软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(4):1665−1691 [doi: 10.13328/j.cnki.jos.007245] [CSTR: 32375.14.jos.007245]

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

多模态信息抽取研究综述*

王永胜, 李培峰, 王中卿, 朱巧明

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

通信作者: 李培峰, E-mail: pfli@suda.edu.cn

摘　要: 多模态信息抽取任务是指从非结构化或半结构化的多模态数据 (包含文本和图像等) 中提取结构化知识. 其研究内容主要包含多模态命名实体识别、多模态实体关系抽取和多模态事件抽取. 首先对多模态信息抽取任务进行分析, 然后对多模态命名实体识别、多模态实体关系抽取和多模态事件抽取这 3 个子任务的共同部分, 即多模态表示和融合模块进行归纳和总结. 随后梳理上述 3 个子任务的常用数据集和主流研究方法. 最后总结多模态信息抽取的研究趋势并分析该研究存在的问题和挑战, 为后续相关研究提供参考.

关键词: 多模态信息抽取; 多模态命名实体识别; 多模态实体关系抽取

中图法分类号: TP18

中文引用格式: 王永胜, 李培峰, 王中卿, 朱巧明. 多模态信息抽取研究综述. 软件学报, 2025, 36(4): 1665–1691. http://www.jos.org.cn/1000-9825/7245.htm

英文引用格式: Wang YS, Li PF, Wang ZQ, Zhu QM. Survey on Multimodal Information Extraction Research. Ruan Jian Xue Bao/Journal of Software, 2025, 36(4): 1665–1691 (in Chinese). http://www.jos.org.cn/1000-9825/7245.htm

Survey on Multimodal Information Extraction Research

WANG Yong-Sheng, LI Pei-Feng, WANG Zhong-Qing, ZHU Qiao-Ming

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Multimodal information extraction is a task to extract structured knowledge from unstructured or semi-structured multimodal data (such as text and images). It includes multimodal named entity recognition, multimodal relation extraction, and multimodal event extraction. This study analyzes multimodal information extraction tasks and summarizes the common part of the above three subtasks, i.e., a multimodal representation and fusion module. Moreover, it sorts out the commonly used datasets and mainstream research methods of the above three subtasks. Finally, it outlines research trends in multimodal information extraction and analyzes the existing problems and challenges in this field to provide a reference for future research.

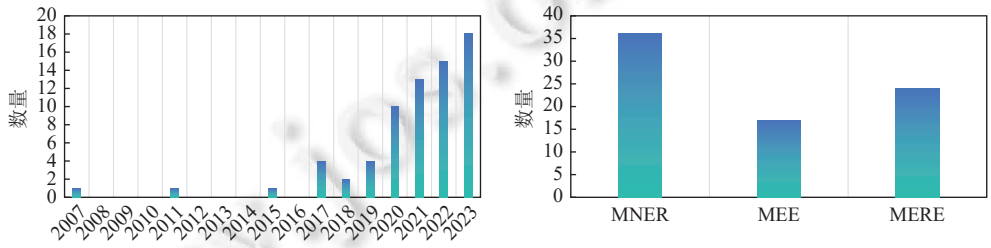
Key words: multimodal information extraction (MIE); multimodal named entity recognition (MNER); multimodal entity relation extraction (MERE)

我们生活在一个多模态信息相互交融的环境中, 每天看到的文字、图片以及听到的声音等都属于一种模态. 尤其是随着社交媒体 (如 X 和微博等) 的快速发展, 近年来社交媒体的内容往往是以文本、图片和音频等多模态的形式联合表示. 信息抽取 (information extraction) 旨在从自然语言文本中抽取出特定实体 (entity)、关系 (relation) 和事件 (event) 等信息, 帮助人们将海量的内容自动分类、提取和重构. 其中, 社交媒体上的文本呈现出简短、包含特殊字符、表达偏口语化和未收录的网络流行词语爆发等特点. 针对这样具有高噪音的文本语料, 传统的基于文本信息抽取的模型面临巨大的挑战 (如未能正确识别话题中所有实体、实体关系和事件触发词等). 一种方法是在模型的输入端增加与文本相关的其他模态信息, 从而增强文本的语义表示, 然后利用基于多模态的方法来提高信息抽取的性能. 这种从非结构化或半结构化的多模态数据 (包含文本和图像等) 中提取结构化知识的

* 基金项目: 国家自然科学基金 (62276177, 61836007); 江苏高校优势学科建设工程项目
收稿时间: 2023-09-13; 修改时间: 2024-02-25, 2024-04-16; 采用时间: 2024-06-26; jos 在线出版时间: 2024-12-09
CNKI 网络首发时间: 2024-12-09

任务被称为多模态信息抽取任务. 目前, 随着多模态数据集的逐步开放和 GPU (graphics processing unit) 算力的大幅提升, 多模态信息抽取已经成为自然语言处理 (natural language processing, NLP) 领域中一个新兴的重要研究方向.

基于上述背景, 本文首先以多模态信息抽取 (multimodal information extraction, MIE)、多模态命名实体识别 (multimodal named entity recognition, MNER)、多模态实体关系抽取 (multimodal entity relation extraction, MERE) 和多模态事件抽取 (multimodal event extraction, MEE) 为主题检索包括中国计算机学会认定的自然语言处理、人工智能和多媒体领域的 A、B、C 类国内外会议论文以及国内外重要期刊论文和高引用率的论文, 然后将检索出的多模态信息抽取相关论文分别按年份和子任务排列, 排列分布如图 1 所示. 由图 1 可以看出: 多模态信息抽取任务发展的时间较短, 多数相关任务于 2020 年前后被提出, 随后研究热度呈现出逐年快速上升的趋势; 多模态信息抽取任务主要的研究工作集中在多模态命名实体识别和多模态实体关系抽取任务上, 其他相关任务 (如多模态事件抽取等) 还处于起步发展阶段. 目前, 多模态研究的其他领域 (如多模态情感分析[1]和视觉问答[2]等) 已有多篇综述, 而多模态信息抽取还缺乏相应的综述工作. 因此, 本文尝试从 NLP 的角度对多模态信息抽取进行归纳总结并分析其存在的挑战, 为后续研究提供参考. 鉴于现有工作主要集中在图片和文本这两种模态上, 本文讨论的多模态主要涉及图片和文本.



(a) 研究热度 (b) 子任务分布

图 1 多模态信息抽取的研究热度及其子任务分布

1 多模态信息抽取任务

多模态信息抽取任务是指从非结构化或半结构化的多模态数据 (包含文本和图像等) 中提取结构化知识. 该任务是知识图谱和自动问答等任务的上游任务, 应用广泛. 由图 1 可以看出, 最早在 2007 年就有学者提出多模态信息抽取任务, 但是受制于有限的计算资源以及低效的特征提取等条件, 该方向没有得到足够的关注. 直到 2017 年, 该方向的研究热度逐步显现, 首先是 Zhang 等人[3]提出通过图片信息来提升文本事件抽取的性能以及 Lu 等人[4]提出通过图片信息来提高命名实体识别的性能. 此时, 得益于机器学习以及深度学习等技术的快速发展, 研究者们有更多的工具来表示文本特征和图片特征. 但是依然面临模态间的对齐和融合问题, 早期的对齐只是简单地通过余弦相似度来衡量[5], 融合方式也只是简单地将两种特征表示直接拼接[6,7]. 随着注意力机制以及图模型等技术的兴起, 各种高效的融合策略[8,9]以及联合抽取方法[10]相继被提出. 模型性能得到进一步提高. 随后, 研究者们注意到现有的多模态数据集存在规模小、标注成本高等特点, 各种基于小样本的方法[11,12]、对比学习的方法[13,14]被提出. 此时, 基于现有规模的多模态数据集、先进的特征表示方法以及高效的融合策略, 多模态信息抽取任务的研究取得了巨大的进展[15−17].

从研究内容来看, 多模态信息抽取任务一般包含多模态命名实体识别、多模态实体关系抽取和多模态事件抽取等子任务. 本节重点围绕上述 3 个子任务展开讨论, 包括各个子任务的定义以及模型评价指标的定义.

1.1 多模态命名实体识别

命名实体识别是指识别出文本中具有特定意义的实体, 并将这些实体按预先定义的实体类型 (如人名、地名、

机构名、事件、货币和百分比等) 进行正确分类[5]. 虽然命名实体识别任务在大多数数据集上已经取得了较大成功[18,19], 但在短文本上定位和分类其中的实体仍然存在挑战. 如图 2 所示: 基于文本的命名实体识别模型很难判断“I love Alibaba”中的“Alibaba”是一个人名还是公司组织名称, 但是根据文本附随的图片可以很容易判断“Alibaba”是一个人名, 其实体类型为“Person”.

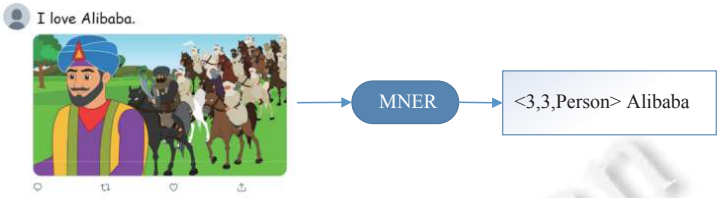


图 2 MNER 任务

MNER 旨在通过输入端引入与文本相关的其他模态信息作为文本态的补充, 从而提高文本实体识别的性能. MNER 是多模态信息抽取任务中一个重要的子任务, 近年来引起了研究者们的广泛关注. 通常, 将 MNER 定义为一个序列标注任务: 给定文本序列 $S=\{w_1,w_2,...,w_n\}$, 文本附随的图片序列 $M=\{p_1,p_2,...,p_n\}$ 和预先定义的实体类型 T, MNER 模型输出三元组列表 $<E_s,E_d,t>$, 每一个三元组都包含一个实体 e 的信息. 其中, w_i 表示一个句子; p_i 表示一张文本附随的图片; $E_s\in[1,n]$, $E_d\in[1,n]$, 分别表示实体 e 的开始位置索引和结束位置索引; $t\in T$ 表示实体 e 的实体类型[20]. 图 2 为 MNER 的一个样例.

1.2 多模态实体关系抽取

和命名实体识别任务类似, 虽然实体关系抽取任务在大多数数据集上取得了较大成功[10,21], 但在短文本上判断两个实体的关系仍然存在挑战. 如图 3 所示: 如果仅仅给出文本序列“Tobey doesn’t like eating with Leonardo.”, 基于文本的实体关系抽取模型可能将“Tobey”和“Leonardo”的关系识别为“family”, 从而导致实体关系抽取的结果错误. 而根据随文图片中出现的“警服”和“警帽”等信息, 结合文本很容易得出“Tobey”和“Leonardo”的关系为“colleague”.

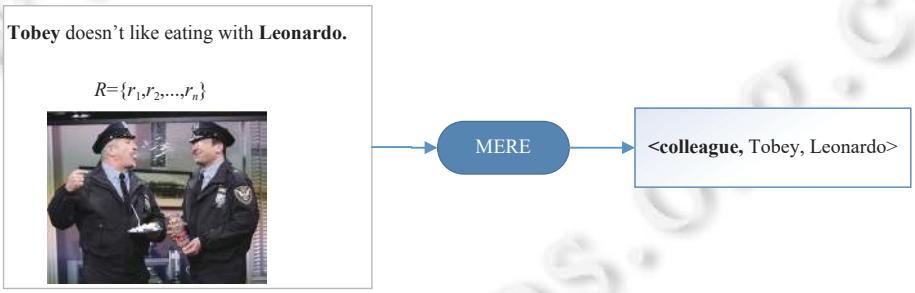


图 3 MERE 任务

MERE[16] 是指在输入端引入图片的情况下, 抽取文本中两个实体之间的语义关系. 给定一个序列 $<e_1,e_2,S,M>$ 和预先定义的实体关系类别集合 R, MERE 的目标是预测两个实体 e_1 和 e_2 之间的关系 r. 其中, $S=\{w_1,w_2,...,w_n\}$, $M=\{p_1,p_2,...,p_n\}$ 分别为给定文本序列和随文的图片序列; $r\in R$. 图 3 为 MERE 任务的一个样例.

1.3 多模态事件抽取

文本事件抽取任务是指从非结构化的自然语言文本中自动抽取用户感兴趣的事件信息并将其以结构化的形式表示[22]. 具体来讲, 是通过模型识别出事件触发词、事件类型和事件元素, 并给这些事件元素分配角色. 事件抽取同样面临诸多挑战, 如图 4 所示: 如果仅给定文本序列“Ford was in his rush to confront members in Toronto”, 由于文本较短, 上下文缺乏证据支持此处的“confront”是“攻击”还是“开会”的意思, 不同含义可能对应着不同的事件类

型以及不同的论元, 从而导致事件抽取的结果错误. 但是, 结合文本附随的图片中出现“圆形会议桌”和“笔记本电脑”等实体信息, 很容易判断文中的“confront”是“开会”的意思.

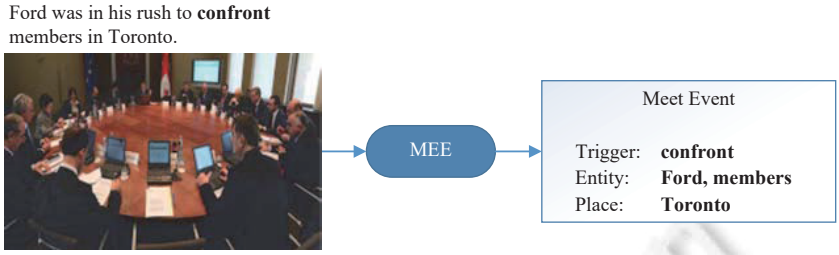


图 4 MEE 任务

多模态事件抽取 MEE 的定义为: 给定文本序列 $S=\{w_1,w_2,...,w_n\}$ 和文本附随的图片序列 $M=\{p_1,p_2,...,p_n\}$, 目标为抽取一个多模态事件触发词集合 $V=\{v_1,v_2,...,v_n\}$ 和一个多模态事件论元集合 $A=\{a_1,a_2,...,a_n\}$. 其中, 每个事件 v_i 可表示为 $v=(y_v,\{g,p\})$, g 和 p 分别为文本触发词和图片事件提及, y_v 表示事件 v 的事件类型. 若 g 和 p 同时存在, 则表示图片事件和文本事件指向同一事件, 此时, 定义该事件为一个多模态事件 (multimodal event); 若只存在 g, 则定义该事件为仅文本事件 (text-only event); 若只存在 p, 则定义该事件为仅图片事件 (image-only event). 相应地, 每个论元 a_i 可表示为 $a=(y_a,\{u,o\})$, 其中 u 表示文本实体, o 表示图片中的目标实体 (一般用矩形框在图片中标出), y_a 表示论元 a 的语义角色. 若图片事件和文本事件指向同一事件, 则合并它们对应的论元为一个多模态事件论元 (multimodal argument); 否则, 分开表示它们对应的论元[13]. 图 4 为 MEE 任务的一个样例.

1.4 评价指标

针对多模态信息抽取任务, 模型性能的主要评估指标包括: 正确率 (accuracy, *Acc*)、准确率 (precision, *Pre*)、召回率 (recall, *Rec*) 以及 $F1$ 值[23]. 正确率表示在预测结果中所有预测正确的样本占总样本的比值, 即:

$$Acc=\frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

其中, TP 表示将正例预测为正的样本数量, TN 表示将负例预测为负的样本数量, FP 表示将负例预测为正的样本数量, FN 表示将正例预测为负的样本数量.

准确率表示将样本中正例预测为正的样本数量占所有样本预测为正的样本数量的比值, 即:

$$Pre=\frac{TP}{TP+FP} \tag{2}$$

召回率表示将样本中正例预测为正的样本数量占所有真实值为正的样本数量的比例, 即:

$$Rec=\frac{TP}{TP+FN} \tag{3}$$

$F1$ 值表示综合考虑准确率和召回率后的数值, 即:

$$F1=\frac{2Pre\times Rec}{Pre+Rec} \tag{4}$$

一般情况下, $F1$ 值可以从宏观上直接评价多模态信息抽取模型性能的优劣, 本文后面的章节中也有多处是直接使用 $F1$ 值来评价模型性能的优劣.

2 多模态表示和融合

在多数多模态相关任务中, 多模态表示和融合都是建立模型的关键步骤之一. 多模态表示通常包含文本特征表示和图片特征表示. 文本特征表示关注如何抽取文本才能获得更好的文本特征, 图片特征表示关注如何抽取图片才能获得更好的图片特征. 多模态融合即通过融合策略将 2 个 (或多个) 模态特征表示整合为 1 个多模态特征表示, 多模态融合关注如何通过其他模态 (如图片) 来增强文本的语义表示以便获得更好的模型性能. 本节将从多

模态信息抽的 3 个子任务所共有的多模态表示 (包含文本特征表示和图片特征表示) 和多模态融合这两个方面进行分析.

2.1 文本特征表示 (字符级和单词级)

文本特征表示通常包含基于卷积神经网络的方法 (convolutional neural network, CNN)、基于循环神经网络的方法 (recurrent neural network, RNN)、基于语法依存树的方法 (syntax dependency tree) 以及基于 Transformer 编码端的方法 (bidirectional encoder representations from Transformers, BERT) 等. 基于卷积神经网络的方法可以获取单词 (或字符) 的局部结构特征; 基于循环神经网络的方法能够获取文本的顺序信息; 基于语法依存树的方法可以建模复杂的语义表示; BERT 方法可以有效捕获文本上下文信息. 由于不同的文本特征表示方法侧重点不同, 可能对整个句子的理解有不同影响, 本节主要围绕上述几种不同的特征表示方法展开分析.

● 基于 CNN 的方法. 早期, 基于 CNN 的方法主要被应用在图像处理领域, 近年来不断有学者尝试将该方法引入文本领域[9,24]. 它可通过卷积操作来捕获句子中的局部特征, 在识别关键词和短语等方面表现出色. 考虑到多模态信息抽取任务中, 数据集中的文本多数直接来源于网络, 经常会有拼写错误、包含未登录词 (out-of-vocabulary, OOV) 以及不正常的大小写等特点, Chen 等人[9]引入字符级编码 (character-level, 即将单词看成一个序列并将单词中的每个字符看成序列中的元素), 然后使用 CNN 来提取字符的特征向量. 通过该方法可以提取单词的局部信息 (如前缀、后缀以及大写等结构特征), 为理解存在拼写错误、OOV 以及不正常的大小写等特点的句子提供帮助.

● 基于 RNN 的方法. 虽然上述方法在某些需要捕获局部特征的任务中表现出色, 但是通过该方法无法捕获文本的位置信息. 因此, Lu 等人[4]引入基于 RNN 的 LSTM (long short-term memory) 来编码文本特征, 采用 LSTM 的编码方式可以捕获句子中的顺序信息. 此外, LSTM 通过引入门控还可以缓解梯度爆炸和梯度消失问题. 但是上述方法只能考虑单词之前的信息, 即第 i 个词所获得的信息均来自第 i 个词以及之前的信息, 无法获得第 i 个词之后的信息. 因此, Wu 等人[25]使用双向 LSTM (bi-directional long short-term memory, Bi-LSTM) 来编码文本特征, 其包含一个前向 LSTM 和一个后向 LSTM, 将双向 LSTM 的输出拼接得到一个新的特征表示 $[\overrightarrow{h_t},\overleftarrow{h_t}]$. 进一步地, 为了同时捕获句子的字符特征以及文本位置特征, Moon 等人[6]将 Bi-LSTM 方法得到的词级别特征表示与基于 CNN 方法得到的字符级特征表示拼接在一起得到最终的词向量表示.

● 基于 Transformer 编码端的方法. 随着预训练技术的快速发展, 不断有学者开始使用 BERT 来提取文本特征[26−29], 如图 5 所示: 分别在给定文本的开始位置和结束位置添加特殊标记“[CLS]”和“[SEP]”, 然后将添加特殊标记后的文本输入到 BERT 编码器中, 输出得到文本的特征表示. BERT 模型在预训练阶段已经学习到了大量知识, 可以有效地捕获文本上下文信息, 因此, 基于 BERT 的方法在各下游任务中均有着出色的表现. 这种文本特征表示方法在 NLP 各任务中被广泛应用[30].

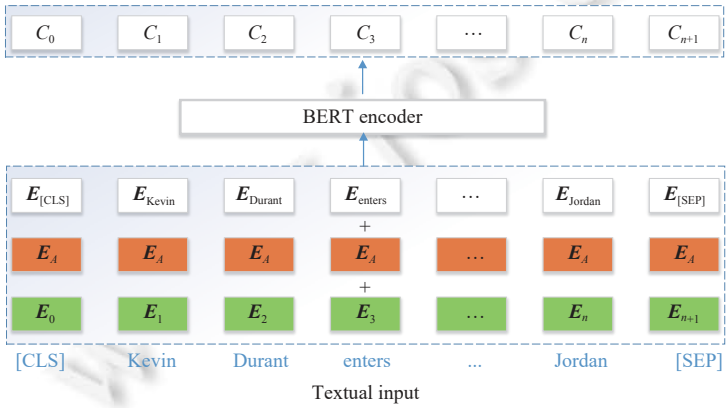


图 5 文本编码器 BERT 的一个样例[26]

● 基于语法依存树的方法. 针对复杂场景 (如实体嵌套), Li 等人[13]使用基于语法依存树的方法提取文本特征. 它将句子的语义结构以树或图的方式呈现, 树 (或图) 的节点表示实体, 其连接的边表示实体之间的关系. 该方法不仅仅关注表层的语法结构, 还关注句子深层的语义结构, 能够帮助模型更好地理解句子含义. 上述各种文本特征表示方法的优缺点比较见表 1.

表 1 各种文本特征表示方法的优缺点比较

文本特征表示	优点	缺点	多模态信息抽取领域的代表性模型
基于CNN的方法	可有效地捕获句子中的局部特征, 方便处理多通道输入(如词向量和字符向量等)	可能会忽略全局上下文信息	文献[9,24]
基于RNN的方法	可捕获句子中的顺序信息以及短期和长期依赖关系	无法并行计算, 计算效率低	OCSGA[25], 文献[4,7]
基于Transformer编码端的方法	可有效地捕获文本上下文信息	无法完全解决长文本处理问题, 依赖大量数据	UMT-BERT-CRF[26], ITA[27], MAF[28]
基于语法依存树的方法	语义表达丰富, 跨语言通用性, 适应多领域应用	复杂且难以自动解析, 存在多义词问题及规模限制等	MEGA[31], WASE[13]

2.2 图片特征表示

多模态信息抽取任务通常是利用图片信息来增强文本的语义表示, 以达到提高信息抽取性能的目的. 然而, 不同粒度的图片特征表示可能对文本的语义增强有不同的影响. 早期的研究主要通过 CNN 的方法以整图方式提取图片特征 (image feature), 但是图片中往往只有部分区域对理解文本有帮助, 以整图方式提取图片特征可能会引入图片噪声, 而且无法获取图片中各实体之间的关系. 随着图像分割和目标检测等技术的发展, 将图片切分为区域的特征表示 (regional area)、基于目标检测工具提取目标实体的特征表示 (object-level feature) 以及基于情境图的方法 (situation graph) 构建树或图结构的特征表示等细粒度方法相继被提出.

● 整图特征表示. 将图片直接输入 CNN 可得到整图的特征表示. 该表示方法可获取图片的全局信息. Lu 等人[4]将图片直接输入 ResNet 中, 然后从其全连接层的前一层输出和最后一个卷积层输出分别得到整个图片特征表示和 7×7 个区域图片特征表示, 然后采用同样的融合策略分别将得到的两种图片特征与文本特征融合, 最后将融合后的多模态特征输入任务模型中. 通过在 Snap Captions 和 Twitter[4]数据集上的实验表明: 与融合区域图片特征表示相比, 融合整图特征表示的模型性能较差. 可能的原因是在融合整图特征时, 引入了图片中不相关的噪声信息.

● 区域特征表示. Lu 等人[4]的实验结果表明: 整图特征表示容易引入与文本不相关的噪声信息, 最终影响模型的性能. 因此, 有学者在提取图片特征时, 首先将整个图片划分为若干区域, 然后针对文本与这些区域图片之间的关系进行建模[6,7,9,13,24,26,28,29,32,33], 这样的好处是在融合时可通过注意力机制只关注与文本相关的区域图片, 过滤掉不相关的区域信息. 其中, Moon 等人[6]使用 GoogLeNet 变体的最后一个隐藏层来表示图片特征; 也有学者使用 16 层 VGGNet 的最后一个池化层得到 7×7 个区域图片特征[7,9,13,24]; 此外, 还有部分学者使用 ResNet 最后一个卷积层得到 7×7 个区域图片特征[26,28,29,32,33].

● 基于目标实体的特征表示. 虽然通过区域特征表示可以在一定程度上减少噪声, 但是在将整图均匀地切分为几个区域图片特征的同时, 可能会将图片中某 1 个或多个目标实体切分开, 破坏了目标实体的整体语义信息. 此外, 依据不同尺度对图片进行区域划分可能对模型性能也有不同的影响. 为了缓解上述问题, 部分学者采用目标检测等技术抽取图片特征时, 仅保留图片中的目标视觉区域, 提出目标级图片特征表示 (object-level feature)[8,15,25,27,34]. 如图 6 所示: 将给定图片输入到目标检测工具 (如 MASK RCNN), 输出图片中实体可能的目标类别以及每个目标类别对应的概率, 最大概率对应的类别即为识别出的目标类别. 该方法的好处是: 一方面, 在融合时, 可通过直接编码图片中各个目标实体的标签得到图片的语义信息, 这样就将图片特征映射到文本空间, 可实现与文本特征的无缝连接; 另一方面, 这种目标级图片特征可与文本实体进行显性对齐, 对多模态事件抽取等复杂任务有积极影响. 其中, 多数工作是直接使用目标检测工具得到目标级图片特征[15,25,34]. 进一步地, Wang 等人[27]通过 3 个辅助任务 (目标检测、图片摘要生成和光学字符识别) 将图片分别生成 3 个文本特征表示, 然后将得到的 3 个文本特征表示

直接连接得到一个图片映射到文本空间的特征表示. 此外, 受 Li 等人[35]启发, MRC (machine reading comprehension) 具备稳定的语言理解能力, 因此, Jia 等人[8] 引入 MRC 框架, 首先为每个文本实体设计了合适的查询规则, 然后基于给定查询使用视觉定位工具得到图片中 Top-k 个候选目标特征. 因为每个文本实体类型对应的查询经过特殊设计, 相较于文献 [15], 此处得到的候选目标特征包含更多的先验知识.

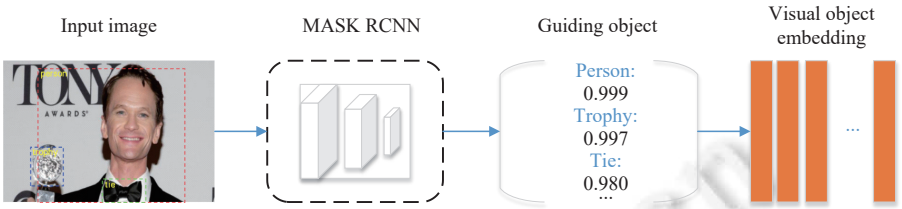


图 6 目标检测工具 MASK RCNN 的一个样例[25]

● 基于情境图的特征表示. 上述图片特征表示方法在动作识别和目标识别等简单任务场景上性能表现较好. 但是对于多模态事件抽取等复杂任务, 常常涉及实体嵌套和图文事件同指等复杂场景, 基于情境图的方法可同时识别图片中的目标实体以及实体之间的关系, 对于理解图片中的整体场景很有帮助. 因此, 有研究工作是将基于目标检测构建的情景图作为图片的特征表示[13,31]. 其中, Li 等人[13]首先采用 VGG-16[36]检测出图片中所有实体作为候选论元, 并采用 MLP (multi-layer perceptron) 预测一个动词编码和所有实体的名词编码, 然后通过比较 imSitu 数据集[37]的所有动词和名词, 对动词编码和名词编码进行分类, 最终构建成中间节点为动词, 邻居节点分别为实体类型和论元角色的基于目标实体的结构图. Zheng 等人[31]也采用类似的流程构建结构图, 最终通过上述方法在任务模型上均取得了较好的性能. 上述各种图片特征表示方法的优缺点比较如表 2 所示.

表 2 各种图片特征表示方法的优缺点比较

视觉特征表示	优点	缺点	多模态信息抽取领域的代表性模型
整图特征表示	可捕获图片的全局信息, 计算效率高	容易引入图片中与文本实体不相关的噪声	文献[4]
区域特征表示	可捕获图片中不同区域的局部特征, 可缓解噪声问题	特征质量依赖区域选择, 能够捕捉到的区域上下文关系有限	UMT-BERT-CRF[26], MAF[28]
基于目标实体的特征表示	可进一步明确图片中目标实体的语义信息, 缓解噪声问题	复杂场景中目标实体识别困难, 且难以捕捉目标实体之间的关系	OCSGA[25], ITA[27]
基于情境图是特征表示	可同时识别图片中的实体及实体之间的关系, 从而建模复杂场景	构建该图难度高, 依赖物体检测和关系推断的准确性	MEGA[31], WASE[13]

2.3 多模态融合

多模态的特征融合也是多模态信息处理的关键步骤之一[5]. 多模态融合是指将多个模态特征表示整合成为一个多模态特征表示. 在多模态信息抽取任务的预测过程中, 单个模态通常不能包含产生精确预测结果所需的全部有效信息, 而多模态交互过程融合了来自两个或多个模态的信息, 可以提升模型预测结果的精度以及模型的鲁棒性[38]. 早期的前融合方法包含将两种模态特征直接相加以及点乘等, 虽然这种融合方法简单易实现, 但是该融合策略中由于两种模态间的交互不充分导致任务模型的性能提升非常有限, 甚至没有提升的效果, 因此, 该方法常常被用来作为多模态模型的基准. 目前主流的工作是通过基于神经网络的方法来实现两种模态的融合, 主要包含: 基于注意力和门控机制的多模态融合、基于图模型的多模态融合以及基于多模态预训练模型的多模态融合方式.

● 图文特征拼接的多模态融合. 通过上述各类特征表示方法可分别得到文本的特征表示和图片的特征表示, 一类最容易考虑到的融合方法便是通过直接拼接的方法将两种特征表示进行融合, 在早期的文献中, 多数研究者通过直接拼接的方法来融合图文特征, 并将该多模态特征得到的结果作为多模态方法的基准. 与纯文本相比, 融合后的多模态特征可在一定程度上弥补短文本缺失的内容. 但正是由于此类方法简单易实现, 这种粗暴的融合方式

导致图文之间的信息交互不充分, 进而导致与单模态模型相比, 多模态模型的性能提升有限, 甚至在某些指标上略差于单模态模型[4,6,7].

● 基于注意力和门控机制的多模态融合. 为了缓解通过上述图文拼接的多模态融合方式导致的图文交互不充分问题, 一些研究者开始使用跨模态注意力机制, 使得融合的过程中模型更关注重要的信息[4,6−9,17,24,26]. 具体来讲, Lu 等人[4]提出了基于视觉注意力模型. 该模型可以通过文本引导的视觉注意力来决定图片中的哪一块区域与文本 h_t 交互. 模型可关注到图片中的重要区域. 但是上述模型仍然不知道文本中哪些词与 h_t 最相关, 因此, Zhang 等人[7]进一步提出了自适应共注意力网络 (adaptive co-attention network), 继续通过图片引导的文本注意力来决定文本中哪些词参与图片的交互, 通过视觉注意力和文本注意力共同捕获不同模态之间的语义交互. 此外, Tong 等人[17]基于注意力机制提出了双循环多模态模型 DRMM (dual recurrent multimodal model). 该模型中包含 N 个交替双重注意力模块 ADA (alternating dual attention), 每个 ADA 模块内部通过注意力机制分别对文本和图片进行更新, 通过 N 个 ADA 模块不断迭代的方式对图文模态进行细粒度交互, 最终得到图片和文本融合后的多模态特征.

虽然通过上述融合策略, 任务模型在多数数据集上取得了较好的性能[6,8]. 但是, 当数据集中存在较多图文不相关的数据对时, 仅通过注意力机制的融合策略对任务模型的性能提升有限. 为了缓解上述问题, 有研究提出在注意力机制上加入门控 (gate) 机制来动态融合文本特征和图片特征, 动态融合的过程可以通过权重来控制[4,7,9,24,26,32,39,40]. 其中, Chen 等人[40]提出一个基于 Transformer 架构的分层视觉前缀融合网络 HVPNeT (hierarchical visual prefix fusion network), 当两种模态融合时, 将预训练模型 ResNet 得到的分层图像特征分别融合到 Transformer 各层中, 具体做法为: Transformer 每个层通过一个注意力模块来融合, 其各个网络层之间的融合通过动态门控机制来控制. 此外, Wang 等人[39]进一步细化 Transformer 的多头注意力层, 基于外部知识提出增强跨模态注意力框架 (refined multimodal attention). 首先通过外部知识来扩展预定义的实体类型标签库, 并通过这些实体标签来识别任务的显著性特征, 然后利用这些特征的显著性分数来增强跨模态注意力的权重, 这样的好处是让模型关注与任务高度相关的特征, 同时也实现了模态间较好的交互.

上述方法主要通过文本中的单个单词来捕获视觉注意力, 忽略了模态内部的交互, 因此, Arshad 等人[24]首先通过自注意力捕获文本态内部的对齐关系, 然后通过文本态的多个单词来引导视觉注意力. Wu 等人[25]将自注意力和引导注意力组合成一个密度共注意力模块 (dense co-attention). Yu 等人[26]组合标准 Transformer 层和跨模态注意力提出了一个统一的多模态 Transformer 框架 (unified multimodal Transformer, UMT). 通过上述方法可实现模态内部和模态间的交互, 使得跨模态融合更充分.

● 基于图模型的多模态融合. 区别于上述基于注意力和门控机制的融合方法, 图模型可在复杂的异构数据上建模且具有一定的关系推理能力和可解释性. Zhang 等人[15]在多模态命实体识别任务中提出了基于图模型的多模态融合方法. 其中, 文本实体 (或者图片中的目标实体) 作为图模型中的节点, 文本实体 (或者图片中的目标实体) 之间的内部关系以及文本实体和图片中的目标实体之间的外部关系作为图模型的边, 通过图模型的边将各个节点连接来实现多模态的语义交互. 具体地, 为了捕获相同模态的上下文关系, 研究人员将任何相同模态的节点都通过模态内的边 (intra-modal edge) 连接; 同时, 为了捕获模态间的关系, 研究人员做出如下规则: 如果图片中的某个目标实体与文本中的名词短语存在强对应关系, 则将该目标实体节点与对应的文本实体节点 (名词短语) 通过模态间的边 (inter-modal edge) 连接; 否则, 通过模态间的边将该目标实体节点与所有文本实体节点连接. 通过上述方法构建了一个完整的图模型, 同时实现了模态内和模态间的交互. 上述方法中, 虽然在模态内部之间建立了上下文关系, 但是缺乏针对性. 为了解决上述问题, Zhao 等人[33]首先通过以下规则建立模态内各实体之间的关系: 若图片中包含相同的目标实体类型, 则两个目标实体之间通过边来连接; 否则, 两个目标实体之间不连接. 然后通过图文摘要描述 (image-caption) 模块将图片转为对应的文本描述, 比较与文本之间的相似度来建立文本内各实体之间的关系, 最终通过上述方法构建的图模型实现了模态间和模态内的交互. 此外, Zheng 等人[31]提出了一个基于双图对齐的多模态神经网络方法 MEGA (multimodal neural network with efficient graph alignment). 该方法首先通过依存树工具和预训练的场景图模型分别构造文本表示图和图片表示图, 然后通过图对齐方法分别从结构相似度和语义相

似度来实现图片和文本之间的关系映射, 这样的好处是可以找到图片中的目标实体与文本中实体的相关性, 使得文本和图片之间得到更深的交互.

● 基于多模态预训练模型的多模态融合. 随着对比学习、强化学习等技术的快速发展, 研究者们通过大量数据可以训练出一个强大的二分类器[29,41], 该分类器可将多模态数据集分为相关的图文数据和不相关的图文数据, 相关的图文数据输入到先进的多模态模型中, 不相关的数据输入先进的单模态数据中. 这类方法与上述门控机制相比, 模型的性能得到进一步提升.

上述各种融合方法的优缺点比较如表 3 所示.

表 3 各种融合方法的优缺点比较

融合方法	优点	缺点	多模态信息抽取领域的代表性模型
图文特征拼接的多模态融合	可在一定程度上弥补短文本中缺失的内容; 方法简单, 易实现	图文模态间交互不充分	Bi-LSTM-CRF + Global Image Vector[4], Bi-LSTM-CRF + Bi-CharLSTM + Inception[6]
基于注意力和门控机制的多模态融合	使得模型更关注不同模态中最重要的信息	模态间数据不平衡以及模态间不相关时性能较差	UMT-BERT-CRF[26], 文献[7], RIVA[32]
基于图模型的多模态融合	可在复杂的异构数据上建模且具有一定的关系推理能力和可解释性	通过不同方法构建的图模型表现出不同的性能, 构建图模型具有较高复杂性	UMGF[15], MEGA[31], R-GCN (w/o Gate)[33]
基于多模态预训练模型的多模态融合	可缓解图文数据不相关进而引入噪声的影响	需要大量数据以及计算资源	RpBERT[29], 文献[41]

3 多模态命名实体识别

3.1 常用数据集

为了推动多模态命名实体识别任务的发展, Zhang 等人[7]首先通过 Twitter API 收集了 2014–2015 年间的 2650 万条推特数据, 然后过滤掉其中的非英文以及不包含对应图片的推特数据, 初次筛选后共保留 430 万条数据. 由于这些推文与个人习惯和爱好强相关, 为了降低人为特异性, 研究人员进一步从初筛数据中随机选取 5 万条推特数据, 并安排两名标注人员使用 BIO 方法[42]独立对其进行标注, 标注实体的类型包含 Person、Location、Organization 和 Misc 这 4 类. 为了进一步提高数据集的质量, 与 CoNLL-2003 数据集类似, 研究人员再次过滤数据集中不包含任何命名实体、其 token 的长度小于 3 以及很难理解其含义的句子, 最终得到由 2116 名用户发布的 8257 个句子, 实体总数为 12800.

同年, Lu 等人[4]通过 Twitter API 收集了 2016–2017 年间以“运动“和“社交”为主题的多模态推特数据. 和 Twitter2015 类似, 该数据集由 3 名标注人员也是使用 BIO 方法单独标注, 同样也是包含 4 种命名实体的类型, 分别为: Person、Location、Organization 和 Misc. 为了保证数据集的高质量, 过滤掉不包含图片的推特数据, 并要求剩下的每条数据包含一个句子和图像对; 其中, 若一个句子与多个图片相关, 则随机选取其中一张图片. 最终得到的数据集共有 7181 个句子, 平均每个句子包含 16 个 token.

考虑到文献 [4] 数据集中的小部分数据涉及用户隐私, 该部分数据已被删除; 同时为了区分文献 [7] 和文献 [4] 中的两个数据集, Yu 等人[26]进一步整理这两个数据集, 并分别命名这两个数据集为 Twitter2015 和 Twitter2017. 通过比较上述数据集发现: 两个数据集都是通过 Twitter API 收集得到; Twitter2017 比 Twitter2015 的规模小; 在内容上, 两个数据集也有所不同, Twitter2017 更偏向运动和社交的主题. Twitter2015 和 Twitter2017 统计信息如表 4 所示, 其中, 各模型在两个数据集上的 SOTA 性能来源于文献 [43].

此外, Lu 等人[4]还提出了图文数据集 Snapchat; Ji 等人[44]提出了中文多模态数据集 CMNER; Wang 等人[45]提出了一个细粒度的多模态命名实体识别任务 FMNERG (fine-grained multimodal named entity recognition and grounding), 并提出了针对该任务的数据集 Twitter-FMNERG; Sui 等人[46]提出了包含中文文本和语音的多模态数据集 CNERTA. 由于后续研究使用这些数据集较少, 本文不作详细展开.

表 4 Twitter2015 和 Twitter2017 统计信息

实体类型	Twitter2015			Twitter2017		
	Train	Dev	Test	Train	Dev	Test
Person	2217	552	1816	2943	626	621
Location	2091	522	1697	731	173	178
Organization	928	247	839	1674	375	395
Misc	940	225	726	701	150	157
数据集中实体类型总数	6176	1546	5078	6049	1324	1351
数据集中的句子总数	4000	1000	3257	3373	723	723
数据集类型	推文			运动、社交等主题推文		
创建数据集的年份	2018			2018		
SOTA性能 (*F*1值) (%)	76.79[43]			89.58[43]		
数据集链接 (处理后)[26]	https://github.com/jefferyYu/UMT					

3.2 主流方法总结

针对 NER 任务的方法通常包含: 基于规则和模板的方法、基于统计机器学习的方法和基于深度学习的方法. 随着基于规则和模板的方法存在规则制定成本高、规则无法普适到其他语料以及基于统计机器学习的方法存在需要大量手工特征工程等缺点越来越突出, 前两种方法逐步被抛弃, 基于深度学习的方法成为后续研究的主流方法. MNER 任务是在 2019 年逐步发展起来的, 因此深度学习也是解决 MNER 任务的主流方法. 本节主要围绕针对 MNER 任务的深度学习方法展开论述, 其研究框架主要分为基于序列标注的方法、基于跨度分类的方法以及基于大模型直接生成的方法这 3 类.

3.2.1 基于序列标注 (token-based) 的方法

基于序列标注的方法是 MNER 任务的主流方法之一, 该方法流程上一般分为 3 个部分[23]: 输入分布式表示层、多模态编码层和标签序列解码层. 其中, 第 1 层通常为单词 (或字符) 和图片信息; 第 2 层通常为多模态特征, 使用 RNN、CNN 和 Transformer 等神经网络技术先分别提取文本特征和图片特征, 然后采用各种融合策略将文本特征和图片特征融合为多模态特征; 最后一层为解码层, 输入多模态特征到解码层即可获得与任务对应的分类标签, 解码层以 CRF (conditional random field) 最为常见. 目前的主流工作一般采用 Deep Learning+CRF 的模型结构 (即在深层神经网络后接入 CRF 层). 针对 MNER 任务的深度学习方法主要包含基于图结构的方法、基于注意力机制的方法和基于多模态预训练模型的方法等.

● 基于图结构的方法

区别于链式结构, 基于图结构的神经网络可以通过各个节点的连接建模模态间以及模态内之间的关系, 为融合文本特征和图片特征提供天然优势. 而包含融合策略的多模态编码层是序列标注模型的关键步骤之一, 因此, 近年来不断有学者提出基于图结构的方法来解决 MNER 任务[15,33]. 其中, Zhang 等人[15]提出了一个统一的多模态图融合方法 UMGF (unified multi-modal graph fusion), 首先通过图模型的各个节点来表示输入的文本和图片信息, 并通过各个节点之间的联系来捕获它们之间的语义关系, 然后通过堆叠多个基于图模型的多模态融合层, 迭代的执行语义交互以便学习各节点的特征表示, 通过上述步骤得到一个基于注意力的多模态表示, 最终将该多模态表示输入解码器 CRF 中得到实体标签. Zhao 等人[33]进一步细化图模型结构, 提出了 R-GCN (relation-enhanced graph convolutional network), 相同模态间的节点不再是全连接, 而是通过图片中目标实体的类型选择性的建立图片节点之间的联系, 通过比较文本与辅助模块之间的相似度来建立文本节点之间的联系, 然后通过 Transformer 层的自注意力机制分别提取图模型中的文本特征和图片特征, 接着通过跨注意力层 (cross-attention Transformer) 得到包含文本态和图片态的一个多模态表示, 最终将该多模态表示输入 CRF 层实现实体分类.

● 基于注意力机制的方法

2017 年, 谷歌提出了 Transformer 模型. 该模型可实现并行计算, 其内部注意力结构也能够解决基于 CNN 方法存在的无法捕获长距离依赖的问题, 模型经过微调后在各个下游任务中表现出色. 因此, 近年来不断有学者基

于 Transformer 的注意力机制提出各种变体来提高 MNER 任务上的性能[4,7,9,24−28,40,47,48]. 其中, Zhang 等人[7]在 LSTM 网络层与 CRF 层之间加入一个自适应共注意力网络模块 (adaptive co-attention network). 模型中的图片注意力模块可以基于预测来捕获与 t 时间步长的单词最相关的图片区域, 文本注意力模块则可以基于预测来捕获与 t 时间步长的单词最相关的其他单词. 该模型在所提数据集上取得了较好的性能表现. 同年, Lu 等人[4]则在 LSTM 网络层之前加入视觉注意力模块 (visual attention model). 该模块可以提取图片中与输入文本最相关的区域图片特征, 忽略掉图片中噪声信息, 然后通过视觉调制门 (visual modulation gate)[49]动态地融合文本特征和区域图片特征, 最终在所提出的数据集 Twitter 上也取得了不错的性能表现. Arshad 等人[24]基于多维注意力 (multi-dimensional attention)[50]提出了一个可同时实现模态内融合和跨模态融合的端到端模型, 首先计算文本中两个词之间的对齐分数, 然后将该分数作为查询继续计算与图片中各区域的对齐分数; 以此类推, 直至提取出所有与文本相关的区域图片特征. 通过该方法提取的区域图片特征与文本相关度更高, 最终 $F1$ 值比基线[7]提高了 2.22%. Chen 等人[9]在此模型基础上进一步细化, 通过引入外部知识库查询图片的属性等知识, 然后将这些外部知识用同样的方法融合到文本特征, 模型性能得到进一步提升. Wang 等人[47]提出了一个基于多模态检索的框架 MoRe (multi-modal retrieval based framework), 首先基于维基百科 (Wikipedia) 建立一个知识库 (knowledge corpus), 然后利用文本检索器检索知识库中最相关的段落, 利用图像检索器检索包含最相关图片的描述性文档; 由于检索得到文本的长度通常很大, 所以将两部分检索结果分别作为基于文本检索模型和基于图片检索模型的输入, 通过 MOE (mixture of experts) 模块来融合两个模型的概率分布, 最后输入 CRF 层做最终的预测. 引入外部知识的好处是可为模型提供更多信息, 缺点是检索过程可能引入噪声, 导致错误传播. 这类方法对知识库的构建以及检索条件的设计要求较高.

虽然上述方法通过各种融合策略提升了 MNER 性能, 但是未考虑到图片中的目标实体与文本中实体的对应关系. 因此, Chen 等人[40]提出一个基于 Transformer 架构的分层视觉前缀融合网络 HVPNeT, 该模型同时考虑了图片的整体特征和图片中的目标实体特征; 整体特征可以表达更抽象的概念, 目标实体特征可提供更多语义知识, 模型中, 将图片特征表示视为可插拔的前缀来引导文本特征表示做出正确预测. 此外, Wu 等人[25]提出 OCSGA (object embeddings+textual representations+self-attention+guide-attention) 模型, 首先通过目标检测得到图片中各目标实体对应的文本标签, 并将该标签编码成和文本同样维度的特征向量, 然后基于密度共注意力层 (dense co-attention module) 找到与文本相关的图片目标特征, 并过滤掉不相关的图片特征, 最终将融合后的特征输入 CRF 层得到实体分类.

尽管上述方法通过引入图片信息有效缓解了实体类型多样化问题, 但是句子中往往仅有部分实体与图片中的目标对应, 导致模型过于重视图片中出现的文本实体而忽视图片中未出现的文本实体, 该现象被称为视觉偏差 (visual bias). 为了缓解上述问题, Yu 等人[26]在 Transformer 基础上进一步引入基于纯文本的辅助模块, 通过该辅助模块的一个转换矩阵引导模型最终做出正确预测. 该方法分别在两个数据集 Twitter2015 和 Twitter2017 上提供了两个强竞争力的基线, 后续多数工作都与这两个基线进行了比较. Wang 等人[27]基于 Transformer 提出了 ITA (image-text alignment) 框架, 考虑到文本表示对于 MNER 任务的重要性更大, 因此, 通过 3 个辅助任务将图片映射到文本空间, 然后与原文本连接, 作为 Transformer 编码层的输入, 这样的好处是将图文表示尽可能统一到同一语义空间, 最终通过 CRF 层实现实体分类. Lu 等人[48]提出一个基于 Transformer 的扁平多模态融合框架 FMIT (flat multi-modal interaction Transformer), 如图 7 所示. 首先利用句子中的名词短语和普通领域词来获取视觉线索, 然后将图片和文本的细粒度语义表示转换为统一的晶格结构, 并设计一种新的相对位置编码来匹配 Transformer 层的不同模态. 此外, 还引入一个实体边界检测模块作为辅助任务, 最终通过 CRF 层实现实体分类. 通过上述方法可缓解视觉偏差问题, 在 Twitter2015 和 Twitter2017 上均取得了较好的性能表现.

上述基于注意力的方法是建立在给定的图文对是相关的基础上, 当图文不相关或者图片模糊导致无法判断图文是否相关时, 基于注意力机制的模型也显得力不从心. 引入与文本不相关的图片信息会增加新的不确定性, 甚至会起到相反的作用, 而 Vempala 等人[51]统计显示: 33.8% 的推文存在图文不相关或者弱相关的情况. 基于以上原

因, Xu 等人[28]提出一个匹配对齐框架 MAF (matching and alignment framework). 该框架包含一个对齐模块 (cross-modal alignment) 和一个匹配模块 (cross-modal matching). 通过对齐模块使得两种模态的特征表示尽可能一致, 通过匹配模块计算图文相似度来决定融合时图片保留的比例. 这两个子模块在很大程度上缓解了给定图文对不相关的问题, 同时也增加了模型的鲁棒性; 与文献 [4,26] 中的方法类似, 然后通过门控机制 (gate mechanism) 来动态融合处理后的两种模态的特征表示, 最后将融合后的多模态特征输入 CRF 层得到实体分类.

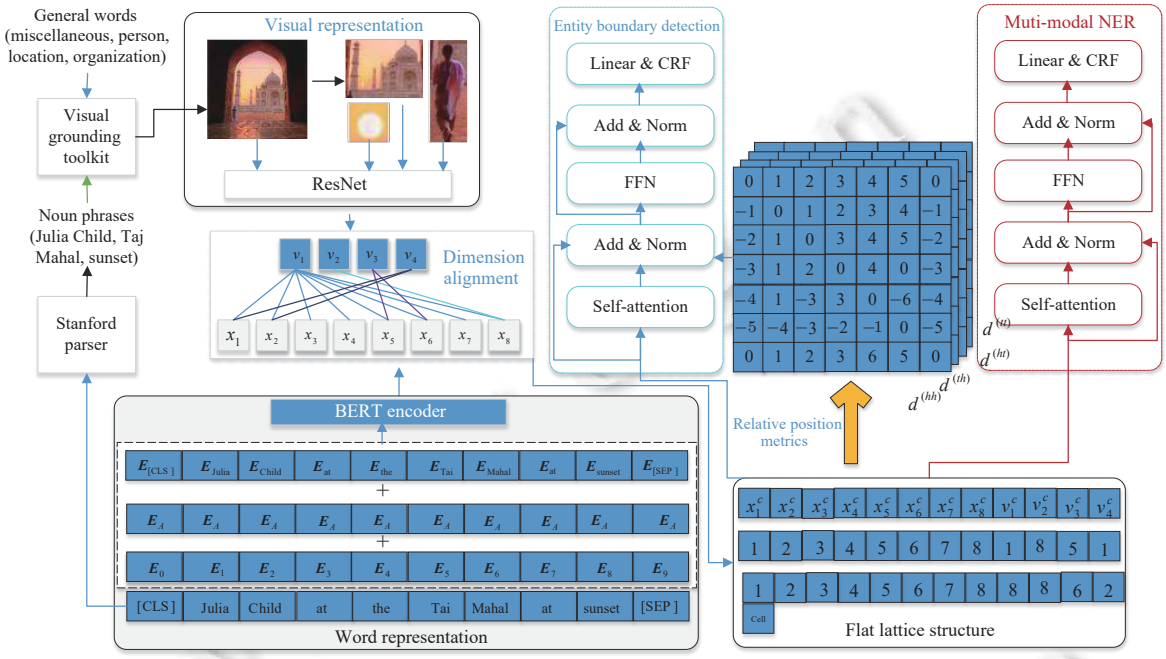


图 7 FMIT 框架图[48]

● 基于多模态预训练模型的方法

随着多模态预训练模型 (如 CLIP (contrastive language-image pre-training) 和 VL-BERT (vision-and-language BERT)) 的快速发展, 基于多模态预训练模型的这种端到端 (end to end) 的方法越来越受到研究者们的欢迎. 近年来, 不断有学者将该类方法应用到 MNER 任务上[29,32,41]. 虽然基于 Transformer 的方法[26]已取得较好的性能, 但是针对其错误结果分析显示: 22% 的实体可以用先进的 MNER 模型正确识别, 但是在最先进的文本 NER 模型上却不能被正确识别; 此外, 12% 的实体可以在仅文本的 NER 模型上正确识别, 但是却不能在最先进的 MNER 模型上识别. 上述结论表明: MNER 模型并不总是优于仅文本 NER 模型. 受此启发, Xu 等人[41]提出首先基于强化学习框架训练一个分类器, 该分类器由 CLIP[52]和 MLP (multilayer perceptron) 两部分组成, 通过分类器将图文相关的数据划分到多模态数据集, 不相关的图文数据中只保留文本态, 并将其划分到单模态数据集, 然后再将这些数据分别输入 MNER 模型和对应的文本态 NER 模型中, 最后组合两部分结果作为最终结果. 实验结果表明: 该模型性能比基线提升明显. Sun 等人[32]首先使用基于半监督的教师-学生方法学习图文对之间的关系, 以训练一个可以判断图文关系是否相关的二分类器, 分类器得到的结果作为门控单元的一个输入来决定提取的图片特征是否传入下一步的视觉-语言上下文模块, 然后通过训练两个子任务 (图文对关系预测和上下文单词预测) 得到一个基于多模态预训练模型的方法 RIVA (relation inference and visual attention), 最终在测试集上, 将 RIVA 的输出输入到 Bi-LSTM-CRF 框架[19]得到实体分类. Sun 等人[29]提出一个基于图文关系传播的 BERT 变体模型 RpBERT (relation-propagation-based BERT), 首先通过视觉语言预训练模型 VL-BERT[53]得到两种多模态的特征表示, 紧接着将其输入到一个全连接层 (作为一个二分类器) 来判断图文对的关系是否相关, 然后再通过关系传播机制得到的掩码矩阵来控制图

片信息的融合, 以融合后的多模态特征表示作为 RpBERT 的输入序列, 通过训练两个子任务 (图文关系对分类和基于关系传播的 MNER) 得到一个端到端的模型 RpBERT. 由于上述基于多模态预训练模型的方法在训练过程中已经将图文关系作为一个子任务, 因此, 对于 MNER 任务, 不再需要考虑图文关系是否匹配的情况, 这种端到端的方法为研究者提供了极大的便利.

显然, 上述多模态预训练模型需要大量多模态数据和强大算力的支持, 其成本较大. 因此, Li 等人[43]基于 ChatGPT 作为隐式知识库提出了两阶段生成框架 PGIM (prompting ChatGPT in MNER). 该框架首先从利用提示模板通过 ChatGPT 生成辅助的细粒度知识, 然后将其与原文本拼接, 拼接后的内容继续输入到下游模型得到最终分类. 该方法可利用 ChatGPT 强大的隐式知识库, 并分别在 Twitter2015 和 Twitter2017 上取得了较好的性能. 此外, 张天明等人[12]针对小样本数据提出了一种融合多模态数据的命名实体识别模型, 借助多模态数据提供额外语义信息, 帮助模型提升预测效果. 该方法主要包含 3 个模块: 首先, 多模态信息提取模块将图片转为文本形式并作为辅助模态信息输入发射模块; 然后, 由发射模块计算输出对应的发射分数, 转移模块输出对应的转移分数; 最后, 基于计算出的发射分数和转移分数之和, 将候选标签序列中概率最高的标签序列作为查询样本对应的标签序列并输出.

3.2.2 基于跨度分类 (span-based) 的方法

针对 MNER 任务, 主流方法是将其视为一个序列标注任务, 但是这类方法往往存在耗时、实体边界预测不准确等缺点. 因此, 近年来, 有学者尝试利用 span-based 方法来解决 MNER 任务[8,39,54]. 该方法的主要思想是: 通过始末位置的概率分布图将潜在实体的所有可能区域或范围枚举出来, 然后利用深度神经网络对其进行分类[23]. 其融合方法主要包含基于注意力机制的前融合方法和基于 POE (product of experts)[55]的后融合方法.

● 基于注意力机制的前融合方法

Yamada 等人[56]通过实验展示了基于跨度分类方法在 NER 任务上的强竞争力. 受此启发, Jia 等人[8]通过设计实体跨度分类预测等子任务提出基于阅读理解的 MNER 模型 (MRC-MNER). 首先通过迁移学习训练一个区域视觉定位模块 (visual grounding model), 并基于该模块得到与查询相关的 Top-k 个区域图片信息, 然后通过文本内交互模块和图文间交互模块实现图文充分融合, 最后通过 3 个子任务 (视觉区域权重评估、实体存在性检测和实体跨度分类预测) 联合训练 MRC-MNER 模型. 由 MRC 框架中特殊的查询设计可提供部分先验信息, 因此, MRC-MNER 模型通常性能较好. 但是该模型中查询设计是关键步骤, 不同的查询设计可能存在较大的性能差异. Wang 等人[39]延续上述工作的思路, 首先将 NER 任务重新表述为识别实体跨度的开始和结束位置索引以及为跨度分配类别标签这样一个新任务, 然后提出一个基于 Transformer 的先进方法 CAT-MNER. 该框架的重点是图片和文本进行融合的多头注意力增强部分以及跨度预测部分, 与其他模型相比, CAT-MNER 结构相对简单且易于实现.

● 基于 POE 的后融合方法

上述方法是以有监督方式来解决 MNER 任务, 但是由于大规模多模态数据集存在标注成本高等特点, 近年来, 半监督或远程监督的方法引起了研究者们的关注. Zhou 等人[54]提出基于跨度的多模态变分自编码器 SMVAE (span-based multimodal variational auto encoder) 来解决 MNER 任务. 该方法首先利用两个变分自编码器分别建模图片的潜在语义表示和文本的跨度水平表示 (span-level token), 然后引用 POE 来融合图片语义表示和文本跨度表示, 最终通过融合后的特征来预测每条文本所有跨度的标签. 这样的好处是可以利用预测概率和多模态特征来重构输入特征表示, 隐式建模跨度标签与多模态特征之间的相关性, 通过这种方式可以利用未标记多模态数据的有用信息来提高 MNER 任务的性能.

3.2.3 基于大模型直接生成的方法

随着语言大模型和多模态大模型相继被提出, 学者们可以直接使用大模型来生成任务的答案[43,57]. 其中, Li 等人[43]分别提出了直接生成实体的基准模型 VanillaGPT、PromptGPT (N=1) 和 PromptGPT (N=10). VanillaGPT 模型首先是将图片转为图片-摘要对, 然后将摘要与原文本组合一起通过调用 GPT-3.5-Turbo 直接生成实体. 与 VanillaGPT 模型不同的是, PromptGPT (N=1) 和 PromptGPT (N=10) 模型除了需要组合原文本外, 还需要组合人工设计的提示模板, 模板的作用是提示模型以便生成更好的答案, 这里的 N 表示为上下文学习选择 Top-N 个相似样

例. Chen 等人[57]除了调用 GPT-3.5-Turbo 直接生成实体外, 还调用了更先进的 GPT-4, 通过输入图片和文本的方式直接生成实体. 该模型被命名为 GPT4. 各个基于大模型直接生成的方法在数据集 Twitter2015 和 Twitter2017 上的表现如表 5 和表 6 所示, 虽然直接调用 GPT-3.5-Turbo 或者更先进的 GPT-4 接口来生成结果更方便, 但是与其他先进模型 (如: 序列标注模型) 相比, 其结果 (*F*1 值) 远远不及预期, 因此, 更好的方案可能是先通过 ChatGPT 或者 GPT-4 获取外部知识, 然后将获取得外部知识与原文本或者原图片信息结合并将结合后的信息输入到下游模型 (如 CRF), 最终实现实体分类.

表 5 多模态命名实体识别模型在数据集 Twitter2015 上的性能比较 (%)

类别	方法	模型	所有类型 (Overall)		
			Pre	*Rec*	*F*1
基于序列标注的方法	基于图结构的前融合	UMGF[15]	74.49	75.21	74.85
		R-GCN[33]	73.95	76.18	75.00
		R-GCN (w/o Gate)[33]	72.50	76.89	74.60
	基于注意力机制的前融合	OCSGA[25]	74.71	71.21	72.92
		文献[9]	74.78	71.82	73.27
		UMT-BERT-CRF[26]	71.67	75.23	73.41
		文献[7]	72.75	68.74	70.69
		文献[24]	73.50	72.33	72.91
		ITA[27]	76.52	74.71	75.60
		MAF [28]	71.86	75.10	73.42
		FMIT (l=3)[48]	75.11	77.43	76.25
		HVPNeT[40]	73.87	76.82	75.32
	基于多模态预训练模型的前融合	文献[41]	71.94	75.13	73.50
		PGIM_BERT[43]	75.84	**77.76**	**76.79**
基于跨度分类的方法	基于注意力机制的前融合	MRC-MNER[8]	**78.10**	71.45	74.63
		CAT-MNER [39]	76.19	74.65	75.41
	基于POE的后融合	文献[54]	74.40	75.76	75.07
基于大模型直接生成的方法	ChatGPT	VanillaGPT[43]	42.96	75.37	54.73
		PromptGPT (N=1)[43]	51.96	75.24	61.47
		PromptGPT (N=10)[43]	58.57	74.07	65.41
	GPT-4	GPT4[57]	—	—	57.98

注: “—”表示原文献中未给出该指标对应的实验结果, 下文所有表中的“—”含意相同

表 6 多模态命名实体识别模型在数据集 Twitter2017 上的性能比较 (%)

类别	方法	模型	所有类型 (Overall)		
			Pre	*Rec*	*F*1
基于序列标注的方法	基于图结构的前融合	UMGF[15]	86.54	84.50	85.51
		R-GCN[33]	86.72	87.53	87.11
		R-GCN (w/o Gate)[33]	85.90	87.57	86.70
	基于注意力机制的前融合	UMT-BERT-CRF[26]	85.28	85.34	85.31
		ITA[27]	86.69	84.77	85.72
		MAF[28]	86.13	86.38	86.25
		FMIT (l=3)[48]	87.51	86.08	86.79
	基于多模态预训练模型的前融合	文献[41]	86.25	86.38	86.32
		PGIM_BERT [43]	**89.09**	**90.08**	**89.58**

表 6 多模态命名实体识别模型在数据集 Twitter2017 上的性能比较 (%) (续)

类别	方法	模型	所有类型 (Overall)		
			Pre	*Rec*	*F*1
基于跨度分类的方法	基于注意力机制的前融合	MRC-MNER[8]	88.78	85.00	86.85
		CAT-MNER [39]	87.04	84.97	85.99
	基于POE的后融合	文献[54]	85.77	86.97	86.37
基于大模型直接生成的方法	ChatGPT	VanillaGPT[43]	52.19	75.03	61.56
		PromptGPT (*N*=1)[43]	56.99	74.77	64.68
		PromptGPT (*N*=10)[43]	72.90	77.65	75.20
	GPT-4	GPT4[57]	—	—	66.61

4 多模态实体关系抽取

4.1 数据集

近年来, 缺乏包含实体关系的大规模多模态数据集成为 MERE 任务发展的首要障碍. 因此, 2021 年, Zheng 等人[16]提出了多模态关系抽取数据集 MNRE. 该数据集主要来源于: Twitter2015、Twitter2017 以及再次在 Twitter 网站上爬取的数据. 同年, Zheng 等人[31]也基于 Twitter2015、Twitter2017 以及再次在 Twitter 网站上爬取的数据提出了数据集 MNRE_MM (文献 [31] 中的数据集与文献 [16] 中的数据集同名, 为避免重名引起歧义, 本文中将文献 [31] 中的数据集命名为 MNRE_MM). 与数据集 MNRE 不同的是: 数据集 MNRE_MM 是由标注人员按照不同主题 (如音乐、体育和社会事件) 进行筛选后得到, 而 MNRE 数据集没有按主题筛选. 两个数据集中的文本都表现为短文本的形式, 与文本相关的图片可以为短文本提供信息补充, 进而提高实体关系识别的精度. 两个数据集的统计信息见表 7.

表 7 MNRE[16]和 MNRE_MM [31]数据集的统计信息

数据集	*Img*	*Sent*	*Ent*	*Rel*	*Inst*	创建数据集的年份	数据集类型	SOTA性能 (*F*1值) (%)	数据集链接
MNRE	10089	14796	20178	31	10089	2021	推文	68.60[47]	https://github.com/thecharm/MNRE/tree/main/Version-1
MNRE_MM	9201	9201	30970	23	15485	2021	音乐、体育等主题推文	84.86[58]	https://drive.google.com/file/d/1gD9ipQgDEDRxaVxkKr8T0gFFQgKyPpa7/view

注: *Img*表示图片的数量, *Sent*表示句子的数量, *Ent*表示实体的数量, *Rel*表示实体关系的数量, *Inst*表示实例的数量

4.2 方法总结

在之前的研究工作中, 尽管实体关系抽取已经取得了较大的成功[59−61], 但这些模型绝大多数是在纯文本语料上训练, 其在包含图文的多模态语料上的性能表现缺乏一般性, 模型不能学习图片提供的信息. 与纯文本实体关系抽取方法类似, 由于基于流水线方法 (Pipeline) 存在错误累计、缺少子任务间的信息交互以及产生无确定实体关系的冗余实体等缺点; 基于特征工程的联合抽取方法 (Joint) 又依赖大量人工提取的特征规则, 存在成本高效率低等缺点. 因此, 基于深度学习的联合抽取方法成为多模态实体关系抽取任务的主流方法.

Zhao 等人[59]认为两个实体的类型可能对实体关系的分类有重要影响, 例如: 如果已经知道两个实体的标签是“位置 (location)”, 那么很容易判定这两个实体的关系为“位于 (located in)”. 因此, 将命名实体识别得到的实体标签看成区别于文本的另一种模态, 这样就将纯文本的单模态任务变成为一个跨模态任务, 并提出联合抽取命名实体和实体关系的方法 CMAN (cross-modal attention network). 通过在 ADE 和 CoNLL04 这两个数据集上的实验显示: 使用该方法, 其实体关系抽取的性能 (*F*1 值) 比基线模型分别提高了 1.9% 和 1.5%.

上述工作本质上还是单模态 (纯文本) 关系抽取, 并未引入文本态之外的其他模态作为输入. 随后, Zheng 等人[16]

引入包含图文的多模态关系抽取任务, 并分别通过 GloVe+CNN[62]、BertNRE[63]和 BERT+CNN[16]的变体方法 (GloVe+CNN(Att.)[16]、BertNRE(Att.)[16]、BERT+CNN(Att.)[16]等) 证明了融合图片信息可提高实体关系抽取的性能. 此外, 进一步引入远程监督方法 PCNN[64]的变体 (PCNN(Lab.)[16]、PCNN(Obj.)[16]、PCNN(Att.)[16]等), 实验结果显示: PCNN 的变体在 MNRE 数据集上的性能均差于 PCNN 方法. 上述对比实验说明: 虽然融合图片信息可提高实体关系抽取的性能, 但并不是所有多模态方法在性能上均优于单模态方法. 为了进一步提高模型性能, Wang 等人[47]提出了一个基于多模态检索的联合抽取框架 MoRe (multi-modal retrieval), 如图 8 所示. 该框架包含文本检索模块和图片检索模块, 与 PURE 模型[61]类似, 然后分别在这两个检索结果上添加特殊标记并单独进行结果预测, 最后将两个预测结果输入一个专家混合模块得到最终的预测结果. 该方法在 MNRE 数据集上取得了较好的性能表现. Zheng 等人[31]认为, MERE 除了需要捕获图片中各目标实体以及文本中各实体之间的相关性外, 还需要关注图片中各目标实体之间的视觉关系到句子中各实体之间的文本关系的映射, 因此, 提出了一个基于双图对齐的多模态神经网络方法 MEGA. 该方法通过图片和文本之间的关系映射可以找到图片中实体与文本中实体的相关性, 然后利用图片中的实体关系来提高文本中实体关系的精度, 在数据集 MNRE_MM 上的实验表明: MEGA 以及各种变体模型均优于纯文本模型, 进一步证明了在关系抽取任务中通过引入相关的图片信息是有效的. 此外, 他们基于情景图工具 (pretrained scene graph tool) 提出了两个多模态实体关系抽取的基准模型: BERT+SG 和 BERT+SG+Att., 得益于有效的图文对齐方法, MEGA 方法在 MNRE_MM 数据集上的各项指标中均取得了较好的结果, 尤其是与先进的多模态基准模型 BERT+SG+Att.相比, MEGA 方法的准确率提升了 5.8%, 这表明图片的有效引入帮助模型缓解了实体关系歧义的问题.

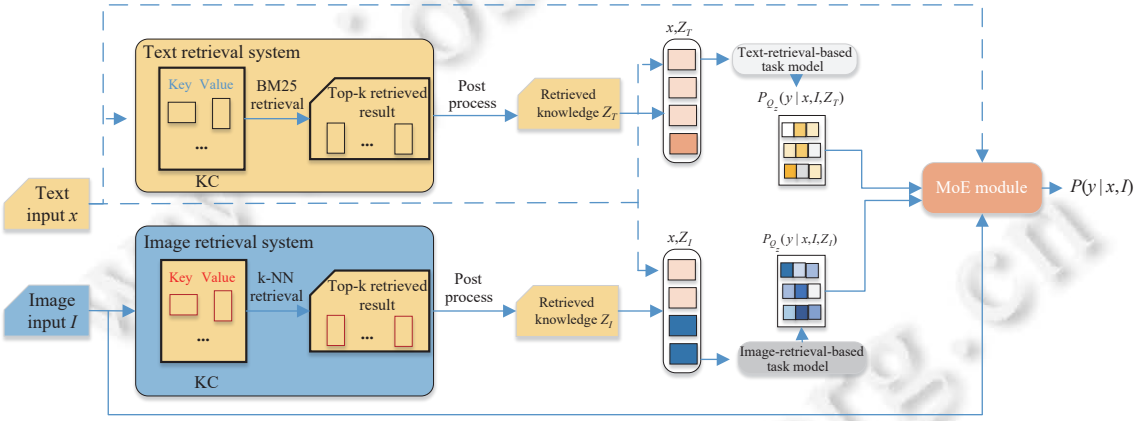


图 8 MoRe 框架图[47]

上述模型并未考虑图文不相关的情况. 因此, Xu 等人[41]基于强化学习首先训练一个二分分类器将数据集中的数据分为单模态和多模态两部分, 然后基于先进模型 MEGA 和 MTB 联合抽取命名实体以及实体之间的关系, 在数据集 MNRE_MM 上取得了较好的性能. 此外, Chen 等人[40]提出了一个分层视觉前缀融合网络联合抽取模型 HVPNeT. 该模型引入了门控机制, 同时考虑了图片的整体特征和图片中的目标实体特征, 最终通过 Softmax 函数得到实体关系分类的概率分布. Chen 等人[57]通过生成思维链 CoT (chain of thought) 引入大模型 LLM (large language model) 强大的常识推理能力, 然后基于该推理提出一种新的条件提示蒸馏方法, 从而增强了模型的性能. Hu 等人[58] 提出了实体-对象和关系-图像对齐预训练任务, 能够从海量未标记图像-字幕对中提取自监督信号来预训练多模态融合模块并提高 MERE 的性能. 该模型在数据集 MNRE_MM 上取得了强有力的性能表现.

随着大模型越来越受到关注, Chen 等人[57]引入了 ChatGPT 和 GPT4 模型, 其中, ChatGPT 直接调用 GPT-3.5-Turbo 来生成结果, GPT4 直接调用 GPT-4 来生成结果. 现阶段, 研究人员通常将这两个模型作为基准模型. 如表 8 所示, 不管是有监督的方法还是远程监督的方法, 它们的性能 ($F1$ 值) 在数据集 MNRE 上的表现均远优于无监督方式下的两个基准模型. 与 MNER 类似, 基于大模型生成的外部知识进行改进是一个值得探索的方向.

表 8 多模态实体关系抽取模型在数据集 MNRE 上的性能表现 (%)

监督方式	模型	模型特点	*Pre*	*Rec*	*F*1
远程监督	PCNN(Obj.)[16]	在文本基础上进一步引入图片中的实体特征	63.85	45.40	53.07
有监督	GloVe+CNN(Att.)[16]	基于CNN方法提取其中的文本模态特征	62.25	46.72	53.38
	BERT+CNN(Att.)[16]	图片模态中使用CNN方法提取图片特征	65.28	61.72	63.45
	BertNRE(Att.)[16]	BERT方法提取其中的文本模态特征	**68.94**	62.47	65.56
	MoRe[47]	基于多模态检索框架来过滤不相关的图文信息	68.23	**68.79**	**68.51**
	GPT4-BERT[57]	引入大模型的强大推理能力提高MERE任务上的性能	—	—	67.88
无监督	ChatGPT[57]	调用GPT-3.5-Turbo来生成结果	—	—	35.20
	GPT4[57]	调用GPT-4来生成结果	—	—	42.11

注: “Obj.”表示目标检测模型得到的图片特征, “Att.”表示通过注意力机制实现两种模态的交互

上述各模型在 MNRE_MM 数据集上的性能表现如表 9 所示.

表 9 多模态实体关系抽取模型在数据集 MNRE_MM 上的性能表现 (%)

类别	模型	模型特点	*Pre*	*Rec*	*F*1
假定数据集中图文相关	BERT+SG[31]	在文本信息基础上进一步引入图片信息	62.95	62.65	62.80
	BERT+SG+Att.[31]	进一步使用注意力机制来衡量图片与文本之间的语义相似度	60.97	66.56	63.64
	MEGA[31]	进一步使用图对齐方法实现图片和文本之间的结构一致性和语义一致性	64.51	68.44	66.41
考虑了数据集中图文不相关	Xu等人[41]	通过高效分类器将数据分别对应到单模态和多模态模型	66.83	65.47	66.14
	HVPNeT[40]	使用动态门控和注意力机制实现模态融合	83.64	80.78	81.85
	Hu等人[58]	提取自监督信号来预训练多模态融合模块	**84.95**	**85.76**	**84.86**

注: SG为Scene Graph的缩写

Wan 等人[11]进一步讨论了多模态社会关系抽取 (multimodal social relation extraction). 与一般的实体关系抽取不同, 该任务旨在抽取日常生活中两个人的社会关系 (父子关系和兄妹关系等), 这种关系比一般的实体关系更难抽取, 因为部分关系还会涉及推理知识, 例如: 文本可得到两个人是家人, 图片信息可得到两个人是女生, 而且年龄相仿, 那么结合图文信息, 可推理得到两人的关系是姐妹. 过去的工作主要集中在单模态[65−69], 忽略了多模态信息之间的高度耦合, 而图片信息可能提供文本表示关系之外的其他关系, 这些信息可为社会关系提供更精确的分类. 受此启发, Wan 等人[11]尝试在文本输入的基础上引入图片模态, 并提出了基于少样本学习 (few-shot) 方法抽取多模态社会关系, 最终取得了较好的性能.

5 多模态事件抽取

5.1 数据集

为了推动多模态事件抽取的发展, Li 等人[13]除了提出多模态事件抽取这个新任务外, 还公布了一个面向多模态事件抽取任务的大规模数据集: M^2E^2. 他们首先选取了 2006–2017 年间的 108 693 篇包含图文的多媒体文章, 然后按以下 3 个规则进一步对这些文章过滤: 1) 倾向选取包含更多事件的文章; 2) 倾向选取包含更多图片的文章 (最少包含 4 张图片); 3) 选取文章的事件类型尽可能平衡. 过滤后共有 235 篇文章 (可能涉及敏感话题, 原文中删除了 10 篇), 由这 235 篇文章组成数据集 M^2E^2. 该数据集共包含 6 167 个句子和 1 014 张图片.

此外, Li 等人也对该数据集的部分数据进行了标注, 其中, 文本部分的标注参考 ACE 事件标注指南, 图片部分的标注参考图片事件标注指南 (http://blender.cs.illinois.edu/software/m2e2/ACL2020_M2E2_annotation.pdf). 标注后的数据集共包含 8 个文本事件类型. 这 8 个文本事件类型均来自 ACE2005 数据集中的事件类型, 之所以重复使用这些已存在的分类是使得事件分类器和论元分类器可以直接接收多模态数据的输入而不需要在已标注的多模态数据集上做联合训练. 同时, 还借助 imSitu 数据集拓展了原 ACE2005 数据集中的论元角色, 这样做的目的是使数

据集 M^2E^2 中的论元信息更丰富. 数据集 M^2E^2 的事件类型和论元角色信息如表 10 所示 (补充的论元角色在表 10 中加粗显示, 如 Instrument 和 Police 等).

表 10 数据集 M^2E^2 的事件类型和论元角色信息[13]

事件类型	论元角色
Movement.Transport	Agent, Artifact, Vehicle , Destination, Origin
Conflict.Attack	Attacker, Target, Instrument, Place
Conflict.Demonstrate	Entity, **Police**, **Instrument**, Place
Justice.ArrestJail	Agent, Person, **Instrument**, Place
Contact.PhoneWrite	Entity, **Instrument**, Place
Contact.Meet	Participant, Place
Life.Die	Agent, Instrument, Victim, Place
Transaction. TransferMoney	Giver, Recipient, Money

按第 1.3 节多模态事件抽取的定义可知: M^2E^2 数据集中共有 1 297 个句子和 391 张图片包含事件; 其中, 1 105 个句子包含仅文本事件, 188 张图片包含仅图片事件, 剩下的 192 个句子可与剩下的 203 张图片组成 309 个多模态事件对 (句子和图片不是一对一的关系). M^2E^2 数据集的统计信息如表 11 所示.

表 11 数据集 M^2E^2 的统计数据[13]

数据源		事件提及		论元角色		数据集类型	创建数据集年份	SOTA性能 (触发词$F1$值) (%)	数据集链接
Sentence	Image	Textual	Visual	Textual	Visual				
6 167	1 014	1 297	391	1 965	1 429	新闻	2020	57.5[70]	http://blender.cs.illinois.edu/software/m2e2

5.2 方法总结

虽然文本事件抽取已取得了较大的成功[71−73], 但其多数模型均是在 ACE2005 数据集上训练得到, 这些模型仍然存在一词多义、缺失部分论元等问题导致事件抽取错误. 为了进一步缓解上述问题, 研究者们引入相关图片来增强文本的语义表示. 目前, 多模态事件抽取的相关工作主要包含两个方向. 1) 图片辅助文本的事件抽取 (通过图片提高文本事件抽取的性能). 2) 多模态事件抽取 (合并文本事件抽取和图片事件抽取为统一的多模态事件抽取). 本节主要围绕上述两方面内容展开分析.

5.2.1 图片辅助文本的事件抽取

文本事件抽取任务是信息抽取任务中的一个经典子任务, 学者们通过检索外部文档知识[74,75]、文本特征工程[76−78]、先进的学习框架[71,79,80]等方法在该任务上取得了较好性能. 但是上述方法中的模型多数是基于文本语料库 ACE2005 上训练, 由于数据集 ACE2005 本身规模并不大且存在长尾现象, 当测试集中存在的多义词触发不常见的事件时, 上述模型基本都是将该多义词判定为常见的事件类型, 导致事件抽取错误. 在多模态视角下, 学者们通过在输入端融合图片信息来训练文本事件抽取模型以提升模型在测试集上的性能表现[3,17], 输入与文本相关的图片可提供话题相应的背景知识, 这些背景知识可缓解由一词多义造成事件触发词类型及事件论元抽取错误的问题.

Bosselut 等人[81]和 Young 等人[82]分别在两个涉及图文事件推理的任务中, 通过引入与图片对应的文本描述提高了图片中事件检测的性能, 实验结果显示: 通过上述模型除了可以找到与主题直接相关的事件, 还找到了部分相关的蕴含事件. 受此启发, 有学者尝试引入图片信息来提高文本事件抽取性能[3,17]. 其中, Zhang 等人[3]首先使用 VAD (visual argument discovery) 以弱监督方式构建一个丰富的视觉背景知识库, 然后根据给定句子中的实体来检索视觉背景知识库中的视觉信息, 将检索到的视觉信息与文本特征融合得到多模态特征, 以得到的多模态特征来训练文本事件分类器, 最后分别在基准数据集 ACE2005 和 ERE 上测试. 测试结果显示: 在数据集 ACE2005 上, 与基准模型 JointIE[77]相比, VAD 在事件触发词和论元抽取上的 $F1$ 值分别获得了 1.8% 和 3.2% 的提升; 在数据集 ERE 上, 与基准模型相比, VAD 在事件触发词和论元抽取上的 $F1$ 值分别获得了 7.1% 和 8.2% 的大幅提升. 通过

上述实验结果表明: 输入中融合相关的图片信息可以提高文本事件抽取的性能, 同时也验证了所提出模型的有效性.

虽然通过上述 VAD 模型提升了文本事件抽取模型的性能, 但是在训练阶段, VAD 模型使用全连接层对两种模态进行融合, 两种模态没有得到深层次的交互. 随着注意力机制的出现, 基于注意力的模型既可以关注与文本相关的图片区域, 还可以关注与图片相关的文本内容. 因此, Tong 等人[17]基于注意力机制提出双循环多模态模型 DRMM (dual recurrent multimodal model), 如图 9 所示. 先通过 BERT 和 ResNet 分别提取文本和图片特征, 然后通过 N 个交替双重注意力模块 ADA (alternating dual attention) 不断迭代来融合图文信息, 最后将融合后的多模态特征用来训练文本事件分类器. 通过在测试集 ACE2005 上实验显示: 该模型的性能 ($F1$) 比 VAD 方法提升了 7%. 上述各模型在数据集 ACE2005 和 ERE[83]上的性能对比如表 12 所示.

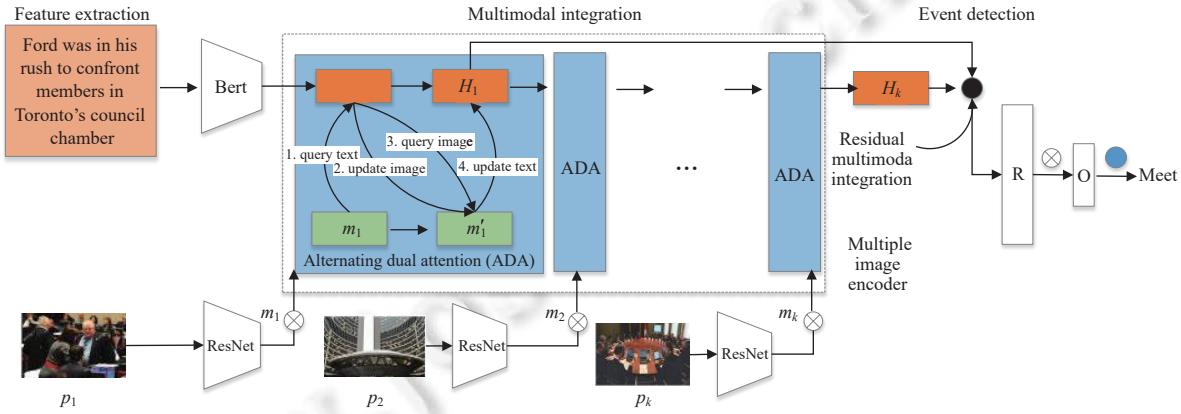


图 9 DRMM 方法总览[17]

表 12 图片辅助文本事件抽取模型的性能对比 (%)

模型	触发词			论元			测试数据集
	Pre	*Rec*	*F*1	*Pre*	*Rec*	*F*1	
VAD[3]	75.1	64.3	69.3	63.3	50.1	55.9	ACE2005
DRMM[17]	**77.6**	**74.1**	**75.8**	**64.8**	**55.7**	**59.9**	
VAD[3]	**48.9**	**44.5**	**46.6**	**31.8**	**25.8**	**28.5**	ERE

5.2.2 多模态事件抽取

上述事件抽取任务仅仅是通过融合另一种模态信息来提高单模态事件抽取的性能, 并没有完整抽取横跨多个模态的多模态事件 (同时包含文本事件抽取和图片事件抽取). 通过增加与文本对应的图片模态不仅可以提高抽取文本事件触发词的性能, 还可以补充文本中缺乏的部分事件论元和事件类型. 由于多模态事件标注工作复杂, 标注数据成本大, 缺乏对齐的大规模图文事件标注数据集. 因此, 多数研究者通过对比学习[13,14,84]、迁移学习[13]等弱监督方法来解决多模态事件抽取任务.

Radford 等人[52]首先在图文匹配任务上提出使用对比学习框架来联合学习文本和图片的特征表示, 然后将模型以零样本的方式迁移到下游任务中. 在 30 多个不同任务的数据集上测试显示: 模型性能与全监督模型取得的性能相当. 实验结果证明了对比学习框架是一个非常高效的图文表示方法. 受此启发, 有学者将对比学习框架引入 MEE 任务中[13,14,84]. 其中 Li 等人[13]提出一个弱对齐的结构编码框架 WASE (weakly aligned structured embedding). 如图 10 所示. 该框架分为训练和测试两个部分. 在训练阶段, 首先分别得到文本表示的结构图信息和图片表示的结构图信息, 然后通过对齐模块在图片描述 (image caption) 数据集上做弱对齐训练; 训练过程中, 要求结构图上图文匹配的节点靠近, 不匹配的节点之间则远离; 通过该方法将图文两种模态尽可能地编码到同一语义空间. 在测试阶段, 为每个句子找到最匹配的图片并融合两个模态的特征, 然后输入文本事件分类器进行事件和论元分类. 类似

地, 为每个图片找到最匹配的句子并融合两个模态的特征然后输入图片事件分类器进行事件和论元分类; 最终若两个事件的事件类型相似度超过阈值, 则合并两个事件为一个多模态事件. 在上述方法中, 首先使用基于目标检测的方法来构建图片模态的结构图, 其多模态事件抽取方法命名为 $WASE_{obj}$, 但是由于该目标检测方法是在 Open Images 数据集上训练得到的, 通过该方法得到的检测结果是一个有限的集合, 无法实现新的实体类型的检测. 因此, 进一步提出了基于注意力机制构建图片模态的结构图, 并将该多模态事件抽取方法命名为 $WASE_{att}$. 两种方法在多模态数据集 M^2E^2 上的性能对比如表 13 所示: 在事件触发词抽取的子任务上, $WASE_{obj}$ 显著优于 $WASE_{att}$; 在事件论元抽取的子任务上, $WASE_{att}$ 的性能要略优于 $WASE_{obj}$. 因此, 在不同的子任务上, 两种方法表现出不同的优势.

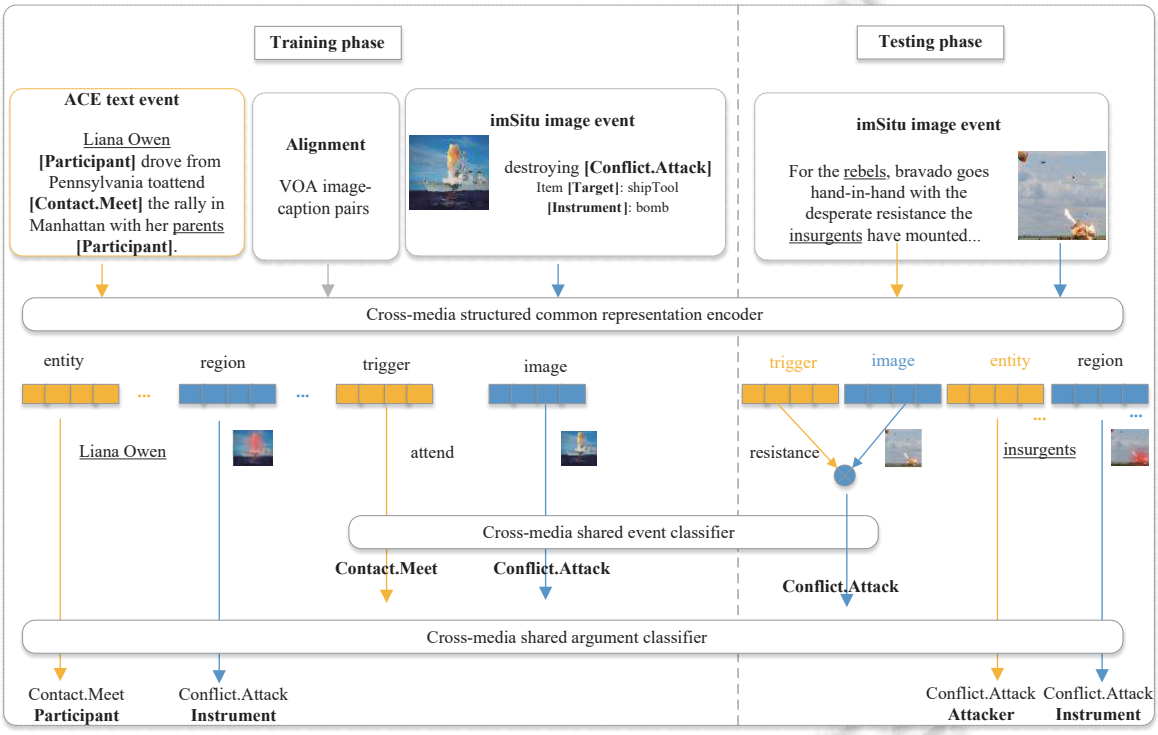


图 10 WASE 框架总览[13]

表 13 多模态事件抽取模型在数据集 M^2E^2 上的性能对比 (%)

模型	触发词			论元		
	Pre	*Rec*	*F*1	*Pre*	*Rec*	*F*1
$WASE_{att}$[13]	38.2	67.1	49.1	18.6	21.6	19.9
$WASE_{obj}$[13]	43.0	62.1	50.8	19.5	18.9	19.2
CLIP-Event[84]	41.3	**72.8**	52.7	21.1	13.1	17.1
UniCL[14]	44.1	67.7	53.4	24.3	22.6	23.4
CAMEL[70]	**55.6**	59.5	**57.5**	**31.4**	**35.1**	**33.2**
GPT3.5[85]	17.78	31.31	19.56	10.77	21.62	12.11
GPT3.5/SC[85]	12.62	17.77	11.49	7.9	8.62	6.95

Liu 等人[14]进一步改进上述弱对齐框架, 提出了统一对比学习框架 UniCL (unified contrastive learning framework). 与 WASE 框架的空间表示部分相同, 首先通过引导匹配的图文对获得比不匹配图文对更高的分数这样的方式来学习图文的公共表示空间; 不同的是, 在图片事件抽取部分, 研究人员基于查询的策略, 将预先定义好

的动作动词分别与图片和文本计算相似度, 并基于权重加权平均融合得到最终的相似度分数, 依据得到的相似度分数匹配得到图片的事件类型和论元; 然后利用贪心算法不断迭代找到最佳匹配的图文事件, 最后逐步合并匹配的图文事件得到所有的多模态事件.

与上述方法不同, Du 等人[70]利用文本-图片 (text-to-image) 和图片-文本 (image-to-text) 这种双向数据增强的方式提出了一个跨模态增强多模态事件学习框架 CAMEL (cross-modality augmented multimedia event learning). 该框架首先利用先进的图像生成工具和文本生成工具分别在纯文本数据集和纯图片数据集上生成对应的模态信息, 然后合并这两个数据集为一个有标记的多模态数据集. 通过该方式可缓解对齐的大规模图文事件标注数据集短缺的问题, 但是在利用生成工具生成对应模态的过程中难免会引入噪声, 因此, 进一步设计了一种增量训练策略, 该策略可缓解人工生成的多模态数据中存在的伪影、幻觉和分布变化的问题, 以此来缓解此类噪声引起模型性能下降的问题. 该方法在数据集 M^2E^2 上取得了较好的性能表现.

上述多模态事件抽取方法在考虑模态间的融合时, 只是考虑了实体级别的对齐, 并没有考虑事件结构级别的对齐. 而同样的实体由于可以对应不同的论元角色分类, 可能代表两个不同的事件. 因此, Li 等人[84]提出考虑事件结构的自监督对比学习框架 CLIP-Event, 通过在 M^2E^2 数据集上的测试结果显示: 与先进模型 UniCL 相比, CLIP-Event 在触发词抽取任务上的召回率提高了 5.1%.

随着大模型的应用越来越广泛, Moghimifar 等人[85]将 ChatGPT 引入到多模态事件抽取任务上并分别提出了 GPT3.5 模型和 GPT3.5/SC (scene description) 模型. 其中, GPT3.5 模型是直接输入文本, 然后调用 GPT-3.5-Turbo 得到; GPT3.5/SC 模型首先使用视觉生成文本工具将图片转为文本, 然后调用 GPT-3.5-Turbo 得到. 由表 13 可以看出: GPT3.5 模型在各项评价指标上的性能均要优于 GPT3.5/SC. 鉴于 GPT3.5 是一个单流模型, 图像生成文本的过程中可能会丢失大量特征, 下一步可以考虑在事件抽取任务中引入双流模型 (如 GPT-4).

此外, 由于视频中包含丰富的动作信息以及更多事件论元信息, 这些动作或论元信息在单帧的图片中可能无法体现. 因此, Chen 等人[86]引入了视觉多媒体事件抽取任务 (video multimedia event extraction, VMEE). 在文本模态上融合视频模态的难点在于: 需要确定视频中事件的时间边界. 相较于图片模态, 视频态的标注工作更复杂. 因此, 提出了一个自监督的训练框架, 首先找到句子-视频片段对中的事件同指, 然后提出一个多模态 Transformer 框架, 利用特定于模态的解码器联合抽取文本模态和视频模态的事件和论元.

6 未来与展望

6.1 研究趋势

总体来看, 现有的多模态信息抽取研究工作主要集中在多模态命名实体识别和多模态实体关系抽取两个子任务上, 多模态事件抽取任务还处于起步发展阶段.

● 多模态命名实体识别任务的研究趋势. 现阶段, 主要的研究工作集中在通过图模型、注意力机制、对比学习和基于上下文学习的推理等方法使得图文两种模态得到充分交互. 融合策略上也从尽可能多的融合图片特征向恰当的融合图片特征转移, 如通过训练高效的数据分类器以及增加基于纯文本的辅助模块等方法来缓解由不匹配的图文给模型性能带来的负影响. 尽管近年来通过各种先进的深度学习方法在该任务上取得了较大成功, 但大部分模型均是在 Twitter2015 和 Twitter2017 两个社交数据集上测试其性能, 缺乏领域性多模态数据集: 如金融、法律 以及自然灾害等领域数据对模型进行评价. 总体来看, 多模态命名实体识别仍然是一个重要且充满挑战的研究任务.

● 多模态实体关系抽取任务的研究趋势. 在多模态实体关系抽取任务的发展初期, 研究者主要通过融合图片信息对短文本的内容进行补充, 通过图片中的背景知识、图片中的实体以及图片中的实体之间的关系来增强短文本的语义表示, 进而提高实体关系抽取模型的性能. 这个阶段主要关注图片特征提取以及图文融合策略等方面. 现阶段, 研究者们更加关注联合抽取命名实体和实体之间的关系. 通过先进的联合抽取技术, 可较大程度上缓解管道方法存在的各种缺点. 但是对于开放域以及特定领域的实体关系, 目前还缺乏相应探究.

● 多模态事件抽取的研究趋势. 类似于上述 MNER 和 MERE 任务, 早期的研究工作专注于通过融合图片信息

来缓解句子中的多义词问题, 进而提高文本事件抽取的性能. 考虑到各模态数据之间的互补性, 图片模态不仅可以增强文本态的语义表示, 还可以通过图片事件抽取来补充文本中可能缺失的部分事件论元和事件类型. 随后, 研究重心逐步从图片辅助文本的事件抽取转向多模态事件抽取. 多模态事件抽取同时涉及文本事件抽取、图片事件抽取以及图文事件同指等工作, 该任务较为复杂, 因此, 多模态事件抽取仍然存在较大的探索空间.

6.2 存在的挑战与展望

通过上文中对多模态信息抽取研究趋势的分析, 本文认为多模态信息抽取的研究工作在大规模多模态标注数据集、多模态数据的特征表示及融合策略、多模态信息抽取的子任务间协同以及融合外部知识等问题上仍然存在挑战. 具体来讲, 在构建面向多模态信息抽取任务的大规模多模态标注数据集、面向多模态信息抽取任务的细粒度图文融合策略、面向多模态信息抽取各子任务之间的推理知识和面向开放域的多模态信息抽取任务这 4 个方面存在如下的挑战.

● 构建面向多模态信息抽取任务的大规模多模态标注数据集. 大规模的多模态标注数据集是多模态研究中不同任务面临的共性问题, 原因是多模态数据集普遍存在标注费用高、耗时长等特点. 相比于单模态数据集, 标注多模态数据时需要同时结合两种模态信息并将其对齐, 然后在理解整体含义的基础上对两种模态进行标注, 工作量大且复杂度高. 因此, 目前的多模态信息抽取研究中普遍存在数据集规模小 (MNER 任务中常用的 Twitter2015 有 8257 个句子, Twitter2017 有 7181 个句子)、标注数据少 (MEE 任务中常用的 M^2E^2 共标注 1297 个句子和 391 张图片) 等问题. 虽然一些针对小样本的先进模型被提出[11,12], 但这只是在小样本情况下提升模型性能的一种方式, 数据集的规模大小将会影响对先进模型的全面评估, 这也是在其他多模态任务上遇到的模型性能超过人工标注性能的一种可能原因. 因此, 构建大规模的多模态标注数据集刻不容缓. 虽然通过表 5、表 6、表 8 和表 13 发现: 直接调用 ChatGPT 在信息抽取任务上表现较差, 但是鉴于 ChatGPT 具有强大的常识推理能力, 而且 GPT-4 可提供双流输入, 调用 GPT-4 对多模态数据集进行初步标注, 然后结合人工检查的方式将为构建大规模的多模态标注数据集提供更大可能.

● 面向多模态信息抽取任务的细粒度图文融合策略. 由于多模态知识最早是被应用于情感分析以及视觉问答等任务上, 并取得了出色的性能表现[1,2]. 目前, 这些领域的多数图文融合策略均在多模态信息抽取任务中被借鉴. 虽然通过上述融合策略在多模态事件抽取任务上取得了较好的成绩[13,17], 但它们的融合策略并没有考虑事件抽取任务具有事件同指、事件结构复杂等特点, 这将导致融合图片信息后事件抽取性能提升不及预期, 其原因可能是在构建多模态公共语义空间时, 只考虑了模态间实体之间的对齐, 而没有考虑实体之间的关系、实体位置以及常识推理等信息, 而这些细粒度信息往往决定事件的论元信息. 因此, 针对多模态事件抽取等具有复杂结构的多模态任务, 我们可能需要增加细粒度的融合策略, 如采用对比学习方法构建公共空间时, 设计负例的同时需要考虑除实体之外的其他方面不相关的多种情况.

● 面向多模态信息抽取各子任务之间的推理知识. 在多模态数据集中, 针对简单的动作场景, 图片与文本存在直观的对应关系, 可简单的通过计算两种模态的相似度判断它们是否相关, 进而对两种模态进行对齐和融合操作; 而对于复杂的应用场景, 如自然灾害的场景中, 文本可能描述的是地震、暴风等事件, 而对应的图片则呈现的是废墟、捐赠或者被损坏的房屋等画面, 此时, 需要引入外部知识来推理文本和对应图片的相关性. 目前, 虽然有学者提出基于大模型的方法尝试解决上述问题[57], 但还不确定这些大模型否具备信息抽取各种子任务之间的推理能力, 如: 句子中的某个实体有很大可能是一个事件的论元; 两个实体是同学关系, 则有很大可能发生毕业、上课等事件等, 针对这一问题, 可考虑多任务架构和强化学习组成一个新型的信息抽取模型, 比如在考虑事件抽取任务时, 利用其他非事件抽取任务的输出作为奖励机制来促进事件抽取模型的性能优化. 这类模型很少在信息抽取领域出现, 但近期在其他任务上已出现类似的设计思路.

● 面向开放域的多模态信息抽取任务. 针对多模态信息抽取任务, 主流的模型基本都是在限定域类别做分类任务[13,17], 虽然在部分子任务上已分出 Other 类别[21,25], 但是这只是将可能存在的其他类型粗糙的划分为 Other 类, 即使通过某种方法提高了模型性能, 仍然需要人工干预来解决 Other 类里的模糊问题. 针对这一问题, 利用多模态信息处理技术实现实体、关系以及事件等体系的自动认知, 将是一项极富挑战且具趣味性的工作. 特别地, 利

用图像的检索和匹配技术, 然后结合外部知识将有助于实现更多可能分类.

7 总 结

近年来, 随着深度学习技术的快速发展, 多模态信息抽取任务迎来研究者们的广泛关注. 本文主要梳理了近 6 年来多模态信息抽取任务相关的重要文章, 详细阐述了多模态信息抽取的研究进程中, 针对短文本部分内容缺失、图文交互不充分、图文不相关可能引入噪声等问题的解决方法. 进一步的, 本文以任务为导向, 将多模态信息抽取任务的研究内容分解为多模态表示和融合、MNER、MERE 以及 MEE 这 4 个部分, 然后分别针对这 4 个部分的方法进行了分析. 最后, 总结了多模态信息抽取任务的研究趋势, 并对多模态信息抽取的研究方向进行了展望, 希望能给相关领域的研究者提供参考.

References:

[1] Zhang YZ, Rong L, Song DW, Zhang P. A survey on multimodal sentiment analysis. Pattern Recognition and Artificial Intelligence, 2020, 33(5): 426–438 (in Chinese with English abstract). [doi: 10.16451/j.cnki.issn1003-6059.202005005]

[2] Bao XG, Zhou CL, Xiao KJ, Qin B. Survey on visual question answering. Ruan Jian Xue Bao/Journal of Software, 2021, 32(8): 2522–2544 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6215.htm [doi: 10.13328/j.cnki.jos.006215]

[3] Zhang TT, Whitehead S, Zhang HW, Li HZ, Ellis J, Huang LF, Liu W, Ji H, Chang SF. Improving event extraction via multimodal integration. In: Proc. of the 25th ACM Int'l Conf. on Multimedia. Mountain: ACM, 2017. 270–278. [doi: 10.1145/3123266.3123294]

[4] Lu D, Neves L, Carvalho V, Zhang N, Ji H. Visual attention model for name tagging in multimodal social media. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Melbourne: ACL, 2018. 1990–1999. [doi: 10.18653/v1/P18-1185]

[5] Wu YZ, Li HR, Yao T, He XD. A survey of multimodal information processing frontiers: Application, fusion and pre-training. Journal of Chinese Information Processing, 2022, 36(5): 1–20 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2022.05.001]

[6] Moon S, Neves L, Carvalho V. Multimodal named entity recognition for short social media posts. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: ACL, 2018. 852–860. [doi: 10.18653/v1/N18-1078]

[7] Zhang Q, Fu JL, Liu XY, Huang XJ. Adaptive co-attention network for named entity recognition in tweets. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence and the 30th Innovative Applications of Artificial Intelligence Conf. and the 8th Symp. on Educational Advances in Artificial Intelligence. New Orleans: AAAI, 2018. 5674–5681. [doi: 10.1609/aaai.v32i1.11962]

[8] Jia MHZ, Shen X, Shen L, Pang JH, Liao LJ, Song Y, Chen M, He XD. Query prior matters: A MRC framework for multimodal named entity recognition. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisboa: ACM, 2022. 3549–3558. [doi: 10.1145/3503161.3548427]

[9] Chen DW, Li ZX, Gu BB, Chen ZG. Multimodal named entity recognition with image attributes and image knowledge. Database systems for advanced applications. In: Proc. of the 26th Int'l Conf. on Database Systems for Advanced Applications. Taipei: Springer, 2021. 186–201. [doi: 10.1007/978-3-030-73197-7_12]

[10] Eberts M, Ulges A. Span-based joint entity and relation extraction with Transformer pre-training. In: Proc. of the 24th European Conf. on Artificial Intelligence. Santiago de Compostela: IOS Press, 2020. 2006–2013. [doi: 10.3233/FAIA200321]

[11] Wan H, Zhang MR, Du JF, Huang ZL, Yang YF, Pan JZ. FL-MSRE: A few-shot learning based approach to multimodal social relation extraction. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 13916–13923. [doi: 10.1609/aaai.v35i15.17639]

[12] Zhang TM, Zhang S, Liu X, Cao B, Fan J. Multimodal data fusion for few-shot named entity recognition method. Ruan Jian Xue Bao/Journal of Software, 2024, 35(3): 1107–1124 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7069.htm [doi: 10.13328/j.cnki.jos.007069]

[13] Li ML, Zareian A, Zeng Q, Whitehead S, Lu D, Ji H, Chang SF. Cross-media structured common space for multimedia event extraction. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 2557–2568. [doi: 10.18653/v1/2020.acl-main.230]

[14] Liu J, Chen YF, Xu JN. Multimedia event extraction from news with a unified contrastive learning framework. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisboa: ACM, 2022. 1945–1953. [doi: 10.1145/3503161.3548132]

[15] Zhang D, Wei SZ, Li SS, Wu HQ, Zhu QM, Zhou GD. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 14347–14355. [doi: 10.1609/aaai.v35i16.17687]

[16] Zheng CM, Wu ZW, Feng JH, Fu Z, Cai Y. MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In: Proc. of the 2021 IEEE Int'l Conf. on Multimedia and Expo (ICME). Shenzhen: IEEE, 2021. 1–6. [doi: 10.1109/ICME51207.2021.9428274]

[17] Tong MH, Wang S, Cao YX, Xu B, Li JZ, Hou L, Chua TS. Image enhanced event detection in news articles. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 9040–9047. [doi: 10.1609/aaai.v34i05.6437]

[18] Chiu JPC, Nichols E. Named entity recognition with bidirectional LSTM-CNNs. Trans. of the Association for Computational Linguistics, 2016, 4: 357–370. [doi: 10.1162/tacl_a_00104]

[19] Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016. 260–270. [doi: 10.18653/v1/N16-1030]

[20] Li J, Sun AX, Han JL, Li CL. A survey on deep learning for named entity recognition. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(1): 50–70. [doi: 10.1109/TKDE.2020.2981314]

[21] Adel H, Schütze H. Global normalization of convolutional neural networks for joint entity and relation classification. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 1723–1729. [doi: 10.18653/v1/D17-1181]

[22] Ahn D. The stages of event extraction. In: Proc. of the 2006 Workshop on Annotating and Reasoning about Time and Events. Sydney: ACL, 2006. 1–8.

[23] Zhang RJ, Dai L, Wang B, Guo P. Recent advances of Chinese named entity recognition based on deep learning. Journal of Chinese Information Processing, 2022, 36(6): 20–35 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2022.06.002]

[24] Arshad O, Gallo I, Nawaz S, Calefati A. Aiding intra-text representations with visual context for multimodal named entity recognition. In: Proc. of the 2019 Int'l Conf. on Document Analysis and Recognition (ICDAR). Sydney: IEEE, 2019. 337–342. [doi: 10.1109/ICDAR.2019.00061]

[25] Wu ZW, Zheng CM, Cai Y, Chen JY, Leung HF, Li Q. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 1038–1046. [doi: 10.1145/3394171.3413650]

[26] Yu JF, Jiang J, Yang L, Xia R. Improving multimodal named entity recognition via entity span detection with unified multimodal Transformer. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 3342–3352. [doi: 10.18653/v1/2020.acl-main.306]

[27] Wang XY, Gui M, Jiang Y, Jia ZX, Bach N, Wang T, Huang ZQ, Huang F, Tu KW. ITA: Image-text alignments for multi-modal named entity recognition. In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: ACL, 2022. 3176–3189. [doi: 10.18653/v1/2022.naacl-main.232]

[28] Xu B, Huang SZ, Sha CF, Wang HY. MAF: A general matching and alignment framework for multimodal named entity recognition. In: Proc. of the 15th ACM Int'l Conf. on Web Search and Data Mining. ACM, 2022. 1215–1223. [doi: 10.1145/3488560.3498475]

[29] Sun L, Wang JQ, Zhang K, Su YD, Weng FS. RpBERT: A text-image relation propagation-based BERT model for multimodal NER. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 13860–13868. [doi: 10.1609/aaai.v35i15.17633]

[30] Huang SZ. Research on general multimodal information extraction for social media [MS. Thesis]. Shanghai: Donghua University, 2022 (in Chinese with English abstract). [doi: 10.27012/d.cnki.gdhuu.2022.002241]

[31] Zheng CM, Feng JH, Fu Z, Cai Y, Li Q, Wang T. Multimodal relation extraction with efficient graph alignment. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 5298–5306. [doi: 10.1145/3474085.3476968]

[32] Sun L, Wang JQ, Su YD, Weng FS, Sun YX, Zheng ZW, Chen YY. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: ACL, 2020. 1852–1862. [doi: 10.18653/v1/2020.coling-main.168]

[33] Zhao F, Li CH, Wu Z, Xing SY, Dai XY. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal NER. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisboa: ACM, 2022. 3983–3992. [doi: 10.1145/3503161.354822]

[34] Zheng CM, Wu ZW, Wang T, Cai Y, Li Q. Object-aware multimodal named entity recognition in social media posts with adversarial learning. IEEE Trans. on Multimedia, 2021, 23: 2520–2532. [doi: 10.1109/TMM.2020.3013398]

[35] Li XY, Feng JR, Meng YX, Han QH, Wu F, Li JW. A unified MRC framework for named entity recognition. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 5849–5859. [doi: 10.18653/v1/2020.acl-main.519]

[36] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. 2015. [doi: 10.48550/arXiv.1409.1556]

[37] Yatskar M, Zettlemoyer L, Farhadi A. Situation recognition: Visual semantic role labeling for image understanding. In: Proc. of the 2016

IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 5534–5542. [doi: 10.1109/CVPR.2016.597]

[38] Lahat D, Adali T, Jutten C. Multimodal data fusion: An overview of methods, challenges, and prospects. Proc. of the IEEE, 2015, 103(9): 1449–1477. [doi: 10.1109/JPROC.2015.2460697]

[39] Wang XW, Ye JB, Li ZX, Tian JF, Jiang Y, Yan M, Zhang J, Xiao YH. CAT-MNER: Multimodal named entity recognition with knowledge-refined cross-modal attention. In: Proc. of the 2022 IEEE Int'l Conf. on Multimedia and Expo (ICME). Taipei: IEEE, 2022. 1–6. [doi: 10.1109/ICME52920.2022.9859972]

[40] Chen X, Zhang NY, Li L, Yao YZ, Deng SM, Tan CQ, Huang F, Si L, Chen HJ. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In: Findings of the Association for Computational Linguistics: NAACL 2022. Seattle: ACL, 2022. 1607–1618. [doi: 10.18653/v1/2022.findings-naacl.121]

[41] Xu B, Huang SZ, Du M, Wang HY, Song H, Sha CF, Xiao YH. Different data, different modalities! Reinforced data splitting for effective multimodal information extraction from social media posts. In: Proc. of the 29th Int'l Conf. on Computational Linguistics. Gyeongju: ACL, 2022. 1855–1864.

[42] Sang EFTK, Veenstra J. Representing text chunks. In: Proc. of the 9th Conf. on European Chapter of the Association for Computational Linguistics. Bergen: ACL, 1999. 173–179. [doi: 10.3115/977035.977059]

[43] Li JY, Li H, Pan Z, Sun D, Wang JH, Zhang WK, Pan G. Prompting ChatGPT in MNER: Enhanced multimodal named entity recognition with auxiliary refined knowledge. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: ACL, 2023. 2787–2802. [doi: 10.18653/v1/2023.findings-emnlp.184]

[44] Ji YZ, Li BB, Zhou J, Li F, Teng C, Ji DH. CMNER: A Chinese multimodal NER dataset based on social media. arXiv:2402.13693, 2024.

[45] Wang JM, Li ZY, Yu JF, Yang L, Xia R. Fine-grained multimodal named entity recognition and grounding with a generative framework. In: Proc. of the 31st ACM Int'l Conf. on Multimedia. Ottawa: ACM, 2023. 3934–3943. [doi: 10.1145/3581783.3612322]

[46] Sui DB, Tian ZK, Chen YB, Liu K, Zhao J. A large-scale Chinese multimodal ner dataset with speech clues. 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. 2807–2818. [doi: 10.18653/v1/2021.acl-long.218]

[47] Wang XY, Cai J, Jiang Y, Xie PJ, Tu KW, Lu W. Named entity and relation extraction with multi-modal retrieval. In: Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi: ACL, 2022. 5925–5936. [doi: 10.18653/v1/2022.findings-emnlp.437]

[48] Lu JY, Zhang DX, Zhang JX, Zhang PJ. Flat multi-modal interaction transformer for named entity recognition. In: Proc. of the 29th Int'l Conf. on Computational Linguistics. Gyeongju: ACL, 2022. 2055–2064.

[49] Miyamoto Y, Cho K. Gated word-character recurrent language model. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Austin: ACL, 2016. 1992–1997. [doi: 10.18653/v1/D16-1209]

[50] Shen T, Zhou TY, Long GD, Jiang J, Pan SR, Zhang CQ. DiSAN: Directional self-attention network for RNN/CNN-free language understanding. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 5446–5455. [doi: 10.1609/aaai.v32i1.11941]

[51] Vempala A, Preoţiuc-Pietro D. Categorizing and inferring the relationship between the text and image of twitter posts. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 2830–2840. [doi: 10.18653/v1/P19-1272]

[52] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. 2021. 8748–8763.

[53] Su WJ, Zhu XZ, Cao Y, Li B, Lu LW, Wei FR, Dai JF. VL-BERT: Pre-training of generic visual-linguistic representations. In: Proc. of the 8th Int'l Conf. on Learning Representations. 2020. 1–16.

[54] Zhou BH, Zhang Y, Song KH, Guo WY, Zhao GQ, Wang HB, Yuan XJ. A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 6293–6302. [doi: 10.18653/v1/2022.emnlp-main.422]

[55] Hinton GE. Training products of experts by minimizing contrastive divergence. Neural Computation, 2002, 14(8): 1771–1800. [doi: 10.1162/089976602760128018]

[56] Yamada I, Asai A, Shindo H, Takeda H, Matsumoto Y. LUKE: Deep contextualized entity representations with entity-aware self-attention. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2020. 6442–6454. [doi: 10.18653/v1/2020.emnlp-main.523]

[57] Chen F, Feng Yj. Chain-of-thought prompt distillation for multimodal named entity recognition and multimodal relation extraction. arXiv:2306.14122, 2023.

[58] Hu XM, Chen JZ, Liu AW, Meng SA, Wen LJ, Yu PS. Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In: Proc. of the 31st ACM Int'l Conf. on Multimedia. Ottawa: ACM, 2023. 5185–5194. [doi: 10.1145/3581783.3611899]

[59] Zhao S, Hu MH, Cai ZP, Liu F. Modeling dense cross-modal interactions for joint entity-relation extraction. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. 2021. 4032–4038. [doi: 10.24963/ijcai.2020/558]

[60] Zheng SC, Wang F, Bao HY, Hao YX, Zhou P, Xu B. Joint extraction of entities and relations based on a novel tagging scheme. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017. 1227–1236. [doi: 10.18653/v1/P17-1113]

[61] Zhong ZX, Chen DQ. A frustratingly easy approach for entity and relation extraction. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 50–61. [doi: 10.18653/v1/2021.naacl-main.5]

[62] Nguyen TH, Grishman R. Relation extraction: Perspective from convolutional neural networks. In: Proc. of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Denver: ACL, 2015. 39–48. [doi: 10.3115/v1/W15-1506]

[63] Soares LB, Fitzgerald N, Ling J, Kwiatkowski T. Matching the blanks: Distributional similarity for relation learning. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 2895–2905. [doi: 10.18653/v1/P19-1279]

[64] Zeng DJ, Liu K, Chen YB, Zhao J. Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 1753–1762. [doi: 10.18653/v1/D15-1203]

[65] Du JF, Pan JZ, Wang S, Qi KX Shen YM, Deng Y. Validation of growing knowledge graphs by abductive text evidences. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 2784–2791. [doi: 10.1609/aaai.v33i01.33012784]

[66] Du YJ, Su FH, Yang AZ, Li XY, Fan YQ. Extracting deep personae social relations in microblog posts. IEEE Access, 2020, 8: 5488–5501. [doi: 10.1109/ACCESS.2019.2960659]

[67] Zhang ZP, Luo P, Loy CC, Tang XO. Learning social relation traits from face images. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 3631–3639. [doi: 10.1109/ICCV.2015.414]

[68] Mori J, Ishizuka M, Matsuo Y. Extracting keyphrases to represent relations in social networks from Web. In: Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence. 2007. 2820–2827.

[69] Bramsen P, Escobar-Molano M, Patel A, Alonso R. Extracting social power relationships from natural language. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: ACL, 2011. 773–782.

[70] Du ZL, Li YX, Guo X, Sun YD, Li BY. Training multimedia event extraction with generated images and captions. In: Proc. of the 31st ACM Int'l Conf. on Multimedia. Ottawa: ACM, 2023. 5504–5513. [doi: 10.1145/3581783.3612526]

[71] 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]

[72] 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]

[73] Wadden D, Wennberg U, Luan Y, Hajishirzi H. Entity, relation, and event extraction with contextualized span representations. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 5784–5789. [doi: 10.18653/v1/D19-1585]

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

[75] Li H, Ji H, Deng HB, Han JW. Exploiting background information networks to enhance bilingual event extraction through topic modeling. In: Proc. of the 1st Int'l Conf. on Advances in Information Mining and Management. 2011. 23–30.

[76] Hong Y, Zhang JF, Ma B, Yao JM, Zhou GD, Zhu QM. Using cross-entity inference to improve event extraction. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: ACL, 2011. 1127–1136.

[77] Li Q, Ji H, Huang L. Joint event extraction via structured prediction with global features. In: Proc. of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia: ACL, 2013. 73–82.

[78] 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.

[79] Feng XC, Huang LF, Tang DY, Ji H, Qin B, Liu T. A language-independent neural network for event detection. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016. 66–71. [doi: 10.18653/v1/P16-2011]

[80] 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]

[81] Bosselut A, Chen JF, Warren D, Hajishirzi H, Choi Y. Learning prototypical event structure from photo albums. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Berlin: ACL, 2016. 1769–1779. [doi: 10.18653/v1/P16-1167]

[82] Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. of the Association for Computational Linguistics, 2014, 2: 67–78. [doi: 10.1162/tacl_a_00166]

[83] Song ZY, Bies A, Strassel S, Riese T, Mott J, Ellis J, Wright J, Kulick S, Ryant N, Ma XY. From light to rich ERE: Annotation of entities, relations, and events. In: Proc. of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. Denver: ACL, 2015. 89–98. [doi: 10.3115/v1/W15-0812]

[84] Li ML, Xu RC, Wang SH, Zhou LW, Lin XD, Zhu CG, Zeng M, Ji H, Chang SF. Clip-event: Connecting text and images with event structures. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 16399–16408. [doi: 10.1109/CVPR52688.2022.01593]

[85] Moghimifar F, Shiri F, Nguyen V, Li YF, Haffari G. Theia: Weakly supervised multimodal event extraction from incomplete data. In: Proc. of the 13th Int'l Joint Conf. on Natural Language Processing and the 3rd Conf. of the Asia-Pacific Chapter of the Association for Computational Linguistics. Nusa Dua: ACL, 2023. 139–145. [doi: 10.18653/v1/2023.ijcnlp-short.16]

[86] Chen B, Lin XD, Thomas C, Li ML, Yoshida S, Chum L, Ji H, Chang SF. Joint multimedia event extraction from video and article. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana: ACL, 2021. 74–88. [doi: 10.18653/v1/2021.findings-emnlp.8]

附中文参考文献:

[1] 张亚洲, 戎璐, 宋大为, 张鹏. 多模态情感分析研究综述. 模式识别与人工智能, 2020, 33(5): 426–438. [doi: 10.16451/j.cnki.issn1003-6059.202005005]

[2] 包希港, 周春来, 肖克晶, 覃飙. 视觉问答研究综述. 软件学报, 2021, 32(8): 2522–2544. http://www.jos.org.cn/1000-9825/6215.htm [doi: 10.13328/j.cnki.jos.006215]

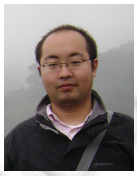
[5] 吴友政, 李浩然, 姚霆, 何晓冬. 多模态信息处理前沿综述: 应用、融合和预训练. 中文信息学报, 2022, 36(5): 1–20. [doi: 10.3969/j.issn.1003-0077.2022.05.001]

[12] 张天明, 张杉, 刘曦, 曹斌, 范菁. 融合多模态数据的小样本命名实体识别方法. 软件学报, 2024, 35(3): 1107–1124. http://www.jos.org.cn/1000-9825/7069.htm [doi: 10.13328/j.cnki.jos.007069]

[23] 张汝佳, 代璐, 王邦, 郭鹏. 基于深度学习的中文命名实体识别最新研究进展综述. 中文信息学报, 2022, 36(6): 20–35. [doi: 10.3969/j.issn.1003-0077.2022.06.002]

[30] 黄世洲. 面向社交媒体的通用多模态信息抽取方法研究 [硕士学位论文]. 上海. 东华大学, 2022. [doi: 10.27012/d.cnki.gdhuu.2022.002241]

王永胜(1990－), 男, 博士生, 主要研究领域为自然语言处理, 信息抽取.



王中卿(1987－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为自然语言处理.

李培峰(1971－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 机器学习.

朱巧明(1963－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为中文信息处理, Web 信息处理.