Supplementary Materials: Backpropagating through Structured Argmax using a SPIGOT

Hao Peng[♢] **Sam Thomson**[♣] **Noah A. Smith**[♢]
[♢] Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣] School of Computer Science, Carnegie Mellon University
{hapeng,nasmith}@cs.washington.edu, sthomson@cs.cmu.edu

1 Implementation Details

Our implementation is based on the DyNet toolkit.[1] We use part-of-speech tags and lemmas predicted by NLTK.[2]

1.1 Syntactic-then-Semantic Parsing Experiment

Each input token is represented as the concatenation a word embedding vector, a learned lemma vector, and a learned vector for part-of-speech, all updated during training. In joint training, we apply early-stopping based on semantic dependency parsing development performance (in labeled F_1). We do not use mini-batch. We set the step size η for SPIGOT to 1.

Semantic dependency parser. We use the pruning techniques in Martins and Almeida (2014), and replace their feature-rich model with neural networks (Peng et al., 2018). We observe that the number of parts surviving pruning is linear in the sentence length ($5.5\times$ on average), with $\sim$99% recall.

We do not deviate far from the hyperparameter setting in Peng et al. (2017), with the only exception being that we use 50-dimensional lemma and part-of-speech embeddings, instead of 25.

Syntactic dependency parser. For the max-margin syntactic parsers used in PIPELINE, STE, and SPIGOT, we use the hyperparameters reported in Kiperwasser and Goldberg (2016), but replace their 125-dimensional BiLSTMs with 200-dimensional ones, and use 50-dimensional POS embeddings, instead of 25. We anneal the learning rate at a rate of 0.5 every 5 epochs.

For the marginal syntactic parser in SA, we follow the use of Adam algorithm (Kingma and Ba, 2015), but set a smaller initial learning rate of 5×10^{-4}, annealed at a rate of 0.5 every 4 epochs. The rest of the hyperparameters stay the same as the max-margin parser.

[1] https://github.com/clab/dynet
[2] http://www.nltk.org/

Hyperparameter	Values
MLP dimension	$\{100, 150, 200, 250, 300\}$
BiLSTM dimension	$\{100, 150, 200, 250, 300\}$
Embedding dropout	$\{0.2, 0.3, 0.4, 0.5\}$
MLP dropout	$\{0.0, 0.1, 0.2, 0.3, 0.4\}$

Table 1: Hyperparameters explored in sentiment classification experiments.

1.2 Semantic Parsing and Sentiment Classification Experiment

The model is trained for up to 30 epochs in the joint training stage. We apply early-stopping based on sentiment classification development accuracy. For semantic dependency parser, we follow the hyperparameters described in §1.1.

Sentiment classifier. We use 300-dimensional GloVe (Pennington et al., 2014) to initialize word embeddings, fixed during training. We use a single-layer BiLSTM, followed by a two-layer ReLU-MLP. Dropout in word embeddings and MLPs is applied, but not in LSTMs. We use Adam algorithm (Kingma and Ba, 2015), and follow the default procedures by DyNet for optimizer settings and parameter initializations. An ℓ_2-penalty of 10^{-6} is applied to all weights. Learning rate is annealed at a rate of 0.5 every 5 epochs. We use mini-batches of 32, and clip the ℓ_2-norm of gradients to 5 (Graves, 2013). We set the step size η for SPIGOT to $\frac{5}{32}$. We explore the same set of hyperparameters based on development performance for all compared models, summarized in Table 1.

References

Alex Graves. 2013. Generating sequences with recurrent neural networks. arXiv:1308.0850.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. ICLR*.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *TACL* 4:313–327.

André F. T. Martins and Mariana S. C. Almeida. 2014. Priberam: A Turbo semantic parser with second order features. In *Proc. of SemEval*.

Hao Peng, Sam Thomson, and Noah A. Smith. 2017. Deep multitask learning for semantic dependency parsing. In *Proc. of ACL*.

Hao Peng, Sam Thomson, Swabha Swayamdipta, and Noah A. Smith. 2018. Learning joint semantic parsers from disjoint data. In *Proc. of NAACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of EMNLP*.