计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.20210445
59(12):2794-2802,2022

# 融合实体外部知识的远程监督关系抽取方法

高建伟　万怀宇　林友芳
(北京交通大学计算机与信息技术学院　北京　100044)
(gaojianwei@bjtu.edu.cn)

# Integrating External Entity Knowledge for Distantly Supervised Relation Extraction

Gao Jianwei, Wan Huaiyu, and Lin Youfang
(*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044)

**Abstract**　Distantly supervised relation extraction aims to find the relational facts from unstructured texts, which is meaningful for many downstream tasks. Although distant supervision can automatically generate labeled training instances, it inevitably suffers from the wrong label problem. Current works mostly focus on the denoising process, trying to generate a more effective bag-level representation by selecting valid sentences. Nevertheless, there is a large amount of entity knowledge that can help the model to understand the relationship between entities, and these kinds of knowledge have not been fully utilized. Based on this observation, in this paper, we propose a novel distantly supervised relation extraction approach that exploits external entity knowledge to enhance the model's expressive ability. In the model, the knowledge-aware word embeddings are generated to enrich the sentence level representations by introducing both structure knowledge from external knowledge graphs and semantic knowledge from corpus. The experimental results demonstrate that our proposed approach outperforms state-of-the-art the methods on both versions of a large-scale benchmark New York Times dataset. Besides, the differences between the two versions of dataset are also investigated through further comparative experiments, in which the dataset with no entity intersection can move effectively reflect model performance.

**Key words**　relation extraction; distant supervision; knowledge representation; word embedding; gating mechanism

**摘　要**　远程监督关系抽取旨在从无结构化的文本当中发现关系事实,它对许多下游任务有着非常重要的意义.虽然远程监督可以自动地生成大量带标签的训练样本,但是自动标注的过程不可避免地会遇到噪声数据的问题.当前的许多研究工作主要把关注点放在降噪的过程当中,尝试通过选择出正确的句子来生成更有效的包级别特征表示.但是在文本语料之外,还存在着大量与实体相关的外部知识没有被充分利用,而这些知识能够帮助模型更好地理解实体之间的关系.基于这一观察,提出了一种新颖的远程监督关系抽取方法,该方法通过利用外部知识图谱当中的结构化知识和文本语料中的语义知识,设计了一种实体知识感知的词嵌入表示方法,来丰富句子级别的特征表达能力.实验结果表明,在2个版本的大规模"纽约时报"基准数据集上,该方法都明显优于其他方法.此外,还通过对比实验进一步探索了2个版本的数据集所存在的差异,其中无实体交集的数据集能够更有效地反映模型性能.

**收稿日期**:2021-05-07;**修回日期**:2022-01-13
**基金项目**:国家重点研发计划项目(2018YFC0830200)
This work was supported by the National Key Research and Development Program of China (2018YFC0830200).
**通信作者**:万怀宇(hywan@bjtu.edu.cn)

**关键词** 关系抽取;远程监督;知识表示;词嵌入;门控机制

**中图法分类号** TP391

作为自然语言处理领域中一个重要的基础研究课题,关系抽取旨在从无结构化的文本当中预测出给定实体对之间的关系事实.例如,从表 1 的第 1 行句子中,我们可以抽取出实体对 Apple 和 Steve Jobs 之间的关系是创始人.

**Table 1 An Example of Sentences in a Bag Labeled by Distant Supervision**

**表 1 通过远程监督进行标记的样本示例**

| 标签 | 实体对包中的句子 | 标签是否正确 |
|---|---|---|
| /founder | Steve Jobs and Wozniak co-founded Apple in 1976. | 是 |
| | Steve Jobs was the co-founder and CEO of Apple. | 是 |
| | Steve Jobs argued with Wozniak, the co-founder of Apple. | 否 |

注:"是"和"否"表示该句子是否真的表达了该关系.

通常,大多数传统的关系抽取模型[1-3]都采用有监督学习的方法来进行训练,然而这一过程往往需要大量的高质量标注样本来进行支撑,非常的耗费人力.Mintz 等人[4]提出使用远程监督的方法来缓解缺乏训练数据的问题,该方法可以通过将知识图谱(knowledge graph, KG)中的实体对与文本中相应的实体对进行对齐来自动生成带标签的训练样本.关系抽取中的远程监督是基于这样的假设来定义的:若在给定 KG 中的 2 个实体之间存在关系事实,那么我们认为所有包含该相同实体对的句子都表达了对应的关系.因此,远程监督方法所具有的这种强有力的假设不可避免地会伴随着错误标记的问题,从而导致了噪声数据的产生.

因此,Riedel 等人[5]提出了一个使用多示例学习(multi-instance learning, MIL)框架的方法来缓解噪声数据的问题,这一方法提出了一种叫作"expressed-at-least-once"的假设来缓解先前约束较强的假设条件.该方法假设,在所有包含相同实体对的句子当中,至少有 1 个确实表达了它们的关系.在多示例学习框架当中,关系抽取的目标从句子级别变成了为包级别,其中每个包是由一组包含相同实体对的句子所组成的集合.此后,有许多研究者都受到该工作的启发,基于 MIL 框架开展了一系列的研究工作来提高模型选择有效句子的能力[6-9].其中,Lin 等人[9]提出的选择性注意力框架使用注意力机制来为句子分配权重,从而能够充分地利用所有句子中所包含的信息.近些年来,基于选择性注意力框架,提出了一系列新的关系抽取模型[10-13],它们大多使用卷积神经网络(convolutional neural network, CNN)来作为句子编码器,证明了这一结构的稳定性和有效性.

然而,尽管上述框架结构被广泛使用在远程监督关系抽取领域,但是传统特征抽取器却忽略了广泛存在于实体之间的知识信息,这导致所捕获的特征有可能会误导选择有效句子的过程.例如,表 1 中句子"Steve Jobs was the co-founder and CEO of Apple."和"Steve Jobs argued with Wozniak, the co-founder of Apple."在句式结构上非常相似.因此先前的模型会从这 2 个句子当中捕获到相似的特征(即认为它们都表达了 Steve Jobs 和 Apple 是创始人的关系).在这样的情况下,如果缺乏实体知识信息,模型就无法很好地辨别出正确的信息来生成有效的包级别的特征表示.

为了解决上述提到的问题,本文通过探索额外的实体知识提出了一种实体知识增强的神经网络结构(entity knowledge enhanced neural network, EKNN).EKNN 模型通过动态地将实体知识与词嵌入融合在一起,从而能够使模型将更多的注意力集中在与句子中给定实体对有关的信息上,提高了模型在句子级别的表达能力.本文的主要贡献有 3 个方面:

1) 提出了一种知识感知的词嵌入方法,将实体中的 2 种知识,即来自语料库的语义知识和来自外部 KG 的结构知识动态地注入到词嵌入中.

2) 在广泛使用的"纽约时报"(New York Times, NYT)数据集[5]上评估了 EKNN 模型.实验结果表明,本文提出的模型在 2 个版本的 NYT 数据集上的表现明显优于其他最新模型.此外,通过额外的对比实验确认了 2 个版本的数据集之间存在的差异.

3) 通过进一步的消融实验,分别探究了 2 种不同的知识在关系抽取任务当中的有效性.

## 1 相关工作

大多数有监督的关系抽取模型[1-3]都会遇到标注数据不足的问题,而手动标记大规模的训练数据既费时又费力.因此,有研究者提出了使用远程监督

的方法来自动完成标记训练数据的工作[4].尽管远程监督在一定程度上缓解了人工标注数据的困难,但仍然会伴随着噪声数据的问题.Riedel 等人[5]和 Hoffmann 等人[6]都提出利用多示例学习的方法来缓解噪声数据的问题,该方法不再使用单个句子作为样本,而是将包含相同实体对的句子所组成的集合看作一个整体来作为样本.

传统的关系抽取方法主要是基于人工设计的特征进行的.近年来,随着深度学习的发展,神经网络模型已经被证明可以有效地捕获句子中的语义特征,并且还避免了由人工特征所引起的误差传递[14-16].Zeng 等人[8]提出了分段池化卷积神经网络(piecewise convolutional neural network, PCNN)从句子中更充分地提取实体之间的文本特征,并选择可能性最大的句子来作为包级别的表示.Lin 等人[9]提出了一个选择性注意力框架,该框架通过对集合内的所有句子进行加权求和来生成包级别的特征表示,这一框架也被之后的许多研究工作所广泛地采用[11-13].Shang 等人[17]则提出了一种基于深度聚类方法的关系抽取模型,通过无监督的深度聚类方法来为噪声句子重新生成可靠的标签,进而缓解噪声问题.此外,也有许多的工作尝试利用实体相关的外部信息来改善模型性能.Han 等人[18]提出了一种用于降噪的联合学习框架,该框架能够在知识图谱和文本之间的相互指导下进行学习.Hu 等人[19]利用知识图谱的结构信息和实体的描述文本来选择有效的句子进行关系提取.然而,这些方法大多数都仅仅考虑将知识信息用于降噪,并没有充分地利用实体知识中所蕴含的丰富的信息.

因此,在本文中同时引入了结构知识和语义知识来生成知识感知的词嵌入向量.通过这样的方法,知识信息可以更加深入地融合到模型中.

## 2 实体知识感知的神经网络模型

在本节中将介绍本文用于远程监督关系抽取的 EKNN 模型的整体框架和细节描述.

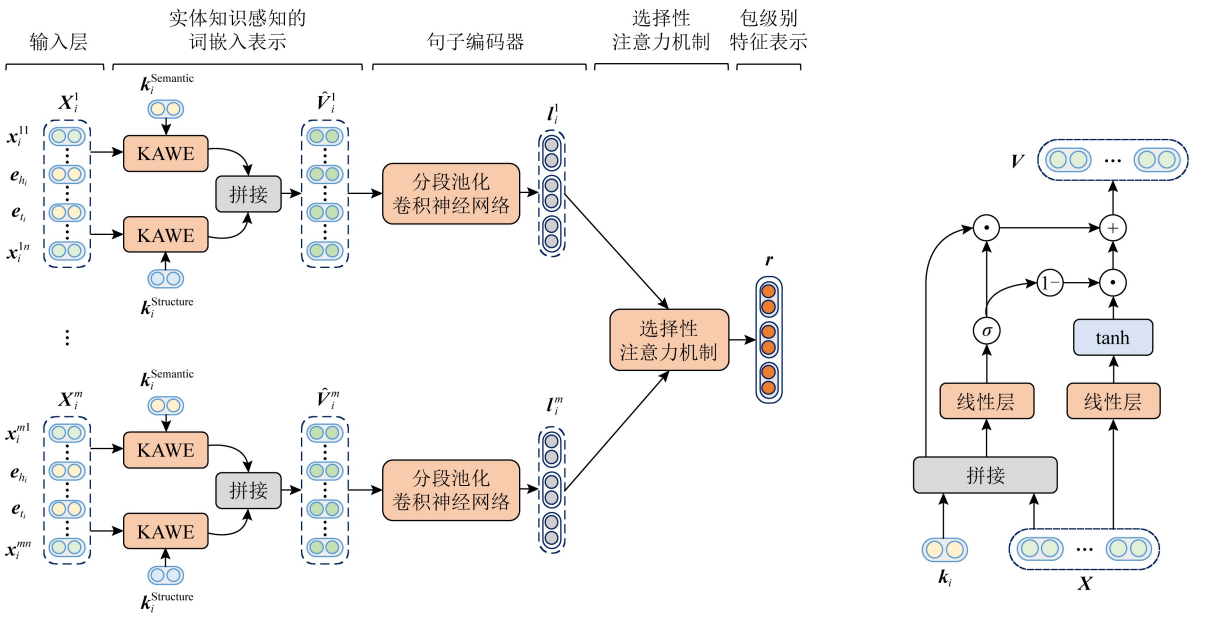
### 2.1 符号定义

定义知识图谱为 $\mathcal{G}=(\mathcal{E},\mathcal{R},\mathcal{F})$,其中 $\mathcal{E},\mathcal{R},\mathcal{F}$ 分别表示实体、关系和关系事实的集合.$(h,r,t)\in\mathcal{F}$ 表示 $h\in\mathcal{E}$ 和 $t\in\mathcal{E}$ 之间存在关系 $r\in\mathcal{R}$.使用 $k$ 来表示关于实体对$(h,t)$的知识.根据多示例学习的定义,令 $\mathcal{B}=\{\mathcal{S}_1,\mathcal{S}_2,\cdots,\mathcal{S}_{|\mathcal{B}|}\}$ 表示实体对包的集合.同时,令 $\mathcal{S}_i=\{s_i^1,s_i^2,\cdots,s_i^m\}$ 表示包 $\mathcal{S}_i$ 中的所有句子,而包 $\mathcal{S}_i$ 的标签是利用远程监督方法根据相应的实体对$(h_i,t_i)$给出的.将包中的每个句子都表示为一个单词序列 $s_i^j=[x_i^{j1},x_i^{j2},\cdots,x_i^{jn}]$.

### 2.2 模型框架

给定一个实体对$(h_i,t_i)$及其实体对包 $\mathcal{S}_i$,关系抽取的目的是预测实体对之间的关系 $r_i$.模型的总体框架如图 1(a)所示,主要有 3 个部分:

1) 知识感知的词嵌入模块.给定单词 $x_i^{jt}$ 和实




(a) EKNN模型架构 (b) 知识感知的词嵌入表示模块(KAWE)

Fig. 1 The framework of our proposed neural architecture EKNN

图 1 本文提出的 EKNN 模型架构图

体知识 $k_i$,我们使用门控机制来动态地生成知识感知的词嵌入 $v_i^{jt}$.

2) 句子语义特征编码.对于给定的单词序列 $s_i^j=[x_i^{j1},x_i^{j2},\cdots,x_i^{jn}]$,它的语义上下文嵌入 $\boldsymbol{l}_i^j$ 由 PCNN[8] 网络编码得到.

3) 选择性注意力机制.给定实体对包 $\mathcal{S}_i$ 中所有句子的语义上下文嵌入表示,包级别的特征表示通过注意力机制计算得到,最终用于预测关系类型.

### 2.3 知识感知的词嵌入模块

关系抽取的目标是预测 2 个实体之间的关系.因此,实体对中所包含的信息是非常重要的.在当前的研究当中,实体对中仍然还有很多隐含的信息尚未得到充分的利用.受这一想法的启发,本文引入了语义信息和结构信息作为实体对的外部知识,以此来丰富传统的词嵌入表示.

#### 2.3.1 语义知识嵌入

词嵌入技术由 Hinton 等人[20] 首次提出,其目的是为了将词语转换为向量空间当中的分布式向量表示,以捕获词语间的句法和语义特征.因此,本文采用词嵌入作为语义信息的来源.给定一个实体对 $(h_i,t_i)$ 及其词嵌入 $(\boldsymbol{e}_{h_i},\boldsymbol{e}_{t_i})$,将实体对的语义知识嵌入定义为

$$\boldsymbol{k}_i^{\text{Semantic}}=(\boldsymbol{e}_{h_i};\boldsymbol{e}_{t_i}),\tag{1}$$

其中,$\boldsymbol{e}_{h_i},\boldsymbol{e}_{t_i}\in\mathbb{R}^{d_\text{w}}$,$\boldsymbol{k}_i^{\text{Semantic}}\in\mathbb{R}^{2d_\text{w}}$.

#### 2.3.2 结构知识嵌入

典型的知识图谱通常是一个具有多种关系类型的有向图,可以将其表示为一系列关系三元组 $(h,r,t)$ 的集合[12].因此,知识图谱通常会包含有丰富的结构信息,可以将其看作本文结构知识的来源.本文使用 TransE[21] 作为知识图谱嵌入模型,以此来得到实体和关系的预训练嵌入向量.给定一个三元组 $(h,r,t)$ 及其嵌入表示 $(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$,TransE 将关系 $r$ 看作是从头实体 $h$ 到尾实体 $t$ 的一种翻译操作,如果 $(h,r,t)$ 存在,则可以假设嵌入向量 $\boldsymbol{t}$ 应该接近于 $\boldsymbol{h}+\boldsymbol{r}$.因此,将实体对的结构知识嵌入定义为

$$\boldsymbol{k}_i^{\text{Structure}}=\boldsymbol{t}_i-\boldsymbol{h}_i,\tag{2}$$

其中,$\boldsymbol{h}_i,\boldsymbol{t}_i\in\mathbb{R}^{d_\text{s}}$,$\boldsymbol{k}_i^{\text{Structure}}\in\mathbb{R}^{d_\text{s}}$.

#### 2.3.3 门控融合

为了能够动态地将实体对的知识与原始的词嵌入融合在一起,本文使用门控机制来生成知识感知的词嵌入表示.

给定一个实体对 $(h_i,t_i)$ 和一个由单词序列所组成的句子 $s_i^j=(x_i^{j1},x_i^{j2},\cdots,x_i^{jn})$,其中每个词 $x_i^{jt}$ 的嵌入表示是通过预训练的词向量 $\boldsymbol{x}_i^{jt}\in\mathbb{R}^{d_\text{w}}$ 初始化得到的.因此单词序列可以定义为 $\boldsymbol{X}=(\boldsymbol{x}_i^{j1},\boldsymbol{x}_i^{j2},\cdots,\boldsymbol{x}_i^{jn})\in\mathbb{R}^{n\times d_\text{w}}$,其中 $n$ 表示单词序列的长度.如图 1(b)所示,将门控机制定义为

$$\boldsymbol{V}=\alpha\boldsymbol{X}^{(\text{e})}+(1-\alpha)\hat{\boldsymbol{X}},\tag{3}$$

$$\hat{\boldsymbol{X}}=\tanh(\boldsymbol{X}\boldsymbol{W}_{g1}^\text{T}+\boldsymbol{b}_{g1}),\tag{4}$$

$$\alpha=sigmoid(\lambda(\boldsymbol{X}^{(\text{e})}\boldsymbol{W}_{g2}^\text{T}+\boldsymbol{b}_{g2})),\tag{5}$$

$$\boldsymbol{X}^{(\text{e})}=(\boldsymbol{x}_i^{(\text{e})j1},\boldsymbol{x}_i^{(\text{e})j2},\cdots,\boldsymbol{x}_i^{(\text{e})jn}),\tag{6}$$

其中,$\forall\,\boldsymbol{x}_i^{(\text{e})jt}=(\boldsymbol{x}_i^{jt};\boldsymbol{k}_i)$,$\boldsymbol{k}_i\in\mathbb{R}^{d_\text{k}}$ 表示知识嵌入,即语义知识或结构知识,$\boldsymbol{W}_{g1}\in\mathbb{R}^{d_\text{h}\times d_\text{w}}$ 和 $\boldsymbol{W}_{g2}\in\mathbb{R}^{d_\text{h}\times(d_\text{w}+d_\text{k})}$ 是可学习的参数,$\lambda$ 是用于控制平滑的超参数.而 $\boldsymbol{V}=(\boldsymbol{v}_i^{j1},\boldsymbol{v}_i^{j2},\cdots,\boldsymbol{v}_i^{jn})\in\mathbb{R}^{n\times d_\text{h}}$ 则表示一个句子中所有单词的知识感知词嵌入表示.

对于每个单词 $x_i^{jt}$,利用结构知识嵌入和语义知识嵌入分别构造了 2 种知识感知的词嵌入表示 $\boldsymbol{v}_i'^{jt}\in\mathbb{R}^{d_\text{h}}$ 和 $\boldsymbol{v}_i''^{jt}\in\mathbb{R}^{d_\text{h}}$.同时,还使用了位置嵌入 $\boldsymbol{p}_i^{jt}$,$\boldsymbol{q}_i^{jt}\in\mathbb{R}^{d_\text{p}}$ 来描述当前词与 2 个实体之间的相对位置信息[14].

最终,句子 $s_i^j$ 中每个词语完整的知识感知词嵌入表示是由上述几种嵌入向量拼接得到的,其定义为

$$\hat{\boldsymbol{V}}=(\hat{\boldsymbol{v}}_i^{j1},\hat{\boldsymbol{v}}_i^{j2},\cdots,\hat{\boldsymbol{v}}_i^{jn})\in\mathbb{R}^{n\times d_\text{v}},\tag{7}$$

其中,$\forall\,\hat{\boldsymbol{v}}_i^{jt}=(\boldsymbol{v}_i'^{jt};\boldsymbol{v}_i''^{jt};\boldsymbol{p}_i^{jt};\boldsymbol{q}_i^{jt})$,$d_\text{v}=d_\text{h}\times2+d_\text{p}\times2$.

### 2.4 句子语义特征编码

本节将详细介绍句子级别的语义上下文嵌入表示模块.给定一个句子 $s_i^j$,对于句子的语义上下文嵌入,采用在上一节中所描述的知识感知词嵌入 $\hat{\boldsymbol{V}}$ 来作为输入,而不是原始词嵌入 $\boldsymbol{X}$.

在许多的关系抽取研究当中,卷积神经网络已经被证明是一个稳定且有效的网络结构[9-13],它在编码句子局部上下文特征的同时,可以捕获细粒度的信息.因此,在本文中使用 Zeng 等人[8] 所提出的分段池化卷积神经网络(PCNN)作为编码器.PCNN 是一个基于传统卷积神经网络所设计的神经网络结构,它设计了一个全新的分段最大池化层来捕获实体之间的结构化信息,以此来适应关系抽取任务.首先,PCNN 在输入序列 $[\hat{\boldsymbol{v}}_i^{j1},\hat{\boldsymbol{v}}_i^{j2},\cdots,\hat{\boldsymbol{v}}_i^{jn}]$ 上滑动窗口大小为 $z$ 的卷积核来得到隐藏层的嵌入表示:

$$\boldsymbol{h}_i^{jt}=f_{\text{CNN}}(\hat{\boldsymbol{v}}_i^{j(t-\frac{z-1}{2})},\hat{\boldsymbol{v}}_i^{j(t-\frac{z+1}{2})},\cdots,\hat{\boldsymbol{v}}_i^{j(t+\frac{z-1}{2})}),\tag{8}$$

其中,$\boldsymbol{h}_i^{jt}\in\mathbb{R}^{d_\text{c}}$.

之后,将分段池化层作用于隐藏层序列 $\boldsymbol{H}_i^j=(\boldsymbol{h}_i^{j1},\boldsymbol{h}_i^{j2},\cdots,\boldsymbol{h}_i^{jn})\in\mathbb{R}^{n\times d_\text{c}}$,最终得到句子级别的语义上下文嵌入表示:

$$\boldsymbol{l}_i^j=\tanh((Pool(\boldsymbol{H}_i^{(1)j});Pool(\boldsymbol{H}_i^{(2)j});Pool(\boldsymbol{H}_i^{(3)j}))), \tag{9}$$

其中,$\boldsymbol{l}_i^j\in\mathbb{R}^{3d_c}$是由 PCNN 网络编码得到的句子语义特征,而 $\boldsymbol{H}_i^{(1)j}$,$\boldsymbol{H}_i^{(2)j}$,$\boldsymbol{H}_i^{(3)j}$ 分别是 $\boldsymbol{H}_i^j$ 的 3 个分段,其中分段的边界是由 2 个实体的位置所决定的.

### 2.5 选择性注意力机制

为了从带噪声的实体对包当中抽取出有效的信息,我们使用选择性注意力机制[9]来为句子计算权重并得到包级别的特征表示.给定一个实体对包 $\mathcal{S}_i=[s_i^1,s_i^2,\cdots,s_i^m]$,使用 PCNN 来为所有的句子计算语义嵌入表示.$\mathcal{S}_i=[s_i^1,s_i^2,\cdots,s_i^m]$中每个句子注意力分数的计算公式定义为:

$$\beta_i^j=\frac{\exp(\gamma_i^j)}{\sum_{j'=1}^{m}\exp(\gamma_i^{j'})}, \tag{10}$$

$$\gamma_i^j=\tanh(\boldsymbol{l}_i^j\boldsymbol{W}_{b1}^{\mathrm{T}})\boldsymbol{W}_{b2}^{\mathrm{T}}, \tag{11}$$

其中,$\boldsymbol{W}_{b1}\in\mathbb{R}_{d_b\times 3d_c}$,$\boldsymbol{W}_{b2}\in\mathbb{R}^{1\times d_b}$是可学习参数,$d_b$是超参数.之后,可以通过上述注意力分数来得到包级别的特征表示用于关系分类,其定义为

$$\boldsymbol{r}=\sum_{j=1}^{m}\beta_i^j\boldsymbol{l}_i^j. \tag{12}$$

最终,特征 $\boldsymbol{r}$ 在经过线性变换后被送入到 Softmax 分类器当中.其计算公式定义为

$$P(r|\mathcal{S};\theta)=\mathrm{Softmax}(\boldsymbol{r}\boldsymbol{M}_r^{\mathrm{T}}+\boldsymbol{b}_r), \tag{13}$$

其中,$\boldsymbol{M}_r$ 是变换矩阵,$\boldsymbol{b}_r$ 是偏置项.同时,与 Lin 等人[9]相同,本文在包级别的特征表示 $\boldsymbol{r}$ 上使用了 dropout[22]来防止过拟合.

### 2.6 模型学习

在训练阶段,本文尝试最小化交叉熵损失函数:

$$J(\theta)=-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\ln P(r_i|\mathcal{S}_i;\theta), \tag{14}$$

其中,$\theta$ 表示模型中的所有参数,$\mathcal{B}=[\mathcal{S}_1,\mathcal{S}_2,\cdots,\mathcal{S}_{|\mathcal{B}|}]$表示实体对包的集合,而$[r_1,r_2,\cdots,r_{|\mathcal{B}|}]$则表示对应的关系标签.本文中所有模型均使用随机梯度下降(stochastic gradient descent, SGD)作为优化算法.

## 3 实验及结果分析

### 3.1 数据集

在 Riedel 等人[5]公开的"纽约时报"NYT 数据集上对本文的模型进行了评估,该数据集是通过将 Freebase 中的关系与 NYT 语料库进行自动地对齐而生成的,它在目前许多最新的远程监督关系抽取研究工作当中被广泛使用.该数据集共包含 53 类关系和一个无关系 NA 标签,无关系标签表示 2 个实体之间没有任何关系.值得一提的是,在当前的许多研究工作当中存在 2 个不同版本的 NYT 数据集,这是由于一次错误的数据集发布所造成的①.这 2 个版本数据集之间的主要区别在于训练集部分不同,它们的测试集是相同的.具体而言,表 2 中列出了这 2 个数据集的一些数据统计情况.本文将这 2 个不同的训练集分别称为 train-570K 和 train-520K.从表 2 中可以看出,train-570K 和测试集在实体对而非句子上存在交集,而 train-520K 是比较干净的训练集,不存在交集.在交集部分,大多数实体对的标签为 NA,这使得模型更容易区分标签为 NA 的样本.因此可以推断出,对于同一模型,使用 train-570K 数据集进行训练的效果会高于 train-520K 数据集.

**Table 2 Statistics of the NYT Dataset**

**表 2 NYT 数据集统计情况**

| 数据集 | 句子数 | 实体对 | 关系事实数量 | 交集占比/% |
|---|---|---|---|---|
| train-570K | 570 088 | 293 003 | 19 429 | 11.8 |
| train-520K | 522 611 | 281 270 | 18 252 | 0 |
| 测试集 | 172 448 | 96 678 | 1 950 | |

注:"交集占比"表示训练集和测试集中实体对的交集情况.

在评价指标方面,本文遵循了目前主流的研究工作[4,9,13],在 NYT 测试集上使用精确率-召回率(P-R)曲线以及该曲线下面积(AUC)和 Top-$N$ 精确度(P@$N$)来作为评估指标.

### 3.2 实验设置

#### 3.2.1 知识图谱嵌入

本文使用 FB40K[23]来作为我们的外部知识图谱,它包含大约 40 000 个实体和 1 318 种关系类型.为了生成预训练的知识图谱嵌入,使用 OpenKE Toolkit[24]进行训练,其中嵌入向量的大小 $d_s=100$,超参数 $margin=5$,学习率设置为 1,迭代轮数为 500 轮.值得注意的是,FB40K 和测试集在实体对上没有的任何交集.因此,在外部知识图谱 FB40K 中不会包含任何出现在测试集中的实体对.

#### 3.2.2 参数设置

在实验中,通过网格法对超参数进行了选择,其中批量大小 $B\in\{50,120,160\}$,CNN 卷积核个数 $d_c\in\{64,128,230,256\}$,隐藏层 $d_b\in\{256,512,$

① https://github.com/thunlp/NRE/commit/77025e5cc6b42bc1adf3ec46835101d162013659

690},平滑参数 $\lambda \in \{30,35,40,45,50\}$,其余参数则与前人工作[9,12-13]保持一致.同时采用了 Lin 等人[9]发布的 50 维的词向量用于初始化设置.表 3 列出了在 2 个版本的训练集上进行实验使用的所有超参数情况.此外,对于优化算法,在 train-570K 和 train-520K 上分别使用学习率为 0.1 和 0.2 的 mini-batch SGD 算法进行训练.

**Table 3 Hyper-parameter Settings in Our Experiments**

**表 3 本实验中的超参数设置**

| 超参数 | 取值 |
|---|---|
| 批量大小 $B$ | 160 |
| 隐藏层维度 $d_h$ | 150 |
| 隐藏层维度 $d_b$ | 512 |
| CNN 卷积核数量 $d_c$ | 230 |
| 词向量维度 $d_w$ | 50 |
| 位置向量维度 $d_p$ | 5 |
| 卷积核窗口大小 $z$ | 3 |
| dropout 率 | 0.5 |
| 平滑参数 $\lambda$ | 45 |

#### 3.2.3 基准模型

为了对本文所提出的 EKNN 模型进行评估,将其与当前最新的基准模型进行了比较:PCNN+ATT[9]是最基础的选择性注意力模型;+HATT[12]采用层次注意力机制,在长尾关系抽取上的效果有很大的提升;+BAG-ATT[13]分别使用包内和包之间的注意力机制来缓解句子级别和包级别的噪声;JointD+KATT[18]设计了一个联合学习框架,通过知识图谱和文本之间的相互指导学习来进行降噪;RELE[19]通过知识图谱的结构化信息来指导标签嵌入(label embedding)的学习从而进行降噪,提高了关系抽取的性能.此外,本文还与传统的基于特征的模型进行了对比,包括 Mintz[4],MultiR[6],MIML[7]等.

### 3.3 对比实验结果与分析

从表 4 中所列出的 P@$N$ 值结果中可以看出,对于 CNN+ATT 和 PCNN+ATT,使用 train-520K 进行训练的 P@$N$ 要明显低于使用 train-570K 进行训练的 P@$N$.这一实验结果与数据集上(表 2)所观察到的现象是一致的,即训练集和测试集之间的实体对存在交集可以在一定程度上提高模型在关系抽取任务上的性能.而对于当前最先进的研究方法,也使用 train-520K 重新进行了训练.与上述实验结果相似,这些模型的结果在 train-520K 也出现了显著的性能下降.而所提出的 EKNN 模型在 2 个训练集上进行训练的结果也有所不同.但是与其他基准方法相比,在 train-570K 和 train-520K 上分别进行训练时,本文的方法在 P@$N$ 指标上仍然要明显地优于其他方法.具体而言,在 P@$N$ 均值这一指标上,相比 PCNN+ATT 模型在 2 个训练集上分别提升了 11.6% 和 5.0%.此外,与最优的基准模型+BAG-ATT 相比,本文所提出的模型在 2 个训练集上也有着显著的性能提升.上述结果证明了本文所提出的远程监督关系抽取方法的有效性.

**Table 4 P@$N$ Values of Different Models on the Two Training Sets**

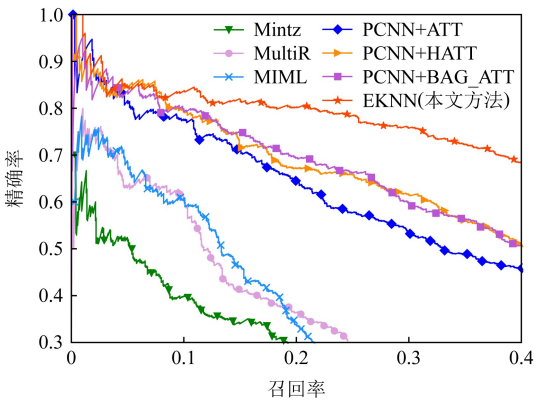
**表 4 各模型在 2 个训练集上的 P@$N$ 值** %

| 对比模型 | train-570K | | | | train-520K | | | |
|---|---|---|---|---|---|---|---|---|
| | $N=100$ | $N=200$ | $N=300$ | 均值 | $N=100$ | $N=200$ | $N=300$ | 均值 |
| CNN+ATT | 79.20 | 74.90 | 70.30 | 74.80 | 76.20 | 68.60 | 59.80 | 68.20 |
| PCNN+ATT | 80.80 | 77.50 | 72.30 | 76.90 | 76.20 | 73.10 | 67.40 | 72.20 |
| PCNN+HATT | 88.00 | 79.50 | 75.30 | 80.90 | | | | |
| PCNN+BAG-ATT | 91.80 | 84.00 | 78.70 | 84.80 | 84.00 | 75.50 | 70.00 | 76.50 |
| EKNN(本文方法) | **90.00** | **89.00** | **86.67** | **88.55** | **85.00** | **76.00** | **70.79** | **77.26** |
| EKNN w/o Structure | 95.00 | 86.00 | 82.00 | 87.67 | 81.00 | 72.50 | 67.33 | 73.61 |
| EKNN w/o Semantic | 85.00 | 89.00 | 87.67 | 87.22 | 80.00 | 76.00 | 72.67 | 76.22 |
| EKNN w/o All | 84.60 | 76.20 | 68.53 | 76.44 | 80.00 | 72.00 | 65.33 | 72.44 |

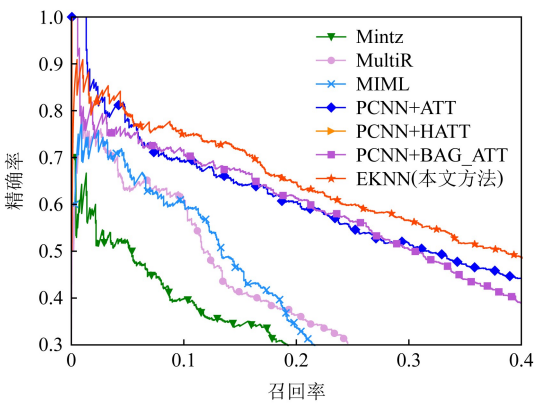
注:加粗数字表示本文模型的结果.

此外,图 2 和表 5 也分别展示了精确率-召回率(P-R)曲线和 AUC 的结果.从图 2 中的 P-R 曲线可以看出,随着召回率的提升,各模型的精确率出现了急剧的下降,这是由于远程监督数据集中的噪声问题所导致的.而从表 5 中的 AUC 值还可以看出,对比 train-570K 上的结果,包括本文模型在内的所有

模型在 train-520K 下进行训练都有不同程度的性能下降,这与表 4 中的 P@$N$ 指标的结果是一致的,也同样验证了在 3.1 节中所作的分析.但是,对比最优的基准模型+BAG-ATT,本文的方法在 train-520K 上仍然有显著提升,这进一步证明了本文所提出模型的性能提升是稳定且有效的.具体而言,本文在 2 个数据集上,+BAG-ATT 的 AUC 指标分别提高了 0.12 和 0.05.




(a) train-570K



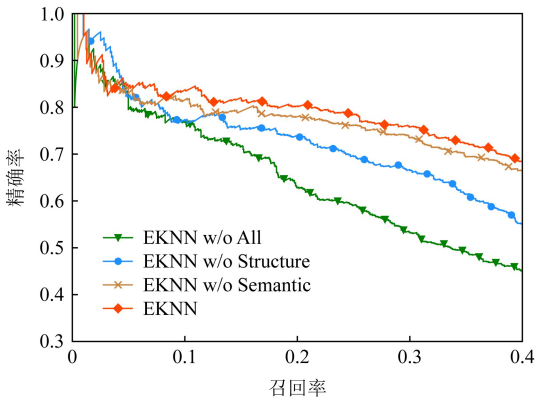

(b) train-520K

Fig. 2 Precision-recall curves of the proposed model and other baseline models

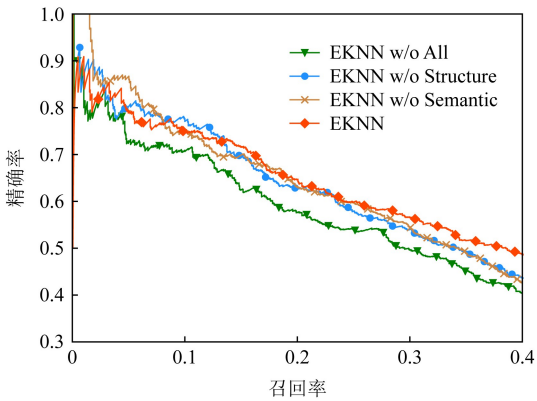
图 2 本文模型与其他模型的精确率-召回率曲线

### 3.4 消融实验结果与分析

为了进一步验证本文所提出的方法中不同模块的有效性,本文进行了充分的消融实验,旨在探索什么样的实体知识对于关系抽取任务更有价值.消融实验的 P@$N$ 指标结果在表 4 的 6~8 行列出,而精确率-召回率曲线和及其相应的 AUC 值在图 3 和表 5 的部分区域进行展示.其中,w/o All 表示去掉本文中设计的所有新模块,相当于最基础的选择性注意力模型[9].在接下来的分析当中,可以将其作为基准和其它模型进行对比.




(a) train-570K




(b) train-520K

Fig. 3 Precision-recall curves for the ablation study

图 3 消融实验的精确率-召回率曲线

**Table 5 AUC Values of Different Models on the Two Training Sets**

**表 5 各模型在 2 个训练集上的 AUC 值**

| 对比模型 | train-570K | train-520K |
|---|---|---|
| PCNN+ATT | 0.380 | 0.341 |
| PCNN+BAG-ATT | 0.422 | 0.345 |
| EKNN(本文方法) | **0.542** | **0.397** |
| EKNN w/o Structure | 0.461 | 0.374 |
| EKNN w/o Semantic | 0.523 | 0.388 |
| EKNN w/o All | 0.381 | 0.342 |

注:加粗数字表示本文模型的结果.

在本文提出的 EKNN 模型中,引入了 2 类实体知识,分别为模型提供语义信息和结构化信息.为了验证它们的有效性,设计了 2 个变种模型.具体而言,w/o Semantic 和 w/o Structure 都表示丢弃其中一类知识而保留另一类.从结果当中可以发现,2 类实体知识都可以丰富模型的表达能力,并显著提高模型性能.以 train-520K 为例,假如去掉整个知识感知词嵌入模块(w/o All),P@$N$ 值和 AUC 指标分

别下降了4.8%和0.055.此外,通过对比w/o Semantic和w/o Structure这2个变种模型,可以了解到在关系抽取任务当中结构化信息比语义信息具有更大的价值,这是由于模型从结构化数据中所学到的隐式嵌入具有更强的推理能力.

### 3.5 不同知识融合方式的对比实验分析

为了验证本文知识融合方法的有效性,本文与当前主流的融合知识的远程监督关系抽取方法进行了对比,实验结果如表6所示:

**Table 6 P@N Values of Different Knowledge Integration Methods**

**表6 不同知识融合方式的P@N值** %

| 对比模型 | $N=100$ | $N=300$ | $N=500$ | 均值 |
|---|---|---|---|---|
| JointD+KATT | 80.60 | 68.70 | 63.70 | 71.00 |
| RELE | 88.00 | 78.60 | 69.80 | 78.80 |
| EKNN(本文方法) | **90.00** | **86.60** | **81.00** | **85.80** |

注:加粗数字表示本文模型的结果.

从实验结果中可以看出,所提出的实体知识感知的词嵌入模块拥有更加优越的性能提升.这是由于JointD+KATT和RELE仅仅考虑了将知识信息用于模型训练和指导降噪的过程,而忽略了实体知识中所蕴含的丰富表示.EKNN模型通过知识信息和词嵌入表示融合的方式,更加深层次地将知识整合进了模型,对实体知识进行了更充分地利用,因而获得了更好的性能表现.

### 3.6 平滑参数λ对模型性能的影响

在实体知识感知的词嵌入表示模块当中,超参数λ用于对知识融合的过程进行平滑控制,图4给出了不同的λ值对于模型性能的影响.从图4中可以看出,当λ值在40~45之间时,模型中的实体知识和词嵌入可以实现相对较好的融合效果,从而提升模型性能.

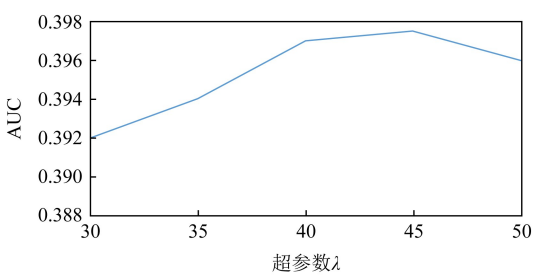



Fig. 4 The effect of hyperparameter λ on model performance

图4 超参数λ对模型性能的影响

## 4 总 结

本文提出了一种用于远程监督关系抽取的神经网络模型EKNN.为了提高模型的表达能力,引入了2类实体知识(即语义知识和结构知识)来动态地生成知识感知的词嵌入.通过丰富的对比实验,证明了本文的模型性能显著优于当前最优的方法.此外,本文还通过对比实验探究了"纽约时报"数据集上2个版本的训练数据之间的差异,结果表明,由于排除了数据集间的实体对交集,train-520K数据集比train-570K数据能够更有效的反映模型性能.

**作者贡献声明**:高建伟负责模型设计以及文章的撰写;万怀宇负责方法概念的提出文章的润色和审阅校对;林友芳负责实验数据的管理、文章的润色和审阅校对.

## 参 考 文 献

[1] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction [J]. Journal of Machine Learning Research, 2003, 3(2): 1083-1106

[2] Culotta A, Sorensen J. Dependency tree kernels for relation extraction [C] //Proc of the 42nd Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2004: 423-429

[3] Mooney R J, Bunescu R C. Subsequence kernels for relation extraction [C] //Proc of the 18th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2006: 171-178

[4] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data [C] //Proc of the Joint Conf of the 47th Annual Meeting of the ACL and the 4th Int Joint Conf on Natural Language Processing of the AFNLP. Stroudsburg, PA: ACL, 2009: 1003-1011

[5] Riedel S, Yao Limin, McCallum A. Modeling relations and their mentions without labeled text [C] //Proc of the 21st European Conf on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2010: 148-163

[6] Hoffmann R, Zhang Congle, Ling Xiao, et al. Knowledge-based weak supervision for information extraction of overlapping relations [C] //Proc of the 49th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2011: 541-550

[7] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction [C] //Proc of the 2012 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA: ACL, 2012: 455-465

[8] Zeng Daojian, Liu Kang, Chen Yubo, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 1753-1762

[9] Lin Yankai, Shen Shiqi, Liu Zhiyuan, et al. Neural relation extraction with selective attention over instances [C] //Proc of the 54th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2016: 2124-2133

[10] Ji Guoliang, Liu Kang, He Shizhu, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 3060-3066

[11] Liu Tianyu, Wang Kexiang, Chang Baobao, et al. A soft-label method for noise-tolerant distantly supervised relation extraction [C] //Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 1790-1795

[12] Han Xu, Yu Pengfei, Liu Zhiyuan, et al. Hierarchical relation extraction with coarse-to-fine grained attention [C] // Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 2236-2245

[13] Ye Zhixiu, Ling Zhenhua. Distant supervision relation extraction with intra-bag and inter-bag attentions [C] //Proc of the 2019 Conf of the North American Chapter of the ACL. Stroudsburg, PA: ACL, 2019: 2810-2819

[14] Zeng Daojian, Liu Kang, Lai Siwei, et al. Relation classification via convolutional deep neural network [C] // Proc of the 25th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2014: 2335-2344

[15] Xu Yan, Mou Lili, Li Ge, et al. Classifying relations via long short term memory networks along shortest dependency paths [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 1785-1794

[16] dos Santos C, XiangBing, Zhou Bowen. Classifying relations by ranking with convolutional neural networks [C] //Proc of the 53rd Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2015: 626-634

[17] Shang Yuming, Huang Heyan, Mao Xianling, et al. Are noisy sentences useless for distant supervised relation extraction?[C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 8799-8806

[18] Han Xu, Liu Zhiyuan, Sun Maosong. Neural knowledge acquisition via mutual attention between knowledge graph and text [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 4832-4839

[19] Hu Linmei, Zhang Luhao, Shi Chuan, et al. Improving distantly-supervised relation extraction with joint label embedding [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2019: 3812-3820

[20] Hinton G E. Learning distributed representations of concepts [C/OL] //Proc of the 8th Annual Conf of the Cognitive Science Society. Amherst, MA: Psychology Press, 1986 [2021-12-19]. http://www.cs.toronto.edu/~hinton/absps/families.pdf

[21] Bordes A, Usunier N, Garcia-Durán A, et al. Translating embeddings for modeling multi-relational data [C] //Proc of the 26th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 2787-2795

[22] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958

[23] Lin Yankai, Liu Zhiyuan, Sun Maosong, et al. Learning entity and relation embeddings for knowledge graph completion [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 2181-2187

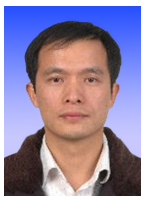
[24] Han Xu, Cao Shulin, Lü Xin, et al. Openke: An open toolkit for knowledge embedding [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 139-144

**Gao Jianwei**, born in 1995. Master. His main research interests include knowledge graph and information extraction.

**高建伟**,1995 年生.硕士.主要研究方向为知识图谱和信息抽取.

**Wan Huaiyu**, born in 1981. PhD, associate professor, PhD supervisor. Member of CCF. His main research interests include social network mining, text information extraction, user behavior analysis, and spatial-temporal data mining.

**万怀宇**,1981 年生.博士,副教授,博士生导师.CCF 会员.主要研究方向为社交网络挖掘、文本信息抽取、用户行为分析和时空数据挖掘.



**Lin Youfang**, born in 1971. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include data mining, machine learning, reinforcement learning, complex networks, intelligent technologies and systems.

**林友芳**,1971 年生.博士,教授,博士生导师.CCF 高级会员.主要研究方向为数据挖掘、机器学习、强化学习、复杂网络、智能技术与系统.