Towards an Efficient Approach for Controllable Text Generation

Iván Martínez-Murillo, Paloma Moreda, Elena Lloret
Dept. of Software and Computing Systems
University of Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
{ivan.martinez, elena.lloret, moreda}@ua.es

Abstract

Since the emergence of Transformers architecture, the Natural Language Generation (NLG) field has advanced at breakneck speed. Large language models (LLMs) have achieved remarkable results in the field of generative artificial intelligence (AI). Nevertheless, they also present some problems worth analysing: not only are they computationally non-viable to academia, but they also have other issues, such us not generating text in a fully controllable way or the phenomenon known as hallucination. Because of this, the purpose of this paper is to outline and set the ideas for a new PhD thesis research. This PhD thesis will aim at advancing the state of the art by discovering new cost-effective, efficient and high-performing approaches to controlled text generation that could perform well in the different NLG tasks. Therefore, the main objective of this PhD thesis is to design a novel and efficient task-agnostic architecture that could obtain equivalent performance of LLMs, while generating text in a controllable way and including external commonsense knowledge.

1 Introduction

Natural language generation (NLG) field is the sub-field within natural language processing (NLP) area that generates natural language to meet a communicative goal (Reiter and Dale, 1997). Traditionally, there was a more classical and global vision about the NLG architecture that implied to divide generation in three stages: (1) macro-planning, (2) micro-planning and (3) surface realisation (Reiter and Dale, 1997). Later, neural networks caused a new trend in NLG, involving what we know nowadays as generative artificial intelligente (AI). Generative AI is a trend that encompasses systems that are constructed applying machine learning algorithms (Sun et al., 2022). Whitin this trend, Transformers (Vaswani et al., 2017) have revolutionised the NLG field owing to the concept of attention. Several proposals based on Transformers have been made, being Large Language Models (LLMs) the ones which better performance have achieved in tasks such as text summarisation or machine-translation, among others (Wolf et al., 2020). Despite this, these models present some issues worth commenting on. On the one hand, bests LLMs, such as GPT4 (estimated to have 1 trillion of parameters) (OpenAI, 2023) or LLaMa (65 billions of parameters) (Touvron et al., 2023) have a huge amount of parameters in their neural networks, which is only available to big companies, such as Google, due to the economic and temporary expense of training that models. On the other hand, these models do not generate text in a fully controlled way, leading to problems, such as hallucination or the lack of commonsense, among others. In fact, hallucination occurs even in the most superior LLMs such as GPT4 (Zhao et al., 2023). Figure 1 shows an example of hallucination in ChatGPT.

Because of this, the purpose of this paper is to set up the ideas for a new PhD thesis in which we will study and present a novel architecture that

[1]Tested in May, 2023

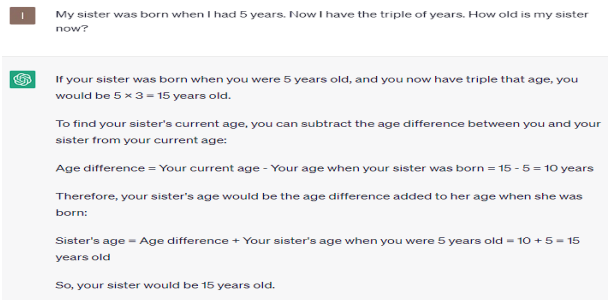

Figure 1: Hallucination example of ChatGPT [1]

could generate text in a more controlled manner, while being more efficient and less expensive. The proposed architecture will also include external commonsense knowledge with the aim of mitigating hallucination.

Therefore, the structure of this paper is organised in the following way: First of all, a commentary on the NLG background and the most common architectures are explained. Secondly, some research questions are introduced. Thirdly, the initial hypothesis about this PhD thesis and its corresponding objectives scheduled within a three years plan are set. Finally, a conclusion with the expected results of this thesis are presented.

2 Background

Research in NLG started by the end of 1970 (McDonald, 2010) and since then, it has advanced substantially. Depending on the type input, NLG can be traditionally classified into 2 main subgroups (Vicente et al., 2015): (1) text-to-text generation (T2T) and (2) data-to-text generation (D2T). Input data in D2T generation can adopt several types including images, voice, binary data, databases and knowledge. Recently, with the emergence of generative AI, the concept of (3) none-to-text is also introduced (Chandu and Black, 2020).

Other classifications are based on the task typology the generation system has been trained for. According to (Dong et al., 2022), NLG tasks are divided into three classes:

1. Text abbreviation: These tasks are devoted to detect the most relevant information in a text and condense that information into a short text, such as text summarization or question generation.

2. Text expansion: These tasks aim at generate completing sentences or texts from some meaningful words. Short text expansion and topic-to-essay generation are examples of this type of task.

3. Text rewriting and reasoning: These task work towards rewriting text into another style or applying reasoning methods, e.g. text style transfer and dialogue generation.

To achieve the communicative goal of the aforementioned tasks, several types of architectures have been proposed along this time. Based on the existing literature concerning NLG, some key papers proposing these architectures have been selected and have been represented in a temporal timeline. Figure 3 shows the evolution of architecture trends in NLG. These architectures can be grouped into three main categories (Gatt and Krahmer, 2018):

1. Modular architectures: This type of architectures follow a sequential scheme, which makes a clear distinction between distinct sub-tasks. The most popular modular architecture was proposed by Reiter (1994), which consists in a pipeline of three phases plus one optional phase, where the input into a sub-task is the output of the preceding sub-task. Figure 2 shows the different sub-tasks in the classical modular architecture.

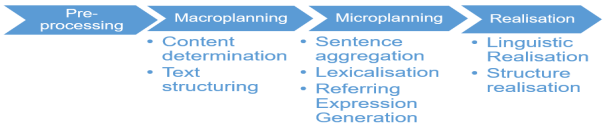

Figure 2: Sub-task division in the modular architecture for the stages proposed by (Reiter and Dale, 1997)

Some examples of this type of architectures can be found in (Mann and Moore, 1981), (Hovy, 1987), (Levelt, 1989), (Nirenburg et al., 1989) and (Reiter, 1994).

2. Planning perspectives: This type of architectures have a similar sub-task division similar to modular architectures, but they are more flexible owing to they allow to combine two or more consecutive sub-tasks into the same task. An example of this combination is to perform text structuring and sentence aggregation sub-tasks in the same tasks. Within this group, some examples of approaches that could be highlighted can be found in (Fikes and Nilsson, 1971), (Appelt, 1985), (Hovy, 1991), (Bateman, 1997),(Koller and Stone, 2007), (Rieser and Lemon, 2009), (Nakatsu and White, 2010) and (Lemon, 2011).

3. Global approaches: This type of architectures do not distinguish between sub-tasks, performing the entire generation process in a single task, having a strong reliance on statistical learn-

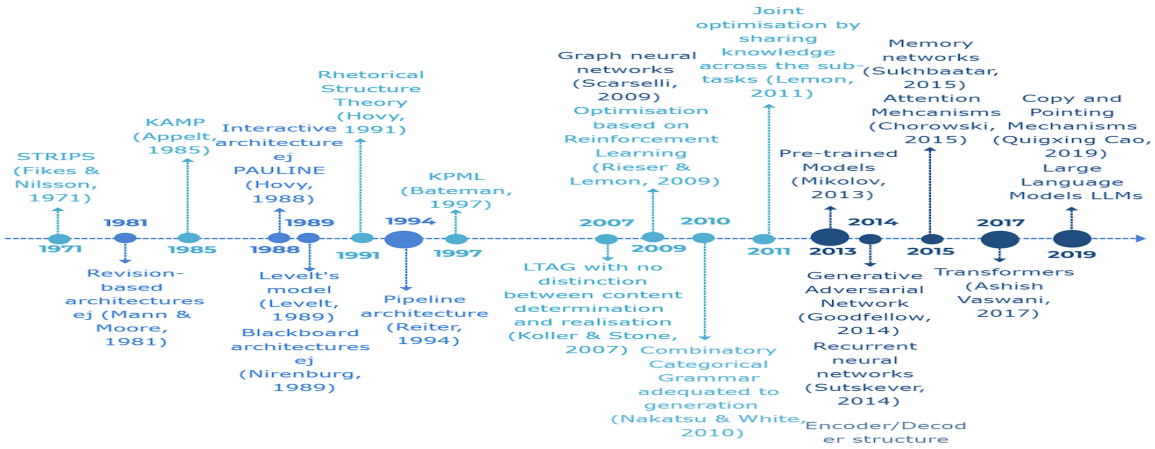

Figure 3: Timeline of NLG architectures (modular architectures in blue, planning perspectives in light blue, and global approaches in dark blue).

ing. Transformers (Vaswani et al., 2017) are an example of architecture within this category. With an encoder/decoder structure and an attention mechanism (Chorowski et al., 2015), Transformers and LLMs have revolutionised the NLG field. Research works that fall under this group are: Graph Neural Networks (Scarselli et al., 2008), Generative Adversarial Nets (Mirza et al., 2014), Recurrent Neural Networks (Sutskever et al., 2014), Pre-trained Models (Mikolov et al., 2013), Memory Networks (Sukhbaatar et al., 2015) and Copy and Pointing Mechanism (See et al., 2017). However, they also present some problems, as mentioned in Section 1.

Considering these problems, there is still some promising future directions to enhance text generation models. Some of the future directions in which this PhD thesis will focus are suggested in the following section.

3 Open Research Questions

In order to advance towards an efficient approach for controllable text generation that can overcome the drawbacks state-of-the-art architectures have, several research questions are suggested and discussed.

What is controllable text generation, and what are the most common techniques to address it? Controllable text generation is the task of generating natural language whose attributes can be controlled (Prabhumoye et al., 2020). These attributes can be stylistics (politeness, sentiment, etc), based on the demographic attributes of the interlocutor (age, gender, etc), or based on the content (including some keywords, entities, order of information, etc).

In order to control text generation, there are three main strategies (Erdem et al., 2022):

1. Via hyperparameters: Language models are trained with huge amounts of texts, which maybe cause that training data is unbalanced. Controlling the generation by hyperparameters could help the model to do a better generalisation of knowledge.

2. Via additional input: This group of methods consist on fine-tuning pre-trained models with additional input in order to adapt a pre-trained model to have a good performance in a more specific.

3. Via conditional training: This term refers to the group of training methods that utilise internal control variables that enrich the generation with specific capabilities.

During development of this PhD thesis, I will study and combine all three groups of approaches to propose a model that could produce text in a controllable way.

What is hallucination, what causes hallucination and which are the best ways to mitigate it? Hallucination in NLG refers to a text generated by a NLG model that is nonsensical or unfaithful to the provided source input (Ji et al., 2023). There are two categories of hallucinations: *intrinsic hallucinations* when the generated text refutes the input text, and *extrinsic hallucinations* when the generated text cannot be proved by the input.

Hallucination can be caused at two stages of the generation: both during the construction of datasets which may contain source-reference divergences, and during the training and inference step caused by the incomprehension to represent information in the encoder and decoder.

To solve this, there are some ways to keep hallucination at a low level. First of all, creating a faithful dataset, or automatically cleaning data from existing datasets. Secondly, by altering the structure of encoders and decoders to make them interpret semantics of the input in a better way. Thirdly, by proposing an optimal training strategy such as reinforcement learning or controllable generation. Finally, including external commonsense knowledge could help the model to mitigate hallucination. This PhD thesis will focus on the analysis of controllable generation techniques to reduce hallucination along with inclusion of external commonsense knowledge.

Is is possible to obtain an architecture that performs equally to LLMs without being as computationally demanding as them? Recently, LLMs have been the most hot topic in the NLG area, achieving a high performance in most of the latest models such as GPT4 (OpenAI, 2023), LLaMa (Touvron et al., 2023) and BLOOM (Scao et al., 2022), among others. Nevertheless, they have one major inconvenient. The time and computational expense needed to train these models are inaccessible to academia, as mentioned in Section 1. Thus, this PhD thesis will analyse and propose cost-effective architectures that could approximate LLMs performance and also solving some issues these models have.

Is there a task-agnostic architecture able to perform well for different tasks? Most of researches in the NLG area are focused on a specific task that while they perform correctly in one task, they underperform in others. Thus, this study will analyse most common task-agnostic techniques in order to propose a model that could achieve a high performance at every task.

4 Objectives

Given the research questions defined in Section 3 that we aim to cover in this thesis, our initial objective is that a cost-effective and efficient NLG approach that implements controllable text generation techniques along with external commonsense knowledge will help to mitigate the problem of hallucination, without worsening the results compared to the best-performing state-of-the-art models and will be able to perform well in different generation tasks.

To complete this objective, the following tasks with its corresponding schedule along three years have been proposed, as it can be seen in Figure 4. The schedule is divided in three sub-groups. In *Group A* the state-of-the-art will be studied. In *group B* an architecture will be proposed and tested. Finally, in *group C* the proposed architecture will be adapted to different NLG tasks.

A1. To analyse the state-of-the-art focused on controllable text generation techniques.

A2. To analyse the state-of-the-art focused on hallucination mitigation techniques.

A3. To analyse the state-of–the-art focused on task-agnostic architectures.

B1. To compare the performance of open-source state-of-the-art architectures using a common benchmark.

B2. To propose a cost-effective architecture that can generate text in a controllable way.

B3. To evaluate the performance of the proposed architecture against state-of-the-art architectures.

C1. To adapt the architecture to some of NLG tasks, e.g., summarisation or text simplification.

C2. To compare results with some architectures oriented to a specific task.

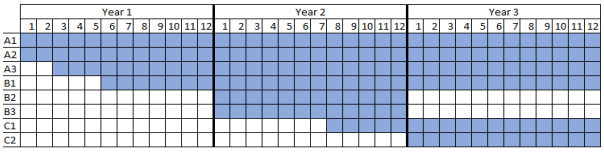

Figure 4: PhD thesis schedule

5 Conclusion

In spite of the great performance LLMs have for NLG, they also present some drawbacks. Thus, there is some room for improvement to advance scientific knowledge in NLG. In light of this, the objective of this PhD thesis is to find a more efficient architecture that could produce text in a controllable way and mitigate as much as possible the phenomena known as hallucination as much as possible by exploiting the use of external commonsense knowledge. Once an architecture is defined, this line of work will focus on adapting that architecture to achieve a cost-effective performance in some NLG tasks, and measuring that performance. We expect to obtain similar and comparable results to state-of-the-art models, but solving the issue of hallucination while using an efficient model that will help to reduce the carbon footprint.

Acknowledgements

This work has been funded by the European Commission ICT COST Action "Multi-task, Multilingual, Multi-modal Language Generation" (CA18231). In addition, the research work conducted is part of the R&D projects "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00), funded by MCIN/ AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"; "CLEAR.TEXT:Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities" (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and "European Union NextGenerationEU/PRTR"; and the project "NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/21)" funded by the Generalitat Valenciana.

References

Appelt, DE. 1985. Planning english sentences. cambridge university press.

Bateman, John A. 1997. Enabling technology for multilingual natural language generation: the kpml development environment. *Natural Language Engineering*, 3(1):15–55.

Chandu, Khyathi Raghavi and Alan W Black. 2020. Positioning yourself in the maze of neural text generation: A task-agnostic survey.

Chorowski, Jan K, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28.

Dong, Chenhe, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Comput. Surv.*, 55(8), dec.

Erdem, Erkut, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Fikes, Richard E and Nils J Nilsson. 1971. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208.

Gatt, Albert and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation.

Hovy, Eduard. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Hovy, Eduard H. 1991. *Approaches to the planning of coherent text*. Springer.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar.

Koller, Alexander and Matthew Stone. 2007. Sentence generation as a planning problem. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 336–343, Prague, Czech Republic, June. Association for Computational Linguistics.

Lemon, Oliver. 2011. Learning what to say and how to say it: Joint optimisation of spoken dialogue management and natural language generation. *Computer Speech & Language*, 25(2):210–221.

Levelt, W. 1989. Speaking: From intention to articulation mit press. *Cambridge, MA*.

Mann, William C and James A Moore. 1981. Computer generation of multiparagraph english text. *American Journal of Computational Linguistics*, 7(1):17–29.

McDonald, David D. 2010. Natural language generation. *Handbook of natural language processing*, 2:121–144.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Mirza, Mehdi, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Ian J Goodfellow, and Jean Pouget-Abadie. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680.

Nakatsu, Crystal and Michael White. 2010. Generating with discourse combinatory categorial grammar. *Linguistic Issues in Language Technology*, 4.

Nirenburg, Sergei, Victor R Lesser, and Eric Nyberg. 1989. Controlling a language generation planner. In *IJCAI*, pages 1524–1530.

OpenAI. 2023. Gpt-4 technical report.

Prabhumoye, Shrimai, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th*

International Conference on Computational Linguistics, pages 1–14, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Reiter, Ehud and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Reiter, Ehud. 1994. Has a consensus nl generation architecture appeared, and is it psycholinguistically plausible?

Rieser, Verena and Oliver Lemon. 2009. Natural language generation as planning under uncertainty for spoken dialogue systems. *Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation*, pages 105–120.

Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Scarselli, Franco, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

See, Abigail, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Sukhbaatar, Sainbayar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. *Advances in neural information processing systems*, 28.

Sun, Jiao, Q Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D Weisz. 2022. Investigating explainability of generative ai for code through scenario-based design. In *27th International Conference on Intelligent User Interfaces*, pages 212–228.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Vicente, Marta, Cristina Barros, Fernando S Peregrino, Francisco Agulló, and Elena Lloret. 2015. La generación de lenguaje natural: análisis del estado actual. *Computación y Sistemas*, 19(4):721–756.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.