強化学習を用いたキャラクタらしさを持つ雑談応答の生成

清水健吾　上垣貴嗣　菊池英明

早稲田大学人間科学研究科

{k.bi.sket@akane.,t-gappy@fuji.,kikuchi@}waseda.jp

概要

雑談対話システムがユーザに「また対話したい」と思わせるための方法として，特定のキャラクタらしさをシステム応答に付与する方法が考えられる．本研究は，強化学習を用いてニューラル対話生成モデルを fine-tuning することで，特定のキャラクタらしい雑談応答を生成できる対話生成モデルを提案する．Fine-tuning の際に，特定のキャラクタらしさを持つ単一の発話データのみを必要とし，対話形式のデータを必要としない点で既存研究と大きく異なる．提案手法に基づく対話システムを評価した結果，対話システムのユーザに特定のキャラクタらしさの印象を与え，「また対話したい」と思わせる効果を確認した．

1　はじめに

雑談対話システムの研究において「ユーザにまた対話したいと思わせられるか」という観点は大きな課題となっている．ユーザに「また対話したい」と思わせられる対話システムを開発することを目的としたコンペティションも開催されている [1, 2]．

また，近年，人と極めて自然な雑談対話を行うことが出来る大規模なニューラル対話生成モデルが登場した [3, 4]．これらの対話生成モデルに基づく対話システムが，よりユーザに「また対話したい」と思わせるための方法として，システム応答に特定のキャラクタらしさを付与する方法が考えられる．

対話生成モデルに特定のキャラクタらしさを付与する研究として，[5] が挙げられる．[5] は，特定のキャラクタらしさを持つ対話データを収集することで，特定のキャラクタらしい発話を生成する対話システムを提案した．しかし，特定のキャラクタらしさを持つ対話形式のデータを収集するには大きなコストを要する．

本研究の目的は，ある特定のキャラクタらしさを持つ応答を生成する手法を提案することである．特に，対象のキャラクタらしさを持つ対話形式のデータではなく，対象のキャラクタらしさを持つ単一の発話データのみを使用して，発話にキャラクタらしさを付与する手法を提案する．具体的には，強化学習を用いて対話生成モデルを fine-tuning することで，システム応答に特定のキャラクタらしさを付与する．また，本研究が提案する手法によって，対話システムのユーザに「このシステムとまた対話したい」と思わせることができるかについても併せて検証する．

2　提案手法

提案手法は以下の 3 つの段階で構成される．

1. 対話生成モデルを学習するために，Twitter 疑似対話データセットとキャラクタツイートデータセットの 2 種類のデータセットを構築する．
2. Twitter 疑似対話データセットを用いて大規模言語モデルを学習することで対話生成モデルを構築する．
3. 強化学習を用いて，構築した対話生成モデルを fine-tuning する．まず，キャラクターツイートデータセットを用いて，「対象のキャラクタらしさ」を算出する報酬モデルを構築する．次に，報酬モデルが算出する報酬を用いて対話生成モデルに強化学習を行う．

2.1　データセットの構築

2.1.1　Twitter 疑似対話データセット

Twitter のツイートとそのツイートに対するリプライの連鎖を疑似的な対話とみなし，Twitter 疑似対話データセットを構築した．収集したツイート・リプ

ライ連鎖に対して URL，ハッシュタグ等の削除や，文字の正規化などの前処理を行った．ツイート・リプライ連鎖の取得は 2020 年 3 月から 2021 年 12 月にかけて行い，最終的に約 292 万件のデータセットを構築した．

2.1.2 キャラクタツイートデータセット

特定のキャラクタらしさを持つツイートを収集し，キャラクタツイートデータセットを構築した．本研究では対象のキャラクタとして，アニメ「けものフレンズ」[1]に登場する「アライグマ」[2]を採用した．アライグマを取り上げた理由は以下の 2 点である．

- Twitter 上にアライグマを模したアカウントが多く存在し，データを集めやすかった．
- 発言内容にキャラクタらしさが表れやすいキャラクタであるため，提案手法の有効性を確認しやすかった．

ツイートを収集した基準を以下に示す．

- 有志がまとめた，Twitter 上に存在するアライグマに模したアカウントの一覧[3]に含まれていること．
- アカウント名に「さん」が含まれ，アカウントのプロフィールに「のだ」が含まれていること．

収集したツイートに対して，2.1.1 節で述べた Twitter 疑似対話データセットと同様の前処理を施した．さらに，感情分析タスク用に fine-tuning した BERT[6][4]を用いてツイートフィルタリングすることで，ポジティブな内容のツイートのみを残した．最終的に約 3.5 万件のデータセットを構築した．

2.2 対話生成モデルの構築

本研究では，対話を生成するモデルとして GPT-2[7] を使用した．事前学習済み GPT-2 として rinna 社が公開しているモデル[5]を使用した．GPT-2 を 2.1.1 節で構築した Twitter 疑似対話データセットを用いて学習することで，日本語版 DialoGPT[8] を構築した．ミニバッチサイズを 16，gradient accumulation steps を 2，学習率を 2.5×10^{-5}，optimizer として AdamW[9] を使用し，4 エポック学習を行った．

2.3 強化学習による fine-tuning

2.2 節で構築した DialoGPT が，対象のキャラクタらしさを持つ応答を生成できるように強化学習によって fine-tuning を行った．

2.3.1 報酬モデル

まず，強化学習における報酬である「対象のキャラクタらしさ」を算出するための報酬モデルを構築した．報酬モデルとして RoBERTa[10] を使用した．事前学習済み RoBERTa として rinna 社が公開しているモデル[6]を使用した．RoBERTa を 2.1.2 節で述べたキャラクターツイートデータセットを用いて fine-tuning した．「対象のキャラクタらしさを持つツイートか否か」の 2 値分類タスクを学習することで，学習したモデルからの出力を「対象のキャラクタらしさ」とみなすことができる．

ミニバッチサイズを 256，学習率を 3.0×10^{-5}，optimizer として AdamW[9] を使用し，1 エポック学習を行った．

2.3.2 Proximal Policy Optimization (PPO)

本研究では，強化学習アルゴリズムの一つである Proximal Policy Optimization (PPO)[11] を用いる．PPO を言語生成タスクに適用する際には，損失関数は以下のように定義される．

$$r(\theta) = \frac{\pi_\theta\left(y_t \mid s_t\right)}{\pi_{\text{old}}\left(y_t \mid s_t\right)}$$

$$loss = -\mathbb{E}_{\hat{Y}\sim\pi_{old}}\left[\min\left(\sum_{t=1}^{T} r(\theta)A\left(s_t, y\right), \right.\right.$$
$$\left.\left.\text{clip}\left(r(\theta), 1-\epsilon, 1+\epsilon\right)A\left(s_t, y\right)\right)\right] \quad (1)$$

式 (1) における $A\left(s_t, y\right)$ は強化学習の報酬を表す．π_{old} は数ステップ過去のモデルを表す．現在のモデルからの出力と過去モデルからの出力の比 $r(\theta)$ と，報酬の値 $A\left(s_t, y\right)$ の積が最大化されるように学習する．このようにすることで，報酬の値が正の時は現在の出力の確率を上げ，反対に報酬の値が負の時は現在の出力の確率を下げるように学習される．ただし，現在のモデルからの出力と過去モデルからの出力の比が大きすぎるものに対してはクリッピングを

1) https://kemono-friends.jp/
2) https://kemono-friends.jp/zoo/common-raccoon/
3) https://twitter.com/i/lists/1115496728734527488
4) https://huggingface.co/daigo/bert-base-japanese-sentiment
5) https://huggingface.co/rinna/japanese-gpt2-medium
6) https://huggingface.co/rinna/japanese-roberta-base

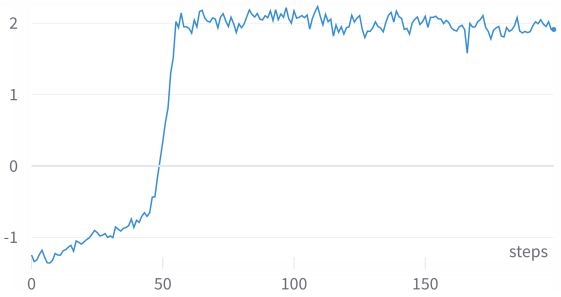

図 1　PPO 学習時の報酬の増減

する．これはモデルの急激な変化を防ぎ，学習を安定させる働きがある．

また，さらに学習を安定させるために報酬 $A(s_t, y)$ を次のように設定する．

$$A(x, y) = r(x, y) - \beta \log \frac{\pi(y \mid x)}{\rho(y \mid x)} \tag{2}$$

式 (2) における ρ は初期状態のモデルを表す．現在のモデルの状態が初期のモデルの状態から離れ過ぎないように，報酬の値 $r(x, y)$ に対してペナルティを与えることで，学習の安定化を図る．

2.3.3　PPO による fine-tuning

2.2 節で構築した DialoGPT に対して，PPO を用いて fine-tuning を行った．報酬として 2.3.1 節で構築した報酬モデルが算出する「対象のキャラクタらしさ」を利用する．「対象のキャラクタらしさ」を最大化するように学習することで，DialoGPT が生成する応答に対象のキャラクタらしさを付与する．

ミニバッチサイズを 8，学習率を 1.41×10^{-5}，optimizer として Adam[12] を使用し，200 ステップ学習した．その他のパラメータは [13] に準拠した．学習時の報酬の増減を図 1 に示す．学習終了後，報酬が安定した 60 ステップ目のモデルの重みを保存した．

また，学習したモデルが生成した実際の対話例を表 1 に示す．

表 1　PPO によって学習した DialoGPT との対話例

User	こんにちはー
System	こんにちはなのだ!今日もよろしくお願いしますのだ!
User	こちらこそよろしくねー！
System	よろしくお願いしますのだ!
User	今日寒いけど、アライグマさんは寒くない？
System	寒いのだー!暖かくして過ごすのだ!
User	アライさんはもふもふしていて暖かそう
System	もふもふなのだー!
User	私はアイスが好きで、今日もアイス食べちゃった。
System	アイス美味しいのだー!アライさんも食べるのだー!

表 2　評価項目

評価項目名	アンケート質問文
自然さ	この対話システムの対話は自然であった。
キャラクタらしさ	この対話システムは「アライさん」だと思う。
対話継続欲求	この対話システムと今後も対話をしたい。

3　評価実験

3.1　実験目的

以下の 2 点を目的に評価実験を行った．

1. 提案システムによる発話が，発話の受け手に対してある特定のキャラクタらしさを感じさせるかを検証する．
2. 提案システムのユーザが「このシステムと今後も対話をしたい」と思わせる効果があるかを検証する．

3.2　実験設定

実験では 2.2 節で構築した fine-tuning を行う前の DialoGPT（ベースラインシステム）と，2.3 節で fine-tuning を行った後の DialoGPT（提案システム）を比較した．

評価項目を表 2 に示す．被験者はクラウドソーシングを利用して 50 名募集した．キャラクタらしさを判断できるように，被験者はアニメ「けものフレンズ」の視聴経験がある者に限定した．被験者は LINE 上で 8 回対話システムと対話を行い，それぞれの対話に対して，表 2 の評価項目について 1（まったくそう思わない）から 5（とてもそう思う）の 5 段階で評価した．なお，被験者はベースラインシステムと提案システムのどちらと対話をしているか知らされなかった．

3.3　実験結果

評価実験の結果を図 2 に示す．キャラクタらしさの評価項目に対してはウィルコクソンの符号順位検定を行い，他の 2 つの評価項目に対しては対応のある t 検定を行った．検定の結果，キャラクタらしさの項目（$W = 15.5, p < .001$）と対話継続欲求の項目

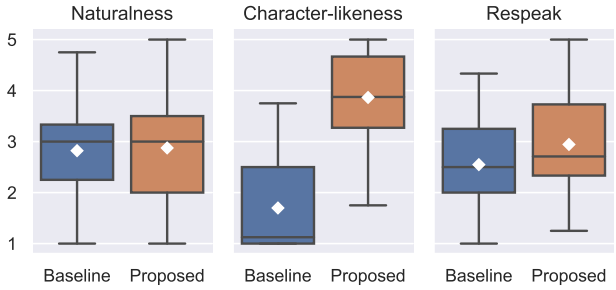

図 2　評価実験の結果（左から順に自然さ，キャラクタらしさ，対話継続欲求を表す．）

表 3　各評価項目の組み合わせごとの無相関検定

評価項目 1	評価項目 2	相関係数
自然さ	キャラクタらしさ	.12
自然さ	対話継続欲求	.70***
キャラクタらしさ	対話継続欲求	.41***

Note. *** $\hat{p} < .001$

($t(49) = 2.26, p = .028$）において，提案システムがベースラインシステムを有意に上回った（有意水準5%未満）．したがって，提案システムが生成する応答が，ユーザに特定のキャラクタの印象を与え，「このたいわシステムとまた対話をしたい。」と思わせる効果を確認した．また，自然さの項目において，ベースラインシステムと提案システムの評価値の同等性を確認するために Two One-Sided Test (TOST) をおこなった．許容する差の閾値は ±0.5 とした．検定の結果，2 つのシステムの評価値間において有意な同等性が確認された（$p = .0043$）．したがって提案手法は，対話の自然さを損なわずに特定のキャラクタらしさを付与できることを確認した．

また，各評価項目間の組合せにおいて無相関検定を行った．相関係数としてはスピアマンの相関係数を用い，p 値にはボンフェローニ補正を施した．各項目の組み合わせごとの検定の結果を表 3 に示す．

検定の結果，自然さと対話継続欲求，キャラクタらしさと対話継続欲求のそれぞれの評価値間で有意な正の相関が確認された．

4　おわりに

本研究の貢献は以下のとおりである．

1. 強化学習を用いて，ある特定のキャラクタらしさを持つ対話応答を生成する手法を提案した．モデルの学習には対象のキャラクタらしさを持つ単一の発話データのみを必要とし，対話形式のデータである必要はない．
2. 本研究が提案した手法で応答を生成することによって，対話システムのユーザに「このシステムと今後も対話したい」と思わせる効果があることを確認した．

本研究で提案した手法によって，より容易に特定のキャラクタらしい対話システムを構築できる可能性が示唆された．今後は，提案手法が他のキャラクタやより高精度で大規模な対話システムに対しても応用可能かを検証する予定である．

参考文献

[1] Raefer Gabriel, Yang Liu, Anna Gottardi, Mihail Eric, Anju Khatri, Anjali Chadha, Qinlang Chen, Behnam Hedayatnia, Pankaj Rajan, Ali Binici, et al. Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. **Alexa Prize Proceedings**, 2020.

[2] 東中竜一郎, 西川寛之, 宇佐美まゆみ, 船越孝太郎, 高橋哲朗, 稲葉通将, 赤間怜奈, 佐藤志貴, 堀内颯太, ドルサテヨルス, 小室允人, 西川寛之, 宇佐美まゆみ. 対話システムライブコンペティション 4. 第 93 回 言語・音声理解と対話処理研究会, pp. 92–100, 2021.

[3] 藤原吏生, 岸波洋介, 今野颯人, 佐藤志貴, 佐藤汰亮, 宮脇峻平, 加藤拓真, 鈴木潤, 乾健太郎. ILYS aoba bot: 大規模ニューラル応答生成モデルとルールベースを統合した雑談対話システム. 第 90 回 言語・音声理解と対話処理研究会, pp. 110–115, 2020.

[4] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. **arXiv preprint arXiv:2109.05217**, 2021.

[5] Ryo Ishii, Ryuichiro Higashinaka, Koh Mitsuda, Taichi Katayama, Masahiro Mizukami, Junji Tomita, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Yushi Aono. Methods for efficiently constructing text-dialogue-agent system using existing anime characters. **Journal of Information Processing**, Vol. 29, pp. 30–44, 2021.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, Vol. 1, pp. 4171–4186, 2019.

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. Technical report, Open AI, 2019.

[8] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. DIALOGPT: Large-scale generative pre-training for conversational response generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 270–278, 2020.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. **arXiv preprint arXiv:1711.05101**, 2017.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. **arXiv preprint arXiv:1707.06347**, 2017.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **International Conference on Learning Representations (ICLR)**, 2015.

[13] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. **arXiv preprint arXiv:1909.08593**, 2019.