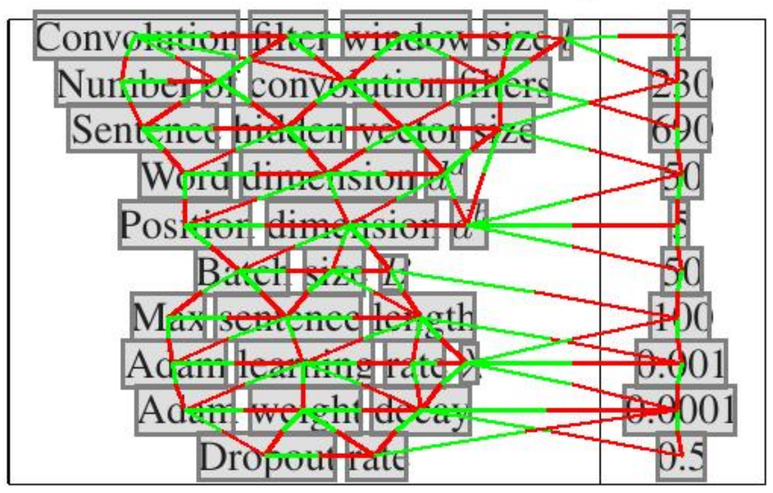
aligning Freebase relations with the New York Times corpus (NYT). The Freebase relations are divided into two parts for training and testing. The training set aligns the sentences from the corpus of the years 2005-2006, and the testing one aligns the sentences from 2007. The dataset contains 53 possible relationships including a special relation type 'NA' which indicates no relation between the mentioned two entities. The resulted training and testing data contain 570,088 and 172,448 sentences, respectively. We further randomly extract 10 percent of relation pairs and the corresponding sentences from the training data as the validation data for model selection and parameter tuning, and leave the rest as the actual training data.

Similar to previous works (Mintz et al. 2009; Lin et al. 2016), we evaluate our model in the held-out testing data. The evaluation compares the extracted relation instances discovered from the test sentences against Freebase relation data. It makes the assumption that the inference model has similar performance in relation instances inside and outside Freebase. We report the precision/recall curves, Precision@N (P@N), and average precision in our experiments.

Parameter Settings

We tune the parameters of maximum sentence length, learning rate, weight decay, and batch size by testing the performance on the validation dataset. For other parameters, we use the same parameters as (Lin et al. 2016). Table 1 shows the major parameters used in our experiments.

Table 1: Parameter settings.



For the initialization of W, we use the following strategy. We define a ratio e, and assign $w_{11} = 1 - e$, and the rest $K - 1$ elements to be $e/(1 - K)$. We do evaluation on the validation dataset and pick $e = 0.1$ in the candidate set $\{0.001, 0.01, 0.1, 0.2, 0.3, 0.5\}$. We pretrain our model for 2 epochs by setting W as an identity matrix and then fine tune our model for another 18 epochs with trainable W.

Comparison with Baseline Methods

To evaluate our proposed approach, we select several baseline methods for comparison by held-out evaluation:

Mintz (Mintz et al. 2009) is a traditional distant supervised model.

MultiR (Hoffmann et al. 2011) is a probabilistic, graphical model for multi-instance learning that can handle overlapping relations.

MIML (Surdeanu et al. 2012) is a method that models both multiple instances and multiple relations.

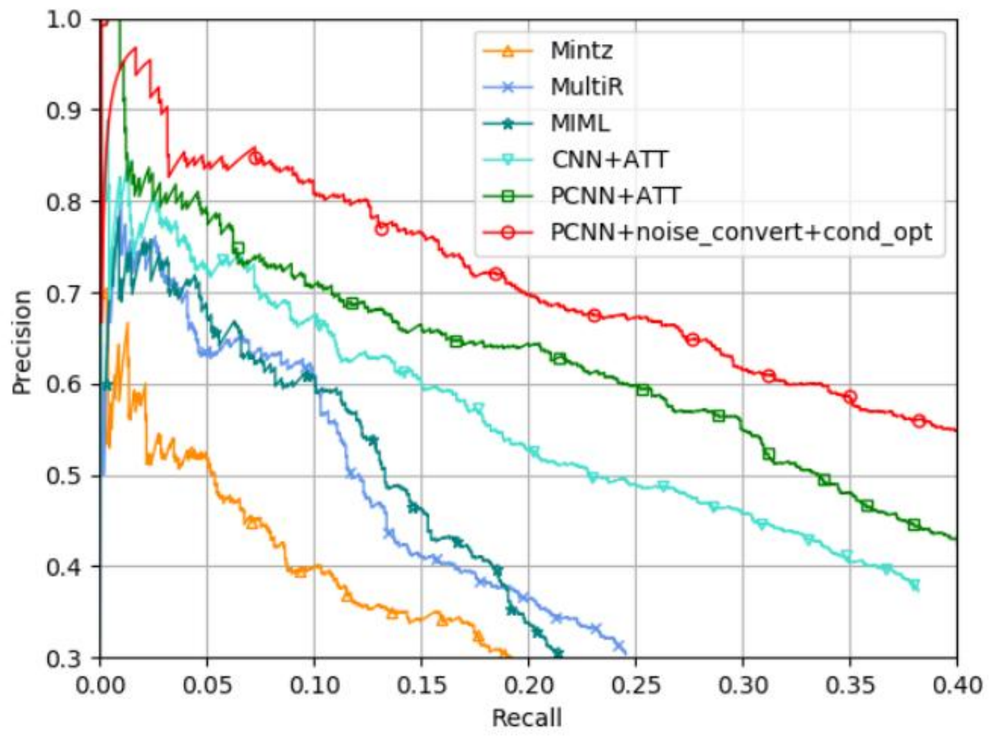


Figure 3: Performance comparison of proposed model and baseline methods. Our model and our implementation of PCNN+ATT both pick the model with the highest accuracy on the validation dataset.

CNN + ATT and **PCNN+ATT** (Lin et al. 2016) are two methods that first represent a sentence by CNN and PCNN respectively, and then use sentence-level selective attention to model a group of sentences with the same entity pair.

For the above baseline methods, we implement the state of the art method **PCNN+ATT**. To make fair comparison, we use the same implementation of the component PCNN in our model and **PCNN+ATT**, and use the same hyper-parameters. For other methods, we use the results from the source code released by the authors.

For our method and **PCNN+ATT**, we run 20 epochs in total, and track the accuracy on the validation dataset. The model is saved for every 200 batches during training. We use the saved model that has the highest accuracy on the validation dataset for making predictions on the testing dataset.

Figure 3 shows the precision/recall curves for all methods, including ours (labeled as **PCNN+noise_convert+cond_opt**). For all of the baseline methods, we can see that PCNN+ATT shows much better performance than others, which demonstrates the effectiveness of the sentence-level selective attention. Although PCNN+ATT has shown significant improvement over other baselines, our method still gains great improvement over PCNN+ATT. Particularly, Table 2 compares the precision@N (P@N) between our model and PCNN+ATT. For PCNN+ATT, we report both the P@N numbers from the authors' original paper and the results based on our implementation. Our method achieves the highest values for P@100, P@200, P@300, with mean value of 9.1 higher than original report of PCNN+ATT, and 7.1 higher than our implementation of PCNN+ATT. To avoid randomness in the best single model, we also compare our method and PCNN+ATT on ensemble of several saved models in one single training run. For each method, the corresponding ensemble model averages the probability scores for each