システム発話間の整合性を重視した発話選択への深層強化学習の適用

System Utterance Selection Considering Consistency between System Utterances with Deep Reinforcement Learning

黒田 佑樹 *　武田 龍　駒谷 和範
Yuki Kuroda　Ryu Takeda　Kazunori Komatani

大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research (SANKEN), Osaka University

Abstract: Our goal is to develop a listening-oriented dialogue system not by heavily depending on understanding results of user utterances but by simply controlling the sequence of system utterances. We previously implemented system utterance selection that considered consistency between the system's utterances using Q-learning. Deep Q-Network (DQN) can be used for taking more dialogue states into account. In this paper, we report our implementation of DQN to the same task. We used one-hot vectors as input representations of the dialogue states used in Q-learning and normalized rewards. We conducted several text dialogues with the trained model and compared the number of dialogue breakdowns with our previous Q-learning model. We also compared the number of episodes required for the training.

1　はじめに

近年，対話そのものを目的とする非タスク指向型の対話システムが盛んに研究されている．本稿では，その中でもユーザの話の聞き役となるシステム [1][2] に注目する．聞き役の対話システムはユーザの話したい，聞いてほしいという欲求を満たすことが期待される．

我々は，システム発話の列をコントロールするだけで，聞き役の対話システムを実現することを目指している [3]．システム発話の選択においては，文脈的に不適切な発話を選択しないこと，また，ユーザをより満足させるような順序で選択することが重要となる．我々は以前に，システム発話間の整合性に注目して，強化学習の手法の 1 つである Q 学習を行った [4]．

今後より多くの状態を考慮して発話選択を行うために，行動価値関数をニューラルネットワークで近似する深層強化学習を導入する．例えば，ユーザの表情や声色等からユーザの状態を推定し，これを強化学習に組み込めば，ユーザの状態に対応した発話選択が可能となる．

本稿では，深層強化学習の手法の 1 種である Deep Q Network[5]（以下 DQN と呼ぶ）を用いる．これによって当該タスクを DQN を用いて適切に学習させた．また，Q 学習を用いた場合と DQN を用いた場合で学習過程の変化を調べた．

*連絡先: 大阪大学産業科学研究所
〒 567-0047 大阪府茨木市美穂ケ丘 8-1
E-mail: y_kuroda@ei.sanken.osaka-u.ac.jp

2　システム発話間の整合性を重視した発話選択

Q 学習では，ある状態である行動をとる価値 (Q 値) が更新されていく．これは，あるターン t での状態 s_t，行動 a_t と，次のターン $t+1$ での状態 s_{t+1}，報酬 r_{t+1}，また，学習率 α，割引率 γ を用いて式 1 のように表される．この行動価値関数は図 1 のように Q テーブルと呼ばれる表形式で表現される．右辺の第 2 項が 0 となる理想の Q 関数であり，これに近づくように学習が行われる．

$$Q(s_{t+1}, a_{t+1}) = Q(s_t, a_t) + \alpha(r_{t+1} + \gamma * max_a Q(s_{t+1}, a) - Q(s_t, a_t)) \quad (1)$$

我々の以前の研究では，システム発話やユーザ発話の特徴を状態として表現し，選択されたシステム発話に応じて人手で設計した報酬を与えていくことで，適切なシステム発話選択ができるような行動価値関数を求めた．

まず状態として，

状態 *1:* システム発話の対話行為

状態 *2:* ユーザ発話内の特定名詞の有無

状態 *3:* システム発話 ID

の 3 つを定義し，学習時には発話選択の直前のシステム発話とユーザ発話からこれらの状態を取得する．

次に，行動として選択されたシステム発話から，

情報 *1:* システム発話の対話行為

情報 *2:* システム発話内の指示語の有無

情報 *3:* システム発話 ID

の 3 つの情報を取得し，図 2 のように状態と照らし合わせて報酬を与える．

ここで報酬を与える基準として,「対話行為の整合性」,「指示語の整合性」,「発話内容の整合性」の 3 点を考える．

「対話行為の整合性」に関しては，連続するシステム発話の対話行為の順番に報酬を与える．例えばシステムが質問をすると，ユーザが何らかの応答を返すことが予想される．この時，システムはすぐ次の質問をするより，何らかの反応（応答）を返した方が良いというような関係を決めておき，それに従って報酬を与えた．

「指示語の整合性」に関しては，システム発話内に指示語がある場合，その指示語が指す内容がユーザ発話に存在するかを確認する．システム発話内に指示語があるのに，直前のユーザ発話に指示語が指す内容（特定名詞）が存在しない場合，負の報酬をつけた．

「発話内容の整合性」に関しては，直前のシステム発話から予想されるユーザ発話と噛み合わないシステム発話の選択を防ぎたい．例えば，図 3 のように,「競技は何をご覧になりますか」のような質問には，ユーザは何らかの競技を答えると予想できる．それに対する，システムの「それは大変ですよね」といったような食い違った反応の発話選択は不適切である．これを防ぐため，内容的に食い違う発話 ID の組み合わせを人手で記述した blacklist を用意しておいて，学習時に blacklist に該当するような発話の組み合わせが選択されたら，負の報酬を付けた．

また，ここまでの 3 つの状態は言語情報のみを用いた適切な発話選択のためのものであったが，それ以外に，最低限のユーザ満足度の考慮のために，ユーザが対話を楽しんでいるかどうかを表す

状態 *4:* ユーザ心象

も状態として加えた．ここでは単純に，ユーザから入力された心象の高低に応じて報酬をつけた．さらに連続して高低が続く場合にはより大きな報酬をつけた．

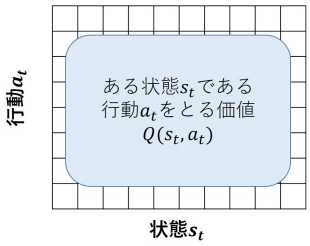

図 1: Q 学習の行動価値関数

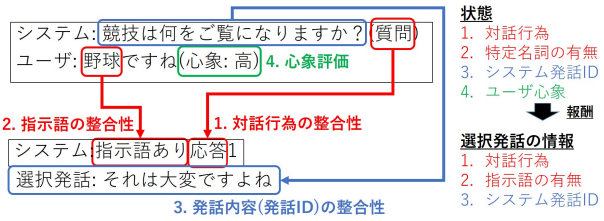

図 2: 整合性を考慮する上で注目する関係

3 深層強化学習 (DQN) を用いた実装

本章では 2 章で説明した，システム発話間の整合性を重視した発話選択を DQN を用いて再現する方法を述べる．DQN を用いた実装では，Q 学習 (図 1) を用いる場合とは違い，行動価値関数を図 4 のようなニューラルネットワークで表現する．そのため，入力層に与える状態の表現方法や，ニューラルネットワークのハイパーパラメータの設定を考える必要がある．本章では，この中でも状態の入力表現，中間層のノード数に着目して設計の詳細を述べる．また，これらを決定する際の試行から得られた，適切な学習のための知見に関しても報告する．

3.1 Deep Q Network

Deep Q Network とは，Q 学習ではテーブル形式で表現されていた行動価値関数を，ニューラルネットによって関数近似したものである．Q 学習では全状態に関して Q 値を更新しなければならないが，そのため，

S: 競技は何をご覧になりますか？
U: 野球です
S: それは大変ですよね
U: 大変ですかね

図 3: 内容的に不適切な発話選択の例 (赤字部分)

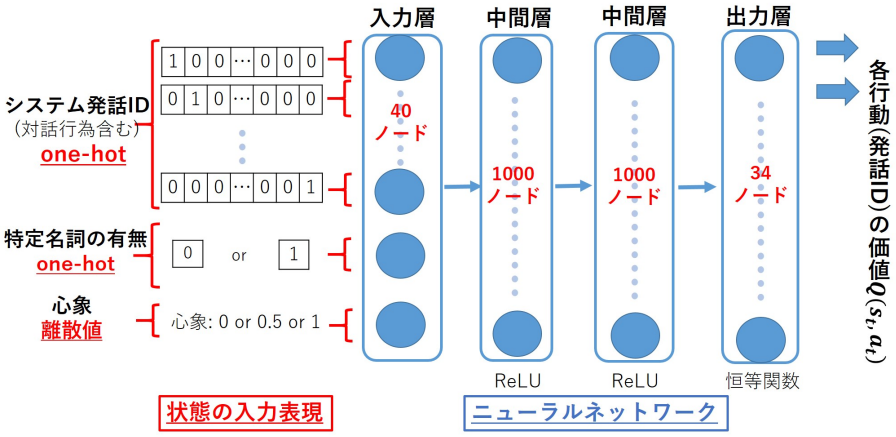

図 4: DQN を用いる場合の設計

DQN では Q 学習に比べて多くの状態を扱うことができる．

DQN で学習されたニューラルネットワークは，Q 学習における状態を入力とし，その状態における各行動の Q 値を出力する．ニューラルネットワークの学習方法としては，損失関数を最小化するように重みを調整する．ここでの損失関数は，あるターン t での状態 s_t，行動 a_t と，次のターン $t+1$ での状態 s_{t+1}，報酬 r_{t+1}，また，学習率 α，割引率 γ を用いて式 2 のように表される．

$$E(s_t, a_t) = (r_{t+1} + \gamma * max_a Q(s_{t+1}, a) - Q(s_t, a_t))^2 \quad (2)$$

また，DQN の学習安定のために以下のような工夫がある．

Experience replay

Q 学習では各ステップごとに学習を行っていたが，ニューラルネットワークを用いる場合，時間的に相関が高い内容を学習してしまい，過学習が起きやすい．そこで，状態，行動，報酬，次の状態を各ステップごとにメモリに蓄積していき，ランダムにサンプリングしてから学習する．

Target Q Network

ニューラルネットワークの更新には式 2 の損失関数が用いられるが，式内の Q 値の計算に毎回更新した直後のニューラルネットワークを用いると学習が安定しない．そのため，Q 値の計算に用いるニューラルネットワークはしばらくの間固定し，一定間隔で更新していく形を取る．

3.2 入力表現

DQN を適用する際には，Q 学習における状態をニューラルネットワークで扱える形にする必要がある．具体的には，2 章で示した 4 種の状態をニューラルネットワークの入力表現にする．

まず,「システム発話の対話行為」に関しては，システム発話 ID がより細かく分類したものにあたるため，考慮しない．

次に,「ユーザ発話内の特定名詞の有無」に関しては，2 値を取るため 0 or 1 で表現した．

また,「ユーザ心象」に関しては，低，中，高の 3 段階に離散化し，それぞれに 0, 0.5, 1 の数値を割り当てた．

「システム発話 ID」の表現方法を決定する際には，発話 ID(1〜38) を正規化して用いる方法と，ID 数分の長さの one−hot ベクトルを用意し，該当 ID の発話が現れたときに 1 をたてる方法の両方を試した．この結果を縦軸が行動に番号を振ったもの，横軸が状態の組み合わせに番号を振ったものである，Q テーブルのヒートマップ図 5,6 に示す．

結果として，図 5 のように，ID を正規化して用いる方法ではほとんどの状態で行動価値が同じになってしまい，適切に学習できているとは言えない．一方，one−hot ベクトルを用いる方法では，図 6 ように，状態ごとの行動価値に適度なばらつきが見られる．これは ID 化された発話間に数学的に連続な関係がないことが原因であると考えられる．

この結果から,「システム発話 ID」の入力表現には one−hot ベクトルを用いる方法を採用した．

3.3 中間層のノード数の設定

中間層のノード数を決定する際には，1,5,10,100,1000 と変え，Q テーブルの変化を見ながら値を選んだ．この結果を，縦軸が行動に番号を振ったもの，横軸が状態の組み合わせに番号を振ったものである，Q テーブルのヒートマップ図 7 に示す．図 7 を見ると，中間層のノード数が 1 や 5 と小さいときには，異なる状態でも同じよ

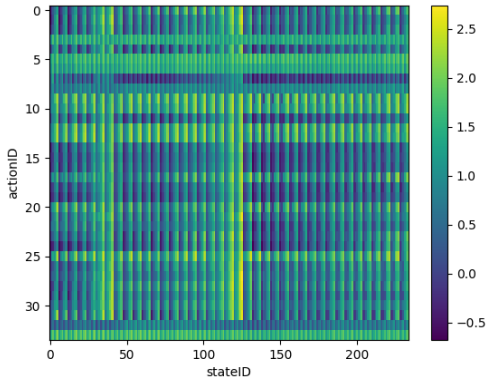

図 5: 発話 ID を正規化して用いた時の Q テーブル

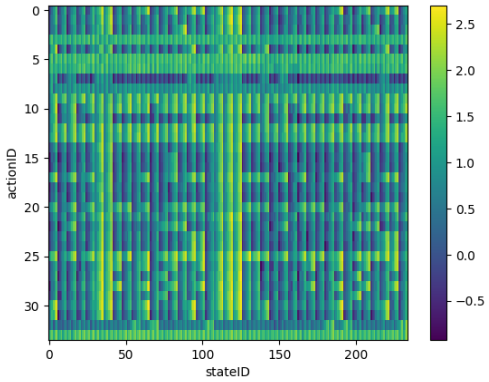

図 6: one−hot ベクトル用いた時の Q テーブル

うな行動の価値が高くなっている．一方，10,100,1000 とノード数を増やしていくと，徐々に状態ごとの行動価値にばらつきが現れ，100〜1000 程度で安定した．これは中間層のノード数を増やすごとに，ニューラルネットワークの表現力が上がっているためと考えられる．

この結果から，中間層のノード数は 1000 とした．

4 Q 学習による実装と DQN による実装の比較

本章では，システム発話間の整合性を重視した発話選択の学習方法として，Q 学習を用いた場合と，DQN を用いた場合の比較を行う．

まず，学習過程における違いを調べると，十分に学習できるまでにかかるエピソード数に違いがあったため，その回数を報告する．

次に，両者の Q テーブルを見比べ，今回実装した DQN による手法で Q 学習を用いた場合と同様の学習ができているかを調べる．また，実際に対話を行い，不適切な発話が選択される数をカウントすることで，システム発話間の整合性を重視した手法の再現ができていることを確かめる．

4.1 実験条件

システム発話のデータは雑談対話コーパス Hazumi1902[6] 収集時に使用された発話から,「スポーツ」,「音楽」,「食事」,「旅行」の話題にラベル付けされたものを用いた．具体的には，それぞれの話題の発話集合に，どの話題でも用いることのできる「default」発話を加えた発話集合を作成し，4 つそれぞれに関して Q テーブルを学習した．発話集合はいずれも 60〜70 発話で構成されている．

Q 学習で十分な学習を行うためには，膨大な数の対話サンプルが必要となる．しかし，そのような対話サンプルを実際に収集するのは現実的ではない．そのため，システム発話に対応するユーザ発話とユーザ心象を出力するようなユーザモデルを用いた．具体的には，ユーザモデルはシステム発話に対して，コーパス収集時と同じ発話を返す．

学習時の行動選択方策としては，確率 epsilon でランダムな行動選択をし，それ以外で Q 値が最大となる行動選択をする epsilon greedy 法を用いた．強化学習においては，学習の初期フェイズではランダムな探索を多く行い，十分な探索が行えたら学習結果を用いた行動選択によって Q 値の安定を図るのが望ましい．そこで，今回の実装では，epsilon を初期値 1，終値 0.1 として全体のエピソード数の 1/4 の学習を終えるまで減衰させ続けた．現在のエピソード数を c_e（current_episode），全体のエピソード数を e（episode）とすると，epsilon は式 3 のように表される．

$$epsilon = max(1 - 0.9 * c\_e * (\frac{e}{4}), 0.1) \qquad (3)$$

学習エピソード数は，10 交換の対話を 1 エピソードとし，それぞれの手法において，十分に学習されるまでとした．強化学習において，十分に学習されたと判断されるタイミングは，得られる報酬の平均が変化しなくなった時である．しかし，学習の初期フェイズでランダム探索が十分でない場合，報酬の平均が低い値で安定してしまうことがある．そこで，本稿では報酬の平均が変化しなくなるエピソード数を検証し，それを十分な学習に必要なエピソード数とした．

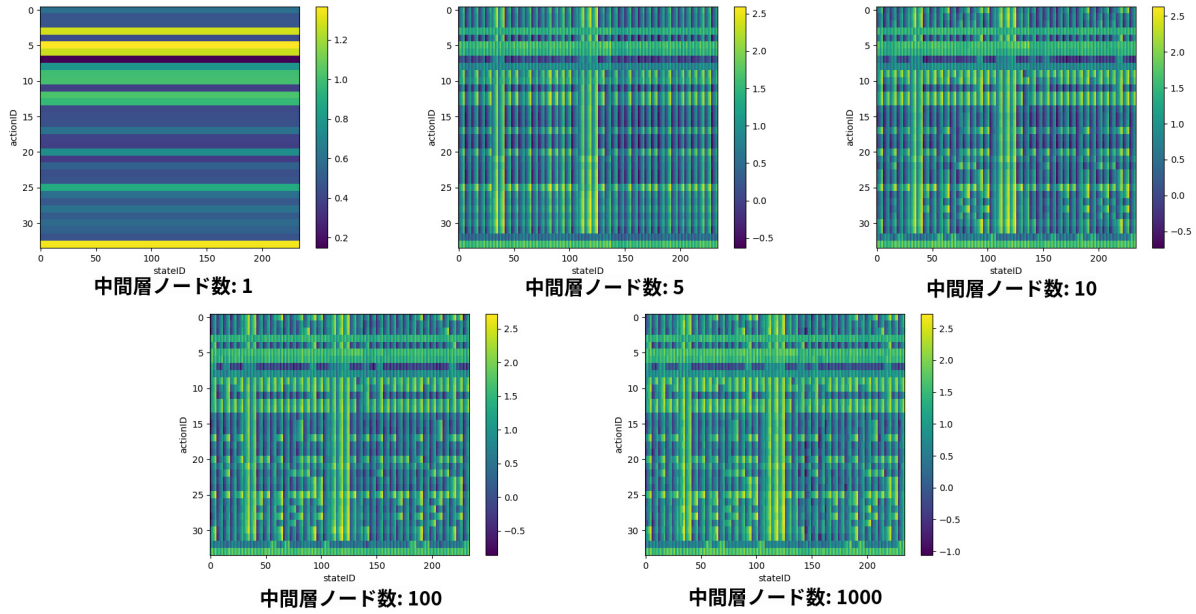

図 7: DQN における中間層ノード数ごとの Q テーブル

4.2 十分な学習に必要なエピソード数の比較

十分な学習に必要なエピソード数を検証する際，Q 学習と DQN で明らかな差が生じた．図 8 にそれぞれの手法において，平均報酬が変化しなくなった時のエピソードごとの報酬推移を示す．結果として，Q 学習では 500000 回，DQN では 50000 回学習させたところで報酬平均の変化が止まった．

このことから，DQN を用いた手法では，Q 学習より少ないエピソード数で十分な学習が行えることがわかる．

4.3 Q テーブルによる比較

Q テーブルを比較することで DQN を用いた実装で，Q 学習を用いた手法の再現ができているか調べた．ただし，DQN の Q 関数はニューラルネットであるため，一旦状態に対応する Q 値をニューラルネットから出力し，テーブル形式に直して比較した．

比較した結果として，同様の Q テーブルが得られていることを確認した．

4.4 対話による比較

話題「スポーツ」に関して 10 交換の対話を 10 セットずつ行い，Q 学習を用いた手法と DQN を用いた手法の不適切な発話が選択される数をカウントした．ここで，適切な発話と不適切な発話の例を図 9 に示す．不適切な発話とは例のように，選択したシステム発話に違和感が感じられるような場合である．

表 1: 対話における不適切な発話選択の数

	適切	不適切
Q 学習を用いたシステム	94	6
DQN を用いたシステム	95	5

結果は表 1 のようにほぼ同数であり，不適切な発話を防ぐという点で同等の性能を再現できていることがわかる．

5 おわりに

本稿では，聞き役対話システムの発話選択の強化学習による定式化において，将来的により多くの状態を考慮できるように，DQN の枠組みの導入を目指した．以前 Q 学習を用いて実装したシステム発話間の整合性を重視した発話選択を，DQN を用いて再現することにより，入力表現や報酬の扱い，パラメータの設定などに関して，当該タスクへの DQN の適用方法の知見を得た．また，Q 学習の場合と学習過程を比較することで，より少ないエピソード数で十分な学習が行えることを確かめた．

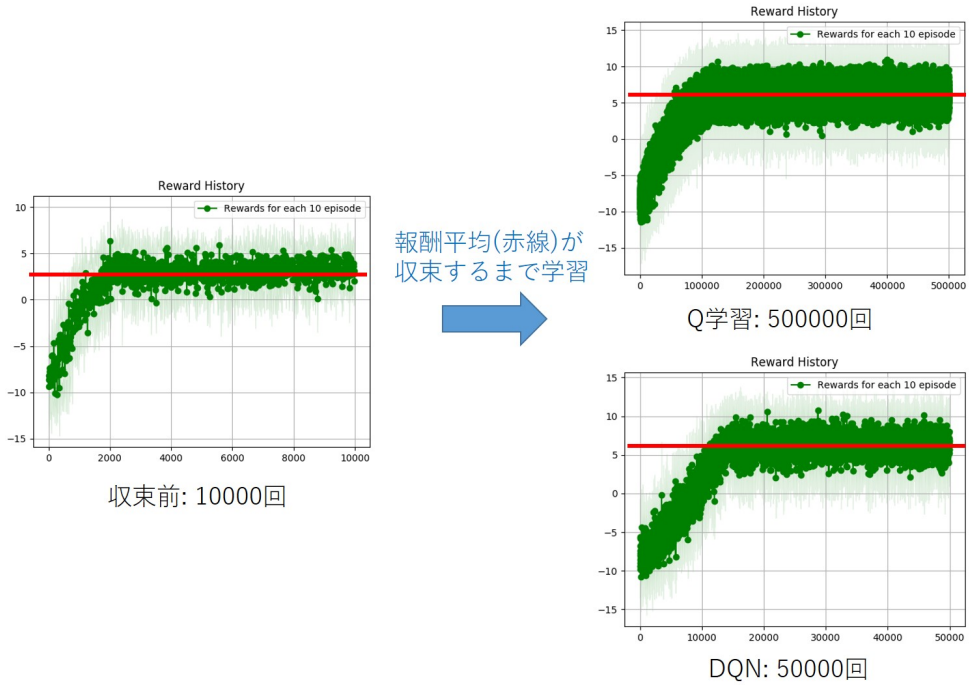

図 8: Q 学習と DQN における報酬平均の収束

【不適切】
S: 競技は何をご覧になりますか？
U: 野球です
S: それは大変ですよね
U: 大変ですかね

【適切】
S: 競技は何をご覧になりますか？
U: 野球です
S: それは面白そうですね
U: ええとても面白いです

図 9: 不適切な発話選択と適切な発話選択の例 (赤字)

参考文献

[1] Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka. Learning to control listening-oriented dialogue using partially observable Markov decision processes. *ACM Trans. Speech Lang. Process.*, Vol. 10, No. 4, January 2014.

[2] 石田真也, 井上昂治, 中村静, 高梨克也, 河原達也. 傾聴対話システムのための発話を促す聞き手応答の生成. *SIG-SLUD*, Vol. B5, No. 01, pp. 1–6, aug 2016.

[3] 西本遥人, 武田龍, 駒谷和範. 対話コーパスに基づく新たなシステム対話行為の設計の検討. *SIG-SLUD*, Vol. B5, No. 02, pp. 101–103, 2019.

[4] 黒田佑樹, 武田龍, 駒谷和範. システム発話間の内容的整合性を用いた強化学習に基づく発話選択. 情報処理学会第 83 回全国大会, pp. 4P–03, march 2021.

[5] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. 2013.

[6] 駒谷和範, 岡田将吾. 複数の主観評定を付与した人システム間マルチモーダル対話データの収集と分析. 信学技報, Vol. 119, No. 179, pp. 21–26, 2019.