

融合记忆网络的细粒度实体分类方法

周 祺,陶 皖⁺,孔 超,崔佰婷

安徽工程大学 计算机与信息学院,安徽 芜湖 241000

+ 通信作者 E-mail: taowan@ahpu.edu.cn

摘 要:细粒度实体分类是在给定实体指称后要求为其分配细粒度类型标签的任务。大多数细粒度实体分类采用远程监督的方法,为实体指称分配知识库中实体所对应的全部类型标签,这会引入无关或具体的噪声标签。在远程监督中,将分配与指称上下文无关的类型标签归为无关噪声标签,分配细粒度标签导致在上下文中实体含义不准确的类型标签归为具体噪声标签。为减轻噪声影响,以往采用人工标注、启发式规则剪枝等方法,但存在效率低、缩减训练集规模导致分类模型整体性能变差等问题。通过引入记忆网络,分类模型能深入学习实体指称上下文与类型标签之间的关联性,增强对相似的指称上下文所对应类型标签的记忆表示,有效减轻无关噪声标签的影响。与此同时,利用变形的层次损失函数有效学习类型标签之间的层次关系,从而缓解具体噪声标签的负面影响。此外,结合L2正则化函数防止训练模型对噪声标签的过拟合。在公开数据集上的实验结果表明,提出的方法能够有效缓解无关噪声标签和具体噪声标签对分类模型的消极影响,在准确率、Macro F1值、Micro F1值上表现均优于以往处理标签噪声的方法。

关键词:细粒度实体分类;噪声处理;记忆网络;类型标签

文献标志码:A **中图分类号:**TP391.1

Fine-Grained Entity Classification Method Fused with Memory Network

ZHOU Qi, TAO Wan⁺, KONG Chao, CUI Baiting

School of Computer and Information, Anhui Polytechnic University, Wuhu, Anhui 241000, China

Abstract: Fine-grained entity classification is a task that requires a fine-grained type label to be assigned to a given entity mention. Most of the existing fine-grained entity classification uses distant supervision method. All of the type labels corresponding to the entities in the knowledge base are assigned to the entity mention, which will introduce irrelevant or specific noise labels. In distant supervision, type labels that are not related to the entity mention context are classified as out-of-context noise labels, and type labels whose assignment of fine-grained labels leads to inaccurate entity meaning in the context are classified as overly-specific noise labels. In order to reduce the impact of noise, manual labeling and heuristic pruning methods have been used in the past, but there are some problems such as low efficiency and reducing the size of the training set, which leads to the deterioration of the overall performance of the classification model. By introducing the memory network, the classification model can deeply learn the correlation between the entity mention context and the type label, enhance the memory representation of the type label

基金项目:国家自然科学基金青年基金项目(61902001);安徽省教育厅高校自然科学重点项目(KJ2019A0158,KJ2019ZD15);国家级大学生创新创业项目(202010363098,201910363076)。

This work was supported by the National Natural Science Foundation for Youth of China (61902001), the Key Natural Science Project of Education Department of Anhui Province (KJ2019A0158, KJ2019ZD15), and the National Innovation and Entrepreneurship Program for College Students (202010363098,201910363076).

收稿日期:2021-03-26 **修回日期:**2021-05-14

corresponding to the similar entity mention context, and effectively reduce the influence of out-of-context noise labels. At the same time, transformative hierarchical loss function is used to effectively learn the hierarchical relationship between type labels, so as to alleviate the negative impact of overly-specific noise labels. In addition, using the L2 regularization function can prevent the model from overfitting noise labels. Experimental results on public datasets show that the proposed method can effectively alleviate the negative effects of out-of-context noise labels and overly-specific noise labels on the classification model, and its performance in accuracy, Macro F1 value and Micro F1 value is superior to previous methods for processing noise labels.

Key words: fine-grained entity classification; noise processing; memory network; type label

命名实体识别(named entity recognition, NER)是一项识别句子中具有特定意义实体的任务。NER借助人工标注的语料库为具有特定意义的实体分配类型标签,其有两个关键的任务:一是要识别出句子中是否有命名实体,二是要判断出命名实体具体所指的目标类型^[1]。作为自然语言处理(natural language processing, NLP)领域中重要的基础工作,其准确率影响许多下游任务,如信息提取、问答系统、句法分析、机器翻译等的可信度及工作效率。然而,传统的命名实体识别只是将实体类型分为几种粗粒度的实体种类,例如:人名、地点、组织、其他。在现有的大型数据库以及知识图谱中,实体类型不再局限于某几种,已达到数十种,甚至成百上千种,而且实体类别之间形成一定的层次关系。如果此时继续沿用粗粒度的实体分类会加大人工标注的难度,给后续的关系抽取、事件抽取等增加判断难度,影响许多NLP下游任务的可信度和处理质量。因此,学者们将粗粒度的命名实体识别推向细粒度实体分类(fine-grained entity typing, FET)领域。

细粒度实体分类通过给定实体指称(entity mention, 又称实体提及,即实体在具体上下文中的一段指代)及其上下文,预测实体可能的类型标签,允许实体指称有多种类型,而且类型之间通常具有一定的层次结构。细粒度的实体类型表示为其他NLP任务提供了更多的语义信息,有助于为后续的关系

抽取和事件抽取等任务增强指示性,提高问答系统、句法分析、实体推荐等下游任务的工作效率和精确率。目前,大多数的细粒度实体分类采用远程监督的方式将实体在知识库中的所有类型标签分配给实体指称,这不可避免地会引入噪声标签。以往处理噪声标签时采用人工标注、启发式剪枝噪声等方法,但人工标注效率低,无法适用于大规模的数据集;剪枝噪声使得训练集规模减小,影响分类模型的整体性能。因此,针对远程监督引入的噪声标签问题,本文将噪声标签划分为两种:无关噪声标签(out-of-context noise labels)和具体噪声标签(overly-specific noise labels),如图1所示。若为句子1分配除musician以外与上下文无关的类型标签,则为无关噪声标签。若为句子2分配任一细粒度标签都使得在句子中的实体含义不够准确,由此将其归为具体噪声标签,在此情况下应选择细粒度实体类型标签的父类型person最佳。

为减轻以上两种噪声标签对细粒度实体分类的负面影响,本文提出利用记忆网络模型增强实体指称的上下文和类型标签之间的关联性。基于语义的相关性,相似的上下文句子中实体指称对应的类型也具有相似性,因此增强对相似的指称上下文所对应类型标签的记忆表示以达到减轻无关噪声标签影响的目的。而且通过使用变形的层次损失函数,深入学习类型标签之间的层次关系,有助于缓解具体

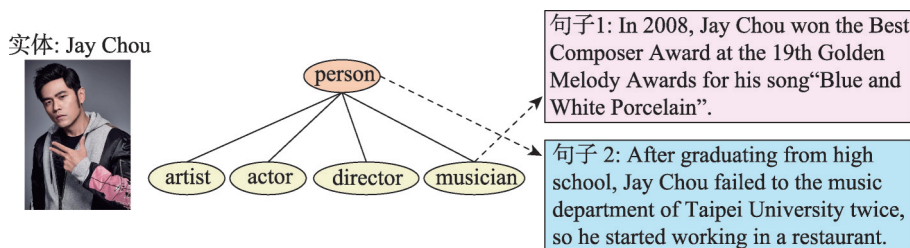


图1 两种噪声标签形式

Fig.1 Two forms of noise labels

噪声标签的消极影响。最后,通过引入L2正则化函数防止模型对噪声标签过拟合,从而提高细粒度实体分类模型的整体性能。

1 相关工作介绍

1.1 细粒度实体分类

首次处理细粒度实体类型分类任务的是Lee等人^[2],利用条件随机场模型(conditional random field, CRF)检测命名实体的边界,并使用最大熵(maximum entropy, ME)对实体进行分类,同时定义147种细粒度的实体类型。Sekine^[3]定义200种实体类型作为细粒度实体类型集的基础。Ling等人^[4]创建了细粒度实体类型数据集FIGER,将原本只具有5~6种类型的实体扩展到具有112种类型标签的实体;并且受到文献[5]启发,通过远程监督利用Wikipedia词条获得实体类型,完成了多类型、多标签的细粒度实体分类任务。Yosef等人^[6]以层次分类法划分了实体的505种类型,在不同级别的数百种类型基础上开发了一个多标签的分级分类系统HYENA(hierarchical type classification for entity names)。Yogatama等人^[7]、Dong等人^[8]使用标签嵌入以实现相关标签之间的信息共享,并且证明了细粒度的类型标签能够改进粗粒度实体分类模型的性能。Corro等人^[9]构建FINET模型,能从实体指称和实体类型中生成感知上下文的候选类型并从中选出最合适类型,此外还创建New York Times和Twitter两个细粒度的实体类型数据集。Shimaoka等人^[10]提出一种基于注意力机制的神经网络模型,通过使用长短期记忆网络(long short-term memory, LSTM)编码实体的上下文,并利用注意力机制将模型的关注点集中在实体指称的上下文表示上。马建红等人^[11]提出了一种基于多通道神经网络的实体识别模型,融合了字词特征和片段特征,利用半马尔科夫随机场完成实体边界识别和实体分类任务。盛剑等人^[12]利用网络词典回标实体,获得命名实体的领域标签,之后再根据实体的上下文确定细粒度类型标签。王红等人^[13]利用注意力机制计算知识图谱中实体与相邻实体的相关性,并且通过图卷积神经网络融合实体特征和关系特征,为实体分类提供了有效帮助。胡新棒等人^[14]利用多头注意力机制挖掘上下文信息,并将先验知识融入实体词典中,以增强模型的学习能力。西尔艾力·色提等人^[15]利用字符级和单词级注意力机制,增强字符与单词之间的关联性和互补性,有利于更好地完成中文的

细粒度命名实体识别任务。

1.2 噪声处理

现有大多数模型通过远程监督为实体从知识库中分配类型标签,由于返回的是实体在知识库中的所有类型标签,这难免会引入噪声标签。为缓解远程监督带来的噪声问题,Lawrence等人^[16]对于含噪声标签的训练数据,构造了一种核Fisher判别算法,将每个标签被翻转的概率与实例相关联,再利用期望最大化算法更新概率并确定判别式的参数。Gillick等人^[17]利用启发式规则剪枝噪声,删除与单个实体关联的同级类型、与粗粒度类型分类器输出标准不一致的类型以及出现次数少于文档中最小次数的类型,但此方法会显著减少训练集的样本量,后期将导致分类模型的整体性能变差。Ren等人^[18]针对噪声标签,创新性地训练集划分为“干净集”和“噪声集”,并采用不同的方法对以上两种训练集进行训练。参考文献[18],Abhishek等人^[19]也将训练数据分为“干净集”和“噪声集”,并提出联合学习实体指称及其上下文表示的模型。通过类型路径划分数数据集的方法取得了一定的成功,但忽略了相同路径的类型标签可能也会出错。为进一步减轻噪声标签的影响,Xu等人^[20]提出了一种基于神经网络端到端的解决方案NFETC(neural fine-grained entity type classification),将噪声标签分为与上下文无关的标签以及过于具体的标签,利用变形的交叉熵损失函数和层次损失函数来处理以上两种噪声标签。Chen等人^[21]提出压缩隐空间簇(compact latent space clustering, CLSC)算法,基于实体指称的相似度图,推导噪声标签的分布,将具有相似语义的实体指称形成紧密的簇,从而缓解噪声标签对分类模型性能的负面影响。考虑到实体和类型标签的语义相关性,Xin等人^[22]提出以无监督的方式,利用标签含义衡量指称上下文与远程监督得到类型标签之间的兼容性,减少由远程监督产生的噪声。为平等对待干净数据和含噪声的数据,Zhang等人^[23]提出一种基于概率自动重标记的方法,统一处理所有训练样本,在无额外监督的情况下解决噪声标签问题。Xia等人^[24]提出一种基于随机森林的高效方法,通过模拟网格的生成和扩展以检测分类噪声,并且引入投票机制有效处理含特征噪声的数据集。Zhang等人^[25]利用自动编码器学习具体类型特征,找出每个类中将含噪声的数据,并将其视为类的离群值,在含噪声标签的数据清理和分类任务中取得了较好的结果。Wei等人^[26]将实例特

征作为辅助信息,将噪声标签问题定义为矩阵重建问题,通过辅助信息的低维映射构建正例标签矩阵,行稀疏矩阵建模负例标签矩阵,两者相结合使得矩阵重建能力加强,有助于提高处理标签噪声时的性能。

2 融合记忆网络的细粒度实体分类模型

细粒度实体分类是一项多类型、多标签的复杂任务。针对远程监督引入的噪声标签问题,提出融合记忆网络的细粒度实体分类模型(neural fine-grained entity type classification-memory network, NFETC-MN),整体的模型架构如图2所示。具体来说,NFETC-MN模型主要分为四部分,分别是实体指称处理器、类型标签处理器、指称上下文处理器以及优化处理器。在此,预定义 T 为包含所有实体类型标签的标签集,通过给定实体指称 m 以及指称上下文 c ,以预测实体指称所对应的细粒度实体类型 t 。

2.1 实体指称处理器

根据每个句子给定的实体指称,其可能具有不同的单词数,因此将实体指称中的单词进行平均嵌入处理。给定实体指称的表示形式为 $m_i = (\omega_1, \omega_2, \dots,$

$\omega_k)$ (k 为实体指称中的单词个数),根据其表示计算实体指称中每个单词的平均词嵌入 e_m ,如式(1)所示。

$$e_m = \frac{1}{k} \sum_{i=1}^k \omega_i \quad (1)$$

2.2 类型标签处理器

为获得知识库中丰富的实体类型信息,采用实体链接将 Freebase 中的实体类型信息链接到具体的实体指称上。若实体链接能够返回实体类型标签,就将其收集存储在 Freebase 的实体类型信息中;若不能链接到相应实体,则将返回值设置为空。由于 Freebase 知识库中的类型信息与预定义的标签集 T 中的表示形式不同,运用与 Zhou 等人^[27]相似的映射方法,将收集的 Freebase 实体类型信息转换成标签集 T 中的类型表示形式,并运用独热编码将细粒度的实体类型标签表示成 e_t 的形式。此外,针对远程监督产生的噪声标签影响,除了使用简单的实体链接以外,受 Dai 等人^[28]的启发,采用共性分数以衡量实体链接的可信度。共性分数^[29]是由 Wikipedia 的锚链接计算得来,能通过给定实体指称估计各实体类型标签的正确概率分数,并在类型标签处理器中以 e_s 表

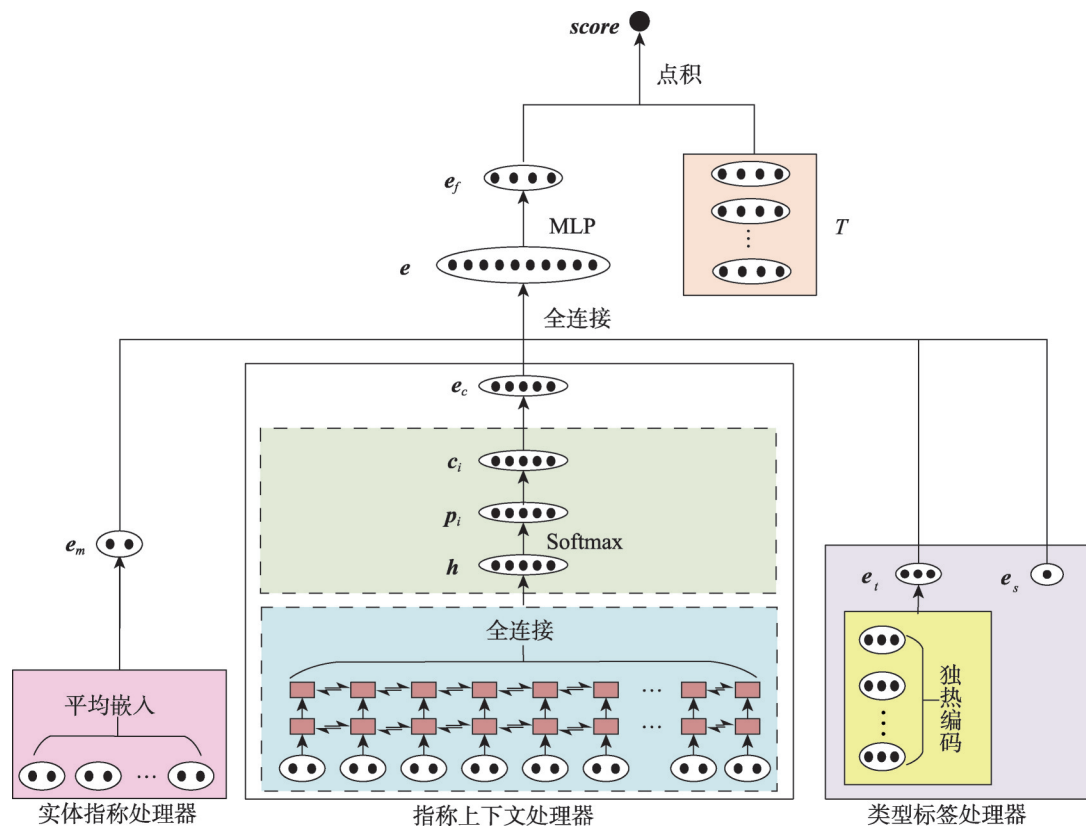


图2 融合记忆网络的细粒度实体分类模型

Fig.2 Fine-grained entity classification model fused with memory network

示共性分数。

2.3 指称上下文处理器

为从实体指称的上下文中获取更多的实体相关信息,首先将上下文向量表示为 $c_1, c_2, \dots, c_{k-1}, c_k, c_{k+1}, \dots, c_n$ (n 为句子中的单词数, c_k 作为实体指称的向量表示)。本文不仅对实体指称左右两边的上下文做词嵌入处理,还将实体指称本身进行词嵌入得到最终的上下文向量。传统的方法对上下文向量进行处理大多采用双层的 LSTM 或单层双向长短期记忆网络 (bi-directional long short-term memory, BiLSTM)。BiLSTM 网络包含两个子网络:一个称为前向 LSTM,另一个称为后向 LSTM。为提取上下文之间更多的关系信息,本文利用两层的 BiLSTM 处理上下文向量。在此,将上下文向量在第一层 BiLSTM、第二层 BiLSTM 的输出分别表示为 h_f 和 h_s ,对于两层的 BiLSTM 输出 h_f 和 h_s ,采用全连接进行计算,如式(2)所示。

$$h = h_f \oplus h_s \quad (2)$$

另外,考虑到指称上下文对选择类型标签的重要性,避免将不必要的标签引入,减轻噪声标签对细粒度实体分类模型整体性能的影响,使用端到端的记忆网络再次对指称的上下文和类型标签进行记忆学习,记忆网络的具体框架结构如图3所示。虽然单独学习实体指称上下文的信息,对分配类型标签有一定的指示作用,但未充分利用实体所对应的类型标签信息。因此,通过添加记忆网络模块,联合学习指称上下文和类型标签,有助于增强指称上下文与类型标签之间的关联性,为分配最优类型标签提供

强指示性。并且在相似的指称上下文中,记忆网络以往分配的最优类型标签也成为细粒度分类模型的记忆基础,有利于为相似实体指称分配最合适的类型标签。

首先,将指称上下文经过两层的 BiLSTM 输出后的向量 h 以及获得的类型标签 e_t 作为记忆网络的输入。在记忆网络的输入空间中,利用内积和 Softmax 函数计算 h 以及 e_t 之间的匹配度,将其最终表示为概率向量 P_i ,如式(3)所示。

$$P_i = \text{softmax}(e_i^T h); \text{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \quad (3)$$

字符级注意力机制专注于提取单词中每个英文字符的语义信息以及位置信息,过程复杂且意义不大。句子级注意力机制能在段落中提取具有更多信息的句子,并对其赋予较高权重,较适合用于段落或文档中。相较于字符级注意力机制、句子级注意力机制,单词级注意力机制能进一步捕获实体指称的上下文语义,同时突出具有指示性单词的重要性并分配其权重。因此为方便记忆网络的计算,提高其记忆性能,对实体指称上下文的两层 BiLSTM 输出 h 的最后一层隐藏层做单词级注意力机制^[30]处理。首先将 h 的 BiLSTM 最后一层隐藏层状态表示为 $\vec{h}_1, \vec{h}_1, \vec{h}_2, \vec{h}_2, \dots, \vec{h}_n, \vec{h}_n$ 以及 $\overleftarrow{h}_1, \overleftarrow{h}_1, \overleftarrow{h}_2, \overleftarrow{h}_2, \dots, \overleftarrow{h}_n, \overleftarrow{h}_n$, 利用两层前向和后向相结合的神经网络计算注意力权重,让记忆网络选择信息更丰富的单词进行训练,如式(4)~式(7)所示。

$$e_i^j = \tanh(W_e [\vec{h}_i^j; \overleftarrow{h}_i^j]) \quad (4)$$

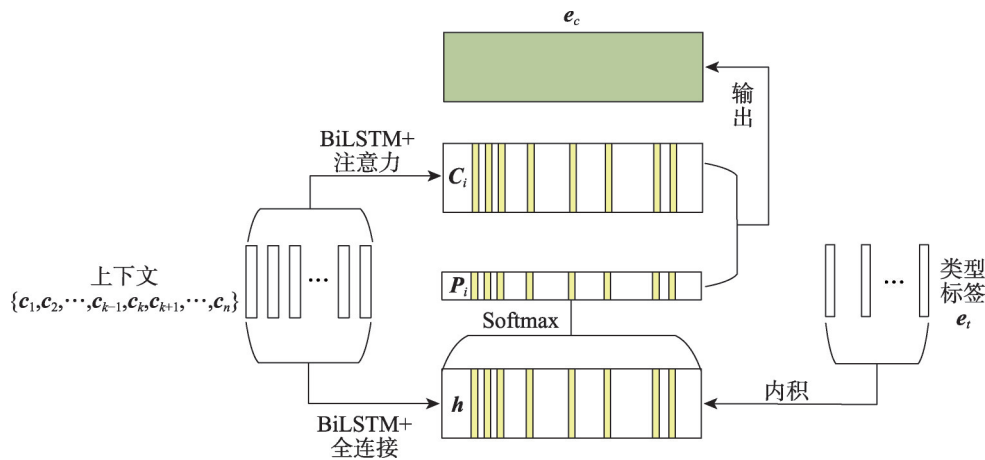


图3 记忆网络的具体框架结构

Fig.3 Concrete frame structure of memory network

$$\widehat{\mathbf{a}}_i^j = \exp(\mathbf{W}_a \mathbf{e}_i^j) \quad (5)$$

$$\mathbf{a}_i^j = \frac{\widehat{\mathbf{a}}_i^j}{\sum_{i=1}^n \widehat{\mathbf{a}}_i^l + \widehat{\mathbf{a}}_i^r} \quad (6)$$

其中, $j=l$ 或 r , $\mathbf{W}_e \in \mathbb{R}^{D_h \times 2 \times D_a}$, $\mathbf{W}_a \in \mathbb{R}^{1 \times D_a}$, D_h 为 BiLSTM 隐藏层的维度, D_a 为注意力维度。通过对上文、下文、实体指称的加权和得到另一种指称上下文的嵌入向量 $\mathbf{C}_i$, 如式(7)所示。利用指称上下文的嵌入向量 $\mathbf{C}_i$ 以及上文计算得来的概率向量 $\mathbf{P}_i$ 相乘的和作为端到端记忆网络的输出部分 $\mathbf{e}_c$, 作为指称处理器的最终表示, 如式(8)所示。

$$\mathbf{C}_i = \mathbf{a}_i^l [\overline{\mathbf{h}}_i^l; \overline{\mathbf{h}}_i^l] + \mathbf{a}_i^r [\overline{\mathbf{h}}_i^r; \overline{\mathbf{h}}_i^r] \quad (7)$$

$$\mathbf{e}_c = \sum_i \mathbf{P}_i \mathbf{C}_i \quad (8)$$

2.4 优化处理器

根据以上实体指称处理器、指称上下文处理器、类型标签处理器计算得出的向量 $\mathbf{e}_m$ 、 $\mathbf{e}_c$ 、 $\mathbf{e}_l$ 、 $\mathbf{e}_r$ 进行全连接处理, 得到的结果表示为 $\mathbf{e}$ 。为优化向量表示、提高后期分类模型的预测性能, 在多层感知器 (multi-layer perceptron, MLP) 中, 对 $\mathbf{e}$ 进行三层训练, 其中第一层为输入层, 第二层为隐层, 第三层为输出层。经过三层 MLP 的处理, 有助于训练模型提取出更具特征的向量, 并将整个句子最终表示为 $\mathbf{e}_f$ 的向量形式。这时, 将预定义标签集 T 中的标签嵌入到与 $\mathbf{e}_f$ 相同的维度空间中得到标签向量 $\mathbf{t}_i$, 之后通过对 $\mathbf{e}_f$ 和 $\mathbf{t}_i$ 进行点积处理, 得到实体指称 m 与实体类型 $\mathbf{t}_i$ 之间的预测分数 score , 如式(9)所示。

$$\text{score}(m, \mathbf{t}_i) = \mathbf{e}_f \cdot \mathbf{t}_i \quad (9)$$

由于远程监督为实体指称分配的是实体在知识库中所有的类型标签, 其中有些类型标签可能与指称上下文没有太大关系。因此, 为减少无关标签噪声的影响, 利用变形的铰链损失函数及惩罚参数 α , 使得较符合语义的类型标签损失值减小, 无关噪声标签的损失值增大。通过选择损失值较小的类型标签, 以减轻无关噪声标签的影响。针对具体噪声标签的问题, 一味地分配细粒度类型标签并不是最佳解决方案, 于是引入变形的层次损失函数, 可以更好地理解类型标签之间的层次结构 (如: 父类型-子类型)。在标签集 T 中, 所有实体的类型标签都具有一定的层次结构, 例如 Jay Chou 对应的类型标签为 /person/artist、/person/actor、/person/director、/person/musician 等。在某些语境下, 在标签集中无法找到实体

指称所对应的细粒度类型标签时, 应分配细粒度标签的父类型最为合适。此时如果坚持为其分配实体所对应的任一细粒度标签, 由于不符合语义环境, 在测试时都会判断为负例标签。因此, 在无法选择对应指称语义的细粒度类型标签时, 利用层次损失函数理解分配类型标签时的路径, 确定 person、location、organization 等父类型地位, 为实体指称分配父类型标签, 以缓解具体噪声标签的影响, 有助于提高细粒度实体分类模型性能。支持向量机 (support vector machine, SVM) 的二分类问题通常采用标准的铰链损失函数 (hinge loss function), 以实现“最大间隔 (max-margin)”分类。对于训练集中数据 x_i , 计算第 j 类得分向量 $f(x_i, \omega)_j$, y_i 表示正确标签, Δ 为阈值, 如式(10)所示。

$$L_i = \sum_{j \neq y_i} \max(0, f(x_i, \omega)_j - f(x_i, \omega)_{y_i} + \Delta) \quad (10)$$

可以看出, 标准铰链损失函数在处理实体类型标签时可能导致训练模型过拟合无关噪声标签, 造成分类模型性能降低。因此采用变形的铰链损失函数进行处理, 通过以不同的参数 α 惩罚与上下文无关的错误标签, 以减少无关噪声标签的影响, 如式(11)所示。

$$L_i = \sum_{\mathbf{t}_i \in T_{\text{true}}} \max(0, 1 - \text{score}(m, \mathbf{t}_i)) + \sum_{\mathbf{t}_i \in T_{\text{not true}}} \alpha \cdot \max(0, 1 + \text{score}(m, \mathbf{t}_i)) \quad (11)$$

其中, m 为实体指称, T_{true} 是分配给 m 的正例类型标签集, $T_{\text{not true}}$ 是分配给 m 的负例类型标签集, α 为参数, 可调整对无关噪声标签的惩罚力度。

受文献[20]启发, 使用层次损失函数有助于理解类型标签之间层次结构, 缓解因过于具体的类型标签而造成的消极影响。因此, 为处理具体噪声标签的影响, 本文利用变形的层次损失函数, 即以铰链损失函数形式预测细粒度实体类型标签父类型, 如式(12)所示。其中, $\mathbf{t}_j$ 表示细粒度实体类型标签 $\mathbf{t}_i$ 的父类型, β 为调整惩罚力度的参数。当预测实体指称的类型时出现过于具体的类型标签, 训练模型可增大 β 以达到有效减轻具体噪声标签消极影响的目的。

$$\text{score}(m, \mathbf{t}_i, \mathbf{t}_j) = \mathbf{e}_f \cdot \mathbf{t}_i + \beta \cdot \mathbf{e}_f \cdot \mathbf{t}_j \quad (12)$$

在远程监督过程中, 从 Freebase 中获得的正确实体类型信息链接到实体指称上时, 可能会造成训练模型过拟合噪声标签。于是在优化损失函数时, 在损失函数中添加 L2 正则项, 对训练模型中的参数 w_i 进行 L2 正则化约束, 可改善训练模型对噪声标签的

过拟合问题,如式(13)所示,其中 λ 是正则化参数。

$$L_{\text{final}} = \frac{1}{n} \sum_i L_i + \lambda \cdot \sum_i w_i^2 \quad (13)$$

3 实验

为验证本文方法的优越性,将该方法与其他几种方法进行了对比实验。本章首先介绍实验所用的数据集、评价指标以及对比方法,然后采用以往先进的实验方法作为基线,评估各方法在处理细粒度实体分类中噪声标签时的性能,重点展示 NFETC-MN 模型在两个数据集上的性能表现。

3.1 数据集

评估 NFETC-MN 使用到两个在细粒度实体分类任务中公开的英文数据集,即以 FIGER 和 BBN 作为实验的基础数据集,其中 FIGER 和 BBN 都是两层细粒度实体分类数据集,具体细节如表 1 所示。

表 1 数据集集中的统计数据

Table 1 Statistics in datasets

统计指标	FIGER	BBN
类型数	113	47
最深层次	2	2
mentions-训练	2 009 898	86 078
mentions-测试	563	12 845
干净 mentions-训练/%	64.46	75.92
干净 mentions-测试/%	88.28	100.00

(1)FIGER: 训练数据为 Wikipedia 中的句子,通过远程监督将 Wikipedia 的语料映射于 Freebase 中相同的类型标签后自动生成训练数据。测试数据主要是来自新闻报道里的句子,并且经过人工标注实体指称后进行验证^[4]。

(2)BBN: 数据集包括华尔街日报中 2 311 篇文章的句子,训练数据和测试数据均为人工标注^[31]。

3.2 评价指标

对于细粒度实体分类任务的评价指标,采用 Ling 等人^[4]提出的准确率(strict accuracy, Strict Acc)、宏观 F1 值(macro F1-measure, Macro F1)以及微观 F1 值(micro F1-measure, Micro F1),其中公式如下所示:

$$Acc = \frac{\sum_{m \in M} \sigma(Y = \hat{Y})}{M} \quad (14)$$

宏观 F1 值(Macro F1)是宏观精确率(macro precision, P_{ma})和宏观召回率(macro recall, R_{ma})的调和平均值。

$$P_{\text{ma}} = \frac{1}{|M|} \sum_{m \in M} \frac{|Y \cap \hat{Y}|}{\hat{Y}} \quad (15)$$

$$R_{\text{ma}} = \frac{1}{|M|} \sum_{m \in M} \frac{|Y \cap \hat{Y}|}{Y} \quad (16)$$

微观平均 F1 值(Micro F1)是微观精确率(micro precision, P_{mi})和微观召回率(micro recall, R_{mi})的调和平均值。

$$P_{\text{mi}} = \frac{\sum_{m \in M} |Y \cap \hat{Y}|}{\sum_{m \in M} \hat{Y}} \quad (17)$$

$$R_{\text{mi}} = \frac{\sum_{m \in M} |Y \cap \hat{Y}|}{\sum_{m \in M} Y} \quad (18)$$

其中, m 为实体指称, M 为其集合, Y 表示实体指称的真实类型标签, $\hat{Y}$ 表示实体指称的预测类型标签。

3.3 实验设置

实验采用 PyTorch 框架,并使用 NVIDIA 的 3070 显卡进行了加速,其他实验设置如下:

(1)预训练词向量: GloVe 词向量模型^[32]是一种广泛使用的词向量模型,以无监督的方式学习单词的向量表示,能有效地捕捉词的语义信息。采用 300 维预训练的 GloVe 词向量表示句子中单词的词向量。

(2)参数设置:模型的参数包括批处理大小 B 、Adam 优化器的学习率 l_r 、BiLSTM 中隐藏层的维数 D_h 、MLP 中隐藏层的维数 D_m 、类型标签嵌入维数 D_t 、惩罚无关噪声标签的参数 α 、惩罚具体噪声标签的参数 β 以及 L2 正则化参数 λ 。对于两个数据集的具体参数设置如表 2 所示。

表 2 超参数设置

Table 2 Hyperparameter setting

超参数	FIGER	BBN
B	256	256
l_r	0.002	0.007
D_h	250	250
D_m	500	500
D_t	500	500
α	1	1
β	0.4	0.4
λ	0.000 1	0.000 2

3.4 对比实验

本节将本文提出的 NFETC-MN 模型与其他先进

表3 细粒度实体分类方法性能比较

Table 3 Performance comparison of fine-grained entity classification methods

单位: %

方法	FIGER			BBN		
	Strict Acc	Macro F1	Micro F1	Strict Acc	Macro F1	Micro F1
Attentive	59.7	80.0	75.4	48.4	73.2	72.4
AFET	53.3	69.3	66.4	68.3	74.4	74.7
AAA	65.8	81.2	77.4	65.5	73.6	75.2
NFETC _{hier}	68.9 ± 0.6	81.9 ± 0.7	79.0 ± 0.7	73.9 ± 1.2	78.8 ± 1.2	79.4 ± 1.1
NFETC-CLSC _{hier}	—	—	—	73.0 ± 0.3	79.8 ± 0.4	79.5 ± 0.3
NFETC-AR _{hier}	70.1 ± 0.9	83.2 ± 0.7	80.1 ± 0.6	74.9 ± 0.6	80.4 ± 0.6	80.3 ± 0.6
NFETC-MN _{hier}	72.2 ± 1.2	83.7 ± 0.6	81.3 ± 0.7	76.8 ± 0.2	86.2 ± 0.2	86.8 ± 0.2

的细粒度实体分类模型进行了对比。

(1) Attentive^[10]: Shimaoka 等人提出基于特征提取的注意力网络模型,但未将训练集中的干净数据与含噪声数据进行区分。

(2) AFET^[18]: 针对细粒度实体分类的噪声标签,根据类型路径将训练数据划分为“干净集”和“噪声集”,再利用部分标签损失对标签嵌入进行训练。

(3) AAA^[19]: Abhishek 等人也将训练数据分为“干净集”和“噪声集”,并提出联合学习实体指称及其上下文表示的神经网络模型。

(4) NFETC_{hier}^[20]: 使用变形的交叉熵损失函数和层次损失函数来处理无关噪声标签以及具体噪声标签。

(5) NFETC-CLSC_{hier}^[21]: 通过标签传播和候选类型约束以推断噪声标签的类型分布,使用压缩隐空间规范远程监督分配的噪声标签,并利用层次损失函数处理具体噪声标签。

(6) NFETC-AR_{hier}^[23]: 基于概率自动重标记数据,能平等对待干净和含噪声样本。在没有额外监督的情况下解决噪声标签问题,并利用层次损失函数处理具体噪声标签。

3.5 实验结果及分析

在 FIGER 和 BBN 两个数据集上,细粒度实体分类任务中各方法的准确率、Macro F1 值以及 Micro F1 值如表 3 所示。通过对比实验结果发现,NFETC-MN 在准确率、Macro F1 值以及 Micro F1 值上均优于以往的细粒度实体分类方法。

NFETC-MN 模型首先利用记忆网络模型,联合学习指称上下文及类型标签,构成记忆表示,为分配符合语义的类型标签增加指示性信息。其中,通过对指称上下文进行两种不同方式的处理以及计算指称上下文向量与类型标签的匹配度,能够进一步为上下文中更具指示性的单词分配更高权重,充分提取指称上下文与类型标签的重要信息。其次,为了

缓解无关噪声标签以及具体噪声标签的消极影响,分别设计变形的铰链损失函数和变形的层次损失函数。通过调整惩罚参数 α 、 β ,增大无关噪声标签以及具体噪声标签的损失值,能够有效减轻噪声标签对细粒度分类模型整体性能的负面影响。

具体来说,融合记忆网络的 NFETC-MN 在 FIGER 数据集上相较于 NFETC-AR_{hier},准确率提高 2 个百分点左右,Macro F1 值无太大差异,Micro F1 值提高 1 个百分点左右。在 BBN 数据集上,相较于 NFETC-AR_{hier},NFETC-MN 的准确率提高 2 个百分点左右,Macro F1 值提高 6 个百分点左右,Micro F1 值提高 6 个百分点左右。综合看来,融合记忆网络的 NFETC-MN 在 BBN 数据集上的表现优于在 FIGER 数据集上的表现,尤其是在 BBN 上的 Macro F1 值和 Micro F1 值,相较于 NFETC-AR_{hier} 增长约 6 个百分点。原因可能有两点:(1)BBN 数据集中实体类型数量少于 FIGER;(2)BBN 是单类型路径的数据集,FIGER 数据集允许实体有多个类型路径,因此细粒度实体分类任务与 FIGER 相比较为简单。

正如表 3 所示,通过增加记忆网络使得细粒度实体分类模型能够更好地学习实体指称上下文以及类型标签表示,并且利用深度学习的 MLP 模型提取到更有效的特征,从而提高 NFETC-MN 模型在准确率、Macro F1 值以及 Micro F1 值上的性能表现。与此同时,采用变形的铰链损失函数以及变形的层次损失函数分别处理无关噪声标签、具体噪声标签,可以改善 NFETC-MN 模型的分类性能。另外,实验结果表明 L2 正则化函数的引入能有效避免训练模型对噪声标签过拟合。

4 结论

本文针对远程监督引入的噪声标签问题,提出

了一种融合记忆网络的实体分类方法(NFETC-MN)。通过引入记忆网络增强对实体指称上下文和类型标签之间的联合学习,建立指称上下文与类型标签之间的关联性。同时使用变形的铰链损失函数和变形的层次损失函数以缓解无关噪声标签和具体噪声标签的消极影响。此外,结合L2正则化函数防止训练模型对噪声标签过拟合。实验结果表明,引入记忆网络模型、变形的层次损失函数以及L2正则化函数,有助于处理细粒度实体分类噪声标签任务。并且该模型在准确率、Macro F1值以及Micro F1值三个评价指标上,均优于以往处理噪声标签方法。

参考文献:

- [1] 冯建周, 马祥聪. 基于迁移学习的细粒度实体分类方法的研究[J]. 自动化学报, 2020, 46(8): 1759-1766.
FENG J Z, MA X C. Research on fine-grained entity classification method based on transfer learning[J]. Acta Automatica Sinica, 2020, 46(8): 1759-1766.
- [2] LEE C, HWANG Y G, OH H J, et al. Fine-grained named entity recognition using conditional random fields for question answering[C]//LNCS 4182: Proceedings of the 3rd Asia Information Retrieval Symposium Information Retrieval Technology, Singapore, Oct 16-18, 2006. Berlin, Heidelberg: Springer, 2006: 581-587.
- [3] SEKINE S. Extended named entity ontology with attribute information[C]//Proceedings of the 2008 International Conference on Language Resources and Evaluation, Marrakech, May 26-Jun 1, 2008: 1-6.
- [4] LING X, WELD D S. Fine-grained entity recognition[C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Jul 22-26, 2012. Menlo Park: AAAI, 2012: 94-100.
- [5] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]//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, Singapore, Aug 2-7, 2009. Stroudsburg: ACL, 2009: 1003-1011.
- [6] YOSEF M A, BAUER S, HOFFART J, et al. HYENA: hierarchical type classification for entity names[C]//Proceedings of the 24th International Conference on Computational Linguistics, Mumbai, Dec 8-15, 2012. India: Indian Institute of Technology Bombay, 2012: 1361-1370.
- [7] YOGATAMA D, GILLICK D, LAZIC N. Embedding methods for fine grained entity type classification[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, Jul 26-31, 2015. Stroudsburg: ACL, 2015: 291-296.
- [8] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, Aug 24-27, 2014. New York: ACM, 2014: 601-610.
- [9] DEL CORRO L, ABUJABAL A, GEMULLA R, et al. FINET: context-aware fine-grained named entity typing[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Sep 17-21, 2015. Stroudsburg: ACL, 2015: 868-878.
- [10] SHIMAOKA S, STENETORP P, INUI K, et al. An attentive neural architecture for fine-grained entity type classification[C]//Proceedings of the 5th Workshop on Automated Knowledge Base Construction, San Diego, Jun 17, 2016. Stroudsburg: ACL, 2016: 69-74.
- [11] 马建红, 张炳斐, 张少光, 等. 基于主动MCNN-SCRF的新能源汽车命名实体识别[J]. 计算机工程与应用, 2019, 55(7): 23-29.
MA J H, ZHANG B F, ZHANG S G, et al. New energy vehicle named entity recognition based on active MCNN-SCRF[J]. Computer Engineering and Applications, 2019, 55(7): 23-29.
- [12] 盛剑, 向政鹏, 秦兵, 等. 多场景文本的细粒度命名实体识别[J]. 中文信息学报, 2019, 33(6): 80-87.
SHENG J, XIANG Z P, QIN B, et al. Fine-grained named entity recognition for multi-scene text[J]. Journal of Chinese Information Processing, 2019, 33(6): 80-87.
- [13] 王红, 林海舟, 卢林燕. 基于Att_GCN模型的知识图谱推理算法[J]. 计算机工程与应用, 2020, 56(9): 183-189.
WANG H, LIN H Z, LU L Y. Knowledge graph inference algorithm based on Att_GCN model[J]. Computer Engineering and Applications, 2020, 56(9): 183-189.
- [14] 胡新棒, 于淑乔, 李邵梅, 等. 基于知识增强的中文命名实体识别[J]. 计算机工程, 2021, 47(11): 84-92.
HU X B, YU G Q, LI S M, et al. Chinese named entity recognition based on knowledge enhancement[J]. Computer Engineering, 2021, 47(11): 84-92.
- [15] 西尔艾力·色提, 艾山·吾买尔, 王路路, 等. 结合单词-字符引导注意力网络的中文旅游文本命名实体识别[J]. 计算机工程, 2021, 47(2): 39-45.
XIERAILI S, AISHAN W, WANG L L, et al. Named entity recognition in Chinese tourism text based on wordcharacter guided attention network[J]. Computer Engineering, 2021, 47(2): 39-45.
- [16] LAWRENCE N D, SCHÖLKOPF B. Estimating a kernel Fisher discriminant in the presence of label noise[C]//Proceedings of the 18th International Conference on Machine Learning, Williamstown, Jun 28-Jul 1, 2001. San Francisco: Morgan Kaufmann, 2001: 306-313.
- [17] GILLICK D, LAZIC N, GANCHEV K, et al. Context-dependent fine-grained entity type tagging[J]. arXiv:1412.1820, 2014.

- [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, Nov 1-4, 2016. Stroudsburg: ACL, 2016: 1369-1378.
- [19] 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, Apr 3-7, 2017. Stroudsburg: ACL, 2017: 797-807.
- [20] 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, Jun 1-6, 2018. Stroudsburg: ACL, 2018: 16-25.
- [21] CHEN B, GU X, HU Y, et al. Improving distantly-supervised entity typing with compact latent space clustering[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Jun 2-7, 2019. Stroudsburg: ACL, 2019: 2862-2872.
- [22] XIN J, ZHU H, HAN X, et al. Put it back: entity typing with language model enhancement[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Oct 31-Nov 4, 2018. Stroudsburg: ACL, 2018: 993-998.
- [23] ZHANG H, LONG D, XU G, et al. Learning with noise: improving distantly-supervised fine-grained entity typing via automatic relabeling[C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence, Yokohama, Jul 2020: 3808-3815.
- [24] XIA S, WANG G, CHEN Z, et al. Complete random forest based class noise filtering learning for improving the generalizability of classifiers[J]. IEEE Computer Architecture Letters, 2019, 31(11): 2063-2078.
- [25] ZHANG W, WANG D, TAN X. Robust class-specific auto-encoder for data cleaning and classification in the presence of label noise[J]. Neural Processing Letters, 2019, 50(2): 1845-1860.
- [26] WEI Y, GONG C, CHEN S, et al. Harnessing side information for classification under label noise[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(9): 3178-3192.
- [27] ZHOU B, KHASHABI D, TSAI C T, et al. Zero-shot open entity typing as type-compatible grounding[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Oct 31-Nov 4, 2018. Stroudsburg: ACL, 2018: 2065-2076.
- [28] DAI H, DU D, LI X, et al. Improving fine-grained entity typing with entity linking[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, Nov 3-7, 2019. Stroudsburg: ACL, 2019: 6209-6214.
- [29] PAN X, CASSIDY T, HERMJAKOB U, et al. Unsupervised entity linking with abstract meaning representation [C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, May 31-Jun 5, 2015. Stroudsburg: ACL, 2015: 1130-1139.
- [30] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Aug 7-12, 2016. Stroudsburg: ACL, 2016: 207-212.
- [31] WEISCHEDEL R, BRUNSTEIN A. BBN pronoun coreference and entity type corpus[EB/OL]. [2021-01-06]. <https://doi.org/10.35111/9fx9-gz10>.
- [32] PENNINGTON J, SOCHER R, MANNING C D. GloVe: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Oct 25-29, 2014. Stroudsburg: ACL, 2014: 1532-1543.



周祺(1997—),女,安徽宿州人,硕士研究生,主要研究方向为云计算、大数据处理。

ZHOU Qi, born in 1997, M.S. candidate. Her research interests include cloud computing and big data processing.



陶皖(1972—),女,安徽芜湖人,硕士,教授,主要研究方向为大数据、数据分析。

TAO Wan, born in 1972, M.S., professor. Her research interests include big data and data analysis.



孔超(1986—),男,山东人,博士,副教授,主要研究方向为网络数据管理、流媒体数据处理、社交网络分析、数据挖掘。

KONG Chao, born in 1986, Ph.D., associate professor. His research interests include web data management, streaming data processing, social network analysis and data mining.



崔佰婷(1999—),女,安徽宿州人,主要研究方向为图嵌入、图挖掘。

CUI Baiting, born in 1999. Her research interests include graph embedding and graph mining.