MoL thesis proposal:
A stochastic decoder for recurrent neural network grammars

Daan van Stigt

April 24, 2018

1 Project description

We propose to combine the recurrent neural network grammar (RNNG) [Dyer et al., 2016], a syntactic model of sentences parametrized by recurrent neural networks (RNNs), with the recently proposed stochastic decoder [Schulz and Aziz, 2018], an RNN augmented with continuous latent variables. The goal of this combination is to enrich the RNNG with latent variables to explicitly model uncertainty and variability. We conjecture that this improves the RNNG both for use as parser and as language model. From this we can move in a number of directions, which we draw out below. We first describe the foundation in more detail.

1.1 Recurrent neural network grammars

The RNNG [Dyer et al., 2016] is a probabilistic model of sentences that explicitly models words together with their hierarchical structure. It is based on a transition-based parser, in which the transition probabilities are parametrized by RNNs that condition on the entire syntactic derivation thus far, thus making no Markov assumptions. The RNNG can be formulated both as a discriminative and generative model and can work with both constituency and dependency grammar formalisms.

The discriminative variant models $p(y|x)$, the conditional probability distribution over parse trees y given a sentence x and can be used for the task of parsing: assigning a tree structure to a given sentence. The generative variant models the joint distribution $p(x, y)$, thus modelling sentences together explicitly with hierarchical structure. The generative model can be evaluated as a parser by finding the tree y that maximizes $p(x, y)$ for given x. Additionally, the generative model can function as a language model. Marginalizing over the set of valid trees $\mathcal{Y}(x)$ for a sentence x we obtain the marginal distribution of the sentence under the model:

$$p(x) = \sum_{y \in \mathcal{Y}(x)} p(x, y). \tag{1}$$

This is a promising application of the generative RNNG: it has the rich parameterization of RNN language models but by explicitly modeling hierarchical structure and marginalizing over it, the RNNG moves beyond the sequential assumption underlying regular RNNs. Experiments moreover show that it can outperform such sequential RNNs in language modelling.

There is a challenge, however. As a result of the unbounded dependencies in the model, the above inference tasks are intractable, and an approximate inference in the form of importance sampling is used to obtain estimates of these. Dyer et al. [2016] use the trained discriminative parser as proposal distribution $q(y|x)$:

$$\begin{aligned} p(x) &= \sum_{y \in \mathcal{Y}(x)} p(x, y) \\ &= \sum_{y \in \mathcal{Y}(x)} q(y|x) \frac{p(x, y)}{q(y|x)} \\ &= \mathbb{E}_{q(y|x)} \left[\frac{p(x, y)}{q(y|x)} \right] \\ &\approx \sum_{i=1}^{N} \frac{p(x, y^{(i)})}{q(y^{(i)}|x)}. \end{aligned}$$

Where the trees $\{y^{(i)}\}_{i=1}^{N}$ are sampled independently from q by ancestral sampling over the sequence of transition actions. How well these samples reflect the true posterior distribution over trees for the given sentence will determine the closeness of this approximation and thus the success of the generative RNNG.

1.2 Stochastic decoder

The stochastic decoder is a recently proposed model that explicitly models variation in the context of encoder-decoder architectures for neural machine translation (NMT). The model is a conditional neural language model augmented with a chain of latent Gaussian variables on the target side, one for each hidden state of the RNN. These latent variables are introduced to explicitly model variation in the training data. Posterior inference is performed with amortized variational inference, taking ideas from the variational auto-encoder [Kingma and Welling, 2013].

We propose to use the stochastic RNN to replace the RNN that parametrizes the RNNG. Thus we aim to train the discriminative parser for more variability. We conjecture that explicitly modeling this variability will make the discriminative parser both more robust as a parser, as well as a better proposal distribution for approximate inference in the generative model. We conjecture that this will improve the generative RNNG, both as parser and as language model.

1.3 Further directions

The above combination is the bedrock of this project. Beyond it, there are a number of directions in which to proceed.

- **Change encoder architecture.** The idea is to replace the RNNs with other neural sequence architectures based on convolutions, or stacked attention layers [Vaswani et al., 2017].
- **Chance priors.** The incorporation of latent variables through the stochastic decoder provides the opportunity to incorporate further inductive bias into the model by our choice of priors. The stochastic RNN in [Schulz and Aziz, 2018] uses Gaussian latent variables. A first extension is to use another distribution, for example to use priors that induce sparsity. For this we can use Dirichlet priors with a choice of hyperparameter that induces sparsity. This could help interpretation of the trained model by making the posterior distribution more like categorical variables.
- **Make latent variables hierarchical.** The stochastic RNN incorporates a chain of local latent variables, one for each hidden state of the RNN, and one initial hidden state. When the initial latent variable is made global—such that all the other latent variables condition on it—we obtain a hierarchical latent variable model, and effectively obtain a sentence-level variable. This sentence level variable can be explored as latent sentence embedding, along the lines of [Bowman et al., 2016a]. We conjecture that the syntactic bias of our model will make the posterior over the latent space be more structurally organized, something that was also observed in the grammar VAE introduced in [Kusner et al., 2017].
- **Learn $p(x, y)$ and $q(y|x)$ jointly.** This is an ambitious objective. The aim here is to jointly learn an approximate posterior $q_\lambda(y|x)$, instead of The posterior is parametrized by variational parameters λ and trained along with the generative model $p(x, y)$ by directly optimizing the following lower bound on the marginal likelihood:

  $$\log p(x) \geq \mathbb{E}_{q_\lambda(y|x)}\left[\log \frac{p(x, y)}{q_\lambda(y|x)}\right]$$

  This is challenging because it involves variational inference for discrete latent variables (the parse actions that produce valid trees), for which we cannot rely on reparametrization [Kingma and Welling, 2013] to compute gradients of this lower bound. Instead, we must resort to policy gradient methods for estimates of this gradient, which are generaly difficult to get to work due to high variance.

  There is some precedence here, luckily. To set of in this direction we can follow [Cheng et al., 2017], in which it is shown how to perform variational inference for the RNNG by adapting methods introduced in [Miao and Blunsom, 2016].

.

2 Related work

- **Genereative parsing.** There exists previous work on generative parsing. [Buys and Blunsom, 2015] propose a generative dependency parser also based on a stack-reduce parsing algorithm. The transitions are parametrized by feedforward neural networks that work on dense feature templates similar to [Chen and Manning, 2014].

- **Tree-structured encoders.** Much research has been done on recursive TreeRNNs, which build vector representations for sentences by following the structure of the sentence's parse tree. The SPINN [Bowman et al., 2016b] is an example from this research that is closest to the RNNG: it combines a shift reduce algorithm with RNNs. However, it is mostly concerned with *encoding* sentences for discriminative tasks such as sentiment classification, and not for generative modelling.

- **Latent tree learning.** This is recent work that attempts to train TreeRNNs neural networks both to parse a sentence and then use the resulting parse to encode the sentence, for example for classification, without access to gold parse trees. An example is [Yogatama et al., 2016] in which reinforcement learning is used to learn trees for sentences that help optimize a downstream tasks such as sentiment classification or textual entailment. This research area has recently been surveyed nicely in [Williams et al., 2017]. If we decide to train the generative RNNG with variational inference, we can look to [Yogatama et al., 2016] for examples.

- **Applications to NMT.** In [Eriguchi et al., 2017] the RNNG is incorporated into the decoder of a standard encoder-decoder model with attention used for neural machine translation (NMT). It is shown that modelling syntax on the target side can improve translation quality.

3 Research questions

- Does the stochastic decoder improve the parsing performance of the discriminative model? Is the parser more robust, especially for out of training domains?

- Does the stochastic decoder make the discriminative RNNG a better proposal distribution for the generative RNNG? Does it improve generative parsing? Does it improve the language model?

- Can we improve interpretability of RNNG by modeling with different priors on the latent states of the stochastic RNN? What can we learn about the trained model using this?
- Can we further improve the RNNG by using discrete variational inference to train generative model with a joint objective?

4 Planning

Here are some guidelines.

- The thesis is 30 EC, which corresponds to 5 months of full-time workload.
- Start: end of April 2018
- End: October 2018
- Total: 6 months (5 + 1 month vacation / spare)

We propose the following timeline for this project:

1. Replicate the original paper. We first implement the discriminative variant and then continue with the generative model. Our implementation will work for both constituency and dependency grammars. We evaluate for parsing and language modelling on the Penn Treebank (for constituency trees) and on universal dependencies (for dependency trees). (1 month - May)
2. Incorporate the stochastic RNN into the RNNG: this give the S-RNNG. A good first step to this end is to implement the stochastic RNN first as language model. Evaluate the S-RNNG in parsing and as language model. (1 month - June)
3. Then explore extensions: either learn approximate posterior jointly using discrete variational inference following [Cheng et al., 2017] and [Miao and Blunsom, 2016]; or change / extend priors of the stochastic decoder. (1-2 months - August/September)
4. Write down results. (1 month)

References

S. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio. *Generating Sentences from a Continuous Space.* 2016a.

S. R. Bowman, J. Gauthier, A. Rastogi, R. Gupta, C. D. Manning, and C. Potts. A fast unified model for parsing and sentence understanding. In *Association for Computational Linguistics (ACL)*, 2016b.

J. Buys and P. Blunsom. Generative incremental dependency parsing with neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 863–869. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-2142. URL `http://www.aclweb.org/anthology/P15-2142`.

D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

J. Cheng, A. Lopez, and M. Lapata. A generative parser with a discriminative recognition algorithm. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, Vancouver, Canada, 2017. Association for Computational Linguistics.

C. Dyer, A. Kuncoro, M. Ballesteros, and N. A. Smith. Recurrent neural network grammars. *CoRR*, abs/1602.07776, 2016. URL `http://arxiv.org/abs/1602.07776`.

A. Eriguchi, Y. Tsuruoka, and K. Cho. Learning to parse and translate improves neural machine translation. *CoRR*, abs/1702.03525, 2017. URL `http://arxiv.org/abs/1702.03525`.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL `http://arxiv.org/abs/1312.6114`.

M. J. Kusner, B. Paige, and J. M. Hernández-Lobato. Grammar variational autoencoder. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1945–1954, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/kusner17a.html`.

Y. Miao and P. Blunsom. Language as a latent variable: Discrete generative models for sentence compression. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 319–328. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1031. URL `http://www.aclweb.org/anthology/D16-1031`.

P. Schulz and W. Aziz. A stochastic decoder for neural machine translation. *Association for Computational Linguistics (ACL)*, 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Kaiser, Łukaszand Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

A. Williams, A. Drozdov, and S. R. Bowman. Do latent tree learning models identify meaningful structure in sentences? *CoRR*, abs/1709.01121, 2017. URL `http://arxiv.org/abs/1709.01121`.

D. Yogatama, P. Blunsom, C. Dyer, E. Grefenstette, and W. Ling. Learning to compose words into sentences with reinforcement learning. *CoRR*, abs/1611.09100, 2016. URL `http://arxiv.org/abs/1611.09100`.