A Supplementary Material

A.1 Datasets

We used the full-text portion of FrameNet 1.5 release[8] for frame-semantic role labeling. We use the same test set as Das et al. (2014), and create a validation set by selecting 8 documents from the train set. The dataset contains 3,139 train sentences with 16,621 target annotations, 387 validation sentences with 2,282 targets, and 2,420 test sentences with 4,427 targets. Each target from a given sentence is treated as an independent training instance. Following Täckström et al. (2015), we only use the first annotation for each target with multiple annotations.

We use the standard splits provided in OntoNotes for the CoNLL 2012 shared task. The dataset contains 115,812 train sentences with 278,026 target annotations, 15,680 validation sentences with 38,377 targets, and 12,217 test sentences with 29,669 targets.

We use the English coreference resolution data from the CoNLL 2012 shared task (Pradhan et al., 2012), containing 2,802, 343 and 348 documents for train, validation, and test respectively.

Syntax OntoNotes contains 115,812 training instances for the syntactic scaffold. There is no overlap between FrameNet and OntoNotes training data.

A.2 Experimental Settings

We used GloVe embeddings (Pennington et al., 2014) for tokens in the vocabulary, with out of vocabulary words being initialized randomly. For frame-SRL, 300 dimensional embeddings were used, and kept fixed during training. For PropBank SRL, we used 100 dimensional embeddings which were updated during training. A 100-dimensional embedding is learned for indicating target positions, following Zhou and Xu (2015). Bidirectional LSTMs with highway connections (Srivastava et al., 2015) between 6 layers are used, each layer containing 300-dimensional hidden states. A dropout of 0.1 is applied to the LSTMs. The feed-forward networks are of dimension 150 and of depth 2, with rectified linear units (Nair and Hinton, 2010). A dropout of 0.2 is applied to the feed-forward networks.

We limit the maximum length of spans to $D = 15$ in FrameNet, resulting in oracle recall of 95% on the development set, and to 13 in Propbank, resulting in an oracle recall of 96%. An identical maximum span length is used for the scaffold task.

For the SRL scaffolds, we randomly sample instances from OntoNotes to match the size of the SRL data, and alternate between training an SRL batch and a scaffold batch. In FrameNet, this amounts to downsampling OntoNotes. For PropBank SRL, this amounts to upsampling syntactic annotations from OntoNotes, since a sentence has a single syntactic tree, but could have multiple target annotations, each of which is a training instance.

The mixing ratio, δ is set to 1.0 (tuned across $\{0.1, 0.5, 1.0, 1.5\}$) for frame and PropBank SRL. We use Adam (Kingma and Ba, 2014) for optimization, at a learning rate of 0.001, and a minibatch of size 32. Our dynamic program formulation for loss computation and inference under the semi-CRF is also batched. To prevent exploding gradients, the 2-norm of the gradient is clipped to 1 before a gradient update (Graves, 2013). All models are trained for a maximum of 20 epochs, and stopped early based on dev F_1.

We extended the `AllenNLP` library,[9] which is built on top of PyTorch.[10] Each experiment was run on a single TitanX GPU.

For the coreference model, we use the same hyperparameters and experimental settings from Lee et al. (2017). The only new hyperparameter needed for scaffolding is the mixing ratio, δ, which we set to 0.1 based on performance on the validation set.

[8] A later release, 1.7 is also available, but for ease of comparison to other published systems we report results on the earlier release.

[9] http://allennlp.org/

[10] http://pytorch.org/