IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol. 003, LIIR075
DOI: https://doi.org/10.52731/liir.v003.075

Implementation of Autoregressive Language Models for Generation of Seasonal Fixed-form Haiku in Japanese

Kodai Hirata [*], Soichiro Yokoyama [*],
Tomohisa Yamashita [*], Hidenori Kawamura [*]

Abstract

This paper describes the implementation of an artificial intelligence haiku generator. We trained language models using existing haiku and literary studies, evaluated model performance using automatically computable evaluation indices such as perplexity, and subjectively evaluated the generated haiku by using a questionnaire. The main contributions of this paper are as follows. First, the effectiveness of a series of model evaluation processes, including automatically calculable evaluation indices and the results of subjective evaluations using questionnaires, is investigated. These processes are effective in the development of haiku generation models. Second, high-quality haiku generation is achieved using high-performance language models such as GPT-2 and BART. The results of the questionnaire survey revealed that it is possible to generate sensible haiku comparable to those written by humans. The insight gained from this study is applicable to other generative tasks.

Keywords: haiku generation, natural language generation, human evaluation, language model.

1 Introduction

Creative activity in arts is a uniquely human activity that originates from the unique human desire to create. The use of artificial intelligence to perform creative activities to the level of human beings attests to the advancement in artificial intelligence.

Novels and other works of fiction are representative of the fields related to creative writing by artificial intelligence; in particular, poetry generation has received considerable research attention. Rita Dove, a famous American author, said, “Poetry is language at its most distilled and most powerful.” Thus, poetry generation is considered to be one of the most difficult creative activities.

Among the various types of poetry, this study focuses on haiku, a written art form that has been popular in Japan for centuries. Haiku is a standard form of poetry based on the following restrictions: the number of syllables must be 17 (each part has 5, 7, 5 syllables), and it must contain only one seasonal word which expresses the scene and feelings of seasons. Some studies[1] show LSTM[2] could generate haiku to some extent so we try to generate haiku using transformer-based model like GPT-2[3], a more advanced model.

[*] University of Electro-Communications, Tokyo, Japan

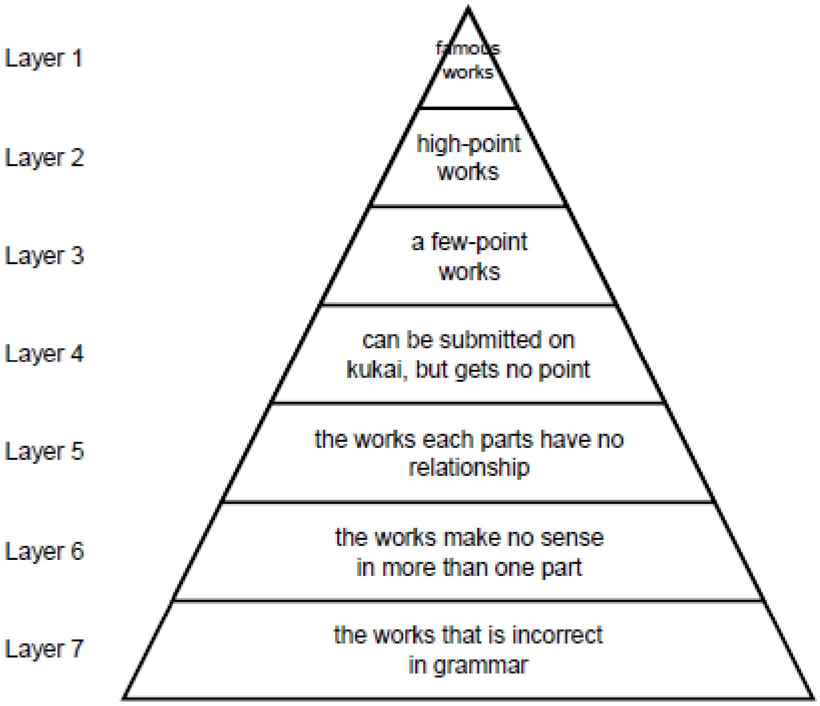

Layer 1
famous works
Layer 2
high-point works
Layer 3
a few-point works
Layer 4
can be submitted on kukai, but gets no point
Layer 5
the works each parts have no relationship
Layer 6
the works make no sense in more than one part
Layer 7
the works that is incorrect in grammar

Figure 1: An example of the quality of haiku in a hierarchical structure. The role of kukai referenced in the diagram is explained in Section 3.

Haiku that are considered “good” by multiple viewers are not necessarily identical because the historical background and knowledge possessed by the individual viewer play a major role in the haiku viewer’s reading of the scene and sentiment. These characteristics make it difficult to evaluate the model as a haiku generation model. On the other hand, many viewers can share the same evaluation of haiku in which interpreting the scene or emotion is difficult. Therefore, we evaluated the most common haiku to convey the author’s scene and presented it as a hierarchical structure, as displayed in Figure 1. According to Figure 1, performing a unified evaluation of the haiku in the upper layer is difficult because many people interpret a haiku differently. However, for a haiku in the lower layer, many people will be able to make the same evaluation.

Against this background, the primary aim of this study is to generate haiku that are classified as Layer 4 or higher, which are haiku that convey a scene in Figure 1, and we trained and evaluated a language model using haiku data. In addition to perplexity, which is a measure of the fluency of a language model, we evaluated the model using the proportion of the model’s generated sentences that satisfy haiku constraints. Furthermore, to verify the possibility of generating haiku that can imagine scenes and emotions, a questionnaire survey of people with haiku experience was conducted to evaluate the quality of the generated haiku from a subjective aspect. As far as we know, no studies conducted the consistent model evaluation in automatic and subjective way.

The contributions of this paper are twofold.

1. We implemented a flow that could be an effective process for training and evaluating haiku generation models using artificial intelligence. The process consists of esti-

mating model performance in terms of indicators that can be automatically computed and subsequently interpreting the actual model performance by comparing these results with the those of subjective evaluations.

2. We confirmed the superiority of transformer-based models in haiku generation.

2 Related Works

Poetry generation by artificial intelligence has been actively studied in various languages. In addition to English and Japanese, there have been studies on French poetry generation[4] using GPT-2[3] and Chinese poetry generation[5] using GPT-2.

In addition to haiku, which is the subject of this study, the generation of waka poetry and other forms of poetry in the Japanese language have been actively studied. Waka is a poem composed of 31 syllables (each part has 5, 7, 5, 7, 7 syllables), and like haiku, it has been popular in Japan since ancient times. There have been several studies on waka generation from keywords using models based on transformer[6] and variational autoencoders. In this study, the generated models are evaluated from objective aspects such as perplexity of the learned models as well subjective aspects by using a questionnaire survey.

There have been attempts to use artificial intelligence to generate haiku even before the development of deep learning. These methods involved haiku generation by combining existing phrases[7] and using blog posts as input[8]. Studies using deep learning include haiku generation using long short-term memory (LSTM)[2][1] and SeqGAN[9][10]. These studies don’t use transformer-based model for haiku generation.

The subjective evaluation of creative works by artificial intelligence is common. In addition, subjective evaluation in various ways is also used in creative text generation tasks[11].

In this study, a deep learning model including transformer-based models is used for haiku generation. A human subjective evaluation is performed on the trained model in addition to an automatic evaluation of perplexity and haiku conditions.

3 Types of Haiku and Evaluation of Generated Haiku

3.1 About Haiku

Haiku is considered as the smallest form of poetry in the world and has been a popular form of poetry in Japan for over 600 years. This study focuses on the generation and evaluation of the most common type of haiku, namely yuuki-teikei haiku in Japanese. According to the association of Japanese classical haiku, there are two basic conditions for seasonal fixed-form haiku. Only some works have deviated from these rules; however, these works are not appropriate as a first step in haiku generation and will not be covered in this study.

- Make it with 17 syllables(each part has 5, 7, 5 syllables)
- Include seasonal word (called kigo in Japanese)

Seasonal words are words that convey scenes from the four seasons, for example cherry blossoms (spring) and frogs (summer). In addition, to convey the author's intention with a small number of characters, there is a common understanding (true intention and true feeling) among haiku regarding the scene and meaning of seasonal words. Haiku composers and readers engage in creative activities based on the assumption of this true intention and true feeling. Thus, richer scenes can be shared using a smaller number of characters. The restriction on the number of syllables and the presence of seasonal words mentioned here are the main characteristics of haiku. Other common haiku techniques include the use of a single hiragana such as ”ya” and ”kana” (called kireji) and the avoidance of words where syllables span between each 5-7-5 haiku part. The key to creating a haiku is to convey a scene with a small number of characters.

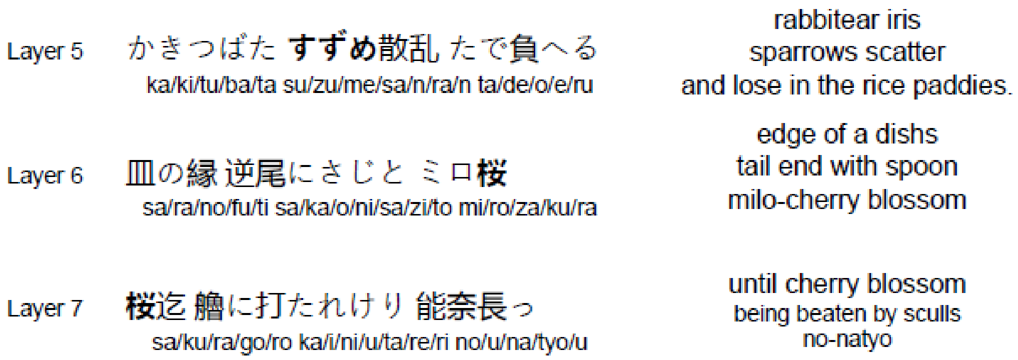
3.2 Evaluation of Haiku between Human

As is true of artistic works in general, it is important to receive recognition from others in order to hone the skills for creating good works of art. In the world of haiku, meetings known as kukai play a role in this. At a kukai, haiku poets bring their own haiku to share their sensitivities with each other by critiquing the haiku of others and voting on the haiku they think is good. Voting is typically conducted without identifying the author, and opinions are exchanged based on the results. This process of voting provides a guideline for improving the quality of haiku because it quantifies the quality of the haiku, at least in a specific kukai. Because a kukai is typically attended by people belonging to one school of haiku, not all people will share the same assessment. However, it still provides important information to hone a poet's haiku-generating skills.

3.3 Evaluation Policy of Generated Haiku Inspired by Kukai

In this paper, we aimed to construct an evaluation method that can give a high rating to the most common seasonal fixed-form haiku that conveys the author's scene. The scores in kukai described in the previous section can be used as guidelines for measuring the performance of the current model in haiku generation by artificial intelligence. Some of the most scored haiku in kukai include haiku that are grammatically inadequate or do not make sense. However, these haiku are considered intentional breakdowns of the basic rules of seasonal fixed-form haiku, and it seems appropriate to set the generation of seasonal fixed-form haiku as the first goal for the generation of haiku by artificial intelligence. Figure 1 shows a pyramid-like diagram of how we think of classifying seasonal fixed-form haiku according to the quality of the work. As mentioned above, not all haiku can be classified as shown in Figure 1, and there may be haiku that are highly evaluated by humans even if they are located at the bottom of the pyramid diagram. However, we believe that it can be used as a policy for considering the quality of works, which is highly subjective, and we adopted it as the policy for evaluating haiku in this study.

Layer 1 in Figure 1 is for haiku that are recognized by many people as masterpieces. In many cases, a haiku poet selects among hundreds or thousands of his or her own works, based on his or her own sense of haiku and the scores of the kukai, and publishes them as a collection. The haiku whose collections have become widely known and are recognized by the general public are those classified in Layer 1.



Layer 5	かきつばた **すずめ**散乱 たで負へる ka/ki/tu/ba/ta su/zu/me/sa/n/ra/n ta/de/o/e/ru	rabbitear iris sparrows scatter and lose in the rice paddies.
Layer 6	皿の縁 逆尾にさじと ミロ**桜** sa/ra/no/fu/ti sa/ka/o/ni/sa/zi/to mi/ro/za/ku/ra	edge of a dishs tail end with spoon milo-cherry blossom
Layer 7	**桜**迄 艪に打たれけり 能奈長っ sa/ku/ra/go/ro ka/i/ni/u/ta/re/ri no/u/na/tyo/u	until cherry blossom being beaten by sculls no-natyo

Figure 2: Examples of haiku and a possible poem meaning is added below each poem by author. The seasonal word of each haiku is displayed in bold.

Layers 2–4 are categorized by their scores in kukai.

- Layer 2: Haikus that are selected by many people and receive high scores in kukai
- Layer 3: Haikus with low scores in kukai
- Layer 4: The haiku as a whole makes sense and can be submitted to kukai; however, it will not receive several points in kukai.

Layers 5–7 contain phrases that are not submitted to the group and are defined as follows:

- Layer 5: Haikus with no apparent connection between parts
- Layer 6: Haikus with one or more clauses that do not make sense
- Layer 7: Haikus that contain one or more grammatical errors in Japanese

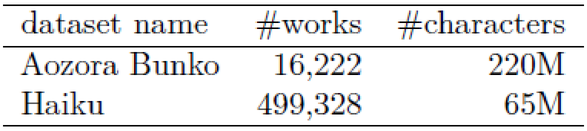
Examples and translations of haiku in Japanese classified as Layers 5–7 are shown in Figure 2. In the example of Layer 5, the parts 5, 7, and 5 alone make sense. However, because the haiku as a whole does not evoke a scene, it is classified as Layer 5. In the example in Layer 6, the“milo-cherry blossom” part is not grammatically incorrect in Japanese, but the word does not evoke a scene. In the example of Layer 7, the prompt“tsu” (small “tsu” in Japanese) appears at the end of the minute. This is classified as Layer 7 because it is grammatically incorrect in Japanese.

4 Dataset and Learning Language Model

4.1 Dataset

The language model was trained using data from the Aozora Bunko collection of works[1] for pretraining and a dataset of existing haiku for fine-tuning. The number of works and the total number of characters for each dataset are presented in Table 1.

Table 1: Dataset for training models.”# works” column indicates the number of literary works included in the Aozora Bunko or the number of haiku.



dataset name	#works	#characters
Aozora Bunko	16,222	220M
Haiku	499,328	65M

In this study, the training process involved pretraining with the Aozora Bunko dataset followed by fine-tuning with the haiku data for the following two reasons. The first reason is that the model performed better than models trained on haiku alone in a preliminary validation, and the second reason is that a large dataset of literary works can be collected.

The Aozora Bunko dataset, which includes a total of 16,222 works, was obtained from GitHub[2]. A total of 499,328 haiku datasets were collected from haiku published on the Internet by removing similar haiku. From the collected haiku, we prepared a dataset of 306,679 works that contained only seasonal fixed-form haiku, using the morphological analyzer MeCab[12]. A sampling study of 40 of these haiku confirmed that 75% of the haiku belong to Layer 4 or higher.

4.2 Learning Language Model

Autoregressive models based on deep learning exhibit high performance in sentence generation. AWD-LSTM[13] using LSTM[2], a type of recursive neural network, and GPT-2[3] using the decoder part of transformer are representative structures of such models. Transformer-based models typically outperform LSTM-based models in general sentence generation. We conducted an experiment to see if a similar trend holds for haiku generation and whether this trend can be captured by both automatic and subjective evaluations.

The models used for haiku generation in this study are AWD-LSTM, one of the most popular LSTM-based models for text generation, and our model is consist of 3 LSTM-blocks. GPT-2, BART[14], one of the most popular transformer-based models for text generation are used for haiku generation in this study. We use GPT-2 small has 12 decoders. BART is originally an encoder-decoder model, but only the decoder part was used this time because it is used as a language model. Therefore, the only difference from GPT-2 is the position of layer regularization in each transformer block. It was implemented using

[1] Aozora Bunko is a dataset of Japanese literary works, most of which have copyrights that expired.
[2] https://github.com/aozorabunko/aozorabunko

Huggingface transformers[15].

During pretraining, the entire data from Aozora Bunko was divided in the ratio 8:1:1 for training, validation, and testing, respectively. During fine-tuning, haiku data was divided into training and testing data at a ratio of 8:2, and 20% of the training data was used as validation data. Tokenization was performed on a letterby-letter basis, with a vocabulary size of 6,542 for all models.

We trained our models on one machine with 4 NVIDIA GeForce RTX 2080ti GPUs. Fine-tuning takes 1 hour (GPT-2) or 2 hours (AWD-LSTM, BART).

5 Experiments

5.1 Experiments Overview

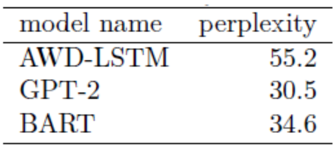
To construct a model capable of generating haiku that can be classified as Layer 4 or higher as per Figure 1 with high accuracy, we conducted the following experiments. Although submitting all the generated haiku to a kukai and collecting evaluations would be ideal, this process is not realistic because the maximum number of works that can be submitted to a single kukai is about five. Therefore, we circulated a questionnaire asking haiku poets whether they would rate the model-generated haiku as good if they were submitted to kukai. Furthermore, because of the time constraints of this survey, the results of the relatively inexpensive automatic evaluation were used to select the models to be included in the survey. Specifically, the process was as follows: First, test perplexity was calculated for the trained model to evaluate its basic performance as a language model for haiku data. Next, for each string generated by each model, the proportion that satisfies the rules of seasonal fixed-form haiku introduced in the previous chapter was calculated using a morphological analyzer, and the proportion was compared with that of the training source data. Based on the results obtained up to this point, the model to be surveyed was determined. Finally, a questionnaire survey on the quality of the generated haiku was conducted to evaluate the performance of the generative model at this stage.

The three models used in the experiment were AWD-LSTM, GPT-2, and BART, and they were trained on haiku data. For GPT-2, which had the lowest perplexity, we created a model trained only with seasonal fixed-form haiku to evaluate differences in performance depending on the training data. In each experiment, a 5-fold cross-validation was performed, and the initial parameters were set with five seed values. Training was performed until the loss on the validation data converged, and the model with the smallest loss on the validation data was used in subsequent experiments. The vocabulary size is 6,452 for all models.

5.2 Automatic Evaluation: Perplexity

The test perplexity of haiku data was calculated for three trained language models as one of the automatic evaluation methods. To ensure uniform distribution of the training data, the haiku data includes some haiku that do not satisfy the requirements for seasonal fixed-form haiku. The results are shown in Figure 2. As we confirmed in Section 3, about 75% of the training data were haiku from Layer 4 and above, so the models with lower test perplexity were better at capturing the features of haiku from Layer 4 and above.

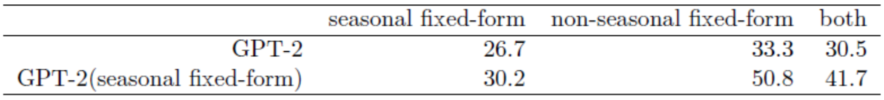
Table 2: Test perplexity for haiku data.



model name	perplexity
AWD-LSTM	55.2
GPT-2	30.5
BART	34.6

From Table 2, it can be seen that the transformer-based models, GPT-2 and BART, capture the characteristics of the haiku in Layer 4 and above better than the LSTM-based model, AWD-LSTM.

Table 3: Test perplexity for haiku test sets with seasonal fixed-form, non-seasonal fixed-form and both. GPT-2 (seasonal fixed-form) represents a model trained only with seasonal fixed-form.



	seasonal fixed-form	non-seasonal fixed-form	both
GPT-2	26.7	33.3	30.5
GPT-2(seasonal fixed-form)	30.2	50.8	41.7

Next, for the GPT-2 model that produced the best results, we trained a model with the training data limited to seasonal fixed-form haiku in order to evaluate the difference in performance depending on the training data. The results are presented in Table 3.

From Table 3, it can be seen that the perplexity is larger for the model trained with only seasonal fixed-form haiku on all test sets than for the model trained with all haiku including non-seasonal fixed-form haiku.

5.3 Automatic Evaluation: Ratio of Satisfying Haiku Rule

For automatic evaluation, we calculated the proportion of strings generated by each model that satisfied the conditions for seasonal fixed-form haiku and the conditions for typical haiku, as described in the previous section, by automatic judgment using the morphological analyzer MeCab.

First, 1,000 strings were generated from each model. For each of these strings, we calculated the percentage of unknown words and diversity conditions required for the models, in addition to the haiku conditions. The conditions for comparison are as follows.

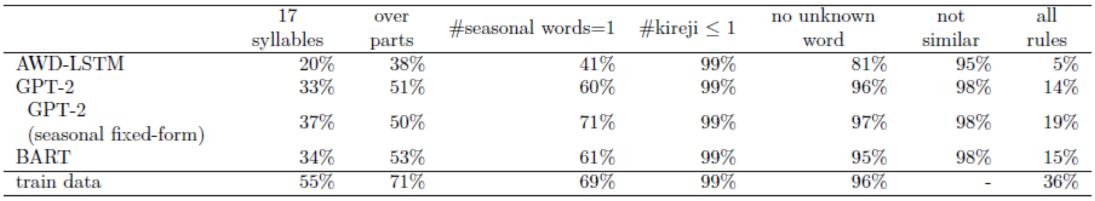
- 17 syllables: Morphological analysis results revealed that it consists of 17 syllables
- over parts syllable: Do not use words that span syllables between each phrase of 5, 7, and 5
- #seasonal words=1: The number of seasonal words is one
- #kireji ≤ 1: The number of kireji is less than one

- no unknown words: Undefined words are not included in the morphological
- not similar: Levenshtein distance from all strings in the training data is less than five

Table 4 presents the results of calculating the percentage of each condition satisfied.

From the Table 4, it can be seen that the two transformer-based models could generate strings that satisfied each condition at about the same rate as the training data, except for the condition regarding the number of kireji. Even the model that learned non-seasonal fixed-form haiku could generate haiku that satisfied each condition at a rate not considerably different from the one that learned only seasonal fixed-form haiku.

Table 4: Percentage of generated sentences that satisfy the haiku rules



	17 syllables	over parts	#seasonal words=1	#kireji ≤ 1	no unknown word	not similar	all rules
AWD-LSTM	20%	38%	41%	99%	81%	95%	5%
GPT-2	33%	51%	60%	99%	96%	98%	14%
GPT-2 (seasonal fixed-form)	37%	50%	71%	99%	97%	98%	19%
BART	34%	53%	61%	99%	95%	98%	15%
train data	55%	71%	69%	99%	96%	-	36%

The results of the experiments show that the two transformer-based models outperform the LSTM-based models in terms of perplexity and generating strings that satisfy the haiku conditions, which can be evaluated automatically.

Models trained on all haiku exhibited lower perplexity than models trained on only seasonal fixed-form haiku and were able to generate strings that satisfied the haiku conditions at about the same level. The reasons for this include the possibility that the amount of data was not sufficient to learn haiku with only seasonal fixed-form haiku and that non-seasonal fixed-form haiku may contain important information for capturing the characteristics of haiku.

5.4 Human Evaluation: Questionnaire on the Quality of Generated Haiku

Finally, a questionnaire survey was conducted on the quality of the generated haiku as seasonal fixed-form haiku to determine if they could be classified as Layer 4 or higher from a subjective perspective. According to the results of previous experiments, we used the model trained with all haiku in this survey.

This study was conducted on 160 haikus, 40 generated by each model and 40 randomly sampled from the training data. Because the purpose of this survey was to investigate the quality of seasonal fixed-form haiku, haiku that satisfied the conditions of 17 syllables and #seasonal words=1 were extracted from the haiku generated by each model and used as the target haiku for the survey. Three haiku poets with more than 10 years of haiku experience and two university students, including the author, with one and two years of haiku experience, respectively, were asked to evaluate the generated haikus.

The following three survey items were set up to classify haiku in Figure 1. The respondents were asked to answer on a three-point scale of 1 (does not apply), 2 (applies a little), and 3 (applies a lot), respectively.

- meaningful: It is a haiku that makes sense as Japanese.
- appreciated seasonal word: Seasonal words are used in accordance with their original meaning, the true intention and true feelings
- kukai: It is a haiku that I would like to vote for as a good haiku in kukai

"Meaningful" and "appreciated seasonal word" are factors that determine whether a haiku is classified as Layer 4 or higher. “Kukai’’ is a prospective item for this study because it is the element that determines whether a haiku is classified as Layer 3 or higher.

The results of the questionnaire are presented in Table 5.

Table 5: Questionnaire result regarding the quality of generated and human haiku. In the questionnaire, respondents were instructed to respond on a three-point scale for each question. The value in the table is the mean of answers.

	meaningful	appreciate seasonal word	kukai
AWD-LSTM	1.5	1.4	1.1
GPT-2	2.2	1.8	1.5
BART	2.1	1.6	1.3
human	2.0	1.7	1.4

"Meaningful" indicates that the two transformer-based models could generate haiku that are comparable to those generated by humans. The two transformerbased models received similar ratings to human haiku for "appreciated seasonal word." Thus, it is clear that the models could use seasonal words, which is important for haiku, in a more appropriate sense. For the third "kukai", the two transformerbased models also exhibited results comparable to human haiku.

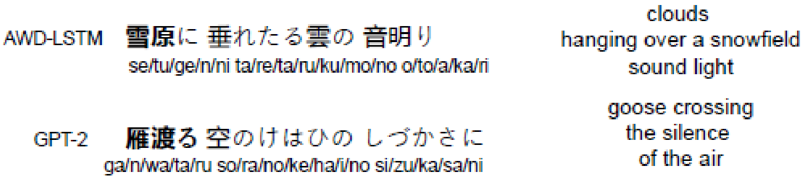
In the present setting, the results of the automatic evaluation of perplexity and haiku conditions and the results of the subjective evaluation based on the questionnaire survey revealed similar trends. The number of experiments that require human intervention is limited. Therefore, an effective model evaluation process is to narrow down promising candidates by using several automatic evaluation indices and subsequently subjecting them to human subjective evaluation, as performed in this study.

5.5 Qualitative Analysis of Questionnaire Results

Finally, we conducted a qualitative analysis of the survey results based on the actual haiku. As a trend among the models, many of the haiku generated by AWD-LSTM were classified as Layer 5, which lacked connection between haiku parts and did not convey a sense of scene. In contrast, many of the haiku generated by GPT-2 conveyed a clear scene, and many of them reached a level that could be submitted to kukai. Figure 3 shows an example of a haiku used in the survey. In the example shown in Figure 3, AWD-LSTM generated the poorly connected phrase “clouds hanging over a snowfield,” and GPT-2 could generate a work that conveys a scene.

The results presented in Table 5 show that the transformer-based model and the human-created haiku were almost equally rated. However, the individual haiku revealed that only a few of the human-created haiku fall into Layers 6 and 7, while some of the model-generated haiku are low-level, falling into Layers 6 and 7. Therefore, additional questionnaire items or subdivision of the rating scale may be necessary for a detailed analysis.

The overall trend of the responses was that the higher the “meaningful” and “appreciated seasonal word” were in the haiku, the higher the “kukai” also tended to be. Thus, it is effective, to some extent, to view the quality of seasonal fixed-form haiku in a hierarchical structure such as that shown in Figure 1. However, some haiku do not fit this trend, thus indicating that not all haiku can be captured in a simple hierarchical structure.



AWD-LSTM 雪原に 垂れたる雲の 音明り
se/tu/ge/n/ni ta/re/ta/ru/ku/mo/no o/to/a/ka/ri

clouds
hanging over a snowfield
sound light

GPT-2 雁渡る 空のけはひの しづかさに
ga/n/wa/ta/ru so/ra/no/ke/ha/i/no si/zu/ka/sa/ni

goose crossing
the silence
of the air

Figure 3: Examples of haikus used in the experiment. Each haiku generated by AWD-LSTM and GPT-2.

6 Conclusion

In this study, three models were trained for haiku generation, with automatic evaluation of perplexity and haiku conditions, and human subjective evaluation was performed through a questionnaire survey. The results showed that the transformerbased model outperformed the LSTM-based model and could generate haiku that were classified as Layer 4 or higher in Figure 1 with higher accuracy. In the future, we would like to obtain a model that can generate haiku above Layer 3 with higher accuracy.

A major issue in creative generation using deep learning models is finding a suitable method to evaluate the generated works. An effective method of model evaluation is to cycle through the indicators that can be automatically calculated and the results of subjective evaluation, which was performed in this study. The results of the study can provide insight into the development of other creative models.

References

[1] Xianchao Wu, Momo Klyen, Kazushige Ito, and Zhan Chen. Haiku generation using deep neural networks. In The Association for Natural Language Processing, 2017.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. Neural Computation, 9(8):1735–1780, 11 1997.

[3] A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In OpenAI Blog, 2019.

[4] Mika Hämäläinen, Khalid Alnajjar, and Thierry Poibeau. Modern French Poetry Generation with RoBERTa and GPT-2. In 13th International Conference on Computational Creativity (ICCC) 2022, Bolzano, Italy, June 2022. ICCC.

[5] Xingxing Zhang and Mirella Lapata. Chinese poetry generation with recurrent neural networks. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 670–680, Doha, Qatar, October 2014. Association for Computational Linguistics.

[6] Vaswani et al. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[7] Naoko Tosa, Hideto Obara, and Michihiko Minoh. Hitch haiku: An interactive supporting system for composing haiku poem. In Scott M. Stevens and Shirley J. Saldamarco, editors, Entertainment Computing - ICEC 2008, pages 209–216, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[8] Rafal Rzepka and Kenji Araki. Haiku generator that reads blogs and illustrates them with sounds and images. In Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, page 2496–2502. AAAI Press, 2015.

[9] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, page 2852–2858. AAAI Press, 2017.

[10] Konishi et al. Generation of haiku preferable for ordinary people by seqgan. In Information Processing Society of Japan, Kansai Branch Section Convention, 09 2017.

[11] Mika Hämäläinen and Khalid Alnajjar. Human evaluation of creative nlg systems: An interdisciplinary survey on recent papers. In Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), page 84–95, United States, 2021. Workshop on Natural Language Generation, Evaluation, and Metrics, GEM Workshop.

[12] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[13] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing LSTM language models. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 2018.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and compre-

hension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[15] Wolf et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October 2020. Association for Computational Linguistics.