7V-08

有季定型句の生成における深層学習モデル評価用データセットの作成と適用

平田 航大 †
北海道大学 大学院情報科学院 †
横山 想一郎 ‡
北海道大学 大学院情報科学研究院 ‡
山下 倫央 §
北海道大学 大学院情報科学研究院 §
川村 秀憲 ¶
北海道大学 大学院情報科学研究院 ¶

表1　句会の選による俳句の質の階層化.

	句会の選
レイヤー1	句会で高得点が入る
レイヤー2	句会でいくつか点数が入る
レイヤー3	句会で点数の入らない作品
レイヤー4	句会には出されないレベル

表2　句節の質による俳句の質の階層化.

	句節の質
レイヤー1	句節同士のつながりがある
レイヤー2	句節同士のつながりがない
レイヤー3	意味の通らない句節が1つ以上ある
レイヤー4	文法的におかしい箇所が1つ以上ある

1　はじめに

芸術分野における創作活動は，人間固有の欲求である知的好奇心や想像意欲を源とする人間特有の活動である．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].

日本で古くから親しまれてきた文章による芸術作品として俳句がある．俳句は，音数が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を基本とする定型詩であり，詠み手の情景や心情が表現される．鑑賞者が俳句から受ける印象は，時代背景や鑑賞者個人の持つ知識が大きな役割を果たすため，複数の鑑賞者が良いと考える俳句は必ずしも同一ではない．人工知能による創作を通じてこうした過程を紐解く上で，俳人がどんな俳句を良いと感じたかというデータが重要な役割を果たす．ここで，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．主観的な側面の強い俳句評価において，句会での得点数は俳人による評価の集合として扱うことができ，このフィードバックをモデル作成に活かすことで先の課題への有効なアプローチになる．

そこで本稿では句会での得点と句節の質に注目した，有季定型句の階層化とそれに基づく深層学習モデル評価用データセットの作成を行った．より大規模，複数人の俳人によるラベル付けを行う前の事前実験を実施し，データセットを用いたモデル評価と生成俳句のアンケート調査を行った．

2　有季定型句の質の階層化

句会の性質と俳人に対するヒアリングから，有季定型句の質を表す指標として「句会の選」と「句節の質」を設定し，階層構造により有季定型句の質を一側面から定義した．表1，2にその階層構造を示す．「句会の選」は句会における点数での分類，「句節の質」は句節間のつながりと句節の意味的，文法的な質で分類される．

3　実験

句会で高得点を獲得する俳句を生成可能なモデルは，表1，2の高いレイヤーの俳句に対し，より高い尤度を付与できると考えられる．そこで各モデルについて，実際の句会で投句された俳句を用いた分類問題に対し算出されるAUC，および生成俳句についてのアンケート調査による結果を比較し，モデルの評価を行った．

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AWD-LSTM，GPT-2，BARTを用いた．3種類の生成モデルを青空文庫データセットと俳句データセットにより訓練し，得られた俳句生成器を評価する．各モデルには，データセットの文章を文字ごとに分割し6,542種類のトークンを割り当てた整数列が入力される．各モデルの詳細を次に示す．

AWD-LSTM　3層のAWD-LSTMにより構成され，2,200万のパラメータ数を持つモデルである．先行研究で精度向上が確認されたfine-tuneと呼ばれる技法 [2] を適用した．

GPT-2　GPT-2の提案論文 [3] でGPT-2 Small（総パラメータ数約9千万）と呼ばれるモデルを用いた．

BART[4]　本来はEncoder-Decoderモデルだが，本研究では言語モデルとして用いるためデコーダ部分のみを用いた．（総パラメータ数約2億）

データセットを表3に示す．俳句データセットはインターネットで公開されている俳句を収集して用いる．事前学習では青空文庫データセット[*1]を8：1：1に分割し，学習，検証，テストデータとする．俳句データセットの学習時は5分割交差検証を実施し，異なる5つのランダムシードで計25回の学習を行う．それぞれ検証データに対する損失が最小のモデルを用いた．

俳人による評価付きのデータセットとして愛媛県松山市で開催された「ふくし句会」に投句された俳句計271句を用い，

[*1] https://github.com/aozorabunko/aozorabunko

Evaluation of a haiku generator with autoregressive models based on deep learning
† Kodai Hirata, 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3　学習に用いるデータセット

データセット名	作品数	総文字数
青空文庫データセット	16,222 作品	約 2.2 億字
俳句データセット	504,068 句	約 6,500 万字

表4　俳句テストデータに対するパープレキシティと分類問題の AUC

モデル名	パープレキシティ	AUC
AWD-LSTM	46.6	0.62
BART	34.3	0.66
GPT-2	30.5	0.70

正例・負例を判定する分類問題を定義した．正例は実際に句会に投句された 271 句であり，「句節の質」はレイヤー 1，「句会の選」は少なくともレイヤー 3 以上に分類されると筆者が判断した俳句である．負例としては正例の各文字列から季語と動詞を同音の異なる単語に置換することで作成した．例として，正例「端居して北海道のスープ飲む」，負例「紙衣して北海道のスープ摘む」のように作成される．

また，各モデルが，より高いレイヤーに分類される俳句を生成可能かどうかを確かめるため，生成俳句の有季定型句としての質についてのアンケート調査を実施した．

各モデルが生成した俳句 40 句と訓練データからランダムサンプリングした俳句 40 句の計 160 句に対して行った．なお今回は有季定型句としての質を調査するアンケートのため，各モデルが生成した俳句についてはあらかじめ 17 音，季語数 1 つの条件を満たす俳句を抽出してアンケートの対象俳句とした．アンケートの対象者は俳句歴 10 年以上の俳人 3 名，俳句歴がそれぞれ 1 年，2 年の，著者を含めた大学生 2 名である．

アンケート項目は以下の 3 つを設定した．それぞれ 0（あてはまらない），1（少し当てはまる），2（当てはまる）の 3 段階で回答をしてもらった．

- 意味が通る: 日本語として意味が通る俳句である
- 適切な季語: 季語が本来の意味である本意・本情に沿って使われている
- 句会で選: 句会で良い俳句として投票したい俳句である

”意味が通る”と”適切な季語” は「句節の質」ラベルのレイヤー 1，「句会の選」ラベルのレイヤー 3 以上に分類される俳句であるかを決める要素の一つである．”句会で選” 「句会の選」レイヤーのどこに分類される俳句であるかを決める要素である．

3.2 実験結果

3.2.1 俳句データに対するパープレキシティ

俳句テストデータに対するパープレキシティを表 4 に示す．一般的な文章生成タスクと同様に Transformer をベースとしたモデルである BART, GPT2 がより低いパープレキシティを達成している．

3.2.2 階層的評価に基づく分類問題に対する AUC

言語モデルが算出した尤度とラベルの間で AUC を算出した結果を表 4 に示す．GPT-2 が算出する尤度が，実際に句会に投句された俳句に対してより高い尤度を算出する傾向にあることがわかる．

3.3 生成俳句の質に関するアンケート調査

表 5 にアンケートの結果を示す．

”意味が通る”，”適切な季語” の項目で，Transformer ベースの 2 モデルが人間の俳句と同程度の質の俳句を生成できていることがわかる．また，”句会で選” の項目についても Transformer ベースの 2 モデルが人間の俳句に匹敵するという結果を示した．

表5　モデルの生成俳句と人間の俳句に対するアンケート調査の結果．表中の値は 3 段階の評価尺度の平均値である．

	意味が通る	適切な季語	句会で選
AWD-LSTM	0.5	0.4	0.1
BART	1.1	0.6	0.3
GPT-2	1.2	0.8	0.5
human	1.0	0.7	0.4

以上より今回の設定においては，パープレキシティと分類問題に対する結果，アンケート調査による主観評価の結果について同様の傾向を示すことがわかった．人間が作業を行う必要のある主観評価については実施可能な実験数が限られる．したがってある側面における俳句の質を測定可能なデータセットを作成し，良い性能を示したモデルを人間の主観評価にかけていく，という流れがモデル評価プロセスの一つとして有効であると考えられる．

4 まとめ・今後の展望

本稿では俳句の質に関する評価用データセットの構築を目指し，俳句の質の階層化と句会データを用いた実験を行った．句会データを用いた分類問題に対する AUC と生成俳句に対するアンケート調査の結果から，評価用データセットを用いた俳句生成器評価の可能性を示した．今後は階層化した有季定型句の質に基づき，より大規模で複数人の俳人によるラベル付きデータを作成し，より正確なモデル評価が可能なデータセットを構築していく．

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] Merity, S., Keskar, N. S. and Socher, R.: Regularizing and optimizing LSTM language models, *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings* (2018).

[3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I.: Language Models are Unsupervised Multitask Learners, Technical report.

[4] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguistics, pp. 7871–7880 (2020).