TEXT-TO-LYRICS GENERATION WITH IMAGE-BASED SEMANTICS AND REDUCED RISK OF PLAGIARISM

Kento Watanabe **Masataka Goto**
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{kento.watanabe, m.goto}@aist.go.jp

ABSTRACT

This paper proposes a text-to-lyrics generation method, aiming to provide lyric writing support by suggesting the generated lyrics to users who struggle to find the right words to convey their message. Previous studies on lyrics generation have focused on generating lyrics based on semantic constraints such as specific keywords, lyric style, and topics. However, these methods had limitations because users could not freely input their intentions as text. Even if such intentions can be given as input text, the lyrics generated from the input tend to contain similar wording, making it difficult to inspire the user. Our method is therefore developed to generate lyrics that (1) convey a message similar to the input text and (2) contain wording different from the input text. A straightforward approach of training a text-to-lyrics encoder-decoder is not feasible since there is no text-lyric paired data for this purpose. To overcome this issue, we divide the text-to-lyrics generation process into a two-step pipeline, eliminating the need for text-lyric paired data. (a) First, we use an existing text-to-image generation technique as a text analyzer to obtain an image that captures the meaning of the input text, ignoring the wording. (b) Next, we use our proposed image-to-lyrics encoder-decoder (I2L) to generate lyrics from the obtained image while preserving its meaning. The training of this I2L model only requires pairs of "lyrics" and "images generated from lyrics", which are readily prepared. In addition, we propose for the first time a lyrics generation method that reduces the risk of plagiarism by prohibiting the generation of uncommon phrases in the training data. Experimental results show that the proposed method can generate lyrics with different phrasing while conveying a message similar to the input text.

1. INTRODUCTION

Automatic lyrics generation methods have been proposed as an important research topic in lyrics information processing [1]. With the aim of supporting users who already know what they want to convey in their lyrics but struggle to find the appropriate words, the methods are used in writing support systems providing them with generated lyrics as a source of new inspiration [2–8]. Most previous studies have focused on lyrics generation that is conditioned by semantic constraints, including specific keywords, lyric style, and topics. For example, Watanabe et al.'s system generates lyrics based on pre-defined topics selected by the user, but the limited range of topics results in similar styles of generated lyrics [2]. Oliveira et al.'s system generates poems based on keywords entered by the user, but it cannot generate poems based on sentences or paragraphs representing the user's intention [3, 4].

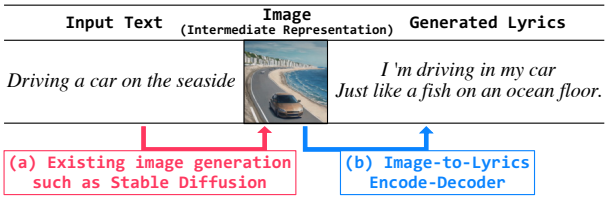


Figure 1. Overview of the proposed text-to-lyrics generation method.

To provide more flexible lyric writing support, we propose generating lyrics based on freely formatted text entered by the user. We believe this approach surpasses the use of semantic constraints such as topics and keywords in terms of flexibility. While existing paraphrase systems [9] can be considered useful for this approach, the paraphrased lyrics may not provide sufficient inspiration because they tend to be similar in wording to the input text. For example, even if a similar phrase "*Driving a car along the coastline*" is generated from the input text "*Driving a car on the seaside*", the user is unlikely to get new inspiration.

Therefore, the aim of this study is to develop a method for generating lyrics that not only have meanings similar to the input text but also use wording different from the input text. For example, if a user freely enters text that represents the content of the lyrics, such as "*Driving a car on the seaside*", our method generates lyrics with different wording, such as "*I'm driving in my car. Just like a fish on an ocean floor.*". As a simple way to achieve this aim, Transformer-based encoder-decoders [10] could be used for generating lyrics from text, but they require large text-lyric paired data for training, which is currently unavailable. To address this issue, we could use text summarization and machine translation to generate text from lyrics and obtain paired data automatically. However, since the generated text

and lyric pairs have similar wording, an encoder-decoder trained using those paired data may generate lyrics with wording similar to the input text.

To achieve text-to-lyrics generation without using any paired text data for training, we propose a two-step pipeline framework: (a) using an existing text analyzer to obtain only the semantic representation from the input text, and (b) generating lyrics from the obtained representation. The core idea of this framework is to leverage a text-to-image generation technique such as Stable Diffusion [11] as the text analyzer. An image generated from the input text can serve as a reasonable intermediate representation that captures the meaning of the text while ignoring the details of its wording (Figure 1 (a)). Using the generated image, our image-to-lyrics encoder-decoder generates semantically related lyrics (Figure 1 (b)). It needs many image-lyric pairs as training data, but we can readily prepare those pairs by generating images from lyrics of many songs. This is an advantage of using text-to-image generation. Another advantage is that it can generate images without regard to the input text's format, i.e., whether it is a word, phrase, sentence, or paragraph. We can thus provide flexible lyric writing support that is not constrained by the format of the input text.

Machine learning-based generation methods may inadvertently output portions of the training data directly without modification. This output can be considered plagiarism in some cases [12, 13]. Therefore, this paper also proposes an anti-plagiarism method to reduce this risk. We assume that generating common phrases (word sequences having high commonness [14]) used in many songs is not plagiarism, and reduce the risk of plagiarism by prohibiting the generation of uncommon phrases used in only a few songs. To the best of our knowledge, this is the first study to include such an anti-plagiarism method in lyrics generation.

Experimental results show that our text-to-lyrics generation method can generate lyrics with meaning similar to the input text but expressed differently. Another experiment shows that lyrics generated without using our anti-plagiarism method would result in plagiarizing uncommon phrases in the training data, but those undesirable phrases can successfully be removed by our method.

2. RELATED WORK

While natural language generation methods such as machine translations and chat systems have been actively studied and their performance greatly improved by deep neural networks (DNNs), automatic lyrics generation has also attracted attention as a research topic [1]. Most studies of lyrics generation have focused on lyric-specific musical constraints such as melody [15–20], rhyme [6, 8, 21–25], and audio signal [26–28]. While these lyric-specific musical constraints are an important aspect of lyrics generation, the main focus of this study is on the controllability of the semantic content of the generated lyrics.

Other studies have focused on lyrics generation that is conditioned by semantic constraints such as input keywords, styles, and topics [2–5, 29–32]. However, although these constraints allow some control over the semantic content of the generated lyrics, there may be differences between the user's intentions and the semantic content of the generated lyrics. To improve the usability of the lyrics generation method as a creative tool, we believe that users should be able to enter freely formatted text (words, phrases, sentences, paragraphs, etc.). Our proposed method therefore allows any text format, giving users greater control over the semantic content of the generated lyrics.

Some studies have proposed methods for generating lyrics that are semantically related to the input text [6, 7]. Ram et al. proposed a fine-tuned T5 model [9] that generates single-line lyrics that follow several lines of input lyrics [6]. This method allows the user not only to enter sentences but also to control the rhyme and syllable count of the generated lyrics by adding special tokens at the end of the input sentence. In contrast to that method, in which the generated lyrics are a continuation of the input lyrics, ours generates lyrics that capture the semantic content of the input text. Zhang et al.'s research motivation is similar to ours, as they have also proposed a method for generating lyrics that capture the semantic content of the input text (which they refer to as passage-level text) [7]. To overcome the problem of the lack of text-lyric paired data for training the text-to-lyrics encoder-decoder, they collected lyrics data and passage-level text data (such as short novels and essays) separately and utilized an unsupervised machine translation framework. Specifically, they prepared two encoder-decoders, one for lyric text and one for passage-level text. They then aligned the latent representation space of these two encoder-decoders to build a text-to-lyrics encoder-decoder. In this paper, we propose a novel approach to develop a text-to-lyrics generation method that requires only lyrics data. While Zhang et al.'s method requires the collection of both lyrics and input texts, ours does not require additional text data, thus simplifying the development of the lyrics generation method.

3. TEXT-TO-LYRICS GENERATION WITH IMAGE-BASED SEMANTICS

As described in Section 1, the proposed text-to-lyrics generation method first generates an image from the input text by leveraging an existing text-to-image generation method. It then generates lyrics from the generated image by using our own image-to-lyrics encoder-decoder that we call *I2L*. Since the image serves as an intermediate representation to extract the meaning of the input text, the generated lyrics can have similar meaning but different wording.

The network structure of the I2L is illustrated in Figure 2. By assuming that one paragraph of lyrics can be represented in a single image, we set the unit of the generated lyrics to a paragraph.

We first uses the animation-style image generation method *Anything V3.0*. [1] to obtain an image having a uniform style. The reasons for using Anything V3.0 here are

[1] A fine-tuned Stable Diffusion model. `https://huggingface.co/Linaqruf/anything-v3.0`

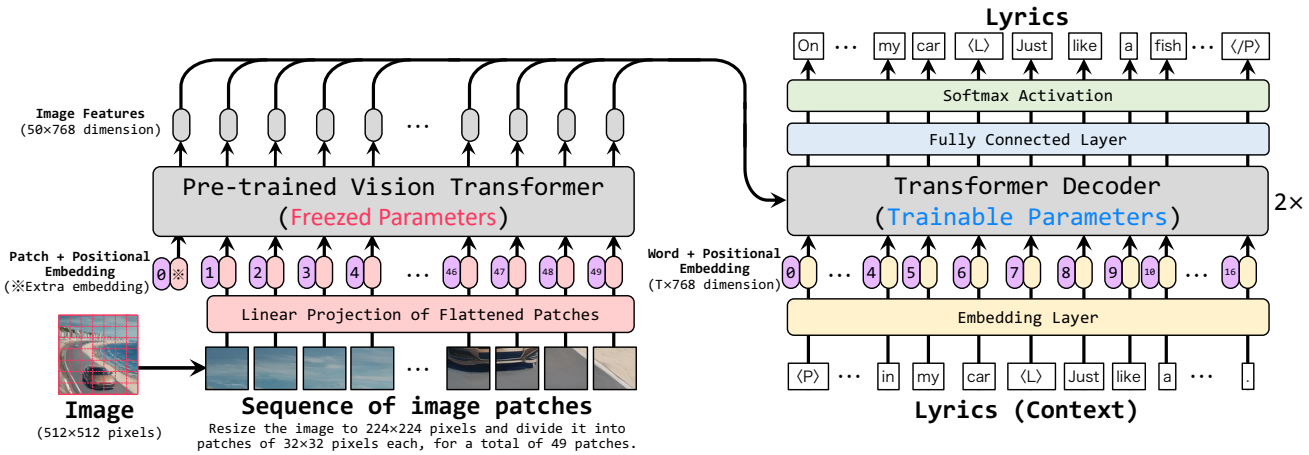


Figure 2. Image-to-lyrics encoder-decoder (I2L) for generating lyrics from an image that is generated from the input text.

(1) it can generate images that represent the input text without prompt engineering, and (2) the use of images with a uniform style facilitates I2L training. The image has a resolution of 512×512 and corresponds to a paragraph of the English lyrics.

As shown in Figure 2, we then segment the generated image into 49 patches and compute the features of the image patches by using a pre-trained Vision Transformer [2] [33] to obtain 50 features (each with 768 dimensions) per image. These 50 image features are fed into the multi-head attention layer of the Transformer decoder [10]. We feed each word in a paragraph into the word embedding and positional embedding layers to compute the word vectors, and feed each word vector into the masked multi-head attention layer of the Transformer decoder. The output of the Transformer decoder is fed into the fully connected layer FC to obtain a vector of vocabulary size dimensions. Finally, we apply the $\mathrm{softmax}$ activation function to this vector to calculate the word probability distribution.

3.1 Parameters

We use 768 as the number of embedding dimensions, 6 as the number of multi-heads, 2 as the number of decoder layers, 1024 as the number of feedforward layer dimensions, and GELU as the activation function. For optimization we use AdamW [34] with a mini-batch size of 8, a learning rate of 0.001, and a warm-up step of one epoch. Training was run for 40 epochs, and the I2L used for testing was the one that achieved the best loss on the development set.

We dare to train our Transformer decoder from scratch using only the lyrics data we have, without reusing available pre-trained large-scale language models (LLMs) such as BERT [35] or GPT-2 [36]. This is because when the training data of LLMs contain copyrighted literary works such as novels, poems, or essays, reusing pre-trained LLMs can result in plagiarizing those works. Since we would like to reduce the risk of plagiarism as described in Section 3.4, we cannot leverage pre-trained LLMs.

[2] https://huggingface.co/google/vit-base-patch32-224-in21k

3.2 Training data

We sample 129,747 English songs from the Music Lyrics Database V.1.2.7 [3] so that each song contains at least three paragraphs. The resulting dataset contains 927,535 paragraphs. This means that we can obtain 927,535 images by using Anything V3.0. We then split these songs into training (90%) and development (10%) sets. We use the top 52,832 words with the highest document-frequency as the vocabulary for training, and convert the other words to a special symbol ⟨unknown⟩. This vocabulary includes ⟨L⟩ tags for line breaks, ⟨P⟩ tags for the beginning of paragraphs, and ⟨/P⟩ tags for the end of paragraphs.

We applied the same procedure not only to the lyrics of English songs but also to the lyrics of 142,772 Japanese songs. This Japanese dataset contains 1,078,500 paragraphs, and the vocabulary size is 50,989 words. To extract word boundaries for Japanese lyrics, we apply the CaboCha parser [37]. Japanese lyrics are pre-translated into English by a Japanese-English translator [4] for use with Anything V3.0. We use these English and Japanese lyrics datasets to train two I2Ls (one for each language).

3.3 Decoding algorithm

We expect that generating and suggesting different variations of lyrics can give users new ideas for writing lyrics. To generate such different variations, we use a sampling method rather than a beam search method. In the sampling method, we sample each word according to the probability distribution calculated by the Transformer decoder. Sampling words according to a probability distribution allows a wide variety of words to be included in the generated lyrics, although some words that make the generated lyrics meaningless may be included. To avoid generating such meaningless lyrics, we use a Top-p sampling method that prohibits sampling words with low generation probabilities [38]. We can generate several lyrics simultaneously by

[3] https://www.odditysoftware.com/page-datasales1.htm

[4] https://huggingface.co/staka/fugumt-en-ja

running Top-p sampling in parallel. The probability distribution for word sampling in Top-p sampling is calculated using the formula $\mathrm{softmax}(\mathbf{z}/\tau)$: where $\mathbf{z}$ is the output of the fully connected layer FC and τ is the temperature parameter. If τ is less than 1, common words with high probability values are more likely to be sampled. In model training we set τ to 1, while in lyrics generation the user can set τ freely.

3.4 Anti-plagiarism method for lyrics generation

One of concerns with lyrics generation based on machine learning is the risk of plagiarism since the generated lyrics may contain phrases that are identical to existing lyrics phrases in training data, potentially leading to copyright infringement issues. To address this issue, we propose a method to reduce the risk of plagiarism in machine learning-based lyrics generation. This method not only allows the generation of new phrases that are not present in the training data, but also permits the use of commonly used phrases such as "*I love you*" in the generated lyrics. In contrast, it prohibits the use of uncommon phrases that we consider to be a form of plagiarism. To achieve this, we create a list of uncommon phrases, $UncommonPhrase$, and prohibit the generation of phrases that are included in this list.

First, we define the uncommon phrases included in $UncommonPhrase$, as well as the new phrases and common phrases that are allowed to be generated. A phrase is defined by a word n-gram, denoted by $\{w_1, ..., w_n\}$, where w is a word. We categorize a phrase as "new", "common", or "uncommon" according to $SN(\{w_1, ..., w_n\})$ defined as the number of songs in which the n-gram occurs in the training data:

- If $SN(\{w_1, ..., w_n\}) = 0$, this n-gram is a new phrase (i.e., it does not appear in the training data).
- If $3 < SN(\{w_1, ..., w_n\})$, this n-gram is a common phrase (i.e., it appears frequently in the training data).
- If $1 \leq SN(\{w_1, ..., w_n\}) \leq 3$, this n-gram is an uncommon phrase (i.e., it appears infrequently in the training data). [5]

Note that there is a possibility of mistaking uncommon phrases for common phrases when duplicate lyrics are contained in the training data, which results in larger SN values than they should be. It could happen when different artists sing the same lyrics, the same lyrics is repeatedly registered, and so on. We therefore identify duplicate lyrics according to the following two criteria: (1) we assume that pairs of lyrics with the same 20-grams are duplicates, and (2) we assume that pairs of lyrics with a normalized edit distance [39] of less than 0.5 are duplicates. To calculate SN accurately, we then concatenate the identified duplicate lyrics and replace those lyrics with the single concatenated lyrics. When lyrics that do not duplicate are mistaken for duplicate lyrics, a common phrase can be mistaken for an uncommon phrase, but it is better than vice versa from the anti-plagiarism viewpoint. This reduced the number of English songs in our lyrics data from 129,747 to 108,497. [6]

Based on this SN criteria, we collect uncommon phrases from our training data. However, it is important to note that even if a word 3-gram is a common phrase, it may become an uncommon phrase when it becomes a word 4-gram. For instance, "*I love you*" is a common 3-gram with a large SN, while "*I love you darling*" is an uncommon 4-gram with a small SN. Therefore we do not use a single value of n but instead consider all values of n within a range from 1 to sufficiently large values. However, it is difficult to store all uncommon phrases in memory because the number of n-grams that have to be listed increases with n. To overcome the memory limitation problem, we propose to use the following procedure to minimize the number of uncommon phrases we need to store in memory: (1) we start by examining 1-grams, then move on to 2-grams, 3-grams, and so on until we have looked at all possible n-grams in the training data. (2) For each target n-gram, we generate all possible sub-n-grams of length 1, 2, ..., $n-1$. If any of these sub-n-grams are already in $UncommonPhrase$, we can skip adding the target n-gram to $UncommonPhrase$ because we know it is uncommon. Otherwise, we add the target n-gram to $UncommonPhrase$. Following this procedure, we collected approximately 22.3M uncommon n-grams with n ranging from 1 to 21 for English lyrics. [7]

After creating $UncommonPhrase$ using the above procedure, we prohibit their generation during Top-p sampling by the following two steps: (1) During word generation, we check whether any sub-n-grams derived from the word sequence $\{w_1, ..., w_t\}$ are included in $UncommonPhrase$. (2) If any of these sub-n-grams are found in $UncommonPhrase$, we prohibit the generation of word w_t by setting its generation probability $P(w_t|\{w_1, ..., w_{t-1}\})$ to zero.

4. QUANTITATIVE EVALUATION

The proposed text-to-lyrics generation method was quantitatively evaluated using two metrics:

Test-set perplexity (PPL): This is a standard evaluation measure for encoder-decoders. The PPL metric measures the degree of predictability of the phrasing in the original text in the test set [40]. A smaller PPL value is better since it indicates that the encoder-decoder has a higher ability to generate lyrics that capture the meaning of the input text.

Normalized edit distance (NED): The normalized edit distance [39] between the generated lyrics and the input text is calculated to evaluate whether the proposed method generates lyrics that differ in wording from the input text. A larger NED is better since it indicates that the generated lyrics have wording more different from the input text.

[5] In this study, we tentatively set the threshold for SN at 3. Since there is no established legal rule, we believe that this threshold will be determined by social consensus in the future. Providing the technical basis for such discussions is also a contribution of this study.

[6] For Japanese lyrics, the number of songs was reduced from 142,772 to 119,595.

[7] For Japanese lyrics, we collected approximately 18.2M uncommon n-grams with n ranging from 1 to 19.

Method	English		Japanese	
	PPL	NED	PPL	NED
I2L (proposed)	**84.86**	**0.78**	**231.49**	**0.92**
S2L	346.73	0.69	306.19	0.86
B2L	544.21	0.71	1051.58	0.66
H2H	163.98	0.68	583.13	0.90

Table 1. Results of quantitative evaluation.

4.1 Experimental dataset

To evaluate the proposed lyrics generation method, we constructed a small test dataset consisting of pairs of lyrics and input text representing the semantic content of the lyrics. Since such a dataset is not available, for English songs, we prepared a test dataset that included plot texts from 20 Disney animated films, taken from Wikipedia, along with their corresponding theme song lyrics. We here assume that the lyrics of each theme song are written based on the content of that film. For Japanese songs, we prepared 51 Japanese animation plot texts and their theme song lyrics.

4.2 Methods Compared

To compare the proposed method with possible different methods, we prepared the following encoder-decoders trained on paired data created in different suitable ways.

Image-to-Lyrics encoder-decoder (I2L) This is the proposed encoder-decoder trained on image-lyric paired data.

Summary-to-Lyrics encoder-decoder (S2L) We converted each lyric paragraph in the training data into a summary using a text summarization method [8] to create summary-lyric paired data. The data is then used to train a Transformer-based summary-to-lyric encoder-decoder.

Back-translated-lyrics-to-Lyrics encoder-decoder (B2L) We translated each lyric paragraph in the training data from English to Japanese to English by using English-Japanese and Japanese-English translation methods [9] to create paired data of the back-translated lyrics and the original lyrics. The data is then used to train a Transformer-based back-translated-lyrics-to-lyrics encoder-decoder.

Half-to-Half encoder-decoder (H2H) Inspired by an existing text-to-lyrics encoder-decoder training method [6], we first split each lyrics paragraph in the training data into first and second halves. We then used this split lyrics data to train a Transformer-based encoder-decoder that generates the second half lyrics from the first half lyrics.

Since the above S2L, B2L, and H2H are also Transformer-based encoder-decoders, their parameter settings are the same as for the proposed I2L. Given one input text, five lyrics were generated by each method. The parameter p for Top-p sampling was set to 0.9 and τ was set to 0.4. The generation process stops when the symbol ⟨/P⟩ (end of paragraph) is generated. For this comparison, we did not use the proposed anti-plagiarism method.

[8] `https://huggingface.co/google/pegasus-xsum` for the English summarization. `https://huggingface.co/tsmatz/mt5\_summarize\_japanese` for the Japanese summarization.

[9] `https://huggingface.co/staka/fugumt-en-ja` for the English to Japanese translation. `https://huggingface.co/staka/fugumt-ja-en` for the Japanese to English translation.

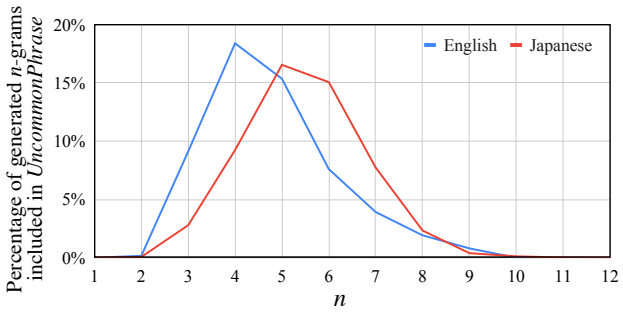


Figure 3. The percentage of generated lyric n-grams that are included in $UncommonPhrase$, a list of phrases that should not be generated (plagiarized). For example, 18.4% at English 4-grams means that among all 4-gram phrases in the generated lyrics, 18.4% are uncommon phrases, though 81.6% are new or common phrases.

4.3 Experimental results

Table 1 indicates that the proposed I2L method had the best PPL in both the English and Japanese experiments and that the NED between the lyrics generated by this method and the input text was the largest ($p_t < 0.05$ based on the paired t-test). As expected, the NEDs were smaller for the S2L and B2L methods, which were trained on paired data where the wording of the input text and lyric pairs was similar. In contrast, although the H2H method can generate lyrics with wording different from the input text, it cannot generate lyrics that are semantically related to the input text like the proposed method can. These findings confirm that image-lyric pairs are more effective than other paired data sets as training data for encoder-decoders generating lyrics that are semantically related to the input text but differ from it in wording.

5. EFFECTIVENESS OF THE PROPOSED ANTI-PLAGIARISM METHOD

We examined whether the absence of the anti-plagiarism method proposed in Section 3.4 results in plagiarizing uncommon phrases found in existing lyrics. In the lyrics generated by the I2L method in Section 4, we calculated the percentage of n-grams included in $UncommonPhrase$.

The results with n ranging from 1 to 12 are shown in Figure 3. The percentage of uncommon 1-grams and 2-grams in the generated lyrics is almost 0%. This indicates that almost all of the generated 1-grams and 2-grams are common phrases used in many existing lyrics, even without the use of the anti-plagiarism method. On the other hand, the percentage of uncommon 3-grams to 8-grams ranged between 3% and 18%. This suggests that many phrases in the generated lyrics may plagiarize if the proposed anti-plagiarism method is not applied. Furthermore, as n increases beyond 9, the n-gram combinations become

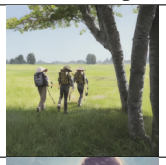
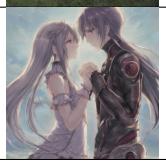
Input text	Image (intermediate representation)	Generated lyrics
A group of explorers are walking through the grass neutral.	⇒ 	⇒ Out in the country, out of sight We've got to get this together right now I'm going out with you today And some day we'll make a lot better way
We meet again I guess our love is forever.	⇒ 	⇒ This is the last time We've been together for long years I'm here with you To be forever yours

Table 2. Examples of lyrics generated by our text-to-lyrics generation method with the anti-plagiarism method.

so vast that the generated n-grams are rarely included in $UncommonPhrase$. These results confirm that our machine learning-based lyrics generation method tends to sample common words, but the generated 3- to 8-gram phrases, even though they are composed of common words, may be uncommon enough to raise suspicion of plagiarism. Using the proposed anti-plagiarism method, in contrast, ensures that uncommon phrases contained in $UncommonPhrase$ are never generated, thereby reducing the risk of plagiarism.

While the proposed anti-plagiarism method is effective, it is important to note that it is not intended to be a foolproof solution that ensures legal compliance. Rather, it is designed to provide a helpful guideline for those who wish to generate original lyrics while reducing the risk of plagiarism. We hope that our approach can contribute to further discussions on a reasonable balance between encouraging creativity and respecting intellectual property rights.

6. QUALITATIVE EVALUATION

Table 2 shows two examples of lyrics generated using the proposed method. Given the input text, our method can generate any number of lines of lyrics, but here four lines are generated by stopping the generation process when four ⟨L⟩ (line break) symbols and the ⟨/P⟩ (end of paragraph) symbol are generated. In the first example, the input text is taken from the SICK dataset [41], while in the second example the input text is taken from lyrics in the RWC Music Database [42]. In both examples, our method can generate lyrics that reflect the content of the input text. In the first example, it generates an image that represents the scene described in the input text and generates corresponding lyrics that reflect the image. In contrast, in the second example, our method generates an image of a person with emotional expression corresponding to the input text and generates lyrics that express the emotion depicted in the image. Other examples can be found in the supplementary material A. [10]

In addition to the quantitative evaluation and the generated examples, we evaluated the similarity between the input text and the generated lyrics through a human evaluator. To prepare the input text in an objective way, we collected the titles of the "Hot 100 Songs" in 2022 on the Billboard year-end charts [11], extracted the first verse from their lyrics, and summarized each verse into a short sentence using ChatGPT. [12] Since 9 songs contained explicit content in either the input text or the generated lyrics, they were excluded for ethical reasons. [13] We then showed the evaluator the input text and the lyrics generated from it, and asked to classify whether the impressions of the two were similar or not. As a result, the impressions of the input text and the generated lyrics were judged to be similar for 52 of the 91 songs, confirming that the proposed method can generate lyrics that express the content of the input text to some extent. In cases where the impressions were classified as dissimilar, most of the input texts contain complex situations or abstract content that is difficult to generate as images. Thus, the limitation of this approach is that it cannot generate lyrics for input texts that are difficult to represent as images. Nevertheless, our method is useful as a writing support tool for many situations where users have intentions that can be represented as images, and is also valuable because it pioneered a novel lyric generation approach. Detailed results of the generated lyrics and the judgments are included in the supplementary material B. [14]

7. CONCLUSION

This paper has described a method for generating lyrics that are similar in meaning to the input text but expressed differently. The contributions of this study are as follows: (1) We proposed a novel two-step pipeline framework. First, we apply text-to-image generation as a text analyzer to extract only the semantic content from the input text. Next, we use our proposed image-to-lyrics encoder-decoder to generate lyrics that capture the semantics of the generated image. (2) We proposed a method to reduce the risk of plagiarism by prohibiting the generation of uncommon phrases in the training data and verified its effectiveness. (3) We quantitatively showed that our proposed method outperforms other methods in generating lyrics for our purpose.

Future work will develop the flexible lyric writing support system incorporating the proposed lyrics generation method.

[10] https://github.com/KentoW/ISMIR2023

[11] https://www.billboard.com/charts/year-end/2022/hot-100-songs/

[12] https://chat.openai.com/chat

[13] As future work, we plan to incorporate a filtering function that uses explicit lyrics detection [43–46].

[14] https://github.com/KentoW/ISMIR2023

8. ACKNOWLEDGMENTS

This work was supported in part by JST CREST Grant Number JPMJCR20D4 and JSPS KAKENHI Grant Number JP20K19878, Japan.

9. REFERENCES

[1] K. Watanabe and M. Goto, "Lyrics information processing: Analysis, generation, and applications," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, 2020, pp. 6–12.

[2] K. Watanabe, Y. Matsubayashi, K. Inui, T. Nakano, S. Fukayama, and M. Goto, "LyriSys: An interactive support system for writing lyrics based on topic transition," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces (ACM IUI)*, 2017, pp. 559–563.

[3] H. G. Oliveira, T. Mendes, and A. Boavida, "Co-PoeTryMe: A co-creative interface for the composition of poetry," in *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*, 2017, pp. 70–71.

[4] H. G. Oliveira, T. Mendes, A. Boavida, A. Nakamura, and M. Ackerman, "Co-PoeTryMe: Interactive poetry generation," *Cognitive Systems Research*, vol. 54, pp. 199–216, 2019.

[5] R. Zhang, X. Mao, L. Li, L. Jiang, L. Chen, Z. Hu, Y. Xi, C. Fan, and M. Huang, "Youling: An AI-assisted lyrics creation system," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, 2020, pp. 85–91.

[6] N. Ram, T. Gummadi, R. Bhethanabotla, R. J. Savery, and G. Weinberg, "Say what? collaborative pop lyric generation using multitask transfer learning," in *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI)*, 2021, pp. 165–173.

[7] L. Zhang, R. Zhang, X. Mao, and Y. Chang, "QiuNiu: A Chinese lyrics generation system with passage-level input," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics - System Demonstrations (ACL)*, 2022, pp. 76–82.

[8] N. Liu, W. Han, G. Liu, D. Peng, R. Zhang, X. Wang, and H. Ruan, "ChipSong: A controllable lyric generation system for Chinese popular song," in *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing)*, 2022, pp. 85–95.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 1–11, 2017.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 674–10 685.

[12] A. Papadopoulos, P. Roy, and F. Pachet, "Avoiding plagiarism in markov sequence generation," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014, pp. 2731–2737.

[13] Q. Feng, C. Guo, F. Benitez-Quiroz, and A. M. Martínez, "When do GANs replicate? on the choice of dataset size," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, 2021, pp. 6681–6690.

[14] T. Nakano, K. Yoshii, and M. Goto, "Musical similarity and commonness estimation based on probabilistic generative models of musical elements," *International Journal of Semantic Computing (IJSC)*, no. 1, pp. 27–52, 2016.

[15] K. Watanabe, Y. Matsubayashi, S. Fukayama, M. Goto, K. Inui, and T. Nakano, "A melody-conditioned lyrics language model," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018, pp. 163–172.

[16] X. Lu, J. Wang, B. Zhuang, S. Wang, and J. Xiao, "A syllable-structured, contextually-based conditionally generation of Chinese lyrics," in *Proceedings of the 16th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, 2019, pp. 257–265.

[17] Y. Chen and A. Lerch, "Melody-conditioned lyrics generation with SeqGANs," in *Proceedings of the 2020 IEEE International Symposium on Multimedia (ISM)*, 2020, pp. 189–196.

[18] Y. Huang and K. You, "Automated generation of Chinese lyrics based on melody emotions," *IEEE Access*, vol. 9, pp. 98 060–98 071, 2021.

[19] X. Ma, Y. Wang, M. Kan, and W. S. Lee, "AI-Lyricist: Generating music and vocabulary constrained lyrics," in *Proceedings of the 29th ACM International Conference on Multimedia (ACM-MM)*, 2021, pp. 1002–1011.

[20] Z. Sheng, K. Song, X. Tan, Y. Ren, W. Ye, S. Zhang, and T. Qin, "SongMASS: Automatic song writing with pre-training and alignment constraint," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021, pp. 13 798–13 805.

[21] G. Barbieri, F. Pachet, P. Roy, and M. D. Esposti, "Markov constraints for generating lyrics with style," in *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)*, vol. 242, 2012, pp. 115–120.

[22] J. Hopkins and D. Kiela, "Automatically generating rhythmic verse with neural networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 168–178.

[23] E. Manjavacas, M. Kestemont, and F. Karsdorp, "Generation of hip-hop lyrics with hierarchical modeling and conditional templates," in *Proceedings of the 12th International Conference on Natural Language Generation (INLG)*, 2019, pp. 301–310.

[24] L. Xue, K. Song, D. Wu, X. Tan, N. L. Zhang, T. Qin, W. Zhang, and T. Liu, "DeepRapper: Neural rap generation with rhyme and rhythm modeling," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021, pp. 69–81.

[25] J. Chang, J. C. Hung, and K. Lin, "Singability-enhanced lyric generator with music style transfer," *Computer Communications*, vol. 168, pp. 33–53, 2021.

[26] O. Vechtomova, G. Sahu, and D. Kumar, "Generation of lyrics lines conditioned on music audio clips," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, 2020, pp. 33–37.

[27] ——, "LyricJam: A system for generating lyrics for live instrumental music," in *Proceedings of the 12th International Conference on Computational Creativity (ICCC)*, 2021, pp. 122–130.

[28] K. Watanabe and M. Goto, "Atypical lyrics completion considering musical audio signals," in *Proceedings of the 27th International Conference on Multimedia Modeling (MMM)*, vol. 12572, 2021, pp. 174–186.

[29] K. Watanabe, Y. Matsubayashi, K. Inui, and M. Goto, "Modeling structural topic transitions for automatic lyrics generation," in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 2014, pp. 422–431.

[30] P. Potash, A. Romanov, and A. Rumshisky, "GhostWriter: Using an LSTM for automatic rap lyric generation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 1919–1924.

[31] M. Ghazvininejad, X. Shi, Y. Choi, and K. Knight, "Generating topical poetry," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, pp. 1183–1191.

[32] H. Fan, J. Wang, B. Zhuang, S. Wang, and J. Xiao, "A hierarchical attention based seq2seq model for Chinese lyrics generation," in *Proceedings of the 16th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, vol. 11672, 2019, pp. 279–288.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[35] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186.

[36] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[37] T. Kudo and Y. Matsumoto, "Japanese dependency analysis using cascaded chunking," in *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, 2002.

[38] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[39] Y. Li and B. Liu, "A normalized Levenshtein distance metric," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[40] C. Manning and H. Schutze, *Foundations of statistical natural language processing*. MIT press, 1999.

[41] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli, "A SICK cure for the evaluation of compositional distributional semantic models," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 216–223.

[42] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Popular, classical and jazz music databases," in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.

[43] H. Chin, J. Kim, Y. Kim, J. Shin, and M. Y. Yi, "Explicit content detection in music lyrics using machine learning," in *Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2018, pp. 517–521.

[44] M. Fell, E. Cabrio, M. Corazza, and F. Gandon, "Comparing automated methods to detect explicit content in song lyrics," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, 2019, pp. 338–344.

[45] E. Egivenia, G. R. Setiawan, S. S. Mintara, and D. Suhartono, "Classification of explicit music content based on lyrics, music metadata, and user annotation," in *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology (SIET)*, 2021, pp. 265–270.

[46] M. Rospocher, "On exploiting transformers for detecting explicit song lyrics," *Entertainment Computing*, vol. 43, p. 100508, 2022.