対話モデルにおけるキャラクター特性の実現法の探索

山田　和志[1] 篠崎　隆志[1,2,3]
[1]近畿大学 理工学部 情報学科
[2]近畿大学 情報学部 情報学科
[3]近畿大学 情報学研究所
tshino@kindai.ac.jp

概要

対話型 AI の爆発的な普及は社会にまさに変革をもたらしているが，さらなる浸透のためには個々の AI が独立したキャラクター特性を持つことが必要と考えられる．本研究では LLM への入力を適切に構成することによって，単一のモデルで複数のキャラクター特性を実現する手法を探索する．日本語特有の語尾によるキャラクター特性の実現について注目し，対話 LLM の入力の構成要素である Persona と History を操作することで，ファインチューニングなしでのキャラクター付けを実現，人間による感性評価によって，Persona と History についての適切な構成を明らかにした．本手法によって，ゲームにおける多数の Non-Player Character（NPC）や，ご当地キャラとの対話を実現できると考えられる．

1 はじめに

1.1 本研究の背景

近年，ChatGPT[1]のような自然言語処理モデルが注目されている．ChatGPT は，OpenAI によって開発された大規模言語モデルで，同社の開発した Generative Pre-trained Transformer（GPT）[2]と呼ばれる言語モデルによる文章生成によって，人間のような対話の生成を実現している．

GPT をはじめとする大規模言語モデル（Large Language Model，以下，LLM）では，基本構造として Transformer[3]というネットワークが使われている．Transformer は深層ニューラルネットワークを用いた言語モデルで，従来のモデルである Long Short Term Memory（LSTM）等で用いられていた再帰構造を避け，代わりに単語間の依存関係を表現する注意機構と呼ばれるメカニズムを導入したアーキテクチャである．Transformer は自然言語処理にとどまらず，画像処理[4]やマルチモーダル信号[5]など，多岐にわたるタスクで優れた性能を発揮している．

対話型 AI によって生成される文章のキャラクター特性は学習に用いたデータに依存しているが，学習のための文章量が豊富な英語の処理においては，入力を調整することによってキャラクター特性をある程度コントロールできることが知られている．日本語の場合においても，雑談対話において LLM を利用することは有効[6]であることが示されており，学習データの前処理を適切に行うことで，その応答の生成にキャラクターの口調や個性が反映されることが示されている[7].

しかし，日本語特有の語尾によるキャラクター特性の表現に関しては，その特殊性から十分な研究がなされていない．本研究では LLM において，会話の語尾を変更させることでキャラクター特性を実現することを試みる．

1.2 本研究の目的

例えば多数の Non-Player Character（NPC）が登場するゲームにおいて，NPC との対話を実現するためには，NPC ごとに特徴付けされたキャラクター特性が必要とされる．しかし，LLM にキャラクターごとに適した会話やセリフを学習させるために必要なファインチューニングは，膨大な手間とリソースを要する．また，キャラクターごとに独立したモデルを用いるとすると，極めて多くのハードウェアリソースを必要とするという問題もある．

そこで，本研究では，GPT のような汎用の大規模言語モデルを使用しつつも，メモリの使用量を抑えたキャラクター特性を実現する方法を探索する．具体的には，ファインチューニングをせずに，単一のモデルで効率良くキャラクター付けを可能にする方法を提案する．対話モデルへの入力を制御することによって，生成される文章の語尾を変更させ，キャラクター特性の実現，その実現度合いを感性評価によって検証する．

2 研究内容

2.1 対話 AI，Persona，History の構成

本研究では, LLM への入力を適切に構成することによって，対話 AI において単一のモデルで複数のキャラクター特性を実現することを目的とする．

「対話 AI」とは，自然言語処理技術と人工知能を組み合わせたシステムである．人間のような対話を理解し，応答することができる人工知能エージェントを指す．最近の研究では，大規模な言語モデルの開発や転移学習の導入により，対話 AI の性能が大きく向上している[8,9]．これによって対話 AI は，コンピュータとユーザーの双方向のコミュニケーションを可能にしており, さらなる品質向上のためにも，個々のユーザーやシナリオ等の前提情報に適応できる柔軟性が求められている．

対話 LLM における一般的な入力の構成では，対話自体の入力以外に，Persona と History という補助的な内容が前文として毎回入力されている[8]．

「Persona」とは，対話 AI においてユーザーまたはエージェントが持つ特性を持つようにするために設計された，入力の冒頭に付加される説明文を表わす．これによって，会話の一貫性や姿勢，ユーザーエクスペリエンスの向上が実現されている．Persona は，特定のトピックに関する知識や言い回し，スタイルなどを包括することがある．例えば会話モデリングのデータセットである「PERSONA-CHAT」では，各ユーザーが異なる Persona を持ち，それに基づいた対話が行われる[10]．実際にモデルに Persona が組み込まれることで，対話の個別化と魅力を高められている．このことからも Persona は対話エージェントがユーザーに対してより個別化されたサービスを提供する手段として重要である．

「History」とは，対話における事前のコンテキストを指す．これは，過去のユーザーの発言やエージェントの応答など，対話の流れ全体を含んだものである．GPT を含む Transformer[3]では self-attention という仕組みを通じて長い文脈を捉えることができることから，History を効果的に取り扱うことによって，対話 AI の一貫性と理解を向上させることが可能となる[8]．特に，長い対話や複雑な質問応答タスクでは, 精確に管理された History を用いることによって，過去の対話コンテキストを正確に把握し，より洗練された対話が可能となる．

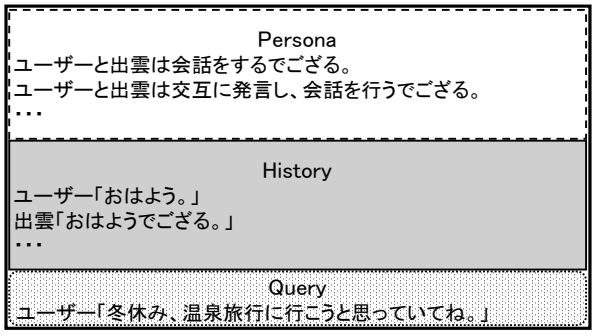

図 1　対話 LLM における一般的な入力の構成

2.2 使用した言語生成モデル

本研究では, rinna 社による GPT2 の日本語 medium モデル（以降，「rinna-2」とする）[11]，LINE ヤフー社による日本語 36 億パラメータモデル（以降，「line-3.6b」とする）[12], およびその対話モデル（以降，「line-3.6b-is」とする）[13]，Preferred Networks 社による日本語 130 億パラメータモデル（以降，「PlaMo-13B」とする）[14]の 4 つの LLM モデルを使用する．近年，極めて多数の LLM がリリースされているが，予備実験での結果や使用する言語モデルの多様性のため，上記のモデルを選定した．

2.3 実験の概要

近年 LLM のサイズは拡大の一途をたどっていることから，ファインチューニングに必要とされるハードウェアリソースの要求仕様も高まっており，個々のキャラクター付けのためにファインチューニングを利用することはさらに困難となりつつある．そこで本研究ではファインチューニングしない単一の LLM で効率良くキャラクター付けを可能にする方法を提案し，これを検証する．より具体的には，日本語特有の会話の語尾を変更することによるキャラクター特性を実現し，評価者による認知の基づいた点数付けによってその有効性を検証する．これによって，将来的には単一のモデルで複数のキャラクター特性を実現し得るようなシステムの実現を目指す．

図 1 に対話 LLM における一般的な入力の構成[8]を示す．なお，ここで「出雲」とあるのは，内部的に設定されたチャットモードのコードネームである．図 1 ではユーザー側の対話自体の入力は最下段の Query にのみ反映されるが，それ以外に Persona と History という補助的な内容が毎回入力されている．本研究では，この Persona と History を操作すること

で，ファインチューニングせずに，単一のモデルでキャラクター付けを実現する方法を検証する．例えば，語尾を「ござる」とすることによって，忍者のようなキャラクター付けをすることを考える．このとき，Persona と History のそれぞれについて，忍者のような「ござる」語尾を付加することによって目的とする特性を実現することを試みる．

なお，開始時に設定された Persona と History の補助的な内容の他に，ユーザーと LLM による対話についても，History を更新しながら入力されるようにしている．実験では，Persona と History の量（トークン数）や使用する学習済み言語モデルを変更することで，それぞれの変更が会話のキャラクター付けにどれだけ寄与するのかを明らかにし，より効率のよいシステムを模索する．

2.3.1 Persona の変更

本研究で用いる Persona は，キャラごとに内容の異なる Persona を用意するのではなく，内容は同じで，それぞれの Persona の語尾を変化させたものを用い，Persona の語尾の変化のみで，会話にどれだけ影響があるかを検証する．なお，LLM への Persona の性格の作成には，16Personalities[15]を参考にした．作成された Persona はいわば対話 AI の特性の定義文であるが，LLM で実現するキャラクター特性に合わせて，これら定義文の語尾を変更する．さらに Persona のトークン数についても変更させ，その影響も検証する．

2.3.2 History の変更

本研究で用いる History は，Persona と同様に，キャラごとに内容の異なる History を用意せず，同一の内容の History について目的とする語尾へと変化させたものである．曖昧な言葉，質問と応答のやりとりなどを含んだ History を用意し，LLM で実現するキャラクター特性に合わせて，応答文の語尾の変更を行う．また，Persona と同様に，History の量の変化による影響も検証する．

2.3.3 会話のやりとり

本研究では，Persona と History の変更によって，キャラクター付けにどのような影響があるかを確かめるため，実験でのユーザー側の発言は出来る限り統一する必要がある．そこで，今回の実験で行った会話は，ユーザーが LLM に，冬休みの旅行として城崎温泉[16]に行く予定を話している，という前提でユーザー側の発言を行うことにより，条件の統一を図った．ただし，LLM の返答によってユーザー側の返答も変える必要があるため必ずしも内容は一定とはならない．そのような場合であっても可能な限り条件を統一するため，会話の流れが自然な範囲で，ユーザー側の返答の具体的な順番は適宜変更しつつ，試行間で同様の情報量の返答が行われるように努めた．

2.3.4 対話の評価

生成された対話は 5 人の評価者によって，会話としての妥当性と，キャラクター特性を，それぞれ 10 点満点，合計 20 点満点で感性評価を行った．1 セットの会話は 10 文の応答を含むことから，ある応答が妥当である場合，もしくは適切なキャラクター特性である場合は 1 点，そうでない場合は 0 点とし，10 文の合計値をその対話における Score とした．

3 実験内容

本章では，どのモデルが語尾の変更に適しているか，また，適していると考えられるモデルで Persona や History の量を変更した場合などについて，それぞれの実験の内容と結果を記述する．検証には 10 会話（入力文と生成文を合わせて 20 文）の対話を用いた．

3.1 モデル探索

表 1 に，rinna-2，line-3.6b，line-3.6b-is，PLaMo-13B の各モデルにおけるキャラクター付けで語尾が「ござる」となるようにした場合の結果を示す．表中で Persona と History はそれぞれのトークン数，GPU は推論時のメモリ使用量を GB で，Score は 5 人の評価者の平均値を示す．

表 1　各モデルの実験結果

#	model	Persona	History	GPU	Score
1	rinna-2	250	113	3	4.9
2	line-3.6b	1005	361	13	15.1
3	line-3.6b-is	1005	361	20	12.2
4	**PLaMo-13B**	**1301**	**462**	**67**	15.5

表 1 から，GPU のメモリ使用量が多いもののPLaMo-13B の Score が高いことがわかる．また line-3.6b についてもメモリ使用量が少ないにも関わらず良好なスコアであることが確認された．#1 の

Persona および History の量が他と著しく異なるのは，入力可能なトークン数の制約による．上記の結果から，以降の実験では PLaMo-13B のみを使用し，Persona と History の量を変更させた場合の性能を検証する．

3.2 Persona と History の有無

表 2 に，PLaMo-13B において入力する Persona と History の量を変更した場合の結果を示す．

表 2　Persona と History の有無

#	Persona	History	GPU	Score
4	**1301**	**462**	**67**	15.5
5	**0**	**462**	**55**	15.3
6	1301	0	62	8.4
7	**258**	**462**	**54**	**16.4**

History のみ（#5），または Persona のみ（#6）では，両方入力した場合（#4）と比べて，メモリの使用量が少なくなっている．しかし，#6 の Score は，その他の実験と比べて，著しく悪いことが分かる．これは，語尾の使い方を，会話の流れを通して予想することができないため，Score が低くなったと考えられる．裏を返すと，語尾の使い方を示すことができる History があれば，ある程度キャラクター付けができると考えられる．そこで History の量を一定としたままで Persona の量を削減した場合（#7）を検証したところ，最も良い Score を記録した．

3.3 History だけ変更

表 2 の#5 から，Persona が無く，History だけという場合でも，ある程度キャラクター付けができることが分かった．そこで今度は，Persona の量を 0 に固定し，History の量のみを変更させた場合での検証を行った．表 3 に PLaMo-13B において Persona を入力せず History の量を変更した場合の結果を示す．

表から History が 113 トークンと著しく短い場合（#10）では，キャラクター特性は反映されにくいことが分かった．一方で History の量を著しく増やしたとしても，それに比例して語尾が反映されるわけではなく頭打ちが見られ（#11, #12），ある程度の History の量があれば十分な性能が見込めることが確認された（#5, #8）．

表 3　History の量を変更した実験結果

#	Persona	History	GPU	Score
5	**0**	**462**	**55**	15.3
8	0	346	54	15.3
9	0	232	53	12.0
10	0	113	53	11.2
11	0	919	59	15.9
12	0	682	58	14.2

4 考察

実験結果から，Persona を 258 トークン程度，history を 462 トークン程度，それぞれ，適したキャラクター特性のものを用意することで，より会話の語尾が変更されるようなキャラクター特性の実現が可能である．また，会話の噛み合わせもよく，かつ，メモリ効率のよいシステムとなる．今回の実験では LLM の生成時パラメータである temperature, top_p, top_k, repetition_penalty などは一定にして行っていたが，これらの変更によって，例えばより temperature が高く，生成文の変動が大きい条件下では，より長い Persona や History が必要である可能性が考えられ，これらについてのさらなる検証が必要である．

さらに，一口に LLM といっても，「質問した内容に対する答え」を生成するもの，「入力した文章の続きを生成するもの」など，モデルの目的は様々なので，「会話」を目的としたモデルなら，よりキャラクター特性が反映されやすい可能性がある．例えば表 1 のように，line-3.6b モデル[12]では少ない GPU メモリ使用量にも関わらず高い Score を実現している．一方で対話へのファインチューニングが行われた場合[13]の方が今回の手法への適合度が低下しており，特に複数のキャラクターを演じさせる場合のモデル選定，あるいはベースとなるモデルの構築法について，さらなる検証が必要とされる．

本研究の結果から，会話の噛み合わせの良い LLM を一つ使用し，キャラクターごとの Persona や History を用意することで，例えばオンラインゲームなどで，複数のキャラクターを個別のキャラクター特性で演技させることができると考えられる．また，ゲーム以外についても，例えばご当地キャラの会話文生成においても提案手法の応用が可能であることが見込まれ，精度の高い演出が可能となることが期待され，幅広い応用が見込まれる．

参考文献

[1] OpenAI, ChatGPT, 2022. (参照 2023-12-20). https://openai.com/blog/chatgpt.

[2] OpenAI, Improving language understanding with unsupervised learning, 2018. (参照 2023-12-20). https://openai.com/research/language-unsupervised.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. **Advances in neural information processing systems 30**, 2017.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. **International Conference on Learning Representations (ICLR)**, 2020.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. **arXiv preprint arXiv:2103.00020**, 2021.

[6] 山崎天, 坂田亘, 川本稔己, 小林滉河, グエントゥン, 上村卓史, 中町礼文, 李聖哲, 佐藤敏紀. ペルソナ一貫性の考慮と知識ベースを統合したHyperCLOVAを用いた雑談対話システム, 人工知能学会研究資料 言語・音声理解と対話処理研究会 93回, p.113-118, 人工知能学会, 2021.

[7] 秋山一馬, 稲葉通将. 小説から生成した擬似対話に基づくキャラクター対話システムの構築, 人工知能学会全国大会論文集 36回, 一般社団法人 人工知能学会, 2022.

[8] Thomas Wolf, Victor Sanh, Julien Chaumond, Clement Delangue. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents, **arXiv preprint arXiv:1901.08149**, 2019.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. Language Models are Few-Shot Learners. **Advances in neural information processing systems 33**, 2020

[10] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, Jason Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too?, **arXiv preprint arXiv:1801.07243**, 2018.

[11] rinna Co., Ltd. 日本語に特化したGPT-2の大規模言語モデルを開発しオープンソース化, 2021. (参照 2023-10-20). https://rinna.co.jp/news/2021/04/20210407.html.

[12] Shun Kiyono, Sho Takase, Toshinori Sato(overlast). 36億パラメータの日本語言語モデルを公開しました, LINE Engineering Blog, 2023. (参照 2023-10-20). https://engineering.linecorp.com/ja/blog/3.6-billion-parameter-japanese-language-model.

[13] Koga Kobayashi, Tomoya Mizumoto. Instruction Tuningにより対話性能を向上させた 3.6B 日本語言語モデルを公開します, LINE Engineering Blog, 2023. (参照 2023-10-20). https://engineering.linecorp.com/ja/blog/3.6b-japanese-language-model-with-improved-dialog-performance-by-instruction-tuning.

[14] Hiroaki Mikami. PLaMo-13Bを公開しました, Preferred Networks Blog, 2023. (参照 2023-10-20). https://tech.preferred.jp/ja/blog/llm-plamo/.

[15] NERIS Analytics Limited. 性格タイプ, (参照 2023-11-01). https://www.16personalities.com/ja/%E6%80%A7%E6%A0%BC%E3%82%BF%E3%82%A4%E3%83%97.

[16] 城崎温泉観光協会. 城崎温泉観光協会, (参照 2023-11-24). https://kinosaki-spa.gr.jp/.