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 $\eta$ 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 ($5.5 \times 10^{-4}$, annealed at a rate of 0.5 every 4 epochs. The rest of the hyperparameters stay the same as the max-margin parser.

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 $\ell_2$-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 $\ell_2$-norm of gradients to 5 (Graves, 2013). We set the step size $\eta$ for SPIGOT to $\frac{32}{5}$. We explore the same set of hyperparameters based on development performance for all compared models, summarized in Table 1.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP dimension</td>
<td>${100, 150, 200, 250, 300}$</td>
</tr>
<tr>
<td>BiLSTM dimension</td>
<td>${100, 150, 200, 250, 300}$</td>
</tr>
<tr>
<td>Embedding dropout</td>
<td>${0.2, 0.3, 0.4, 0.5}$</td>
</tr>
<tr>
<td>MLP dropout</td>
<td>${0.0, 0.1, 0.2, 0.3, 0.4}$</td>
</tr>
</tbody>
</table>

Table 1: Hyperparameters explored in sentiment classification experiments.
References


