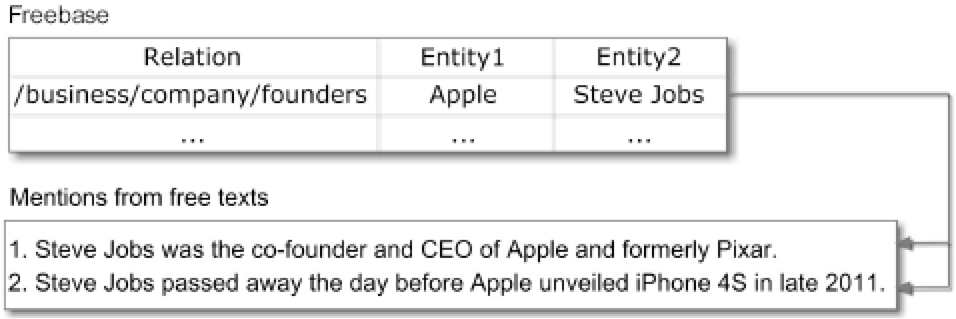
Distantly Supervised Relation Extraction

Definition



Freebase

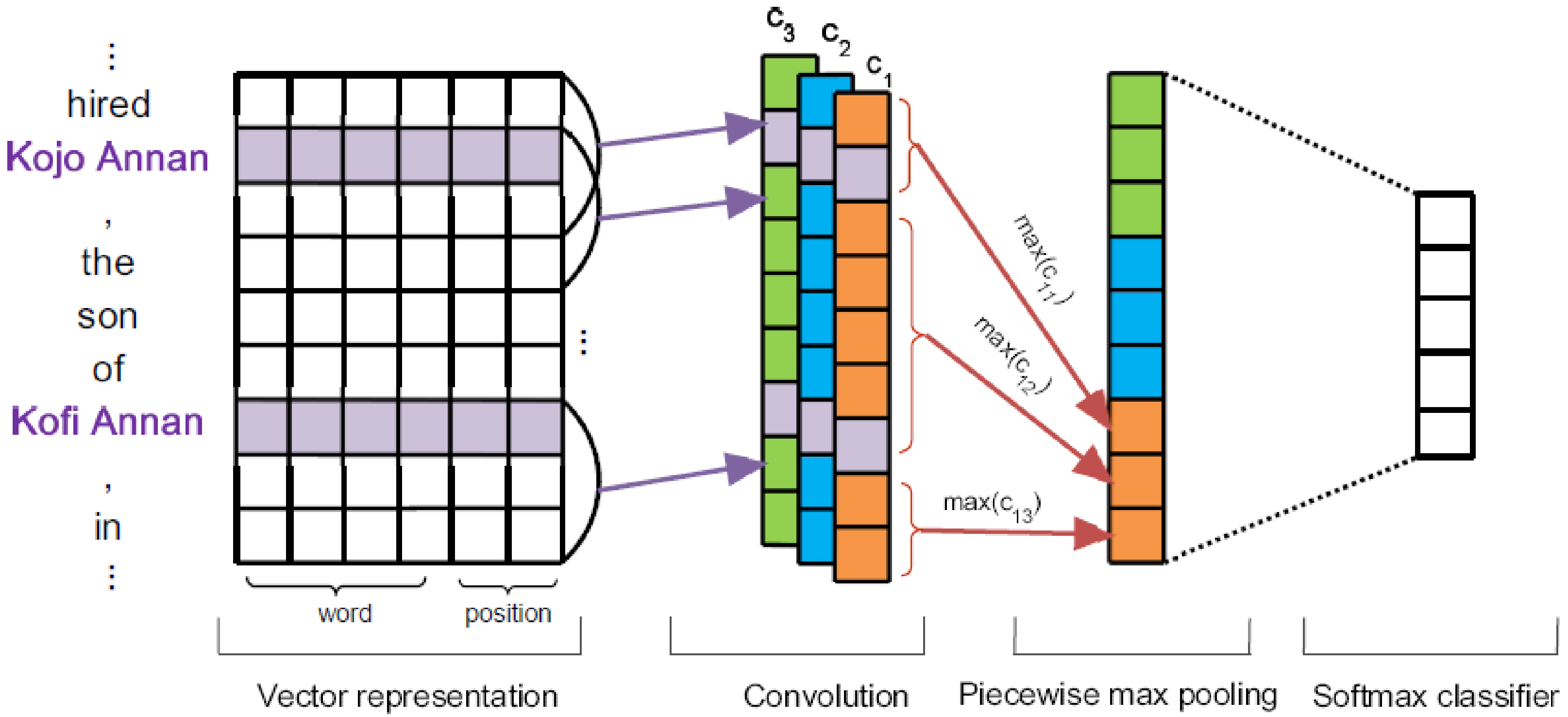
Relation	Entity1	Entity2
/business/company/founders	Apple	Steve Jobs
...	...	...

Mentions from free texts

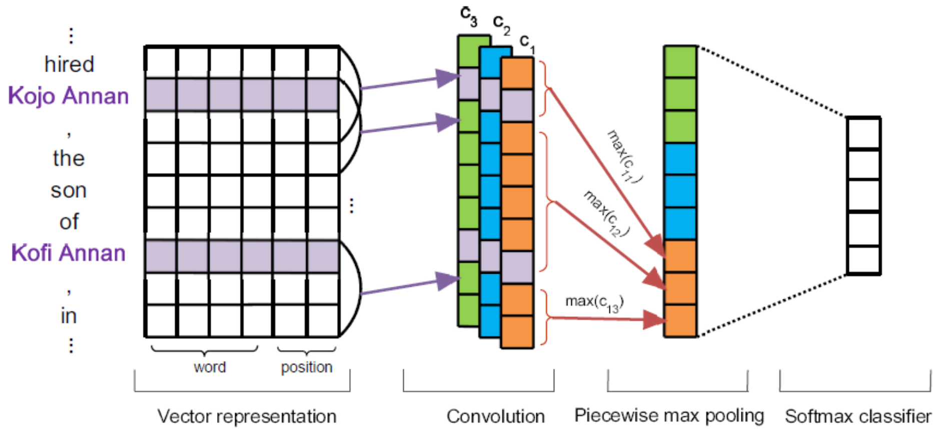
1. Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.
2. Steve Jobs passed away the day before Apple unveiled iPhone 4S in late 2011.

(Mintz et al., 2009) proposes distant supervision to automatically generate training data via aligning KBs and texts. They assume that if two entities have a relation in KBs, then all sentences that contain these two entities will express this relation. For example, (Microsoft, founder, Bill Gates) is a relational fact in KB. Distant supervision will regard all sentences that contain these two entities as active instances for relation founder.

Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks



Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks



$\mathbf{S}$ to be a sequence $\{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_s\}$

$\mathbf{w} \in \mathbb{R}^m \ (m = w * d).$

$c_j = \mathbf{w}\mathbf{q}_{j-w+1:j}$

the index j ranges from 1 to $s+w-1$

$c_{ij} = \mathbf{w}_i \mathbf{q}_{j-w+1:j} \quad 1 \le i \le n$

$\mathbf{C} \ = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n\} \in \mathbb{R}^{n \times (s+w-1)}$

n = number of filter

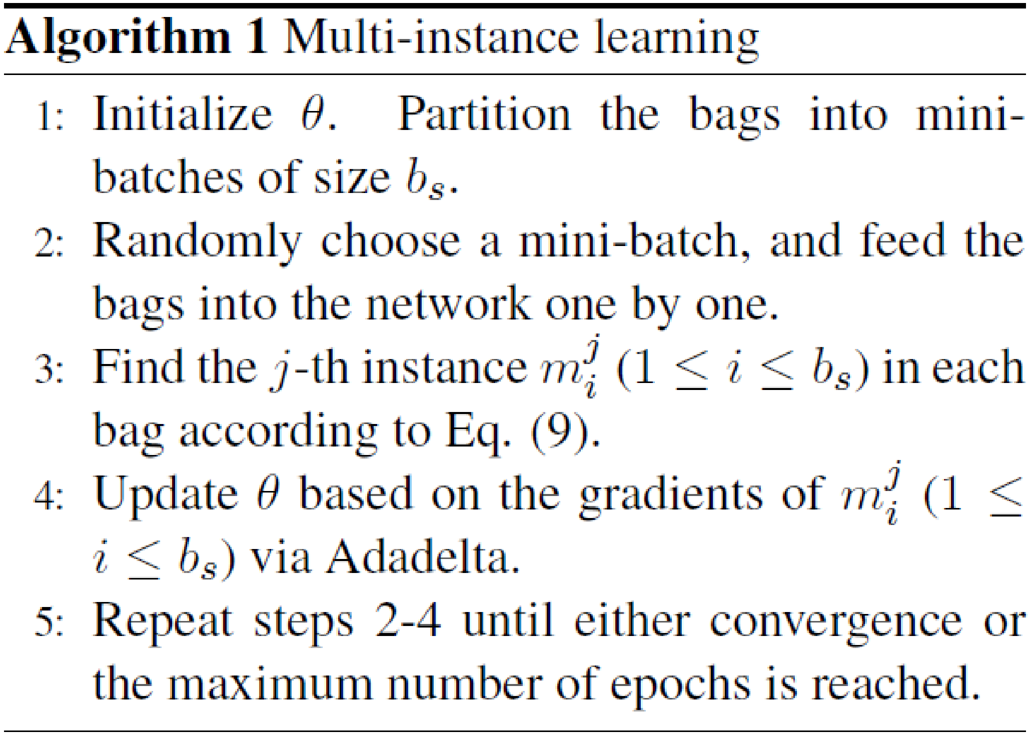
$\{\mathbf{c}_{i1}, \mathbf{c}_{i2}, \mathbf{c}_{i3}\}$

$$p_{ij} = \max(\mathbf{c}_{ij}) \quad 1 \le i \le n, \ \ 1 \le j \le 3$$

$\mathbf{g} = \tanh(\mathbf{p}_{1:n}) \qquad \mathbf{g} \in \mathbb{R}^{3n}.$

$\mathbf{o} = \mathbf{W}_1 \mathbf{g} + b \qquad \mathbf{W}_1 \in \mathbb{R}^{n_1 \times 3n}$

Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks



Algorithm 1 Multi-instance learning

```
1: Initialize $\theta$. Partition the bags into mini-batches of size $b_s$.
2: Randomly choose a mini-batch, and feed the bags into the network one by one.
3: Find the $j$-th instance $m_i^j$ $(1 \leq i \leq b_s)$ in each bag according to Eq. (9).
4: Update $\theta$ based on the gradients of $m_i^j$ $(1 \leq i \leq b_s)$ via Adadelta.
5: Repeat steps 2-4 until either convergence or the maximum number of epochs is reached.
```

$$p(r|m_i^j;\theta) = \frac{e^{o_r}}{\sum\limits_{k=1}^{n_1} e^{o_k}}$$

$$J(\theta) = \sum_{i=1}^{T} \log p(y_i|m_i^j;\theta)$$

$$j^* = \arg\max_j p(y_i|m_i^j;\theta) \quad 1 \leq j \leq q_i$$

A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction

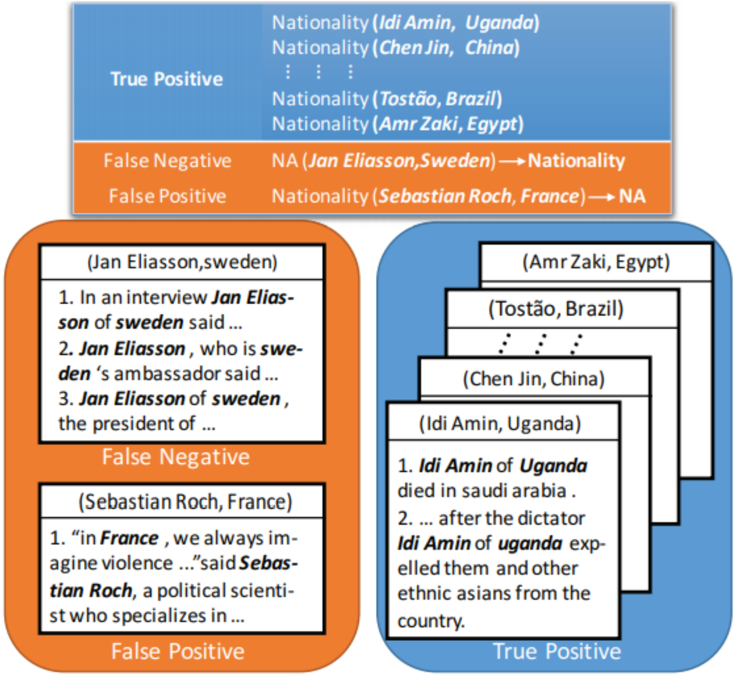

True Positive
Nationality (Idi Amin, Uganda)
Nationality (Chen Jin, China)
Nationality (Tostão, Brazil)
Nationality (Amr Zaki, Egypt)
False Negative
NA (Jan Eliasson,Sweden) → Nationality
False Positive
Nationality (Sebastian Roch, France) → NA
(Jan Eliasson,sweden)
1. In an interview Jan Eliasson of sweden said ...
2. Jan Eliasson , who is sweden 's ambassador said ...
3. Jan Eliasson of sweden , the president of ...
False Negative
(Sebastian Roch, France)
1. "in France , we always imagine violence ..."said Sebastian Roch, a political scientist who specializes in ...
False Positive
(Amr Zaki, Egypt)
(Tostão, Brazil)
(Chen Jin, China)
(Idi Amin, Uganda)
1. Idi Amin of Uganda died in saudi arabia .
2. ... after the dictator Idi Amin of uganda expelled them and other ethnic asians from the country.
True Positive

Figure 1: An example of soft-label correction on *Nationality* relation. We intend to use syntactic/ semantic information of correctly labeled entity pairs (blue) to correct the false positive and false negative instances (orange) during training.

At-least-one: At-least-one assumption is a down sampling method which assumes at least one sentence in the bag will express the relation between two entities, and select the most likely sentence in the bag for training and prediction. To be more specific, the weight of the selected sentence is 1 while those of other sentences in the bag are all 0.

Selective Attention: Lin et al. (2016) proposes selective attention mechanism to reduce weights of noisy instances within the entity-pair bag.

$$\mathbf{s} = \sum_i \alpha_i \mathbf{x}_i ; \alpha_i = \frac{\exp\left(\mathbf{x}_i \mathbf{A} \mathbf{r}\right)}{\sum_k \exp\left(\mathbf{x}_k \mathbf{A} \mathbf{r}\right)} \tag{4}$$

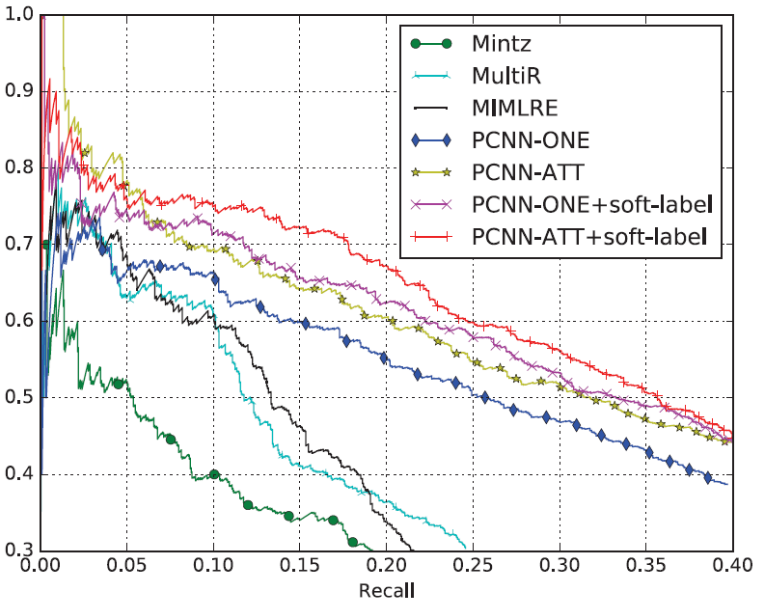
where α_i is the weight of sentence vector $\mathbf{x}_i$, $\mathbf{A}$ and $\mathbf{r}$ are diagonal and relation query parameters.

A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction

$$r_i = \arg\max(\mathbf{o} + \max(\mathbf{o})\mathbf{A} \odot L_i)$$

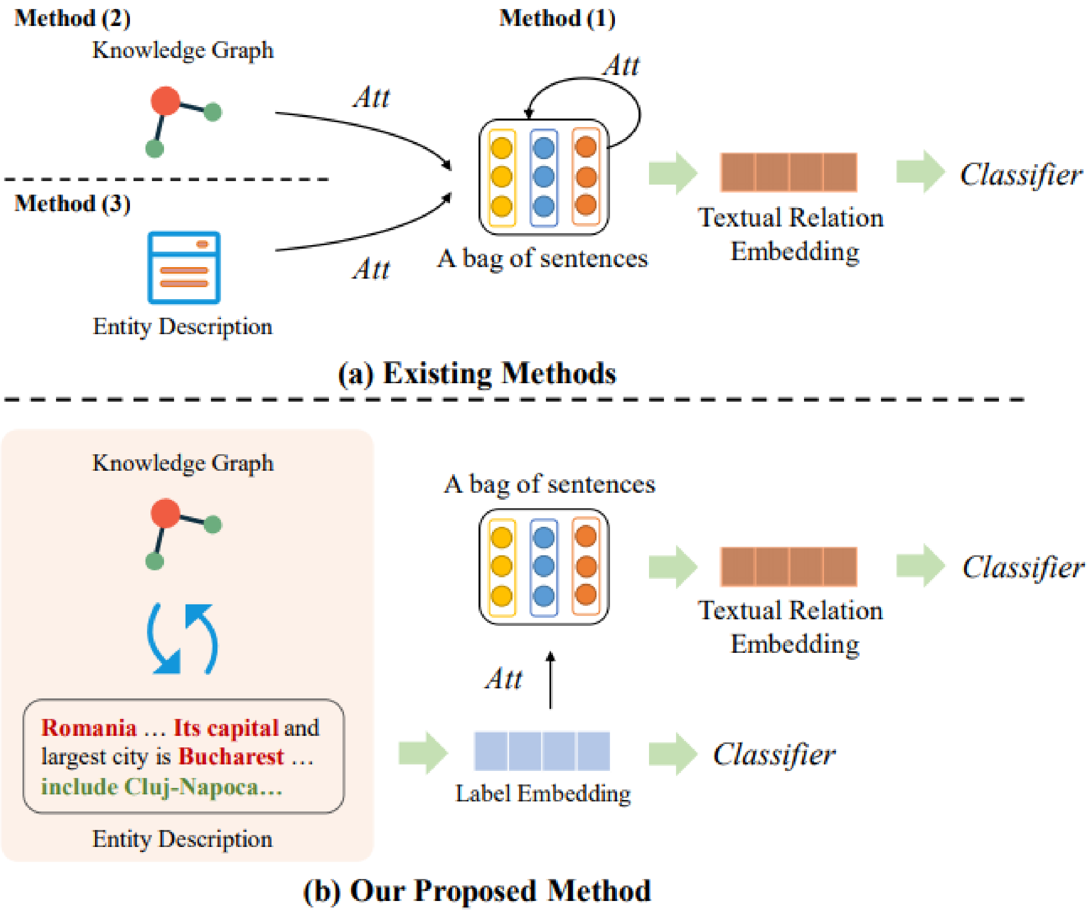
$$\mathbf{o}_t = \frac{\exp\left(\mathbf{M}\mathbf{s}_t + \mathbf{b}\right)}{\sum_k \exp\left(\mathbf{M}\mathbf{s}_k + \mathbf{b}\right)}$$

$$J(\theta) = \sum_{i=1}^{n} \log p(r_i|\mathbf{s}_i;\theta)$$

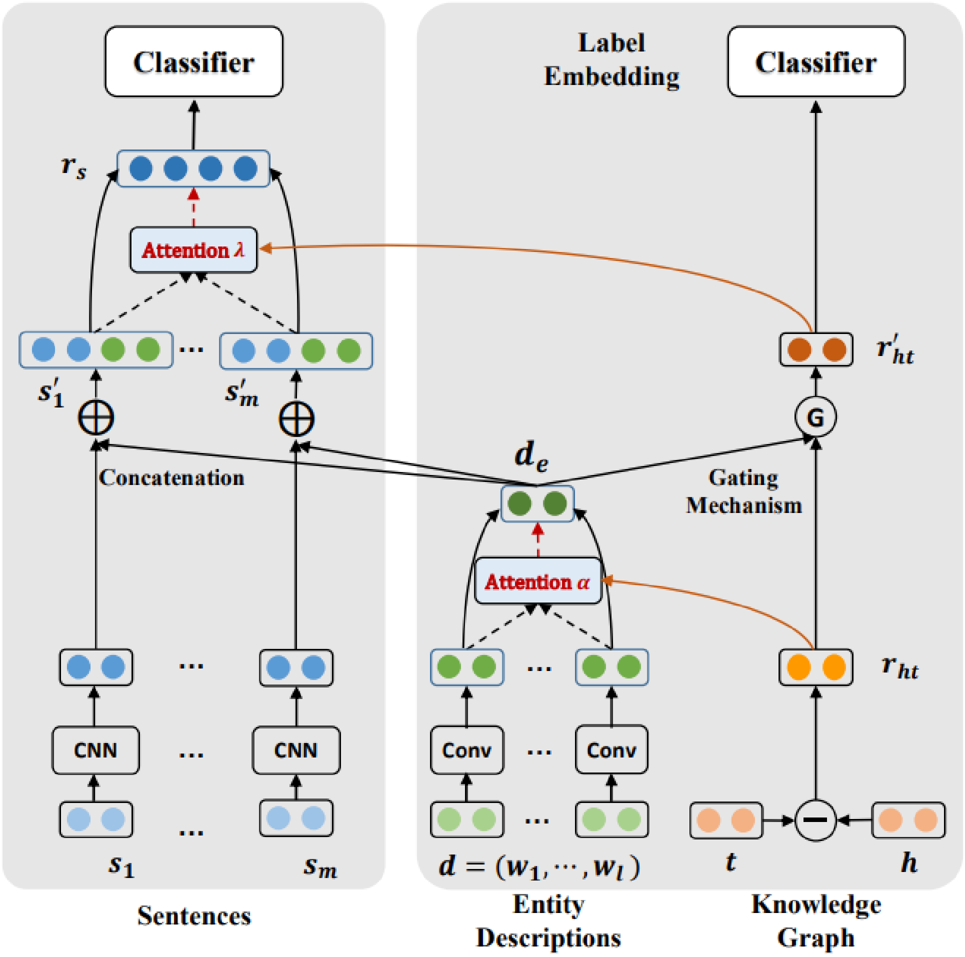


Settings	One				Two				All			
P@N(%)	100	200	300	Avg	100	200	300	Avg	100	200	300	Avg
ONE	73.3	64.8	56.8	65.0	70.3	67.2	63.1	66.9	72.3	69.7	64.1	68.7
+soft-label	77.0	72.5	67.7	72.4	80.0	74.5	69.7	74.7	84.0	81.0	74.0	79.7
ATT	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
+soft-label	**84.0**	**75.5**	**68.3**	**75.9**	**86.0**	**77.0**	**73.3**	**78.8**	**87.0**	**84.5**	**77.0**	**82.8**

Improving Distantly-Supervised Relation Extraction with Joint Label Embedding

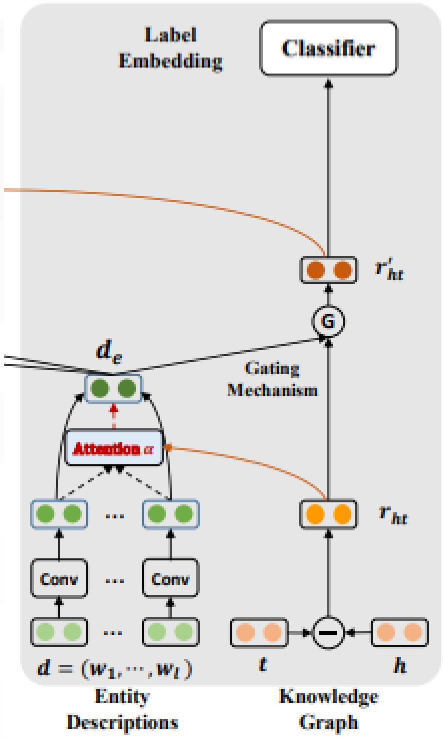


Improving Distantly-Supervised Relation Extraction with Joint Label Embedding



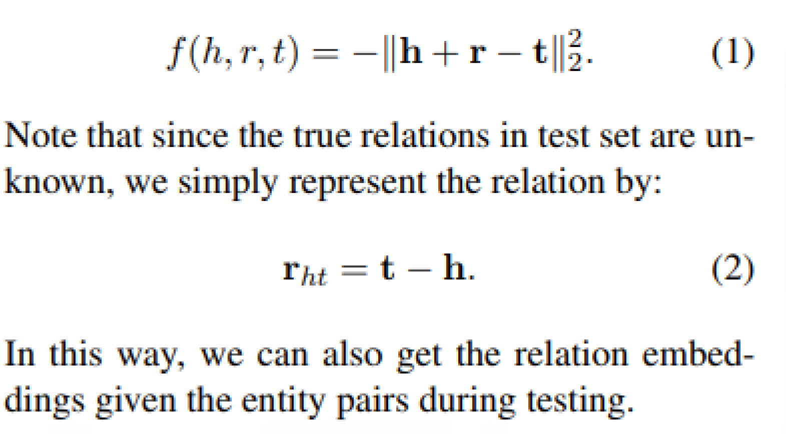
Improving Distantly-Supervised Relation Extraction with Joint Label Embedding

Joint Label Embedding



KG Embedding

TransE algorithm



$$f(h, r, t) = -\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2. \qquad (1)$$

Note that since the true relations in test set are unknown, we simply represent the relation by:

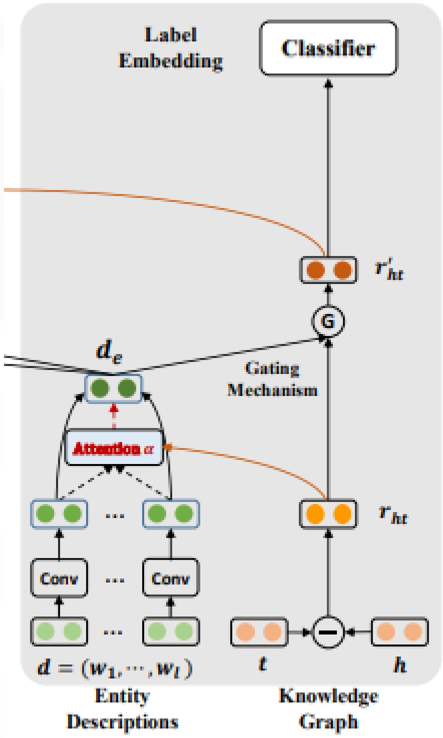
$$\mathbf{r}_{ht} = \mathbf{t} - \mathbf{h}. \qquad (2)$$

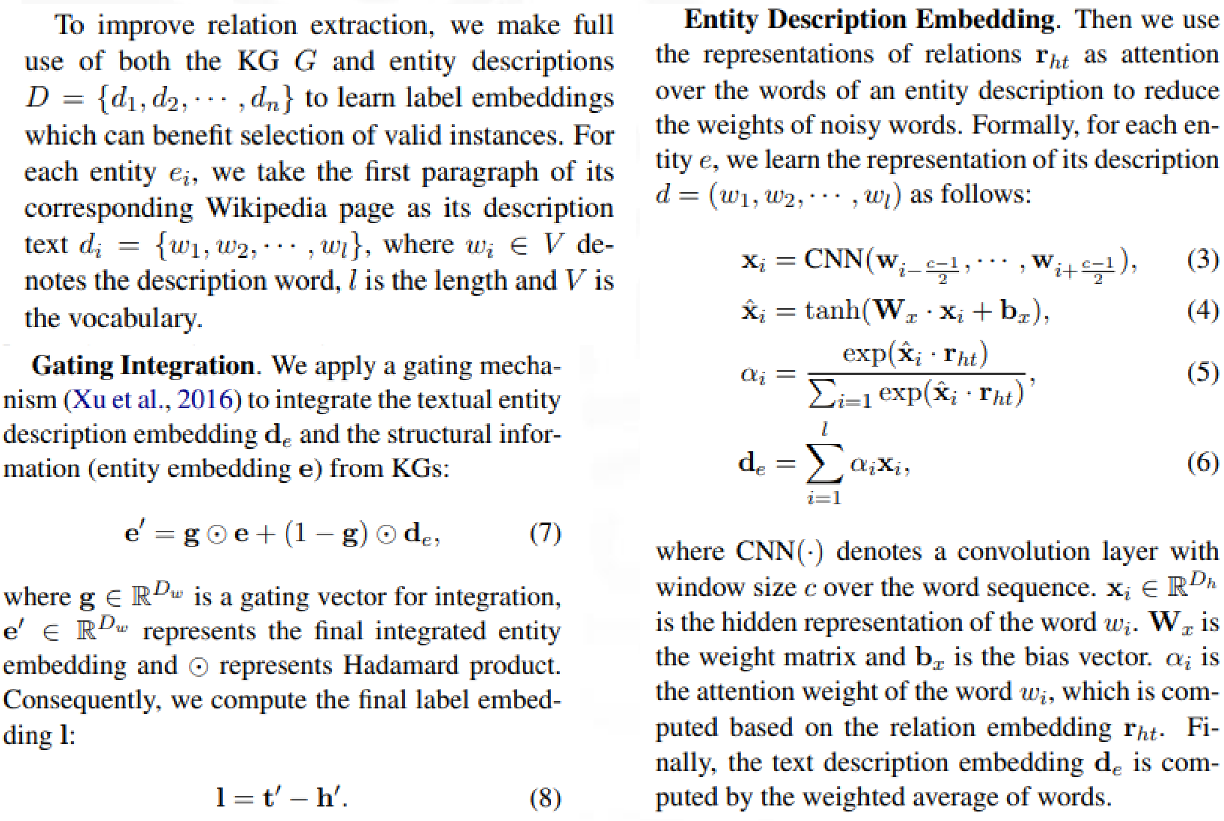
In this way, we can also get the relation embeddings given the entity pairs during testing.

Improving Distantly-Supervised Relation Extraction with Joint Label Embedding

Joint Label Embedding

Entity Description Embedding




To improve relation extraction, we make full use of both the KG G and entity descriptions $D = \{d_1, d_2, \cdots, d_n\}$ to learn label embeddings which can benefit selection of valid instances. For each entity e_i, we take the first paragraph of its corresponding Wikipedia page as its description text $d_i = \{w_1, w_2, \cdots, w_l\}$, where $w_i \in V$ denotes the description word, l is the length and V is the vocabulary.

Gating Integration. We apply a gating mechanism (Xu et al., 2016) to integrate the textual entity description embedding $\mathbf{d}_e$ and the structural information (entity embedding $\mathbf{e}$) from KGs:

$$\mathbf{e}' = \mathbf{g} \odot \mathbf{e} + (1 - \mathbf{g}) \odot \mathbf{d}_e, \qquad (7)$$

where $\mathbf{g} \in \mathbb{R}^{D_w}$ is a gating vector for integration, $\mathbf{e}' \in \mathbb{R}^{D_w}$ represents the final integrated entity embedding and $\odot$ represents Hadamard product. Consequently, we compute the final label embedding $\mathbf{l}$:

$$\mathbf{l} = \mathbf{t}' - \mathbf{h}'. \qquad (8)$$

Entity Description Embedding. Then we use the representations of relations $\mathbf{r}_{ht}$ as attention over the words of an entity description to reduce the weights of noisy words. Formally, for each entity e, we learn the representation of its description $d = (w_1, w_2, \cdots, w_l)$ as follows:

$$\mathbf{x}_i = \mathrm{CNN}(\mathbf{w}_{i-\frac{c-1}{2}}, \cdots, \mathbf{w}_{i+\frac{c-1}{2}}), \qquad (3)$$

$$\hat{\mathbf{x}}_i = \tanh(\mathbf{W}_x \cdot \mathbf{x}_i + \mathbf{b}_x), \qquad (4)$$

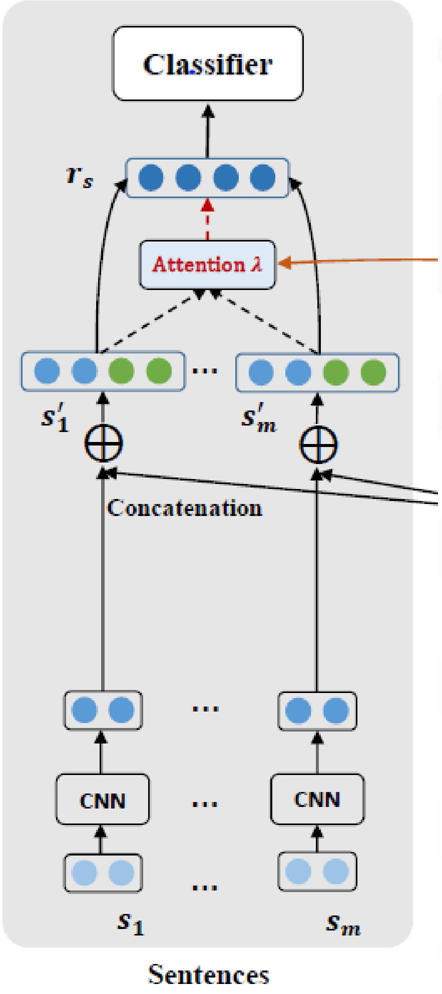
$$\alpha_i = \frac{\exp(\hat{\mathbf{x}}_i \cdot \mathbf{r}_{ht})}{\sum_{i=1} \exp(\hat{\mathbf{x}}_i \cdot \mathbf{r}_{ht})}, \qquad (5)$$

$$\mathbf{d}_e = \sum_{i=1}^{l} \alpha_i \mathbf{x}_i, \qquad (6)$$

where $\mathrm{CNN}(\cdot)$ denotes a convolution layer with window size c over the word sequence. $\mathbf{x}_i \in \mathbb{R}^{D_h}$ is the hidden representation of the word w_i. $\mathbf{W}_x$ is the weight matrix and $\mathbf{b}_x$ is the bias vector. α_i is the attention weight of the word w_i, which is computed based on the relation embedding $\mathbf{r}_{ht}$. Finally, the text description embedding $\mathbf{d}_e$ is computed by the weighted average of words.

Improving Distantly-Supervised Relation Extraction with Joint Label Embedding

Neural Relation Extraction



$$\hat{\mathbf{w}}_i = \mathbf{w}_i \oplus \mathbf{p}_{i1} \oplus \mathbf{p}_{i2},$$

$$\mathbf{z}_i = \mathrm{CNN}(\hat{\mathbf{w}}_{i-\frac{c-1}{2}}, \cdots, \hat{\mathbf{w}}_{i+\frac{c-1}{2}}), \tag{11}$$

$$[\mathbf{s}]_j = \max\{[\mathbf{z}_1]_j, \cdots, [\mathbf{z}_n]_j\}, \tag{12}$$

$$\mathbf{s}' = \mathbf{s} \oplus \mathbf{d}_h \oplus \mathbf{d}_t.$$

$$\hat{\mathbf{s}}_i = \tanh(\mathbf{W}_s \mathbf{s}'_i + \mathbf{b}_s),$$

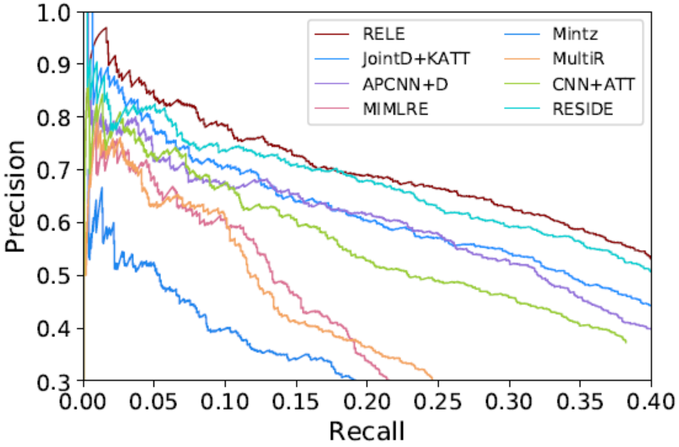
$$\lambda_i = \frac{\exp(\mathbf{l} \cdot \hat{\mathbf{s}}_i)}{\sum_{j=1}^{m} \exp(\mathbf{l} \cdot \hat{\mathbf{s}}_i)},$$

$$\bar{\mathbf{s}} = \sum_{i=1}^{m} \lambda_i \mathbf{s}'_i, \qquad \mathrm{P}(y|B) = \mathrm{Softmax}(\mathbf{M}_s \bar{\mathbf{s}} + \mathbf{b}_s),$$

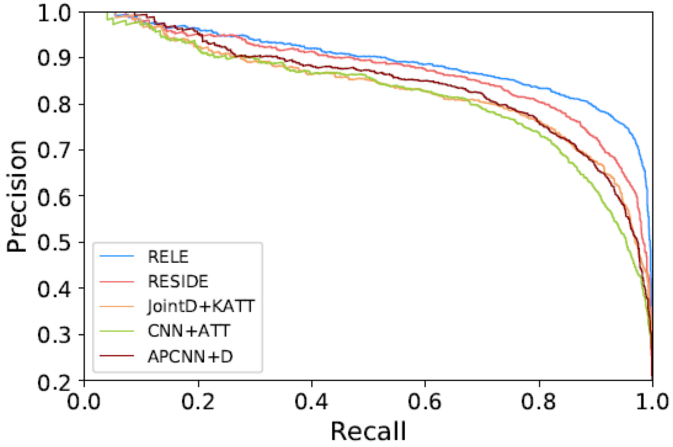
$$L_1 = -\sum_{i=1}^{N} \log \mathrm{P}(y_i|D, G),$$

$$L_2 = -\sum_{i=1}^{N} \log \mathrm{P}(y_i|B_i),$$

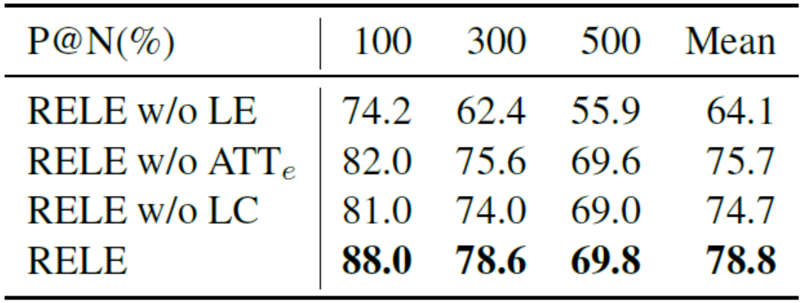
Improving Distantly-Supervised Relation Extraction with Joint Label Embedding



(a) NYT-FB60K Dataset



(b) GIDS-FB8K Dataset



P@N(%)	100	300	500	Mean
RELE w/o LE	74.2	62.4	55.9	64.1
RELE w/o ATT$_e$	82.0	75.6	69.6	75.7
RELE w/o LC	81.0	74.0	69.0	74.7
RELE	**88.0**	**78.6**	**69.8**	**78.8**

Table 3: Evaluation results P@N of variant models on NYT-FB60K dataset.