シミュレータを使用した実験の検討

1. 実験概要

2019年度実施したSCCFの特性調査実験では、カテゴリ数が状態行動数の削減のために6個であることや、2回前までの発話内容を格納するなど、状態行動数が少ない状態での学習ができることを確認した。

本実験では、状態行動数を増やし、深層強化学習アルゴリズムをSCCFに適用することで学習が可能であるかについて、性能を調査する。

ただし、実験を簡単にすることで、サイクルを早くするため、IDAは実際の利用者ではなく、エージェントを評価するシミュレータを対象として実験を実施する。

対象アルゴリズム

- REINFORCE
- DQN（DQNの派生系も実装する予定です）
- そのほか未定

2. エージェントのパラメータ/実験環境

- 状態: 過去T回までの発話したカテゴリ
- 行動: 次に発話するカテゴリ
- エージェントが発話するカテゴリの総数: C
- エージェントが持つ、過去の発話履歴の長さ: T (Tステップ前までの発話を保存)

よって、エージェントのもつ状態行動の総数は

$$\text{状態・行動} = C^T \cdot C = C^{T+1}$$

3. シミュレータのパラメータ設定

- シミュレータがエージェントが与える報酬:

  $$r \in \{1, 0, -1\}$$

  DQNの場合、報酬値をClipping(1,0,-1の値に絞る)することで学習が安定化するため。

  2019年度のREINFORCEの実験で性能の良かった実験設定（報酬設定B）

- シミュレータがエージェントに与える報酬がrである確率

  $P(r|s,n)$：特定の発話$c \in C$がn回連続で続いた時、さらにcが発話された条件のもと、報酬rを与える確率

具体例:

$$
\begin{leftline}
P_x(r=1|c,n) = [\\
\quad \text{ここに示す、配列の}1\text{番目から、} n=1, n=2, n=3, \ldots \\
\quad [0.5, 0.9, 0.9, \ldots](2\text{回目以降}\alpha = 1.0)(c=1) \\
\quad [0.4, 0.45, 0.1, \ldots](3\text{回目以降：}\alpha = 0.22)(c=2) \\
\quad [0.5, 0.15, \ldots](3\text{回目以降：}\alpha = 0.3)(c=3) \\
\quad [0.35, 0.5, \ldots](3\text{回目以降：}\alpha = 0.3)(c=4) \\
\quad [0.45, \ldots](3\text{回目以降：}\alpha = 0.1)(c=5) \\
\quad [0.4, \ldots](3\text{回目以降：}\alpha = 0.1)(c=6) \\
]
\end{leftline}
$$

ただし、ここでα：減衰係数

（上記において...と示している箇所は、その一つ前の数に減衰係数をかけた値として続く）

例：$P_x(r=1|c=2,n) = [0.4, 0.45, 0.1, 0.1 \cdot 0.2, 0.1 \cdot 0.2^2, 0.1 \cdot 0.2^3, \ldots]$

また、減衰係数がかかる前までの、値が記載されている箇所は、2019年度の実験で、実際に実験協力者が与えた報酬確率を用いている（実測値を有効数字2桁で四捨五入）

減衰係数の決め方

減衰係数がかかるまでに2つデータがある箇所：直前の勾配を利用

例：$P_x(r=1|c=1,n) = [0.4.0.45, 0.1, \ldots]$ この時 $a = 0.1/0.45 = 0.22$

ただし、直前の勾配が1以上の時：$\alpha = 0.2$

データがない箇所：$\alpha = 0.1$

4. 実際に利用するシミュレータのパラメータ

以上の設定の元、SCCF特性調査実験において、特徴的だった実験協力者4名のデータから、パラメータを用意する。

ただし、報酬rを与える確率の合計値が1にならない箇所もあるが、この時はルーレット選択で対処する。

- user1
  Chatbot-REINFORCEでの実験結果:1,2,3,4,5,6位と順に発話確率が高くなった

```
P(r|c,n) =
[ (r = 1)
    [0.5, 0.9, 0.9 ... (2回目以降 α=1.0)] (c=1)
    [0.4, 0.45, 0.1, ...] (3回目以降: α=0.2) (c=2)
    [0.5, 0.15, ...] (3回目以降: α=0.3) (c=3)
    [0.35, 0.5, ...] (3回目以降: α=0.3) (c=4)
    [0.45, ... ](3回目以降: α=0.1) (c=5)
    [0.4, ... ](3回目以降: α=0.1) (c=6)
],
[ (r = 0)
    [0.45, 0.1, 0.1, ... ] (2回目以降α=0.1)
    [0.55, 0.3, ...] (3回目以降: α=0.55)
```

```
    [0.5, 0.1, ...](3回目以降: α=0.1)
    [0.55, 0.5, ...] (3回目以降: α=0.9)
    [0.4, ... ](3回目以降: α=0.1) (m=5)
    [0.5, ... ](3回目以降: α=0.1) (m=6)
],
[ (r = -1)
    [0.001, ...] α=0.1
    [0.05, 0.2, ...] α=0.2
    [0.02, ...] α=0.1
    [0.1, ...] α=0.1
    [0.15, ...] α=0.1
    [0.1, ...] α=0.1
]
```

- user4
  Chatbot-REINFORCEでの実験結果: 1,3位のカテゴリの発話確率が増加

  ```
  調査中
  ```

- user7
  Chatbot-REINFORCEでの実験結果: 2,4位のカテゴリの発話確率が増加

  ```
  調査中
  ```

- user8
  Chatbot-REINFORCEでの実験結果: 3,4位のカテゴリの発話確率が増加

  ```
  調査中
  ```