78.1% の平均正解率が得られたが，分類語彙表を用いた場合の平均正解率には及ばなかった．しかし，何も用いない場合と比較すると平均正解率は上昇していることにより，単語クラスタを用いることである程度の効果は得られる．また，索引語・文書行列を用いた場合の平均正解率と比較すると，次元数 3000 においてはほぼ同じ結果であったが，次元数 1000 と 5000 においては若干正解率が良くなっている．これより，語義判別問題に対しては，索引語・文書行列を用いるよりも共起関係行列を用いる

単語数	1000	3000	5000
上昇	21	17	20
減少	16	17	16

表 1: 分類語彙表よりも正解率が上昇・減少した単語数

次元数	1000	3000	5000	なし	分類語彙表
正解率	0.78105	0.782	0.781	0.7738	0.78785

表 2: 共起行列を用いた場合の動詞語義判別問題における平均正解率

さめうら　におの浜　…　雨ごい　汚物　加入堂山　隔日　渇水干害　間欠　紀の川　吉野川　給水　近木川　栗林公園原酒　減圧　減水　枯死　湖底　降雨　香里ケ丘　高梁川　今堅田　昆陽　昆陽池　三庄　散水　四国地建　蛇口　取水　受水槽　小河内少雨　焼け石　上水道　浄水　神流川　水がめ　水もの　水ガメ水位　…　用水　利根川　流量

図 2: 索引語・文書行列でのクラスタ例 1

ばい煙　造営　煙突　汚染　黄砂　気管支炎　…　公害　鉱業　採掘　作出　疾病　重化学　浸潤　水銀　水島　西淀川　大気　脱硫　窒素　二酸化　廃水　排ガス　排出　肺水しゅ　粉じん　有症　硫黄

図 3: 索引語・文書行列でのクラスタ例 2

方が効果的であると考えられる．

これまでは，得られたクラスタを語義判別問題に利用して数値的な評価を行った．ここでは，クラスタ自身が意味的なまとまりをなしているか検討するために，抽出されたクラスタの一例を示しながらどのような単語がまとまりをなしているかを比較する．図 2 および，図 3 は索引語・文書行列を用いた場合に得られたクラスタの例，図 4 および，図 5 は共起関係行列を用いた場合に得られたクラスタの例を表している．

依願　官職　紀久男　技師　教員　建彦　研修生　公務員　厚遇孝昌　事務職　女王蜂　昌行　職員　制服　正社員　船員　僧りょ達生　団員　智司　日勤　飛車　婦女　保母

図 4: 共起関係行列でのクラスタ例 1

安価　安上がり　以前　可能　家屋　格安　割安　割高　共有最適　手ごろ　手軽

図 5: 共起関係行列でのクラスタ例 2

図 2 に示すクラスタには，「さめうら」や「吉野川」といった場所を表す単語の他，「減水」や「水位」などのような「渇水」に関係する単語が存在している．このような単語の集合には意味的な共通性というよりも，「渇水」に関する記事内容に出現する単語が多く含まれていることが分かる．また，図 3 のクラスタには「黄砂」，「粉じん」などといった「公害」に関する単語が多く含まれている．このように，索引語・文書行列を用いてクラスタリングをした場合，特定の文書内容に関連のある単語，すなわち「関連語」がクラスタとなって現れていると考えられる．

これに対し，共起関係行列で得られたクラスタの例では，図 4 に含まれる「官職」や「事務職」といったように，職名を表す単語が多く存在している．図 5 では，「安価」や「手ごろ」といった「安さ」を意味する単語が多く集まっている．これらの例から，共起

次元数	1000	3000	5000
正解率	0.7806	0.7828	0.7804

表 3: 索引語・文書行列を用いた場合の動詞語義判別問題における平均正解率

関係行列を用いることにより，意味的にまとまりのある単語がクラスタとなっていると考えられる．ただ，誤差もあるため職名とは無関係の単語も含まれていることもあり，クラスタ自身の正確さを改善することは今後に向けて重要な課題として残っている．

これらのクラスタ例より，情報検索においてシソーラスを用いる場合は，関連語がクラスタとなるように索引語・文書行列により得られるシソーラスを用いる方が有効であると考えられる．また，文書要約や語義判別など同義語を考慮するためにシソーラスを用いる場合は，意味的なまとまりを持つように共起関係行列により得られたシソーラスを用いる方が有効であると考えられる．それぞれの目的に応じて用いるシソーラスを変更することで，より有効なシソーラスの利用が可能となることが語義判別実験においても表れている．

4 おわりに

本稿では，文書要約の支援を目的としたシソーラスの自動構築を行うために，大規模な単語集合において意味的につながりやすい単語を手がかりとしてクラスタリングを行う手法を提案した．新聞記事1年分のデータから共起関係行列を求め，単語クラスタリングを行い，クラスタリング結果から語義判別実験により評価を行ったところ，何も用いない場合と比較し単語クラスタを用いることである程度の効果が得られることが分かった．また，索引語・文書行列を用いた場合と比較して若干正解率が良くなっていることより，語義判別問題に対しては，索引語・文書行列よりも共起関係行列を用いる方が効果的であった．

語義判別問題への応用だけではなく，クラスタとなっている単語がどのような意味のまとまりをなしているかを検討したところ，情報検索においてシソーラスを用いる場合は，索引語・文書行列により得られるシソーラスを用いる方が有効で，語義判別などにおいては共起関係行列を用いる方が有効であった．このように，それぞれの目的に応じて用いるシソーラスを変更することで，より有効的なシソーラスの利用が可能となることが本実験において明らかとなった．

今後の課題としては，クラスタリングを行う際の手がかりとして「名詞」＋「動詞」の組を加えたり，対象単語が複数のクラスタに割り付けるように改良することによりクラスタの精度を高めことがあげられる．また，現在のクラスタリングではクラスタ間に関連性を持っていないため，階層的なクラスタリングを利用できるように改良を行いたい．

参考文献

[1] J. R. Bellegarda, J. W. Butzberger, and Y.-L. Chow. A novel word clustering algorithm based on latent semantic indexing. In *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96)*, pages 172–175, 1996.

[2] E. Chicholm and T. G. Kolda. New term weighting formulas for the vector space method in information retrieval. Technical report, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 1998.

[3] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. Technical report, IBM Almaden Research Center, 1999.

[4] 徳永 健伸. **情報検索と言語処理**. 東京大学出版会, 1999.

[5] 北 研二. **確率的言語モデル**. 東京大学出版会, 1999.

[6] 新納 浩幸, 佐々木 稔. 情報検索手法を利用した語義判別問題の高速解法. 情報処理学会自然言語処理研究会, 152–9, pp. 57–62, 2002.

[7] 黒橋 禎夫, 白井 清昭. Senseval-2 日本語タスク. 電子情報通信学会言語理解とコミュニケーション研究会, NLC-36〜48, pp. 1–8, 2001.

[8] 川前 徳章, 青木 輝勝, 安田 浩. 統計的モデルを用いた単語クラスタリング. 情報処理学会自然言語処理研究会, 144–8, pp. 55–60, 2001.