Structured Minimally Supervised Learning for Neural Relation Extraction

Fan Bai and **Alan Ritter**
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH
{bai.313, ritter.1492}@osu.edu

Abstract

We present an approach to minimally supervised relation extraction that combines the benefits of learned representations and structured learning, and accurately predicts sentence-level relation mentions given only proposition-level supervision from a KB. By explicitly reasoning about missing data during learning, our approach enables large-scale training of 1D convolutional neural networks while mitigating the issue of label noise inherent in distant supervision. Our approach achieves state-of-the-art results on minimally supervised sentential relation extraction, outperforming a number of baselines, including a competitive approach that uses the attention layer of a purely neural model.[1]

1 Introduction

Recent years have seen significant progress on tasks such as object detection, automatic speech recognition and machine translation. These performance advances are largely driven by the application of neural network methods on large, high-quality datasets. In contrast, traditional datasets for relation extraction are based on expensive and time-consuming human annotation (Doddington et al., 2004) and are therefore relatively small. Distant supervision (Mintz et al., 2009), a technique which uses existing knowledge bases such as Freebase or Wikipedia as a source of weak supervision, enables learning from large quantities of unlabeled text and is a promising approach for scaling up. Recent work has shown promising results from large-scale training of neural networks for relation extraction (Toutanova et al., 2015; Zeng et al., 2015).

There are, however, significant challenges due to the inherent noise in distant supervision. For example, Riedel et al. (2010) showed that, when learning using distant supervision from a knowledge base, the portion of mis-labeled examples can vary from 13% to 31%. To address this issue, another line of work has explored *structured* learning methods that introduce latent variables. An example is MultiR (Hoffmann et al., 2011), which is based on a joint model of relations between entities in a knowledge base and those mentioned in text. This structured learning approach has a number of advantages; for example, by integrating inference into the learning procedure it has the potential to overcome the challenge of missing facts by ignoring the knowledge base when mention-level classifiers have high confidence (Ritter et al., 2013; Xu et al., 2013). Prior work on structured learning from minimal supervision has leveraged sparse feature representations, however, and has therefore not benefited from learned representations, which have recently achieved state-of-the-art results on a broad range of NLP tasks.

In this paper, we present an approach that combines the benefits of structured and neural methods for minimally supervised relation extraction. Our proposed model learns sentence representations that are computed by a 1D convolutional neural network (Collobert et al., 2011) and are used to define potentials over latent relation mention variables. These mention-level variables are related to observed facts in a KB using a set of deterministic factors, followed by pairwise potentials that encourage agreement between extracted propositions and observed facts, but also enable inference to override these soft constraints during learning, allowing for the possibility of missing information. Because marginal inference is intractable in this model, a MAP-based approach to learning is applied (Taskar et al., 2004).

Our approach is closely related to recent work on Structured Prediction Energy Networks

[1] The code and data are publicly available on Github: https://github.com/bflashcp3f/PCNN-NMAR

(SPENs) (Belanger and McCallum, 2016); the key differences are the application to minimally supervised relation extraction (as opposed to multi-label classification) and the inclusion of latent variables with deterministic factors, which we demonstrate enables effective learning in the presence of missing data in distant supervision. Our proposed method achieves state-of-the-art results on minimally supervised sentential relation extraction, outperforming a number of baselines including one that leverages the attention layer of a purely neural model (Lin et al., 2016).

2 A Latent Variable Model for Neural Relation Extraction

In this section we present our model, which combines continuous representations with structured learning. We first review the problem setting and introduce notation, next we present our approach to extracting feature representations which is based on the piecewise convolutional neural network (PCNN) model of Zeng et. al. (2015) and includes positional embeddings (Collobert et al., 2011). Finally we describe how this can be combined with structured latent variable models that reason about overlapping relations and missing data during learning.

2.1 Assumptions and Problem Formulation

Given a set of sentences, $\mathbf{s} = s_1, s_2 \ldots, s_n$ that mention a pair of knowledge base entities e_1 and e_2 (dyad), our goal is to predict which relation, r, is mentioned between e_1 and e_2 in the context of each sentence, represented by a set of hidden variables, $\mathbf{z} = z_1, z_2, \ldots z_n$. Relations are selected from a fixed set drawn from a knowledge base, in addition to *NA* (no relation). Minimally supervised learning is more difficult than supervised relation extraction, because we do not have direct access to relation labels on the training sentences. Instead, during learning, we are only provided with information about what relations hold between e_1 and e_2 according to the KB. The problem is further complicated by the fact that most KBs are highly incomplete (this is the reason we want to extend them by extracting information from text in the first place), which effectively leads to false-negatives during learning. Furthermore, there are many overlapping relations between dyads, so it is easy for a model trained using minimal supervision from a KB to confuse these relationships. All of these issues are addressed to some degree by the structured learning approach that we present in Section 2.3. First, however we present our approach to feature representation based on convolutional neural networks.

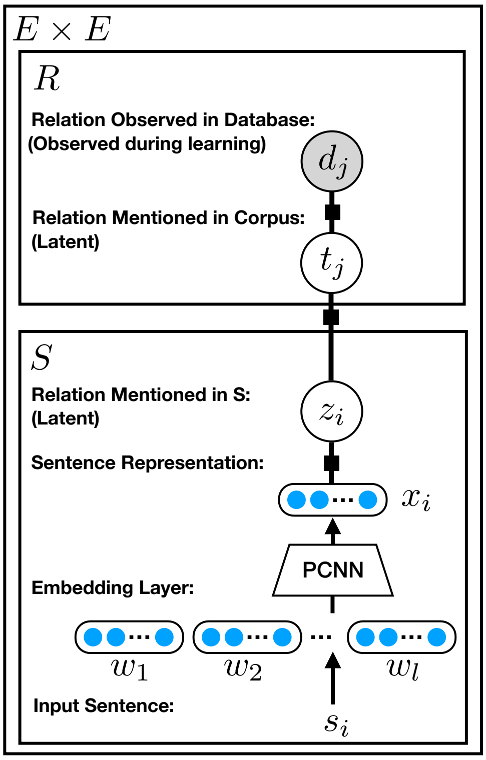

Figure 1: Plate representation of our proposed model. Plates represent replication; $E \times E$ is the number of entity pairs in the dataset, S is the number of sentences mentioning each entity pair and R is the number of relations. Arrows represent functions from input to output. Latent variables are represented as unshaded nodes. Factors over variables are represented as boxes.

2.2 Mention Representation

In the following section we review the Piecewise CNN (PCNN) architecture, first proposed by Zeng et. al. (2015), which is used as the basis for our feature representation.

Input Representation: A sentence, s_i consisting of l words is represented by two types of embeddings: word embeddings, E_i, and position embeddings, P_i relative to the entity pair. Following Lin et. al. (2016), word embeddings were initialized by running Word2Vec on the New York Times corpus and later fine-tuned; position embeddings encode the position of the word relative to KB entities, e_1 and e_2, mentioned in the sentence. The form of input sentence representation is $w_1, w_2, \cdots, w_l$, where $w_i \in \mathbb{R}^d$. The dimension of embedding at each word position is equal to the word embedding dimension plus two times the position embedding size (one position is encoded for each entity).

Convolution: Given an input sentence representation, we perform 1D convolution within a window of length l to extract local features. Assume we have d_f convolutional filters ($F = \{f_1, f_2, \cdots, f_{d_f}\}, f_i \in \mathbb{R}^{l \times d}$). The output of the i-th convolutional filter within the j-th window is:

$$c_{ij} = f_i \cdot w_{j-l+1:j} + b \quad (1 \leq j \leq m + l - 1)$$

Where b is a bias term. We use zero padding when the window slides out of the sentence boundaries.

Piecewise Max Pooling: The output of the convolutional layer c_i is separated into three parts (c_{i1}, c_{i2}, c_{i3}) using the positions of the two entities in the sentence. Max pooling over time is then applied to each of these parts, followed by an elementwise tanh. The final sentence vector is defined as follows:

$$[x]_{ik} = \tanh(\max_j(c_{ikj})) \quad (1 \leq i \leq d_f, 1 \leq k \leq 3)$$

2.3 Structured Minimally Supervised Learning

Our proposed model is based on the PCNN representations described above, in addition to a latent variable model that reasons about missing data and ambiguous relations during learning and is illustrated in Figure 1. The embedding for sentence i, is used to define a factor over the ith input sentence and latent relation mention variable z_i:

$$\phi_{\text{PCNN}}(s_i, z_i) = e^{x_i \cdot \theta_{z_i}}$$

where x_i is the representation for sentence s_i, as encoded by the piecewise CNN.

Another set of factors, ϕ_{OR}, link the sentence-level mention variables, z_i, to aggregate-level variables t_j, representing whether relation j is mentioned between e_1 and e_2 in text. This is modeled using a deterministic OR:

$$\phi_{\text{OR}}(\mathbf{z}, t_j) = \mathbf{1}_{\neg t_j \oplus \exists i: j = z_i}$$

where $\mathbf{1}_x$ is an indicator function that takes the value 1 when x is true. The choice of deterministic OR can be interpreted intuitively as follows: if a proposition is true according to t_j, then it must be extracted from at least one sentence in the training corpus, on the other hand, if it is false, no sentences in the corpus can mention it.

Finally, we incorporate a set of factors that penalize disagreement between observed relations in the KB, d_j, and latent variables t_j, which represent whether relation j was extracted from the text. The penalties for disagreement with the KB are hyperparameters that are adjusted on held-out development data and incorporate entity frequency information from the KB, to model the intuition that more popular entities are less likely to have missing facts:

$$\phi_{\text{A}}(t_j, d_j) = \begin{cases} e^{-\alpha_T}, & \text{if } t_j = 0 \text{ and } d_j = 1 \\ e^{-\alpha_D}, & \text{if } t_j = 1 \text{ and } d_j = 0 \\ 1, & \text{otherwise} \end{cases}$$

Putting everything together, the (unnormalized) joint distribution over $\mathbf{t}$, $\mathbf{d}$ and $\mathbf{z}$ conditioned on sentences $\mathbf{s}$ mentioning a dyad is defined as follows:

$$\begin{aligned} P(\mathbf{d}, \mathbf{t}, \mathbf{z} | \mathbf{s}) &\propto \prod_{i=1}^{|\mathbf{s}|} \phi_{\text{PCNN}}(s_i, z_i) \times \Big(\prod_{j=1}^{|\mathbf{r}|} \phi_{\text{OR}}(\mathbf{z}, t_j) \phi_{\text{A}}(t_j, d_j) \Big)^{\mu} \\ &= \exp(S_\theta(\mathbf{s}, \mathbf{z}, \mathbf{t}, \mathbf{d})) \end{aligned} \quad (1)$$

Here, μ is a tunable hyperparameter to adjust impact of disagreement penalty, and $S_\theta(\cdot)$ is the model score for a joint configuration of variables, which corresponds to the log of the unnormalized probability.

A standard conditional random field (CRF) formulation would optimize model parameters, θ so as to maximize marginal probability of the observed KB relations, $\mathbf{d}$ conditioned on observed sentences, $\mathbf{s}$:

$$P(\mathbf{d}|\mathbf{s}) = \sum_{\mathbf{z}, \mathbf{t}} P(\mathbf{d}, \mathbf{t}, \mathbf{z}|\mathbf{s})$$

Computing gradients with respect to $P(\mathbf{d}|\mathbf{s})$ (and marginalizing out $\mathbf{z}$ and $\mathbf{t}$) is computationally intractable, so instead we propose an approach that uses maximum-a-posteriori (MAP) parameter learning (Taskar et al., 2004) and is inspired by the latent structured SVM (Yu and Joachims, 2009).

Given a large text corpus in which a set of sentences, $\mathbf{s}$ mention a specific pair of entities (e_1, e_2) and a set of relations $\mathbf{d}$ hold between e_1 and e_2, our goal is to minimize the structured hinge loss:

$$L_{\text{SH}}(\theta) =$$

$$\max \left\{ 0, \begin{array}{l} \max\limits_{\mathbf{z}_e^*,\mathbf{t}_e^*,\mathbf{d}_e^*} \left[S_\theta(\mathbf{s}, \mathbf{z}_e^*, \mathbf{t}_e^*, \mathbf{d}_e^*) + l_{\text{Ham}}(\mathbf{d}_e^*, \mathbf{d}) \right] \\ \qquad - \max\limits_{\mathbf{z}_g^*,\mathbf{t}_g^*} \left[S_\theta(\mathbf{s}, \mathbf{z}_g^*, \mathbf{t}_g^*, \mathbf{d}) \right] \end{array} \right\} \tag{2}$$

Where $l_{\text{Ham}}(\mathbf{d}_e^*, \mathbf{d})$ is the Hamming distance between the bit vector corresponding to the set of observed relations holding between (e_1, e_2) in the KB and those predicted by the model. Minimizing $L_{\text{SH}}(\theta)$ can be understood intuitively as adjusting the parameters so that configurations consistent with observed relations in the KB, $\mathbf{d}$, achieve a higher model score than those with a large hamming distance from the observed configuration. $\mathbf{z}_e^*$ corresponds to the most confusing configuration of the sentence-level relation mention variables (i.e. one that has a large score and also a large Hamming loss) and $\mathbf{z}_g^*$ corresponds to the best configuration that is consistent with the observed relations in the KB.

This objective can be minimized using stochastic subgradient descent. Fixing $\mathbf{z}_g^*$ and $\mathbf{z}_e^*$ to their maximum values in Equation 2, subgradients with respect to the parameters can be computed as follows:

$$\nabla_\theta L_{\text{SH}}(\theta) = \begin{cases} \mathbf{0} & \text{if } L_{\text{SH}}(\theta) \leq 0, \\ \nabla_\theta S_\theta(\mathbf{s},\mathbf{z}_e^*,\mathbf{t}_e^*,\mathbf{d}_e^*) & \\ -\nabla_\theta S_\theta(\mathbf{s},\mathbf{z}_g^*,\mathbf{t}_g^*,\mathbf{d}) & \text{otherwise} \end{cases} \tag{3}$$

$$= \begin{cases} \mathbf{0} & \text{if } L_{\text{SH}}(\theta) \leq 0, \\ \sum_i \nabla_\theta \log \phi_{\text{PCNN}}(s_i, z_{e,i}^*) & \\ -\sum_i \nabla_\theta \log \phi_{\text{PCNN}}(s_i, z_{g,i}^*) & \text{otherwise} \end{cases} \tag{4}$$

Because the second factor of the product in Equation 1 does not depend on θ, it is straightforward to compute subgradients of the scoring function, $\nabla S_\theta(\cdot)$, with fixed values of $\mathbf{z}_g^*$ and $\mathbf{z}_e^*$ using backpropagation (Equation 4).

Inference: The two inference problems, corresponding to maximizing over hidden variables in Equation 2 can be solved using a variety of solutions; we experimented with A* search over left-to-right assignments of the hidden variables. An admissible heuristic is used to lower-bound the maximum score of each partial hypothesis by maximizing over the unassigned PCNN factors, ignoring inconsistencies. This approach is guaranteed to find an optimal solution, but can be slow and memory intensive for problems with many variables. In preliminary experiments on development data, we found that local-search (Eisner and Tromble, 2006) using both relation type and mention search operators (Liang et al., 2010; Ritter et al., 2013) usually finds an optimal solution and also scales up to large training datasets; we use local search with 30 random restarts to compute argmax assignments for the hidden variables, $\mathbf{z}_g^*$ and $\mathbf{z}_e^*$, in all our experiments.

Bag-Size Adaptive Learning Rate: Since the search space of the MAP inference problem increases exponentially as the number of hidden variables goes up, it becomes more difficult to find the exact argmax solution using local search, leading to increased noise in the computed gradients. To mitigate the search-error problem in large bags of sentences, we dynamically modify the learning rate based on the number of sentences in each bag as follows:

$$\lambda_i = \begin{cases} \lambda, & \text{if } |\mathbf{s_i}| < \beta_1 \\ \lambda \times \frac{\beta_1}{|\mathbf{s}_i|}, & \text{if } \beta_1 \leq |\mathbf{s}_i| \leq \beta_2 \\ \lambda \times (\frac{\beta_1}{|\mathbf{s}_i|})^2, & \text{otherwise} \end{cases}$$

where λ_i is the learning rate for ith training entity pair and β_1/β_2 are two tunable bag-size thresholds. In Table 3 and Table 4, we see that this strategy significantly improves performance, especially when training on the larger NYTFB-280K dataset. We also experimented with this method for PCNN+ATT, but found that its performance did not improve.

3 Experiments

In Section 2, we presented an approach that combines the benefits of PCNN representations and structured learning with latent variables for minimally supervised relation extraction. In this section we present the details of our evaluation methodology and experimental results.

Datasets: We evaluate our models on the NYT-Freebase dataset (Riedel et al., 2010) which was created by aligning relational facts from Freebase with the New York Times corpus, and has been used in a broad range of prior work on minimally supervised relation extraction. Several versions of this dataset have been used in prior work; to facilitate the reproduction of prior results, we experiment with two versions of the dataset used by Riedel et. al. (2010) (henceforth NYTFB-68K) and Lin et. al. (2016) (NYTFB-280K). Statistics of these datasets are presented in Table 8. A more detailed discussion about the differences between datasets used in prior work is also presented in Appendix B.

Dataset	NYTFB-68K (Riedel et. al. 2010)	NYTFB-280K (Lin et. al. 2016)
Entity pairs	67,946	280,275
Sentences	120,290	523,312

Table 1: Number of entity pairs and sentences in the training portion of Riedel's HELDOUT dataset (NYTFB-68K) and Lin's dataset (NYTFB-280K).

Window length l	3
Number of convolutional filters d_f	230
Word embedding dimension d_w	50
Position embedding dimension d_p	5
Batch size B	1

Table 2: Untuned hyperparameters in our experiments.

Hyperparameters: Following Lin et. al. (2016), we utilize word embeddings pre-trained on the NYT corpus using the word2vec tool, other parameters are initialized using the method described by Glorot and Bengio (2010). The Hoffmann et. al. sentential evaluation dataset is split into a development and test set and grid search on the development set was used to determine optimal values for the learning rate λ among $\{0.001, 0.01\}$, KB disagreement penalty scalar μ among $\{100, 200, \cdots, 2000\}$ and β_1/β_2 bag size threshold for the adaptive learning rate among $\{10, 15, \cdots, 40\}$. Other hyperparameters with fixed values are presented in Table 2.

Neural Baselines: To demonstrate the effectiveness of the our approach, we compare against col-less universal schema (Verga et al., 2016) in addition to the PCNN+ATT model of Lin et. al. (2016). After training the Lin et. al. model to predict observed facts in the KB, we use its attention layer to make mention-level predictions as follows:

$$p(r_j|x_i) = \frac{\exp(r_j \cdot x_i)}{\sum_{k=1}^{n_r} \exp(r_k \cdot x_i)}$$

Where r_j indicates the vector representation of jth relation.

Structured Baselines: In addition to initializing convolutional filters used in the $\phi_{\text{PCNN}}(\cdot)$ factors randomly and performing structured learning of representations as in Equation 4, we also experimented with variants of MultiR and DNMAR, which are based on the structured perceptron (Collins, 2002), using fixed sentence representations: both traditional sparse feature representations, in addition to pre-trained continuous representations generated using our best-performing reimplementation of PCNN+ATT. For the structured perceptron baselines, we also experimented with variants based on MIRA (Crammer and Singer, 2003), which we found to provide consistent improvements. More details are provided in Appendix A.

3.1 Sentential Evaluation

In this work, we are primarily interested in mention-level relation extraction. For our first set of experiments (Tables 3 and 4), we use the manually annotated dataset created by (Hoffmann et al., 2011). Note that sentences in the Hoffman et. al. dataset were selected from the output of systems used in their evaluation, so it is possible there are high confidence predictions made by our systems that are not present. Therefore, we further validate our findings, by performing a manual inspection of the highest confidence predictions in Table 5.

NYTFB-68K **Results:** As illustrated in Table 3, simply applying structured models (MultiR and DNMAR) with pre-trained sentence representations performs competitively. MIRA provides consistent improvements for both sparse and dense representations. PCNN+ATT outperforms most latent-variable models on the sentential evaluation, we found this result to be surprising as the model was designed for extracting proposition-level facts. Col-less universal schema does not perform very well in this evaluation; this is likely due to the fact that it was developed for the KBP slot filling evaluation (Ji et al., 2010), and only uses the part of a sentence between two entities as an input representation, which can remove important context. Our proposed model, which jointly learns sentence representations using a structured latent-variable model that allows for the possiblity of missing data, achieves the best overall performance; its improvements over all baselines were found to be statistically significant according to a paired bootstrap test (Efron and Tibshirani, 1994; Berg-Kirkpatrick et al., 2012).[2]

NYTFB-280K **Results:** When training on the larger dataset provided by Lin et. al. (2016), linguistic features are not available, so only neural representations are included in our evaluation. As illustrated in Table 4, PCNNNMAR also achieves the best performance when training on the larger dataset; its improvements over the baselines are statistically significant. The AUC of most mod-

[2] p-value is less then 0.05.

Model	Name	DEV	TEST
Fixed Sentence Representations	MultiR_sparse (Hoffmann et al., 2011)	66.2	63.2
	MultiR_sparse_MIRA	75.3	71.6
	MultiR_continuous	74.2	68.7
	MultiR_continuous_MIRA	80.3	72.5
	DNMAR_sparse (Ritter et al., 2013)	77.9	70.1
	DNMAR_sparse_MIRA	77.5	72.1
	DNMAR_continuous	80.2	70.0
	DNMAR_continuous_MIRA	82.2	74.2
Jointly Learned Representations	PcnnNmar	82.9	81.0
	PcnnNmar (bag size adaptive learning rate)	**85.5**	**83.1**
Baselines	col-less universal schema (Verga et al., 2016)	63.4	61.1
	PCNN+ATT (Lin et al. (2016) code)	81.4	76.4
	PCNN+ATT (our reimplementation with parameter tuning)	83.5	75.5

Table 3: AUC of sentential evaluation precision / recall curves for all models trained on NytFb-68k. Continuous sentence representation works as well as human-engineered sentence representation, and MIRA consistently helps structured perceptron training. PCNN+ATT performs competitively while our PcnnNmar (AdapLR) is statistically significantly better (p-value of bootstrap is less than 0.05)

els decreases on the Hoffmann et. al. sentential dataset when training on NytFb-280k. This is not surprising, because the Hoffmann et. al. dataset is built by sampling sentences from positive predictions of models trained on NytFb-68k; changing the training data causes a difference in the ranking of high-confidence predictions for each model, leading to the observed decline in performance against the Hoffmann et. al. dataset. To further validate our findings, we also manually inspect the models' top predictions as described below.

Manual Evaluation: Because the Hoffmann et. al. sentential dataset does not contain the highest confidence predictions, we also manually inspected each model's top 500 predictions for the most frequent 4 relations, and report precision @ N to further validate our results. As shown in Table 5, for NytFb-68k, PCNN+ATT performs comparably on `/location/contains`[3] and `/person/company`, whereas our model has a considerable advantage on the other two relations. For NytFb-280k, our model performs consistently better on all four relations compared with PCNN+ATT. When training on the larger NytFb-280k dataset, we observe trend of increasing mention-level P@N for PcnnNmar, however the performance of PCNN+ATT appears to decrease. We investigate this phenomenon further below.

Performance at Extracting New Facts: To explain PCNN+ATT's drop in mention-level performance after training on the larger NytFb-280k dataset, our hypothesis is that the larger KB-supervised dataset not only contains more true positive training examples but also more false negative examples. This biases models toward predicting facts about popular entities, which are likely to exist in Freebase. To provide evidence in support of this hypothesis, we divide the manually annotated dataset from Hoffmann et. al. into two categories: mentions of facts found in Freebase, and those that are not; this distribution is presented in the Table 6. In Table 7, we present a breakdown of model performance on these two subsets. For PCNN+ATT, although the AUC of in-Freebase mentions on the test set increases after training on the larger NytFb-280k, its Out-Of-Freebase AUC on both dev and test sets drops significantly, which clearly illustrates the problem of increasing false negatives during training. In contrast, our model, which explicitly allows for the possibility of missing data in the KB during learning, has relatively stable performance on the two types of mentions, as the amount of weakly-supervised training data is increased.

3.2 Held-Out Evaluation

In Section 3.1, we evaluated the results of minimally supervised approaches to relation extraction by comparing extracted mentions against human judgments. An alternative approach, which has been used in prior work, is to evaluate a model's performance by comparing predictions against held out facts from a KB. Taken in isolation, this approach to evaluation can be misleading, because it penalizes models that extract many new facts that do not already appear in the knowledge base. This is undesirable, because the whole

[3] `/location/contains` is the most frequent relation in the Hoffmann et. al. dataset.

Model	Name	DEV	TEST
Fixed Sentence Representations	MultiR_continuous	72.4	66.7
	MultiR_continuous_MIRA	74.6	73.4
	DNMAR_continuous	73.1	68.0
	DNMAR_continuous_MIRA	75.6	68.7
Jointly Learned Representations	PcnnNmar	78.1	75.4
	PcnnNmar (bag size adaptive learning rate)	**82.9**	**83.1**
Baselines	col-less universal schema (Verga et al., 2016)	60.3	57.5
	PCNN+ATT (Lin et al. (2016) code)	67.9	72.1
	PCNN+ATT (our reimplementation with parameter tuning)	78.2	74.8

Table 4: AUC of sentential evaluation precision / recall curves for all models trained on NytFb-280k. Our proposed PcnnNmar (AdapLR) still performs the best, and the advantage over baselines is also statistically significant (p-value of bootstrap is less than 0.05).

Relation	N	PCNN+ATT	PcnnNmar (AdapLR)
	NytFb-68k		
/location/contains	100	**1.00**	0.99
	500	0.97	**0.98**
/person/place_lived	100	0.76	**0.98**
	500	0.63	**0.78**
/person/nationality	100	0.62	**0.89**
	500	0.43	**0.54**
/person/company	100	0.98	0.98
	500	0.72	**0.78**
	NytFb-280k		
/location/contains	100	0.98	**0.99**
	500	0.82	**0.99**
/person/place_lived	100	0.58	**0.98**
	500	0.57	**0.84**
/person/nationality	100	0.70	**0.91**
	500	0.35	**0.56**
/person/company	100	0.59	**0.95**
	500	0.40	**0.68**

Table 5: Top: P@N of 4 most frequent relations for models trained on NytFb-68k. Bottom: P@N of 4 most frequent relations for models trained on NytFb-280k. Both models can perform well on `/location/contains` relation while PcnnNmar (AdapLR) is consistently better over other relations.

Category	True	False	Total
	DEV		
In-Freebase	102	180	282
Out-Of-Freebase	58	96	154
	TEST		
In-Freebase	113	192	305
Out-Of-Freebase	41	99	140

Table 6: Top: Sentence distribution in Hoffmann et. al. (2011) sentential evaluation DEV dataset. Bottom: Sentence distribution in Hoffmann et. al. (2011) sentential evaluation TEST dataset. There are substantial Out-Of-Freebase mentions which are manually labelled as correct relational mentions.

Model	Dataset	InFB	OutFB
	DEV		
PCNN+ATT	NytFb-68k	78.0	89.6
	NytFb-280k	77.1	77.0
	Change	-0.9	**-12.6**
PcnnNmar(AdapLR)	NytFb-68k	81.4	90.4
	NytFb-280k	77.7	90.6
	Change	**-3.7**	+0.2
	TEST		
PCNN+ATT	NytFb-68k	74.8	75.9
	NytFb-280k	81.9	56.8
	Change	+7.1	**-19.1**
PcnnNmar(AdapLR)	NytFb-68k	81.9	85.4
	NytFb-280k	83.1	81.5
	Change	+1.2	**-3.9**

Table 7: Top: Comparison of AUCs of In-Freebase and Out-Of-Freebase mentions on sentential DEV set for PCNN+ATT and PcnnNmar (AdapLR) with two datasets. Bottom: Comparison of AUCs of In-Freebase and Out-Of-Freebase mentions on sentential TEST set for PCNN+ATT and PcnnNmar (AdapLR) with two datasets. PCNN+ATT has significant drops about Out-Of-Freebase mentions on both sentential DEV and TEST set after training on the larger NytFb-280k which explains why its overall AUC performances go down while PcnnNmar (AdapLR) does not have such problem.

point of an information extraction system is to extract *new* facts that are not already contained in a KB. Furthermore, sentential extraction has the benefit of providing clear provenance for extracted facts, which is crucial in many applications. Having presented these limitations of the held-out evaluation metrics, however, we now present results using this approach to facilitate comparison to prior work.

Figure 2 presents precision-recall curves against held out facts from Freebase comparing PcnnNmar to several baselines and Figure 3 presents results on the larger NytFb-280k dataset. All models perform better according to the held out evaluation metric when training on the larger dataset, which is consistent with our hypothesis, presented at the end of Section 3.1. Our structured model with learned representations, PCNNNMAR (AdapLR), has lower precision when recall is high. This also fits with our hypothesis, as systems that explicitly model missing data will extract many correct facts that do not appear in the KB, resulting in an under-estimate of precision according to this metric.

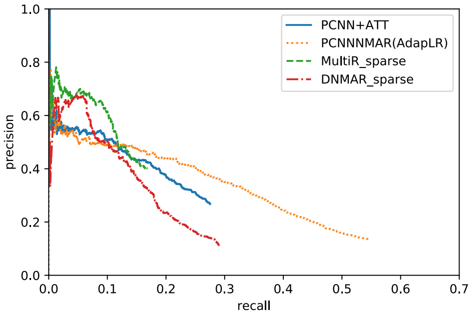

Figure 2: Held-out evaluation precision / recall curves for PCNN+ATT, MultiR, DNMAR and our proposed model PCNNNMAR (AdapLR) on NYTFB-68K.

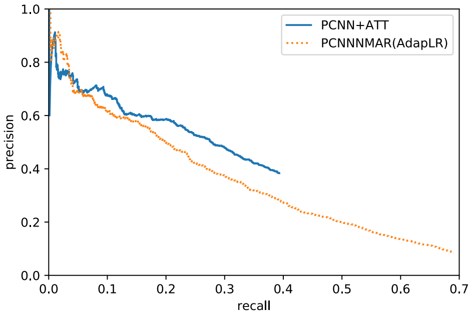

Figure 3: Held-out evaluation precision / recall curves for all NN-based models on NYTFB-280K.

4 Related Work

Knowledge Base Population: There is a long line of prior work on learning to extract relational information from text using minimal supervision. Early work on semantic bootstrapping (Hearst, 1992; Brin, 1998; Agichtein and Gravano, 2000; Carlson et al., 2010; Gupta and Manning, 2014; Qu et al., 2018), applied an iterative procedure to extract lexical patterns and relation instances. These systems tend to suffer from the problem of semantic drift, which motivated work on distant supervision (Craven et al., 1999; Snyder and Barzilay, 2007; Wu and Weld, 2007; Mintz et al., 2009), that explicitly minimizes standard loss functions, against observed facts in a knowledge base. The TAC KBP Knowledge Base Population task was a prominent shared evaluation of relation extraction systems (Ji et al., 2010; Surdeanu, 2013; Surdeanu et al., 2010, 2012). Recent work has explored a variety of new neural network architectures for relation extraction (Wang et al., 2016; Zhang et al., 2017; Yu et al., 2015), experimenting with alternative sentence representations in our framework is an interesting direction for future work. Recent work has also shown improved performance by incorporating supervised training data on the sentence level (Angeli et al., 2014; Beltagy et al., 2018), in contrast our approach does not make use of any sentence-level labels during learning and therefore relies on less human supervision. Finally, prior work has explored a variety of methods to address the issue of noise introduced during distant supervision (Wu et al., 2017; Yaghoobzadeh et al., 2017; Qin et al., 2018).

Another line of work has explored open-domain and unsupervised methods for IE (Yao et al., 2011; Ritter et al., 2012; Stanovsky et al., 2015; Huang et al., 2016; Weber et al., 2017). Universal schemas (Riedel et al., 2013) combine aspects of minimally supervised and unsupervised approaches to knowledge-base completion by applying matrix factorization techniques to multi-relational data (Nickel et al., 2011; Bordes et al., 2013; Chang et al., 2014). Rows of the matrix typically model pairs of entities, and columns represent relations or syntactic patterns (i.e., syntactic dependency paths observed between the entities).

Structured Learning with Neural Representations: Prior work has investigated the combination of structured learning with learned representations for a number of NLP tasks, including parsing (Weiss et al., 2015; Durrett and Klein, 2015; Andor et al., 2016), named entity recognition (Cherry and Guo, 2015; Ma and Hovy, 2016; Lample et al., 2016) and stance detection (Li et al., 2018). We are not aware of any previous work that has explored this direction on the task of minimally supervised relation extraction; we believe structured learning is particularly crucial when learning from minimal supervision to help address the issues of missing data and overlapping relations.

5 Conclusions

In this paper we presented a hybrid approach to minimally supervised relation extraction that combines the benefits of structured learning and learned representations. Extensive experiments show that by performing inference during the learning procedure to address the issue of noise in distant supervision, our proposed model achieves state-of-the art performance on minimally supervised mention-level relation extraction.

Acknowledgments

Funding was provided by the National Science Foundation under Grant No. IIS-1464128, the Defense Advanced Research Projects Agency (DARPA) via the U.S. Army Research Office (ARO) and under Contract Number W911NF-17-C-0095 and the Office of the Director of National Intelligence (ODNI) and Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C0114. The content of the information in this document does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation here on.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D Manning. 2014. Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

David Belanger and Andrew McCallum. 2016. Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992.

Iz Beltagy, Kyle Lo, and Waleed Ammar. 2018. Improving distant supervision with maxpooled attention and sentence-level supervision. *arXiv preprint arXiv:1810.12956*.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

Sergey Brin. 1998. Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases*, pages 172–183. Springer.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3. Atlanta.

K. Chang, W. Yih, B. Yang, and C. Meek. 2014. Typed tensor decomposition of knowledge bases for relation extraction. In *Empirical Methods in Natural Language Processing (EMNLP).*

Colin Cherry and Hongyu Guo. 2015. The unreasonable effectiveness of word representations for twitter named entity recognition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 735–745.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3(Jan):951–991.

Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. *LREC*.

Greg Durrett and Dan Klein. 2015. Neural crf parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 302–312, Beijing, China. Association for Computational Linguistics.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Jason Eisner and Roy W Tromble. 2006. Local search with very large-scale neighborhoods for optimal permutations in machine translation. In *Proceedings of the HLT-NAACL Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*.

Sonal Gupta and Christopher Manning. 2014. Improved pattern learning for bootstrapped entity extraction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 98–108.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.

Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R Voss, Jiawei Han, and Avirup Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010)*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Chang Li, Aldo Porco, and Dan Goldwasser. 2018. Structured representation learning for online debate stance prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Percy Liang, Michael I Jordan, and Dan Klein. 2010. Type-based mcmc. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 573–581. Association for Computational Linguistics.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1064–1074.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

M. Nickel, V. Tresp, and H. Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning (ICML)*, pages 809–816.

Pengda Qin, Weiran XU, and William Yang Wang. 2018. Dsgan: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Meng Qu, Xiang Ren, Yu Zhang, and Jiawei Han. 2018. Weakly-supervised relation extraction by pattern-enhanced embedding learning. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84.

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter.

In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12.

Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics (TACL)*, 1:367–378.

Benjamin Snyder and Regina Barzilay. 2007. Database-text alignment via structured multilabel classification. In *IJCAI*, pages 1713–1718.

Gabriel Stanovsky, Ido Dagan, et al. 2015. Open ie as an intermediate structure for semantic tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Mihai Surdeanu. 2013. Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *TAC*.

Mihai Surdeanu, David McClosky, Julie Tibshirani, John Bauer, Angel X Chang, Valentin I Spitkovsky, and Christopher D Manning. 2010. A simple distant supervision approach for the tac-kbp slot filling task. In *TAC*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 455–465. Association for Computational Linguistics.

Ben Taskar, Carlos Guestrin, and Daphne Koller. 2004. Max-margin markov networks. In *Advances in neural information processing systems*, pages 25–32.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 886–896, San Diego, California. Association for Computational Linguistics.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Noah Weber, Niranjan Balasubramanian, and Nathanael Chambers. 2017. Event representations with tensor-based compositions. *AAAI*.

David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015. Structured training for neural network transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 323–333.

Fei Wu and Daniel S Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50. ACM.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783.

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2.

Yadollah Yaghoobzadeh, Heike Adel, and Hinrich Schütze. 2017. Noise mitigation for neural entity typing and relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 1183–1194.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural svms with latent variables. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Mo Yu, Matthew R Gormley, and Mark Dredze. 2015. Combining word embeddings and feature embeddings for fine-grained relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

A MIRA

Prior work on minimally supervised structured learning has made use of sparse feature representations in combination with perceptron-style parameter updates. We found these updates result in poor performance on held-out development data, however, when using fixed, pre-trained continuous sentence representations. Perhaps this is not surprising because, intuitively, the margin of the dataset is likely to be smaller when using lower dimensional, continuous representations, leading to a larger mistake bound for convergence of the perceptron. To address this, we applied the the **M**argin **I**nfused **R**elaxation **A**lgorithm (Crammer and Singer, 2003), as described below. In Section 3.1, we show empirically that MIRA is crucial for achieving good performance when using continuous representations, and consistently improves performance when using sparse features as well.

As discussed above, we have $\hat{\mathbf{z}}^{\text{KB}}$ the most likely sentence extractions conditioned on the KB and $\hat{\mathbf{z}}$, the MAP assignment to $\mathbf{z}$, ignoring the KB. MIRA updates parameters of the PCNN factors as follows:

$$\theta_j \quad = \quad \theta_j + \tau \cdot \left(F_j(x_i, \hat{z}_i^{\text{KB}}) - F_j(x_i, \hat{z}_i) \right)$$

here τ is an adaptive learning rate that scales the update to the smallest step size that achieves 0 loss on each mention-level classification:

$$\tau \quad = \quad \min\left(C, \frac{1 - \theta \cdot \left(F(x_i, \hat{z}_i^{\text{KB}}) - F(x_i, \hat{z}_i) \right)}{2||x_i||^2} \right)$$

θ is the concatenation of parameters θ_j across relations j, and similarly $F(\cdot)$ is the concatenation of PCNN features across relations. C is a hyperparameter that truncates large steps and helps to prevent overfitting.

B Differing Versions of the NYT-Freebase Corpus Used in Prior Work

We evaluate our models on the NYT-Freebase dataset (Riedel et al., 2010) which was created by aligning relational facts from Freebase with the New York Times corpus, and has been used in a broad range of prior work on minimally supervised relation extraction. Originally, Riedel et. al. created two separate datasets for their HeldOut and Manual evaluations. In the HeldOut dataset, Freebase entity pairs are divided into two parts, one for training and one for testing. Training dyads are aligned to the 2005-2006 portion of the NYT corpus while testing dyads are aligned to the year 2007. In the Manual evaluation data, **all** Freebase entity pairs are matched against the 2005-2006 articles and used as training instances. Testing data in the Riedel et. al. Manual evaluation consists of dyads found within sentences in the 2007 NYT articles, for which at least one entity does not appear in Freebase; their models' predictions on this data were annotated manually. The Riedel et. al. data splits ensure it is not possible to have overlapping train/test entity pairs in either the HeldOut or Manual evaluation.

As neural models with many parameters typically benefit significantly from larger quantities of training data, Lin et. al. (2016) added training data from the Riedel et. al. Manual-Train dataset into their training dataset. This modification of the training data leads to overlap in the entity pairs in the Lin et. al. training/test split. We found 11,424 entity pairs appearing in both training and test sets, however no sentences appear in both the training and test sets, as the matched NYT articles came from different time periods. In all our evaluations we remove these overlapping entity pairs from the training set, to ensure the models are not simply memorizing KB facts that appear in the training data. Figure 4 shows that after removing these shared entity pairs from the training data, performance of the Lin et. al. PCNN+ATT model does not change very much when evaluating against held out facts from Freebase.

We name two versions of the NYT-Freebase dataset according to the number of training entity pairs they include. Table 8 shows that NytFb-280k training set has around 4 times the number of sentences and entity pairs as NytFb-68k, and the proportions of multi-sentence entity pairs in NytFb-280k is higher. In Table 9, we can see that the distribution of relations in the two datasets are comparable, but NytFb-280k has much more entity pairs for each relation. Also, Figure 5 tells us that NytFb-280k has a wider bag-size range and more large training bags.

C Variations on Structured Hinge Loss

Since we use the hinge loss as the loss function in our proposed PcnnNmar model, the way that the hamming loss is calculated decides how we solve

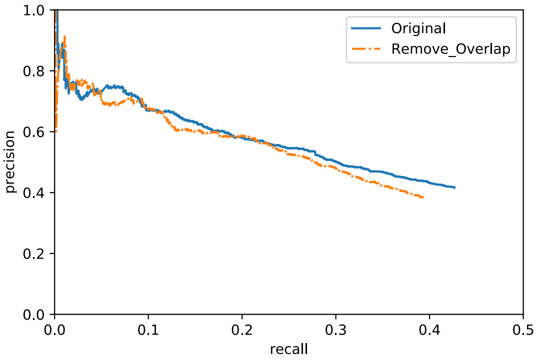

Figure 4: Held-out evaluation precision / recall curves for PCNN+ATT model on original NYTFB-280K and its shared-entity-pairs-removed version.

Dataset	NYTFB-68K	NYTFB-280K
	(Riedel et. al. 2010)	(Lin et. al. 2016)
Entity pairs	67,946	280,275
Sentences	120,290	523,312
Distinct sent.	96,340	340,970
Relations	52	53

Table 8: Number of entity pairs and sentences in the training portion of Riedel's HELDOUT dataset (NYTFB-68K) and Lin's dataset (NYTFB-280K).

Relation	NYTFB-68K		NYTFB-280K	
	# EPs	percent	# EPs	percent
NA	63596	93.12	263372	93.52
/location/contains	2147	3.14	7760	2.76
/person/place_lived	581	0.85	2300	0.86
/person/nationality	436	0.64	2553	0.87
/person/place_of_birth	370	0.54	1400	0.49
/person/company	357	0.52	1417	0.50

Table 9: Distribution of the most frequent relations in the training set of NYTFB-68K and NYTFB-280K.

Method		DEV	TEST
0/1 loss	normal	82.7	79.3
	AdapLR	84.6	80.6
relation-level	normal	83.1	79.7
	AdapLR	85.1	78.5
mention-level	normal	82.9	81.0
	AdapLR	**85.5**	**83.1**

Table 10: AUC of sentential evaluation precision / recall curves for PCNNNMAR with three loss functions trained on NYTFB-68K. Mention-loss hamming loss has obvious advantage over other two loss functions.

the argmax problem in loss-augmented search. In our experiments, we explore three ways to compute the loss: 0/1 loss, relation-level hamming loss and mention-level hamming loss. Table 10 shows that mention-level hamming loss has obvious advantage on AUC performance over other two methods. Although theoretically relation-level hamming loss should be better, it is really hard to find the exact argmax solution in loss-augmented inference with local search while we can easily get it with mention-level hamming loss.

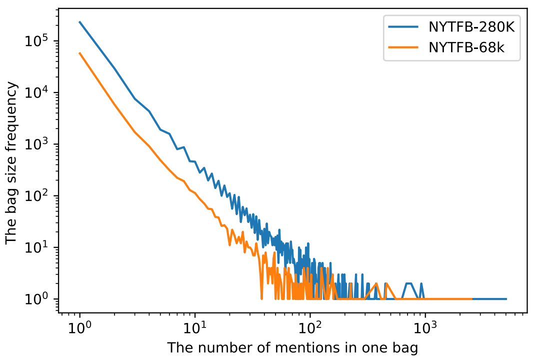

Figure 5: Distribution of bag size in the training set of the NYTFB-68K and NYTFB-280K.