Vol.2021-IFAT-143 No.2
2021/7/30

Transformerによる言語モデルを用いた俳句生成とその評価

平田 航大[1,a)]　横山 想一郎[2]　山下 倫央[2]　川村 秀憲[2]

概要：本稿では，人工知能技術の芸術分野の応用として俳句に着目し，Transformer をベースとした言語モデルである GPT-2 を用いて俳句生成を行う．過去の俳句作品や青空文庫で学習を行い，生成された俳句に対して，季語数，音数，切れ字数に制約を持つ有季定型句の制約を評価指標として用いる．また，品詞の並びが適切である，俳句の意味が通じる，といった人間による主観評価も行う．これらの評価指標を用いて，俳句生成における言語モデルの有効性を検証する．

1. はじめに

芸術分野における創作活動は，人が生きていくために直接的に必要なことでないが，知的好奇心や創作意欲は人間が持っている特殊な能力であり，人間だけが行う行動である．人はなぜ芸術を創作するのか．また，それを人工知能に行わせようという試みにどのような意味があるのか．これらの問いは哲学的で明確な答えを見つけることは容易ではないが，人工知能を使って何かを創作することに強い興味が抱かれている．「創作する」こと，もしくは「創作することに興味を持つこと」は知能の本質に深くかかわっていると考えられる．人工知能で芸術作品を生成するという取り組みは，知能とは何か，人とは何かを考えるときの，大事な糸口を含んでいる [1].

そのため，本稿では，人工知能技術の芸術分野の応用として，俳句生成を取り上げる．文章によって表現される芸術のうち，日本で古くから親しまれているものの一つに俳句がある．俳句には一定の制約があり，音数が５・７・５の１７音で構成されること，「季語」と呼ばれる四季を表す単語を含むことが挙げられる．小説などの他の文学作品に比べ，俳句は日本人固有の感性や価値観を反映した文学であるという点，音数などの量的制約が強い点，古くから親しまれているために作品の絶対数が多い点などが特徴的である．これらの特徴の中で，俳句には明確な制約があることや作品の絶対数が多いといった部分は，特に人工知能による生成を考えた時には生成作品の評価がしやすいことや学習データが確保しやすいことが利点となる．

近年，深層学習の発展により人工知能による文章生成が様々なタスクに応用されている．特に 2017 年に発表されたモデルである Transformer[2] が提案されて以降，Transformer をベースとした事前学習モデルの開発が進められてきた．Transformer をベースとした代表的な事前学習モデルとしては BERT[3] が挙げられる．BERT は 2018 年に Google から発表され，Transformer を何層も重ねたモデルアーキテクチャを含み，大規模な事前学習を行うことで，データが少量の場合にも高い性能を発揮することを示したモデルである．BERT の発表後，2019 年に OpenAI から GPT-2[4] が発表された．GPT-2 は自己回帰モデルの一種であり，文章生成タスクにおいて BERT などの Masked 言語モデルに比べて高い性能を発揮するという特徴がある．

これまでにも俳句の自動生成・自動評価に関する研究は行われてきている．近年は Wu らの研究 [5] のように深層学習モデルを利用することが多くなってきている．Transformer ベースのモデルを用いて俳句生成を行った研究はいままでに行われていない．また，他の芸術作品に対しては行われているような人間による主観評価が俳句に対して提要された研究は行われていない．

そこで本稿では Transformer ベースのモデルである GPT-2 を用いた俳句生成を行い，事前学習として青空文庫のデータ [6] を用いる場合と俳句データのみで学習した場合のモデルが生成した俳句に対して，俳句としての制約を満たす割合を比較する．さらに，生成俳句に対して，品詞の並びが適切である，俳句の意味が通じる，といった人間による主観評価を行うことで，モデルの有効性を検証する．

[1] 北海道大学 大学院情報科学院
[2] 北海道大学 大学院情報科学研究院
[a)] hiratako@ist.hokudai.ac.jp

2. 関連研究

既存の俳句生成のアプローチについて先行研究を挙げる．

先行研究において，俳句生成においては二つの方法が挙げられる．テンプレートを活用して俳句を生成する方法と，生成モデルを使って俳句を生成する方法である．テンプレートを活用する方法として，土佐ら [7] は俳句に用いられるフレーズをあらかじめデータベース化しておき，ユーザが選択したフレーズと関連性が深いフレーズを組み合わせることで俳句を生成する手法を提案している．あらかじめ決められたフレーズを用いるなどのテンプレート型手法は日本語としての破綻が少なくなるなどの利点がある．一方でテンプレートから外れた俳句は生成できない，テンプレートの作成・拡張にコストがかかるなどの欠点がある．

生成モデルを試用する方法として，Wu ら [5] はニューラルネットワーク，LSTM，seqGAN などの深層学習モデルを用いた俳句生成を行っている．各モデルの評価について，Wu らは言語モデルのパープレキシティを用いている．太田ら [8] は深層学習モデルを用い，入力した単語列と同じ季節の季語を選択するようなモデルを構築した．また，生成した俳句の評価として拍数を用いている．筆者ら [9] は LSTM を用いた俳句生成を行った．生成した俳句の評価としては音数，季語数，切れ字数などの俳句のルールを用いている．生成モデルを用いた俳句生成では多様な俳句の生成が期待できる一方，日本語として正しくない文章が出力される可能性があるなどの欠点がある．

本研究は生成モデルである GPT-2 を用いた俳句生成を行い，生成俳句の評価として，有季定型句の制約を用いた機械的評価と人間の主観による評価を行う．

3. 俳句の概要

本節では，俳句の満たす基本的な条件と GPT-2 による生成の対象となる有季定型句について述べる．

3.1 俳句の基本条件

俳句は日本における伝統的な文学であり，音数が 17 音で構成されることが制約の一つとして存在することから世界最小の詩とされている．公益社団法人日本伝統俳句協会の定義[*1] によると，俳句の定義は以下の二つであるとされている．

- 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の音数に分けたそれぞれの部分は上五，中七，下五と呼ばれる．また，このほかにも俳句特有の技法として「切れ」やというものがある．切れは俳句中に「や」「かな」「けり」などの切れ字と呼ばれる語を用いることにより，前の単語を詠嘆したり，場面を切れ字の前後で切り替えるなどの様々な役割を果たす．

[*1] 公益社団法人日本伝統俳句協会 俳句入門講座-1, https://haiku.jp/tsukuru/2430/

3.2 有季定型句の制約条件

本研究では，まず，有季定型句という季語を含み俳句や五音・七音・五音の並びになっている俳句を生成・評価の対象とする．人間が詠んだ俳句には，有季定型句ではない自由律俳句にも優れた句であると評価された句は多くある．しかし，人間が自由律俳句を詠むときには，俳句には有季定型句という型があることをきちんと踏まえたうえで，あえて型から外すという選択を取っていることが多い．そのため，まずは基本に忠実な有季定型句を生成することを目指す．

有季定型句は以下のような制約条件を満たしている俳句である．

3.2.1 音数が 17 音

俳句が 17 音から構成されている．

3.2.2 句またがりでない

上五，中七，下五の複数にわたって単語がまたがっているとき句またがりであるという．句またがりは効果的に使用すれば句に独特の効果をもたらすことも多い．ただし，本稿では，句またがりは応用的な技法と捉えて，句またがりでないことを条件の一つとする．

3.2.3 季語を 1 つ含む

俳句には季節を表す単語である季語が一句に対し一つ含まれることが一般的である．例えば有名な松尾芭蕉の句である「古池や蛙飛び込む水の音」という俳句には「蛙」という季語が含まれる．

3.2.4 切れ字数が一つ以下である

切れ字は俳句特有の技法で，詠嘆や場面の切り替えなどの多彩な役割がある．「古池や蛙飛び込む水の音」という俳句には「や」という切れ字が使われており，切れ字の前後で場面が切り替えるなどの効果があるとされる．切れ字は一句の中に二つ以上切れ字を使うことは高度な表現法であるため，本稿では，切れ字数が一つ以下であることを条件の一つとする．なお，上五，中七，下五の末尾が以下の条件に当てはまる場合にその単語を切れ字と判定する．

- 助詞の「や」
- 助詞「か」と助詞の「な」の連続
- 助動詞の「けり」

4. 実験

本章では俳句データを GPT-2 に学習させた際に，前述した俳句の条件を満たす俳句の割合，言語モデルのパープレキシティ，過学習が起きていないかという点を比較・検証する．青空文庫データでの事前学習後，俳句データでファインチューニングを行うモデルと，俳句データのみで学習

表 1 青空文庫の作品データによる事前学習の設定

学習パラメータ	設定値
モデルアーキテクチャ	GPT-2
モデルのパラメータ数	約 1 億 1700 万
学習データ	12,978 作品
バッチ当たりのトークン長	512
バッチサイズ	16
学習エポック数	20

表 2 俳句データによる学習の設定

学習パラメータ	設定値
モデルアーキテクチャ	GPT-2
モデルのパラメータ数	約 1 億 1700 万
学習データ	407,439 句
バッチ当たりのトークン長	可変
バッチサイズ	512
学習エポック数	30

したモデルの 2 パターンのモデルを用意し，比較対象として訓練データを加えた計 3 パターンに対して比較を行う．各モデルの実装は PyTorch[*2]を用いて行った．

4.1 実験設定

俳句データを用いて深層学習を用いた言語モデル GPT-2 を訓練し，得られた言語モデルにより生成される文章のうち，有季定型句の定義を満たす文章の割合を検証する．

GPT-2 をはじめとする言語モデルを小規模な学習コーパスを用いて訓練する際には，ウィキペディアの記事などの大規模コーパスを事前学習した後に，本来の学習コーパスを用いてファインチューニングを行う手法が高い性能を示すことが知られている．本稿では事前学習の対象のコーパスとして，著作権フリーの文学作品を集めた青空文庫の全作品データを用いる．事前学習の後に俳句データを用いた訓練したモデルと，事前学習を行わず俳句データのみを訓練したモデルの出力を比較する．本稿では，青空文庫を事前学習した後に俳句データにより訓練したモデルを「青空文庫＋俳句」と表記し，俳句データのみを用いて訓練したモデルを「俳句のみ」と表記する．また，GPT-2 を提案する論文では，埋め込み表現の次元数と中間層の数を変えることで学習パラメータ数を調整した，4 種類の規模のモデルが提案されている．本稿では約 1 億 1700 万のパラメータを持つ，最も小規模なモデルを用いる．

青空文庫データからは作品ごとに本文を抽出し，脚注や読み仮名などの削除の前処理を施した．12,978 作品を訓練データとし，1,622 作品を検証・テストデータとした．俳句の学習にはインターネットで公開されている約 50 万句の俳句を用いる．この中には，小林一茶，高浜虚子，松尾芭蕉，正岡子規による俳句が含まれる．これを訓練データ，検証データ，テストデータに 8:1:1 の割合で分割し，検証・テストデータから訓練データに含まれる俳句との最小編集距離が 5 以下のものを削除した．得られた俳句の訓練，検証，テストデータはそれぞれ 407,439 句，48,323 句，48,306 句である．文章のトークン化は GPT-2 の提案論文と同様に Byte-level Byte Pair Encoding により行う．単語ごとの分かち書きに形態素解析器 MeCab および辞書 UniDic を用い，語彙数を 32,000 として青空文庫の訓練データによる語彙の学習を行った．

[*2] https://pytorch.org/

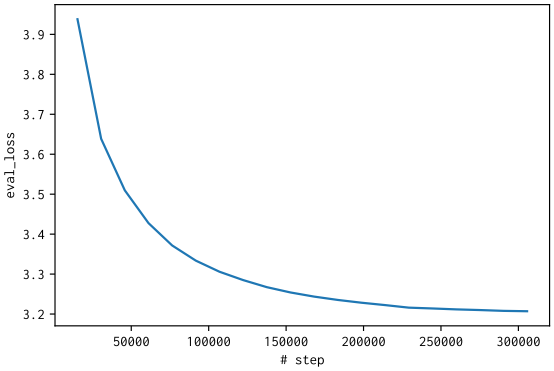

図 1 事前学習の学習過程

青空文庫による事前学習および俳句データによる学習の実験設定を表 1, 2 にそれぞれ示す．青空文庫による事前学習の際は，コーパスを 512 トークンごとに分割し，1 ステップの学習では 16 のデータを用いてパラメータを更新する．俳句データによる学習では，1 ステップの学習に 512 句を用いる．バッチのトークン長は，バッチに含まれる俳句をトークン列化した際の最長サイズにより決定される．学習エポック数は，学習データに対する損失が収束する値を設定した．青空文庫による事前学習は 1 試行のみ実施し，得られたモデルに対するファインチューニングは乱数のシード値を変更して 5 試行を実施した．また，俳句データのみを用いたモデルの学習についても 5 試行を実施した．

4.2 実験結果

4.2.1 損失およびパープレキシティによる評価

青空文庫データの事前学習の学習過程について，横軸に学習ステップ数，縦軸に検証データに対する損失をプロットしたものを図 1 に示す．検証データに対する損失が 3.2 付近に収束したことがわかる．このとき，青空文庫の検証データに対するパープレキシティは 24.5 であった．

次に，俳句データを用いて訓練したときの検証データに対する損失の推移を示す．青空文庫データを事前学習の後に俳句データを用いてファインチューニングした際の損失の推移を図 3 に，俳句データのみを用いて学習した際の推移を図 2 にそれぞれ示す．ファインチューニングの際には，検証データに対する損失が減少した後で増加に転じていることから，この後の実験では検証データに対する損失

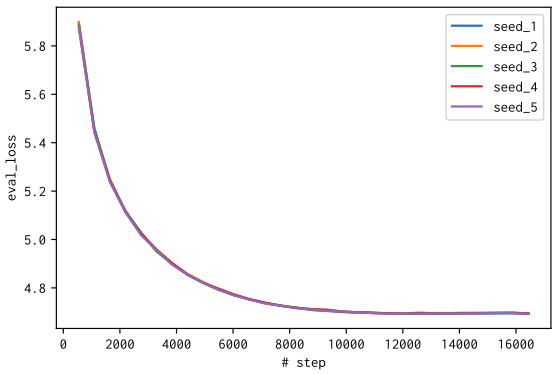

図 2 俳句データのみで学習した場合の学習過程

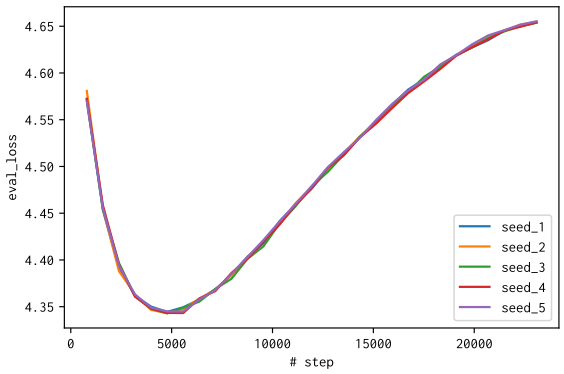

図 3 ファインチューニングの学習過程

表 3 言語モデルの俳句テストデータに対するパープレキシティ

モデル名	1	2	3	4	5
青空文庫＋俳句	76.9	76.9	76.9	76.8	76.9
俳句のみ	94.0	94.1	93.7	93.9	93.9

が最も小さいステップ数におけるモデルを使用する．

表 3 に俳句テストデータに対する両モデルのパープレキシティを示す．青空文庫による事前学習の後に俳句データによる訓練を行ったモデルのパープレキシティの平均は 93.9，俳句のみで学習を行ったモデルは 76.9 となっており，対応なし t 検定の結果有意差 5 %で有意な差が認められた．したがって，青空文庫による事前学習を行うことで，より汎化性能の高い言語モデルを得られたことがわかる．

4.2.2 有季定型句の制約条件を満たす文字列出力の割合による評価

青空文庫データでの事前学習を行ったモデルと行わないモデルから出力文字列を 1000 個ずつサンプリングし，得られた文字列が第 3 章で述べた有季定型句の 4 つの制約条件を満たす割合を調査した．更に，未知語を含む文字列および訓練データとして与えた俳句と類似した文字列は，本稿で検証する俳句の言語モデルの出力として望ましくないことから，こうした文字列を除外するための制約条件を設定する．文字列がこうした制約条件を満たすことは，次の方法により機械的に判定した．

音数（音数が 17 音であること） MeCab[10] を使用して文字列を形態素に分解したのち，MeCab 用辞書である UniDic[11] から各形態素の音数を取得し，累計の音数が 17 音になること．

句またがり（句またがりの俳句でないこと） 文字列を形態素解析したときに 5 音目もしくは 1 2 音目をまたぐ形態素が存在しないこと．

未知語（未知語を含まないこと） 文字列を形態素解析した際に，未知語と判定される形態素を含まないこと．

季語数（季語を 1 つのみ含むこと） インターネット上の季語データベース[*3]から収集した 8,642 語のうち，文字列の中に 1 語のみが含まれていること．

切れ字数（切れ字を含む数が 1 追加であること） 文字列を形態素解析したときに切れ字と判定される単語が 1 回以下しか出現しないこと．

非類似句（学習データのなかに類似句が含まれないこと） 訓練データとの最小編集距離が 5 よりも大きい文字列であること．

これを訓練データにおける各種条件の割合と比較することで，モデルが訓練データの特徴を学習できているかどうかを調査した．表 4 に実験の結果を示す．表の「全条件」という項目には 6 条件すべてを満たす文字列の割合を示している．なお，訓練データに関しては，全条件の判定の際に類似句に関する条件を除外している．

表 4 の結果より，本稿で設定する制約条件に対しては，俳句のみで学習したモデルの方が，未知語を含まないこと以外のすべての条件で訓練データにより近い割合で条件を満たす俳句を生成できていることが分かった．ここから，訓練データに含まれる俳句の特徴を捉えるという観点で，俳句データのみで学習を行ったモデルの方が良い性能を示すといえる．これは，散文に分類される作品が多い青空文庫作品での事前学習が，韻文である俳句の特徴を捉える際の妨げになっているためと考えられる．また，非類似句の割合は両モデルとも 9 割以上となっており，訓練データとまったく同じような文字列を出力するようなモデルではないことがわかる．

4.2.3 出力された文字列に対する主観評価

次に，事前学習を行ったモデルと行わないモデルが生成する文字列について，俳句の観点からの主観評価による検証を行う．主観評価の方法は [12][13] などにならい，文字列を見た被験者がいくつかの質問に対して回答する形式で行った．[12] などで採用されている質問項目を俳句に適用し，以下の 3 点について，満たす，満たさない，判断が難しいの 3 つの選択肢から被験者の回答を得た．

[*3] http://www.haiku-data.jp/kigo.php, 7 月 7 日閲覧

表 4 制約条件を満たす俳句の割合

対象データ	音数	句またがり	未知語	季語数	切れ字数	非類似句	全条件
「青空文庫＋俳句」からサンプリングした文字列	24.4%	46.9%	95.9%	61.7%	95.4%	97.8%	9.95%
「俳句のみ」からサンプリングした文字列	29.1%	54.5%	94.9%	62.0%	95.4%	98.2%	13.1%
訓練データに含まれる文字列	32.0%	58.7%	96.1%	71.0%	96.2%	0%	19.5%

表 5 主観評価の結果の例

評価対象	Fluent	Meaningful	Poetic
後の月闇は静かに座禅草	3	3	3
山垣の奥があかるくて豆の花	3	3	1
霜来り護符の白さに野良の富士	3	1	3
ひとり住むは祭笛でありにけり	1	1	1

表 6 主観評価のアンケート結果

モデル名	Fluent	Meaningful	Poetic
青空文庫＋俳句	96.5%(0%)	58.0%(7.5%)	66.0%(0%)
俳句のみ	91.0%(0%)	34.0%(5.0%)	75.0%(0%)
訓練データ	93.0%(0.5%)	59.5%(7.5%)	78.5%(0%)

- Fluent
  品詞の並びが日本語として正しい．また，未知語を含まない．
- Meaningful
  句として意味が通る．情景が矛盾なく想像できる．
- Poetic
  人間が見た時に 5・7・5 のリズムで読むことができる．また，季語数が 1 つ，切れ字数が 1 つ以下という条件を満たす．

実施した主観評価の結果の一部を表 5 に示す．表に示された番号は，条件を満たす（3），判断が難しい（2），満たさない（1）に対応する．

Fluent に関しては文章としての意味は考慮せず，品詞の並びや単語単位で違和感を与える日本語がないかという観点で評価を行った．例えば表 5 の例を参照すると，「ひとり住むは祭笛でありにけり」という俳句以外は品詞の並び，単語として違和感を与える日本語はないと判断したため，Fluent を満たすと判定した．

Meaningful に関しては句として意味が通るか，情景が矛盾なく想像できるかという観点で評価を行った．例えば表 5 の例を参照すると，「後の月闇は静かに座禅草」と「山垣の奥があかるくて豆の花」という俳句は情景が矛盾なく浮かぶと判断したために Meaningful を満たすと判定した．一方で「霜来り護符の白さに野良の富士」という俳句は，例えば「野良の富士」という部分の情景が浮かびづらいと判断し，Meaningful を満たさないと判定した．

Poetic に関しては季語数，切れ字数，音数などの有季定型句の制約条件を満たしているかという観点で評価した．Poetic は第 4 章では MeCab などを用いて機械的に判定した俳句の諸条件について，人間の判定とどの程度乖離があるかを調査するために設けた項目である．

主観評価の集計結果を表 6 に示す．表内の数値は各条件を満たす文字列の割合を示す．カッコ内の数字は判断が難しいと回答された文字列の割合を示す．事前学習を行ったモデルの方がより訓練データに近い割合で Fluent と Meaningful を満たす文字列を生成したことが分かった．Poetic に関しては前節での結果と同じく，俳句データのみで学習したモデルの方が訓練データに近い数値を残すことが分かった．ただし前節での制約条件の全条件を満たす文字列の割合と本節での Poetic を満たす文字列の割合を比較すると，本節の主観評価で有季定型句の制約条件を満たすと判断された文字列の割合がより大きいことがわかる．これは形態素解析器と辞書を用いて機械的に音数などのルールを判定する場合，読み方が複数ある単語や特殊な読みをする単語の正確な解析が難しいことが一つの要因であると考えられる．例えば「夜」という単語は俳句の上では「よる」と読む場合と「よ」と読む場合が文脈によって異なるため，判定が難しい．これに対処するには，読み方が複数ある場合，すべてのパターンを保持しておき，一つでも 17 音になる読みがあれば条件を満たすと判定するなどの工夫が考えられる．

5. おわりに

本稿では Transformer ベースのモデルである GPT-2 を用いた俳句生成を行った．青空文庫データでの事前学習を行ったモデルと俳句データのみで学習したモデルを比較し，俳句生成における事前学習の効果について検証した．散文である青空文庫を事前学習に用いることで，俳句の各条件を満たす俳句の割合は減少するものの，日本語としての破綻がなく意味の通る俳句を生成できる割合を増加させることを確かめることができた．

謝辞 本研究においては有限会社マルコボ. コムの方々から，俳句雑誌の編集者の目線から数々の指摘をいただきました．ここに深く感謝の意を表します．

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] 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).
[3] 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).

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

[5] Wu, X., Klyen, M., Ito, K. and Chen, Z.: Haiku generation using deep neural networks, *Proceedings of the Twenty-third Annual Meeting of the Association for Natural Language Processing,(Tsukuba, Japan March 2017)* (2017).

[6] 青空文庫：青空文庫，`https://github.com/aozorabunko/aozorabunko` (accessed 2021-01-18).

[7] Tosa, N., Obara, H. and Minoh, M.: Hitch haiku: An interactive supporting system for composing haiku poem, *International Conference on Entertainment Computing*, Springer, pp. 209–216 (2008).

[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] 米田航紀，横山想一郎，山下倫央，川村秀憲：LSTM を用いた俳句自動生成器の開発，人工知能学会全国大会論文集 第 32 回全国大会 (2018)，一般社団法人 人工知能学会，pp. 1B2OS11b01–1B2OS11b01 (2018).

[10] 工藤　拓，山本　薫，松本裕治：Conditional Random Fields を用いた日本語形態素解析, 情報処理学会研究報告. NL, 自然言語処理研究会報告, Vol. 161, pp. 89–96（オンライン），入手先 ⟨https://ci.nii.ac.jp/naid/110002911717/⟩ (2004).

[11] 照晃　岡，Teruaki, O.：CRF 素性テンプレートの見直しによるモデルサイズを軽量化した解析用 UniDic : unidic-cwj-2.2.0 と unidic-csj-2.2.0，言語資源活用ワークショップ発表論文集 = Proceedings of Language Resources Workshop， No. 2, pp. 144–153（オンライン），DOI: info:doi/10.15084/00001515 (2017).

[12] Yan, R.: i, Poet: Automatic Poetry Composition through Recurrent Neural Networks with Iterative Polishing Schema, Technical report.

[13] Zhang, X. and Lapata, M.: Chinese Poetry Generation with Recurrent Neural Networks, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Association for Computational Linguistics, pp. 670–680 (online), DOI: 10.3115/v1/D14-1074 (2014).