

# Twitter データによる性別を付与した雑談対話システムの構築

## Gendered Open Dialogue System Using Twitter Data

大城 昂希<sup>†</sup>  
Koki Oshiro

姜 東植<sup>‡</sup>  
Dongshik Kang

### 1 はじめに

Transformer[1] の登場により、大規模な言語モデルを用いることで人間のような自然な応答が行える対話システムが登場している [2]。対話システムには特定タスクに関する対話に特化したタスク指向対話システムと、複数のジャンルに渡って対話を行う雑談対話システムが存在する。この Transformer を用いた雑談対話システムは大量の対話データを学習させることにより履歴を考慮した流暢な対話が行えるようになる。しかしこの雑談対話システムの問題の一つに、応答文のキャラクタを一貫させることの難しさが挙げられる。この問題により、対話の途中でキャラクタが変化してしまうことで対話の継続意欲が下がる可能性が考えられる。

また、我々は、様々な人に雑談対話システムを用いてもらうために、各個人の趣味嗜好に合った雑談対話システムを提供することを目的として研究を進めている。そのために、単一モデルで複数のキャラクタ性を持ち、適切なキャラクタでの応答ができるモデルの開発を行っている。

そこで、本研究では、性別属性を付与した雑談対話システムを構築することを目的とし、データの収集とモデルの学習を行った。今回、モデルにキャラクタ性が付与できているか確かめるために本システムにより複数回応答を生成し、その可能性を目視により検証した。

### 2 関連研究

特定のキャラクタ性を対話システムに学習させるためには、そのようなキャラクタ性を持つ対話形式のデータ

を用いて訓練を行う手法が考えられる。しかし、特定のキャラクタ性を持つ対話形式のデータの収集には膨大な時間がかかるとともに、対象のキャラクタ性によってはデータが十分に得られない事も考えられる。この問題に対して、清水ら [3] はキャラクタらしさをモデルに付与するために、対話形式のデータではなく、個々の独立した発話データを使用できることを示した。

### 3 データセットの構築

基本的には清水ら [3] が実施した手法に基づいてデータセットを構築した。

#### 3.1 データの収集と雑談対話データセットの構築

データの収集には Twitter API を利用し、2021 年 9 月から 2022 年 4 月まで実施した。Twitter では、あるツイートに対してリプライをすることで対話が行えるようになっている。今回このツイートとリプライの連鎖を雑談対話として収集した。雑談対話データセットの構築手法を以下に示す。

1. Twitter API を用いてツイートデータを収集
2. ツイートに対してデータ整形を適用 (3.1.1 項)
3. ツイート、リプライ連鎖から対話データを構築
4. 不適切な対話データをフィルタリング (3.1.2 項)

フィルタリング前の対話データ数は 1 億 4030 万件であり、フィルタリング後は 2143 万件となった。

#### 3.1.1 発話データ整形処理

発話データの整形には以下の処理を行った。また、ここでの顔文字判定には風間ら [4] の手法を用いた。

1. URL、ハッシュタグ、メンションの削除
2. html 特殊文字 (&gt; など) の置換
3. NFKC 正規化
4. 英小文字化
5. 絵文字および顔文字を句点に置換

<sup>†</sup> 琉球大学大学院理工学研究科工学専攻 Graduate School of Computer science & Intelligent systems, University of the Ryukyus

<sup>‡</sup> 琉球大学工学部工学科知能情報コース Computer science & Intelligent systems, University of the Ryukyus

6. 特定文字の 2 回以上の繰り返し短縮(っあいうえお!?) など)
7. 同一文字の 3 回以上の繰り返し短縮
8. 不必要な空白の除去
9. 句読点や記号の不適切な重複除去

### 3.1.2 対話データのフィルタリング処理

対話データのフィルタリングには以下の処理を行った。

1. 句点が 3 つ以上
2. 公序良俗に反する単語を含む [5]
3. 対話に参加するユーザ数が 3 名以上
4. ASCII 文字コード領域の文字が 50% 以上
5. 異なり文字の割合が 50% 以下
6. 形態素数が 5 未満か 30 より多い

### 3.2 男女判別データセットの構築

Twitter から得られるデータには性別情報が付与されていない。そこで、wikipedia の男性語<sup>\*1</sup>、女性語<sup>\*2</sup>の項を参考に以下の方法で男性らしさ、女性らしさをスコア化した。

1. 女性語および男性語を形態素解析器にかけ、語のリストを取得
2. ツイートに対して女性語を含む場合は正のスコアを、男性語であれば負のスコアを付与
3. ユーザごとにツイートのスコアの平均を取り、女性らしさ、男性らしさのスコアとして付与

使用した女性語と男性語のリストを表 1 に示す。付与するスコアは品詞によって異なり、代名詞は 2pt、終助詞と感動詞は 1pt とした。上記の手順で得られたユーザごとの性別スコアから 0pt より大きいものを女性らしいユーザ、0pt より小さいものを男性らしいユーザとする用にバイナリ化した。この時点で女性らしいツイートが 638 万件、男性らしいツイートが 63 万件得られた。データ数に 10 倍ほどの差があるため、女性らしいツイートのデータをスコアが高い方からサンプリングし男性らしいツイートのデータ数と合わせた。最終的に 126 万件のデータセットを構築した。

<sup>\*1</sup> <https://ja.wikipedia.org/wiki/男性語>

<sup>\*2</sup> <https://ja.wikipedia.org/wiki/女性語>

表 1 女性語と男性語のリスト

#### (a) 女性語

品詞	単語
代名詞	あたい、わたくし、あなた、わたし、あたくし、あたし、あんた
終助詞	わ、かしら、よ、ね、かしらん
感動詞	きゃーっ、あら、ん、おほほ、きゃー、きゃっ、いやー

#### (b) 男性語

品詞	単語
代名詞	俺、おれ、僕、ぼく、俺、わし、己等、おいら、自分、じぶん、お前、おまえ、てめえ、あいつ
終助詞	ぜ
感動詞	おい

## 4 日本語版 DialoGPT の学習

日本語版 DialoGPT[2] には rinna 社が公開している事前学習済み GPT-2 モデル<sup>\*3</sup>[6] を使用した。また、fine-tuning には 3.1 節で構築した Twitter 雑談対話データセットを用いた。

batch size を 512、gradient accumulation steps を 16、learning rate を  $1 \times 10^{-5}$ 、num optim steps を  $7 \times 10^5$  に設定し、それ以外は [2] に準拠した。今回は学習時間の制約から 28,000 ステップ時点のモデルを用いることにした。

表 2 に学習した DialoGPT の応答例を示す。Sys+ 数字となっているのは、モデルに応答を複数回生成させたものであり、どれも同じ入力で生成したものを掲示している。

## 5 強化学習による DialoGPT の fine-tuning

本研究では雑談対話システムにキャラクター性として性別を付与することを目的としている。そのため、報酬として女性らしさを選択し fine-tuning を行った。

### 5.1 報酬モデルの構築

報酬モデルには rinna 社が公開している RoBERTa[6] を使い、性別判別データセットを利用して classification task で fine-tuning を行った。ハイパーパラメータは HuggingFace[7] の TrainingArguments のデフォルト値

<sup>\*3</sup> <https://huggingface.co/rinna/japanese-gpt2-medium>

表 2 学習した DialoGPT の応答例

User	おはよう！
Sys0	おはようございます。今日は暖かいですね。
Sys1	おはよう!(今更)
Sys2	おはようございます.unk .unk
Sys3	おはようございます。今日も 1 日前進しましよ う!
Sys4	おはよう!メンションありがと。
User	今日は天気良くて暖かいな
Sys0	うん!晴れてて良かった!
Sys1	え、そうなん?
Sys2	うんうん!いいね!
Sys3	そうよね!今日は曇ってないからいいよね
Sys4	うん!お出かけしてますー?

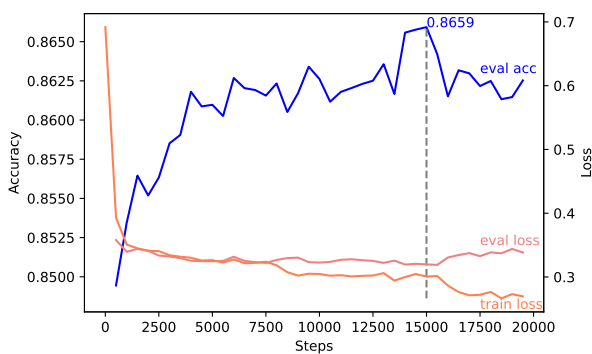


図 1 報酬モデルの学習過程

を用いた。

報酬モデルの学習経過を図 1 に示す。図 1 を見ると 15,000 step 時点で Accuracy が最大の 0.8659 になり、Loss が十分に下がっていることが確認できる。そこで、本実験での報酬モデルには 15,000step 時点の RoBERTa モデルを用いることにした。

## 5.2 Proximal Policy Optimization (PPO) での fine-tuning

Proximal Policy Optimization (PPO)[8] は強化学習アルゴリズムの 1 つであり、長期的な報酬を最大化するように方策を直接最適化するアルゴリズムである。特徴として、クリッピングを用いることで過去の方策から大きく変更されることを防ぎ、学習を安定化させている。

batch size を 64、forward batch size を 64、learning rate を  $1 \times 10^{-6}$ 、steps を 600 とし、その他のパラメー

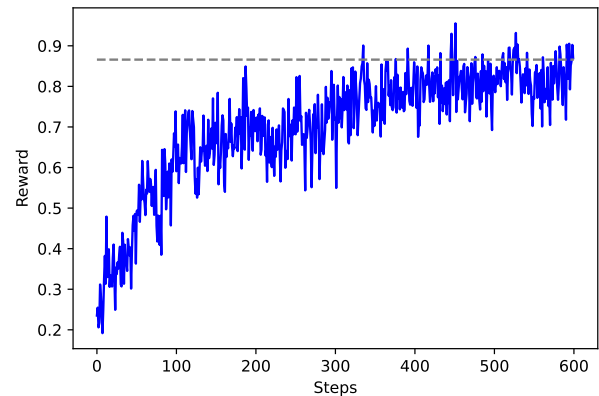


図 2 PPO での学習過程

表 3 PPO で fine-tuning した後の応答例

User	おはよう！
Sys0	username さん!おはようございます。
Sys1	おはようございます。今日も、いい日ですね。
Sys2	おはよう御座います!今日も 1 日頑張っていま しょうね。
Sys3	おはよう、そしておやすみ～
Sys4	username おはよう。今日も可愛いね。
User	今日は天気良くて暖かいな
Sys0	今日は今日はお祭りだね。
Sys1	今日はめちゃくちゃ眠いよ。
Sys2	いいね!体調気をつけてね。
Sys3	そうね。
Sys4	これから寒くなるね。

タは trl<sup>\*4</sup>のデフォルト値を使用した。

PPO での fine-tuning の学習過程を図 2 に示した。また、PPO での学習後の応答例を表 3 に示した。username となっている箇所は人名が入っていたためマスクしたものである。

## 6 結果と考察

今回、雑談対話データセットで学習させた DialoGPT モデル (ベースモデル) と性別判別データセットで fine-tuning した DialoGPT モデル (新規モデル) を用い、同様の入力に対してどのような応答が得られるか確認した。その結果、ベースモデルでは時々男性らしい応答が生成されていたが、新規モデルではほとんどの場合女性らしい応答となっていることが確認できた。また、ペー

\*4 <https://github.com/lvwerra/trl>

モデルと新規モデルで共に入力に対して妥当でないと思われる応答をすることが多々あった。この問題に対しては、両モデルの学習が足りていない可能性と、我々が考慮している文脈がモデルに対して十分に与えられていない可能性が考えられる。これらについては今後より詳しく調査・実験を行う必要がある。

## 7 おわりに

本研究では性別属性を付与した雑談対話システムを構築するために、清水らの研究 [3] を参考に Twitter から収集したデータを整形し、それを用いて DialoGPT に女性らしい応答を学習させた。我々の本実験においては女性らしい応答が多く生成されることが確認できたが、その有効性を客観的データにより示すことはできていない。今後はモデルや学習手法の有効性をアンケートなどによる調査で検証し、より多様なキャラクター性を備えたモデルを構築する。

## 参考文献

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *arXiv:1706.03762*, Jun 2017.
- [2] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, JJ (Jingjing) Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *arXiv:1911.00536*, November 2019.
- [3] 清水 健吾, 上垣 貴嗣, and 菊池 英明. 強化学習を用いてキャラクターらしさを付与した雑談応答の生成. **人工知能学会研究会資料言語・音声理解と対話処理研究会**, 94:06, 2022.
- [4] 風間 一洋, 水木 栄, and 榊 剛史. Twitter における顔文字を用いた感情分析の検討. **人工知能学会全国大会論文集**, JSAI2016:3H3OS17a4-3H3OS17a4, 2016.
- [5] 稲葉 通将. おーぶん 2 ちゃんねる対話コーパスを用いた用例ベース対話システム. In **第 87 回言語・音声理解と対話処理研究会 (第 10 回対話システムシンポジウム)**, 人工知能学会研究会資料 *SIG-SLUD-B902-33*, pages 129–132, 2019.
- [6] 趙 天雨 and 沢田 慶. 日本語自然言語処理における事前学習モデルの公開. **人工知能学会研究会資料言語・音声理解と対話処理研究会**, 93:169–170, 2021.
- [7] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.
- [8] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv:1707.06347*, 2017.