Syllable-level lyrics generation from melody exploiting character-level language model

Zhe Zhang[1], Karol Lasocki[2], Yi Yu[1], Atsuhiro Takasu[1]

National Institute of Informatics, SOKENDAI[1]
Aalto University[2]

{zhe, yiyu, takasu}@nii.ac.jp, karolasocki@gmail.com

Introduction

Generating lyrics from a given melody is a subjective and creativity-driven process that does not have a definitive correct answer. Recognizing the importance of subjective and creativity-driven generation processes is essential for advancing the development of AI. In this work, we explore the generation of lyrics from simplified symbolic melodies consisting of 20 notes. Our aim is to maintain the alignment between the syllables of the lyrics and the corresponding melody notes during the inference stage. To achieve this, we propose a melody-encoder-syllable-decoder Transformer architecture, which generates syllables sequentially in accordance with the melody.

Main Contributions

1. Training a melody-encoder-syllable-decoder Transformer model to generate lyrics syllable by syllable, ensuring semantic correlation with individual notes in the melody.
2. Exploiting the fine-tuned character-level pre-trained language models for refining candidate syllables generated by the Transformer decoder to ensure the coherence and correctness in the generated lyrics, overcoming the difficulty of unavailable pre-trained syllable-level language models.
3. Designing a beam search and re-ranking technique to integrate the fine-tuned language model with the Transformer decoder to predict re-ranked lyrics candidates.

Proposed Methods

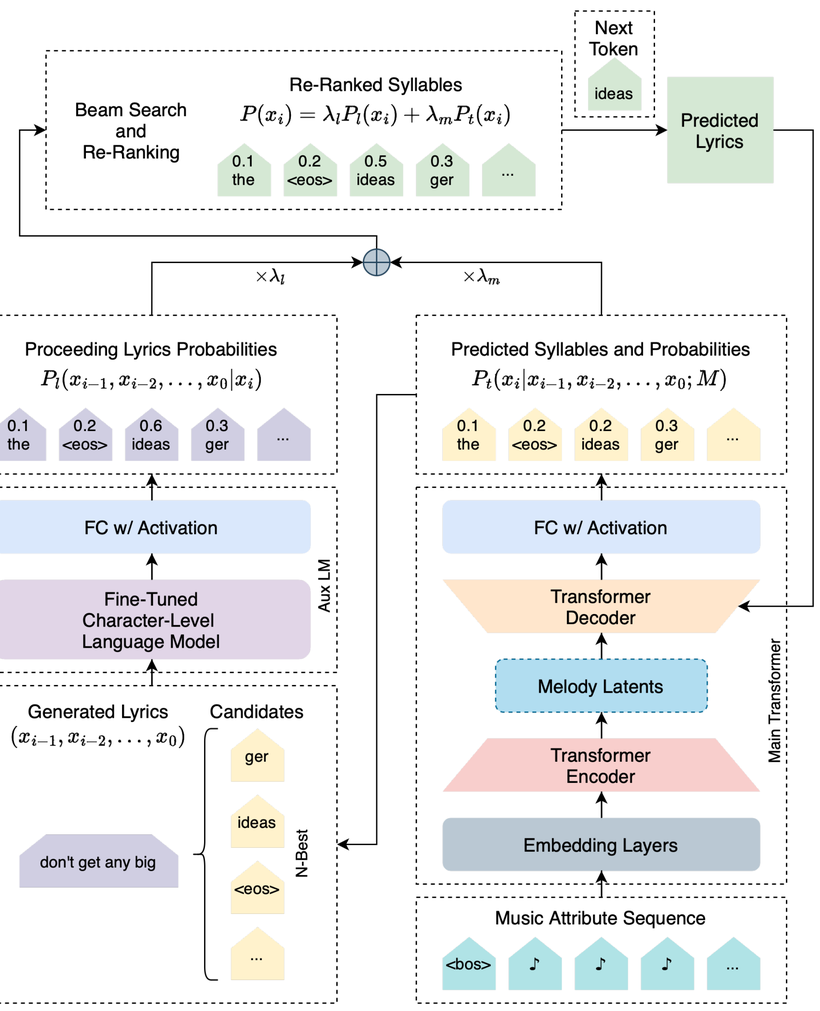


Figure 1: Transformer-based melody-encoder-syllable-decoder architecture exploiting character-level language model.

As shown in Figure 1, the Transformer on the right side generates the candidate syllable tokens based on the encoded melody latent representations M and previously generated lyrics. The fine-tuned language model on the left evaluates the probability of the candidates based on the given lyrics generated, which aims to improve the coherence and correctness of the generated lyrics. In such a way, our model refines semantic meanings within generated syllable-level lyrics by using a fine-tuned language model and re-ranking candidates by beam search.

Objective Evaluation

Table 1 shows the evaluation results of our model (Transformer + LM) and the baselines. We selected the recently published semantic dependency network (SDN) as a strong baseline, which already surpassed some methods like LSTM-GAN, SeqGAN, and RelGAN. We also implemented the original Transformer as another baseline. The BLEU and ROUGE metrics are slightly worse for the proposed model, however, the difference is insignificant enough to judge that our approach stays relatively close to ground truth in terms of the modeled syllable distribution.

Metric	SDN [Duan et al.]	Transformer	Transformer + LM
ROUGE F score (1,2,L)	0.1301, 0.0008, 0.0981	0.1476, 0.0354, 0.1248	0.1439, 0.0289, 0.1186
Sentence BLEU (2,3,4-gram)	0.0171, 0.0074, 0.0049,	0.0637, 0.0454, 0.0374	0.0576, 0.0386, 0.0308
BERT Scores (Precision, Recall, F1)	0.8771, 0.8870, 0.8819	0.967, 0.968, 0.967	0.967, 0.969, 0.968

Table 1: Objective metrics on the validation dataset.

ChatGPT Evaluation

Metrics	Ground-truth		Transformer		Trans.+LM	
	1st	2nd	1st	2nd	1st	2nd
Naturality	6	6	3	5	**4**	**7**
Correctness	7	7	4	6	**5**	**8**
Coherence	5	5	3	4	3	**6**
Originality	4	4	2	3	**3**	**5**
Poetic Value	4	5	2	4	2	**6**
Overall	5.2	5.4	2.8	4.4	**3.4**	**6.4**

Table 2: Results of the ChatGPT evaluation of generated lyrics on a scale from 1 to 10.

We evaluate the quality and correctness of generated lyrics via LLMs, since they are objective and have a vast linguistic knowledge. We asked the GPT-3 to evaluate our generated lyrics with the following prompts. In addition, we informed ChatGPT that the lyrics are syllable-split, lowercase, and without punctuation in the second round of evaluation. We show the results from both runs in Table 2. In both cases, the proposed method is able to outperform the baseline. This also verified the effectiveness of our proposed methods with language models.

I will send you three sets of generated candidate lyrics for 20-note melodies. I want you to evaluate them in terms of naturality, correctness, coherence (staying on topic), originality, and poetic value. Try to give numerical scores to all three candidate methods of lyric generation. I will send them in separate messages, please evaluate them after the third message. Is it clear?

Subjective evaluation

In addition to text-based evaluation, we performed a subjective evaluation by synthesizing audible samples with 11 participants. Evaluation results show that our proposed model achieves an improvement based on the Transformer baseline. Comparison between human evaluation and ChatGPT evaluation show that ChatGPT gives similar results to human evaluation. This potential consistency between human evaluation and ChatGPT evaluation makes it promising for future research on ChatGPT-based evaluation, improving evaluation efficiency and reduce human resource costs.

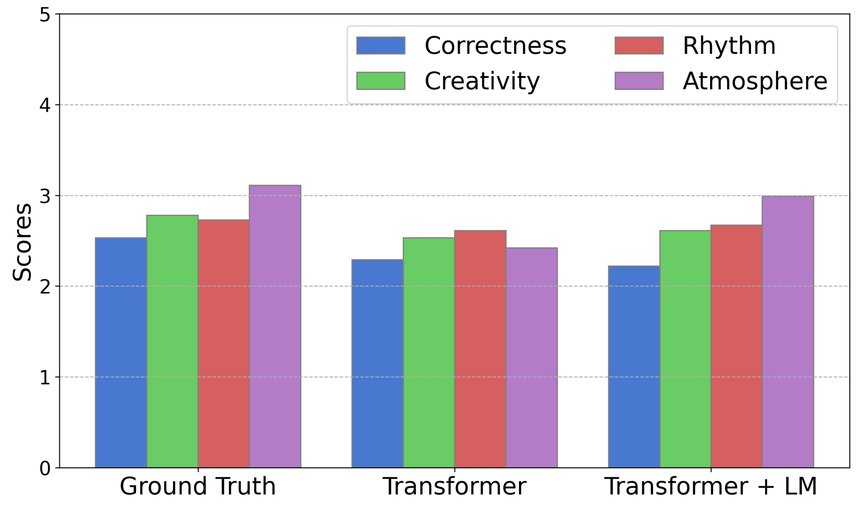
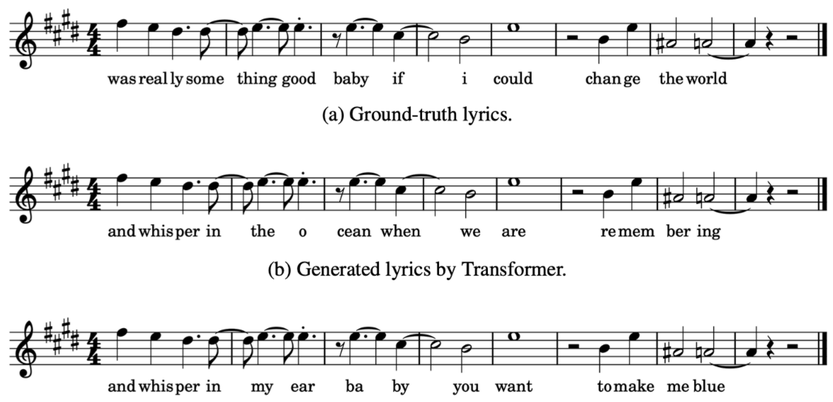


Figure 2: Results of subjective evaluation of lyrics generation from melody.

Sheet Music Examples

Examples of generated lyrics accompanied by the input melody are shown in Figure 3, which show that the lyrics generated by our model can better capture the characteristics of musical lyrics.



(a) Ground-truth lyrics.

(b) Generated lyrics by Transformer.

(c) Generated lyrics by Transformer + LM.

Figure 3: Sheet music with generated lyrics.

Conclusions

- **Innovative Methodology:** We introduced a unique method that enhances the prediction capabilities of a syllable-level, melody-conditioned lyrics generation Transformer. This method involves the creation of a specialized dataset for fine-tuning character-level language models, aimed at refining syllable-level semantic meanings.
- **Semantic Refinement and Re-ranking Algorithm:** Our approach includes an innovative algorithm for re-ranking candidate tokens during the beam search process, improving the naturality, correctness, and coherence of the generated lyrics in relation to the conditioning melodies.
- **Future Directions:** We outline plans for future research, including the development of a syllable-level language model pre-trained on extensive data corpora. This work will also explore the application of fine-tuned character-level language models for enhancing the generation of lyrics conditioned on melodies, promising further advancements in the field.