Exploring a Masked Language Model for Creative Text Transformation

Hugo Gonçalo Oliveira
CISUC, Department of Informatics Engineering
University of Coimbra, Portugal
hroliv@dei.uc.pt

Abstract

We explore a masked-language model based on BERT for shifting the meaning of text towards a target theme. Content words in the original text are masked and the model provides a list of filling candidates, out of which one is selected based on its similarity to the theme and constraints regarding morphology and metre. Experimentation is performed with Portuguese song lyrics and trade-offs between grammaticality, semantics, form, and novelty are analysed. We confirm that BERT is a useful tool for creative text transformation.

Introduction

Many creative systems rely on some kind of inspiration set, but only some assumedly aim at transforming a single existing artefact with a pre-defined purpose. When it comes to linguistic creativity, transformation-based approaches have been adopted for the generation of song lyrics (Bay, Bodily, and Ventura, 2017; Gonçalo Oliveira, 2020) and headlines (van Stegeren and Theune, 2019; Mendes and Gonçalo Oliveira, 2020). In this domain, transformation consists of replacing parts of the text, often words, with others that meet desired constraints. Applications include advertising campaigns and others, such as the creation of parodies, either celebrating an event / someone, or making fun of them. And starting with a well-known text or melody makes the result more memorable, also helping to spread the message.

Following recent trends on using neural language models in natural language processing (NLP) and generation tasks, we explore the language-modelling capabilities of BERT (Devlin et al., 2019) for transforming creative text. BERT is a bidirectional language model based on a Transformer neural network, trained for the prediction of masked words, considering both their left and their right context. Given some text, ideally well-known, and a theme, we rely on BERT for providing replacement candidates for some words. Yet, instead of always using the first candidate, in an attempt to control the semantics of the resulting text, we consider the similarity between the resulting sequences and the theme. This is the logic underlying Zorro, a creative system that transforms Portuguese text according to a given theme. It is applied to song lyrics and, besides semantic similarity, two constraints are tested, for better consistency and aesthetics, namely morphology and metre.

As it happens for other creative artefacts, several subjective aspects must be considered during evaluation. Therefore, we had to rely on human opinions for the manual evaluation of grammaticality, semantic coherence and overall appreciation. Whenever possible, we also explored automatic and semi-automatic procedures, namely for assessing: relatedness to the theme, given by the semantic similarity between human-given titles and the original theme; novelty towards the original lyrics, given by the textual overlap between the new and the original lyrics.

Aforementioned aspects are compared for a set of lyrics transformed with different constraints, confirming some initial expectations. Results of human evaluation confirm the subjectivity of the task and, overall, show moderate success. While it is true that the meaning of the lyrics shifts towards the selected themes, trade-offs exist, and increasing the similarity with the theme, by considering more candidates, often results in more grammatical issues. Moreover, overall appreciation is lower than for the original song lyrics. Nevertheless, we believe to have confirmed that BERT is not only a powerful model for NLP, but has also potential for Computational Creativity.

In the remainder of the paper, we overview previous work on the generation of creative text with neural language models, as well as on the transformation of text. We then describe the proposed approach, followed by the experimentation setup, including implementation details. Before concluding, evaluation results are presented and discussed, together with insights and examples.

Related Work

In the last 20 years, poetry generation, including song lyrics, has arguably been one of the most active research topics in Computational Creativity, with a broad range of computational techniques explored (Gonçalo Oliveira, 2017; Lamb, Brown, and Clarke, 2017). Lately, we have seen a growing application of models based on neural networks for this task. For instance, recurrent neural networks with LSTM layers were used for generating rap lyrics (Potash, Romanov, and Rumshisky, 2015). An advantage of such approaches is that they do not require handcrafted rules.

Specifically, the introduction of Transformers (Vaswani et al., 2017) made possible the development of large language models, like GPT (Radford et al., 2019), with powerful gen-

Proceedings of the 12th International Conference on Computational Creativity (ICCC '21)
ISBN: 978-989-54160-3-5

eration capabilities. This model was used for poetry generation, namely Chinese classic poetry (Liao et al., 2019), after pre-training on a corpus of this kind of text. Using only poems with a specific format, GPT could be fine-tuned for that format, characterised by the number of lines, their metre and rhymes. However, the latter features are only captured implicitly. So, unless specific symbols are introduced for length, rhyme and others (Li et al., 2020), not all formal constraints are guaranteed to be met.

Even if previous approaches may result in fluent text that matches desired formal and aesthetic requirements, they give few or no control on the message to transmit. To deal with this limitation, a Transformer-based autoencoder can learn to map input words to related text, while matching the target form (Nikolov et al., 2020). Content words can be stripped from original lyrics, and the model trained to reconstruct the latter from part of the former and their synonyms. Generating lyrics on a new topic becomes a matter of changing the input words to others related to the target topic.

While the previous work generates new text from scratch, others assume the goal of transforming a given text with some creative purpose. Here, another Transformer-based language model, BERT (Devlin et al., 2019), can be useful for suggesting replacements. BERT is less used in generation tasks, but it is very powerful when it comes to predicting words based on their context. A BERT model is often pre-trained in two tasks, Masked Language Model (MLM) and Next Sentence Prediction, but can be fine-tuned for many downstream tasks, like Question Answering or Natural Language Inference. In the MLM task, about 15% of the tokens in the training sequences are masked (i.e., replaced by the token `[MASK]`) and the model is trained to predict each, considering both their left and right context. This makes such a model different from conditional language models.

In fact, the work of Nikolov et al. (2020) has a post-processing step where words ending each line are masked, and BERT used for providing suitable replacements, out of which those maximising rhymes are selected.

Though not using neural networks, several works have tackled the transformation of text. This approach has been applied to the automatic generation of poetry (Toivanen et al., 2012), also including song lyrics, where words have been replaced by: key concepts extracted from daily news, always considering syntactic and metre constraints (Gatti et al., 2017); words selected according to a theme, emotion, meter, or rhyme (Bay, Bodily, and Ventura, 2017); words whose relation to a new theme is analogous to the relation between the original words and the original song title (Gonçalo Oliveira, 2020). In the last two works, distributional representations of words (word embeddings) were used for computing semantic similarity.

Moreover, transformation-based approaches were applied to the adaptation of human-produced news headlines to new contexts, by replacing the nouns in the headlines with nouns from the context (van Stegeren and Theune, 2019), or replacing pairs of content words, based on analogy (Mendes and Gonçalo Oliveira, 2020).

In our work, we combine the earlier approach with the previous constraints for text transformation, i.e., we use BERT for suggesting a large list of replacement candidates, and further constrain the selection, based on a target theme, as well as morphology and metre constraints.

Approach

We exploit the capability of BERT predicting masked words in the transformation of text. For this purpose, we mask the content words in the text we want to transform and use BERT for obtaining replacement candidates. The first candidate will be BERT's best prediction and will result in the most fluent sentence. On the other hand, the result will always be the same and, except for the words to replace, we will not have any control in the result.

But BERT can provide a ranked list of replacement candidates. So, we can explore this list and select the candidate that best suits our goal, in this case, shifting the meaning of the text towards a given theme. Moreover, other constraints can be applied to the candidates, namely on morphology and on their metre. We consider four configurations:

- **Basic:** selects the candidate that maximises the semantic similarity with the theme.
- **Morphology:** besides similarity, for increasing syntactic consistency, only considers candidates that may have the same part-of-speech (PoS) and are inflected (number, gender, tense) as the replaced word.
- **Metre:** besides similarity, to agree with the original metre, only considers candidates with the same number of syllables and stress position as the replaced word. For line-ending words, it gives priority to candidates that rhyme with the original word.
- **Morphology and Metre:** combines both of the previous (hereafter, MM).

Constraints are applied to any list of candidates, and only nouns, verbs and adjectives in the original text are masked. This is done both for keeping some syntactic consistency and resemblance with the original text, but also because similarity between function words (e.g., prepositions and determiners) and any other would be meaningless. We should add that, even though the method tries to replace all content words, it might not find a replacement that matches all the morphology and meter constraints. Only in that case, the original word is kept.

BERT can also be used for computing the semantic similarity. We use it for representing both the theme and the resulting sequences, after each replacement, and then compute their cosine. The higher the cosine, the higher the similarity.

Experimentation Setup

Even though the proposed approach is applicable to any language, we adopted it in the development of Zorro, a creative system for the transformation of Portuguese text, ideally poems. This section describes its implementation and illustrates the process with an example.

Implementation

Implementation of Zorro is based on Python and relies on the `transformers` library from Hugging Face,[1] for loading BERT; and on NLTK (Loper and Bird, 2002), for PoS tagging. As our language model, we use BERTimbau (Souza, Nogueira, and Lotufo, 2020) base, a pre-trained BERT model for Portuguese, with 12 layers, that encodes text sequences in 768-sized vectors. In addition to NLTK, LABEL-Lex (Ranchhod, Mota, and Baptista, 1999), a morphology lexicon for Portuguese, is used for identifying content words and checking the PoS of the candidates. Syllable division and rhyme identification were performed according to a set of rules adapted from Tra-la-Lyrics (Gonçalo Oliveira, Cardoso, and Pereira, 2007).

Example

To illustrate the proposed approach, we consider the lyrics of the song *Efectivamente* (in English, 'actually'), by the Portuguese band GNR. This song starts with the following four lines, roughly translated by those following them:

Adoro o campo as árvores e as flores
Jarros e perpétuos amores
Que fiquem perto da esplanada de um bar
Pássaros estúpidos a esvoaçar

I love the countryside, the trees and the flowers
Arums and perpetual loves
That stay close to the terrace of a bar
Stupid birds flying

In order to transform these lyrics according to a theme, each line is first PoS-tagged. Then, for each content word:

1. It is replaced by the mask token `[MASK]`;
2. A ranked list of filling candidates is obtained from BERT;
3. Similarity between the theme and the sequence resulting from filling the mask with each candidate is computed.[2] If it is the highest so far, it is kept.

For the theme *criatividade computacional* ('computational creativity'), and considering the first 500 filling candidates, with no additional constraint, here is how the previous steps would be instantiated for the first content word (in English, '[I] love'):

1. Masked sequence:
   `[MASK] o campo as árvores e as flores`
2. List of candidates:
   `[Sobre, E, Entre, É, Para, São, ... ]`
   ('about', 'and', 'between', 'is', 'for', 'are')
3. Maximum similarity with *criatividade computacional*:
   `Usando` ('using')

   Selected replacement: `Usando`

Then, for the second content word ('country'):

1. Masked sequence:
   `Usando o [MASK] as árvores e as flores`
2. List of candidates:
   `[jardim, espaço, amor, verde, olhar ... ]`
   ('garden', 'space', 'love', 'green', 'look')
3. Maximum similarity with *criatividade computacional*:
   `Usando o projeto` ('using the project')

   Selected replacement: `projeto` ('project')

Going through every line and content word results in the following four lines, roughly translated to those after them:

```
Usando o projeto as tecnologias e
   as nuvens
simples e inteligente criativa
Que oferece perto da eficiência de
   um recurso
Recursos matemáticos a engenharia
```

Using the project, the technologies and the clouds
simple and intelligent creative
That offers close to the efficiency of a resource
Mathematical resources to engineering

Although we did not use the first candidate, the text consistency is still ok and, more importantly, several words related to the theme are used (e.g., project, technologies, intelligent, creative, resource, engineering). The main issue is that the metre of the new text does not match the original, meaning that it cannot be sung with the same melody.

To fix the latter, we constrain the candidates by considering only those with the same number of syllables and stress position as the original words. Plus, whenever such a word exists, we try to end each line with a word that rhymes with the original word, in an attempt to keep the original rhyming scheme. Once these constraints are added, we get:

```
Usando o design as máquinas e as cores
simples e legítimo conceito
Que gira perto da novidade de um lar
Círculo contínuo a executar
```

Using the design, the machines and the colours
simple and legitimate concept
That revolves near the novelty of a home
Continuous circle to perform

Words in the domain of the theme are still used (e.g., design, machines, colours, concept, novelty), the rhythm now matches the original lyrics, and the third line rhymes with the fourth. To increase the probability of the first and second lines rhyming, more replacement candidates have to be considered. This is the result with the top-5000:

```
Software o design as máquinas e as cores
Ino e inúmeros motores
Que monta perto da teoria de um ar
máquinas contínua a modificar
```

Software the design, the machines and the colours
Ino and countless engines
That assembles close to the theory of an air
continuous machines modifying

Increasing the search space made it possible to have the first two lines rhyme, also increasing the probability of using words in the domain of the theme (e.g., new words like 'software' and 'modifying'). But this was done at the expense of less syntactic consistency and stranger words,

[1] `https://huggingface.co/transformers/`

[2] We use vector representations of the `[CLS]` tokens, returned by the *feature-extraction* pipeline of the `transformers` library.

e.g.: in the first line, the word *software* is used as a verb; in the second, an unknown word, *Ino*, possibly a suffix, is used; in the fourth line, there is a number inconsistency, as *máquinas* ('machines') is in the plural, but its modifier *contínua* ('continuous') is in the singular. While the first candidates should suit the syntactic structure well, matching the PoS and inflection, the lower we get in the rank, more and more words for which this does not happen will appear. This is especially true for poetry, where less common syntactic structures are frequent and some sentences are broken into different lines.

A way to minimise the latter issues would be constraining the candidates to only those that match both the PoS and the inflection of the original words they will replace. This is done with the help of the lexicon, which should have all the possible PoS of each candidate. Adding this constrain to the previous configuration results in the following text:

```
desenho o design as máquinas e as cores
robôs e inúmeros rumores
Que contem perto da teoria de um ar
cálculos cenários a utilizar
```

drawing the design, the machines and the colours
robots and countless rumours
Which count close to the theory of an air
scenarios calculations to be used

Not only it uses words in the domain of the theme, but it follows the original rhythm, has two rhymes, and no syntactic issues, except for an odd last line. Semantic coherence could also be better, especially for the third line.

For illustrating purposes, we show the result of the last configuration in the same lyrics, but now with a different theme, *liga dos campeões* (Champions League):

```
conjunto o clube as líderes e as cores
Clubes e teóricos cantores
Que tenham perto da temporada de um mar
títulos estádios a classificar
```

set the club, the leaders and the colours
Singing clubs and theorists
That have close to the season of a sea
stadium titles to be classified

Evaluation

Experimentation showed interesting results regarding syntax, semantics and metre, all important in a poem. However, these are all subjective aspects, for which we cannot rely exclusively on the opinion of a single person, especially if they were involved in the development of the system.

So, a sample of lyrics was created for the evaluation of the proposed approach. We report on its assessment, starting with a human evaluation, but also applying automatic measures for computing the similarity and novelty.

Data Sample

Results were produced for evaluation purposes, with:

- Lyrics of **five** well-known Portuguese pop-rock songs: *Efectivamente*, by GNR; *Contentores* (containers), by Xutos & Pontapés; *Estou Além* (I'm beyond), by António Variações; *O Anzol* (the fishing hook), by Rádio Macau; *Cavalos de Corrida* (racehorses), by UHF.
- Using **five** different themes, namely *criatividade computacional* (computational creativity), plus four trending topics: *portal das finanças* (the website where Portuguese contributors declare their taxes), *liga dos campeões* (Champions League), *festival da eurovisão* (Eurovision Song Contest), *plano de vacinação* (vaccination plan).
- Following the **four** different strategies: Basic, Morphology, Metre, MM.
- Considering the top-500 (more conservative) and the top-5000 first replacement candidates.

This results in 200 different combinations ($5\times5\times4\times2$) and thus 200 different new lyrics.

Human Evaluation

We resorted to the crowdsourcing platform Amazon Mechanical Turk (AMT), where human workers were instructed to perform a task consisting of: (1) reading a presented poem (i.e., one of the lyrics in the sample); (2) using Likert scales for rating it according to three main aspects (grammaticality, semantic coherence, overall appreciation) and providing a suitable title in a text field. Instructions and questions were written in Portuguese, the same language as the text. The translation of the questions was:

1. At the grammar level, the text has: 1-Many issues, 3-Some issues, 5-No issues.
2. On the transmitted message, the text: 1-Does not make any sense; 3-Has some coherence; 5-Is perfectly coherent.
3. Considering its contents, a good title for the poem is: ...
4. My overall appreciation of the poem was: 1-Hated it; 3-Interesting; 5-Loved it.

For each of the 200 lyrics, questions were answered by two different workers, who were not told that the text had been automatically transformed. Given titles are important for assessing meaningfulness, i.e., the more semantically similar they are to the themes, the better the theme is expressed by the lyrics.

Initially, it was also our intention to gather human opinions on two additional aspects of the new lyrics: rhythm and novelty, both considering the original lyrics. Yet, for this to work well, the workers had to previously know the melody of the original songs and their lyrics. All songs are by Portuguese artists, and even if they are popular enough for being known by a vast Portuguese audience, their popularity is mostly restricted to Portugal and Portuguese people. This meant that we would have to restrict the evaluation to workers located in Portugal. We actually started to do it, but had no answers for some time, which made us open the survey to other Portuguese-speaking countries, namely Brazil. For this reason, it is expected that workers will not be familiar with the original lyrics and thus unapt for as-

sessing the rhythm and novelty of the new lyrics.[3] While an approximation to the latter was obtained automatically (see Novelty below), we ended up not assessing the rhythm. Still, from our observation of the results, we can assume that, for a large majority of cases, the Metre constraint guarantees that all replacements have the same metre as the original words they are replacing. On the other hand, this is not always the case of rhymes. We now look at the results for each human-assessed aspect.

Grammaticality Table 1 summarises the scores regarding the grammaticality of the lyrics transformed by each configuration with a different number of top candidates considered (Top-k). It shows the proportion of scores below, equal and above 3, followed by the median score. The latter suggests that, except perhaps for the Morphology and Metre configuration with 500 candidates (hereafter, MM-500), there are no substantial differences among configurations. Yet, even if differences are low, score distribution is in line with three expectations: grammaticality is harmed by a higher number of candidates and by the metre constraint; with the morphology constraint on and 500 candidates, the proportion of lyrics with scores >3 is higher, even when the metre constraint is added. In fact, the highest proportion (62%) is when both constraints are used. Here, we speculate that a correct metre contributes to better reading, possibly hiding grammar issues. At the same time, when 5,000 candidates are considered, there is a chance of selecting less natural replacements, leading to less fluent text and exposing grammar issues. This, however, does not seem to make a difference for the Basic configuration, as the proportion is the same for the three categories (<3, 3, >3), either considering 500 or 5,000 candidates.

Config	Top-k	<3	3	>3	Med
Basic	500	24%	40%	36%	3.0
Basic	5000	24%	40%	36%	3.0
Morphology	500	16%	38%	46%	3.0
Morphology	5000	28%	42%	30%	3.0
Metre	500	34%	28%	38%	3.0
Metre	5000	46%	30%	24%	3.0
Morph+Metre	500	22%	16%	62%	4.0
Morph+Metre	5000	38%	28%	34%	3.0

Table 1: Grammaticality in lyrics transformed by each configuration, considering a different number of top candidates.

Semantics Table 2 is on the semantic-related aspects. It has the scores given by the workers for semantic coherence, followed by two similarity values which are average cosines between the BERT representations of the worker-given titles, respectively with the theme of the transformed lyrics and with the original title of the song. Since it is virtually impossible for the workers to guess exactly the words of the theme, computing the semantic similarity is a shortcut for quantifying the proximity between the meaning of the title and the theme. BERT encodings are suitable for this, as they can represent whole sequences in a vector of real numbers, based on the meaning of the words in context.

Medians are not very different than for grammaticality, i.e., combining morphology and metre with the top-500 candidates gets slightly higher than the other configurations. Scores also suggest that morphology constraints contribute to semantic coherence more than the metre.

However, we were also assessing to what extent the lyrics are related to the theme, i.e., their semantic similarity with the theme. Since there is not a threshold for which we can consider that two texts are similar enough, we look at the similarity values relatively. We can compare average similarities of the given title with the theme, for different configurations, and also with the original title of the song. As we were trying to shift the meaning of the lyrics, similarity of the given title should be higher for the theme than for the original title.

Even though average similarity values are all very close, the previous expectation is confirmed for every configuration. However, some differences are very low and not significant when the standard deviation is considered. This happens especially for the MM configuration, the one with highest scores for grammaticality and semantic coherence. In addition to what was said about this configuration, we started to question whether the application of both constraints was preventing the replacement of all words, leaving large sequences of the original lyrics, and thus a better grammaticality, semantic coherence and a meaning that is also closer to the original. In the following sections, we show examples of lyrics transformed with this configuration and compute their overlap with the original lyrics, to confirm that this is not exactly the case.

Out of curiosity, we looked at the most given titles to find that a total of nine lyrics got the title "*Agora*" (now). Though not very related to any of our themes, it is the starting word of all lines in the lyrics of *Cavalos de Corrida* and, as an adverb, it is never replaced. A similar situation occurred for the title "*Debaixo do Sol*" (under the sun), given four times, for being the end of the lyrics of *O Anzol*, and kept intact with metre constraints on. Other common titles that are related to the themes include: *Prevenção* (prevention), *Vírus* (virus) and *Gripe* (flu), all related to vaccination plan, respectively given six, five and four times; "*Futebol*" (football), related to Champions League, given five times; and *Finanças* (finances), part of one of the themes, given four times.

There was a single case for which the given title was exactly the same as the theme and it was *Liga dos Campeões*, given to the lyrics of *O Anzol*, transformed with Basic-500. Other titles highly similar (cosine > 0.87) to the themes were: *Os Clubes Campeões* (champion clubs) and *Campeão dos Campeões* (champion of the champions), respectively for the lyrics of *Efectivamente* transformed with Basic-500, and *Cavalos de Corrida* with Basic-5000; *Inovação Computacional* (computational innovation) and *Design de Inteligência* (design of intelligence), respectively for the lyrics of *Contentores* transformed by Basic-5000 and *Cavalos de Corrida* with Morph-5000. This analysis showed that, for some original lyrics and themes (computational creativity,

[3] A link for the original song could be included, but having to listen to the song first would make the task slower. Plus, since we would have no control, many workers could still not listen to it.

Config	Top-k	Coherence <3	Coherence 3	Coherence >3	Med	Sim(theme)	Sim(original)	Diff
Basic	500	28%	42%	30%	3.0	0.716±0.07	0.629±0.08	(+0.088)
Basic	5000	34%	42%	24%	3.0	0.740±0.08	0.615±0.08	(+0.125)
Morphology	500	20%	38%	42%	3.0	0.698±0.06	0.627±0.09	(+0.072)
Morphology	5000	20%	44%	36%	3.0	0.707±0.07	0.634±0.07	(+0.073)
Metre	500	38%	36%	26%	3.0	0.682±0.07	0.650±0.07	(+0.032)
Metre	5000	40%	46%	14%	3.0	0.670±0.08	0.628±0.07	(+0.042)
Morph+Metre	500	20%	30%	50%	3.5	0.648±0.07	0.647±0.07	(+0.001)
Morph+Metre	5000	42%	34%	24%	3.0	0.665±0.08	0.656±0.08	(+0.011)

Table 2: Semantic coherence and average similarity of given titles with the theme and the original title.

champions league) given titles were closer to the theme than for others.

Overall Appreciation Table 3 is on the overall appreciation, which, in the end, is probably what matters the most. In this aspect, we cannot say that the results were very positive, as all configurations have a great proportion of lyrics with scores 1 and 2. When comparing configurations, we see that the metre constraints alone, which favor words based on their length and stress, lead to the worst appreciation. Sometimes, a better rhythm and sound can compensate for some ungrammatical text. Yet, we recall that the workers did not know the original song nor its rhythm, meaning that metre would hardly play a role in their judgement.

Config	Top-k	<3	3	>3	Med
Basic	500	36%	48%	16%	3.0
Basic	5000	48%	40%	12%	3.0
Morphology	500	30%	58%	12%	3.0
Morphology	5000	34%	46%	20%	3.0
Metre	500	52%	38%	10%	2.0
Metre	5000	50%	28%	22%	2.5
Morph+Metre	500	16%	36%	48%	3.0
Morph+Metre	5000	50%	34%	16%	2.5

Table 3: Overall appreciation for each configuration.

MM-500 was again the configuration with more lyrics rated 4 or 5 and less below 3. As discussed earlier, the impossibility of replacing some words could end up favouring this configuration. Still, looking at some lyrics by MM-500, we see that many content words are indeed replaced. In Figure 1, we show the three top-scored (average=4.5) lyrics. For each, the titles given and their similarity to the theme, plus the average scores for grammaticality (Gram) and semantic coherence (Sem) are shown. We note that, in the first two, a single line from the original lyrics is kept intact (*Para outro mundo*). In addition, two other lines are the same in both, despite the different theme. In the third lyrics, two lines are equal to the original (*Porque até aqui eu só*, *Quem não conheci*). Moreover, four and seven lines have a single word replaced, respectively in the first two and in the third. Rough translations for these lyrics are in the Appendix of the paper. Further ahead on the paper, we analyse the novelty of the new lyrics based on their overlap with the original.

A total of 14 lyrics had the lowest overall appreciation (average= 1.5), out of which, one was produced with MM-500 and five with MM-5000, three of them in Figure 2. The remaining were by Basic-500 (1), Morph-500 (1), Metre-500 (4) and Metre-5000 (2). Semantic coherence is lower in all three examples shown. In the first and third, we note a higher presence of odd line constructions and words (e.g., *els*, *log*, *cm*).

Novelty

The higher average similarity between given titles and themes was already a hint that some original words were being replaced, resulting in a meaning shift. Still, we attempted to quantify the novelty, based on the overlap between the original and the transformed lyrics. Some overlap is expected, and even desired for increased familiarity, but the resulting lyrics should not just be a copy of the original.

In order to approximate novelty towards the original, we adopted ROUGE (Lin, 2004), a set of metrics commonly used for assessing automatic summarisation and machine translation, based on the overlap between the generated and the original text. The main difference here is that we are aiming for lower ROUGE scores, meaning that overlap with the original text is also lower. In the scope of Computational Creativity, ROUGE has previously been used for assessing variation in automatically-generated poems (Gonçalo Oliveira et al., 2017).

We consider the overlap between uni (R-1), bi (R-2), and tri (R-3) grams, as well as the longest common subsequence (R-LCS). Alone, the obtained values are probably not so meaningful, but they enable us to compare different configurations. For each metric considered, Table 4 shows the average values for the simplest configuration (Basic-500) and the difference for all the others. First of all, values confirm that, for any configuration, several changes are made to the original lyrics. Otherwise, they would be 1 or close. They further confirm some initial expectations, even if by low margins. For instance, novelty is higher for the Basic strategy, which does not constrain the replacements, meaning that all content words end up being replaced. It is also higher with the morphology than with the metre constraints alone, showing that it is more difficult to find words with a certain number of syllables and stress than with a target PoS and inflection. In fact, when applying only the morphology constraints, ROUGE values are almost the same as for the Basic configuration. A final confirmation is that novelty is higher when considering the top-5000 candidates, for which chances of finding suitable replacements increase.

Original: *Contentores*
Theme: *festival da eurovisão*
Titles: *a despedida* (the farewell, 0.72), *adeus romantico* (romantic goodbye, 0.69)
Gram: 4, **Sem:** 4

A crise grega e chegada nos corredores
Adeus aos meus sapatos que me vou
Para outro mundo
Num astro surgido num projecto judicial
Não faze nada mal isto onde vou
P'lo teatro mundo
Mudaram todas as dores
fazem sozinho os leitores
E numa crise impossível
vendo a Lisboa Central
Não faze nada mal

Original: *Contentores*
Theme: *portal das finanças*
Titles: *eu me vou* (i'm leaving, 0.68), *impostos* (taxes, 0.70)
Gram: 3.5, **Sem:** 3

A folha falsa e travada nos corredores
Adeus aos meus impostos que me vou
Para outro mundo
Num banco externo num governo comercial
Não guarda nada mal isto onde vou
P'lo inteiro mundo
Mudaram todas as dores
usam sozinho os valores
E numa crise impossível
vendo a revista geral
Não entra nada mal

Original: *Estou Além*
Theme: *criatividade computacional*
Titles: *ao me conhecer* (meeting me, 0.68), *arte de gerar* (art of generating, 0.68)
Gram: 4.5, **Sem:** 4.5

Não preciso transformar
Este sistema de utilidade
A Arte de gerar
P'ra não tornar tarde
Não sou do que tem que eu uso
Será desta dimensão
Mas porque faz que eu abuso
Quem faz dar-me a mão
Vou continuar a explorar
A quem eu me porto dar
Porque até aqui eu só
Tendo quem quem eu nunca vi
Porque eu só acho quem
Quem não conheci

Figure 1: Lyrics with high overall appreciation, all with MM-500 configuration.

Original: *Cavalos de Corrida*
Theme: *festival da eurovisão*
Config: MM-500
Titles: *hipocresia* (hypocrisy, 0.59), *evento de comida* (food event, 0.59)
Gram: 3.5, **Sem:** 2.5

Agora há que a Europa terminou, e os media se seguem num esforço
Agora há que todos eles jogaram, a mitologia em jogo
Agora há que eles fazem os media, destruindo todas as leis
Agora há que els seguem ao buraco e fazem-no por qualquer preço
Agora, agora, agora, agora, tu dá um evento de comida

Original: *Contentores*
Theme: *plano de vacinação*
Config: MM-5000
Titles: *vírus* (virus, 070), *o vírus* (the virus, 0.70)
Gram: 3.5, **Sem:** 2.5

A gripe paga e consulta nos portadores
Adeus aos meus arquivos que me vou
Para outro fundo
Num vírus tratado num esquema oriental
Não trata nada mal isto onde vou
P'lo completo fundo
indicam todas as dores
devem sozinho os menores
E numa lista compatível
Plano a Saúde natal
Não trata nada mal

Original: *O Anzol*
Theme: *criatividade computacional*
Config: MM-5000
Titles: *medir* (measuring, 0.80), *medição* (measurement, 0.80)
Gram: 2, **Sem:** 2

Ai, eu já medi
pensar mover o top em log de azul
Pra cm estrutural
Mas só depois medi
faz que plural já ele tens, monta alguém
Que monta ideia mental
Eu não sou se hei de medir
Ou gerar o padrão
Já não cm nada de novo aqui
Debaixo do sol

Figure 2: Lyrics with low overall appreciation.

Config	Top-k	R-1	R-2	R-3	R-LCS
Basic	500	0.535±0.02	0.250±0.01	0.113±0.01	0.525±0.01
Basic	5000	-0.002±0.01	0.000±0.01	0.000±0.01	-0.004±0.01
Morphology	500	+0.006±0.01	+0.005±0.01	+0.001±0.00	+0.009±0.01
Morphology	5000	+0.005±0.00	+0.001±0.01	0.000±0.00	+0.004±0.01
Metrics	500	+0.053±0.02	+0.081±0.05	+0.080±0.04	+0.058±0.03
Metrics	5000	+0.042±0.02	+0.055±0.04	+0.053±0.04	+0.045±0.03
Morph+Metrics	500	+0.080±0.02	+0.118±0.05	+0.110±0.04	+0.087±0.03
Morph+Metrics	5000	+0.056±0.02	+0.077±0.05	+0.073±0.05	+0.061±0.03

Table 4: ROUGE scores of the Basic-500 configuration and differences of the others when compared to it.

Scoring the original lyrics

Since workers were not from Portugal, they would not recognise the original lyrics. Therefore, we thought that it would be interesting to mix the original lyrics among the automatically transformed. This enabled us to gather opinions for the former, on the same aspects we were considering for the latter, also giving us another term of comparison.

Table 5 shows the scores obtained by the original lyrics of the songs used in this experimentation. They are based on four human opinions for each of the five original lyrics. The

semantic similarity of the given title was computed against the original title of the song.

Results are interesting to show that, according to human opinions, original songs still have issues. This happens for grammaticality, less expected, but also for semantics and, especially, for the overall appreciation, which is also the most subjective. Our interpretation is that song lyrics are not always easy to interpret, also due to the presence of figurative language. Together with the position of the line breaks, this might also have a minor impact on grammaticality. Not to mention that all lyrics were in European Portuguese, but scored by Brazilian workers.

Regarding the average similarity between given and original titles, it is only higher than similarity between given titles for the MM configuration and their themes. This suggests that, even though replacements are made in this configuration, relatedness between the target themes and the transmitted message could be stronger.

Conclusion

We have explored the application of a popular masked language model, BERT, in the transformation of creative text, namely Portuguese song lyrics. We took advantage of the mask-filling feature of this model and further constrained the filling predictions, according to their similarity to a given theme, PoS and metre.

There are trade-offs between increasing the similarity with a theme and keeping the text grammatical, so, finding the right balance can be tricky. Our evaluation was limited to a language, a set of original lyrics and constraints, in any case leading to an overall appreciation of the transformed lyrics below the one for the original. Towards more successful results, testing with different parameters is required, and, depending on the final purpose, human curation might still be necessary in the end. A possibly erroneous conclusion was that matching the metre does not contribute to higher appreciation, but this might be due to our impediments on properly assessing singability, due to lack of workers that knew the original songs.

We cannot say that we are completely satisfied, but interesting results were obtained. Enough to believe that BERT should be seen as a useful tool in the transformation of creative text. For stronger conclusions, further experimentation is needed, e.g., in other languages, considering other texts, different numbers of candidates and possibly other constraints. This also applies to alternative ways of using BERT, e.g., on computing the similarity of each candidate to the theme (e.g., just the token, right context followed by the token, full resulting sequence), or of getting sequence representations (e.g., from different layers, considering the contextual token embeddings), possibly after fine-tuning the model to some task where creative text is used.

Acknowledgments

This work was supported by national funds through the FCT – Foundation for Science and Technology, I.P., within the scope of the project CISUC – UID/CEC/00326/2020 and by European Social Fund, through the Regional Operational Program Centro 2020.

References

Bay, B.; Bodily, P.; and Ventura, D. 2017. Text transformation via constraints and word embedding. In *Proceedings 8th International Conference on Computational Creativity*, ICCC 2017, 49–56.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. ACL.

Gatti, L.; Özbal, G.; Stock, O.; and Strapparava, C. 2017. Automatic generation of lyrics parodies. In *Proceedings 25th ACM International Conference on Multimedia*, 485–491.

Gonçalo Oliveira, H.; Hervás, R.; Díaz, A.; and Gervás, P. 2017. Multilingual extension and evaluation of a poetry generator. *Natural Language Engineering* 23(6):929–967.

Gonçalo Oliveira, H. R.; Cardoso, F. A.; and Pereira, F. C. 2007. Tra-la-Lyrics: an approach to generate text based on rhythm. In *Proceedings 4th International Joint Workshop on Computational Creativity*, 47–55. IJWCC 2007.

Gonçalo Oliveira, H. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings 10th International Conference on Natural Language Generation*, INLG 2017, 11–20. ACL.

Gonçalo Oliveira, H. 2020. WeirdAnalogyMatic: Experimenting with analogy for lyrics transformation. In *Proceedings 11th International Conference on Computational Creativity*, ICCC 2020, 228–235. ACC.

Lamb, C.; Brown, D. G.; and Clarke, C. L. 2017. A taxonomy of generative poetry techniques. *Journal of Mathematics and the Arts* 11(3):159–179.

Li, P.; Zhang, H.; Liu, X.; and Shi, S. 2020. Rigid formats controlled text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 742–751. Online: Association for Computational Linguistics.

Liao, Y.; Wang, Y.; Liu, Q.; and Jiang, X. 2019. GPT-based generation for classical Chinese poetry. *arXiv preprint arXiv:1907.00151*.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: ACL.

Loper, E., and Bird, S. 2002. NLTK: The natural language toolkit. In *Proceedings ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 63–70.

Grammar				Semantic Coherence					Appreciation			
<3	**3**	**>3**	**Med**	**<3**	**3**	**>3**	**Med**	**Sim(title)**	**<3**	**3**	**>3**	**Med**
10%	5%	85%	4.0	10%	10%	80%	4.0	0.670±0.08	5%	30%	65%	4.0

Table 5: Scores for the original lyrics.

Mendes, R., and Gonçalo Oliveira, H. 2020. TECo: Exploring word embeddings for text adaptation to a given context. In *Proceedings 11th International Conference on Computational Creativity*, ICCC 2020, 185–188. ACC.

Nikolov, N. I.; Malmi, E.; Northcutt, C.; and Parisi, L. 2020. Rapformer: Conditional rap lyrics generation with denoising autoencoders. In *Proceedings 13th International Conference on Natural Language Generation*, INLG 2020, 360–373. ACL.

Potash, P.; Romanov, A.; and Rumshisky, A. 2015. GhostWriter: Using an LSTM for automatic rap lyric generation. In *Proceedings 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2015, 1919–1924. ACL.

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

Ranchhod, E.; Mota, C.; and Baptista, J. 1999. A computational lexicon of Portuguese for automatic text parsing. In *Proceedings of SIGLEX99 Workshop: Standardizing Lexical Resources*. ACL.

Souza, F.; Nogueira, R.; and Lotufo, R. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Proceedings Brazilian Conference on Intelligent Systems (BRACIS 2020)*, volume 12319 of *LNCS*, 403–417. Springer.

Toivanen, J. M.; Toivonen, H.; Valitutti, A.; and Gross, O. 2012. Corpus-based generation of content and form in poetry. In *Proceedings of the 3rd International Conference on Computational Creativity*, ICCC 2012, 211–215.

van Stegeren, J., and Theune, M. 2019. Churnalist: Fictional headline generation for context-appropriate flavor text. In *Proceedings 10th International Conference on Computational Creativity*, ICCC 2019, 65–73. UNC Charlotte, North Carolina, USA: ACC.

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.

Appendix

For better understanding of their contents, Figures 3 and 4 show rough English translations of the Portuguese lyrics in Figures 1 and 2, respectively.

Original: *Contentores*
Theme: *festival da eurovisão*
Titles: *a despedida* (the farewell, 0.72), *adeus romantico* (romantic goodbye, 0.69)
Gram: 4, **Sem:** 4

The Greek crisis and arrival in the corridors
Farewell to my shoes cause I am going
To another world
In a star appearing in a judicial project
This where I go does not make anything bad
To the theatre world
All the colours have changed
make alone the readers
And in an impossible crisis
Seeing Central Lisbon
Do nothing wrong

Original: *Contentores*
Theme: *portal das finanças*
Titles: *eu me vou* (i'm leaving, 0.68), *impostos* (taxes, 0.70)
Gram: 3.5, **Sem:** 3

The fake and locked sheet in the corridors
Farewell to my taxes, I'm going
To another world
In a foreign bank in a commercial government
It doesn't save that bad this where I'm going
All over the world
All the pains have changed
They use the values alone
And in an impossible crisis
I sell the general magazine
Does not enter that bad

Original: *Estou Além*
Theme: *criatividade computacional*
Titles: *ao me conhecer* (meeting me, 0.68), *arte de gerar* (art of generating, 0.68)
Gram: 4.5, **Sem:** 4.5

No need to transform
This utility system
The art of generating
So that it doesn't become late
I am not of what there is I use
Is it from of this dimension
But why does it make me abuse
Who makes me give my hand
I will continue to explore
To whom I give myself
Because until here I only
I have who, who I've never seen
Because I only find who
Who I haven't met

Figure 3: Lyrics with high overall appreciation, all with MM-500 configuration.

Original: *Cavalos de Corrida*
Theme: *festival da eurovisão*
Config: MM-500
Titles: *hipocresia* (hypocrisy, 0.59), *evento de comida* (food event, 0.59)
Gram: 3.5, **Sem:** 2.5

Now that Europe is over, and the media follow in an effort
Now they've all played, the mythology at stake
Now they make the media, destroying all the laws
Now they follow in the hole and they do it at any price
Now, now, now, now, you give a food event

Original: *Contentores*
Theme: *plano de vacinação*
Config: MM-5000
Titles: *vírus* (virus, 070), *o vírus* (the virus, 0.70)
Gram: 3.5, **Sem:** 2.5

Influenza pay and consultation in the bearers
Farewell to my archives, I'm leaving
To another background
In a virus treated in an oriental scheme
It does not treat badly this where I go
To the complete fund
indicate all the pains
should alone the underaged
And in a compatible list
Home Health Plan
It does not treat that badly

Original: *O Anzol*
Theme: *criatividade computacional*
Config: MM-5000
Titles: *medir* (measuring, 0.80), *medição* (measurement, 0.80)
Gram: 2, **Sem:** 2

Oh, I've already measured
think about moving the top in log from blue
For structural cm
But only then I measured
make plural you already have, it assembles someone
Who assembles a mental idea
I am not whether to measure
Or generate the pattern
There's nothing new here anymore
Under the sun

Figure 4: Lyrics with low overall appreciation.