软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.007147]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于元属性学习的事件检测*

贺瑞芳[1,2], 马劲松[1,2], 黄孝家[1,2], 张仕奇[3], 白 洁[3]

[1](天津大学 智能与计算学部, 天津 300350)

[2](天津市认知计算与应用重点实验室, 天津 300350)

[3](中国电子科技集团公司第五十四研究所, 河北 石家庄 050081)

通信作者: 贺瑞芳, E-mail: rfhe@tju.edu.cn

摘 要: 事件检测旨在识别非结构化文本中的事件触发词, 并将其分类为预定义的事件类别, 可用于知识图谱构建及舆情监控等. 然而, 其中的数据稀疏和不平衡问题严重影响了事件检测系统的性能和可用性. 现有大多数方法没有很好地解决这一问题, 这源于其将不同类别的事件独立看待, 并通过分类器或空间距离对触发词进行识别和分类. 尽管有研究考虑事件大类下子类的事件元素存在关联性, 采用多任务学习进行互增强, 但忽略了不同类别事件触发词之间的共享属性. 已有相关建模事件类别关系的工作需要大量的规则设计和数据标注, 导致作用域局限, 泛化性不强. 因此, 提出一种基于元属性的事件检测方法. 其旨在学习不同类别样本中包含的共享内在信息, 包括: (1) 构造触发词的特殊符号表示并通过表示向量的映射来提取触发词的类别无关语义; (2) 拼接触发词表示, 类别的样本语义表示和类别的标签语义表示, 输入一个可训练的相似度度量层, 从而建模关于触发词和事件类别的公用相似度度量. 通过学习以上两种信息以缓解数据稀疏和不平衡的影响. 此外, 将样本的类别无关语义集成到分类方法中, 并构建完整的融合模型. 在 ACE2005 和 MAVEN 数据集上通过不同程度稀疏和不平衡情景下的实验证明所提出方法的有效性, 并建立传统和少样本设置之间的联系.

关键词: 事件检测; 类别无关语义; 度量学习; 少样本学习

中图法分类号: TP18

中文引用格式: 贺瑞芳, 马劲松, 黄孝家, 张仕奇, 白洁. 基于元属性学习的事件检测. 软件学报. http://www.jos.org.cn/1000-9825/7147.htm

英文引用格式: He RF, Ma JS, Huang XJ, Zhang SQ, Bai J. Event Detection Based on Meta-attribute Learning. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7147.htm

Event Detection Based on Meta-attribute Learning

HE Rui-Fang[1,2], MA Jin-Song[1,2], HUANG Xiao-Jia[1,2], ZHANG Shi-Qi[3], BAI Jie[3]

[1](College of Intelligence and Computing, Tianjin University, Tianjin 300350, China)

[2](Tianjin Key Laboratory of Cognitive Computing and Applications, Tianjin 300350, China)

[3](The 54th Research Institute of CETC, Shijiazhuang 050081, China)

Abstract: Event detection (ED) aims to detect event triggers in unstructured text and classify them into pre-defined event types, which can be applied to knowledge graph construction, public opinion monitoring, and so on. However, the data sparsity and imbalance severely impair the system's performance and usability. Most existing methods cannot well address these issues. This is due to that during detection, they regard events of different types as independent and identify or classify them through classifiers or space-distance similarity. Some work considers the correlation between event elements under a broader category and employs multi-task learning for mutual enhancement; they overlook the shared properties of triggers with different event types. Research related to modeling event connections requires designing lots of rules and data annotation, which leads to limited applicability and weak generalizability. Therefore, this study

* 基金项目: 国家自然科学基金 (62376192, 61976154)
收稿时间: 2023-06-06; 修改时间: 2023-08-28; 采用时间: 2023-12-19; jos 在线出版时间: 2024-09-11

proposes an event-detection method based on meta-attributes. It aims to learn the shared intrinsic information contained in samples across different event types, including (1) extracting type-agnostic semantics of triggers through semantic mapping from the representations of special symbols; (2) concatenating the semantic representations of triggers and samples in each event type as well as the label embedding, inputting them into a trainable similarity measurement layer, thereby modeling a public similarity metric related to triggers and event categories. By combining these representations into a measuring layer, the proposed method mitigates the effects of data sparsity and imbalance. Additionally, the full fusion model is constructed by integrating the type-agnostic semantic into the classification method. Experiments on ACE2005 and MAVEN datasets under various degrees of sparsity and imbalance, verify the effectiveness of the proposed method and build the connection between conventional and few-shot settings.

Key words: event detection; type-agnostic semantics; metric learning; few-shot learning

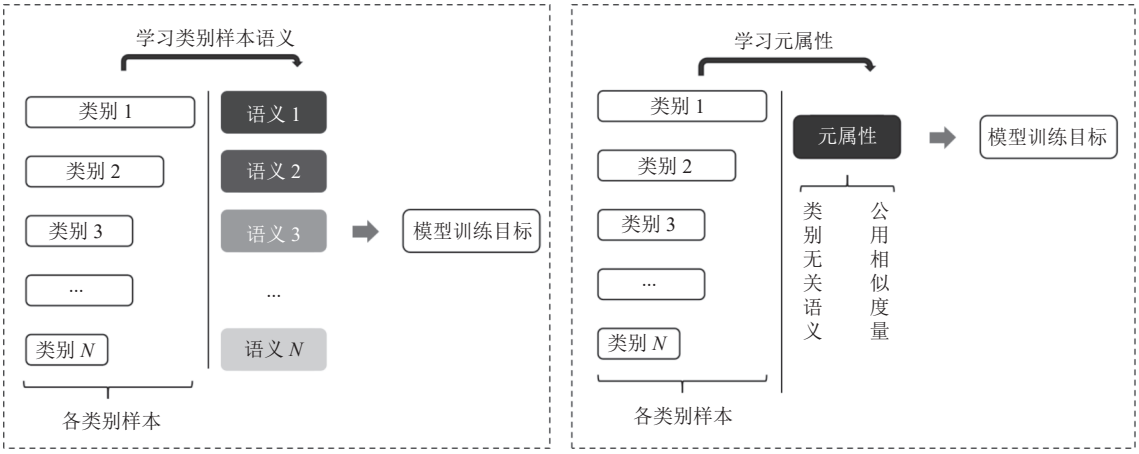
事件检测 (event detection, ED) 旨在识别非结构化文本中的触发词 (即标记特定事件发生的单词), 并将其分类为预定义的事件类别, 其有助于诸多应用, 例如知识图谱构建[1]和对话系统中的意图检测[2]等.

由于文本表示的发展[3,4]和丰富知识库 (如词汇[5]和常识知识[6]) 的引入, 事件检测取得了优异的性能[7–10]. 此外, 为了适应新的事件类别, 一些研究还提出了基于原型和片段训练的少样本方法[11–14], 并取得了较大的进步. 但由于固有的数据稀疏和不平衡, 在相对较大的类别集中有效地检测训练样本稀缺的事件类别仍是一项艰巨的任务. 以 ACE2005 数据集为例, 其标注的触发词所占比例不到整个语料的 2% (5649/301229), 其中, 一些事件类别的比例甚至更低, 例如, “袭击 (attack)”类别事件的样本数量为 1 629, 而“无罪释放 (release-parole)”“引渡 (extradite)”和“赦免 (pardon)”这 3 种事件总计只有 16 个样本, 仅为前者的 1%. 在实际场景中, 很难同时获得足够的训练样本和均衡的类别分布, 这将导致某些事件类别很难被正确识别和分类.

现有的大多数方法都不能很好地解决这个问题. 这源于其本质上将不同类别的事件独立看待, 对模型训练能够起到作用的只是每个事件类别各自的特征: 例如, 固定类别集的事件检测方法使用带标注的样本训练分类器对每种类别样本特征的“记忆”[7,8,15], 基于原型的少样本事件检测方法学习每种类别及其所包含样本的表示, 并通过空间距离相似度[11,16]对测试样本进行识别和分类. 分类器的“记忆”或类别表示与判别的性能主要受样本数量和比例的影响: 样本太少无法提供足够的信息, 而不平衡的样本可能会导致表示和分类偏离或偏向某些类别. 文献 [17] 考虑同一事件大类下的事件子类, 其事件元素存在高度的相互关联性, 采用多任务学习方法进行互增强的联合学习以缓解数据稀疏, 但忽视了事件类别之间的关联. 尽管一些研究对事件类别之间的关系进行了建模[9,13,18,19], 但此类方法要求大量额外的规则设计和数据标注, 导致作用域局限, 泛化性不强.

相比之下, 所有类别的样本都拥有一些共同的与类别特性无关的属性可供模型学习. 如果一个事件类别只有几个样本, 它仍然可以从其他类别样本提供的信息中获益. 本文称这种属性为元属性 (meta-properties). 考虑触发词的两个属性: (1) 每个触发词都可以抽象为一种类别, 即“事件”, 无论它具体是什么类别的事件; (2) 每个触发词与相同类别的触发词更相似, 而不是其他类别的触发词. 这些属性由不同类别的样本共享, 不受不平衡分布的影响, 其在模型训练时能够对梯度更新等过程起到同等贡献的作用. 为了更好地理解元属性, 我们将元属性模型训练与传统方法对比如图 1 所示, 对于相同的模型训练目标, 相比于独立学习每个类别的特征 (图 1(a)), 学习元属性 (图 1(b)) 受数据分布的影响更小, 这有助于缓解数据稀疏和不平衡.

因此, 本文提出一种缓解训练数据标注稀疏和类别分布不均衡的元属性学习事件检测方法, 尝试以神经网络层参数的形式对触发词的上述两个元属性进行建模. 对于属性 (1), 该模型将每个触发词替换为一个特殊符号 (即 [trigger]), 保留其上下文, 并通过类别无关投影层的多层感知机 (multi-layer perception, MLP) 网络使触发词和特殊符号的表示尽可能相似. 这种学到的表示可以看作是“触发词”的类别无关语义, 而不是具体类别的具象特征, 这种语义能够更容易地与非事件词的语义进行区分. 对于属性 (2), 该模型在事件类别和输入样本之间建立了一个可学习的度量模型. 对于每个事件类别, 模型获取其所有样本的表示计算均值作为类别的样本语义表示, 并拼接其标签 (即事件类别的名称) 的语义表示作为补充. 标签表示的语义信息对于样本非常稀少以至于无法准确表示其语义的类别至关重要. 对于要预测类别的输入样本, 模型将其表示与上述类别的样本表示和标签表示拼接, 并通过度量网络, 将组合转换为相似度分数从而进行分类. 最后, 为了解决“非事件”缺乏明确的类别语义的问题, 本文将样本的类别无关语义集成到了分类方法中, 并构建了完整的融合模型.



(a) 训练模型时独立建模类别样本语义 (b) 通过学习元属性训练模型

图 1 与传统模型训练对比的元属性学习示意图 (其中方块颜色深浅代表语义学习质量)

综上所述, 主要贡献包括: 1) 提出基于元属性的事件检测方法, 包括通过学习类别无关语义进行触发词识别以缓解数据稀疏, 以及通过学习样本和每种事件类别的相似度度量来进行触发词分类以缓解数据分布不平衡; 2) 在 ACE2005 和 MAVEN 数据集上的实验结果验证了本文方法的有效性以及样本数量和比例的影响; 3) 通过控制样本的数量和比例来建立传统设置和少样本设置之间的联系.

本文第 1 节介绍事件检测的相关方法和研究现状. 第 2 节介阐述基于元属性学习的事件检测模型. 第 3 节对所提出方法的总体性能和各部分性能进行了验证. 第 4 节通过构建不同的稀疏和类别不平衡程度场景验证了所提模型的有效性. 最后总结全文.

1 相关工作

1.1 一般性事件检测

作为自然语言处理的一个重要子任务, 已有的传统事件检测方法大致可以分为两大类, 即基于特征和基于表示的方法. 基于特征的方法[20–22]依赖手动设计的判别特征来建立统计模型. 研究者利用多种策略将有助于挖掘分类的特征和线索, 例如将单词的词性和语法依赖关系等, 转换为特征向量, 用传统的统计机器学习方法检测事件触发词及其事件类别. 但此类方法需要人工设计繁琐复杂的特征, 且缺少泛化能力.

随着深度学习的发展, 现有基于表示的事件检测方法表现出了优越的性能. 具体包括: (1) 改善文本表示的方法, 利用单词嵌入和各种神经网络结构, 如卷积神经网络 (convolutional neural network, CNN)[23], 循环神经网络 (recurrent neural network, RNN)[24–26], 图卷积网络 (graph convolutional network, GCN)[19,27–30], 生成对抗网络 (generative adversarial network, GAN)[31,32]和预训练语言模型 (pre-trained language models, PLMs)[33–35]等来获得丰富的表示. Wang 等人[8]提出了对比预训练方法使模型充分学习文本的语义表示和语法结构, 从而提高事件触发词和论元抽取性能. (2) 数据增强的方法, 采用半监督等方法引入额外的训练数据以提高模型的性能和鲁棒性. Wang 等人采用了教师-学生架构来过滤半监督数据中的噪声[15]. Pouran 等人[36]利用 GPT-2 自动生成的训练数据来提高模型的性能. (3) 建模知识的方法, 引入外部知识以丰富可用的信息, 例如远程监督[7]利用已有的知识库来对齐文本, 自动标注数据用于模型的训练, 知识蒸馏框架[32,37]训练一个知识感知的教师 (teacher) 模型来指导训练轻量化的学生 (student) 模型等. Lu 等人[38]提出一种 Delta 表示方法来建模触发词的判别知识和泛化知识. Liu 等人[39]提出显式地建模并识别触发词依赖的类别和上下文依赖的有歧义的事件类别, 从而充分利用文本自身提供的线索. Wang 等人[40]将事件类别作为自然语言查询, 通过注意力机制来检测触发词和论元. 王捷等人[35]在建模触发词表示时利用门控机制融入基于图神经网络的依存信息和基于自注意力的上下文信息. 陈佳丽等人[41]构建了基于

BERT 事件检测和实体识别的多任务方法来提高触发词检测的性能. 上述方法性能强大, 但仅限于固定域, 并且依赖于固定的事件类别集合以及大量的训练样本.

1.2 少样本事件检测

近年来, 少样本事件检测 (few-shot event detection, FSED) 被提出并引起了关注. 少样本事件检测的核心思想是为新的事件类别引入一些示例, 用来获取事件类别的原型 (prototype)[42]. 样本的事件类别可以根据其表示和原型之间的相似度来确定. 早期工作致力于改善原型的表示, 例如 Deng 等人[9]定义了少样本事件检测任务, 并利用动态记忆网络来学习事件类别原型的语义表示. Lai 等人[12]提出了簇内匹配和簇间信息, 为少样本事件检测提供更多的训练信号. 之后, 研究人员探索各种信息, 如事件类别关系[13]和词汇知识[14], 以增强原型的表示. Huang 等人[43]利用半监督学习建模事件类别的向量化表示用于事件检测和自动类别归纳. 后续一些工作着重提高少样本方法在不同任务之间的泛化性能, 例如 Lai 等人[16]提出对跨任务信息进行建模, 以解决采样偏差和离群点问题. Lai 等人[44]使用迁移学习和表示正则化方法解决了事件检测的少样本学习模型的不良泛化问题, 将开放域词义消歧知识迁移到少样本学习事件检测模型中, 并且利用句子的句法依赖图来强化文本表示, 以提高其对新事件类别的泛化表示. Zheng 等人[45]提出利用空间距离关系将训练样本转换为样本对以丰富训练数据, 并采用 Poincaré嵌入在原型网络的基础上更好地表示层次化的类标签, 从而提高泛化性能.

综上, 大部分已有工作主要关注如何改善文本表示, 或从训练数据本身以及外部知识库引入知识以强化触发词的识别和分类线索, 而忽略了在表示学习和分类过程中不同事件类别之间数据量分布的差异, 这些方法很容易受到数据稀疏和不平衡的损害. 因此, 提出对触发词的元属性进行建模, 旨在通过学习共享属性来提升模型的泛化性.

1.3 缓解数据稀疏和不平衡

数据稀疏和不平衡是深度学习方法长久以来面临的一个问题. 其中, 不平衡又称为长尾分布, 是指在数据集存在多个类别时, 少量类别占训练总样本的绝大部分 (通常称为头部类), 而大部分类别占训练样本的小部分; 数据稀疏则是指正负样本中正样本占比极低的情形, 例如事件检测中相比非触发词, 触发词占比很低. 其本质上属于类别不平衡的一种特殊情况. 缓解数据不平衡的方法主要包括: (1) 基于采样的方法, 例如过采样 (对样本少的类别进行重复采样)、欠采样 (在样本多的类别中进行部分采样) 等. Ren 等人[46]针对 Softmax 分类器在长尾分布中有偏估计问题, 提出一种 Meta-Softmax 方法来估计每个类别的最佳训练样本采样率, 使训练样本分布更符合测试样本分布, Xu 等人[47]提出通过逐步调整标签空间, 划分头部类和尾部类, 动态构建类别样本的平衡采样以便于分类; (2) 基于损失的方法, 在损失项中对不同类别分配不同权重, 例如 Lin 等人[48]提出 Focal Loss, 在训练中对分类效果较好的类分配更低的交叉熵损失权重. Tian 等人[49]提出了一种困难类挖掘损失, 通过重塑交叉熵损失, 使其动态加权每个类的损失. Alshammari 等人[50]探索了长尾分布场景中的权重衰减, 提出在使用交叉熵损失学习特征和类平衡损失时调整权值衰减和 MaxNorm. 然而, 采样和损失的调整相对更依赖于任务和数据集分布, 为此本文从模型训练角度提出了一种元属性学习方法, 通过建模元属性适应并缓解不同程度的数据稀疏和不平衡.

2 模型方法

所提出的基于元属性学习的事件检测方法如图 2 所示, 包括: (1) 学习用于触发词识别的元属性以缓解数据稀疏, 以及 (2) 学习用于触发词分类的元属性以解决不平衡. 具体而言. 其中触发词分类方法本质上是完整的事件检测, 这里将非事件视为了预定义的类别, 并且集成了触发词识别组件. 因此, 本文提出基于元属性的事件检测方法, 试图对触发词的上述两个属性进行建模.

在图 2 中, 左半部分代表学习触发词识别的元属性. 右半部分代表学习触发词分类的元属性. 虚线框内的模块同时出现在训练和测试中, 虚线框外的部分则仅出现在训练中. 实线箭头表示变量的流向, 虚线箭头表示复制. 图 2 中 tap 表示类别无关投影层. 两部分之间的虚线表示分类方法融合了左侧虚线椭圆中提出的类别无关语义. 由于图 2 的空间有限, transfer-money 在下标中被缩写为 transfer. 下标 other 表示非事件类别

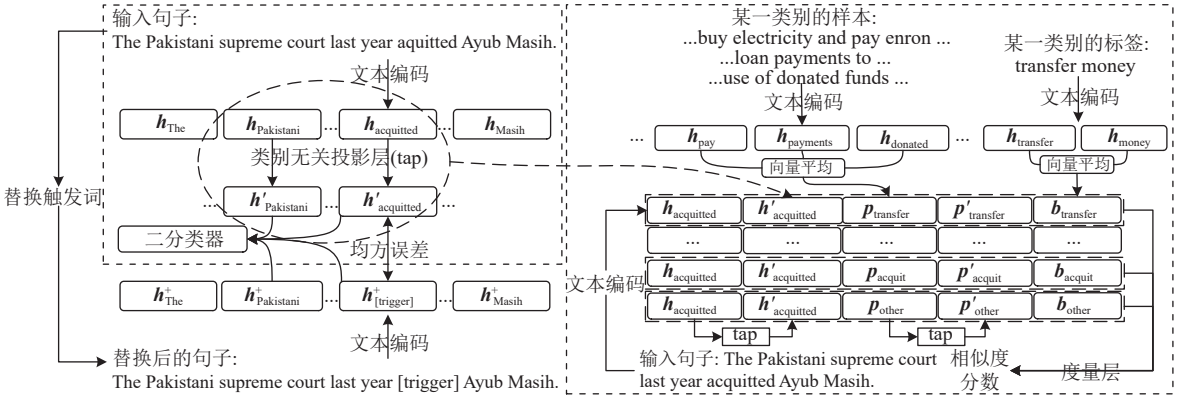


图 2 基于元属性学习的事件检测模型

在本文中, 样本指的是上下文中的触发词或非事件词. 将所有样本视为候选触发词 (称为输入样本, 记为 $\boldsymbol{x}$). E_φ 表示文本编码过程, 可以是任何先进的文本表示技术, 例如预训练语言模型 BERT[3]或 RoBERTa[4]等. 如公式 (1) 所示, 通过文本编码获取所有样本的表示向量 (记为 $\boldsymbol{h}$), 文本编码用于后续步骤.

公式 (1) 中, *PLM* 表示编码所使用的语言模型 (本文中为 BERT), φ 是其参数, X 表示样本 $\boldsymbol{x}$ 所在的句子, $i(\boldsymbol{x})$ 表示样本在句子中的索引位置 n, 符号 $[n]$ 表示在序列中取第 n 个元素.

$$\boldsymbol{h} = E_\varphi(\boldsymbol{x}) = PLM\varphi(X)_{[i(\boldsymbol{x})]} \tag{1}$$

2.1 用于触发词识别的元属性: 类别无关语义建模

在本节中, 模型重点学习触发词识别的元属性, 即: 每个触发词都可以抽象为一种类别, 即“事件”, 无论它是什么类别的事件. 模型试图抽取这样的语义, 并通过类别无关投影层将其用于区分触发词和非触发词. 具体过程如下.

● 触发词的类别无关语义: 每个触发词都包含与事件相关的语义, 通常充当句子中的关键谓词或宾语. 因此, 即使在一个句子中掩盖掉具体的触发词, 也可以预测这一位置所出现的词可能暗示某种事件的发生. 预训练语言模型可以通过掩盖某些特定的单词来预测其所在位置可能的所有单词, 原因是强大的注意力机制可以借助上下文获得特定位置单词的表示. 受此启发, 将所有触发词替换为一个特殊符号, 例如 [trigger] (记为 $R(\boldsymbol{x})$), 保留其上下文, 然后使用其表示作为类别无关的触发词语义 (记为 $\boldsymbol{h}^+$), 如公式 (2) 所示.

$$\boldsymbol{h}^+ = E_\varphi(R(\boldsymbol{x})) \tag{2}$$

● 类别无关投影层: 在这个过程中, 触发词的表示试图模仿类别无关的触发词语义 $\boldsymbol{h}^+$. 模型试图通过一个称为类别无关投影层的多层感知机 (MLP) 来约束触发词的表示向量, 使其与特殊符号的表示相似, 如公式 (3) 所示 (MLP 记为 $F_\varphi^{a\times b}$, 其中 a 和 b 分别是输入和输出的维数, φ 是其参数. 常数 Δ 表示向量的维数. 投影后的表示向量记为 $\boldsymbol{h}'$).

$$\boldsymbol{h}' = F_{\theta_{\text{tap}}}^{\Delta\times\Delta}(\boldsymbol{h}) \tag{3}$$

计算均方误差 (mean square error, MSE) 损失 (记为 M) 作为优化目标, 如公式 (4) 所示, 以更新类别无关投影层 (记为 $F_{\theta_{\text{tap}}}^{\Delta\times\Delta}$) 的参数 θ_{tap}. 包括非触发词在内的所有样本都在训练和测试期间投影, 而只有触发词参与了训练期间的损失计算.

$$L_{\theta_{\text{tap}}} = M(\boldsymbol{h}', \boldsymbol{h}^+) \tag{4}$$

● 基于类别无关抽象语义表示的触发词识别训练: 对这两种表示向量 ($\boldsymbol{h}'$ 和 $\boldsymbol{h}^+$) 进行二分类, 作为触发词识别. 在此过程中, 如公式 (5) 所示, 计算交叉熵损失 (记为 M, 其中第 1 项经过 Softmax 操作) 作为优化目标, 以更新二分类器的参数 (记为 θ_{idn}). 符号 I_x 是一个指示符号, 如果样本 $\boldsymbol{x}$ 是触发词, 则其值为 1, 否则为 0.

$$L_{\theta_{\text{idn}}} = C(F^{\Delta\times 2}_{\theta_{\text{idn}}}(h'),(I_{\boldsymbol{x}},1-I_{\boldsymbol{x}})) + C(F^{\Delta\times 2}_{\theta_{\text{idn}}}(h^{+}),(0,1)) \tag{5}$$

样本 $\boldsymbol{x}$ 为触发词的判别函数记为 $trig(\boldsymbol{x})$, 如公式 (6) 所示, 其中运算符⌊⌋表示向下取整, $trig(\boldsymbol{x})$ 为 1 时表示 $\boldsymbol{x}$ 是触发词, 为 0 则表示 $\boldsymbol{x}$ 是非触发词.

$$trig(x) = \left\lfloor 2F^{\Delta\times 2}_{\theta_{\text{idn}}}(h') \right\rfloor \tag{6}$$

最后, 通过将公式 (4) 和公式 (5) 中的联合训练损失相加, 构建整体识别损失函数, 如公式 (7) 所示.

$$L_{\theta_{\text{tap}},\theta_{\text{idn}}} = L_{\theta_{\text{tap}}} + L_{\theta_{\text{idn}}} \tag{7}$$

该模型通过学习“事件”的共同抽象语义, 理解触发词的作用机制来识别触发词. 这种作用机制对于任何特定的类别都是一致的. 它避免了模型退化为一个简单的基于词汇记忆的分类器.

2.2 用于触发词分类的元属性: 类别共享的相似度度量建模

在本节中, 模型重点学习触发词分类的元属性, 即: 触发词更类似于相同类别的触发词, 而不是其他类别的触发词. 模型建立了一种关于样本和每个事件类别相似度的度量模型, 并利用它来确定每个样本的类别. 此类度量由各种类别共享, 这有助于缓解数据不平衡. 具体过程如下.

●类别的样本表示: 为了计算输入样本和事件类别之间的相似度, 首先获取某一类别的所有样本 (记为 $x_k \in X_k$, 其中 X_k 表示类别 k 中的样本集合) 的表示, 并使用其平均值作为事件类别的表示 (记为 $\boldsymbol{p}_k$), 如公式 (8) 所示.

$$\boldsymbol{p}_k = \frac{1}{|X_k|}\sum_{x_k\in X_k} E_{\varphi}(x_k) \tag{8}$$

● 类别的标签表示: 由于某些类别的标签 (记为 l_k) 提供了额外的类别相关信息, 有助于提供补充语义信息, 特别是对于样本非常稀少, 无法准确表示类别语义的类别. 因此, 还获得了标签的表示 (记为 $\boldsymbol{b}_k$), 如公式 (9) 所示.

$$\boldsymbol{b}_k = E_{\varphi}(l_k) \tag{9}$$

● 类别共享的相似度度量: 计算样本 $\boldsymbol{x}$ 和事件类别 k 的相似度 (记为 $S_{x\cong k}$), 即通过度量层计算上述 3 种表示 $\boldsymbol{h}$, $\boldsymbol{p}_k$ 和 $\boldsymbol{b}_k$ 的组合, 如公式 (10) 所示. 运算 $[\cdot,\cdot]$ 表示向量的拼接. 样本的预测结果是与其相似度最高的类别. 在此过程中, 计算交叉熵损失作为优化目标, 以更新度量层的参数 (记为 θ_{meas}), 如公式 (11) 所示. 符号 $P_{x\cong k}$ 是一个指示符, 如果样本 x 的标注类别为 k, 则其值为 1, 否则为 0.

$$S_{x\cong k} = F^{3\Delta\times 1}_{\theta_{\text{tap}}}([\boldsymbol{h},\boldsymbol{p}_k,\boldsymbol{b}_k]) \tag{10}$$

$$L_{\theta_{\text{meas}}} = C(S_{x\cong k}, P_{x\cong k}) \tag{11}$$

因此, 样本 $\boldsymbol{x}$ 的类别 $Y(\boldsymbol{x})$ 可以表示为公式 (12).

$$Y(x) = \text{argmax}_k S_{x\cong k} \tag{12}$$

在公用的相似度度量影响下, 富样本类别 (sample-rich type, SRT) 的样本可以减少与少样本类别 (sample-scarce type, SST) 的样本之间的相似度, 反之亦然. 这种训练目标可以通过两种事件类别 (SRT 和 SST) 来实现. 因此, 所有类别的样本对模型训练的贡献相等, 被错误分类的概率更小.

2.3 类别无关抽象语义增强的事件检测模型

为了构建完整的事件检测模型, 将触发词识别组件集成到分类方法中, 即另外将样本表示 (公式 (3)) 和类别表示 (公式 (13)) 的类别无关语义拼接到公式 (12) 中的组合中, 以替换公式 (10).

$$\boldsymbol{p}'_k = F^{\Delta\times\Delta}_{\theta_{\text{tap}}}(\boldsymbol{p}_k) \tag{13}$$

$$S_{x\cong k} = F^{5\Delta\times 1}_{\theta_{\text{tap}}}([\boldsymbol{h},\boldsymbol{h}',\boldsymbol{p}_k,\boldsymbol{p}'_k,\boldsymbol{b}_k]) \tag{14}$$

完整事件检测模型的最终损失函数通过将公式 (4) 和公式 (10) 中的损失相加来构建, 用于联合训练, 如公式 (15) 所示.

$$L_{\theta_{\text{tap}},\theta_{\text{meas}}} = L_{\theta_{\text{tap}}} + L_{\theta_{\text{meas}}} \tag{15}$$

在融合模型的作用下, 只有当输入样本和类别的原始语义和类别无关语义都一致时, 才被视为相同的类别. 它减少了将事件触发词分类为错误事件类别或“非事件”类别, 或将非事件词分类为任何一种事件类别的概率.

3 模型总体性能的实验及分析

本节通过与已有方法和退化模型进行对比实验来验证所提出方法的总体性能. 对缓解数据稀疏和不平衡问题的有效性验证则放在第 4 节.

3.1 数据集

在两个英文数据集 ACE2005 和 MAVEN[51]上评估了本文提出的模型. ACE2005 包含 13 672 个标注句子, 包括 3 994 个事件提及, 分布在 599 篇文章中. MAVEN 包含 40 473 个标注句子. ACE2005 数据集定义了 33 种事件类别, MAVEN 数据集定义了 168 种事件类别. 两个数据集中大部分类别的样本都很稀少. 参照基线模型中的实验设置, 将 ACE2005 中的文章数分别划分为 529/30/40, 用于训练/验证/测试. 对于 MAVEN 数据集, 本实验采用官方的划分[51].

3.2 评价指标

本文采用的评价指标为 $F1$ 分数, 其计算方法为准确率 (precision, P) 和召回率 (recall, R) 的 2 倍调和平均, 如公式 (16) 所示. 其中, 对于二分类任务, 准确率 (P) 表示预测的正样本中实际为正样本的比例, 召回率 (R) 表示所有标注的正样本中被预测为正样本的比例. 对于 N (N>2) 分类任务, 则将每一个预测样本分解为 N 个二分类任务计算平均准确率、召回率和 $F1$ 分数. 若对样本数计算平均, 则称为微观 (micro-) 准确率、召回率和 $F1$ 分数, 若对类别数计算平均, 则称为宏观 (macro-) 准确率、召回率和 $F1$ 分数. 其中微观分数更能反应模型整体性能, 而宏观分数更能反映模型应对不平衡场景的能力.

$$F1 = \frac{2PR}{P+R} \tag{16}$$

对于触发词分类, 本文报告了事件类别的 micro-$F1$ 分数和 macro-$F1$ 分数. 触发词的类别和位置与标注都匹配, 才认为预测结果正确. 对于触发词识别, 本实验报告了 binary-$F1$ 分数.

编码器是预训练语言模型 BERT-base-uncased, 由 12 层带有 12 个注意力头, 隐藏嵌入维度为 768 的 Transformer 组成. 批次大小设置为 8, 学习率设置为 1E–5, 最大序列长度设置为 80. 模型在一台 NVIDIA Tesla V100 上训练.

3.3 对比模型

为了观察本文方法的整体性能, 我们与以下几类基线方法进行了比较.

(1) 不引入外部资源的方法

● PLMEE[30]提出利用预训练语言模型进行事件检测和自动数据生成. 注意到 PLMEE 中的事件检测模块实际上是 BERT 模型, 本实验中重新实现了 BERT 表示+分类作为基线模型之一.

● CDSI (combination of dependency and semantic information)[41]利用图卷积网络建模依存信息, 以及利用自注意力机制建模语义信息, 并采用门控机制将两者融合进行事件检测.

● SS-VQ-VAE[43]提出利用半监督向量量子化变分自动编码器框架来学习每个已知和未见过类别的离散潜在类别表示.

● M-Mode[34]提出一个上下文选择掩码泛化模型, 用于挖掘上下文特定的模式以进行学习, 并使模型具有鲁棒性.

● MLBiNet[26]提出了一个多层解码器结构用于同时捕获事件和语义信息的文档级关联.

● SaliencyED (SL)[39]将所有触发词类别划分为触发词依赖型和语境依赖型分别进行建模以充分挖掘分类线索.

● CorEd-BiLSTM[19]设计了一种基于图的自适应类别编码器来捕获实例级相关性.

(2) 引入外部资源的方法

● DMBERT[15]提出应用对抗训练机制来迭代识别信息型实例并过滤掉有噪声的实例.

- MSBERT+Entity[35]基于共享的 BERT 模型表示构建包括事件检测和实体识别在内的多任务学习框架.
- EKD[7]利用知识抽取框架从 FrameNet 引入开放的域触发词知识.
- CAKD[37]设计了一种基于触发词识别强化与可适配分类器的事件检测方法.

ACE2005 数据集上的所有基线结果均来自原始论文, MAVEN 上的基线结果来自文献 [51].

3.4 整体性能对比

ACE2005 和 MAVEN 上的总体性能如表 1 所示. 其中, binary-*F*1 分数是触发词识别的性能. 符号*表示基于外部资源的方法. 基线方法 (PLMEE 除外) 的结果来自原始论文[7,15,19,26,30,34,35,37,39,41,43]. 符号“N/A”表示原文未报告该项结果或由于数据集划分不同等原因不适合比较.

表 1 标准划分数据集上的对比实验和退化实验结果 (%)

方法类型	评价指标	ACE2005			MAVEN	
		micro-*F*1	macro-*F*1	binary-*F*1	micro-*F*1	macro-*F*1
不引入外部资源的方法	PLMEE[30]	74.7	48.0	79.8	65.0	53.7
	CDSI[41]	73.9	N/A	76.3	N/A	N/A
	SSVQVAE [43]	76.7	N/A	80.2	N/A	N/A
	M-Model[34]	74.8	N/A	N/A	N/A	N/A
	MLBiNet [26]	**78.6**	N/A	N/A	N/A	N/A
	SaliencyED (SL)[39]	75.1	N/A	N/A	66.5	59.2
	CorED-BiLSTM[19]	77.5	N/A	N/A	67.5	60.3
引入外部资源的方法	DMBERT[15]*	75. 1	N/A	N/A	**67.1**	N/A
	MSBERT+Entity[35]*	76.5	N/A	79.2	N/A	N/A
	EKD[7]*	**78.6**	N/A	N/A	N/A	N/A
	CAKD[37]*	78. 1	N/A	79.4	N/A	N/A
本文方法及退化模型	Ours (Full)	78.1	**52.6**	**85.2**	66.9	**59.9**
	–w/o TAS	76.8	50.4	81.3	66.5	58.8
	–w/o LR	76.5	49.4	80.7	66.2	56.0
	–w/o Both	75.5	48.7	80.1	65.6	53.3

可以看出, 本文所提出的完整模型取得了比较满意的性能, 其识别性能也优于基于表示和二分类的识别方法. 原因可能是本文所提出的模型可以利用触发词的元属性来提高触发词识别和分类的性能. 本文所提出的完整事件检测方法没有超过一些基线, 但达到了相似的性能, 例如 ACE2005 上的 EKD 和 MAVEN 数据集上的 DMBERT, 原因可能是其受益于难以复制的丰富外部资源和强大计算资源 (用于支持更大的 batch-size).

本文所提出的模型在多个基线模型中表现优于或达到了类似的性能, 这表明了它的有效性. 除了使用 PLM 作为编码器的模型外, 之前的大多数工作主要集中在词汇相关问题上, 包括歧义和未见过的/稀疏标注的触发词. 然而, 由于能够捕获上下文信息, 大规模预训练语言模型改进了单词表示, 并且与以前的工作相比取得了相似或更好的性能. 因此, 在基于序列的模型中, 基于 PLM 的模型 DMBERT、M-Model 和所提出的模型优于大多数基于 LSTM 的模型 (包括基于 Elmo 的 Delta) 和基于图的模型. 与这些基于 PLM 的模型相比, 本文所提出的方法在 *F*1 分数上提高了或达到了类似的性能, 这表明该方法的有效性是由于元属性学习的成功, 而不仅是一个强力的编码器.

与 PLMEE、DMBERT 和 M-Model 比, 本文的工作主要集中在如何通过缓解数据不平衡问题来提高性能. 本文所提出的模型取得了比较满意的性能, 可能有两个原因: 1) 模型可以将学习到的关系度量从富样本类迁移到少样本类, 充分利用了富样本类的优势, 从而提高了触发词分类的整体性能. 2) 模型进一步利用了普通 MLP 分类器在富样本类上的优势和在少样本类上的可学习度量, 实现了全面的改进.

3.5 退化实验分析

为了验证本文所提出的模型中每个模块的有效性, 本实验通过比较以下退化模型进行退化实验.

(1) Ours (Full): 完整模型, 包括类别无关语义, 类别的标签表示和可学习的度量.

(2) –w/o TAS: 从 (1) 中删除类别无关语义.

(3) –w/o LR: 从 (1) 中删除类别的标签表示.

(4) –w/o Both: 同时删除前两者作为基线模型, 忽略数据稀疏和不平衡问题.

表 1 报告了退化实验的结果, 从中获得了以下观察结果.

● 类无关投影层的作用: 可以看出, 与没有类别无关投影的模型相比, 完整模型获得了更好的性能, 特别是在触发词识别性能方面, 这表明了触发词的类别无关语义所提供信息的有效性. 原因可能是类别无关语义包含类别事件触发词的统一特征, 它能够提供关于单词是否属于事件触发词的判别信息, 有助于避免将非事件单词预测为事件触发词的错误.

● 标签表示的作用: 可以看出, 完整模型比没有标签表示的模型获得更好的性能, 显示了标签提供的信息的有效性. 原因可能是标签是稳定不变的, 包含事件类别的核心语义信息. 由于本文所提出的度量层是一个可学习的度量, 而不是一个简单的余弦相似度, 因此它受益于这些语义信息.

3.6 在不平衡类别上的表现

本实验计算了预测结果的 micro-$F1$ 和 macro-$F1$ 分数, 如表 1 所示. 宏观 $F1$ 得分受数据不平衡的影响更大. 本实验将本文所提出的完整模型与两个基线进行比较: PLMEE 模型和–w/o LR 模型.

结果表明, 本文所提出的完整模型同时实现了微观 $F1$ 分数和宏观 $F1$ 分数的提高, 显示了其从富样本类向少样本类迁移的有效性. 原因可能是: 1) PLMEE 模型无法学习事件类别之间的共同指导. 为了提高总体分类精度, 它倾向于将样本正确地分类在富样本类上, 导致在少样本类上的性能相对不理想. 2) –w/o LR 模型通过可学习的度量, 学习的比较能力使富样本类和少样本类的分类相互促进. 然而, 如果没有类别的标签表示, 由于少样本类样本的多样性较低, 其预测结果并不令人满意. 3) 在完整模型中, 类别表示通过标签表示得到了增强, 并获得了相对高质量的语义表示, 特别是对少样本类有效.

4 对缓解数据稀疏和不平衡问题的有效性验证

本节通过探索 3 个问题来验证所提出方法的有效性.

● 问题 1: 学习事件的元属性对缓解数据稀疏和不平衡问题有无帮助?

● 问题 2: 样本数和样本比例如何影响最终的事件检测效果?

● 问题 3: 传统的完全监督实验设置和少样本实验设置怎样建立关联?

因此, 本节建立了一组平行实验, 通过控制每个实验中样本的数量和比例来观察性能的差异.

4.1 数据集与评价指标

本节在两个英文数据集上进行了实验验证, 分别是 ACE2005 数据集和 MAVEN 数据集[51]. 为了充分探索数据稀疏和不平衡的情况, 本实验根据以下原则重新划分了数据集.

● 验证集中的样本数分布是平衡的, 即每个类别的样本数是一个常量 (记为 N_{eval}). 原因是一种类别的样本过多更容易影响微观 $F1$ 分数, 而样本过少更容易影响宏观 $F1$ 分数.

● 训练集中每个类别 k 的样本数 (记为 N^k_{train}) 的计算公式为 $N^k_{\text{train}}=\eta(N^k_{\text{total}}-N^{\min}_{\text{train}}-N_{\text{eval}})+N^{\min}_{\text{train}}$, 其中 $0\leqslant\eta\leqslant 1$ 是一个固定的比例, 称为平衡系数, N^k_{total} 是类别 k 在标注语料中的样本总数, $N^{\min}_{\text{train}}$ 是每个类别样本数量的固定最小值. η 的值越低训练样本总数越低, 但样本分布更平衡.

● 样本数小于 $N_{\text{eval}}+N^{\min}_{\text{train}}$ 的类别不足以有效地学习和评估, 因此被丢弃.

表 2 显示了这两个数据集及其划分的信息. 确定样本数后, 具体的样本将随机选择, 训练集和验证集之间没有样本重叠.

本实验报告了微观 (micro-, 用于触发词分类) 和二分类 (binary-, 用于触发词识别) 准确率、召回率和 $F1$ 分

数. 如果触发词的类别和位置与标注都匹配, 则认为预测结果是正确的. 随后获得的每个结果都是 10 次运行的平均值. 本实验采用了在标准划分验证集上优化的超参数 (见第 3.1 节).

表 2 实验数据集

设置	ACE2005	MAVEN
预定义事件类别数	33	168
验证样本数	10	45
最小训练样本数	2	5
丢弃的类别数	4	18

4.2 基线方法

本质上, 基于元属性的事件检测方法是一种基于文本表示的联合模型. 因此, 本实验将其与以下几条重新实现的基线模型进行比较, 这些基线模型具有更好的代表性.

● 表示+分类器 (RC)[30], 在获得样本的表示后, 该方法使用 MLP+Softmax 多分类器将其分类为事件类别之一或非事件类别.

● 表示+识别+分类器 (RIdC), 仅使用 MLP+Softmax 多分类器将样本分类为一种事件类别, 由二分类器确定它是否为触发词.

● 表示+原型网络 (RP)[42], 通过计算样本表示与特定类别样本表示的平均值之间的余弦相似度, 将样本分类为事件类别之一或非事件类别.

● 表示+识别+原型网络 (RIdP), 类似 RP, 只将样本分类为一种事件类别, 并通过二分类器确定它是否是触发词.

此外, 本实验还构建了以下退化模型.

● –w/o Proj: 删除类别无关语义的组件, 即删除公式 (12) 中的 $\boldsymbol{h}'$ 和 $\boldsymbol{p}'_k$, 仅使用公式 (9) 的组合.

● –w/o Label: 进一步从–w/o Proj 模型中删除标签表示的组件, 即删除公式 (9) 中的 $\boldsymbol{b}_k$.

度量层的大小随着组件的添加或移除而相应变化.

本实验还将触发词识别组件 (第 2.1 节) 与基于表示+二分类器 (记为 Idn) 的方法进行了比较. 整个实验采用预训练的 BERT 模型 BERT-base-cased 用于文本编码.

4.3 实验结果及分析

表 3 和表 4 分别报告触发词分类和触发词识别的结果. 对于 η 的每个设置, Ours (Full) 表示本文提出的触发词分类方法. 粗体表示相同设置中的最高结果. 在表 4 中, 对于每个 η 设置, Ours (I) 表示触发词识别组件. 粗体表示相同设置中的最高结果. 此外, 本实验将类别分为两类: 样本数小于平均样本数的类别称为少样本类别 (SST), 否则称为富样本类别 (SRT). 这里的“充足”和“稀缺”是相对的. 表 5 报告了两种类别 (分别记为 Rich 和 Scarce) 的 micro-$F1$ 分数的分类结果. 本实验比较了不同场景下的各种方法, 并分析了样本数量和样本比例变化的影响.

4.3.1 方法间的对比

对于问题 1, 从表 3–表 5 中获得的观察结果如下.

● 当表示保持一致时, 本文所提出的方法在 $F1$ 分数上的性能优于其他方法, 无论是最终检测还是触发词识别.

● 本文所提出的识别组件在 ACE2005 数据集上表现相对较好, 该数据集中数据稀疏更为严重. 此外, 本文所提出的完整模型比–w/o Proj 退化模型性能更好.

● 本文所提出的完整模型比–w/o Proj 模型实现了更高的召回率, 这表明在类别无关语义的指导下, 在分类过程中可以成功识别更多的触发词, 这对于解决数据稀疏问题至关重要.

● 与其他方法相比, 在大多数情况下, 本文所提出的方法在少样本类别和富样本类别之间的差距较小, 并且在某些情况下本文所提出的方法对样本稀缺型的表现甚至更好.

● 去掉类别无关的语义和标签表示, 只保留最简单的模型, 它仍然比基线方法性能更好. 这证明改进不仅来自

上述两个部分.

● 触发词识别的结果甚至低于 ACE2005 数据集上的完全检测结果. 这表明, 作为二分类, 触发词识别也是一项具有挑战性的任务, 原因可能是很难捕获“事件”的统一特征, 尤其是在数据稀疏情况下.

表 3　多种不平衡程度下的完整事件检测性能对比 (%)

η	方法	ACE2005			MAVEN		
		P	R	$F1$	P	R	$F1$
1.0	Ours (Full)	86.7	**85.6**	**86.2**	**78.1**	**76.6**	**77.2**
	–w/o Proj	**89.6**	78.5	83.6	76.5	68.4	72.2
	–w/o Label	85.9	74.2	79.6	72.6	64.7	68.4
	RC	81.7	59.2	68.7	70.5	58.2	63.8
	RIdC	83.7	45.9	59.3	71.2	49.3	58.2
	RP	4.3	60.5	8.0	6.1	51.0	10.9
	RIdP	53.0	39.0	44.9	40.1	38.2	39.1
0.1	Ours (Full)	78.0	**64.6**	**70.6**	**71.4**	**69.8**	**70.6**
	–w/o Proj	78.7	59.5	67.7	71.1	66.5	69.0
	–w/o Label	74.7	56.3	64.2	70.8	64.5	67.5
	RC	80.2	48.3	60.3	68.5	53.9	60.3
	RIdC	**81.1**	38.6	52.3	69.2	47.0	56.0
	RP	4.1	60.4	7.6	6.1	50.9	11.0
	RIdP	49.5	38.0	43.0	39.4	39.6	39.5
0.01	Ours (Full)	61.3	**45.8**	**52.4**	60.3	**55.0**	57.4
	–w/o Proj	69.9	39.4	50.4	61.7	54.0	**57.5**
	–w/o Label	69.3	40.0	49.9	61.0	51.7	55.9
	RC	83.8	27.4	41.3	65.8	48.0	55.5
	RIdC	**86.4**	24.1	37.7	**66.2**	41.1	50.7
	RP	3.0	45.8	5.7	5.7	47.8	10.2
	RIdP	33.9	29.4	31.5	37.5	35.3	36.4

表 4　多种不平衡程度下的触发词识别性能对比 (%)

η	方法	ACE2005			MAVEN		
		P	R	$F1$	P	R	$F1$
1.0	Ours (I)	70.8	**55.6**	**62.3**	74.8	**72.7**	**73.7**
	Idn	**76.8**	47.9	59.0	**78.0**	67.3	72.2
0.1	Ours (I)	69.5	**54.6**	**61.4**	72.8	**74.5**	**73.6**
	Idn	**72.7**	49.7	59.0	**76.5**	69.2	72.6
0.01	Ours (I)	61.2	**51.5**	**55.9**	74.9	**71.2**	**73.0**
	Idn	**74.6**	36.2	48.8	**76.8**	67.9	72.0

综上所述, 本实验得出结论: 学习类别无关语义的元属性有助于解决数据稀疏问题, 而进一步学习相似度度量的元属性则有助于处理数据不平衡问题.

4.3.2　样本数量和比例的影响

对于问题 2, 从表 5 中获得的观察结果如下.

● 样本绝对稀缺且占比例低是本质的不利因素. 结果表明, 随着平衡系数的下降, RC 和 RIdP 的结果比本文所提出的方法下降得更显著, 特别是在样本丰富的类别上. 这显示了样本数的影响.

● 当样本数量有限时, 增加比例是有利的. 随着平衡系数的降低, 数据分布更加均衡. 富样本类别的样本比例下降, 导致该类别的性能下降. 虽然少样本类别的样本数量略有减少, 但样本比例有所增加, 本文所提出的方法的性能并没有变得相对较差.

● 对于样本数量, 足够是最好的, 而不是越多越好. 随着平衡系数的增加, 富样本类别的样本数量更多, 但其性

能增长缓慢; 少样本类别的样本增长较慢, 但其性能增长相对较快. 原因可能是, 尽管少样本类别中样本的比例相对较低, 但数量足够, 样本充足的类别不会随着样本数量的增加而表现得更好. 同时, 富样本类别中样本过多, 使得绝对少样本类别的比例降低, 导致整体性能变差. 这揭示了长尾现象的本质, 并说明应使数据分布尽可能平衡和充分.

表 5　富样本与少样本类别的实验性能对比 (%)

η	方法	ACE2005		MAVEN	
		Rich	Scarce	Rich	Scarce
1.0	Ours (Full)	**83.8**	**87.5**	**78.6**	**74.6**
	–w/o Proj	**83.8**	83.2	74.8	70.5
	–w/o Label	80.5	79.2	70.0	65.0
	RC	67.8	69.2	64.3	62.6
	RIdC	62.5	57.1	59.7	55.8
	RP	13.1	6.6	14.9	7.8
	RIdP	50.2	42.6	41.5	34.8
0.1	Ours (Full)	69.0	**71.6**	**71.8**	**68.2**
	–w/o Proj	**69.1**	65.9	69.7	67.7
	–w/o Label	62.3	65.3	68.7	65.5
	RC	62.7	56.7	61.6	57.6
	RIdC	55.5	51.7	57.7	51.9
	RP	10.8	6.5	15.5	7.3
	RIdP	45.1	42.4	42.4	35.0
0.01	Ours (Full)	**51.9**	**52.7**	**60.1**	51.5
	–w/o Proj	51.8	49.7	59.1	**54.5**
	–w/o Label	50.5	49.5	57.9	51.9
	RC	45.2	35.5	56.5	49.1
	RIdC	41.2	32.8	52.7	44.7
	RP	11.0	4.5	16.3	6.4
	RIdP	36.6	36.2	39.9	31.7

4.3.3　极端少样本设置下的表现

对于问题 3, 本实验答案是, 随着 $\eta \to 0$, 实验设置从传统设置演变为少样本设置, 其中所有类别只有最小数量的训练样本. 本实验报告了不同数量的评估类别下的性能, 包括 5 类, 15 类和所有类别, 从而进一步分析样本分布均衡时类别数引起的比例变化的影响. 结果如表 6 所示, 其中, #*N* 表示 *N* 分类.

表 6　极少样本设置下多种类别数的实验性能对比 (%)

方法	ACE2005			MAVEN		
	#5	#15	#All	#5	#15	#All
Ours (Full)	57.9	50.8	43.4	62.4	58.1	49.6
–w/o Proj	57.1	47.7	41.1	61.9	56.6	48.8
–w/o Label	56.4	46.6	38.5	60.1	54.7	48.3
RC	46.8	32.8	23.8	55.2	50.5	46.0
RIDC	44.9	30.8	22.2	52.1	47.2	42.7
RP	2.5	4.2	5.1	1.0	2.2	9.5
RIdP	30.4	31.5	27.9	16.1	26.7	33.4

可以看出, 一些基线方法几乎无法在 ACE2005 数据集上进行有效的事件检测, 并且随着事件类别的增加, 性能都有所下降. 这表明极少数样本不足以训练有效的事件检测模型. 与其他方法相比, 本文所提出的方法实现了更好的性能, 并且随着事件类别的减少, 性能有了更显著的提高. 这可能是因为更多的类别提供了更多的样本, 而本文所提出的方法可以从元属性中受益.

5 总 结

本文提出了一种基于元属性的模型来缓解事件检测中的数据稀疏性和不平衡性. 它以神经网络层参数的形式建模事件触发词的两种元属性, 包括事件触发词的类别无关抽象语义表示和类别内部的统一相似度度量, 并将其作为最终的训练目标. 通过在不同的稀疏性和不平衡程度下的实验, 验证了该方法的有效性, 并建立了常规设置和少样本设置之间的联系. 通过综合实验得出 3 个关键结论: (1) 学习样本的元属性有助于提高触发词的识别和分类, 并缓解数据的稀疏性和不平衡性; (2) 对于样本有限的类别, 通过适当减少其他富样本来增加其样本比例可以提高其性能; (3) 通过控制样本和类别的数量和比例, 可以统一常规和少样本实验设置.

References:

[1] Rudnik C, Ehrhart T, Ferret O, Teyssou D, Troncy R, Tannier X. Searching news articles using an event knowledge graph leveraged by Wikidata. In: Proc. of the Companion of the 2019 World Wide Web Conf. San Francisco: ACM, 2019. 1232–1239. [doi: 10.1145/3308560.3316761]

[2] Wu D, Ding L, Lu F, Xie J. SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 1932–1937. [doi: 10.18653/v1/2020.emnlp-main.152]

[3] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[4] Liu YH, Ott M, Goyal N, Du JF, Joshi M, Chen DQ, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.

[5] Liu SL, Chen YB, He SZ, Liu K, Zhao J. Leveraging FrameNet to improve automatic event detection. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016. 2134–2143. [doi: 10.18653/v1/P16-1201]

[6] Ding X, Liao K, Liu T, Li ZY, Duan JW. Event representation learning enhanced with external commonsense knowledge. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: ACL, 2019. 4894–4903. [doi: 10.18653/v1/D19-1495]

[7] Tong MH, Xu B, Wang S, Cao YX, Hou L, Li JZ, Xie J. Improving event detection via open-domain trigger knowledge. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 5887–5897. [doi: 10.18653/v1/2020.acl-main.522]

[8] Wang ZQ, Wang XZ, Han X, Lin YK, Hou L, Liu ZY, Li P, Li JZ, Zhou J. CLEVE: Contrastive pre-training for event extraction. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. ACL, 2021. 6283–6297. [doi: 10.18653/v1/2021.acl-long.491]

[9] Deng SM, Zhang NY, Li LQ, Hui C, Huaixiao T, Chen MS, Huang F, Chen HJ. OntoED: Low-resource event detection with ontology embedding. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. ACL, 2021. 2828–2839. [doi: 10.18653/v1/2021.acl-long.220]

[10] Liao JZ, Zhao X, Li XY, Zhang LL, Tang JY. Learning discriminative neural representations for event detection. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2021. 644–653. [doi: 10.1145/3404835.3462977]

[11] Deng SM, Zhang NY, Kang JJ, Zhang YC, Zhang W, Chen HJ. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In: Proc. of the 13th Int'l Conf. on Web Search and Data Mining. Houston: ACM, 2020. 151–159. [doi: 10.1145/3336191.3371796]

[12] Lai VD, Nguyen TH, Dernoncourt F. Extensively matching for few-shot learning event detection. In: Proc. of the 1st Joint Workshop on Narrative Understanding, Storylines, and Events. ACL, 2020. 38–45. [doi: 10.18653/v1/2020.nuse-1.5]

[13] Cong X, Cui SY, Yu BW, Liu TW, Wang YB, Wang B. Few-shot event detection with prototypical amortized conditional random field. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. ACL, 2021. 28–40. [doi: 10.18653/v1/2021.findings-acl.3]

[14] Shen SR, Wu TT, Qi GL, Li YF, Haffari G, Bi S. Adaptive knowledge-enhanced Bayesian meta-learning for few-shot event detection. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. 2021. 2417–2429. [doi: 10.18653/v1/2021.findings-acl.214]

[15] Wang XZ, Han X, Liu ZY, Sun MS, Li P. Adversarial training for weakly supervised event detection. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1. Minneapolis: ACL, 2019. 998–1008. [doi: 10.18653/v1/N19-1105]

[16] Lai V, Dernoncourt F, Nguyen TH. Learning prototype representations across few-shot tasks for event detection. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 5270–5277. [doi: 10.18653/v1/2021.emnlp-

main.427]

[17] He RF, Duan SY. Joint Chinese event extraction based multi-task learning. Ruan Jian Xue Bao/Journal of Software, 2019, 30(4): 1015−1030 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5380.htm [doi: 10.13328/j.cnki.jos.005380]

[18] Liu SL, Chen YB, Liu K, Zhao J, Luo ZC, Luo W. Improving event detection via information sharing among related event types. In: Proc. of the 16th China National Conf. on Chinese Computational Linguistics and 5th Int'l Symp. on Natural Language Processing Based on Naturally Annotated Big Data. Nanjing: Springer, 2017. 122−134. [doi: 10.1007/978-3-319-69005-6_11]

[19] Sheng JW, Sun R, Guo S, Cui SY, Cao JX, Wang LH, Liu TW. CorED: Incorporating type-level and instance-level correlations for fine-grained event detection. In: Proc. of the 45th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Madrid: ACM, 2022. 1122−1132. [doi: 10.1145/3477495.3531956]

[20] Ji H, Grishman R. Refining event extraction through cross-document inference. In: Proc. of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Columbus: ACL, 2008. 254−262.

[21] Liao SS, Grishman R. Using document level cross-event inference to improve event extraction. In: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala: ACL, 2010. 789−797.

[22] Giuglea AM, Moschitti A. Semantic role labeling via FrameNet, VerbNet and PropBank. In: Proc. of the 21st Int'l Conf. on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Sydney: ACL, 2006: 929−936. [doi: 10.3115/1220175.1220292]

[23] Chen YB, Xu LH, Liu K, Zeng DJ, Zhao J. Event extraction via dynamic multi-pooling convolutional neural networks. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing. Beijing: ACL, 2015. 167−176. [doi: 10.3115/v1/P15-1017]

[24] Duan SY, He RF, Zhao WL. Exploiting document level information to improve event detection via recurrent neural networks. In: Proc. of the 8th Int'l Joint Conf. on Natural Language Processing. ACL, 2017. 352−361.

[25] Nguyen TH, Cho K, Grishman R. Joint event extraction via recurrent neural networks. In: Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016. 300−309. [doi: 10.18653/v1/N16-1034]

[26] Lou DF, Liao ZL, Deng SM, Zhang NY, Chen HJ. MLBiNet: A cross-sentence collective event detection network. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. ACL, 2021. 4829−4839. [doi: 10.18653/v1/2021.acl-long.373]

[27] Nguyen TH, Grishman R. Graph convolutional networks with argument-aware pooling for event detection. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI Press, 2018. 5900−5907. [doi: 10.1609/aaai.v32i1.12039]

[28] Cui SY, Yu BW, Liu TW, Zhang ZY, Wang XB, Shi JQ. Edge-enhanced graph convolution networks for event detection with syntactic relation. In: Proc. of the 2020 Findings of the Association for Computational Linguistics. ACL, 2020. 2329−2339. [doi: 10.18653/v1/2020.findings-emnlp.211]

[29] Lai VD, Nguyen TN, Nguyen TH. Event detection: Gate diversity and syntactic importance scores for graph convolution neural networks. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 5405−5411. [doi: 10.18653/v1/2020.emnlp-main.435]

[30] Yan HR, Jin XL, Meng XB, Guo JF, Cheng XQ. Event detection with multi-order graph convolution and aggregated attention. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: ACL, 2019. 5766−5770. [doi: 10.18653/v1/D19-1582]

[31] Hong Y, Zhou WX, Zhang JL, Zhou GD, Zhu QM. Self-regulation: Employing a generative adversarial network to improve event detection. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 515−526. [doi: 10.18653/v1/P18-1048]

[32] Liu J, Chen YB, Liu K. Exploiting the ground-truth: An adversarial imitation based knowledge distillation approach for event detection. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 6754−6761. [doi: 10.1609/aaai.v33i01.33016754]

[33] Yang S, Feng DW, Qiao LB, Kan ZG, Li DS. Exploring pre-trained language models for event extraction and generation. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 5284−5294. [doi: 10.18653/v1/P19-1522]

[34] Liu J, Chen YB, Liu K, Jia YT, Sheng ZC. How does context matter? On the robustness of event detection with context-selective mask generalization. In: Proc. of the 2020 Findings of the Association for Computational Linguistics. ACL, 2020. 2523−2532. [doi: 10.18653/v1/2020.findings-emnlp.229]

[35] Wang J, Hong Y, Chen JL, Yao JM. Event detection by shared BERT and gated multi-task learning. Journal of Chinese Information

Processing, 2021, 35(10): 101–109 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2021.10.012]

[36] Pouran Ben Veyseh A, Lai V, Dernoncourt F, Nguyen TH. Unleash GPT-2 power for event detection. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. ACL, 2021. 6271–6282. [doi: 10.18653/v1/2021.acl-long.490]

[37] Song DD, Xu J, Pang JH, Huang HY. Classifier-adaptation knowledge distillation framework for relation extraction and event detection with imbalanced data. Information Sciences, 2021, 573: 222–238. [doi: 10.1016/j.ins.2021.05.045]

[38] Lu YJ, Lin HY, Han XP, Sun L. Distilling discrimination and generalization knowledge for event detection via delta-representation learning. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 4366–4376. [doi: 10.18653/v1/P19-1429]

[39] Liu J, Chen YF, Xu JN. Saliency as evidence: Event detection with trigger saliency attribution. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022. 4573–4585. [doi: 10.18653/v1/2022.acl-long.313]

[40] Wang SJ, Yu M, Chang SY, Sun LC, Huang LF. Query and extract: Refining event extraction as type-oriented binary decoding. In: Proc. of the 2022 Findings of the Association for Computational Linguistics. Dublin: ACL, 2022. 169–182. [doi: 10.18653/v1/2022.findings-acl.16]

[41] Chen JL, Hong Y, Wang J, Zhang JL, Yao JM. Combination of dependency and semantic information via gated mechanism for event detection. Journal of Chinese Information Processing, 2020, 34(8): 51–60 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2020.08.007]

[42] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4080–4090.

[43] Huang LF, Ji H. Semi-supervised new event type induction and event detection. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 718–724. [doi: 10.18653/v1/2020.emnlp-main.53]

[44] Lai VD, van Nguyen M, Nguyen TH, Dernoncourt F. Graph learning regularization and transfer learning for few-shot event detection. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2021. 2172–2176. [doi: 10.1145/3404835.3463054]

[45] Zheng JM, Cai F, Chen WY, Lei WQ, Chen HH. Taxonomy-aware learning for few-shot event detection. In: Proc. of the 2021 Web Conf. 2021. Ljubljana: ACM, 2021. 3546–3557. [doi: 10.1145/3442381.3449949]

[46] Ren JW, Yu CJ, Sheng SN, Ma X, Zhao HY, Yi S, Li HS. Balanced Meta-Softmax for long-tailed visual recognition. In: Proc. of the 34th Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 351.

[47] Xu Y, Li YL, Li JF, Lu CW. Constructing balance from imbalance for long-tailed image recognition. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 38–56. [doi: 10.1007/978-3-031-20044-1_3]

[48] Lin TY, Goyal P, Girshick R, He KM, Dollár P. Focal loss for dense object detection. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2980–2988. [doi: 10.1109/ICCV.2017.324]

[49] Tian JJ, Mithun NC, Seymour Z, Chiu HP, Kira Z. Striking the right balance: Recall loss for semantic segmentation. In: Proc. of the 2022 Int'l Conf. on Robotics and Automation. Philadelphia: IEEE, 2022. 5063–5069. [doi: 10.1109/ICRA46639.2022.9811702]

[50] Alshammari S, Wang YX, Ramanan D, Kong S. Long-tailed recognition via weight balancing. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 6897–6907. [doi: 10.1109/CVPR52688.2022.00677]

[51] Wang XZ, Wang ZQ, Han X, Jiang WY, Han R, Liu ZY, Li JZ, Li P, Lin YK, Zhou J. MAVEN: A massive general domain event detection dataset. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 1652–1671. [doi: 10.18653/v1/2020.emnlp-main.129]

附中文参考文献:

[17] 贺瑞芳, 段绍杨. 基于多任务学习的中文事件抽取联合模型. 软件学报, 2019, 30(4): 1015–1030. http://www.jos.org.cn/1000-9825/5380.htm [doi: 10.13328/j.cnki.jos.005380]

[35] 王捷, 洪宇, 陈佳丽, 姚建民. 基于共享 BERT 和门控多任务学习的事件检测方法, 中文信息学报, 2021, 35(10): 101–109. [doi: 10.3969/j.issn.1003-0077.2021.10.012]

[41] 陈佳丽, 洪宇, 王捷, 张婧丽, 姚建民. 利用门控机制融合依存与语义信息的事件检测方法. 中文信息学报, 2020, 34(8): 51–60. [doi: 10.3969/j.issn.1003-0077.2020.08.007]

贺瑞芳(1979－), 女, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为自然语言处理, 社会媒体挖掘, 机器学习.

马劲松(1997－), 男, 硕士生, 主要研究领域为自然语言处理, 事件抽取.

黄孝家(2000－), 男, 硕士生, 主要研究领域为自然语言处理, 事件抽取.

张仕奇(1997－), 男, 硕士, 主要研究领域为数据分析与挖掘.

白洁(1981－), 男, 硕士, 主要研究领域为大数据及智能技术应用.