Hierarchical Relation-Guided Type-Sentence Alignment for Long-Tail Relation Extraction with Distant Supervision

Yang Li[1], **Guodong Long**[1], **Tao Shen**[1] and **Jing Jiang**[1]
[1]Australian AI Institute, School of Computer Science, FEIT, University of Technology Sydney
yang.li-17@student.uts.edu.au,
{guodong.long,tao.shen,jing.jiang}@uts.edu.au

Abstract

Distant supervision uses triple facts in knowledge graphs to label a corpus for relation extraction, leading to wrong labeling and long-tail problems. Some works use the hierarchy of relations for knowledge transfer to long-tail relations. However, a coarse-grained relation often implies only an attribute (e.g., domain or topic) of the distant fact, making it hard to discriminate relations based solely on sentence semantics. One solution is resorting to entity types, but open questions remain about how to fully leverage the information of entity types and how to align multi-granular entity types with sentences. In this work, we propose a novel model to enrich distantly-supervised sentences with entity types. It consists of (1) a pairwise type-enriched sentence encoding module injecting both context-free and -related backgrounds to alleviate sentence-level wrong labeling, and (2) a hierarchical type-sentence alignment module enriching a sentence with the triple fact's basic attributes to support long-tail relations. Our model achieves new state-of-the-art results in overall and long-tail performance on benchmarks.

1 Introduction

Human-curated knowledge graphs (KGs), play a critical role in many downstream tasks but suffer from the incompleteness (Xiong et al., 2018; Yao et al., 2019). As a remedy, relation extraction is to distinguish the relation between two entities according to their semantics in text, but a major obstacle is a lack of sufficient labeled corpus. Fortunately, distant supervision can be used to annotate a raw text corpus via KGs for relation extraction, a.k.a. distantly supervised relation extraction (DSRE). This is based on a strong assumption that a sentence containing two entities will express the semantics of their relation in a KG (Riedel et al., 2010).

The assumption cannot always hold, leading to the wrong labeling problem. For example, both "*Jobs founded Apple*" and "*Jobs ate Apple*" are labeled with "/BUSINESS/COMPANY/FOUNDERS" according to a KG triple fact (*Steven Jobs*, /BUSINESS/COMPANY/FOUNDERS, *Apple Inc*). A basic technique for this problem is selective attention (Zeng et al., 2015; Lin et al., 2016; Ji et al., 2017) under multi-instance learning framework (Riedel et al., 2010; Hoffmann et al., 2011). Given a bag of sentences with the same entity pair, it learns to select correct one(s) by an end-to-end attention. The other major challenge is known as the long-tail problem, caused by domain mismatching during distant supervision. That is, many relation labels correspond only to a limited number of training sentences in the corpus (Ye et al., 2019). For example, in a DSRE benchmark, the distant supervision is an encyclopedic KG (i.e., Freebase (Bollacker et al., 2008)) while the corpus is news articles from the New York Times (NYT), so relations, like "/PEOPLE/PERSON/RELIGION", scarcely appear. As illustrated by Li et al. (2020b) and Zhang et al. (2019), more than 70% of relation labels in NYT can be regarded as *long-tail relations*.

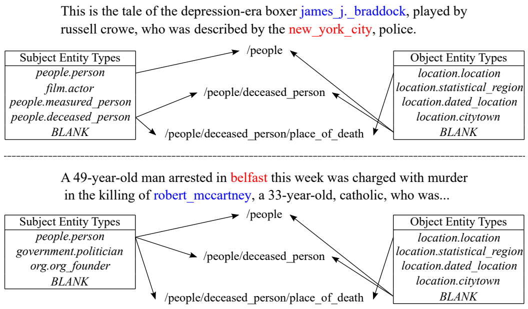
This is the tale of the depression-era boxer james_j._braddock, played by russell crowe, who was described by the new_york_city, police.
Subject Entity Types
people.person
film.actor
people.measured_person
people.deceased_person
BLANK
/people
/people/deceased_person
/people/deceased_person/place_of_death
Object Entity Types
location.location
location.statistical_region
location.dated_location
location.citytown
BLANK
A 49-year-old man arrested in belfast this week was charged with murder in the killing of robert_mccartney, a 33-year-old, catholic, who was...
Subject Entity Types
people.person
government.politician
org.org_founder
BLANK
/people
/people/deceased_person
/people/deceased_person/place_of_death
Object Entity Types
location.location
location.statistical_region
location.dated_location
location.citytown
BLANK

Figure 1: Two sentences with the same long-tail relation. For each sentence, multi-granular relations from top to bottom are pointed by its best pairwise types, which indicates not all pairwise types provide the same contribution. Blue is subject entity, and red is object entity. The 1st sentence relies on the direct pairwise types due to its relation-irrelevant semantics while the 2nd sentence integrates its relation-relevant semantics and pairwise types to enhance its representation.

To mitigate the long-tail problem, some works

Findings of the Association for Computational Linguistics: NAACL 2022, pages 316 - 326
July 10-15, 2022 ©2022 Association for Computational Linguistics

(Han et al., 2018; Zhang et al., 2019; Li et al., 2020b) resort to the hierarchy of relations for knowledge transfer from data-rich relations to the long-tail ones since the relations have coarse-grained overlap. They focus on interactive operations between hierarchical relations and intra-bag sentences, including relation-to-sentence attention (Han et al., 2018) as a hierarchical extension of selective attention, and sentence-to-relation attention (Li et al., 2020b) enriching sentences with multi-granular relations. As such, they achieve knowledge transfer by learning to distinguish coarse-grained relations for sentences with sufficient data, which provides a latent constraint for the long-tail relations. However, a coarse-grained relation usually denotes the only basic attribute of the distant oracle triple fact in KG, so a sentence scarcely contains its semantics and we can only imply the relation via background information. Again, true-labeled "*Jobs founded Apple*", does not explicitly contain any semantics of its coarse-grained relation "/BUSINESS/COMPANY", but we can directly reason it from the predicate *founded* and type of *Apple*. Thus, it is a challenge for a hierarchical DSRE model to correctly imply coarse-grained relations based solely on sentences, not to mention the existence of the wrong labeling problem.

A direct yet promising way to overcome this challenge is to incorporate extra information for entities (Vashishth et al., 2018; Hu et al., 2019; Chu et al., 2020). One popular source is the entity types, i.e., an entity's "IS A" attributes in KG, which characterizes the entity from multiple perspectives (Chen et al., 2020). As Figure 1 shows, although the 1st sentence's semantics is irrelevant to relation, the pairwise types *people.deceased_person* and *location.location* directly align with the fine grained relation. However, existing works (Vashishth et al., 2018; Chu et al., 2020) ignore this potential of explicit structured types information.

In this work, we aim to improve DSRE by exploiting structured information in the entity types from both pairwise and hierarchical perspectives to alleviate the wrong labeling and the long-tail problems respectively. To this end, we first propose a *context-free type-enriched embedding* module to generate word embeddings with pairwise types associated with the entity pair in a bag. As shown in Figure 1, even without the corresponding semantic support, pairwise types can provide direct attributes of entities to align with the relation. Besides, we develop a *context-related type-sentence alignment* module to generate robust sentence representation with pairwise types. Since entities have specific characteristic in certain semantics, we leverage semantics to select proper pairwise types and then enrich sentence representation, as the 2nd sentence in Figure 1 shows. Such an alignment is enhanced by a guidance from the relation to auto-seek for associations between pairwise types and sentences.

At the meantime, hierarchical information has been proven crucial in knowledge transfer for long-tail relations (Han et al., 2018; Zhang et al., 2019; Li et al., 2020b). Thereby, we naturally extend the base alignment module into a hierarchy by proposing a *hierarchical type-sentence alignment* module. An intuitive example in Figure 1 shows that different grained relations are pointed by various granular pairwise types. This indicates that these pairwise types contain hierarchical semantics, which makes it feasible to extend base alignment into hierarchy. Thus, the strong association between pairwise types and coarse-grained relations can improve knowledge transfer for long-tail relations.

We conduct extensive experiments on two popular benchmarks, NYT-520k and NYT-570k, showing that our model achieves new state-of-the-art overall and long-tail performance. Further analyses reveal insights into our model.

2 Approach

Task Definition. Given a bag of sentences $\mathcal{B} = \{s_1, \ldots, s_N\}$ containing a pair of subject $e^{(s)}$ and object $e^{(o)}$ entities, the distant supervision (Mintz et al., 2009) assigns the sentence bag with a relation label r according to KG triple fact. The goal of relation extraction is to predict the relation label $\hat{r}$ of an entity pair based on the corresponding sentences bag $\mathcal{B}$. Labels of coarse-grained relations, $[r^{(1)}, \ldots, r^{(M)}]$, can be derived from the mention of r. For instance, when $r =$ /BUSINESS/COMPANY/FOUNDERS, $r^{(1)} =$ /BUSINESS/COMPANY and $r^{(2)} =$ /BUSINESS. In the following, we will detail our approach, as illustrated in Figure 2.

2.1 Context-Free Type-Enriched Word Emb

Following most previous DSRE works, we first tokenize each sentence $s_j \in \mathcal{B}$ and employ a word2vec method (Mikolov et al., 2013) to derive a sequence of word embeddings by looking up a learnable matrix $\boldsymbol{W}^{(emb)} \in \mathbb{R}^{d_e \times |\mathbb{V}|}$, i.e., $\tilde{\boldsymbol{X}}^j = [\tilde{\boldsymbol{x}}_1^j, \ldots, \tilde{\boldsymbol{x}}_n^j] \in \mathbb{R}^{d_e}$, where $\mathbb{V}$ denotes word

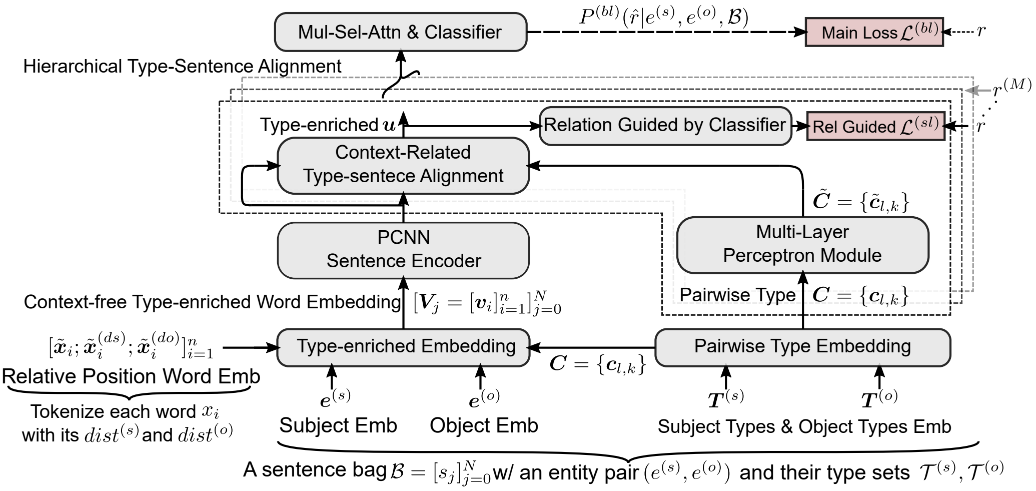
Mul-Sel-Attn & Classifier
$P^{(bl)}(\hat{r}|e^{(s)}, e^{(o)}, \mathcal{B})$
Main Loss $\mathcal{L}^{(bl)}$
r
Hierarchical Type-Sentence Alignment
$r^{(M)}$
Type-enriched $\boldsymbol{u}$
Relation Guided by Classifier
Rel Guided $\mathcal{L}^{(sl)}$
r
Context-Related Type-sentece Alignment
$\tilde{\boldsymbol{C}} = \{\tilde{\boldsymbol{c}}_{l,k}\}$
Multi-Layer Perceptron Module
PCNN Sentence Encoder
Context-free Type-enriched Word Embedding
$[\boldsymbol{V}_j = [\boldsymbol{v}_i]_{i=1}^{n}]_{j=0}^{N}$
Pairwise Type
$\boldsymbol{C} = \{\boldsymbol{c}_{l,k}\}$
$[\tilde{\boldsymbol{x}}_i; \tilde{\boldsymbol{x}}_i^{(ds)}; \tilde{\boldsymbol{x}}_i^{(do)}]_{i=1}^{n}$
Type-enriched Embedding
$\boldsymbol{C} = \{\boldsymbol{c}_{l,k}\}$
Pairwise Type Embedding
Relative Position Word Emb
Tokenize each word x_i with its $dist^{(s)}$and $dist^{(o)}$
$\boldsymbol{e}^{(s)}$
Subject Emb
$\boldsymbol{e}^{(o)}$
Object Emb
$\boldsymbol{T}^{(s)}$
$\boldsymbol{T}^{(o)}$
Subject Types & Object Types Emb
A sentence bag $\mathcal{B} = [s_j]_{j=0}^{N}$w/ an entity pair $(e^{(s)}, e^{(o)})$ and their type sets $\mathcal{T}^{(s)}, \mathcal{T}^{(o)}$

Figure 2: Our proposed model, called **Hi**erarchical **R**elation-guided Type-Sentence **A**lignment **M**odel (HiRAM), for DSRE.

vocabulary. j denotes the index of a sentence in the bag and n denotes the sentence length. In the sequel, we omit j if no confusion is caused. Then, as a common practice in DSRE (Zeng et al., 2014), a word's relative distances to both the subject and object entities (a.k.a relative positions) also play significant roles. The distances are first denoted as two integers ($dist^{(s)}$ and $dist^{(o)} \in \mathbb{Z}$) and then embedded into two learnable vectors ($\tilde{\boldsymbol{x}}_i^{(ds)}$ and $\tilde{\boldsymbol{x}}_i^{(do)} \in \mathbb{R}^{d_p}$). Therefore, the updated sequence of word embeddings is $\boldsymbol{X}^j = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]$, where $\boldsymbol{x}_i = [\tilde{\boldsymbol{x}}_i; \tilde{\boldsymbol{x}}_i^{(ds)}; \tilde{\boldsymbol{x}}_i^{(do)}] \in \mathbb{R}^{d_w}$, $[;]$ denotes vector concatenation, and $d_w := d_e + 2d_p$.

Previous works (Li et al., 2020a,b) also found that explicitly enriching each word with both entity embeddings (i.e., $\boldsymbol{e}^{(s)}$ and $\boldsymbol{e}^{(o)}$) in a context-free manner is important to DSRE's success. However, many entities scarcely appear in the raw corpus and have multi-characteristics (e.g., *Apple* could be a fruit or a company). Thus, the model is hard to distinguish the relations only via sentence semantics.

Therefore, we leverage entity types to characterize entities' attributes. That is, given an entity e, its types are defined as a set of type mentions, i.e., $\mathcal{T} = \{t_1, t_2, \dots\}$. However, previous works (Chu et al., 2020) directly concatenate the entity types of both $e^{(s)}$ and $e^{(o)}$, completely regardless of potentials of explicit structured information of types. As demonstrated by Krompaß et al. (2015), a relation in KG is usually constrained by the entity types of $e^{(s)}$ and $e^{(o)}$ simultaneously (i.e., pairwise types), instead of their individuals. We thereby propose a pairwise type embedding module to enrich the word embedding $\boldsymbol{X}$ also in a context-free manner.

Type and Pairwise Type Embedding. First, given an entity type set $\mathcal{T} = \{t_1, t_2, \dots\}$ (either $\mathcal{T}^{(s)}$ for subject or $\mathcal{T}^{(o)}$ for object), we tokenize each type mention t_j into a sequence of words, then embed the words by looking up $\boldsymbol{W}^{(emb)}$, and lastly derive the type embedding $\boldsymbol{t}_j$ by applying a mean-pooling to the word embeddings of the mention. The embedding of the entire type is

$$\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \dots] \in \mathbb{R}^{|\mathcal{T}| \times d_e}. \tag{1}$$

As such, we subsequently define the embedding of the pairwise type by considering a combination of every subject $\forall t_l^{(s)} \in \mathcal{T}^{(s)}$ and object type $\forall t_k^{(o)} \in \mathcal{T}^{(o)}$. Instead of sole semantics via a vector concatenation, we take into account the prior structured information in each type pair by leveraging a translational scheme (Bordes et al., 2013). Hence, we represent each type pair $(t_l^{(s)}, t_k^{(o)})$ as

$$\begin{aligned} \boldsymbol{c}_{l,k} &= [\tilde{\boldsymbol{c}}_{l,k}^{(sem)}; \tilde{\boldsymbol{c}}_{l,k}^{(str)}] \in \mathbb{R}^{4d_e}, \\ \text{where, } \tilde{\boldsymbol{c}}_{l,k}^{(sem)} &= \boldsymbol{t}_l^{(s)} \odot \boldsymbol{W}^{(sem)} \boldsymbol{t}_k^{(o)}, \\ \text{and } \tilde{\boldsymbol{c}}_{l,k}^{(str)} &= \boldsymbol{t}_k^{(o)} - \boldsymbol{t}_l^{(s)}. \end{aligned} \tag{2}$$

Here, "$\odot$" denotes Hadamard product, and $\boldsymbol{W}^{(sem)}$ denotes a learnable projection. $\tilde{\boldsymbol{c}}_{l,k}^{(sem)}$ aims to capture the prior semantic relation in the pair (Nickel et al., 2011) since not all types combinations are valid in the whole dataset. $\tilde{\boldsymbol{c}}_{l,k}^{(str)}$ aims to measure its structured relation. Lastly, we denote all the embeddings of pairwise types as

$$\boldsymbol{C} = \{\boldsymbol{c}_{l,k}\}_{\forall l \in [1, |\mathcal{T}^{(s)}|], \forall k \in [1, |\mathcal{T}^{(o)}|]}, \tag{3}$$

where $\boldsymbol{C} \in \mathbb{R}^{4d_e \times m}$ and $m = |\mathcal{T}^{(s)}| \cdot |\mathcal{T}^{(o)}|$.

Type-Enriched Word Embedding. However, an open question still remains about how to operate on variable-length embeddings of pairwise types, $\boldsymbol{C}$, to enrich each word embedding, $\boldsymbol{x}_j \in \boldsymbol{X}$, in a context-free manner. Inspired by self-attentive sentence encoding (Lin et al., 2016), we present a bag-level type-attentive module, which compresses $\boldsymbol{C}$ into a single vector representation to facilitate type-enriching. Intuitively, such self-attentive module is focused on the prior knowledge of the type pair in the corpus. Formally, we first generate a global query (Lin et al., 2016) with structured information of both entities and types to retrieve possible prior pairwise types, i.e.,

$$\tilde{\boldsymbol{q}}^{(f)} = [\boldsymbol{e}^{(o)};\text{Pool}(\boldsymbol{T}^{(o)})] - [\boldsymbol{e}^{(s)};\text{Pool}(\boldsymbol{T}^{(s)})], \quad (4)$$

followed by a standard Bilinear-based attention,

$$\boldsymbol{q}^{(f)} = \boldsymbol{C} \cdot \text{softmax}(\boldsymbol{C}^T \boldsymbol{W}^{(sa)} \boldsymbol{q}^{(f)}) \in \mathbb{R}^{4d_e}, \quad (5)$$

where "$\cdot$" denotes matrix multiplication and $\boldsymbol{W}^{(sa)}$ is a learnable weight matrix. Lastly, we use a gate as in (Li et al., 2020b) to derive the context-free type-enriched word embedding, i.e.,

$$\boldsymbol{g}_i^{(gf)} = \text{Sigmoid}(\text{MLP}([\boldsymbol{x}_i; \boldsymbol{q}^{(f)}]; \theta^{(gf1)})), \quad (6)$$

$$\boldsymbol{x}_i^{(gf)} = \text{MLP}([\boldsymbol{x}_i; \boldsymbol{q}^{(f)}]; \theta^{(gf2)}), \quad (7)$$

$$\boldsymbol{v}_i = \boldsymbol{g}_i^{(gf)} \odot \boldsymbol{x}_i + (\boldsymbol{1} - \boldsymbol{g}_i^{(gf)}) \odot \boldsymbol{x}_i^{(gf)}, \quad (8)$$

where MLP denotes a multi-layer perceptron (MLP) module. Hence, word embeddings for s are updated to $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_n] \in \mathbb{R}^{d_w \times n}$.

2.2 Context-Related Type-Sent Alignment

Sentence Encoding. In DSRE, piecewise convolutional neural network (PCNN) (Zeng et al., 2015) is used for sentence embedding. 1D-CNN (Kim, 2014) is first invoked over $\boldsymbol{V}$ for contextualized representations. Then a piecewise max-pooling performs over the output sequence to obtain sentence-level embedding with highlighted entity positions:

$$\boldsymbol{H} = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_n] = \text{1D-CNN}(\boldsymbol{V}; \theta^{(cnn)}),$$

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

where $\boldsymbol{H}^{(1)}$, $\boldsymbol{H}^{(2)}$ and $\boldsymbol{H}^{(3)}$ are three consecutive parts of $\boldsymbol{H}$ by dividing $\boldsymbol{H}$ w.r.t. the indices of subject $e^{(s)}$ and object $e^{(o)}$ entities. Consequently, $\boldsymbol{s} \in \mathbb{R}^{d_h}$ is the resulting sentence-level embedding.

Type-Sentence Alignment. Consider that types are not comprehensive enough to align with multi-granular relations, we leverage semantic context to select valid pairwise types for generating robust sentence representation. Hence, we first calculate alignment scores between a sentence $\boldsymbol{s} \in \mathbb{R}^{d_h}$ and the embeddings of pairwise types $\boldsymbol{C} \in \mathbb{R}^{4d_e \times m}$ by using a simple Bilinear layer, i.e.,

$$\tilde{\boldsymbol{C}} = \text{MLP}(\boldsymbol{C}; \theta^{(p)}) \in \mathbb{R}^{d_h \times m}, \quad (9)$$

$$\boldsymbol{a} = \text{softmax}(\tilde{\boldsymbol{C}}^T \boldsymbol{W}^{(al)} \boldsymbol{s}) \in \mathbb{R}^{m}. \quad (10)$$

Then, we enrich the sentence embedding with the aligned type pairs via another gating mechanism:

$$\boldsymbol{z} = \tilde{\boldsymbol{C}} \cdot \boldsymbol{a} \quad (11)$$

$$\boldsymbol{g} = \text{Sigmoid}(\text{MLP}([\boldsymbol{s}; \boldsymbol{z}]; \theta^{(g)})), \quad (12)$$

$$\tilde{\boldsymbol{u}} = \boldsymbol{g} \odot \boldsymbol{s} + (1 - \boldsymbol{g}) \odot \boldsymbol{z}. \quad (13)$$

Lastly, following previous success (Li et al., 2020b; Devlin et al., 2019), we leverage a residual connection (He et al., 2016) with layer normalization (Ba et al., 2016) to derive the final context-related type-enriched sentence embedding, i.e.,

$$\boldsymbol{u} = \text{LayerNorm}(\boldsymbol{s} + \tilde{\boldsymbol{u}}; \theta^{(lm)}). \quad (14)$$

Relation-Guided Alignment at the Sentence Level. Due to the severe wrong labeling problem at the sentence level, previous DSRE works usually skip over sentence-level relation supervisions. Fortunately, empowered by the proposed context-free type enrichment and context-related type-sentence alignment, we can utilize the sentence-level relation label even if the relation label is wrong. The reason for this is that, a sentence has already been equipped with structured background to support sentence-level relation even if the sentence semantics cannot deliver the relation. We applied an MLP-based neural classifier to the type-enriched sentence embedding, $\boldsymbol{u}$, to determine the relation at the sentence level, i.e.,

$$P^{(sl)}(\hat{r}|\boldsymbol{u}) = \text{softmax}(\text{MLP}(\boldsymbol{u}; \theta^{(sl)})), \quad (15)$$

where, $P^{(sl)}(\hat{r}|\boldsymbol{u})$ is a categorical distribution over all possible relations. Hence, the training objective is to minimize the cross-entropy loss,

$$\mathcal{L}^{(sl)} = -\sum_{\mathcal{D}} \sum_{\mathcal{B}} \log P^{(sl)}(\hat{r} = r|\boldsymbol{u}), \quad (16)$$

where $\mathcal{D}$ denotes a DSRE dataset consisting of sentence bags $\mathcal{B}$. The guidance from the sentence-level relation leads to strong type-sentence alignment (as illustrated in §3.1 and §3.2). As a result, the sentence-level wrong labeling problem is alleviated. In contrast, previous works w/ sentence-level relation supervisions (Li and Roth, 2002) suffer from the confirmation bias problem (Chen et al., 2019) caused by the sentence-level wrong labeling.

2.3 Hierarchical Type-Sentence Alignment

Inspired by former works (Han et al., 2018; Zhang et al., 2019; Li et al., 2020b) for handling long-tail relations, we also extend our basic model into hierarchy. However, the basic attributes contained by coarse-grained relation are irrelevant to the semantics in sentences. Thus, instead of direct operating on the hierarchy of relations (i.e., from fine-grained r to coarse-grained $[r^{(1)} \dots r^{(M)}]$ relations), we leverage coarse-grained entity types describing the domain/type properties of the entities in the triple facts to enrich each sentence via the guidance from coarse-grained relation.

Formally, we adapt the relation-guided type-sentence alignment (§2.2) into hierarchy, which shares a high-level inspiration with multi-head attention (Vaswani et al., 2017). First, we reuse the architecture from Eq.(9-14) by defining

$$\begin{aligned} \boldsymbol{a}^{(l)}, \tilde{\boldsymbol{C}}^{(l)} &= \text{TS-Align}^{(l)}(\boldsymbol{s}, \boldsymbol{C}), \ \forall l \in [1, M], \\ \boldsymbol{u}^{(l)} &= \text{TS-Integrate}^{(l)}(\boldsymbol{a}^{(l)}, \tilde{\boldsymbol{C}}^{(l)}, \boldsymbol{s}), \end{aligned} \tag{17}$$

where $\text{TS-Align}()$ denotes Eq.(9-10) to obtain type-sentence alignment $\boldsymbol{a}^{(l)}$ and $\text{TS-Integrate}()$ denotes Eq.(11-14) to generate enriched sentence representation $\boldsymbol{u}^{(l)}$ at level l. Note that, these modules are parameter-untied from each other. Then, we update the sentence-level relation-guided loss in Eq.(16) to its hierarchical version, i.e.,

$$\mathcal{L}^{(sl)} = -\sum_{\mathcal{D},\mathcal{B},l \in [1,M]} \log P^{(sl)}(\hat{r}^{(l)} = r^{(l)} | \boldsymbol{u}^{(l)}) \tag{18}$$

Again, learnable parameters of the sentence-level classifiers across l are also untied. Lastly, we obtain the hierarchical type-enriched representation, i.e.,

$$\boldsymbol{u}^{(h)} = [\boldsymbol{u}; \boldsymbol{u}^{(1)}; \dots; \boldsymbol{u}^{(M)}] \in \mathbb{R}^{(1+M)d_h}. \tag{19}$$

Different to previous works (Han et al., 2018; Zhang et al., 2019; Li et al., 2020b) focusing on hierarchical relation embeddings, our work explores the constraints by pairwise types for relations to mitigate sentence-level wrong labeling and uses the hierarchy of entity types on par with that of the relation to improve long-tail performance.

2.4 Relation Classification and Objectives

Lastly, we put the sentences back into the bag and derive bag-level embedding for the final relation classification. Hence, for a bag $\mathcal{B} = [s_1, ...s_N]$, we can obtain sentence embeddings of all the sentences $\boldsymbol{U}^{(h)} = [\boldsymbol{u}_1^{(h)}, \dots, \boldsymbol{u}_N^{(h)}]$, where $\boldsymbol{u}_j^{(h)}$ is hierarchical type-enriched sentence encoding derived from Eq.(19). To preserve the hierarchical information learned in $\boldsymbol{u}_j^{(h)}$, we proposed to apply multiple selective modules to its different parts, i.e.,

$$\begin{aligned} &\boldsymbol{b} = \text{Mul-Sel-Attn}(\boldsymbol{U}^{(h)}) = [\boldsymbol{b}^{(0)}; \boldsymbol{b}^{(1)}; \dots; \boldsymbol{b}^{(M)}], \\ &\boldsymbol{b}^{(0)} = \text{Sel-Attn}([\boldsymbol{u}_1; \dots, \boldsymbol{u}_N]), \\ &\boldsymbol{b}^{(l)} = \text{Sel-Attn}([\boldsymbol{u}_1^{(l)}; \dots, \boldsymbol{u}_N^{(l)}]), \ \forall l \in [1, M]. \end{aligned}$$

where, $\text{Sel-Attn}()$ represents the selective attention among the sentences in each granular relation, and $\text{Mul-Sel-Attn}()$ represents the selective attention among the multi-granular bag representations. For bag representation, $\boldsymbol{b}^{(0)}$ denotes the finest grained and $\boldsymbol{b}^{(l)}$ denotes coarser grained. Lastly, we use an MLP-based classifier upon $\boldsymbol{b}$ to derive a bag-level categorical distribution, i.e.,

$$P^{(bl)}(\hat{r} | e^{(s)}, e^{(o)}, \mathcal{B}). \tag{20}$$

Meanwhile, the corresponding training loss is

$$\mathcal{L}^{(bl)} = -\sum_{\mathcal{D}} P^{(bl)}(\hat{r} = r | e^{(s)}, e^{(o)}, \mathcal{B}). \tag{21}$$

Therefore, the final training objective is to minimize a linear combination of both sentence-level in Eq.(16) and bag-level (in Eq.(21)) losses, i.e.,

$$\mathcal{L} = \mathcal{L}^{(bl)} + \beta \mathcal{L}^{(sl)}. \tag{22}$$

3 Experiments

Datasets. We evaluate our HiRAM on DSRE benchmarks, New York Times – NYT (Riedel et al., 2010), including NYT-520K and NYT-570K. NYT datasets have 53 distinct relations, including an *NA* class denoting the unavailable relation between entity pairs. Each relation includes two coarse-grained relations (i.e., $M = 2$), and the number of relations from fine to coarse are 53, 36 and 9. NYT-520K and NYT-570K have the same testing set containing 172,488 sentences, with 96,678 entity pairs. The only difference is that there is an overlap of 11,416 entity pairs between training and testing in NYT-570K. Thus, NYT-520K has severer wrong labeling and long-tail problems.

P@N (%)	**One**				**Two**				**All**				**AUC**
	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean	
Comparative Approaches													
CNN+ATT (Lin et al., 2016)	76.2	65.2	60.8	67.4	76.2	65.7	62.1	68.0	76.2	68.6	59.8	68.2	-
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	0.341
CoRA (Li et al., 2020b)	78.0	69.0	66.0	71.0	79.0	72.0	66.3	72.4	81.0	74.0	68.3	74.4	0.344
RESIDE (Vashishth et al., 2018)	80.0	75.5	69.3	74.9	83.0	73,5	70.6	75.7	84.0	78.5	75.6	79.4	-
InSRL (Chu et al., 2020)	-	-	-	-	-	-	-	-	-	-	-	-	0.451
HiRAM	**93.0**	**89.0**	**83.0**	**88.3**	**93.0**	**88.5**	**84.0**	**88.5**	**93.0**	**88.5**	86.0	**89.2**	**0.484**
Ablations													
HiRAM w/o Hierarchy in §2.3	88.0	84.5	83.0	85.2	90.0	86.0	85.0	87.0	90.0	86.5	85.0	87.2	0.450
HiRAM w/o CF in §2.1	78.0	75.5	74.3	75.9	87.0	76.5	74.0	79.2	87.0	77.5	74.7	79.7	0.425
HiRAM w/o Rel Guidance in Eq. 16	89.0	86.0	76.7	83.9	93.0	88.0	81.7	87.6	93.0	87.0	86.7	88.9	0.482
HiRAM w/ TC	84.0	82.0	75.3	80.4	85.0	81.5	79.7	82.1	89.0	82.5	78.0	83.2	0.462
RoBERTa (Liu et al., 2019)	44.0	46.5	43.3	44.6	38.0	39.5	38.7	38.7	33.0	36.5	37.7	35.7	0.301
RoBERTa w/ CF	80.0	76.0	74.0	76.7	81.0	78.5	76.0	78.5	81.0	76.0	75.0	77.3	0.488
RoBERTa w/ HiRAM	85.0	83.0	79.3	82.4	86.0	85.5	81.3	84.3	89.0	86.0	81.7	85.6	0.518

Table 1: Model Evaluation and ablation study on NYT-520K. "P@N" denotes precision values for the entity pairs with the top-100, -200 and -300 prediction confidences by randomly keeping one/two/all sentence(s) in each bag. The abbreviation "CF" represents **C**ontext-**F**ree embedding in §2.1; "TC" represents **T**ype **C**oncatenation replacing CF. "RoBERTa" directly predicts relations via [CLS] token. "RoBERTa w/ CF" adds context-free type-enriched word embedding module on the output of RoBERTa to generate sentences representation. "RoBERTa w/ HiRAM" denotes the combination of HiRAM and RoBERTa.

P@N (%)	**One**				**Two**				**All**				**AUC**
	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean	
Comparative Approaches													
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	0.42
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	0.42
SeG (Li et al., 2020a)	94.0	89.0	85.0	89.3	91.0	89.0	87.0	89.0	93.0	90.0	86.0	89.3	0.51
CoRA (Li et al., 2020b)	94.0	90.5	82.0	88.8	98.0	91.0	86.3	91.8	**98.0**	92.5	88.3	92.9	0.53
HiRAM	**96.0**	**91.5**	**85.7**	**91.1**	**98.0**	**94.5**	**89.3**	**93.9**	**98.0**	**95.0**	**92.3**	**95.8**	**0.580**

Table 2: Model Evaluation on NYT-570K, published by PCNN+HATT (Han et al., 2018)

.

Evaluation Metrics. Following previous works (Lin et al., 2016; Han et al., 2018; Zhang et al., 2019; Li et al., 2020b; Chu et al., 2020), we use area under precision-recall curve (AUC) and top-N precision (P@N) to measure models' performance with the disturbance of wrong labeling, and use Hits@K to measure the performance on long-tail relations. AUC measures the ability of relation classification, while P@N measures the precision of high-confidence predictions ranked by the model.

Settings. For both versions of NYT datasets, d_e, d_p, d_w, d_h and M are 50, 5, 60, 690, and 2 respectively. The types number of each entity is various but we set an upper limit and pad BLANK as a choice. We use AdaDelta (Zeiler, 2012) with 0.1 learning rate. Batch size is 160 with 15 epochs and 5-th is the best, dropout probability is 0.5, weight decay of L2-reg is 10^{-5}. We use random initialization or RoBERTa-base to initialize our models.

Comparative Approach. We compare our HiRAM with many strong competitors, including **(1) PCNN+ATT** (Lin et al., 2016) proposes a selective attention to alleviate wrong labeling. **(2) PCNN+HATT** (Han et al., 2018) extends selective attention with hierarchical relations. **(3) RESIDE** (Vashishth et al., 2018) leverages side KGs' information to improve DSRE. **(4) PCNN+BAG-ATT** (Ye and Ling, 2019) proposes intra-bag and inter-bag attentions to handle the wrongly labeled sentences. **(5) PCNN+KATT** (Zhang et al., 2019) integrates externally pre-trained graph embeddings with relation hierarchies for long-tail relations. **(6) SeG** (Li et al., 2020a) focuses on one-sentence bags and proposes entity-aware embedding. **(7) CoRA** (Li et al., 2020b) transfers multi-granular relations features into sentences in hierarchies for long-tail relations. **(8) InSRL** (Chu et al., 2020) integrates sentence, entity description and types together via intact space representation learning.

3.1 Overall Performance on Benchmarks

As shown in Tables 1 and 2, HiRAM outperforms former baselines on NYT-570K. Different from CoRA's poor performance on NYT-520K, HiRAM achieves a new state-of-the-art on both popular

# Training Instance	**<100**			**<200**		
Hits@K (Macro)	**10**	**15**	**20**	**10**	**15**	**20**
PCNN+ATT (Lin et al., 2016)	<5.0	7.4	40.7	17.2	24.2	51.5
PCNN+HATT* (Han et al., 2018)	29.6	51.9	61.1	41.4	60.6	68.2
PCNN+KATT* (Zhang et al., 2019)	35.3	62.4	65.1	43.2	61.3	69.2
CoRA* (Li et al., 2020b)	66.6	72.0	87.0	72.7	77.3	89.4
CoRA (Li et al., 2020b)	66.6	66.6	75.9	71.7	72.7	80.3
HiRAM	**72.2**	**96.3**	**96.3**	**77.3**	**96.9**	**96.9**
HiRAM w/o Hierarchy in §2.3	50.0	88.9	92.6	59.1	90.9	93.9
HiRAM w/o CF in §2.1	66.6	88.9	92.6	72.7	90.9	93.9
HiRAM w/o Rel Guidance in Eq. 16	55.6	66.7	88.9	63.6	72.7	90.9
HiRAM w/ TC	72.2	77.7	88.9	77.3	81.8	90.9
RoBERTa (Liu et al., 2019)	0	0	0	0	0	11.6
RoBERTa w/ HiRAM	38.8	61.1	66.6	50.0	54.5	72.7

Table 3: Hits@K (Macro) tests only on the relations whose number of training instance $<$ 100/200. "Hits@K" denotes whether a test sentence bag whose gold relation label $r^{(0)}$ falls into top-K relations ranked by their prediction confidences."Macro" denotes macro average is applied regarding relation labels. "*" denotes the model is trained on NYT-570K.

Case Sentence 1: although the regime of president **bashar_al-assad** hails from an obscure offshoot of shiism – the alawites – syria is nearly three-quarters sunni, with alawites, members of other **muslim** sects and ...		
$r^{(2)}$: /people	$r^{(1)}$: /people/person	$r^{(0)}$: /people/person/religion
Case Sentence 2: having so many operating systems makes it expensive to make software **, said faraz_hoodbhoy**, the chief executive of camera phones save and share multimedia content.		
$r^{(2)}$: /business	$r^{(1)}$: /business/company	$r^{(0)}$: /business/company/founder

Table 4: Two cases with long-tail relations are mis-classified by previous works whereas HiRAM is competent. Analysis of the attention probability shown in Figure 3 proves the utility of context-related type-sentence alignment with relation guidance.

benchmarks in P@N and AUC. Compared with InSRL integrating both clean entity types' concatenation and accurate entity descriptions, HiRAM increases the AUC score by nearly 7%, verifying the capability of our specific model designer.

3.2 Ablation Study

We conduct an ablation study on NYT-520K, as shown at the bottom of Table 1. Compared to HiRAM, "HiRAM w/o Hierarchy" drops 6% in AUC. "HiRAM w/o Rel Guidance" performs well on P@N and AUC but has huge gap in P@One, which represents that the relation-Guided alignment in hierarchy can empower sentence representation with less data in Multi-instance Learning. Meanwhile, top-n precision of "HiRAM w/o CF" drops by nearly 10.5%. To prove the superiority of our specific design, we replace the pairwise type in §2.1 with simple type concatenation. The AUC score of "HiRAM w/ TC" decreases by 4.5% and nearly 5.6% of top-n precision. To further emphasize our word embedding §2.1 is module-agnostic, we combine RoBERTa (Liu et al., 2019) with our module respectively. As the bottom panel shows, "RoBERTa w/ CF" makes great progress, and "RoBERTa w/ HiRARM" achieves the best performance among three RoBERTa-related experiments. However, due to the strong ability of RoBERTa model, the wrong labeling problem hurt the performance severely, especially in P@N.

3.3 Performance on Long-Tail Relations

Since former baselines are mainly trained on NYT-570K, we reproduce CoRA on NYT-520K for fair comparison as shown in Table 3. HiRAM achieves a new state-of-the-art result in Hits@K with 20% superiority. Removing hierarchy module in §2.3, the performance of "HiRAM w/o Hierarchy" decreases by nearly 30% on Hits@10 but is better than baselines in other settings, verifying the importance of hierarchical model for long-tail relations. The huge decline of "HiRAM w/o Rel Guidance" verifies the necessity of relation guidance. Due to lacks of plenty reliable training data, RoBERTa is hard to handle the long-tail problem but our specific modules further increase its performance.

3.4 Case Study and Error Analysis

Firstly, we conduct a case study to qualitatively analyze the effect of our model in §2.3 The case study of two samples are shown in Table 4 and the type-sentence alignment distribution is shown in

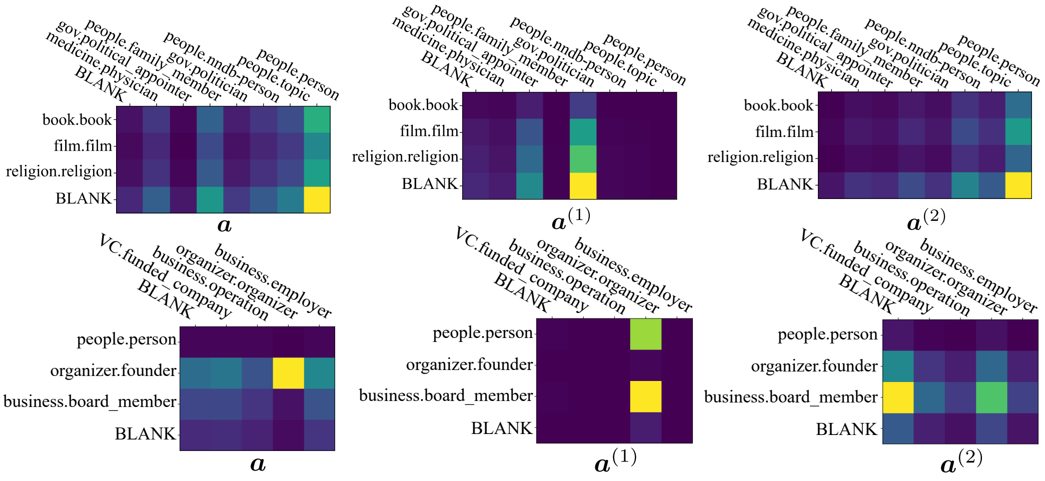

Figure 3: Each heatmap represents the distribution of type-sentence alignment $\boldsymbol{a}$ in Eq.(10) and $\boldsymbol{a}^{l}$ in Eq.(17). The horizontal axis represents the types of subject entity, and the vertical axis represents the types of object entity. The top row, from left to right, represents three alignment distributions of first case, and the bottom row represents three alignment distributions of second case, as Table 4 shows. Notice that "VC" is the abbreviation of venture captial.

Figure 3. Secondly, we investigate the possible reasons for the misclassifications of HiRAM.

Distribution of Type-Sentence Alignment. For the first case, despite the failure in expressing the long-tail relation "/PEOPLE/PERSON/RELIGION", the selected pairwise types are sufficient to predict this relation. As the top row of Figure 3 shows, *people.person* with *BLANK* helps to identify the character of subject entity, and *religion.religion* with high alignment score can provide direct attributes. For the second case, the semantics is implicitly related to its long-tail relation "/BUSINESS/COMPANY/FOUNDER". The proper pairwise types are selected by coarser relation guidance, like (*organizer.organizer*, *organizer.founder*).

Error Analysis. To analyse the implicit reasons for wrong predictions, we have manually checked several randomly-sampled error test examples. 1) Most of error cases are annotated as /PEOPLE/PERSON/PLACE_OF_BIRTH because the semantics and the relation may be completely irrelevant and the types are hard to maintain people's birth place. 2) Mean pooling in Eq.(4) might not be the most optimal way to replace entity itself when the entity has too many characters.

4 Related Work

Wrong Labeling Problem. Many works (Liu et al., 2016; Ji et al., 2017; Ye and Ling, 2019; Li et al., 2020a) propose various extensions of vanilla selective attention (Lin et al., 2016). Ye and Ling (2019) combine intra-/inter-bag level selective attention for DSRE. For one-sentence bags, Li et al. (2020a) design the entity-aware embedding in a context-free manner with a gate mechanism.

Long-tail Relations. Knowledge transfer via hierarchical relations is effective. Han et al. (2018) design relation-to-sentence attention in hierarchies, and Li et al. (2020b) modify it to sentence-to-relation attention. Many works (Vashishth et al., 2018; Hu et al., 2019; Chu et al., 2020) resort to extra knowledge, i.e., entity description and entity types. Entity description (Hu et al., 2019; Chu et al., 2020) mainly stems from the Wikipedia page, which contains factual statements of the relation with other entities. Such oracle knowledge can boost DSRE performance but is impractical.

5 Conclusion

In this work, we propose a new model, HiRAM, training on a single Titan XP, except for RoBERTa w/ RTX 6000, to alleviate wrong labeling and long-tail problems in DSRE. For the wrong labeling problem, we propose a context-free type-enriched word embedding to enrich each word with prior knowledge and a context-related type-sentence alignment module to complement sentences with semantics-fitted pairwise types. For the long-tail problem, we extend the base alignment into the hierarchy to utilize the multi-granular entity types. The experiments with extensive analyses show the superiority of HiRAM.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Bo Chen, Xiaotao Gu, Yufeng Hu, Siliang Tang, Guoping Hu, Yueting Zhuang, and Xiang Ren. 2019. Improving distantly-supervised entity typing with compact latent space clustering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2862–2872. Association for Computational Linguistics.

Shuang Chen, Jinpeng Wang, Feng Jiang, and Chin-Yew Lin. 2020. Improving entity linking by modeling latent entity type information. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7529–7537. AAAI Press.

Zhendong Chu, Haiyun Jiang, Yanghua Xiao, and Wei Wang. 2020. Insrl: A multi-view learning framework fusing multiple information sources for distantly-supervised relation extraction. *CoRR*, abs/2012.09370.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2236–2245. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

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-Volume 1*, pages 541–550. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3819–3827. 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, February 4-9, 2017, San Francisco, California, USA*, pages 3060–3066. AAAI Press.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Denis Krompaß, Stephan Baier, and Volker Tresp. 2015. Type-constrained representation learning in knowledge graphs. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I*, volume 9366 of *Lecture Notes in Computer Science*, pages 640–655. Springer.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *ACL*.

Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2020a. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second In-*

novative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8269–8276. AAAI Press.

Yang Li, Tao Shen, Guodong Long, Jing Jiang, Tianyi Zhou, and Chengqi Zhang. 2020b. Improving long-tail relation extraction with collaborating relation-augmented attention. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1653–1664. International Committee on Computational Linguistics.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional LSTM model and inner-attention. *CoRR*, abs/1605.09090.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 1003–1011. The Association for Computer Linguistics.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 809–816. Omnipress.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III*, volume 6323 of *Lecture Notes in Computer Science*, pages 148–163. Springer.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha P. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 1257–1266. Association for Computational Linguistics.

Ashish Vaswani, Shazeer, Noam, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *The Neural Information Processing Systems*.

Shengwu Xiong, Weitao Huang, and Pengfei Duan. 2018. Knowledge graph embedding via relation paths and dynamic mapping matrix. In *Advances in Conceptual Modeling - ER 2018 Workshops Emp-ER, MoBiD, MREBA, QMMQ, SCME, Xi'an, China, October 22-25, 2018, Proceedings*, volume 11158 of *Lecture Notes in Computer Science*, pages 106–118. Springer.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *CoRR*, abs/1909.03193.

Qinyuan Ye, Liyuan Liu, Maosen Zhang, and Xiang Ren. 2019. Looking beyond label noise: Shifted label distribution matters in 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3839–3848. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2810–2819. Association for Computational Linguistics.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1753–1762. The Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2335–2344. ACL.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3016–3025. Association for Computational Linguistics.