Automatic Generation of Paragraph Templates Using a Textual Energy Approach: Applications in Literary Text Production

Luis-Gil Moreno Jiménez, Juan-Manuel Torres-Moreno
Laboratoire Informatique d'Avignon, Université d'Avignon
339 Chemin des Meinajariès, CEDEX 9, 84911 Avignon, France
{luis-gil.moreno-jimenez,juan-manuel.torres}@univ-avignon.fr

Abstract

In this paper, we present the results of preliminary experiments using the Textual Energy measure to be used in Automatic Text Generation tasks (ATG). Textual Energy calculates the similarity among the sentences of a document, using intuitive ideas coming from Mechanical statistical like the associative memories and the energy of a system. Using this approach we intend to generate a set of selected sentences having a semantic and structural coherence. In our experiment, the number of selected sentences was manually determined. In particular, the experiments were performed using sets of 4 sentences. Then, the selected sentences can be employed for paragraph generation using Canned Text-like techniques. We have performed an important number of experiments, and we found interesting results that we present in this paper. These results allow us to conclude that it is possible to generate a set of sentences, as paragraphs-like, through methods, avoiding as much as possible undesirable phenomena, such hallucination, which have been recently found in ATG models based on Deep learning Neural Networks.

1 Introduction

Automatic Text Generation (ATG) is an area that has been extensively studied by scientists in the Natural Language Processing domain (NLP) (Kumar et al. 2022; Sridhara et al. 2010; Szymanski and Ciota 2002). Recently, models based on Neural Networks (NN) have achieved important results in this field (Qu et al. 2020; Du, Xing, and Pei 2021). However, they have also shown unexpected behaviors, for example generating grammatically correct text but out-of-context or completely isolated from what the user expected as a response. This phenomenon, sometimes known as hallucination (Ji et al. 2022), is difficult to correct given the structural complexity that characterizes NN-based models.

In this paper, we conduct some preliminary experiments with Textual Energy for the constitution of paragraphs by selecting sentences from an analyzed document. The objective is that the constituted paragraph may be used as a template for the production of a new paragraph, following the principle of homosyntax. Homosyntax consists in producing a new text respecting the syntactic structure of another text. Several works are based on homosyntax for ATG tasks using techniques such as Canned Text (Oliveira and Cardoso 2015; Molins and Lapalme 2015). So far, the present experiments are in an exploratory phase and any evaluation has been carried out. Nevertheless, looking our results we consider that it is possible to conceive our proposal as a system for text generation without depending on NN techniques, facilitating the manipulation of the model architecture and providing an important control over phenomena such as hallucination.

The paper is structured as follows: In the Section 2 we present some works related to Automatic Text Generation tasks based on NN and templates, which have served as inspiration for our experiments. Then, in Section 3 we explain the concept of textual energy and what it implies. The experiments and some examples are presented in the section 4. Finally, we discuss our conclusions and future works in the section 5.

2 Related works

In the last few decades, researchers in NLP have dedicated themselves to proposing better models for ATG tasks. Actually, the most of the models are based on the use of Artificial Neural Networks (ANN), such as transformers, which in recent years have demonstrated a great ability to perform tasks in different areas (Radford et al. 2019). For ATG, we have models such as those proposed in (Ormazabal et al. 2022), where the researchers have proposed a model capable of generating rhymes using a language model based on transformers. The experiments were conducted in Spanish and the evaluation was performed manually by mixing the artificially generated rhymes with rhymes written by humans obtaining that 30% of the rhymes meet the evaluated criteria, the low performance of the model is due to the fact that most of the rhymes have the same word as rhyme.

In (Ta et al. 2023), the authors proposed a model based on an encoder-decoder neural network for the generation of short descriptions of documents. For their experiments and evaluations, the authors performed an exhaustive experimentation using the wikipedia database and employing various metrics such as ROUGE, BLEU and pre-trained models such as BERT. Some proposals based on GPT models have shown interesting results, for example, in (Bena and Kalita 2020) the authors used a pre-trained GPT-2 model for poetry generation, the intention was to generate poems with a sentimental charge so the authors used a corpus of dream stories to train their model. The results showed that 87.5% and 85%

percent of the generated poems contained emotional charges associated with sadness and joy emotions. In (Van de Cruys 2020), Van de Cruys proposed an algorithm for text generation using a recurrent neural network (RNN) and a set of keywords. The keywords are used by the RNN to determine the context of the text that is produced.

The works previously described report interesting results, all have in common the use of methods based on NN. These methods have a very complex structure and sometimes can limit the manipulation of its own structure in order to improve the results or fix superficial errors as those exposed in (Ormazabal et al. 2022). For that reason, we searched others works using NN methods for only very specific tasks e.g. semantic analysis for vocabulary selection. In (Moreno-Jiménez, Torres-Moreno, and Wedemann 2020) the authors propose a model for the generation of literary sentences, this model is based on the canned text method, generating a semi-lexical template that are subsequently altered with a vocabulary associated with a defined context, the experiments have been conducted in French, Spanish and Portuguese, performing manual evaluations under criteria such as: literary perception, coherence and grammaticality, with results above 70% in all cases.

Another work based on templates is presented in (Agirrezabal et al. 2013). The authors have generated a language model analyzing POS tags[1] from several corpora in order to produce templates or structures whose elements are subsequently replaced by a selected vocabulary. This method has obtained a precision of 75% after a manual evaluation.

We have found in ATG models based on templates, an opportunity to reduce the use of NN-based methods. Our proposal focuses on the stage prior to text generation. Through Textual Energy we intend to compose a new paragraph, that can be then used as a template, by selecting the sentences with the closest semantic content.

3 Textual Energy

A magnetic system may studied as a set of N small magnets called *spins* (Hopfield 1982; Hertz, Krogh, and Palmer 1991). These spins can turn according to several directions. The simplest case is represented by the Ising model which considers only two possible directions: up (↑, +1) or down (↓, -1 or 0). The Ising model is used in several systems which can be described by binary variables (Fernández, SanJuan, and Torres-Moreno 2007). A system of N binary units has $\nu = 1, ..., 2^N$ possible configurations (patterns). In the Hopfield model the spins correspond to the neurons, interacting with the Hebb learning rule[2]:

$$J^{i,j} = \sum_{\mu=1}^{P} s_\mu^i s_\mu^j \tag{1}$$

s^i et s^j are the states of neurons i and j.

[1]POS tags provide all the grammatical information of a parsed text. Information about the POS tag structure can be found on `https://freeling-user-manual.readthedocs.io/en/latest/tagsets/`

[2]The connections are proportionals to the correlation between neurons' states (Hertz, Krogh, and Palmer 1991).

Documents are pre-processing with classical algorithms of filtering[3], normalization and lemmatisation. A representation in bag of words produces a matrix $S_{[P\times N]}$ of frequencies/absences consisting of $\mu = 1, \cdots, P$ sentences (lines); $\vec{\sigma}_\mu = \{s_\mu^1, \cdots, s_\mu^i, \cdots, s_\mu^N\}$ and a vocabulary of $i = 1, \cdots, N$ terms (columns).

$$S = \begin{pmatrix} s_1^1 & s_1^2 & \cdots & s_1^N \\ s_2^1 & s_2^2 & \cdots & s_2^N \\ \vdots & \vdots & \ddots & \vdots \\ s_P^1 & s_P^2 & \cdots & s_P^N \end{pmatrix} \quad s_\mu^i = \begin{cases} \text{TF}^i & \text{if term } i \text{ exists} \\ 0 & \text{elsewhere} \end{cases} \tag{2}$$

Because the presence of the word i represents a spin $s^i \uparrow$ with a magnitude given by its frequency TF^i (its absence by ↓ respectively), a sentence $\vec{\sigma}_\mu$ is therefore a chain of N spins. We differ from (Hopfield 1982) on two points: S is a whole matrix (its elements take absolute frequential values) and we use the elements $J^{i,i}$ because this autocorrelation makes possible to establish the interaction of the word i among the P sentences, which is important in NLP. Hebb's rule may be applied to calculate the interactions between N terms:

$$J = S^T \times S \tag{3}$$

Each element $J^{i,j} \in J_{[N\times N]}$ is equivalent to the calculation of (1). This is know as Textual Energy of interaction. This concept was introduced by (Fernández, SanJuan, and Torres-Moreno 2007), and it can be expressed:

$$E = -\frac{1}{2} S \times J \times S^T \,;\, E_{\mu,\nu} \in E_{[P\times P]} \tag{4}$$

$E_{\mu,\nu}$ represents the energy between patterns μ and ν.

4 Experiments and Examples

In this section we describe the experiments conducted using Textual Energy (TE) and show some examples of the results obtained. For training we have selected a set of $P = 132$ random sentences in Spanish from the LiSSS corpus (Moreno-Jiménez and Torres-Moreno 2020). The value of P was determined empirically in order to control the consumption of computational resources, increasing the value of P would exponentially increase the computational cost to calculate the S matrix.

Subsequently, four experiments were performed, as a result, four S matrices were generated. For each matrix, one different version of the P sentences was used. The versions of the P sentences are described below, and exemplified in the Table 1.

1. The first version contains the entire vocabulary of sentences.
2. The second version contains only lexical words (verbs, adjectives and nouns) in the lemmatized format.
3. For the third version, the lexical words were replaced by their POS tags, the non-lexical words were eliminated.
4. For the last version, the lexical words were concatenated with theirs POS tags as follows : `lexical_word+POS_tag`.

ver	**Sentence**
1	*"La soledad de este paraíso terrenal es un precioso bálsamo para mi alma."*
2	soledad paraíso terrenal ser precioso bálsamo alma.
3	NCFS NCMS AQCS VSIP3S AQMS NCMS NCFS
4	soledad+NCFS paraíso+NCMS terrenal+AQCS ser+VSIP3S precioso+AQMS bálsamo+NCMS alma+NCFS

Table 1: Sample of one sentence in the different versions. Sentence taken from the work "*The Sorrows of Young Werther*" of Johann Wolfgang von Goethe.

Our motivation for extending our experiments by means of these four experiments is explained by the following. We argue that by analyzing lexical words we retain the most important semantic content of the text, and eliminate all the noise that functional words might cause. On the other hand, we consider that by analyzing only POS tags we are able to project the morphosyntactic similarity of the studied text. Our fourth experiment was a measure to merge the previous hypotheses, projecting through the TE the semantic content and morphology structure similarities of the analyzed sentences.

The confusion matrix projected in Figure 1 shows the distribution of TE values in the matrix S_1, calculated from version 1 of the P sentences. When the dots tend to yellow, it means that the TE values are higher. Each row on the X-axis represents a sentence within the set of P sentences, $f_p, p = 1, 2, 3, ..., P$, and on the Y-axis we have the TE values of each P sentence corresponding to the f_p. The diagonal represents the TE of each sentence f_p with respect to itself.

Although in Figure 1 the diagonal is clearly visible, we can appreciate that around it there are many sentences with similar values of TE. Considering that our objective is, given a sentence, select a set of sentences with similar TE values. Evidently, the task of selection becomes complicated when most of the sentences have similar values of TE.

One solution to this problem was through the experiments performed with the others versions of the P sentences showed in Table 1. In Figure 2 we illustrate the TE values corresponding to the matrix S_2, computed with lexical words. We can observe how only some sentences are easily distinguished thanks to their TE values, facilitating the selection of the f_n phrases. The results obtained from sentences with POS tags are less interesting since they present a very similar TE values, we consider that this is due to the poor variety that exists in the POS tags structures. In order to save more space in the manuscript, we have decided not to illustrate this matrix.

The most logic solution would be to continue the experiments using the S_2 matrix, however, we decided to perform

[3]Filtering of numbers and stop-words.

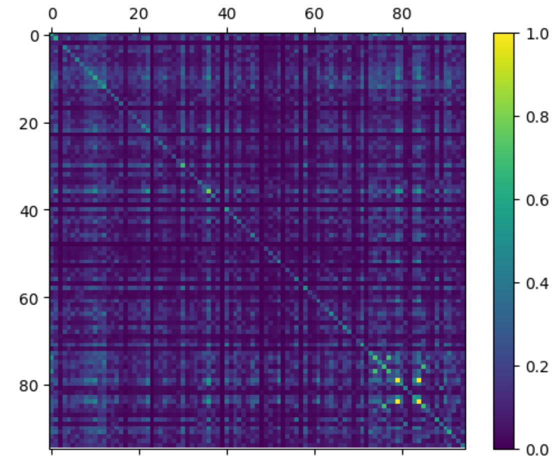

Figure 1: Textual Energy from S_1 matrix, calculated with all vocabulary.

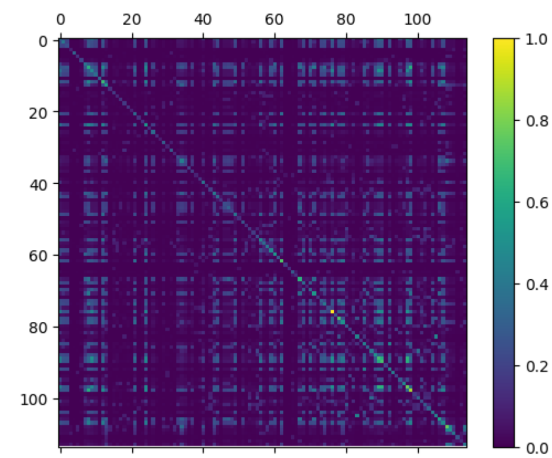

Figure 2: Textual Energy from S_2 matrix, calculated with lexical words.

another experiment, using the version 4, a combination between the lexical words and the POS tags of the P sentences, as a result we have obtained the matrix S_4 illustrated in Figure 3. In this figure we can see how the diagonal is easily observable, observing also only the most important sentences with the highest TE values. We consider that this scenario is the ideal for our objectives, since in addition to facilitating the selection process, we integrate in a single matrix the semantic and grammatical information of the analyzed sentences.

To compare both matrices representing the best scenarios in a greater degree of detail, we choose a random sentence and observe its TE distribution with respect to the rest of the sentences, the results are shown in Figures 4 and 5 for the matrices S_4 and S_2 respectively.

We can observe that the sentence in the S_4 matrix presents a TE distribution that allows to easily differentiate the best cases, unlike the S_2 matrix, where the most of the sentences

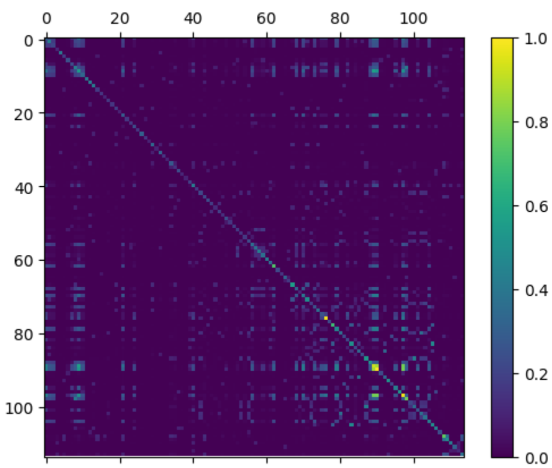

Figure 3: Textual Energy from S_4 matrix, calculated with lexical words and POS tags.

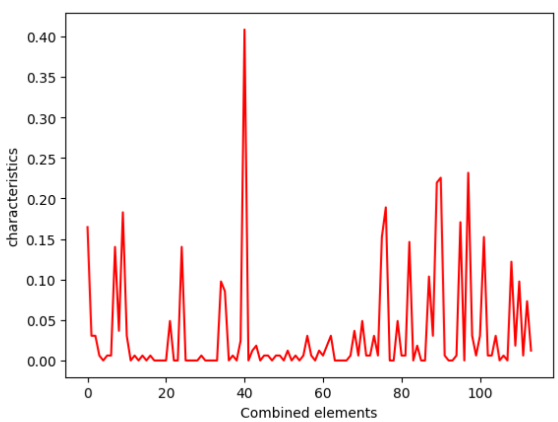

Figure 4: Distribution of Textual Energy of a sentence using S_4 matrix

have very similar TE values.

Below we list a segmented paragraph produced with 3 sentences in Spanish selected automatically using matrix S_4. The paragraph production process starts from the random selection of a sentence fo, from fo we select the sentence with the highest TE according to the S_4 matrix values, f_1; this sentence will be the first one in the paragraph, subsequently, f_1 becomes fo and we repeat the process to determine f_2. The process is repeated n times, where n is the number of sentences desired in the paragraph.

- **¿Por qué será que lo que colma de felicidad al hombre es al mismo tiempo la fuente de sus desgracias?** – *Why is it that what fills man with happiness is at the same time the source of his misfortunes?*
- **¿Es que tenía que ser así, que lo que hace la felicidad del hombre sea también la fuente de su desdicha?** – *Was it meant to be so, that what makes man's happiness is also the source of his unhappiness?*

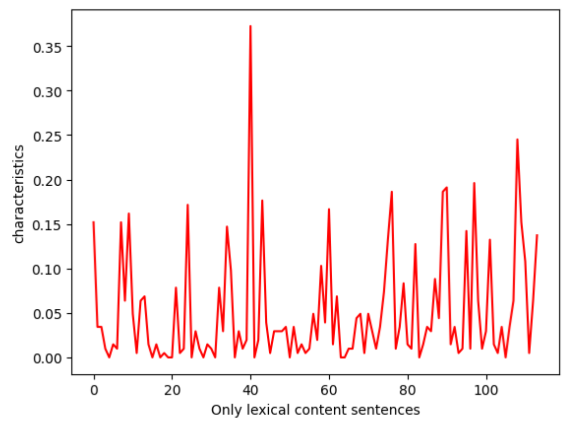

Figure 5: Distribution of Textual Energy of a sentence using S_4 matrix

- **¿Es preciso que lo que constituye la felicidad del hombre sea también la fuente de su miseria?** – *Must what constitutes man's happiness also be the source of his misery?*

Our hypothesis is that sentences with a similar semantic content and vocabulary can be useful for automatic text generation. This hypothesis arises from observing the results obtained by the researchers in (Moreno-Jiménez, Torres-Moreno, and Wedemann 2020). In that work, the authors experimented with templates to generate literary sentences. For the generation of the sentences, a vocabulary is automatically selected based on the semantics of a context and the lexical words of the analyzed template. We consider that our proposal can be useful in this type of models to automatically determine the set of sentences with the necessary characteristics that allow them to be used as templates for the generation of paragraphs.

5 Conclusions

In this paper we carried out several experiments with Textual Energy applying four types of processing to a set of sentences extracted from the LiSSS corpus. The objective was to select the best set of sentences to serve as input for the generation of templates suitable for Automatic Text Generation tasks. In the examples shown in Section 4 we can see how the sentences selected from matrix S_4 have sentences where the vocabulary and the idea conveyed are similar. We found that by calculating the Textual Energy of sentences using lexical words and POS tags, we were able to find sentences with the appropriate characteristics that could be useful for automatic paragraph generation. As future work we consider extending the size of the corpus of sentences used in our experiments. We also plan to experiment with sentences in languages other than Spanish, such as French or Portuguese. Finally, in order to validate our hypothesis, some experiments for automatic paragraph generation are also planned. **Acknowledgments,** this work was partially funded by the Laboratoire Informatique d'Avignon and Agorantic program (France).

References

[Agirrezabal et al. 2013] Agirrezabal, M.; Arrieta, B.; Astigarraga, A.; and Hulden, M. 2013. Pos-tag based poetry generation with WordNet. In *14th European Workshop on Natural Language Generation (ENLG)*, 162–166. Sofia, Bulgaria: Association for Computational Linguistics.

[Bena and Kalita 2020] Bena, B., and Kalita, J. 2020. Introducing aspects of creativity in automatic poetry generation. *arXiv preprint arXiv:2002.02511*.

[Du, Xing, and Pei 2021] Du, H.; Xing, W.; and Pei, B. 2021. Automatic text generation using deep learning: providing large-scale support for online learning communities. *Interactive Learning Environments* 0(0):1–16.

[Fernández, SanJuan, and Torres-Moreno 2007] Fernández, S.; SanJuan, E.; and Torres-Moreno, J. M. 2007. Textual energy of associative memories: Performant applications of enertex algorithm in text summarization and topic segmentation. In Gelbukh, A., and Kuri Morales, Á. F., eds., *MICAI 2007: Advances in Artificial Intelligence*, 861–871. Berlin, Heidelberg: Springer Berlin Heidelberg.

[Hertz, Krogh, and Palmer 1991] Hertz, J.; Krogh, A.; and Palmer, G. 1991. *Introduction to the theorie of Neural Computation*. Redwood City, CA: Addison Wesley.

[Hopfield 1982] Hopfield, J. 1982. Neural networks and physical systems with emergent collective computational abilities. *National Academy of Sciences* 9:2554–2558.

[Ji et al. 2022] Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; and Fung, P. 2022. Survey of hallucination in natural language generation. *ACM Comput. Surv.*

[Kumar et al. 2022] Kumar, L. A.; Renuka, D. K.; Rose, S. L.; Wartana, I. M.; et al. 2022. Deep learning based assistive technology on audio visual speech recognition for hearing impaired. *International Journal of Cognitive Computing in Engineering* 3:24–30.

[Molins and Lapalme 2015] Molins, P., and Lapalme, G. 2015. JSrealB: A bilingual text realizer for web programming. In *ENLG '15*, 109–111. Brighton, UK: ACL.

[Moreno-Jiménez, Torres-Moreno, and Wedemann 2020] Moreno-Jiménez, L.-G.; Torres-Moreno, J.-M.; and Wedemann, R. S. 2020. Literary natural language generation with psychological traits. In *NLPIS '20, LNCS*, volume 12089, 193–204. Cham: Springer.

[Moreno-Jiménez and Torres-Moreno 2020] Moreno-Jiménez, L.-G., and Torres-Moreno, J.-M. 2020. LiSSS: A new multi-annotated multi-emotion corpus of literary spanish sentences. *Computación and Sistemas* 24:1139–1147.

[Oliveira and Cardoso 2015] Oliveira, H. G., and Cardoso, A. 2015. Poetry generation with PoeTryMe. In *CCR-TCM '15*, volume 7. Paris: Atlantis Thinking Machines.

[Ormazabal et al. 2022] Ormazabal, A.; Artetxe, M.; Agirrezabal, M.; Soroa, A.; and Agirre, E. 2022. Poelm: A meter-and rhyme-controllable language model for unsupervised poetry generation. *arXiv preprint arXiv:2205.12206*.

[Qu et al. 2020] Qu, Y.; Liu, P.; Song, W.; Liu, L.; and Cheng, M. 2020. A text generation and prediction system: Pre-training on new corpora using bert and gpt-2. In *2020 IEEE 10th international conference on electronics information and emergency communication (ICEIEC)*, 323–326. IEEE.

[Radford et al. 2019] 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.

[Sridhara et al. 2010] Sridhara, G.; Hill, E.; Muppaneni, D.; Pollock, L.; and Vijay-Shanker, K. 2010. Towards automatically generating summary comments for java methods. In *IEEE/ACM International Conference on Automated Software Engineering*, 43–52. Antwerp, Belgium: ACM.

[Szymanski and Ciota 2002] Szymanski, G., and Ciota, Z. 2002. Hidden markov models suitable for text generation. In Mastorakis, N.; Kluev, V.; and Koruga, D., eds., *WSEAS '02*, 3081–3084. Athens: WSEAS - Press.

[Ta et al. 2023] Ta, H. T.; Rahman, A. B. S.; Majumder, N.; Hussain, A.; Najjar, L.; Howard, N.; Poria, S.; and Gelbukh, A. 2023. Wikides: A wikipedia-based dataset for generating short descriptions from paragraphs. *Information Fusion* 90:265–282.

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