Song Lyric Generation

By Michael Phaneuf and Dylan Edwards

I. Introduction

Writing song lyrics is generally viewed as a very human expression of feelings. They often have deep personal meanings that could only be understood by a human. We present a way to generate lyrics without the help of human creativity. As technology becomes a bigger and bigger part of our every day lives it only seems natural that it would find its way into our leisure activities. Technology is already a huge part of music production and could be a powerful tool to help artists write lyrics and find inspiration. We are interested in taking a genre such as country, rock, etc. and using machine learning to generate lyrics that fit the genre and make some sense semantically and grammatically.

II. Related Work

A. DopeLearning: A Computational Approach to Rap Lyrics Generation[1]

Researchers in Espoo, Finland created a rap lyric generation method to capture both the creativity and structure of rap lyrics. They first develop a prediction model to determine the next line given existing lyrics and a set of candidate next lines. They then use this prediction model to combine lines from existing songs to produce lyrics that rhyme and hold meaning. They use a couple machine learning techniques to achieve these goals including the RankSVM algorithm and a deep neural network. They find that their prediction model correctly predicts next lines 17% of the time and when used for the purpose of creating rhyme density, outperforms the best human rappers by 21%. Our problem differs from their work in a few key ways. They narrowed their lyric generation to specifically rap lyrics, while we include several other genres. They focus on specific aspects of rap as opposed to other genres, which allows them to generate lyrics better suited for that genre of music. They also look at their lyric generation method as a way for artists to get ideas or finish already created lyrics. Ours takes just a genre and generates all the lyrics from there.

B. AC/DC Lyric Generation[2]

Kirt Connor is known for scraping lyrics from specific artists and then using them to generate a new song that sounds like it was written by that artist. One example is the AC/DC song he made. He scraped AC/DC lyrics from the Genius Lyrics database and then used a Markov Chain to create a new song that seems like it could have been written by AC/DC. This approach differs from the approach we take in multiple ways. He trains his models on data from one specific artist to generate lyrics in their style. Instead of taking a genre name and generating lyrics to fit that genre, he takes an artist and generates lyrics to fit that artist. He uses a Markov Chain to generate his lyrics, while we use a language model with the underlying architecture being a neural network.

III. Methodology

A. Dataset

We were given an already cleaned dataset to work with that came from scraping lyrics from multiple websites. There are four genres of lyrics including country, metal, pop, and rock. The dataset consists of 100 song lyrics for each genre, all in English.

B. Model Creation

We first load all of the data of whatever genre we are fitting our model to. We then turn each word into integer-index representations. We use these to create all of the trigram and bigram word sequences. Finally, we make all the sentences equal length by padding them, since this is required by the machine learning library we use, Keras.

We define two LSTM models. One using bigrams with their target word as input sequences and one using trigrams with their target word as input sequences. We want to investigate whether the length of input sequences impacts the quality of our generated lyrics. The only difference between the two models is the length of the input sequences.

C. GloVe Word Embeddings

GloVe is a word vector technique similar to Word2vec. We choose GloVe over Word2vec because it goes beyond just local statistics and incorporates global statistics such as word co-occurence when constructing word vectors. These word embeddings help our model create an output that makes sense semantically. Word vectors are important because they add context to the equation. They group similar words together so that even if our model has not seen a specific sequence of words it can still output that sequence if it makes sense semantically.

IV. Results and Analysis

Our goal is to take a genre and generate lyrics that make sense and fit the given genre. Here is one example that our trigram model generated given the genre country:

[1] Malmi, Eric, et al. “Dopelearning: A Computational Approach to Rap Lyrics Generation.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, https://doi.org/10.1145/2939672.2939679.

[2] https://brobible.com/culture/article/new-acdc-song-great-balls-ai/

My happiness depends on you and whatever you decide to do jolene jolene
And so he became my grandchild for he was my daughter's wife
I just want to thank you lord thank you lord thank you lord
And I don't feel at home in this world anymore oh lord you
I slashed a hole in all four tires maybe next time he'll think
But I'm alive and well in Tennessee me the first time saying this
I hope you think of me and i'm back on all of that
Yeah it would be easy not to miss you wonder about who's with
Who will bear my light to them whom shall i send here i
Just as easy as can be just the way you laugh like you

Clearly this is not perfect, but it has distinct country themes such as religion, heartbreak, and Tennessee. It also makes sense semantically and grammatically, for the most part. To further evaluate the results of our models we asked a few random people to look at our generated lyrics for each model and genre and tell us which they thought was the better song and how it compared to a real song in that genre.

Our first respondent picked the trigram model for all genres except for metal and told us that each of the lyrics clearly fit the genres they were intended for, but it was clear they were not written by a talented musician. Respondent number two picked the trigram model for metal and rock, but choose the bigram model for pop and country. He told us that some of the lyrics seemed like they could have been written by a real musician, but overall you can tell it is not a real song. Finally, our third respondent picked the trigram model for metal and rock, but choose bigram for pop and country. Similar to the other respondents, she said that the lyrics mostly made sense and fit the intended genre, but were clearly not written by a professional.

Based on these responses we conclude that our models perform well and that both models generate lyrics similarly well. None of our respondents chose the same model for every genre. Our generated lyrics clearly fit the category that they are intended for and mostly make sense grammatically and semantically. It would be unrealistic to expect our generated lyrics to be indistinguishable from a real song as language processing technology is simply not advanced enough yet. Constructing a full song of lyrics that make sense and sound good together is an art form that can not yet be accomplished by computers.

V. THREATS TO VALIDITY

There are a few threats to the validity of our results. There are very limited evaluation metrics for this type of model. The quality of song lyrics is very subjective and people's opinions may differ drastically. Our dataset is also somewhat small compared to what might be ideal. We only have lyrics from 100 songs for each genre. Our result might be more convincing if we had a larger dataset. We also only asked three people for their opinions on our generated lyrics. Ideally we would be able to ask hundreds or thousands of people so that we have more evaluations to look at and more conclusions to draw from.

VI. CONCLUSION

We set out to create a language model that takes a musical genre and outputs lyrics that fit with the given genre. We constructed two LSTM models, one with bigram input sequences and one with trigram input sequences. We use GloVe embeddings to add semantic information to our models. Our model successfully outputs lyrics that fit with each genre. We asked three random people to look at the lyrics generated by each of our models and tell us which they liked better and how they compared to real lyrics. There was no clear cut winner in terms of which model was better, but everyone agreed that the lyrics fit their intended genres. It was also agreed upon that the lyrics clearly were not written by a professional, however this is something we expected. We are happy with the results of our model as it generates lyrics that make sense semantically and grammatically and fit specific genres. The fact that we are using humans to evaluate our lyrics implies that the validity of our results may be jeopardized. The quality of song lyrics is very subjective and opinions can vary drastically among individuals. We would need to collect a lot more data and test our results among a much larger group of people to make our conclusions more robust. There is a lot of opportunity for future work on this subject. This type of model could be improved upon and used to give artists creative inspiration. At some point we might even see computer generated songs topping the musical charts.