The 30th Symposium on Information Theory
and its Applications (SITA2007)
Kashikojima, Mie, Japan, Nov. 27–30, 2007

スペルミスを伴う形態素解析に関する一考察
A Note on Morphological Analysis with Misspellings

前田康成* 吉田秀樹 * 藤原祥隆 *
Yasunari MAEDA Hideki YOSHIDA Yoshitaka FUJIWARA
松嶋敏泰†
Toshiyasu MATSUSHIMA

Abstract— In this research we study morphological analysis with misspellings problem based on statistical decision theory. We propose a Bayes optimal method which minimizes an error rate with reference to a Bayes criterion. But the computational complexity of the Bayes optimal method is an exponential order. So we also propose an approximate method in order to reduce the computational complexity.

Keywords— morphological analysis, misspelling, statistical decision theory

1 はじめに

形態素解析は大別すると文法規則などのルールに基づく方法と確率モデルに基づく方法に分けられるが，本研究では後者の確率モデルに基づく方法を対象とする．確率モデルに基づく形態素解析に関する従来研究 [2, 5, 6] では言語モデルとして隠れマルコフモデルや多重マルコフ連鎖などが採用されており，本研究でもその中から隠れマルコフモデルを採用する．また，従来研究では特にスペルミスを伴う形態素解析の問題設定は検討されていない．従来研究でもスペルミスの修正問題は形態素解析問題と別に検討されている．よって，実際にスペルミスの混入の可能性がある文書の形態素解析を行う場合には先にスペルミスを修正した後に形態素解析を行うことになる．しかし，統計的決定理論 [1] に基づいて考えれば，目的が形態素解析のみであればスペルミスの修正候補を 1 つに限定してしまうことは，形態素解析の精度を低下させてしまう可能性がある．そこで，本研究ではスペルミスを伴う形態素解析問題を統計的決定理論の視点から 1 つの決定問題として検討する．

従来研究 [2, 5] では英語のように単語ごとに分かち書きされる言語または単語ごとに分かち書きされた言語データを対象とした形態素解析方法が提案されている．また，従来研究 [6] では，日本語のような分かち書きされない言語または分かち書きされていない言語データを対象とするために従来研究 [2, 5] で提案されている形態素解析方法の拡張を行うとともに，未知語への対応なども行われている．本研究ではスペルミスを伴う形態素解析の基本的な考え方に焦点を絞るため，英語のように単語ごとに分かち書きされる言語または単語ごとに分かち書きされた言語データを仮定し，未知語も無いと仮定する．

形態素解析では隠れマルコフモデルを支配する真のパラメータは未知なので，各単語の品詞が既知の複数の文の学習データを利用して未知パラメータの学習を行う．従来研究 [2, 5, 6] では，形態素解析を未知パラメータの推定と，品詞の推定という二つの問題に分けて検討している．そのため，未知パラメータの推定方法の選択理由が不明確である．そこで，本研究では未知パラメータの推定と品詞の推定を合わせて 1 つの形態素解析問題として統計的決定理論の視点から見直す．

また，形態素解析方法の評価の仕方としては，形態素解析を間違えてしまう確率である誤り率を品詞系列単位（文単位）で評価する場合と，単語単位で評価する場合が考えられるが，本研究では前者の品詞系列単位の場合を検討する．その上で，ベイズ基準のもとで誤り率を最小にするベイズ解を導出する．しかし，ベイズ解の算出に必要な計算量は膨大で実用向きではない．そこで，計算量を軽減した近似アルゴリズムを導出する．

2 形態素解析と従来研究

2.1 形態素解析

最初にいくつかの定義を行う．t_i，$t_i \in T$ は品詞を示し，$T=\{t_1, t_2, \ldots, t_{|T|}\}$ は要素数が有限で既知の品詞集合である．品詞 t_i は隠れマルコフモデルの状態に相当する．w_i，$w_i \in W$ は単語を示し，$W=\{w_1, w_2, \ldots, w_{|W|}\}$ は要素数が有限で既知の単語集合である．各単語は文字列で表現される．a_i，$a_i \in A$ は文字を示し，$A=\{a_1, a_2, \ldots, a_{|A|}\}$ は要素数が有限で既知の文字集合である．$l(w_i)$ は単語 w_i の長さ（文字数）を示す．$G_1(w_i)$ は単語 w_i と同じ長

* 〒 090-8507 北海道北見市公園町 165 番地北見工業大学 工学部 情報システム工学科 Dept. of Computer Sciences Kitami Institute of Technology 165 Koen-cho, Kitami-shi, Hokkaido 090-8507 Japan.
† 〒 169-8555 東京都新宿区大久保 3-4-1 早稲田大学 基幹理工学部 応用数理学科 Dept. of Applied Mathematics School of Fundamental Science and Engineering Waseda University Okubo 3-4-1, Shinjuku-ku, Tokyo 169-8555 Japan.

さで 1 文字だけ異なる文字列の集合である．

$$G_1(w_i) = \{a^{l(w_i)} \in A^{l(w_i)} : d_H(w_i, a^{l(w_i)}) = 1\}, \quad (1)$$

ただし，$d_H(w_i, a^{l(w_i)})$ は単語 w_i と文字列 $a^{l(w_i)}$ のハミング距離を示す．本研究ではスペルミスを伴う形態素解析の基本的な考え方に焦点を絞るためにスペルミスは各単語からハミング距離が 1 の文字列に限定して検討する．$p(t_i \mid \theta)$ は文の先頭において品詞 t_i が生起する確率を示し，隠れマルコフモデルの初期状態が t_i となる確率に相当する．$p(t_j|t_i, \theta)$ は品詞 t_i の次に品詞 t_j が生起する確率を示し，隠れマルコフモデルの状態 t_i から状態 t_j への遷移確率に相当する．$p(w_j \mid t_i, \psi)$ は品詞 t_i の単語の中で単語 w_j が生起する確率を示し，隠れマルコフモデルの状態 t_i で単語 w_j が生起する確率に相当する．$p(a^{l(w_i)} \mid w_i, \phi)$ は単語 w_i が文字列 $a^{l(w_i)}$ になる確率を示す．

$$p(a^{l(w_i)} \mid w_i, \phi) = \begin{cases} \frac{p(G_1|\phi)}{l(w_i)(|A|-1)}, & d_H(w_i, a^{l(w_i)}) = 1; \\ 1 - p(G_1 \mid \phi), & d_H(w_i, a^{l(w_i)}) = 0, \end{cases} \quad (2)$$

ただし，$a^{l(w_i)}$，$a^{l(w_i)} \in A^{l(w_i)}$ は単語 w_i と等しいか，または単語 w_i に 1 文字のスペルミスが加わった文字列である．ある単語 w_i があった時に，単語 w_i にスペルミスが加わるかどうかは確率分布 $p(G_1 \mid \phi) = \phi$ に従う．確率 $p(G_1 \mid \phi)$ でスペルミスが加わり，確率 $1 - p(G_1 \mid \phi)$ でスペルミスが加わらない．スペルミスが加わる場合には，式 (2) に従って単語 w_i の代わりに単語 w_i とハミング距離が 1 のいずれかの文字系列が生起する．θ，$\theta \in \Theta$，ψ，$\psi \in \Psi$ と ϕ，$\phi \in \Phi$ は確率分布 $p(t_i|\theta)$，$p(t_j \mid t_i, \theta)$，$p(w_j \mid t_i, \psi)$，$p(a^{l(w_i)} \mid w_i, \phi)$，$p(G_1 \mid \phi)$ を支配するパラメータであり，真のパラメータ θ^*，$\theta^* \in \Theta$，ψ^*，$\psi^* \in \Psi$ と ϕ^*，$\phi^* \in \Phi$ は未知である．

$(x^N, y^N, z^N)^n = (x^{N_1}, y^{N_1}, z^{N_1}) \cdots (x^{N_n}, y^{N_n}, z^{N_n})$ は未知のパラメータ θ^*，ψ^*，ϕ^* について学習するための学習データ，n は学習データ数である．$(x^{N_i}, y^{N_i}, z^{N_i})$ は i 番目の学習データで，x^{N_i} が品詞系列，y^{N_i} が単語系列，z^{N_i} が単語系列 y^{N_i} に対する観測文字列である．

$$(x^{N_i}, y^{N_i}, z^{N_i}) = (x_{i,1} x_{i,2} \cdots x_{i,N_i}, y_{i,1} y_{i,2} \cdots y_{i,N_i}, z_{i,1} z_{i,2} \cdots z_{i,N_i}). \quad (3)$$

N_i は x^{N_i}，y^{N_i} 中の品詞数および単語数を示す．$x_{i,j}$ は x^{N_i} 中の j 番目に並んでいる品詞，$y_{i,j}$ は y^{N_i} 中の j 番目に並んでいる単語，$z_{i,j}$ は単語 $y_{i,j}$ に対する観測文字列を示す．スペルミスを伴う形態素解析では通常は観測文字列しか与えられないが，学習データ中の品詞系列及び単語系列は既知とする．学習データの生起確率は式 (4) のようになる．

$$\begin{aligned} & p((x^N, y^N, z^N)^n \mid \theta, \psi, \phi) \\ = & \prod_{i=1}^{n} \Big(p(x_{i,1}|\theta) p(y_{i,1}|x_{i,1}, \psi) \\ & \prod_{j=2}^{N_i} \big(p(x_{i,j}|x_{i,j-1}, \theta) p(y_{i,j}|x_{i,j}, \psi) \big) \\ & \prod_{k=1}^{N_i} p(z_{i,j}|y_{i,j}, \phi) \Big). \end{aligned} \quad (4)$$

$(x'^{N'}, y'^{N'}, z'^{N'})$ は品詞系列 $x'^{N'}$ と単語系列 $y'^{N'}$ が未知で，単語系列に対する観測文字列 $z'^{N'}$ が既知の新規に形態素解析を行いたいデータを示す．N' は品詞系列の品詞数および単語系列の単語数を示す．新規データの生起確率は式 (5) のようになる．

$$\begin{aligned} & p(x'^{N'}, y'^{N'}, z'^{N'} \mid \theta, \psi, \phi) \\ = & \; p(x_1'|\theta) p(y_1'|x_1', \psi) \\ & \prod_{i=2}^{N'} \big(p(x_i'|x_{i-1}', \theta) p(y_i'|x_i', \psi) \big) \\ & \prod_{j=1}^{N'} p(z_j'|y_j', \phi). \end{aligned} \quad (5)$$

上記より，スペルミスを伴う形態素解析問題とは学習データ $(x^N, y^N, z^N)^n$ と新規データの観測文字列 $z'^{N'}$ を受け取ったもとで，品詞系列 $x'^{N'}$ を推定する問題である．

2.2 従来研究

従来から数多くの形態素解析方法に関する研究が行われているが，従来研究では形態素解析方法への入力文にはスペルミスが含まれていないことが暗に仮定されている．実用に際しては，スペルミスの混入の可能性がある場合には，先にスペルミスの修正を行った後に形態素解析を行うことになる．そこで，本節では形態素解析方法に関する従来研究とスペルミスの修正方法に関する従来研究について説明する．

形態素解析の従来研究では，形態素解析問題を隠れマルコフモデルの未知パラメータの推定と，品詞系列の推定という 2 つの別々の問題として検討している．いろいろな従来研究があるが，基本的には最尤推定法でパラメータ推定した結果をビタビアルゴリズム [4] に代入して，品詞系列の推定を行っている．ビタビアルゴリズムは形態素解析における品詞系列に相当するような系列を推定する際に系列単位での評価を仮定して符号理論の分野で導出されたアルゴリズムである．形態素解析の従来研究では隠れマルコフモデルの未知パラメータの推定値として，最尤推定法による推定値や，平滑化を行った推

定値などが利用されているが，未知パラメータの推定と品詞系列の推定という 2 つの問題に分けてしまっているため，何故その推定値を利用するかについて明確な根拠がない．

そこで，本研究では未知パラメータの推定と品詞系列の推定を合わせて 1 つの形態素解析問題として統計的決定理論の視点から見直す．

他方，スペルミスの修正方法に関する従来研究では文脈を利用してスペルミスを修正する方法が検討されており，これらの従来研究 [3, 7] の中では形態素解析で利用されているのと同じ隠れマルコフモデルが言語モデルとして利用されている．特に従来研究 [7] ではスペルミスの修正を目的にしているが，修正候補と一緒に品詞系列の推定結果も出力するアルゴリズムが提案されいる．よって，このスペルミスの修正方法はスペルミスを伴う形態素解析問題に対する 1 つの対処法も与えている．しかし，統計的決定理論に基づいて考えれば，目的が形態素解析のみであればスペルミスの修正候補を 1 つに限定してしまうことは，形態素解析の精度を低下させてしまう可能性がある．

そこで，本研究ではスペルミスを伴う形態素解析問題を形態素解析のみを目的とした決定問題として統計的決定理論に基づいて以下で検討する．

3 スペルミスを伴う形態素解析

3.1 ベイズ最適解の導出

まず最初にスペルミスを伴う形態素解析問題の定式化を行う．この場合の損失関数は式 (6) のようになる．

$$L(\widehat{x'}^{N'}(z'^{N'},(x^N,y^N,z^N)^n),x'^{N'}) = \begin{cases} 1, & \widehat{x'}^{N'}(z'^{N'},(x^N,y^N,z^N)^n) \neq x'^{N'}; \\ 0, & \widehat{x'}^{N'}(z'^{N'},(x^N,y^N,z^N)^n) = x'^{N'}, \end{cases} \tag{6}$$

ただし，$\widehat{x'}^{N'}(z'^{N'},(x^N,y^N,z^N)^n)$ は学習データと新規データの観測文字列を受け取ったもとで品詞系列の推定結果を返す決定関数であり，$L(\widehat{x'}^{N'}(z'^{N'},(x^N,y^N,z^N)^n),x'^{N'})$ は品詞系列の推定結果に対する $0-1$ 損失である．

リスク関数は式 (7) のようになる．

$$\begin{aligned} &R(\widehat{x'}^{N'}(z'^{N'},(x^N,y^N,z^N)^n),\theta,\psi,\phi) \\ &= \sum_{(x^N,y^N,z^N)^n \in (T^N W^N (A^{l(y)})^N)^n} \\ &\quad \sum_{(x'^{N'},y'^{N'},z'^{N'}) \in (T^{N'} W^{N'} (A^{l(y')})^{N'})} \\ &\quad p((x^N,y^N,z^N)^n|\theta,\psi,\phi) p(x'^{N'},y'^{N'},z'^{N'}|\theta,\psi,\phi) \\ &\quad L(\widehat{x'}^{N'}(z'^{N'},(x^N,y^N,z^N)^n),x'^{N'}). \end{aligned} \tag{7}$$

ベイズリスクは式 (8) のようになる．

$$\begin{aligned} &BR(p(\theta),p(\psi),p(\phi)) \\ &= \int_\Theta \int_\Psi \int_\Phi p(\theta)p(\psi)p(\phi) \\ &\quad R(\widehat{x'}^{N'}(z'^{N'},(x^N,y^N,z^N)^n),\theta,\psi,\phi) d\phi d\psi d\theta, \end{aligned} \tag{8}$$

ただし，$p(\theta)$，$p(\psi)$，$p(\phi)$ は各パラメータの事前分布である．

式 (9) が式 (8) のベイズリスクを最小にし，ベイズ基準のもとで誤り率を最小にする形態素解析方法である．

$$\begin{aligned} Bd(z'^{N'},(x^N,y^N,z^N)^n) = \arg\max_{\widehat{x'}^{N'} \in T^{N'}} \sum_{\widehat{y'}^{N'} \in ((z' \cup G_1(z')) \cap W)^{N'}} \\ \int_\Theta p(\theta|(x^N)^n) p(\widehat{x_1'}|\theta) d\theta \\ \int_\Psi p(\psi|(x^N,y^N)^n) p(\widehat{y_1'}|\widehat{x_1'},\psi) d\psi \\ \prod_{i=2}^{N'} \Big(\int_\Theta p(\theta|(x^N)^n,\widehat{x'}^{i-1}) p(\widehat{x_i'}|\widehat{x_{i-1}'},\theta) d\theta \\ \int_\Psi p(\psi|(x^N,y^N)^n,\widehat{x'}^{i-1},\widehat{y'}^{i-1}) p(\widehat{y_i'}|\widehat{x_i'},\psi) d\psi \Big) \\ \prod_{j=1}^{N'} \widehat{p}(z_j'|\widehat{y_j'}), \end{aligned} \tag{9}$$

ただし，

$$\widehat{p}(z_j'|\widehat{y_j'}) = \begin{cases} \frac{\int_\Phi p(\phi|(y^N,z^N)^n,\widehat{y}^{j-1}z'^{j-1}) p(G_1|\phi) d\phi}{l(\widehat{y_j'})(|A|-1)}, & \widehat{y_j'} \neq z_j'; \\ 1 - \int_\Phi p(\phi|(y^N,z^N)^n,\widehat{y'}^{j-1}z'^{j-1}) p(G_1|\phi) d\phi, & \widehat{y_j'} = z_j'. \end{cases} \tag{10}$$

式 (9) の積分計算は共役事前分布であるディレクレ分布をパラメータ θ，ψ，ϕ の事前分布として採用することによって，例えば，以下のように容易に計算できる．

$$\begin{aligned} &\int_\Psi p(\psi|(x^N,y^N)^n) p(\widehat{y_1'}|\widehat{x_1'},\psi) d\psi \\ &= \frac{F(\widehat{x_1'}y_1'|(x^N,y^N)^n) + \xi(\widehat{y_1'}|\widehat{x_1'})}{\sum_{w \in W} \left(F(\widehat{x_1'}w|(x^N,y^N)^n) + \xi(w|\widehat{x_1'}) \right)}, \end{aligned} \tag{11}$$

ただし，$F(\widehat{x_1'}w|(x^N,y^N)^n)$ は学習データ中で品詞 $\widehat{x_1'}$ の状態で単語 w が生起した回数を示し，$\xi(w|\widehat{x_1'})$ は $p(w|\widehat{x_1'},\psi)$ に対するディレクレ分布のパラメータを示す．

3.2 近似アルゴリズムの提案

式 (9) で推定される品詞系列はベイズ基準のもとで誤り率を最小にする推定結果である．しかし，ディレクレ分布を事前分布として採用して積分計算を容易に計算できるようにしても，ベイズ最適な品詞系列の推定に必要な計算量は膨大である．よって，以下で計算

量を軽減した近似アルゴリズムを提案する．ベイズ最適な式 (9) では新規データに対する事後分布の更新部分の計算量が大きい．そこで，新規データに対する事後分布の更新は行わずに学習データに対する事後分布による予測分布を未知パラメータの推定値として用いる．つまり，式 (9) の $p(\theta|(x^N)^n, \widehat{x'}^{i-1})$ を $p(\theta|(x^N)^n)$ で近似し，$p(\psi|(x^N,y^N)^n, \widehat{x'}^{i-1}\widehat{y'}^{i-1})$ を $p(\psi|(x^N,y^N)^n)$ で近似し，式 (10) の $p(\phi|(y^N,z^N)^n, \widehat{y'}^{j-1}\widehat{z'}^{j-1})$ を $p(\phi|(y^N,z^N)^n)$ で近似する．$p(t_i \mid \theta^*)$，$p(t_j \mid t_i, \theta^*)$，$p(w_j \mid t_i, \psi^*)$，$p(a^{l(w_i)} \mid w_i, \phi^*)$，$p(G_1 \mid \phi^*)$ の近似による推定値をそれぞれ $\widehat{p}_{pos}(t_i)$，$\widehat{p}_{pos}(t_j|t_i)$，$\widehat{p}_{pos}(w_j|t_i)$，$\widehat{p}_{pos}(a^{l(w_i)}|w_i)$，$\widehat{p}_{pos}(G_1)$ と表記する．例えば，$\widehat{p}_{pos}(t_j|t_i)$，$\widehat{p}_{pos}(a^{l(w_i)}|w_i)$ は以下のように計算できる．

$$\widehat{p}_{pos}(t_j \mid t_i) = \int_\Theta p(\theta|(x^N)^n) p(t_j \mid t_i, \theta) d\theta = \frac{F(t_i t_j|(x^N)^n) + \xi(t_j|t_i)}{\sum_{t_k \in T}\big(F(t_i t_k|(x^N)^n) + \xi(t_k|t_i)\big)}. \tag{12}$$

$$\widehat{p}_{pos}(a^{l(w_i)} \mid w_i) = \begin{cases} \frac{\widehat{p}_{pos}(G_1)}{l(w_i)(|A|-1)}, & w_i \neq a^{l(w_i)}; \\ 1 - \widehat{p}_{pos}(G_1), & w_i = a^{l(w_i)}, \end{cases} \tag{13}$$

ただし，

$$\widehat{p}_{pos}(G_1) = \int_\Phi p(\phi|(y^N,z^N)^n) p(G_1 \mid \phi) d\phi = \frac{F(G_1|(y^N,z^N)^n) + \xi(G_1)}{\sum_{i=1}^n N_i + \xi(G_1) + \xi(\neg G_1)}, \tag{14}$$

$F(G_1|(y^N,z^N)^n)$ は学習データ中でスペルミスが生起した回数，$\neg G_1$ はスペルミスが生起しない事象（単語にスペルミスが加わらなかった事象）を示す．上記のように学習データに対する事後分布による予測分布を未知パラメータの推定値として採用することにより，式 (9) に対する近似解が式 (15) で得られる．

$$\widehat{x_{pos}'}^{N'}(z'^{N'},(x^N,y^N,z^N)^n) = \arg\max_{\widehat{x'}^{N'} \in T^{N'}} \sum_{\widehat{y'}^{N'} \in ((z' \cup G_1(z')) \cap W)^{N'}} \widehat{p}_{pos}(\widehat{x_1'}) \widehat{p}_{pos}(\widehat{y_1'}|\widehat{x_1'}) \widehat{p}_{pos}(z_1'|\widehat{y_1'}) \prod_{i=2}^{N'} \big(\widehat{p}_{pos}(\widehat{x_i'}|\widehat{x_{i-1}'}) \widehat{p}_{pos}(\widehat{y_i'}|\widehat{x_i'}) \widehat{p}_{pos}(z_i'|\widehat{y_i'})\big). \tag{15}$$

この近似解はベイズ最適解を導出する際に利用されるのと同じ予測分布を推定値として採用しており，ベイズ流の考え方に基づいた近似解になっている．式 (15) の近似解を実際に求める際に動的計画法を用いると符号理論におけるビタビアルゴリズムに似たアルゴリズムが導出される．この近似アルゴリズムではビタビアルゴリズムと同様にトレリス線図を用いる．

時点 1 から時点 N' までは各時点ごとに $|T|$ 個のノード（品詞），時点 $N'+1$ は end ノードのみを持つトレリス線図を考える．$|T| = 4$ の場合のトレリス線図の例を図 1 に示す．時点 i のノード $\widehat{x_i'}$ から時点 $i+1$ のノード $\widehat{x_{i+1}'}$ への遷移に対応するメトリックは次式による．

$$m_i(\widehat{x_{i+1}'}|\widehat{x_i'}) = \log \sum_{\widehat{y_i'} \in ((z_i' \cup G_1(z_i')) \cap W)} \widehat{p}_{pos}(\widehat{y_i'}|\widehat{x_i'}) \widehat{p}_{pos}(z_i'|\widehat{y_i'}) + \log \widehat{p}_{pos}(\widehat{x_{i+1}'}|\widehat{x_i'}). \tag{16}$$

時点 1 から時点 $N'+1$ の各ノードにおいてメトリックの足し合わせが最大になる遷移を選択することによって，学習データに対する事後分布による予測分布を用いた場合に生起確率が最大となる品詞系列を見つけることができる．逐次的な遷移の選択は次式による．

$$M_i(\widehat{x_i'}) = \max_{\widehat{x_{i-1}'} \in T} (M_{i-1}(\widehat{x_{i-1}'}) + m_{i-1}(\widehat{x_i'}|\widehat{x_{i-1}'})). \tag{17}$$

式 (17) によって定まる $\widehat{x_{i-1}'}$ を次式で保持する．

$$ps_i(\widehat{x_i'}) = \arg\max_{\widehat{x_{i-1}'} \in T} (M_{i-1}(\widehat{x_{i-1}'}) + m_{i-1}(\widehat{x_i'}|\widehat{x_{i-1}'})). \tag{18}$$

また，時点 1 から時点 i までの部分的に最適なパスを次式で保持する．

$$path_i(\widehat{x_i'}) = path_{i-1}(ps_i(\widehat{x_i'}))\widehat{x_i'}. \tag{19}$$

上記の各式を用いて実際に学習データに対する事後分布による予測分布を用いた場合に生起確率が最大となる品詞系列を見つけるアルゴリズムを以下に示す．

Step 1: 時点 1 の全ノードについて，
$M_1(\widehat{x_1'}) = \log \widehat{p}_{pos}(\widehat{x_1'})$，$path_1(\widehat{x_1'}) = \widehat{x_1'}$ とする．

Step g $(2 \leq g \leq N')$: 時点 g の全ノードについて $M_g(\widehat{x_g'})$，$ps_g(\widehat{x_g'})$，$path_g(\widehat{x_g'})$ を求める．

Step $N'+1$: end ノードについて $M_{N'+1}(end)$，$ps_{N'+1}(end)$，$path_{N'+1}(end)$ を求める．ただし，

$$m_{N'}(end|\widehat{x_{N'}'})$$

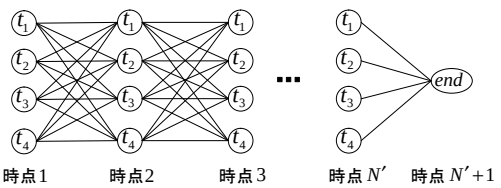

図 1: トレリス線図の例

$$= \log \sum_{\widehat{y_{N'}}' \in ((z_{N'}' \cup G_1(z_{N'}')) \cap W)} \widehat{p}_{pos}(\widehat{y_{N'}}'|\widehat{x_{N'}}')\widehat{p}_{pos}(z_{N'}'|\widehat{y_{N'}}'), \quad (20)$$

$$path_{N'+1}(end) = path_{N'}(ps_{N'+1}(end)). \quad (21)$$

最終的に $path_{N'+1}(end)$ として，学習データに対する事後分布による予測分布を未知パラメータの推定値として用いた場合に生起確率が最大となる品詞系列が求まる．

4 考察と今後の課題

本研究では，スペルミスを生起する確率分布の真のパラメータと言語モデルである隠れマルコフモデルの真のパラメータが未知のスペルミスを伴う形態素解析問題を研究対象とした．

形態素解析方法に関する従来研究は多いが，従来研究では入力文に対してスペルミスが含まれないことを暗に仮定している．また，スペルミスの修正方法に関する従来研究の中には修正候補と一緒に品詞系列の推定結果も出力する方法があるが，スペルミスの修正と品詞系列の推定を同時に行っているため，品詞系列の推定精度が低くなっている可能性がある．そこで，本研究では形態素解析のみを目的としてスペルミスを伴う形態素解析問題を統計的決定理論に基づいて検討した．

最初に有限の学習データに対して誤り率をベイズ基準のもとで最小にするベイズ最適な解を導出したが，その計算量が膨大なため，計算量を軽減した近似アルゴリズムを導出した．近似アルゴリズムでは形態素解析の精度に関する保証は無いが，ベイズ最適解の中でも利用されている学習データに対する事後分布による予測分布を利用しているので，統計的決定理論の考え方に基づいた近似アルゴリズムになっている．

今後の課題としては，本研究で提案した近似アルゴリズムの精度を言語の実データを用いた評価実験で検証することや，単語単位での評価の場合の検討を進めることが挙げられる．

参考文献

[1] Berger, J.: Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, New York (1985).

[2] Church, K.: A stochastic parts program and noun parser for unrestricted text, Proc. 2nd Conf. on Applied Natural Language Processing, pp.136-143 (1988).

[3] Golding, A. and Schabes, Y.: Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction, ACL-96, pp.71-78 (1996).

[4] Lin, S. and Costello, D.: Error Control Coding, Pearson Prentice Hall, New Jersey (1983).

[5] Manning, C. and Schütze, H.: Foundations of Statistical Natural Language Processing, The MIT Press, Massachusetts (1999).

[6] 永田昌明：統計的言語モデルと N-best 探索を用いた日本語形態素解析法，情報処理学会論文誌，Vol.40, No.9, pp.3420-3431 (1999).

[7] Nagata, M.: Context-Based Spelling Correction for Japanese OCR, COLING-96, pp.806-811 (1996).