

# GPT2 モデルを用いた感情を考慮する日本語対話生成

竹原 和輝 全 昌勤

神戸大学大学院 システム情報学研究科

[231x041x@stu.kobe-u.ac.jp](mailto:231x041x@stu.kobe-u.ac.jp), [quanchqin@gold.kobe-u.ac.jp](mailto:quanchqin@gold.kobe-u.ac.jp)

## 概要

対話生成 AI の感情表現力や感情理解力を向上させることで、AI がより人間らしいコミュニケーションが可能になり、ユーザーの体験が改善されることがわかっている。本研究では単語ごとの感情情報を登録した単語感情辞書を作成し、日本語学習済み GPT2 モデルにおいて感情情報を付与する層を追加することで感情表現力や感情理解力の向上を目指した。

感情付与を行わなかったモデルと感情付与を行ったモデルを、ChatGPT を利用して比較し、感情付与によって感情表現能力が向上することを示した。

## 1 はじめに

現在、最も広く認知されている対話生成 AI である ChatGPT は、自然で人間らしい対話生成を実現したが、依然として人間の応答と完全に同等とは評価されていない。

対話生成 AI のさらなる発展において重要な課題の一つは、AI に感情を理解させ、適切に応答させる能力の向上である。人間の感情や心情を把握し、適切に反応する能力は、対話の自然さを高め、ユーザーの満足度を向上させられることが報告されている [1]。

従来の感情情報を考慮した対話生成は、感情をあるなしの 2 値で考慮していたり、文単位の感情を抽出した研究が多いため、感情の強弱に対応したり、複数の感情を考慮することができなかった。本研究では、この課題に対応するため、各単語の感情情報を確率として抽出する単語感情辞書生成アルゴリズムを提案し、それを使用した対話生成モデルの評価を行う。

対話生成 AI の評価においては Perplexity や BERTScore が広く用いられているが、これらは生成文章の自然さや感情表現などを測る指標ではない。

こういった主観的なものの評価においてはアンケート調査が用いられるが、同じ条件で実施するのが

難しいことや、回答の真偽が定かではないなどの問題がある。

そこで対話生成の研究 [2] で ChatGPT を利用した評価が行われるようになってきている。本研究では ChatGPT を使って 2 つの感情付与モデルと感情無しモデルを比較し、感情付与を行った方が感情表現力や対話の自然さが向上することを示した。

## 2 関連研究

感情情報を考慮した対話生成に関する研究は、深層学習を用いた対話生成 AI の実現以降、活発に進められてきた。

Zhou ら [3] が提案した Emotional-Chatting-Machine は、ニューラルネットワークを基盤とする対話生成モデルに感情情報を組み込む初期の試みである。これは生成したい感情を手動で指定するシステムであり、感情情報はあるかないかの 2 値で考慮されていた。

近年になって感情強度を測定する研究や、相手の対話に対して適切な感情の文章を生成する感情制御の研究も行われている。Gupta と Yang [4] の研究では、Twitter 上の文章を「怒り」「恐怖」「喜び」「悲しみ」の 4 つの感情に分類し、各感情の強度を自動的に測定するシステムを構築した。また、Zhang ら [5] の研究では、6 種類の感情カテゴリに対応する応答を生成し、その中から最適な応答を選択する手法が提案された。

日本国内でも感情を考慮した対話生成の研究は活発に行われている。片山ら [6] の研究では、Transformer と LSTM を組み合わせることで、ユーザーからの感情指定なしに感情的な対話を可能とするシステムを開発した。また本間ら [7] の研究では、Transformer を基盤とした感情対話生成手法が提案されている。この手法はシンプルな構造ながらも手動での感情指定を必要とせず、感情を考慮した自然な対話生成を実現している。本研究では、本間らが提案した Emotional-Embedding の概念を参考に、より高度な感情表現を可能とする対話生成モデルの構

案を試みる。

### 3 提案手法

#### 3.1 Emotional-Embedding-GPT2

本論文で提案する Emotional-Embedding-GPT2 は、従来の GPT2 モデル[8]に Emotional-Embedding 層を追加したモデルである。

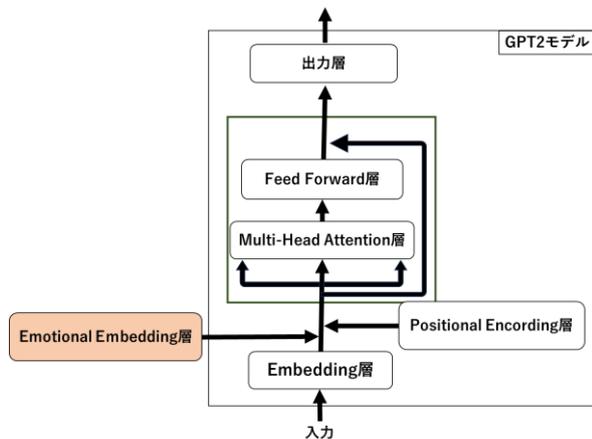


図1 Emotional-Embedding-GPT2 の概要

Emotional-Embedding 層では Embedding 層でベクトルに変換された単語の末尾を、単語の感情情報に置き換えるという処理を行っている。

単語のベクトルが

$$[a_1, a_2, a_3, \dots, a_{1022}, a_{1023}, a_{1024}]$$

であり、単語の感情ベクトルが

$$[e_1, e_2, e_3]$$

となる時、Emotional-Embedding 層通過後の単語ベクトルは

$$[a_1, a_2, a_3, \dots, a_{1021}, e_1, e_2, e_3]$$

となる。

#### 3.2 単語感情辞書の構築

本論文で提案する単語感情辞書の構築アルゴリズムは、日本語感情表現辞書(JIWC)[9]の手法を参考にしており、感情ラベルが付いた文章のデータセットがあることを前提としている。構築手順を以下に示す。

**手順 1** 単語ごとに感情の数と等しい次元を持つベクトルを用意する

**手順 2** 文章の感情ラベルに従って、単語の感情値を加算する

**手順 3** データセット内の各感情ラベルの文章の数

を数え、単語の感情ベクトルを割って正規化する  
**手順 4** 総和が 1 になるように変換する。

以下に処理の例を示す。

**手順 1**

本研究では喜怒哀の 3 感情のデータセットを使用したため、データセットに登場する全ての単語に 3 次元のベクトルを用意する。

**手順 2**

今日/は/ピクニック/に/行き/まし/た [喜び]  
今日 [喜び : 0→1, 怒り : 0, 哀しみ : 0]  
は [喜び : 0→1, 怒り : 0, 哀しみ : 0]  
ピクニック [喜び : 0→1, 怒り : 0, 哀しみ : 0]  
に [喜び : 0→1, 怒り : 0, 哀しみ : 0]  
行き [喜び : 0→1, 怒り : 0, 哀しみ : 0]  
まし [喜び : 0→1, 怒り : 0, 哀しみ : 0]  
た [喜び : 0→1, 怒り : 0, 哀しみ : 0]  
これをデータセット内の全文章に対して行う。

**手順 3**

データセット内にある喜びの文章の数が 50、怒りの文章の数が 20、哀しみの文章の数が 30 の場合、全ての単語の感情ベクトルに対して以下の計算を行い正規化する

[喜びの数値/50, 怒りの数値/20, 哀しみの数値/30]

**手順 4**

単語の感情ベクトルの総和で割り、確率に変換する  
今日 [0.1 0.2 0.2] → 今日 [0.2, 0.4, 0.4]

表 1 単語感情ベクトルの例

番号	単語	喜び	怒り	哀しみ
1	楽しい	0.949	0.028	0.022
2	の	0.291	0.363	0.345
3	惚気	0.513	0.436	0.050
4	運動	0.439	0.191	0.470
5	エグい	0.428	0.114	0.459
6	東京	0.234	0.510	0.256
7	汗	0.161	0.424	0.414
8	夫	0.235	0.444	0.321
9	ケーキ	0.640	0.159	0.201
10	焼きたて	0.765	0.000	0.235
11	猫	0.229	0.444	0.327
12	愛猫	0.620	0.000	0.380
13	政治	0.070	0.611	0.320

### 3.3 対話生成モデル

本研究では rinna 社が公開している `japanese-gpt2-medium` を対話生成用にファインチューニングと Emotional-Embedding 層の追加を行い、Emotional-Embedding-model, Emotional-Response-model, Non-Emotion-model の 3 種類の対話生成モデルを作成した。以下に各モデルの概要を示す。

#### • Emotional-Embedding-model

本研究で使用したデータセットは、会話文+返答文+返答文の感情ラベルの組で構成されている。

Emotional-Embedding-model(以下 EEmodel)は返答文+返答文の感情ラベルの組で構築した単語感情辞書を使用したモデルであり、先行研究で同様の構造で対話生成をした際に文章の自然さが向上するという報告が上がっているが、感情表現能力については言及されていない。

#### • Emotional-Response-model

Emotional-Response-model(以下 ERmodel)は会話文+返答文の感情ラベルの組で構築した単語感情辞書を使用したモデルであり、返答の際に適切な感情を表現できるようになると考えられる。

#### • Non-Emotion-model

Non-Emotion-model(以下 model)は、Emotional-Embedding 層を追加せず、対話生成用のファインチューニングのみを行ったモデルである

実験ではこの 3 つのモデルの比較を行った。

## 4 モデル評価実験

EEmodel, ERmodel, NEmodel の 3 つのモデルが生成する対話の自然さと感情表現能力を、ChatGPT を利用して比較、評価を行った

### 4.1 対話の自然さ評価

ChatGPT に[日常]、[質問]、[告知]、[批判]、[感想]の 5 種類の Twitter 風投稿をそれぞれ 10 個ずつ生成させ、これらの投稿に対して 3 つのモデルを用いて応答文を生成し、生成された文章が自然であるかどうかを ChatGPT に 5 段階で評価してもらった。

投稿文と生成結果の例を以下に示す

表 2 自然さ評価の対話例

ChatGPT	Non-Emotion	Emotional-Embedding	Emotional-Response
1 朝のコーヒーがないと1日が始まらないなあ	コンビニにもあるんですよ	コンビニにもありますけどね	コーヒー飲んだら仕事の能率が落ちるからやめろ
2 仕事終わりに友達とご飯行くの楽しみ	ええ!めっちゃ美味しそう!	めっちゃめっちゃ楽しみにしてるね!	飯でも行きませんか?
3 最近、天気が良くて気持ちいいね	おはようございます!気持ちいいですね	おはようございます。気持ちいいですね	おはようございます。良い天気ですね。
4 やっと週末!ゆっくりできる時間が最高	えへへ。ゆっくりできるのもあと少しですね。	えみちゃんおはよ!ゆっくり休んでね!	え?!そんなに暇なの?
5 あのドラマの最終回、どう思った?涙止まらなかった...	ほんとに泣けてくるよね。	最終回は泣かなかったけど泣いた(/)	最終回まであと2回。最終回まで泣いていいの?って思ってしまった。

### 4.2 感情表現能力評価

ChatGPT に喜んだ返答が予測される投稿、怒った返答が予測される投稿、悲しんだ返答が予測される投稿をそれぞれ 10 個ずつ生成させた。これらの投稿を 3 つのモデルに入力し、返答を生成させた後、3 つのモデルが生成した返答の感情を ChatGPT に判断してもらい、想定していた感情と一致した精度を調べた。

ChatGPT が生成した投稿文と生成結果の例を以下に示す。

表 3 感情表現能力評価の対話例

ChatGPT	Non-Emotion	Emotional-Embedding	Emotional-Response
1 大好きなアーティストのライブチケット当選した!夢みたい!	当選おめでとうございます!素敵な日になりますように!	めっちゃ嬉しいですよ!	当選おめでとうございます!素敵なライブになりますように!
2 今日、新しい友達ができた!人生が少し豊かになった気がする。	友達ってすごく素敵ですよ!私も友達作りたいです!	ええ友達!ええ!ええ!	え嬉しい友達できたの!?!すごいね!おめでとう!
3 猫を飼い始めました!毎日が癒しの連続。	猫ちゃんおはようございます癒しパワーで癒されますね	猫ちゃんに癒やされてますか?	猫ちゃんも癒しですね
4 あの本、遂に完読!次はどの本にしようかな。	あの本は絶対に買う価値あります!絶対に!	待ってるよ!	え?!それはうれしい!ありがとございます!次は「死」かな。
5 今日のランニング、自己ベスト更新!やったー!	ナイスラン!記録おめでとうございます!	よかったですね!おめでとうございます!	やったー!記録更新ですねおめでとうございます!

### 4.3 実験結果と考察

表 4 は対話モデルが生成した文章の自然さを ChatGPT に評価してもらった結果である。各項目は 10 質問×5 点の 50 点満点で、%形式で表示している。

表 4 自然さの評価結果

自然さ	日常	質問	告知	批判	感想	合計
EEmodel	72%	72%	66%	64%	78%	70%
ERmodel	50%	60%	68%	68%	70%	63%
NEmodel	58%	70%	64%	48%	64%	61%

表 5 は対話モデルの感情表現能力を ChatGPT に評価してもらった結果である。ChatGPT が推測した生成文章の感情と、想定される返答文の感情が一致していた割合である。

表 5 感情表現能力の評価結果

感情	喜	怒	哀	合計
EEmodel	30%	30%	60%	40%
ERmodel	90%	30%	70%	63%
NEmodel	40%	30%	40%	37%

単語の感情付与を行った 2 つのモデルと、感情付与を行わなかったモデルの結果を比較すると、生成された文章の自然さおよび感情表現能力のいずれにおいても、感情付与を行ったモデルの方が向上していることがわかる。特に、生成文章の自然さにおいては通常の感情付与を行ったモデルが、感情表現能力においては返答感情の付与を行ったモデルが顕著に向上している。

通常の感情付与モデルは、単語の感情情報を学習しており、感情なしのモデルと比較して情報量が多いため、文章の生成能力が向上していると考えられる。一方、返答感情付与を行ったモデルは、元の文章に対してどのような感情で返答すべきかを学習できており、感情表現能力において他のモデルよりも優れた結果を示している。

しかし、返答感情付与を行ったモデルにおいて、喜びの項目では他のモデルより高い評価を得た一方で、怒りに関しては他のモデルと変わらない結果となっている。このことから、喜びと怒り、哀しみの感情においては差異が明確であり、異なる感情として出力することが比較的容易であるのに対し、哀しみと怒りの 2 つの感情間では差が小さく、異なる感情として出力することが難しかったと考えられる。

## 5 おわりに

本研究では、感情ラベルが付与された文章データ

セットを用いて単語感情辞書を作成し、その辞書から得られた感情ベクトルを GPT-2 モデルに付与することによる対話生成を行った。さらに、通常の返答感情と、感情に基づいた 2 種類の方法で単語感情辞書を作成し、感情無しで学習を行ったモデルとそれら 2 つのモデルを ChatGPT を使用して比較した。その結果、感情を付与したモデルが生成した文章において、自然さや感情表現能力が向上することが示された。

現状の課題として、考慮している感情の範囲が限定的であることが挙げられる。プルチックの感情の輪においては、人間の基本感情は怒り、喜び、受容、驚き、恐れ、悲しみ、嫌悪、期待の 8 つとされているが、現行モデルではこれらの感情すべてを考慮することができておらず、その結果として、人的な感情の学習に不十分な部分があると考えられる。

また、使用したデータセットに関しては、Twitter の文章をそのまま利用したことにより、データの質が低い可能性がある点や、感情分類器が古いため、単語感情辞書のラベル精度が不十分である点が挙げられる。対話生成の精度は、文章の感情を判別する感情分類器の性能に密接に関連しており、感情分類器の技術向上に伴い、本手法の精度も向上すると予想される。

今後の評価方法に関する課題として、本論文で行った ChatGPT による評価が人間による評価にどの程度近いのかを検証する必要がある。具体的には、ChatGPT による評価と人間によるアンケート結果を比較することにより、ChatGPT 評価の信頼性を担保できれば、対話評価の確立した指標としての役割を果たす可能性があると考えられる。

## 謝辞

本研究を進めるにあたり、お忙しい中貴重なお時間を割いてくださった羅教授に厚く御礼申し上げます

## 参考文献

- [1] H. Prendinger and M. Ishizuka: “The empathic companion: A character-based interface that addresses users’ affective states,” *Applied Artificial Intelligence*, 2005.
- [2] Yunfan Shao, Linyang Li, Junqi Dai, Xipeng Qiu, “Character-LLM: A Trainable Agent for Role-Playing”, *EMNLP2023*, (2023)
- [3] Zhou, H., Huang, M., Zhang, T., Zhu, X. and Liu, B., “Emotional chatting machine: Emotional Conversation Generation with Internal and External Memory”, *AAAI-18*, (2018)
- [4] Gupta, R.K. & Yang, Y., “CrystalFeel at SemEval-2018 Task 1: Understanding and Detecting Emotion Intensity using Affective Lexicons”, *Proceedings of the 12th International Workshop on Semantic Evaluation*, (2018)
- [5] Zhang, R, Wang, Z, Mai, D, Huang, X, Jiang, J, Zhao, D, Feng, Y, & Hong, Y., “Building emotional conversation systems using multi-task Seq2Seq learning”, *NLPCC 2017*, (2018)
- [6] 片山 晋, 米澤 拓郎, 河口 信夫, “テキスト対話システムにおける感情制御を伴う応答生成手法”, *知能と情報* 2022年34巻3号, (2022)
- [7] 本間 心, 武田 敦志, “日本語を対象とする感情を考慮した対話応答生成に関する研究”, *言語処理学会 第26回年次大会発表論文集*, (2020)
- [8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, “Language Models are Unsupervised Multitask Learners”, <https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf>, (2019)
- [9] 柴田大作, 若宮翔子, 伊藤薫, 荒牧英治, “JIWC: クラウドソーシングによる日本語感情表現辞書の構築”, *言語処理学会 第23回年次大会発表論文集*, (2017)

# A 付録

## 文章の自然さ評価のプロンプトと対話例

### • ChatGPT 指示文

今から twitter で行われた会話を伝えます。Tweet 文に対しての応答の文章が自然な文章かどうかを 5 段階で判定してください。

ここで言う自然さとは『ツイッター上で行われた会話として適切かどうか』であるため、『意味が伝わる文章か』『返答の内容が適切か』などを判断材料とし、『日本語としての正確さ』『一般的な会話かどうか』などは考慮しないでください。

また、一連の会話がどのような背景で行われたかはわからないため、状況や関係性を減点対象としないでください。

### • 対話例

🗨️ 了解しました。それでは、AさんのツイートとBさんの返答をお願いします。

質問文 ついにやった！あのプロジェクトが無事に成功したよ！  
1 えっおめでとうございます！  
2 えっまじがすごい！  
3 おおめでとうございます！ついに達成ですな疲れ様です！

🗨️ (質問文) ついにやった！あのプロジェクトが無事に成功したよ！

- 1: 喜び
- 2: 喜び
- 3: 喜び

🗨️ 🗨️ 🗨️ 🗨️

## 文章の感情評価のプロンプトと対話例

### • ChatGPT 指示文

#### • 文章生成

Twitter のツイートっぽく、返信相手が『喜び』『怒り』『哀しみ』の感情を発露しそうな文章をそれぞれ 10 個考えてください。ただし、質問文は禁止とします。

※補足 Twitter の会話に疑問文に対して答えるシチュエーションが少なく、疑問文に対して上手く答えられないのでこのようにしている。

#### • 感情分析

今から A さんのツイートと、ソレに対する B さんの返答を 3 パターン伝えます。それぞれの返答が『喜び』『怒り』『哀しみ』『中立 or その他』のどの感情を持って発言されたか判断してください。

解答は以下のフォーマットで表示してください

(質問文)

1: (感情)

2: (感情)

3: (感情)

### • 対話例

🗨️ 了解しました。それでは、AさんのツイートとBさんの返答をお願いします。

質問文 ついにやった！あのプロジェクトが無事に成功したよ！  
1 えっおめでとうございます！  
2 えっまじがすごい！  
3 おおめでとうございます！ついに達成ですな疲れ様です！

🗨️ (質問文) ついにやった！あのプロジェクトが無事に成功したよ！

- 1: 喜び
- 2: 喜び
- 3: 喜び

🗨️ 🗨️ 🗨️ 🗨️