Atypical Lyrics Completion Considering Musical Audio Signals

Kento Watanabe and Masataka Goto, National Institute of Advanced Industrial Science and Technology (AIST)

◆ Lyrics Completion Task

We propose a lyrics completion task to recommend candidate words for a blank in a given sentence.

INPUT
The [*blank*] in my heart.

Lyrics Completion System

OUTPUT
The [*pain*] in my heart.
The [*place*] in my heart.
The [*deep*] in my heart.
The [*fire*] in my heart.
…

What Words Should be Suggested?

◆ Lyrics Completion Task

Suggesting words that are **(1) atypical** but **(2) suitable for the musical audio signals**.

1. **Atypicality** is important in writing lyrics in a creative way; to make lyrics attractive, lyric writers usually consider both typical and atypical phrases.

   However, previous research on automatic lyrics generation has used language models (LSTM-LM, GPT2, GPT3, Transformer…) that **predict highly frequent (i.e., typical) words.**

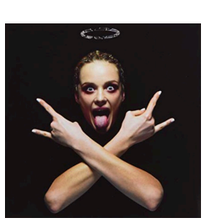
   ☹ The *pain* in my heart.
   Many songs use this phrase…

   ☺ The agony in my heart
   This is unique!

2. Lyrics depend on the moods of music.
   (e.g., “*love*” is often used in ballad songs, “*kill*” is often used in metal songs).

➔ … *love you* …

Ballad song



➔ … *kill you* …

Metal song

Previous Studies and Our Study

◆ Previous study of general word completion task

- Generating highly frequent (i.e., typical) words.

INPUT
The [*blank*] in my heart.
Draft

Language model

OUTPUT
The [*pain*] in my heart.
The [*place*] in my heart.
The [*deep*] in my heart.
…
Typical words

◆ This study

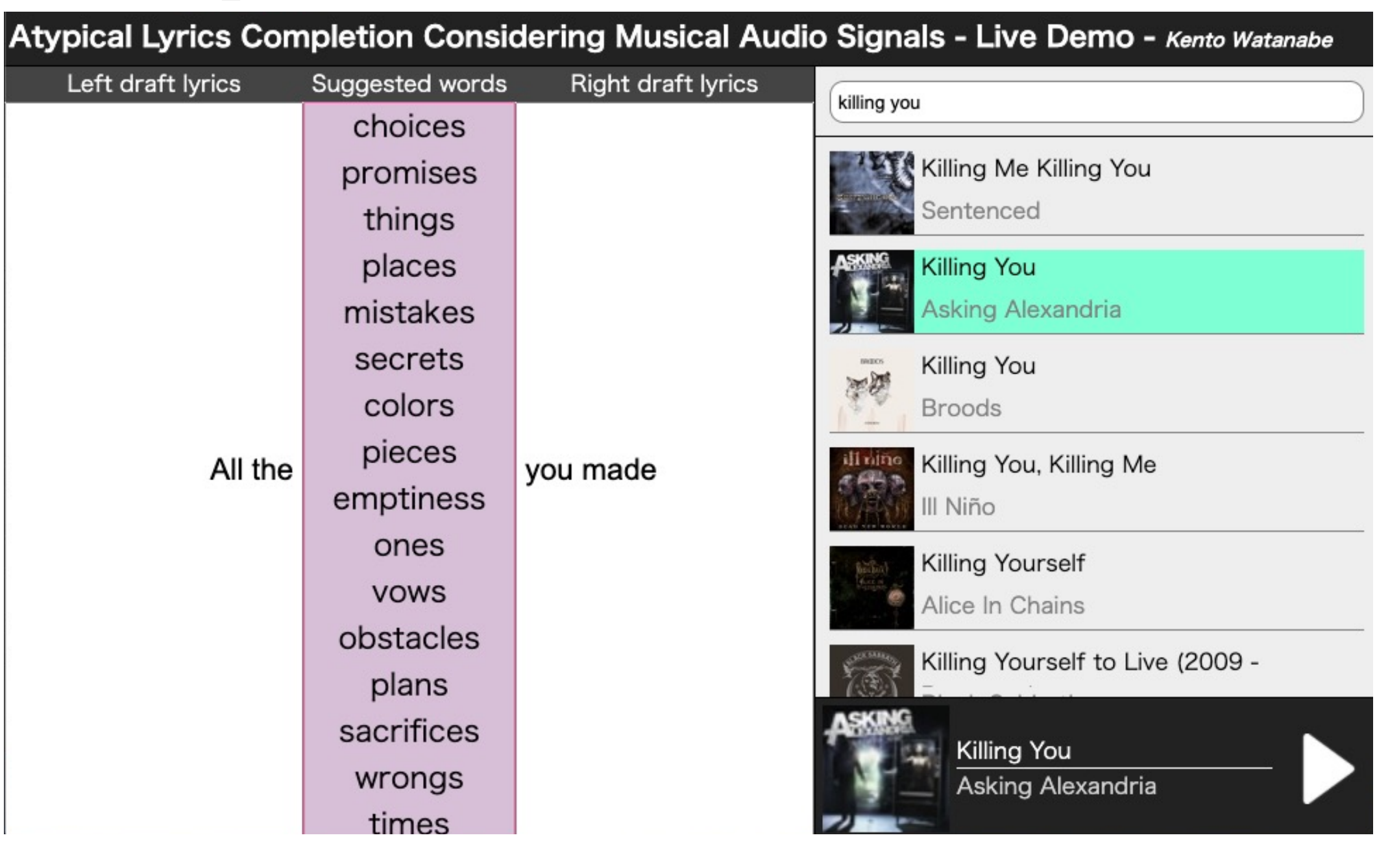
- Suggesting words that are (1) atypical but (2) suitable for musical audio signals.

INPUT
The [*blank*] in my heart
and
Draft and **audio**

Lyrics Completion System

OUTPUT
The [*agony*] in my heart.
The [*dagger*] in my heart.
The [*plague*] in my heart.
…
Atypical words

Demo System



Atypical Lyrics Completion Considering Musical Audio Signals - Live Demo - *Kento Watanabe*

All the [choices / promises / things / places / mistakes / secrets / colors / pieces / emptiness / ones / vows / obstacles / plans / sacrifices / wrongs / times] you made

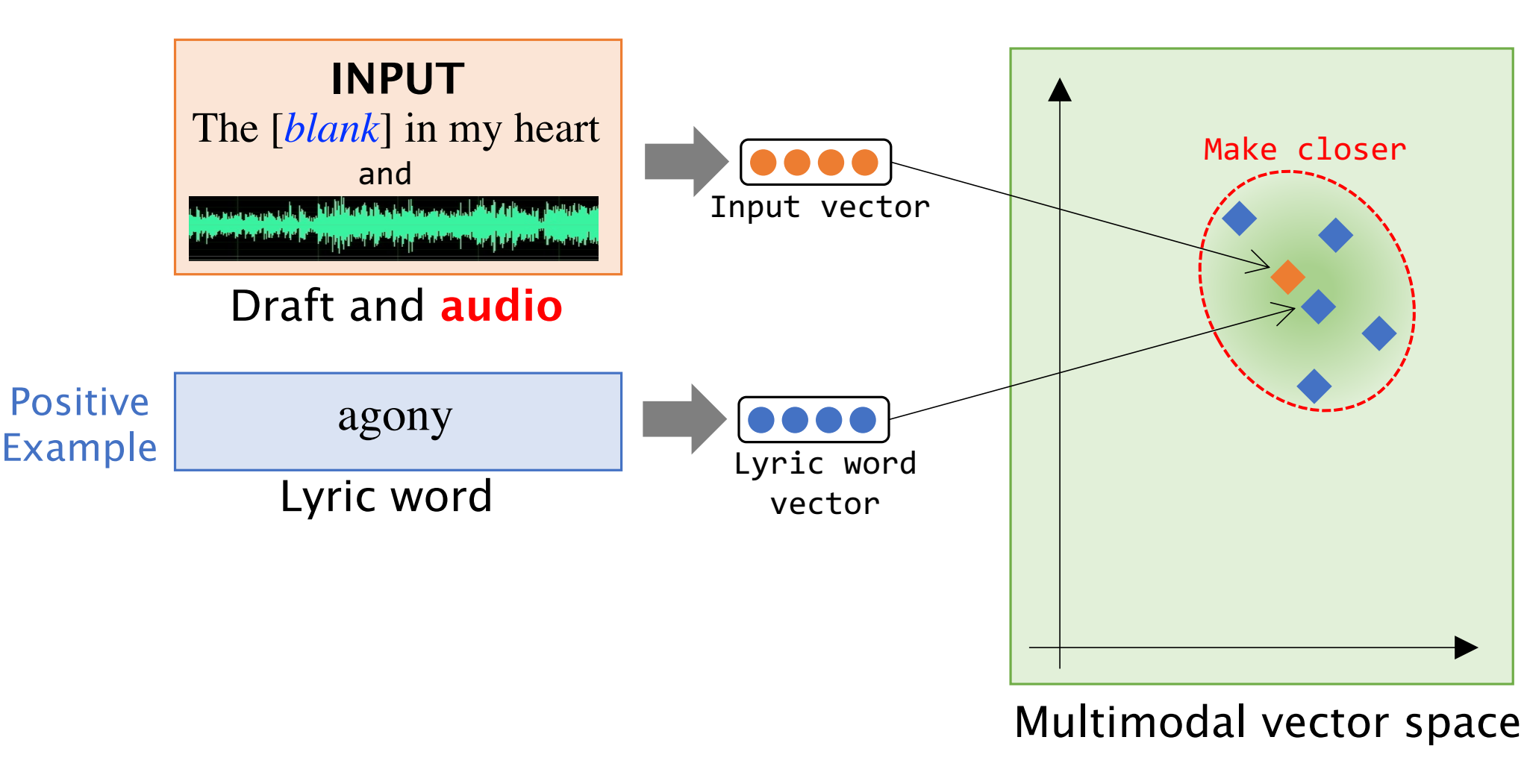
killing you

- Killing Me Killing You — Sentenced
- Killing You — Asking Alexandria
- Killing You — Broods
- Killing You, Killing Me — Ill Niño
- Killing Yourself — Alice In Chains
- Killing Yourself to Live (2009 -

Killing You
Asking Alexandria

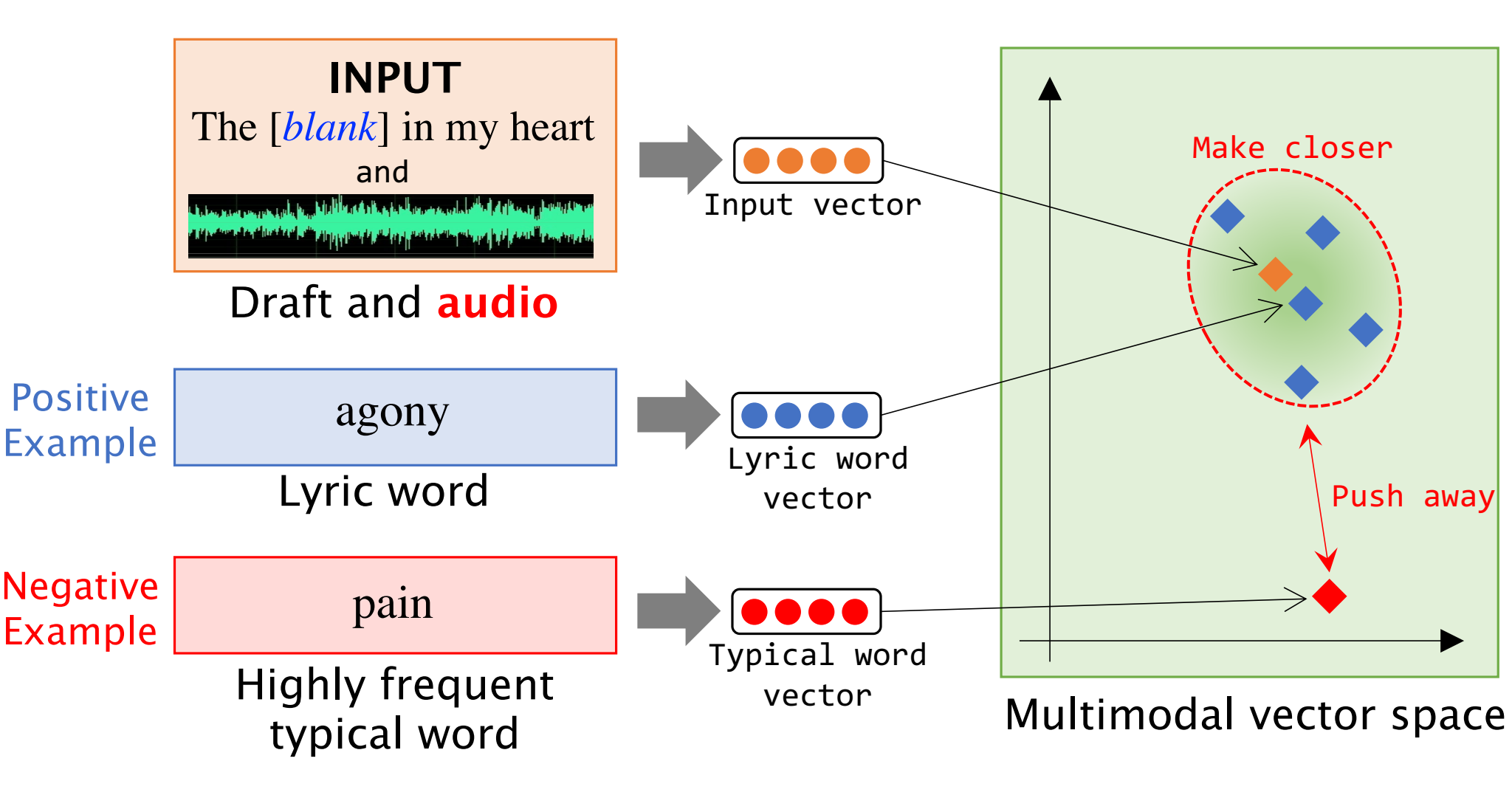
Key Idea

- We propose a **multimodal vector space model: Lyrics-context2vec**.
  1. We encode both of the input draft sentence with a blank and audio signals of the song into an input vector, and also encode a lyric word appropriate for the blank into a lyric word vector.
  2. We train the model to make those vectors located near each other.



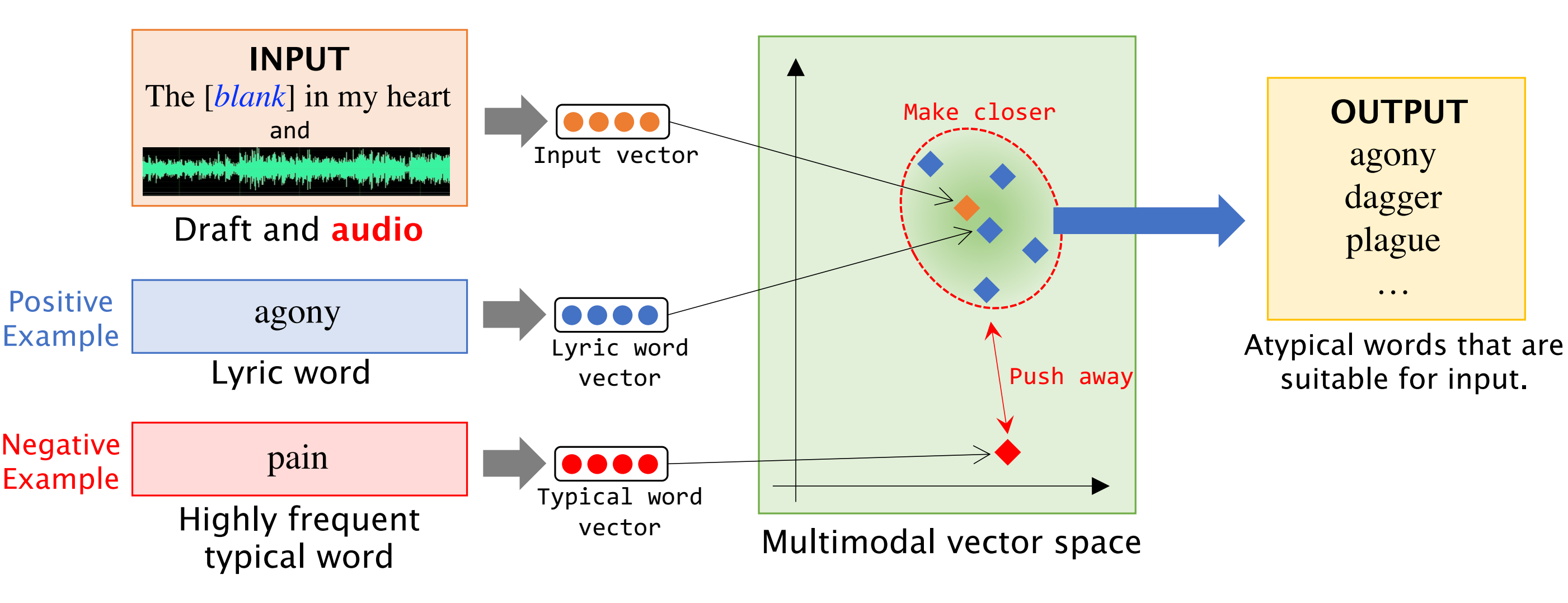
Key Idea

- We propose a **multimodal vector space model: Lyrics-context2vec**.
  3. A highly frequent word is encoded into a typical word vector.
  4. We use it as a negative example to make it located far away from the input vector.



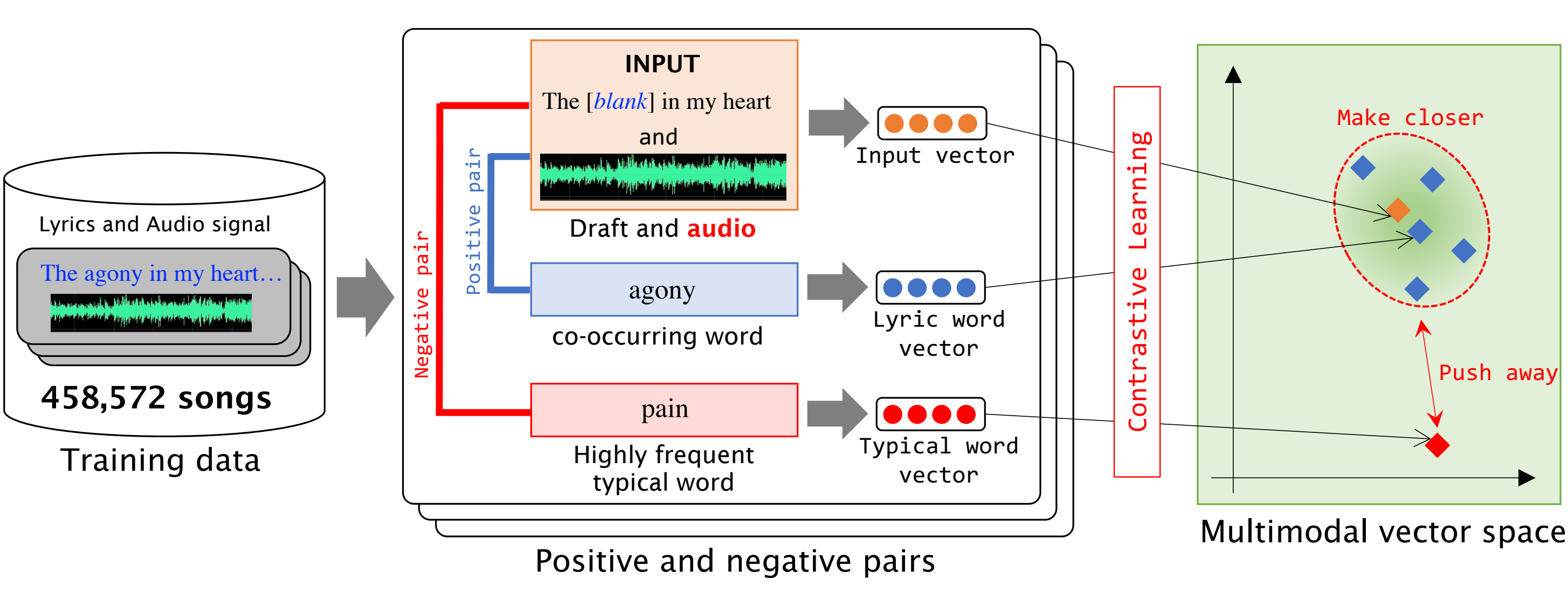
Key Idea

- We propose a **multimodal vector space model: Lyrics-context2vec**.
  5. Only atypical words are suggested since they are close to the input vector of a draft sentence and audio signals.



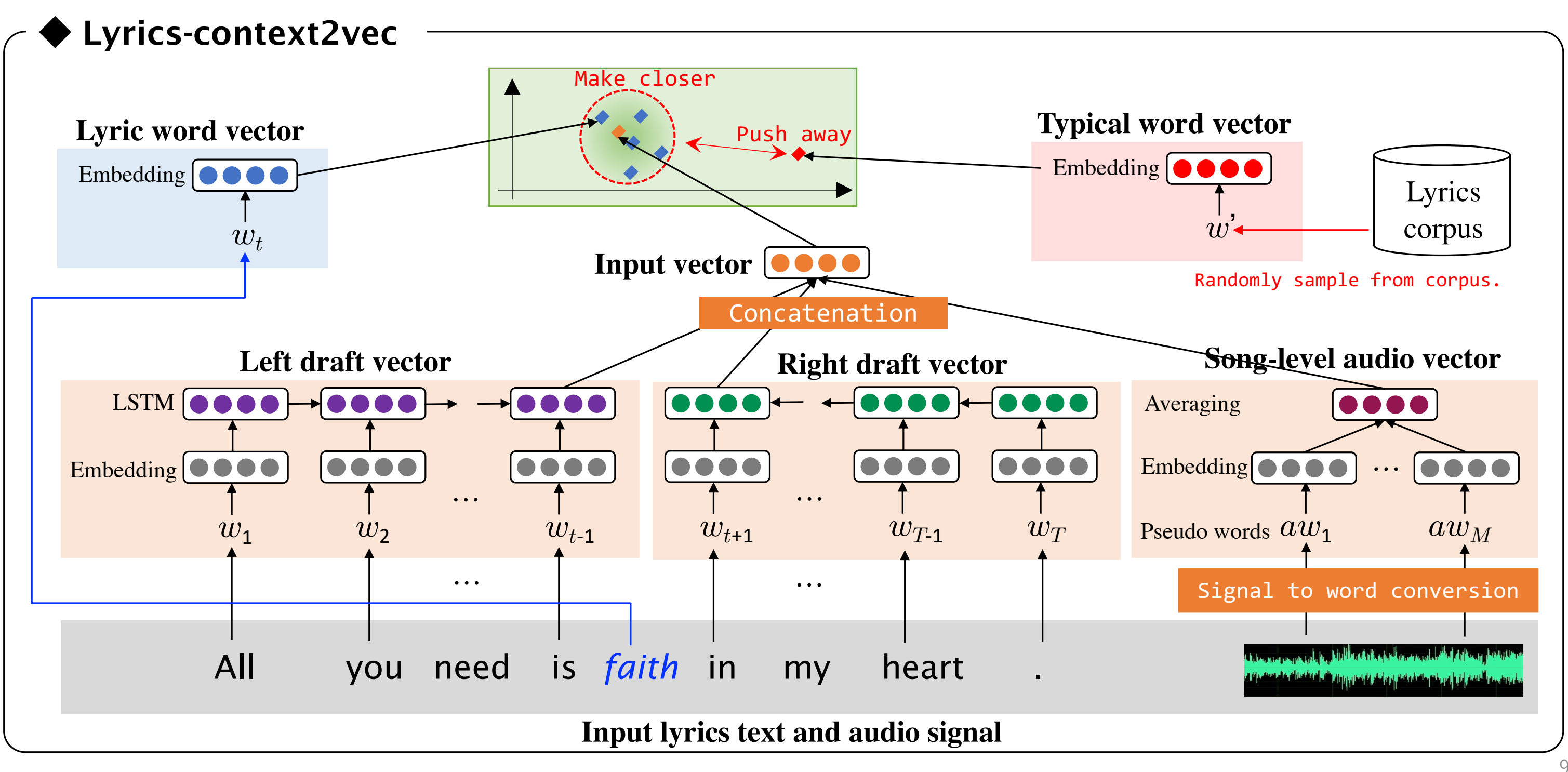
Training Lyrics-context2vec

We used contrastive learning by extracting positive and negative pairs from a large-scale dataset.



Model Construction

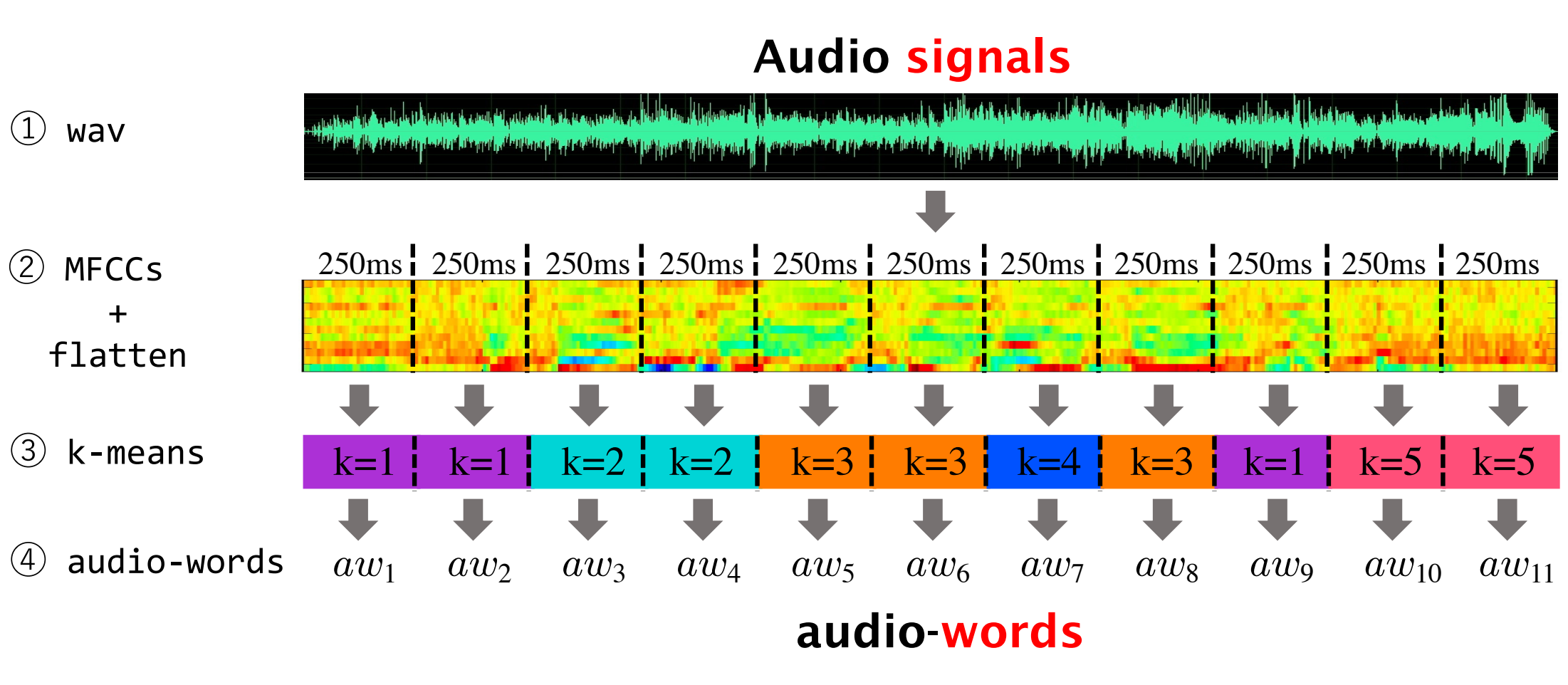
◆ Lyrics-context2vec



Signal to word conversion

◆ Bag of Audio-words

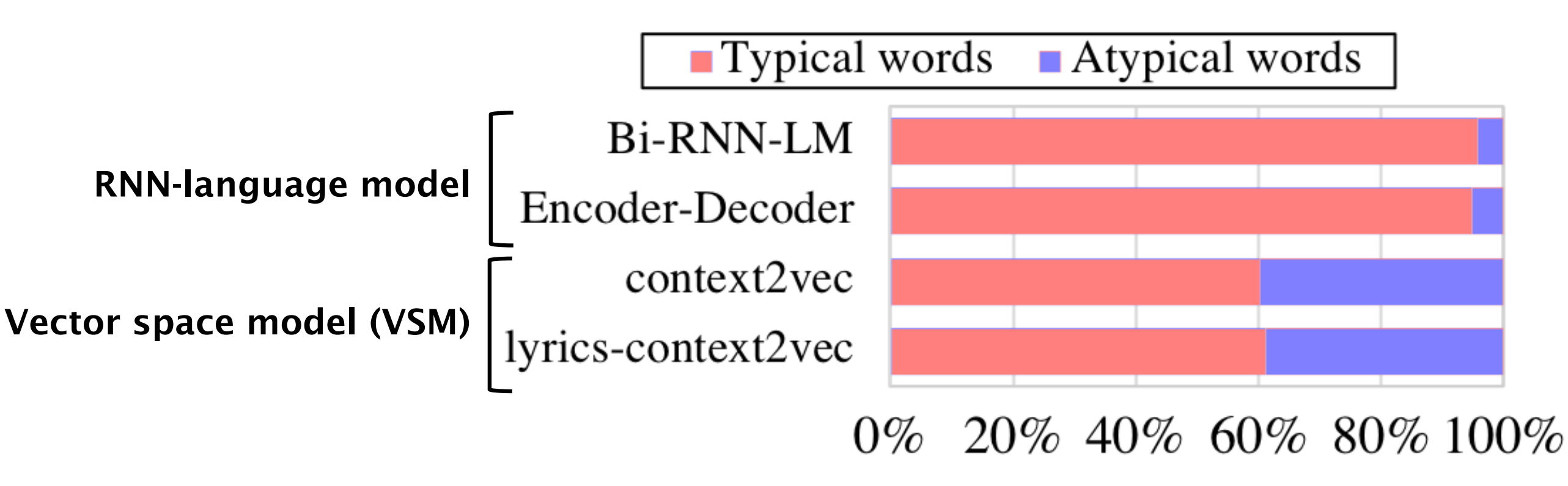
To embed audio signals, we use a discrete symbol called an audio-word.



The audio can be used with NLP techniques by converting the signal to symbols.

Experiments (1)

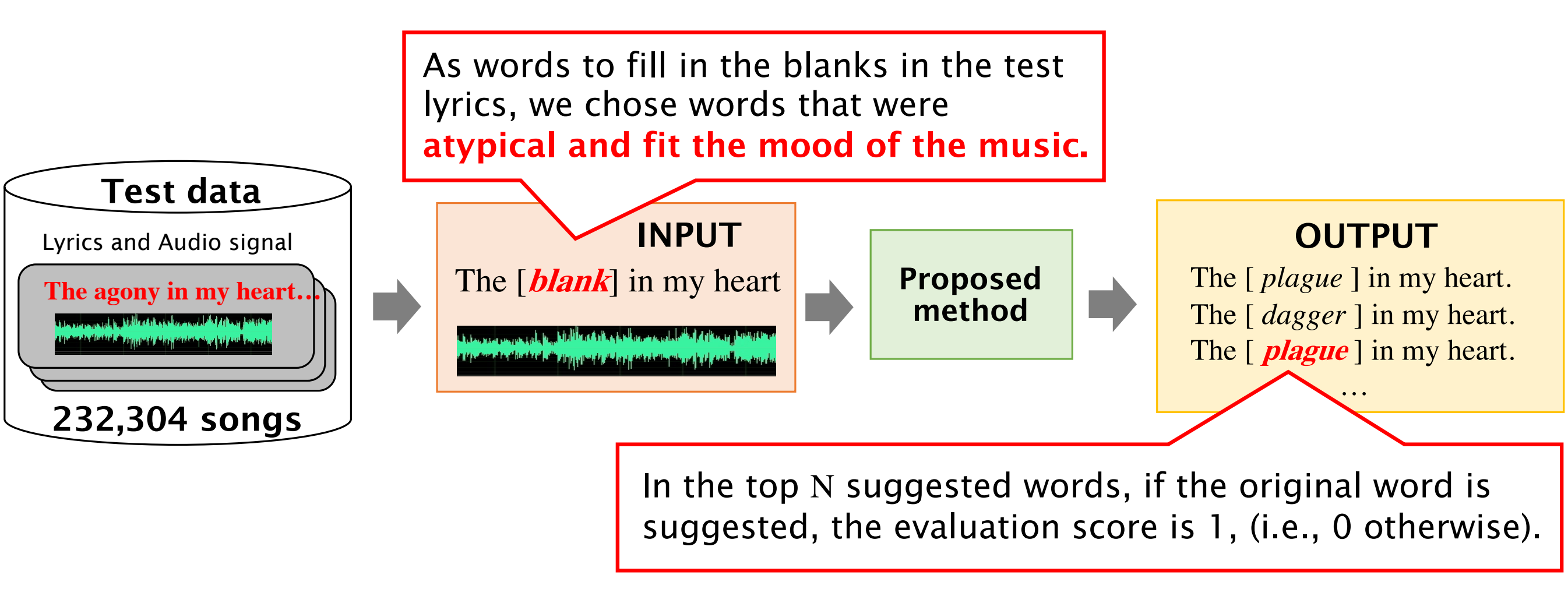
- We calculated how many of the top 20 words suggested by each model were typical or atypical.
- We calculated the document frequency of words and assumed that the top 10% of them are typical words (i.e., the remaining 90% are atypical words).



- VSMs suggested many more atypical words than RNN-LM type models did.
- Contrastive learning contributes to suppression of typical word completion while RNN-LM type models maximize the probability of predicting typical words.

Experiments (2)

We evaluated whether the proposed method could fill in the blanks of the test lyrics with the original words.

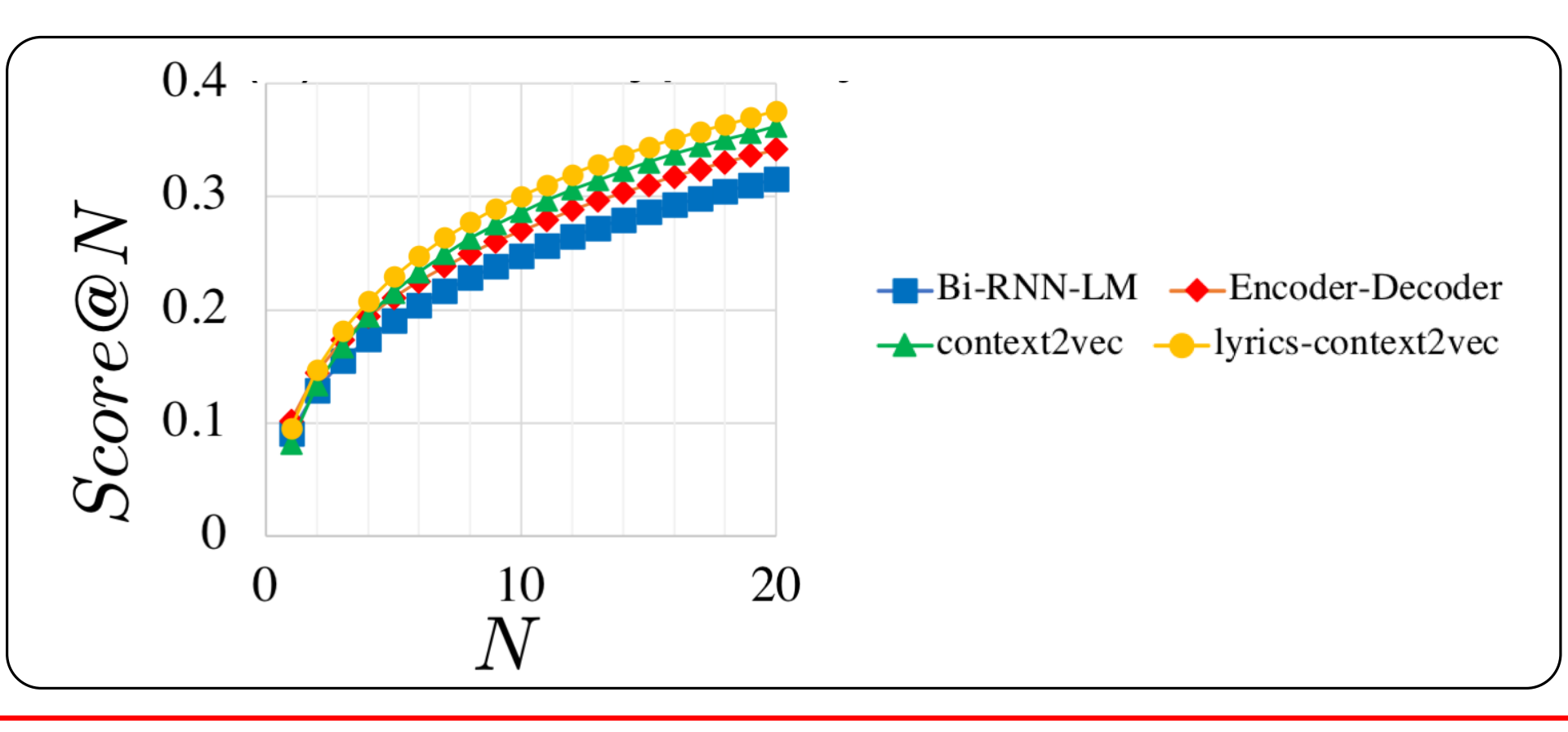


As words to fill in the blanks in the test lyrics, we chose words that were **atypical and fit the mood of the music.**

In the top N suggested words, if the original word is suggested, the evaluation score is 1, (i.e., 0 otherwise).

Experiments (2)

We evaluated whether the proposed method could fill in the blanks of the test lyrics with the original words.



- **Lyrics-context2vec predicted atypical words suitable for music audio better than all other models.**
- **Our model captures both the atypicality and the relationship between a music audio and words simultaneously.**

Examples

Input draft sentence: **The [*blank*] in my heart.**
Input Audio: **Killing time (Metal song)** 🔊

Method: LSTM-based Encoder-Decoder (previous approach)

pain, deep, burning, fire, hole, world, dead, darkness, feeling, beauty, devil, silence, drowning, shadows, words, dream, demons, power, wind, thunder

Method: Lyrics-context2vec (our approach)

hole, devil, burning, emptiness, **blackness**, pain, darkness, demons, **dagger**, hatred, **tremors**, void, fire, **agony**, holes, **essence**, **coldness**, **plague**, **needles**, deep

- **Our lyrics-context2vec can successfully suggest some rare words in bold fonts, such as "dagger" and "agony".**
- **Lyrics-context2vec successfully suggested explicit and negative words, such as "darkness", "agony", and "coldness". Those words are suitable for the mood of the input metal song.**

Conclusion

◆ Task

- We propose a lyrics completion task to recommend candidate words for a blank in a given sentence.
- Our task aims to suggest words that are **(1) atypical** and **(2) suitable for the musical audio signal**.

◆ Method

- We proposed **lyrics-context2vec, a multimodal vector space model** that suggests atypical but appropriate words for the given music audio and draft sentence.
- Input vector and output word vector located near each other.
- Vectors of highly frequent word located far away from the input vector.

◆ Findings

- Our contrastive learning strategy contributes to suggesting atypical words.
- Embedding audio signals contributes to suggesting words suitable for the mood of the provided music audio.