Latent Structure Models for Natural Language Processing

André F.T. Martins[♫♋] **Tsvetomila Mihaylova**[♫] **Nikita Nangia**[♈] and **Vlad Niculae**[♫]
[♫]Instituto de Telecomunicações, Lisbon, Portugal
[♈]Center for Data Science, New York University, US
[♋]Unbabel, Lisbon, Portugal
andre.martins@unbabel.com, tsvetomila.mihaylova@gmail.com,
nikitanangia@nyu.edu, vlad@vene.ro

Link to materials:
https://deep-spin.github.io/tutorial/

1 Description

Latent structure models are a powerful tool for modeling compositional data, discovering linguistic structure, and building NLP pipelines (Smith, 2011). Words, sentences, paragraphs, and documents represent the fundamental units in NLP, and their discrete, compositional nature is well suited to combinatorial representations such as trees, sequences, segments, or alignments. When available from human experts, such structured annotations (like syntactic parse trees or part-of-speech information) can help higher-level models perform or generalize better. However, linguistic structure is often **hidden** from practitioners, in which case it becomes useful to model it as a latent variable.

While it is possible to build powerful models that obliviate linguistic structure almost completely (such as LSTMs and Transformer architectures), there are two main reasons why modeling it is desirable: first, incorporating **structural bias** during training can lead to better generalization, since it corresponds to a more informed and more appropriate prior. Second, discovering hidden structure provides better **interpretability**: this is particularly useful when used in conjunction with neural networks, whose typical architectures are not amenable to interpretation. The learnt structure offers highly valuable insight into how the model organizes and composes information.

This tutorial will cover recent advances in latent structure models in NLP. In the last couple of years, the general idea of **hidden linguistic structure** has been married to **latent representation learning** via neural networks. This has allowed powerful modern NLP models to learn to uncover, for example, latent word alignments or parse trees, jointly, in an **unsupervised** or **semi-supervised** fashion, from the signal of higher-level downstream tasks like sentiment analysis or machine translation. This avoids the need for preprocessing data with off-the-shelf tools (e.g., parsers, word aligners) and engineering features based on their outputs; and it is an alternative to techniques based on parameter sharing, transfer learning, multi-task learning, or scaffolding (Swayamdipta et al., 2018; Peters et al., 2018; Devlin et al., 2019; Strubell et al., 2018), as well as techniques that incorporate structural bias directly in model design (Dyer et al., 2016; Shen et al., 2019).

The proposed tutorial is about such **discrete latent structure models**. We discuss their motivation, potential, and limitations, then explore in detail three strategies for designing such models:

- Reinforcement learning;
- Surrogate gradients;
- End-to-end differentiable methods.

A challenge with structured latent models is that they typically involve computing an "argmax" (i.e. finding a best scoring discrete structure such as a parse tree) in the middle of a computation graph. Since this operation has null gradients almost everywhere, gradient backpropagation cannot be used out of the box for training. The methods we cover in this tutorial differ among each other by the way they handle this issue.

Reinforcement learning. In a stochastic computation graph, such methods seek the hidden discrete structures that minimize an expected loss on a downstream task (Yogatama et al., 2017); similar to maximizing an expected reward in reinforcement learning with discrete actions. Estimated stochastic gradients are typically obtained with a combination

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 1–5
Florence, Italy, July 28 - August 2, 2019.

of Monte Carlo sampling and the score function estimator (a.k.a. REINFORCE, Williams, 1992). Such estimators often suffer from instability and high variance, requiring care (Havrylov et al., 2019).

Surrogate gradients. Such techniques usually involve approximating the gradient of a discrete, argmax-like mapping by the gradient of a continuous relaxation. Examples are the straight-through estimator (Bengio et al., 2013) and the structured projection of intermediate gradients optimization technique (SPIGOT; Peng et al. 2018). In stochastic graphs, surrogate gradients yield biased but lower-variance gradient estimators compared to the score function estimator. Related is the Gumbel softmax (Jang et al., 2017; Maddison et al., 2017; Choi et al., 2018; Maillard and Clark, 2018), which uses the reparametrization trick and a temperature parameter to build a continuous surrogate of the argmax operation, which one can then differentiate over. Structured versions were recently explored by Corro and Titov (2019a,b). One limitation of straight-through estimators is that backpropagating with respect to the sample-independent means may cause discrepancies between the forward and backward pass, which biases learning.

End-to-end differentiable approaches. Here, we directly replace the argmax by a continuous relaxation for which the exact gradient can be computed and backpropagated normally. Examples are structured attention networks and related work (Kim et al., 2017; Maillard et al., 2017; Liu and Lapata, 2018; Mensch and Blondel, 2018), which use marginal inference, or SparseMAP (Niculae et al., 2018a,b), a new inference strategy which yields a sparse set of structures. While the former is usually limited in which the downstream model can only depend on local substructures (not the entire latent structure), the latter allows combining the best of both worlds. Another line of work imbues structure into neural attention via sparsity-inducing priors (Martins and Astudillo, 2016; Niculae and Blondel, 2017; Malaviya et al., 2018).

This tutorial will highlight connections among all these methods, enumerating their strengths and weaknesses. The models we present and analyze have been applied to a wide variety of NLP tasks, including sentiment analysis, natural language inference, language modeling, machine translation, and semantic parsing. In addition, evaluations specific to latent structure recovery have been proposed (Nangia and Bowman, 2018; Williams et al., 2018). Examples and evaluation will be covered throughout the tutorial. After attending the tutorial, a practitioner will be better informed about which method is best suited for their problem.

2 Type of Tutorial & Relationship to Recent Tutorials

The proposed tutorial mixes the **introductory** and **cutting-edge** types. It will offer a gentle introduction to recent advances in structured modeling with **discrete** latent variables, which were not previously covered in any ACL/EMNLP/IJCNLP/NAACL related tutorial.

The closest related topics covered in recent tutorials at NLP conferences are:

- Variational inference and deep generative models (Aziz and Schulz, 2018); [1]
- Deep latent-variable models of natural language (Kim et al., 2018).[2]

Our tutorial offers a complementary perspective in which the latent variables are structured and discrete, corresponding to linguistic structure. We will briefly discuss the modeling alternatives above in the final discussion.

3 Outline

Below we sketch an outline of the tutorial, which will take three hours, separated by a 30-minutes coffee break.

1. Introduction (30 min)
   - Why latent variables?
   - Motivation and examples of latent structure in NLP
   - Continuous vs. discrete latent variables
   - Bypassing latent variables
     - Pipelines / external classifiers
     - Transfer learning / parameter sharing
     - Multi-task learning
   - Challenges: gradients of argmax
   - Categorical versus structured: the simplex and the marginal polytope
2. Reinforcement learning methods (30 min)

[1] https://github.com/philschulz/VITutorial
[2] http://nlp.seas.harvard.edu/latent-nlp-tutorial.html

- SPINN: parsing and classification with shared parameters
- Stochastic computation graphs
- The Score Function Estimator and REINFORCE (application: RL-SPINN with unsupervised parsing)
- Example: the ListOps diagnostic dataset benchmark
- Actor-critic methods & variance reduction

3. Surrogate gradient methods (30 min)
   - Unstructured: straight-through estimators
   - Structured: SPIGOT
   - Sampling categoricals with Gumbel-argmax
   - Gumbel-softmax: reparametrization and straight-through variants
   - Example: Gumbel Tree-LSTM to compose tree structures
   - Perturb-and-MAP / Perturb-and-parse

   Coffee break (30 min)

4. End-to-end differentiable formulations (60 min)
   - Attention mechanisms & hidden alignments
   - Sparse and grouped attention mechanisms
   - Structured attention networks
   - Example: dense / sparse differentiable dynamic programming
   - SparseMAP
   - Relationships with gradient approximation
   - Example: Natural language inference with latent structure (matchings and trees)

5. Closing Remarks and Discussion (30 min)
   - Is it Syntax? Addressing if existing methods learn recognizable grammars
   - Alternative perspectives:
     - Structural bias in model design
     - Deep generative models with continuous latent variables
   - Current open problems and discussion.

4 Breadth

We aim to provide the first unified perspective into multiple related approaches. Of the 31 referenced works, only 6 are co-authored by the presenters. In the outline, the first half presents exclusively work by other researchers and the second half present a mix of our own work and other people's work.

5 Prerequisites and reading

The audience should be comfortable with:

- **math**: basics of differentiability.
- **language**: basic familiarity with the building blocks of structured prediction problems in NLP, e.g., syntax trees and dependency parsing.
- **machine learning**: familiarity with neural networks for NLP, basic understanding of backpropagation and computation graphs.

6 Instructors

André Martins[3] is the Head of Research at Unbabel, a research scientist at Instituto de Telecomunicações, and an invited professor at Instituto Superior Técnico in the University of Lisbon. He received his dual-degree PhD in Language Technologies in 2012 from Carnegie Mellon University and Instituto Superior Técnico. His research interests include natural language processing, machine learning, deep learning, and optimization. He received a best paper award at the Annual Meeting of the Association for Computational Linguistics (ACL) for his work in natural language syntax, and a SCS Honorable Mention at CMU for his PhD dissertation. He is one of the co-founders and organizers of the Lisbon Machine Learning Summer School (LxMLS). He co-presented tutorials at NAACL in 2012, EACL in 2014, and EMNLP in 2014. He co-organized the NAACL 2019 Workshop on Structured Prediction for NLP (http://structuredprediction.github.io/SPNLP19) and the ICLR 2019 Workshop "Deep Reinforcement Learning Meets Structured Prediction".

Tsvetomila Mihaylova[4] is a PhD student in the DeepSPIN project at Instituto de Telecomunicações in Lisbon, Portugal, supervised by André Martins. She is working on empowering neural networks with a planning mechanism for structural search. She has a master's degree in Information Retrieval from the Sofia University, where she was also a teaching assistant in Artificial Intelligence. She is part of the organizers of a shared task in SemEval 2019.

[3] https://andre-martins.github.io
[4] https://tsvm.github.io

Nikita Nangia[5] is a PhD student at New York University, advised by Samuel Bowman. She is working on building neural network systems in NLP that simultaneously do structured prediction and representation learning. This work focuses on finding structure in language without direct supervision and using it for semantic tasks like natural language inference and summarization.

Vlad Niculae[6] is a postdoc in the DeepSPIN project at the Instituto de Telecomunicações in Lisbon, Portugal. His research aims to bring structure and sparsity to neural network hidden layers and latent variables, using ideas from convex optimization, and motivations from natural language processing. He earned a PhD in Computer Science from Cornell University in 2018. He received the inaugural Cornell CS Doctoral Dissertation Award, and co-organized the NAACL 2019 Workshop on Structured Prediction for NLP (http://structuredprediction.github.io/SPNLP19).

References

Wilker Aziz and Philip Schulz. 2018. Variational inference and deep generative models. In *Proc. ACL Tutorial Abstracts*.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *preprint arXiv:1308.3432*.

Jihun Choi, Kang Min Yoo, and Sang-goo Lee. 2018. Learning to compose task-specific tree structures. In *Proc. AAAI*.

Caio Corro and Ivan Titov. 2019a. Differentiable Perturb-and-Parse: Semi-supervised parsing with a structured variational autoencoder. In *Proc. ICLR*.

Caio Corro and Ivan Titov. 2019b. Learning latent trees with stochastic perturbations and differentiable dynamic programming. In *Proc. ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent neural network grammars. In *Proc. NAACL-HLT*.

Serhii Havrylov, Germán Kruszewski, and Armand Joulin. 2019. Cooperative learning of disjoint syntax and semantics. In *Proc. NAACL-HLT*.

[5] https://woollysocks.github.io

[6] https://vene.ro

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparametrization with Gumbel-Softmax. In *Proc. ICLR*.

Yoon Kim, Carl Denton, Loung Hoang, and Alexander M Rush. 2017. Structured attention networks. In *Proc. ICLR*.

Yoon Kim, Sam Wiseman, and Alexander Rush. 2018. Deep latent variable models of natural language. In *Proc. EMNLP Tutorial Abstracts*.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *TACL*, 6:63–75.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *Proc. ICLR*.

Jean Maillard and Stephen Clark. 2018. Latent Tree Learning with Differentiable Parsers:Shift-Reduce Parsing and Chart Parsing. In *Proc. ACL*.

Jean Maillard, Stephen Clark, and Dani Yogatama. 2017. Jointly learning sentence embeddings and syntax with unsupervised tree-LSTMs. *preprint arXiv:1705.09189*.

Chaitanya Malaviya, Pedro Ferreira, and André FT Martins. 2018. Sparse and constrained attention for neural machine translation. In *Proc. ACL*.

André FT Martins and Ramón Fernandez Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proc. ICML*.

Arthur Mensch and Mathieu Blondel. 2018. Differentiable dynamic programming for structured prediction and attention. In *Proc. ICML*.

Nikita Nangia and Samuel Bowman. 2018. ListOps: A diagnostic dataset for latent tree learning. In *Proc. NAACL Student Research Workshop*.

Vlad Niculae and Mathieu Blondel. 2017. A regularized framework for sparse and structured neural attention. In *Proc. NeurIPS*.

Vlad Niculae, André FT Martins, Mathieu Blondel, and Claire Cardie. 2018a. SparseMAP: Differentiable sparse structured inference. In *Proc. ICML*.

Vlad Niculae, André FT Martins, and Claire Cardie. 2018b. Towards dynamic computation graphs via sparse latent structure. In *Proc. EMNLP*.

Hao Peng, Sam Thomson, and Noah A Smith. 2018. Backpropagating through structured argmax using a SPIGOT. In *Proc. ACL*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. NAACL-HLT*.

Yikang Shen, Tan Shawn, Alessandro Sordoni, and Aaron Courville. 2019. Ordered neurons: Integrating tree structures into recurrent neural networks. In *Proc. ICLR*.

Noah A Smith. 2011. *Linguistic Structure Prediction*. Synth. Lect. Human Lang. Technol. Morgan & Claypool.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proc. EMNLP*.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A Smith. 2018. Syntactic scaffolds for semantic structures. In *Proc. EMNLP*.

Adina Williams, Andrew Drozdov, and Samuel R Bowman. 2018. Do latent tree learning models identify meaningful structure in sentences? *TACL*, 6:253–267.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256.

Dani Yogatama, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Wang Ling. 2017. Learning to compose words into sentences with reinforcement learning. In *Proc. ICLR*.