Online Learning of Bayes Risk-Based Optimization of Dialogue Management for Document Retrieval Systems with Speech Interface

Teruhisa Misu[1], Komei Sugiura[1], Tatsuya Kawahara[1,2], Kiyonori Ohtake[1], Chiori Hori[1], Hideki Kashioka[1] and Satoshi Nakamura[1]

[1] National Institute of Information and Communications Technology (NICT), MASTAR Project, Keihannna Science City, Japan
teruhisa.misu@nict.go.jp
[2] Academic Center for Computing and Media Studies Kyoto University, Kyoto, Japan

Abstract. We propose an efficient online learning method of dialogue management based on Bayes risk criterion for document retrieval systems with a speech interface. The system has several choices in generating responses. So far, we have optimized the selection as minimization of Bayes risk based on reward for correct information presentation and penalty for redundant turns. In this paper, this framework is extended to be trainable by online learning. by maximum likelihood estimation of success probability of a response generation. Effectiveness of the proposed framework was demonstrated through an experiment with a large amount of utterances of real users. The online learning method was then compared with the method using reinforcement learning and discussed in terms of convergence speed.

1 Introduction

There are quite a few choices in spoken dialogue systems for handling user utterances and generating responses that involve parameter tuning. Since a subtle change in these choices may affect the behavior of the entire system, they are usually manually tuned by experts. In addition, every time the system is updated, such as when knowledge bases are added or changes are made to the ASR modules, the parameters must be re-tuned. Due to the high cost of such tuning, there have been many studies that have addressed the automatic optimization of dialogue management [Bohus et al., 2006,Lemon and Pietquin, 2007,Kim et al., 2008b], and most of these have dealt with database retrieval tasks[Young et al., 2007,Kim et al., 2008a]. The dialogue process in these studies has been designed using the formulation of Markov decision processes (MDPs) and trained by reinforcement learning (RL) [Roy et al., 2000,Levin et al., 2000,Singh et al., 2002]. The dialogue process in these frameworks needs to be mapped into a finite number of states. Since a list of database slots and a definite set of keywords are prepared a priori and manually in relational database (RDB) query tasks, the dialogue process is easily managed based on these. Such mapping is straightforward. For example, in a train information task, the combination of statuses of database slots, (such as “blank”, “filled” and “confirmed”) can be used as one dialogue state.

However, they cannot be directly applied to document retrieval tasks with a speech interface, where there is no relational structure in the document and every word is used in matching. The system has several choices for generating responses. Confirmation is needed to eliminate any misunderstandings caused by ASR errors, but users easily become irritated with too many redundant confirmations. Although there have been several studies dealing with dialogue management in call routing systems [Levin and Pieraccini, 2006,Horvitz and Paek, 2006], these methods cannot be applied to complex decision making processes in information guidance tasks. For example, our navigation system classifies user utterances into two types of information queries and factoid wh-questions, and generates appropriate responses to respective inputs. Unlike conventional question-answering (QA) tasks, in which all user inputs are assumed to be wh-questions, such as the TREC QA Track [3], it is often difficult to tell whether an utterance is an information query or a wh-question[Rosset et al., 2006]. In addition, there is not necessarily an exact answer to a wh-question in a document set. Therefore, it is not always optimal to respond to a wh-question with only its answer.

In order to manage the choices in response generation efficiently, we have proposed a dialogue management minimizing Bayes risk, based on reward for correct information presentation and penalty for redundant turns as well as the score of document retrieval and answer extraction [Misu and Kawahara, 2008]. However, this method requires a large amount of training data for parameter optimization. The strategy must be retrained when the reward/penalty parameters were updated. In this paper, we extend this framework to be trainable online. The proposed method is intended for efficient learning with a small number of samples. It is based on maximum likelihood estimation using Fisher's scoring algorithm, and optimality of the dialogue strategy is guaranteed even when the reward or penalty parameters were re-tuned after the learning. We also compare the proposed method with a method using reinforcement learning and discussed in terms of convergence speed.

2 Dialogue Management and Response Generation in Document Retrieval System

"Dialogue Navigator for Kyoto City" is a document retrieval system with a spoken dialogue interface. This system can retrieve information from a document set about sightseeing spots of Kyoto city. The knowledge bases (KBs) of this system are Wikipedia [4] documents concerning Kyoto and the official tourist information of Kyoto (810 documents, 220K words in total). This system is also capable of handling user's specific questions, such as "Who built this shrine?". An example dialogue of the system is shown in Figure 1.

2.1 Choices in Generating Responses

In this paper, among the many choices available in [Misu and Kawahara, 2008], we focus on choices related to the generation of responses and confirmations. Confirmation is indispensable to avoid inappropriate documents from being presented especially when the score for retrieval is low. It may also be "safer" to present the entire document

[3] http://trec.nist.gov/data/qamain.html

[4] http://wikipedia.org/

U1: Please tell me about the Golden Pavilion. (Information query)
S1: The Golden Pavilion is one of the buildings in the Rokuon-ji in Kyoto, and is the main attraction of the temple sites. The entire pavilion except the basement floor is covered with pure gold leaf.
U2: When was it built? (WH-Question)
S2: The Golden Pavilion was originally built in 1397 to serve as a retirement villa for Shogun Yoshimitsu.
U3: Then, please tell me its history. (Information query)
.....

Fig. 1. Example dialogue of Dialogue Navigator for Kyoto City

than to present a specific answer to the user's wh-question, when the score for answer extraction is low. For example, in the example dialogue in Fig. 1, if the system cannot find the exact answer of "1,397" for U2, the system can present a document about the history of the pavilion that may include information about the construction.

This kind of choices in conventional studies were made based on combinations of empirical knowledge, such as the ASR performance and the task type. However, hand-crafting heuristic rules is usually costly, and subtle changes in choices can seriously affect the performance of the whole system. Therefore, we propose a formulation where the above choices are optimized through online learning.

2.2 Optimization of Responses based on Bayes Risk

In the following subsections, we review the Bayes risk-based dialogue management that we have proposed in [Misu and Kawahara, 2008].

Bayes risk $L(d_j|\mathbf{W})$ is minimized in general pattern classification to determine the optimal class d_j for an input $\mathbf{W}$. In the Bayes classifier this is defined by

$$L(d_j|\mathbf{W}) = \sum_{i=1}^{n} l(d_j|d_i)p(d_i|\mathbf{W}), \tag{1}$$

where $(d_1, d_2, \ldots, d_n)$ denotes the given classes and $p(d_i|\mathbf{W})$ denotes the posterior probability for class d_i of $\mathbf{W}$. $l(d_j|d_i)$ is the loss function and represents the loss of predicting class d_j when the true class is d_i.

These classes $(d_1, d_2, \ldots, d_n)$ in our document retrieval task correspond to all documents. We assume the loss function among classes is the same, and we extend the framework to reward (negative loss; $l(d_j|d_i) < 0$) appropriate classifications.

$$l(d_j|d_i) = \begin{cases} -Rwd & \text{if } j = i \\ Penalty & \text{otherwise} \end{cases} \tag{2}$$

As a result, from Eq. (1), we obtain the Bayes risk $L(d_j|\mathbf{W})$ to determine document d_j for input W:

$$L(d_j|\mathbf{W}) = -Rwd * p(d_j|\mathbf{W}) + Penalty * (1 - p(d_j|\mathbf{W})). \tag{3}$$

In the spoken dialogue system, there are several choices in the manner of response or action to the user's request. Thus, we can define the Bayes risk for each response

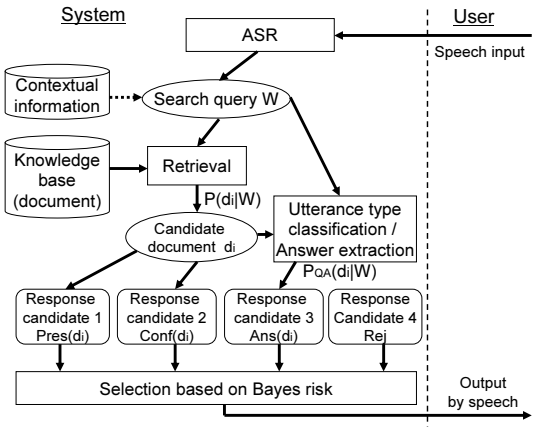

Fig. 2. Overview of Bayes risk-based dialogue management

candidate. The Rwd and $Penalty$ values are determined depending on the manner of response, and are defined by the degree of benefit to the user based on the correct information presentation and the loss caused by redundant time:

$$L(Res_i(d_j)|\mathbf{W}) = -Rwd_{Res_i} * p(d_j|\mathbf{W}) + Penalty_{Res_i} * (1 - p(d_j|\mathbf{W})). \quad (4)$$

The optimal choice is made by selecting the response that has the minimal amount of risk.

2.3 Generation of Response Candidates

Bayes risk-based dialogue management [Misu and Kawahara, 2008] is accomplished by comparing the possible responses hypothesized by varying the conditions for generating the search queries for KB retrieval and the manner of response and then selecting an appropriate response from the set of these responses. This paper focuses on response generation, and we do not deal with optimization in search query generation in [Misu and Kawahara, 2008]. That is, among the retrieved candidate documents by varying the manner in which the N-best hypotheses of ASR are used (the 1st, 2nd, or 3rd hypothesis, or all of them) and choosing whether to use contextual information, the candidate document is fixed to the hypothesis with the maximum likelihood (=matching score) of retrieval.

The possible response set $\mathbf{Res}$ includes answering $Ans(d_i)$, presentation $Pres(d_i)$, confirmation $Conf(d_i)$, and rejection $Rej(d_i)$. $Ans(d_i)$ denotes the user's specific wh-question being answered, which is generated by extracting one specific sentence that includes an answer named entity (NE) to the wh-question. $Pres(d_i)$ denotes a simple presentation of document d_i, which is actually made by summarizing it. $Conf(d_i)$ is an explicit confirmation[5] for presenting document d_i. Rej denotes a rejection: the system gives up making a response from document d_i and request the user for a rephrasal. This flow is illustrated in Fig. 2.

[5] We adopted an explicit confirmation such as "Do you want to know the *document's title*?"

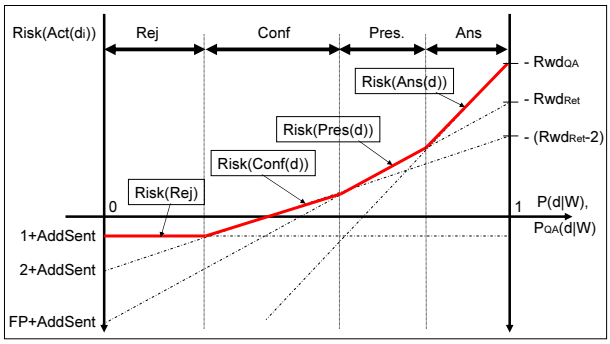

Fig. 3. Success rates v.s. risks for response candidates

2.4 Definition of Bayes Risk for Candidate Response

The dialogue management we propose is accomplished by comparing and then selecting from possible responses hypothesized by varying the condition. We define the Bayes risk based on the reward for success, the penalty for failure, and the probability of success, which is approximated by the confidence measure of the document matching (Sec. 2.5), for response candidates. That is, a reward is given depending on the manner of response (Rwd_{Ret} or Rwd_{QA}) when the system presents an appropriate response. On the other hand, when the system presents an incorrect response, a penalty is given based on extraneous time, which is approximated by the total number of sentences in all turns before the appropriate information is obtained. For example, the penalty for a confirmation is 2 {system's confirmation + user's approval}, and that of a rejection is 1 {system's rejection}. When the system presents incorrect information, the penalty for a failure $FailurePenalty$ (FP) is calculated, which consists of an improper presentation, the user's correction, and the system's request for rephrasal. Penalty for additional sentences to complete a task ($AddSent$) is also given as extraneous time before accessing the appropriate document when the user rephrases a information query/wh-question. The value of $AddSent$ is calculated as an expected number of additional sentences before accessing the correct response assuming the probability for success by rephrasal was p. The $AddSent$ for a retrieval is calculated as

$$\begin{aligned} AddSent &= FP + p * 1 + (1-p)(FP + p * 1 + (1-p)(\dots)) \\ &\cong \frac{(FP+p)}{p}. \end{aligned}$$

In the experiment described in this paper, we use the success rate in the field trial [Misu and Kawahara, 2007]. In particular, $p = 0.6$ is used. Thus, $AddSent$ depends on variable FP.

The Bayes risk for the response candidates is formulated as follows using the success rate of document retrieval $p(d_i|\mathbf{W})$, success rate of answer extraction $p_{QA}(d_i|\mathbf{W})$, and the reward pair (Rwd_{Ret} and Rwd_{QA}; $Rwd_{Ret} < Rwd_{QA}$) for successful presentations as well as the FP for inappropriate presentations.

– **Answering wh-question using document** d_i

$$Risk(Ans(d_i)) = -Rwd_{QA} * p_{QA}(d_i|\mathbf{W}) + (FP + AddSent) * (1 - p_{QA}(d_i|\mathbf{W}))$$

User utterance: When did the shogun order to build the temple?
(Previous query:) Tell me about the Silver Pavilion.

Response candidates:

Document with the largest score:
$\rightarrow p(\text{Silver Pavilion history}) = 0.4$
$\rightarrow p_{QA}(\text{Silver Pavilion history}) = 0.2$: In 1485
- $Risk(Ans(\text{Silver Pavilion history; In1485})) = 8.4$
- $Risk(Pres(\text{Silver Pavilion history})) = 6.4$
- $\mathbf{Risk(Conf(Silver\ Pavilion\ history))} = 4.8$

* Rejection
- $Risk(Rej) = 9.0$

↓

Response: Conf(Silver Pavilion history)
"Do you want to know the history of the Silver Pavilion?"

Fig. 4. Example of Bayes risk calculation

– **Presentation of document** d_i (without confirmation)

$$Risk(Pres(d_i)) = -Rwd_{Ret} * p(d_i|\mathbf{W}) + (FP + AddSent) * (1 - p(d_i|\mathbf{W}))$$

– **Confirmation for presenting document** d_i

$$Risk(Conf(d_i)) = (-Rwd_{Ret} + 2) * p(d_i|\mathbf{W}) + (2 + AddSent) * (1 - p(d_i|\mathbf{W}))$$

– **Rejection**
Since the success rate is 0 in this case, $Risk(Rej)$ is given as follows.

$$Risk(Rej) = 1 + AddSent$$

Figure 3 shows the relation between success rates and risks for response candidates. The risks of four response candidates are illustrated. Note that, the x-axis is $p(d_i|\mathbf{W})$ for $Pres(d_i)$, $Conf(d_i)$ and $p_{QA}(d_i|\mathbf{W})$ for $Ans(d_i)$. The optimal response candidate is determined for $p(d_i|\mathbf{W})$ and $p_{QA}(d_i|\mathbf{W})$ as shown in the bold line.

Figure 4 shows an example of calculating a Bayes risk (where $FR = 6$, $Rwd_{Ret} = 5$, $Rwd_{QA} = 30$). In this example, since the answer to the user's question does not exist in the knowledge base, the score of answer extraction is low. Therefore, the system chooses a confirmation before presenting the entire document.

2.5 Confidence Measure of Information Retrieval and Question-Answering

We adopted a standard vector space model to calculate the matching score between a user utterance (=ASR result) and the document in the KB. That is, the vector of the document $\mathbf{d} = (x_1, x_2, \ldots, x_n)^T$ was created by using the occurrence counts of nouns in the document. The vector for the user utterance $\mathbf{W} = (w_1, w_2, \ldots, w_n)^T$ is also created from the ASR result. Here, x_i and w_i are occurrence counts of noun i. The matching score $Match(\mathbf{W}, \mathbf{d})$ is calculated as the product of these two vectors.

$$Product(W, d) = \sum x_i \cdot w_i \tag{5}$$

The ASR confidence measure $CM(w_i)$ [Lee et al., 2004] is also used as a weight for the occurrence count. The matching score $Match(\mathbf{W}, \mathbf{d}_i)$ is then transformed into a confidence measure $p(d_i)$ using a logistic sigmoid function. This is used as an approximation of $p(d_i|\mathbf{W})$[6].

$$p(d_i) = \frac{1}{1 + \exp\{-\theta_1 * Match(\mathbf{W}, \mathbf{d}) - \theta_2\}} \tag{6}$$

Here, θ_1 and θ_2 are parameters of the sigmoid function ($\theta_1 > 0$). The score of question-answering $QAScore$ [Misu and Kawahara, 2008] is also transformed into a likelihood $p_{QA}(d_i|\mathbf{W_i})$ using another sigmoid function which is defined using another parameters of θ_3 and θ_4 ($\Theta = (\theta_1, \ldots, \theta_4)$).

3 Online Learning of Bayes Risk-based Dialogue Management

3.1 Parameter Optimization by Online Learning

The Bayes risk-based dialogue strategy is trained by updating the parameters of sigmoid functions $\Theta = (\theta 1, \cdots, \theta_4)$ so as to appropriately estimate the success rate of retrieval and question-answering. For tractable inputs, the system will learn to present documents or answers more efficiently. In contrast, for intractable inputs, such as erroneous or out-of-system inputs, the system will learn to make confirmations or gives up as quickly as possible (appropriate action for such queries is "rejection"). Thus, training with several dialogue sessions should lead to optimal decisions being made considering the current success rate of retrieval.

The proposed method is also expected to adapt to changes in the data, by periodically conducting parameter updates. This is one of the advantages of using the proposed method, as compared to the previous works [Levin and Pieraccini, 2006,Horvitz and Paek, 2006].

The training procedure can be described in four steps.

1. (At each step t) Generate response candidates $Pres(d_i)$, $Conf(d_i)$, $Ans(d_i)$, and Rej from document d_i that has the largest likelihood $p(d_i)$.
2. Generate response $Res_t(d_i)$ (or select response candidate with minimum risk) for d_i and calculate actual reward/penalty.
3. Update parameters Θ. θ_1 and θ_2 are updated when the input is an information query, and θ_3 and θ_4 is updated for wh-questions. This is elaborated in the following subsections.
4. Return to step 1. $t \leftarrow t + 1$

3.2 Optimization using Maximum Likelihood Estimation

We adopt the maximum likelihood estimation for learning the parameters Θ used for probability function [Kurita, 1994]. Let the set of the learning samples be $\{< Match_p, C_p > | p = 1, \ldots, t\}$, where a teaching signal C_p is given as a binary of 1

[6] This corresponds to a logistic regression of the success rate.

(success) or 0 (failure)[7]. If we assume the output of Eq. (6) (relabeled as z_p) as an estimate of the conditional probability given an input $Match_p$, the log-likelihood l for the samples is given by the following cross entropy error function:

$$l = \sum_{p=1}^{t} \{C_p \ln z_p + (1 - C_p) \ln(1 - z_p)\}. \quad (7)$$

The maximum likelihood estimate (MLE) of the weights is computed as one that maximizes this log-likelihood for previous t samples [Kurita, 1994]. The MLE weights are used as Θ^{t+1}.

As an algorithm to solve the maximum likelihood equations numerically, we adopt the Fisher's scoring algorithm that considers the variance in the score via the second derivative of the log of the likelihood function with respect to Θ (or Fisher information that is often used in natural gradient approaches of RL [Peters and Schaal, 2008]). This is a type of Newton's method, and the MLE is calculated quickly (in less than five seconds) by matrix calculation. We demonstrate that the optimal value is obtained with a small number of samples.

3.3 Optimization using Steepest Descent

The parameters Θ can be optimized by the steepest descent method that simply uses the first derivative. The parameters Θ is updated using the following equation in order to minimize the mean square error between the estimated risk and the actual reward/penalty.

$$\Theta^{t+1} = \Theta^t + \delta \frac{\partial}{\partial \Theta} (ARP - Risk(Res_t(d_i))^2$$

Here, δ is a learning rate, which is empirically set to 0.001.

3.4 Online Learning Method using Reinforcement Learning

The optimal decisions can also be obtained using reinforcement learning (RL)[8]. The goal of the online learning using RL is to estimate the value $Q(S, A)$ of each response (or action) $\mathbf{A} = (Pres(d_i), Conf(d_i), Ans(d_i), Rej)$ for state space. In a document retrieval task, since the matching score $Match(W, d)$, which corresponds to the state space S in this task, can take any positive number, we need to train the value $Q(S, A)$ for the continuous state space. We thus represent the values of responses for the current state by a function approximation system instead of a lookup table [Singh et al., 2002]. It should be noted that the POMDP solution technique using belief states with a delayed reward (e.g. [Williams and Young, 2007]) is similar to a RL for the continuous state space.

We approximate $Q(S, A)$ with triangle functions τ given by

$$\tau_m(S) = \begin{cases} 1 - |\frac{S}{\lambda} - m\lambda| & \text{if } |\frac{S}{\lambda} - m\lambda| < 1 \\ 0 & \text{otherwise} \end{cases}$$

[7] We regard a retrieval as successful if the system presented (or confirmed) the appropriate document/NE for the query.

[8] In this task, a reward or a penalty is given as a sooner reward. This problem corresponds to a multi-armed bandit problem with a continuous state space.

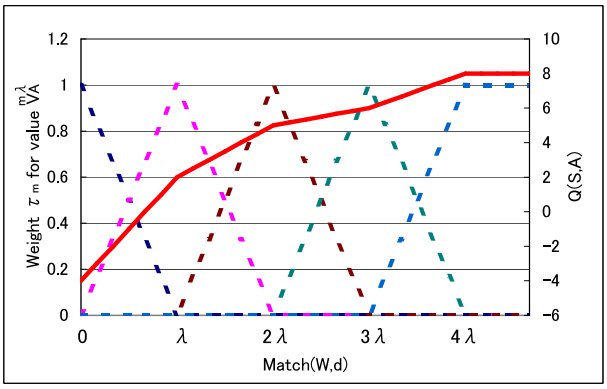

Fig. 5. Example of a value function

and values $\mathbf{V}_A = (V_A^0, V_A^\lambda, \ldots, V_A^{n\lambda})$ for grid points. Thus, the value function $Q(S, A)$ is represented as follows:

$$Q(S,A) = \begin{cases} \sum_{m=0}^{n} V_A^{m\lambda} \cdot \tau_m(S) & \text{if } S < n\lambda \\ V_A^{n\lambda} & \text{if } S \geq n\lambda \end{cases}$$

Here, λ denotes the grid size and n denotes the number of grid points. Figure 5 illustrates an example of a value function $Q(S, a)$ for an action a, where five grid points and triangle functions $n = 4$ are used to approximate the function ($\mathbf{V}_A = (-4, 2, 5, 6, 8)$ are used.). There is another plane whose x-axis is $QAScore$; $Q(QAScore, Ans(d_i))$ is calculated in the plane. The optimal choice is made by selecting the response that has the minimum value of $Q(S, A)$ or $Q(QAScore, A)$.

The values $\mathbf{V}$ are updated through online learning by the following procedures: For each step t, the system generates an action to execute a_e^t based on the ϵ-greedy strategy. That is, the best response that has the minimum value is selected for a probability $1 - \epsilon$, and responses is randomly selected for a probability ϵ, which was set to 0.2. Value parameters $\mathbf{V}_{a_e}$ of the selected response a_e^t were updated using the temporal difference (TD) algorithm:

$$\begin{aligned} V_{a_e}^{n\lambda\ (t+1)} &= V_{a_e}^{n\lambda\ (t)} + \delta\ \mathbf{TDError} \frac{\partial Q(S, a_e)}{\partial V_{a_e}^{n\lambda}} \\ &= V_{a_e}^{n\lambda\ (t)} + \delta\ (R_{a_e} - Q(S, a_e)) \cdot \tau_n(S). \end{aligned}$$

Here, R_{a_e} denotes the actual reward/penalty for the selected response a_e. The parameters of λ, n and δ were empirically set to $\lambda = 1.5$, $n = 6$ and $\delta = 0.001$ based on the result of a preliminary experiment.

4 Evaluation of Online Learning Methods

We evaluated the online learning methods. The set of 1,416 utterances (1,084 information queries and 332 wh-questions) is used in this evaluation. We trained the dialogue strategy by optimizing the parameters. We evaluated the improvement by using a 10-fold cross validation by splitting the utterance set into ten (set-1, $\cdots$, set-10), that is,

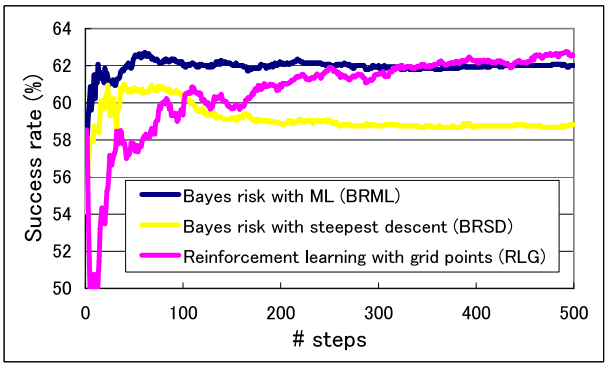

Fig. 6. Number of step vs. Success rate of information access

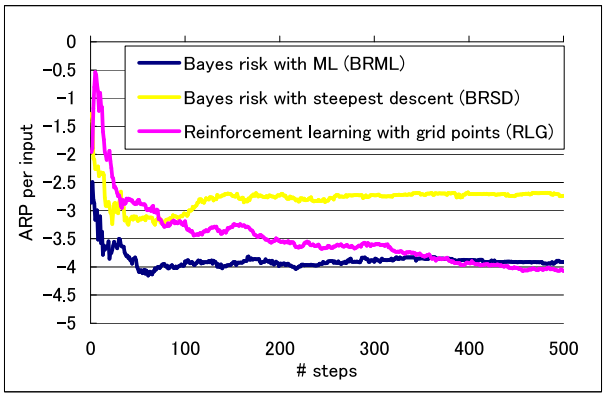

Fig. 7. Number of step vs. Actual reward/penalty (ARP)

one set was used as a test set to evaluate performance, and the other nine were used as training data. Since the method using RL has a random factor in the ϵ-greedy strategy, the result of the method is an average of the 10 trials.

The evaluation measures were the success rate and the average number of sentences for information access. We regard a retrieval as successful if the system presented (or confirmed to present) the appropriate response for the utterance. The actual reward/penalty ARP (or R) is obtained by assigning 1 into $p(d_i|\mathbf{W})$ of equations in section 2.4 for success and 0 for failure, for the response candidates[9]. We rewarded correct presentation by 10 ($Rwd_{Ret} = 10$) and correct question-answering by 30 ($Rwd_{QA} = 30$) considering difference in the number of samples in the test set. The FP was set to 6 based on typical recovery patterns observed in the field trial [Misu and Kawahara, 2007]. All parameters (Θ or $\mathbf{V}$) were initialized to zero.

Fig. 6 shows the relationship between the number of steps t for learning and the success rate of information access, and Fig. 7 shows the relationship between the number of steps t and the average number of expected ARP per query obtained by the step

[9] These values were calculated using the manually labeled correct responses.

t strategy at that time[10]. A small number of ARP implies a better dialogue strategy. By using Bayes risk-based strategy using ML estimation using Fisher's scoring algorithm (BRML), we could achieve a significantly higher performance than that of the steepest descent method (BRSD). The eventual performance of BRML was almost comparable to that of the method using RL with grid points (RLG).

We then evaluated the performance in terms of convergence speed. The BRML was converged very quickly with almost 50 samples. This convergence speed is one of the advantages of the BRML method such as when developing a new system or adapting it to changes in the tendency in the data. In contrast, the convergence speed of the RLG was considerably slower than that of the BRML requiring 500 steps. Of course other techniques, such as the natural gradient approach [Peters and Schaal, 2008] may improve the speed, but training by RLG requires a large number of iterations, especially when dealing with a continuous state space. One reason for this is that the RLG considers each response action as an independent one using no a priori knowledge about the dependency between responses. However, this assumption is not true (at least between "Presentation", "Confirmation" and "Rejection"). For example, if the system is rewarded by "Confirmation", it is supposed to obtain a better reward by "Presentation". In contrast, if it is penalized by confirmation, the penalty is supposed to be less with rejection. In the method using Bayes risk, the values of responses are optimized simultaneously in one step via the estimation of the success rate of retrieval. Thus, the method can estimate the risk of response with a fewer number of parameters. For these reasons, we consider that the training by BRML was converged with a small number of steps.

The target of the optimization in BRML is parameters of the logistic sigmoid function that estimate posterior probability of success, and it does not depend on the values of reward and penalty. This means that the optimality of the dialogue strategy by the proposed method is guaranteed. For example, if we replace the rewards by $Rwd_{Ret} = 0$ and $Rwd_{QA} = 0$, we will obtain a dialogue strategy that minimize the number of sentences in all turns before the appropriate information is obtained without parameter retuning. This property is an important advantage over the other approaches that require the whole training process using the re-tuned parameters.

5 Conclusion

We have proposed an online learning method of dialogue framework to generate an optimal response based on Bayes risk. Experimental evaluations by real user utterances demonstrated that the optimal dialogue strategy can be obtained with a small number of training samples. Although we only implemented and evaluated a simple explicit confirmation that asks the user whether the retrieved document is a correct one or not, the proposed method is expected to incorporate more various responses in document retrieval tasks, such as a clarification request and an implicit confirmation.

We used only two parameters of the matching score and bias for the logistic regression (Eq. (6)) to estimate the success rate, but this can be easily extended to incorporate various feature parameters, such as a difference of score (margin) with the second best candidate or a system's previous response. The Bayes risk-based strategies presented

[10] The response with the minimum risk is selected in the Bayes risk-based strategies and the response with the minimum value is selected in the strategy using RL.

in this paper assume a sooner reward, and cannot be directly applied to a dialogue task where a reward/penalty is given as a delayed reward. However, we can optimize the entire dialogue by introducing a cumulative future reward and the optimization process of WFSTs [Hori et al., 2008].

References

[Bohus et al., 2006] Bohus, D., Langner, B., Raux, A., Black, A., and Rudnicky, M. E. A. (2006). Online Supervised Learning of Non-understanding Recovery Policies. In *Proc. Spoken Laguage Technology Workshop (SLT)*, pages 170–173.

[Hori et al., 2008] Hori, C., Ohtake, K., Misu, T., Kashioka, H., and Nakamura, S. (2008). Dialog Management using Weighted Finite-state Transducers. In *Proc. Interspeech*, pages 211–214.

[Horvitz and Paek, 2006] Horvitz, E. and Paek, T. (2006). Complementary Computing: Policies for Transferring Callers from Dialog Systems to Human Receptionists. *User Modeling and User Adapted Interaction*, 17(1-2):159 – 182.

[Kim et al., 2008a] Kim, D., Sim, H., Kim, K., Kim, J., Kim, H., and Sung, J. (2008a). Effects of User Model on POMDP-based Dialogue Systems. In *Proc. Interspeech*, pages 1169–1172.

[Kim et al., 2008b] Kim, K., Lee, C., Jung, S., and Lee, G. (2008b). A Frame-based Probabilistic Framework for Spoken Dialog Management using Dialog Examples. In *Proc. of the 9th sigdial workshop on discourse and dialog*.

[Kurita, 1994] Kurita, T. (1994). Iterative weighted least squares algorithms for neural networks classifiers. *New Generation Computing*, 12.

[Lee et al., 2004] Lee, A., Shikano, K., and Kawahara, T. (2004). Real-Time Word Confidence Scoring using Local Posterior Probabilities on Tree Trellis Search. In *Proc. ICASSP*, pages 793–796.

[Lemon and Pietquin, 2007] Lemon, O. and Pietquin, O. (2007). Machine Learning for Spoken Dialogue Systems. In *Proc. Interspeech*, pages 247–255.

[Levin and Pieraccini, 2006] Levin, E. and Pieraccini, R. (2006). Value-based Optimal Decision for Dialog Systems. In *Proc. Spoken Laguage Technology Workshop (SLT)*, pages 198–201.

[Levin et al., 2000] Levin, E., Pieraccini, R., and Eckert, W. (2000). A Stochastic Model of Human-machine Interaction for Learning Dialog Strategies. *IEEE Trans. on Speech and Audio Processing*, 8:11–23.

[Misu and Kawahara, 2007] Misu, T. and Kawahara, T. (2007). Speech-based Interactive Information Guidance System using Question-Answering Technique. In *Proc. ICASSP*.

[Misu and Kawahara, 2008] Misu, T. and Kawahara, T. (2008). Bayes risk-based dialogue management for document retrieval system with speech interface. In *Proc. COLING, Vol.Posters & Demo*, pages 59–62.

[Peters and Schaal, 2008] Peters, J. and Schaal, S. (2008). Natural Actor-Critic. *Neurocomputing*, 71(7-9):1180–1190.

[Rosset et al., 2006] Rosset, S., Galibert, O., Illouz, G., and Max, A. (2006). Integrating Spoken Dialog and Question Answering: the Ritel Project. In *Proc. Interspeech*, pages 1914–1917.

[Roy et al., 2000] Roy, N., Pineau, J., and Thrun, S. (2000). Spoken Dialogue Management using Probabilistic Reasoning. In *Proc. of 38th Annual Meeting of the ACL*, pages 93–100.

[Singh et al., 2002] Singh, S., Litman, D., Kearns, M., and Walker, M. (2002). Optimizing Dialogue Managment with Reinforcement Learning: Experiments with the NJFun System. *Journal of Artificial Intelligence Research*, 16:105–133.

[Williams and Young, 2007] Williams, J. and Young, S. (2007). Scaling POMDPs for Spoken Dialog Management. *IEEE Trans. on Speech and Audio Processing*, 15(7):2116–2129.

[Young et al., 2007] Young, S., Schatzmann, J., Weilhammer, K., and Ye, H. (2007). The Hidden Information State Approach to Dialog Management. In *Proc. ICASSP*.