Vol.2022-ICS-206 No.11
2022/3/10

RoBERTaを用いた俳句評価器の構築と性能評価

花野 愛里咲[1,a)]　横山 想一郎[2]　山下 倫央[2]　川村 秀憲[2]

概要：本稿では深層言語モデルの RoBERTa を用いて俳句候補群から俳句として認識できる候補を選択する俳句評価器を構築する．人間が詠んだ俳句を正例，人工的に作成した俳句のルールを満たす文字列を負例として二値分類タスクをモデルに適用することで，俳句として成立する候補を選択する機能の獲得を目指す．実験では，人間が詠んだ俳句と人工的に作成した俳句の識別においては俳句評価器が高い精度で識別可能であることが示され，俳句生成器が生成した俳句に対しては俳句評価器の実用における可能性を示すことができた．俳人から選ばれる句の判定は現状の俳句評価器では難易度が高いタスクであることが確認できた．

1. はじめに

1.1 研究背景

近年，人工知能技術を用いた芸術作品の創作が盛んに行われている．その対象は，絵画，音楽，文学と多岐に渡っている [1]．創作においては，鑑賞者が良いと思う作品を生成することだけではなく，その中から鑑賞者が良いと思う作品を選ぶことも重要である．

本稿では，創作の対象として俳句を取り上げる．俳句は，17 音という短い音数で情景や心情を表現する．俳句の創作においては，詠み手が情景や心情を俳句で表現する過程や俳句から情景や心情を想像する仕組みを理解し，再現することが課題である．

人工知能技術を俳句に適用した先行研究として，深層言語モデルを用いて数百万句の俳句を生成した事例が報告されている [2]．数十万句の俳句データ，深層言語モデル，大規模な計算機リソースを駆使することで，大量に俳句を生成することは可能になった．一方で，俳句候補群から受け手に最も受け入れられる句を選句するということは容易ではない．選句において，特定の個人からの高い評価や多くの読者の支持を得られる俳句を選ぶためには評価者の選好・属性を考慮する必要があり，万人が共通して良いと感じるような俳句を選ぶことは難しい．また，生成された俳句の中には俳句として成立しないような質の低い句が含まれている．このような句を含んでいる候補の中から選句することは人間にとって余計に時間やストレスなどの負担がかかる．

1.2 研究目的

選句における人間への負担を軽減するために，俳句候補群の中から俳句として成立しない句を取り除き，できるだけ質の高い俳句だけを残すことができるような俳句評価器の構築を本研究の目的とする．そのための方法として，俳句評価器にマスク化言語モデル RoBERTa[3] を採用し，人間が作成した俳句を正例，俳句のルールを満たすように人工的に作成した文字列を負例として二値分類タスクを適用することで，俳句として成立する候補を選択する機能の獲得を目指す．

1.3 本稿の構成

本稿では，第 2 章に関連研究について説明する．第 3 章では本稿で取り扱う俳句について説明する．第 4 章では俳句評価器について説明する．第 5 章では俳句評価器に関する実験について説明する．最後に第 6 章で結論を述べる．

2. 関連研究

2.1 芸術作品の自動生成

近年では，人工知能技術が絵画や音楽，文学などの芸術作品の創作に活用されている．その中でも文章生成に関する研究として，星新一の作品を用いて約 8000 字以内の短い小説であるショートショートを生成する試み [4] や，Transformer[5] と variational autoencoder（VAE）[6] に基づくモデルを用いた和歌の自動生成 [7] に関する研究がある．創作の対象は様々であるが，本稿では俳句を取り扱う．

[1] 北海道大学 工学部
School of Engineering, Hokkaido University
[2] 北海道大学 大学院情報科学研究院
Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido, Japan
[a)] hanano@ist.hokudai.ac.jp

その理由として，俳句は 17 音という短さから生成・評価にかかるコストが他の文学作品よりも少ないことが挙げられる．

2.2 俳句の自動生成・評価

俳句の自動生成に関する研究として，太田ら [8] は，注意機構付きの系列変換モデルをベースにして，詠みたい俳句に含まれる単語列を入力系列として与えることにより俳句の生成を行っている．再帰層に Long Short-Term Memory（LSTM）[9] を用いている．正しい韻律で生成されやすくするための拍数素性と，入力系列と同じ季節の季語を出やすくするための季節素性，入力系列の季節の季語以外が出力されないような制約を注意機構付きの系列変換モデルに加えて比較検討をしている．この研究により，それぞれの素性や制約は生成された俳句に正しく反映されることが確認された．この研究では季語の有無と拍数で生成した俳句に対する評価を行っているが，これらの条件を既に満たしている俳句に対して質の評価を行う点が本研究の異なる点である．

2.3 言語モデル

深層学習を導入することによって自然言語処理技術は大きく向上し，翻訳や文書分類，質問応答といった自然言語処理タスクで従来よりも高い精度を達成した．中でも，言語モデルは発展に大きく貢献している．あらゆる自然言語処理タスクで当時の SoTA を達成した BERT[10] の登場以降，その派生として数多くのモデルが出現し，BERT の精度を超えるモデルも登場した．その中の 1 つに RoBERTa[3] というモデルがある．RoBERTa は，BERT における事前学習の設定や学習に用いるデータを変更したモデルである．事前学習の際に BERT は学習を行う前にあらかじめ文章の一部をマスキングするため，学習に用いるデータには重複したものが含まれるが，RoBERTa では学習の度に異なるマスキングを行うように変更した．この手法により，RoBERTa は BERT と同等，もしくはそれ以上の性能を発揮した．そのほかの変更点として，学習時のバッチサイズの増加や BERT の事前学習で用いられていた NSP の廃止がある．また，テキストのエンコーディングの方法を文字単位の Byte-Pair Encoding(BPE) からバイト単位の BPE に変更した．これらの設定の変更に加えて，事前学習用のデータセットと学習回数を増やしたことで，RoBERTa は GLUE と RACE[11] のいずれのタスクにおいても BERT の精度を上回った．RoBERTa が BERT の精度を上回ったタスクの 1 つに文書分類タスクがある．本研究では，俳句として成立するか否かという文書分類タスクとして俳句の評価を行うため，文書分類タスクで優れた精度を達成した RoBERTa を用いる．

3. 俳句とは

本章では，本研究で扱う俳句の概要について述べる．俳句は世界最短の定型詩であり，公益社団法人日本伝統俳句協会 [12] によると，俳句とは以下のルールを満たすものである．

- 5・7・5 の 17 文字 (音) で作る
- 季節の言葉 (季題) を入れる

5・7・5 で構成される俳句を定型俳句といい，その中でも季題を含む俳句を有季定型句と呼ぶ．季題は「季語」とも呼ばれ，それぞれの季語は特有の背景的な意味を有する．この背景的な意味は「本意・本情」と呼ばれ，歳時記などを通じて俳句を詠む人々の間で共有されている．季語には，その季語自体が持つイメージを上手く活用することで広がりのある余韻の深い世界を表現することができるという効果があり，[13] 俳句が 17 音という短い音数で心情や情景を伝えるために重要な役割を果たしている．また，句切れや意味，内容，リズムの切れ目を表す「切れ」を作るために，「や」「かな」「けり」などの「切れ字」を入れることが多い．切れ字には詠嘆や感動をより深くさせる効果がある．例えば，「古池や蛙飛び込む水の音」という俳句は，「蛙」という春の季語と切れ字の「や」を含んでいる．季語を含み，5・7・5 の 17 音であるため，この俳句は有季定型句である．定型俳句に対して，5・7・5 の型にとらわれない俳句を自由律俳句という．自由律俳句の例として，「分け入っても分け入っても青い山」という俳句がある．俳句の 5・7・5 の初めの 5 音を「上五」，真ん中の 7 音を「中七」，最後の 5 音を「下五」という．言葉が上五と中七，または中七と下五にまたがるものを「句またがり」，または「破調の句」という．これらの句は 5・7・5 のリズムを崩すことで独特な効果をもたらすと言われている．例えば，「大学のさびしさ冬木（ふゆき）のみならず」という俳句は，「冬木」という冬の季語を含み，「冬木のみならず」という言葉が中七と下五にまたがっているため，有季定型句で句またがりの句とされる．

4. 俳句評価器

この章では，本研究で提案する俳句評価器について説明する．俳句生成器が生成する俳句の中には，日本語として不自然であったり，意味が通らないような句が存在する．本研究ではこのような俳句として成立しない文字列を検出し，俳句生成器が生成した俳句候補群から取り除くことで，俳句として成立する候補の獲得を目的とする．そのために，人間が作った俳句をインターネットから収集したものを正例（俳句として成立する例）として用い，人工的に作成した俳句のルールを満たす文字列を負例（俳句として成立しない例）として用いる．ここで，学習データの作成方法は 2 通り考えられる．1 つ目は，俳句生成器が生成し

た俳句に俳句として成立するかどうかアノテーションを手動で行う方法である．この方法は，人間が 1 つ 1 つの俳句にアノテーションを行うため，時間や労力の面でコストがかかり十分な量のデータを得ることができないが，本研究の評価の対象となる俳句生成器が生成した俳句とデータの分布が一致するという点でデータの質は高いと言える．2 つ目は，既存のデータセットから人工的に自動で作成する方法である．この方法は，アノテーションにかかるコストが低く大量にデータを得ることができるが，本来の目的とは異なる分布から得たデータであるためデータの質は 1 つ目の方法よりは下がる．本研究では，言語モデルを用いて文書分類タスクとして俳句の評価を行うため，大量のラベル付きデータを得ることを重視し，人工的にデータを作成する方法を採用した．入力として与えられた文字列が正例か負例かを識別する二値分類タスクを RoBERTa で学習することで俳句評価器を構築する．多くの自然言語処理のタスクで SoTA を獲得した BERT よりも，文書分類タスクにおいて高い精度を達成した RoBERTa を採用した．

4.1 データセット

負例として，以下の 3 種類のデータセットを作成した．

- 交換データセット : 正例の俳句に含まれる名詞や季語，助詞，動詞，上五，下五のいずれかを同じ種類の別の単語に交換して生成する．例えば，正例の「みちのくの青田に降りる山の雲」という俳句の「青田」という季語をランダムに選ばれた季語「ざくろ」に交換することで「みちのくのざくろに降りる山の雲」という文字列を生成する．交換データセットは，意味が通らない俳句を検出することを目的とする．意味が伝わるように人間が作成した俳句の中からランダムに単語を選び変えるため，意味が通りにくい文字列となる．以下に交換データの例を示す．括弧内は単語交換前の俳句である．
  - 蝉の声きゝそめてより山路かな（鶯(うそ)の声きゝそめてより山路かな）
  - 雲水ののびきてうつる春田かな（犬の首のびきてうつる春田かな）
  - 春月に竹の騒げる嵐かな（春月に竹の騒げる阿波路かな）
  - 烈風になだれて狂ふ樅(もみ)の聖（烈風になだれて狂ふ樅の雪）
  - 田も畑もまだ目覚めざる甲高車（田も畑もまだ目覚めざる初電車）
- 散文データセット : 日本語コーパスである青空文庫 [14]，CC100[15]，Wiki-40b[16] を使用して，これらに含まれる有季定型句を満たす部分文字列を抽出して生成する．青空文庫に含まれる既存の俳句との重複を避けるため，正例として使用する俳句データとの最小編集距離が 5 以下の文字列を除外する．例えば，青空文庫の作品の中の「春の来るのがどのくらい祝福であるかをお察しする。」という文から季語である「春」を含み 17 音になる「春の来るのがどのくらい祝福で」という文字列を抽出する．散文データセットは，詩的要素を含まない俳句を検出することを目的とする．俳句は 17 音という限られた音数で伝えたいことを表現するために，省略を上手く活用し，読み手の想像を膨らませ感動を与えるとされている．しかし，散文データセットの作成元の日本語コーパスは説明的な文章であるため，省略が重要という俳句の特徴を捉えていない文字列となる．こういった文字列は俳句として成立しないと言える．以下に散文データの例を示す．
  - 掛乞の来てしまひたる三時かな
  - 泣いてゐる窓の硝子にさす月も
  - 乾燥し寒い地域の大半で
  - 台風の頻発などが発生し
  - 披露したところ多くの審査員
- ランダムデータセット : 青空文庫のデータを形態素解析器 MeCab のラッパーライブラリである fugashi を用いて単語に分割し，品詞ごとに別々のリストに単語とその音数を格納した後，ランダムに品詞を選び，その品詞に対応するリストからランダムに単語を抽出し，17 音以上になるまで単語を結合する処理を繰り返す．作成した文字列が 17 音ちょうどで季語を含んでいたら処理を停止し，季語を含んでいない場合は文字列の中から単語を 1 つランダムに選択し，その単語と同じ音数の季語に置き換える．作成した文字列が 17 音よりも大きくなった場合は，生成された文字列の音数の合計が 17 音以下になるようにランダムに単語を 1 つ取り除き，17 音になったら先ほどと同様の処理を行い，17 音よりも小さくなった場合は 17 音になるように足りない音数と同じ音数の季語と入れ替える．もし指定回数内に 17 音以下にならなければ最初から文字列を作り直す．例えば，青空文庫の作品から合計 17 音になるようにランダムに選択した名詞「家族連れ」，接続詞「しかし」，形容詞「程近い」，「はかない」を結合する．次にこれらの 4 つの単語からランダムに選択された「家族連れ」を同じ音数の季語「冬の雲」に置き換えて「冬の雲しかし程近いはかない」という文字列を生成する．ランダムデータは，単語の並びが不自然な文字列を検出することを目的とする．3 つの負例の中で最も意味が破綻しているデータセットである．以下にランダムデータの例を示す．
  - 片付かかゝわら海栗思い止まら
  - ワタシ涼風着飾らつつがなかっ

表 1　事前学習のパラメータ設定

パラメータ名	設定値
モデル	roberta-base
学習データ	12,978 作品
エポック数	20
バッチサイズ	4
ドロップアウト率	0.1

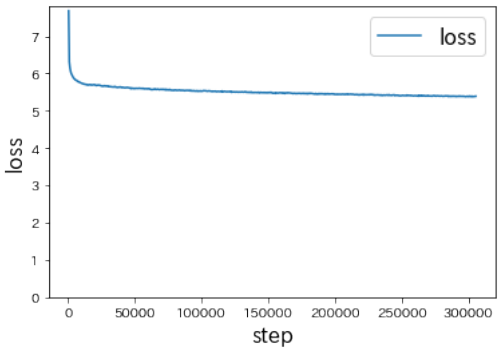

図 1　事前学習の様子

– 雪突きかかろう行き渡ら呼出し
– して小田巻しめよ潜り込ま小麦
– 悪賢くかしこそ五月の節句

4.2 モデル

本研究では，俳句評価器としてマスク化言語モデルである RoBERTa を使用した．正例と負例を識別する分類タスクを解く前に，青空文庫のデータの一部をマスクし，マスクされた部分に入る単語を予測するというタスクを適用することで事前学習を行った．事前学習のパラメータ設定を表 1，学習の様子を図 1 に示す． 青空文庫のデータで事前学習したモデルに正例と負例を識別するよう学習し，学習データに用いる負例が異なる 5 つのモデルを作成した．これらのモデル名を「交換モデル」「散文モデル」「ランダムモデル」「混合モデル①」「混合モデル②」とする．すべてのモデルで学習に用いる正例は俳句データを 40 万句使用し，検証に用いる正例の数は 4 万句である．交換モデル，散文モデル，ランダムモデルは学習に用いる負例としてそれぞれ交換データ，散文データ，ランダムデータを 40 万句，検証では 4 万句使用した．混合モデル①は学習に用いる負例として，3 つの負例を合計 40 万句（交換データが 133,334 句，他 133,333 句），検証では 3 つの負例を合計 4 万句（交換データが 13,334 句，他 13,333 句）使用し，混合モデル②は学習に用いる負例として，3 つの負例をそれぞれ 40 万句，検証ではそれぞれ 4 万句使用した．そのため，混合モデル②が学習に用いた負例の数が 120 万句と最も多いということになる．過学習を防ぐ早期停止を導入し，1 エポック終了ごとに計算される検証データに対する損失が最も低いモデルを早期停止のモデルとした．各モデルのパラメータ設定を表 2 に示す．交換モデルはドロップアウト率が 0.3 の場合よりも 0.1 の場合の方が検証データに対する損失の最小値が低かったためドロップアウト率が 0.1 のモデルを採用した．モデルの実装には Hugging Face の Transformers[17] を使用した．

表 2　ファインチューニングのパラメータ設定

パラメータ名	設定値
モデル	roberta-base
学習データ	800,000 句（混合モデル②は 1,200,000 句）
エポック数	50
バッチサイズ	512
ドロップアウト率	0.3（交換モデルは 0.1）

5. 実験

本章では，俳句評価器の性能評価に関する以下の 3 つの実験について説明する．

(1) 学習データと同様の方法で作成した検証データに対する評価
(2) 俳句生成器が生成した俳句に対する評価
(3) 俳人の評価との一致度合いによる評価

本研究では俳句生成器が生成した俳句に対して評価を行うことが目標であるため，2 つ目の実験における評価基準が本研究の目的に合ったものであるが，俳句評価器の構築の際の学習データは既存の俳句や日本語コーパスに手を加えて人工的に作成したデータであり，俳句生成器が生成した俳句とは異なる分布から得たデータを用いているため，学習データと同じ性質のデータに対する汎化性能の検証も必要である．よって，1 つ目の実験を行う．3 つ目の俳人の評価との一致度合いによる評価実験における俳句評価器の使用方法は，本研究の目的である俳句として成立しない句を取り除くこととは異なるが，俳句評価器を用いることで得られる俳句候補群は俳句として成立する句であると考え，これらの句と俳人から評価を得られる句との相関関係の有無を調査する．

5.1 学習データと同様の方法で作成した検証データに対する評価

5.1.1 実験目的

この実験では，学習データと同様の方法で作成した検証データを用いて作成したモデルの性能評価を行う．

5.1.2 実験方法

評価指標には混同行列と Precision@k を用いる．本研究では，俳句評価器によって出力された評価値上位の俳句候補群に意味が破綻している文字列や日本語として不自然な単語の並びになっている文字列が含まれることを避けるため，正例と予測されたものの中で負例の数が少ないモデルを最も良いモデルとし，Precision@k によりモデルの性能

表 3 交換モデル（早期停止なし）の混同行列

データセット	正例と予測された数	負例と予測された数
正例（俳句）	23,766	16,234
負例（交換）	7,233	32,767
負例（散文）	10,924	29,076
負例（ランダム）	224	39,776

表 4 交換モデル（早期停止あり）の混同行列

データセット	正例と予測された数	負例と予測された数
正例（俳句）	32,931	7,069
負例（交換）	17,616	22,384
負例（散文）	23,945	16,055
負例（ランダム）	1,982	38,018

表 5 散文モデル（早期停止なし）の混同行列

データセット	正例と予測された数	負例と予測された数
正例（俳句）	38,433	1,567
負例（交換）	38,453	1,547
負例（散文）	957	39,043
負例（ランダム）	19,667	20,333

表 6 散文モデル（早期停止あり）の混同行列

データセット	正例と予測された数	負例と予測された数
正例（俳句）	38,673	1,327
負例（交換）	38,728	1,272
負例（散文）	1,217	38,783
負例（ランダム）	21,520	18,480

表 7 ランダムモデル（早期停止なし）の混同行列

データセット	正例と予測された数	負例と予測された数
正例（俳句）	39,879	121
負例（交換）	39,624	376
負例（散文）	37,509	2,491
負例（ランダム）	70	39,930

評価を行う．

5.1.3 実験結果・考察

早期停止を導入しない交換モデルと早期停止を導入した交換モデルによる混同行列をそれぞれ表 3，表 4，早期停止を導入しない散文モデルと早期停止を導入した散文モデルによる混同行列をそれぞれ表 5，表 6，早期停止を導入しないランダムモデルと早期停止を導入したランダムモデルによる混同行列をそれぞれ表 7，表 8，早期停止を導入しない混合モデル①と早期停止を導入した混合モデル①による混同行列をそれぞれ表 9，表 10，早期停止を導入しない混合モデル②と早期停止を導入した混合モデル②による混同行列をそれぞれ表 11，表 12 に示す．混合行列の算出に使用したデータセットは検証データに用いたデータと同じものであり，俳句データ，交換データ，散文データ，ランダムデータをそれぞれ 4 万句含む．モデルによって出力された俳句に対する評価値が 0.5 以上であれば正例，0.5 未満であれば負例と判定される．各モデルの accuracy と真陰性率を表 13，precision，recall を表 14 に示す．

表 8 ランダムモデル（早期停止あり）の混同行列

データセット	正例と予測された数	負例と予測された数
正例（俳句）	39,835	165
負例（交換）	39,425	475
負例（散文）	37,531	2,469
負例（ランダム）	76	39,924

表 9 混合モデル①（早期停止なし）の混同行列

データセット	正例と予測された数	負例と予測された数
正例（俳句）	35,267	4,733
負例（交換）	25,089	14,911
負例（散文）	1,018	38,982
負例（ランダム）	52	39,948

表 10 混合モデル①（早期停止あり）の混同行列

データセット	正例と予測された数	負例と予測された数
正例（俳句）	36,933	3,067
負例（交換）	29,647	10,353
負例（散文）	2,073	37,927
負例（ランダム）	66	39,934

表 11 混合モデル②（早期停止なし）の混同行列

データセット	正例と予測された数	負例と予測された数
正例（俳句）	31,004	8,996
負例（交換）	14,079	25,921
負例（散文）	621	39,379
負例（ランダム）	10	39,990

表 12 混合モデル②（早期停止あり）の混同行列

データセット	正例と予測された数	負例と予測された数
正例（俳句）	33,348	6,652
負例（交換）	22,497	17,503
負例（散文）	1,203	38,797
負例（ランダム）	28	39,972

表 13 各モデルの accuracy と真陰性率

モデル	早期停止	accuracy			真陰性率		
		俳句-交換	俳句-散文	俳句-ランダム	俳句-交換	俳句-散文	俳句-ランダム
交換		70.6%	66.0%	79.4%	81.9%	72.6%	99.4%
交換	○	69.1%	61.2%	88.6%	55.9%	40.1%	95.0%
散文		49.9%	96.8%	73.4%	3.8%	97.6%	50.8%
散文	○	49.9%	96.8%	71.4%	3.1%	96.9%	46.2%
ランダム		50.3%	52.9%	99.7%	0.9%	6.2%	99.8%
ランダム	○	50.3%	52.8%	99.6%	1.1%	6.1%	99.8%
混合①		62.7%	92.8%	94.0%	37.2%	97.4%	99.8%
混合①	○	59.1%	93.5%	96.0%	25.8%	94.8%	99.8%
混合②		71.1%	87.9%	88.7%	64.8%	98.4%	99.9%
混合②	○	63.5%	90.1%	91.6%	43.7%	96.9%	99.9%

表 5〜表 8 より，散文モデルとランダムモデルは交換データを正しく分類できている数が少ないことから，交換データを正しく負例と分類するためには学習データとして用いる必要があると言える．また，ランダムデータはどのモデルでも正しく負例と分類できている数が多いことから，学習データとして用いなくても分類が容易なデータであると考えられる．

表 14　各モデルの precision と recall

モデル	早期停止	precision			recall		
		俳句-交換	俳句-散文	俳句-ランダム	俳句-交換	俳句-散文	俳句-ランダム
交換		76.6%	68.5%	99.0%	59.4%	59.4%	59.4%
交換	○	65.1%	57.8%	94.3%	82.3%	82.3%	82.3%
散文		49.9%	97.5%	66.1%	96.0%	96.0%	96.0%
散文	○	49.9%	96.9%	64.2%	96.6%	96.6%	96.6%
ランダム		50.1%	51.5%	99.8%	99.6%	99.6%	99.6%
ランダム	○	50.1%	51.4%	99.8%	99.5%	99.5%	99.5%
混合①		58.4%	97.1%	99.8%	88.1%	88.1%	88.1%
混合①	○	55.4%	94.6%	99.8%	92.3%	92.3%	92.3%
混合②		68.7%	98.0%	99.9%	77.5%	77.5%	77.5%
混合②	○	59.7%	96.5%	99.9%	83.3%	83.3%	83.3%

表 15　precision@k による評価

モデル	早期停止	k=500	k=5000	k=50000
交換		0.376	0.594	0.534
交換	○	0.776	0.681	0.504
散文		0.494	0.469	0.445
散文	○	0.400	0.418	0.448
ランダム		0.428	0.388	0.042
ランダム	○	0.404	0.421	0.403
混合①		0.840	0.786	0.618
混合①	○	0.784	0.724	0.590
混合②		0.874	0.865	0.656
混合②	○	0.816	0.755	0.606

早期停止の導入による効果については，ランダムモデルに関しては早期停止なしのモデルよりも早期停止ありのモデルの方が負例を正しく負例と判定できた数がわずかに増加したが，他のモデルでは早期停止なしのモデルよりも早期停止ありのモデルの方が正例を正しく正例と判定できた数は増加したものの，負例を正しく負例と判定できた数は減少し，早期停止の導入による効果はほとんどのモデルで確認することができなかった．

次に，Precision@k を用いたモデルの評価を行う．混同行列を算出した際に用いた検証データにおいて，各モデルの評価値の上位 k 番目までに含まれる正例の割合を算出した．調査した k の値は 500,5000,50000 の 3 つである．本研究においては，正例を誤って負例と判定することはある程度許容されるが，負例を誤って正例と判定することは許容できないため，k の値をデータセット全体と比べて小さい値にした．

各モデルの Precision@k による結果を表 15 に示す．表 15 より，早期停止なしの混合モデル②は 4 万句のなかからランダムに 1 句選択する場合，選択したものが俳句として成立する（正例）である確率が 25 %であるが，このモデルを俳句評価器として用い上位 500 句の中から選ぶ場合ではその確率が約 87 %に向上することがわかる．このことから，混合モデル②を俳句評価器として用いることに有用性があると言える．次に，各モデルの評価値上位 500 句に含まれる各データセットの数を表 16 に示す．

表 16　上位 500 句に含まれる各データセットの数

モデル	早期停止	データセット			
		俳句	交換	散文	ランダム
交換		188	23	287	2
交換	○	388	50	62	0
散文		247	234	4	15
散文	○	200	297	0	3
ランダム		214	235	51	0
ランダム	○	202	282	16	0
混合①		420	74	6	0
混合①	○	392	96	12	0
混合②		437	58	5	0
混合②	○	408	87	5	0

表 17　評価値の違いによる負例の比較

評価値上位 500 句に含まれる負例	評価値上位 500 句に含まれない負例
春雨や寝返りもせぬ膝の猫	代表で出場し生年男子
思ひ出し笑ひをつめては炉灰掻く	暑二天一磨かさばさついシカモ
子の頬に妻の香もある桜桃忌	台風の頻発などが発生し
ひとりできに死ねさうもない寒の星	犠牲者の数はどんどん減少し
曇天へ煙直ぐ下がる野焼かな	月以上従事したもの二学校
碓氷の窓に見上る岡の桜哉	して南ジョージア湾の捕鯨基地
しばらくは庭をしてをり初嵐	披露したところ多くの審査員
田の水を見せて穂田のさかりかな	ラーメンと蕎麦はしつこく存在し
はつ雪や鴛をあふ人鴛の人	場合でもキャンセル料は発生し
掌に浮くて掌のいろとなる落花かな	分離した後もセルビア代表と
武蔵野や粟に咲入る草の花	警告に注意しスキーパトロール
おほよそのこと見えてをり濁端居	影響はほとんど無いと判断し
松籟にひかりあがりし寒雁かな	スタン麻痺状態中は発動し
一人居ののよく通る道黍畑	指定した月の月間ページビュー
桶に鴛つて海立つ春の鴎かな	使用する年齢層はびっくりし

表 16 において，早期停止なしの混合モデル②による評価値上位 500 句に含まれる 5 つの散文データのうち，4 つが青空文庫と Wiki-40b に載っている俳句作品であることが確認された．そのため，これらの 4 句は負例であるが評価値の上位に来ることは俳句評価器としての実用には問題ないと言える．表 17 に早期停止なしの混合モデル②による評価値の上位 500 句に含まれる負例と，評価値の上位 500 句に含まれない負例の例を示す．表 17 の右側の評価値上位 500 句以内に含まれない負例を見ると，日本語として不自然な単語の並びをしているものや文章の途中から始まっているような文字列など意味が通らず明らかに俳句としてみなすことができないような文字列が多く含まれていることが確認できる．これに対し，表の左側の評価値の上位 500 句に含まれる負例は，「一人居ののよく通る道黍畑」の句のように一部単語の並びに不自然な点が見られるものが含まれているものの，ほとんどは俳句としてみなせるものが多い．このことから，実験 1 で作成した早期停止なしの混合モデル②は俳句として成り立たないような文字列を取り除くことができていると言える．

表 18　俳句生成器が生成した俳句に対する各モデルの precision@k

モデル	早期停止	K=10	K=20	K=30	K=40	K=50
交換		0.200	0.300	0.400	0.450	0.460
交換	○	0.500	0.550	0.566	0.525	0.480
散文		0.400	0.350	0.400	0.350	0.400
散文	○	0.500	0.450	0.433	0.400	0.400
ランダム		0.600	0.450	0.366	0.375	0.380
ランダム	○	0.600	0.450	0.400	0.375	0.400
混合		0.600	0.600	0.500	0.450	0.440
混合	○	0.800	0.650	0.566	0.525	0.480
混合		0.600	0.550	0.430	0.400	0.460
混合	○	0.700	0.550	0.466	0.425	0.360

5.2　俳句生成器が生成した俳句に対する評価

5.2.1　実験目的

俳句評価器が生成した俳句に対する俳句評価器の実用性を検証する．

5.2.2　実験方法

俳句評価器が生成した句をランダムに 100 句取得し，意味が通る俳句を正例，意味が通らない俳句を負例としてアノテーションを行った．評価指標に Precision@k を用いて俳句評価器の性能評価を行う．

5.2.3　実験結果・考察

結果を表 18 に示す．また，100 句中 41 句が正例となった．表 18 より，混合モデル②（早期停止なし）は 100 句の中からランダムに選択すると意味が通る俳句は 41 %含まれるが，評価値の上位 10 %を選ぶことで 60 %まで向上するという結果が得られた．しかし，実験 1 の検証データに対する評価よりは効果が見られなかった．今回は 100 句しかサンプリングしていないため，今後の課題としてはアノテーションを行う句の数を増やして再実験を行うことが挙げられる．

5.3　俳人による評価付きデータを用いた性能評価

5.3.1　実験目的

俳句生成器が生成した俳句に対して俳人がランク付けを行ったデータを用いて，俳句評価器による評価と俳人による評価の一致度合いを確認する．本研究の目的は作成したモデルを俳句評価器として用いた場合に俳句として成り立たないものを除くことであるが，除いた結果得られた俳句として成立する句と俳人から評価を得られるような句には相関があるのか確認する．

5.3.2　実験設定

俳句イベントを利用して，俳句評価器が生成した俳句に対する俳人の評価付きデータを収集した．収集したデータは 2019 年と 2021 年の 2 回分のイベントの俳句データである．どちらのイベントでも，俳句評価器が生成した俳句を俳人 1 人につき 300 句配布し，イベントのテーマである「恋」にまつわる俳句の中で最も良いと感じた俳句で☆特選 1 句，テーマにかかわらず良いと感じた俳句で特選 5 句，並選 30〜60 句を選んでもらった．配布した 300 句の俳句は俳人ごとに異なる．配布した俳句はイベントのテーマに沿ったキーワード（「あなた」,「初恋」など）を含む．イベントの参加者数は，2019 年は 26 名，2021 年は 24 名であった．2019 年のイベントで使用した俳句データは，LSTM により構成された俳句生成器が生成した俳句の中から，人間が作成した俳句を正例，正例の中の任意の 2 つの単語を交換した文字列を負例として分類タスクを適用し LSTM により構成された俳句評価器で評価値が上位の俳句を使用した．2021 年のイベントで使用した俳句データは，GPT-2 により構成された俳句生成器が生成した俳句の中から，各キーワードごとに GPT-2 対数尤度が上位 5000 番目までの俳句を取得し，日本語として意味が通じない俳句を取り除いたものを使用した．2 つのイベントで収集した俳人の評価付き俳句データの例を表 19 に示し，それぞれの選における句数を表 20 に示す．

表 19　俳句イベントで収集した俳句データの例

	2019 年のイベントの俳句	2021 年のイベントの俳句
☆特選	初恋の焚火の跡を通りけり	香水を深めて嘘をつきはじむ
特選	初恋や鏡の中に猫がゐる	初恋の人をかくして牡丹雪
並選	笹鳴や水に心のありにけり	配達の二人来てゐる障子かな
選なし	座禅草艶なる村の暗さかな	初恋をくりかへしゆく紙帳かな

表 20　俳句イベントで収集した俳句データの数

	2019 年	2021 年
☆特選	26	24
特選	130	120
並選	592	1,069
選なし	7,052	5,987
合計	7,800	7,200

評価指標には AUC を用いる．AUC は閾値を変えたときの真陽性率と偽陽性率の関係をプロットした ROC 曲線下の面積を表す．並選，特選，☆特選以上の評価を持つ俳句をそれぞれ正例とし，俳句に対して評価器が算出した正例に対する予測確率が大きい順に俳句を並び替えて AUC を算出する．AUC は値が 1.0 に近いほど俳人による評価と一致していることを示し，0.5 に近いほどランダムな分類器と同程度の性能を示す．

5.3.3　実験結果・考察

表 21 と表 22 にそれぞれ 2019 年と 2021 年のイベントで収集した俳句データを用いて算出した AUC の結果を示す．混合モデルはより多くの負例を検出することができた混合モデル②の方で検証した．

表 21 と表 22 より，2019 年と 2021 年の両方のイベントで，正例が並選以上の場合と特選以上の場合のどちらにおいても最も AUC が高いモデルが交換モデルであることがわかる．2019 年の俳句データは，単語が交換された俳句を負例として学習した俳句評価器が算出した評価値が上位であったものの中から選んだ俳句であるため，同様な方法で

表 21　2019 年のイベントの俳句データの AUC 算出結果

モデル	早期停止	AUC 算出における正例の対象		
		並選以上	特選以上	☆特選以上
交換		0.600	0.630	0.517
交換	○	0.598	0.598	0.500
散文		0.498	0.487	0.515
散文	○	0.491	0.507	0.560
ランダム		0.500	0.500	0.500
ランダム	○	0.501	0.501	0.485
混合②		0.596	0.582	0.554
混合②	○	0.555	0.547	0.510

表 22　2021 年のイベントの俳句データの AUC 算出結果

モデル	早期停止	AUC 算出における正例の対象		
		並選以上	特選以上	☆特選以上
交換		0.543	0.570	0.567
交換	○	0.551	0.552	0.477
散文		0.498	0.488	0.412
散文	○	0.472	0.488	0.469
ランダム		0.500	0.500	0.500
ランダム	○	0.502	0.502	0.505
混合②		0.544	0.561	0.509
混合②	○	0.518	0.512	0.474

学習を行った交換モデルによる AUC が高くなったと考えられる．混合モデル②は実験 1 で負例を最も多く検出できることを示したが，AUC の値は 0.5 に近いことから，俳人との評価の一致度合いはあまり見られなかった．

6. 結論

本研究では，人間が作った俳句を正例，人工的に作成した俳句のルールを満たす文字列を負例として二値分類タスクを深層言語モデルに適用することで，俳句として成立しない句を除去し，俳句として成立する候補を獲得するための俳句評価器を開発した．1 つ目の実験では，モデルの学習に用いたデータと同様の方法で作成した検証データを用いて俳句評価器の評価を行った．負例として使用するデータの種類と量が最も多いモデルが他のモデルよりも高い精度で負例を検出できることを確認し，明らかに俳句として成立しない句には低い評価値を付けることができていることが確認できた．2 つ目の実験では，俳句生成器が生成した俳句を用いて俳句評価器の評価を行った．俳句生成器が生成した俳句に対しては実験 1 の検証データに対する評価ほど良い結果は見られなかったものの，俳句評価器を使う方が俳句として成立する候補を得られる確率が上がることが確認できた．3 つ目の実験では，俳人による評価付きのデータを用いて俳句評価器による評価と俳人による評価の一致度合いを確認したが，評価が一致することは確認できなかった．現状では，俳人から選ばれる句の判定機能は獲得できておらず，選句における負担を減らすための今後の課題である．

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] 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, pp. 467–474（オンライン），DOI: 10.11517/jjsai.34.4_467 (2019).
[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv*, (online), available from ⟨http://arxiv.org/abs/1907.11692⟩ (2019). cite arxiv:1907.11692.
[4] 松原　仁，佐藤理史，赤石美奈，角　薫，迎山和司，中島秀之，瀬名秀明，村井　源，大塚裕子：コンピュータに星新一のようなショートショートを創作させる試み，人工知能学会全国大会論文集，Vol. JSAI2013, pp. 2D11–2D11（オンライン），DOI: 10.11517/pjsai.JSAI2013.0_2D11 (2013).
[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention Is All You Need, *CoRR*, Vol. abs/1706.03762 (online), available from ⟨http://arxiv.org/abs/1706.03762⟩ (2017).
[6] Kingma, D. P. and Welling, M.: Auto-Encoding Variational Bayes (2014).
[7] Takeishi, Y., Niu, M., Luo, J., Jin, Z. and Yang, X.: WakaVT: A Sequential Variational Transformer for Waka Generation, *CoRR*, Vol. abs/2104.00426 (online), available from ⟨https://arxiv.org/abs/2104.00426⟩ (2021).
[8] 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，技術報告 1，奈良先端科学技術大学院大学, 奈良先端科学技術大学院大学, 奈良先端科学技術大学院大学 (2018).
[9] Sundermeyer, M., Schlter, R. and Ney, H.: LSTM neural networks for language modeling, *Proc. Interspeech 2012*, pp. 194–197 (online), DOI: 10.21437/Interspeech.2012-65 (2012).
[10] Devlin, J., Chang, M., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *CoRR*, Vol. abs/1810.04805 (online), available from ⟨http://arxiv.org/abs/1810.04805⟩ (2018).
[11] Lai, G., Xie, Q., Liu, H., Yang, Y. and Hovy, E.: RACE: Large-scale ReAding Comprehension Dataset From Examinations, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 785–794 (online), DOI: 10.18653/v1/D17-1082 (2017).
[12] : 日本伝統俳句協会，https://haiku.jp/. (Accessed on 02/04/2022).
[13] : 俳句入門講座-1，`https://haiku.jp/tsukuru/2745/`. (Accessed on 02/04/2022).
[14] : aozorabunko/aozorabunko - GitHub, `https://github.com/aozorabunko/aozorabunko`. (Accessed on 02/04/2022).
[15] : CC-100: Monolingual Datasets from Web Crawl Data, `https://data.statmt.org/cc-100/`. (Accessed on 02/04/2022).
[16] : wiki40b — TensorFlow Datasets, `https://www.tensorflow.org/datasets/catalog/wiki40b`. (Accessed on 02/04/2022).
[17] : Transformers, `https://huggingface.co/docs/transformers/index`. (Accessed on 02/04/2022).