Atypical Lyrics Completion Considering Musical Audio Signals

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

Abstract

This paper addresses the novel task of lyrics completion for creative support. Our proposed task aims to suggest words that are (1) atypical but (2) suitable for musical audio signals. Previous approaches focused on fully automatic lyrics generation tasks using language models that tend to generate frequent phrases, despite the importance of atypicality for creative support. In this study, we propose a novel vector space model and hypothesize that embedding multimodal aspects (words, draft sentences, and music audio) in a unified vector space contributes to capturing (1) the atypicality of words and (2) the relationships between words and the moods of music audio. To test our hypothesis, we used a large-scale dataset to investigate whether the proposed model suggests atypical words.

1 Introduction

Lyrics are important in conveying emotions and messages in popular music, and the recently increasing popularity of user-generated content on video sharing services makes writing lyrics popular even for novice writers. Lyrics writers, however, unlike the writers of prose text, need to create attractive phrases suitable for the given music. Writing lyrics is thus not an easy job.

Its difficulty has motivated a range of studies for automatic lyrics generation (Oliveira et al., 2007; Potash et al., 2015; Watanabe et al., 2018). For example, Watanabe et al. train a Recurrent Neural Network Language Model (RNN-LM) that generates fluent lyrics while maintaining compatibility between the boundaries of lyrics and melody structures. However, even if LMs generate perfect lyrics, a fully automatic generation system cannot support writers because it ignores their intentions.

In this study, for creative support instead of lyrics generation, we design a lyrics completion task that

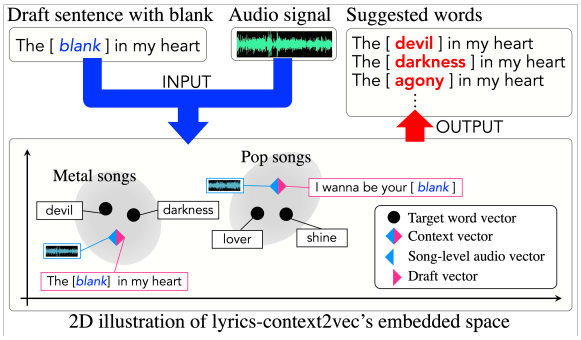

Figure 1: Overview of lyrics completion task.

recommends candidate words for the blank in a given sentence (Fig. 1). Specifically, we focus on the following two properties of lyrics. (1) Lyrics sometimes depend on the *moods* of music audio (e.g., "death" is often used in metal songs) (Watanabe and Goto, 2019). We propose a task in which a system recommends words suitable for the mood of a given song excerpt represented as an audio signal. (2) *Atypicality* is important in writing lyrics; to make lyrics attractive, writers consider both typical and atypical phrases. However, previous study on lyrics generation has used LMs that predict highly frequent (i.e., *typical*) words (Barbieri et al., 2012; Potash et al., 2015; Watanabe et al., 2017, 2018). Creative support systems need to recommend unusual and rare (i.e., *atypical*) words while maintaining the fluency of the sentence.

We therefore propose a multimodal vector space model (VSM), *lyrics-context2vec*, that, given a draft sentence with a blank, suggests atypical words while maintaining the relationship with the mood of the music audio. With lyrics-context2vec, input vectors (i.e., combinations of music audios and draft sentences) and output vectors (i.e., atypical words) are located near each other in a unified high-dimensional vector space (Fig. 1). This model suggests atypical words because we use typical words

as negative examples in its training.

The contributions of this study are summarized as follows: (1) We propose, for creative support, a novel multimodal vector space model that captures the relationship between atypical words and the mood of music audio. (2) We demonstrate that our model suggests words suitable for the mood of the input musical audio signal. (3) We demonstrate that our model suggests words more atypical than those suggested by RNN-LMs.

This paper is a short version of our MMM 2021 paper (Watanabe and Goto, 2021).

2 Lyrics-audio data

To model the relationship between lyrics and moods of music audio, we obtained 458,572 songs, each consisting of a pair comprising a text file of English lyrics and an audio file of a music excerpt[1]. Here each text file contains all sentences of the lyrics of a song, and each audio file is a music excerpt (30 sec) that was collected from the Internet and provided for trial listening. We embedded the moods of audio signals as well as the words of lyrics into a unified vector space without using coarse metadata (e.g., genre tags).

2.1 Bag-of-Audio-Words

To represent the mood feature of a short music excerpt, we use a discrete symbol called an *audio-word* (Liu et al., 2010). The bag-of-audio-words (BoAW) creation procedure is as follows. (1) Each music excerpt is downsampled to 22,050 Hz. (2) *LibROSA*, a python package for music and audio analysis, is used to extract 20-dimensional mel-frequency cepstral coefficients (MFCCs) with the FFT size of 2048 samples and the hop size of 512 samples. This result is represented as an MFCC matrix (20×1280). (3) The MFCC matrix is divided into 128 submatrices (20×10) without overlap. (4) To create a vocabulary of k audio-words, we apply the k-*means*++ algorithm to all the divided MFCCs of all the songs. In other words, each k-th cluster corresponds to an audio-word (aw). In this study we made 3000 audio-words.

3 Atypical word completion model

We propose a multimodal vector space model *lyrics-context2vec* that, given a music audio signal and a draft sentence with a blank, suggests atypical words while maintaining the relationship with the mood of the music audio. Specifically, lyrics-context2vec suggests the best N atypical words $w^1, \ldots, w^N$ that could fit with the context. Here we assume two types of contexts: (1) the words on the left and right sides of the blank and (2) the BoAW converted from the audio signal.

There are two technical problems in recommending atypical words suitable for the music audio. First, since most LMs learn to predict highly frequent words, it is hard to suggest atypical words that are important for creative support. Second, how to model the relationship between words and musical audio signals is not obvious.

To address the first problem, we focus on the negative sampling strategy in word2vec (Mikolov et al., 2013). This strategy was proposed for the purpose of approximation because computation of loss function is time-consuming. We, however, use negative sampling for the purpose of suppression of typical word recommendation because we want to suggest *atypical* words for creative support. Since negative examples are drawn from the distribution of highly frequent words, it is expected that input vectors of contexts are located far from vectors of typical words. It is not obvious that the negative sampling contributes to suggesting atypical words.

To address the second problem, we utilize the mechanism of *lyrics2vec* proposed by Watanabe and Goto. In lyrics2vec, co-occurring audio-words and lyric words are located near each other under the assumption that *some words of lyrics are written depending on the musical audio signal*.

3.1 Model construction

Lyrics-context2vec is based on lyrics2vec and context2vec (Melamud et al., 2016). Formally, context2vec is a vector space model that encodes left draft words $w_1, \ldots, w_{t-1}$ and right draft words $w_{t+1}, \ldots, w_T$ into latent vectors z_1 and z_2, respectively, using two Recurrent Neural Networks (RNNs). Then the target word vector $v(w_t)$ and a vector that is nonlinearly transformed from the latent vectors are mapped closely into a unified vector space. The loss function of context2vec E_{c2v} is defined so that the inner product of the target word vector $v(w_t)$ and the nonlinearly transformed vector is maximized:

$$E_{c2v} = -\log \left(v(w_t)^\top \cdot \mathrm{MLP}([z_1; z_2]) \right) - \sum_{s=1}^{S} \log \left(-v(w_s^\prime)^\top \cdot \mathrm{MLP}([z_1; z_2]) \right); \quad (1)$$

[1] Lyrics were provided by a lyrics distribution company.

where $\sigma(\cdot)$ is a sigmoid function. To obtain an x-dimensional word vector representation, we define an embedding function $v(\cdot)$ that maps the target word to an x-dimensional vector. S is the number of negative examples $w_s^{\emptyset}$. $[z_1; z_2]$ denotes a concatenation of latent vectors z_1 and z_2. $\mathrm{MLP}(\cdot)$ stands for multilayer perceptron. In this loss function, negative examples $w_s^{\emptyset}$ are sampled from the distribution $P(w_s^{\emptyset}) = D(w_s^{\emptyset})^{0.75} / \sum_{w' \in V} (D(w'^{\emptyset})^{0.75})$ where V is the vocabulary and $D(w^{\emptyset})$ is the document frequency of a word $w^{\emptyset}$. In other words, since frequent words tend to be sampled as negative examples, we expect that a draft sentence vector and the vector of highly frequent typical words are located far away from each other. When computing word completion, our system displays target words with high cosine similarity to the input context vector $\mathrm{MLP}([z_1; z_2])$.

Then we extend context2vec to suggest atypical words suitable for both the music audio and the draft sentence by embedding three aspects (i.e., target words, draft sentences, and song-level audio). We concatenate song-level audio and draft vectors and define the loss function E so that the concatenated vector $[z_1; z_2; \frac{1}{M} \sum_{m=1}^{M} u(aw_m)]$ is located close to the target word vector $v(w)$:

$$E = -\log \left(v(w_t)^{\top} \cdot [z_1; z_2; \frac{1}{M} \sum_{m=1}^{M} u(aw_m)] \right) - \sum_{s=1}^{S} \log \left(-v(w_s^{\emptyset})^{\top} \cdot [z_1; z_2; \frac{1}{M} \sum_{m=1}^{M} u(aw_m)] \right); \quad (2)$$

where we define the dimension of draft vectors $z_1; z_2$ as d and define an embedding function $u(\cdot)$ that maps the context word/audio-word to a d-dimensional vector. M is the number of audio-words in the song. We define the average of audio-word vectors as a song-level audio vector.

4 Experiments

To evaluate whether lyrics-context2vec can suggest (1) atypical words and (2) words suitable for music audio, we designed word completion tasks. The input of these tasks is $T - 1$ draft words $w_1; \ldots; w_{t-1}; w_{t+1}; \ldots; w_T$ of each sentence in a test song. Therefore the model needs to fill in the t-th blank with a word. We used the following $Score$ to evaluate the performance of models in the lyrics completion task: $Score@N = \sum_{r \in R} \mathbb{1}(r \in \{h^1; \ldots; h^N\}) / |R|$, where r denotes the correct word and $|R|$ is the number of blanks in the test data. $h^1; \ldots; h^N$ are the top N suggested words. $\mathbb{1}(\cdot)$ is the indicator function. In this study we calculated $Score@N$, with N ranging from 1 to 20 under the assumption that our support system suggests 20 words to users.

Here it is important to define which word in each sentence is the correct word r. We defined four types of correct answers:

Typicality We defined a randomly chosen word in each sentence of the test song as the correct word r. In this metric, high-frequency words tend to be chosen as the correct answer. In other words, this metric is a measure of typical word completion.

Atypicality We first calculated the document frequency of words of the test song and then defined the minimum-document-frequency word in each sentence as the correct word. This metric is a measure of atypical word completion.

Music+Typicality In each sentence of the test song, we extracted the word most similar to the music audio of the song by using the pre-trained lyrics2vec that was proposed by Watanabe and Goto. If the document frequency of the extracted word was more than 1,000, we defined this word as the correct word for the sentence and did not use the other words. This metric is a measure of prediction of *typical* words suitable for the music audio.

Music+Atypicality We extracted the word most similar to the music audio of the song as with Music+Typicality. If the document frequency of the extracted word was less than or equal to 1,000, we defined this word as the correct word for the sentence and did not use the other words. This metric is a measure of prediction of *atypical* words suitable for the music audio of the song.

4.1 Comparison methods

To investigate the effect of our lyrics-context2vec, we compared the following four models. (1) *Bi-RNN-LM*, a bidirectional RNN-LM trained with lyrics without audio. (2) *Encoder-Decoder*, a Bi-RNN-LM in which the song-level audio vector $\frac{1}{M} \sum_{m=1}^{M} u(aw_m)$ is input to the initial RNN state. (3) *Context2vec* (Melamud et al., 2016). (4) *Lyrics-context2vec*, the proposed model. The RNN-LMs (Bi-RNN-LM and Encoder-Decoder) predict words with high predicted probability in the blank, and the VSMs (context2vec and lyrics-context2vec) predict the most similar words in the blank. Examples of words suggested by the models are available at a web page (https://kentow.github.io/

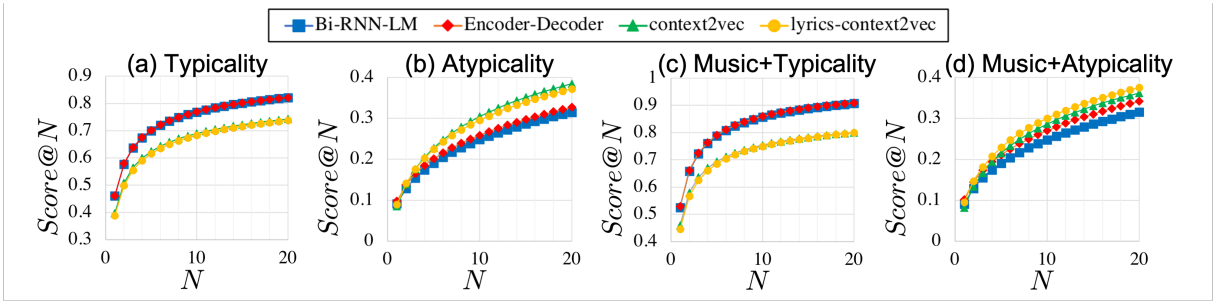

Figure 2: Results of the lyrics completion tasks.

nlp4musa2021/).

4.2 Settings

We randomly split our dataset into 80-10-10% divisions to construct the training, validation, and test data. In all models, we utilized Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) as the RNN layer. We chose $d = 300$ for the dimension of the audio-word vector $u(\cdot)$ and the dimension of the LSTM hidden state z. We chose $x = 900$ for the dimension of the target word vector $v(\cdot)$. We used negative sampling with $S = 20$ negative examples. We used Adam (Kingma and Ba, 2015) with an initial learning rate of 0.001 for parameter optimization. The model used for testing was the one that achieved the best Music+Atypicality score on the validation set.

4.3 Results

Figure 2(a) shows the result of the typical word completion task (Typicality). As shown in this figure, RNN-LMs achieved higher scores than VSMs. This is because the RNN-LMs are trained to maximize the probability of generating highly frequent phrases. Interestingly, we can see that there is no difference between the scores of Bi-RNN-LM and Encoder-Decoder. This indicates that audio information does not contribute to predicting typical words. Typical words were thus expected to be correlated with draft sentences rather than audio.

Regarding the task of predicting the typical words suitable for music audio (Fig. 2 (c)), we can observe results similar to those for the task Typicality. This reinforces the fact that typical words can be predicted from only the draft sentence, without using audio information.

Regarding the atypical word completion (Fig. 2 (b)), VSMs achieved higher scores than RNN-LMs. This indicates that negative sampling contributes to suppression of typical word completion. Overall, for atypical word completion tasks it is desirable to use a VSM with negative sampling rather than a LM aimed at generating typical phrases.

Regarding the main task Music+Atypicality (Fig. 2 (d)), lyrics-context2vec predicted atypical words suitable for music audio better than any of the other models. This means that our model captures both the atypicality and the relationship between a music audio and words simultaneously. Moreover, we can see that lyrics-context2vec performs better than context2vec and that Encoder-Decoder performs better than Bi-RNN-LM. This indicates that using audio information contributes to suggesting atypical words suitable for the music audio.

5 Conclusion

We proposed lyrics-context2vec, a multimodal vector space model that suggests atypical but appropriate words for the given music audio and draft sentence. In the vector space of lyrics-context2vec, a vector corresponding to an atypical word in a song and a song-level audio vector corresponding to an audio excerpt of the song are located near each other. We trained the models to suggest atypical words by embedding the highly frequent word vector away from the song-level audio vector.

In the experiment, we used a large-scale dataset to investigate whether the proposed model suggests atypical but appropriate lyrics. Several findings were obtained from experiment results. One is that the negative sampling contributes to suggesting atypical words. Another is that embedding audio signals contributes to suggesting words suitable for the mood of the music audio. We conclude that embedding multiple aspects into a vector space contributes to capturing atypicality and relationship with audio.

Acknowledgements. The authors appreciate SyncPower Corporation for providing lyrics data. This work was supported in part by JST ACCEL Grant Number JPMJAC1602, JST CREST Grant Number JPMJCR20D4, and JSPS KAKENHI Grant Number 20K19878, Japan.

References

Gabriele Barbieri, François Pachet, Pierre Roy, and Mirko Degli Esposti. 2012. Markov constraints for generating lyrics with style. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 115–120.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Yang Liu, Wan-Lei Zhao, Chong-Wah Ngo, Chang-Sheng Xu, and Han-Qing Lu. 2010. Coherent bag-of audio words model for efficient large-scale video copy detection. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 89–96.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, pages 3111–3119.

Hugo R. Gonçalo Oliveira, F. Amialcar Cardoso, and Francisco C. Pereira. 2007. Tra-la-lyrics: an approach to generate text based on rhythm. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pages 47–55.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. GhostWriter: Using an LSTM for automatic Rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1924.

Kento Watanabe and Masataka Goto. 2019. Query-by-Blending: A music exploration system blending latent vector representations of lyric word, song audio, and artist. In *Proceedings of the 20th annual conference of the International Society for Music Information Retrieval*, pages 144–151.

Kento Watanabe and Masataka Goto. 2021. Atypical lyrics completion considering musical audio signals. In *Proceedings of the 27th International Conference on Multimedia Modeling*, volume 12572, pages 174–186.

Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. 2018. A melody-conditioned lyrics language model. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 163–172.

Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, Tomoyasu Nakano, Satoru Fukayama, and Masataka Goto. 2017. LyriSys: An interactive support system for writing lyrics based on topic transition. In *Proceedings of the 22nd Annual Meeting of the Intelligent User Interfaces Community*, page 559—563.