

利用类型语义表示进行标签降噪的细粒度实体分类^①

席鹏弼^{②***} 靳小龙^{***} 白 硕^{***} 程学旗^{***}

(* 中国科学院计算技术研究所网络数据科学与技术重点实验室 北京 100190)

(** 中国科学院大学计算机科学与技术学院 北京 100408)

(*** 恒生电子股份有限公司 杭州 310053)

摘 要 细粒度实体分类(FET)任务的训练数据往往利用已有知识库中的知识通过远程监督方法进行生成,生成过程中不可避免地引入多余的噪音标签。现有考虑训练数据中噪音问题的方法通常只建模训练数据和标注类型的概率分布,对细粒度类型的语义信息学习不足,造成在标注了多个细粒度类型的训练数据上选择了与实体上下文不相关的类型进行模型的学习。本文提出一种利用细粒度类型的语义表示进行标签降噪的细粒度实体分类方法。首先利用训练数据中具有唯一细粒度类型路径的数据学习一部分细粒度类型的表示,进而结合细粒度类型间的关系信息学习其他细粒度类型的表示;其次在标注了细粒度类型的训练数据中选取与实体上下文的语义信息最相似的细粒度类型为目标类型,从数据集中选择 *Top-K* 个相似数据进行细粒度类型语义信息的聚合;最后在聚合信息上学习最终的细粒度实体分类模型。实验结果表明,该方法可以有效地从标注了细粒度类型的训练数据中选出与实体上下文的语义信息最相符的细粒度类型,达到提升细粒度实体分类准确率的效果。

关键词 实体分类;细粒度类型;多标签降噪;多标签分类

随着人工智能技术的兴起,各类应用系统需要构建更多语义较为精确的领域知识图谱。细粒度实体分类(fine-grained entity typing, FET)任务为实体确定若干个树形结构的、语义信息更为精确的类型,在领域知识图谱构建过程中发挥着关键的作用。

在细粒度实体分类任务中,知识图谱构建中预先定义的细粒度类型通常比较多,如 FINGER 数据集^[1]有 112 个类型、文献[2]构建的数据集有 505 个类型。同时,细粒度类型之间存在着树形层级结构关系,在语义上并不完全互斥^[3],信息粒度比命名实体抽取任务的类型更细^[4]。人工标注训练数据需要花费更长的时间理解实体及上下文信息及细粒度类型的信息,因此通过人工标注训练数据方法

形成该任务的训练数据成本高,难以满足应用对训练数据的需求^[5]。当前通用的方法是采用知识库远程监督生成训练数据^[6]。例如,该任务公开的数据集 BBN、OntoNotes、Wiki、DBpedia 普遍采用知识库 Freebase、Wikipedia 和 DBpedia 中实体的细粒度类型远程监督生成^[4,7-8]。远程监督方法^[4]先采用实体链接算法,将文本中标注的实体链接到知识库中,如 Freebase、DBpedia 等;接着通过知识库中实体的细粒度类型对文本中的实体进行类型标注,形成训练数据。在这个过程中,由于知识库中实体的细粒度类型缺少细粒度类型对应的上下文信息,生成的训练数据中会标注与实体所在的上下文中语义不相符的噪音细粒度实体类型^[9]。

① 国家自然科学基金(U1911401,61772501,62002341,U1836206)资助项目。

② 男,1981 年生,博士生;研究方向:知识图谱;联系人,E-mail: xipengbi@ict.ac.cn。

(收稿日期:2022-11-16)

现有考虑训练数据中噪音问题的细粒度实体分类研究存在对候选的细粒度实体类型的分析不足的情况,为后续的模型学习带来了在候选细粒度类型集中选取噪音标签进行细粒度实体分类监督学习的问题。具体地,采用频次删除训练数据的标注细粒度类型中频次较低细粒度类型的方法,完全没有考虑实体所在的上下文信息^[10];模型训练中选择预测概率最高的类型作为训练数据唯一正确的细粒度类型,带来 Confirmation bias 问题^[11],即当模型预测概率最高的类型为错误类型时,后续的模型优化过程中会在此基础上持续最大化该错误类型的得分^[12-13]。文献[12]使用 WordNet 外部词典学习细粒度类型文本的语义信息。因为训练数据中细粒度类型的语义信息是对应标注实体及实体上下文的语义信息,不同来源的类型其文本的语义信息不一定一致,难以有效降噪。

考虑训练数据中噪音问题的细粒度实体分类研究的主要难点有以下几点。(1)细粒度实体分类任务的数据集中缺少学习细粒度类型集上所有类型语义信息的数据。数据集只标注了实体及其上下文和对应的多个细粒度类型,因可能出现和上下文不相符的标签情况,难以确定类型和数据的对应关系,甚至树形结构的细粒度类型集中的部分类型没有训练数据。(2)树形结构的细粒度类型之间在语义上并不完全独立,存在语义上的包含关系,分类模型难以学习到类型之间的语义上的信息。(3)训练数据降噪过程使细粒度实体类型的训练数据规模有较大减少,造成细粒度实体类型的训练数据过少,细粒度类型语义难以学习的问题。本文对降噪处理前后的细粒度实体分类任务常用的数据集对比统计,结果发现,常用数据集 Wiki、OntoNotes 和 BBN 的训练数据的总数分别减少了 36%、29% 和 21%。

最近有研究通过训练数据中实体指称词的表示学习部分细粒度实体类型的表示^[14],并构建图学习实体指称词的特征。该研究说明通过训练数据中的实体及上下文可以学习到训练数据中对应的细粒度类型的表示。在知识推断中可以用实体关系三元组学习实体的语义信息,同样,可以用该任务训练数据生成过程中使用的知识库中的细粒度实体类型间关

系三元组学习细粒度实体类型的语义信息。

针对上述细粒度实体分析的难点问题研究和技术的整理,本文提出实体类型相似过滤细粒度实体分类(entity type similarity filtering fine-grained entity typing, ETSF-FET)模型。该方法首先通过数据集中的具有唯一细粒度类型路径的数据,基于 ELMo 预训练模型学习数据中细粒度类型的表示^[15],结合细粒度类型间的关系信息数据^[16],学习预定义细粒度类型集中剩下的细粒度类型表示;然后对数据集中的噪音标签数据,在候选细粒度类型集中选取与实体的语义信息最相似的细粒度实体类型作为监督学习细粒度类型;最后针对降噪处理造成的细粒度实体类型的训练数据规模有较大减少的情况,设计相似实体上下文信息的语义聚合方法,补充细粒度实体类型训练数据信息,从数据集层面学习更多的信息,建模实体在细粒度类型集上的分类模型。本文的主要贡献概括如下。

(1)提出了细粒度实体分类任务中通过训练数据中实体及上下文的嵌入表示、实体对应的细粒度类型、远程监督知识库中的细粒度类型间关系进行细粒度类型嵌入表示的学习方法。

(2)针对训练数据降噪造成细粒度实体类型的训练数据少的问题,提出了基于语义聚合的细粒度实体类型特征学习方法,通过数据集中相似实体及其上下文信息的语义聚合,学习细粒度实体类型的语义特征。

(3)提出了基于细粒度类型语义表示进行训练数据标签降噪的细粒度实体分类模型 ETSF-FET。先采用前面提出的细粒度类型嵌入表示的学习方法生成细粒度类型的表示;接着通过实体和上下文信息的表示和细粒度类型表示的相似度打分,为数据集中噪音数据选择最符合实体上下文的细粒度类型;然后采取前面提出的基于语义聚合的细粒度实体类型特征学习方法学习语义信息;最后,用最符合实体上下文的细粒度类型和基于语义聚合方法学习的语义信息训练细粒度实体分类模型。

(4)用细粒度实体分类任务的常用 Wiki、OntoNotes、BBN 数据集,对 ETSF-FET 方法进行细粒度实体分类实验,性能指标较当前方法有所提升。表

明通过 ETSF-FET 方法有效地在候选类型中选取了更符合上下文信息的细粒度类型,达到提升细粒度实体分类准确率的效果。

1 相关工作

本文从训练数据降噪方法这个角度整理现有工作,具体有实体及上下文特征学习类研究、细粒度类型间关系信息应用类研究、数据集信息增强类研究 3 个方面的研究工作。

早期的研究人员主要关注实体及其上下文特征的学习,通过训练数据学习细粒度类型对应的训练数据的更多特征,采用学习到的特征,将细粒度类型的表示映射到同一高维空间中,建模特征和类型之间的相似性打分,选取细粒度类型集中得分最好的类型为预测类型。文献[17-19]通过训练数据生成使用的知识库、训练数据中细粒度类型的共现特征为细粒度类型集上的所有类型设计距离计算函数,并将实体及上下文特征、细粒度类型的表示映射到同一空间,根据实体及上下文和细粒度类型的相似性打分,选取得分最高细粒度类型进行模型训练。文献[20]采用长短期记忆(long short-term memory, LSTM)神经网络模型提取实体及上下文特征,并将细粒度类型表示映射到相同空间,采用迁移学习方法,结合多个训练数据中获取的实体及上下文特征进行细粒度类型相似度打分,选取得分高的细粒度类型进行模型训练。

随着对问题研究的深入,研究人员发现细粒度实体类型间的树形结构信息可以用于训练数据降噪。文献[21]先对实体指称词进行扩展,采用双向 LSTM 神经网络学习实体上下文表示,结合正则化的层级损失函数,利用细粒度类型的树形结构关系,细粒度类型的得分为细粒度类型本身的得分与其所有祖先类型的得分加和。通过细粒度类型的得分在训练数据标注类型集中取得分最高的类型进行模型训练。文献[22]采用双向 LSTM 学习实体及上下文表示的基础上,同时用随机行走算法建模细粒度类型集的树型结构中的细粒度类型之间关系信息,形成细粒度类型在训练数据中的权重矩阵。具体来

说,先将预定义细粒度类型集通过补充虚拟节点规范化为只能将树上的叶子节点作为候选类型集;然后随机初始化细粒度类型标签的上下文相关权重;接着通过最大化随机游走过程的预期回报,对训练数据中标注的所有细粒度类型进行权重调整,选取训练数据标注类型中具有最大权重的类型作为正确类型进行监督模型学习。

最近,研究人员采用数据集中实体及上下文的聚类信息进行训练数据降噪。文献[23]采用原型网络学习数据集中细粒度类型的信息。文献[24,25]先对多类型的训练数据进行聚类,借助聚类方法从该部分训练数据中获取部分语义信息。具体地,先将带噪音标签的训练数据中的标签用指定值覆盖,接着对实体及上下文中提取的特征进行隐含语义压缩聚类。该方法在模型训练过程中既使用了带噪音标签的训练数据,又避免了训练数据中的噪音标签对模型学习的影响。文献[26]提出了基于数据集实体上下文信息相似关系图分类方法。先通过训练数据和细粒度类型的对应关系计算部分细粒度类型表示;接着在数据集上通过细粒度类型的表示和实体指称词上下文信息之间相似值,在超过阈值的实体指称词之间建立相似关系,构建相似关系图。再对训练数据中的实体及上下文表示分 2 阶段处理,第 1 阶段基于实体指称词之间的相似关系图对基于 ELmo 语音模型生成的实体指称词的表示进行特征提炼调优;第 2 阶段对提取的特征进行多标签分类。

本文参考了实体及上下文特征学习、细粒度类型间关系信息应用及数据集信息增强等方面的最新研究成果,设计了细粒度类型语义表示的学习方法,将细粒度类型的表示信息应用于细粒度实体分类任务的训练数据的标签降噪处理,并针对训练数据降噪造成的训练数据稀疏问题,设计基于实体及上下文相似度的语义聚合方法增强训练数据中的细粒度类型语义信息,进行细粒度实体分类。

2 模型介绍

本节提出一种利用细粒度类型语义标签进行标签降噪的细粒度实体分类方法,该方法由训练数据

降噪、信息聚合和细粒度实体分类 3 部分组成。

首先介绍模型中用到的各类信息的数学表示。本文将实体指称词所在的句子表示为 $s = w_1, w_2, \dots, w_n$; 实体指称词表示为 $m = \{w_l, \dots, w_r\}$, 其中 l, r 分别为实体指称词开始和结束的位置。对于数据集中的第 i 个训练数据 $instance_i$, 先用实体抽取方法识别出句子中的实体指称词 m_i , 接着通过实体链接方法将实体指称词链接到知识库中的实体 a_i , 然后分配 a_i 在知识库中的细粒度类型集 Y_i 作为训练数据 $instance_i$ 中实体指称词 m_i 的细粒度类型集。在形式上, 一个标注的数据集可以表示为三元组 $D = \{(m_i, c_i, Y_i)\}$, 其中 m_i 是数据集的第 i 个实体指称词, c_i 是 m_i 的上下文, Y_i 是 m_i 的候选类型集, 数据集上所有的细粒度类型形成细粒度类型集 T , 其中 $Y_i \subset T$ 。为了对远程监督方法标注的细粒度类型集进行更进一步分析, 将 Y_i 中细粒度类型路径 $Y_{i,j}$ 的粒度最细的类型表示为该细粒度类型路径的目标类型 $Y_{i,j}^t$, Y_i 中所有类型的目标类型构成类型集 Y_i^t 。远程监督对训练语料进行标注时, Y_i 有 2 种可能。一种是 Y_i 中的目标类型 Y_i^t 只包含细粒度类型树上的 1 条路径, 即 $|Y_i^t| = 1$ 。借

鉴现有的研究成果, 训练数据中的这类标注数据被视为干净数据, 表示为 $D_{\text{clean}} = \{(m_i, c_i, Y_i)\}$, 当且仅当 $|Y_i^t| = 1$ 。另一种是 Y_i 中的目标类型 Y_i^t 多于 1 条, 这类数据被视为噪音数据, 表示为 D_{noisy} 。为了学习实体的特征, 考虑噪音问题的细粒度实体分类任务的挑战就是训练数据中存在噪音数据情况下, 如何利用标注数据集获得性能较好的细粒度实体分类模型。

本文的模型框架如图 1 所示。从上到下依次分为训练数据降噪、类型语义信息聚合和细粒度实体分类 3 个步骤。训练数据降噪包括细粒度类型表示学习、基于实体和细粒度类型语义相似度的细粒度类型过滤。具体地, 先通过预训练 Elmo 语言模型得到实体指称词的嵌入表示和实体上下文的嵌入表示, 用训练集中干净数据 D_{clean} 中实体和细粒度类型的对应关系进行部分细粒度类型的学习。再使用 TransE 模型的关系推理方法, 通过远程监督方法使用的知识库获取细粒度类型之间关系 $relation$, 组成细粒度类型关系三元组 $(head_{\text{type}}, relation, tail_{\text{type}})$ 信息, 其中三元组的头尾元素 $head_{\text{type}} \in T, tail_{\text{type}} \in T$ 。对任务数据集预定义的细粒度类型集上缺失嵌

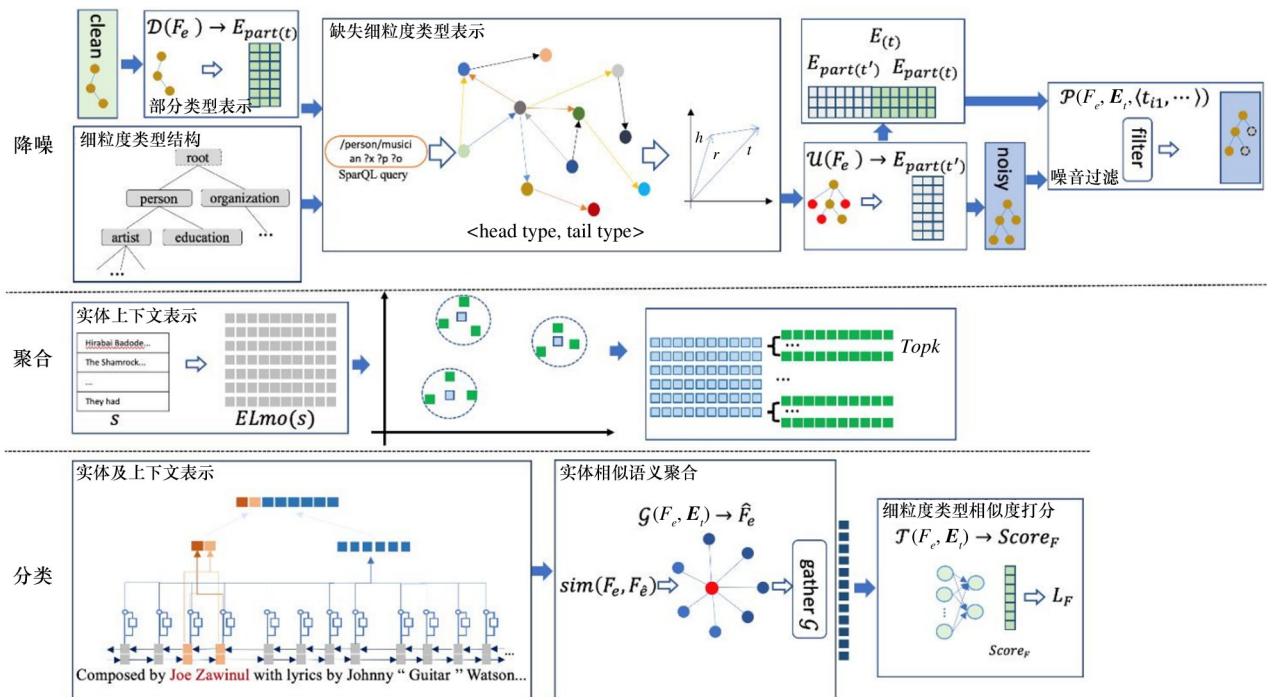


图 1 ETSF-FET 模型的整体框架

入表示的部分细粒度类型进行嵌入表示推理补全。基于远程监督生成细粒度类型中的正确类型由实体的上下文语义信息决定的观察,通过实体上下文信息和远程监督方法标注候选类型集 Y_i 中细粒度类型嵌入表示之间的相似性度量,选取与实体上下文信息最相似的细粒度类型为正确细粒度类型,降低训练数据中的噪音细粒度类型标签信息;针对现有工作在降噪中大量减少了细粒度类型训练数据的问题,本文采用了语义相似类型信息聚合方法,通过训练数据集中相似信息聚合实体指称词的细粒度类型信息,增强实体指称词特征中细粒度类型语义信息,降低训练数据减少对实体指称词特征学习的影响;最后进行细粒度实体分类,采用为噪音数据选择最符合实体上下文的细粒度类型 Y_i^s 的训练数据,结合信息聚合增强细粒度类型语义信息的实体指称词特征,训练 2 层全连接神经网络模型进行细粒度类型分类。

2.1 训练数据降噪

在细粒度实体分类任务数据集中,细粒度类型作为实体指称词的标签,和实体所在句子相比较可用于学习语义的信息少,所以本文通过数据中细粒度类型对应的实体指称词和其上下文信息学习细粒度类型表示。同时,数据集中干净数据 D_{clean} 中实体指称词和其上下文信息与细粒度类型是一一映射关系,但噪音数据 D_{noisy} 实体指称词和其上下文信息与细粒度类型存在一对多的映射关系,同时学习会引入噪音信息,为了防止这个问题本文设计不同的方法学习类型表示。

2.1.1 基于干净数据的部分细粒度类型表示学习

本文采用干净数据 D_{clean} 中实体指称词和其上下文信息与细粒度类型的对应关系计算训练数据中干净数据出现的部分细粒度类型 $t_{c, \text{part}}$ 的表示。具体而言,先采用 ELMo 语言模型进行向量化表示后生成实体所在句子的向量 $\{e_{w1}, \dots, e_{wn}\}$, 其中 $e_{wi} \in R^d$ 是句子中词的向量表示。然后通过句子中词向量的均值作为句子中实体的向量表示 e_{a_c} , 细粒度类型 t_c 的表示 e_{t_c} 通过数据集中所有的实体的向量表示求均值计算得到。计算如式(1)和(2)所示。

$$e_{a_c} = \frac{1}{n} \sum_{i=0}^n e_{wi} \quad (1)$$

$$e_{t_c} = \frac{1}{|C_t|} \sum_{t_c \in C_t} e_{a_c} \quad (2)$$

2.1.2 基于类型间关系数据的细粒度类型表示学习

本文从知识图谱的本体层面^[27]分析训练数据生成过程使用的知识库中的细粒度实体类型,发现细粒度类型之间除了树形层级结构之间的 is_a 关系,还存在其他关系。具体如图 2 所示。

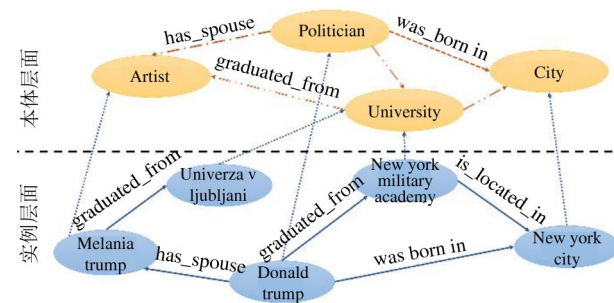


图 2 远程监督知识库中细粒度类型间关系

这个关系为细粒度实体类型表示的学习提供了数据。因此对于没有在干净数据 D_{clean} 中存在对应实体的部分细粒度类型 $t_{o, \text{part}}$, 式(1)、(2)的方法会引入噪音信息。本文采用数据集中生成过程中远程监督方法使用的知识库中细粒度类型之间关系数据 G_t , 结合上节式(1)和(2)中获得 $t_{c, \text{part}}$, 采用 TransE 模型学习部分细粒度类型集 $t_{o, \text{part}}$ 的表示, 对三元组 $(\text{head}_{\text{type}}, r, \text{tail}_{\text{type}})$, 将 r 的向量表示 E_r 作为头实体向量表示 $E_{\text{head}_{\text{type}}}$ 和尾实体向量表示 $E_{\text{tail}_{\text{type}}}$ 之间的平移。学习目标函数如式(3)所示。

$$f(\text{head}_{\text{type}}, r, \text{tail}_{\text{type}}) = - \| \text{head}_{\text{type}} + r - \text{tail}_{\text{type}} \|_2 \quad (3)$$

在学习过程中,本工作主要学习干净数据 D_{clean} 没有出现的细粒度实体类型 $t_{o, \text{part}}$ 。因此,本文调整了 TransE 方法头尾实体表示的更新策略,在细粒度类型的嵌入表示的更新步骤中,只更新 $t_{o, \text{part}}$ 的细粒度类型的嵌入表示 e_{t_o} 。将上一节学习得到的 e_{t_c} 和本节获得的 e_{t_o} 合并,得到细粒度类型集 T 中所有细粒度类型的嵌入表示 E_t 。

2.1.3 基于实体和细粒度类型语义相似度的细粒度类型过滤

已有研究工作的 D_{noisy} 处理,因没有根据训练数据中细粒度类型的语义信息与实体所在句子的语义

信息的一致性进行处理,仅选取模型预测得分最高的细粒度类型作为候选细粒度类型集的目标类型,故无法对 D_{noisy} 中的训练数据选取正确的细粒度类型进行模型训练。同时,该处理会造成训练数据多的类型,学习了更多的实体上下文信息而被较大的概率选中^[17,21],模型易受到细粒度类型标注数据分布的影响。本文通过计算实体指称词及其所在上下文的语义信息和细粒度类型的表示信息相似性,在 D_{noisy} 的标注类型集合 Y 中选择适合实体指称词及其上下文语义信息细粒度类型,作为实体指称词的目标类型进行模型训练。具体算法如算法 1 所示。先采用式(1)、(2)基于实体指称词及其文本上下文信息的向量化表示、细粒度类型间关系信息获取细粒度类型表示,进一步和采用 ELmo 语言模型学习获得实体所在句子上下文信息相似度计算,获取噪音数据 D_{noisy} 中的实体指称词 m_i 的目标类型。该算法的时间复杂度为 $O(l_n h)$, l_n 为噪音数据的大小, h 为训练数据标注细粒度类型个数最大值。

2.2 信息聚合

2.2.1 实体指称词及上下文的表示学习

细粒度实体分类任务常用的训练数据生成过程如图 3 所示。

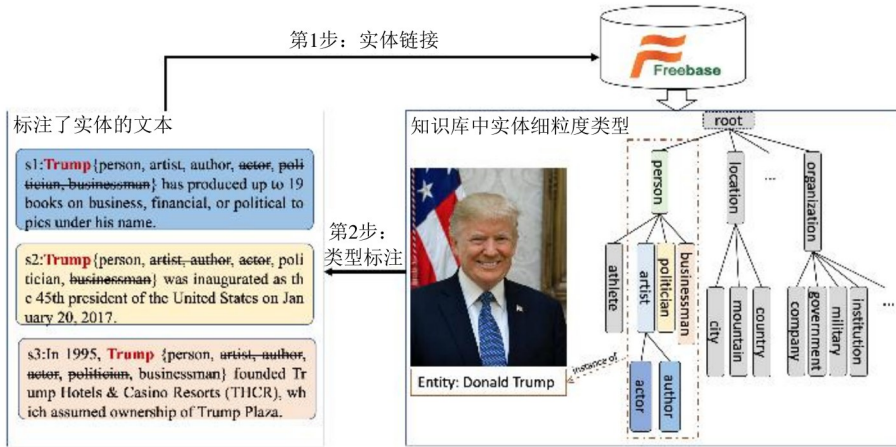


图 3 细粒度实体分类任务训练数据的生成过程

实体指称词及上下文中细粒度类型相关语义信息。本文经过实验对比了文献[28]采用 ELMo 预训练模型进行词向量表示方法和文献[21]采用双向的 LSTM 进行词向量表示方法,选用了本文方法上性能较好的双向 LSTM 进行词向量表示方法。和文献[21]

算法 1 实体指称词的目标类型获取算法

输入:训练集中的噪音数据集 D_{noisy} , 细粒度类型的嵌入表示 E_i

输出:过滤了噪音标签的训练数据 D_{filtered}

```

1  实体上下文和细粒度类型最大相似度  $topsim \leftarrow 0$ 
2  目标细粒度类型  $target \leftarrow root$ 
3  for  $m, c, Y \leftarrow D_{\text{noisy}}$  do
4       $sem_{mc} \leftarrow mean(ELMO_m + ELMO_c)$ 
5       $target \leftarrow y_{i,0}$ 
6      for  $y_{i,j} \leftarrow Y_i$  do
7           $emb_{y_{ij}} \leftarrow \text{look-up } E_i$ 
8           $sim \leftarrow \cos(sem_{mc}, emb_{y_{ij}})$ 
9          if  $sim > topsim$  then
10              $topsim \leftarrow sim$ 
11              $target \leftarrow y_{i,j}$ 
12         end if
13     end for
14      $D_{\text{filtered}} \leftarrow (m, c, target)$ 
15 end for
    
```

本文观察发现,在考虑噪音问题的细粒度实体分类任务中,实体指称词的上下文信息对实体的细粒度类型提供重要信息。基于这个观察,实体指称词及上下文的表示学习采用注意力机制,重点学习

方法不同,本文使用了全部的上下文信息,并使用注意力机制进行实体指称词表示学习。先采用 Glove 词向量对训练数据进行向量化,句子 s 向量表示为 $\{e_{w_1}, \dots, e_{w_n}\}$ 。实体指称词的表示学习由采用注意力的实体指称词表示 m_{att} 和采用实体指称词均值表

示 m_{ave} 拼接生成,即 $m = [m_{\text{att}}; m_{\text{ave}}]$, 其中 m_{att} 表示学习采用式(4)计算, 其中实体指称词的注意力权重 a_i 采用式(5)计算。

$$m_{\text{att}} = \sum_l^r a_i \text{LSTM}(e_{w_i}) \quad (4)$$

$$a_i = \text{softmax}(v \tan(W e_{w_i})) \quad (5)$$

实体指称词上下文的表示学习采用了实体指称词的上下文的词注意力表示 c_{att} 。其中 c_{att} 表示学习采用式(6)计算。

$$c_{\text{att}} = \sum_0^n \bar{a}_i \text{BiLSTM}(e_{w_i}) \quad (6)$$

其中实体指称词上下文的注意力权重 $\bar{a}_i$ 采用式(7)计算。

$$\bar{a}_i = \text{softmax}(v^c \tan(W^c e_{w_i})) \quad (7)$$

最后将两部分表示拼接形成实体指称词及上下文的嵌入表示 $F = [m; c_{\text{att}}]$ 。

2.2.2 语义相似信息聚合

对噪音数据 D_{noisy} 的处理中,在噪音训练数据的标注类型集中选取唯一的细粒度类型做为目标类型,造成标注训练数据少的部分细粒度类型训练数据严重不足问题,引起在降噪数据集上细粒度实体分类模型的性能指标差于原始数据集。针对这个问题,本文在实体指称词和上下文进行基于语义相似度的分类前,通过数据集中 $Top-K$ 个相似句子,对实体指称词的语义特征进行相似语义信息聚合,让数据集三元组中的 e_m, e_c 具有更多的类型语义信息。信息聚合方法如算法 2 所示。可以看到,经过实体指称词和其上下文信息的相似语义聚合,实体的特征表示更新为 $F'_e = [udpsem_{m,c}] \in R^{2d_w}$, 其中 d_w 表示词嵌入的维度。在该处理过程中,参数 K 作为信息聚合中参与相似信息聚合的句子数, K 的选择比较重要。根据该算法解决的问题,参数 K 的确定采用了以下方法:先统计训练数据集中细粒度类型标注数据的数量;然后根据标注数据的规模分布特征,让细粒度类型的训练数据数少的部分细粒度类型在 K 的作用下能达到细粒度类型的训练数据数量的中位数的量级。

2.3 细粒度实体分类模型

2.3.1 实体特征和细粒度类型的相似度得分

在分类处理中,以上节通过相似语义信息聚合

算法 2 实体指称词和上下文的语义聚合算法

输入:数据集 D , 信息聚合的规模参数 K , 权重参数 α , 数据集实体及上下文特征表示 F_e

输出:语义聚合后的数据集 D_{gather}

```

1  与实体上下文相似的  $K$  个实体句子聚合 entity_sim-
   topk  $\leftarrow \emptyset$ 
2  目标细粒度类型  $target \leftarrow root$ 
3  for  $id, m, c, Y_i \leftarrow D$  do
4       $sem_{mc} \leftarrow Mean(m_{\text{ELMO}} + c_{\text{ELMO}})$ 
5      entity_simtopk  $\leftarrow id$ , look-up ( $sem_{mc}, K, D$ )
6  end for
7  for  $id, m, c, Y_i \leftarrow D$  do
8       $udpsem_{mc} \leftarrow F_e[id]$ 
9       $udpsem_{mc} \leftarrow udpsem_{mc} \times (1 - \alpha) +$ 
10          $Mean(entity\_simtopk[id]) \times \alpha$ 
11   $D_{\text{gather}} \leftarrow id, udpsem_{mc}$ 
12 end for
```

形成的实体指称词及上下文的表示 F'_e 为输入,本文采用了 2 个全连接神经网络进行实体指称词特征学习,计算过程为

$$\bar{F}_e = FFNN(\bar{W}F'_e) \quad (8)$$

以标签降噪得到的实体在上下文中最合适的细粒度类型 Y_i^s 为对应的细粒度类型,通过 softmax 分类器建立映射关系。计算过程如式(9)、(10)所示:

$$\hat{p}(y | F) = \text{softmax}(W\bar{F}_e + b) \quad (9)$$

$$\hat{y} = \arg \max_y \hat{p}(y | F) \quad (10)$$

其中 $y \in Y_i^s$, W 为分类器的可训练的特征映射参数矩阵, b 为分类器的可学习的偏置参数。

2.3.2 模型优化

在训练过程中,本文采用多标签分类的交叉熵损失函数进行模型优化,并且针对细粒度实体类型的层级结构,结合文献[28]实验发现的细粒度类型集合层级树上不同级别之间细粒类型预测结果不能平行处理实验结论。本文根据细粒度类型的层级特征,设计细粒度类型层级相关的参数,让模型按照层级树从上到下的优先级学习细粒度类型的语义信息。目标函数具体为

$$L(\theta) = -\frac{1}{N} \left(\sum_{y \in Y} \beta_{\text{Lev}(y)} \log(\hat{p}(y | F'_e)) \right) + \alpha \|\theta\|^2 \quad (11)$$

其中, $\beta_{\text{Lev}(y)}$ 为细粒度类型层级相关的权重参数;

$Lev(y)$ 为细粒度类型在层级树上的层级函数,例如,图 3 中的“/person”类型, $Lev(/person) = 1$ 。图中的“/person/politician”类型, $Lev(/person/politician) = 2$, α 是损失函数 $L2$ 正则项参数。

3 实验

3.1 实验设置与评价

本文使用细粒度实体分类任务上比较通用的 3 个数据集 Wiki、OntoNotes 和 BBN 进行实验。Wiki 数据集通过 Wikipedia 远程监督生成,数据集的细粒度类型通过 Freebase 知识库中实例数超越指定阈值的细粒度类型形成。OntoNotes 细粒度实体分类数据集通过 Freebase 知识库中导出的类型,人工标注了 133 000 个新闻数据。BBN 细粒度实体分类数据集对 2 311 篇文章通过 Freebase 知识库的类型标注形成。数据集提出后出现了多种版本,为了对比实验的严谨性,本文采用文献[17]实验中使用的版本,其中数据集的统计信息如表 1 所示。

表 1 各数据集统计信息

数据集名称	Wiki	OntoNotes	BBN
类型总数	113	89	56
1 层类型数	37	4	17
2 层类型数	76	43	39
3 层类型数	0	42	0
训练集的用例总数	2 000 000	251 039	86 078
训练集干净的用例数	1 289 154	183 139	39 134
训练集带噪音的用例数	710 846	67 900	46 944
测试集的用例总数	563	8 963	12 845

为了推理补全细粒度类型集上部分类型的语义表示,本文采用文献[20]提出的方法,根据数据集生成中远程监督使用的知识库,获取了各个数据集上细粒度类型知识库中细粒度类型间的关系数据,形成细粒度类型之间的关系数据 G_r 。关系数据统计信息如表 2 所示。

本文模型的参数包括实体及上下文的分布式表示维度 d 、细粒度类型分布式表示维度 d_t 、细粒度类型间关系表示维度 d_r 、相似信息聚合权重超参 α 、

模型训练的迭代次数 p 和损失函数中细粒度类型层级权重参数 $\beta_{Lev(y)}$ 。其中实验使用的最佳参数如表 3 所示。

表 2 各数据集类型间关系统计

数据集名称	数据集类型数	获取的细粒度类型间关系数	关系的种类
Wiki	113	421 257	4 343
OntoNotes	89	374 864	3 837
BBN	56	118 533	2 943

表 3 模型使用参数

数据集名	d	d_t	d_r	α	p	$\beta_{Lev(y)}$
BBN	300	1 024	1 024	0.5	30	[2,1]
OntoNotes	300	1 024	1 024	0.6	40	[3,2,1]
Wiki	300	1 024	1 024	0.5	40	[2,1]

在模型评估方面,本文沿用了文献[4]提出的严格准确率(Acc)、宏平均 F1(Ma-F1)和微平均 F1(Mi-F1)3 个性能指标。严格准确率是测试集中实体的正确类型和预测类型完全相同的数据量在整个测试集中的占比。宏平均 F1 是测试集上每个实体的 F1 值在整个测试集上的平均值。微平均 F1 为先计算整个测试集上准确率和召回率的平均值,然后再计算 F1。对于这 3 个评价指标,结果值越大,模型细粒度实体分类的性能越好。严格准确率越大,表示测试集上预测结果和人工标注结果相同的实体越多;宏平均 F1 越大,表示测试集上每个实体预测的正确类型越多;微平均 F1 越大,表示整个测试集所有实体的预测的细粒度类型越多。

3.2 实验结果分析

3.2.1 细粒度类型的嵌入表示进行标签降噪的细粒度实体分类分析

为了分析本文提出模型的有效性,在 Wiki、OntoNotes、BBN 数据集上进行了实验,实验结果如表 4 所示。从表中可以看出,本文的方法(表中的 ETSF-FET)实验结果优于现有的研究,证明 ETSF-FET 在带有噪音标签的训练数据中,通过实体上下文语义信息和细粒度类型语义嵌入表示的相似度计算,在远程监督生成多个细粒度类型标注类型集合中选取

表4 模型性能指标

模型	Wiki			OntoNotes			BBN		
	Acc	Ma-F1	Mi-F1	Acc	Ma-F1	Mi-F1	Acc	Ma-F1	Mi-F1
AFET 2016 ^[18]	0.694	0.745	0.743	0.493	0.682	0.612	0.504	0.666	0.652
Anand 2017 ^[20]	0.658	0.812	0.774	0.522	0.685	0.633	0.604	0.741	0.757
NFETC 2018 ^[21]	0.689	0.819	0.790	0.602	0.764	0.702	—	—	—
NFETC-CLSC 2019 ^[24]	—	—	—	0.628	0.778	0.720	0.719	0.798	0.795
NDP 2019 ^[22]	0.677	0.818	0.780	0.580	0.712	0.648	0.727	0.764	0.777
Chen 2020 ^[28]	0.655	0.805	0.781	0.587	0.730	0.681	0.752	0.797	0.805
Box 2021 ^[29]	—	0.816	0.770	—	0.773	0.709	—	0.787	0.780
ETSF-FET	0.700	0.818	0.790	0.632	0.784	0.727	0.755	0.799	0.803

了最符合实体上下文语义信息的细粒度类型,降低了细粒度实体分类训练数据的噪音标签,提升了细粒度实体分类任务的性能。

3.2.2 细粒度类型嵌入表示进行训练数据标签降噪效果分析

为了分析本文提出的基于细粒度类型语义表示进行训练数据标签降噪方法的有效性,本文选用文献[21]的研究工作开源的代码。在 Wiki、OntoNotes、BBN 数据集上,将数据集中的训练数据和验证数据先采用本文

的方法进行数据标签降噪处理,然后直接运行文献[21]开源的方法。实验结果如表5所示,可以看到采用本文提出的基于细粒度类型表示进行标签降噪处理的训练数据,能让已有的研究工作提升细粒度实体分类的性能。在 BBN 数据集中,实验性能接近最近的研究工作。这证明本文所提的细粒度实体分类标签降噪方法具有通用性,其不仅在本文自己的方法上提升了分类性能,而且可通用于其他方法,并且对已有研究工作在细粒度实体分类性能上有不小的提升。

表5 所提训练数据降噪方法处理在已有开源方法上的性能实验

模型	Wiki			OntoNotes			BBN		
	Acc	Ma-F1	Mi-F1	Acc	Ma-F1	Mi-F1	Acc	Ma-F1	Mi-F1
NFETC 2018 ^[21]	0.689	0.819	0.790	0.602	0.764	0.702	—	—	—
NDP 2019 ^[22]	0.677	0.818	0.780	0.580	0.712	0.648	0.727	0.764	0.777
Chen 2020 ^[28]	0.655	0.805	0.781	0.587	0.730	0.681	0.752	0.797	0.805
NFETC 2018-FilterData	0.694	0.819	0.795	0.612	0.768	0.710	0.744	0.792	0.795

3.2.3 基于细粒度类型间的关系数据学习的类型嵌入表示进行标签降噪分析

在训练数据降噪过程中细粒度类型的表示提供了关键信息。因部分细粒度类型没有训练数据,无法通过训练数据形成细粒度类型的表示,本文结合远程监督知识库中细粒度类型间的关系数据 G_r 获得细粒度类型的表示。为了验证细粒度类型间的关系数据 G_r 为细粒度类型表示提供语义信息的有效性,本文只使用细粒度类型间的关系数据 G_r 学习细粒度类型的表示进行实验。实验的性能指标如表6

所示。从表中可以看出,在 OntoNotes 数据集上性能有明显提升,证明细粒度类型间的关系数据为训练数据降噪过程中细粒度类型的表示学习提供了有效的语义信息。同时结合数据集的细粒度类型数分析发现,OntoNotes 数据集中大于1层的类型数据占比最高,证明细粒度类型间的关系数据能为细粒度层级数上大于1层的类型提供更多的有效语义信息。

3.2.4 语义信息聚合实验分析

训练数据的降噪过程造成了细粒度类型的训练数据减少,引发训练数据少的部分细粒度类型的语

表 6 细粒度类型间关系的学习嵌入表示进行细粒度实体分类

模型	Wiki			OntoNotes			BBN		
	Acc	Ma-F1	Mi-F1	Acc	Ma-F1	Mi-F1	Acc	Ma-F1	Mi-F1
NFETC 2018 ^[21]	0.689	0.819	0.790	0.602	0.764	0.702	—	—	—
NDP 2019 ^[22]	0.677	0.818	0.780	0.580	0.712	0.648	0.727	0.764	0.777
Chen 2020 ^[28]	0.655	0.805	0.781	0.587	0.730	0.681	0.752	0.797	0.805
Box 2021 ^[29]	—	0.816	0.770	—	0.773	0.709	—	0.787	0.780
ETSF-FET-RFET	0.627	0.749	0.713	0.622	0.769	0.712	0.627	0.676	0.690

义信息学习不足的问题。本文提出通过训练数据中语义上相似的实体指称词及上下文信息,聚合数据中的细粒度类型语义信息,缓解训练数据降噪过程引起的负面影响。为了验证本文提出的类型语义信息聚合方法的有效性,在 BBN 数据集上进行了实验,实验性能指标如表 7 所示。

表 7 相似信息聚合语义学习实验性能

模型	BBN		
	Acc	Ma-F1	Mi-F1
NFETC-CLSC 2019 ^[24]	0.719	0.798	0.795
NDP 2019 ^[22]	0.727	0.764	0.777
Chen 2020 ^[28]	0.752	0.797	0.805
ETSF-FET	0.755	0.799	0.803
ETSF-FET-NoGather	0.743	0.795	0.797

从表中可以看出,本文提出的方法去掉类型语义信息聚合学习模块后的模型 ETFS-FET-NoGather,虽然比 2020 年以前的方法降噪类细粒式实体分类方法好,但是难以超过 2020 年文献[28]的实验,证明类型语义信息聚合方法能为训练数据降噪过程引起的部分细粒度类型语义学习补充有用的语义信息,再次提升细粒式实体分类的效果。

4 结 论

针对细粒度实体分类任务的训练数据降噪过程中对实体上下文和细粒度类型标签之间的语义相似性学习不足、细粒度类型的训练数据减少等问题,本文提出了一种基于细粒度类型语义表示进行标签降噪的细粒度实体分类方法。该方法先基于 ELMo 预训练模型学习数据集中干净数据对应的部分细粒度

类型的嵌入表示,并用该嵌入表示和细粒度类型间的关系数据,学习细粒度类型集中其余细粒度类型的表示;然后通过计算训练数据中标注的细粒度类型和实体指称词及上下文信息的相似性进行训练数据降噪;接着通过数据集中数据之间的相似性,选取 *Top-K* 相似数据进行细粒度类型语义信息的聚合;最后,在经过降噪和聚合处理的数据上训练细粒度实体分类模型。实验结果表明,本文提出的方法可以有效地减少细粒度实体分类任务的训练数据中的噪音标签信息,提升细粒度实体分类的性能。

缺乏高质量标注训练数据是细粒度实体分类任务一直面临的问题,最近有研究人员尝试用小样本学习技术探索解决方案并取得了一些研究成果。本文认为训练数据中的噪音数据虽然存在噪音标签,但这些噪音标签不完全是噪音,只是在实体指称词所在的上下文中不表示该部分标签表达的语义信息。找到合理的方法合理利用这些信息可以让细粒度实体分类任务应用更多的远程监督生成的训练数据,提升细粒度实体分类的性能,降低该任务训练数据标注的人力成本。同时,本文在训练数据存在噪音情况下进行了最大程度用好标注数据方面的探索,从用好训练数据的角度为细粒度实体分类任务探究解决方案,可以作为今后在细粒度实体分类任务上研究的探索方向。

参考文献

[1] LING X, WELD D. Fine-grained entity recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021,26(1):94-100.

[2] YOSEF M A, BAUER S, HOFFART J, et al. Hyena: hierarchical type classification for entity names[C]//The 24th International Conference on Computational Linguis-

- tics. Mumbai, India: ICCL, 2012:1361-1370.
- [3] WU Y, TSENG B L, SMITH J R. Ontology-based multi-classification learning for video concept detection[C] // 2004 IEEE International Conference on Multimedia and Expositions. Taipei, China: IEEE, 2004:1003-1006.
- [4] LING X, WELD D S. Fine-grained entity recognition[C] // Proceedings of the AAAI Conference on Artificial Intelligence. Toronto, Canada: AAAI, 2012:94-100.
- [5] ZHANG T, XIA C, LU C T, et al. MZET: memory augmented zero-shot fine-grained named entity typing[EB/OL]. (2020-11-02) [2022-11-11]. <https://arxiv.org/pdf/2004.01267.pdf>.
- [6] CRAVEN M, KUMLIEN J. Constructing biological knowledge bases by extracting information from text sources[C] // Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology. Stockholm, Sweden: ISMB, 1999:77-86.
- [7] GILLICK D, LAZIC N, GANCHEV K, et al. Context-dependent fine-grained entity type tagging [EB/OL]. (2016-08-01) [2022-11-11]. <https://arxiv.org/pdf/1412.1820.pdf>.
- [8] SHIMAOKA S, STENETORP P, INUI K, et al. Neural architectures for fine-grained entity type classification [EB/OL]. (2017-02-21) [2022-11-11]. <https://arxiv.org/pdf/1606.01341.pdf>.
- [9] SMIRNOVA A, CUDR' E-MAUROUX P. Relation extraction using distant supervision: a survey [J]. ACM Computing Surveys, 2018,51(5):1-35.
- [10] ANDREA R M, EGENHOFER M J. Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure[J]. International Journal of Geographical Information Science, 2004,18(3):229-256.
- [11] TARVAINEN A, VALPOLA H. Mean teachers are better role models: weight-averaged, consistency targets improve semi-supervised deep learning results[EB/OL]. (2018-04-16) [2022-11-11]. <https://arxiv.org/pdf/1703.01780.pdf>.
- [12] ONOE Y, DURRETT G. Fine-grained entity typing for domain independent entity linking[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5):8576-8583.
- [13] DU B, XINYAO T, WANG Z, et al. Robust graph-based semisupervised learning for noisy labeled data via maximum correntropy criterion[J]. IEEE Transactions on Cybernetics, 2018,49(4):1440-1453.
- [14] ALI M A, SUN Y, LI B, et al. Fine-grained named entity typing over distantly supervised data based on refined representations[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020,34(5):7391-7398.
- [15] NIU G, LI B, ZHANG Y, et al. AutoETER: automated entity type representation for knowledge graph embedding [EB/OL]. (2020-10-06) [2022-11-11]. <https://arxiv.org/pdf/2009.12030.pdf>.
- [16] XIANG Y, ZHANG Z, CHEN J, et al. OntoEA: ontology-guided entity alignment via joint knowledge graph embedding[EB/OL]. (2021-05-24) [2022-11-11]. <https://arxiv.org/pdf/2105.07688.pdf>.
- [17] REN X, HE W, QU M, et al. Label noise reduction in entity typing by heterogeneous partial-label embedding [C] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: Association for Computing Machinery, 2016:1825-1834.
- [18] REN X, HE W, QU M, et al. AFET: automatic fine-grained entity typing by hierarchical partial-label embedding[C] // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA: EMNLP, 2016:1369-1378.
- [19] REN X, WU Z, HE W, et al. Cotype: joint extraction of typed entities and relations with knowledge bases[C] // Proceedings of the 26th International Conference on World Wide Web. Perth, Australia: WWW, 2017:1015-1024.
- [20] ABHISHEK A, ANAND A, AWEKAR A. Fine-grained entity type classification by jointly learning representations and label embeddings[C] // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain: EACL, 2017:797-807.
- [21] XU P, BARBOSA D. Neural fine-grained entity type classification with hierarchy-aware loss[C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA: NAACL-HLT, 2018:16-25.
- [22] WU J, ZHANG R, MAO Y, et al. Modeling noisy hierarchical types in fine-grained entity typing: a content-

- based weighting approach [C] // Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2019:5264-5270.
- [23] 任权. 基于原型网络的细粒度实体分类方法[J]. 中文信息学报, 2020,34(12):65-72.
- [24] 陈博. 基于半监督学习的远程监督细粒度实体分类技术研究[D]. 杭州: 浙江大学, 2020.
- [25] CHEN B, GU X, HU Y, et al. Improving distantly-supervised entity typing with compact latent space clustering [EB/OL]. (2019-04-12) [2022-11-11]. <https://arxiv.org/pdf/1904.06475.pdf>.
- [26] ALI M A, SUN Y, LI B, et al. Fine-grained named entity typing over distantly supervised data based on refined representations [EB/OL]. (2020-04-07) [2022-11-11]. <https://arxiv.org/pdf/2004.03554.pdf>.
- [27] HAO J, CHEN M, YU W, et al. Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts [C] // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, USA: Association for Computing Machinery, 2019:1709-1719.
- [28] CHEN T, CHEN Y, VAN DURME B. Hierarchical entity typing via multi-level learning to rank [C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL, 2020:8465-8475.
- [29] LIU Q, LIN H, XIAO X, et al. Fine-grained entity typing via label reasoning [EB/OL]. (2021-09-13) [2022-11-11]. <https://arxiv.org/pdf/2109.05744v1.pdf>.

Fine grained entity type classification using type semantic representation for noisy label reduction

XI Pengbi^{***}, JIN Xiaolong^{***}, BAI Shuo^{***}, CHENG Xueqi^{***}

(^{*} Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

(^{**} School of Computer Science & Technology, University of Chinese Academy of Sciences, Beijing 101408)

(^{***} Hundsun Technologies Inc, Hangzhou 310053)

Abstract

The training data of fine-grained entity typing (FET) is usually generated by the distant supervision based on knowledge base, this process inevitably introduces noise type labels. The existing work mostly models the probability distribution of the training data and annotation types, and lacks the semantic learning of fine-grained types, causing the problem of the usage of types unrelated to the entity context during models learning. This paper proposes a fine-grained entity classification method for label noise reduction based on the semantic representation of fine-grained types. First, it learns the representation of some fine-grained types from the data with a unique fine-grained type path in the training set, and learns the representation of the rest fine-grained types by the combination of the relationship information between fine-grained types. Second, select the fine-grained entity type in the training data annotation fine-grained type set that is most similar to the semantic information of the entity context as target types, then, select *Top-K* similar sentences from the dataset to aggregate fine-grained semantic information. Last, it learns final fine-grained entity classification model based on the aggregated information. Experimental results and analysis on datasets demonstrate that our model effectively selects the fine-grained type that best matches the semantic information of the entity context from the fine-grained types set annotated in the training data, and is able to achieve the effect of improving the accuracy of fine-grained entity.

Key words: entity typing, fine grained type, multi-label noise reduction, multi-labels classification