Vol.2010-SLP-82 No.23
2010/7/24

相談型対話のモデル化と対話戦略の最適化

翠　輝久[†1]　杉浦孔明[†1]　大竹清敬[†1]
堀　智織[†1]　柏岡秀紀[†1]
河井　恒[†1]　中村　哲[†1]

本稿では，ユーザがシステムから情報提示・推薦を受けながら候補を選択する相談型の音声対話システムの枠組みについて述べる．嗜好に合った候補を選択する際には，多くの要因を考慮する必要がある．システムを利用するユーザは，そのような要因の全てを必ずしも把握しているわけではないため，システムはユーザに対して情報推薦を行い，知識のギャップを埋める必要がある．本研究では，このように複数の候補の中からユーザに適した候補を選択する相談型対話のモデルを提案する．京都観光案内タスクにおいて，観光スポットを決定する対話システムの実装を行い，被験者実験を行った．さらに，被験者実験で得られた対話データから，ユーザのシミュレータを作成し，自然政策勾配法を用いた強化学習を用いて対話戦略の最適化を行った．

Modeling Spoken Consulting Dialogue and its Optimization by Reinforcement Learning

Teruhisa Misu,[†1] Komei Sugiura,[†1]
Kiyonori Ohtake,[†1] Chiori Hori,[†1]
Hideki Kashioka,[†1] Hisashi Kawai[†1]
and Satoshi Nakamura[†1]

This paper addresses a spoken dialogue framework that helps users make decisions. Various decision criteria are involved in selecting from a given set of alternatives. Users often do not have a definite goal or criteria for selection, and thus the system has to bridge the knowledge gap and also recommend an appropriate alternative together with the reason for the recommendation through a dialogue. In this paper, we present a model for such consulting dialogue. In order to evaluate the model, we implement a trial sightseeing guidance system and conduct a user experiment. Then, we optimize the dialogue strategy through reinforcement learning with a natural policy gradient approach using a user simulator trained from the collected dialogue data.

1. はじめに

情報検索型の音声対話システムをユーザが利用する際のユーザの対話目的は，検索した情報を得ることではなく，意思決定のための手段であることがある[1)]．例えば，レストランを検索する対話システムを利用するユーザの真の目的は，レストランの価格帯などの情報を調べることではなく，候補の中から価格を基に自分の嗜好と合致したレストランを決めることであるかも知れない．

本研究では，ユーザが対象のドメインに関する知識を十分に有していないために，対話システムを利用して意思決定を行うために必要な情報を収集 (システムに相談) する状況を扱う．このような状況では，ユーザはシステムがどのような情報を提供できるかを知らないだけではなく，ユーザ自身の嗜好（どのような要素を重視して意思決定を行えばいいのか）にも気づいていない可能性がある．また、このような利用シーンではシステム側も同様に，ユーザがどのような要素を重視するかに関してほとんど知識がないことが多い．そこで，システムはユーザの (潜在的な) 嗜好を推定した上で，ユーザが興味を持つ情報を推薦する必要がある．ただし，その際には対話の長さとのトレードオフを考慮する必要がある．

本研究では，このような相談型の対話において，ユーザの知識と嗜好を考慮した対話状態のモデルを提案する．試作対話システムを利用して収集した対話データを利用してユーザシミュレータを構築し，強化学習により対話戦略の最適化を行った．

2. 相談対話のモデル化

2.1 情報案内・提示に基づく相談対話

ユーザがシステムから提示された複数の候補の中から，1 つの候補を選択する状況を考える．例えば，カーナビゲーションシステムが提示した複数のレストランの中から候補を一つ選択するような状況がこれに該当し，実世界においてもしばしば起こりうる状況である．本研究では，ユーザが自身の京都観光に対する知識が乏しい状況の下で，訪れる観光スポットを選択する状況を考える．我々はこれまで，このような状況を想定した人間同士の対話の収集を行い，ユーザが自身の希望を伝達し，ガイドが希望に合った観光スポットを提案し，ユーザがそのスポットの評価を行うという流れで意思決定が行われていることを確認した[2)]．本研究では，相談対話におけるこれらの事象を対象とする．

2.2 意思決定支援システムとしての相談対話

本研究で想定する相談型の音声対話システムは，意思決定支援システムの一種であると考えられる．意思決定支援タスクは，オペレーションリサーチの研究分野において多くの研究事例があり，代表的な手法として階層分析法 (AHP 法)[3)] が提案されている．AHP 法では，問題の要素を「最終目標」,「評価基準」,「代替案」の 3 階層に分け，ユーザの各評価基準に対する局所重み (重要度) を推定することにより最適な意思決定を行う．我々の扱う観光行為の決定支援を行う場合には，最終目的はユーザ自身の嗜好にあった観光スポットを決定す

†1 情報通信研究機構, MASTAR プロジェクト
NICT, MASTAR Project

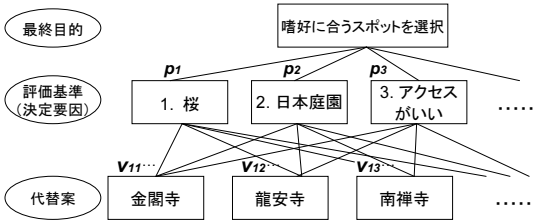

図 1　観光スポット推薦対話における階層構造

ることであり，代替案はシステムが紹介できる観光地のリストである．評価基準には我々が京都観光案内対話コーパスで定義している決定要因[2] を用いる．決定要因は観光行為を確定する上で，その要因になり得るもの (たとえば，桜が有名であるとか，混雑していないなど) を指し．コーパスに付与している発話行為タグ設計の際に意識している事象の 1 つである．これらの要素を用いて作成した観光スポット推薦対話の階層構造の例を図 1 に示す．

ユーザにとっての最適な決定を行うためにまず，評価基準に対する重み $\mathbf{P}_{user} = (p_1, p_2, \ldots, p_M)$，および各代替案に対する各評価基準の観点からの局所重み $\mathbf{V}_{user} = (v_{11}, v_{12}, \ldots, v_{1M}, \ldots, v_{NM})$ を決定する．最適な候補の決定は，優先度 $\sum_{m=1}^{M} p_m v_{km}$ が最大となる代替案 k を選択することで実現される．一般的な AHP 法では，評価基準や代替案に対する一対比較により上記の重みを決定がする．しかしながら，このような手法をそのまま音声対話システムに適用することは難しい．ユーザにとってシステムが提示可能な候補やドメイン知識は，対話を通じて初めて知ることができる情報である場合も多く，対話開始時点で全てが既知であることは少ない[4]．また，システムによっては多数の候補 (代替案) や評価基準を扱う場合も多い．そのような状況下で，一対比較を行うのは非常に多くのやり取りが必要となるため，現実的ではない．

そこで本研究では，システムが持つ情報をユーザに推薦しながら，ユーザの (対話開始時点でユーザ自身気づいていない潜在的なものを含む) 嗜好やドメインに対する知識を推定する枠組みを考える．

3. 音声インタフェースによる意思決定支援システム

3.1 システムの概観

我々が想定する音声対話システムは二つの機能を備えている．一つ目は，ユーザが要求した情報に対して，適切な情報を提供する機能である．システムはユーザが要求した場合に，当該の観光スポットに関する説明や決定要因に関する説明を提示する．二つ目は，ユーザに情報を推薦する機能である．システムはユーザから要求された情報を提供した後に，意思決定を支援するための情報を提供する．(例: システムが提供できる情報を伝達，現在の話題に対する詳細な情報を提供) 以上の戦略に基づくシステムの処理の流れは以下の通りである．なお、4 節において，(4) を強化学習により最適化する．

(1)　ユーザの発話を認識する
(2)　音声認識結果からユーザが要求した観光スポットや決定要因を特定する

表 1　知識ベースの例

スポット名	決定要因	評価	説明文
清水寺	桜で有名	1	境内には約 1,000 本のサクラがあります．とくに本堂から見下ろす一面の桜は絶景です．
	景色がいい	1	清水の舞台は斜面の上に建てられ，ここから望む市街の風景は見事です．
	混雑しない	0	知名度が高く人気のあるため，常にこんでいます．特に観光シーズンは，身動きが取れないぐらいです．
	...	...	...
...			

(3)　(2) の結果に基づいて，情報を提供する
(4)　現在の話題に関する情報を推薦する．

3.2 知識ベース

ユーザが選択可能な京都の観光スポット 15 箇所，決定のための評価基準として 10 種類の決定要因からなるデータベースを整備した．情報検索を行う音声対話システムが扱うエントリ数と比較して候補の数が少ないが，本研究での研究対象はユーザが必須条件 (例えば「京都駅付近の観光地など」) を満たす候補を比較・評価しながら決定を行うプロセスであり，このような状況において，候補数がそれほど多くないこと (15 候補) は実世界でも起こりうると考える．決定のための基準として 10 種類の決定要因からなるデータベースを整備した．今回使用する決定要因は，我々が整備している京都観光案内コーパス中から，観光スポットを決定するために利用されることが多いものを選択した．これらの一覧を表 4 に示す．本来決定要因には依存関係があると考えられるが，(e.g. “桜” や “紅葉” が見たいのは “自然” が見たいため) すべての決定要因は独立かつ並列な関係であると仮定した．

前述の観光スポットに対して，スポットが各決定要因に該当するかどうかの評価 (“1” もしくは “0”) を人手により付与した．また，説明文は，Web から説明の根拠となる文を検索し，話し言葉調に文体を変換することで作成した．試作したシステムで用いる知識ベースの一部を表 1 に示す．

3.3 システムによる情報推薦

ユーザに対する情報推薦の内容は，以下の 6 手法を用いて決定する．ここで，システムの情報推薦の対話行為 (行動) は，発話行為 ca(=推薦手法) と意味内容 sc から構成されるものとする．発話の意味内容は観光スポットと決定要因からなり，推薦手法ごとに決められたルールに基づいて決定される．

(1)　**現在話題のスポットに関する情報推薦 (手法 1)**
直前に説明したスポットについて，詳細な説明を推薦する．具体的には，説明中のスポットに関して，知識ベース中の評価が “1” である決定要因を複数個ランダムに選んでユーザに推薦する．

(2)　**現在話題の決定要因に関する情報推薦 (手法 2)**
直前に説明した決定要因に関連した，別の観光スポットを推薦する．具体的には，当該の決定要因に関する評価が “1” の観光スポットを複数個選択してユーザに推薦する．

(3)　**オープンプロンプト (手法 3)**
特に情報を推薦せず，オープンプロンプトを提示する．

表 2 システムとの対話例

S1:	京都観光案内システムです，お好みの観光スポットを推薦します.
U1:	仁和寺の桜について教えて.
S2-Ans.:	御室桜は，樹高が低く単弁の香り高い白花を根元から咲かせる珍しい桜です．開花時期が遅く，京都の春の終わりを飾ります.
S2-Rec.:	仁和寺は，庭園や，紅葉が有名で，イベントがあります．何か説明しましょうか？ $(Method\ 1\{(Spot_8), (Det_1, Det_6, Det_{10})\})$
...	
U5:	ここの景色はどうですか？
S6-Ans.:	三門の上からは，京都市内が一望できます.
S6-Rec.:	景色が綺麗なところですと，清水寺，伏見稲荷大社，鞍馬寺などが紹介できます. $(Method\ 2\{(Spot_2, Spot_6, Spot_8), (Det_4)\})$
U6:	清水寺.
S7-Ans.:	清水の舞台は斜面の上に建てられ，ここから望む市街の風景は見事です.
S7-Rec.:	他にも世界遺産や，散策できるところ，歴史的に有名なところなどが説明できます. $(Method\ 4\{(NULL), (Det_3, Det_8, Det_9)\})$.
...	
U16:	では，南禅寺に行きます.
S16:	南禅寺に決定しました．他に何か知りたいことはありますか？

表 3 試作システムの推薦手法とユーザの発話内容の関係

	手法 1 (%)	手法 2 (%)	手法 3 (%)
システムの推薦内容	71.5	30.0	-
決定要因	7.0	25.0	38.6
スポット名	16.8	16.7	24.8
スポット名と決定要因	1.4	6.7	2.0
その他（決定の伝達，ドメイン外発話など）	3.3	21.6	34.6

(4) **決定要因の提示 1 (手法 4)**
システムが説明可能な決定要因を提示する．提示内容は，システムが推定するユーザの知識 $\mathbf{K}_{sys}$ が低い順に選択する．($\mathbf{K}_{sys}$, $\mathbf{P}_{sys}$, p_m, $Pr(p_m = 1)$ は，4.2 節で定義する．)

(5) **決定要因の提示 2 (手法 5)**
手法 4 同様に，システムが説明可能な決定要因を提示する．提示内容は，システムが推定するユーザの興味が $\mathbf{K}_{sys}$ が高い順に選択する．

(6) **ユーザが興味があると推定されるスポットの推薦 (手法 6)**
システムが推定するユーザの興味 $\mathbf{P}_{sys}$ に基づいて，$\sum_{m=1}^{M} Pr(p_m = 1) \cdot e_{k,m}$ が最大となるスポット k を選択肢ユーザに提示する．この手法は，推薦システムで利用される協調フィルタリング[5] の考えに基づいたものである．本手法による推薦は，システムがユーザの嗜好を正しく推定できている場合に有効であると考えられるが，推定が不十分な場合には，関係のない情報を提示する可能性が高い．

上記の推薦手法を発話行為 ca_{sys} と，意味内容 sc_{sys} からなる対話行為表現 ($ca_{sys}\{sc_{sys}\}$) により記述する．(例:$Method1\{(Spot_5), (Det_3, Det_4, Det5)\}$, $Method3\{NULL, NULL\}$) これらの機能を備えた対話システムの対話例を図 2 に示す．

表 4 ユーザの嗜好と知識に関する分析

決定要因名	決定に重視する割合 (%)	実際に発話した割合 (%)	システム推薦前に発話した割合 (%)
庭園で有名	34.7	47.2	22.2
混雑していない	19.4	41.7	1.4
世界遺産	48.6	50.0	2.7
景色がいい	48.6	22.2	1.4
アクセスがいい	16.7	19.4	19.4
紅葉が有名	37.5	47.2	18.1
桜が有名	33.3	51.4	13.9
歴史で有名	43.1	31.9	12.5
散策できる	45.8	38.9	1.4
イベントがある	29.2	36.1	8.3

4. 相談対話の対話戦略の最適化

我々はこれまでに，ユーザシミュレータを構築する統計データを収集するために，試作システムを用いた被験者実験を行った[6]．実験の分析結果から、ユーザの行動や，意思決定がユーザの嗜好や知識に影響されれることを確認し，また対話システムの対話戦略を改善することで，ユーザがよりよい選択ができることを確認した．試作システムではユーザの嗜好や知識の推定は行っていないため，手法 1-3 のみをランダムに利用して推薦を行った．これらの推薦手法を行った直後のユーザ発話の分布を図 4 に示す．本節では，このような相談対話の対話戦略を最適化するために，対話状態を表すモデルを提案し，それに基づいて対話戦略 (=推薦手法の選択方法) を最適化することにより，ユーザがよりよい意思決定を行えることを示す．

4.1 ユーザシミュレーションのためのモデル化

4.1.1 ユーザモデル

最初に，ユーザのシミュレーションを行うために，知識ベクトル $\mathbf{K}_{user}$，嗜好ベクトル $\mathbf{P}_{user}$，局所重み行列 $\mathbf{V}_{user}$ の 3 要素からなるユーザのモデルを導入する．本研究では簡単のため，ユーザの嗜好ベクトル $\mathbf{P}_{user} = (p_1, p_2, \ldots, p_M)$ の要素は，1/0 の 2 値からなるパラメータであると仮定する．すなわち，ユーザがある決定要因 m に興味があり (もしくは潜在的に興味があり)，観光スポットを決定する際に重視する場合に p_m は “1” をとるものとする．ユーザが，(ユーザ自身も気づいていない) 潜在的な嗜好を持っている状態を表現するために，ユーザの知識ベクトル $\mathbf{K}_{user} = (k_1, k_2, \ldots, k_M)$ を導入する．ユーザが，システムが決定要因 m を扱えることを知っている，もしくはシステムが決定要因を推薦した場合にベクトルの要素 k_m は，“1” をとる．これらのベクトルを用いることで，例えば，決定要因 m が，ユーザが潜在的に興味を持っている要因であるが，ユーザはそれに気づいていないという状態は ($k_m = 0, p_m = 1$) で表現できる．この設定は，ユーザ自身が対話開始時点から観測可能なゴール状態を持っていると仮定した従来研究 (例えば 7)) と対照的なものである．ユーザの決定要因 m の観点からのスポット n に対する局所重み v_{nm} は (ユーザは，システムから提示された情報のみから判断すると仮定して)，システムが推薦手法 1, 2, 6 を用いてユーザにスポットの評価を知らせた場合に “1” をとるものとする．

Vol.2010-SLP-82 No.23
2010/7/24

4.1.2 ユーザシミュレータ

被験者実験のユーザの統計データと，ユーザの知識・嗜好の状態に基づいたユーザシミュレータを構築した．システムの行動 a^t_{sys} に対するユーザの発話行為 ca^t_{user}，意味内容 sc^t_{user} は以下の式に基づいて生成される．

$$Pr(ca^t_{user}, sc^t_{user}|ca^t_{sys}, sc^t_{sys}, \mathbf{K}_{user}, \mathbf{P}_{user}) = Pr(ca^t_{user}|ca^t_{sys}) \cdot Pr(sc_{user}|\mathbf{K}_{user}, \mathbf{P}_{user}, ca^t_{user}, ca^t_{sys}, sc^t_{sys})$$

すなわち，ユーザの発話行為 ca_{user} は，表 3 の条件付き確率 $Pr(ca^t_{user}|ca^t_{sys})$ に基づいてサンプリングする．推薦手法 4-6 に対するユーザの行動選択には，推薦手法 1 による確率を用いる．ユーザ発話の意味内容 sc_{user} は，ユーザの知識下にあるユーザの嗜好に基づいて決定される．sc は，ユーザが知っている ($k_m = 1$) 決定要因の中から，ユーザが興味の有無に基づいて (コーパスの統計に基づいて) サンプリングする．

4.2 対話状態の記述

前節では，ユーザの対話状態の状態記述方法を定義した．しかしながら，システムはユーザの内部状態 ($\mathbf{P}_{user}, \mathbf{K}_{user}, \mathbf{V}_{user}$) を直接観測することはできないため，ユーザとのインタラクションから推定する必要がある．そのため，このモデルは部分観測マルコフ決定過程 (POMDP) であるといえる．POMDP の状態を解決し，問題をマルコフ決定過程 (MDP) として扱うために，システムが推定するユーザの知識・嗜好の状態を表す確率分布 $\mathbf{K}_{sys} = (Pr(k_1 = 1), Pr(k_2 = 1), \ldots, Pr(k_M = 1))$ および $\mathbf{P}_{sys} = (Pr(p_1 = 1), Pr(p_2 = 1), \ldots, Pr(p_M = 1))$ を導入する[⋆1]．

また，ステップ (ターン)$t+1$ における対話状態 DS^{t+1} は，直前の対話状態 DS^t とユーザ・システム間のインタラクション $I^t = (a^t_{sys}, a^t_{user})$ のみに依存するものとする．このような近似は，対話制御を扱うの多くの研究において採用され，ダイナミックベイジアンネットワークに基づく記述が行われている[8),9)]．本研究の対話状態に対するベイジアンネットワークによる記述を図 2 に示す．

システムが推定するユーザの対話の状態は，確率分布として表現され，インタラクションが行われるごとに更新される．これは，従来研究の多く（例えば 10)）が，ユーザをいくつかの固定のタイプに分類していたのに対して，ユーザのタイプを確率分布として表現することに相当する．システムが想定するユーザの嗜好 $\mathbf{P}_{sys}$ は，直前の対話状態 DS^t を事前分布として，以下のベイズ則を適用することによって更新される．[⋆2]

$$Pr(p_m = 1|I^t) = \frac{Pr(I^t|p_m = 1)Pr(p_m = 1)}{Pr(I^t|p_m = 1)Pr(p_m = 1) + Pr(I^t|(p_m = 0))Pr(1 - Pr(p_m = 1))}$$

ここで，右辺の $Pr(I^t|p_m = 1), Pr(I^t|(p_m = 0))$ は，試作システムによる被験者実験により得られた統計値を利用する．(例えば，ユーザが推薦手法 3 のプロンプトの直後に特定の決定要因 k の情報を求めた場合には，手法 1 の直後に求めた場合よりも p_k が “1” である確率が高い．) ユーザの知識の推定値 k_m は，システムがユーザに決定要因を推薦した場合，もしくはユーザがシステムに当該決定要因を要求した場合に，“1” に更新される．このようにして更新された事後確率は次のインタラクション I^{t+1} を用いた対話状態更新の事前確率として用いられる．状態更新の例を図 3 に示す．

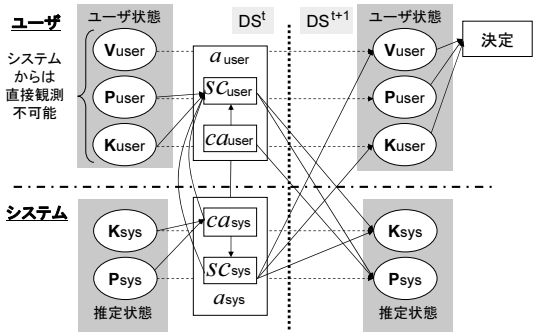

図 2 ベイジアンネットワークによるモデル記述

推定状態の事前確率:
- 知識: $\mathbf{K}_{sys} = (0.22, 0.01, 0.02, 0.18, \ldots)$
- 嗜好: $\mathbf{P}_{sys} = (0.37, 0.19, 0.48, 0.38, \ldots)$

インタラクション（状態観測）:
- システムの情報推薦: $a_{sys} = Method1\{(Spot_5), (Det_1, Det_3, Det4)\}$
  「仁和寺 ($Spot_5$) に関しては，庭園情報 (Det_1)，世界遺産情報 (Det_3)，紅葉情報 (Det_4) が説明できます.」
- ユーザ発話: $a_{user} = Accept\{(Spot_5), (Det_3)\}$
  「世界遺産について (Det_3) 教えて.」

推定状態の事後確率:
- 知識: $\mathbf{K}_{sys} = (1.00, 0.01, 1.00, 1.00, \ldots)$
- 嗜好: $\mathbf{P}_{sys} = (0.26, 0.19, 0.65, 0.22, \ldots)$

ユーザの知識獲得:
- 決定要因に対する知識: $\mathbf{K}_{user} \leftarrow \{k_1 = 1, k_3 = 1, k_4 = 1\}$
- 局所重み: $\mathbf{V}_{user} \leftarrow \{v_{51} = 1, v_{53} = 1, v_{54} = 1\}$

図 3 対話状態更新の例

4.3 報酬関数

ユーザが選択した観光スポットが持つ属性と，ユーザの嗜好との一致率を基に報酬関数を設計する．ユーザは現在の対話状態における知識 $\mathbf{K}_{user}$ と局所重み $\mathbf{V}_{user}$ の下で，最も優先度 $\sum_m k_k \cdot p_k \cdot v_{km}$ が高いスポット k を選択するものとする．報酬 $\mathbf{R}$ は，ユーザが決定

⋆1 本来なら，局所重み v_{nm} に対する重みを導入することが望ましいが，本研究では v_{nm} はシステムがユーザに対して推薦した場合に “1” をとるという仮定を導入しているため，局所重みの推定は行わない．

⋆2 各決定要因に興味があるかを，ユーザに対して明示的に尋ねることもできるが，本研究では暗黙に興味を推定することを目指す．また，仮にユーザがそのような質問に肯定的に回答した場合であっても，p_m は必ずしも “1” ではないと考えられる．

したスポット k が，ランダムにスポットを決定した場合と比較して，どれだけよい選択であるかに基づいて与えられる．

$$R = \sum_{m=1}^{M} p_m \cdot e_{k,m} - \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} p_m \cdot e_{n,m}$$

例えば，決定したスポットが，ユーザの嗜好のうち 3 つを満たし，ランダムにスポットを決定した場合の平均の一致数が 1.3 であった場合，報酬は 1.7 となる．

4.4 強化学習による最適化

4.4.1 MDP の定義

推薦手法の選択方法は，強化学習により最適化できる．まず，マルコフ決定過程 (MDP)**(S, A, R)** を以下のように定義する．対話状態の特徴ベクトルに相当する状態パラメータ $\mathbf{S} = (s_1, s_2, \ldots, s_I)$ は，現在の対話状態 DS^t から，以下の 29 個の特徴量を抽出することにより定義する．

(1) ターン数を表すパラメータ．(ノコギリ関数を利用することで，5 つのパラメータでターン数を表現する．)
(2) 直前のユーザの発話行為 (1 if $a_{user}^{t-1} = x_i$, otherwise 0)
(3) 直前のシステムの発話行為 (1 if $a_{sys}^{t-1} = y_j$, otherwise 0)
(4) ユーザの決定要因に対する知識 ($\sum_{n=1}^{N} Pr(k_n = 1)$)
(5) システムが提示したスポット・決定要因数 (本研究の仮定の下では結果的に $\sum_{n=1}^{N} \sum_{m=1}^{M} v_{nm}$ と一致)
(6) ユーザが決定要因を重視する確率の期待値 ($Pr(k_n = 1) \times Pr(p_n = 1)$) (各決定要因ごと計 10 パラメータ)

システムの行動集合 $\mathbf{A}$ は，3.3 節で定義した 6 つの推薦手法であり，報酬 $\mathbf{R}$ には，4.3 節で定義した報酬関数を用いる．

4.4.2 システムの行動選択のための政策

システムの行動 a_{sys} (ca_{sys}) は，以下のソフトマックス政策に基づいて選択される．

$$\pi(a_{sys} = k|\mathbf{S}) = Pr(a_{sys} = k|\mathbf{S}, \mathbf{\Theta}) = \frac{\exp(\sum_{i=1}^{I} s_i \cdot \theta_{ki})}{\sum_{j=1}^{J} \exp(\sum_{i=1}^{I} s_i \cdot \theta_{ji})}$$

パラメータ $\mathbf{\Theta} = (\theta_{11}, \theta_{12}, \ldots \theta_{1I}, \ldots, \theta_{JI})$ は，J (行動数) $\times$ I (特徴量数) 個のパラメータからなる．パラメータ θ_{ji} は，行動 j の i 番目の特徴量に対する重みであり，行動 j の選択されやすさを決定する．この $\mathbf{\Theta}$ が，強化学習による最適化の対象である．

4.5 自然政策勾配法による対話戦略の最適化

政策を最適化する手法として，自然政策勾配法の一つである Natural Actor Critic (NAC)[11] を利用する．政策勾配法では，状態 S に対する価値関数を直接推定したり，行動価値関数 $Q(S, A)$ を推定することは行わない代わりに，更新前の政策により得られた対話エピソードの報酬を増加させるように自然勾配法により政策 π を直接更新する．

4.6 対話シミュレーションによる評価実験

各シミュレーション対話ごとに，シミュレーション話者 ($\mathbf{P}_{user}, \mathbf{K}_{user}, \mathbf{V}_{user}$) をサンプリングする．擬似話者は，4 つの嗜好を持つものと仮定する (=嗜好ベクトル $\mathbf{P}_{user}$ の 4 つの要素が “1”，残りの要素が “0”)．嗜好の選択には，被験者実験を行った後に行ったアンケートにより調べたユーザの嗜好の分布 (c.f. 表 4) を用いた．ユーザの知識 $\mathbf{K}_{user}$ についても同様に，予備実験においてユーザがシステム推薦前に発話した割合に基づいて設定した．ユーザの局所重み $\mathbf{V}_{user}$ は，ユーザが予備知識を持たないと仮定し，すべてを “0” に初期化した．システム側のパラメータについても同様に，予備実験の結果に基づいてシステムが推定するユーザの嗜好 $\mathbf{P}_{sys}$ と知識 $\mathbf{K}_{sys}$ を初期化した．

シミュレーションを行うに際して，以下の仮定を置いた．システムは，ユーザの発話の音声認識・理解誤りを行わず，その時点での政策 π に基づいて推薦内容を決定する．ユーザは，20 ターン対話を継続するものとし，4.1.2 節のユーザシミュレータに基づき応答を生成する．システムは，ターン毎に 4.3 節の報酬関数に基づいて報酬を与えられる．以上の条件で対話のシミュレーションを行い，2,000 対話ごとに政策 (パラメータ $\mathbf{\Theta}$) を NAC により更新した．

4.7 実験結果

まず最初に，政策反復による報酬の改善について調べた．本研究での手法にはランダム要素が含まれるために，実験結果はすべて 5 回の試行の平均である．図 4 に，行ったシミュレーション対話数 (2,000 対話を 1batch とする) と，2, 5, 10, 15, 20 ターン後の報酬の関係を示す．また，ユーザがドメインに関するすべての知識を持っている場合に決定を行った場合[⋆1]を (Oracle) として併記する．システムの政策は 30,000 対話で収束した．

学習されたパラメータ $\mathbf{\Theta}$ の値を比較・分析することにより，対話戦略を分析した[⋆2]．手法 4，5 では，開始からの対話のターン数が少ないことを表すパラメータに対する重みが大きく，手法 2, 6 においてターン数が多いことを表すパラメータの重みが大きいことが分かった．この結果は，最初にユーザに決定要因に対する知識を与え，ユーザの嗜好を推定した上で，具体的な候補を提示する対話戦略が学習されたことを表している．

次に，学習された対話戦略を，以下の 2 つのベースライン手法と比較した．

(1) **推薦なし (B1)**
システムは要求された情報の提示のみを行い，推薦は行わない．これは，常に手法 3 を選択する場合と等価である．
(2) **ランダムに推薦 (B2)**
システムは，選択可能な 6 手法からランダムに推薦手法を選択する．これは，パラメータ $\mathbf{\Theta}$ の初期値 (すべて 0) における戦略と等価である．

表 5 に，これらのベースライン手法との比較結果を示す．NAC により最適化した対話戦略は，ベースライン手法と比較して有意に大きな報酬を得ることができた ($n = 500, p < .01$)．

さらに，決定するスポットの適合度と対話の長さのトレードオフの問題を考える．ユーザ

⋆1 これには最低 50 ターンは必要である．
⋆2 パラメータ θ_{ji} の大きさはアクション決定を行う際の重要度の指標であると解釈できる

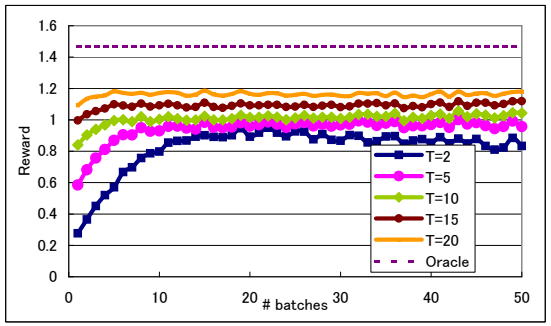

図 4　対話数とターン数毎の報酬の関係

表 5　ベースライン手法との報酬の比較

	報酬 (標準偏差)			
政策	T = 5	T = 10	T = 15	T = 20
NAC	0.96 (0.53)	1.04 (0.51)	1.12 (0.50)	1.19 (0.48)
B1	0.02 (0.42)	0.13 (0.54)	0.29 (0.59)	0.34 (0.59)
B2	0.46 (0.67)	0.68 (0.65)	0.80 (0.61)	0.92 (0.56)

表 6　ベースライン手法との報酬の比較 (推薦にペナルティを与える場合)

	報酬 (標準偏差)			
政策	T = 5	T = 10	T = 15	T = 20
NAC	0.79 (0.63)	0.69 (0.61)	0.59 (0.58)	0.46 (0.55)
B1	0.02 (0.42)	0.13 (0.54)	0.29 (0.59)	0.34 (0.59)
B2	0.30 (0.67)	0.31 (0.66)	0.25 (0.62)	0.14 (0.58)

にとって，次に尋ねたい事項が明確に決まっている場合に情報推薦されることや，既知の内容を繰り返し推薦されることは，わずらわしいものとなる．そこで，推薦行為にペナルティを考慮した上での，対話戦略の最適さを考える．手法 3 以外の推薦手法に 0.05 のペナルティを与える評価関数により対話戦略を学習し評価を行った．この結果を表 6 に示す．

ランダムに推薦手法を選択する手法と比較して，提案法により学習した対話戦略は対話が長引いている場合にも，報酬の減少量が少ない．これは，学習された対話戦略では不必要な推薦を避けているものと考えられる．提案手法により得られた報酬は，ベースライン手法と比較して統計的に有意であった ($p < .01$)．

5. むすび

本稿では，音声対話を通じた情報提示・推薦に基づいて，ユーザが候補の集合の中から候補を選択する支援をする枠組みについて述べた．ユーザの知識と嗜好の両方を考慮する対話状態のモデルを提案し，強化学習により最適化を行い，ベースライン手法と比較してユーザがよりよい意思決定を行えることを確認した．

本研究で利用した推薦手法の数は少なく，単純なものであるが，12) で提案されているシステムの応答の文生成の最適化など，提案手法には多くの拡張が考えられる．また，ユーザの局所重み行列を相槌などのユーザの反応から推定すること[13] も考えられる．さらに，ユーザ発話の音声認識誤りを考慮したモデルの拡張を行う予定である．

参考文献

1) Polifroni, J. and Walker, M.: Intensional Summaries as Cooperative Responses in Dialogue: Automation and Evaluation, *Proc. ACL/HLT*, pp.479–487 (2008).
2) Ohtake, K., Misu, T., Hori, C., Kashioka, H. and Nakamura, S.: Annotating Dialogue Acts to Construct Dialogue Systems for Consulting, *Proc. The 7th Workshop on Asian Language Resources*, pp.32–39 (2009).
3) Saaty, T.: *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*, Mcgraw-Hill (1980).
4) 駒谷和範，池田智志，福林雄一朗，尾形哲也，　奥乃博：音声対話システムにおける文法検証結果と発話履歴に基づくヘルプメッセージ候補のランキング，情報処理学会研究報告.
5) Breese, J., Heckerman, D. and Kadie, C.: "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", *"Proc. the 14th Annual Conference on Uncertainty in Artificial Intelligence"*, pp.43–52 (1998).
6) 　翠輝久，大竹清敬，　堀智織，柏岡秀紀，　中村哲：京都観光案内タスクにおける観光地情報を推薦する音声対話システムの構築と実験，人工知能学会研究会資料，SIG-SLUD-A902-1 (2009).
7) Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H. and Young, S.: Agenda-based User Simulation for Bootstrapping a POMDP Dialogue System, *Proc. HLT/NAACL* (2007).
8) Pietquin, O. and Dutoit, T.: A probabilistic framework for dialog simulation and optimal strategy learning, *IEEE Trans. on Audio, Speech and Language Processing*, Vol.14, No.2, pp.589–599 (2006).
9) Thomson, B., Schatzmann, J. and Young, S.: Bayesian Update of Dialogue State for Robust Dialogue Systems, *Proc. ICASSP*, pp.4937–4940 (2008).
10) Komatani, K., Ueno, S., Kawahara, T. and Okuno, H.: User Modeling in Spoken Dialogue Systems to Generate Flexible Guidance, *User Modeling and User-Adapted Interaction*, Vol.15, No.1, pp.169–183 (2005).
11) 八谷大岳，杉山　将：強くなるロボティック・ゲームプレイヤーの作り方，毎日コミュニケーションズ (2008).
12) Rieser, V. and Lemon, O.: Natural Language Generation as Planning Under Uncertainty for Spoken Dialogue Systems, *Proc. 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (2009).
13) Kawahara, T., Toyokura, M., Misu, T. and Hori, C.: Detection of Feeling Through Back-Channels in Spoken Dialogue, *Proc. Interspeech*, pp.1696–1696 (2008).