Lyrics Generation supported by Pre-trained Models

Matheus Augusto G. Rodrigues, Alcione de P. Oliveira, Alexandra Moreira, Maurilio de A. Possi
Federal University of Vicosa
Vicosa - Minas Gerais - Brazil
{mathinvoker,xandramoreira}@gmail.com, {alcione,maurilio}@ufv.br

Abstract

Advancements in neural network architectures have improved the quality of several tasks in computational linguistics. Among the tasks benefited we can mention question and answer systems, dialogue systems, opinion mining and the automatic generation of texts, just to mention a few. Despite the advances, there is still room for contributions, since there are still open problems. In the case of text generation, especially in the musical genre, there are challenges for the production of texts that involve poetry and idioms. In particular, some of these challenges are linked to the treatment of metaphors and metonymy and the generation of paraphrases. This paper presents an analysis of the generation of excerpts of lyrics based on a pre-trained GPT-2 neural network model, after fine-tuning with two lyrics corpora, one in English and one in Portuguese. An analysis of the spelling, syntax and semantics of the generated texts are presented, as well as the discussion about the attempt to find a pattern in the sections generated by the implemented tool. Research demonstrates the potential for using such models in the generation of poetic texts.

Introduction

Natural Language Processing (NLP) has made significant progress over the past decade, building on recent deep neural network architectures (Deng and Liu 2018). Tasks within the scope of NLP have reached new levels in the state-of-the-art, such as machine translation, opinion mining, dialogue and question and answer systems and text generation. Today, it is possible to acquire, in commercial stores, personal assistants that interact through natural language with some fluidity, such as Alexa, Siri and Google Assistant (Hoy 2018).

Among the tasks related to NLP, one that has attracted the attention of researchers is the natural language generation (NLG). This is an attractive task as it can be combined with other tasks, generating more natural and fluid dialogue and question/answer systems. It is also an important task, in isolation, aiming to produce unpublished texts, with performance equivalent to that of human beings (Gatt and Krahmer 2018). As (Gatt and Krahmer 2018) have stressed, defining exactly what NLG is is more challenging than it initially appears. Although the final product is composed of natural language expressions, what was used as the input for producing the text can vary enormously. Just to cite a few examples, there are systems that produce texts from images (Vinyals et al. 2015), from a single other text, as in the case of text summarization (Tas and Kiyani 2007), or from a neural model previously trained using a linguistic Corpus (Radford et al. 2019). This work follows in the latter case. More recently, thanks to advances in hardware and architecture of neural networks, it was possible to develop neural models with billions of parameters, as is the case of GPT-2 (Radford et al. 2019) and GPT-3 (Brown et al. 2020), both developed by OpenAI. These models were trained with a high computational cost and, consequently, a high energy cost (Strubell, Ganesh, and McCallum 2020) and, therefore, need to be incorporated into other systems, through transfer learning, so justify the cost of their training and avoid major environmental impacts.

In this paper, it is verified whether it is possible to use one of these models, more specifically the GPT-2, in the task of generating a natural language of poetic text/song lyrics. The choice of the aforementioned model was due to the fact that, at the time the work in question began, GPT-2 was the state of the art among language models. The characteristics of the implemented model based on the samples (song lyrics, poetry) generated from the execution of the model are discussed. The result is evaluated with respect to the textual structure, the patterns detected in the generated texts, the proximity of the samples with the lyrics of songs composed by human beings, among other criteria. Algorithms were implemented to generate textual samples using simplified models of GPT-2 provided to the public by OpenAI, generating a total of 10 samples for further analysis and discussion. The model was trained with a corpus of English lyrics prepared by (Rodrigues, Oliveira, and Moreira 2019). A corpus of lyrics in Portuguese extracted from the website Vagalume (https://www.vagalume.com.br/) was also used, for the purpose of analyzing the potential of architecture in the generation of lyrics in Portuguese.

The article is organized as follows: in the next section we discuss the works related to this research. After that, we present the approach taken, along with the resources used for research. Next we present the results achieved and, finally, we present the conclusions of the work.

Related Works

Park and Ahn, (Park and Ahn 2018) present a model for automatic generation of sentences from keywords using LSTM (Long Short-Term Memory) recurrent neural networks organized in the form of adversarial networks (Generative Adversarial Network - GAN). The model also includes a self-attention module. The authors use synonymous keywords as input to the model in order to improve the number of distinct sentences generated. The model proposed by the authors performed better than models that do not use GANs. Contrary to the model proposed in this article, the authors did not deal with poetic texts nor did they use pre-formed models.

YI et al. (Yi et al. 2018), addressed two problems in the automatic poetry generation, which are the lack of diversity and the mismatch of the loss assessment, which are caused by neural models based on maximum likelihood estimation. To deal with these problems, they used reinforcement learning and directly modeled features and used them as explicit rewards to guide the gradient update. The model was based on GRU (Gated Recurrent Unit) recurrent neural networks. The authors worked with Chinese poetry and the results surpassed the state of the art. The difference from the current work is that they did not work with English or Portuguese and did not use pre-trained templates.

Van de Cruys (Van de Cruys 2020), proposes a model based on recurrent neural networks of the GRU type for generating poetic text. The model was trained exclusively on standard non-poetic text, being used to generate poems in English and French and, according to the authors, the system produced results compatible with the state of the art for poetry generation. Despite dealing with poetic texts, the difference in relation to current work is the non-use of Poetic Corpus for training and the fact that they do not use pre-trained models.

Materials and methods

One of the areas of research that was driven by deep learning architectures was the study of text generation from the grammatical structures captured by these new methods. According to Piccialli et al. (Piccialli, Marulli, and Chianese 2017), the field of natural language generation consists of creating texts that provide information contained in other types of sources (numerical data, graphics, taxonomies and ontologies or even other texts), with the aim of making these texts indistinguishable from those created by humans. It is crucial that a text compose a whole that has a certain meaning and that can be interpreted by human beings. Automatic text generation makes it possible to increase the production of textual material that can have different purposes, such as production of teaching material, production of technical manuals, assistance in the production of scientific dissemination material, automatic generation of propaganda, etc. Among the text formats that are most present in our daily lives, song lyrics and poetry stand out. As with any other textual type, it is also possible to obtain interesting results in the automatic generation of musical and poetic content using NLP tools, which will be dealt with in this paper. Linguistic structures such as rhymes, intonation, quotations, among others, make the objective of generating such texts somewhat challenging.

The model for generating song lyrics proposed in this paper is based on the fine-tuning of a pre-trained model, a technique called transfer learning. This strategy has the benefit of reusing all the computational and energy costs invested in pre-training. The pre-trained model adopted was the GPT-2 (Radford et al. 2019) provided to the public by OpenAI. GPT-2 is a pre-trained multitasking model that adopts the Transformer architecture (Vaswani et al. 2017) and has about 1.5 billion parameters. Because it is multitasking, the model can be used in several natural language processing tasks, such as named entity recognition, question and answer systems, translation, summarization and natural language generation. The Transformer architecture, proposed by in (Vaswani et al. 2017), is an architectural model based solely on attention mechanisms and which dispenses with the use of recurrent networks and convolutional networks. In addition to adopting the Transformer architecture, the GPT-2 is suitable for use in unsupervised model learning transfer in a zero-shot configuration. In the paper by (Radford et al. 2019), the GPT-2 is employed in several NLP tasks, without performing fine-tuning, and yet it achieves state-of-the-art results in 7 out of 8 of the datasets tested. In the work presented here, the objective is to use the GPT-2 model in order to verify if it is possible to generate musical texts that are grammatically correct and that make sense as a whole. A task for which the GPT-2 was not previously trained.

Some GPT-2 models have been made publicly available to test their effectiveness across a range of tasks. Versions vary according to the number of model parameters. There is a version with 117, 345, 762 and 1542 million parameters. The greater the number of parameters, the greater the potential for performance in the tasks, nonetheless, the greater the model, the greater the demand on the hardware where the model will be executed. In the case of the research described in this paper, models 345M and 762M were used to compare the generative potential of the models.

To perform the fine-tuning, two Corpus were used in two different languages: English and Portuguese. The corpora were generated through Web Scraping on music sites, and the corpus in English ($corpus_1$) was described in (Rodrigues, Oliveira, and Moreira 2019). The corpus in Portuguese ($corpus_2$) was created by extracting the lyrics from the Vagalume website. The $corpus_2$ cleaning process consisted only of tokenization and lemmatization.

Regarding the $corpus_1$, after the lemmatization, 12,355,270 tokens and 175,412 types were counted. One expects, in the case of song lyrics, to have a far greater number of tokens than words, because there are many repetitions like what happens in the choruses. Regarding $corpus_2$, after lemmatization, 3,761,958 tokens and 96,358 types were counted in 24,783 songs.

The *corpora* were used to perform the fine-tuning of the model. Still, care must be taken with the size of the *corpus*, because if it's too small and the fine-tuning is performed for a long time, over fitting may occur. In our case, this problem has not happened, since, for example, $corpus_1$ consists of

more than 12 million tokens.

In the next section, the results of fine-tuning experiments using the corpora are presented. The experiments were performed on a machine with 4 2080 Ti GPUs and 116 GB of RAM for sample generation. It is worth mentioning that the GPT-2 learns to predict the words according to the context, so the performance does not depend directly on the language, but on the quality of the datasets.

Results

Four experiments were performed, summarized in Table 1. The experiments were divided according to the Corpus used and the size of the model. The experiments performed with the larger model (762M) could not be run for a long time due to hardware limitation.

Table 1: Experiments performed. The experiments were separated according to the size of the model (millions of parameters) and the corpus used in the fine-tuning.

experiment number	Model	Corpus	number of samples
1	345M	2	71
2	762M	2	18
3	345M	1	98
4	762M	1	6

To measure the evolution of the generated text, a metric related to semantic cohesion was used. There are many proposals that can be used to measure the semantic cohesion of a text (Newman et al. 2010). In this research, the intrinsic perplexity measure of the generated text was used, based on the probability of the bigrams of the original corpus. The choice of this metric was based on the ease of use and because it is based on the probability of co-occurrence. The lower the result of the perplexity calculation, the greater the semantic cohesion of the text excerpt. The quality of the generated lyrics was not measured, as this is a subjective criterion. The perplexity formula used is expressed in Equation 1.

$$p = \sqrt[N]{\prod_{i-1}^{N} \frac{1}{p(x_{i+1}|x_i)}} \quad (1)$$

where: N is the number of tokens of parsed text; x_i is the i-th token of the text. If the bigram does not occur in the corpus, the unigram probability multiplied by a smoothing factor of 0.4 is used instead.

Due to lack of space, we will only show a single output emitted by the model for the English language. The other outputs will only be commented.

First, we present the performance of the 345M model in the Portuguese corpus. A total of 71 songs were generated in a period of 4h37min of model execution. After some samples were generated, it was already possible to notice an evolution in semantic cohesion, with regard to the existence of words and their placement. The general context of the songs could already be detected (often the model used the theme "love"), but it was noted that some sentences contained spelling errors or that the sentences were incoherent. At the end of the training, interesting results are observed in the text generation. Some examples even have musical markings, such as the chorus and repetition marking. In addition, it was noted that the subjects addressed in the lyrics became increasingly diverse. In the second experiment performed, the same corpus was used, but the model was changed to the 762M. Eighteen songs were generated and, in general, it was noted that the issues addressed in the samples generated by the model were similar to those generated in the previous experiment. For the most part, subjects such as love and religion were more prominent.

Initially, the 345M model was tested for English. 98 samples were generated in a period of 9 hours. Right at the beginning of the sample generation, it was already possible to notice that the syntactic and semantic quality surpassed the previous results. Table 2 shows one output of the 345M.

Unfortunately, due to lack of computational resources, it was not possible to generate many examples of the corpus in English for the 762M model. Therefore, only 6 samples were generated. Despite that, the samples generated were interesting.

Table 2: Output of model 345M tuned with $corpus_1$, first sample.

What an incredible moment of your life
You walked out of the movie theater
And into a rainstorm
On a moonlight night
Who knew you were the one you wanted to be
Oh the sky is blue
The moon is clear
You don't remember me, but what I do know is
You were the star of the world
I knew the way you could be
You were the star of the star
I knew you were the star of the sun

It is noticeable that some phrases lose their meaning, such as "Don't let me let you hide in your heart" or the phrase "My love I know you are in your heart". The phrase could be replaced by "My love I know you are in [my] heart".

Table 3 shows the perplexity result of each generated sample presented in this paper. These samples were selected because we believe they are the most representative. It may be noted that there is a tendency to decrease perplexity while the model is running. Nonetheless, changing the size of the model does not seem to bring significant benefit. The best result with $corpus_2$ was obtained in $sample_5$ with model 762M, obtaining perplexity 54. However, the second best result with $corpus_2$ was achieved in $sample_2$ with model 345M, resulting in perplexity 78. In the case of $corpus_2$, the best result was obtained in $sample_8$ with the 345M model, obtaining perplexity 50. However, other perplexity values of other samples generated with $corpus_2$ were close to this value. In an informal analysis of the generated samples, an improvement in the results is noticed with the processing time and with the size of the model. To have a better view of the influences of model size and time on the results, a larger number of samples would be needed, but the hard-

ware costs for executing these models prevented obtaining a larger number of samples.

Table 3: Perplexity of $corpus_1$ and $corpus_2$ samples. The lower the value, the lower the perplexity and hence the greater the coherence of the text..

Sample	Model	Perplexity	Corpus
1	345M	1485	$corpus_2$
2	345M	78	$corpus_2$
3	345M	166	$corpus_2$
4	345M	138	$corpus_2$
5	762M	54	$corpus_2$
6	762M	102	$corpus_2$
7	345M	112	$corpus_1$
8	345M	50	$corpus_1$
9	345M	82	$corpus_1$
10	762M	74	$corpus_1$

Conclusions

All GPT-2 models analyzed were able to generate syntactically correct musical/poetic texts with semantic coherence, even with some limitations of hardware resources. It was also noted that the tool made some spelling mistakes, which also occur in the original corpus, nonetheless, in general it presented a cohesive result, both syntaxically and semantically. In this sense, the quality of the generated samples becomes subjective in the eyes of those who analyze them, as there are still no metrics capable of automatically analyzing the quality of the texts generated from an artistic perspective. The Perplexity metric was adopted to evaluate the generated texts, but we recognize that this metric is insufficient to measure the quality of a poetic text and serves only as a starting point for the analysis of this type of text. It is worth mentioning that it is not the project's goal to replace humans with regard to musical composition, but to act, as a complement to the final product. Despite this, it is believed that the tool has unique potential for generating text and is generic enough to be used in other text genres and for other purposes.

As part of future work, it is necessary to define more appropriate metrics to handle this type of information. It is also necessary to increase the number of experiments to be conducted, in order to generate enough results for further analysis. Finally, another interesting future work would be the analysis of the incorporation of semantic content into datasets capable of helping models to produce better quality poetic texts.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and FAPEMIG.

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

Deng, L., and Liu, Y. 2018. *Deep learning in natural language processing*. Springer.

Gatt, A., and Krahmer, E. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61:65–170.

Hoy, M. B. 2018. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly* 37(1):81–88.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 100–108.

Park, D., and Ahn, C. W. 2018. Lstm encoder-decoder with adversarial network for text generation from keyword. In *International Conference on Bio-Inspired Computing: Theories and Applications*, 388–396. Springer.

Piccialli, F.; Marulli, F.; and Chianese, A. 2017. A novel approach for automatic text analysis and generation for the cultural heritage domain. *Multimedia Tools and Applications* 76(8):10389–10406.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Rodrigues, M. A. G.; Oliveira, A. d. P.; and Moreira, A. 2019. Development of a song lyric corpus for the english language. In *International Conference on Applications of Natural Language to Information Systems*, 376–383. Springer.

Strubell, E.; Ganesh, A.; and McCallum, A. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13693–13696.

Tas, O., and Kiyani, F. 2007. A survey automatic text summarization. *PressAcademia Procedia* 5(1):205–213.

Van de Cruys, T. 2020. Automatic poetry generation from prosaic text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2471–2480.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Yi, X.; Sun, M.; Li, R.; and Li, W. 2018. Automatic poetry generation with mutual reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3143–3153.