Neurosymbolic Large Neighbourhood Search

Arnaud Delage-Reid
Polytechnique Montréal, Canada

Gilles Pesant
Polytechnique Montréal, Canada
IVADO, Canada

Amal Zouaq
Polytechnique Montréal, LAMA-WeST Lab, Canada
MILA, Canada

1 Introduction and Motivation

Neurosymbolic AI (NeSy) aims to combine the pattern recognition strengths of neural models with the formal guarantees of symbolic reasoning. This synergy is essential for generative tasks where output must not only be fluent but also respect strict structural constraints (e.g., molecule discovery or safety-critical text generation [3]). While Large Language Models (LLMs) excel at natural-language semantics, they often struggle with hard structural rules. Conversely, Constraint Programming (CP) excels at ensuring structural satisfaction but faces challenges in capturing complex linguistic nuances [6].

Current hybrid architectures using CP for sequential generation [4, 6, 2] typically rely on left-to-right decoding. However, two major hurdles remain: (i) in some applications such as text generation, variable domains can be extremely large, which may slow down domain filtering or general computation over domain values; (ii) sequences generated in a left-to-right manner may insufficiently anticipate the satisfaction of structural relationships, ending awkwardly or even aborting (the *procrastination problem* [4]). Other methods exist but lack the satisfaction guarantees provided by CP, requiring extensive training, preprocessing, or bespoke mechanisms to represent task-specific constraints.

2 Proposed Framework: NeSyLNS

We introduce Neurosymbolic Large Neighbourhood Search (NeSyLNS), an architecture that reframes constrained generation as an iterative refinement process using the Large Neighbourhood Search (LNS) paradigm [7]. Instead of rigid sequential generation, NeSyLNS starts from an initial (potentially poor) valid solution and iteratively improves it through a *Destroy and Repair* cycle.

This approach specifically addresses the previously identified challenges:

- **Efficiency:** By fixing a majority of variables during the *Destroy* phase, we drastically reduce the active search space, accelerating the CP component.
- **Bidirectionality:** We utilize a Masked Language Model (MLM) instead of an autoregressive LLM. MLMs are natively suited for iterative infilling, allowing the framework to repair any part of the sequence—including "procrastinated" errors—with full bidirectional context [1].

Figure 1 illustrates our approach applied to a constrained text generation task such as generating a solution with exactly 12 words. Starting from a valid but poor sentence, we destroy it by masking certain words, biased by their unlikeliness according to the MLM. Then, we repair the sentence by placing it in a CP model and asking our solver to fill in the masked words. Thanks to the CPBP framework [5], we can inject the probability distributions from the MLM directly into the symbolic solver. This ensures that the repair remains biased toward semantically likely solutions while maintaining a 100% guarantee of constraint satisfaction.

32nd International Conference on Principles and Practice of Constraint Programming (CP 2026).
Editor: Nicolas Beldiceanu; Article No. 12; pp. 12:1–12:3
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

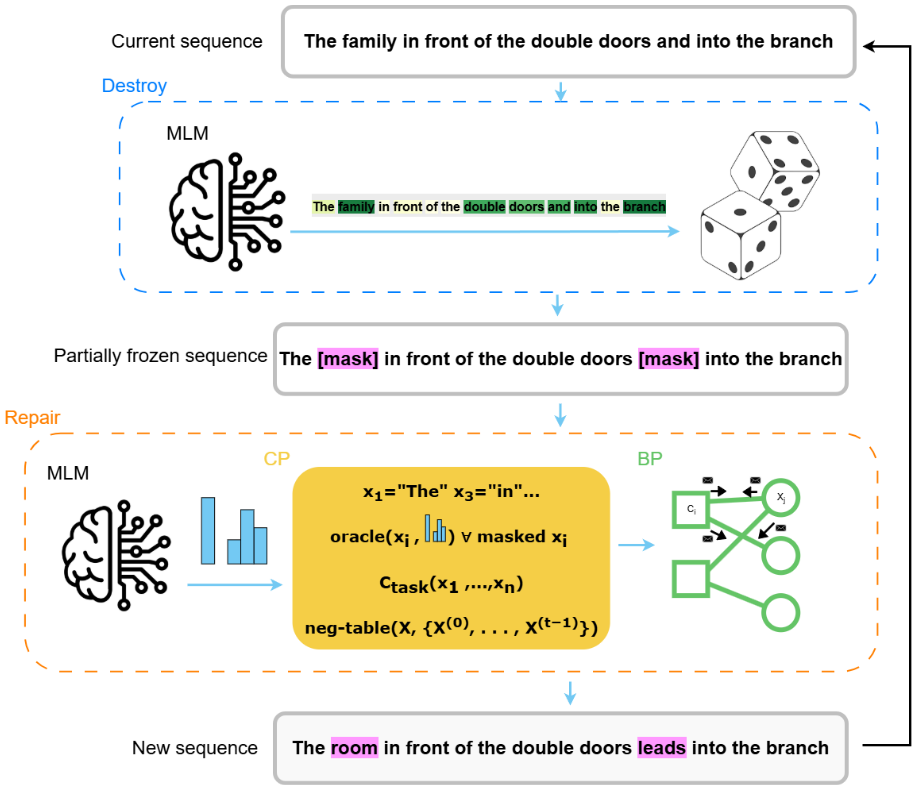

Figure 1 Architecture of the NeSyLNS framework. The MLM identifies variables to mask (Destroy), and the CPBP solver explores the local neighbourhood to find valid, high-probability replacements (Repair).

3 Discussion and Future Work

Our experiments across text generation and molecule discovery show that NeSyLNS circumvents the need for task-specific fine-tuning, offering a robust and general-purpose approach to constrained generative AI. The results demonstrate that our method manages to create diverse solutions even in highly constrained settings. In these circumstances, we show that BP-based inference is essential for quality. We also demonstrate that neural masking consistently outperforms random strategies by more effectively targeting semantically weak components, with only modest computational overhead.

The core contribution of this work lies in integrating soft semantic guidance within a framework that guarantees strict satisfaction of symbolic constraints. More broadly, NeSyLNS suggests a shift from purely sequential generation toward iterative, globally-informed refinement, enabling tighter interaction between neural and symbolic reasoning.

Despite its strengths, the current architecture relies on the availability of an initial valid solution and operates on fixed-length sequences, limiting exploration of the solution space. Addressing these limitations is part of ongoing work, drawing inspiration from local search metaheuristics: (i) relaxing neighbourhoods to allow temporarily invalid intermediate states, and (ii) introducing sequence expansion and reduction operators for variable-length generation. Finally, the versatility of CP combined with masked autoencoder models [8] opens promising avenues beyond sequence generation, including applications to structured and multimodal domains such as graphs, images, or hybrid data representations.

References

1 Samuel Belkadi, Libo Ren, Nicolo Micheletti, Lifeng Han, and Goran Nenadic. Generating Synthetic Free-text Medical Records with Low Re-identification Risk using Masked Language Modeling. In Abteen Ebrahimi, Samar Haider, Emmy Liu, Sammar Haider, Maria Leonor Pacheco, and Shira Wein, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–206, Albuquerque, USA, April 2025. Association for Computational Linguistics. `doi:10.18653/v1/2025.naacl-srw.20`.

2 Alexandre Bonlarron, Florian Régin, Elisabetta De Maria, and Jean-Charles Régin. Large language model meets constraint propagation. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 10036–10044. International Joint Conferences on Artificial Intelligence Organization, August 2025. `doi:10.24963/ijcai.2025/1115`.

3 Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. AI Safety in Generative AI Large Language Models: A Survey, July 2024. `arXiv:2407.18369`, `doi:10.48550/arXiv.2407.18369`.

4 Virasone Manibod, David Saikali, and Gilles Pesant. Constrained Sequential Inference in Machine Learning Using Constraint Programming. In *Thirty-Fourth International Joint Conference on Artificial Intelligence*, volume 1, pages 2628–2636, September 2025. `doi:10.24963/ijcai.2025/293`.

5 Gilles Pesant. From support propagation to belief propagation in constraint programming. *Journal of Artificial Intelligence Research*, 66:123–150, 2019. `doi:10.1613/JAIR.1.11487`.

6 Florian Régin, Elisabetta De Maria, and Alexandre Bonlarron. Combining Constraint Programming Reasoning with Large Language Model Predictions. In Paul Shaw, editor, *30th International Conference on Principles and Practice of Constraint Programming (CP 2024)*, volume 307 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 25:1–25:18, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. `doi:10.4230/LIPIcs.CP.2024.25`.

7 Paul Shaw. Using constraint programming and local search methods to solve vehicle routing problems. In Michael J. Maher and Jean-Francois Puget, editors, *Principles and Practice of Constraint Programming - CP98, 4th International Conference, Pisa, Italy, October 26-30, 1998, Proceedings*, volume 1520 of *Lecture Notes in Computer Science*, pages 417–431. Springer, 1998. `doi:10.1007/3-540-49481-2\_30`.

8 Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, and In So Kweon. A Survey on Masked Autoencoder for Visual Self-supervised Learning. In *Thirty-Second International Joint Conference on Artificial Intelligence*, volume 6, pages 6805–6813, August 2023. `doi:10.24963/ijcai.2023/762`.