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

2003－ＮＬ－158　(24)
2003／11／7

単語の意味関係を抽出する統計モデル選択についての考察
－ベイズ統計に基づく単語の意味関係抽出－

川前 徳章　鈴木英明　水野修

ＮＴＴ情報流通プラットフォーム研究所
〒180-8585　東京都武蔵野市緑町 3-9-11
E-mail:　{kawamae.noriaki,suzuki.hideaki,mizuno.osamu}@lab.ntt.co.jp

あらまし　本稿の目的は、文書集合から抽出できる単語間の意味関係とそれに必要な統計モデルの対応を明らかにすることにある。既に、文書集合から抽出できる単語間の意味関係を取り扱う統計モデルはいくつか提案されてきたが、それらの統計モデルとそれらを適用することで抽出できる単語の意味関係、及び、単語間の意味の具体的な関係は論じられてこなかった。そこで、本稿では抽出可能な単語間の意味関係を具体、抽象、同概念、類義語の四つに分類し、これらの意味関係にある単語を抽出するのに最適な統計モデルの提案を行う。提案は、単語の重み付けの導入、ベイズ統計に基づいた統計モデルの二点について行う。更に、提案手法をニュース記事の文書集合を適用した結果、文書集合から抽出される単語間の意味関係と統計モデルの対応を比較することで、目的にあった単語の意味関係を抽出できることを実験によって確認した。

キーワード　EM アルゴリズム，PLSI，SAM，ベイズ統計，語彙意味表現，単語行列

The Choice of Statistical Model to Extract Lexical Meanings of Words
－Extraction of Lexical Meaning Based on Bayesian Statistics－

Noriaki KAWAMAE　Hideaki SUZUKI　and　Osamu MIZUNO

NTT Information Sharing Platform Laboratories
3-9-11,Midori-cho Musashino-shi,Tokyo 180-8585 Japan
E-mail:　{kawamae.noriaki,suzuki.hideaki,mizuno.osamu}@lab.ntt.co.jp

Abstract　The purpose of this paper is to clarify the correspondence among the lexical meaning relationship between words extracted from documents and the statistical model to extract these relationship. There are some statistical models proposed to deal with the lexical meaning relationship between words extracted from documents. However, some discussions have left about not only the correspondence among the statistical model and the lexical meaning relationship between words based on their model, but also specification of these lexical meaning relationship. We define four classes of meaning relationship as specification, abstraction, same topic and synonym, and propose the statistical model fitting to these relations in this paper. Our proposed method is composed of the dealing with the method of words weighting and the statistical model based on Bayesian Statistics. Additionally, we apply the proposed models to documents of news articles, compare the correspondents between the statistical model and the lexical meaning relationships among words based on these methods and confirm that we could extract the lexical meaning relationships meets our needs.

Keyword　EM algorithm,　PLSI,　SAM,　Bayesian Statistics,　Lexical meaning,　Word matrix

1. はじめに

1.1. 情報収集の効率化

多くの情報が文書として電子化され、ネットワーク上に蓄積されている。それら電子化された文書から、その文書に記される意味を自動的に抽出することが実現できれば、機械に文書や我々の発言の内容を理解させ、あるいは理解させた結果を我々が利用することが

可能となり、我々の情報収集をより効率化することが期待できる。

現在、情報収集として最も利用されているものの一つにサーチエンジンなどの検索システムが存在する。これらの検索システムは、データベースと同様なキーワードマッチングの技術に基づいているため、入力したキーワードを含む文書しか検索できない。従って、既存の検索システムで文章の内容によって文章を検索することは困難であり、我々の情報収集が非効率的になるおそれがある。また Semantic Web による情報収集の効率化を支援するアプローチも提案されている[1]。ただし、Semantic Web の実現にはオントロジー等の準備が必要であり、そのコスト等を考慮すると実現までには多くの問題が残されている。

以上より、文書から自動的に意味抽出をするためには、検索システムにおいてキーワードに加え内容による検索や、Semantic Web においてオントロジーの自動構築を実現しなければならない。この実現のために、現在は文書から意味関係を単語単位で抽出するアプローチがとられている。

1.2. 文書からの意味抽出

検索システムなどの計算機が文書の内容を扱うために、文書を構成する意味を抽出し、この意味を軸とした意味空間において相対的な文書や単語間の内容の類似性を表現することが有効であることが確率・統計的自然言語処理の既存研究で述べられている[4]。既存研究は意味の抽出は統計モデルを用いることで実現できると考えているが、利用した統計モデルと抽出された単語の意味関係の対応が明らかにされてこなかった。また意味関係にしても、その概念的な関係については明らかになっていない問題がある。

1.3. 統計モデルと単語の意味

本研究の目的は、文書集合から抽出できる単語間の意味関係とそれに必要な統計モデルの対応を明らかにすることにある。本稿ではこれらの意味関係にある単語を抽出するのに最適な手法の提案を行う。具体的には、単語間の意味関係を具体、抽象、同概念、類義語の四つに分類し、重み付けを導入する方法と、ベイズ統計に基づいた統計モデルを用いる方法を組合わせる。これにより、統計モデルとそれによって抽出できる単語の意味関係の対応が明らかになり、必要とする単語間の意味関係を自動的に抽出できるようになる。その結果、計算機を用いて文書の内容をより高い精度で検索できるようになり、ユーザの検索の効率化だけでなく Semantic Web の実現に寄与し、我々の情報収集を効率化することが期待できる。

2. 既存研究における統計モデル

2.1. PLSI

Hofmann[2]は単語・文書の共起行列を Aspect Model[3]を用いて確率的に圧縮する PLSI(Probabilistic Latent Semantic Indexing)を提案した。PLSI は意味的なクラス隠れ変数 $c \in C$ の存在を仮定し、文書 d における単語 w の生起確率を次のように定式化し、この式を解くことで意味を軸とした空間に、文書及び単語の確率分布を配置することができ、この空間内の類似性が意味の類似性を表すことを提案している。

$$P(d,w)=\sum_{c\in C}P(d|c)P(c)P(w|c) \qquad (1)$$

2.2. SAM

持橋[6]らにより提案された SAM(Semantic Aggregate Model)は、単語が文書集合に分かれているとは限らないことに着目し、文書によらず単語間共起関係に基づいた式を次のように提案している。ここでは PLSI と同様に $c \in C$ は意味クラスが存在を仮定し、意味的に関係のある単語 w_i と単語 w_j の共起確率を求めている。

$$P(w_i,w_j)=\sum_{c\in C}P(w_i|c)P(c)P(w_j|c) \qquad (2)$$

観測した単語の組の共起回数を $N(w_i,w_j)$として、以下に示すデータの尤度関数を最大にすることで確率分布を求める。

$$L=\sum_{i=1}^{n}\sum_{j=1}^{n}N(w_i,w_j)\log P(w_i,w_j) \qquad (3)$$

尤度関数の最大化は EM アルゴリズムを次のように用いることで最尤推定することで求めることが出来る。

E step

$$P(c|w_i,w_j)=\frac{P(w_i|c)P(c)P(w_j|c)}{\sum_{c\in C}P(w_i|c)P(c)P(w_j|c)} \qquad (4)$$

M step

$$P(c)=\sum_{i=1}^{n}\sum_{j=1}^{n}N(w_i,w_j)P(c|w_i,w_j) \qquad (5)$$

$$P(w_i|c)=\frac{\sum_{j=1}^{n}N(w_i,w_j)P(c|w_i,w_j)}{P(c)} \qquad (6)$$

これより単語の意味クラスへの帰属確率 $P(c|w_i)$をベイズの定理より次のように求めることで、単語の持つ意味を確率的に表現することが出来る。

$$P(c|w_i)=\frac{P(w_i|c)P(c)}{P(c)} \propto P(w_i|c)P(c) \qquad (7)$$

3. ベイズ統計を用いた単語の意味の抽出

3.1. 手法の概要

本稿では具体化、抽象化、同概念、類義語という単語間の意味関係を抽出と、それぞれに最適な統計モデルの提案を行う。単語の重み付け方法の導入、ベイズ統計に基づいた統計モデルの二点について、これらを選択的に用いることで、必要とする意味関係にある単語を抽出することが出来るようになる。その詳細を以下に述べる。

3.2. 重み付け

計算機で文書内容の取り扱いのために統計モデルを適用するには、まず文書を形態素解析し、文書を単語集合として表現する。次にこれらの単語に数値を与え、文書集合を行列形式で表現することができる。行列は文書と各文書を構成する単語の数値を要素とした単語・文書行列と、単語間の共起関係に着目した単語・単語行列の二通りを設計することが出来る。数値の与え方により、同じ形式のでもそれらの行列は異なってくる。

文書における単語の出現頻度はその文書の内容と関係が深いと考えられるが、同じ内容でも文書によって異なってくる。このような背景から単語の出現頻度をそのまま用いるのでなく、重みをつけて用いることが考えられる。本稿では出現頻度に偏りがある単語が、単語の意味クラスの抽出に重要と考え、単語に次のような重みを与える。PLSIにおいて、単語の生起確率をそのまま用いていたが、ここではその生起確率にこの重みを掛けて利用することを提案する。

$$w_{ij} = p_{ij} \cdot g_j = p_{ij} \cdot \left(1-\left(-\frac{1}{\log_2 n_j}\sum_{i\in n_j} np_{ij}\log_2 np_{ij}\right)\right) \qquad (8)$$

w_{ij}：文書 i の単語 j に対する重み
p_{ij}：文書 i における単語 j の生起確率
g_j：単語 j の重み
n_j：単語 j を含む文書の数
np_{ij}：正規化された文書 i における単語 j の生起確率

3.3. ベイズ統計に基づいた統計モデル

単語・文書行列及び単語・単語行列から意味クラスと文書、単語間の数値関係を抽出するためにベイズ推定に基づいた統計モデルを提案する。ベイズ推定はベイズの定理に基づいた未知母数の推定方法である。本稿でも単語の生起には、PLSI，SAM同様に単語 w_i と単語 w_j が共起するときに、その背後に共通する意味クラス $c\in C$ の存在を仮定し、その意味クラスの推定にベイズ推定を用いる。単語を観測事象、意味クラスを推定する未知母数とすることで、ベイズ推定が適用できる。意味クラス $c\in C$ に関する確率密度関数 $P(c)$ を準備して損出関数を次のように定義する。

$$L(c)=-\frac{1}{n}\sum_{i=1}^{n}\log P(w_i|c)-\frac{1}{n}\sum_{j=1}^{n}\log P(w_j|c)-\frac{1}{n}\log P(c) \qquad (9)$$

次にこれをエネルギー関数とするギブス分布を次のようにおくことができる。ここで β は逆温度定数である。

$$\rho(c)=\frac{1}{Z}\exp(-\beta nL(c)) \qquad (10)$$

ここで Z は $\rho(c)$ が確率密度関数になるための正規化定数である。このギブス分布を用いて、単語 w_i と単語 w_j の共起確率は次の確率密度関数より求めることが出来る。

$$P(w_i,w_j)=\sum_{c\in C}P(w_i|c)\rho(c)P(w_j|c) \qquad (11)$$

SAMと同様に観測した共起回数を $N(w_i,w_j)$ として、以下に示すデータの尤度関数を最大にする。

$$L=\sum_{i=1}^{n}\sum_{j=1}^{n}N(w_i,w_j)\log P(w_i,w_j) \qquad (12)$$

データの尤度関数を最大化はEMアルゴリズムを用いることで次のように最尤推定することで求めることが出来る。$\beta=1$ である場合、$\rho(c)=P(c|w_i,w_j)$ となり、SAMと同じ事後確率密度関数になる。

E step

$$P(c|w_i,w_j)=\frac{P(w_i|c)^{\beta}P(c)P(w_j|c)^{\beta}}{\sum_{c\in C}P(w_i|c)^{\beta}P(c)P(w_j|c)^{\beta}} \qquad (13)$$

M step

$$P(c)=\sum_{i=1}^{n}\sum_{j=1}^{n}N(w_i,w_j)P(c|w_i,w_j) \qquad (14)$$

$$P(w_i|c)=\frac{\sum_{j=1}^{n}N(w_i,w_j)P(c|w_i,w_j)}{P(c)} \qquad (15)$$

これより SAM と同様にベイズの定理により、意味クラスの確率 $P(c)$ と意味クラスにおける単語の生起確率 $P(w_i|c)$ から単語の意味クラスへの帰属確率分布 $P(c|w_i)$ を求めることで、単語の持つ意味を確率的に表

現することが出来る。

単語の意味を各意味クラスにおける確率分布で表現できると、単語間の意味の類似性を、それらの持つ確率分布の距離尺度である KL 情報量で数値化することが出来る。但し、この距離は非対称であり、単語 w_i から見た単語 w_j の距離と単語 w_j から見た単語 w_i の距離は異なってくる。単語 w_i から見た単語 w_j の距離を KL 情報量で数値化すると次のようになる。

$$KL(w_i|w_j)=\sum_{c\in C}P(c|w_i)\log\frac{P(c|w_i)}{P(c|w_j)} \quad (16)$$

[6]では類似度をこの KL 情報量を用いて次のような変形をして用いている。

$$\begin{aligned} e^{-D(w_i|w_j)} &= \exp\left(-\sum_{c\in C}P(c|w_i)\log\frac{P(c|w_i)}{P(c|w_j)}\right) \\ &= \exp\sum_{c\in C}P(c|w_i)\log\frac{P(c|w_i)}{P(c|w_j)} \\ &= \exp E_{P(c|w_i)}\left[\log\frac{P(c|w_i)}{P(c|w_j)}\right] \\ &= E_{P(c|w_i)}\left[\frac{P(c|w_i)}{P(c|w_j)}\right] \end{aligned} \quad (17)$$

4. 実験

4.1. 実験の目的と概要

実験の目的は行列の設計方法と、重み、適用した統計モデルとそれによって抽出できる単語の意味関係との対応を明らかにすることである。今回の実験で文書集合は日本語情報検索システム評価用テストコレクション BMIR-J2 を用いた。まず、BMIR-J2 を茶筅[3]を用いて形態素解析を行った。形態素解析によって、新聞データは単語に分割される。今回の実験で利用する単語は品詞が名詞、未知語のみを用いた。これらの単語の異なり数は 28613 個である。次にこれらの単語から単語・文書行列と単語・単語行列を作成し、これらの行列に統計モデルを適用した結果を比較評価する。

4.2. 実験結果

表 1 は今回の実験に利用した統計モデルによって抽出した意味クラスの分布の様子を示したものである。今回の実験では意味クラスの数を 100 とし、分布の様子を意味クラスエントロピーとして数値化し、温度定数 β は 0.5 と 2 の二通りを用意した。表 1 で単語・単語行列とあるものは各行列に対してベイズ統計を適用したものである。以下、同様の表記を取る。なお、ここで意味クラスエントロピーを算出するのに用いた式は次の通りである。

$$e(c)=1-\frac{H(c)}{\log|C|}=1+\frac{\sum_{c\in C}P(c|C)\log P(c|C)}{\log|C|} \quad (18)$$

この式により、意味クラスエントロピーは意味クラスの分布が偏れば 1 に近くなり、分布が均質になれば 0 に近くなる。

表 1 統計モデル別の意味クラスの分布の様子

統計モデル	意味クラス エントロピー
SAM	0.062528
単語・単語行列（β=0.5）	0.0472834
単語・単語行列（β=2）	0.144282
PLSI（β=0.5）	0.0424705
PLSI（β=1）	0.038268
PLSI（β=2）	0.0663217
重み付け PLSI（β=0.5）	0.045510
重み付け PLSI（β=2）	0.184444

表 1 より、意味クラスエントロピーを 1 に近づけるために、行列はベイズ統計を単語・単語行列あるいは単語・文書行列に適用し、及び逆温度定数 β を高くすることがよいことが分かる。この理由は[6]にもあるように単語・単語行列が文書間を超えた意味クラスを抽出できることによるものと考えられる。逆温度定数を下げることは、統計物理学において温度を冷却させることに相当する。例えば、熱した金属を徐々に冷却していくと、内部熱エネルギーが最低の状態に落ち着くことに相当する。この実験では、温度を下げること、つまり逆温度定数を上げることで、特定の意味クラスに分布が集中することで意味クラスの意味クラスエントロピーを 1 に近づけることができたと解釈できる。ただし、あまりにも急速に温度を下げると内部エネルギーが局所的に最低、つまり、特定の意味クラスのみに分布が集中することになる。

表 2 にモデル毎の単語の Semantic Entropy の計算例を示す。Semantic Entropy は次の式で与えられる。

$$e(w_i)=1-\frac{H(w)}{\log|C|}=1+\frac{\sum_{c\in C}P(c|w_i)\log P(c|w_i)}{\log|C|} \quad (19)$$

Semantic Entropy の定義からその値が高い単語ほど帰属する意味クラスに偏りがあり、低いほど全ての意味クラスに帰属する。この結果から、単語の Semantic Entropy により意味クラスの特定が可能になることが分かる。

表 2 モデル毎の単語の Semantic Entropy の高い単語上位 10 件

SAM	単語・単語行列 (β=2)	PLSI (β=2)	重み付け PLSI (β=2)
領内	的確	いろいろ	いろいろ
連銀	適当	かん	お客様
連結	哲郎	くじ	かぎ
労	徹	こだわり	けが
労災	撤去	アフリカ	たくさん
網谷	撤退	クルマ	におい
累積	鉄	サンプル	ガイドライン
利一郎	典型	セントラルア	クルー
利下げ	天下り	パート	クロ
涙	展示	タイムリミット	グ
		タンク	

図 1 から 3 までは任意に抽出した各単語の意味クラスへの帰属確率 P(c|w)の分布の様子を示す。実際のクラスは 100 あるが、ここでは各単語に共通する 10 の意味クラスについての確率分布である。同じ番号の意味クラスは同じ意味クラスを示している。この図から各単語は各意味クラスにおいて異なった P(c|w)の分布の形状を取っていることが分かる。

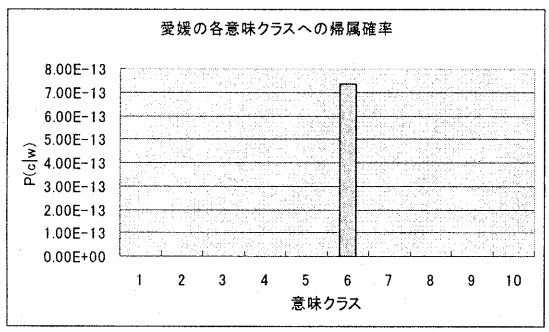

図 1 単語「愛媛」の各意味クラスへの帰属確率の分布

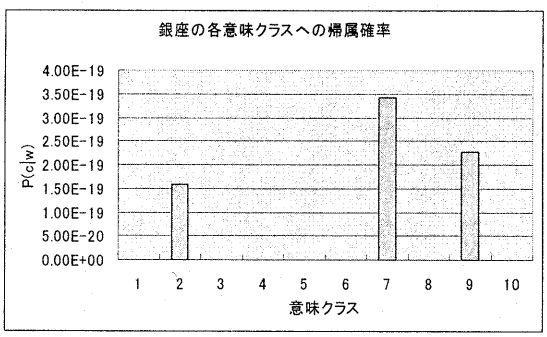

図 2 単語「宇宙開発事業団」の各意味クラスへの帰属確率の分布

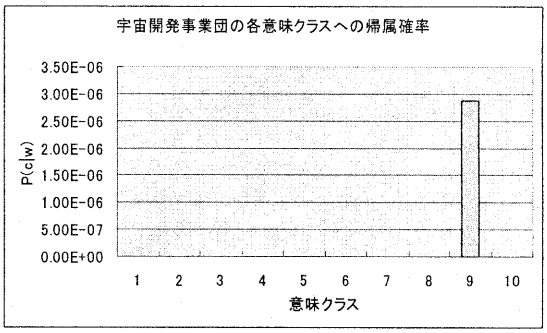

図 3 単語「宇宙開発事業団」の各意味クラスへの帰属確率の分布

表 3 単語「株」に対しての各単語の類似度

統計モデル	株券	株式会社	株主	株数
SAM	0.54502	0.43278	0.47494	0.51449
	0.42620	0.40890	0.42260	0.49600
単語単語行列 β=0.5	0.99801	0.99727	0.99763	0.99800
	0.99800	0.99727	0.99760	0.99797
単語単語行列 β=2	4.3199e-20	7.5204e-29	8.4246e-24	7.1074e-33
	0.023296	0.037070	2.0308e-67	1.4066e-52
PLSI β=0.5	0.98959	0.99151	0.99034	0.99301
	0.98961	0.99172	0.99051	0.99306
PLSI β=1	2.1001e-03	0.089300	0.022566	0.15401
	0.054068	0.11004	0.12691	0.28400
PLSI β=2	2.9224e-147	6.0125e-72	2.5452e-53	2.2437e-75
	3.6534e-38	3.1508e-36	5.7163e-19	2.2180e-19
重み付け PLSI β=0.5	0.98501	0.99118	0.98336	0.98828
	0.98535	0.99132	0.98392	0.98890
重み付け PLSI β=2	8.0564e-100	2.8205e-35	1.4573e-51	4.1883e-51
	5.2789e-51	0.11976	0.18269	1.7067e-18

表 3 に単語「株」に対して単語「株券」、「株式会社」、「株主」及び「株数」の類似度を式 (19) によって測定したものを示す。各統計モデルに対して、上の行は「株」から各単語に対しての類似度であり、下の列は各単語から「株」に対しての類似度になっている。全統計モデルは共通して、同じ単語間でも距離は非対称になっている。SAM の「株券」を例にとると、「株」から「株券」は「株券」から「株」よりも類似度が高いために、近いということになる。また、逆温度定数を低くすると単語間の類似度は共に高くなり、対称性は失われるが、逆温度定数を高くすると、単語間の類似度は極端に非対称になる。これは表 1 の結果で述べたように、温度を低くすることによって単語が特定の意味クラスへの帰属確率が高くなったためであると考えられる。この類似度の値は概念の上下関係に対応するものと考えられる。

表 4　統計モデル別の単語「愛媛」に対しての類似度の高い単語

SAM	単語単語 行列 β =0.5	単語単語 行列 β =2	PLSI β=0.5	PLSI β=1	PLSI β=2	重み付け PLSI β=0.5	重み付け PLSI β=2
領土	領土	落下	四国	創出	チャンネル	府	商い
路	路	躍進	キロ	直結	厳重	暮れ	お客様
漏れ	漏れ	李	部	キロ	無償	港	銅
老齢	老齢	問屋	納入	方面	景観	トン	著
論調	論調	要旨	私的	やり方	出力	焼却	好感
論理	論理	誘発	板	社会	評判	取り組み	北九州
和	和	夜間	要旨	超	フィリピン	層	カーブ
和解	和解	利率	歳	走行	代金	うち	増発
和男	和男	門戸	乗客	豊田	左	キロ	松山
和田	和田	余剰	正	相当	茨城	地	運転手

表 5 に単語「愛媛」に対し、式(19)の類似度を測定する式を用いて、類似度が高い単語上位 10 個を示したものである。更に、他の単語についてどのような意味関係にあるかを調べた結果、次のようなことが明らかになった。各統計モデルによって抽出できる単語間の意味関係は殆どが同概念あるいはそれ以外に相当することが分かる。具体的な意味関係に関しては統計モデルより逆温度定数に依存する。逆温度定数を高めに設定すれば、単語間の類似度の非対称性が明らかになり、単語の概念の上下関係、つまり具体的か抽象的かが明らかになる。また適用する統計モデルの違いについて比較すると通常の PLSI よりは重みを付けた PLSI の方が抽出できる意味関係が具体、抽象、同概念、類義語のいずれかの関係にある確率が高くなる。これは提案した重みによって単語間の持つ意味クラスについての情報が顕在化され、更に逆温度定数を高めにすることで、表 1 の結果に示すように意味クラスエントロピーが 1 に近づく、つまりいくつかの意味クラスに特徴的に帰属することでこのような結果が得られたと考えられる。抽象については具体の関係を逆に使うことで抽出することが出来る。同概念と類義語に関しては単語文書行列を用いる重み付け PLSI が最も割合が高くなる。SAM などの単語・単語行列では共起する数が優先するので、この意味関係は埋没してしまいがちになるが、単語文書行列を用いるとこの意味関係が抽出できるためと考えられる。ここでも逆温度定数が高いほうが抽出できる割合が高くなる。

5. 考察

各単語の意味クラスへの帰属確率 P(c|w)に付いて実験結果を眺めると、どの単語も値の高い意味クラス c が存在する。これはどの意味にも相当しない、いわば雑音のクラスであると考えられる。単語間の意味関係をより高い精度で扱うためにはこの雑音クラスを取り除いておく必要があると考えられる。また、出現していない単語の共起回数及び出現回数をどのようなに補正するかによって、単語間の意味関係も変化すると考えられる。

6. まとめ

本稿では統計モデルと単語の意味関係を明らかにするために、いくつかの行列設計、単語の重み付けの導入とベイズ推定を用いた統計モデルを提案し、比較実験を行った。実験の結果、行列設計により、抽出される単語の意味が異なり、提案した重み付けが抽出に有効であり、ベイズ統計においてパラメタの逆温度定数により抽出する単語の概念関係を制御できることが明らかになった。今後はより単語間の意味が明らかになるための逆温度定数についての検討を行いたい。

文　献

[1] Tim Berners Lee.: Semantic Web Road map. World Wide Web Consortium, September 1998.

[2] Thomas Hofmann: Probabilistic Latent Semantic Indexing, Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval

[3] Thomas Hofmann, Jan Puzicha, Michael I. Jordan: Learning from Dyadic Data. NIPS 1998: 466-472

[4] 北　研二：確率的言語モデル，東京大学出版会，2001

[5] 茶筅:http://chasen.aistnara.ac.jp/index.html.ja

[6] 持橋　大地，松本　裕治：意味の確率的表現，自然言語処理 No.147 - 012SAM，2002.