Atypical Lyrics Completion Considering Musical Audio Signals

Kento Watanabe(✉) and Masataka Goto(✉)

National Institute of Advanced Industrial Science and Technology (AIST),
Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki, Japan
{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 (e.g., "I love you"), despite the importance of atypicality for creative support. In this study, we propose a novel vector space model with negative sampling strategy and hypothesize that embedding multimodal aspects (words, draft sentences, and musical audio signals) 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 multimodal vector space model suggests atypical words. Several findings were obtained from experiment results. One is that the negative sampling strategy contributes to suggesting atypical words. Another is that embedding audio signals contributes to suggesting words suitable for the mood of the provided music audio.

Keywords: Lyrics completion · Natural language processing · Multi-modal embedding

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. Thus, writing lyrics is not an easy job.

This difficulty has motivated a range of studies for computer-assisted lyrics writing [9,10,13]. For example, Watanabe et al. (2018) train a Recurrent Neural Network Language Model (RNN-LM) that generates fluent lyrics while maintaining compatibility between the boundaries of lyrics and melody structures. Those studies, however, aim to generate lyrics fully automatically. Even if language models generate perfect lyrics, a fully automatic generation system cannot support writers because it ignores their intentions.

J. Lokoč et al. (Eds.): MMM 2021, LNCS 12572, pp. 174–186, 2021.
https://doi.org/10.1007/978-3-030-67832-6_15

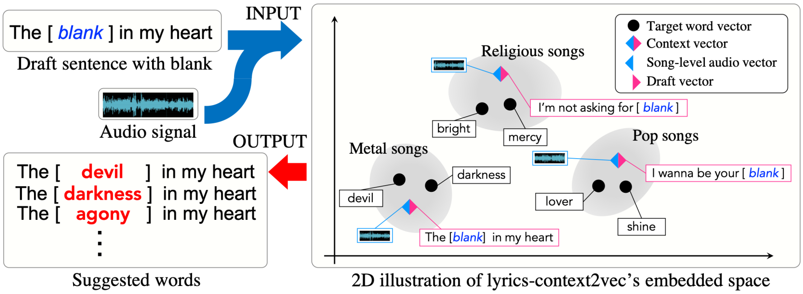
The [blank] in my heart
Draft sentence with blank
INPUT
Audio signal
OUTPUT
The [devil] in my heart
The [darkness] in my heart
The [agony] in my heart
⋮
Suggested words
Target word vector
Context vector
Song-level audio vector
Draft vector
Religious songs
I'm not asking for [blank]
bright
mercy
Metal songs
Pop songs
darkness
devil
I wanna be your [blank]
lover
shine
The [blank] in my heart
2D illustration of lyrics-context2vec's embedded space

Fig. 1. Overview of lyrics completion task. Our model predicts words similar to the input draft sentences and musical audio signals.

In this study, for creative support instead of lyrics generation, we design a lyrics completion task to recommend 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., 'love" is often used in ballad songs, "kill" is often used in metal songs). However, no text completion system suggests words that are appropriate for the mood of music audio. Thus, 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 research on automatic lyrics generation has used language models that predict highly frequent (i.e., *typical*) words. Creative support systems need also 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 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.

2 Related Work

We first discuss the related work on vector space models focusing on music. Weston et al. (2011) proposed a model for embedding acoustic signals, artist tags, and social tags in a unified vector space [15]. They designed several relationships based on assumptions such as "*the songs created by an artist are correlated*". Lopopolo and van Miltenburg (2015) and Karamanolakis et al. (2016) used a bag-of-audio-words (BoAW) approach and vectorized audio words by utilizing social tags [3,6]. Their studies shared with ours the motivation of embedding multiple aspects in vector spaces but dealt only with audio and metadata without lyrics even though lyrics are an important element that conveys messages and emotions of music. Yu et al. (2019) and Watanabe and Goto (2019) embedded different aspects (i.e., lyric word and song sound) into a unified vector space [12,16]. To the best of our knowledge, there has been no study modeling the relationship between a draft sentence of lyrics (a sentence with a blank) and music audio simultaneously.

We then discuss the related work on automatic lyrics generation. Barbieri et al. (2012), Potash et al. (2015), and Watanabe et al. (2018) proposed models that generate lyrics under a range of constraints provided in terms of topic, rhyme, rhythm, part-of-speech, and so on [1,10,14]. Oliveira et al. (2007) and Watanabe et al. (2018) proposed language models that generate singable lyrics based on melody segment positions [9,13]. However, they used language models that tend to generate typical words and did not focus on the atypicality of lyrics.

This paper thus can be considered the first work that tackles the novel lyrics completion task by dealing with both of those relationship and atypicality.

3 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 short popular-music excerpt (30 s, 44.1 kHz) that was collected from the Internet and originally provided for trial listening. In this study, we embedded the moods of audio signals as well as the words of lyrics directly into a unified vector space without using metadata such as genre tags because those tags are too coarse. The total duration of all the excerpted audio files was more than 159 days.

3.1 Bag-of-Audio-Words

To represent the mood feature of a short music excerpt, we use a discrete symbol called an *audio-word (aw)* [5]. 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

[1] In our experiments, English lyrics text were provided by a lyrics distribution company.

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.

4 Atypical Word Completion Model Considering Audio

In this section 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 N-best atypical words $w^1, ..., w^N$ that could fit with the context. Here we assume three types of contexts: (1) the words on the left side of the blank, (2) the words on the right side of the blank, and (3) the BoAW converted from the audio signal.

This model is useful for creative support because it helps a user (lyrics writer) come up with new ideas for a song by looking at atypical words suitable for it. There are two technical problems in recommending atypical words suitable for the music audio. First, since most statistical models (e.g., RNN-LM) 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 [8]. 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 [12]. 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* (e.g., words about love tend to be used in ballad songs).

4.1 Model Construction

Lyrics-context2vec is based on lyrics2vec and context2vec [7]. Formally, context2vec is a vector space model that encodes left draft words $w_1, ..., w_{t-1}$ and right draft words $w_{t+1}, ..., w_T$ into latent vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, respectively, using two Recurrent Neural Networks (RNNs). Then the target word vector $\boldsymbol{v}(w_t)$ 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

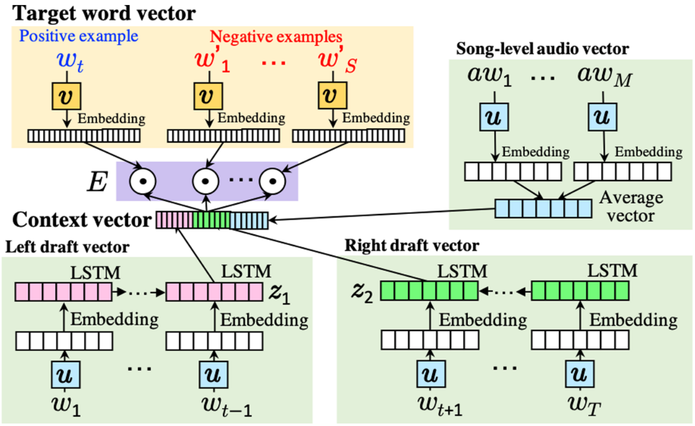

Fig. 2. Overview of the proposed lyrics-context2vec model.

defined so that the inner product of the target word vector $\boldsymbol{v}(w_t)$ and the non-linearly transformed vector is maximized:

$$E_{c2v} = -\log\sigma\Big(\boldsymbol{v}(w_t)^\top \cdot \mathrm{MLP}([\boldsymbol{z}_1, \boldsymbol{z}_2])\Big) - \sum_{s=1}^{S} \log\sigma\Big(-\boldsymbol{v}(w'_s)^\top \cdot \mathrm{MLP}([\boldsymbol{z}_1, \boldsymbol{z}_2])\Big), \quad (1)$$

where $\sigma(\cdot)$ is a sigmoid function. To obtain an x-dimensional word vector representation, we define an embedding function $\boldsymbol{v}(\cdot)$ that maps the target word to an x-dimensional vector. S is the number of negative examples w'_s. $[\boldsymbol{z}_1, \boldsymbol{z}_2]$ denotes a concatenation of latent vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. MLP$(\cdot)$ stands for multilayer perceptron (MLP). In this loss function, negative examples w'_s are sampled from the distribution $P(w'_s) = D(w'_s)^{0.75} / \sum_{w' \in V}(D(w')^{0.75})$ where V is the vocabulary and $D(w')$ is the document frequency of a word w'. 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 MLP $([\boldsymbol{z}_1, \boldsymbol{z}_2])$.

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). The structure of this extended model is illustrated in Fig. 2. We concatenate song-level audio and draft vectors and define the loss function E so that the concatenated vector $[\boldsymbol{z}_1, \boldsymbol{z}_2, \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{u}(aw_m)]$ is located close to the target word vector $\boldsymbol{v}(w)$:

$$\begin{aligned} E = &-\log\sigma\Big(\boldsymbol{v}(w_t)^\top \cdot [\boldsymbol{z}_1, \boldsymbol{z}_2, \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{u}(aw_m)]\Big) \\ &- \sum_{s=1}^{S} \log\sigma\Big(-\boldsymbol{v}(w'_s)^\top \cdot [\boldsymbol{z}_1, \boldsymbol{z}_2, \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{u}(aw_m)]\Big), \end{aligned} \quad (2)$$

where we define the dimension of draft vectors $\boldsymbol{z}_1, \boldsymbol{z}_2$ as d and define an embedding function $\boldsymbol{u}(\cdot)$ that maps the context word/audio-word to a d-dimensional vector. Thus the dimension of target word vectors $\boldsymbol{v}(\cdot)$ is $x = 3d$. M is the number of audio-words in the song. We define the average of audio-word vectors as a song-level audio vector.

In the original context2vec the concatenated vector connects to an MLP, but lyrics-context2vec uses the concatenated vector directly without an MLP. This is because it is useful for lyrics writers to be able to flexibly change the contexts (draft sentence and music audio) to obtain the suggested words. For example, even if a user provides only the left draft vector $\boldsymbol{z}_1$, the lyrics-context2vec can suggest appropriate words by computing the cosine similarity between the word vector $\boldsymbol{v}(w_t)$ and the concatenated vector $[\boldsymbol{z}_1, \mathbf{0}, \mathbf{0}]$[2]. Models with an MLP cannot provide this flexibility since all the three vectors are always required to compute an MLP. We therefore do not use an MLP in the proposed model.

5 Experiments

In order 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 $w_1, ..., w_{t-1}, w_{t+1}, ..., w_T$ 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:

$$Score@N = \frac{\sum_{r \in R} \mathbb{1}(r \in \{h^1, ..., h^N\})}{|R|}, \tag{3}$$

where r denotes the correct word and $|R|$ is the number of blanks in the test data. $h^1, ..., h^N$ are the top N words suggested by the model. $\mathbb{1}(\cdot)$ 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 designed 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

[2] $\mathbf{0}$ is the zero vector that has all components equal to zero.

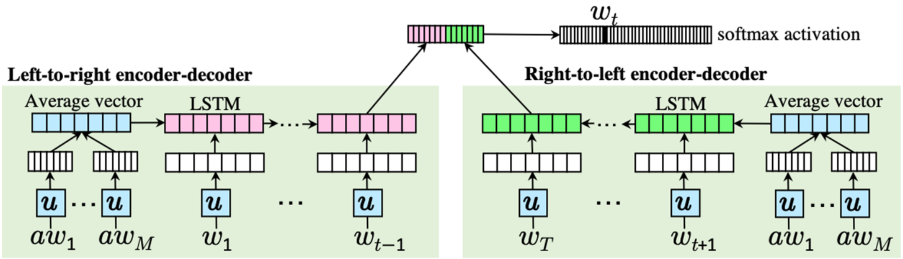

Fig. 3. Overview of Encoder-Decoder model.

was proposed by Watanabe and Goto [12]. 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 of the song.

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.

5.1 Comparison Methods

To investigate the effect of our lyrics-context2vec, we compared the following four models. (1) *Bi-RNN-LM*, a standard bidirectional RNN-LM trained with lyrics without audio information. (2) *Encoder-Decoder*, a Bi-RNN-LM in which the song-level audio vector $\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{u}(aw_m)$ is input to the initial RNN state (Fig. 3). (3) *Context2vec*, a context2vec [7] without a multilayer perceptron (MLP). (4) *Lyrics-context2vec*, the proposed model.

The RNN-LM type models (Bi-RNN-LM and Encoder-Decoder) predict words with high predictive probability in the blank, and the VSM-type models (context2vec and lyrics-context2vec) predict the most similar words in the blank.

5.2 Settings

Dataset. We randomly split our dataset into 80-10-10% divisions to construct the training, validation, and test data. From those, we used the words whose frequency was more than 20 and converted the others to a special symbol ⟨unk⟩.

Parameters. In all models, we utilized Long Short-Term Memory (LSTM) [2] as the RNN layer. We chose $d = 300$ for the dimension of the audio-word vector $\boldsymbol{u}(\cdot)$ and the dimension of the LSTM hidden state $\boldsymbol{z}$. We chose $x = 900$ for the dimension of the target word vector $\boldsymbol{v}(\cdot)$. We used negative sampling with $S = 20$ negative examples. We used a categorical cross-entropy loss for outputs of RNN-LM type models. We used Adam [4] with an initial learning rate of 0.001 for parameter optimization and used a mini-batch size of 100. Training was run for 10 epochs, and the model used for testing was the one that achieved the best Music+Atypicality score on the validation set. In this study, we utilized the pre-trained lyrics2vec [12] for Music+Typicality and Music+Atypicality; this lyrics2vec was trained with the same parameters as lyrics-context2vec.

5.3 Results

Figure 4(a) shows the result of the typical word completion task (Typicality). As shown in this figure, RNN-LM type models achieved higher scores than VSM type models. This is because the RNN-LM type models 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. We speculated that typical words are strongly correlated with draft sentences rather than audio.

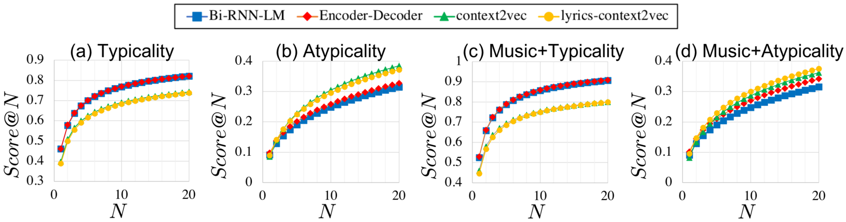

Fig. 4. Results of the lyrics completion tasks.

Regarding the task of predicting the typical words suitable for music audio (Fig. 4(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. 4(b)), VSM type models achieved higher scores than RNN-LM type models. 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 language model aimed at generating highly frequent phrases.

Regarding the main task Music+Atypicality (Fig. 4(d)), lyrics-context2vec predicted atypical words suitable for music audio better than all 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 Encoder-Decoder performs better than Bi-RNN-LM. This indicates that using audio information contributes to suggesting atypical words suitable for the music audio.

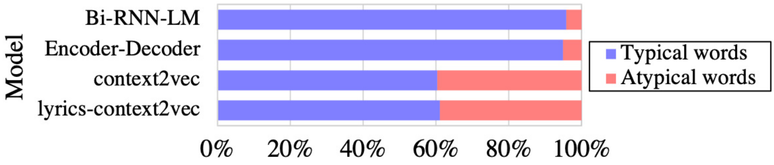

Fig. 5. Ratio of the suggested typical/atypical words.

Table 1. Effect of multilayer perceptron (MLP).

Score@20	The proposed lyrics-context2vec w/o MLP	lyrics-context2vec with MLP
Typicality	0.738	**0.742**
Atypicality	**0.372**	0.351
Music+Typicality	0.799	**0.807**
Music+Atypicality	**0.376**	0.351

5.4 Ratio of Suggested Atypical Words

In all completion tasks, we calculated how many of the top 20 words suggested by each model were typical or atypical. We calculated the document frequency of words and assumed that the top 10% of them are typical words (i.e., the remaining 90% are atypical words).

Figure 5 shows the ratio of typical/atypical words suggested by each model. As we can see in this figure, VSM type models suggested many more atypical words than RNN-LM type models did. This result confirms our intuition that negative sampling contributes to suppression of typical word completion while RNN-LM type models maximize the probability of predicting typical words.

6 Discussions

6.1 Effect of Multilayer Perceptron (MLP)

For the purpose of developing a support system that allows the user to flexibly change the contexts (draft sentence and music audio), we omitted the MLP from our lyrics-context2vec even though an MLP is used in the original context2vec. Here we investigate whether excluding the MLP has a negative impact on word completion tasks. To check the effect of an MLP, we compared the performance of lyrics-context2vec with that of the model with an MLP.

Table 1 summarizes the results. Regarding typical word completion tasks Typicality and Music+Typicality, the model without an MLP achieved almost the same performance as the model with an MLP. Interestingly, the model without an MLP improved the performance of atypical word completion tasks Atypicality and Music+Atypicality. We thus confirmed that excluding the MLP does not have a negative impact for our purposes.

Table 2. The suggested words for draft sentences and songs. We highlight atypical words in bold. Here we calculated the document frequency of words and assumed that the top 10% of them are typical words and the remaining 90% are atypical words. Words are shown in descending order of similarity or prediction probability.

Draft	Music audio	Model	Suggested words
I [ ] you	No audio	Bi-RNN-LM	love, need, know, want, remember, thought, miss, promise, understand, believe, see, like, think, tell, told, **appreciate**, worship, wanted, loved, hate
		context2vec	love, **appreciate**, adore, **implore**, **guarantee**, promise, followed, miss, want, worship, **trusted**, thank, **approached**, believed, need, remember, watched, **entertain**, **assure**, recognize
	Killing Time (Metal)	Encoder-Decoder	**despise**, know, want, thought, remember, followed, wish, **await**, believe, need, love, **summon**, watched, promise, understand, will, **suffocate**, **defy**, **implore**, **destroyed**
		lyrics-context2vec (the proposed model)	**despise**, **await**, **trusted**, **appreciate**, remember, followed, worship, **implore**, believed, thank, hate, **reject**, **assure**, adore, promised, **consume**, **warned**, **possess**, **beckon**, **destroyed**
	Amazing Grace (Pop)	Encoder-Decoder	love, want, need, adore, know, remember, thought, believe, hear, miss, promise, found, will, worship, thank, followed, understand, have, loved, believed
		lyrics-context2vec (the proposed model)	adore, **appreciate**, love, **overheard**, promise, remember, **enfold**, **await**, **forsake**, promised, **hypnotize**, recognize, need, followed, missed, **surround**, thank, follow, deliver, thought
The [ ] in my heart	No audio	Bi-RNN-LM	pain, deep, place, thunder, fire, devil, beating, voices, hole, poison, wind, sun, darkness, burning, love, feeling, song, world, beat, light
		context2vec	**tremors**, pain, devil, **sparkles**, **pounding**, **dagger**, deep, echo, magic, poison, **conflicts**, echoes, diamonds, **toxins**, hunger, hole, **bloodlust**, burning, demon, **blackness**
	Killing Time (Metal)	Encoder-Decoder	pain, deep, burning, fire, hole, world, dead, darkness, feeling, beauty, devil, silence, drowning, shadows, words, dream, demons, power, wind, thunder
		lyrics-context2vec (the proposed model)	hole, devil, burning, emptiness, **blackness**, pain, darkness, demons, **dagger**, hatred, **tremors**, void, fire, **agony**, holes, **essence**, **coldness**, **plague**, **needles**, deep
	Amazing Grace (Pop)	Encoder-Decoder	deep, pain, dream, wind, song, thunder, fire, tears, music, feeling, stars, burning, sunshine, silence, hole, darkness, love, answer, beauty, words
		lyrics-context2vec (the proposed model)	**pains**, ringing, **brightness**, deep, **cuckoo**, angels, **teardrops**, **ache**, roses, **falcon**, music, wind, **waltzes**, troubles, pain, **lump**, **birdie**, melody, **elements**, devil

6.2 Examples of Suggested Words

In the results of lyrics completion tasks, we observed that our lyrics-context2vec suggests an atypical word suitable for the provided music audio. In order to interpret this observation intuitively, we investigated words suggested when draft sentences were fixed and the input music audio was changed. Table 2 shows the top 20 words suggested by each model. In this table, atypical words whose document frequency is among the lowest 90% are shown in bold. We can see that the bolded rare words (e.g., “guarantee” and “entertain”) appear more often in word sets suggested by VSM type models than in word sets suggested by RNN-LM type models. This observation supports our claim that negative sampling suppresses typical word completion.

Regarding the calm song *Amazing Grace*, lyrics-context2vec and Encoder-Decoder tended to suggest emotional and positive words (e.g., “missed” and “brightness”). On the other hand, regarding the metal song *Killing Time*, lyrics-context2vec and Encoder-Decoder tended to suggest explicit and negative expressions (e.g., “destroyed” and “darkness”). This indicates that both RNN-LM and VSM type models with song-level audio vectors successfully suggest words suitable for the mood of the provided music audio. These results are consistent with the result of the lyrics completion tasks in Sect. 5. The audio and words used in this table are available at an anonymized web page (https://kentow.github.io/mmm2021/index.html).

7 Conclusion and Future Work

This paper addresses the novel task of lyrics completion for creative support. Our proposed task aims to suggest words that are (1) atypical and (2) suitable for the musical audio signal. Previous work focused on fully automatic lyrics generation using language models that tend to predict highly frequent phrases (e.g., “I love you”), despite the importance of atypicality in creative support.

In this study, 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. Moreover, we trained the models to suggest atypical words by embedding the highly frequent word vector away from the song-level audio vector. No previous study has ever conducted such an analysis of the word completion task focusing on atypicality and relationship with music audio.

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

We plan to incorporate the proposed lyrics-context2vec model into a writing support system and conduct a user study evaluating that system. We also plan to investigate the behavior of our method when using powerful language models such as Transformers [11] instead of LSTMs. Future work will also include application of this model to different types of texts in which atypical words are effective, such as poetry and advertising slogans.

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

References

1. Barbieri, G., Pachet, F., Roy, P., Esposti, M.D.: Markov constraints for generating lyrics with style. In: Proceedings of the 20th European Conference on Artificial Intelligence, pp. 115–120 (2012)
2. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
3. Karamanolakis, G., Iosif, E., Zlatintsi, A., Pikrakis, A., Potamianos, A.: Audio-based distributional representations of meaning using a fusion of feature encodings. In: INTERSPEECH, pp. 3658–3662 (2016)
4. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015) (2015)
5. Liu, Y., Zhao, W.L., Ngo, C.W., Xu, C.S., Lu, H.Q.: 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, pp. 89–96 (2010)
6. Lopopolo, A., van Miltenburg, E.: Sound-based distributional models. In: Proceedings of the 11th International Conference on Computational Semantics, pp. 70–75 (2015)
7. Melamud, O., Goldberger, J., Dagan, I.: context2vec: Learning generic context embedding with bidirectional LSTM. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, pp. 51–61 (2016)
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems, pp. 3111–3119 (2013)
9. Oliveira, H.R.G., Cardoso, F.A., Pereira, F.C.: Tra-la-lyrics: an approach to generate text based on rhythm. In: Proceedings of the 4th International Joint Workshop on Computational Creativity, pp. 47–55 (2007)
10. Potash, P., Romanov, A., Rumshisky, A.: GhostWriter: using an LSTM for automatic Rap lyric generation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1919–1924 (2015)
11. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, pp. 5998–6008 (2017)
12. Watanabe, K., Goto, M.: 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, pp. 144–151 (2019)
13. Watanabe, K., Matsubayashi, Y., Fukayama, S., Goto, M., Inui, K., Nakano, T.: 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, pp. 163–172 (2018)

14. Watanabe, K., Matsubayashi, Y., Inui, K., Nakano, T., Fukayama, S., Goto, M.: 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, pp. 559–563 (2017)
15. Weston, J., Bengio, S., Hamel, P.: Multi-tasking with joint semantic spaces for large-scale music annotation and retrieval. J. New Music Res. **40**(4), 337–348 (2011)
16. Yu, Y., Tang, S., Raposo, F., Chen, L.: Deep cross-modal correlation learning for audio and lyrics in music retrieval. ACM Trans. Multimed. Comput. Commun. Appl. **15**(1), 1–16 (2019)