文書検索型音声対話システムにおける応答生成の最適化戦略のオンライン学習

翠 輝久　　大竹 清敬　　堀 智織　　柏岡 秀紀　　中村 哲

情報通信研究機構 MASTAR プロジェクト

teruhisa.misu@nict.go.jp

1　はじめに

自然言語テキストで記述された文書を検索・提示する音声対話システムにおける対話戦略のオンライン学習を用いた最適化手法を提案する．これまで我々は文書検索タスクを対象に，検索要求のみではなく，特定の情報・事実を求める質問応答機能を利用することで，インタラクティブな情報提示を行う対話システム「京都版ダイアログナビ」を構築してきた [1].

円滑な音声対話を実現するためには，音声認識誤りに対する確認が不可欠である．しかし，不必要な確認はユーザにとってわずらわしいものであり，必要最小限の回数にとどめることが望ましい．また，京都版ダイアログナビは，ユーザの発話を検索要求・質問の 2 つの発話タイプに分類し，それぞれに対応した応答を生成する．実際のユーザの発話には検索要求か質問のいずれであるか判断が困難な場合も多く，また入力される質問の全てに対して，必ず回答が存在するとも限らない．そのため，ユーザの発話が質問である可能性が高い場合でも，質問の回答のみを提示するのが最適な応答であるとは限らない．

我々はこれらの問題に対して，システムが生成可能な応答集合に対して，それぞれのベイズリスクを計算し，応答候補選択の最適化を行う手法を提案してきた [2]．しかし，この手法においては，大量の学習データを利用してパラメータを設定する必要があった．そこで本稿では，この手法を拡張したオンライン学習手法を提案する．さらに，強化学習による最適化手法を示し，収束速度の観点から議論する．

2　情報検索システムにおける対話制御・応答生成

京都版ダイアログナビは，音声入力による文書検索システムであり，質問応答機能を利用することで，単純な情報提示に加えて，ユーザの聞き逃しや関連情報に対する要求などに対処することができる．このシステムにおける対話例を Fig.1 に示す．

このようなシステムにおいて，誤った内容の情報提示を避けるために音声認識やマッチングの尤度が低い場合には，確認を行うことが望ましい．また，質問に対する回答抽出の確信度が低い場合には，質問の回答のみを提示するのではなく，文書全体を提示することが有効な場合もありうる．

Ｕ１：	銀閣寺について教えてください．(検索要求)
Ｓ１：	銀閣寺は京都市左京区にある東山文化を代表する臨済宗相国寺派の寺院です． 銀閣寺の歴史やみどころについて何か知りたいことはありますか？
Ｕ２：	いつ建てられましたか？(質問)
Ｓ２：	銀閣寺の建立は 1482 年です．
Ｕ２：	じゃあ，庭園を説明して．
……	

図 1: 「京都版ダイアログナビ」の対話例

我々はこれらの問題に対して，ベイズリスク最適化の枠組みの下でこれらの問題を統合的に扱い，最適な候補を選択・解釈し，応答生成を行う方法を提案した [2]．本稿では，ベイズリスクに基づく対話制御で考慮した選択肢の中から，応答方法を選択する部分のみを扱う．すなわち，最も検索のスコアの高い文書 d が得られた場合に，最適な応答方法を選択する問題を考える．

検索結果の文書 d を用いて生成する応答 $Act(d)$ は，以下の 4 つからなるものとする．一つ目は，文書の（確認なしでの）提示 $Pres(d)$ であり，文書 d を要約して応答を作成する．二つ目は文書 d を提示することに対する確認 $Conf(d)$ であり，文書のタイトルを基に「金閣寺でよろしいでしょうか？」のような確認を生成する．三つ目は，質問に対する回答の提示 $Ans(d)$ であり，文書 d を基に，ユーザの質問に対する回答を含む一文を提示する．四つ目は，リジェクト $Rej(d)$ であり，文書 d から応答を生成するのを諦めて，ユーザに再発話を求める．

3　ベイズリスクに基づく応答選択

本節では，ベイズリスクに基づく応答生成 [2] について概説する．応答候補に対するベイズリスクは，ユーザに所望の情報を提示したときにシステムが得る報

酬と，確認や誤った情報を提示することによるペナルティに基づいて定義する．すなわち，ユーザが要求している情報を正しく提示した場合には，応答内容に応じた報酬を与える．逆に，誤った内容を提示した場合や，候補のリジェクトを行った場合には，システムがその応答を行ったことにより，余分に費やす時間（＝ユーザが正しい回答を得られるまでに必要な文数で近似）に応じたペナルティを与える．ペナルティは正しい候補を提示した場合には 0 であるが，その他の場合には応答内容に応じた正の値をとる．たとえば，確認を行う場合には，［システムの確認＋ユーザの回答］の 2 発話分，誤った情報を提示した場合には情報提示の失敗に対するペナルティ FP と，ユーザが再発話をしてからシステムが正しい回答を提示するまでの文数の期待値 $AddSent$ がペナルティとして与えられる．なお，正解提示までに必要な文数の期待値 ($AddSent$) は，システムが正しい回答を提示できる確率を 60%とし，ユーザの所望の情報を提示するまでに必要な文数を求めることにより計算した [2].

3.1 ベイズリスクの定義

システムが生成する各応答候補に対するベイズリスクは，ユーザの入力 $\mathbf{W}$ に対して文書 d を提示する場合の成功確率 $p(d|\mathbf{W})$，ユーザの質問に対する回答の成功確率 $p_{QA}(d|\mathbf{W})$，情報提示の失敗に対するペナルティ FP，報酬 Rwd_{Ret}，Rwd_{QA} ($Rwd_{Ret} < Rwd_{QA}$) を用いて以下のように記述できる．

- 文書 d を用いてユーザの質問に回答

$$\begin{aligned} Risk(Ans(d)) = & -Rwd_{QA} * p_{QA}(d|\mathbf{W}) \\ & +(FP + AddSent) * (1 - p_{QA}(d|\mathbf{W})) \end{aligned}$$

- 文書 d を（確認なしで）提示

$$\begin{aligned} Risk(Pres(d)) = & -Rwd_{Ret} * p(d|\mathbf{W}) \\ & +(FP + AddSent) * (1 - p(d|\mathbf{W})) \end{aligned}$$

- 文書 d を提示することに対する確認

$$\begin{aligned} Risk(Conf(d)) = & (-Rwd_{Ret} + 2) * p(d|\mathbf{W}) \\ & +(2 + AddSent) * (1 - p(d|\mathbf{W})) \end{aligned}$$

- リジェクト

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

3.2 検索と質問応答の確信度

ユーザの発話と知識ベース中の文書との類似度を計算するために，単語ベースのベクトル空間モデルを採用する．文書ベクトル $\mathbf{d} = (x_1, x_2, \ldots, x_n)^T$ は，文書中に含まれる名詞に対して，タイトルに重み付けをした出現回数により作成する．ユーザ発話から作成する検索クエリベクトル $\mathbf{W} = (w_1, w_2, \ldots, w_n)^T$ も同様に，音声認識結果中の名詞に対して音声認識の信頼度で重み付けして作成する．なお，x_i, w_i は名詞 i の出現回数である．以上の手順で作成した検索クエリベクトル W と，文書ベクトル d を用いて内積類似度 $Match(W, d)$ を計算する．

内積類似度 $Match(\mathbf{W}, \mathbf{d})$ は，以下のロジスティックシグモイド関数により確信度 $p(d)$ に変換する．この確信度を $p(d|\mathbf{W})$ の近似として用いる．

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

ここで，θ_1，θ_2 はシグモイド関数のパラメータである ($\theta_1 > 0$)．質問応答のスコア $QAScore$ [2] も同様に，別のパラメータ θ_3，θ_4 をもつシグモイド関数により質問に対する回答の確信度 $p_{QA}(d)$ に変換する．($\Theta = (\theta_1, \ldots, \theta_4)$)

4 オンライン学習による対話戦略の最適化

4.1 パラメータの最適化

ベイズリスクに基づく対話戦略の最適化は，文書検索と質問応答の確信度を推定するシグモイド関数のパラメータ $\Theta = (\theta 1, \cdots, \theta_4)$ を最適化することにより行う．すなわち，回答が存在する検索要求・質問に対しては，システムは可能な限り少ない文数で正しい回答を提示するための対話戦略を学習する．逆に，音声認識誤りやシステム想定外発話であるため回答が存在しない発話に対しては，システムは確認を生成したりできるだけ早く対話を切り上げて，再発話を促す対話戦略を学習する．(このような発話に対するシステムの最適な応答はリジェクトである．) このため，回答提示までの文数の期待値の合計を最適化することにより，音声認識精度や検索の成功率を考慮した最適な対話戦略を学習できると期待される．

また，提案手法ではパラメータの更新はユーザ発話が入力されるごとに行われるため，入力される発話の傾向の変化に適応できると期待される．この点が，従来研究で行われきた，応答の成功率を推定する識別器を学習して対話制御の最適化を行う手法 [3] に対する，本手法の利点であるといえる．

具体的な学習の手順は以下の通りである．

1. (t ステップ目において) 文書 d を用いて，応答候補 $Pres(d)$，$Conf(d)$，$Ans(d)$，$Rej(d)$ を生成する．

2. リスク最小となる応答候補を選択することで，応答 $Res_t(d)$ を生成し，実際の報酬/ペナルティを観測する．
3. パラメータ Θ を更新する．すなわち，ユーザの発話が検索要求であった場合には θ_1，θ_2 を質問である場合には θ_3，θ_4 を更新する．
4. 1. に戻る． $t \leftarrow t+1$

4.2 最尤推定による最適化

確信度を求めるために用いるシグモイド関数のパラメータ Θ を最適化するために最尤推定を用いる [4]．学習サンプル $\{< Match_p, C_p > | p = 1, \ldots, t\}$ に対して，成功 (1)，失敗 (0) の教師信号 C_p が与えられているものとする．ここで，式 (1) の出力 (以下 z_p と記述する) を $Match_p$ が与えられた場合の条件付確率の推定値であると考えると，サンプル集合の対数尤度 l は下記の交差エントロピー誤差関数によって与えられる．

$$l = \sum_{p=1}^{t} \{C_p \ln z_p + (1 - C_p) \ln(1 - z_p)\}. \qquad (2)$$

ここでのパラメータ Θ の最尤推定量は，直前 t サンプルの対数尤度を最大化することにより計算される．このように求めたパラメータを Θ^{t+1} として用いる．

シグモイド関数の最尤推定値を数値的に計算するために，尤度関数の Θ による 2 次微分 (フィッシャー情報量) を用いるフィッシャーのスコアリングアルゴリズムを用いる [4]．これは，ニュートン法の一種であり，最尤推定値を行列計算により短時間 (数秒程度) で求めることができる．

4.3 強化学習によるオンライン学習

最適な応答選択のオンライン学習を強化学習により行うこともできる．強化学習によるオンライン学習の目標は，状態空間 S における各応答候補 (行動集合 $\mathbf{A} = (Pres(d), Conf(d), Ans(d), Rej(d))$ の行動価値関数 $Q(S, A)$ を学習することである．文書検索タスクにおける状態 S は，文書検索のスコア $Match(W, d)$ に相当する．文書検索のスコアは，任意の正の値を取りうるため，(有限個の離散状態ではなく) 連続状態に対する行動価値関数 $Q(S, A)$ を学習する必要がある．そのため，行動価値関数を (有限の状態集合に対するルックアップテーブルではなく) 関数近似より定める必要がある．

行動価値関数 $Q(S, A)$ を以下の式で与えられるノコギリ関数と，対応するグリッド点の行動価値 $\mathbf{V}_A = (V_A^0, V_A^\lambda, \ldots, V_A^{n\lambda})$ により近似する．

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

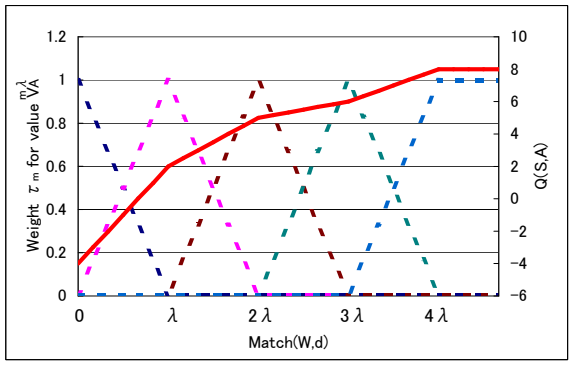

図 2: 価値関数の例

これらの情報を用いて，行動価値関数は $Q(S, A)$ 以下の式により記述される．

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

ここで，λ はグリッド幅であり，n はグリッド点の数である．Fig.2 に行動 A に対する行動価値関数の例を示す．この例では，5 個のグリッド点 ($n = 4$) により $\mathbf{V}_A$ の関数近似を行っており，それぞれの点での行動価値は $(-4, 2, 5, 6, 8)$ である．実際には，質問応答スコア $QAScore$ を x 軸とするもう一つの状態空間があり，質問応答の行動価値関数 $Q(QAScore, Ans(d))$ はこの空間において計算される．最適な応答は現在の状態 S における最大の行動価値 ($Q(S, A)$ もしくは $Q(QAScore, A)$) をもつ行動を選択することで得られる．

価値関数 $\mathbf{V}$ は以下の手順によりオンラインで学習する．ステップ t において，システムは ϵ-greedy 戦略により行動 a_e^t を生成する．すなわち，最大の価値を持つ行動を $1 - \epsilon$ の確率で生成し，$\epsilon(0.2$ に設定) の確率でランダムに行動を生成する．実際に行動 a_e^t を行うことで得られる報酬/ペナルティを用いて，行動 a_e^t の価値関数 $\mathbf{V}_{a_e}$ を以下の TD アルゴリズムにより更新する．

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

ここで，R_{a_e} は，実際に行動 a_e を行った際に得られた報酬/ペナルティである．パラメータ λ，n，δ は経験的に $\lambda = 1.5$，$n = 6$，$\delta = 0.001$ と定めた．

5 オンライン学習手法の評価

検索要求・質問 1,416 発話 (検索要求 1,084 発話，質問 332 発話) を用いてオンライン手法の評価を行った．なお，それぞれの学習において，発話データの 10-fold

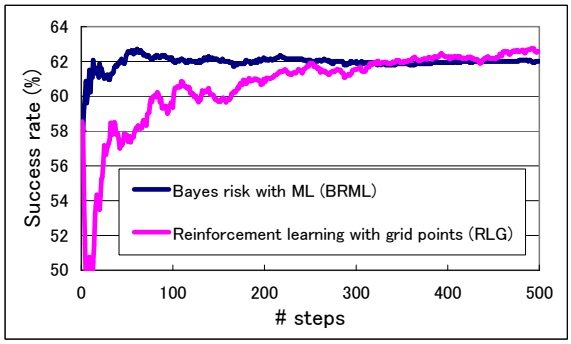

図 3: オンライン学習ステップ数と応答成功率の関係

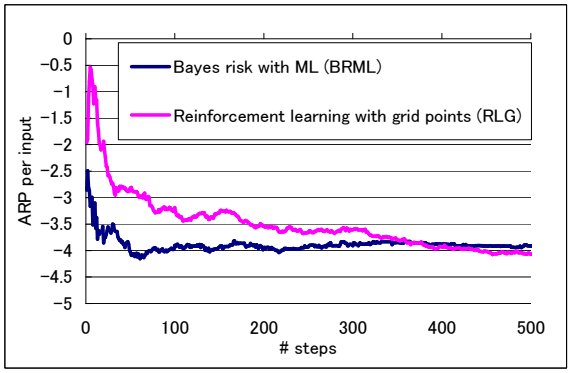

図 4: オンライン学習ステップ数と発話あたりの報酬/ペナルティの関係

クロスバリデーションによる学習・評価を行った．強化学習を用いた学習には，ϵ-greedy 選択によるランダム要因が存在するため，10 回の試行の平均により評価した．

手法の評価尺度として応答成功率と実際の報酬/ペナルティARP を用いた．なお，ARP は，3 節のベイズリスクの式の成功確率 $p(d|\mathbf{W})$ に 1(成功) または 0(失敗) を代入することにより求めた[1]．強化学習の報酬 R にも ARP を利用した．なお，評価データの数や手法の難易度を考慮して検索成功時の報酬として $10(Rwd_{Ret} = 10)$ を，質問応答成功時の報酬として $30(Rwd_{QA} = 30)$ を用いた．誤った応答に対するペナルティFP は，フィールドテストにおける典型的な誤り回復パターンに基づいて 6 を用いた．また，パラメータ $(\Theta, \mathbf{V})$ の初期値として 0 を与えた．

学習ステップ数と，その時点のパラメータを用いて評価セットの応答を選択した場合の応答成功率および 1 発話あたりの報酬/ペナルティ(ARP) の関係を Fig.3, 4 に示す．フィッシャースコアリングアルゴリズムを用いた最尤推定による学習を行った場合 (BRML) と強化学習により最適化を行った場合 (RLG) の性能は、ほぼ同等であった．

次に，学習手法を収束速度を評価した．最尤推定により学習を行った場合 (BRML) には，約 50 サンプルで収束した．これに対して強化学習により最適化を行った場合 (RLG) には収束までに必要なステップ数は約 500 であった．この原因の一つとして，強化学習において行動は (行動間の依存関係などの事前知識を用いず) それぞれ独立のものであると仮定していることが挙げられる．しかしながら，この仮定は (少なくとも，今回のタスクで扱った “提示”，“確認”，“リジェクト” に関しては) 正しくない．例えば，“確認” により報酬が得られた場合には，“提示” により，より大きな報酬が得られるはずである．また，“確認” により，ペナルティが与えられた場合には，リジェクトを行えばより軽いでペナルティで済んだはずである．ベイズリスクに基づく手法においては，これらの行動の報酬/ペナルティの最適化が “成功率” という尺度を通じて同時に行われていることになる．この点が，強化学習を用いた場合と比較して収束速度が早い理由であるといえる．

[1] この判断には人手により付与したラベルを用いた．

6 おわりに

音声入力により情報検索・質問応答を行う対話システムにおいて，ベイズリスクに基づいて最適な応答を生成する対話戦略のオンライン学習による最適化手法を提案した．1,416 発話による評価の結果，初期値からの学習において少数の学習サンプルで最適な戦略が得られることを確認した．

謝辞

本研究を行うにあたり，多大な協力を協力を頂いた京都大学の河原達也先生，NICT の杉浦孔明氏に感謝します．

参考文献

[1] 翠輝久, 河原達也. 質問応答・情報推薦機能を備えた音声による情報案内システム. 情報処理学会論文誌, Vol. 48, No. 11, pp. 3078–3086, 2007.

[2] T. Misu and T. Kawahara. Bayes Risk-based Dialogue Management for Document Retrieval System with Speech Interface. *Speech Communication*, Vol. 52, No. 1, pp. 61–71, 2009.

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

[4] T. Kurita. Iterative weighted least squares algorithms for neural networks classifiers. *New Generation Computing*, Vol. 12, pp. 375–394, 1994.