文を入力とした俳句の自動生成

大山野乃子[1] 杉本徹[2]
[1]芝浦工業大学大学院 理工学研究科 [2]芝浦工業大学 工学部
{ma23044,sugimoto}@shibaura-it.ac.jp

概要

本研究は，GPT-2 モデルを 2 回ファインチューニングすることによって，入力文の内容に合うような俳句を生成する手法を提案する．自分の身近な体験を述べた文から俳句を生成し，それを鑑賞してもらうことでユーザの日々の発見を増やし，より生活を楽しませることを目的としている．俳句と鑑賞文およびそれらから抽出した単語を用いてファインチューニングを行い，モデルを構築した．生成された俳句について，人手の評価を行った結果，俳句中の単語と鑑賞文を用いてファインチューニングすることが入力文の内容と合う俳句を生成するために有効であることが分かった．

1 はじめに

俳句とは一般に 575 であり 17 拍の短詩である．小説や SNS の文章よりも短く，より短時間で鑑賞することができ，また深読みし楽しむこともできるコンテンツである．俳句を作るメリットについて，俳人は，生きづらい日々の中でも俳句は日々の中の小さな喜びの見つけ方を教えてくれる[1]と述べている．このように，俳句を作り鑑賞することで普段の生活の機微をより多角的に発見し，生活を楽しくすることができる．しかし普段俳句を作る機会のない人は，そのような経験をすることがない．

近年コンピュータで俳句を自動生成する試みがなされており[2]，特に SeqGAN，LSTM，GPT-2[3]といった深層学習モデルを利用した研究が多く行われている．小西ら[4]は，SeqGAN を用い，俳人が詠んだ俳句を生成器に，素人が詠んだ俳句を識別器に入力し，後者に似た俳句を生成している．また太田ら[5]は，LSTM を用いて，拍数や季語を表す素性ベクトルを使い，韻律や季語の制約のある俳句を生成する手法を提案している．平田ら[6]は，GPT-2 を用い，俳句だけでなく青空文庫データでモデルを学習し，俳句を生成している．また我々[7]は，GPT-2 を用いて，印象ラベルと紐づいた俳句を学習し，特定の印象を与える俳句を生成する手法を提案した．これらの研究の多くはランダムな内容の俳句を生成する．

一方，実用上は内容を指定して俳句を生成したい場合もある．内容を指定した俳句生成の研究例として，米田ら[8]は，LSTM を用いて生成した俳句の中から入力画像に合う句を選別することでモチーフ画像に適合する俳句を生成する方法を提案している．しかし，文を入力とする俳句生成の研究は存在しない．

本研究では，入力文の内容に合うような俳句を生成する手法を提案する．大規模な俳句および俳句の鑑賞文データを利用して，GPT-2 モデルをファインチューニングすることによって，文と単語および俳句を関連付けるモデルを構築する．

本研究の成果を用いて，普段俳句に親しくない人に対して，自分の身近な経験を書いた文から俳句を作成し，それを鑑賞する機会を提供することによって生活の素敵を発見し楽しんでもらうことが期待される．

俳句の定義について，近年は無季自由律に寛容な意見が主流であるため，本研究では拍数や季語や切れ字の有無は考慮しないこととする．

2 提案手法

事前学習済み言語生成モデルに，2 回のファインチューニングを行い，文から俳句を生成できる状態にする．鑑賞文データの量の少なさを補うために，1 回目に「俳句」と「俳句中の単語」データ，2 回目に「俳句」と「鑑賞文」と「単語（俳句中もしくは鑑賞文中）」データで学習した．

2.1 1 回目のファインチューニング

まず，GPT-2 モデルを「俳句」と「俳句中の単語」の紐づきデータで学習する．

「俳句」データとして，伊藤園お〜いお茶新俳句大賞[9]や現代俳句データベース[10]などのウェブサイトや，句集などから収集した 101,695 句を利用する．「俳句中の単語」については，それぞれの俳句を MeCab で形態素解析し，名詞・動詞・形容詞・副詞を抽出した．その結果 1 つの俳句から抽出した単語数は最大 11 単語となった. 表 1 に抽出した例を示す（俳句の出典は[9]）．

表 1　「俳句」と「俳句中の単語」の例

俳句	単語
居眠りをするのは心が春だから	居眠り，心，春
七五三飛ぶ子踊る子ふざける子	七，五，三，飛ぶ，子，踊る，子，ふざける，子

学習時には，『<s>[SEP]居眠り<s>心<s>春<s><s><s><s><s><s><s><s><s>[SEP]居眠りをするのは心が春だから</s>』のように，「俳句中の単語」，「俳句」の順に整形しモデルに入力した．

モデルの学習には Google Colaboratory を用い，GPT-2 モデルは rinna 社の japanese-gpt2-medium[11]モデルを利用した. これは Japanese CC-100 や日本語 Wikipedia を学習したモデルで，パラメータ数は約 3 億である．バッチサイズは 1，エポック数は 2 で学習させた．その結果，パープレキシティは元のモデルの約 18 から 3.67 と下がり，学習ができているといえる．

2.2 2 回目のファインチューニング

次に，1 回目のファインチューニングをしたモデルを，「俳句」と「鑑賞文」と「単語」のデータでファインチューニングする．

ここでの「鑑賞文」とは，俳句を鑑賞し情景や連想したことを述べた文章である．HAIKU 日本大賞[12]や伊藤園お〜いお茶新俳句大賞[9]などのウェブサイトや句集から収集した．量を増やすため，他者の評も自作についての評も区別せず，また俳句の鑑賞文が複数の文からなる文章である場合は文章に含まれる 1 文ごとに俳句と対応させた．結果，2,120 句 9,077 文となり，9,077 組の俳句と鑑賞文の対データとなった．表 2 に俳句と鑑賞文の例を示す（俳句の出典は[9]）．

表 2　俳句と鑑賞文の例

俳句	鑑賞文
田舎では星が降ります頭上注意	キラキラかがやく星の光は、今の都会の街中では見られない
	しかし、その都市を離れて、たとえば自分の故郷に里帰りしてみたとき、まるで降るような星空に、おどろきと久しぶりに胸のときめきを感じてしまう
	思わず頭上注意と叫びたい気持ちになる。その眺めを人に伝えたいからだ

ファインチューニングに用いる「単語」の抽出方法について，以下の 2 種類のモデルを構築した．

2.2.1 俳句中の単語を利用するモデル

1 つ目のモデルは，「俳句」と「鑑賞文」と「俳句中の単語」を用いて 2 回目のファインチューニングを行うモデルである．

「俳句中の単語」は，1 回目のファインチューニングと同様に俳句を MeCab で形態素解析し，名詞・動詞・形容詞・副詞を抽出したものである．

学習時には，『<s>[SEP]キラキラかがやく星の光は、今の都会の街中では見られない[SEP]田舎<s>星<s>降る<s>頭上<s>注意<s><s><s><s><s><s>[SEP]田舎では星が降ります頭上注意</s>』のように，「鑑賞文」と「俳句中の単語」と「俳句」の順番でデータを整形し入力した．

「俳句」と「俳句中の単語」でファインチューニングしたモデルに，9,077 組の「俳句」と「鑑賞文」と「俳句中の単語」データで 2 回目のファインチューニングを行った．バッチサイズは 1，エポック数は 6 で学習させた結果，パープレキシティは 1.77 と下がっている．

2.2.2 鑑賞文中の単語を利用するモデル

2 つ目のモデルは，「俳句」と「鑑賞文」と「鑑賞文中の単語」を用いて 2 回目のファインチューニングを行うモデルである．

「鑑賞文中の単語」は鑑賞文から単語を抽出したものである．MeCab を利用し鑑賞文中の名詞・動詞・形容詞・副詞を抽出したところ，「鑑賞文」から抽出した単語数は最大 23 単語となった. 1 回目のファインチューニングの 11 単語以内に合わせるため，以下のように利用する単語の選別を行った．

1. Google 1-gram の出現回数が 2 億以上の単語を省く
2. 11 単語より多かった場合，Google 1-gram の出現頻度が高い順に 11 単語を採用する

表 3 に抽出した例を示す（俳句の出典は[9]）．

表 3　「俳句」と「鑑賞文」と「鑑賞文中の単語」の例

俳句	鑑賞文（1 文）	単語
ロボットのむねの歯車春を待つ	外は雪でしょうか	外，雪
名月やきれいな音のでる薬缶	物思いにふけりながら名月を見ていたら、湯を沸かしていた薬缶が、突然笛のようなきれいな音をたてました	物思い, 耽る, 名月, 居る，湯，沸かす，薬缶，突然，笛，音，立てる

学習時には，『<s>[SEP]物思いにふけりながら名月を見ていたら、湯を沸かしていた薬缶が、突然笛のようなきれいな音をたてました[SEP]物思い<s>耽る<s>名月<s>居る<s>湯<s>沸かす<s>薬缶<s>突然<s>笛<s>音<s>立てる[SEP]名月やきれいな音のでる薬缶</s>』のように，「鑑賞文」と「鑑賞文中の単語」と「俳句」の順番で入力した．

俳句中の単語を利用するモデルと同じく，「俳句」と「俳句中の単語」でファインチューニングしたモデルに，9,077 組の「俳句」と「鑑賞文」と「鑑賞文中の単語」データで 2 回目のファインチューニングを行った．バッチサイズは 1，エポック数は 6 で学習させた結果，パープレキシティは 1.82 と下がっている．

2.3 俳句の生成

2 回のファインチューニングを行ったモデルを用いた俳句の生成は，次のように行う．まずユーザが文を入力すると，システムは MeCab を用いて入力文中の名詞・動詞・形容詞・副詞を抽出する．そしてファインチューニング済みのモデルに『<s>[SEP]コンビニの横の桜がいつのまにか散っていた[SEP]コンビニ<s>横<s>桜<s>何時<s>間<s>散る<s>居る<s><s><s><s><s>[SEP]』のように，「入力文」と「入力文中の単語」を入力し，俳句を生成する．

生成された俳句に対して次のような選別を行う．まず類似の句を避けるため学習元データの俳句との最小編集距離が 5 以下の句を省く．また主観であるが，短すぎると意味をなさず長すぎると文らしくなってしまうため，読み仮名が 7〜27 拍でない句も省く．

3　評価実験

3.1 評価方法

提案手法により生成した俳句の適切さを評価するために，アンケートを行った．

評価に用いる入力文は，実験協力者 5 人に対して「自分が俳句に表したい内容（日常の出来事や自分の気持ち等）を述べた 10〜30 文字ぐらいの長さの文を書いてください」と依頼して集めた 13 文の中から，文の長さのバランスを考慮して主観で 8 文を選んだ．入力文 1 文につき，俳句中の単語を利用するモデルと鑑賞文中の単語を利用するモデルを用いて，それぞれ 4 句ずつランダムサンプリングにより生成した合計 64 句を実験協力者に鑑賞し評価してもらった．表 4 にアンケートの質問と選択肢を示す．

表 4　アンケートで尋ねた質問と選択肢

質問	選択肢
俳句として自然だと思うか	自然だと思う どちらかと言えば自然だと思う どちらかと言えば自然だと思わない 自然だと思わない
入力文の内容と合っているか	合っている どちらかと言えば合っている どちらかと言えば合っていない 合っていない
その俳句を気に入ったか	気に入った どちらかと言えば気に入った どちらかと言えば気に入らない 気に入らない

アンケートには Google フォームを用い，俳句を提示する順番は入力文ごとシャッフルさせた．

また，この実験は俳句に親しくない人を対象とするため，俳句結社に所属していない人を対象とし，合計 12 人から回答を得た．

3.2 評価結果

それぞれの質問における肯定的な回答（「自然だと思う」，「どちらかと言えば自然だと思う」など）の割合を図 1 に示す．

表 5　評価の高かった俳句の例

質問項目	入力文	生成された俳句	モデル
俳句として自然だと思うか	動いたり動かなかったりする食洗器に日々振り回されている	動かない食洗器に我慢の二秒	鑑賞文中の単語利用
入力文の内容と合っているか	今日のハンバーグはおいしく作れたのでうれしかった	今日のハンバーグおいしかったね「うれしいな」	俳句中の単語利用
	8 月のお盆休みに、サークルの友達と沖縄の海を満喫した	夏休み友達 4 人で沖縄旅行	俳句中の単語利用
その俳句を気に入ったか	庭の柿の木に十数年ぶりに実がなって嬉しかった	庭に柿の木十数年の実の嬉しさよ	俳句中の単語利用
	動いたり動かなかったりする食洗器に日々振り回されている	どうにかして皿洗いをサボれないものかと悩む猫	鑑賞文中の単語利用

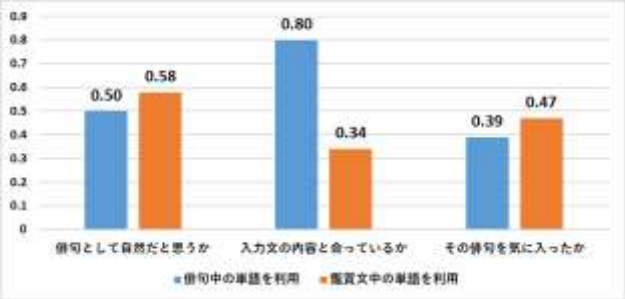

図 1　肯定的な回答の割合

3.3 考察

入力文の内容と合っているかについて，俳句中の単語を利用するモデルを用いて生成した俳句では 8 割が肯定的な回答であり，鑑賞文中の単語を利用するモデルと大きな差が見られた．そこで，入力文の単語が出力された俳句にどれだけ含まれているか，類似度の平均を調べた. 学習時と同様に MeCab を利用し名詞・動詞・形容詞・副詞を抽出し，gensim ライブラリの Word2Vec を用いて，入力文の単語ごとに類似度が最高の単語との類似度を求めて平均をとった. 結果, 俳句中の単語を利用するモデルは 0.78, 鑑賞文中の単語を利用するモデルは 0.50 と差があった．これは，学習元データである鑑賞文と俳句は内容が異なるものも多く，生成された俳句も入力文と内容が異なるようになったと考えられる．

その俳句を気に入ったかについて，鑑賞文中の単語を利用するモデルで生成した俳句の方が少し高くなった．生成された俳句を目視で確認すると，鑑賞文中の単語を利用するモデルで生成した俳句の方が，入力文から離れた言い回しを使っており，意外性があるといえる．

また俳句として自然であるかについては，評価の低い句を目視で確認すると，内容が破綻しているものや，俳句よりも文らしいリズム感になっているものが多かった．今回どちらのモデルも肯定的な回答が 5 割ほどと低かったため，内容が破綻せず，俳句のようなリズム感のある文字列を生成するモデルを構築することが今後の課題である．

表 5 に，評価が高かった俳句の例を示す．

4　おわりに

本研究では，俳句を鑑賞することでユーザの日々の発見を増やし，より生活を楽しませることを目的とし，文を入力としその文に合った内容の俳句を自動生成する手法を提案した．GPT-2 モデルを 1 回目は学習データに大規模な「俳句」と「俳句中の単語」，2 回目は少量の「俳句」と「鑑賞文」と「単語」を用いてファインチューニングすることによって，文と単語および俳句を関連付けるモデルを構築した．

「単語」として「俳句中の単語」と「鑑賞文中の単語」を用いたモデルを構築し，2 つのモデルから生成された俳句について人手の評価を行った結果，俳句中の単語を利用するモデルは，鑑賞文中の単語を利用するモデルより入力文の内容と合っている俳句を生成することができた．俳句を気に入ったかについては, 鑑賞文中の単語を利用するモデルの方が，俳句中の単語を利用するモデルより少し気に入りやすい結果となった．

今後の展望について，入力文と内容が合う俳句を生成することができたので，それに加えて多くの人が気に入ることを両立するような俳句を生成するモデルを構築することを目指す．

参考文献

[1] 夏井いつき．“夏井いつきの日々是「肯」日”．清流出版．（2020）

[2] 横山想一郎，山下倫央，川村秀憲．“深層学習を用いた俳句の生成と選句”．人工知能，Vol.34 No.4，pp467-474（2019）

[3] Alec Radford，Jeffrey Wu，Rewon Child，David Luan，Dario Amodei，Ilya Sutskever．“Language Models are Unsupervised Multitask Learners”．OpenAI（2019）

[4] 小西文昂，廣田敦士，松尾星吾，家原瞭，小原宗一郎，加賀ゆうた，鶴田穣士，脇上幸洋，金尻良介，深田智，田中一晶，岡夏樹．“SeqGANを用いた一般人に好まれやすい俳句の生成”．情報処理学会関西支部　支部大会（2017）

[5] 太田瑶子，進藤裕之，松本裕治．“深層学習を用いた俳句の自動生成”．情報処理学会研究報告，Vol.2018-NL-235 No.1，pp1-8（2018）

[6] 平田航大，横山想一郎，山下倫央，川村秀憲．“Transformer による言語モデルを用いた俳句生成とその評価”．情報処理学会研究報告，Vol.2021-IFAT-143 No.2，pp1-6（2021）

[7] 大山野乃子，杉本徹．“特定の印象を与える俳句の自動生成”．情報処理学会第 85 回全国大会，pp2-765-766（2023）

[8] 米田航紀，横山想一郎，山下倫央，川村秀憲．“深層学習を用いたモチーフ画像に基づく俳句生成”．人工知能学会第二種研究会資料，社会における AI 研究会，SIG-SAI-031-03（2018）

[9] 伊藤園．“伊藤園お～いお茶新俳句大賞”．https://itoen-shinhaiku.jp/（2023/1/11 アクセス）

[10] 現代俳句協会．“現代俳句データベース”．https://www.haiku-data.jp/index.php（2023/1/11 アクセス）

[11] Hugging Face．https://huggingface.co/rinna/japanese-gpt2-medium（2023/1/11 アクセス）

[12] 特定非営利活動法人 HAIKU 日本，“HAIKU 日本大賞”．https://haikunippon.net/（2023/1/11 アクセス）