NUS National University of Singapore
School of Computing

ACM multimedia
Chengdu, China OCT 20-24 2021

AI-Lyricist:
Generating Music and Vocabulary Constrained Lyrics

Xichu Ma, Ye Wang, Min-Yen Kan, Wee Sun Lee

Jul 29, 2021

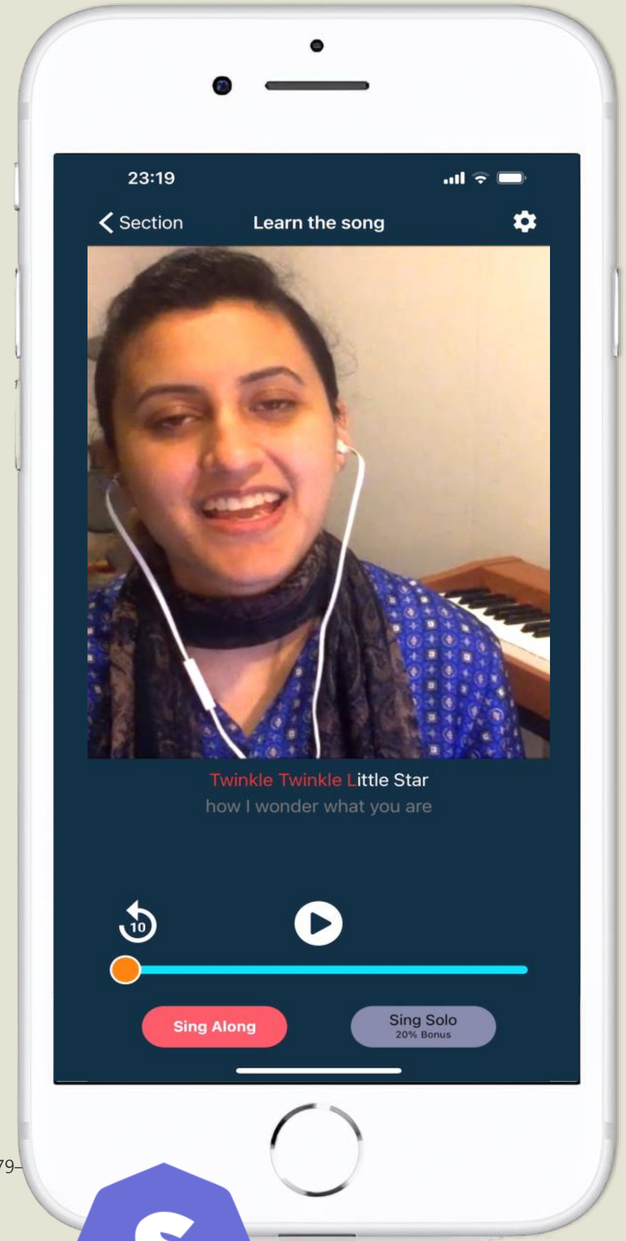
Language Learning via Singing

- Singing benefits learning foreign languages.

- **SLIONS**: transforming this idea into practice:

  Watching pre-recorded song videos;

  Singing songs following the teacher's examples;



[1] Dania Murad, Riwu Wang, Douglas Turnbull, and Ye Wang. 2018. SLIONS: A Karaoke Application to Enhance Foreign Language Learning. In 2018 ACM Multimedia Conference on Multimedia Conference. ACM, 1679–
[2] Dwayne Engh. 2013. Why Use Music in English Language Learning? A Survey of the Literature. English Language Teaching 6, 2 (2013), 113–127.
[3] Douglas Fisher. 2001. Early language learning with and without music. Reading Horizons 42, 1 (2001), 39.
[4] Arla J Good, Frank A Russo, and Jennifer Sullivan. 2015. The efficacy of singing in foreign-language learning. Psychology of Music 43, 5 (2015), 627–640.
[5] Suzanne L Medina. 1990. The Effects of Music upon Second Language Vocabulary Acquisition. (1990).
[6] Susan Bergman Miyake. 2004. Pronunciation and music. Sophia Junior College Faculty Bulletin 20, 3 (2004), 80.
[7] Andrés Roberto Rengifo. 2009. Improving pronunciation through the use of karaoke in an adult English class. Profile Issues in Teachers Professional Development 11 (2009), 91–106.
[8] Wanda T Wallace. 1994. Memory for music: Effect of melody on recall of text. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 6 (1994), 1471.
[9] Judith Weaver Failoni. 1993. Music as Means To Enhance Cultural Awareness and Literacy in the Foreign Language Classroom. Mid-Atlantic Journal of Foreign Language Pedagogy 1 (1993), 97–108.

SLIONS

Limitations

Songs suitable for language learning is limited, lacking diversity.

Songs with suitable lyrics may not match users' music preferences.

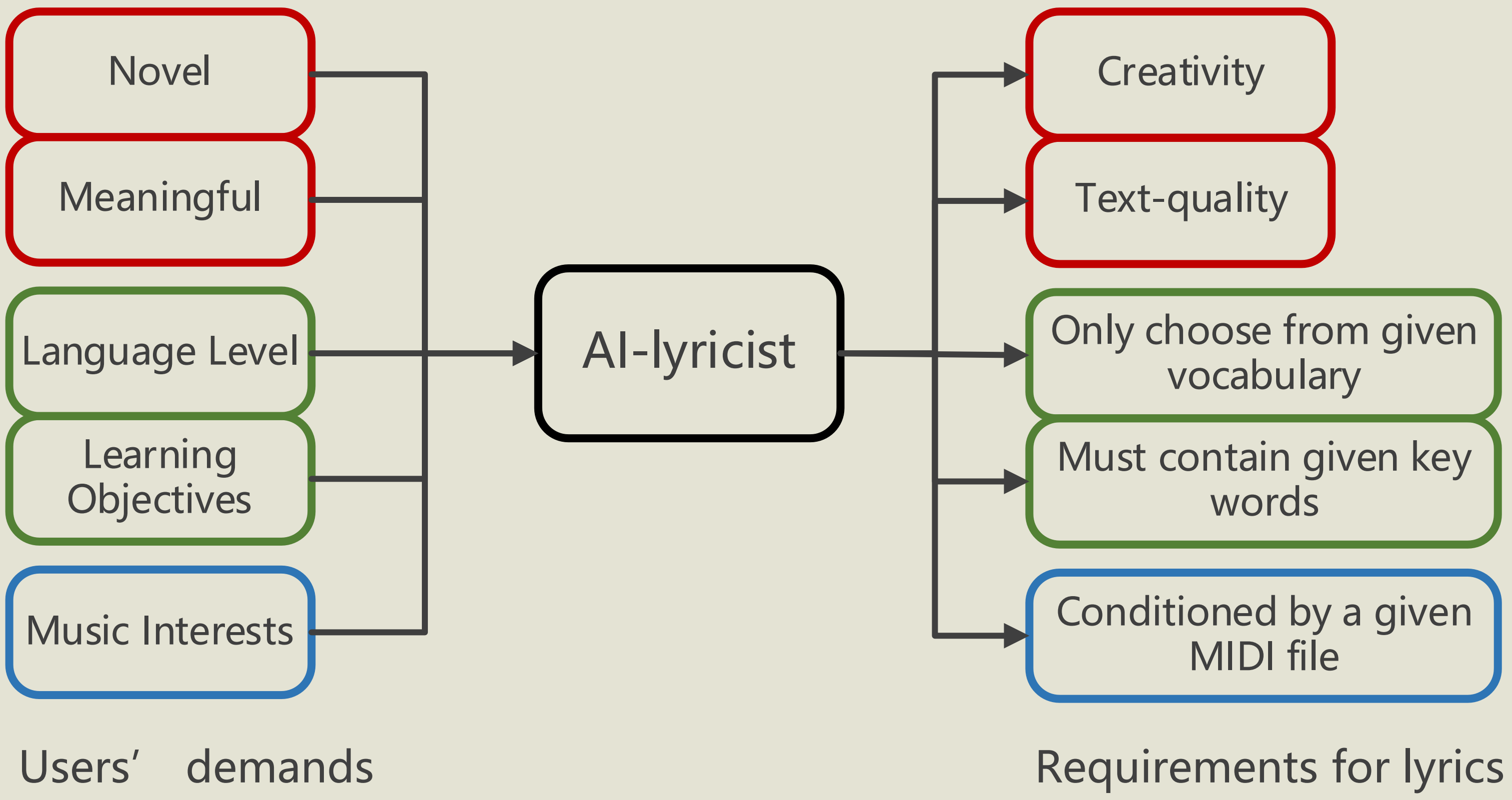
Lyrics do not match learners' linguistic ability and learning objectives.

Problem Statement

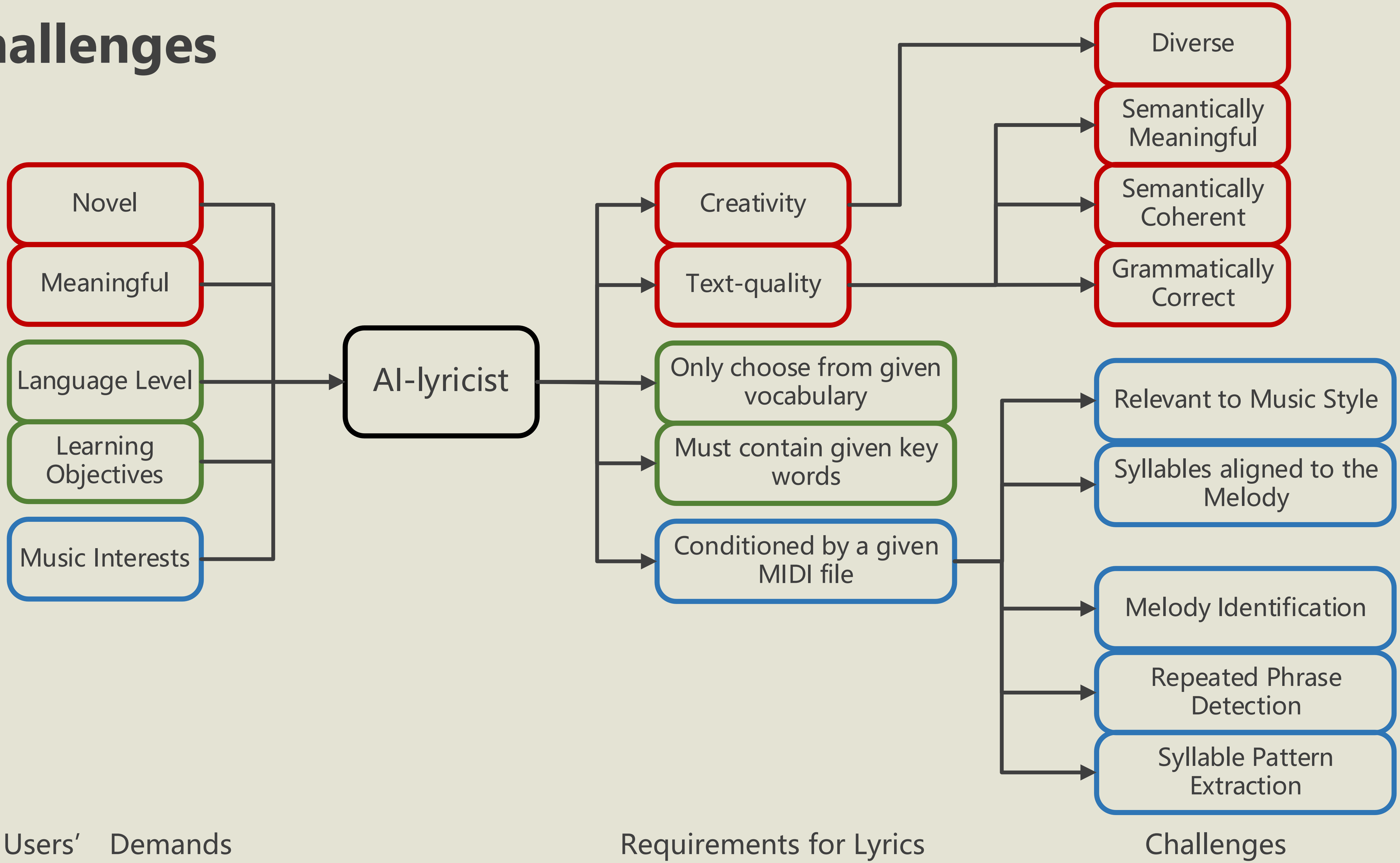
- Generating **novel** yet **meaningful** lyrics that match the users’ :

  - **language level** (mastered vocabulary) and **learning objectives** (new words);

  - **music interests** (reflected by a MIDI file they prefer).



Challenges

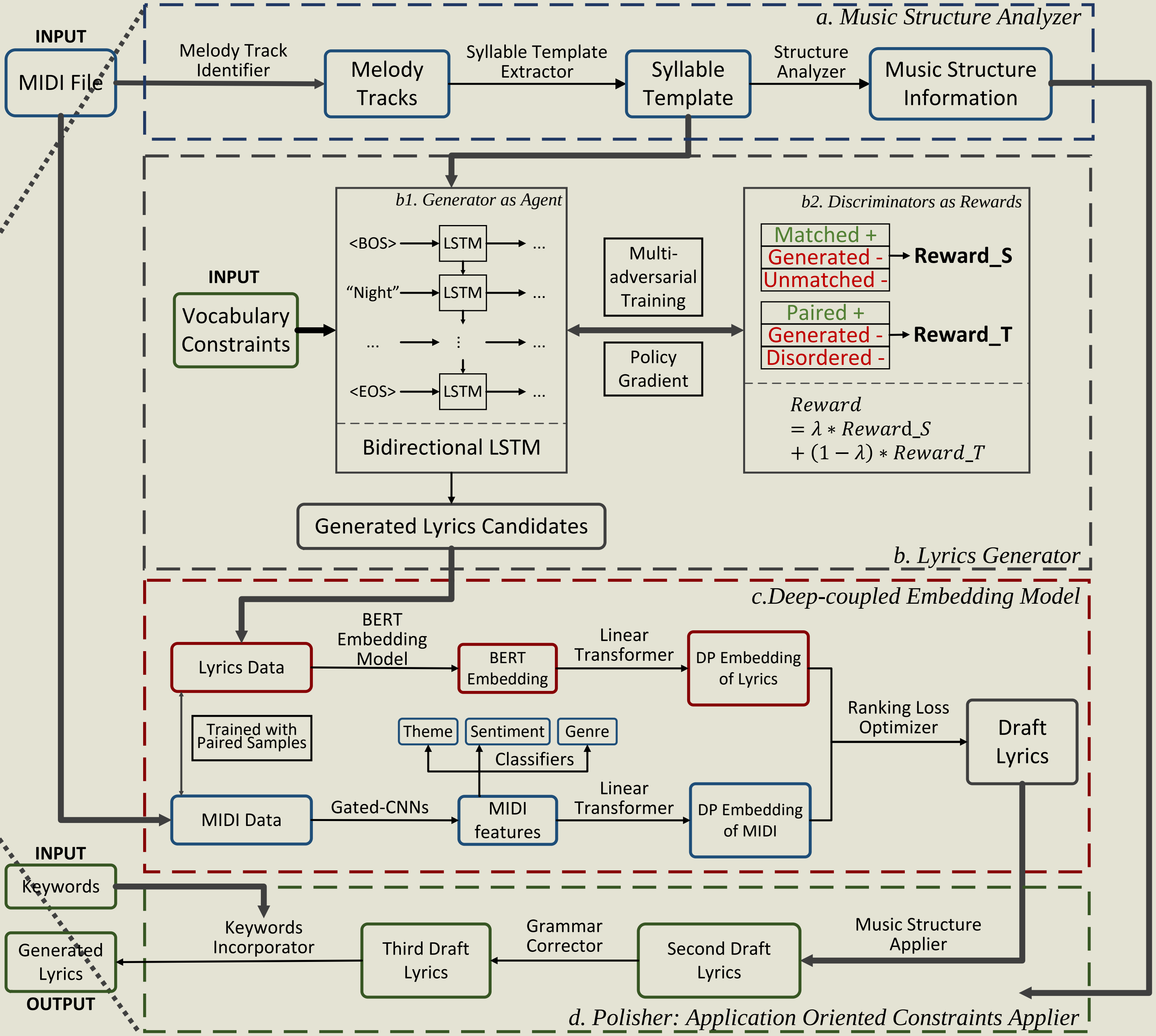

Novel
Meaningful
Language Level
Learning Objectives
Music Interests
AI-lyricist
Creativity
Text-quality
Only choose from given vocabulary
Must contain given key words
Conditioned by a given MIDI file
Diverse
Semantically Meaningful
Semantically Coherent
Grammatically Correct
Relevant to Music Style
Syllables aligned to the Melody
Melody Identification
Repeated Phrase Detection
Syllable Pattern Extraction
Users' Demands
Requirements for Lyrics
Challenges

System Architecture

- Creativity
- Semantically Meaningful
- Semantically Coherent
- Grammatically Correct
- Given vocabulary
- Given key words
- Relevant to Music Style
- Syllables aligned to the Melody
- Melody Identification
- Repeated Phrase Detection
- Syllable Pattern Extraction

AI-lyricist

- Music Structure Analyzer
- SeqGAN-based Lyrics Generator
- Deep Coupled Music–lyrics Embedding Model
- Polisher



Music Structure Analyzer

Creativity

Semantically Meaningful

Semantically Coherent

Grammatically Correct

Given vocabulary

Given key words

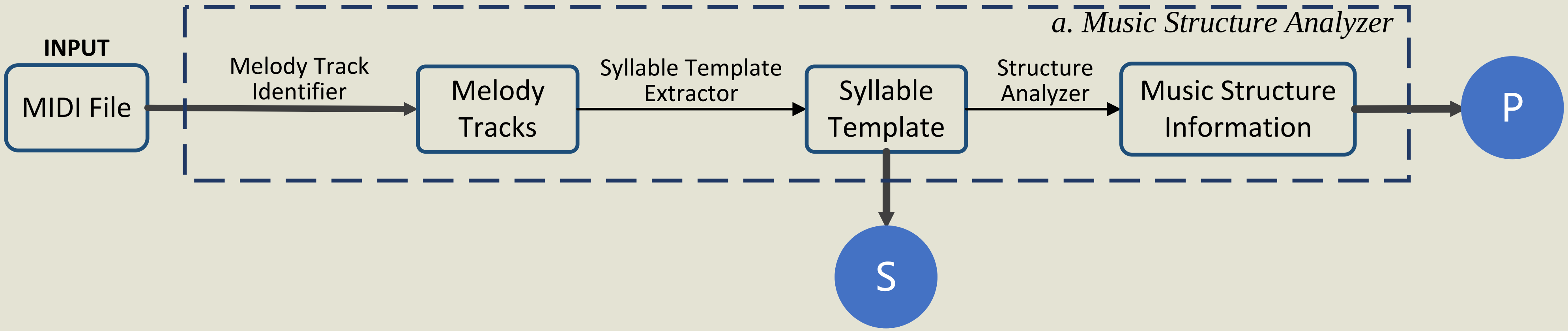
Relevant to Music Style

Syllables aligned to the Melody

Melody Identification

Repeated Phrase Detection

Syllable Pattern Extraction

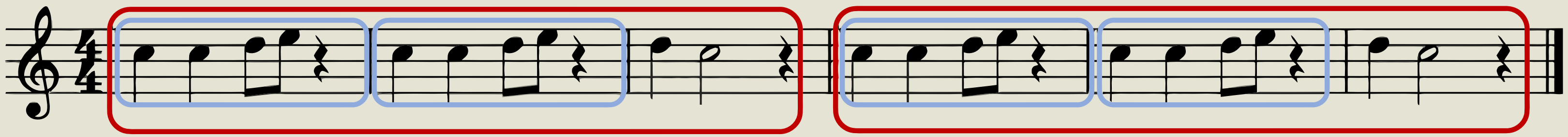


Melody Identification:

$$f_6 = Estimated\ Syllable\ Number \in \left[\frac{1.7 \times sd(s) \times tm(bpm)}{60}, \frac{2.3 \times sd(s) \times tm(bpm)}{60}\right]$$

$$Melody\ channel = \underset{i}{argmax} \sum_j w_j f_j(track_i)$$

Repeated Phrase Detection and Syllable Pattern Extraction:

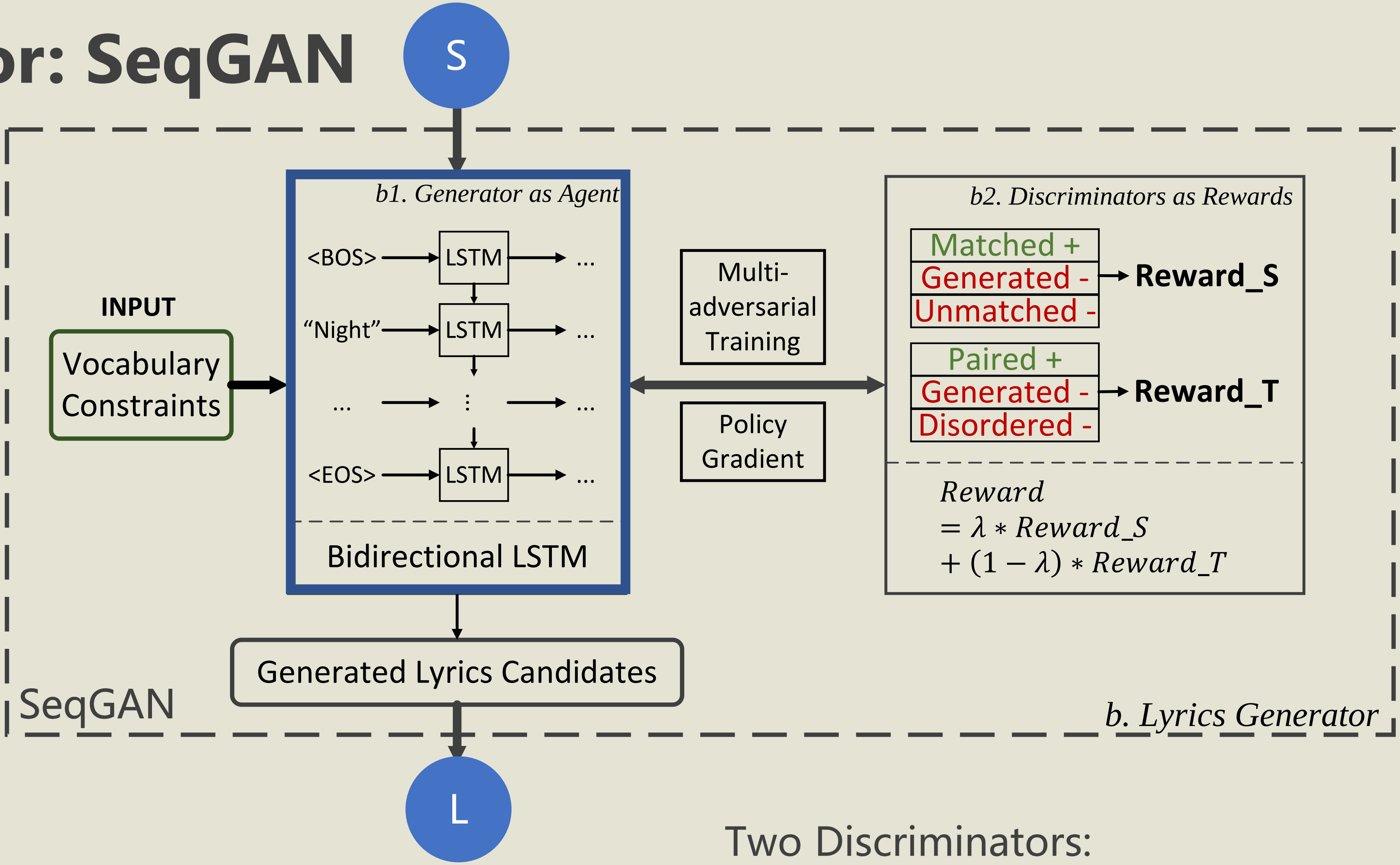


Granularity: 1 phrase   Syllable template: [**4, 4, 2, 4, 4, 2**]

Music Structure: [**[[0,3],3],** [[0,1],2], [[3,4],2]]

Lyrics Generator: SeqGAN

- Creativity
- Semantically Meaningful
- Semantically Coherent
- Grammatically Correct
- Given vocabulary
- Given key words
- Relevant to Music Style
- Syllables aligned to the Melody
- Melody Identification
- Repeated Phrase Detection
- Syllable Pattern Extraction



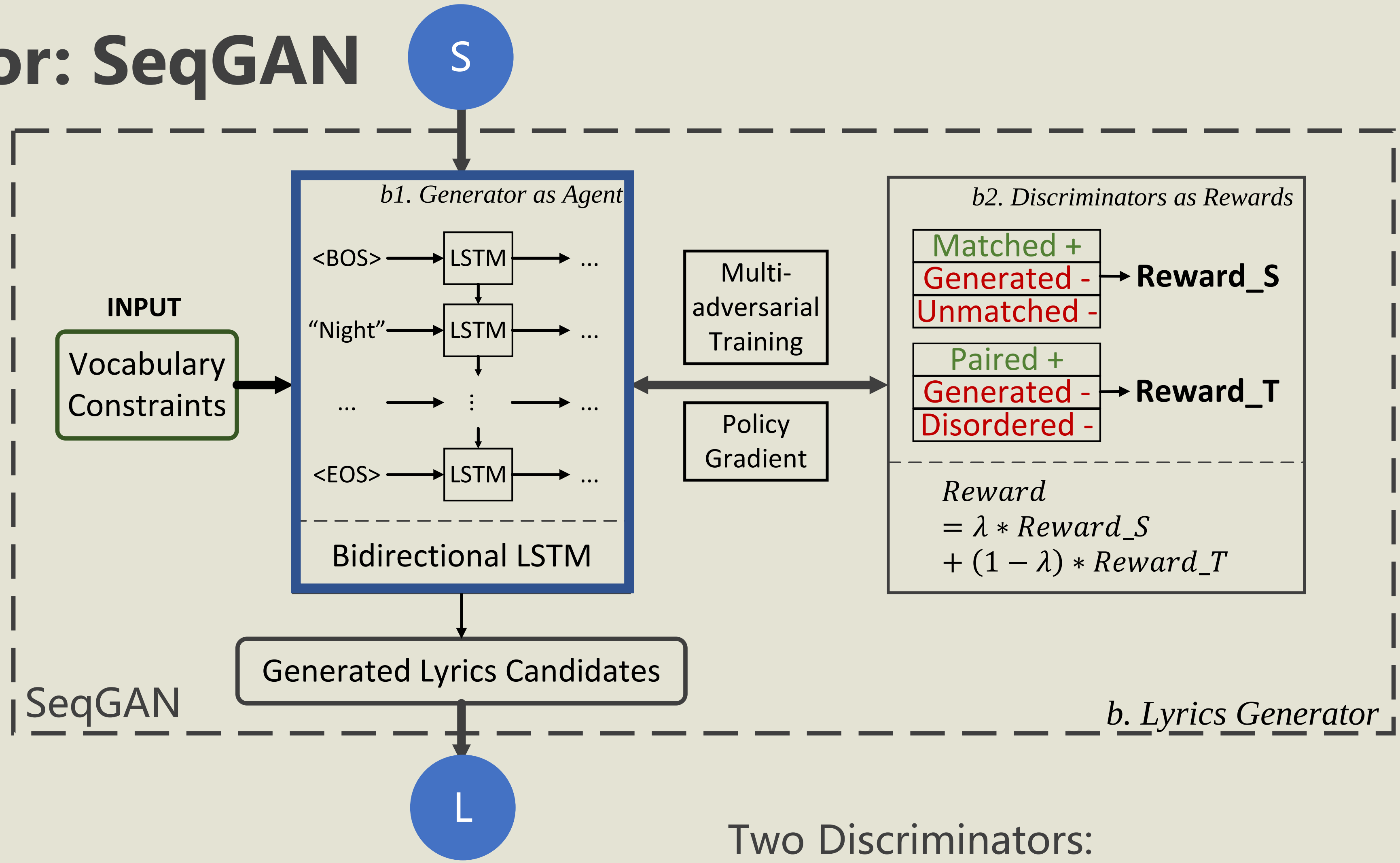
Two Discriminators:

- Syllable Discriminator
- Text-quality Discriminator

[1] Lantao Yu, Weinan Zhang, JunWang, and Yong Yu. 2017. SeqGAN: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

Lyrics Generator: SeqGAN

- Creativity
- Semantically Meaningful
- Semantically Coherent
- Grammatically Correct
- Given vocabulary
- Given key words
- Relevant to Music Style
- Syllables aligned to the Melody
- Melody Identification
- Repeated Phrase Detection
- Syllable Pattern Extraction



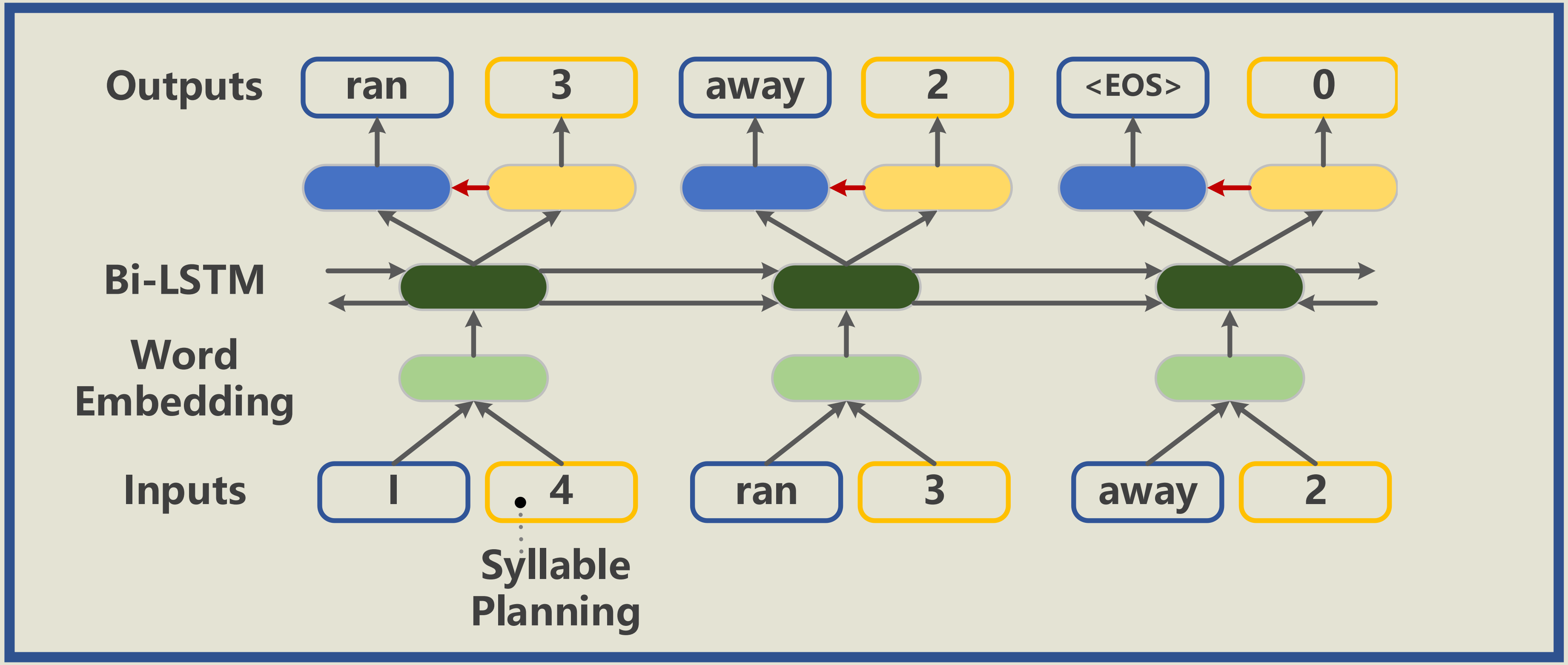
Two Discriminators:

- Syllable Discriminator
- Text-quality Discriminator

[1] Lantao Yu, Weinan Zhang, JunWang, and Yong Yu. 2017. SeqGAN: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

Lyrics Generator: Syllable Awareness

- Creativity
- Semantically Meaningful
- Semantically Coherent
- Grammatically Correct
- Given vocabulary
- Given key words
- Relevant to Music Style
- Syllables aligned to the Melody
- Melody Identification
- Repeated Phrase Detection
- Syllable Pattern Extraction



Syllable Planning:

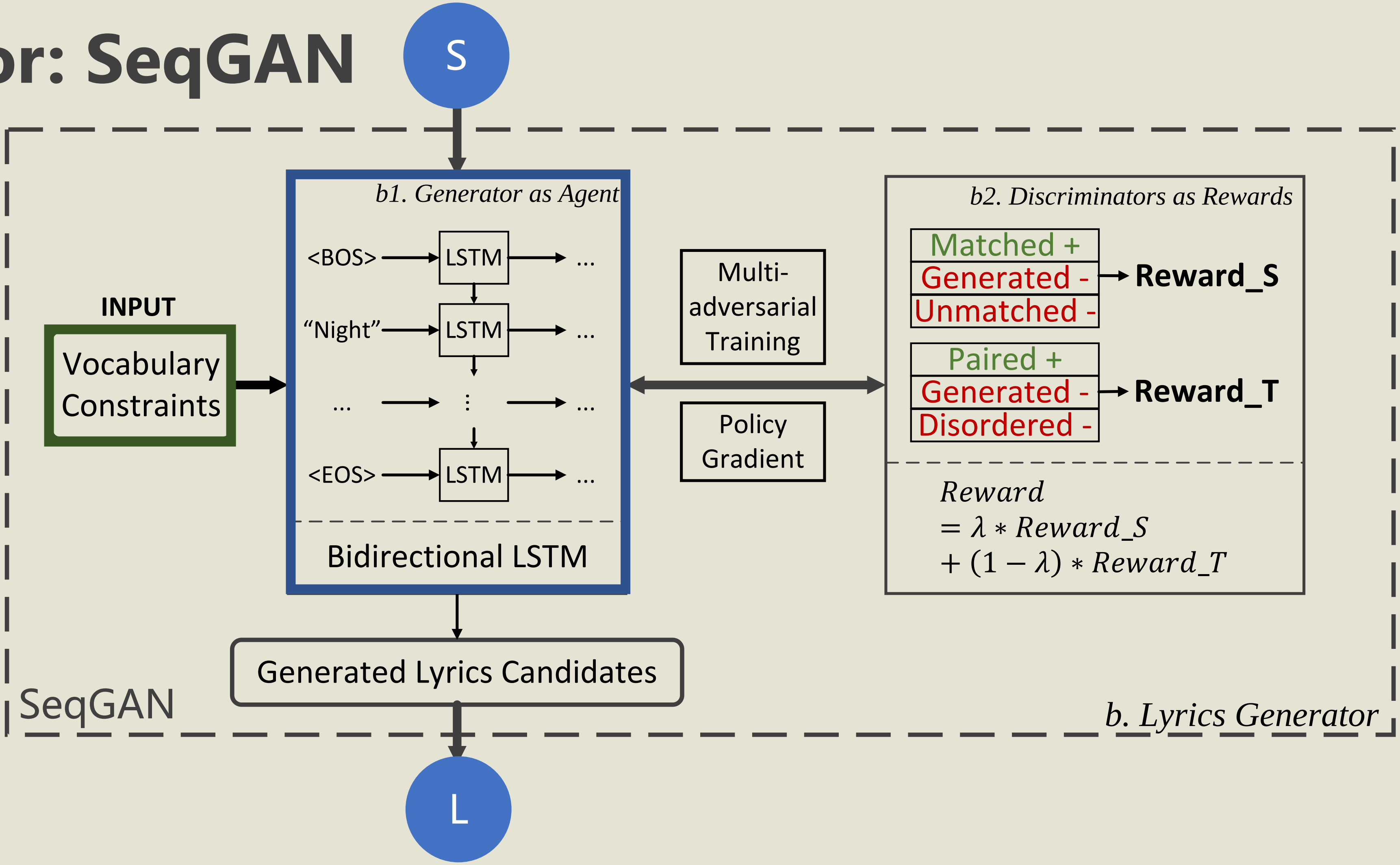
Input: How many syllables the remaining sentence should have?

Output: How many syllables the remaining sentence should have after picking this word?

[1] Lantao Yu, Weinan Zhang, JunWang, and Yong Yu. 2017. SeqGAN: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

Lyrics Generator: SeqGAN

- Creativity
- Semantically Meaningful
- Semantically Coherent
- Grammatically Correct
- Given vocabulary
- Given key words
- Relevant to Music Style
- Syllables aligned to the Melody
- Melody Identification
- Repeated Phrase Detection
- Syllable Pattern Extraction

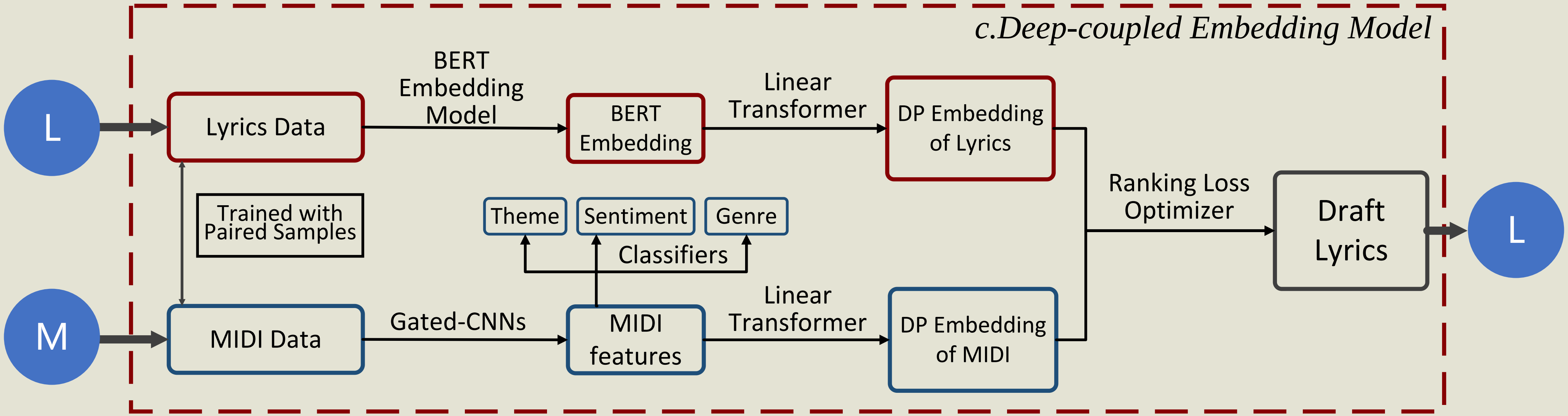


Given Vocabulary Set as a mask

[1] Lantao Yu, Weinan Zhang, JunWang, and Yong Yu. 2017. SeqGAN: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

Deep-coupled Music-Lyrics Embedding

- Creativity
- Semantically Meaningful
- Semantically Coherent
- Grammatically Correct
- Given vocabulary
- Given key words
- Relevant to Music Style
- Syllables aligned to the Melody
- Melody Identification
- Repeated Phrase Detection
- Syllable Pattern Extraction



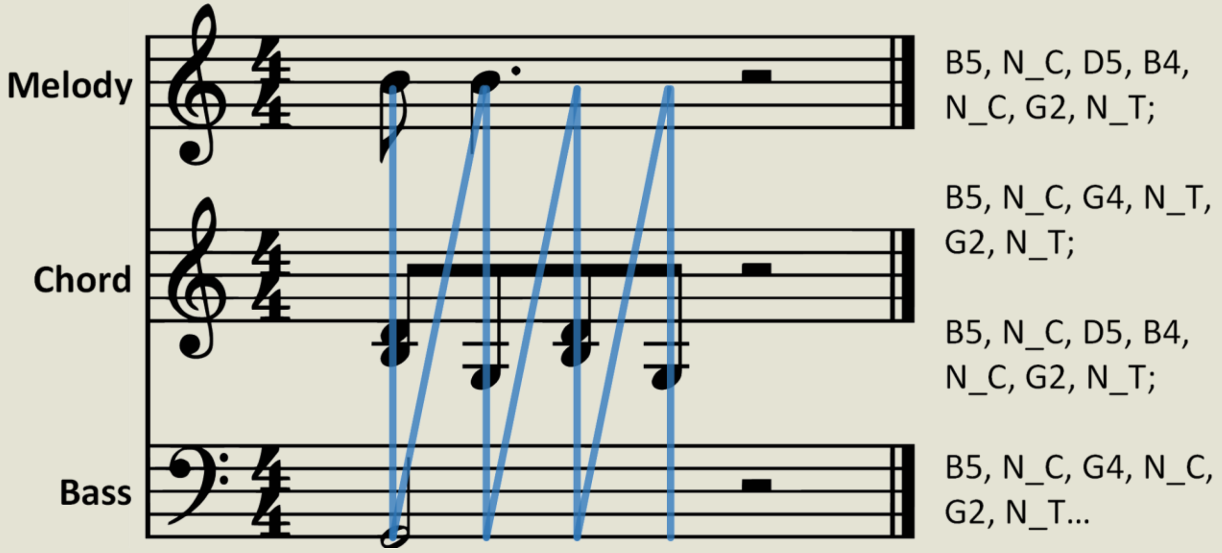
MIDI File Encoding:

Scanning Line:

timestep by timestep

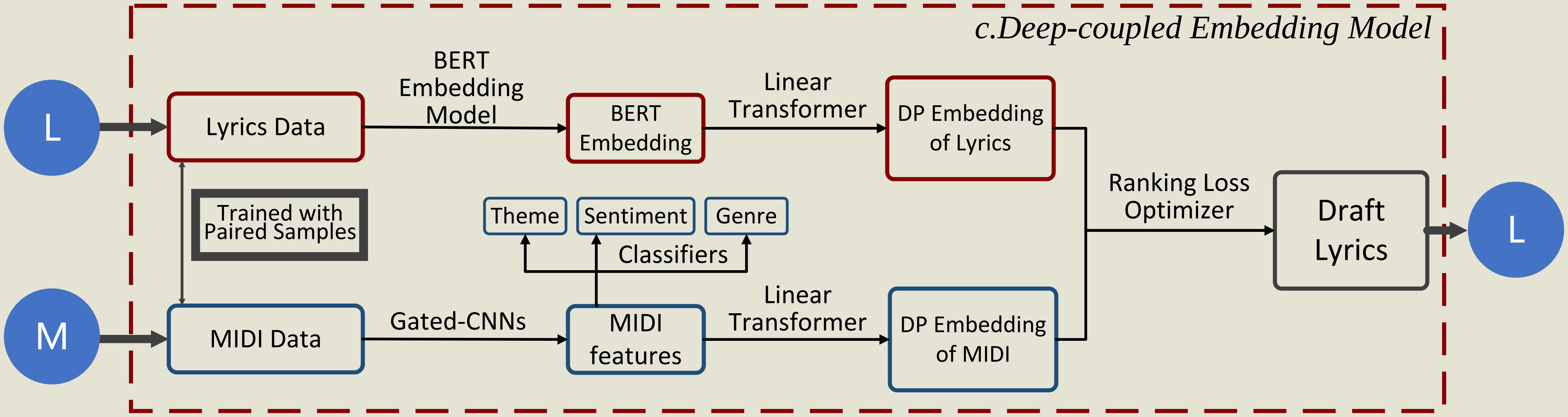
channel by channel

from high to low pitch



Deep-coupled Music-Lyrics Embedding

- Creativity
- Semantically Meaningful
- Semantically Coherent
- Grammatically Correct
- Given vocabulary
- Given key words
- Relevant to Music Style
- Syllables aligned to the Melody
- Melody Identification
- Repeated Phrase Detection
- Syllable Pattern Extraction



Projecting lyrics features and music features into the same space, so that:
Lyrics and music matched in style have similar embeddings.

Training:

Lyrics features: BERT Embedding x Linear Transformation → μ

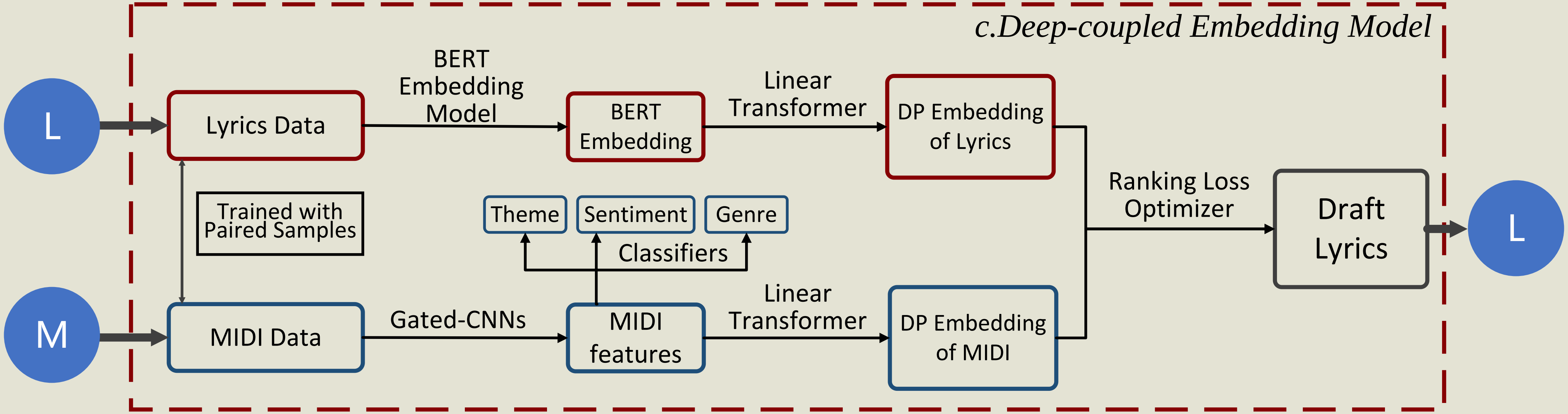
Music features: Concat (Theme, Sentiment, Genre features) x Linear Transformation → ν

$$Loss = \sum_{\mu}\sum_{n} max(0, m - \mu \cdot \nu + \mu \cdot \nu_n) + \sum_{\nu}\sum_{n} max(0, m - \nu \cdot \mu + \nu \cdot \mu_n)$$

[1] Bei Liu, Jianlong Fu, Makoto P Kato, and Masatoshi Yoshikawa. 2018. Beyond narrative description: Generating poetry from images by multi-adversarial training. In 2018 ACM Multimedia Conference on Multimedia Conference. 783-791..

Deep-coupled Music-Lyrics Embedding

- Creativity
- Semantically Meaningful
- Semantically Coherent
- Grammatically Correct
- Given vocabulary
- Given key words
- Relevant to Music Style
- Syllables aligned to the Melody
- Melody Identification
- Repeated Phrase Detection
- Syllable Pattern Extraction



Projecting lyrics features and music features into the same space, so that: Lyrics and music matched in style have similar embeddings.

Generating:

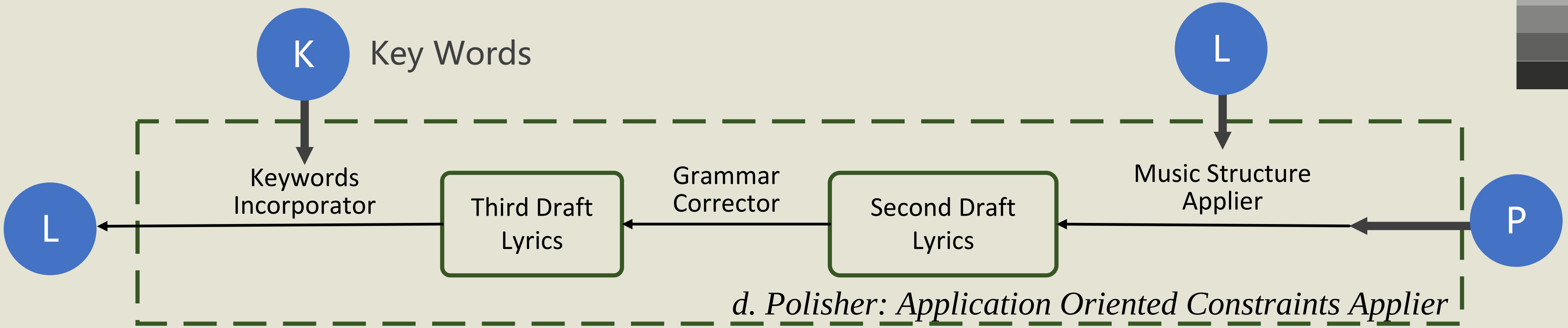
Input: MIDI file, {a group of candidate generated lyrics}

Output: The candidate having the most similar embedding with the input MIDI file

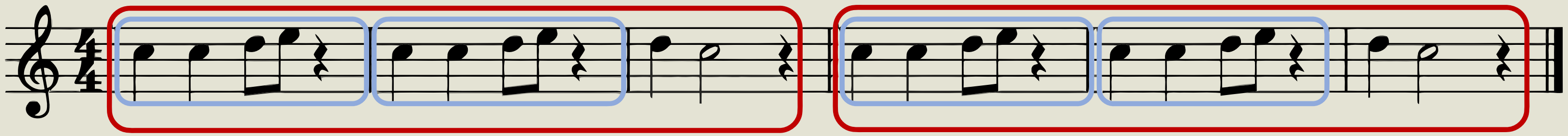
[1] Bei Liu, Jianlong Fu, Makoto P Kato, and Masatoshi Yoshikawa. 2018. Beyond narrative description: Generating poetry from images by multi-adversarial training. In 2018 ACM Multimedia Conference on Multimedia Conference. 783-791..

Polisher

Creativity
Semantically Meaningful
Semantically Coherent
Grammatically Correct
Given vocabulary
Given key words
Relevant to Music Style
Syllables aligned to the Melody
Melody Identification
Repeated Phrase Detection
Syllable Pattern Extraction



Music Structure Applier: Copy the generated lyrics to every repeated music sections.



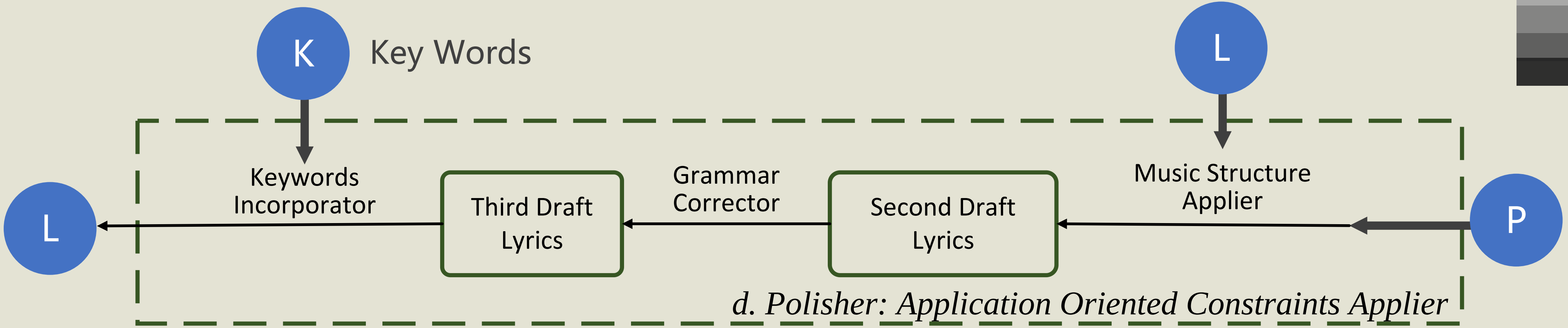
Granularity: 1 phrase   Syllable template: [**4, 4, 2, 4, 4, 2**]

Music Structure: [**[[0,3],3],** [[0,1],2], [[3,4],2]]

E.g. "I ran away. I ran away. And you? I ran away. I ran away. And you?"

Polisher

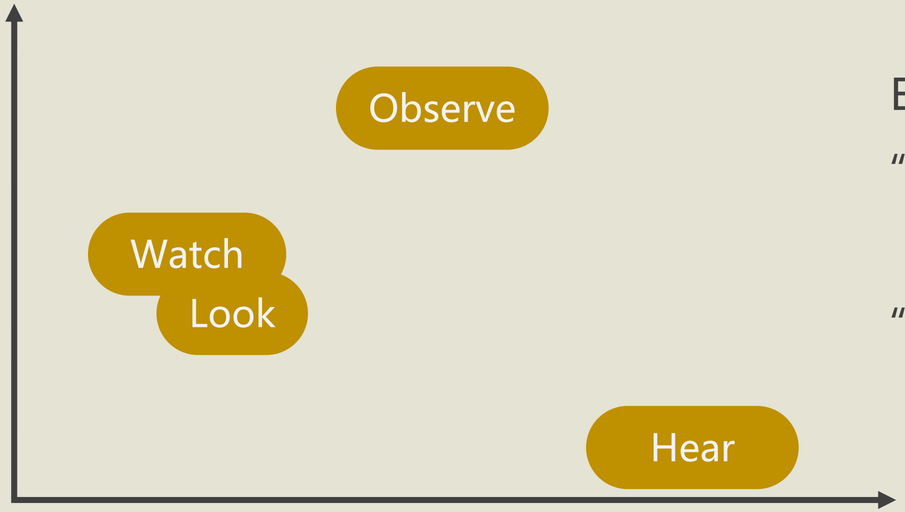
Creativity
Semantically Meaningful
Semantically Coherent
Grammatically Correct
Given vocabulary
Given key words
Relevant to Music Style
Syllables aligned to the Melody
Melody Identification
Repeated Phrase Detection
Syllable Pattern Extraction



Grammar Corrector: Grammarly (C) APIs

Keywords Incorporator:

Substitutes the given keywords to the words having the most similar embeddings regarding both **semantic meaning** and **syllable number**



E.g. Keyword is "look"

"**Watch**! the distant hills and listen to the music."

↓

"**Look**! the distant hills and listen to the music."

Grammar Corrector: https://github.com/grammarly/gector

Experiment: Objective Evaluation

$Novelty - n = m_{n\_IF}/m_n$ How often the generator uses infrequent words/phrases?

$Dist - n = m_{n\_U}/m_n$ Info Density: How many unique words/phrases are used?

$Relevance$ The music-lyrics relevance calculated by the deep-coupled embedding model.

$BLEU$ Text quality.

$Application - oriented\ Satisfaction$ Whether the generated lyrics is singable?
Containing keywords and matching vocabulary constraints?

Experiment：Subjective Evaluation

Participants: 60 native English speakers

Platform: Amazon MTurk

Procedures:

(1) A participant **reads the lyrics** generated from the same music by the 5 compared models and **rates the lyrics** on a 0-5 scale regarding their fluency, coherence, meaningfulness and poetic aesthetics respectively.

(2) A participant **listens to a singing sample** synthesized with the music and lyrics generated by the 5 models, and **rates the lyrics** on a 0-5 scale regarding the syllable alignment to the melody and their relevance respectively.

Experiment: Results

a. Objective Evaluation	Novelty-2	Novelty-3	Dist-1	Dist-2	BLEU-1	BLEU-2	BLEU-3	Relevance	Overall	App Satisfaction
MT	0.390	0.190	9.16e-2	0.254	0.22	1.09e-2	3.71e-7	66.52	0.538	N
ED	6.02e-3	1.07e-3	4.81e-2	0.980	0.03	2.55e-5	1.53e-6	60.67	0.187	N
SeqGAN	5.21e-2	8.5e-2	0.132	0.417	0.17	9.61e-3	5.25e-4	68.22	0.554	N
PRE-M2L-B	0.410	0.213	6.82e-3	0.135	0.26	1.40e-2	3.30e-4	66.13	0.590	Y
POST-M2L	1.97e-4	0.0	0.164	0.363	1.92e-3	3.53e-6	~0.0	59.03	0.159	N
POST-M2L-S	8.09e-2	0.330	0.115	0.251	3.91e-2	3.47e-3	4.31e-4	68.84	0.482	Y
POST-M2L-T	0.224	0.144	0.150	0.473	3.67e-2	2.28e-3	6.61e-5	74.82	0.464	Y
POST-M2L-B	0.169	0.158	0.108	0.406	0.21	1.08e-2	3.36e-4	70.59	$\boxed{0.601}$	$\boxed{Y}$

b. Subjective Evaluation	Fluency	Coherence	Meaningfulness	Aesthetics	Syllable Alignment	Relevance	Overall
MT	2.43	2.40	2.60	2.47	2.87	2.87	2.61
ED	2.87	2.83	2.6	2.87	3.43	2.77	2.90
SeqGAN	2.83	2.80	2.63	2.67	3.23	3.00	2.86
PRE-M2L-B	3.67	3.38	3.46	3.40	4.00	3.67	3.59
POST-M2L-B	4.37	4.47	4.37	4.17	4.53	4.50	$\boxed{4.40}$

Demo: Auto Mode & Interactive Mode

Input MIDI: Imagine.mid **Keywords**: [“know” , “see”] **Interactive Mode**

Automatic mode: Generating the whole piece of lyrics

Interactive mode: Generating sentence by sentence. The user chooses one sentence from top 5 candidates.

Summary

First attempt to generate lyrics that match both the style and syllable pattern of multi-channel music.

Utilize a SeqGAN based generator to generate meaningful and coherent lyrics approaching human songwriting.

Propose Polisher module to constrain lyrics generation with mandatory keywords and a vocabulary set.

Thank you!

AI-Lyricist: Generating Music and Vocabulary Constrained Lyrics

Xichu Ma, Ye Wang, Min-Yen Kan, Wee Sun Lee

Q & A

Sound & Music Computing Lab, NUS
https://smcnus.comp.nus.edu.sg/