Self-Attention Enhanced Selective Gate with Entity-Aware Embedding for Distantly Supervised Relation Extraction

Yang Li,[1] Guodong Long,*[1] Tao Shen,[1] Tianyi Zhou,[2] Lina Yao,[3] Huan Huo,[1] Jing Jiang[1]
[1]Centre of Artificial Intelligence, FEIT, University of Technology Sydney
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington
[3]School of Computer Science and Engineering, University of New South Wales
{yang.li-17, tao.shen}@student.uts.edu.au
tianyizh@uw.edu, lina.yao@unsw.edu.au
{guodong.long, huan.huo, jing.jiang}@uts.edu.au

Abstract

Distantly supervised relation extraction intrinsically suffers from noisy labels due to the strong assumption of distant supervision. Most prior works adopt a selective attention mechanism over sentences in a bag to denoise from wrongly labeled data, which however could be incompetent when there is only one sentence in a bag. In this paper, we propose a brand-new light-weight neural framework to address the distantly supervised relation extraction problem and alleviate the defects in previous selective attention framework. Specifically, in the proposed framework, 1) we use an entity-aware word embedding method to integrate both relative position information and head/tail entity embeddings, aiming to highlight the essence of entities for this task; 2) we develop a self-attention mechanism to capture the rich contextual dependencies as a complement for local dependencies captured by piecewise CNN; and 3) instead of using selective attention, we design a pooling-equipped gate, which is based on rich contextual representations, as an aggregator to generate bag-level representation for final relation classification. Compared to selective attention, one major advantage of the proposed gating mechanism is that, it performs stably and promisingly even if only one sentence appears in a bag and thus keeps the consistency across all training examples. The experiments on NYT dataset demonstrate that our approach achieves a new state-of-the-art performance in terms of both AUC and top-n precision metrics.

1 Introduction

Relation extraction (RE) is one of the most fundamental tasks in natural language processing, and its goal is to identify the relationship between a given pair of entities in a sentence. Typically, a large-scale training dataset with clean labels is required to train a reliable relation extraction model. However, it is time-consuming and labor-intensive to annotate such data by crowdsourcing. To overcome the lack of labeled training data, Mintz et al. (2009) presents a distant supervision approach that automatically generates a large-scale, labeled training set by aligning entities in knowledge graph (e.g. Freebase (Bollacker et al. 2008)) to corresponding entity mentions in natural language sentences. This approach is based on a *strong assumption* that, any sentence containing two entities should be labeled according to the relationship of the two entities on the given knowledge graph. However, this assumption does not always hold. Sometimes the same two entities in different sentences with various contexts cannot express a consistent relationship as described in the knowledge graph, which certainly results in wrongly labeled problem.

Bag consisting of one sentence	**Label**	**Correct**
After moving back to *New York*, *Miriam* was the victim of a seemingly racially motivated attack ...	place_lived	True
... he faced, walking *Bill Mueller* and giving up singles to Mark Bellhorn and *Johnny Damon*.	place_lived	False

Table 1: Two examples of one-sentence bag, which are correctly and wrongly labeled by distant supervision respectively.

To alleviate the aformentioned problem, Riedel, Yao, and McCallum (2010) proposes a multi-instance learning framework, which relaxes the strong assumption to *expressed-at-least-one* assumption. In plainer terms, this means any possible relation between two entities hold true in at least one distantly-labeled sentence rather than all of the them that contains those two entities. In particular, instead of generating a sentence-level label, this framework assigns a label to a *bag* of sentences containing a common entity pair, and the label is a relationship of the entity pair on knowledge graph. Recently, based on the labeled data at bag level, a line of works (Zeng et al. 2015; Du et al. 2018; Lin et al. 2016; Han et al. 2018; Ye and Ling 2019) under selective attention framework (Lin et al. 2016) let model implicitly focus on the correctly labeled sentence(s) by an attention mechanism and thus learn a stable and robust model from the noisy data.

However, such selective attention framework is vulnerable to situations where a bag is merely comprised of one

* Corresponding author

single sentence labeled; and what is worse, the only one sentence possibly expresses inconsistent relation information with the bag-level label. This scenario is not uncommon. For a popular distantly supervised relation extraction benchmark, e.g., NYT dataset (Riedel, Yao, and McCallum 2010), up to 80% of its training examples (i.e., bags) are one-sentence bags. From our data inspection, we randomly sample 100 one-sentence bags and find 35% of them is incorrectly labeled. Two examples of one-sentence bag are shown in Table 1. These results indicate that, in training phrase the selective attention module is enforced to output a single-valued scalar for 80% examples, leading to an ill-trained attention module and thus hurting the performance.

Motivated by aforementioned observations, in this paper, we propose a novel **Se**lective **G**ate (SeG) framework for distantly supervised relation extraction. In the proposed framework, 1) we employ both the entity embeddings and relative position embeddings (Zeng et al. 2014) for relation extraction, and an entity-aware embedding approach is proposed to dynamically integrate entity information into each word embedding, yielding more expressively-powerful representations for downstream modules; 2) to strengthen the capability of widely-used piecewise CNN (PCNN) (Zeng et al. 2015) on capturing long-term dependency (Yu et al. 2018), we develop a light-weight self-attention (Lin et al. 2017; Shen et al. 2018) mechanism to capture rich dependency information and consequently enhance the capability of neural network via producing complementary representation for PCNN; and 3) based on preceding versatile features, we design a selective gate to aggregate sentence-level representations into bag-level one and alleviate intrinsic issues appearing in selective attention.

Compared to the baseline framework (i.e., selective attention for multi-instance learning), SeG is able to produce entity-aware embeddings and rich-contextual representations to facilitate downstream aggregation modules that stably learn from noisy training data. Moreover, SeG uses gate mechanism with pooling to overcome problem occurring in selective attention, which is caused by one-sentence bags. In addition, it still keeps a light-weight structure to ensure the scalability of this model.

The experiments and extensive ablation studies on New York Time dataset (Riedel, Yao, and McCallum 2010) show that our proposed framework achieves a new state-of-the-art performance regarding both AUC and top-n precision metrics for distantly supervised relation extraction task, and also verify the significance of each proposed module. Particularly, the proposed framework can achieve AUC of 0.51, which outperforms selective attention baseline by 0.14 and improves previous state-of-the-art approach by 0.09.

2 Proposed Approach

As illustrated in Figure 1, we propose a novel neural network, i.e., SeG, for distantly supervised relation extraction, which is composed of following neural components.

2.1 Entity-Aware Embedding

Given a bag of sentences[1] $B^k = \{s_1^k, \dots, s_{m^k}^k\}$ where each sentence contains common entity pair (i.e., head entity e_h^k, and tail entity e_t^k), the target of relation extraction is to predict the relation y^k between the two entities. For a clear demonstration, we omit indices of example and sentence in remainder if no confusion caused. Each sentence is a sequence of tokens, i.e., $s = [w_1, \dots, w_n]$, where n is the length of the sentence. In addition, each token has a low-dimensional dense-vector representation, i.e., $[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n] \in \mathbb{R}^{d_w \times n}$, where d_w denotes the dimension of word embedding.

In addition to the typical word embedding, relative position is a crucial feature for relation extraction, which can provide downstream neural model with rich positional information (Zeng et al. 2014; 2015). Relative positions explicitly describe the relative distances between each word w_i and the two targeted entities e_h and e_t. For i-th word, a randomly initialized weight matrix projects the relative position features into a two dense-vector representations w.r.t the head and tail entities, i.e., $\boldsymbol{r}_i^{e_h}$ and $\boldsymbol{r}_i^{e_t} \in \mathbb{R}^{d_r}$ respectively. The final low-level representations for all tokens are a concatenation of the aforementioned embeddings, i.e., $\boldsymbol{X}^{(p)} = [\boldsymbol{x}_1^{(p)}, \cdots, \boldsymbol{x}_n^{(p)}] \in \mathbb{R}^{d_p \times n}$ in which $\boldsymbol{x}_i^{(p)} = [\boldsymbol{v_i}; \boldsymbol{r}_i^{e_h}; \boldsymbol{r}_i^{e_t}]$ and $d_p = d_w + 2 \times d_r$.

However, aside from the relative position features, we argue that the embeddings of both the head entity e_h and tail entity e_t are also vitally significant for relation extraction task, since the ultimate goal of this task is to predict the relationship between these two entities. This hypothesis is further verified by our quantitative and qualitative analyses in later experiments (Section 3.2 and 3.3). The empirical results show that our proposed embedding can outperform the widely-used way in prior works (Ji et al. 2017).

In particular, we propose a novel entity-aware word embedding approach to enrich the traditional word embeddings with features of the head and tail entities. To this end, a position-wise gate mechanism is naturally leveraged to dynamically select features between relative position embedding and entity embeddings. Formally, the embeddings of head and tail entities are denoted as $\boldsymbol{v}^{(h)}$ and $\boldsymbol{v}^{(t)}$ respectively. The position-wise gating procedure is formulated as

$$\boldsymbol{\alpha} = \text{sigmoid}(\lambda \cdot (\boldsymbol{W}^{(g1)} \boldsymbol{X}^{(e)} + \boldsymbol{b}^{(g1)})), \quad (1)$$

$$\tilde{\boldsymbol{X}}^{(p)} = \tanh(\boldsymbol{W}^{(g2)} \boldsymbol{X}^{(p)} + \boldsymbol{b}^{(g2)}), \quad (2)$$

$$\boldsymbol{X} = \boldsymbol{\alpha} \cdot \boldsymbol{X}^{(e)} + (1 - \boldsymbol{\alpha}) \cdot \tilde{\boldsymbol{X}}^{(p)}, \quad (3)$$

$$\text{where, } \boldsymbol{X}^{(e)} = [\boldsymbol{x}_i^{(e)}]_{i=1}^n, \ \forall \boldsymbol{x}_i^{(e)} = [\boldsymbol{v}_i; \boldsymbol{v}^{(h)}; \boldsymbol{v}^{(t)}], \quad (4)$$

in which $\boldsymbol{W}^{(g1)} \in \mathbb{R}^{d_h \times 3d_w}$ and $\boldsymbol{W}^{(g2)} \in \mathbb{R}^{d_h \times d_p}$ are learnable parameters, λ is a hyper-parameter to control smoothness, and $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \in \mathbb{R}^{d_h \times n}$ containing the entity-aware embeddings of all tokens from the sentence.

2.2 Self-Attention Enhanced Neural Network

Previous works of relation extraction mainly employ a piecewise convolutional neural network (PCNN) (Zeng et al.

[1] "sentence" and "instance" are interchangeable in this paper.

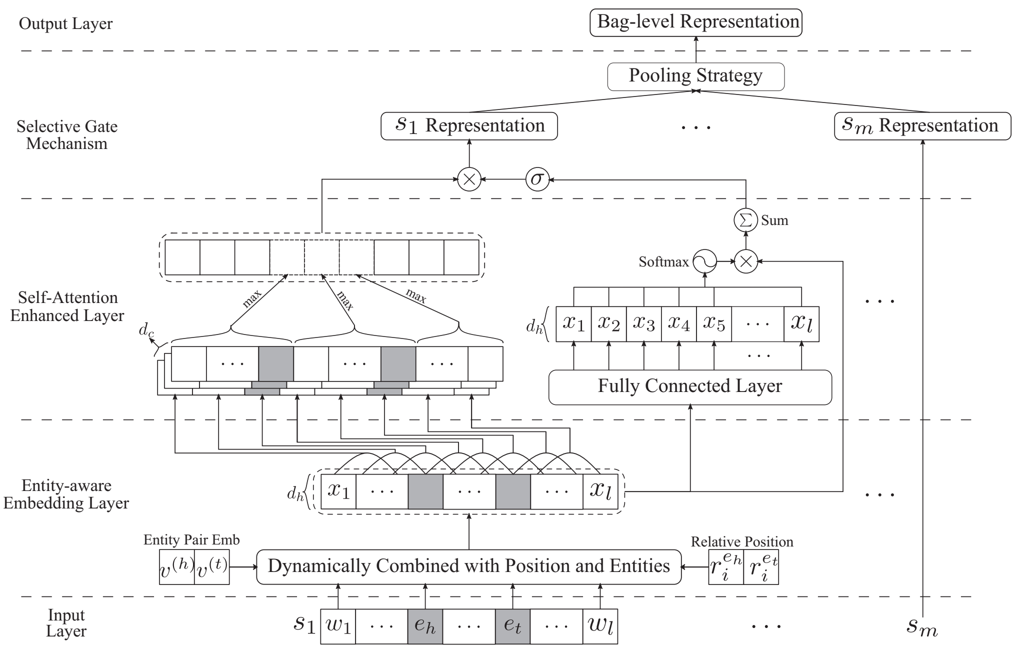

Figure 1: The framework of our approach (i.e. SeG) that consisting of three components: 1) entity-aware embedding 2) self-attention enhanced neural network and 3) a selective gate. Note, tokens e_h and e_t with gray background mean the head entity and tail entity of this sentence.

2015) to obtain contextual representation of sentences due to its capability of capturing local features, less computation and light-weight structure. However, some previous works (Vaswani et al. 2017) find that CNNs cannot reach state-of-the-art performance on a majority of natural language processing benchmarks due to a lack of measuring long-term dependency, even if stacking multiple modules. This motivates us to enhance the PCNN with another neural module, which is capable of capturing long-term or global dependencies to produce complementary and more powerful sentence representation.

Hence, we employ a self-attention mechanism in our model due to its parallelizable computation and state-of-the-art performance. Unlike existing approaches that sequentially stack self-attention and CNN layers in a cascade form (Yu et al. 2018; Wu et al. 2019), we arrange these two modules in parallel so they can generate features describing both local and long-term relations for the same input sequence. Since each bag may contain many sentences (up to 20), a light-weight networks that can can efficiently process these sentences simultaneously is more preferable, such as PCNN that is the most popular module for relation extraction. For this reason, there is only one light-weight self-attention layer in our model. This is contrast to Yu et al. (2018) and Wu et al. (2019) who stack both modules many times repeatedly. Our experiments show that two modules arranged in parallel manner consistently outperform stacking architectures that are even equipped with additional residual connections (He et al. 2016)). The comparative experiments will be elaborated in Section 3.1 and 3.2.

Piecewise Convolutional Neural Network This section provides a brief introduction to PCNN as a background for further integration with our model, and we refer readers to Zeng et al. (2015) for more details. Each sentence is divided into three segments w.r.t. the head and tail entities. Compared to the typical 1D-CNN with max-pooling (Zeng et al. 2014), piecewise pooling has the capability to capture the structure information between two entities. Therefore, instead of using word embeddings with relative position features $\boldsymbol{X}^{(p)}$ as the input, we here employ our entity-aware embedding $\boldsymbol{X}$ as described in Section 2.1 to enrich the input features. First, 1D-CNN is invoked over the input, which can be formally represented as

$$\boldsymbol{H} = \text{1D-CNN}(\boldsymbol{X}; \boldsymbol{W}^{(c)}, \boldsymbol{b}^{(c)}) \in \mathbb{R}^{d_c \times n}, \qquad (5)$$

where, $\boldsymbol{W}^{(c)} \in \mathbb{R}^{d_c \times m \times d_h}$ is convolution kernel with window size of m (i.e., m-gram). Then, to obtain sentence-level representation, a piecewise pooling performs over the output sequence, i.e., $\boldsymbol{H}^{(c)} = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_n]$, which is formulated as

$$\boldsymbol{s} = \tanh([\text{Pool}(\boldsymbol{H}^{(1)}); \text{Pool}(\boldsymbol{H}^{(2)}); \text{Pool}(\boldsymbol{H}^{(3)})]). \qquad (6)$$

In particular, $\boldsymbol{H}^{(1)}$, $\boldsymbol{H}^{(2)}$ and $\boldsymbol{H}^{(3)}$ are three consecutive parts of $\boldsymbol{H}$, obtained by dividing $\boldsymbol{H}$ according to the positions of head and tail entities. Consequently, $\boldsymbol{s} \in \mathbb{R}^{3d_c}$ is the resulting sentence vector representation.

Self-Attention Mechanism To maintain efficiency of proposed approach, we adopt the recently-promoted self-attention mechanism (Liu et al. 2016; Lin et al. 2017;

Shen et al. 2019; Li et al. 2018; Liu et al. 2019) for compressing a sequence of token representations into a sentence-level vector representation by exploiting global dependency, rather than computation-consuming pairwise ones (Vaswani et al. 2017). It is used to measure the contribution or importance of each token to relation extraction task w.r.t. the global dependency. Formally, given the entity-aware embedding $\boldsymbol{X}$, we first calculate attention probabilities by a parameterized compatibility function, i.e.,

$$\boldsymbol{A} = \boldsymbol{W}^{(a2)}\sigma(\boldsymbol{W}^{(a1)}\boldsymbol{X} + \boldsymbol{b}^{(a1)}) + \boldsymbol{b}^{(a2)}, \tag{7}$$

$$\boldsymbol{P}^{(A)} = \mathrm{softmax}(\boldsymbol{A}), \tag{8}$$

where, $\boldsymbol{W}^{(a1)}, \boldsymbol{W}^{(a2)} \in \mathbb{R}^{d_h \times d_h}$ are learnable parameters, $\mathrm{softmax}(\cdot)$ is invoked over sequence, and $\boldsymbol{P}^{(A)}$ is resulting attention probability matrix. Then, the result of self-attention mechanism can be calculated as

$$\boldsymbol{u} = \sum \boldsymbol{P}^{(A)} \odot \boldsymbol{X}, \tag{9}$$

in which, $\sum$ is performed along sequential dimension and $\odot$ stands for element-wise multiplication. And, $\boldsymbol{u} \in \mathbb{R}^{d_h}$ is also a sentence-level vector representation which is a complement to PCNN-resulting one, i.e., $\boldsymbol{s}$ from Eq.(6).

2.3 Selective Gate

Given a sentence bag $B = [s_1, \dots, s_m]$ with common entity pair, where m is the number of sentences. As elaborated in Section 2.2, we can obtain $\boldsymbol{S} = [\boldsymbol{s}_1, \dots, \boldsymbol{s}_m]$ and $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_m]$ for each sentence in the bag, which are derived from PCNN and self-attention respectively.

Unlike previous works under multi-instance framework that frequently use a selective attention module to aggregate sentence-level representations into bag-level one, we propose a innovative selective gate mechanism to perform this aggregation. The selective gate can mitigate problems existing in distantly supervised relation extraction and achieve a satisfactory empirical effectiveness. Specifically, when handling the noisy instance problem, selective attention tries to produce a distribution over all sentence in a bag; but if there is only one sentence in the bag, even the only sentence is wrongly labeled, the selective attention mechanism will be low-effective or even completely useless. Note that almost 80% of bags from popular relation extraction benchmark consist of only one sentence, and many of them suffer from the wrong label problem. In contrast, our proposed gate mechanism is competent to tackle such case by directly and dynamically aligning low gating value to the wrongly labeled instances and thus preventing noise representation being propagated.

Particularly, a two-layer feed forward network is applied to each $\boldsymbol{u}_j$ to sentence-wisely produce gating value, which is formally denoted as

$$g_j = \mathrm{sigmoid}(\boldsymbol{W}^{(g1)}\sigma(\boldsymbol{W}^{(g2)}\boldsymbol{u}_j + \boldsymbol{b}^{(g2)}) + \boldsymbol{b}^{(g1)}), \quad \forall j = 1, \dots, m, \tag{10}$$

where, $\boldsymbol{W}^{(g1)} \in \mathbb{R}^{3d_c \times d_h}$, $\boldsymbol{W}^{(g2)} \in \mathbb{R}^{d_h \times d_h}$, $\sigma(\cdot)$ denotes an activation function and $g_j \in (0, 1)$. Then, given the calculated gating value, an mean aggregation performs over sentence embeddings $[\boldsymbol{s}_j]_{j=1}^m$ in the bag, and thus produces bag-level vector representation for further relation classification. This procedure is formalized as

$$\boldsymbol{c} = \frac{1}{m}\sum_{j=1}^{m} g_j \cdot \boldsymbol{s}_j \tag{11}$$

Finally, $\boldsymbol{c}$ is fed into a multi-layer perceptron followed with $|C|$-way $\mathrm{softmax}$ function (i.e., an MLP classifier) to judge the relation between head and tail entities, where $|C|$ is the number of distinctive relation categories. This can be regarded as a classification task (Long et al. 2012). Formally,

$$\boldsymbol{p} = \mathrm{softmax}(\mathrm{MLP}(\boldsymbol{c})) \in \mathbb{R}^{|C|}. \tag{12}$$

2.4 Model Learning

We minimize negative log-likelihood loss plus L_2 regularization penalty to train the model, which is written as

$$L_{NLL} = -\frac{1}{|\mathcal{D}|}\sum\nolimits_{k=1}^{|\mathcal{D}|} \log \boldsymbol{p}^k_{(i=y^k)} + \beta||\theta||_2^2 \tag{13}$$

where $\boldsymbol{p}^k$ is the predicted distribution from Eq.(12) for the k-th example in dataset $\mathcal{D}$ and y^k is its corresponding distant supervision label.

3 Experiments

To evaluate our proposed framework, and to compare the framework with baselines and competitive approaches, we conduct experiments on a popular benchmark dataset for distantly supervised relation extraction. We also conduct an ablation study to separately verify the effectiveness of each proposed component, and last, case study and error analysis are provided for an insight into our model.

Dataset In order to accurately compare the performance of our model, we adopt New York Times (NYT) dataset (Riedel, Yao, and McCallum 2010), a widely-used standard benchmark for distantly supervised relation extraction in most of previous works (Lin et al. 2016; Zeng et al. 2015; Han et al. 2018; Du et al. 2018), which contains 53 distinct relations including a null class *NA* relation. This dataset generates by aligning Freebase with the New York Times (NYT) corpus automatically. In particular, NYT dataset contains 53 distinct relations including a null class *NA* relation referred to as the relation of an entity pair is unavailable. There are 570K and 172K sentences respectively in training and test set.

Metrics Following previous works (Zeng et al. 2015; Lin et al. 2016; Han et al. 2018; Du et al. 2018), we use precision-recall (PR) curves, area under curve (AUC) and top-N precision (P@N) as metrics in our experiments on the held-out test set from the NYT dataset. To directly show the perfomance on one sentence bag, we also calculate the accuracy of classification (Acc.) on non-NA sentences.

Approach	**One**				**Two**				**All**			
P@N (%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
Comparative Approaches												
CNN+ATT (Lin et al. 2016)	72.0	67.0	59.5	66.2	75.5	69.0	63.3	69.3	74.3	71.5	64.5	70.1
PCNN+ATT (Lin et al. 2016)	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
PCNN+ATT+SL (Liu et al. 2017)	84.0	75.5	68.3	75.9	86.0	77.0	73.3	78.8	87.0	84.5	77.0	82.8
PCNN+HATT (Han et al. 2018)	84.0	76.0	69.7	76.6	85.0	76.0	72.7	77.9	88.0	79.5	75.3	80.9
PCNN+BAG-ATT (Ye and Ling 2019)	86.8	77.6	73.9	79.4	91.2	79.2	75.4	81.9	91.8	84.0	78.7	84.8
SeG (*ours*)	**94.0**	**89.0**	**85.0**	**89.3**	**91.0**	**89.0**	**87.0**	**89.0**	**93.0**	**90.0**	**86.0**	**89.3**
Ablations												
SeG w/o Ent	85.0	75.0	67.0	75.6	87.0	79.0	70.0	78.6	85.0	80.0	72.0	79.0
SeG w/o Gate	87.0	85.5	82.7	85.1	89.0	87.0	84.0	86.7	90.0	88.0	85.3	87.7
SeG w/o Gate w/o Self-Attn	86.0	85.0	82.0	84.3	88.0	86.0	83.0	85.7	90.0	86.5	86.0	87.5
SeG w/o ALL	81.0	73.5	67.3	74.0	82.0	75.0	72.3	76.4	81.0	75.0	72.0	76.0
SeG+ATT w/o Gate	89.0	83.5	75.7	82.7	90.0	83.5	77.0	83.5	92.0	82.0	76.7	83.6
SeG+ATT	88.0	81.0	75.0	81.3	87.0	82.5	77.0	82.2	90.0	86.5	81.0	85.8
SeG w/ stack	91.0	88.0	85.0	88.0	91.0	87.0	85.0	87.7	92.0	89.5	86.0	89.1

Table 2: Precision values for the top-100, -200 and -300 relation instances that are randomly selected in terms of one/two/all sentence(s).

Training Setup For a fair and rational comparison with baselines and competitive approaches, we set most of the hyper-parameters by following prior works (Lin et al. 2017; Han et al. 2018), and also use 50D word embedding and 5D position embedding released by (Lin et al. 2016; Han et al. 2018) for initialization, where the dimension of d_h equals to 150. The filters number of CNN d_c equals to 230 and the kernel size m in CNN equals to 3. In output layer, we employ dropout (Srivastava et al. 2014) for regularization, where the drop probability is set to 0.5. To minimize the loss function defined in Eq.13, we use stochastic gradient descent with initial learning rate of 0.1, and decay the learning rate to one tenth every 100K steps.

Baselines and Competitive Approaches We compare our proposed approach with extensive previous ones, including feature-engineering, competitive and state-of-the-art approaches, which are briefly summarized in the following.

- **Mintz** (Mintz et al. 2009) is the original distantly supervised approach to solve relation extraction problems with distantly supervised data.
- **MultiR** (Hoffmann et al. 2011) is a graphical model within a multi-instance learning framework that is able to handle problems with overlapping relations.
- **MIML** (Surdeanu et al. 2012) is a multi-instance, multi-label learning framework that jointly models both multiple instances and multiple relations.
- **PCNN+ATT** (Lin et al. 2016) employs a selective attention over multiple instances to alleviate the wrongly labeled problem, which is the principal baseline of our work.
- **PCNN+ATT+SL** (Liu et al. 2017) introduces an entity-pair level denoising method, namely employing a soft label to alleviate the impact of wrongly labeled problem.

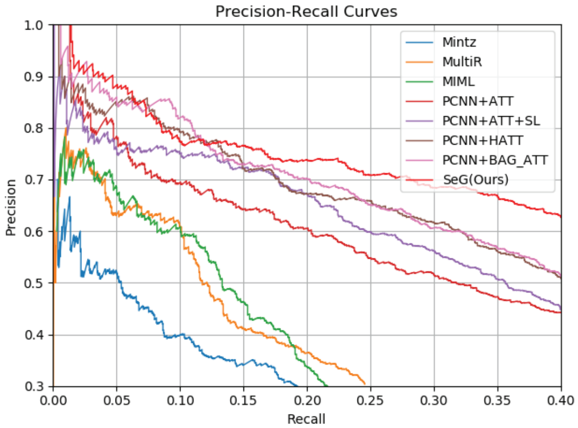

Figure 2: Performance comparison for proposed model and previous baselines in terms of precision-recall curves

- **PCNN+HATT** (Han et al. 2018) employs hierarchical attention to exploit correlations among relations.
- **PCNN+BAG-ATT** (Ye and Ling 2019) uses an intra-bag to deal with the noise at sentence-level and an inter-bag attention to deal with noise at the bag-level.

3.1 Relation Extraction Performance

We first compare our proposed SeG with aforementioned approaches in Table 2 for top-N precision (i.e., P@N). As shown in the top panel of the table, our proposed model SeG can consistently and significantly outperform baseline (i.e., PCNN+ATT) and all recently-promoted works in terms of all P@N metric. Compared to PCNN with selective attention (i.e., PCNN+ATT), our proposed SeG can significantly improve the performance by 23.6% in terms of P@N mean

Approach	AUC
PCNN+HATT	0.42
PCNN+ATT-RA+BAG-ATT	0.42
SeG (ours)	0.51

Table 3: Model comparison regarding the AUC value. The comparative results are reported by Han et al. (2018) and Ye and Ling (2019) respectively.

Approach	AUC	Acc.
PCNN	0.36	83%
PCNN+ATT	0.35	78%
SeG(ours)	0.48	90%

Table 4: Model that is trained and tested on extracted one sentence bags from NYT dataset comparison regarding the AUC value and Acc., where Acc. is accuracy on non-NA sentences.

for all sentences; even if a soft label technique is applied (i.e., PCNN+ATT+SL) to alleviate wrongly labeled problem, our performance improvement is also very significant, i.e., 7.8%.

Compared to previous state-of-the-art approaches (i.e., PCNN+HATT and PCNN+BAG-ATT), the proposed model can also outperform them by a large margin, i.e., 10.3% and 5.3% , even if they propose sophisticated techniques to handle the noisy training data. These verify the effectiveness of our approach over previous works when solving the wrongly labeled problem that frequently appears in distantly supervised relation extraction.

Moreover, for proposed approach and comparative ones, we also show AUC curves and available numerical values in Figure 2 and Table 3 respectively. The empirical results for AUC are coherent with those of P@N, which shows that, our proposed approach can significantly improve previous ones and reach a new state-of-the-art performance by handling wrongly labeled problem using context-aware selective gate mechanism. Specifically, our approach substantially improves both PCNN+HATT and PCNN+BAG-ATT by 21.4% in aspect of AUC for precision-recall.

3.2 Ablation Study

To further verify the effectiveness of each module in the proposed framework, we conduct an extensive ablation study in this section. In particular, *SeG w/o Ent* denotes removing entity-aware embedding, *SeG w/o Gate* denotes removing selective gate and concatenating two representations from PCNN and self-attention, *SeG w/o Gate w/o Self-Attn* denotes removing self-attention enhanced selective gate. In addition, we also replace the some parts of the proposed framework with baseline module for an in-depth comparison. *SeG+ATT* denotes replacing mean-pooing with selective attention, and *SeG w/ stack* denotes using stacked PCNN and self-attention rather than in parallel.

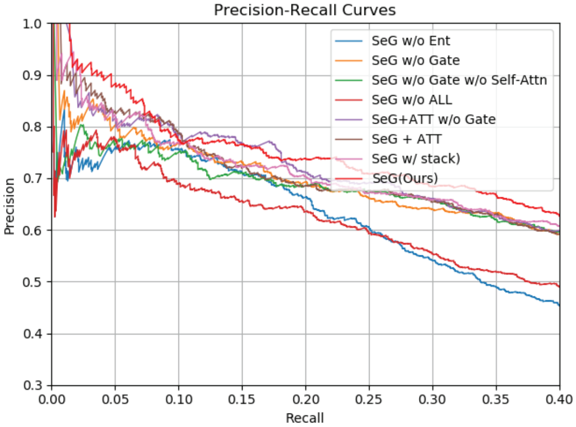

Figure 3: Performance comparison for ablation study under precision-recall curves

The P@N results are listed in the bottom panel of Table 2, and corresponding AUC results are shown in Table 5 and Figure 3. According to the results, we find that our proposed modules perform substantially better than those of the baseline in terms of both metrics. Particularly, by removing entity-aware embedding (i.e, SeG w/o Ent) and self-attention enhanced selective gate (i.e., SeG w/o Gate w/o Self-Attn), it shows 11.5% and 1.8% decreases respectively in terms of P@N mean for all sentences. Note that, when dropping both modules above (i.e., SeG w/o ALL), the framework will be degenerated as selective attention baseline (Lin et al. 2016), which again demonstrates that our proposed framework is superior than the baseline by 15% in terms of P@N mean for all sentences.

To verify the performance of selective gate modul when handling wrongly labeled problem, we simply replace the selective gate module introduced in Eq.(11) with selective attention module, namely, SeG+Attn w/o Gate, and instead of mean pooling in Eq.(11), we couple selective gate with selective attention to fulfill aggregation instead mean-pooling, namely, SeG+Attn. Across the board, the proposed SeG still deliver the best results in terms of both metrics even if extra selective attention module is applied.

Lastly, to explore the influence of the way to combine

Approach	AUC
SeG (ours)	0.51
SeG w/o Ent	0.40
SeG w/o Gate	0.48
SeG w/o Gate w/o Self-Attn	0.47
SeG w/o ALL	0.40
SeG + ATT w/o Gate	0.47
SeG + ATT	0.47
SeG w/ stack	0.48

Table 5: Ablation study regarding precision-recall AUC value.

Bag	Sentence	Relation	SeG (Ours)	SeG w/o Ent	SeG w/o GSA
B1	**Yul Kwon**, 32, of **San Mateo**, Calif., winner of last year's television contest "Survivor" and ...	*/people/person/place_lived*	Correct	Wrong	Wrong
B2	Other winners were Alain Mabanckou from Congo, **Nancy Huston** from **Canada** and Léonora Miano from Cameroon.	*/people/person/nationality*	Correct	Correct	Wrong
B3	... production moved to **Connecticut** to film interiors in places like Stamford, Bridgeport, Shelton, **Ridgefield** and Greenwich.	*/location/location/contains*	Correct	Wrong	Correct
B4	... missionary **George Whitefield**, according to The Encyclopedia of **New York City**.	*NA*	Correct	Wrong	Correct

Table 6: A case study where each bag contains one sentence. *SeG w/o GSA* is an abbreviation of *SeG w/o Gate w/o Self-Attn*.

PCNN with self-attention mechanism, we stack them by following the previous works (Yu et al. 2018), i.e., SeG w/ Stack. And we observe a notable performance drop after stacking PCNN and self-attention in Table 5. This verifies that our model combining self-attention mechanism and PCNN in parallel can achieve a satisfactory result.

To further empirically evaluate the performance of our method in solving one-sentence bag problem, we extract only the one-sentence bags from NYT's training and test sets, which occupy 80% of the original dataset. The evaluation and comparison results in Table 4 show that compared to PCNN+ATT, the AUC improvement (+0.13) between our model and PCNN+ATT on one-sentence bags is higher than the improvement of full NYT dataset, which verifies SeG's effectiveness on one-sentence bags. In addition, PCNN+ATT shows a light decrease compared with PCNN, which can also support the claim that selective attention is vulnerable to one-sentence bags.

3.3 Case Study

In this section, we conduct a case study to qualitatively analyze the effects of entity-aware embedding and self-attention enhanced selective gate. The case study of four examples is shown in Table 6.

First, comparing Bag 1 and 2, we find that, without the support of the self-attention enhanced selective gate, the model will misclassify both bags into *NA*, leading to a degraded performance. Further, as shown in Bag 2, even if entity-aware embedding module is absent, proposed framework merely depending on selective gate can also make a correct prediction. This finding warrants more investigation into the power of the self-attention enhanced selective gate; hence, the two error cases are shown in Bags 3 and 4.

Then, to further consider the necessity of entity-aware embedding, we show two error cases for SeG w/o Ent whose labels are */location/location/contains* and *NA* respectively in Bag 3 and 4. One possible reason for the misclassification of both cases is that, due to a lack of entity-aware embedding, the remaining position features cannot provide strong information to distinguish complex context with similar relation position pattern w.r.t the two entities.

3.4 Error Analysis

To investigate the possible reasons for misclassification, we randomly sample 50 error examples from the test set and manually analyze them. After human evaluation, we find the errors can be roughly categorized into following two classes.

Lack of background We observe that, our approach is likely to mistakenly classify relation of almost all the sentences containing two place entities to */location/location/contains*. However, the correct relation is */location/country/capital* or */location/country/administrative_divisions*. This suggests that we can incorporate external knowledge to alleviate this problem possibly caused by a lack of background.

Isolated Sentence in Bag Each sentence in a bag can be regarded as independent individual and do not have any relationship with other sentences in the bag, which possibly leads to information loss among the multiple sentences in the bag when considering classification over bag level.

4 Conclusion

In this paper, we propose a brand-new framework for distantly supervised relation extraction, i.e., selective gate (SeG) framework, as a new alternative to previous ones. It incorporates an entity-aware embedding module and a self-attention enhanced selective gate mechanism to integrate task-specific entity information into word embedding and then generates a complementary context-enriched representation for PCNN. The proposed framework has certain merits over previously prevalent selective attention when handling wrongly labeled data, especially for a usual case that there are only one sentence in the most of bags. The experiments conduct on popular NYT dataset show that our model SeG can consistently deliver a new benchmark in state-of-the-art performance in terms of all P@N and precision-recall AUC. And further ablation study and case study also demonstrate the significance of the proposed modules to handle wrongly labeled data and thus set a new state-of-the-art performance for the benchmark dataset. In the future, we plan to incorporate an external knowledge base into our framework, which may further boost the prediction quality by overcoming the problems with a lack of background information as discussed in our error analysis.

Acknowledgements

This research was funded by the Australian Government through the Australian Research Council (ARC) under grants LP180100654 partnership with KS computer. We also acknowledge the support of NVIDIA Corporation and Google Cloud with the donation of GPUs and computation credits respectively.

References

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 1247–1250. AcM.

Du, J.; Han, J.; Way, A.; and Wan, D. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. *arXiv preprint arXiv:1809.00699*.

Han, X.; Yu, P.; Liu, Z.; Sun, M.; and Li, P. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2236–2245.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; and Weld, D. S. 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-Volume 1*, 541–550. Association for Computational Linguistics.

Ji, G.; Liu, K.; He, S.; and Zhao, J. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Li, Z.; Wei, Y.; Zhang, Y.; and Yang, Q. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; and Sun, M. 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)*, 2124–2133.

Lin, Z.; Feng, M.; Santos, C. N. d.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A structured self-attentive sentence embedding. In *The International Conference on Learning Representations*.

Liu, Y.; Sun, C.; Lin, L.; and Wang, X. 2016. Learning natural language inference using bidirectional lstm model and inner-attention. *arXiv preprint arXiv:1605.09090*.

Liu, T.; Wang, K.; Chang, B.; and Sui, Z. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1790–1795.

Liu, L.; Zhou, T.; Long, G.; Jiang, J.; and Zhang, C. 2019. Learning to propagate for graph meta-learning. In *Neural Information Processing Systems (NeurIPS)*.

Long, G.; Chen, L.; Zhu, X.; and Zhang, C. 2012. TCSST: transfer classification of short & sparse text using external data. In *CIKM 2012*, 764–772.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 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*, 1003–1011. Association for Computational Linguistics.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 148–163. Springer.

Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; and Zhang, C. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Shen, T.; Zhou, T.; Long, G.; Jiang, J.; and Zhang, C. 2019. Tensorized self-attention: Efficiently modeling pairwise and global dependencies together. In *NAACL*, 1256–1266.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 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*, 455–465. Association for Computational Linguistics.

Vaswani, A.; Shazeer; Noam; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *The Neural Information Processing Systems*.

Wu, F.; Fan, A.; Baevski, A.; Dauphin, Y. N.; and Auli, M. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Ye, Z.-X., and Ling, Z.-H. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. *arXiv preprint arXiv:1904.00143*.

Yu, A. W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *The International Conference on Learning Representations*.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J.; et al. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, 2335–2344.

Zeng, D.; Liu, K.; Chen, Y.; and Zhao, J. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1753–1762.