Table 1: Semi-supervised setting for the citation-field extraction task.

| No. | GE | PG | DVN | R-SPEN | SG-SPEN | SG-SPEN-sup | DVN-sup |
|---|---|---|---|---|---|---|---|
| 5 | 54.7 | 55.6 | 50.5 | 55.0 | **65.5** | 53.0 | 57.4 |
| 10 | 57.9 | 67.7 | 60.6 | 65.5 | **71.7** | 62.4 | 61.9 |
| 50 | 68.0 | 76.5 | 67.7 | 81.5 | **82.9** | 81.6 | 81.4 |

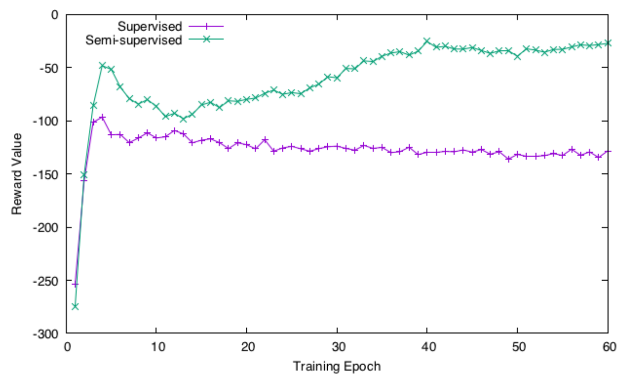


Figure 1: The test reward value of SG-SPEN's outputs trained in the supervised setting and semi-supervised settings with five labeled data points.

We appreciate the reviewers' comments and concerns. Two common questions of reviewers are comparisons to fully-supervised and semi-supervised settings that we address here. The reported results here and in the main paper are with respect to labeled test data on the reported task-specific measure, which is the same as the reward function for shape parsing and multi-label classification. For citation-field extraction, the reward function and accuracy measure are different. The citation reward function is based on domain knowledge and is noisy, and the ground-truth label may not have the highest reward value.

**Fully-supervised setting**: For multi-label classification, DVNs achieve 44.7 $F_1$ score for Bibtex and 37.1 $F_1$ score for Bookmarks, while SG-SPENs achieve 44.0 $F_1$ score for Bibtex and 38.4 $F_1$ score for Bookmarks. Since for this task, the reward function is the oracle $F_1$ score, the performance of SG-SPENs is on a par with the fully supervised setting on Bibtex and Bookmarks. For citation-field extraction, we train SG-SPEN and DVN with token-level accuracy as the reward function. SG-SPEN achieves 91.0% and DVN achieves 90.5% token-level accuracy. We also trained SG-SPEN with domain-knowledge based citation reward function, which resulted in 90.6% token-level accuracy. For shape parsing, we trained the neural shape parser in the supervised setting as described in Sharma et al. (2018), which resulted in 60.0% intersection over union (IOU) comparing to 56.3% IOU of SG-SPEN without labeled data. Neural shape parser requires more labeled training data for better generalization.

**Semi-supervised setting**: We study the citation-field extraction task in the semi-supervised setting with 1000 unlabeled and 5, 10, and 50 labeled data points. SG-SPEN can be extended for the semi-supervised setting by using the ground-truth label instead of the output of the search whenever it is available. Similarly, for R-SPEN, we can evaluate the rank-based objective using a pair of model's prediction and ground truth output when available. For DVNs, if the ground truth label is available, we use adversarial sampling as suggested by Gygli et al. (2017). We also reported the result of PG training with EMA baseline when the model is pre-trained with the labeled data. We reported the performance of GE based on Mann & McCallum (2010). We also reported the results of SG-SPENs when they are only trained with the labeled data using the citation reward function (SG-SPEN-sup). Since the citation reward function is based on domain knowledge and is noisy, DVNs struggle in matching the energy values with the noisy rewards, so we also trained DVNs with token-level accuracy (not available for the unlabeled data) as the reward function (DVN-sup) for the reference.

SG-SPEN's performance is better than the other baselines in the presence of limited labeled data. However, since the training objective of R-SPEN and SG-SPEN are similar for the labeled data (both use rank-based objective), as we increase the number of labeled data, their performance become closer. DVNs also benefit from the labeled data, but it is very sensitive to noisy reward functions (see DVN and DVN-sup in Table 1). To better understand the behavior of SG-SPEN in the semi-supervised setting, we compare the reward value of test data for SG-SPENs during training with five labeled data in the fully-supervised and semi-supervised settings (see Figure 1). The unlabeled data helps SG-SPEN to better generalize to unseen data.