GramError: a quality metric for machine generated songs.

DAVIES, C., WIRATUNGA, N., MARTIN, K.

2018

GramError: A Quality Metric for Machine Generated Songs

Craig Davies, Nirmalie Wiratunga[0000−0003−4040−2496], and Kyle Martin[0000−0003−0941−3111]

Robert Gordon University, Aberdeen, Scotland
{c.davies, n.wiratunga, k.martin}@rgu.ac.uk

Abstract. This paper explores whether a simple grammar-based metric can accurately predict human opinion of machine-generated song lyrics quality. The proposed metric considers the percentage of words written in natural English and the number of grammatical errors to rate the quality of machine-generated lyrics. We use a state-of-the-art Recurrent Neural Network (RNN) model and adapt it to lyric generation by re-training on the lyrics of 5,000 songs. For our initial user trial, we use a small sample of songs generated by the RNN to calibrate the metric. Songs selected on the basis of this metric are further evaluated using "Turing-like" tests to establish whether there is a correlation between metric score and human judgment. Our results show that there is strong correlation with human opinion, especially at lower levels of song quality. They also show that 75% of the RNN-generated lyrics passed for human-generated over 30% of the time.

Keywords: Natural Language Generation · Quality Metric · Recurrent Neural Network

1 Introduction

Artificial Intelligence (AI) has in recent years proved to be effective in a variety of Natural Language Processing (NLP) applications such as neural translation [2] and caption generation [12]. Much of this has been driven by the success of sequence-to-sequence learning algorithms [11]. Model learning and evaluation commonly relies on metrics (such as BLEU and NIST) which were developed for machine translation tasks [10, 3]. These metrics compare the machine-written output of a translation task to a human-written translation and rate the quality based on the similarity of the two translations, with a more similar translation said to be of higher standard. This makes it less suitable for creative Natural Language Generation (NLG) tasks, as it is virtually impossible to get a human and a machine to write creatively in a way that can be compared in any meaningful form on the basis of exact matching.

In our work, we explore the role of lyrical composition and consider semantics such as sentence structure, language formation and punctuation to formulate an alternative evaluation metric for song generation. With this in mind, we introduce a simple metric, GRAMERROR, that analyses the lyrical composition of songs. We show that such a metric correlates well with human evaluation of song lyrics. The metric proposed in this paper is studied using songs generated by a sequence-to-sequence model trained

on a corpus of songs from the classic rock genre. We created user tests to allow us to compare human judgments with metric scoring. Here judgments refer to the ability to identify whether a song was written by a machine. We found a strong correlation between increasing metric generated scores with decreasing human judgment accuracy. Our results also suggest that 75% of the lyrics generated by the Recurrent Neural Network (RNN) passed for human-generated over 30% of the time when compared adversarially to human-written song lyrics. These results demonstrate the utility of the lyrical composition metric proposed in our work.

In this paper, we explore work on lyric generation and discuss why existing automated metrics are unsuitable for rating the quality of work produced by NLG systems (Section 2). Our proposed metric appears in Section 3 followed by the song generation model in Section 4. The evaluation in Section 5 presents the utility of GRAMERROR following a user survey and two user evaluation trials. Conclusions appear in Section 6.

2 Related Work

There have been several attempts at using NLG systems for song lyric generation [9, 8]. Templates [6] are commonly used to guide the NLG process where the focus is on using existing music and breaking the beat down into rhythm patterns (beats, rests). Thereafter random words are selected to fit the rhythm whilst the generative grammar strategy writes words that best fit pre-written grammar templates. Another approach is referred to as a Generate and Test strategy since sentences are generated on the basis of the grammar templates and tested on the basis of a best fit ranking. Here fit relates to rhythm, rhyme (with other lines) and number of syllables. Oliveira's PoeTryMe [7] is a typical example of this strategy.

Machine evaluation of song quality is often challenging due to the multi-faceted criteria. Work in NLG (although less complex than song analysis) has successfully adopted overlap metrics such as BLEU and NIST [10, 1]. Although these metrics were not originally intended for use in creative NLG, BLEU has been used to rate generated poetry [5] and results suggest a good correlation with human judgment quality, but only for low BLEU scores (between 0.11 and 0.15). One problem with using such metrics is the reliance on reference texts (for the comparison). Even if such a resource existed, creative writing is not about matching previous content from the past.

3 A Grammar Based Quality Metric

In formulating a quality metric for lyric comparison, we take inspiration from findings reported in [8], where generative models with a focus on grammar were shown to outperform more complex models. The strong performance of a grammar-based model performed suggested that a metric which considered the quantity and severity of grammatical errors machine-written song lyrics would be effective. In our work we utilise the popular online grammar correction system "Grammarly" [1] to evaluate text and categorise errors into "major grammatical errors" and "minor grammatical errors". Thereafter a metric is used to aggregate these errors to form the quality score.

[1] www.grammarly.com/

Major grammatical errors typically relate to punctuation errors or spelling mistakes, whilst minor grammatical errors are related to poor sentence structure. Accordingly we combine both these error counts using a weighted formula as follows:

$$\text{GRAMERROR} = \alpha pE - \beta MG - \gamma mG - C \quad (1)$$

Here pE is the percentage of words in the song written in natural English, MG is the number of Major grammatical errors, mG is the number of minor grammatical errors and C is a constant that manages the sensitivity of the score. Here α, β and γ are mixing weights used to help calibrate the metric. This is informed by our findings from an initial user survey discussed in Section 5. After learning the weights we assign a value to C such that the metric is capped at 100. This offers greater granularity when compared to a lower cap such as 10, which would lack the sensitivity to measure the calibre of songs that might be similar in quality. The metric has no lower limit.

4 Generating Song Lyrics with an RNN

A Recurrent Neural Net (RNN) was used to generate song lyrics. The recurrent nature of this model means that the output at each step is fed back into the network and provides important contextual information based on word locality, improving the semantic coherence of its output. As our aim was to test how well the proposed metric correlates with human opinion, we adapted a pre-trained RNN from an existing NLG system which generated plays in the style of William Shakespeare by retraining it on a Kaggle dataset scraped from the Lyrics Freak website [2]. This dataset contains the English lyrics for 57,650 songs and informed two aspects of our work. To retrain the RNN to suit the task of lyric generation, we extracted 5,000 classic rock songs to form the training set. Classic rock was selected because the RNN proved susceptible to repetition in the training set, meaning repetitive genres (i.e. pop music) demonstrated poorer performance.

Due to the size of this dataset, we used a mini-batchs to iteratively train the model. We also extended the network by adding a further hidden layer as we found it improved the quality of generated lyrics. Thus we constructed an RNN with 4 hidden layers, 3 of which used tanh activation functions while the final layer used a softmax classifier to feed into a cross-entropy loss function using Adam optimiser. We then re-trained the model for 20 epochs on the training set of 5,000 classic rock songs. Note that we needed few epochs to retrain the system as we were adapting an already competent network to the task of lyric generation. The sequence length of the data was set at 50. This parameter identifies the max length of a sequence that can be generated by the RNN. The batch size was set to 256. These parameters were selected following extensive empirical evaluation.

5 Evaluation

The aims of our evaluation were two-fold. Firstly, we wished to identify appropriate weights for our proposed metric. Secondly, we aimed to measure the correlation between the metric and human judgment of lyric quality. Accordingly we defined a two

[2] https://www.kaggle.com/mousehead/songlyrics

stage evaluation. In the first stage we used a survey to measure the influence of major and minor grammatical errors on human judgment of song quality. This allowed us to calibrate the mixing weights of GRAMERROR and improve its scoring. In the second stage we performed 'Turing-like' tests to investigate whether our metric was indicative of when machine-generated lyrics would be confused for human-written songs.

Test Survey To determine how well GRAMERROR correlates with human opinion, users were recruited to complete a survey on song quality. Users were presented with 5 bot-written songs and asked to answer four questions using one of six possible answers. The questions were evenly weighted with the most positive answer worth 25 points. The highest achievable score was therefore 100 (designed to match the maximum score GRAMERROR could give), enabling us to measure the discrepancy between GRAMERROR scoring and human ratings to improve calibration of weights in our equation.

The Turing-like Tests We adopted a two-stage 'Turing-like' evaluation to investigate whether lyrics scored highly by our metric would be more likely to be mistaken for human-written. In the first stage, users were presented with 13 songs (4 human-written, 9 machine-generated) in isolation and asked to identify whether the song was written by a bot or a human. We used 8 songs generated by our RNN and a song presented in [4]. We observed that it received the highest possible score by GRAMERROR, thus giving it the best chance of passing the Turing test and forming an upper bound for our tests.

The second stage only used machine-generated songs which passed for human a threshold percentage of times (30%) in the first stage. Users were presented with a bot-written song alongside human-written lyrics and were asked to identify which was which. We could then measure the correlation between GRAMERROR and human opinion when lyrics were in isolation and when presented alongside human-written songs.

5.1 Results

The Test Survey In total, 13 users responded to the first survey. When compared to an uncalibrated (i.e. all mixing weights set to zero) version of GRAMERROR, the total discrepancy between the proposed metric and human ratings was 154 points. We observed that GRAMERROR more accurately reflected human opinion at the lower end of the rating scale, showing deviation of just 3 points on the lowest ranked song, but differing by 28 points on the highest rated song.

These results indicated that GRAMERROR should rate songs more strictly. Iteratively testing weights for each variable in the metric, we found the optimal formula:

$$\text{GRAMERROR} = 4pE - 2MG - 2mG - 300 \tag{2}$$

suggesting that human evaluators assign equal weight to major and minor grammatical errors. The constant C was set to 300, capping the metric's score at 100. After this recalibration the total discrepancy between GRAMERROR scores and user ratings was reduced to 114. Figure 1 shows the comparison between the recalibrated GRAMERROR scores and the human rating for five of the songs.

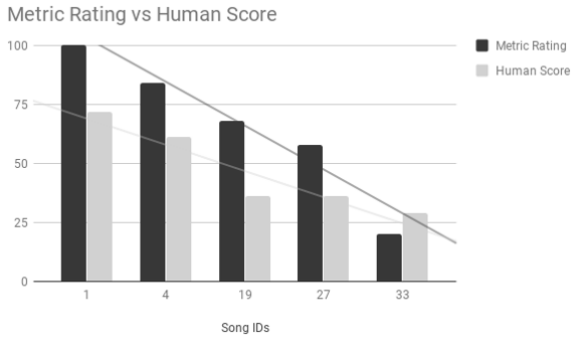

Fig. 1: Rating from Metric scoring vs Human Ratings

The Turing-like Tests In the first stage, 14 users completed the survey. Their responses are summarised in Figure 2. We observe that a song's GRAMERROR score correlated well with how often it was thought to be human-written when presented in isolation. Four of the machine-generated songs were labeled as human-written by 30% or more users, including the three songs highest rated by GRAMERROR. The three songs which received the lowest score from our metric were labeled as human by less than 20% of users. Dropout was low, with only 3 users failing to complete the survey after starting.

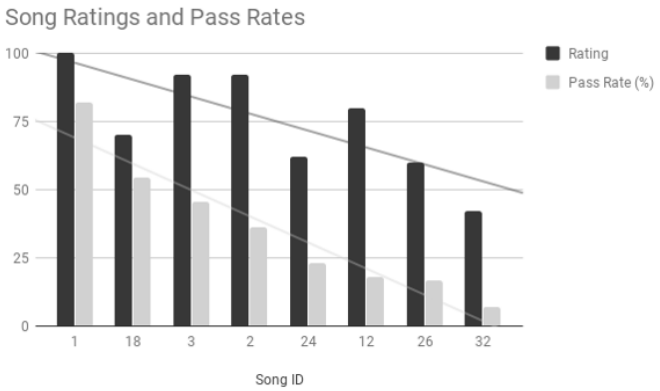

Fig. 2: Comparing Automated ratings and Test 1 Pass Rates

The second Turing-like test demonstrated similar results. Using the four machine-generated songs which had been mislabeled as human most often in the last test, we observed that three of the four machine-written songs were passed for human by over

30% of tested users once more. This suggests that GRAMERROR's scoring is robust to whether songs are viewed in isolation or alongside human-written lyrics. The response rate for this test was high - 62 people started the test and only 4 failed to finish.

6 Conclusions

In this paper we introduced a new grammar focused metric, GRAMERROR, for measuring NLG. When this metric is used to rate song lyrics, it matches human judgment reasonably well, with a difference of 9% between the metric and user opinion at lower levels of quality and a difference of 28% at higher levels of quality. The metric also correlates well with how likely a set of song lyrics is to pass for a set of human-written song lyrics, both when in isolation and when alongside a human-written counterpart.

References

1. Adeyanju, I., Wiratunga, N., Lothian, R., Sripada, S., Lamontagne, L.: Case retrieval reuse net (cr2n): An architecture for reuse of textual solutions. In: Case-Based Reasoning Research and Development. pp. 14–28. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)
2. Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing. pp. 1724–1734. ACL (2014)
3. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proc. of the Second Int. Conf. on Human Language Technology Research. pp. 138–145. Morgan Kaufmann Publishers Inc. (2002)
4. Ghazvininejad, M., Shi, X., Choi, Y., Knight, K.: Generating topical poetry. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. pp. 1183–1191. ACL (2016)
5. He, J., Zhou, M., Jiang, L.: Generating chinese classical poems with statistical machine translation models. In: Proc. of the Twenty-Sixth AAAI Conf. on Artificial Intelligence. pp. 1650–1656. AAAI'12, AAAI Press (2012)
6. Oliveira, H.G.: Automatic generation of poetry: an overview. Universidade de Coimbra (2009)
7. Oliveira, H.G.: Poetryme: a versatile platform for poetry generation. Computational Creativity, Concept Invention, and General Intelligence **1**, 21 (2012)
8. Oliveira, H.G.: Tra-la-Lyrics 2.0: Automatic Generation of Song Lyrics on a Semantic Domain. Journal of Artificial General Intelligence **6**, 87–110 (2015)
9. Oliveira, H.G., Cardoso, F.A., Pereira, F.C.: Exploring different strategies for the automatic generation of song lyrics with tra-la-lyrics. In: Proc. of 13th Portuguese C. on Artificial Intelligence. pp. 57–68 (2007)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. ACL '02, ACL, Stroudsburg, PA, USA (2002)
11. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. pp. 3104–3112. NIPS'14, MIT Press, Cambridge, MA, USA (2014)
12. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. pp. 3156–3164 (2015)