Génération contrôlée

Projet pour le Master MIND - François Yvon

Sorbonne Université, CNRS, ISIR

Les algorithmes de génération automatique de texte sont au cœur du fonctionnement des grands modèles de langue. Durant l'inférence, ces algorithmes sont en charge de calculer, token après token, la meilleure réponse possible à une requête formulée par un prompt utilisateur. La versatilité des requêtes et des types de réponse possibles impose que ces algorithmes soient capables de prendre en charge des contraintes diverses, d'une complexité variable : produire des réponses dans une langue donnée, dans un style prédéfini, en évitant certains mots "tabous", en respectant des contraintes de bonne formation linguistique ou de cohérence sémantique, en simulant une forme de raisonnement, en garantissant la conformité de la sortie, en tenant compte de la syntaxe d'un langage de programmation, en générant un code qui s'exécute correctement, etc. Ces contraintes peuvent être graduelles (par ex. quand il s'agit de style ou de "toxicité") ou bien binaires (par ex. pour la satisfaction de contraintes syntaxiques).

Parmi les nombreuses techniques pour contraindre la génération à être conforme, nous nous focaliserons sur des méthodes qui opèrent durant l'inférence, sans aucune intervention sur les paramètres du modèle – en laissant donc de côté les méthodes d'affinage ou de renforcement, qui ne fournissent pas de garanties exactes sur la bonne formation des séquences générées. Les méthodes étudiées reposent sur des stratégies d'échantillonnage séquentiel, le défi étant de sélectionner le prochain token parmi un ensemble restreint de choix possibles (la contrainte interdit certains tokens), tout en s'assurant que la distribution de probabilité dans laquelle ces échantillons sont choisies est la plus proche possible de la "vraie" distribution a posteriori sur toutes les séquences conformes. Dans un ordre croissant de difficulté, on étudiera les approches de (Miao et al, 2019), de (Zhang et al, 2020), de (Yang and Klein, 2021) et de (Lew et al, 2023), dans le vue d'aboutir à une présentation cohérente, à une implantation et à une comparaison expérimentale sur divers cas simples. Le cas échéant, on poursuivra avec les travaux plus récents de (Zhao et al, 2024) et de (Lipikin et al, 2025).

Références :

— Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. CGMH : constrained sentence generation by metropolis-hastings sampling. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence

and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'19/IAAI'19/EAAI'19), Vol. 33. AAAI Press, Article 839, 6834–6842. https ://doi.org/10.1609/aaai.v33i01.33016834

— Maosen Zhang, Nan Jiang, Lei Li, and Yexiang Xue. 2020. Language Generation via Combinatorial Constraint Satisfaction : A Tree Search Enhanced Monte-Carlo Approach. In Findings of the Association for Computational Linguistics : EMNLP 2020, pages 1286–1298, Online. Association for Computational Linguistics.

— Kevin Yang and Dan Klein. 2021. FUDGE : Controlled Text Generation With Future Discriminators. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, pages 3511–3535, Online. Available from https ://aclanthology.org/2021.naacl-main.276/.

— Alexander K. Lew, Tan Zhi-Xuan, Gabriel Grand, Vikash Mansinghka. 2023. Sequential Monte Carlo Steering of Large Language Models using Probabilistic Programs. Proc. ICLR 2023, available from https ://openreview.net/forum ?id=Ul2K0qXxXy#all.

— Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Baker Grosse. Probabilistic inference in language models via twisted sequential Monte Carlo. In Proceedings of the International Conference on Machine Learning, 2024. URL https ://proceedings.mlr.press/v235/zhao24c.html.

— Benjamin Lipkin, Benjamin LeBrun, Jacob Hoover Vigly, João Loula, David R. MacIver, Li Du, Jason Eisner, Ryan Cotterell, Vikash Mansinghka, Timothy J. O'Donnell, Alexander K. Lew, and Tim Vieira (2025). Fast controlled generation from language models with adaptive weighted rejection sampling. In the proceedings of COLM (Outstanding Paper Award).