Name	Value
Encoder BiLSTM hidden layer size	600
Dependency LSTM hidden layer size	200
The dimensions of embeddings p,q	100, 128
MLPs hidden layer size	4000
Dropout rate in MLPs	0.5
Max transitions during reinforcement learning	10

Table 2: Hyper-parameters in our experiments.

Model	DM	PAS	PSD	Avg.
Peng+ 17 Freda3	90.4	92.7	78.5	88.0
Wang+ 18 Ens.	90.3	91.7	78.6	86.9
Peng+ 18	91.6	-	78.9	-
IPS	91.1	92.4	78.6	88.2
IPS +ML	91.2	92.5	78.8	88.3
IPS +RL	91.6‡	**92.8**‡	79.2‡	88.7‡
IPS +ML +RL	**92.0**‡	**92.8**‡	**79.3**‡	**88.8**‡

Table 3: Labeled parsing performance on in-domain test data. Avg. is the micro-averaged score of three formalisms. ‡ of the *+RL* models represents that the scores are statistically significant at $p < 10^{-3}$ with their non-*RL* counterparts.

makes training quite unstable. Therefore we fix the BiLSTM parameters during policy gradient. In our multi-task learning set-up, we apply multi-task learning of the shared stacked BiLSTMs (Søgaard and Goldberg, 2016; Hashimoto et al., 2017) in supervised learning. We use task-specific MLPs for the three different linguistic formalisms: DM, PAS and PSD. We train the shared BiLSTM using multi-task learning beforehand, and then we fine-tune the task-specific MLPs with policy gradient. We summarize the rest of our hyper-parameters in Table 2.

4 Experiments

We use the SemEval 2015 Task18 (Oepen et al., 2015) SDP dataset for evaluating our model. The training corpus contains 33,964 sentences from the WSJ corpus; the development and in-domain test were taken from the same corpus and consist of 1,692 and 1,410 sentences, respectively. The out-of-domain test set of 1,849 sentences is drawn from Brown corpus. All sentences are annotated with three semantic formalisms: DM, PAS and PSD. We use the standard splits of the datasets (Almeida and Martins, 2015; Du et al., 2015). Following standard evaluation practice in semantic dependency parsing, all scores are *micro-averaged* F-measures (Peng et al., 2017; Wang et al., 2018) with labeled attachment scores (LAS).

Model	DM	PAS	PSD	Avg.
Peng+ 17 Freda3	85.3	**89.0**	76.4	84.4
Peng+ 18	86.7	-	77.1	-
IPS +ML	86.0	88.2	77.2	84.6
IPS +ML +RL	**87.2**‡	88.8‡	**77.7**‡	**85.3**‡

Table 4: Labeled parsing performance on out-of-domain test data. Avg. is the micro-averaged score of three formalisms. ‡ of the *+RL* models represents that the scores are statistically significant at $p < 10^{-3}$ with their non-*RL* counterparts.

The system we propose is the IPS parser trained with a multi-task objective and fine-tuned using reinforcement learning. This is referred to as *IPS+ML+RL* in the results tables. To highlight the contributions of the various components of our architecture, we also report ablation scores for the IPS parser without multi-task training nor reinforcement learning (*IPS*), with multi-task training (*IPS+ML*) and with reinforcement learning (*IPS+RL*). At inference time, we apply heuristics to avoid predicting circles during decoding (Camerini et al., 1980); see Supplementary Material, §A.1. This improves scores by 0.1 % or less, since predicted circles are extremely rare. We compare our proposed system with three state-of-the-art SDP parsers: Freda3 of Peng et al. (2017), the ensemble model in Wang et al. (2018) and Peng et al. (2018). In Peng et al. (2018), they use syntactic dependency trees, while we do not use them in our models.[7]

The results of our experiments on in-domain dataset are also shown in Table 3. We observe that our basic *IPS* model achieves competitive scores in DM and PAS parsing. Multi-task learning of the shared BiLSTM (*IPS+ML*) leads to small improvements across the board, which is consistent with the results of Peng et al. (2017). The model trained with reinforcement learning (*IPS+RL*) performs better than the model trained by supervised learning (*IPS*). These differences are significant ($p < 10^{-3}$). Most importantly, the combination of multi-task learning and policy gradient-based reinforcement learning (*IPS+ML+RL*) achieves the best results among all IPS models and the previous state of the art models, by some margin. We also obtain similar results for the out-of-domain

[7] Dozat and Manning (2018) report *macro-averaged* scores instead, as mentioned in their ACL 2018 talk, and their results are therefore not comparable to ours. For details, see the video of their talk on ACL2018 that is available on Vimeo.