

基于适应性自训练的少样本关系抽取建模

陈洪辉¹ 郑建明¹ 蔡 飞¹ 韩 穗²

¹(国防科技大学系统工程学院 长沙 410073)

²(国防科技大学气象海洋学院 长沙 410073)

(chenhonghui@nudt.edu.cn)

Modeling of Few-Shot Relation Extraction Based on Adaptive Self-Training

Chen Honghui¹, Zheng Jianming¹, Cai Fei¹, and Han Yi²

¹(College of System Engineering, National University of Defense Technology, Changsha 410073)

²(College of Meteorology and Oceanography, National University of Defense Technology, Changsha 410073)

Abstract Relation extraction (RE) is a basic task in natural language processing, which supports plenty of downstream tasks, e.g., dialogue generation, machine reading comprehension, etc. In real life, due to the continuously emerging new relation labels, the speed and cost of human annotation cannot catch up with the data quantity that the training of the traditional supervised RE models demands. Facing this practical challenge, the neural snowball model proposes a bootstrapping method that transfers the RE knowledge from limited labeled instances to iteratively annotate unlabeled data as to increase the amount of labeled data, thereby improving the classification performance of the model. However, the fixed threshold selection and the equally treated unlabeled data make the neural snowball model vulnerable to noise data. To solve these two defects, an adaptive self-training relation extraction (Ada-SRE) model is proposed. In specific, for the fixed-threshold issue, Ada-SRE proposes an adaptive threshold module by the meta learning of threshold, which can provide an appropriate threshold for each relation category. For the equally-treated issue, Ada-SRE designs a gradient-feedback strategy to weight each selected example, avoiding the interference of noise data. The experimental results show that compared with the neural snowball model, Ada-SRE has a better relation extraction ability.

Key words self-training; relation extraction; gradient feedback; few-shot learning; meta learning

摘要 关系抽取 (relation extraction, RE) 是自然语言处理中的一项基础任务, 可以支撑许多下游任务, 例如对话生成和机器阅读理解等。在现实生活中, 由于新关系类别不断涌现, 人工标注的成本和速度无法满足传统基于有监督学习的关系抽取模型的训练要求。面对这种现实挑战, 神经雪球提出一种自助采样的方法, 通过对有限标注数据的信息迁移, 不断为无标注数据打上标签, 增加标注数据量, 从而提升模型分类性能。然而, 固定的阈值选择以及同等对待入选的无标注数据使得神经雪球模型容易受到噪声数据的影响。为了解决这 2 个缺陷, 基于适应性自训练的关系抽取 (adaptive self-training relation extraction, Ada-SRE) 模型由此提出。具体地, Ada-SRE 基于元学习的思想提出自适应阈值模块, 能够为每个关系类别提供合适的阈值选择。另外, Ada-SRE 还提出基于梯度反馈的赋权策略, 为每个入选的示例提供相应的权重, 避免噪声数据的干扰。实验结果表明, 相比于神经雪球模型, Ada-SRE 有更好的关系抽取能力。

关键词 自训练; 关系抽取; 梯度反馈; 少样本学习; 元学习

收稿日期: 2022-03-14; 修回日期: 2022-09-16

基金项目: 湖南省研究生科研创新项目 (CX20190034, CX20210068)

This work was supported by the Postgraduate Scientific Research Innovation Project of Hunan Province (CX20190034, CX20210068).

通信作者: 郑建明 (zhengjianming12@nudt.edu.cn)

中图法分类号 TP391

关系抽取^[1](relation extraction, RE)通过识别非结构化文本中实体之间的关系类别, 获取信息三元组(头实体, 关系, 尾实体). 例如, 给定上下文句子“Newton was served as the president of the Royal Society”, 对于头实体 Newton 和尾实体 Royal Society, 关系抽取模型可以识别出这 2 个实体之间包含“member of”关系. 关系抽取任务是自然语言处理中的一项基础任务, 可以支撑大量下游任务, 例如: 知识图谱补全^[2]、问答系统^[3]、对话生成^[4], 等. 传统的关系抽取方法(基于高斯核^[5-6]、基于特征^[7]以及基于神经网络^[8]等)依赖大量标注数据实现模型的收敛. 然而, 在现实应用中, 新的实体和关系类别不断涌现, 标注足够训练数据的速率远远赶不上新类别的涌现速率^[9-10]^[1], 使得关系抽取任务陷入少样本困境.

针对该挑战, 现有模型普遍采用元学习(meta learning)的训练范式^[11-14]. 其核心在于将数据丰富的关

系类别分解成一系列 N -类 K -样本(N -way K -shot)的元任务, 用以模拟数据匮乏的应用场景, 如图 1 所示. 通过元任务训练, 关系类别间的元知识得以获取, 进而快速泛化到数据匮乏的关系类别中. 尽管元学习的范式取得了明显的进展, 但它们仅仅侧重于标注数据, 而少样本任务中由于标注数据量稀少导致的过拟合风险仍然存在. 为了应对该风险, 神经雪球(neural snowball, NS)模型^[15]着眼于现实应用中广泛存在且容易获得的无标注数据. 该模型将元学习和自助采样相结合, 通过将现有标注数据中的知识迁移到无标注数据中, 为无标注数据打上伪标签, 从而实现标注数据量的持续增加. 由此, 神经雪球模型不仅可通过元学习破除传统关系抽取模型难以解决少样本的局限性, 还可通过无标注数据的知识迁移极大提升标注数据量, 显著降低模型过拟合的风险.

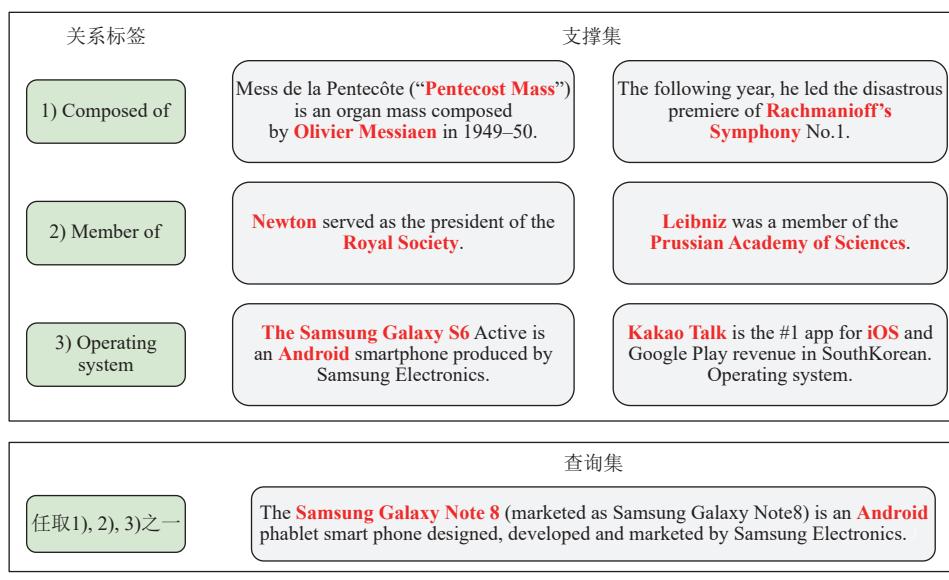


Fig. 1 Example of 3-way 2-shot relation extraction meta task

图 1 3-类别 2-样本关系抽取元任务的示例

然而, 回溯神经雪球模型, 我们发现其存在 2 种缺陷:

1) 当无标注数据打伪标签时, 神经雪球模型是通过一个固定相似度阈值来判定无标注数据是否属于特定关系类别. 可是不同关系类别在相似度阈值

的选择上存在差异. 如图 2 所示, 当相似度阈值固定在 0.60 时, 虽然可以很好地区分“board member”和“owned by”这 2 种关系, 但却无法辨别“publisher”和“item operated”的关系. 因此, 设定固定阈值会不可避免地打错标签, 使得新入选的无标注数据包含噪声.

^① 例如, FewRel^[9]是 2018 年标注的大型关系抽取数据集, 它从 Wikidata^[10]中抽取 100 种关系类别. 然而, 这个数量远少于当时 920 种类别的数据量. 更为严峻地, Wikidata 中无标注关系类别数量到现在为止已经增长到 6 000 余种.

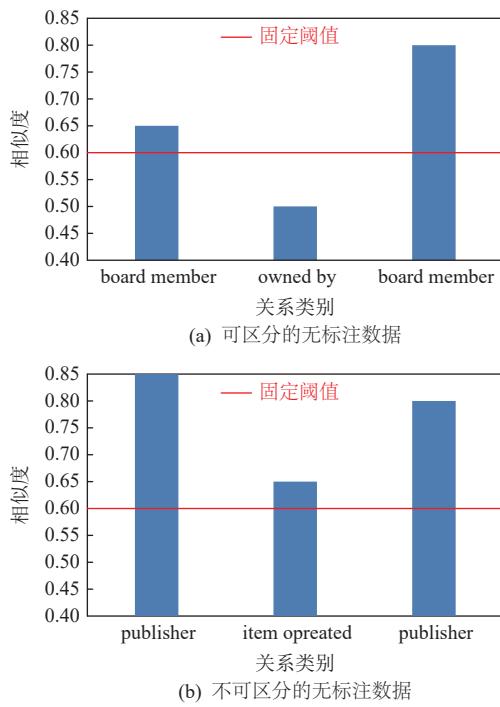


Fig. 2 Similarity scores in unlabeled data

图2 无标注数据相似度得分

2) 神经雪球模型对新入选的无标注数据予以相同的置信度, 并将其纳入关系抽取模型的训练当中. 然而, 这种不加区分的训练, 会进一步放大噪声数据的影响, 势必会造成关系分类器的性能偏移.

针对上述2种缺陷, 本文提出一种基于适应性自训练的少样本关系抽取模型(adaptive self-training relation extraction, Ada-SRE). 具体地, 对于缺陷1), 我们从模型无关的元学习^[16](model-agnostic meta-learning, MAML)中得到启发, 提出一种自适应阈值模块. 不同于传统的有监督学习, MAML通过模型无关的元任务训练来获取模型参数的梯度停靠点. 由此, 即使在只有少量的新关系样本的情况下, 模型也能实现快速泛化. 借助于MAML的快速泛化能力, 自适应阈值模块同样可通过元任务训练去学习关系阈值的梯度停靠点. 当出现新的关系类别时, 能够通过较少的梯度更新获取最适合当前关系类别的阈值. 值得注意的是, MAML仅仅适用于标注数据, 而自适应阈值模块面向的则是无标注数据, 用于确定特定关系类别的阈值选择. 对于缺陷2), 我们提出一种基于梯度反馈的赋权策略, 该策略通过梯度反向传播计算费雪信息矩阵(Fisher information matrix), 获取对新加入的无标注数据的权重分配. 其中, 置信度低的无标注数据权重低, 从而降低噪声数据的影响.

我们在公开的少样本关系抽取数据集上进行一

系列的实验分析. 实验结果证明在少样本场景下, 与基准模型相比, Ada-SRE取得明显的性能提升. 此外, 无标注数据预测分布证明自适应阈值模块在提升召回率的同时有效降低噪声数据的引入, 而消融实验证明各个模块的有效性.

1 相关工作

本节首先介绍传统的关系抽取模型的原理与算法; 接着分析少样本环境下的关系抽取任务, 并讨论现有模型的利弊.

1.1 关系抽取

关系抽取致力于识别在给定句子中的2个实体间的关系. 传统关系抽取建模属于有监督学习, 需要大量标注数据进行训练^[1], 它们可以划分为3种类别: 基于高斯核^[5-6]、基于特征^[7]以及基于神经网络^[8]. 其中, 基于高斯核的关系抽取模型主要设计一组核函数计算2个关系示例之间的相似度. 这些核函数包含句法序列核^[17]、句法树核^[18]、句法依赖树核^[19]、依赖图路径核^[20]以及句法组合树核^[21]. 而基于特征的关系抽取模型关注于为关系分类器生成一组特征, 例如: 词特征^[22]、句法特征^[23]以及语义特征^[24]. 不同于需要人工参与的基于高斯核和基于特征的方法, 基于神经网络的关系抽取模型关注于利用神经网络抽取关系特征实现端到端训练(end-to-end training), 例如: 卷积神经网络^[25]、图卷积神经网络^[26]、递归神经网络^[27]等.

然而, 这些传统的关系抽取模型无法应对不断涌现的新关系类别和标注样本匮乏的困境. 本文聚焦于解决少样本关系抽取建模问题.

1.2 少样本关系抽取

少样本关系抽取建模一般是通过元学习的训练范式得以实现. 其中, 每个元任务(或称之为片段)包含一个支撑集和一个查询集, 如图1所示, 元任务利用支撑集获取关系知识, 并在查询集上进行验证对网络参数进行更新.

这一训练范式可以划分为基于度量^[13,28-29]、基于模型^[30-31]和基于优化^[16,32-33]的3种模式. 具体地, 基于度量的元训练模型通过设计度量准则衡量支撑集和查询集之间的相似度; 基于模型的元训练模型聚焦于设计模型的结构实现少样本任务; 而基于优化的元训练模型尝试初始化网络参数, 使得模型能在一步或者几步的梯度更新过程中快速拟合到新任务上. 常用的元学习模型已经能在少样本关系抽取任务上

得到应用并取得较好的性能,例如:图神经网络^[12]、神经注意力元学习网络^[14]、原型网络^[13]和元网络^[11]。此外, Gao 等人^[34]应用混合注意力机制突出重要示例和特征。Ye 等人^[35]利用子图分析查询集和支撑集之间的交互匹配信息,并且采用注意力机制聚合支撑示例从而计算关系原型。Soares 等人^[36]采用点积来衡量向量相似度。Qu 等人^[37]利用一个全局关系图谱指引每个关系原型的计算。

然而,这些元学习模型在关系抽取建模中对于无标注数据并未很好地利用。虽然神经雪球模型^[15]利用自助采样技术对关系抽取任务中的无标注任务进行初步探索,但是人为手动的阈值以及不加区分的无标注数据,促使我们对神经雪球模型做进一步改进,并提出一种基于适应性自训练的少样本关系抽取模型。

2 适应性自训练的少样本关系抽取建模

本节首先介绍对于少样本关系抽取任务的相关定义;接着对于神经雪球模型中存在的2个缺陷,分别提出2种相应的策略,也就是自适应阈值模块和基于梯度反馈的赋权策略。

2.1 相关问题定义

2.1.1 任务定义

给定一个句子示例 x ,它是由1组词序列和1对标记的头尾实体 e_h, e_t 组成,关系抽取任务致力于通过句子示例 x 预测这对头尾实体 e_h, e_t 之间的关系标签 r 。因为要衡量关系抽取模型对于新关系类别的抽取能力,遵循神经雪球模型中的任务设定(不同于传统的有监督学习和元学习模式)。给定一组大规模标注的关系抽取数据集 D^l ,一组新关系类别标注数据集 D^s (标注样本的数量较少,遵循 N -类 K -样本范式)和对应的查询集 D^q ,一组无标注数据集 D^u ,本文的目的是要检验关系抽取模型能否从包含现有关系类型、新关系类型以及未知关系类型的查询集中准确定位 D^q

中关系类别的能力。

2.1.2 神经雪球模型

为了应对这种应用困境, Gao 等人^[15]提出神经雪球模型。如图3所示,神经雪球模型一共分为预训练和新关系类别样本收集2个阶段。其中,预训练阶段:关系孪生网络和关系分类器共同在 D^l 进行预训练获取在旧类别上的关系抽取能力。而新关系类别增量阶段可划分为4个步骤。

相同头尾实体可能蕴含相同关系类别^[15],基于这个假设,我们通过 D^s 中的实体对对 D^u 进行检索,符合该假设的无标注数据聚合形成候选集 A_1 。具体地:

1) 利用关系孪生网络对候选集 A_1 中的示例进行相似度排序,选取前 k_1 个示例。此外,过滤出相似度小于阈值 α 的示例,获得初选的示例。

2) 利用初选的示例对关系分类器进行微调,提升关系分类器识别新关系类别的能力。

3) 利用微调后的关系分类器从 D^u 中再次选择前 k_2 个示例,并且过滤出预测分数小于阈值 β 的示例,获得候选集 A_2 。

4) 重复步骤2)~3),直到从 D^u 中再也找不到可标注的示例。

对于新关系类别样本收集阶段,我们可以清楚地发现,阈值的选择是决定无标注数据抽取质量的关键。然而,如图2所示,不同关系类别的相似度阈值的选择存在极大差异。以固定阈值简单地确定数据的标签并不能挖掘无标注数据的潜能。例如:过高的阈值选择降低无标注数据的标注召回率,并不能增加新关系类别的样本数量;而过低的阈值选择又增加噪声数据标签的可能性,毒害关系孪生网络和关系分类器。另外,对于所引入的无标签数据,神经雪球模型施以相同的置信度,增加了噪声样本影响模型性能的风险。

鉴于这2种缺陷,本文针对固定阈值问题,提出自适应阈值模块为每个关系类别获取相应阈值;而对于相同置信度问题,提出基于梯度反馈的赋权策

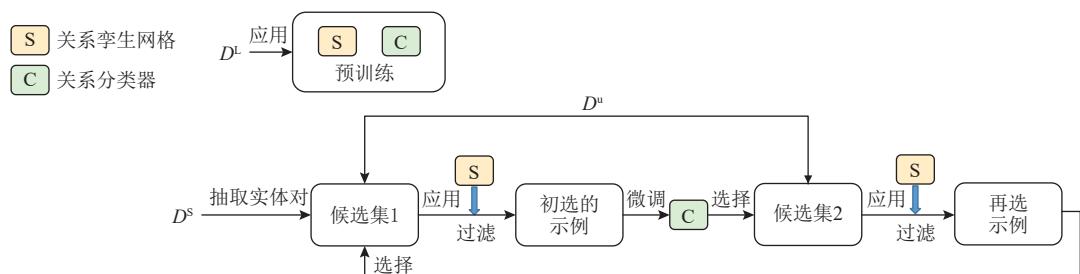


Fig. 3 Neural snowball model

图3 神经雪球模型

略,减少噪声数据的影响.

2.2 自适应阈值模块

在标注样本较少的情况下,无法利用传统的有监督学习直接获得每个关系类别对应的阈值.不过,元学习可以通过少量样本准确定位测试样本所对应的标签,这为我们提供解决该缺陷的一种思路.

具体地,在模型无关的元学习中,MAML通过学习模型的未知标签的泛化能力,获得标签的通用梯度停靠点.因此,当遇到新关系类别时,MAML可以通过该梯度停靠点和少量的标注样本,快速泛化到新关系类别对应的模型权重.因此,本文将标签阈值类比于MAML中的梯度,致力于学习所有类别的通用阈值停靠点,使得模型能够在少量样本情况下,快速泛化到未知标签的最佳阈值点.

正式地,假设阈值 α 和关系分类器 f_c 在第 t 步的参数为 w^t ,通过相似度计算,可以将无标注数据集 D^u 分割成2部分,即 $D^u \rightarrow D^{se} + D^{uns}$,其中 D^{se} 是已选集合而 D^{uns} 是未选集合.接下来,对于新入选的 D^{se} ,可以用来更新 f_c 的参数 w^t ,即

$$w^{t+1} = w^t - \lambda_w \frac{\partial L_{f_c}^{w^t}(D^{se})}{\partial w^t}, \quad (1)$$

其中 λ_w 是参数学习率, $L_{f_c}^{w^t}(\cdot)$ 是关系分类器在参数 w^t 下的损失函数,而 w^{t+1} 是时刻 $t+1$ 关系分类器更新后的参数.

接下来,通过在查询集 D^q 的损失函数计算,可以确定阈值 α 的更新方向,即

$$\alpha = \alpha + \lambda_\alpha (L_{f_c}^{w^{t+1}}(D^q) - L_{f_c}^{w^t}(D^q)), \quad (2)$$

其中 λ_α 是阈值 α 的学习率.值得注意的是,当 $L_{f_c}^{w^{t+1}}(D^q) < L_{f_c}^{w^t}(D^q)$ 时,说明已选集合 D^{se} 对关系分类器 f_c 有促进作用.因此,阈值 α 应该进一步减小,扩大已选集合 D^{se} ;反之, $L_{f_c}^{w^{t+1}}(D^q) > L_{f_c}^{w^t}(D^q)$ 则说明过小的阈值 α 导致已选集合 D^{se} 中充满噪声,需要进一步增大阈值 α ,降低 D^{se} 的噪声率.

以上是阈值 α 的一种软更新策略,基于这一准则,还可以采用一种硬更新策略.具体而言,将未选集合 D^{uns} 按照相似度从大到小排列,获得降序未选集合 D^{uns} ;将已选集合 D^{se} 按照相似度从小到大排列,获得升序已选集合 D^{se} .因此,阈值硬更新策略可表示为

$$\alpha = \begin{cases} T_P(D_{des}^{uns}), & L_{f_c}^{w^{t+1}}(D^q) \leq L_{f_c}^{w^t}(D^q), \\ T_P(D_{asc}^{se}), & L_{f_c}^{w^{t+1}}(D^q) > L_{f_c}^{w^t}(D^q), \end{cases} \quad (3)$$

其中 $T_P(\cdot)$ 是定位函数,它在样本序列中定位第 P 个样本,返回其相似度大小.

最后,基于式(1)~(3)并结合MAML的训练范式,

可以在少量样本的参与下,为每个关系类别标签快速获得其对应的阈值.

将自适应模块的训练过程总结,如算法1所示.

算法1. 自适应阈值模块训练算法.

输入: D^L, D^u , 内外循环学习率 λ_{in} 和 λ_{out} ;

输出: 阈值 α .

① 随机初始化阈值 α ;

② while not done do

③ 采样: $D_i^s, D_i^q \sim p(D^L)$; /*遵循N类K样本

从 D^L 中采样1组元任务*/

④ for all D_i^s, D_i^q do

⑤ 更新阈值

$$\alpha'_i = \alpha + \lambda_{in} \frac{\partial L_\alpha(D_i^s, D_i^q)}{\partial \alpha};$$

⑥ end for

⑦ 更新阈值

$$\alpha = \alpha + \lambda_{out} \frac{\partial \sum_{D_i^s, D_i^q \sim p(D^L)} L_{\alpha_i}(D_i^s, D_i^q)}{\partial \alpha};$$

⑧ end while

其中 $L_{\alpha_i}(D_i^s, D_i^q)$ 表示在阈值 α_i 下输入为 D_i^s, D_i^q 的损失函数.

2.3 基于梯度反馈的赋权策略

虽然自适应阈值模块可以解决不同标签的相似度阈值的选择,但是对新入选的 D^{se} 无差别相信会增大噪声数据的影响.因此,基于梯度反馈的赋权策略致力于获取 D^{se} 的权重,降低噪声数据的影响.

假定 $D^{se} = \{s_1, s_2, \dots, s_n\}$,其中 s_i 表示入选的第 i 个无标注示例.根据定义,费雪信息矩阵可以衡量对参数估计的准确程度.从另一个角度,可以由此确定所入选的无标注示例对参数估计的贡献程度.因此,改进费雪信息矩阵用于确定无标注示例 s_i 的权重 m_i .具体地,首先用无标注示例 s_i 更新 f_c 的参数 w ,即

$$w_i = w - \lambda_w \frac{\partial L_w(s_i)}{\partial w}. \quad (4)$$

之后,基于更新后的参数集合 w_i ,利用查询集 D^q 计算权重,即

$$m_i = \sum_{\theta \in S_w} \left(\frac{\partial \ln L_{w_i}(D^q)}{\partial \theta} \right)^2, \quad (5)$$

其中 S_w 是当前关系分类模型的参数集合.进一步地,将获得的权重序列进行归一化可以获得无标注示例的归一化权重序列,即

$$\{m'_1, m'_2, \dots, m'_n\} = softmax\{m_1, m_2, \dots, m_n\}. \quad (6)$$

最后,结合新学习的权重 m'_i ,式(2)中关系分类

器参数更新可以表示为

$$w = w - \lambda_w \sum_{s_i \in D^c} m'_i \frac{\partial L_w(s_i)}{\partial w}. \quad (7)$$

特别地,在测试阶段时,当没有查询集可供梯度反馈时,从支撑集 D^s 中随机选取10%的示例用于梯度反馈.

3 实验设置

3.1 数据集和评价准则

本文遵循神经雪球模型^[15]中的数据集设定,采用FewRel^[9]和Re-TACRED^[38]作为实验数据集.其中, FewRel从维基百科中共抽取100种关系类别,每种关系类别包含70 000个示例.同时, FewRel将自身划分成3个子集,分别为64类别的训练集、16类别的验证集以及20类别的测试集.与之相比, Re-TACRED则是从一个大规模关系抽取数据集中改进所得,它包含91 467个句子,分布于40种关系类别中.同样地,通过类别划分, Re-TACRED可以划分为30类别的训练集、5类别的验证集以及5类别的测试集.

此外,为了获得更为公正的评价,本文仍然采用神经雪球模型中收集到的大规模无标注语料库,它包含899 996个示例和464 218对实体对.同样地,我们进一步将训练集分割成训练集A和训练集B.在训练阶段,用训练集A作为 D^L .然后,在每步评价中,从验证集和测试集中无放回采样1种关系作为新关系 r ,并选择 k 个示例作为种子组成 D^s ,而查询集 D^q 是从训练集B和验证/测试集中采样得到的.值得注意的是, D^q 不仅包含 D^L 中标签数据丰富的关系类别,还可能来源于支撑集 D^s 中样本数目较少的关系类别,以及先前训练不存在的关系类别.这种设定相较于传统的少样本关系抽取更具有挑战性,而且更加符合现实世界的应用,因为在现实世界语料库并不会局限于一定关系数量或类别.

3.2 基准模型

为了验证基于Ada-SRE在少样本关系抽取建模上的性能,本文选择8种具有挑战性的方法作为基准模型.

1)基于分布式语义的自举关系抽取(bootstrapping relationship extractors with distributional semantics, BREDS)^[39]模型是原始雪球(snowball)模型^[40]的变形,它通过词嵌入进行模式选择;

2)微调分类器(fine-tuning classifier, FC)^[15]是通过微调关系分类器所获得的;

3)关系孪生网络(relational siamese network, RSN)^[15]

在预测查询集时,通过孪生网络计算与查询集中示例标签的相似度大小;

4)远监督(distant supervision, DS)模型^[15]是通过关系孪生网络和关系分类模型进行采样的关系抽取模型;

5)增量元自训练关系抽取(incremental meta self-training relation extraction, MetaSRE)模型^[41]采用一种元学习网络以防止模型因标签噪声而导致的漂移,并通过迭代自训练增强其鲁棒性;

6)多重引用图(multiple reference graph, MRefG)模型^[42]通过构建包含实体、动词以及语义引用在内的引用图,以实现语义和词法层面有效链接标注数据和无标注数据;

7)神经雪球(neural snowball, NS)模型^[15]是对传统雪球模型的改进,结合关系孪生网络和关系分类模型二者神经网络特点;

8)本文提出的基于适应性自训练的少样本关系抽取模型(Ada-SRE),包含软更新(Ada-SRE_S)和硬更新(Ada-SRE_H)策略.

3.3 参数设置

类似文献[15],在验证集上进行参数调节.其中编码器,应用BERT^[5]作为基础编码器进行研究.至于模型微调,在网格搜索后,我们采用训练轮数为50次,批次大小为10,学习率为0.05,而负样本损失系数为0.2.在BERT^[5]上微调采用相同的参数,除了学习率和负样本系数分布分别设为0.01和0.5.在神经雪球过程中,每个阶段从无标注数据集加入的示例同样设定为5.但是不同于文献[15]中的人为阈值设定,我们每次利用自适应阈值模块进行计算,其中学习率分别设定为 $\lambda_w=0.005, \lambda_{in}=0.001, \lambda_{out}=0.001$.此外,对于硬更新策略,我们将更新步数设定为1.

3.4 实验问题

本文将分别从5个实验问题用于验证模型的有效性.

1)在总体性能方面,Ada-SRE是否可以超过现有的基准模型;

2)添加自适应阈值模块后,关系孪生网络的性能能否获得提升;

3)随着迭代步数的增加,Ada-SRE是否都能保持对于神经网络模型的优势;

4)当应用硬更新于自适应阈值模块,定位函数 $T_P(\cdot)$ 中位置参数 P 是否对模型性能产生影响;

5)对于自适应阈值模块和基于梯度反馈赋权模

块,哪一模块对 Ada-SRE 的性能影响更大.

我们将在第 4 节对这 5 个研究问题依次分析,探讨少样本场景下关系抽取性能.

4 实验分析与讨论

根据实验问题,本节依次进行实验分析与讨论.

4.1 整体性能

表 1 和表 2 分别展示了少样本关系抽取模型在数据集 FewRel 和 Re-TACRED 上的整体性能. 整体而言,随着种子示例的增加,所有模型的精确率、召回率以及 $F1$ 值($F1$ score)都得以提升. 这说明种子示例的增加,增强了模型的有监督信号,进而使得模型性能也有所提升. 特别地,Ada-SRE 模型在任意种子示例的参与下仍能取得最好的模型性能,这证明 Ada-

SRE 模型的有效性.

在基准模型当中,传统基于统计性指标的 BREDS 在众多模型中表现最为欠佳. 这是因为传统的有监督模型在样本量较少的情况下,无法规避在测试集中的过拟合风险. FC 和 RSN 虽然可以取得较为可观的召回率,然而无差别地引入无标注数据,反而会引入噪声数据,相对应的精确率也会不可避免地降低. 而 NS 是众多基准模型中表现最为突出的一个,获得了最佳的精确率和 $F1$ 值.

将最优基准模型 NS 和 Ada-SRE 加以比较,我们不难发现,无论是什么评价指标,Ada-SRE 都更胜一筹,获得全面提升,在召回率方面的提升尤为显著. Ada-SRE 在种子示例为 5, 10 和 15 时,在 FewRel 和 Re-TACRED 上分别取得了 36.6%, 16.5%, 23.8% 和 33.20%, 25.16%, 29.38% 的性能提升. 这说明 Ada-

Table 1 Results of Few-Shot Relation Extraction Models on FewRel

表 1 少样本关系抽取模型在 FewRel 上的结果

模型	5 个种子示例			10 个种子示例			15 个种子示例			%
	精确率	召回率	$F1$ 值	精确率	召回率	$F1$ 值	精确率	召回率	$F1$ 值	
BREDS	33.71	11.89	17.58	28.29	17.02	21.25	25.24	17.96	20.99	
FC	50.85	16.66	25.10	59.87	55.19	57.43	<u>81.60</u>	58.92	68.43	
RSN	39.07	<u>51.39</u>	44.47	42.42	54.93	47.87	44.10	52.73	48.03	
DS	38.06	51.18	43.66	38.45	<u>76.12</u>	51.09	35.48	<u>80.33</u>	49.22	
MetaSRE	55.03	38.56	45.35	60.39	57.35	58.83	76.35	65.82	70.70	
MRefG	53.78	37.35	44.08	58.54	53.14	55.71	72.53	62.57	67.18	
NS	<u>56.87</u>	40.43	<u>47.26</u>	<u>60.50</u>	62.20	<u>61.34</u>	78.13	66.87	<u>72.06</u>	
Ada-SRE _s	60.34	54.80	57.44	61.86	74.52	67.60	82.65	80.23	81.42	
Ada-SRE _h	59.24	55.21	57.15	61.37	73.23	66.78	81.45	82.78	82.11	

注:下划线表示每列中表现最好的基准模型,加粗表示每列中性能最好的模型.

Table 2 Results of Few-Shot Relation Extraction Models on Re-TACRED

表 2 少样本关系抽取模型在 Re-TACRED 上的结果

模型	5 个种子示例			10 个种子示例			15 个种子示例			%
	精确率	召回率	$F1$ 值	精确率	召回率	$F1$ 值	精确率	召回率	$F1$ 值	
BREDS	29.85	13.45	18.54	26.40	16.52	20.32	24.28	15.69	19.06	
FC	43.69	33.02	37.61	50.77	39.25	44.27	<u>60.30</u>	40.86	48.71	
RSN	32.53	<u>47.29</u>	38.55	35.28	44.06	39.18	37.41	45.92	41.23	
DS	33.31	46.53	38.83	34.84	<u>52.69</u>	41.94	36.34	<u>56.37</u>	44.19	
MetaSRE	47.49	35.48	40.62	52.74	42.84	47.28	58.57	43.56	49.96	
MRefG	47.15	35.59	40.56	52.39	42.95	47.20	58.49	42.78	49.42	
NS	<u>48.10</u>	36.57	<u>41.55</u>	<u>54.48</u>	43.17	<u>48.17</u>	59.68	44.38	50.91	
Ada-SRE _s	53.40	48.71	50.92	57.70	54.03	55.80	61.38	56.25	58.70	
Ada-SRE _h	50.38	47.61	48.95	58.67	53.52	55.83	62.95	57.42	60.23	

注:下划线表示每列中表现最好的基准模型,加粗表示每列中性能最好的模型.

SRE 可以高质量地从无标注数据集中收集新的训练示例, 在一定程度上避免了 NS 由于固定阈值和同等权重等因素带来的负面影响。此外, 在 Ada-SRE 中, 软更新在种子示例较少的情况下表现更佳, 硬更新则是在种子示例较多的情况下有更优性能。

4.2 自适应阈值模块的提升效能

为了检验在 NS 和 Ada-SRE 中关系孪生网络的抽取能力, 在 FewRel 上随机采样 1 个关系以及所属关系的 5 个类别作为支撑集, 而余下的数据作为查询集。将训练过的 NS, Ada-SRE_S 和 Ada-SRE_H 中的关系孪生网络抽取出来, 分别作为查询集中的每个示例计算相似度大小, 并且采用 P@N(precision at top-N) 作为评价指标。

从表 3 中可以看出, 任意一个模型的关系孪生网络在训练后, 只要给定 5 个种子示例都能达到一定的分类准确度。特别地, 在测试集 P@5 列中, 所有模型都可以达到 80% 以上的分类性能。这说明每增加 5 个示例, 其中至少有 4 个示例是准确的, 而这也是我们每次从无标注数据集选 5 个示例的原因。

Table 3 RSN Results of Different Models

表 3 不同模型关系孪生网络的结果 %

模型	训练集				测试集			
	P@5	P@10	P@20	P@50	P@5	P@10	P@20	P@50
NS	83.60	80.66	76.03	61.98	82.15	78.64	72.57	55.10
Ada-SRE _S	85.78	82.73	78.32	62.22	85.12	81.24	75.32	60.24
Ada-SRE _H	84.35	81.91	79.03	61.85	84.33	80.18	75.89	59.25

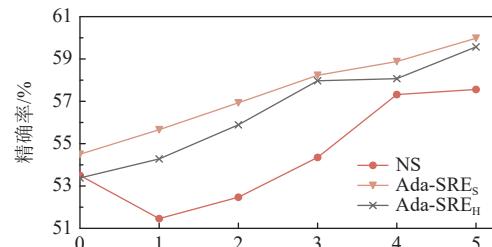
注: 加粗表示每列中性能最好。

此外, 相比于 NS 的关系孪生网络, 添加自适应阈值和梯度反馈模块的 Ada-SRE 的关系孪生网络可以取得更好的分类效果, 这进一步证明选择合适阈值和合适权重的有效性。

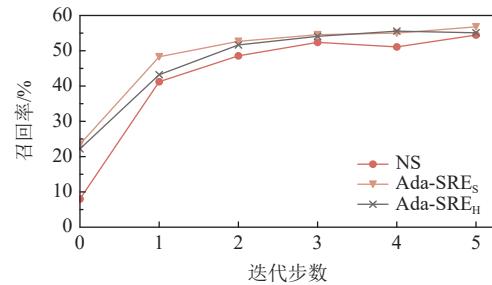
4.3 迭代步数的影响

为了进一步分析迭代步数对于 NS 和 Ada-SRE 的影响, 在 FewRel 中随机选取 1 个新加入的关系 chairperson, 并基于 5 个种子示例分析迭代步数对模型分类性能的影响。由于空间所限制, 这里只展现 1 种关系的分类性能。图 4 中分别展现各个模型在不同迭代步数下精确率和召回率的变化。

总体而言, 随着迭代步数的增加, 各个模型的精确率和召回率也随之增加。特别地, 召回率的增长尤为明显, NS 从起初的召回率为 0 到迭代 5 次后增长到接近 60%。而召回率的提升也进一步带动模型在精确率方面的提升。而随着迭代步数增长到 5, 召回率



(a) 精确率随迭代步数的变化



(b) 召回率随迭代步数的变化

Fig. 4 Model performance on different iteration steps

图 4 不同迭代步数的模型性能

的提升也渐缓, 这说明模型从无标注数据集挖掘的信息也逐渐趋向于饱和。

此外, 将 NS 与 Ada-SRE 相比可以发现, 无论是在精确率还是召回率方面, Ada-SRE 始终保持着对 NS 的领先地位。特别地, Ada-SRE 在初始召回率比 NS 领先近 20%, 这也证明 Ada-SRE 在提升无标注数据集召回质量方面的有效性。而至于软硬更新之间的差异, Ada-SRE_S 也能保持对 Ada-SRE_H 的微弱领先。

4.4 参数敏感性

在 Ada-SRE_H 中, 定位函数 $T_P(\cdot)$ 中位置参数 P 对模型性能的影响很大。因此, 分析位置参数 P 对模型的 $F1$ 值的影响。

从表 4 中可以看出, 无论有多少种子示例的参与, 当位置参数 $P=1$ 时, Ada-SRE_H 的性能都达到最优。特别地, 当位置参数 P 上升到 2 时, Ada-SRE_H 性能开始急剧下降。这种现象可能是因为当迭代步数一定时, 模型从无标注数据集中挖掘信息的能力已经达到极限, 这时定位 2 个以上的无标注数据集会不必要地

Table 4 F1 Score of Ada-SRE_H Under Different P

表 4 不同 P 下 Ada-SRE_H 的 $F1$ 值

P	5 个种子	10 个种子	15 个种子
0	47.26	61.34	72.06
1	57.15	66.78	82.11
2	40.35	49.14	64.34
3	24.22	32.02	45.79

注: 加粗表示每列中性能最好。

引入多余的噪声,造成模型性能的下降.

4.5 消融实验

为了验证哪一个模块对模型性能影响更大,将模型的各个子模块去除或者替换成 Ada-SRE 中原有的模块,用符号“--”表示. 例如:“Ada-SRE_S--weight”表示应用软更新策略的 Ada-SRE 去除梯度反馈模块,而将每个无标注示例视为同等重要. 特别地,无论是什么样的更新策略,去除自适应阈值模块之后,Ada-SRE 都只剩下梯度反馈模块,因此我们都将其表示为“Ada-SRE--threshold”.

然后,在种子示例为 5, 10, 15 时分别检验模型的性能,总结于表 5. 从表 5 中可以发现, Ada-SRE_S 和 Ada-SRE_H 同时去除梯度反馈模块之后, 模型性能都会获得相应的下降, 这证明梯度反映模块的有效性. 另外, 当只保留自适应阈值模块后, 软更新策略始终会优于硬更新策略. 除此之外, 当模型去除自适应阈值模块, 也就是只保留梯度反馈模块, 模型的性能下降得尤为明显. 这一现象反映在提升少样本关系分类性能方面, 自适应阈值模块相比于梯度反映模块能够发挥更大的作用.

Table 5 F1 Score on Ablation Study

模型	表 5 消融实验下的 F1 值			%
	5 个种子	10 个种子	15 个种子	
Ada-SRE _S --weight	56.52	65.23	80.67	
Ada-SRE _H --weight	55.83	64.51	78.85	
Ada-SRE--threshold	49.59	60.74	73.56	

注: 加粗表示每列中性能最差.

5 总结与未来展望

本文聚焦于现实应用中广泛存在的少样本关系抽取任务, 通过分析现有神经雪球模型在设计中存在的缺陷, 提出适应性自训练算法 Ada-SRE 进行改进. 针对神经雪球模型设定统一固定阈值的缺陷, 本文基于元学习的思想提出自适应阈值模块, 能够为每个关系类别提供合适的阈值选择. 而针对神经雪球模型对入选的无标注示例施以相同置信度问题, 本文提出基于梯度反馈的赋权策略, 为每个人选的示例提供特定的权重, 避免噪声数据的干扰. 此外, 通过在 2 个公开数据集上进行实验分析, 发现 Ada-SRE 能够取得优于当前最优的神经雪球模型更好的实验结果.

在未来的工作, 我们将从 2 个方面开展工作: 一

方面, 将探究本该入选但低于阈值的无标注示例和入选示例的关系, 在保持模型精确率的同时, 进一步提升模型的召回率; 另一方面, 鉴于提示学习(prompt learning)在少样本领域的快速发展, 将探究提示学习在无标注数据参与下的关系抽取任务上的应用.

作者贡献声明: 陈洪辉提出了模型的算法思路和实验问题; 郑建明基于提出的实验问题完成实验并撰写论文; 蔡飞负责论文的修改和校对, 并提出指导意见; 韩毅参与了论文校对和实验数据分析整理工作.

参 考 文 献

- [1] Wang Hailin, Lu Guoming, Yin Jin, et al. Relation extraction: A brief survey on deep neural network based methods [C] //Proc of the 4th Int Conf on Software Engineering and Information Management. New York: ACM, 2021: 220–228
- [2] Xia Wei, Wang Shanlei, Yin Zidu, et al. Mutual information based modeling and completion of correlations in knowledge graphs[J]. *Journal of Frontiers of Computer Science and Technology*, 2018, 12(7): 1064–1074 (in Chinese)
(夏维, 王珊蕾, 尹子都, 等. 基于互信息的知识图谱实体关联关系建模与补全[J]. *计算机科学与探索*, 2018, 12(7): 1064–1074)
- [3] Zhang Fangrong, Yang Qing. Research on entity relation extraction method in knowledge-based question answering[J]. *Computer Engineering and Applications*, 2020, 56(11): 219–224 (in Chinese)
(张芳容, 杨青. 知识库问答系统中实体关系抽取方法研究[J]. *计算机工程与应用*, 2020, 56(11): 219–224)
- [4] Zhou Mengjia. Research on relation extraction for dialogue text [D]. Wuhan: Wuhan University, 2021 (in Chinese)
(周孟佳. 面向对话文本的关系抽取研究[D]. 武汉: 武汉大学, 2021)
- [5] Despina C, Grigoris T. Improving distantly-supervised relation extraction through BERT-based label and instance embeddings[J]. *IEEE Access*, 2021, 9: 62574–62582
- [6] Garg S, Galstyan A, Steeg G, et al. Kernelized hashcode representations for relation extraction [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 6431–6440
- [7] Gormley M, Yu Mo, Dredze M. Improved relation extraction with feature-rich compositional embedding models [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 1774–1784
- [8] 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
- [9] Han Xu, Zhu Hao, Yu Pengfei, et al. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 4803–4809

- [10] Vrandecic D, Krötzsch M. Wikidata: A free collaborative knowledgebase[J]. *Communications of the ACM*, 2014, 57(10): 78–85
- [11] Munkhdalai T, Yu Hong. Meta networks [C] //Proc of the 34th Int Conf on Machine Learning. New York: PMLR, 2017: 2554–2563
- [12] Satorras V, Estrach J. Few-shot learning with graph neural networks [C/OL] //Proc of the 6th Int Conf on Learning Representations. 2018[2022-08-24].<https://openreview.net/forum?id=BJj6qGbRW>
- [13] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning [C] //Proc of Annual Conf on Neural Information Processing Systems 2017. New York: PMLR, 2017: 4077–4087
- [14] Mishra N, Rohaninejad M, Chen Xi, et al. A simple neural attentive meta-learner [C/OL] //Proc of the 6th Int Conf on Learning Representations. 2018[2022-08-24].<https://openreview.net/forum?id=B1DmUzWAW>
- [15] Gao Tianyu, Han Xu, Xie Ruobing, et al. Neural snowball for few-shot relation learning [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 7772–7779
- [16] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks [C] //Proc of the 34th Int Conf on Machine Learning. New York: PMLR, 2017: 1126–1135
- [17] Ghosh D, Muresan S. Relation classification using entity sequence kernels [C] //Proc of the 24th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2012: 391–400
- [18] Leeuwenberg A, Buzmakov A, Toussaint Y, et al. Exploring pattern structures of syntactic trees for relation extraction [C] // Proc of the 13th Int Conf on Formal Concept Analysis. Berlin: Springer, 2015: 153–168
- [19] Cho C, Choi Y. Dependency tree positional encoding method for relation extraction [C] // Proc of the 36th ACM/SIGAPP Symp on Applied Computing. New York: ACM, 2021: 1012–1020
- [20] Shi Yong, Xiao Yang, Quan Pei, et al. Distant supervision relation extraction via adaptive dependency-path and additional knowledge graph supervision[J]. *Neural Networks*, 2021, 134: 42–53
- [21] Reichartz F, Korte H, Paass G. Composite kernels for relation extraction [C] //Proc of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th Int Joint Conf on Natural Language Processing. Stroudsburg, PA: ACL, 2009: 365–368
- [22] Bhamare B, Prabhu J. A supervised scheme for aspect extraction in sentiment analysis using the hybrid feature set of word dependency relations and lemmas[J]. *PeerJ Computer Science*, 2021, 7: e347
- [23] Ravi S, Larochelle H. Optimization as a model for few-shot learning [C/OL] //Proc of the 5th Int Conf on Learning Representations. 2017[2022-08-24].<https://openreview.net/forum?id=rJY0-Kcll>
- [24] Yu Yang, Wang Guohua, Ren Haopeng, et al. Incorporating bidirection-interactive information and semantic features for relational facts extraction (student abstract) [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 15947–15948
- [25] Wen Haixu, Zhu Xinhua, Zhang Lanfang, et al. A gated piecewise CNN with entity-aware enhancement for distantly supervised relation extraction[J]. *Information Processing & Management*, 2020, 57(6): 102373–102373
- [26] Niu Weicai, Chen Quan, Zhang Weiwen, et al. GCN2-NAA: Two-stage graph convolutional networks with node-aware attention for joint entity and relation extraction [C] //Proc of the 13th Int Conf on Machine Learning and Computing. New York: ACM, 2021: 542–549
- [27] Peng Yifan, Rios A, Kavuluru R, et al. Chemical-protein relation extraction with ensembles of SVM, CNN, and RNN models [J]. arXiv preprint, arXiv: 1802.01255, 2018
- [28] Prabhudesai M, Lal S, Patil D, et al. Disentangling 3D prototypical networks for few-shot concept learning [C/OL] //Proc of the 9th Int Conf on Learning Representations. 2021 [2022-08-24].<https://openreview.net/forum?id=Lr-u0b42he>
- [29] Sung F, Yang Yongxin, Zhang Li, et al. Learning to compare: Relation network for few-shot learning [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2018: 1199–1208
- [30] Obamuyide A, Vlachos A. Model-agnostic meta-learning for relation classification with limited supervision [C] //Proc of the 57th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 5873–5879
- [31] Santoro A, Bartunov S, Botvinick M, et al. Meta-learning with memory-augmented neural networks [C] //Proc of the 33rd Int Conf on Machine Learning. New York: JMLR, 2016: 1842–1850
- [32] George V, Morar V, Yang Weiwei, et al. Learning without gradient descent encoded by the dynamics of a neurobiological model [J]. arXiv preprint, arXiv: 2103.08878, 2021
- [33] Huang Posen, Wang Chenglong, Singh R, et al. Natural language to structured query generation via meta-learning [C] //Proc of the 2018 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2018: 732–738
- [34] Gao Tianyu, Han Xu, Liu Zhiyuan, et al. Hybrid attention-based prototypical networks for noisy few-shot relation classification [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 6407–6414
- [35] Ye Zhixiu, Ling Zhenhua. Multi-level matching and aggregation network for few-shot relation classification [C] //Proc of the 57th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 2872–2881
- [36] Soares L, FitzGerald N, Ling J, et al. Matching the blanks: Distributional similarity for relation learning [C] //Proc of the 57th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 2895–2905
- [37] Qu Meng, Gao Tianyu, Xhonneux L, et al. Few-shot relation extraction via Bayesian meta-learning on relation graphs [C] //Proc of the 37th Int Conf on Machine Learning. New York: PMLR, 2020: 7867–7876
- [38] Stoica G, Platanios E, Póczos B. Re-TACRED: Addressing shortcomings of the TACRED dataset [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 13843–13850
- [39] Batista D, Martins B, Silva M. Semi-supervised bootstrapping of

- relationship extractors with distributional semantics [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 499–504
- [40] Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections [C] //Proc of the 5th ACM Conf on Digital Libraries. New York: ACM, 2000: 85–94
- [41] Hu Xuming, Zhang Chenwei, Ma Fukun, et al. Semi-supervised relation extraction via incremental meta self-training [C] //Proc of the 2021 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 487–496
- [42] Li Wanli, Qian Tieyun, Chen Xu, et al. Exploit a multi-head reference graph for semi-supervised relation extraction [C] //Proc of the 31st Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2021: 1–7



Chen Honghui, born in 1969. PhD, professor, PhD supervisor. His main research interests include recommendation systems, information retrieval and natural language processing.
陈洪辉, 1969 年生. 博士, 教授, 博士生导师. 主要研究方向为推荐系统、信息检索和自然语言处理.



Zheng Jianming, born in 1993. PhD candidate. His main research interests include event representation and few-shot learning.
郑建明, 1993 年生. 博士研究生. 主要研究方向为事例表示和少样本学习.



Cai Fei, born in 1984. PhD, associate professor. His main research interests include recommend system, information retrieval, and query formulation.
蔡飞, 1984 年生. 博士, 副教授. 主要研究方向为推荐系统、信息检索和查询表达.



Han Yi, born in 1993. PhD, lecturer. His main research interests include event representation and few-shot learning.
韩毅, 1993 年生. 博士, 讲师. 主要研究方向为事例表示和少样本学习.