RLHF を用いた「面白い」短歌の自動生成の試み

羽根田賢和 [1]　浦川通 [2]　田口雄哉 [2]　田森秀明 [2]　坂口慶祐 [1,3]
[1] 東北大学　[2] 株式会社朝日新聞社　[3] 理化学研究所
haneda.kento.t6@dc.tohoku.ac.jp　keisuke.sakaguchi@tohoku.ac.jp
{urakawa-t,taguchi-y2,tamori-h}@asahi.com

概要

自然言語処理技術の発展は小説などの文学作品の自動生成を現実的なものとしている．一方で短歌におけるモーラ数のような制約を満たしつつ，高い芸術性を有する文学作品を生成することは，既存の言語モデルでは容易ではない．本研究ではこの課題に対する解決策として RLHF による学習を用いた短歌の自動生成モデルを提案する．人間の短歌に対する評価を反映した報酬モデルの学習とこれを用いた生成モデルの強化学習により，短歌らしい系列長を有し，かつ人間に「面白い」と判断されやすいような短歌の自動生成が可能となった．

1　はじめに

自然言語処理技術の発展に伴い，大規模言語モデルは対話生成をはじめとした多種多様なテキストの生成が可能となっている．ChatGPT に代表されるこれらの生成モデルの利用法として，小説や詩といった文学作品の自動生成が挙げられる．Belouadi ら [1] は，ByGPT5 というモデルを用いて，韻や拍子といった特有の制約を満たす高品質な四行詩の自動生成を達成している．また，Takeishi ら [2] の研究では日本の文学作品の自動生成への取り組みとして，Transformer を用いた和歌の自動生成が試みられている．このような文学作品の生成モデルは創作活動の手助けや鑑賞の多様化に寄与すると期待される．

日本における文学作品の主要な形態の一種として短歌が挙げられる．短歌は 5-7-5-7-7 の 5 句 31 音からなる定型詩であり，古来より広く親しまれている．一方で短歌は定型詩である関係上，成立にはモーラ数という制約を満たす必要があり，この性質が短歌の自動生成における課題 [3] となっている．ChatGPT をはじめとした大規模言語モデルは文字数のカウントといったタスクを苦手とすることが報告されており [4, 5]，定型詩としての制約を満たす短歌を生成させることは容易ではない．

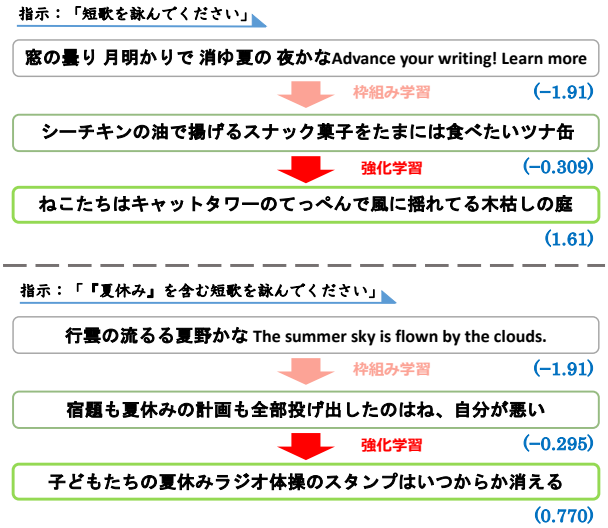

図 1　学習の経過と生成された短歌の例．カッコ内の数値は学習済み報酬モデルによるスコア．

また「面白い」短歌を生成することが困難であることも大きな課題 [6] である．一般に文学作品は良し悪しの定量的な評価が難しい．俳句における季語のように，内容への制約が存在しない短歌では評価の自由度が高く，この問題が顕著である．文学作品として生成された短歌には，独創性や写実性など鑑賞対象として豊かな表現が求められる．しかし既存の言語モデルの多くには雑談のような自然な会話を目指す学習がなされており，モーラ数の制約を満たしたとしても，芸術的な短歌は生成されにくい．

本研究では，これらの課題に対する解決策として “Reinforcement Learning from Human Feedback”（RLHF，人間からのフィードバックを用いた強化学習）を用いた短歌生成モデルを提案する．一連の学習を通じて短歌らしい系列長をモデルに学習させるとともに，人間の好みを反映させることにより，定量的に判定することが難しい短歌の「良し悪し」を人間の好みに沿うか否かという軸で表現することが

可能となる．この手法により，使用者に「良い」と感じさせやすい，言い換えれば「面白い」短歌が生成されることが期待できる．またこのようなモデルの学習を通じて，短歌鑑賞の多様化や作歌の補助への貢献を目指すのみならず，人の心を動かす短歌の本質的な要素の探求を試みる．

2 RLHF

RLHF は人間の価値基準を報酬として利用する強化学習の一手法である [7, 8]．モデルの生成物に対して人間の価値基準におけるスコア付けを行い，それをモデルにフィードバックすることで，より高いスコアがつく生成を行うようモデルを学習させる．これにより言語モデルが人間にとって好ましい生成を学び，生成物の質が向上することが期待できる．RLHF は，報酬モデルの学習，枠組み学習，強化学習という 3 つのステップに分けられる．

報酬モデルの学習　RLHF の学習では，生成物に対し人間の基準により報酬を与えモデルにフィードバックするという過程が必要となる．一方で学習過程での生成全てに対し，直接人間が一貫した基準でスコアを付けることは現実的ではないため，自動的にスコア付けを行う報酬モデルを作成する．テキストとそれに対する人間の評価をモデルに与えて学習を行い，入力文に対して相対的なスコアを付けることを可能とする．この報酬モデルを擬似的なアノテータとして用いることで後述の強化学習を行う．

枠組み学習　短歌のように満たすべき制約を持つ場合は，事前学習のみが行われたベースモデルでの直接の生成が難しい．一例として図 1 に学習前のベースモデルでの生成を示す．生成された文には英語が含まれており，短歌としての枠組みを全く満たしていないことが確認できる．このような場合，生成の目的に合わせたモデルの教師あり学習により，後に行う強化学習の効率を高める必要がある．本研究ではこれを「枠組み学習」と呼称する．

強化学習　強化学習ではまず枠組み学習を行ったモデルにより生成を行い，生成文を報酬モデルに入力として与える．この出力として得られた報酬スコアを生成モデルにフィードバックし，スコアを最大化するような生成を目指す学習を PPO アルゴリズム [9] をもとに行う．PPO は方策ベースの強化学習手法であり，大幅な方策の更新を妨げること（clipping）で安定した学習を可能とする．

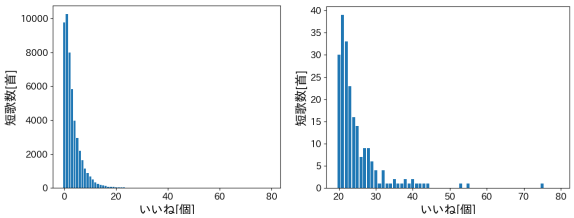

図 2　Utakata の各短歌へのいいね数と短歌の数．

以上が RLHF の具体的なプロセスである．本研究ではこの RLHF を短歌の生成に適用し，その有効性を検証する．具体的には (1) 学習済み報酬モデルによる短歌に対しての自動スコア付けの可能性 (2) 学習済み報酬モデルと枠組み学習済み生成モデルを用いた強化学習による短歌の自動生成の質の向上，の二点に関し実験と評価を行う．

3 実験

本研究では短歌の自動生成に RLHF を適用し，その有効性の検証を行った．また一連の実験に先立ち学習用のデータセットの作成を行った．

3.1 データセット

本研究では各モデルの学習に際し，多数の短歌のデータが必要となる．特に報酬モデルの学習には人手により評価されたデータが求められる．

そこで我々はこの条件を満たすデータセットとして，短歌専用の投稿サイトである Utakata[1)] を利用した．Utakata には 2023 年 12 月時点で 7 万首以上の短歌が投稿されており，十分なデータが確保できると判断した．特筆すべき点として Utakata には「いいね」機能が付いている点が挙げられる．「いいね」の数は不特定多数の鑑賞者による評価の結果であるため，これを各短歌への一般的な評価とみなし，モデルの学習に利用した．学習データとして計 68,996 首の短歌をクローリングした．

図 2 に示す通り，Utakata において各短歌の「いいね」の数には大きく偏りがある．ほとんどの短歌は「いいね」数が 0 ないしは一桁の値であり，10 人以上から評価を得た短歌は数千首に留まる．「いいね」数がさらに多い短歌となると該当する短歌の数は指数関数的に減少していく．一方で我々は十分に優れていると判断されうる短歌を学習に用いる必要があり，これはつまり多くの短歌は学習データとして適していないことを意味している．したがって今回

1）https://utakatanka.jp

表 1　報酬モデルによる *reward* スコア．スコアは絶対的な値ではなく相対的な値である．

	平均値	中央値	標準偏差
$\text{tanka}_{\uparrow}$	0.247	0.302	1.04
$\text{tanka}_{\downarrow}$	-0.639	-0.781	0.948

は，報酬モデルの学習では計 4,980 首を，枠組み学習と強化学習では報酬モデルの学習に使用したものを含む 11,299 首を利用した．使用した短歌の条件は各実験の項で述べる．

3.2　報酬モデルの学習

報酬モデルの学習にはベースのモデルとして，日本語で事前学習されており分類問題に強い rinna 社 RoBERTa[2] を用いた[3]．学習時には多くの人に評価された短歌 $\text{tanka}_{\uparrow}$ とそうでない短歌 $\text{tanka}_{\downarrow}$ のペアをモデルに与え，式 (1) に則り loss の計算を行った．$reward_{\uparrow}$，$reward_{\downarrow}$ はそれぞれ $\text{tanka}_{\uparrow}$，$\text{tanka}_{\downarrow}$ に対してモデルが出力したスコアを意味する．

$$\text{loss} = -\log\left(\frac{1}{1+e^{-(reward_{\uparrow}-reward_{\downarrow})}}\right) \tag{1}$$

学習データに関し $\text{tanka}_{\uparrow}$ として Utakata 上で「いいね」数が 10 以上の短歌を，$\text{tanka}_{\downarrow}$ として「いいね」数が 1 の短歌を用いた．「いいね」がついていない短歌に関しては学習時にノイズとなりうることを考慮し，本研究では用いないこととした．最終的に $\text{tanka}_{\uparrow}$ としての 2,490 首，$\text{tanka}_{\downarrow}$ としての 2,490 首をペアとして報酬モデルの学習を行った．

3.3　生成モデルの枠組み学習と強化学習

生成モデルの学習にはベースとするモデルとして，LINE が公開している japanese-large-lm-instruction-sft の 3.6B[4] を使用し，Utakata 上で「いいね」数が 5 以上であるような短歌，計 11,299 首を学習に用いた．モデルは「短歌を詠んでください」といった指示を入力として与えられた際に，出力として短歌を生成するように枠組み学習がなされた．

学習データの量をさらに多く確保し，また実際に利用する際に多様な入力を受け付けるために，学習データとした短歌をもとに指示文の自動生成を行った．具体的にはキーワード抽出ライブラリである YAKE! [10] を用い元々の短歌のテーマに関する生成を行わせる指示文や，下の句のみを与え上の句を生成させる指示文を作成した．詳細は付録 B に記す．各短歌に対してそれぞれ 8 種類の指示文の作成を試み，計 76,233 件のデータを学習に利用した．

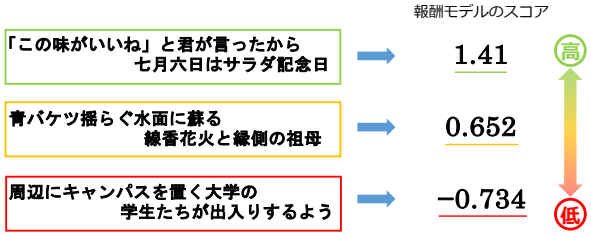

図 3　短歌へのスコア付けの例．上から順に歌人である俵万智の短歌 [11]，本論文著者である羽根田の短歌，Wikipedia より抽出した短歌の形式を満たす文字列．

以上のように訓練された報酬モデルと生成モデルを用いて強化学習を行った．学習データは枠組み学習で利用したものと同じものを利用した．強化学習には Hugging face の提供するライブラリである TRL[5] を使用し，PPO による学習を行った．

4　結果

4.1　報酬モデル

$\text{tanka}_{\uparrow}$，$\text{tanka}_{\downarrow}$ 各 100 首ずつのテストデータを用いたスコアの測定結果は表 1 のようになった．$\text{tanka}_{\uparrow}$ に対するスコアの平均値や中央値が $\text{tanka}_{\downarrow}$ に対するものを大きく上回っており，報酬モデルのスコアが人間による評価と同様の傾向を示していることが確認できる．また，$\text{tanka}_{\uparrow}$ と $\text{tanka}_{\downarrow}$ のペアそれぞれに対して，$reward_{\uparrow}$ が $reward_{\downarrow}$ を上回った割合は 77%であり，期待値が 50%であることを考慮すると，短歌の質の差をある程度判別できていると考えられる．

実際に短歌にスコアをつけた例が図 3 である．一つ目の例のように，社会的に評価されている短歌には高いスコアをつけている．一方で短歌であることを意図して詠まれているものの一般からの評価を受けていない短歌や，短歌の形式を偶然満たした文章へのスコアは低くなっていることが確認できる．

一方で明らかに短歌の形式を満たしていない文字列を入力した際，スコアが想定されるほど低くない場合があることが確認されている．例えば「今夜のメニューは焼肉」といった短文の入力に対しては，-0.0223 というスコアを出力し，これは $\text{tanka}_{\downarrow}$ へ

2）https://huggingface.co/rinna/japanese-roberta-base
3）モデルの有効性の検証のため東北大 BERT，早稲田大 RoBERTa での実験も行った．詳細は付録 A に記す．
4）https://huggingface.co/line-corporation/japanese-large-lm-3.6b-instruction-sft
5）https://github.com/huggingface/trl

表 2　枠組み学習による生成の変化．モーラ数は短歌の定型である 31 に近づく方が好ましい．

	モーラ数		スコア	
	平均値	中央値	平均値	中央値
枠組み学習前	92.1	73.0	-0.342	-0.096
枠組み学習後	**30.5**	**30.0**	-0.495	-0.832

表 3　強化学習による生成の変化．

	モーラ数		スコア	
	平均値	中央値	平均値	中央値
強化学習前	30.5	30.0	-0.495	-0.832
強化学習後	31.3	30.0	**0.724**	**0.879**

のスコアの平均値や中央値よりも高い値である．このことは，入力の長さがこの報酬モデルでのスコアに与える影響が小さいことを示唆している．

4.2　枠組み学習

枠組み学習による生成の変化は表 2 のようにまとめられる．学習により向上した点として，生成文の長さが挙げられる．学習前の生成では 100 モーラを超える長文が目立っていた一方，学習後ではその平均値は 30.7 と短歌としての制約である 31 モーラに迫っている．また学習前の生成ではアルファベットや記号が含まれた生成[6]が多く見られたが，学習後にはこれらはほとんど見られず，系列長と合わせ，より短歌らしい生成をするように変化していた．

一方で，枠組み学習後のモデルが生成した短歌を報酬モデルにより評価したところ，学習の前後でスコアは向上しないことが確認された．この結果の要因として，枠組み学習時には報酬モデルの学習時よりも Utakata 上での評価が低い短歌を多く利用したことが考えられる．また指示文に沿った生成が行えていない例も見られ，上の句を生成させる指示のような満たすことの難しい指示が全体としてのスコアの向上を妨げた可能性がある．

これらの結果は報酬モデルが系列長に対して敏感ではないという上述の結論の裏付けとなり，また，単純なファインチューニングでは短歌としての質を向上させることが困難であることを示唆している．

4.3　強化学習

報酬モデルと枠組み学習済みモデルを用いて強化学習を行った結果が表 3 のようになる．テストデータでの平均スコアは 0.724 となっており，強化学習前よりもスコアが向上している．これは表 1 に示した Utakata での評価が高い tanka$_{\uparrow}$ へのスコアを超えるような値であり，人間から評価を得やすいような短歌が生成されていることを意味している．

また生成テキストのモーラ数からも短歌らしい生成がなされていることが確認できる．生成物の平均モーラ数は 31.3 となっており，強化学習が生成系列長に対して悪影響を与えていないことがわかる．

モデルにより生成された短歌とそれぞれへのスコアの例を図 1 に示す．強化学習前は系列長は短歌らしい一方で，意味の取れない文章や情景描写に欠ける文章が多く生成されていた．しかし強化学習を経ることによりほとんどの生成物で意味が通るようになり，さらに豊かな情景描写や擬音語・体言止めなどの表現技法も見られた．句切れのような意味や調子の切れ目も散見されるようになり，全体としてより短歌らしい生成がなされるようになった．

一方で短歌らしいリズムを持った生成は強化学習後においても多くは見られなかった．枠組み学習により全体のモーラ数では短歌に近い生成が可能になったもの，ほとんどの場合で 5-7-5-7-7 という短歌のリズムは満たされておらず，これは強化学習後においても同様の傾向として確認された．

5　おわりに

本研究では人間の価値基準を用いた強化学習の手法である RLHF を用いて，短歌の自動生成における質の向上を目指した．短歌投稿サイト上の短歌と各短歌への評価をもとに，擬似的な評価者としての報酬モデルの学習を行った．学習により短歌の良し悪しの判断が可能となった報酬モデルと，枠組み学習を行った生成モデルを用いた強化学習では，短歌らしい系列長を持ち，かつ報酬モデルでのスコアが高い短歌の生成が可能であることが確認された．

一方で報酬モデルによるスコアと人手での評価との一致度は調査する必要がある．また短歌内のどういった要素がスコアに影響を与えているかへの検証も今後の課題である．短歌のリズムを満たすといった要素をはじめ，より質の高い生成のために，報酬モデルや強化学習のさらなる検討も必要である．これらの調査を通じて自動生成された短歌の質の向上を目指すとともに，人間の心を動かす本質的な短歌の要素を探求したいと考えている．

6）　これらのアルファベットや記号はモーラ数のカウントには含めていない．

謝辞

本研究は JSPS 科研費 JP21K21343，JP22H00524 の助成を受けたものです．本研究の進行にあたり，多くのご助言，ご協力を賜りました Tohoku NLP グループの皆様に感謝を申し上げます．また，研究活動に際し，有益なコメントを含む多くのサポートをいただいた東北大学坂口・乾・徳久研究室の松﨑孝介氏，吉田遥音氏に深く感謝申し上げます．

参考文献

[1] Jonas Belouadi and Steffen Eger. ByGPT5: End-to-End Style-conditioned Poetry Generation with Token-free Language Models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7364–7381, Toronto, Canada, July 2023. Association for Computational Linguistics.

[2] Yuka Takeishi, Mingxuan Niu, Jing Luo, ZhongYi Jin, and Xinyu Yang. WakaVT: A Sequential Variational Transformer for Waka Generation. **Neural Processing Letters**, Vol. 54, pp. 731 – 750, 2021.

[3] 浦川通, 新妻巧朗, 田口雄哉, 田森秀明, 岡崎直観, 乾健太郎. モーラを考慮した fine-tuning による口語短歌生成. 言語処理学会第 28 回年次大会, 2022.

[4] R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. **ArXiv**, Vol. abs/2309.13638, , 2023.

[5] Hong Wang, Xuan Luo, Weizhi Wang, and Xifeng Yan. Bot or Human? Detecting ChatGPT Imposters with A Single Question. **ArXiv**, Vol. abs/2305.06424, , 2023.

[6] 浦川通, 新妻巧朗, 田口雄哉, 田森秀明, 岡崎直観, 乾健太郎. 短歌における言語モデルの実応用–歌人の視点を通した生成と作歌支援の実践から–. 言語処理学会第 29 回年次大会, 2023.

[7] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, NIPS'17, p. 4302–4310, Red Hook, NY, USA, 2017. Curran Associates Inc.

[8] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, 2017.

[10] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. YAKE! Keyword extraction from single documents using multiple local features. **Information Sciences**, Vol. 509, pp. 257–289, 2020.

[11] 俵万智. サラダ記念日 : 俵万智歌集 . 河出書房新社, 1987.

A　報酬モデルの選定

報酬モデルの選定・学習に際して，rinna 社 RoBERTa の他に，東北大 BERT[7)]，早稲田大 RoBERTa[8)]での実験を行った．図 4 はハイパーパラメータチューニングの様子を一部示したものである．

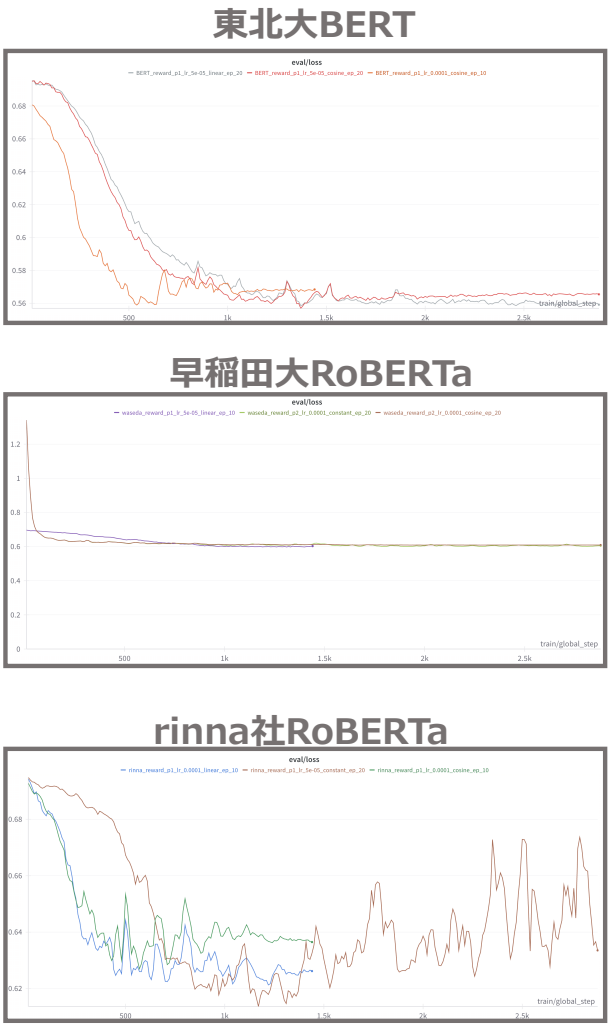

図 4　報酬モデルのハイパーパラメータチューニングの様子.

ハイパーパラメータチューニングでは，学習率やスケジューラー，エポック数などを変更して，各モデルで調整を行った．また直接 Utakata 上の短歌どうしを $\text{tanka}_{\uparrow}$，$\text{tanka}_{\downarrow}$ として与えるのではなく，Wikipedia から抽出した，短歌の形式を偶然満たした文字列を $\text{tanka}_{\downarrow}$ として与える学習を行ったのちに，$\text{tanka}_{\downarrow}$ として Utakata 上の短歌を与える二段階の学習（p2）なども試行した．最終的に検証データでの loss が下がったモデルをいくつか選択し，テストデータでの loss や accuracy なども測定した上で総合的に判断し，本研究では以降 rinna 社 RoBERTa を用いることとした．なお，本研究で用いたのは最大学習率 1e-4 で線形スケジューラーを用いて 10 エポック学習を行ったものであり，二段階の学習をしていない（p1）rinna 社 RoBERTa である．

7)　https://huggingface.co/cl-tohoku/bert-base-japanese-v2

8)　https://huggingface.co/nlp-waseda/roberta-base-japanese

B　指示文の自動生成

生成モデルの学習において指示文のバリエーション確保のため，各短歌につきそれぞれ 8 種類の指示文を自動生成した．これらの指示文とベースとなった短歌をペアとして学習に用いた．指示文のパターンと具体例を以下に示す．各具体例は「青バケツ揺らぐ水面に蘇る線香花火と縁側の祖母」という短歌をベースとした場合である．

(1) ベーシックな指示文
　例：「短歌を詠んでください」
(2) Yake!によるキーワード抽出を用いた指示文
　例：「『バケツ』を含む短歌を詠んでください」
(3) 先頭のワードのみを提示した指示文
　例：「『青』から始まる短歌を詠んでください」
(4) 先頭の 5 モーラを隠し，埋めさせる指示文
　例：「『揺らぐ水面に蘇る線香花火と縁側の祖母』の先頭に 5 文字を加えて短歌を完成させてください」
(5) 末尾の 7 モーラを隠し，埋めさせる指示文
　例：「『青バケツ揺らぐ水面に蘇る線香花火と』に続けて 7 文字を加えて短歌を完成させてください」
(6) 下の句を与え上の句を生成させる指示文
　例：「『線香花火と縁側の祖母』に上の句を加えて短歌を完成させてください」
(7) 上の句を与え下の句を生成させる指示文
　例：「『青バケツ揺らぐ水面に蘇る』に下の句を続けて短歌を完成させてください」
(8) 短歌の一部をランダムに隠し埋めさせる指示文
　例：「『青バケツ揺らぐ？？？？？線香花火と縁側の祖母』の？を埋めて短歌を完成させてください」

なお一部の短歌に関して，上の句と下の句の分離が難しいなどの理由から，すべての指示が作成できていない場合がある．