

大規模言語モデルと Chain of Thought Prompting を用いた俳句の評価

富澤 峻己[†] 横山 想一郎[‡] 山下 倫央[§]
北海道大学大学院情報科学院[†] 北海道大学大学院情報科学研究所[‡] 北海道大学大学院情報科学研究所[§]
川村 秀憲[¶]
北海道大学大学院情報科学研究所[¶]

1 はじめに

人工知能技術の発展により、創作活動や芸術作品における新しい可能性が拓かれている。伝統的な芸術の領域である絵画や音楽、文学において、人工知能が生み出す作品は多くの注目を集め、それらの作品が人間の感性や創造性とのように関連するか、またどのような新しい価値を提供するのかについての議論が活発化している。創作活動を人工知能に行わせるという取り組みは、知能とは何か、人とは何かという問いに対する重要な糸口を含んでいると考えられている [1]。

このような状況の中で、俳句という日本固有の短詩形式にも人工知能技術が応用され始めている。俳句は、5-7-5 の 17 音で構成されること、「季語」と呼ばれる単語を一つだけ含むことなどの制約の中で季節や自然、人々の生活を詠むものであり、その短さゆえに奥深い意味や美しさが凝縮されている。鑑賞者が俳句から受ける印象は自身の経験や知識、感性に大きく依存するため、複数の鑑賞者が良いと考える俳句や良いとする理由は必ずしも同一ではない。人工知能を用いて俳句を評価する過程を紐解く上で、俳人が何を持ってその俳句が良いと感じたかというデータが重要な役割を果たす。

本研究では、俳句文化の一つである句会に注目する。句会では俳人たちが投句した俳句に対し、良いと感じた俳句に匿名投票を行い、各々の批評を述べ合う。各句会は趣味趣向の異なる俳人で構成されるため、各句会で良いとされる俳句の傾向は異なる。本研究では、単一の句会に注目して俳句が評価される要因を特定しやすくする。このフィードバックを評価機構作成に活かすことで人工知能を用いた俳句の評価という課題への有効なアプローチとなる。

以上を踏まえ、本稿ではある単一の句会における俳句を評価する際に特に重要だと考えられる 6 つの評価項目を決定し、各項目に対して大規模言語モデルと Chain of Thought Prompting を用いた俳句の評価の有効性を検証した。俳人の協力してもらい作成した俳句を評価する推論過程を few-shot 事例として与え各項目を 4 段階で評価する実験を行い、実験結果を俳人による評価と比較した。

本論文は以下の通り構成される。第 2 章では決定した評価項目と Chain of Thought Prompting を用いた提案手法について述べる。第 3 章では使用した俳句データとモデルについて述

べ実験結果をまとめた後、モデルが生成した推論過程について考察し、第 4 章を本論文のまとめとする。

2 Chain of Thought Prompting を用いた俳句の評価

2.1 評価項目

これまで研究の一環として参加した句会での自身の経験と俳人に対するヒアリングから、俳句を評価する際の基準として以下の 6 つの項目を決定した。

- 意味が通るか: 俳句中の表現として理解できるか
- 共感性があるか: 自身の過去の経験や考えから共感できるか
- 作者の心情が読み取れるか: 作者の主観的な感情が見られるか
- 詩的表現があるか: 通常の文章では用いられないが俳句中の表現として効果的に用いられているものはあるか
- 季語を活かしているか: 季語を上手く活用しているか
- 語句の対比があるか: 俳句中の複数の語句間に関係が見られるか

各項目に対する評価の度合いは、俳人に対するヒアリングから以下の 4 段階で設定した。それぞれ、1 (非常に当てはまる)、2 (当てはまる)、3 (どちらとも言えない)、4 (当てはまらない) の 4 段階である。

2.2 評価機構

モデルが俳人に近い思考過程で評価に至ることを目的に、Chain of Thought Prompting による評価機構を提案する。

2.1 節の各項目の評価に至るまでの推論過程を含む質問-回答事例を俳人監修の下で 3 句分作成し、few-shot 事例としてモデルに与えた。表 1 に few-shot 事例を含むプロンプトの一例を示す。

俳句に季語が含まれる場合は、あらかじめ作成した季語の辞書を用いて抽出し、季語の読みがなや対応する季節などの 4 つの情報を入力に追加するようプロンプトを構成した [2]。批評文は特徴的な表現を含む 200 字程度の俳句の説明文である。

3 実験

大規模言語モデルと Chain of Thought Prompting を用いて 2.1 節で示した 6 つの項目に対して評価を行った。実験で使用した俳句と同様の俳句をアンケート調査により俳人に評価してもらい、実験結果とアンケート結果を比較した。

3.1 実験設定

文章生成タスクで一般的に用いられるモデルとして OpenAI API の GPT-3.5 系の text-davinci-003 と GPT-4 系の 2024 年 1 月現在の最新モデルである gpt4-1106-preview を用いた [3]。

本実験の対象の俳句として、愛媛県松山市で月に 1 度開催されている「ふくし句会」に投句された俳句から 30 句を選択し用いた。30 句の内訳は、有季定型句 25 句、無季 3 句、季重

Evaluation of Haiku using Large-scale Language Models and Chain of Thought Prompting

[†] Syunki Tomizawa, Graduate School of Information Science and Technology, Hokkaido University

[‡] Soichiro Yokoyama, Faculty of Information Science and Technology, Hokkaido University

[§] Tomohisa Yamashita, Faculty of Information Science and Technology, Hokkaido University

[¶] Hidenori Kawamura, Faculty of Information Science and Technology, Hokkaido University

表1 few-shot 事例を含むプロンプトの例

評価項目：意味が通るか 説明：俳句中の表現として理解できるか
 以下は質問・回答の例です。
 Q: 「洗ひたる硯の海の涼しさよ」という俳句は意味が通りますか？
 以下の4段階で評価してください。
 1: 非常に通る 2: 通る 3: どちらとも言えない 4: 通らない
 A: まず、季語リストを用いてこの俳句の季語を取得します。
 俳句：洗ひたる硯の海の涼しさよ
 << extract-kigo >>
 季語：涼し 読み：すずし 季節：三夏 カテゴリ：時候
 次に、俳句を形態素解析します。
 形態素解析：[洗ひ, 名詞], [たる, 助動詞], [硯, 名詞], [の, 助詞],
 [海, 名詞], [の, 助詞], [涼し, 形容詞], [さ, 接頭辞], [よ, 助詞] 切れ字：なし
 この俳句の特徴的な語句は「硯の海」です。
 硯の海とは、硯の墨汁をためておくくぼんだ部分のことです。
 墨池ではなく硯の海としたことで海そのものへのイメージも膨らみます。
 この俳句からは硯の海を洗っている状況が鮮明に浮かびます。
 また、硯を洗っている状況を季語「涼し」を用いて表現することで、
 気持ちの良い情緒を感じさせます。
 以上から4段階で評価します。
 評価：1 (非常に通る)
 Q: 「白波の立ち上がりたる夜光虫」という俳句は意味が通りますか？
 以下の4段階で評価してください。
 1: 非常に通る 2: 通る 3: どちらとも言えない 4: 通らない

表2 選択肢1と選択肢2を区別した場合の各モデルの正答率

	1 意味	2 共感性	3 心情	4 詩的表現	5 季語	6 語句
text-davinci-003	0.50	0.73	0.83	0.77	0.53	0.40
gpt-4-1106	0.43	0.77	0.83	0.80	0.60	0.60

なり2句である。基本的には有季定型句を対象としているが、本実験で扱う評価項目に対しては無季や季重なるの句を言語モデルがどのように評価するか確認することは効果的であると判断した。

実験の対象である俳句30句それぞれに対して2.1で示した6つの評価項目をどの程度満たすかを上記の4段階で選択する形式のアンケートを作成し、回答してもらった。アンケートの回答者は、愛媛県松山市の俳人12名である。

3.2 実験結果

アンケート結果とモデルの出力結果の比較を表2に示す。表中の値は各モデルの正答率であり、正答率はアンケートの結果が最も大きい割合を占めた評価と同様の評価をモデルが出力した割合である。また、評価の選択肢1(意味が非常に通る)と選択肢2(意味が通る)を区別しなかった場合の比較を表3に示す。アンケートのフィードバックにおいて、選択肢1と選択肢2の区別が難しいという意見が多かったため、区別しない場合についても結果を分析した。表2を見ると、モデル間で大きな正答率の差は見られず、どちらのモデルも共感性、作者の心情について比較的高い正答率を示した。

6つの評価項目の中でも特に俳句特有の理解が求められる季語を活かしているか、語句の対比はあるかについてはどちらのモデルも正答率が他の項目と比べて低い結果となった。また、表3の結果を見ると、表2で正答率の低かった意味が通るか、共感性があるか、語句の対比があるかという3つの評価項目に関して正答率の向上が見られる。この結果より、選択肢1と選択肢2を区別した場合の各モデルの誤答は明らかに間違いではなく、当てはまるかと評価した上でどの程度当てはまるのかという判断にアンケート結果と差があったということがわかった。季語を活かしているかという質問に対しては、無季の俳句に関してはtext-davinci-003は3(どちらとも言えない)、gpt-4-1106-previewは4(活かしていない)と評価し、季重なるの俳句に関しては1(非常に活かしている)、または2(活かしている)と評価した。

表3 選択肢1と選択肢2を区別しない場合の各モデルの正答率

	1 意味	2 共感性	3 心情	4 詩的表現	5 季語	6 語句
text-davinci-003	0.93	0.87	0.83	0.87	0.83	0.67
gpt-4-1106	0.90	0.87	0.87	0.87	0.80	0.60

表4 俳句「夏雲や転校生はスラッガー」を正しく評価した場合の推論過程。

この俳句は「夏雲」という季語と「転校生はスラッガー」という具体的な情景を組み合わせています。夏の雲を見上げながら、新しくやってきた転校生が野球で活躍する姿を想像させる内容です。多くの人が学生時代に経験するかもしれない転校生の新鮮さや、スポーツでの才能を発見するワクワク感を表現しています。この俳句は、夏の爽やかな情景と子供たちのスポーツを楽しむ様子を通じて、読者に共感を呼び起こす可能性があります。特に野球やスポーツに親しんだ経験がある人には、その感覚がより強く共感されるかもしれません。以上から「夏雲や転校生はスラッガー」という俳句は共感性があると言えます。評価：2 (ある)

表5 俳句「羊水を泳いだ記憶ハンモック」を誤って評価した場合の推論過程。

この俳句には「ハンモック」という言葉が含まれていますが、これは一般的な季語リストには含まれていない可能性があります。季語は日本の伝統的な四季に基づいた自然や風物詩に関連する言葉であり、ハンモックはそのような季節感を直接的には表現していないため、季語としての機能は果たしていないと考えられます。ただし、ハンモックが夏のイメージを連想させる場合もあるため、非公式に季節感を表現しているとも言えます。しかし、俳句の伝統的なルールに則ると、季語としては認められない可能性が高いです。そのため、この俳句は季語を活かしているとは言えず、評価は4(活かしていない)となります。

3.3 考察

評価に至るまでのモデルの推論過程の批評文章について、正しく推論した例と誤って推論した例をそれぞれ表4と表5に示す。表4は「夏雲や転校生はスラッガー」という俳句の意味や浮かばれる情景を適切に批評した上で評価項目である共感性の有無に触れ、アンケート回答の割合が最も大きかった2(共感性がある)と評価している。表5は季語「ハンモック」を季語取得ツールで取得しているにも関わらず、季語として認めていないためアンケート結果に反して4(活かしていない)と評価している。季語は、一見季節感が感じられない語句であっても季語と決められているものなので、抽出した季語をモデルが季語として扱うために意味合いなどの詳細な定義を与える必要があることがわかった。

4 おわりに

本稿では、俳句において特に重要だと思われる6つの評価項目に対して大規模言語モデルとChain of Thought Promptingを用いて評価する実験を行い、結果を俳人の評価データと比較した。意味や作者の心情など一般的な日本語文章でも用いられる項目に対しては比較的高い精度を示したが、季語を活かしているかや語句の対比など俳句ならではの項目には改善の余地が見られた。

参考文献

- [1] 川村秀憲, 山下倫央, 横山想一郎: 人工知能が俳句を詠む: AI一茶くんの挑戦, オーム社(2021).
- [2] Inaba, T., Kiyomaru, H., Cheng, F. and Kurohashi, S.: MultiTool-CoT: GPT-3 Can Use Multiple External Tools with Chain of Thought Prompting, ACL2023 (2023).
- [3]: GPT-4 Technical Report, OpenAI (2023).