G-09

SeqGAN を用いた一般人に好まれやすい俳句の生成

Generation of Haiku Preferable for Ordinary People by SeqGAN

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

Bungo Konishi　Hirota Atsushi　Seigo Matsuo　Ryo Iehara　Soichiro Obara
Yuta Kaga　Joji Tsuruda　Yukihiro Wakigami　Ryosuke Kanajiri　Chie Fukada
Kazuaki Tanaka　Natsuki Oka

1. はじめに

俳句とは，五音・七音・五音の十七音で構成された日本語の定型詩である．17 世紀に，松尾芭蕉らが俳句の芸術性を高め，現代においても広く親しまれている近代文芸の一種となった．本来の俳句は，季語を持つものを指すが，現代においては，季語を持たないものも俳句に含めることが一般的であり，本研究においても，季語を持たないものも俳句と呼ぶことにする．

1989 年(平成元年)からは，一般人から俳句を公募する伊藤園お〜いお茶新俳句大賞と呼ばれる創作俳句コンテストが始まったことにより，子どもから大人まで誰でも気軽に自身の作品を発表することができるようになった．平成 28 年第 28 回大会の応募句数は，187 万句を超え，年々上昇傾向にある[1]．このことからも，俳句は，今注目を集めている文化だといえるだろう．

そんな俳句を，機械により自動生成することができれば，機械が詠んだ俳句を人間が楽しむといった新たな娯楽が生まれ，人間が機械により親しみを持つきっかけになるのではないかと考えている．そこで，本研究は，俳句や文学に精通していない一般人がより好む俳句を生成するシステムの構築を目標とする．

2. 関連研究

Yu ら[2]は深層学習を用いた下記の 4 つの手法を用いて，文字レベルでの俳句の自動生成を行った．

1. Vanilla LSTM
2. Multi LSTM
3. R-CNN
4. SeqGAN

1. Vanilla LSTM と 2. Multi LSTM は，リカレントニューラルネットワーク(RNN)の一つである LSTM[3]を用いた言語モデルである．これはニューラル確率的言語モデル(NPLM)と呼ばれ，統計的言語モデルをニューラルネットワーク(NN) で構成するものである．統計的言語モデルとは，ある単語列$w_1, w_2, \ldots, w_i$ が与えられた時に，その次にくる単語w_{i+1}を予測するモデルである．3. R-CNN は，character embedding によって，Bag-of-Words のようなベクトルを生成し，それを畳み込みニューラルネットワーク(CNN)を用いた NPLM で，俳句を生成する手法である．最後に，4. SeqGAN は Wu ら[4]が提案した学習モデルである．SeqGAN では，生成器(G)と識別器(D) の 2 種類の NN を用いる．まず，人が詠んだ俳句を用いて NPLM を学習させることにより生成器 G を作成する．次に，生成器が生成した俳句と人が詠んだ俳句とを CNN を用いた識別器 D によって比較する．生成器が識別器を騙すことができれば正の報酬，騙すことができなければ負の報酬が生成器 G に与えられる．

Yu らの行った生成俳句の評価は，Perplexities と呼ばれる，分岐数を表す尺度を用いた評価にとどまっており，生成された俳句の人間による主観的評価は行われていない．

本研究では，Wu らの提案した SeqGAN を改良し，俳句や文学に精通していない一般人による主観的評価がより高い俳句の生成を目指す．

3. 実験

3.1 提案手法

本研究では，G と D に別々の種類のデータを与えて事前学習を行うことを考える．提案手法における SeqGAN の構成を図 1 に示す．

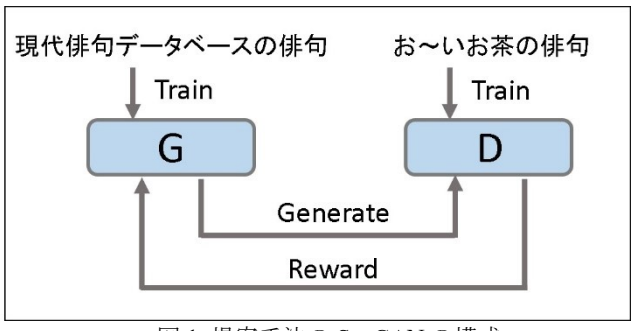

図 1. 提案手法の SeqGAN の構成

本研究においては，G に現代的な俳句データを与えて，D には，お〜いお茶新俳句大賞の受賞作品の俳句データを与えて事前学習を行う．つまり，G は現代的な俳句を学習するが，G によって生成されたデータを，D はお〜いお茶の俳句の観点から評価する．

D に与えるお〜いお茶の俳句に近い俳句を生成するため，G に与えるデータによって，文法や俳句の基本を学習し，D に与えるデータによって，お〜いお茶らしい俳句を生成することができるのではないかという仮説を立て，検証を行う．

本研究で用いた 2 つの NN ハイパーパラメータを表 1 に示す．ハイパーパラメータは，処理時間を考慮しつつ，決定した．なお，最適化手法は Adam[5]を用いた．

† 京都工芸繊維大学，Kyoto Institute of Technology

表 1. NN ハイパーパラメータ

G/D	パラメータの種類	パラメータの値
G	Word Embedding 次元	64
	隠れ層の次元	64
	事前学習のエポック数	64
D	Word Embedding 次元	64
	事前学習のエポック数	64

3.2 データセット

事前学習のための俳句データは，伊藤園お〜いお茶新俳句大賞の受賞作品から 15,805 句，現代俳句協会の現代俳句データベース[6]から 17,554 句を取得した．

取得した俳句を，系列長の統一・句切れの明示のために，形態素解析エンジン MeCab[1] を用いて品詞ごとに分割し，データの整形を行う．

整形のために，特殊な符号「\t」，「<EOH>」(End Of Haiku)，「<p>」を用意する．俳句を構成する単語ごとに半角スペースで区切り，五音・七音・五音の句切れには，「\t」を挿入する．俳句の最後には，「<EOH>」を挿入し，その後，合計 20 単語になるまで「<p>」という符号でパディングを行う．データの整形の具体例として，松尾芭蕉の俳句「古池や 蛙飛び込む 水の音」を整形した例を図 2 に示す．

古池 や ¥t 蛙 飛び 込む ¥t 水 の 音 <EOH> <p> <p> <p> <p> <p> <p> <p> <p> <p>

図 2. 俳句データの整形例

MeCab によって分割された単語は，語彙データベースとして，単語ごとに ID を割り当て，登録する．語彙データベースの一部を図 3 に示す．図 3 中の左側には登録された単語を，右側にはそれに対応する ID が書かれている．

ふらふら 0
と 1
¥t 2
死に 3
ゐ 4
し 5
風 6
が 7
起き 8
上る 9
… …

図 3. 語彙データベースの一部

次に，お〜いお茶の俳句，現代俳句データベースの俳句の各語彙数及び，共通語彙数を表 2 に示す．

表 2. お〜いお茶と現代俳句データベースの俳句の語彙数

種類	語彙数
お〜いお茶	15,020
現代俳句データベース	17,232
合計	24,245
共通	8,008

表 2 より，お〜いお茶と現代俳句データベースは，半数近くの語彙を共有していることがわかる．この共通語彙が多いことを利用して，共通語彙のみを使用している俳句に統一して評価を行うこととする．

3.3 評価方法

実験条件として，下記の 4 種類の俳句を用意した．

・お〜いお茶新俳句大賞の受賞作品
・現代俳句データベースの俳句
・提案手法により生成された俳句
・従来手法により生成された俳句

従来手法では，現代俳句データベースとお〜いお茶の俳句，合計 33,359 句の俳句を G，D 両方に事前学習のためのデータとして与える．

提案手法では，現代俳句データベースの俳句を G のデータとして与え，G によって生成された俳句とお〜いお茶の俳句を D のデータとして与える．

学習後，従来手法・提案手法共に，ランダムに先頭の単語を決定し，10,000 句の俳句を生成した．生成された俳句は，単純に訓練データから学習したものであるため，うまく俳句の形式を学習できず，五・七・五の形式に従わないものも含まれている．そこで，生成された 10,000 句からさらに，形態素解析ツール MeCab を用い，五・七・五の形式に一致するもののみを抽出した．また，形式の統一のため，お〜いお茶，現代俳句データベースの俳句についても字余り [1]，字足らず [2] の俳句は除外した．また，提案手法は，G の事前学習に，お〜いお茶の俳句データのみを与えるため，出力される俳句も G の語彙のみを使用する傾向にある．そこで，使用される語彙を統一するために，全種類において，共通の語彙のみが使用されている俳句に限定した．最終的に抽出された俳句の例を各方法 3 句ずつ表 3 に示す．

表 3. 各手法における俳句の例

種類	俳句
お〜いお茶	落ち込むな 負けても汗が 金メダル
	忘れない 寡黙な人の 優しい目
	雨の中 一期一会の 梅見かな
現代俳句データベース	衣更 鏡の中の 日曜日
	コスモスよ ぼくはねむたく なりました
	紫陽花は 色を旅して おりました
提案手法	かげろうの 雪が減りたり 姉のなみ
	背白くず 母の背中が かくれんぼ
	あの空の 空気のように 僕の道
従来手法	その夜に 列の向こうに お度上がり
	年風呂と 色まで父は 寒さかな
	どんぐりの 鍋は幼い 夏休み

これらの 4 種類の俳句を評価するために，俳句や文学を専攻していない大学生 8 名を対象としたアンケート調査を行った．4 種類の俳句からそれぞれランダムで 10 句ずつ取得し，合計 40 句の俳句に対して下記 4 項目について，5 段階の SD 法による調査を行った．

・意味が通るか，通らないか
・ゆるんだ感じがするか，緊張感のある感じがするか

[1] 日本語形態素解析システム, http://taku910.github.io/mecab/

[2] 五七五の定められた字数より多いこと
[3] 五七五の定められた字数より少ないこと

・新しい感じがするか，古い感じがするか
・好きか，好きではないか

「意味が通るか」の質問は，各俳句がそもそもどの程度意味が通じるのかを調べるための質問である．「ゆるんだ感じがするか」，「新しい感じがするか」，「好きか」については，現代俳句データベースが現代的な俳句であり，お〜いお茶の俳句はお〜いお茶らしい俳句であることに着目し，一般の大学生がこの 2 種類の俳句を見たときに，差が生まれそうな質問として設定した．

4．結果

得られた調査結果を質問項目毎に ANOVA を用いて，統計的分析を行った．各質問項目における，アンケート回答者の回答結果を図 4 に示す．

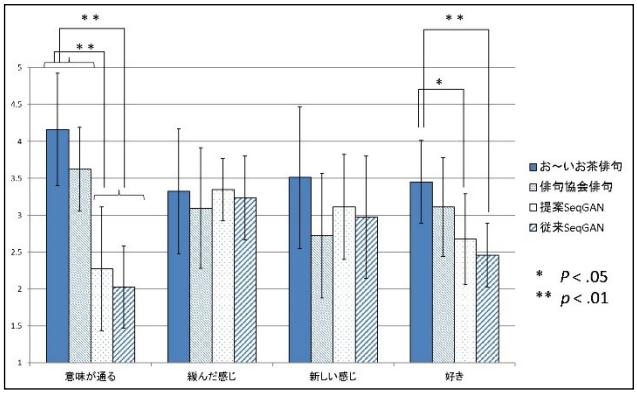

図 4. 評価実験の結果

「意味が通るか」の項目において，お〜いお茶と提案手法の間に有意な差が見られ(df = 15, t = 5.71, p < .01)，お〜いお茶と従来手法の間にも有意な差が見られた(df = 15, t = 6..29, p < .01)．また，「好きか」の項目についても，お〜いお茶と提案手法の間に有意な差が見られ(df = 15, t = 2.84, p < .05)，お〜いお茶と従来手法の間にも有意な差が見られた(df = 15, t = 3.54, p < .01)．「緩んだ感じか」の項目においては，どの手法間においても，有意な差は見られなかった．また，「新しい感じか」の項目についても，どの手法間においても有意な差は見られなかった．

5．考察

図 4 より，提案手法・従来手法ともに意味が通らない俳句が多く，表 3 に挙げた俳句の例を見ても，機械により生成された俳句は意味が通らないものとなっている．これは，今回用いた生成器は，2 つとも NN を用いた言語モデルである為，単語間のつながりにしか注目していない．また，今回用いた俳句データは，お〜いお茶の俳句，現代俳句データベースの俳句を合わせても 33,359 句しかなく，40,432,211 句の俳句データベースを使用している Wu らの研究と比較するとかなり少ない．俳句生成の質を向上させるには，俳句のデータ量を増やす必要があるだろう．

図 4 の「好き」の項目より，アンケート回答者は，従来手法で生成した俳句と提案手法で生成した俳句よりも，お〜いお茶の俳句を，好むことが示された．今回用いたお〜いお茶の俳句は，受賞作品であるから，多くの人が好む良い俳句である．従って，アンケート回答者の多くが好きだと判断することは想定通りである．ただ，「好きか」の質問において，お〜いお茶の俳句に対して，従来手法は 5%水準で有意差があったが，提案手法は 1%水準で有意差があったことから，提案手法は，僅かではあるが，従来手法よりもお〜いお茶らしい俳句が生成できている可能性が示唆された．

6．おわりに

本研究では，一般人がより好む俳句を生成することを目的とし，生成器 G と識別器 D に同じ俳句データを与えて事前学習を行う従来手法(SeqGAN)と，G に現代的な俳句，D に一般人に好まれやすいお〜いお茶新俳句大賞の受賞作品の俳句を与えて事前学習を行う提案手法の 2 つの手法を用いて俳句の自動生成を行った．提案手法は，従来手法と比較して，一般の大学生が好む俳句を生成できる可能性を示せたが，意味の通らない俳句も依然として多い．意味が通らない俳句が多く出力されている原因の 1 つとして，データセットの総数が少ないことが挙げられる．今後はさらに多くの俳句データを収集するのと同時に，単語の意味をある程度扱うことができる word2vec などの自然言語処理技術を導入することを検討している．

参考文献

[1]「伊藤園. お〜いお茶新俳句大賞」，<https://www.itoen.co.jp/new-haiku/about/history/index.html> 2017 年 7 月 27 日アクセス．

[2] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. SeqGAN: Sequence generative adversarial nets with policy gradient. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17),* pp. 2852-2858, 2017.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation,* 9(8) pp. 1735-1780, 1997.

[4] Xianchao Wu, Momo Klyen, Kazushige Ito, and Zhan Chen. Haiku generation using deep neural networks. 言語処理学会 第 23 回年次大会 発表論文集，pp. 1133-1136, 2017.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[6]「現代俳句協会. 現代俳句データベース」，< http://www.haiku-data.jp/> 2017 年 7 月 27 日アクセス．