Latent Variable for NLP Models as VAEs/Diffusion Processes

Joseph Le Roux

November 28, 2024

1 Context

Probabilistic models for NLP tasks often make use of latent random variables to enrich the *generative story* at the expense of more complex learning and prediction algorithms. This has been used for many tasks in NLP such as AMR semantic parsing e.g. [LT18] or [Zho+20], Machine Translation e.g. [CRA19] or [Yan+19], and Topic Modelling [CLN24] *inter alia*. Some researchers even claim that LLMs can indeed be considered as latent variable models [Wan+23]. For this internship we want to revisit the early works on latent variables in NLP, in particular constituent parsing, with modern latent variable formalisms developed for Machine Learning in the last 10 years.

Hybridization of context-free grammars with latent probabilistic ML models (L-PCFGs) have been pioneered 20 years ago [MMT05], studied in several publications [PK08] [Coh+12] [LRF13] and more recently in [ZZT18] . In all these models, non-terminal symbols which correspond to grammatical categories are enriched with latent random variables that encode some sort of specialization. For instance, the *determiner* category can be equipped with a latent variable that encodes subtypes such as demonstratives or possessives. The precise meaning of this specialization is often difficult to intrepret beyond POS tag specialization as they are learned from data in an unsupervised fashion.

These approaches learn model parameters (*i.e.* rule weights) using a variant of Expectation-Maximization (EM) relying on the fact that marginal probabilities can be computed efficiently using the inside-outside algorithm. On the other hand, this can be seen as a constraint on the probabilistic modelling since it imposes that EM, more precisely the Expectation step, can be computed exactly and efficiently.

Recent ML probabilistic latent variable models, such as VAEs [KW14], use richer models but avoid this constraint by approximating expectations over latent values with a Monte-Carlo estimator based on an inference network. This approach is called *variational inference*. In this case, training optimizes a bound on the log-likelihood (ELBo). Some recent papers investigate ELBo training for unsupervised context-free grammars, see [ZBN20] or [KDR19], but not in the case of L-PCFGs.

Finally, recent latent variable models are expressed as diffusion processes [Aus+21], either continuous or discrete. Training amounts to parametrize a denoising neural network optimizing a ELBo loss, or more sophisticated functions based on score functions [LME24]. While there are already NLP tasks that have taken advantage of this new models, it remains a challenge to incorporate diffusion in a structured prediction task such as parsing.

2 Internship Description

For this internship, we propose to work on PCFG modelling, either as a variational inference instance, or possibly as a diffusion process. This includes the definition of inference

and learning for such models, and implementation for experiments on standard treebanks.

For variational inference, the study will start by the design of the encoder and the decoder of the VAE architecture. For diffusion, the study will start with the design of the forward diffusion process, in the variable space or in the latent space. Some recent work addresses diffusion in an unsupervised setting [SHL24], and this can be a potential avenue for future improvement.

3 Application

We are looking for a candidate with either NLP background (master level) with very good knowledge of Machine Learning methods for NLP, or with a strong ML background willing to adapt recent models to NLP tasks. We expect proficiency with python and deep learning libraries such as pytorch. Knowledge of parsing more generally graph-based approaches to NLP tasks is a plus.

The internship work will be carried out at LIPN at Université Sorbonne Paris Nord, on site (no remote), with possibilities of extension to a three-year Ph.D. funding (2025-2028). This internship is funded by the ANR Project SEMIAMOR (2024-2028).

For additional information, please contact leroux@lipn.fr. If you are interested please attach to your application email a CV, a cover letter and a transcript of your Master level marks.

References

[Aus+21] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. "Structured Denoising Diffusion Models in Discrete State-Spaces". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 17981–17993. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/958c530554f78bcd8e97125b70e6973d-Paper.pdf.

[CLN24] Sunrit Chakraborty, Rayleigh Lei, and XuanLong Nguyen. *Learning Topic Hierarchies by Tree-Directed Latent Variable Models*. 2024. arXiv: 2408.14327 [math.ST]. URL: https://arxiv.org/abs/2408.14327.

[Coh+12] Shay B. Cohen, Karl Stratos, Michael Collins, Dean P. Foster, and Lyle Ungar. "Spectral Learning of Latent-Variable PCFGs". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 223–231. URL: https://aclanthology.org/P12-1024.

[CRA19] Iacer Calixto, Miguel Rios, and Wilker Aziz. "Latent Variable Model for Multimodal Translation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 6392–6405. DOI: 10.18653/v1/P19-1642. URL: https://aclanthology.org/P19-1642.

[KDR19] Yoon Kim, Chris Dyer, and Alexander Rush. “Compound Probabilistic Context-Free Grammars for Grammar Induction”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2369–2385. DOI: 10.18653/v1/P19-1228. URL: https://www.aclweb.org/anthology/P19-1228.

[KW14] Diederik P. Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *Proceedings of ICLR*. Ed. by Yoshua Bengio and Yann LeCun. 2014. URL: http://arxiv.org/abs/1312.6114.

[LME24] Aaron Lou, Chenlin Meng, and Stefano Ermon. “Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution”. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp. Vol. 235. Proceedings of Machine Learning Research. PMLR, July 2024, pp. 32819–32848. URL: https://proceedings.mlr.press/v235/lou24a.html.

[LRF13] Joseph Le Roux, Antoine Rozenknop, and Jennifer Foster. “Combining PCFG-LA Models with Dual Decomposition: A Case Study with Function Labels and Binarization”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1158–1169. URL: https://www.aclweb.org/anthology/D13-1116.

[LT18] Chunchuan Lyu and Ivan Titov. “AMR Parsing as Graph Prediction with Latent Alignment”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 397–407. DOI: 10.18653/v1/P18-1037. URL: https://aclanthology.org/P18-1037.

[MMT05] Takuya Matsuzaki, Yusuke Miyao, and Jun’ichi Tsujii. “Probabilistic CFG with Latent Annotations”. In: *ACL*. 2005. URL: https://aclanthology.org/P05-1010.

[PK08] Slav Petrov and Dan Klein. “Sparse Multi-Scale Grammars for Discriminative Latent Variable Parsing”. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 867–876.

[SHL24] Sebastian Sanokowski, Sepp Hochreiter, and Sebastian Lehner. “A Diffusion Model Framework for Unsupervised Neural Combinatorial Optimization”. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp. Vol. 235. Proceedings of Machine Learning Research. PMLR, July 2024, pp. 43346–43367. URL: https://proceedings.mlr.press/v235/sanokowski24a.html.

[Wan+23] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. “Large Language Models Are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: https://openreview.net/forum?id=BGvkwZEGt7.

[Yan+19] Xuewen Yang, Yingru Liu, Dongliang Xie, Xin Wang, and Niranjan Balasubramanian. “Latent Part-of-Speech Sequences for Neural Machine Translation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 780–790. DOI: 10.18653/v1/D19-1072. URL: https://aclanthology.org/D19-1072.

[ZBN20] Hao Zhu, Yonatan Bisk, and Graham Neubig. “The Return of Lexical Dependencies: Neural Lexicalized PCFGs”. In: *Transactions of the Association for Computational Linguistics* 8 (2020). Ed. by Mark Johnson, Brian Roark, and Ani Nenkova, pp. 647–661. DOI: 10.1162/tacl_a_00337. URL: https://aclanthology.org/2020.tacl-1.42.

[Zho+20] Qiji Zhou, Yue Zhang, Donghong Ji, and Hao Tang. “AMR Parsing with Latent Structural Information”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 4306–4319. DOI: 10.18653/v1/2020.acl-main.397. URL: https://aclanthology.org/2020.acl-main.397.

[ZZT18] Yanpeng Zhao, Liwen Zhang, and Kewei Tu. “Gaussian Mixture Latent Vector Grammars”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1181–1189. DOI: 10.18653/v1/P18-1109. URL: https://aclanthology.org/P18-1109.