4P-03

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

黒田 佑樹 † 武田 龍 ‡ 駒谷 和範 ‡

† 大阪大学 工学部電子情報工学科 ‡ 大阪大学 産業科学研究所

1. はじめに

用意した話題についてユーザに話してもらう，聞き役の雑談対話システムの実現を目指している．これに際して，発話の候補集合を用意し，その中からシステム発話を適切に選択するというアプローチを採っている．このような対話システムでは，システムの質問に対するユーザ応答のパターンはある程度限定できる．このため，ユーザ発話の解析に偏重せず，システム発話の順序をコントロールすることで，対話を成立させることを狙う．

対話システムにおいて，適切なシステム発話を選択するという問題は，強化学習で定式化できる．我々のグループの以前の研究では，対話行為を設計してシステム発話集合を分類し [1]，この対話行為に基づき状態と行動を設計していた．ここでは対話行為単位で行動を選択した後，その中に含まれる発話をランダムに選択していたため，不適切なシステム発話が選択されることがあった．

本稿ではまずシステム発話の内容に基づいてシステム発話集合を人手で分類し，状態と行動を再設計する．次に新たな状態行動セットに，システム発話の内容の順序の整合性に基づく報酬を設計する．これらにより，システム発話が適切に選択される回数を増加させる．

2. 現有システムの分析と新たな設計

破綻の生じやすい対話行為順に含まれる対話行為を意味内容レベルで細分化して，状態や行動を再設計する．それぞれの状態や行動に関して内容的整合性に則した報酬を与えて強化学習を行うことで，破綻の削減を図る．

2.1 内容的破綻の生じやすい対話行為順

まず本研究でベースとした，西本らの手法 [1] に基づくシステム（以降これを従来システムと呼ぶ）で得られる結果を分析した．このシステムは，少数のシステム発話を表 1 に示す 8 つの対話行為に分類し，それに基づく強化学習により発話を選択する．ここでの特定話題は，従来システムで扱われている話題に関する内容の発話を指す．また default とはあらゆる話題に使える一般的な発話を指す．

この結果，内容的破綻が生じやすい対話行為順として，以下の 3 種類のパターンがあった．

1. (c) 指示語なし質問 → (b) 指示語あり質問 (特定話題) or (a) 指示語あり質問 (default)
2. (b) 指示語あり質問 (特定話題) or (c) 指示語なし質問 → (d) 指示語あり応答 (特定話題+default) or (g) 感謝
3. (a) 指示語あり質問 (default) → (d) 指示語あり応答 (特定話題+default) or (g) 感謝

それぞれの対話行為順で生じる破綻の例を図 1 に示す．ここで S と U はそれぞれシステムとユーザの発話を表す．システム発話末尾の () 内はその対話行為を示す．赤字は破綻箇所である．パターン 1 の破綻は，主に指示語の指す対象がユーザ発話中に存在しないことが原因である．パターン 2 の破綻は，主に指示語の指す対象がその後のシステムの反応とかみ合っていないことが原因である．パターン 3 は破綻の直前のシステム発話を見るだけでは破綻かどうかわからないが，その一つ前のシステム発話を見るとパターン 1 と同様の理由で破綻している．これは (a) 指示語あり質問（default）の意味が，直近の (c) 指示語なし質問もしくは (b) 指示語あり質問（特定話題）に依存するためである．

表 1: [1] で用いられていた対話行為 8 種類

対話行為	説明
qs_o_d	(a) 指示語あり質問 (default)
qs_o_s	(b) 指示語あり質問 (特定話題)
qs_x_s	(c) 指示語なし質問 (特定話題)
re_o_m	(d) 指示語あり応答 (特定話題+default)
re_x_d	(e) 指示語なし応答 (default)
re_x_m	(f) 指示語なし応答 (特定話題+default)
thank	(g) 感謝
io	(h) 情報提供

【パターン 1】
S: 楽器は何を演奏するのですか？(qs_x_s)
U: ピアノを弾きます
S: その歌手のどういうところが好きなんですか？(qs_o_s)
U: どの歌手？

【パターン 2】
S: 競技は何をご覧になりますか？(qs_x_s)
U: 野球です
S: それは大変ですよね (re_o_m)
U: 大変ですかね

【パターン 3】
S: おすすめの曲はありますか？(qs_x_s)
U: スピッツの楓という曲がおすすめですよ
S: 具体的に教えてください？(qs_o_d)
U: 歌詞が詩的で素敵なんですよね
S: それは大変ですよね。(re_o_m)
U: 大変ですかね

図 1: 3 つの破綻パターンの例

2.2 意味内容を考慮した強化学習の設計

パターン 1 や 2 の破綻を防ぐためには，まず破綻が生じやすい対話行為内の発話を，他の対話行為とは独立の状態，行動とする．次に，あるシステム発話の後にこのシステム発話が来ると不自然であるという関係を人手で与える．この関係に負の報酬を付けて学習することで，

System Utterance Selection based on Reinforcement Learning using Consistency between Utterance: *Yuki Kuroda, Ryu Takeda, and Kazunori Komatani (Osaka Univ.)*

不適切なシステム発話の選択を防ぐ．

パターン 3 の破綻を防ぐには，(a) 指示語あり質問 (default) より前の，(c) 指示語なし質問もしくは (b) 指示語あり質問 (特定話題) に注目する必要がある．まず，(a) 指示語あり質問 (default) の発話と，その直近に現れた (c) 指示語なし質問もしくは (b) 指示語あり質問 (特定話題) の発話の組み合わせを状態とする．(c)，(b) と，(d) 指示語あり応答（特定話題）もしくは (g) 感謝との関係をパターン 1，2 と同様に関係付けることで，同様に不適切なシステム発話の選択を防ぐ．

これらにより，破綻を防ぐより詳細な発話選択の戦略を強化学習する．

3. 強化学習の実装の詳細

発話集合は雑談対話コーパス Hazumi1902[3] 収集時に使用された発話から,「スポーツ」の話題と，どの話題にも用いることのできる「default」発話を抜き出した計 67 発話である．

2.1 節に示した 3 パターンの関係を表現するため，破綻が起きやすい対話行為順の後ろの対話行為を発話ごとに分け，それぞれを行動とする．それ以外の対話行為内の発話はランダムに選んでも問題ないため 1 つの対話行為を 1 つの行動とする．(a) 指示語あり質問 (default)，(b) 指示語あり質問 (特定話題)，(d) 指示語あり応答 (特定話題+default，(g) 感謝は役割がほぼ重複する発話を除いてそれぞれ 3 個，7 個，17 個，3 個の発話があり，これら以外の対話行為は 4 個あるため，行動数は合わせて 34 となる．

状態も行動同様破綻が生じやすい対話行為順の前の対話行為を発話ごとに分け，それ以外は 1 対話行為 1 状態とする．(a) 指示語あり質問 (default)，(b) 指示語あり質問 (特定話題)，(c) 指示語なし質問は役割がほぼ重複する発話を除いてそれぞれ 11 個，5 個，7 個の発話があり，これら以外の対話行為は 5 個ある．またこれに加えて 2.2 節で述べたように (a) 指示語あり質問 (default) 内の発話と，その直近の (b) 指示語あり質問 (特定話題) か (c) 指示語なし質問の発話の組み合わせを状態として表現する必要がある．これが 7*(11+5) 個であるため，合計の状態数は 140 となる．

報酬は連続する 2 つのシステム発話の整合性に対して与える．具体的には各システム発話にこの発話の後に来てはいけないという発話集合（blacklist）を設け，各システム発話が選択されたときにその前のシステム発話が blacklist 内にあれば負の報酬を与える．ただし 2.2 節でも述べたように，(a) 指示語あり質問 (default) を参照する場合は，その前に (b) 指示語あり質問 (特定話題) か (c) 指示語なし質問の発話があればそちらを参照する．

強化学習には Q 学習を用いた．Q 学習の Q 値は式 (1) で更新され，状態行動セットを表す Q テーブルに Q 値が記述される．

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha(r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1})) \quad (1)$$

ユーザモデルは選択されたシステム発話に対し，コーパス収集時と同じユーザ応答を返す．

システムの発話選択では，まず従来手法で対話行為を選択し，その対話行為内の発話で，本手法の選択戦略内において Q 値が高いものが確率的に選択した．

表 2: 主観評価による破綻数の比較

	○	△	×
従来システム	88	6	6
提案システム	94	3	3

4. 評価実験

4.1 実験条件

新たな状態行動報酬設計を用いた強化学習を実装し，テキスト対話を行って評価した．システム発話とユーザ発話の 1 対を 1 交換，10 交換を 1 セットの対話とし，10 セット 100 交換を評価の対象とした．

評価基準は東中らの研究 [2] を参考にした．○は当該システム発話のあと対話を問題無く継続できる．△は当該システム発話のあと対話をスムーズに継続することが困難である．×は当該システム発話のあと対話を継続することが困難であるという基準である．これら○△×のアノテーションの数と全交換数に対する割合で手法を評価した．

比較対象は，従来システムにおける発話集合を「スポーツ」「default」に絞ったものとした．

4.2 実験結果と考察

従来システムと提案システムの評価を表 2 に示す．△と×のアノテーションの数に関していずれの項目でも提案システムの方が少なかった．全アノテーション数に対する割合は，△は従来システムで 6%提案システムで 3%，×は従来システムで 6%提案システムで 3%，△と×の合計は従来システムで 12%提案システムで 6%であった．このことより，提案システムの方がより破綻の少なくなる発話を選択できているといえる．

実際に，従来システムでは図 1 のパターン 2 と同様の破綻が起きたが，提案システムでは見られなかった．これはシステム発話間の内容的整合性を考慮した設計が破綻削減に有効であったことを示している

今回はシステム発話間の内容的整合性を人手で設定したが，発話数が増加するとそれは困難になる．今後は自動でシステム発話間の内容的整合性を設定できるような手法が必要となる．

参考文献

[1] 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.

[2] 東中竜一郎, 船越孝太郎. Project Next NLP 対話タスクにおける雑談対話データの収集と対話破綻アノテーション. *SIG-SLUD*, Vol. B4, No. 02, pp. 45–50, 2014.

[3] 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.