A note on estimation of states in HMM with unknown parameters

Yasunari Maeda, Fumito Masui and Masakiyo Suzuki

Kitami Institute of Technology, 165 Koen-cho, Kitami-shi, Hokkaido 090-8507 Japan

Summary
Estimation of states in HMM(Hidden Markov Model) with unknown parameters is very important topic in many fields. In this paper we propose an optimal estimation method which minimizes an error rate with reference to the Bayes criterion. We also propose approximate method in order to reduce computational complexity.

Key words:
estimation of states, Hidden Markov Model, statistical decision theory, error rate

1. Introduction

Estimation of states in HMM(Hidden Markov Model) with unknown parameters is one of important topics in many fields. For example in natural language processing, a morphological analysis is equal to estimation of states in HMM with unknown parameters. And multi-topic document classification is also equal to estimation of states in HMM with unknown parameters. In this research we study estimation of states in HMM with unknown parameters under the condition that a learning data is given. And the learning data consists of pairs of a known state in HMM and a known symbol in HMM.

In previous research[3][4] of morphological analysis estimation of states in HMM with unknown parameters has been divided into two problems, estimating the unknown parameters of HMM and estimating the states. MLE(Maximum Likelihood Estimate) was used in the previous research basically, but there was no reason why MLE was used. There was no theoretical guarantee when the number of data for learning was finite.

In this research we treat estimating the unknown parameters and estimating the states as one problem based upon statistical decision theory.[1] And we propose Bayes optimal method which minimizes an error rate with reference to a Bayes criterion and approximate method in order to reduce computational complexity.

Manuscript received April 5, 2011
Manuscript revised April 20, 2011

2. Definitions and Previous Research

2.1 Definitions

First, we describe some definitions. Let t_i, $t_i \in T$ be a state of HMM. T, $T = \{t_1, t_2, \ldots, t_{|T|}\}$ is a set of states. w_i, $w_i \in W$ is a symbol which occurs at each state. W, $W = \{w_1, w_2, \ldots, w_{|W|}\}$ is a set of symbols. $p(t_i|\theta)$ is an initial state probability of HMM. $p(t_j|t_i,\theta)$ is a state transition probability of HMM. The probability distributions $p(t_i|\theta)$ and $p(t_j|t_i,\theta)$ are dominated by a parameter θ, $\theta \in \Theta$. And a true parameter θ^*, $\theta^* \in \Theta$ is unknown. $p(w_j|t_i,\psi)$ is the probability of an event that the symbol w_j occurs at the state t_i. The probability distribution $p(w_j|t_i,\psi)$ is dominated by a parameter ψ, $\psi \in \Psi$. And a true parameter ψ^*, $\psi^* \in \Psi$ is unknown.

$(x^N, y^N)^n$, $(x^N, y^N)^n = (x^{N_1}, y^{N_1})(x^{N_2}, y^{N_2}) \cdots (x^{N_n}, y^{N_n})$ is data for learning the unknown parameters θ^* and ψ^*. n is the number of data. (x^{N_i}, y^{N_i}), $(x^{N_i}, y^{N_i}) = (x_{i,1}x_{i,2}\cdots x_{i,N_i}, y_{i,1}y_{i,2}\cdots y_{i,N_i})$ is the i th data in $(x^N, y^N)^n$. x^{N_i} is the string of states in the i th data. And y^{N_i} is the string of symbols in the i th data. N_i is the length of x^{N_i} and y^{N_i}. $x_{i,j}$ is the j th state in x^{N_i}, and $y_{i,j}$ is the j th symbol in y^{N_i}. All x^N and y^N in $(x^N, y^N)^n$ are known.

The probability of an event that the learning data $(x^N, y^N)^n$ occur is described as follows:

$$p((x^N, y^N)^n|\theta,\psi) = \prod_{i=1}^{n} p(x_{i,1}|\theta)p(y_{i,1}|x_{i,1},\psi)\prod_{j=2}^{N_i} p(x_{i,j}|x_{i,j-1},\theta)p(y_{i,j}|x_{i,j},\psi). \quad (1)$$

$(x'^{N'}, y'^{N'})$ is a new pair of a string of states $x'^{N'}$ and a string of symbols $y'^{N'}$. $y'^{N'}$ is known, but $x'^{N'}$ is

unknown. The probability of an event that $(x'^{N'}, y'^{N'})$ occurs is described as follows:

$$p(x'^{N'}, y'^{N'}|\theta,\psi) = p(x'_1|\theta)p(y'_1|x'_1,\psi)\prod_{i=2}^{N'} p(x'_i|x'_{i-1},\theta)p(y'_i|x'_i,\psi). \quad (2)$$

In this research the task of estimation of states in HMM with unknown parameters is estimating the new string of states $x'^{N'}$ under the condition that the learning data $(x^N, y^N)^n$ and the new string of symbols $y'^{N'}$ are given.

2.1 Previous Research

In the previous research the task of estimation of states in HMM with unknown parameters has been divided into two problems, estimating the unknown parameters and estimating the new string of states $x'^{N'}$. In the basic method of the previous research $x'^{N'}$ was estimated by eqs.(3).

$$d_{ML}(y'^{N'}, (x^N, y^N)^n) = \arg\max_{\hat{x}'^{N'} \in T^{N'}} \hat{p}_{ML}(\hat{x}'_1)\hat{p}_{ML}(y'_1|\hat{x}'_1) \prod_{i=2}^{N'} \hat{p}_{ML}(\hat{x}'_i|\hat{x}'_{i-1})\hat{p}_{ML}(y'_i|\hat{x}'_i), \quad (3)$$

where $\hat{p}_{ML}$ was MLE. In eqs.(3) MLE was used, but the reason why MLE was used was not precise. There was no theoretical guarantee when the number of data for learning was finite.

In this paper we treat estimating the unknown parameters and estimating the new string of states $x'^{N'}$ as one problem based upon statistical decision theory. In the task of estimation of states in HMM with unknown parameters there are three kinds of purposes, estimation of the new string of states $x'^{N'}$, estimation of x'_i in $x'^{N'}$ and estimation of existence of state t_i in $x'^{N'}$. In this paper we study the first purpose.

3. Estimation of The New String of States based upon Statistical Decision Theory

3.1 Bayes Optimal Method

A loss function when a purpose is estimation of the new string of states is given by

$$L(\hat{x}'^{N'}(y'^{N'}, (x^N, y^N)^n), x'^{N'}) = \begin{cases} 1, & \hat{x}'^{N'}(y'^{N'}, (x^N, y^N)^n) \neq x'^{N'}; \\ 0, & \hat{x}'^{N'}(y'^{N'}, (x^N, y^N)^n) = x'^{N'}, \end{cases} \quad (4)$$

where $\hat{x}'^{N'}(y'^{N'}, (x^N, y^N)^n)$ is a decision function which returns an estimate of $x'^{N'}$.

A risk function is given by

$$R(\hat{x}'^{N'}(y'^{N'}, (x^N, y^N)^n), \theta, \psi) = \sum_{(x^N, y^N)^n \in (T^N, W^N)^n} \sum_{(x'^{N'}, y'^{N'}) \in (T^{N'}, W^{N'})} p((x^N, y^N)^n|\theta,\psi)p(x'^{N'} y'^{N'}, |\theta,\psi) L(\hat{x}'^{N'}(y'^{N'}, (x^N, y^N)^n), x'^{N'}). \quad (5)$$

This is equal to an error rate which is the probability of an event that the $\hat{x}'^{N'}(y'^{N'}, (x^N, y^N)^n)$ returns a wrong estimate. A Bayes risk is given by

$$BR(p(\theta), p(\psi)) = \int_{\theta\in\Theta}\int_{\psi\in\Psi} p(\theta)p(\psi)R(\hat{x}'^{N'}(y'^{N'}, (x^N, y^N)^n), \theta, \psi)d\psi d\theta, \quad (6)$$

where $p(\theta)$ is a prior density function for θ, $p(\psi)$ is a prior density function for ψ. The Bayes optimal decision which minimizes the Bayes risk is given by

$$\begin{aligned} Bd(y'^{N'}, (x^N, y^N)^n) = \arg\max_{\hat{x}'^{N'} \in T^{N'}} & \int_{\theta\in\Theta} p(\theta|(x^N)^n)p(\hat{x}'_1|\theta)d\theta \\ & \int_{\psi\in\Psi} p(\psi|(x^N, y^N)^n)p(y'_1|\hat{x}'_1,\psi)d\psi \\ & \prod_{i=2}^{N'}\left(\int_{\theta\in\Theta} p(\theta|(x^N)^n, \hat{x}'^{i-1})p(\hat{x}'_i|\hat{x}'_{i-1},\theta)d\theta\right. \\ & \left.\int_{\psi\in\Psi} p(\psi|(x^N, y^N)^n, \hat{x}'^{i-1}, y'^{i-1})p(y'_i|\hat{x}'_i,\psi)d\psi\right). \end{aligned} \quad (7)$$

When a Dirichlet distribution is used as the prior density for θ and ψ, the integration calculation in eqs.(7) is easily calculated. For example $\int_\psi p(\psi|(x^N, y^N)^n)p(y'_1|\hat{x}'_1,\psi)d\psi$ is calculated as follows:

$$\int_{\psi\in\Psi} p(\psi|(x^N, y^N)^n)p(y'_1|\hat{x}'_1,\psi)d\psi = \frac{F(\hat{x}'_1 y'_1|(x^N, y^N)^n) + \xi(y'_1|\hat{x}'_1)}{\sum_{w\in W}(F(\hat{x}'_1 w|(x^N, y^N)^n) + \xi(w|\hat{x}'_1))}, \quad (8)$$

where $F(\hat{x}'_1 w|(x^N, y^N)^n)$ is the number of times that the symbol w occurs at the state $\hat{x}'_1$ in the learning data $(x^N, y^N)^n$, $\xi(w|\hat{x}'_1)$ is the parameter of the Dirichlet distribution for $p(w|\hat{x}'_1,\psi)$.

The Bayes optimal decision by eqs.(7) can be calculated using a dynamic programming(DP) method. At first we introduce a DP-tree. Fig.1 is an example of DP-tree. A root node of the DP-tree is a null string. There are $|T|$ nodes at the depth of 1, and each node represents $\hat{x}'_1$, $\hat{x}'_1 \in T$. There are $|T|^i$ nodes at the depth of i, and each node represents $\hat{x}'^i$, $\hat{x}'^i \in T^i$.

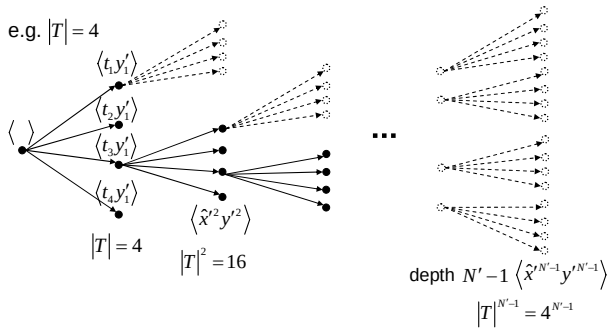

Fig.1 An example of DP-tree.

The Bayes optimal decision can be calculated by continuing calculation at each node from the depth of $N'-1$ to 0 as follows.

Step 1. At each node at the depth of $N'-1$ calculate as follows:

$$\begin{aligned} &qt\left(\hat{x}'^{N'-1} y'^{N'-1}\right) \\ &= \arg\max_{\hat{x}'_{N'} \in T} \int_{\theta\in\Theta} p\left(\theta \middle| \left(x^N\right)^n, \hat{x}'^{N'-1}\right) p\left(\hat{x}'_{N'} \middle| \hat{x}'_{N'-1}, \theta\right) d\theta \\ &\int_{\psi\in\Psi} p\left(\psi \middle| \left(x^N, y^N\right)^n, \hat{x}'^{N'-1}, y'^{N'-1}\right) p\left(y'_{N'} \middle| \hat{x}'_{N'}, \psi\right) d\psi , \end{aligned} \tag{9}$$

where $qt\left(\hat{x}'^{N'-1} y'^{N'-1}\right)$ is a function which returns an estimate of $x'_{N'}$.

$$\begin{aligned} &qp\left(\hat{x}'^{N'-1} y'^{N'-1}\right) \\ &= \max_{\hat{x}'_{N'} \in T} \int_{\theta\in\Theta} p\left(\theta \middle| \left(x^N\right)^n, \hat{x}'^{N'-1}\right) p\left(\hat{x}'_{N'} \middle| \hat{x}'_{N'-1}, \theta\right) d\theta \\ &\int_{\psi\in\Psi} p\left(\psi \middle| \left(x^N, y^N\right)^n, \hat{x}'^{N'-1}, y'^{N'-1}\right) p\left(y'_{N'} \middle| \hat{x}'_{N'}, \psi\right) d\psi , \end{aligned} \tag{10}$$

where $qp\left(\hat{x}'^{N'-1} y'^{N'-1}\right)$ holds the probability of the estimate by eqs.(9).

$$q\hat{x}\left(\hat{x}'^{N'-1} y'^{N'-1}\right) = qt\left(\hat{x}'^{N'-1} y'^{N'-1}\right), \tag{11}$$

where $q\hat{x}\left(\hat{x}'^{N'-1} y'^{N'-1}\right)$ holds the string of states.

Step g ($2 \le g \le N'-1$). At each node at the depth of $N'-g$ calculate as follows:

$$\begin{aligned} &qt\left(\hat{x}'^{N'-g} y'^{N'-g}\right) \\ &= \arg\max_{\hat{x}'_{N'-g+1} \in T} \int_{\theta\in\Theta} p\left(\theta \middle| \left(x^N\right)^n, \hat{x}'^{N'-g}\right) p\left(\hat{x}'_{N'-g+1} \middle| \hat{x}'_{N'-g}, \theta\right) d\theta \\ &\int_{\psi\in\Psi} p\left(\psi \middle| \left(x^N, y^N\right)^n, \hat{x}'^{N'-g}, y'^{N'-g}\right) p\left(y'_{N'-g+1} \middle| \hat{x}'_{N'-g+1}, \psi\right) d\psi \\ &qp\left(\hat{x}'^{N'-g+1} y'^{N'-g+1}\right). \end{aligned} \tag{12}$$

$$\begin{aligned} &qp\left(\hat{x}'^{N'-g} y'^{N'-g}\right) \\ &= \max_{\hat{x}'_{N'-g+1} \in T} \int_{\theta\in\Theta} p\left(\theta \middle| \left(x^N\right)^n, \hat{x}'^{N'-g}\right) p\left(\hat{x}'_{N'-g+1} \middle| \hat{x}'_{N'-g}, \theta\right) d\theta \\ &\int_{\psi\in\Psi} p\left(\psi \middle| \left(x^N, y^N\right)^n, \hat{x}'^{N'-g}, y'^{N'-g}\right) p\left(y'_{N'-g+1} \middle| \hat{x}'_{N'-g+1}, \psi\right) d\psi \\ &qp\left(\hat{x}'^{N'-g+1} y'^{N'-g+1}\right). \end{aligned} \tag{13}$$

$$\begin{aligned} &q\hat{x}\left(\hat{x}'^{N'-g} y'^{N'-g}\right) \\ &= qt\left(\hat{x}'^{N'-g} y'^{N'-g}\right) q\hat{x}\left(\hat{x}'^{N'-g} qt\left(\hat{x}'^{N'-g} y'^{N'-g}\right) y'^{N'-g+1}\right). \end{aligned} \tag{14}$$

Step N'. At the root node calculate as follows:

$$\begin{aligned} &qt(\ ) = \arg\max_{\hat{x}'_1 \in T} \int_{\theta\in\Theta} p\left(\theta \middle| \left(x^N\right)^n\right) p\left(\hat{x}'_1 \middle| \theta\right) d\theta \\ &\int_{\psi\in\Psi} p\left(\psi \middle| \left(x^N, y^N\right)^n\right) p\left(y'_1 \middle| \hat{x}'_1, \psi\right) d\psi \, qp\left(\hat{x}'_1 y'_1\right). \end{aligned} \tag{15}$$

$$q\hat{x}(\ ) = qt(\ ) q\hat{x}\left(qt(\ ) y'_1\right). \tag{16}$$

Finally eqs.(16) yields the same result as eqs.(7). If the Dirichlet distribution is used as the prior density, the integration calculation is done easily. But the computational complexity of the Bayes optimal method is still very big. The number of four arithmetic operations in the Bayes optimal method is about $13\sum_{i=0}^{N'-1}|T|^{i+1}$, it is an exponential order on N'. So, we propose an approximate method in order to reduce the computational complexity.

3.2 Approximate Method

In the approximate method predictive distributions calculated by using posterior density $p\left(\theta \middle| \left(x^N\right)^n\right)$ and $p\left(\psi \middle| \left(x^N, y^N\right)^n\right)$ are used as estimates for the unknown parameters. The predictive distributions are also used in eqs.(7). So there is a precise reason to use the predictive distributions. At first we describe some definitions. Let $\hat{p}_{pos}(t_i)$, $\hat{p}_{pos}\left(t_j \middle| t_i\right)$ and $\hat{p}_{pos}\left(w_j \middle| t_i\right)$ be the predictive distributions as follows:

$$\hat{p}_{pos}(t_i) = \int_{\theta\in\Theta} p\left(\theta \middle| \left(x^N\right)^n\right) p\left(t_i \middle| \theta\right) d\theta = \frac{F\left(t_i \middle| x^n_{\cdot,1}\right) + \xi(t_i)}{\sum_{t_j \in T}\left(F\left(t_j \middle| x^n_{\cdot,1}\right) + \xi\left(t_j\right)\right)}, \tag{17}$$

where $x^n_{\cdot,1} = x_{1,1}x_{2,1}\cdots x_{n,1}$, $F\left(t_j \middle| x^n_{\cdot,1}\right)$ is the number of times that an initial state of each string of states in the learning data $\left(x^N, y^N\right)^n$ is equal to the state t_j, $\xi\left(t_j\right)$ is the parameter of the Dirichlet distribution for $p\left(t_j \middle| \theta\right)$.

$$\begin{aligned} \hat{p}_{pos}\left(t_j \middle| t_i\right) &= \int_{\theta\in\Theta} p\left(\theta \middle| \left(x^N\right)^n\right) p\left(t_j \middle| t_i, \theta\right) d\theta \\ &= \frac{F\left(t_i t_j \middle| \left(x^N\right)^n\right) + \xi\left(t_j \middle| t_i\right)}{\sum_{t_k \in T}\left(F\left(t_i t_k \middle| \left(x^N\right)^n\right) + \xi\left(t_k \middle| t_i\right)\right)}, \end{aligned} \tag{18}$$

where $F\left(t_i t_k \middle| \left(x^N\right)^n\right)$ is the number of times that a transition from the state t_i to the state t_k occurs in the learning data $\left(x^N, y^N\right)^n$, $\xi\left(t_k \middle| t_i\right)$ is the parameter of the Dirichlet distribution for $p\left(t_k \middle| t_i, \theta\right)$.

$$\begin{aligned} \hat{p}_{pos}\left(w_j \middle| t_i\right) &= \int_{\psi\in\Psi} p\left(\psi \middle| \left(x^N, y^N\right)^n\right) p\left(w_j \middle| t_i, \psi\right) d\psi \\ &= \frac{F\left(t_i w_j \middle| \left(x^N, y^N\right)^n\right) + \xi\left(w_j \middle| t_i\right)}{\sum_{w_k \in W}\left(F\left(t_i w_k \middle| \left(x^N, y^N\right)^n\right) + \xi\left(w_k \middle| t_i\right)\right)}, \end{aligned} \tag{19}$$

where $F\left(t_i w_k \middle| \left(x^N, y^N\right)^n\right)$ is the number of times that the symbol w_k occurs at the state t_i in the learning data $\left(x^N, y^N\right)^n$, $\xi\left(w_k|t_i\right)$ is the parameter of the Dirichlet distribution for $p\left(w_k|t_i,\psi\right)$.

The approximate method is equal to a Viterbi algorithm in coding theory.[2] The Viterbi algorithm is also used in the previous research[3][4]. But MLEs are used as estimates for the unknown parameters in the previous research. Fig.2 is an example of trellis diagram.

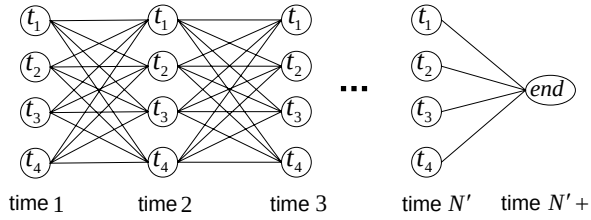

Fig.2 An example of trellis diagram.

At each time(from time 1 to time N') there are $|T|$ states on the trellis diagram. The trellis diagram ends at time $N'+1$. At time $N'+1$ there is only an end state. There is a metric $m_i\left(t'|t\right)$ for each branch on the trellis diagram as follows:

$$\begin{aligned} m_i\left(t'|t\right) &= m_i\left(\hat{x}'_{i+1} = t' \middle| \hat{x}'_i = t\right) \\ &= \log \hat{p}_{pos}\left(y'_i|t\right)\hat{p}_{pos}\left(t'|t\right) = \log \hat{p}_{pos}\left(y'_i|t\right) + \log \hat{p}_{pos}\left(t'|t\right). \end{aligned} \tag{20}$$

Let $M_i(t)$, $ps_i(t)$ and $path_i(t)$ be as follows:

$$M_i(t) = \max_{\hat{x}'_{i-1}\in T}\left(M_{i-1}\left(\hat{x}'_{i-1}\right) + m_{i-1}\left(t|\hat{x}'_{i-1}\right)\right). \tag{21}$$

$$ps_i(t) = \arg\max_{\hat{x}'_{i-1}\in T}\left(M_{i-1}\left(\hat{x}'_{i-1}\right) + m_{i-1}\left(t|\hat{x}'_{i-1}\right)\right). \tag{22}$$

$$path_i(t) = path_{i-1}\left(ps_i(t)\right)t. \tag{23}$$

Using eqs.(20), eqs.(21), eqs.(22) and eqs.(23) we can find an approximate estimate for the new string of states $x'^{N'}$. The approximate method is described as follows:

Step 1. Let $M_1(t)$ and $path_1(t)$ be as follows:

$$M_1(t) = \log \hat{p}_{pos}(t) \quad , \forall t \in T. \tag{24}$$

$$path_1(t) = t \quad , \forall t \in T. \tag{25}$$

Step g ($2 \le g \le N'$). Using eqs.(20), eqs.(21), eqs.(22), eqs.(23), eqs.(24) and eqs.(25) calculate $M_g(t)$, $ps_g(t)$ and $path_g(t)$ for all $t \in T$.

Step $N'+1$. Calculate $M_{N'+1}(end)$, $ps_{N'+1}(end)$ and $path_{N'+1}(end)$, where

$$m_{N'}\left(end|t\right) = \log \hat{p}_{pos}\left(y'_{N'}|t\right). \tag{26}$$

$$path_{N'+1}(end) = path_{N'}\left(ps_{N'+1}(end)\right). \tag{27}$$

Finally eqs.(27) yields the approximate estimate for the new string of states. The number of four arithmetic operations in the approximate method is about $3|T|^2 N'$. So, the computational complexity of the approximate method is proportional to N'.

4. Numerical Experiments

We investigate performance of our proposed approximate method by simulations. Fig.3 and Fig.4 are results of the simulations.

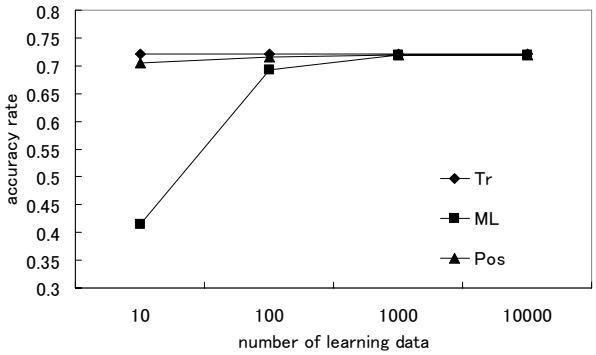

Fig.3 Comparison(type1).

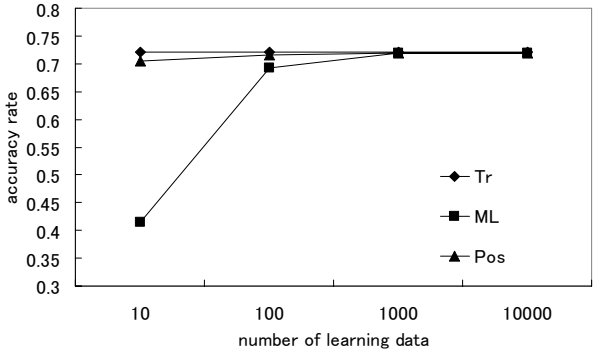

Fig.4 Comparison(type2).

We compared three methods, Pos is our proposed approximate method, ML is the method of previous research by eqs.(3), Tr is a method which knows true parameters of HMM(The true parameters are used in eqs.(3)). We examined two types of conditions. In type1 and type2 $|T| = 5$, $|W| = 5$, length of each string in learning data and new string is 10, comparison times is 10000, a Jeffrey's prior is used as the prior density for θ and ψ. True parameters in type1 are

$$\begin{bmatrix} p\left(t_1|\theta^*\right) \\ p\left(t_2|\theta^*\right) \\ p\left(t_3|\theta^*\right) \\ p\left(t_4|\theta^*\right) \\ p\left(t_5|\theta^*\right) \end{bmatrix} = \begin{bmatrix} 0.8 \\ 0.05 \\ 0.05 \\ 0.05 \\ 0.05 \end{bmatrix}, \tag{28}$$

$$\begin{vmatrix} p(t_1|t_1,\theta^*) & p(t_2|t_1,\theta^*) & p(t_3|t_1,\theta^*) & p(t_4|t_1,\theta^*) & p(t_5|t_1,\theta^*) \\ p(t_1|t_2,\theta^*) & p(t_2|t_2,\theta^*) & p(t_3|t_2,\theta^*) & p(t_4|t_2,\theta^*) & p(t_5|t_2,\theta^*) \\ p(t_1|t_3,\theta^*) & p(t_2|t_3,\theta^*) & p(t_3|t_3,\theta^*) & p(t_4|t_3,\theta^*) & p(t_5|t_3,\theta^*) \\ p(t_1|t_4,\theta^*) & p(t_2|t_4,\theta^*) & p(t_3|t_4,\theta^*) & p(t_4|t_4,\theta^*) & p(t_5|t_4,\theta^*) \\ p(t_1|t_5,\theta^*) & p(t_2|t_5,\theta^*) & p(t_3|t_5,\theta^*) & p(t_4|t_5,\theta^*) & p(t_5|t_5,\theta^*) \end{vmatrix}$$

$$= \begin{vmatrix} 0.9 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.9 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.9 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.9 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.9 \end{vmatrix},$$

(29)

$$\begin{vmatrix} p(w_1|t_1,\theta^*) & p(w_2|t_1,\theta^*) & p(w_3|t_1,\theta^*) & p(w_4|t_1,\theta^*) & p(w_5|t_1,\theta^*) \\ p(w_1|t_2,\theta^*) & p(w_2|t_2,\theta^*) & p(w_3|t_2,\theta^*) & p(w_4|t_2,\theta^*) & p(w_5|t_2,\theta^*) \\ p(w_1|t_3,\theta^*) & p(w_2|t_3,\theta^*) & p(w_3|t_3,\theta^*) & p(w_4|t_3,\theta^*) & p(w_5|t_3,\theta^*) \\ p(w_1|t_4,\theta^*) & p(w_2|t_4,\theta^*) & p(w_3|t_4,\theta^*) & p(w_4|t_4,\theta^*) & p(w_5|t_4,\theta^*) \\ p(w_1|t_5,\theta^*) & p(w_2|t_5,\theta^*) & p(w_3|t_5,\theta^*) & p(w_4|t_5,\theta^*) & p(w_5|t_5,\theta^*) \end{vmatrix}$$

$$= \begin{vmatrix} 0.9 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.9 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.9 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.9 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.9 \end{vmatrix}.$$

(30)

True parameters in type2 are

$$\begin{bmatrix} p(t_1|\theta^*) \\ p(t_2|\theta^*) \\ p(t_3|\theta^*) \\ p(t_4|\theta^*) \\ p(t_5|\theta^*) \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}, \quad (31)$$

$$\begin{vmatrix} p(t_1|t_1,\theta^*) & p(t_2|t_1,\theta^*) & p(t_3|t_1,\theta^*) & p(t_4|t_1,\theta^*) & p(t_5|t_1,\theta^*) \\ p(t_1|t_2,\theta^*) & p(t_2|t_2,\theta^*) & p(t_3|t_2,\theta^*) & p(t_4|t_2,\theta^*) & p(t_5|t_2,\theta^*) \\ p(t_1|t_3,\theta^*) & p(t_2|t_3,\theta^*) & p(t_3|t_3,\theta^*) & p(t_4|t_3,\theta^*) & p(t_5|t_3,\theta^*) \\ p(t_1|t_4,\theta^*) & p(t_2|t_4,\theta^*) & p(t_3|t_4,\theta^*) & p(t_4|t_4,\theta^*) & p(t_5|t_4,\theta^*) \\ p(t_1|t_5,\theta^*) & p(t_2|t_5,\theta^*) & p(t_3|t_5,\theta^*) & p(t_4|t_5,\theta^*) & p(t_5|t_5,\theta^*) \end{vmatrix}$$

$$= \begin{vmatrix} 0.9 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.9 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.9 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.9 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.9 \end{vmatrix},$$

(32)

$$\begin{vmatrix} p(w_1|t_1,\theta^*) & p(w_2|t_1,\theta^*) & p(w_3|t_1,\theta^*) & p(w_4|t_1,\theta^*) & p(w_5|t_1,\theta^*) \\ p(w_1|t_2,\theta^*) & p(w_2|t_2,\theta^*) & p(w_3|t_2,\theta^*) & p(w_4|t_2,\theta^*) & p(w_5|t_2,\theta^*) \\ p(w_1|t_3,\theta^*) & p(w_2|t_3,\theta^*) & p(w_3|t_3,\theta^*) & p(w_4|t_3,\theta^*) & p(w_5|t_3,\theta^*) \\ p(w_1|t_4,\theta^*) & p(w_2|t_4,\theta^*) & p(w_3|t_4,\theta^*) & p(w_4|t_4,\theta^*) & p(w_5|t_4,\theta^*) \\ p(w_1|t_5,\theta^*) & p(w_2|t_5,\theta^*) & p(w_3|t_5,\theta^*) & p(w_4|t_5,\theta^*) & p(w_5|t_5,\theta^*) \end{vmatrix}$$

$$= \begin{vmatrix} 0.75 & 0.1 & 0.025 & 0.025 & 0.1 \\ 0.1 & 0.75 & 0.1 & 0.025 & 0.025 \\ 0.025 & 0.1 & 0.75 & 0.1 & 0.025 \\ 0.025 & 0.025 & 0.1 & 0.75 & 0.1 \\ 0.1 & 0.025 & 0.025 & 0.1 & 0.75 \end{vmatrix}.$$

(33)

Accuracy rates on Fig.3 and Fig.4 are given by

$$\text{accuracy rate} = \frac{\text{the number of true estimates}}{\text{the number of all estimates}}. \quad (34)$$

When the number of learning data is big the accuracy rates of Pos and ML are equal to the rate of Tr. When the number of learning data is small the accuracy rates of Pos and ML are smaller than the rate of Tr. But the rate of Pos is bigger than the rate of ML.

5. Conclusion

In this research we proposed the Bayes optimal method for estimation of states in HMM with unknown parameters. In order to reduce the computational complexity we also proposed the approximate method using the predictive distributions calculated by using the posterior density $p\left(\theta\middle|\left(x^N\right)^n\right)$ and $p\left(\psi\middle|\left(x^N,y^N\right)^n\right)$. The predictive distributions are also used in the Bayes optimal method. So, our approximate method is based upon an idea of statistical decision theory.

We also studied basic properties of our approximate method from some simulations. As further works we want to study more about properties of our approximate method.

References

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

[2] S. Lin, and D. Costello, Error Control Coding, Pearson Prentice Hall, New Jersey, 1983.

[3] M. Nagata, "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm," Proc. 15th International Conference on Computational Linguistics, pp.201-207, 1994.

[4] M. Nagata, "A Japanese Morphological Analysis Method Using a Statistical Language Model and an N-best Search Algorithm," J. IPS Japan, Vol.40, No.9, pp.3420-3431, 1999.