Modeling Structural Topic Transitions for Automatic Lyrics Generation

Kento Watanabe[1], **Yuichiroh Matsubayashi**[1], **Kentaro Inui**[1], and **Masataka Goto**[2]

[1]Graduate School of Information Sciences Tohoku University, JAPAN
[2]National Institute of Advanced Industrial Science and Technology (AIST), JAPAN
[1]{kento.w, y-matsu, inui}@ecei.tohoku.ac.jp
[2]m.goto@aist.go.jp

Abstract

By adopting recent advances in music creation technologies, such as digital audio workstations and singing voice synthesizers, people can now create songs in their personal computers. Computers can also assist in creating lyrics or generating them automatically, although this aspect has been less thoroughly researched and is limited to rhyme and meter. This study focuses on the structural relations in Japanese lyrics. We present novel generation models that capture the topic transitions between units peculiar to the lyrics, such as verse/chorus and line. These transitions are modeled by a Hidden Markov Model (HMM) for representing topics and topic transitions.

To verify that our models generate context-suitable lyrics, we evaluate the models using a log probability of lyrics generation and fill-in-the-blanks-type test. The results show that the language model is far more effective than HMM-based models, but the HMM-based approach successfully captures the inter-verse/chorus and inter-line relations. In the result of experimental evaluation, our approach captures the inter-verse/chorus and inter-line relations.

1 Introduction

Recent music creation technologies such as digital audio workstations and singing voice synthesizers (Kenmochi and Oshita, 2007) have become immensely popular among enthusiasts of automatically created or vocally synthesized music. These technologies assist individuals with their musical creativity and thereby have promoted automatic song generation. To date, many individual and group musical amateurs have created songs and commercial activities. To satisfy the demand for composer-supportive automatic composition systems and services, various systems, including Orpheus (Fukayama et al., 2012), have been developed. Furthermore, as musical composition becomes easier, there is a growing need for automatic lyrics generation.

However, lyrics generation has yet to be thoroughly explored in the natural language processing field. While several works have tackled lyrics generation based on lyric-specific characteristics, current methods are limited to local contexts, such as single sentences, which cannot capture the overall structure of the generated lyrics (Barbieri et al., 2012; Ramakrishnan A et al., 2009; Reddy and Knight, 2011; Wu et al., 2013; Greene et al., 2010).

The contribution of our study is twofold: (1) To more comprehensively understand lyrics generation, we examine the characteristics or rules by which people identify Japanese lyrics writing and survey some previous methods. (2) Based on the survey, we construct three generation models as an initial step toward our aim. We focus on two types of information that are essential for lyrics creation: a language model for lyrics and topic transitions for passages.

Experiments revealed that the language model is far more effective than models capturing topic transitions. However, by capturing the topic transitions, we achieve consistency among the topics.

28th Pacific Asia Conference on Language, Information and Computation pages 422–431

2 Related Work

Previous studies have attempted to reproduce characteristics specific to song lyrics, such as syntax, rhythm, rhyme, and the relation between melody and text. Barbieri et al. (2012) adopted a Markov process to create lyrics satisfying the structural constraints of rhyme and meter. They also ensured syntactical correctness by a part-of-speech template and computed the semantic relatedness between a target concept and the generated verse/chorus by a Wikipedia link-based measure. Our model extends Barbieri et al.'s approach to capture not only the semantic relatedness but also the verse-chorus transitions.

Ramakrishnan A et al. (2009) generated melodic lyrics in a phonetic language (in their case, Tamil). First, they labeled an input melody with appropriate syllable categories using conditional random fields and then filled the syllable pattern with words. Reddy and Knight (2011) developed a language-independent model based on a Markov process that finds the rhyme schemes in poetry and the model stanza dependency within a poem. However, rhyme transition in their model is used to generate a stanza; the overall flow of the poem is not captured.

Some researchers have generated lyrics using statistical machine translation. Wu et al. (2013) applied stochastic transduction grammar induction algorithms to generate a fluent rhyming response to the hip hop challenges allowing various patterns of meter. Using a finite-state transducer, Greene et al. (2010) assigned a syllable-stress pattern to every word in each line, subject to metrical constraints. Moreover, they generated English love poetry and translated Italian poetry into English following a user-defined rhythmic scheme.

Although these works capture lyric-specific characteristics to some extent, the structural relations are limited to lines or local word contexts. To the best of our knowledge, no existing method accounts for the semantic relations among large structures, such as verses and choruses.

Inter-text structural relations are frequently considered in text summarization and conversation modeling. The summarization technique of Barzilay and Lee (2004) captures topic transitions in the text span by a hidden Markov model (HMM), referred to as a *content model*. Using HMM and a large amount of tweet data, Ritter et al. (2010) and Higashinaka et al. (2011) modeled the transition of speech acts in an unsupervised manner.

3 Survey on Lyric Writing Techniques

To create a comprehensive model for lyrics generation, we first investigated the characteristics or rules by which people proceed with lyrics writing in general. We surveyed five textbooks on Japanese lyrics writing (Endo, 2005; Takada, 2007; Aku, 2009; Ueda, 2010; Taguchi, 2012) and identified the common features as follows.

3.1 Consistency of Entire Lyrics

The lyrics preferably follow a consistent theme. Authors usually desire to convey a message in their lyrics, and they reflect their theme in their lyric topics. Frequently, the theme is indirectly expressed through a concrete story composed of who, what, when, where, and why information. Each lyric should be consistent in writing style, such as the point of view (first or third person), gender, and date.

3.2 Lyrics and Melody

Lyrics and melody are mutually dependent and influence each other during the creation process. Which comes first depends on the situation. If developing the melody first, the writer must concentrate on achieving a suitable melody through rhythm, phonetic length, and lyrical structure. They should also match the word intonation and accents to the melody to ensure that their lyrics can be both sung and heard.

Most songs contain some common melodies. However, listeners may experience dissonance when simultaneously hearing upbeat and downbeat melodies. Thus, the writer needs to share the tone and atmosphere of his/her lyrics in the same melody.

3.3 Musical Structure of the Lyrics

The structural units of lyrics are verse, bridge, and chorus. Each unit repeatedly appears and shares the same musical phrases. Consequently, rhythm and meter are common to shared among the same type of units. In addition, same-type units are often created as semantically similar topics, such as scene and emotion, or contrastive topics, such as different seasons and feelings.

In general, each unit plays a typical role in the storyline. For example, verses often describe a concrete scene or complementary topic that emphasizes a message in the following chorus. Furthermore, the lines of a single verse/chorus may relay a suitable order of topic transitions. In the example as follows, the first and second lines collectively describe a concrete scene. This description is followed by the protagonist's reaction to the scene in the third and fourth lines.

Example of relations between lines

(On the way home, it began to snow.)
帰り道 降り始めた雪 ——————— Scene
(It is touching your shoulder and melting.)
あなたの肩に 触れて 溶けてゆく —— Scene
(Time flies. Today went by fast, too.)
今日もまた あっという間だね —— Sentiment
(The weekend with you is almost over.)
あなたとの週末 終わってしまうの – Sentiment

Excerpt from "Everlasting" by Mayo Okamoto

3.4 Balance of Contents

Emotion and scene are often combined in a verse/chorus. For instance, if a verse/chorus expresses emotions alone, such as "I love you" and "I want you", the lyrics are insufficiently balanced to convey the theme. To ensure that their lyrics are easily understood and arouse empathy in listeners, writers should adopt lead-in scenes such as "The road has been long" and "Reflections in a pool". Similarly, maintaining the balance between *subjective* and *objective*, *concrete* and *abstract*, *positive* and *negative*, and *universal* and *novel* will prevent egocentricity in the lyrics.

3.5 Figure of Speech

Lyrical content is frequently emphasized by figures of speech such as rhyme, metaphor, double meaning, double negatives, interrogatives, onomatopoeia, inversion, repetition, and rewording. The grammatical patterns of the lyrical sentence construction markedly differ from those in the general text.

4 Lyrics Generation Task

As noted in the previous section, songwriters incorporate various features, such as theme, structure, and

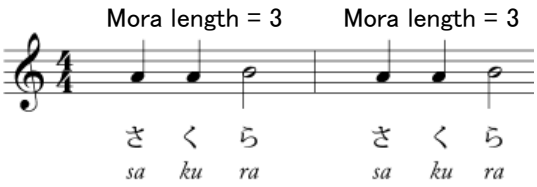

Figure 1: Example of a mora composed of musical notes. Japanese writers usually compose lyrics in such a manner that they can be easily sung by the singer. For example, the melody sequence "A-A-B" corresponds to "sa-ku-ra" (meaning "cherry blossom").

the lyrics-melody relation, into their lyrics, some of which are decided in advance. These predetermined features provide natural inputs to a lyrics generation task.

Some previously defined lyrics generation tasks account of the structural features and lyric-melody relation by inputting rhyme and meter. In our approach, the melody is replaced by the *mora* length of each phrase. In phonology, a mora specifies the period of a sound unit. For example, in Japanese, the mora length of the phrase *"帰り道, (ka-e-ri-mi-chi)" (On the way home)* is 5, whereas that of *"降り始めた雪, (fu-ri-ha-ji-me-ta-yu-ki)" (it began to snow)* is 8. In a Japanese song, a mora often corresponds to a musical note as shown in Figure 1.

In summary, if the input is provided as $M^{line} = [M_0^{phrase}, M_1^{phrase}] = [5, 8]$, the task generates lyrics such as *"帰り道 降り始めた雪" (On the way home, it began to snow)*.

Now, consider that the input includes partially composed lyrics. In this scenario, the system partially supports lyrical writer. For example, if the writer has completed a verse but is unsure of the chorus, the system can generate a chorus that is consistent with the completed verse. The experiments reported in Section 6 confirm that our models correctly capture the suitable topic transitions.

Our lyrics generation task is formally depicted in Figure 2. The task accepts the inputs as follows: (1) previously written parts of the target lyrics including an unwritten line and (2) sequences of mora length $M^{line} = [M_0^{phrase}, M_1^{phrase}, ...]$, each corresponding to the mora length of a line to be generated. The

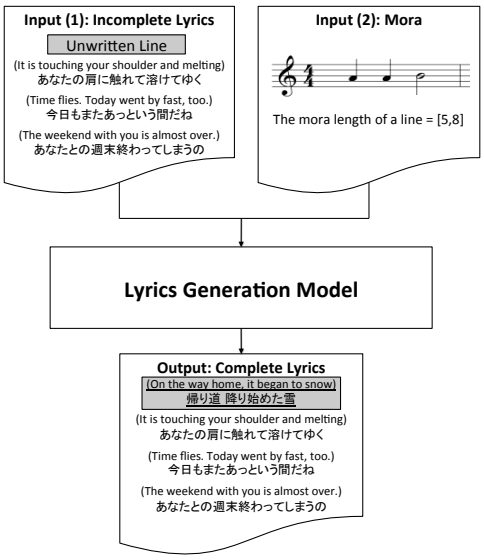

Figure 2: Lyrics Generation Task.

output of our lyrics generation task is a line that satisfies the restriction of the input mora.

5 Proposed Method

This section introduces the three generation models that capture some of the features introduced in previous sections.

In the models, (1) we utilize an n-gram language model assuming that lyrics are characterized by fluent, easily sung word orderings. In our models, the n-gram model is conditioned by the appropriate mora length.

Also, (2) we use a state-transition model assuming that the line and verse/chorus are generated from a consistent, context-dependent word set. Recall from Subsection 3.3 that each line and verse/chorus are often created as semantically related topics, and that topic transitions between lines and verses/choruses follow an appropriate order. We expand the content model (Barzilay and Lee, 2004) which originally estimates the topic transitions in documents using a hidden Markov model by assuming that each sentence has its hidden state representing its topic, to capture the inter-verse/chorus and inter-line relations.

Using these two components, we create three models illustrated in Figure 3: Although the model (a) employs a tri-gram model, the models (b) and (c) employ a bi-gram model to avoid data-sparsity due to the additional conditional parameter, the hidden state. We explain the details of each model in the next section.

5.1 Lyrics Generation Model

The inputs of the lyrics generation model are demonstrated in Figure 4. The positions of the verse/chorus, line, phrase, and word that should be generated are defined by i, j, k, and l, respectively. The mora lengths of the line that should be generated is assigned into the variable $M_{i,j}^{line}$. In Figure 4, the second line in the second verse/chorus should be generated, and the mora length of this line is given as input. We assigned **Line**$_i$ and **Verse** to the previously written lines and verses/choruses of the target lyrics including an unwritten line. These inputs were applied to the three generation models as shown in Figure 3.

(a) The first proposed model is the tri-gram language model $P(w_l|w_{l-1}, w_{l-2}, m_l)$ with mora restrictions (Equation 1), which assumes that a word is generated from its predecessors to satisfy the condition of fluent, easily sung lyrics. Note that m_l in this model is the mora length of the word and not the phrase; therefore, the model output is a sequence of the mora word lengths. For example, if the input is a mora length of the phrase $M_{i,j,k}^{phrase}$ = 7, the model should first generate a sequence of the mora word lengths $[m_0, m_1, m_2, m_3] = [3, 1, 2, 1]$, followed by a word sequence $[w_0, w_1, w_2, w_3] = [$“あなた (a-na-ta)”, “の (no)”, “肩 (ka-ta)”, “に (ni)” $(your\ shoulder)]$. Therefore, we specified that a sequence of words with the mora length **m** is generated with some probability $P(\mathbf{m}|M_{i,j,k}^{phrase})$. Thus, we have $\mathbf{m} = [m_0, ..., m_l, ...]$.

$$P(Line_{i,j}|M_{i,j}^{line}) = \prod_{k=0}^{|M_{i,j}^{line}|} P(\mathbf{m}|M_{i,j,k}^{phrase}) \prod_{l=0}^{|\mathbf{m}|} P(w_l|w_{l-1}, w_{l-2}, m_l) \quad (1)$$

(b), (c) The second and third proposed models is implemented for generating a consistent lyric. In generating a consistent lyric as described in Subsection 3.3, the topic transitions between lines and verses/choruses must be estimated. In this

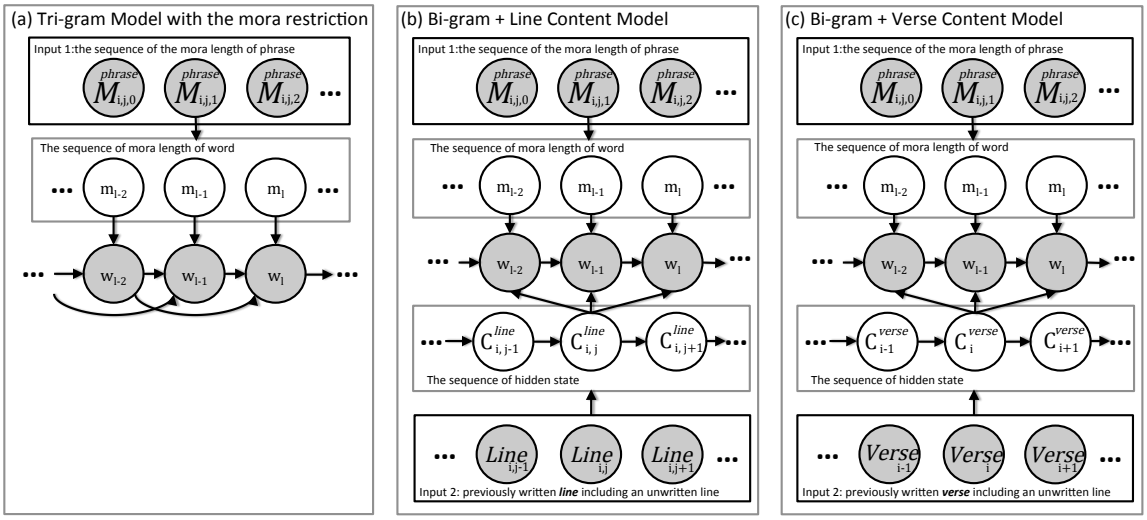

Figure 3: Lyrics generation model: (a) The n-gram language model generates a word given the line's mora length $M_{i,j}^{line} = [M_{i,j,0}^{phrase}, M_{i,j,1}^{phrase}, ...]$. The sequence of the mora lengths of the word $[m_0, m_1, ...]$ is generated from $M_{i,j,k}^{phrase}$. (b), (c) Based on the content model (Barzilay and Lee, 2004), the generation model captures the relations between lines and verses/choruses. The transition sequence of the hidden state $[C_{i,0}^{line}, ..., C_{i,j}^{line}, ...]$ or $[C_0^{verse}, ..., C_i^{verse}, ...]$ is estimated by specifying the context (already composed lines or already composed verses/choruses) and applying the Viterbi algorithm. Finally, the words are generated from each hidden state $C_{i,j}^{line}$ and C_i^{verse}, the mora length of the word m_l, and the previous word w_{l-1}.

study, the hidden state transitions between lines and verses/choruses in Japanese lyrics were learned by a content model (Barzilay and Lee, 2004). The features of the content model were bag-of-word-unigram containing the top 5,000 words in the training set, determined in a preliminary experiment. The hyper parameter in the content model training was set to 0.01. Next, we obtained the sequence of hidden states $\mathbf{C}_i^{line} = [C_{i,0}^{line}, ..., C_{i,j}^{line}, ...]$ and $\mathbf{C}^{verse} = [C_0^{verse}, ..., C_i^{verse}, ...]$; these are the topic transitions obtained by the Viterbi algorithm given the preferably written parts $\mathbf{Line}_i = [Line_{i,0}, ..., Line_{i,j}, ...]$ and $\mathbf{Verse} = [Verse_0, ..., Verse_i, ...]$ including an unwritten line. Finally, we specify the word generation probabilities $P(w_l|w_{l-1}, C_i^{verse}, m_l)$ and $P(w_l|w_{l-1}, C_{i,j}^{line}, m_l)$ to generate a word belonging to the hidden state C_i^{verse} and $C_{i,j}^{line}$ (Equations 2 and 3). In this study, a fluent, easily sung lyric has been generated from the previous word w_{l-1}, the mora length m_l of the word, and the hidden state C_i^{verse} or $C_{i,j}^{line}$. In contrast, the algorithms of Barzilay and Lee (2004), Ritter et al. (2010), and Higashinaka et al. (2011) use only the hidden state.

$$P(Line_{i,j}|C_{i,j}^{line}, M_{i,j}^{line}) = \prod_{k=0}^{|M_{i,j}^{line}|} P(\mathbf{m}|M_{i,j,k}^{phrase}) \prod_{l=0}^{|\mathbf{m}|} P(w_l|w_{l-1}, C_{i,j}^{line}, m_l) \quad (2)$$

$$P(Line_{i,j}|C_i^{verse}, M_{i,j}^{line}) = \prod_{k=0}^{|M_{i,j}^{line}|} P(\mathbf{m}|M_{i,j,k}^{phrase}) \prod_{l=0}^{|\mathbf{m}|} P(w_l|w_{l-1}, C_i^{verse}, m_l) \quad (3)$$

Although it appears that this method is restricted to the hidden state estimation for only one unwritten line, it is possible to extend this method for multipul unwritten lines by repeatedly applying the Viterbi algorithm after generating one line.

5.2 Model Estimation

Our generation model is estimated by maximum likelihood (Equations 4 and 5). The $count(*, w)$ returns the number of the occurrences of the word

The position of verse/chorus that should be generated: $i = 1$

The position of line that should be generated: $j = 1$

The position of phrase that should be generated: k

The position of word that should be generated: l

The mora length of line: $M^{line}_{1,1} = [M^{phrase}_{1,1,0}, M^{phrase}_{1,1,1}, M^{phrase}_{1,1,2}] = [7, 3, 5]$

The previously written lines including an unwritten line: $\textbf{Line}_1 = [Line_{1,0}, Line_{1,1}, Line_{1,2}, Line_{1,3}]$

$Line_{1,0}$ =“帰り道 降り始めた雪” (On the way home, it began to snow.)
$Line_{1,1}$ =Unwritten Line
$Line_{1,2}$ =“今日もまた あっという間だね” (Time flies. Today went by fast, too.)
$Line_{1,3}$ =“あなたとの週末 終わってしまうの” (The weekend with you is almost over.)

The previously written verses/choruses including an unwritten line: $\textbf{Verse} = [Verse_0, Verse_1, Verse_2]$

$Verse_0$ =“何も言わず そっと 肩を抱き寄せて, ...” (Please hug me softly, ...)
$Verse_1 = \textbf{Line}_1 = [Line_{1,0}, Line_{1,1}, Line_{1,2}, Line_{1,3}]$
$Verse_2$ =“5 分だけ, あと 10 分だけ, ...” (Just 5 minutes, just 10minutes, ...)

Excerpt from “Everlasting” by Mayo Okamoto

Figure 4: Example of the model input.

w (or a hidden state), and the W_{m_l} is the word set with the mora length m_l. To avoid the word sparseness problem, these probabilities are smoothed by Good-Turing discounting using SRILM (a toolkit for building and applying statistical language models) (Stolcke, 2002).

$$P_{ML}(w_l|w_{l-1}, w_{l-2}, m_l) = \frac{count(w_{l-2}, w_{l-1}, w_l)}{\sum_{w \in W_{m_l}} count(w_{l-2}, w_{l-1}, w)} \quad (4)$$

$$P_{ML}(w_l|C, w_{l-1}, m_l) = \frac{count(C, w_{l-1}, w_l)}{\sum_{w \in W_{m_l}} count(C, w_{l-1}, w)} \quad (5)$$

The generated sequence of the mora word lengths is simply estimated by maximum likelihood (Equation 6).

$$P(\mathbf{m}|M^{phrase}_{i,j,k}) = \frac{count(\mathbf{m})}{\sum_{\mathbf{m} \in M^{phrase}_{i,j,k}} count(\mathbf{m})} \quad (6)$$

6 Experiments

6.1 Evaluation Measure

The evaluation measure is another open problem in lyrics generation. Barbieri et al. (2012) and Wu et al. (2013) evaluated the generated lyrics by human annotation. However, because manual evaluations are expensive and time consuming, they are limited to a small number of test instances. Furthermore, the evaluation measures of an artistic quality strongly depend on the individuals; therefore, to achieve an evaluation measure of an adequate quality, copious annotation is required.

We evaluated our generation model by two different measures: log probability of the original line and fill-in-the-blanks-type testing. In the log-probability measure, we assumed that among all possible lines, the original line is generated with the highest probability. To calculate the log probability, the topic transition $[C_0, C_1, ...]$ was predetermined by providing the lines or verses/choruses and applying the Viterbi algorithm. The log probability, $log\big(P(Line)\big)$ of generating each line was then calculated as the logarithm of Equations 1, 2, and 3.

The fill-in-the-blanks-type test evaluates whether the correct line, selected from two candidates, is inserted into a given hidden line. One of the candidates is a correct answer randomly selected from the original song. The other candidate is an incorrect answer with the same mora length as the line from the original song but is randomly selected from another song. The candidate scoring the highest

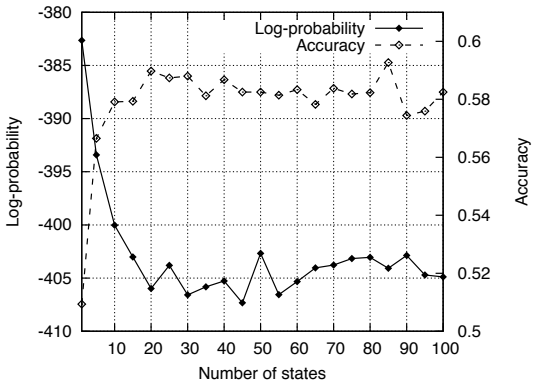

Figure 5: Log probability of whole lyrics generation (black circles) and the accuracy of a fill-in-the-blanks-type test evaluated on the development set, with the content model restricted to verses and choruses (gray diamonds).

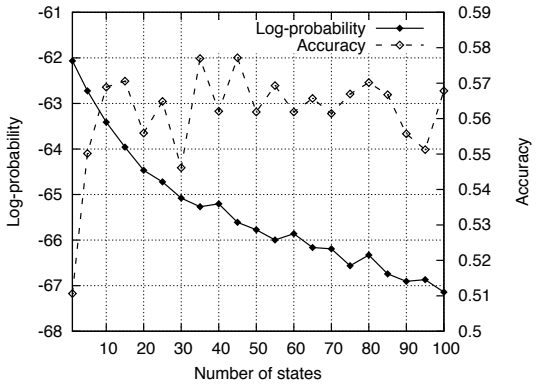

Figure 6: Log probability of verse/chorus generation (black circles) and the accuracy of a fill-in-the-blanks-type test evaluated on the development set, with the content model operated in the line mode (gray diamonds).

log probability is predicted as the correct answer. This measure checks whether the proposed model correctly captures topic transitions in each line or verse/chorus.

6.2 Dataset

The experiments were performed on Japanese popular music lyrics covering various genres, such as *Enka* [1] and *1970s pop*. Because our algorithm has limited capacity for calculating the mora length, foreign language songs were excluded in advance. The dataset contains 24,000 songs, 136,703 verses/choruses, 411,583 lines, and 61,118 words. We allocated 20,000 songs to the training set and reserved 2,000 songs each for development and testing.

6.3 Number of Hidden States

Prior to evaluation, the numbers of the hidden states in the two content models needed to be strategically selected to optimize the accuracy of the fill-in-the-blanks-type test on the development set. Figures 5 and 6 show the average log probabilities and accuracies of the fill-in-the-blanks-type test when applying this test to each content model. Note that the log probability shown in Figure 5 is $log\big(P(Lyrics)\big)$, calculated as the sum of the logarithm of Equation 2 over the entire lyrics. Similarly, the log probability shown in Figure 6 is the $log\big(P(verse/chorus)\big)$, calculated as the sum of the logarithm of Equation 3 in an entire verse or chorus.

In each case, the log probability decreases as the number of states increases because the bi-gram counts face data sparsity. However, the accuracy of the fill-in-the-blanks-type test monotonically increases and almost saturates at 10 states. Consequently, we specified 10 states in each content model.

6.4 Evaluation

Table 1 lists the average log probability, $log\big(P(Line)\big)$, of line generation in each model, evaluated on the development and test data. Although the content models partially include a bi-gram language model, the tri-gram model yielded the best performance. This result indicates the superior effectiveness of line generation by the language model than by the contents models.

Nonetheless, the content models capture a suitable order of lines. The average accuracy of the fill-in-the-blanks-type test is tabulated in Table 2. In this task, the counter-candidate line is randomly selected from another song and thus almost grammatical in construct. The main clue for accurate selection is a semantic relation between the topics. In this situation, the accuracy of the tri-gram model is equiva-

[1] *Enka* is a genre of a Japanese traditional ballad.

Model	Dev	Test
Tri-gram Model	-28.28	-29.02
Content model for Line	-38.36	-38.89
Content model for Verse	-33.03	-33.84

Table 1: Log probability of line generation.

Model	Dev	Test
Tri-gram Model	50.80%	47.97%
Content model for Line	56.89%	56.14%
Content model for Verse	57.91%	56.69%

Table 2: Accuracy of the fill-in-the-blanks-type test.

lent to the chance rate. On the other hand, both content models significantly improve the performance of lyrics generation.

We also qualitatively analyzed the obtained hidden states and their transitions in the content models. Table 3 illustrates the state transition table in the verse/chorus mode with 10 states. To clarify the discussion, we manually assigned easy-to-understand labels to the states and representative words to each hidden state (see Table 4).

Our content model in the verse/chorus mode concurrently learns two types of hidden states (Table 3). The first type corresponds to specific music genres; the second corresponds to specific tendencies of word appearances. The model successfully captures music genres with particular stylistic and vocabulary characteristics, such as *1970s pop*, *Enka*, and *modern songs*. Once a current state shifts into one of these states, it rarely shifts to another state. This indicates that the model generates suitable words that consistently fit the target genre.

Secondly, some states successfully capture the topics where the transition probabilities between them have some tendency; state transition probabilities are not random but instead biased against semantically related contents. As seen in Table 3, this type embraces five states, namely, *Scene*, *Memory*, *Sorrow & Love*, *Dream & Future*, and *Life & World*. In these topics, (1) the self-transition is the most likely one. (2) The transition probability from *START* to *Scene* is relatively high, and the transition probability from *Scene* to *END* is relatively low compared with the ones from others to *END*. (3)

Interpretation	Representative words
Scene	town (*machi*), room (*heya*), city (*tokai*), sunset (*yuuhi*), run (*hashiru*),
Memory	remember (*wasurenai*), memory (*omoide*), met (*deatta*), nostalgic (*natsukasii*),
Sorrow & Love	love (*koi*), express (*iu*), cry (*naku*), affections, sentiment, mind (*kimochi*),
Life & World	live (*ikiru*), future (*mirai*), bravery (*yuuki*), destination (*yukusaki*), reality (*genjitsu*),
Dream & Future	dream (*yume*), future (*mirai*), new (*atarashii*), world (*world*), one (*hitotsu*),
1970s pop	lie (*uso*), romance (*romansu*), rose (*bara*) lullaby (*rarabai*), kiss (*kuchiduke*),
Enka *traditional ballads*	human life (*jinsei*), harbor (*minato*), sake (*sake*), old home (*kokyou*),
Modern Song	paradise (*paradaisu*), cute (*kawaii*), drama (*dorama*), dance (*danse*),

Table 4: Representative Words in each Semantic Class.

The transition probabilities from *Memory* are almost even except for the self-transition and the transition to specific music genres (*1970s pop*, *Enka*, and *Modern song*). Therefore, *Memory* tends to play the role of an intermediate content in a lyrics. (4) The transition probability from *Sorrow & Love*, *Life & World*, and *Dream & Future* to *END* is relatively high. Thus the last verse/chorus in whole lyrics tends to become these three states. (5) *Life & World* and *Dream & Future* are strongly correlated. This indicates that the words representing hopes and bright futures tend to appear side by side.

7 Conclusion and Future Works

In this study, we presented content models for automatic lyrics generation that capture topic transitions in individual lines or verses/choruses. The content models are less capable of computing original line probabilities than the tri-gram model but better capture the inter-verse/chorus and inter-line relations. Currently, each model is separately constructed but the result suggested that combining these models would improve topic consistency. A multi-modal approach combining musical and lyrics information is also worthy of consideration. Some previous researchers have generated lyrics from musical information (Mihalcea and Strapparava, 2012; Hannah Davis, 2014). Musical information other than mora (such as rhyme, rhythm, melody, and chord) will be incorporated in the next version of our structured model.

		State after transition								
		END	Scene	Memory	Sorrow & Love	Life & World	Dream & Future	1970s pop	Enka	Modern Song
State before transition	START		**26.6%**	11.3%	2.4%	9.0%	**11.5%**	10.3%	**17.4%**	10.3%
	Scene	9.0%	**37.7%**	**12.9%**	3.7%	7.6%	9.7%	**10.8%**	1.2%	5.0%
	Memory	9.1%	**11.6%**	**38.1%**	**12.6%**	10.2%	11.0%	5.0%	0.3%	1.1%
	Sorrow & Love	**25.1%**	5.1%	12.5%	**32.1%**	6.1%	**13.2%**	4.6%	0.1%	0.5%
	Life & World	**14.0%**	5.1%	6.3%	4.5%	**49.2%**	**13.5%**	1.3%	0.2%	5.0%
	Dream & Future	**18.5%**	5.5%	6.9%	7.0%	**11.9%**	**45.5%**	2.5%	0.5%	0.7%
	1970s pop	**16.1%**	**10.5%**	3.6%	4.9%	2.0%	4.5%	**52.0%**	1.3%	2.9%
	Enka	**28.8%**	0.9%	0.3%	0.0%	0.4%	0.8%	**1.4%**	**65.5%**	1.3%
	Modern	**12.7%**	6.5%	1.2%	0.6%	**8.1%**	1.2%	3.9%	1.4%	**62.1%**

Table 3: Transition table between the hidden states of a verse/chorus. The vertical axis represents the hidden states before transition. The horizontal axis represents the hidden states after transition. Each cell contains the transition probabilities between the hidden states. The top three transition probabilities are shown in bold. To simplify the table, we omit hidden states that are erroneously reached from the start state.

References

Yu Aku. 2009. *Sakushi Nyumon (Introduction to Writing the Lyrics)*. Iwanami Shoten.

Gabriele Barbieri, Franois Pachet, Pierre Roy, and Mirko Degli Esposti. 2012. Markov constraints for generating lyrics with style. In Luc De Raedt, Christian Bessire, Didier Dubois, Patrick Doherty, Paolo Frasconi, Fredrik Heintz, and Peter J. F. Lucas, editors, *ECAI*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 115–120. IOS Press.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL*, pages 113–120.

Kozo Endo. 2005. *Sakushi Hon (Bool for Writing the Lyrics)*. Shinko-Music Entertainment.

Satoru Fukayama, Daisuke Saito, and Shigeki Sagayama. 2012. Assistance for novice users on creating songs for japanese lyrics. In *in Proceedings of the International Computer Music Association 2012*.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 524–533, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saif M. Mohammand Hannah Davis. 2014. Generating music from literature. pages 1–10. Workshop on Computational Linguistics for Literature.

Ryuichiro Higashinaka, Noriaki Kawamae, Kugatsu Sadamitsu, Yasuhiro Minami, Toyomi Meguro, Kohji Dohsaka, and Hirohito Inagaki. 2011. Building a conversational model from two-tweets. In *Workshop on Automatic Speech Recognition and Understanding*, pages 330–335.

Hideki Kenmochi and Hayato Oshita. 2007. Vocaloid — commercial singing synthesizer based on sample concatenation. In *in Proceedings of Interspeech 2007*, pages 4009–4010.

Rada Mihalcea and Carlo Strapparava. 2012. Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 590–599, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ananth Ramakrishnan A, Sankar Kuppan, and Sobha Lalitha Devi. 2009. Automatic generation of tamil lyrics for melodies. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, CALC '09, pages 40–46, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sravana Reddy and Kevin Knight. 2011. Unsupervised discovery of rhyme schemes. In *ACL (Short Papers)'11*, pages 77–82. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 172–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. pages 901–904.

Shun Taguchi. 2012. *Omoidori Ni Sakushi Ga Dekiru Hon (Book tha You Can Write the Lyrics)*. Rittor-Music.

Motoki Takada. 2007. *Sakushi No Kotsu Ga Wakaru (You Understand the How to Write the Lyrics)*. Chuo Art Publishing.

Tatsuji Ueda. 2010. *Yoku Wakaru Sakushi No Kyoukasho (The Textbook of the Lyrics)*. Ymaha Music Media.

Dekai Wu, Karteek Addanki, Markus Saers, and Meriem Beloucif. 2013. Learning to freestyle: Hip hop challenge-response induction via transduction rule segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 102–112, Seattle, Washington, USA, October. Association for Computational Linguistics.