統計的決定理論に基づく既存名詞シソーラスへの未知語登録方法に関する一考察

前田　康成[†*a)]

A Note on Positioning Unknown Words in an Existing Thesaurus Based upon Statistical Decision Theory

Yasunari MAEDA[†*a)]

あらまし　近年，人工知能の自然言語処理の分野において数多くのシソーラスが構築され，情報検索や機械翻訳などに利用されている．これらの既存シソーラスに対しては未登録の単語（未知語）を新たに登録する必要がある．しかし，従来の未知語登録方法は理論的には何の保証もない．そこで，本研究では統計的決定理論に基づいて未知語登録問題を考え直し，未知語を既存シソーラスの間違ったノードに登録してしまう確率である誤り率をベイズ基準のもとで最小化する未知語登録方法を提案する．更に，実際に既存シソーラスを用いた未知語登録実験を行い，実問題において提案方法が従来方法よりもより多くの未知語を既存シソーラスの正しいノードに登録できることを示す．

キーワード　シソーラス，未知語，統計的決定理論，ベイズ基準，誤り率

1. まえがき

近年，人工知能の自然言語処理の分野において，単語を意味的に分類したシソーラス [4], [7] が数多く構築されている．これらのシソーラスは，情報検索や機械翻訳などの多くの分野において利用されている．また，現在実用化されている既存シソーラスはすべて人手によって作成されており，既存シソーラスに未登録の単語（以下では，未知語と呼ぶ）を新たに登録するような維持管理作業も人手によって行われている．しかし，このような人手によるシソーラスの構築及び維持管理作業には膨大なコストがかかり，自動化の要求が大きい．そこで，本研究では既存シソーラスへの未知語登録を自動で行う自動未知語登録方法を研究対象にする．なお，以下では単に未知語登録方法と呼ぶ．

未知語登録方法及びその基本となる単語間の距離を扱った従来研究には様々なものがある [2], [5], [6], [10]〜[13], [15], [16] が，大きく二つに分類できる．一つ目は，木構造を有すシソーラスの各ノードの特徴ベクトルと未知語の特徴ベクトルの類似度をベクトル間の余弦を用いて算出し，類似度の高いノードに未知語を登録するベクトル空間法 [16] に基づく方法である [12], [13], [15]．ベクトル空間法は情報検索の分野で古くから実用化されている．特徴ベクトル間の余弦の値で比較することはシソーラスの各ノードと未知語との相関の程度を比較していると解釈できる [13]．しかし，相関の大きなノードに未知語を登録することに対しては，何ら理論的な保証はない．

二つ目は，単語の生起や共起の仕方に確率モデルを導入した方法である [2], [5], [6], [10], [11]．これらの方法では，確率分布の推定値を真の確率分布と仮定して未知語登録を行っており，その登録精度には理論的な保証はない．

そこで，本研究では，未知語を間違ったノードに登録してしまう確率である誤り率を損失関数として導入し，統計的決定理論 [1], [14] に基づいて定式化を行い，ベイズ基準のもとで誤り率を最小化するという意味で最適な未知語登録方法を提案する．また，一例にすぎないが，既存シソーラスを用いた未知語登録実験を行うことによって，統計的決定理論に基づく提案方法が

† NTT サイバースペース研究所，横須賀市
NTT Cyber Space Laboratories, 1–1 Hikarinooka Yokosuka-shi, 239–0847 Japan
* 現在，NTT 東日本技術部
a) E-mail: maeda.y@east.ntt.co.jp

従来のベクトル空間法に基づく方法よりもより多くの未知語を既存シソーラスの正しいノードに登録できることを実問題で示す．更に，従来の確率モデルを導入した方法と提案方法との関係について考察する．

まず，**2.** で未知語登録問題について述べ，**3.** で従来方法について述べる．次に **4.** で統計的決定理論に基づく未知語登録問題について述べ，定式化を行う．**5.** で実際にベイズ基準のもとで誤り率を最小化する未知語登録方法を提案する．**6.** で既存シソーラスを用いた未知語登録実験を行い，提案方法が実問題においてベクトル空間法に基づく従来方法よりもより多くの未知語を既存シソーラスの正しいノードに登録できることを示す．更に，提案方法と確率モデルを導入した従来方法との関係について考察する．最後に，**7.** でまとめを行う．

2. 未知語登録問題

2.1 シソーラス

シソーラスとは，単語を意味的に分類した分類体系である．シソーラスの多くは木構造を有し，名詞の集合を分類した名詞シソーラスや，用言の集合を分類した用言シソーラスなどがある．また，木構造の葉のみに単語が属す分類シソーラスと，根及び中間ノードにも単語が属す上位下位シソーラスがある [8].

本研究では，木構造を有す名詞シソーラスを研究対象とし，分類シソーラスと上位下位シソーラスの区別は特に行わない．以下では，木構造を有す名詞シソーラスのことを単にシソーラスと呼ぶ．なお，シソーラスに登録されている名詞には一つのノードのみに属す名詞と，複数のノードに属す名詞がある．ノードにはそのノードに属す名詞の集合の意味を示す概念と呼ばれるラベルが付与されている．図 1 に上位下位シソーラスである NTT シソーラス [4] の木構造の一部を示す．ノードを示す四角の中に付与されているのが概念で，各ノードの脇に付与されている番号は各ノードまたは各概念に対応する通し番号である．NTT シソーラスにおいて図 1 中のノードに属している名詞のリストの一部を図 2 に示す.

2.2 未知語登録問題の概要

本研究で扱う未知語登録問題とは，未知語を既存シソーラスに登録する問題である．なお，ここで述べる未知語とは，既存シソーラスには登録されていないが名詞であることがわかっている単語である．

未知語登録問題を考える場合には，ほとんどの従来研究がそうであるように，“単語の意味は，どのような単語と共起するかという観点から特徴づけられる”という Harris の分布仮説 [3] に基づいて考えるのが一般的である．そこで，本研究でも分布仮説に基づいて未知語登録問題を考えることにする．なお，二つの単語の共起の定義としては，その二つの単語が同一文中にともに存在すればいいという文内共起や，文の意味

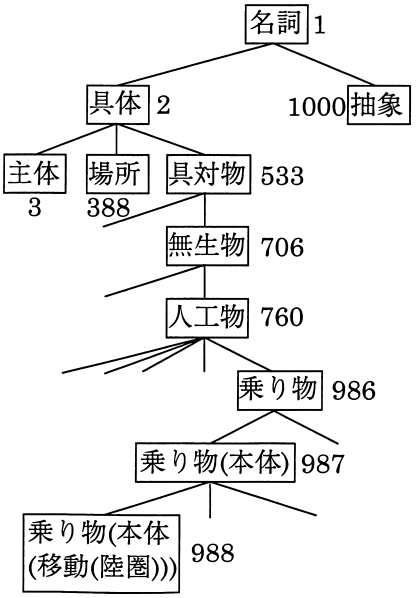

図 1　NTT シソーラスの木構造
Fig. 1　Tree structure of NTT thesaurus.

986　乗り物
空便　初便　先便　増便
定期便　便　夜行便
987　乗り物(本体)
乗りもの　乗り物　乗物
988　乗り物(本体(移動(陸圏)))
愛車　愛用車　青電車
赤電車　網台車　一番列車
稲車　インテグレーター
うば車 乳母車　駅馬車
エクスプレス …
オートバイ　…　私鉄　市電
自転車　自動車　市バス　…

図 2　NTT シソーラスのノードに属す名詞
Fig. 2　Nouns in nodes of NTT thesaurus.

内容まで見て係り受け関係が成立しているもののみ共起とみなすものなどいろいろあるが，本研究では特に定めない．また，本研究では，名詞と動詞の共起のみを考慮する．

分布仮説に基づくと，シソーラスの同一ノードに属す名詞は動詞との共起の仕方が似ていると考えられる．すなわち，名詞と動詞の共起の仕方はシソーラスのノードの数だけ存在し，未知語の共起の仕方もその中の一つであると考えられる．しかし，未知語の共起の仕方がそのうちのどれであるかは未知である．そこで，各ノードごとに動詞との共起の仕方を学習して，未知語の動詞との共起の仕方と比較し，未知語の共起の仕方と最も近い共起の仕方をするノードに未知語を登録することによって，未知語登録問題を解決できる．ここで最も重要なことは，いかにして共起の仕方を比較するかである．従来方法では，未知語の登録精度を理論的に保証するような比較ができていない．そこで，本研究では間違ったノードに登録してしまう確率である誤り率をベイズ基準のもとで最小化するという意味で理論的に保証のある比較方法を未知語登録方法として提案する．

3. 従来方法

3.1 ベクトル空間法に基づく従来方法

ベクトル空間法に基づく従来方法は，シソーラスの各ノードの特徴ベクトルと未知語の特徴ベクトルの類似度をベクトル間の余弦を用いて算出し，類似度の高いノードに未知語を登録するベクトル空間法を基本としている．最も単純なベクトル空間法では，特徴ベクトルは名詞と動詞の共起頻度によるベクトルである．ノードの特徴ベクトルの各要素は，そのノードに属す名詞の動詞との共起頻度を足し合わせたもので，未知語の特徴ベクトルの各要素は，未知語と動詞の共起頻度そのものである．実際には，単にベクトル空間法を用いるだけではなく，様々な言語的な性質を加味することによって登録精度を向上させる工夫がなされるが，本研究では従来研究のベクトル空間法に相当する部分が未知語登録問題の最も重要な部分であると考える．そこで，従来方法としてベクトル空間法を以下で詳しく説明する．

まず，いくつかの定義を行う．$noun_i$ はシソーラスに既に登録されている名詞を示し，$NOUN$, $NOUN = \{noun_1, noun_2, \cdots, noun_{|NOUN|}\}$ は要素数が有限の名詞集合である．なお，$|\cdot|$ は集合の要素数を示す．$node_i$ はシソーラスのノードを示し，$NODE$, $NODE = \{node_1, node_2, \cdots, node_{|NODE|}\}$ は要素数が有限のノード集合である．$unknown$ は未知語を示す．$verb_i$ は共起を考慮する動詞を示し，$VERB$, $VERB = \{verb_1, verb_2, \cdots, verb_{|VERB|}\}$ は要素数が有限の動詞集合である．(w, z), $w \in NODE$, $z \in VERB$ は一つの学習データを示す 2 項組であり，ノード w と動詞 z が共起したことを示す．$(w, z)^N$ は N 個の学習データからなる系列であり，$w^N z^N$, $w_1 z_1 w_2 z_2 \cdots w_N z_N$ と表記することもある．なお，学習データを生成するために用いるもともとの文章のデータの中では名詞 $noun_i$ と動詞 $verb_j$ が共起しているが，学習データを生成する時点で名詞と動詞の 2 項組 $(noun_i, verb_j)$ をノードと動詞の 2 項組 $(node_k, verb_j)$ に変換する．なお，ノード $node_k$ は名詞 $noun_i$ が属すノードであり，複数のノードに属す場合は複数の 2 項組に変換する．$\left(node^*, y^M\right)$, $node^* \in NODE$, $y \in VERB$ は未知語 $unknown$ が属すノード $node^*$ と未知語 $unknown$ と共起した M 個の動詞 y の系列 y^M の 2 項組を示す．しかし，$node^*$ は未知であり，実際に観測される未知語データは未知語 $unknown$ と未知語 $unknown$ と共起した動詞 y の系列 y^M の 2 項組 $\left(unknown, y^M\right)$ である．すなわち，未知語登録問題とは，学習データ $(w, z)^N$ と未知語データ $\left(unknown, y^M\right)$ を観測したもとで未知語 $unknown$ の属すノード $node^*$ を推定する問題である．

従来方法では，次式によって未知語を登録するノードが決定される．

$$\begin{aligned} & d_{\cos}\left((w, z)^N, \left(unknown, y^M\right)\right) \\ & = \arg\max_{node_i} \cos(vec(node_i), vec(unknown)) \\ & = \arg\max_{node_i} \frac{vec(node_i) \cdot vec(unknown)}{\| vec(node_i) \| \| vec(unknown) \|}, \end{aligned} \tag{1}$$

ただし，

$$\begin{aligned} & vec(node_i) \\ & = \Big(co\left((node_i, verb_1) \mid (w, z)^N\right), \\ & \quad co\left((node_i, verb_2) \mid (w, z)^N\right), \\ & \quad \cdots, co\left((node_i, verb_{|VERB|}) \mid (w, z)^N\right)\Big), \\ & vec(unknown) \end{aligned} \tag{2}$$

$$= \Big(co\left(verb_1 \mid y^M\right), co\left(verb_2 \mid y^M\right), \cdots, co\left(verb_{|VERB|} \mid y^M\right) \Big), \quad (3)$$

$d_{\cos}\left((w,z)^N, \left(unknown, y^M\right)\right)$ は学習データ $(w,z)^N$ と未知語データ $\left(unknown, y^M\right)$ を引数にとり，未知語 $unknown$ の登録すべきノードを決定する関数を示し，$vec(node_i)$ はノード $node_i$ の特徴ベクトル，$co\left((node_i, verb_j) \mid (w,z)^N\right)$ は学習データ $(w,z)^N$ 中の $(node_i, verb_j)$ の数でノード $node_i$ と動詞 $verb_j$ が共起した回数を示し，$vec(unknown)$ は未知語 $unknown$ の特徴ベクトル，$co\left(verb_i \mid y^M\right)$ は未知語データ $\left(unknown, y^M\right)$ の y^M 中の $verb_i$ の数で未知語 $unknown$ と動詞 $verb_i$ が共起した回数を示し，cos はベクトル間の余弦の値を求める関数，$vec_A \cdot vec_B$ はベクトル vec_A, vec_B 間の内積，$\| vec \|$ はベクトル vec のノルムを示す．

式 (1) で示されるように，従来方法では未知語の特徴ベクトル $vec(unknown)$ との余弦の値が最大になる特徴ベクトル $vec(node_i)$ に対応するノード $node_i$ に未知語 $unknown$ を登録する．

図 3 で説明すると，ベクトル空間において，各ベクトルの長さは無視して未知語 $unknown$ の特徴ベクトルとなす角度が最小の特徴ベクトルに対応するノード $node_i$ に未知語 $unknown$ を登録する．これは，未知語 $unknown$ を動詞との共起の仕方の相関が最大となるノード $node_i$ に登録していると解釈できる [13]．しかし，相関が最大となるノード $node_i$ に未知語 $unknown$ を登録することに対しては，理論的な保証はない．

なお，上記のような単純に共起頻度を用いるベクトル空間法以外に，各共起頻度に重み付けを行う TF·IDF 法を導入したベクトル空間法も提案されており，情報検索等の分野において実用化されているのは TF·IDF 法を導入したベクトル空間法である [16]．TF·IDF 法を導入したベクトル空間法では，式 (2) 及び式 (3) の特徴ベクトルの第 i 要素に $\log \dfrac{|NODE|}{a(verb_i)}$ を掛け合わせたものを特徴ベクトルとして採用し，式 (1) で未知語の登録を行う．ただし，$a(verb_i)$ は動詞 $verb_i$ との共起頻度が 1 以上のノードの数である．未知語登録問題に限らず一般的に単なる共起頻度によるベクトル空間法よりも，TF·IDF 法を導入したベクトル空間法の方が精度が高いことを示す事例が報告されているが，その精度に理論的な保証はない．

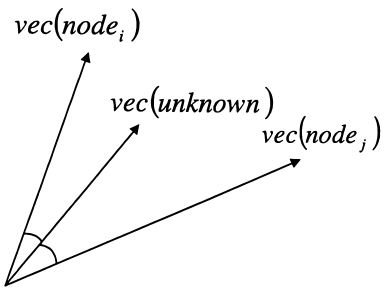

図 3　ベクトル空間における未知語登録
Fig. 3　Positioning unknown words in the vector space.

3.2　確率モデルを導入した従来方法

確率モデルを導入した従来方法としては，必ずしも未知語登録問題を扱っているわけではなく，単語間の距離として提案されているものもあるが，ここでは未知語登録問題に適用した場合を想定して，二つの従来方法について説明する．

まず，いくつかの定義を行う．ノード $node_i$ の生起する確率分布を $p(node_i \mid \theta)$，ノード $node_i$ が生起したもとで動詞 $verb_j$ が生起する確率分布を $p(verb_j \mid node_i, \theta)$ と表す．$p(node_i \mid \theta)$ と $p(verb_j \mid node_i, \theta)$ はともに連続パラメータ θ によって支配され，パラメータ集合を Θ とし，真のパラメータ θ^*, $\theta^* \in \Theta$ は未知とする．

Naïve-Bayes 法 [2] を未知語登録問題に適用すると，次式のような未知語登録方法 $d_{NB}\left(y^M\right)$ が考えられる．

$$d_{NB}\left(y^M\right) = \arg\max_{node_i} \hat{p}(node_i) \prod_{j=1}^{M} \hat{p}(y_j \mid node_i), \quad (4)$$

ただし，$\hat{p}(node_i)$ は $p(node_i \mid \theta^*)$ の最ゆう法等による推定値，$\hat{p}(y_j \mid node_i)$ は $p(verb_j \mid node_i, \theta^*)$ の最ゆう法等による推定値を示す．詳細は **6.** で述べるが，式 (4) による未知語登録方法は推定値を真のパラメータと仮定して，真のパラメータ既知の場合に誤り率を最小にする方法に推定値を代入している．この方法の登録精度には，理論的な保証はない．

次にカルバック・ライブラー（KL）情報量 [5], [10], [11] を未知語登録問題に適用すると，次式のような未知語登録方法 $d_{KL}\left(y^M\right)$ が考えられる．

$d_{KL}\left(y^M\right)$

$$= \arg\min_{node_i} D(\hat{p}(\cdot \mid unknown) \parallel \hat{p}(\cdot \mid node_i))$$
$$= \arg\min_{node_i} \sum_{verb_j} \hat{p}(verb_j \mid unknown) \cdot \log \frac{\hat{p}(verb_j \mid unknown)}{\hat{p}(verb_j \mid node_i)}, \quad (5)$$

ただし，$\hat{p}(verb_j \mid unknown)$ は未知語 $unknown$ が属すノード $node^*$ が生起したもとで動詞 $verb_j$ が生起する確率分布 $p(verb_j \mid node^*, \theta^*)$ の最ゆう法等による推定値，$\hat{p}(verb_j \mid node_i)$ は $\hat{p}(verb_j \mid node_i, \theta^*)$ の最ゆう法等による推定値，$D(\hat{p}(\cdot \mid unknown) \parallel \hat{p}(\cdot \mid node_i))$ は確率分布 $\hat{p}(\cdot \mid unknown)$，$\hat{p}(\cdot \mid node_i)$ 間の KL 情報量を示す．式 (5) による未知語登録方法では，確率分布の推定値による KL 情報量でノードと未知語の距離を測定して，最も未知語 $unknown$ に近いノード $node_i$ に未知語 $unknown$ を登録する．しかし，その登録精度には理論的な保証はない．

以上のように，従来方法には理論的な精度保証がない．そこで，本研究では統計的決定理論に基づいて，未知語 $unknown$ を誤ったノード $node_i$ に登録してしまう確率である誤り率をベイズ基準のもとで最小化するという意味で最適な未知語登録方法を以下で提案する．

4. 統計的決定理論に基づく未知語登録問題

本研究では，統計的決定理論 [1], [14] に基づいて，未知語登録問題を考え直す．

4.1 事前確率密度関数の定義

ここでは，本研究で新たに導入する事前確率密度関数の定義を行う．その他の記号に関しては，**3.** の従来方法における定義と同様であり，学習データ $(w, z)^N$ の生成方法も **3.** と同様である．

$p(\theta)$ はパラメータ θ の事前確率密度関数を示す．

4.2 統計的決定理論に基づく未知語登録問題の概要

本研究では，未知語を間違ったノードに登録してしまう確率である誤り率を統計的決定理論に基づいて最小化するという意味で最適な未知語登録方法を提案する．

まず，真のパラメータ θ^* によって支配される $p(node_i \mid \theta^*)$ と $p(verb_j \mid node_i, \theta^*)$ に基づいて学習データ $(w, z)^N$ と，未知語 $unknown$ が属すノード $node^*$ と未知語 $unknown$ と共起する M 個の動詞 y の系列 y^M の 2 項組 $\left(node^*, y^M\right)$ が生起する．しかし，実際には真のパラメータ θ^* と未知語 $unknown$ が属すノード $node^*$ は未知で，学習データ $(w, z)^N$ と未知語データ $\left(unknown, y^M\right)$ が観測される．未知語登録問題は学習データ $(w, z)^N$ と未知語データ $\left(unknown, y^M\right)$ を観測したもとで，未知語 $unknown$ の属すノード $node^*$ を推定する問題として解釈できる．

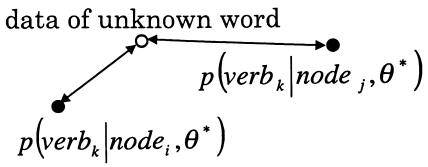

図 4 確率分布空間における未知語登録
Fig. 4 Positioning unknown words in the probability space.

図 4 で説明してみると，シソーラスのノードと動詞の共起の仕方を表す確率分布 $p(verb_k \mid node_i, \theta^*)$ 等が点在する確率分布空間に未知語データ（data of unknown word）を落とし込み，未知語データと最も近い分布に対応するノードを未知語 $unknown$ の属すノード $node^*$ の推定結果として出力する．なお，実際には θ^* も未知なので，$p(verb_k \mid node_i, \theta^*)$ 等も推定する必要がある．本研究では未知語 $unknown$ の属すノード $node^*$ を推定する際に，間違ったノード $node_i$ を推定結果として出力してしまう確率である誤り率を損失関数として導入し，統計的決定理論に基づいて，誤り率の最小化を図る．最適性の基準にはミニマックス基準，ベイズ基準など様々な基準が存在するが，本研究ではベイズ基準を採用する．

4.3 統計的決定理論に基づく未知語登録問題の定式化

以下で，統計的決定理論に基づく未知語登録問題の定式化を行う．

4.3.1 損失関数

式 (6) で示される損失関数は，パラメータ θ によって支配される $p(node_i \mid \theta)$ と $p(verb_j \mid node_i, \theta)$ に基づいて，未知語 $unknown$ が属すノード x ($x = node^*$.) と未知語 $unknown$ と共起する M 個の動詞 y の系列 y^M の 2 項組 $\left(x, y^M\right)$ が生起する場合に，決定関数 $d\left(y^M\right)$ を用いて間違ったノード $node_i$ をノード x の推定結果として出力する確率である誤り率を表す．定性的には，パラメータ θ によって支配される状況

で，未知語データ $\left(unknown, y^M\right)$ を観測した場合に間違ったノード $node_i$ に未知語 $unknown$ を登録する確率を意味する．

$$\begin{aligned} & L\left(d\left(y^M\right), \theta\right) \\ & = \sum_{x \in NODE} \sum_{y^M \in VERB^M} p(x, y^M \mid \theta) I\left(d\left(y^M\right)\right) \\ & = \sum_{x \in NODE} \sum_{y^M \in VERB^M} p(x \mid \theta) \\ & \quad \cdot \prod_{i=1}^{M} p(y_i \mid x, \theta) I\left(d\left(y^M\right)\right), \end{aligned} \tag{6}$$

ただし，$d\left(y^M\right)$ は未知語 $unknown$ と共起して観測された動詞の系列 y^M を引数にとり，未知語 $unknown$ の属すノード x の推定結果を出力する決定関数，$I\left(d\left(y^M\right)\right)$ は $d\left(y^M\right)$ が正しいノードを出力すれば 0，間違ったノードを出力すれば 1 を返すインディケータを示す，

$$I\left(d\left(y^M\right)\right) = \begin{cases} 1, & d\left(y^M\right) \neq x; \\ 0, & d\left(y^M\right) = x. \end{cases} \tag{7}$$

4.3.2 リスク関数

式 (8) で示されるリスク関数は，パラメータ θ によって支配される $p(node_i \mid \theta)$ と $p(verb_j \mid node_i, \theta)$ に基づいて学習データ $(w, z)^N$ が生起し，決定関数 $d\left(y^M\right)$ を用いる場合の損失関数の期待値，言い換えると誤り率の学習データに関する期待値を表す．

$$\begin{aligned} & R\left(d\left(y^M\right), \theta\right) \\ & = \sum_{(w,z)^N} p((w,z)^N \mid \theta) L\left(d\left(y^M\right), \theta\right) \\ & = \sum_{(w,z)^N} \prod_{i=1}^{N} p(w_i \mid \theta) p(z_i \mid w_i, \theta) L\left(d\left(y^M\right), \theta\right). \end{aligned} \tag{8}$$

4.3.3 ベイズリスク

式 (9) で示されるベイズリスクは，パラメータ θ の事前確率密度関数 $p(\theta)$ に関するリスク関数の期待値を表す．

$$BR(p(\theta)) = \int_{\Theta} p(\theta) R\left(d\left(y^M\right), \theta\right) d\theta. \tag{9}$$

ベイズリスクを最小にするノード x の推定結果を出力する式 (10) で示される決定 $BD(p(\theta))$ がベイズ決定であり，言い換えると，誤り率をベイズ基準のもとで最小化するという意味で最適な未知語登録方法である．

$$BD(p(\theta)) = \arg \min_{d\left(y^M\right)} BR(p(\theta)). \tag{10}$$

5. において，式 (10) を満足する最適な未知語登録方法を提案する．

5. 提案方法

誤り率をベイズ基準のもとで最小化するという意味で最適な未知語登録方法を提案するにあたって，まず，式 (9) のベイズリスクを書き下してみる．

$$\begin{aligned} & BR(p(\theta)) \\ & = \int_{\Theta} p(\theta) R\left(d\left(y^M\right), \theta\right) d\theta \\ & = \int_{\Theta} p(\theta) \sum_{(w,z)^N} \prod_{i=1}^{N} p(w_i \mid \theta) p(z_i \mid w_i, \theta) \\ & \quad \sum_{x} \sum_{y^M} p(x \mid \theta) \prod_{j=1}^{M} p(y_j \mid x, \theta) I\left(d\left(y^M\right)\right) d\theta \\ & = \sum_{(w,z)^N} \prod_{i=1}^{N} \Bigg(\int_{\Theta} p(\theta \mid w^{i-1} z^{i-1}) p(w_i \mid \theta) d\theta \\ & \quad \int_{\Theta} p(\theta \mid w^{i} z^{i-1}) p(z_i \mid w_i, \theta) d\theta \Bigg) \\ & \quad \sum_{y^M} \sum_{x} \Bigg(\int_{\Theta} p(\theta \mid w^N z^N) p(x \mid \theta) d\theta \\ & \quad \prod_{j=1}^{M} \Bigg(\int_{\Theta} p(\theta \mid w^N z^N, x, y^{j-1}) p(y_j \mid x, \theta) d\theta \Bigg) \\ & \quad I\left(d\left(y^M\right)\right) \Bigg), \end{aligned} \tag{11}$$

ただし，$p(\theta \mid w^{i-1} z^{i-1})$, $p(\theta \mid w^{i} z^{i-1})$, $p(\theta \mid w^N z^N)$, $p(\theta \mid w^N z^N, x, y^{j-1})$ はパラメータ θ の事後確率密度関数を示し，$p(\theta \mid w^0 z^0) = p(\theta)$, $w^i z^{i-1} = w_1 z_1 \cdots w_{i-1} z_{i-1} w_i$ である．式 (11) において，学習データ $(w, z)^N$ と未知語データ $\left(unknown, y^M\right)$ を受け取ったもとで，実際にどのノード $node_i$ を未

知語 $unknown$ を登録するノードとして出力するかによってベイズリスクの値を変化させるのは

$$\sum_x \left(\int_\Theta p(\theta \mid w^N z^N) p(x \mid \theta) d\theta \prod_{j=1}^{M} \left(\int_\Theta p(\theta \mid w^N z^N, x, y^{j-1}) p(y_j \mid x, \theta) d\theta \right) I\left(d\left(y^M\right)\right) \right)$$

の部分である．よって，最適な未知語登録方法 $d_{Bayes}\left(y^M\right)$ は次式で示される．

$$d_{Bayes}\left(y^M\right) = \arg \max_{x \in NODE} \int_\Theta p(\theta \mid w^N z^N) p(x \mid \theta) d\theta \prod_{i=1}^{M} \int_\Theta p(\theta \mid w^N z^N, x, y^{i-1}) p(y_i \mid x, \theta) d\theta. \tag{12}$$

以上より，学習データ $(w,z)^N$ と未知語データ $\left(unknown, y^M\right)$ を受け取ったもとで，上記式 (12) を用いて未知語 $unknown$ を登録するノード $node_i$ を決定することによって，誤り率がベイズ基準のもとで最小になる．

また，パラメータ θ の事前確率密度関数 $p(\theta)$ としてベータ分布を仮定することにより，式 (12) 中の積分の計算は式 (13) 及び式 (14) で示されるように容易になる．

$$\int_\Theta p(\theta \mid w^N z^N) p(x \mid \theta) d\theta = \frac{co(x \mid w^N) + \beta(x)}{\sum_x (co(x \mid w^N) + \beta(x))}, \tag{13}$$

ただし，$co(x \mid w^N)$ は w^N 中の x の数で学習データ $(w,z)^N$ 中でノード x が生起した回数を示し，$\beta(x)$ は $p(x \mid \theta)$ に対応するベータ分布のパラメータを示す．

$$\int_\Theta p(\theta \mid w^N z^N, x, y^{i-1}) p(y_i \mid x, \theta) d\theta = \frac{co\left(xy_i \mid w^N z^N\right) + co(y_i \mid y^{i-1}) + \beta(y_i \mid x)}{\sum_{y_i} \left(co\left(xy_i \mid w^N z^N\right) + co(y_i \mid y^{i-1}) + \beta(y_i \mid x) \right)}, \tag{14}$$

ただし，$co(xy_i \mid w^N z^N)$ は $w^N z^N$ 中の xy_i の数で学習データ $(w,z)^N$ 中でノード x と動詞 y_i が共起した回数を示し，$co(y_i \mid y^{i-1})$ は y^{i-1}，$y_1 y_2 \cdots y_{i-1}$ 中の y_i の数で未知語データの一部分 $(unknown, y^{i-1})$ 中で未知語 $unknown$ と動詞 y_i が共起した回数を示し，$\beta(y_i \mid x)$ は $p(y_i \mid x, \theta)$ に対応するベータ分布のパラメータを示す．

6. 従来方法と提案方法の比較

6.1 ベクトル空間法に基づく従来方法との比較

ここでは，実際に既存シソーラスを用いた未知語登録実験を行うことによって，一例にすぎないが，ベクトル空間法に基づく従来方法と提案方法の比較を行う．シソーラスには NTT シソーラス [4]，学習データ及び未知語データには EDR コーパス [9] の共起辞書を用いた．NTT シソーラスは NTT が作成したシソーラスで，今回は 12 万語の一般名詞からなる一般名詞シソーラスを用いた．また，EDR コーパスは 22 万文からなる文章のデータベースで，今回は係り受け関係にある単語対を抽出した共起辞書を用いた．実験方法は，まず NTT シソーラスに既登録の名詞 1000 語を未知語（$unknown$）と仮定して抽出する．抽出方法には 2 段抽出法を採用し，第 1 段で既登録語数が 10 以上のノードについて等確率抽出を行い，第 2 段でノード中の名詞について非復元の等確率抽出を行った．未知語として抽出された 1000 語には，アミノ酸，アルバイト，映画館，会長，海岸，古墳，公害，宗教，消しゴム，草木，電話，日の丸，方程式，理髪店，隕石などの名詞が含まれる．次に，NTT シソーラスに既登録の残りの名詞（$NOUN$）と EDR コーパス頻出動詞上位 500 語（$VERB$）との共起回数を共起辞書を参照して算出し，学習データを作成する．次に，NTT シソーラスから取り出しておいた 1000 語の未知語について，学習データと同様に EDR コーパス頻出動詞上位 500 語との共起回数を共起辞書を参照して算出し，1000 個の未知語データを作成する．次に，学習データと各未知語データをもとに従来方法と提案方法を用いて，各未知語に対する登録ノードを出力する．また，式 (1) の従来方法及び式 (12) の提案方法では，それぞれの尺度が最大となる第 1 位の候補のみ出力しているが，この実験では第 1 位の候補から第 10 位の候補まで出力することにした．なお，提案方法に関しては，事前確率密度関数として導入したベータ分布のパラメータを一様分布に設定した．これは，全くの無

知を表すためである [14]．図 5 に実験結果を示す．

図中の Cos は共起頻度のみによるベクトル空間法，TF·IDF は TF·IDF 法を導入したベクトル空間法，Bayes は提案方法に対応する．ここでの正解とは，未知語が元の NTT シソーラスにおいて登録されていたノードを指し，複数のノードに登録されていた場合には，出力したノードがその中のどれか一つと一致すれば正解とみなしている．横軸は考慮した累積の候補数，縦軸は考慮している候補の中に一つでも正解があれば 1，一つも正解がなければ 0 とした場合の，1000 語の未知語についてのその総和の 1000 に対する割合である正解率を示す．図 5 では，提案方法（Bayes）の正解率が共起頻度によるベクトル空間法（Cos）よりも常に 17%以上高く，TF·IDF 法を導入したベクトル空間法（TF·IDF）に対しても 10〜16%高くなっており，明らかに提案方法が従来方法よりも優れた結果を示している．なお，今回の実験は EDR コーパスに依存しているので，未知語ごとに M の値が異なり，M が 10 未満の低頻度の未知語から M が 100 以上の高頻度の未知語まで様々な未知語が存在した．しかし，今回の実験では未知語の頻度による登録精度に対する有意な影響は特に見受けられなかった．

以上の結果より明らかなように，既存シソーラスを用いた未知語登録実験において，提案方法は従来方法よりもより多くの未知語を既存シソーラスの正しいノードに登録できている．

6.2 確率モデルを導入した従来方法との比較

ここでは，確率モデルを導入した従来方法と提案方法の関係について考察する．

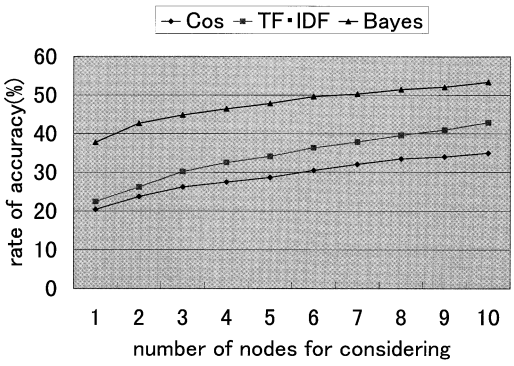

図 5　従来方法（Cos, TF·IDF）と提案方法（Bayes）の比較

Fig. 5　Comparison of proposed algorithm (Bayes) with previous algorithms (Cos, TF·IDF).

未知語登録問題においては，ノードの生起等の確率分布を支配する真のパラメータ θ^* は未知であるが，仮に既知であった場合を考えてみる．θ^* が既知の場合には，式 (8) のリスク関数を最小にする未知語登録方法が最適な未知語登録方法 $d_{\theta^*}\left(y^M\right)$ であり，次式で示される．

$$d_{\theta^*}\left(y^M\right) = \arg\max_{node_i} p(node_i \mid \theta^*) \prod_{j=1}^{M} p(y_j \mid node_i, \theta^*). \tag{15}$$

θ^* が既知であれば，式 (15) によって誤り率を最小にできる．

Naïve-Bayes 法による未知語登録方法の式 (4) と式 (15) を比較すると，Naïve-Bayes 法ではパラメータの推定値を真のパラメータと仮定して，真のパラメータが既知の場合に誤り率を最小にする未知語登録方法を採用していることがわかる．しかし，統計的決定理論ではパラメータの推定誤差を最小にする推定法と，推定したパラメータを使って何らかの行動を行う場合の推定法は異なってくる．よって，Naïve-Bayes 法を適用した未知語登録方法の登録精度には理論的な保証はない．一方，提案方法では，ベイズ基準のもとで誤り率を最小化するような推定値を導出しているので，最適性が保証されている．

次に，式 (5) による KL 情報量を用いた未知語登録方法について考える．式 (5) におけるパラメータの推定値として，最ゆう法を用いた場合を想定すると，式 (5) は次式と同値になる．

$$d_{KL}\left(y^M\right) = \arg\max_{node_i} \prod_{j=1}^{M} \hat{p}(y_j \mid node_i). \tag{16}$$

Naïve-Bayes 法の式 (4) と式 (16) を比較すると，式 (16) が，Naïve-Bayes 法において各ノードの生起確率を等確率と仮定した場合と同値であることがわかる．しかし，各ノードの生起確率を等確率と仮定することはあまり現実的ではない．よって，KL 情報量を用いた未知語登録方法は，推定法として最ゆう法を用いた場合には，Naïve-Bayes 法に更に各ノードの生起確率が等確率という非現実的な仮定を加えた未知語登録方法であり，その登録精度には何ら理論的な保証がないことがわかる．

7. むすび

本研究では，既存シソーラスへの未知語登録問題を研究対象にした．従来から，シソーラスの維持管理技術の一つとして，未知語登録問題は研究されてきた．しかし，その登録精度には何ら理論的な保証はない．

そこで，本研究では統計的決定理論に基づき，未知語を間違ったノードに登録してしまう確率である誤り率をベイズ基準のもとで最小化するという意味で最適な未知語登録方法を提案した．更に，実際に既存シソーラスを用いた未知語登録実験を通して，実問題において提案方法がベクトル空間法に基づく従来方法よりもより多くの未知語を既存シソーラスの正しいノードに登録できることを示した．更に，提案方法と確率モデルを導入した従来方法の関係を推定法の違い等の点から確認した．

今後の研究課題としては，今回は名詞と動詞の共起を考慮したが，その他の品詞の共起データによる未知語登録方法や，シソーラスの構造や言語的な性質を加味した未知語登録方法の新たな定式化も考えたい．また，情報検索や機械翻訳などでのシソーラスの利用を目的にした場合の情報検索や機械翻訳などに適した未知語登録方法の定式化を目的別に考えたい．このような目的別の定式化を行うことによって，それぞれの目的に適したシソーラスの自動構築技術にもつながると考える．

文　献

[1] J.O. Berger, Staristical Decision Theory and Bayesian Analysis, Springer-Verlag, New York, 1980.

[2] A. Fujii, K. Inui, T. Tokunaga, and H. Tanaka, “Selective sampling for example-based word sense disambiguation,” Computational Linguistics, vol.24, no.4, pp.573–597, Dec. 1998.

[3] Z.S. Harris, Mathematical structures of language, Wiley, New York, 1968.

[4] 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.

[5] L. Lee and F. Pereira, “Distributional similarity models: Clustering vs. nearest neighbors,” Proc. 37th Annual Meeting of the Association for Computational Linguistics, pp.33–40, Maryland, USA, June 1999.

[6] D. Lin, “An information-theoretic definition of similarity,” Proc. Int. Conf. on Machine Learning, pp.296–304, Madison, USA, July 1998.

[7] G.A. Miller, “WordNet: A lexical database for English,” Commun. ACM, vol.38, no.11, pp.39–41, Nov. 1995.

[8] 長尾　真，自然言語処理，岩波書店，1996.

[9] 日本電子化辞書研究所，EDR 電子化辞書利用マニュアル第 2.1 版，1994.

[10] F. Pereira, N. Tishby, and L. Lee, “Distributional clustering of English words,” Proc. 31st Annual Meeting of the Association for Computational Linguistics, pp.183–190, Ohio, USA, June 1993.

[11] P. Resnik, “Semantic classes and syntactic ambiguity,” Proc. ARPA Human Language Technology Workshop, pp.278–283, 1993.

[12] H. Schutze, Ambiguity Resolution in Language Learning, CSLI Publications, California, 1997.

[13] H. Schutze, “Automatic word sense discrimination,” Computational Linguistics, vol.24, no.1, pp.97–123, March 1998.

[14] 繁桝算男，ベイズ統計入門，東京大学出版会，1985.

[15] 浦本直彦，“コーパスに基づくシソーラス，” 情処学論，vol.37, no.12, pp.2182–2189, Dec. 1996.

[16] I.H. Witten, A. Moffat, and T.C. Bell, Managing Gigabytes, Van Nostrand Reinhold, New York, 1994.

（平成 11 年 10 月 5 日受付，12 年 1 月 4 日再受付）

前田　康成　（正員）

平 7 早大・理工・工業経営卒．平 9 同大大学院理工学研究科修士課程了．同年，日本電信電話（株）入社．NTT サイバースペース研究所を経て，現在，NTT 東日本技術部勤務．機械学習，統計的決定理論，自然言語処理の研究に従事．