Distantly Supervised Relation Extraction with Sentence Reconstruction and Knowledge Base Priors

Fenia Christopoulou[1], **Makoto Miwa**[2,3], **Sophia Ananiadou**[1]

[1]National Centre for Text Mining,
Department of Computer Science, The University of Manchester, United Kingdom
[2]Toyota Technological Institute, Nagoya, 468-8511, Japan
[3]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Japan
{efstathia.christopoulou, sophia.ananiadou}@manchester.ac.uk
makoto-miwa@toyota-ti.ac.jp

Abstract

We propose a multi-task, probabilistic approach to facilitate distantly supervised relation extraction by bringing closer the representations of sentences that contain the same Knowledge Base pairs. To achieve this, we bias the latent space of sentences via a Variational Autoencoder (VAE) that is trained jointly with a relation classifier. The latent code guides the pair representations and influences sentence reconstruction. Experimental results on two datasets created via distant supervision indicate that multi-task learning results in performance benefits. Additional exploration of employing Knowledge Base priors into the VAE reveals that the sentence space can be shifted towards that of the Knowledge Base, offering interpretability and further improving results[1].

1 Introduction

Distant supervision (DS) is a setting where information from existing, structured knowledge, such as Knowledge Bases (KB), is exploited to automatically annotate raw data. For the task of relation extraction, this setting was popularised by Mintz et al. (2009). Sentences containing a pair of interest were annotated as positive instances of a relation, if and only if the pair was found to share this relation in the KB. However, due to the strictness of this assumption, relaxations were proposed, such as the at-least-one assumption introduced by Riedel et al. (2010): Instead of assuming that all sentences in which a known related pair appears express the relationship, we assume that at least one of these sentences (namely a *bag* of sentences) expresses the relationship. Figure 1 shows example bags for two entity pairs.

[1]Source code is available at https://github.com/fenchri/dsre-vae

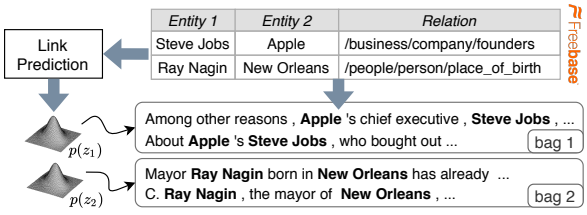


Figure 1: Example of the bag-level setting in distantly supervised relation extraction and the main idea of our approach. Sentences are adapted from the NYT10 dataset (Riedel et al., 2010).

The usefulness of distantly supervised relation extraction (DSRE) is reflected in facilitating automatic data annotation, as well as the usage of such data to train models for KB population (Ji and Grishman, 2011). However, DSRE suffers from noisy instances, long-tail relations and unbalanced bag sizes. Typical noise reduction methods have focused on using attention (Lin et al., 2016; Ye and Ling, 2019) or reinforcement learning (Qin et al., 2018b; Wu et al., 2019). For long-tail relations, relation type hierarchies and entity descriptors have been proposed (She et al., 2018; Zhang et al., 2019; Hu et al., 2019), while the limited bag size is usually tackled through incorporation of external data (Beltagy et al., 2019), information from KBs (Vashishth et al., 2018) or pre-trained language models (Alt et al., 2019). Our goal is not to investigate noise reduction, since it has already been widely addressed. Instead, we aim to propose a more general framework that can be easily combined with existing noise reduction methods or pre-trained language models.

Methods that combine information from Knowledge Bases in the form of pre-trained Knowledge Graph (KG) embeddings have been particularly effective in DSRE. This is expected since they capture broad associations between entities,

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 11–26
June 6–11, 2021.

thus assisting the detection of facts. Existing approaches either encourage explicit agreement between sentence- and KB-level classification decisions (Weston et al., 2013; Xu and Barbosa, 2019), minimise the distance between KB pairs and sentence embeddings (Wang et al., 2018) or directly incorporate KB embeddings into the training process in the form of attention queries (Han et al., 2018; She et al., 2018; Hu et al., 2019). Although these signals are beneficial, direct usage of KB embeddings into the model often requires explicit KB representations of entities and relations, leading to poor generalisation to unseen examples. In addition, forcing decisions between KB and text to be the same makes the connection between context-agnostic (from the KB) and context-aware (from sentences) pairs rigid, as they often express different things.

Variational Autoencoders (VAEs) (Kingma and Welling, 2013) are latent variable encoder-decoder models that parameterise posterior distributions using neural networks. As such, they learn an effective latent space which can be easily manipulated. Sentence reconstruction via encoder-decoder networks helps sentence expressivity by learning semantic or syntactic similarities in the sentence space. On the other hand, signals from a KB can assist detection of factual relations. We aim to combine these two using a VAE together with a bag-level relation classifier. We then either force each sentence's latent code to be close to the Normal distribution (Bowman et al., 2016), or to a prior distribution obtained from KB embeddings. This latent code is employed into sentence representations for classification and is responsible for sentence reconstruction. As it is influenced by the prior we essentially inject signals from the KB to the target task. In addition, sentence reconstruction learns to preserve elements that are useful for the bag relation. To the best of our knowledge, this is the first attempt to combine a VAE with a bag-level classifier for DSRE.

Finally, there are methods for DSRE that follow a rather flawed evaluation setting, where several test pairs are included in the training set. Under this setting, the generalisability of such methods can be exaggerated. We test these approaches under data without overlaps and find that their performance is severely deprecated. With this comparison, we aim to promote evaluation on the amended version of existing DSRE data that can prevent memorisation of test pair relations. Our contributions are threefold:

- Propose a multi-task learning setting for DSRE. Our results suggest that combination of both bag classification and bag reconstruction improves the target task.
- Propose a probabilistic model to make the space of sentence representations resemble that of a KB, promoting interpretability.
- Compare existing approaches on data without train-test pair overlaps to enforce fairer comparison between models.

2 Proposed Approach

2.1 Task Description

In DSRE, the bag setting is typically adopted. A model's input is a pair of named entities e_1, e_2 (mapped to a Knowledge Base), and a bag of sentences $B = \{s_1, s_2, \ldots, s_n\}$, where the pair occurs, retrieved from a raw corpus. The goal of the task is to identify the relation(s), from a predefined set R, that the two entities share, based on the sentences in the bag B. Since each pair can share multiple relations at the same time, the task is considered a multi-label classification problem.

2.2 Overall Framework

Our proposed approach is illustrated in Figure 2. The main goal is to create a joint learning setting where a bag of sentences is encoded and reconstructed and, at the same time, the bag representation is used to predict relation(s) shared between two given entities. The architecture receives as input a bag of sentences for a given pair and outputs (i) predicted relations for the pair and (ii) the reconstructed sentences in the bag. The two outputs are produced by two branches: the left branch, corresponding to bag classification and the right branch, corresponding to bag reconstruction. Both branches start from a shared encoder and they communicate via the latent code of a VAE that is responsible for the information used in the representation and reconstruction of each sentence in the bag. Naturally, both branches have an effect on one another during training.

2.3 Bag Reconstruction

Autoencoders (Rumelhart et al., 1986) are encoder-decoder neural networks that are trained in an unsupervised manner, i.e., to reconstruct their input

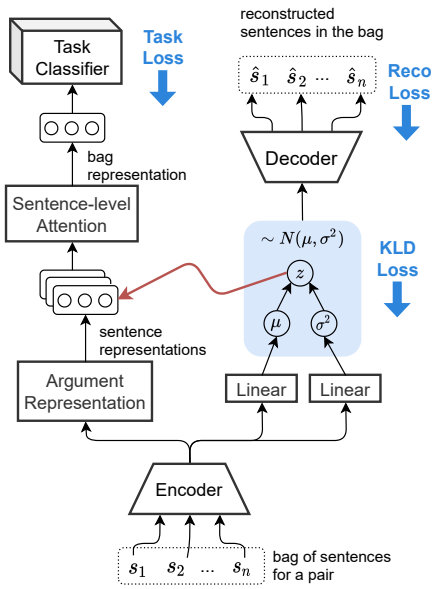


Figure 2: Schematic of the model architecture.

(e.g. a sentence). They learn an informative representation of the input into a dense and smaller feature vector, namely the latent code. This intermediate representation is then used to fully reconstruct the original input. Variational Autoencoders (VAE) (Kingma and Welling, 2013) offer better generalisation capabilities compared to the former by sampling the features of the latent code from a prior distribution that we assume to be similar to the distribution of the data.

2.3.1 Encoder

We form the input of the network similarly to previous work. Each sentence in the input bag is transformed into a sequence of vectors. Words and positions are mapped into real-valued vectors via word embedding $\mathbf{E}^{(w)}$ and position embedding layers $\mathbf{E}^{(p)}$, similarly to Lin et al. (2016). The concatenation of word ($\mathbf{w}$) and position ($\mathbf{p}$) embeddings $\mathbf{x}_t = [\mathbf{w}_t; \mathbf{p}_t^{(e_1)}; \mathbf{p}_t^{(e_2)}]$ forms the representation of each word in the input sentence. A Bidirectional Long-Short Term Memory (BiLSTM) network (Hochreiter and Schmidhuber, 1997) acts as the encoder, producing contextualised representations for each word.

The representations of the left-to-right and right-to-left passes of the BiLSTM are summed to produce the output representation of each word t, $\mathbf{o}_t = \overrightarrow{\mathbf{o}_t} + \overleftarrow{\mathbf{o}_t}$, as well as the representations of the last hidden $\mathbf{h} = \overrightarrow{\mathbf{h}} + \overleftarrow{\mathbf{h}}$ and cell states $\mathbf{c} = \overrightarrow{\mathbf{c}} + \overleftarrow{\mathbf{c}}$ of the input sentence. We use the last hidden and cell states of each sentence s to construct the parameters of a posterior distribution $q_\phi(\mathbf{z}|\mathbf{s})$ using two linear layers,

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{W}_\mu[\mathbf{h}; \mathbf{c}] + \mathbf{b}_\mu, \\ \boldsymbol{\sigma}^2 &= \mathbf{W}_\sigma[\mathbf{h}; \mathbf{c}] + \mathbf{b}_\sigma, \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are the parameters of a multivariate Gaussian, representing the feature space of the sentence. This distribution is approximated via a latent code $\mathbf{z}$, using the reparameterisation trick (Kingma and Welling, 2013) to enable back-propagation, as follows:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

This trick essentially forms the posterior as a function of the normal distribution.

2.3.2 Decoder

The decoder network is a uni-directional LSTM network, that reconstructs each sentence in the input bag. The input is formed in two steps. Firstly, the latent code $\mathbf{z}$ is given as the initial hidden state of the decoder $\mathbf{h}'_0$ via a linear layer transformation. Secondly, the same latent code is concatenated with the representation of each word $\mathbf{w}_t$ in the input sequence of the decoder.

$$\mathbf{h}'_0 = \mathbf{W}\mathbf{z} + \mathbf{b}, \quad \mathbf{x}'_t = [\mathbf{w}_t; \mathbf{z}], \tag{3}$$

A percentage of words in the decoder's input is randomly replaced by the UNK word to force the decoder to rely on the latent code for word prediction, similar to Bowman et al. (2016).

2.3.3 Learning

The optimisation objective of the VAE, namely Evidence Lower BOund (ELBO), is the combination of two losses. The first is the reconstruction loss that corresponds to the cross entropy between the actual sentence s and its reconstruction $\hat{s}$. The second is the Kullback-Leibler divergence (D_{KL}) between a prior distribution $p_\theta(\mathbf{z})$, which the latent code is assumed to follow, and the posterior $q_\phi(\mathbf{z}|\mathbf{h})$, which the decoder produces,

$$\begin{aligned} L_{\text{ELBO}} = \mathbb{E}_{z \sim q_\phi(z|h)} &\left[\log(p_\theta(\mathbf{h}|\mathbf{z}))\right] \\ &- D_{\text{KL}}\left(q_\phi(\mathbf{z}|\mathbf{h})||p_\theta(\mathbf{z})\right) \end{aligned} \tag{4}$$

The first loss is responsible for the accurate reconstruction of each word in the input, while the second acts as a regularisation term that encourages the posterior of each sentence to be close to

the prior. Typically, an additional parameter β is introduced in front of the D_{KL} to overcome KL vanishing, a phenomenon where the posterior collapses to the prior and the VAE essentially behaves as a standard autoencoder (Bowman et al., 2016).

2.4 Bag Classification

Moving on to the left branch of Figure 2, in order to represent a bag we first need to represent each sentence inside it. We realise this using information produced by the VAE as follows.

2.4.1 Sentence Representation

Given the contextualised output of the encoder $\mathbf{o}$, we construct entity representations $\mathbf{e}_1$ and $\mathbf{e}_2$ for a given pair in a sentence by averaging the word representations included in each entity. A sentence representation $\mathbf{s}$ is formed as follows:

$$\mathbf{e}_i = \frac{1}{|e_i|} \sum_{k \in e_i} \mathbf{o}_k, \quad \mathbf{s} = \mathbf{W}_v[\mathbf{z}; \mathbf{e}_1; \mathbf{e}_2], \tag{5}$$

where $|e_i|$ corresponds to the number of words inside the mention span of entity e_i and $\mathbf{z}$ is the latent code of the sentence that was produced by the VAE, as described in Equation (2).

2.4.2 Bag Representation

In order to form a unified bag representation B for a pair, we adopt the popular selective attention approach introduced by Lin et al. (2016). In particular, we first map relations into real-valued vectors, via a relation embedding layer $\mathbf{E}^{(r)}$. Each relation embedding is then used as a query over the sentences in the bag, resulting in $|R|$ bag representations for each pair,

$$a_r^{(s_i)} = \frac{\exp\left(\mathbf{s}_i^\top \mathbf{r}\right)}{\sum_{j \in B} \exp\left(\mathbf{s}_j^\top \mathbf{r}\right)}, \quad \mathbf{B}_r = \sum_{i=1}^{|B|} a_r^{(s_i)} \mathbf{s}_i, \tag{6}$$

where $\mathbf{r}$ is the embedding associated with relation r, $\mathbf{s}_i$ is the representation of sentence $s_i \in B$, $a_r^{(s_i)}$ is the weight of sentence s_i with relation r and $\mathbf{B}_r$ is the final bag representation for relation r.

During classification, we select the probability of predicting a relation category r, using the bag representation that was constructed when the respective relation embedding $\mathbf{r}$ was the query. Binary cross entropy loss is applied on the resulting predictions,

$$\begin{aligned} p(r = 1|B) &= \sigma(\mathbf{W}_c\, \mathbf{B}_r + \mathbf{b}_c), \\ L_{\text{BCE}} &= -\sum_r y_r \log p(r|B) \\ &\quad + (1 - y_r) \log(1 - p(r|B)), \end{aligned} \tag{7}$$

where $\mathbf{W}_c$ and $\mathbf{b}_c$ are learned parameters of the classifier, σ is the sigmoid activation function, $p(r|B)$ is the probability associated with relation r given a bag B and y_r is the ground truth for this relation with possible values 1 or 0.

2.5 Knowledge Base Priors

In the scenario where no KB information is incorporated into the model, we simply assume that the prior distribution of the latent code $p_\theta(\mathbf{z})$ is a standard Gaussian with zero mean and identity covariance $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

To integrate information about the nature of triples into the bag-level classifier, we create KB-guided priors as an alternative to the standard Gaussian. In particular, we train a link prediction model, such as TransE (Bordes et al., 2013), on a subset of the Knowledge Graph that was used to originally create the dataset. Using the link prediction model, we obtain entity embeddings for the subset KB. A KB-guided prior can thus be constructed for each pair, as another Gaussian distribution with mean value equal to the KB pair representation and covariance as the identity matrix,

$$p_\theta(\mathbf{z}) \sim \mathcal{N}(\boldsymbol{\mu}_{\text{KB}}, \mathbf{I}), \text{ with } \boldsymbol{\mu}_{\text{KB}} = \mathbf{e}_h - \mathbf{e}_t, \tag{8}$$

where $\mathbf{e}_h$ and $\mathbf{e}_t$ are the vectors for entities e_{head} and e_{tail} as resulted from training a link prediction algorithm on a KB.

The link prediction algorithm is trained to make representations of pairs expressing the same relations to be close in space. Hence, by using KB priors we try to force the distribution of sentences in a bag to follow the distribution of the pair in the KB. If one of the pair entities does not exist in the KB subset, the mean vector of the pair's prior will be zero, resulting in a standard Gaussian prior. Finally, KB priors are only used during training. Consequently, the model does not use any direct KB information during inference.

2.6 Training Objective

We train jointly bag classification and sentence reconstruction. The final optimisation objective

is formed as,

$$L = \lambda\, L_{\text{BCE}} + (1 - \lambda) L_{\text{ELBO}}, \qquad (9)$$

where λ corresponds to a weight in $[0, 1]$. We weigh the classification loss more than the ELBO to allow the model to better fit the target task.

3 Experimental Settings

3.1 Datasets

We experiment with the following two datasets:

NYT10. The widely used New York Times dataset (Riedel et al., 2010) contains 53 relation categories including a negative relation (NA) indicating no relation between two entities. We use the version of the data provided by the OpenNRE framework (Han et al., 2019), which removes overlapping pairs between train and test data. The dataset statistics are shown in Table 1. Additional information can be found in Appendix A.1.

For the choice of the Knowledge Base, we use a subset of Freebase[2] that includes 3 million entities with the most connections, similar to Xu and Barbosa (2019). For all pairs appearing in the test set of NYT10 (both positive and negative), we remove all links in the subset of Freebase to ensure that we will not memorise any relations between them (Weston et al., 2013). The resulting KB contains approximately 24 million triples.

WIKIDISTANT. The WikiDistant dataset is almost double the size of the NYT10 and contains 454 target relation categories, including the negative relation. It was recently introduced by Han et al. (2020) as a cleaner and more well structured bag-level dataset compared to NYT10, with fewer negative instances.

For the Knowledge Base, we use the version of Wikidata[3] provided by Wang et al. (2019b) (in particular the transductive split[4]), containing approximately 5 million entities. Similarly to Freebase, we remove all links between pairs in the test set from the resulting KB, which contains approximately 20 million triples after pruning.

3.2 Evaluation Metrics

Following prior work, we consider the Precision-Recall Area Under the Curve (AUC) as the primary metric for both datasets. We additionally report Precision at N (P@N), that measures the percentage of correct classifications for the top N most confident predictions.

Dataset	Split	Instances	Bags	NA (%)
NYT10 # Relations: 53	Train	469,290	252,044	93.4
	Val.	53,321	28,109	93.5
	Test	172,448	96,678	97.9
WIKIDISTANT # Relations: 454	Train	1,050,246	575,620	64.8
	Val.	29,145	14,748	70.6
	Test	28,897	15,509	72.0

Table 1: Datasets statistics. 'NA' correponds to the 'no relation' category.

3.3 Training

To obtain the KB priors, we train TransE on the subsets of Freebase and Wikidata using the implementation of the DGL-KE toolkit (Zheng et al., 2020) for 500K steps and a dimensionality equal to the dimension of the latent code. The main model was implemented with PyTorch (Paszke et al., 2019). We use the Adam (Kingma and Ba, 2014) optimiser with learning rate 0.001. KL logistic annealing is incorporated only in the case where the prior is the Normal distribution to avoid KL vanishing (Bowman et al., 2016). Early stopping is used to determine the best epoch based on the AUC score on the validation set. Words in the vocabulary are initialised with pre-trained, 50-dimensional GloVe embeddings (Pennington et al., 2014).

We limit the vocabulary size to the top 40K and 50K most frequent words for NYT10 and WIKIDISTANT, respectively. To enable fast training, we use Adaptive Softmax (Grave et al., 2017). The maximum sentence length is restricted to 50 for NYT10 and 30 words for WIKIDISTANT. Each bag in the training set is allowed to contain maximum 500 sentences selected randomly. For prediction on the validation and test sets, all sentences (with full length) are used.

3.4 Baselines

In this work we compare with various models applied on the NYT10 dataset: **PCNN-ATT** (Lin et al., 2016) is one of the first neural models that uses a PCNN encoder and selective attention over the instances in a bag, similar to our approach. **RESIDE** (Vashishth et al., 2018), utilises syntactic, entity and relation type information as additional input to the network to assist classification. **JOINT**

[2] https://developers.google.com/freebase
[3] https://www.wikidata.org/
[4] https://deepgraphlearning.github.io/project/wikidata5m

Method	Encoder	NYT 520K				NYT 570K			
		AUC (%)	P@N (%)			AUC (%)	P@N (%)		
			100	200	300		100	200	300
Baseline	BiLSTM	34.94	74.0	67.5	67.0	43.59	84.0	77.0	75.3
+ $p_\theta(z) \sim \mathcal{N}(0, I)$		38.59	74.0	74.5	71.6	44.64	80.0	76.0	75.6
+ $p_\theta(z) \sim \mathcal{N}(\mu_{\text{KB}}, I)$		42.89	83.0	75.5	73.0	45.52	81.0	77.5	73.6
PCNN-ATT (Lin et al., 2016)	PCNN	32.66	71.0	67.5	62.6	36.25	76.0	72.5	64.0
JOINT NRE (Han et al., 2018)	CNN	30.62	60.0	57.0	55.3	40.15	75.8	-	68.0
RESIDE (Vashishth et al., 2018)	BiGRU	35.80	80.0	69.0	65.3	41.60	84.0	78.5	75.6
INTRA-INTER BAG (Ye and Ling, 2019)	PCNN	34.41	82.0	74.0	69.0	42.20	91.8	84.0	78.7
DISTRE (Alt et al., 2019)	GPT-2	42.20	68.0	67.0	65.3	-	-	-	-

Table 2: Performance comparison between different methods on the NYT10 test set for the two different versions of the dataset. Results in the 520K column are re-runs of existing implementations, except for DISTRE. Results on the 570K column are taken from the respective publications.

Method	AUC (%)	P@N (%)		
		100	200	300
Baseline	28.54	94.0	93.0	88.3
+ $p_\theta(z) \sim \mathcal{N}(0, I)$	30.59	96.0	93.5	89.3
+ $p_\theta(z) \sim \mathcal{N}(\mu_{\text{KB}}, I)$	29.54	92.0	89.0	90.0
PCNN-ATT (Han et al., 2020)	22.20	-	-	-
w/o non KB-prior pairs (72% of training pairs preserved)				
Baseline	26.16	88.0	85.0	82.6
+ $p_\theta(z) \sim \mathcal{N}(0, I)$	27.46	90.0	88.0	84.6
+ $p_\theta(z) \sim \mathcal{N}(\mu_{\text{KB}}, I)$	28.38	94.0	95.0	89.3

Table 3: Performance comparison on the WIKIDISTANT test set.

NRE (Han et al., 2018) jointly trains a textual relation extraction component and a link prediction component by sharing attention query vectors among the two. **INTRA-INTER BAG** (Ye and Ling, 2019) applies two attention mechanisms inside and across bags to enforce similarity between bags that share the same relations. **DISTRE** (Alt et al., 2019) uses a pre-trained Transformer model, instead of a recurrent or convolutional encoder, fine-tuned on the NYT10 dataset.

We report results on both the filtered data (520K) that do not contain train-test pair overlaps, as well as the non-filtered version (570K) to better compare with prior work[5]. With the exception of DISTRE, all prior approaches were originally applied on the 570K version. Hence, performance of prior work on the 520K version corresponds to re-runs of existing implementations (via their open-source code). For the non-filtered version, results are taken from the respective publications[6].

For the WIKIDISTANT dataset, we compare with the **PCNN-ATT** model as this is the only model currently applied on this data (Han et al., 2020). We also compare our proposed approach with two additional baselines. The first baseline model (Baseline) does not use the VAE component at all. In this case the sentence representation is simply created using the last hidden state of the encoder, $\mathbf{s} = [\mathbf{h}; \mathbf{e}_1; \mathbf{e}_2]$, instead of the latent code. The second model ($p_\theta(z) \sim \mathcal{N}(0, I)$) incorporates reconstruction with a standard Gaussian prior and the final model ($p_\theta(z) \sim \mathcal{N}(\mu_{\text{KB}}, I)$) corresponds to our proposed model with KB priors.

4 Results

The results of the proposed approach versus existing methods on the NYT10 dataset are shown in Table 2. The addition of reconstruction further improves performance by 3.6 percentage points (pp), while KB priors offer an additional of 4.3pp. Compared with DISTRE, our model achieves comparable performance, even if it does not use a pre-trained language model. As we observe from the precision-recall curve in Figure 3, our model is competitive with DISTRE for up to 35% of the recall range but for the tail of the distribution a pre-trained language model has better results. This can be attributed to the world knowledge it has obtained via pre-training, which is much more vast than a KB subset. Overall, for the reduced version of the dataset VAE with KB-guided priors surpasses the entire recall range of all previous methods. For the 570K version, our model is superior to other approaches in terms of AUC score, even for the baseline. We speculate this is because we incorpo-

[5]More information about the two versions can be found in Appendix A.1

[6]For PCNN-ATT we re-run both the 520K and the 570K versions using the OpenNRE toolkit.

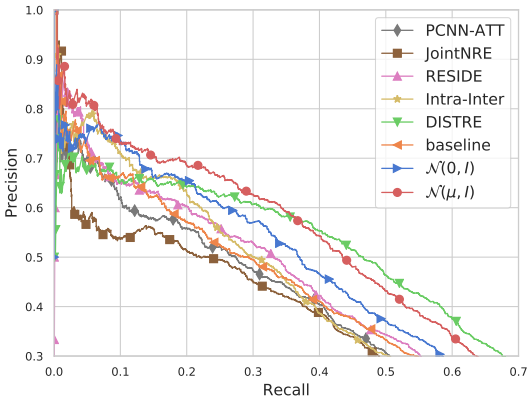


Figure 3: Precision-Recall curves for the NYT10 (520K version) test set.

rate argument representations into the bag representation. As a result, overlapping pairs between training and test set have learnt strong argument representations.

Regarding the results on the WIKIDISTANT dataset in Table 3, once again we observe that reconstruction helps improve performance. However, it appears that KB priors have a negative effect. We find that in the NYT10 dataset 96% of the training pairs are associated with a prior. Instead, this portion is only 72% for WIKIDISTANT. The reason for this discrepancy could be the reduced coverage that potentially causes a confusion between the two signals[7]. To test this hypothesis, we re-run our models on a subset of the training data, removing pairs that do not have a KB prior. As observed in the second half of Table 3, priors do seem to have a positive impact under this setting, indicating the importance of high coverage in prior-associated pairs. We use this setting for the remainder of the paper.

5 Analysis

We then check whether the latent space has indeed learned some information about the KB triples, by visualising the t-SNE plots of the priors, i.e. the $\boldsymbol{\mu}_{\text{KB}}$ vectors as resulted from training TransE (Equation (8)) and the posteriors, i.e. the $\boldsymbol{\mu}$ vectors as resulted from the VAE encoder (Equation (1)).

Figure 4a illustrates the space of the priors in Freebase for the most frequent relation categories in the NYT10 training set [8]. As it can be observed, the separation is obvious for most categories, with a few overlaps. Relations *place of birth*, *place lived* and *place of death* appear to reside in the same region. This is expected as these relations can be shared by a pair simultaneously. Another overlap is identified for *contains*, *administrative divisions* and *capital*. Again, these are similar relations found between certain entity types (e.g. location, province, city). Figure 4b shows the t-SNE plot for a collection of latent vectors (random selection of 2 sentences in a positive bag). The space is very similar to that of the KB and the same overlapping regions are clearly observed. A difference is that it appears to be less compact, as not all sentences in a bag express the exact same relation.

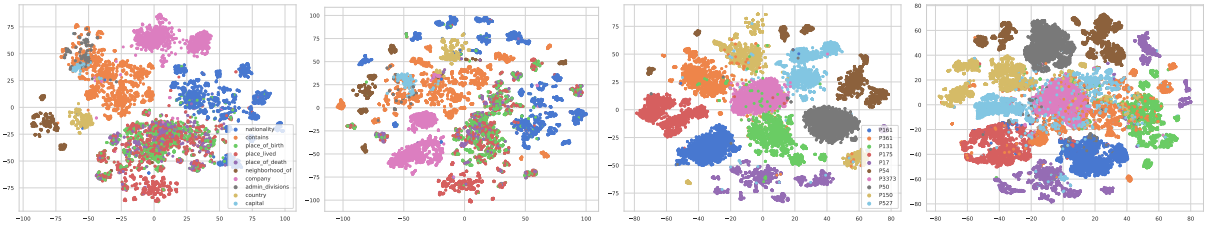
Similar observations stand for Wikidata priors, as shown in Figure 4c. By looking at the space of the posteriors, we can see that although for most categories separation is achieved, there are 2 relations that are not so well separated in the posterior space. We find that *has part* (cyan) and *part of* (orange) are opposite relations, that TransE can effectively learn thanks to its properties. However, the model appears to not be able to fully separate the two. These relations are expressed in the same manner, by only changing the order of the arguments. As there is no restriction regarding the argument order in our model directionality can sometimes be an issue.

Finally, in order to check how the prior constraints affect sentence reconstruction, we illustrate reconstructions of sentences in the validation set of the NYT10 in Table 4 and WIKIDISTANT in Table 5. In detail, we give the input sentence to the network and employ greedy decoding using either the mean of the latent code or a random sample.

Manual inspection of reconstruction reveals that KB-priors generate longer sentences than the Normal prior by repeating several words (especially the UNK). In fact, VAE with KB-priors fails to generate plausible and grammatical examples for NYT10, as shown in Table 4. Instead, reconstructions for WIKIDISTANT are slightly better, due to the less noisy nature of the dataset. In both cases, we see that the reconstructions contain words that are useful for the target relation, e.g. words that refer to places such as *new york*, *new jersey* for the relation *contains* between *bay village* and *ohio*, or sport-related terms (football, team, league) for the *statistical leader* relationship between *wayne rooney* and *england national team*.

[7] If a pair does not have a KB prior it will be assigned the Normal prior instead.

[8] We plot t-SNEs for the training set instead of the validation/test sets because the WIKIDISTANT validation set contains too few pairs belonging to the top-10 categories. NYT10 validation set t-SNE can be found in the Appendix A.5



(a) Prior Freebase (b) Posterior Freebase (c) Prior Wikidata (d) Posterior Wikidata

Figure 4: T-SNE plots of: (a), (c) pair representations obtained from a TransE model (priors) on a subset of Freebase and Wikidata for the 10 most frequent classes in each dataset, (b), (d) the latent codes (μ) for sentences of each training set, when using KB priors.

	INPUT	she graduated from _ *college* in *new concord* , ohio	growing up in *bay village* , *ohio* , steinbrenner haunted the county fairs , riding in pony races .
$\mathcal{N}(\mathbf{0}, \mathbf{I})$	MEAN	he graduated from the university of california and received a master 's degree in education .	he was born in _ , england , and grew up in the united states
	SAMPLE	he graduated from the university of california and received a master 's degree in education .	he was born in # , and then moved to new york
$\mathcal{N}(\boldsymbol{\mu}_{\text{KB}}, \mathbf{I})$	MEAN	the bridegroom , # , is a professor of the university of california at berkeley , and a professor of english ...	the _ , which is based in new york , and the _ ...
	SAMPLE	the _ , a _ of the university of california , berkeley , and the author of " the _ of the world " ...	the _ , which is based in new jersey , and the _ ...

Table 4: Sentence reconstruction examples from the NYT10 validation set, using different priors. _ corresponds to the UNK word and # indicates a number.

	INPUT	*wayne rooney* plays as a striker for manchester united and the *england national team*	ng 's first role was in the # *michael hui* comedy film " *the private eyes* " .
$\mathcal{N}(\mathbf{0}, \mathbf{I})$	MEAN	_ 's first game was the first time in the game against the new york yankees .	the film was adapted into the # film ' the _ ' , directed by _ .
	SAMPLE	he made his debut for the club in the # fa cup final against arsenal at wembley stadium .	in # , he appeared in ' the _ ' , a # film adaptation of the same name by _ .
$\mathcal{N}(\boldsymbol{\mu}_{\text{KB}}, \mathbf{I})$	MEAN	he was a member of the club 's first team , and was a member of the club 's _ club	_ 's first film was ' the _ ' , starring _ and starring _ .
	SAMPLE	he made his debut in the russian professional football league for fc _ ...	_ , who was the first female actress to win the academy award for best actress .

Table 5: Sentence reconstruction examples from the WIKIDISTANT validation set using different priors. _ corresponds to the UNK word and # indicates a number.

6 Related Work

Distantly Supervised RE. Methods developed for DSRE have been around for a long time, building upon the idea of distant supervision (Mintz et al., 2009) with the widely used NYT10 corpus by Riedel et al. (2010). Methods investigating this problem can be divided into several categories. Initial approaches were mostly graphical models, adopted to perform multi-instance learning (Riedel et al., 2010), sentential evaluation (Hoffmann et al., 2011; Bai and Ritter, 2019) or multi-instance learning and multi-label classification (Surdeanu et al., 2012). Subsequent approaches utilised neural models, with the approach of Zeng et al. (2015) introducing Piecewise Convolutional Neural Networks (PCNN) into the task. Later approaches focused on noise reduction via selection of informative instances using either soft constraints, i.e., attention mechanisms (Lin et al., 2016; Ye and Ling, 2019; Yuan et al., 2019), or hard constraints by explicitly selecting non-noisy instances with reinforcement (Feng et al., 2018; Qin et al., 2018b,a; Wu et al., 2019; Yang et al., 2019) and curriculum learning (Huang and Du, 2019). Noise at the word level was addressed in Liu et al. (2018a) via sub-tree parsing on sentences. Adversarial training has been shown to improve DSRE in Wu et al. (2017), while additional unlabelled examples were exploited to assist classification with Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) in Li et al. (2019). Recent methods use additional information from external resources such as entity types and relations (Vashishth et al., 2018), entity

descriptors (Ji et al., 2017; She et al., 2018; Hu et al., 2019) or Knowledge Bases (Weston et al., 2013; Xu and Barbosa, 2019; Li et al., 2020b).

Sequence-to-Sequence Methods. Autoencoders and variational autoencoders have been investigated lately for relation extraction, primarily for detection of relations between entity mentions in sentences. Marcheggiani and Titov (2016) proposed discrete-state VAEs for link prediction, reconstructing one of the two entities of a pair at a time. Ma et al. (2019) investigated conditional VAEs for sentence-level relation extraction, showing that they can generate relation-specific sentences. Our overall approach shares similarities with this work since we also use VAEs for RE, though in a bag rather than a sentence-level setting. VAEs have also been investigated for RE in the biomedical domain (Zhang and Lu, 2019), where additional non-labelled examples were incorporated to assist classification. This work also has commonalities with our work but the major difference is that the former uses two different encoders while we use only one, shared among bag classification and bag reconstruction. Other SEQ2SEQ methods treat RE as a sequence generation task. Encoder-decoder networks were proposed for joint extraction of entities and relations (Trisedya et al., 2019; Nayak and Ng, 2020), generation of triples from sequences (Liu et al., 2018b) or generation of sequences from triples (Trisedya et al., 2018; Zhu et al., 2019).

VAE Priors. Different types of prior distributions have been proposed for VAEs, such as the VampPrior (Tomczak and Welling, 2018), Gaussian mixture priors (Dilokthanakul et al., 2016), Learned Accept/Reject Sampling (LARs) priors (Bauer and Mnih, 2019), non-parametric priors (Goyal et al., 2017) and others. User-specific priors have been used in collaborative filtering for item recommendation (Karamanolakis et al., 2018), while topic-guided priors were employed for generation of topic-specific sentences (Wang et al., 2019a). In our approach we investigate how to incorporate KB-oriented Gaussian priors in DSRE using a link prediction model to parameterise their mean vector.

7 Conclusions

We proposed a probabilistic approach for distantly supervised relation extraction, which incorporates context agnostic knowledge base triples information as latent signals into context aware bag-level entity pairs. Our method is based on a variational autoencoder that is trained jointly with a relation classifier. KB information via a link prediction model is used in the form of prior distributions on the VAE for each pair. The proposed approach brings close sentences that contain the same KB pairs and it does not require any external information during inference time.

Experimental results suggest that jointly reconstructing sentences with relation classification is helpful for distantly supervised RE and KB priors further boost performance. Analysis of the generated latent representations showed that we can indeed manipulate the space of sentences to match the space of KB triples, while reconstruction is enforced to keep topic-related terms.

Future work will target experimentation with different link prediction models and handling of non-informative sentences. Finally, incorporating large pre-trained language models (LMs) into VAEs is a recent and promising study (Li et al., 2020a) which can be combined with KBs as injecting such information into LMs has been shown to further improve their performance (Peters et al., 2019).

Acknowledgements

This research was supported by BBSRC Japan Partnering Award [Grant ID: BB/P025684/1] and based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). The authors would like to thank the anonymous reviewers for their instructive comments.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy. Association for Computational Linguistics.

Fan Bai and Alan Ritter. 2019. Structured Minimally Supervised Learning for Neural Relation Extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3057–3069, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthias Bauer and Andriy Mnih. 2019. Resampled priors for variational autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 66–75. PMLR.

Iz Beltagy, Kyle Lo, and Waleed Ammar. 2019. Combining distant and direct supervision for neural relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1858–1867, Minneapolis, Minnesota. 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*, volume 26, pages 2787–2795. Curran Associates, Inc.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc.

Prasoon Goyal, Zhiting Hu, Xiaodan Liang, Chenyu Wang, and Eric P Xing. 2017. Nonparametric variational auto-encoders for hierarchical representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5094–5102.

Édouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. 2017. Efficient softmax approximation for GPUs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1302–1310. PMLR.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 745–758, Suzhou, China. Association for Computational Linguistics.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 169–174, Hong Kong, China. Association for Computational Linguistics.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

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.

Linmei Hu, Luhao Zhang, Chuan Shi, Liqiang Nie, Weili Guan, and Cheng Yang. 2019. Improving distantly-supervised relation extraction with joint label embedding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3821–3829, Hong Kong, China. Association for Computational Linguistics.

Yuyun Huang and Jinhua Du. 2019. Self-attention enhanced CNNs and collaborative curriculum learning for distantly supervised relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 389–398, Hong Kong, China. Association for Computational Linguistics.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3060–3066. AAAI Press.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human*

Language Technologies, pages 1148–1158, Portland, Oregon, USA. Association for Computational Linguistics.

Giannis Karamanolakis, Kevin Raji Cherian, Ananth Ravi Narayan, Jie Yuan, Da Tang, and Tony Jebara. 2018. Item recommendation with variational autoencoders and heterogeneous priors. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*, DLRS 2018, page 10–14, New York, NY, USA. Association for Computing Machinery.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020a. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online. Association for Computational Linguistics.

Pengshuai Li, Xinsong Zhang, Weijia Jia, and Hai Zhao. 2019. GAN driven semi-distant supervision for relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3026–3035, Minneapolis, Minnesota. Association for Computational Linguistics.

Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2020b. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8269–8276.

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.

Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia. 2018a. Neural relation extraction via inner-sentence noise reduction and transfer learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2204, Brussels, Belgium. Association for Computational Linguistics.

Yue Liu, Tongtao Zhang, Zhicheng Liang, Heng Ji, and Deborah L. McGuinness. 2018b. Seq2rdf: An end-to-end application for deriving triples from natural language text. *CEUR Workshop Proceedings*, 2180.

Fenglong Ma, Yaliang Li, Chenwei Zhang, Jing Gao, Nan Du, and Wei Fan. 2019. Mcvae: Margin-based conditional variational autoencoder for relation classification and pattern generation. In *The World Wide Web Conference*, WWW '19, page 3041–3048, New York, NY, USA. Association for Computing Machinery.

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Transactions of the Association for Computational Linguistics*, 4:231–244.

Mike Mintz, Steven Bills, Rion Snow, and Daniel 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*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8528–8535.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018a. 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)*, pages 496–505, Melbourne, Australia. Association for Computational Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147, Melbourne, Australia. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg. Springer Berlin Heidelberg.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. *Learning Internal Representations by Error Propagation*, page 318–362. MIT Press, Cambridge, MA, USA.

Heng She, Bin Wu, Bai Wang, and Renjun Chi. 2018. Distant supervision for relation extraction with hierarchical attention and entity descriptions. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

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, Jeju Island, Korea. Association for Computational Linguistics.

Jakub Tomczak and Max Welling. 2018. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223.

Bayu Distiawan Trisedya, Jianzhong Qi, Rui Zhang, and Wei Wang. 2018. GTR-LSTM: A triple encoder for sentence generation from RDF data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1627–1637, Melbourne, Australia. Association for Computational Linguistics.

Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240, Florence, Italy. Association for Computational Linguistics.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium. Association for Computational Linguistics.

Guanying Wang, Wen Zhang, Ruoxu Wang, Yalin Zhou, Xi Chen, Wei Zhang, Hai Zhu, and Huajun Chen. 2018. Label-free distant supervision for relation extraction via knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2246–2255, Brussels, Belgium. Association for Computational Linguistics.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019a. Topic-guided variational autoencoders for text generation. *arXiv preprint arXiv:1903.07137*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *arXiv preprint arXiv:1911.06136*.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1366–1371, Seattle, Washington, USA. Association for Computational Linguistics.

Shanchan Wu, Kai Fan, and Qiong Zhang. 2019. Improving distantly supervised relation extraction with neural noise converter and conditional optimal selector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7273–7280.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Peng Xu and Denilson Barbosa. 2019. Connecting language and knowledge with heterogeneous representations for neural relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3201–3206, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaijia Yang, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2019. Exploiting noisy data in distant supervision relation classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3216–3225, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2810–2819, Minneapolis, Minnesota. Association for Computational Linguistics.

Yujin Yuan, Liyuan Liu, Siliang Tang, Zhongfei Zhang, Yueting Zhuang, Shiliang Pu, Fei Wu, and Xiang Ren. 2019. Cross-relation cross-bag attention for distantly-supervised relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 419–426.

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.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3016–3025, Minneapolis, Minnesota. Association for Computational Linguistics.

Yijia Zhang and Zhiyong Lu. 2019. Exploring semi-supervised variational autoencoders for biomedical relation extraction. *Methods*, 166:112 – 119. Deep Learning in Bioinformatics.

Da Zheng, Xiang Song, Chao Ma, Zeyuan Tan, Zihao Ye, Jin Dong, Hao Xiong, Zheng Zhang, and George Karypis. 2020. Dgl-ke: Training knowledge graph embeddings at scale. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 739–748, New York, NY, USA. Association for Computing Machinery.

Yaoming Zhu, Juncheng Wan, Zhiming Zhou, Liheng Chen, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. 2019. Triple-to-text: Converting rdf triples into high-quality natural languages via optimizing an inverse kl divergence. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 455–464.

A Appendix

A.1 The NYT10 Dataset

As described in Bai and Ritter (2019), the NYT10 dataset has been released in several versions. The original one, follows the setting of Riedel et al. (2010), where two sets of data were created. Later versions (Lin et al., 2016) merged the two sets in order to construct a larger dataset. This merging resulted into $570,300$ instances for training. However, in this version of the data exist overlaps in pairs between the training and the test set. The amount of overlaps is significant and accounts for $47,477$ instances, which is approximately 27.5% of the testing instances. The version was corrected later on but there still remain methods that use the non-filtered data. Recently, Han et al. (2019) released a finalised version removing the overlaps, resulting in $522,611$ total training instances. In our experiments we evaluate the proposed model on both versions.

It is also important to note that NYT10 has been used by the community in two settings: bag-level and sentence-level. In the bag-level setting, a pair's relation is defined based on a bag of sentences that contain the pair. On the contrary, in the sentence-level setting a pair's relation is predicted for each sentence. Training data are obtained using distant supervision, while test data are manually annotated (Hoffmann et al., 2011).

A.2 Data Pre-processing Details

We found that the dataset includes several duplicate instances, i.e. the exact same sentence with the exact same pair. We remove such cases from our training data since they can bias the training process. However, they are preserved on the validation and test sets for a fair comparison with other methods. We convert the dataset to lowercase and replace all digits with the hash character (#). We randomly select 10% of the training bags as our validation set.

		Train	Validation	Test
Processed	Instances	400,100	53,319	172,448
	Bags	248,352	28,108	96,678
	Facts	16,338	1,823	1,950
	Negatives	233,092	26,301	94,917
	Instances	469,290	53,321	-
	Bags	252,044	28,109	-
	Duplicates	62,327	-	-
	Outliers	5,570	-	-

Table 6: Statistics of the NYT10 (520K version) dataset.

Sentence Length Filtering. We restrict the length of a sentence to 50 words for the NYT10 dataset and to 30 for the WIKIDISTANT dataset. If at least one of the arguments of a pair is located in a span after the maximum sentence length, then the sentence is resized to contain the words from the first argument until the second. We also add a maximum number of 5 words to the left and 5 words to the right if the total length allows. If the length of the resized sentence is still larger than

		Train	Validation	Test
Processed	Instances	434,453	62,333	172,448
	Bags	258,843	29,303	96,678
	Facts	17,387	1,942	1,950
	Negatives	242,644	27,374	94,917
	Instances	507,755	-	-
	Bags	262,649	-	-
	Duplicates	66,130	-	-
	Outliers	5,856	-	-

Table 7: Statistics of the NYT10 (570K version) dataset.

		Train	Validation	Test
Processed	Instances	1,000,765	29,145	28,897
	Bags	572,215	14,748	15,509
	Facts	201,356	4,333	4,333
	Negatives	370,859	10,415	11,176
	Instances	1,050,246	-	-
	Bags	575,620	-	-
	Duplicates	43,978	-	-
	Outliers	5,503	-	-

Table 8: Statistics of the WIKIDISTANT dataset.

the maximum sentence length, the sentence is removed from the training set. The reason for this choice is that we want to construct contextualised argument representations. Without the arguments inside the sentence, such representations cannot be formed. We call such removed sentences *outliers*. Outliers are not removed for the validation and test sets. Relevant statistics are shown in Tables 6, 7 and 8.

Vocabulary construction. In order to construct the word vocabulary, we use the unique sentences contained in the training set, as resulted from the removal of duplicate instances and the sentence length filtering. Since each sentence in the dataset can contain multiple pairs, it is repeated for each pair. Using non-unique sentences can lead to counting larger frequencies for certain words and producing a misleading vocabulary. We restrict the vocabulary to contain the 40K most frequent words for NYT10, with a coverage of 97.78% in the training set and to 50K for WIKIDISTANT with a coverage of 96%. Other words are replaced with the UNK token.

A.3 Hyper-parameter Settings

DSRE Models. Table 9 shows the parameters used for training the model on the NYT10 and WIKIDISTANT dataset. In the VAE setting Adaptive Softmax (Grave et al., 2017) was incorporated instead of regular Softmax for faster training. We used three clusters by splitting the vocabulary in $\lfloor \frac{|V|}{15} \rfloor$ and $\lfloor \frac{3|V|}{15} \rfloor$ words.

Parameter	NYT	WIKI
Batch size	128	128
Max bag size	500	500
Learning rate	0.001	0.001
Weight decay	10^{-6}	10^{-6}
Gradient clipping	10	5
Optimiser	Adam	Adam
Early stopping patience	5	5
Task loss weight λ	0.8, 0.9	0.9
Word embedding $\mathbf{E}^{(w)}$ dim.	50	50
Relation embedding $\mathbf{E}^{(r)}$ dim.	64	128
Position embedding $\mathbf{E}^{(p)}$ dim.	8	8
Latent code z dim.	64	64
Teacher force	0.3	0.3
Encoder dim.	256	256
Encoder layers	1	1
Decoder dim.	256	256
Decoder layers	1	1
Input dropout	0.3	0.3
Word dropout	0.3	0.1

Table 9: Models hyper-parameters for each dataset.

Knowledge Base Embeddings. In order to train KB entity embeddings we used the DGL-KE toolkit (Zheng et al., 2020). We use the same set of hyper-parameters for both Freebase and Wikidata as shown in Table 10. For Freebase we select $5,000$ triples as the validation set, while for Wikidata we use the validation set provided in the transductive setting ($5,136$ triples).

Parameter	Value
Model	TransE_l2
Emb. size	64
Max train step	500,000
Batch size	1024
Negative sample size	256
Learning rate	0.1
Gamma	10.0
Negative adversarial sampling	True
Adversarial temperature	1.0
Regularisation coefficient	10^{-7}
Regularisation norm	3

Table 10: Knowledge Base Embeddings hyper-parameters.

A.4 WIKIDISTANT Relation Categories

Since WIKIDISTANT contains 454 relations, their labels are used directly from the WikiData properties[9]. Here, we add explanations about the top 10 most frequent categories used in Figures 4c, 4d.

P17	country
P3373	sibling
P131	located in the administrative territorial entity
P54	member sports team
P175	performer
P161	cast member
P361	part of
P50	author
P150	contains administrative territorial entity
P527	has part

Table 11: Explanations of the top 10 most frequent WIKIDISTANT relation categories.

A.5 Additional Plots

Figure 5 illustrates the t-SNE plot of the latent space for the NYT10 validation set. We observe similar clusters to that of the KB (Figure 4a).

Figure 6 illustrates the PR-curves for the non-filtered version of the NYT10 dataset (570K). Here, KB-priors perform comparably with Normal prior but mostly improve the tail of the distribution (after 50% of the recall range). We could not obtain the PR curve for the JOINTNRE method, thus it is not present in the figure.

[9] https://www.wikidata.org/wiki/Wikidata:List_of_properties

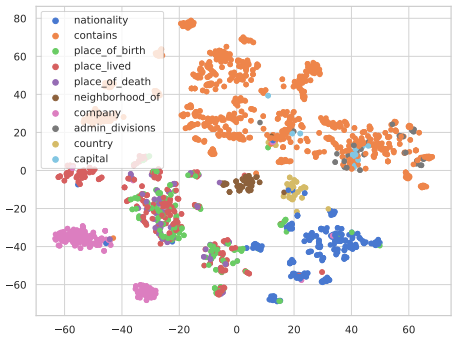


Figure 5: t-SNE plot of the latent vector (μ) for the NYT10 (520K) validation set, when using KB priors during training.

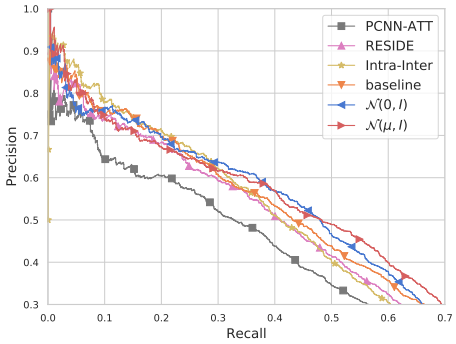


Figure 6: Precision-Recall curves for the NYT10 (570K) test set.