计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220300099

基于深度学习的社交网络舆情信息抽取方法综述

王　剑[1]　彭雨琦[2]　赵宇斐[2]　杨　健[3]
1 郑州大学计算机与人工智能学院、软件学院　郑州 450000
2 郑州大学网络空间安全学院　郑州 450000
3 云南省智慧城市网络空间安全重点实验室　云南 玉溪 653100

摘　要　随着社交媒体平台的快速发展，舆情信息得以在极短的时间内大范围传播，如果不对舆情信息加以管理和控制，将对网络环境乃至社会环境造成巨大威胁。信息抽取技术因其语义化和精准性成为舆情分析和管理的第一步，也是最关键的一步。近年来，随着深度学习的发展，其自动学习潜在特征、组合特征的能力使信息抽取各个子任务的准确率都得到了很大的提高。文中结合社交网络舆情的特点和深度学习技术在信息抽取领域的应用，对基于深度学习的社交网络舆情信息抽取方法进行了系统的梳理和总结。首先整理了社交网络舆情信息的组织方式，详细阐述了舆情信息抽取的框架、评价指标，然后对现有的基于深度学习的舆情信息抽取模型进行了全面的回顾和分析，讨论了现有方法的适用性及局限性，最后对未来的研究趋势进行了展望。

关键词：社交网络；社交媒体；舆情信息；信息抽取；深度学习
中图法分类号　TP391

Survey of Social Network Public Opinion Information Extraction Based on Deep Learning

WANG Jian[1], PENG Yu-qi[2], ZHAO Yu-fei[2] and YANG Jian[3]
1 School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450000, China
2 School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450000, China
3 Yunnan Key Laboratory of Smart City in Cyberspace Security, Yuxi Normal University, Yuxi, Yunnan 653100, China

Abstract　With the rapid development of social media platforms, public opinion information can be widely disseminated in a very short period of time. If the information of public opinion is not managed and controlled, it will pose a great threat to the network environment and even the social environment. Information extraction technology has become the first and the most significant step in public opinion analysis and management due to its semantization and accuracy. Over the last few years, with the development of deep learning, its ability to automatically learn potential features and combine these features has dramatically improved the accuracy of each sub-task of information extraction. This paper systematically composes and summarizes the methods of extracting information by combining the characteristics of social media public opinion and deep learning technology. Firstly, we sort out the organization of public opinion information in social networks, elaborate the framework and evaluation indexes of public opinion information extraction. Then we conduct a comprehensive review and analysis of existing deep learning-based public opinion information extraction models, discuss the applicability and limitations of existing methods. Finally, the future research trends is prospected.

Keywords　Social network, Social media, Public opinion, Information extraction, Deep learning

1　引言

随着互联网技术和社交媒体平台的发展，网民可以随时随地通过自己的社交软件参与舆情事件的讨论、发表观点、表达态度。然而，一些舆情事件被有预谋的个人、组织机构经过社交媒体平台添油加醋或者颠倒黑白地宣传后，成为了网民发泄情绪的导火索，不断发酵后形成负面舆情或者谣言，使社会冲突日益尖锐，不利于国家的长治久安。社交网络舆情是网民观点、态度以及情绪的集合，近年来已有大量学者[1-3]结合多维度数据挖掘、自然语言处理、知识图谱等技术进行舆情信息管理，对舆情进行监控和可视化分析，及时发现舆情热点，对舆情信息提前预警、控制以及引导，从而避免大范围的

到稿日期：2022-03-10　**返修日期**：2022-04-10
基金项目：国家自然科学基金(61972133)；云南省智慧城市网络空间安全重点实验室开放课题项目(202105AG070010)
This work was supported by the National Natural Science Foundation of China(61972133) and Opening Foundation of Yunnan Key Laboratory of Smart City in Cyberspace Security(202105AG070010).
通信作者：王剑(iejwang@zzu.edu.cn)

舆情危机和负面舆情的发展，达到节约网络资源、净化网络环境的目的。但是，当前新闻网站和社交媒体中充斥着大量具有歧义性、非规范性的非结构化信息，同时自然语言还蕴含着大量知识积累以及思维推理过程，因此快速、精准且有效地对海量的社交数据进行舆情信息抽取并将其转换为可以直接查询的结构化信息是进行舆情分析的第一步，也是关键的一步，这也是当前舆情分析和知识图谱交叉领域的研究热点。

信息抽取(Information Extraction，IE)指从结构化、半结构化、非结构化的文本数据中抽取出特定的事实[4]，这些信息常用三元组〈h：头实体，R：关系，t：尾实体〉的形式来进行结构化描述，以便于构建知识图谱或者直接查询，其中 h 代表主实体，t 代表客实体，R 代表两实体之间的关系。信息抽取根据文本所属领域的不同分为限定领域的信息抽取和开放域的信息抽取。限定领域的信息抽取限制关系类别，如金融、医疗、教育等领域，而开放域的信息抽取旨在抽取所有可能的三元组，这些三元组可以用于知识图谱的构建、问答系统、信息检索等任务。

信息抽取技术涉及信息检索、自然语言处理、数据挖掘等多种技术，主要包括 4 个子任务：实体抽取、实体链接、关系抽取和事件抽取。实体抽取识别文本中专有名称和有意义的数量短语并加以分类，但是这样抽取出的实体可能会存在歧义或多个实体描述同一个语义的情况，这就需要将实体链接到已有的知识库或知识图谱中进行实体消歧；关系抽取根据实体抽取的输出识别实体之间的语义关系并将关系分类，构成〈h，R，t〉这样的三元组；事件抽取相当于一种多元关系抽取，通过触发词识别事件类型，抽取出事件要素(包括实体和时间)，并判别其在事件中的角色，如“郑州大学的学生在上个月摧毁了一台计算机”中触发词为“摧毁”，事件要素为{学生，计算机}，要素角色为学生{$role$=“破坏者”}、计算机{$role$=“目标”}、上个月{$role$=“时间”}、郑州大学{$role$=“地点”}。

信息抽取方法分为基于规则的模式匹配和基于机器学习的方法。基于规则的信息抽取方法依赖于手工设计的模板，需要较强的专业知识，效率低下且会耗费大量人力和财力；基于机器学习的方法将任务转化为聚类、分类或者序列标注的问题，选取特征训练模型虽然提高了信息抽取的效率，但存在误差传播的问题，会影响信息抽取的效果，另一方面，基于机器学习的信息抽取需要大规模的标注语料，否则会出现严重的数据稀疏的问题。近年来，深度学习技术的发展为信息抽取提供了行之有效的解决方案，其基本优点是使用了分布式表示，可以为单词、短语提供精确的匹配。除此之外，基于深度学习的模型可以通过非线性激活函数从输入中自动学习特征表示，学习隐藏的特征，避免了特征工程中的误差积累和传播问题，使信息抽取各个子任务的准确率都得到了很大的提高。

大数据环境下的舆情具有自由性、实时性、隐匿性、互动性、多元性、广泛性、非理性、突发性和高效性的特点[5]，舆情相关数据具有跨平台、多模态和噪声大的特点[6]，除此之外，舆情分析涉及新闻学、社会学、心理学、计算机科学等众多学科，基于深度学习的信息抽取由于其自动学习特征表示等优点自然而然地成为了有效的研究工具。本文与现有的介绍信息抽取的综述文章不同，不仅根据舆情属性总结了舆情信息的组织方式，还将基于深度学习方法提升信息抽取任务性能的研究成果作为重点梳理对象，对信息抽取的每个子任务按步骤进行分类，结合舆情特点整理其中的关键问题和解决方法。本文第 2 节介绍了舆情信息组织方式、舆情信息抽取框架、常用语料库以及评价指标；第 3 节分别介绍了深度学习技术应用于信息抽取 4 个子任务的优势以及不同的方法使用的网络结构；第 4 节介绍了舆情信息抽取的现存问题与未来发展趋势；最后总结全文。

2 舆情信息抽取

利用信息抽取技术进行社交网络舆情及舆情事件的监管和分析是目前网络舆情学术界和产业届共同关注的研究热点，涌现出了许多有用的研究成果。例如，文献[7]通过分析评论中高频词的随机网络结构，建立评论的实体以及情感倾向的识别规则，最终构成舆情知识图谱，以供决策者快速掌握网络舆情；文献[8]将天气预报和社交媒体的事件检测相结合，近乎实时地评估极端天气导致的突发事件的演变及在此基础之上发酵的舆情，以便提取可能对公民、急救人员和决策者有用的信息；文献[9]将负面推文中的实体类型被提及的次数进行排序，对新冠疫情期间网民的情绪进行深刻分析；文献[10]抽取中国社交媒体平台关于政策的评论信息，构建情感知识图谱，为地方政府了解舆情从而及时调整各项政策提供了理论依据。

2.1 舆情信息组织方式

社交媒体中的信息不仅嘈杂，还涉及政治、经济、科技、教育、文化、卫生等众多领域。除此之外，由于舆情信息较短且缺少标注数据，因此舆情信息抽取应结合特定领域和开放领域的信息抽取方法。为了对舆情有一个全面的把握，更好地将抽取到的信息组织成为三元组形式，便于后续的信息存储与分析，本文通过总结最近的研究成果以及舆情特点，根据舆情主体、客体、本体 3 个方面将舆情信息的抽取分为以用户信息为中心、以事件信息为中心和以情感信息为中心的信息组织方式，使用信息抽取技术，将抽取出来的信息统一以三元组〈h，R，t〉进行形式化描述，相关舆情实体和实体关系类别如表 1 和表 2 所列。

表 1 舆情实体类别

Table 1 Public opinion entity categories

舆情主体	舆情客体	舆情本体	舆情载体
民众	食品/卫生/突发事件等	用户评论/态度/情感倾向	网站/公众号/社交软件

表 2 舆情实体关系类别

Table 2 Public opinion entity relation categories

事件主体与事件	事件与事件	用户与用户
发起/关注/参与	引发/导致	关注/评论/转发/点赞

2.1.1 以用户信息为中心的组织方式

以用户信息为中心的信息抽取出来后可以被组织为〈h：用户 1，R：关注/转评赞，t：用户 2〉的形式，其中 h，t={用户/意见领袖/官方媒体}。网络舆情的参与主体是社交网络中的用户，这些用户在舆情传播中既是接收者也是传播者，构成了

社交网络中的节点，其认知能力的差异和对信息的理解程度容易受意见领袖的影响，从而导致观点两极分化，推动舆情不断演化，最终达到不可控制的地步。因此，以用户信息为主体的信息抽取方法旨在抽取用户实体以及用户关系，在此基础之上结合社会网络分析法对舆情信息的传播进行分析，可以进一步挖掘意见领袖以及预测传播路径。例如，Wei 等[11]围绕“重庆万州公交车坠江”事件，抽取发表评论的用户及评论内容信息，构建主题图谱，分析推动舆情事件发展的核心用户及其相互关系；Chen 等[12]以实体抽取和关系构建技术为基础，构建了网络舆情主题图谱模型，分析了用户影响力和舆情演化趋势。

2.1.2　以事件信息为中心的组织方式

在网络舆情的研究中，与公众利益有关的公共事件是舆情客体，网络舆情变化通常是由热点社会事件引起的，公众在社交网络中对事件表达的意见往往能够推动事件的发展和演化。以事件为中心的舆情信息抽取是近年来的研究热点，旨在从非结构化的信息中抽取出用户感兴趣的事件，根据事件的要素、类型和要素角色可以将事件信息组织为〈h：事件 1，R：导致/引发/造成，t：事件 2〉或〈h：人物，R：动作，t：事件〉的形式，可以用于构建事理图谱[13-14]，分析事件之间的时序关系和因果关系，从而动态追踪事件情节和舆情演化。社交网络舆情中的事件元素如下。

（1）触发词（Event Trigger）：表示事件发生的核心词，多为动词或名词。

（2）事件类型（Event Type）：表示事件所属类别。

（3）事件要素（Event Argument）：表示事件的参与元素。

（4）要素角色（Argument Role）：指事件要素在事件中扮演的角色[15]。

2.1.3　以情感信息为中心的组织方式

以情感信息为中心抽取出的信息可组织为〈h：情感实体，R：情感倾向，t：评价对象实体〉或〈h：情感实体 1，R：情感倾向，t：情感实体 2〉的形式。公众对公共事件的评价性意见是舆情本体，具有情感倾向的观点，尤其是具有愤怒和消极情绪的言论更容易被传播和转发，从而引发大范围的舆情危机。鉴别舆情发展趋势的一个重要切入点便是情感分析，因此使用自然语言处理技术对社交网络文本进行分析，进而将情感词、情感关系抽取出来组成三元组形式并用于舆情分析是一个很好的思路。例如，Qiu 等[16]针对突发事件中负面网络舆情恶性传播的问题，通过分析用户的情感倾向和影响力综合构建突发事件情感图谱，将突发事件各个时期的情感图谱进行可视化分析，致力于降低消极情感对社会造成的影响。

2.2　舆情信息抽取框架

上文介绍了舆情信息的抽取目标和组织方式，本节将介绍舆情信息抽取的框架。信息抽取包括 4 个子任务，即实体抽取、实体链接、关系抽取和事件抽取，但是这 4 个子任务间并没有严格的顺序关系，现有模型大多采用两种设计路线。

（1）流水线（Pipeline）方式，即先进行实体抽取，将实体两两配对再进行关系分类或者分析句子中是否存在事件，最后进行事件抽取。

（2）联合（Joint）抽取方式，将实体抽取结合关系抽取或者事件抽取共同编码，并采用同一个模型进行端到端的训练。

流水线方法虽然使任务处理变得简单，但是忽略了各个任务之间的关系，导致了误差累积（即实体抽取的结果不准或者缺漏，会极大地影响后续的关系抽取和事件抽取）、实体冗余（即关系分类时存在多余的候选实体，这些实体之间没有关系，会提升错误率）等问题，而联合抽取模型缓解了误差累积的问题，被认为比流水线方法更优越，但是其训练和推断时的误差不容忽视，因此在设计模型时应结合具体场景选择合适的框架。具体的舆情信息抽取框架如图 1 所示。

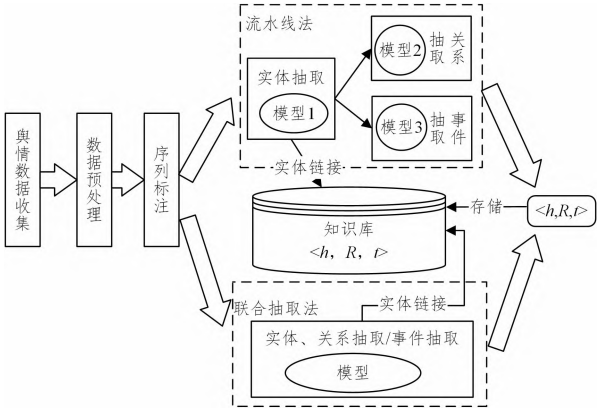

图 1　舆情信息抽取框架图

Fig. 1　Framework of public opinion information extraction

2.2.1　舆情实体抽取

实体抽取也称为命名实体识别（Named Entity Recognization，NER），其任务是在给定的文本中识别不同类型的实体，如人名、地名、公司、位置、时间等，NER 是关系抽取、事件抽取的基础，也是知识图谱构建、问答系统搭建等下游任务的第一步，具体包括实体边界检测和实体分类两个子任务。总体任务可以描述为输入一个标记序列 $s=\langle w_1, w_2, \cdots, w_n\rangle$，输出一个三元组〈$I_s$，$I_e$，$t$〉的列表，列表中每个三元组代表 s 的一个命名实体。其中，$I_s \in [1, N]$为命名实体的起始索引；$I_e \in [1, N]$为结束索引；t 指从预定义类别中选择实体类型。

社交网络中的舆情实体抽取技术是一项具有挑战性的任务，原因如下。

（1）构词灵活：首先实体可能由多个词组成，具有组合性，例如“中国北京市朝阳区”包含 3 个地名，抽取地点时会产生“中国”“北京市”“朝阳区”“中国北京市”“北京市朝阳区”5 种抽取方案；其次实体有多种表达方式，如“中国首都”和“北京市”都表示北京市。

（2）实体类型多样：社交网络每天产生的数据呈井喷式增长，涉及广泛的领域，人们接触到的命名实体是一个开放的集合，需要不断更新。

（3）类别模糊：如地名和机构名，有些地名本身就是机构；对新词的边界把握也很模糊。

值得注意的是，中文社交网络中的实体抽取相比英文实体抽取需要额外的分词步骤，再加上中文社交媒体平台文本的不规范性和标注数据不足的问题，使得中文社交网络的舆情实体抽取面临更大的挑战。

2.2.2　舆情实体链接

实体链接任务研究的对象包含人名、地名和机构名在内

的命名实体，将非结构化文本中的表述指向其代表的真实世界实体，关联到对应的知识库具体实体中，主要解决实体名的歧义性和多样性问题[17]。实体链接任务总体可以描述为文档 D 包含一组已识别的实体 $M=\{M_1,M_2,\cdots,M_m\}$，目标知识库包含一组实体 $E=\{E_1,E_2,\cdots,E_n\}$，将 M 中的每个实体 M_i 链接到目标实体 E_i 中。该任务主要包括以下 3 个步骤。

(1)候选实体生成(Candidate Entity Generation, CEG)：找到一个尽可能小、尽可能包含目标实体的集合，通常采用名称字典的构建方法，其主要思想是充分利用维基百科提供的重定向页面、排歧页面、锚文本等构建实体名称与所有可能链接的实体的映射关系字典，利用字典生成候选实体集合。

(2)实体消歧(Entity Disambiguation, ED)：将实体提及(Mention)和候选实体进行匹配，与文本匹配任务类似，第一步是构造候选实体提及和候选实体的嵌入表示，因为实体提及和候选实体都是非常短的文本，蕴含的语义非常少，所以可以借助描述信息辅助匹配度的计算；第二步是利用语义表示计算候选实体的匹配度并排序。

(3)无链接指代预测(Unlinkable Mention Prediction, UMP)：知识库中不存在实体提及对应的实体的情况下将对应实体用一个“NIL”来链接。

舆情信息涉及范围较为宽泛，确保实体语义信息在特定语境下的指向唯一性是舆情信息抽取的关键，因此实体链接对于舆情信息抽取来说十分重要。由于实体链接是一个相对下游的任务，性能受限于 NER 任务的准确性，而中文的实体链接任务还受到中文分词任务的影响，上游任务的错误会带来不可避免的噪声。同时，实体间还存在多样性和歧义性问题(多样性指同一个实体对应多个名称，歧义性指同一个名称有多个含义)，给实体链接任务带来了很大的困难。

2.2.3 舆情关系抽取

舆情信息中的关系不仅包括实体之间的关系，还包括事件之间的时序关系、因果关系。由于目前事件关系抽取的研究相对较少，对于事件的因果关系抽取也是使用实体关系抽取方法，因此本文将着重介绍实体关系抽取任务。实体关系抽取的任务是识别句子中实体与实体之间的语法以及语义关系，具体可以描述为针对文档 D 中的实体 h，t，识别出关系 R 并将其组织为 $\langle h,R,t\rangle$ 的三元组结构。根据是否在同一个模型中开展实体抽取和关系分类，关系抽取可以采取流水线和联合学习两种方式[18]。

流水线式的关系抽取指先对输入的句子进行实体抽取，将识别出的实体分别组合，然后进行关系分类，这两个子过程是前后串联、完全分离的。联合学习指在一个模型中实现实体抽取和关系分类子过程。流水线学习让每一个子过程都更灵活，使关系抽取更容易，因此多数基于深度学习的关系抽取方法都默认实体对是给定的，在此基础上进行关系分类的工作。然而，此类方法忽视了两个子任务之间的关系，如“李华在华为上班”，识别出“李华”是人名实体，“华为”是公司实体后，可以确定两者具有“受雇于”的关系；同样，这句话具有“受雇于”关系可以帮助确定“李华”和“华为”的实体类别。联合学习的方法目前大多采用标注策略来实现，通过设计特定的标注策略并使两个子过程共享网络底层参数的方法来解决上述问题，取得了不错的效果。

2.2.4 舆情事件抽取

舆情事件抽取指从引起网民大范围讨论的社交媒体平台的文本中收集真实世界的事件，并以结构化的形式表示出来，然而社交媒体中的舆情事件抽取存在许多困难，如时间信息与事件发生时间存在误差，社交网络舆情事件的真实性难以检测，社交网络中的文本有字数限制而没有足够的上下文进行事件抽取等。

事件抽取相当于多元关系抽取，涉及事件的识别和分类，通常被认定为一个多分类问题，依赖于实体识别、共指消解、关系抽取等任务的结果，总体任务可以描述为将文档 D 中的每个单词进行事件分类并填入设定好的角色槽中。目前针对基于深度学习的事件抽取任务的解决方案大多是利用句子级特征、文档级特征，或者建模为其他任务，如机器阅读理解(Machine Reading Comprehension, MRC)、问答(Question Answering, QA)、序列标注(Sequence Labeling)等。事件抽取任务根据事件范围可以分为开放域和特定域的事件抽取。当没有预定义的事件模式时，开放域的事件抽取旨在通过提取事件关键字(触发词和要素)将类似的事件分类或聚类，从而获取与主题相关的一系列事件，归纳出通用的事件模式；而特定领域的事件抽取(如医疗、金融领域)可以借助与专业术语相关的知识库预定义事件模式，包括 4 个子任务，即触发词识别、触发词分类、参数识别、参数角色分类，其中事件分类和触发词识别可以合并为事件检测任务，参数识别和参数角色识别可以合并为参数提取任务。目前特定领域的事件抽取技术较为成熟，而包含大量嘈杂文本数据的社交网络舆情的事件抽取因其覆盖面广、细粒度且动态演变的特点，应当结合两者共同完成舆情事件抽取任务。

2.3 相关数据集

舆情相关数据集通常来自于用户的生成内容，包括用户评论、用户的历史推文等，研究者们通常设定关键词和时间区间，应用社交平台提供的高级搜索 API 或者遍历所有微博用户，收集指定时间段的推文后筛选出需要的数据，再通过词性标注、序列标注等策略将其进一步处理为信息抽取所需要的数据格式。因涉及用户隐私，公开的舆情语料库较为缺乏，本文搜集了部分舆情相关语料库以及适用于舆情领域的信息抽取数据集进行介绍。

(1)Weibo-COV[19]：一个持续维护的舆情语料库，来自微博平台上的与新冠疫情(COVID-19)有关的 40 893 832 条微博，包含用户信息、用户评论等。

(2)XUNRED[20]：一个适用于舆情领域的已标注的有监督的中文关系抽取数据集，来自百度百科、百度贴吧平台，对用户评论涉及的实体及关系进行了标注。

(3)Weibo NER[21]：一个中文 NER 数据集，包含 2 259 434 条微博，标记了人、组织、位置以及政府部门 4 种实体。

2.4 评价指标

信息抽取 4 个子任务的常用评价指标为精确率(Precision)、召回率(Recall)、F 值(F measure)，具体如式(1)—式(3)所示，但是这 4 个子任务间的计算方式有所不同，下面将详细介绍在具体任务中的准确率和召回率的计算方式。

$$Precision=\frac{TP}{TP+FP} \tag{1}$$

$$Recall=\frac{TP}{TP+FN} \tag{2}$$

$$F_\beta=\frac{(\beta^2+1)Precision\times Recall}{Precision\times Recall} \tag{3}$$

其中，TP(True Positive)表示被判定为正样本，实际也是正样本；FP(False Positive)表示被判定为正样本，实际是负样本；FN(False Negative)表示被判定为负样本，实际是正样本；β是一个调节准确率与召回率比重的参数。在实际测试中，一般认为准确率与召回率同等重要，β值设置为1，由此式(3)可以表示为：

$$F_1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{4}$$

2.4.1　实体抽取

实体抽取的结果既要保证实体边界识别正确，也要保证实体类型正确，主要有两种评估方法：精确评估和宽松评估。精确评估中的 TP 表示能正确识别的实体；FP 表示能识别出实体但类别或边界判定出现错误；FN 表示没有被识别的实体，具体如式(5)、式(6)所示：

$$Precision=\frac{\text{正确识别实体类型且边界对应的个数}}{\text{待进行实体抽取的词数}} \tag{5}$$

$$Recall=\frac{\text{正确识别实体类型且边界对应的个数}}{\text{实际应被识别的实体个数}} \tag{6}$$

宽松评估认为，如果一个实体被预测为正确的类型，则不管它的边界是否与真实标签对应，预测就是正确的，实际应用中宽松匹配的评估应用较少，在此不再赘述。

2.4.2　实体链接

现有的针对实体链接的研究大多只针对实体消歧这一单个阶段进行优化，因此评价指标通常采用其子任务实体消歧的评价指标 $Precision$、$Recall$、F_1 值进行评估，而实体消歧由于其针对的问题、采用的算法和数据集不同，难以进行统一的比较，本文采用文献[22]提出的一种模糊召回指标，使用细粒度的评价来解决实体链接任务缺乏共识的问题，具体如式(7)、式(8)所示：

$$Precision=\frac{\text{正确消歧的实体指称}}{\text{待消歧的实体指称总数}} \tag{7}$$

$$Recall=\frac{\text{含有正确实体的候选集数}}{\text{待消歧的实体候选集数}} \tag{8}$$

2.4.3　关系抽取

关系抽取中的评价指标也使用 $Precision$、$Recall$、F_1 值进行评估，其中 $Precision$、$Recall$ 和通用方法的计算方法有所不同，具体如式(9)、式(10)所示：

$$Precision=\frac{\text{被正确抽取的属于关系 }R\text{ 的实体对个数}}{\text{所有被抽取为关系 }R\text{ 的实体对个数}} \tag{9}$$

$$Recall=\frac{\text{被正确抽取的属于关系 }R\text{ 的实体对个数}}{\text{实际应被抽取的属于关系 }R\text{ 的实体对个数}} \tag{10}$$

2.4.4　事件抽取

对于单一事件抽取任务，如突发事件、门户网站、金融资讯等使用 F_1 值进行评估，其中 TP 表示正确识别的事件触发词，FP 表示能识别出事件但类别出现错误，FN 表示没有被正确识别的事件，具体如式(11)、式(12)所示：

$$Precision=\frac{\text{能正确识别事件且类型正确的实体数}}{\text{待识别的事件实体总数}} \tag{11}$$

$$Recall=\frac{\text{正确识别事件且类型正确的实体数}}{\text{实际应识别的事件实体总数}} \tag{12}$$

3　基于深度学习的舆情信息抽取方法

社交媒体数据因其固有的噪声使信息抽取任务具有挑战性，除了不正确的语法结构外，还存在拼写不一致和大量非正式缩略语。近年来，基于深度学习的信息抽取模型占据主导地位，与早期的模式匹配方法和传统的机器学习方法相比，深度学习方法用到的神经网络提供了强大的语义组合能力，有利于发现隐藏的特征并自动提取特征，从而避免特征工程的误差积累和传播。本节将结合最新的研究成果，详细介绍不同的神经网络结构和特征以及这些特征如何应用于信息抽取任务的方法。

3.1　基于深度学习的舆情实体抽取

早期的NER基于专家手工构建的命名实体识别规则，将规则和文本相匹配来抽取实体，其缺点也很明显，只能在特定的领域内使用，难以迁移和泛化。随着机器学习技术的发展，学者利用标注的数据进行训练，采用隐马尔可夫模型(Hidden Markov Mode，HMM)、最大熵隐马尔可夫模型(Maxmium Entropy Markov Mode，MEMM)、支持向量机(Support Vector Machine，SVM)、条件随机场(Conditional Random Fields，CRF)等方法会产生误差传播的问题，因此研究者开始考虑使用深度学习技术。基于深度学习的实体抽取模型是端到端的，主要有3个步骤。

(1)分布式表示(Distributed Representation)：选择字级别、词级别、词缀级别或组合这3个粒度的特征表示法，将其映射到低维空间表示成稠密实值向量，常用方法包括基于Word2Vec，GloVe等模型的静态词向量表示法和基于BERT(Bidirectional Encoder Representation from Transformers)等模型的动态预训练语言模型的表示方法。

(2)上下文编码(Context Encoder)：用特征提取器提取特征表示输出向量的文本特征，常用的三大主流方法为基于CNN的方法、基于RNN的方法和基于Transformer的方法。

(3)标签解码：接收语义表示，输出标注序列。常见的解码方式为：MLP+softmax、CRF、RNN以及指针网络。

3.1.1　基于循环神经网络(RNNs)

循环神经网络及其变体长短期记忆网络(Long Short Term Memory，LSTM)和门循环单元(Gate Recurrent Unit，GRU)结构被证明能利用整个序列的信息，此外，LSTM能够有效地适应不同领域的数据，适用于NER任务，如文献[23]首次将BiLSTM应用于NER任务中，实验发现BiLSTM-CRF的搭配较其他模型的准确率都高，成为了最常见的NER模型，文献[24]采用多个独立的LSTM单元，通过模型之间的正则化提高了各个LSTM单元之间的多样性，大大减少了参数量，其 F_1 值相比文献[23]提高了0.27%。在应用方面，文献[25]为减少事件定位检测的工作量，利用NER提取Twitter中与尼泊尔地震相关的组织、人、位置等命名实体，通过对GRU、LSTM、BiLSTM和激活函数的对比实验得出带有softmax的LSTM结果最好，F_1 可达到92%。

3.1.2　基于卷积神经网络(CNNs)

虽然基于RNN的方法在NER任务中取得了显著的

效果，但是 LSTM 的循环结构阻碍了利用 GPU 的并行性，降低了模型效率，相比之下，多层 CNN 可以在不降低模型效率的情况下有效地捕捉长期信息。如文献[26]提出了结合 CRF 的残差扩张卷积神经网络(RD-CNN-CRF)，该方法将中文 NER 任务看作一个字符级别的序列标注任务，利用 RD-CNN 获取上下文特征，以避免中文分词错误引入的噪声，最后利用 CRF 捕获相邻标签的相关性，从而获得整个句子的最优标注序列。该方法与基于 RNN 的方法相比，在 CCKS-2017 task2 中取得了 F_1 值为 91.32%的成绩。随着注意力机制的兴起，文献[27]提出了 ALL CNN 模型，使用多个具有不同尺寸的核和残差结构构建了一个多层 CNN 模型，并将这些 CNN 层的结果进行融合，获得了不同尺度的上下文信息，此外还设计了注意力机制用于捕捉全局上下文特征。该方法在 CCKS2017 和 CCKS2019 中的表现均超越了基于 Bi-LSTM 的方法。

3.1.3 混合 CNN 和 RNN 的方法

为进一步提升深度学习的性能，学者将 CNN 在提取局部字符特征方面的优势和 RNN 提取全局语义特征方面的优势相结合，以构建混合网络模型的实体抽取。文献[28]采用多任务学习法，分别使用 CNN 获取字符级表示，使用 BiLSTM 编码经过拼接的 POS-tag 嵌入和单词嵌入，从而获得词级别的表示，再加上额外的词典特征表示，最后输入多任务网络中，同时预测命名实体的边界和类型，该方法在 W-NUT17 数据集上取得了最好的成绩。文献[29]通过各种基线对比实验，研究词嵌入、字符特征和词特征及其组合对实体抽取的有效性，并利用 CNN-BiLSTM-CRF 模型进行实体识别，准确率得到了有效的提升。在应用方面，文献[30]使用 BIO 标注方式将社交网络中的实体分为 11 类，采用 BiLSTM 获取词级特征，采用 CNN 获取字符级特征，提取 Twitter 中的关于交通状况的实体，F_1 值达到 78.9%。除此之外，社交媒体短文本的特点使得学者们将目光拓宽至图片，企图通过结合视觉语境为社交媒体短文本提供辅助信息，因此多模态 NER 任务引起了越来越多学者的研究兴趣。目前成功的方法都是聚焦于文字区域和图像区域的对齐以及文本信息与视觉信息的融合，如文献[31-33]依赖于结合注意力机制的不同融合技术，将字级别、词级别和图片类型的信息融合，取得了不错的效果。

3.1.4 基于 Transformer 的方法

由于 Transformer 采用自连接的自注意力结构建模上下文，弥补了 RNN 无法解决远程依赖的缺点，此外 Transformer 与 RNN 相比具有更好的并行能力，因此更适用于大型语料库。文献[34]发现，Transformer 在 NER 中的性能不如在其他 NLP 任务中好，分析其原因后提出了 TENER 模型，基于 Transformer 建模字符级特征和词级特征，并通过对比发现 TENER 有着比 BiLSTM 和 CNN 更好的字符编码性能。文献[35]使用基于 Transformer 的注意力机制的语义增强模块对语义信息进行编码，利用 Gate 模块将语义信息聚合到标签过程中，以缓解社交媒体数据的稀疏性问题，分别在中文数据集 WB 和 3 个英文数据集上达到了新水平。文献[36]以完型填空的方式预训练双向 Transformer 模型，在 CoNLL2003 上达到了 93.5%的效果。文献[37]使用 BERT 预训练词向量，使用 Softmax 和 Dice Loss 分类，在 OneNotes5.0 数据集上取得了 92.07%的效果。

在中文 NER 方面，格结构被证明对利用词信息和避免分词的错误传播有很大的好处。格结构是一个有向无环图，其中每个节点是一个字符或一个词，它们不是按顺序排列的，单词的第一个和最后一个字符决定了它的位置。基于此结构，文献[38]提出了 FLAT 模型，为字符或单词分配两个位置索引，即头部位置和尾部位置，利用 Transformer 的自我注意机制对格结构进行编码，通过实验证明了该模型的优越性。

表 3 实体抽取模型不同实验结果对比

Table 3 Comparison of entity extraction model results

	年份	模型					领域	数据集	F_1/%
		分布式表示			上下文编码	标签解码			
		词级	字级	混合					
文献[23]	2015	Senna	—	Spelling	LSTM	CRF	通用	CoNLL 2003	88.76
文献[24]	2018	skip-n-gram	BiLSTM	—	并行 RNN	Softmax	通用	CoNLL 2003	91.48
文献[25]	2021	Glove	—	—	LSTM	Softmax	社交媒体	Twitter-Nepal Earthquake	92.00
文献[26]	2019	Embedding	—	—	RD-CNN	CRF	中文医学	CCKS2017 task2	91.32
文献[27]	2021	Wor2Vec	—	—	ALL CNN	CRF	中文医学	CCKS-2017 CCKS-2019	90.49 85.13
文献[28]	2019	Twitter Wor2Vec	CNN	POS-tag	BiLSTM	CRF	社交媒体	WNUT-2017	41.86
文献[29]	2020	Glove	CNN	—	BiLSTM	CRF	通用	CoNLL 2003	91.10
文献[30]	2018	Glove	CNN	POS-tag	BiLSTM	Softmax	社交媒体	Twitter-traffic	78.90
文献[31]	2018	Glove	Bi-CharLSTM	CNN-image	Attention+Bi-LSTM	CRF	社交媒体	SnapCaptions	52.40
文献[34]	2019	ELMo	Transformer	POS-tag	Transformer	CRF	通用	CoNLL2003 OntoNotes5.0	92.71 89.93
文献[35]	2020	Glove	—	—	Transformer	CRF	社交媒体	WNUT-2017 WB	49.45 48.41
文献[36]	2019	BERT	CNN	—	GRU	CRF	通用	CoNLL2003	93.50
文献[37]	2020	—	—	POS-tag	BERT	Softmax+Dice Loss	通用	OntoNotes5.0 CoNLL2003	92.07 93.33
文献[38]	2020	—	—	Position	FLAT	CRF	通用	OntoNotes Weibo	75.70 63.42

3.2　基于深度学习的舆情实体链接

前文提到实体链接的3个步骤为CEG，ED和UMP，由于实体链接的CEG和UMP使用的技术近年来没有太大变化，因此本节主要集中在实体消歧（ED）部分，介绍了不同的神经网络结构和特征如何帮助候选实体进行排序。

传统的基于机器学习的实体消歧包括无监督的基于向量空间模型和基于信息检索的方法，以及有监督的二进制分类方法、学习排序法、概率法、图论法，但是这些方法的局限性在于需要大量细致而繁琐的特征工程，并且依赖特定领域的知识库。基于深度学习的实体链接方法在一定程度上弥补了这个缺陷，一般步骤是首先生成实体提及和一组候选实体的各种嵌入，包括词嵌入（Word Embedding）、实体提及嵌入（Mention Embedding）、实体嵌入（Entity Embedding），然后使用嵌入来计算特征，这些特征包括实体流行度（Prior Popularity）、实体表面相似度（Surface Form Similarity）、类型相似度（Type Similarity）、实体描述（Entity Description）、上下文相似度（Context Similarity）、主题一致性（Topical Coherence）等，最后将特征输入到特定的算法（一般为多层感知机MLP、网页排名算法PageRank、强化学习算法等）中，将候选实体进行排序，从而得到相似度最高的实体作为实体链接结果。

根据模型使用的特征的不同，本文将现有研究成果分为基于上下文相关特征和无关特征的实体消歧，现有的模型大多组合这两种特征，共同完成实体消歧任务。具体的模型对比如表4所列。上下文相关特征利用实体提及周围文本的上下文信息，以独立地解决每个实体指称的歧义问题。上下文无关特征提倡使用文档中与主题保持一致性的目标实体，通过计算不同目标实体之间的主题一致性、实体关联度、转移概率和实体流行度等特征进行消歧，值得注意的是，上下文无关特征是几乎所有传统的实体消歧任务都会用到的特征，随着深度学习的发展，这些特征可以直接通过神经模型来学习。

3.2.1　基于上下文相关特征的实体消歧

基于深度学习的实体消歧任务最直接的想法是计算从实体提及的上下文中衍生的表示与目标知识库中候选实体相关的表示之间的相似度。为了建立上下文相关性的特征模型，一般方法是使用不同神经网络结构将实体提及的上下文编码为上下文嵌入，然后利用不同的计算方法计算生成上下文嵌入和实体嵌入的相似性得分。下面将介绍基于不同神经网络的上下文嵌入。

（1）基于RNNs

基于RNN的模型以及一些RNN的变体，如LSTM和GRU，使用递归神经单元捕捉单词之间的依赖关系，因此特别适用于学习相关实体上下文的特征。文献[39-41]使用前向LSTM和后向LSTM分别对实体提及的上文和下文进行编码，结合实体的文本描述、实体类型等其他信息生成一个统一的嵌入，输入进算法中进行实体提及和候选实体的语义匹配。文献[42-43]都在实体链接任务中使用了前向GRU和后向GRU分别对上文和下文进行编码，同样取得了很好的效果。

（2）基于CNNs

基于CNN的模型包含多个卷积层，每个卷积层用于提取每个上下文单词周围的局部特征，并且每个卷积层的输出大小取决于上下文中的单词数量，通过组合由卷积层提取的局部特征向量构造上下文嵌入。文献[44]使用CNN来生成隐藏的向量序列，然后通过一个非线性函数对这些向量序列进行转换，再通过拼接生成上下文嵌入。

（3）基于注意力机制

注意力机制上下文嵌入学习中对上下文词语给予不同的注意权重，基于注意力的实体链接模型可以计算上下文词语与注意向量的相关程度，这些模型以上下文词的加权和作为上下文嵌入。文献[45-46]设定，如果一个单词至少与一个实体密切相关，则将其作为注意向量用于上下文嵌入，达到了很好的效果。

3.2.2　基于上下文无关特征的实体消歧

与计算实体提及的上下文相关特征不同，上下文无关特征仅基于实体提及和候选实体本身对不同候选实体进行打分和排序，多侧重于知识库中的候选实体的特征，下面将对常用的特征进行列举。

（1）实体表面（Form）相似度

通过比较实体提及和候选实体的名称是否完全匹配、实体提及是否以候选实体为前缀或后缀、实体提及是否完全包含候选实体、实体提及所包含的单词首字母序列是否与候选实体包含的首字母相同、实体提及和候选实体共同包含的单词数目等，对实体提及和候选实体进行相似度计算。

（2）实体类型（Type）相似度

利用实体类型相似度这一特征指通过对比实体提及的NER类型和候选实体在知识图谱中的类型是否一致来决定是否链接，如“苹果”很有可能指的是一家公司而不是水果，因此可以用相关类型限制对应的实体，如Raiman[47]提出了DeepType模型，仅使用实体类型特征就完成了实体链接任务，他们通过在维基百科的本体上选择一组关系来限制候选实体的类型选择，然后利用Bi-LSTM分类器获得实体类型的条件概率。但是，现有的预训练实体嵌入只学习文本中的底层语义信息，忽略了细粒度的实体类型信息，导致链接实体的类型与所提及的上下文不兼容，针对这一问题，文献[48]将单词和细粒度实体类型嵌入到同一向量空间中，对单词向量进行预训练以注入类型信息，然后用包含细粒度类型信息的词向量对实体嵌入进行再训练，该方法在数据集上的F_1值分别提高了0.82%和0.42%。

（3）实体流行度（Prior Popularity）

实体流行度是实体消歧值得利用的一个重要特征，如网球运动员“李娜”比歌手“李娜”更出名，因此在大多数情况下人们提到的“李娜”，更有可能指的是网球运动员“李娜”。文献[49]单独使用实体流行度这一特征就在实体链接任务中达到了最高85%的准确率，其通过计算实体提及在目标知识库中出现的概率，说明了实体流行度这一特征的有效性。

表 4 实体链接模型实验结果对比

Table 4 Comparison of entity linking model results

	年份	特征					研究特点	模型	数据集	F_1/%
		上下文	非上下文							
			Form	Type	Pop	Description				
文献[39]	2017	LSTM	—	√	—	√	整合多种信息为每个实体学习统一的密集表示	CDTE	CoNLL ACE05 Wiki	82.9 85.6 89.0
文献[40]	2017	LSTM+Attention	—	—	—	√	第一个使用 LSTM 和注意力机制进行实体消歧的研究，提出了 Pair-Linking 链接算法	NeuPL	ACE04 DBpedia	92.9 82.8
文献[41]	2017	LSTM	√	—	—	—	零样本、跨语言	—	ChineseTAC 2015 CoNLL 2003	87.4 94.0
文献[42]	2017	RNN+Attention	√	—	—	—	适用于嘈杂文本	ARNN	PPRforNED[50] CoNLL-YAGO	87.3 83.3
文献[43]	2018	GRU+ATTN+FEATS	√	—	—	—	适用于嘈杂文本、字符级上下文特征，导致性能下降	—	WikilinksNED	74.9
文献[44]	2019	CNN	√	—	—	√	基于循环随机游走的端到端的实体链接	RRWEL	AVG	88.43
文献[45]	2019	ETHZ-Attn[49] Berkeley-CNN[51]	— —	√ √	√ √	— —	动态上下文扩充(DCA)应用到全局实体链接中	DCA	AIDA-B	94.64 92.72
文献[46]	2020	—	—	√	—	—	构造细粒度语义词典融入实体嵌入中	FGS2EE	AVG	86.32
文献[47]	2018	—	—	√	—	—	优化知识库的 Type 系统，将符号信息整合到神经网络推理过程中	DeepType	WKD-30 CoNLL-YAGO TAC KBP 2010	92.37 94.87 90.85
文献[48]	2021	—	—	√	—	—	将细粒度类型信息融入实体嵌入中	—	AVG	86.32
文献[49]	2017	Embedding+Attention	—	—	√	—	文档级实体消歧，将词和实体嵌入至统一向量空间中	—	AIDA-B	92.22

注：表中选取最优的实验结果作为总结，其中 AVG 代表该模型基于 MSNBC，AQUAINT，ACE2004，CWEB，WIKI 这 5 个数据集进行实验，结果取平均值

3.3 基于深度学习的舆情关系抽取

文献[52]指出，现有的基于深度学习的关系抽取可以分为 4 类，分别是有监督关系抽取、半监督关系抽取、远程监督关系抽取和无监督关系抽取，本节将结合社交网络中舆情信息的特点介绍适用于舆情实体关系抽取的有监督关系抽取和远程监督关系抽取。

3.3.1 有监督的关系抽取

目前有监督的关系抽取主要有两类框架，即流水线法和联合抽取法，前文已经介绍过，此处不再赘述。发表于 NAACL2021 的文献[53]指出，分别学习实体和关系的上下文特征比联合学习更有效，该文献仅采用 Pipeline 方式就超越了各种 Joint 模型，达到了目前最好的效果，但是流水线法的误差传播问题仍然存在，联合关系抽取的研究仍具有意义，本节将详细介绍基于这两种框架的深度学习技术的研究成果。

(1)流水线法

基于流水线的方法的关系抽取通常默认是在实体抽取的基础上进行关系分类，过程可以描述为：输入已经标注好实体对的句子，经过模型处理后把实体关系或事件关系三元组作为预测结果输出[18]。基于流水线的关系抽取通常使用三大主流方法：基于 CNN 的方法、基于 RNN 的方法和基于 BERT 的方法。下面将对相关研究成果进行总结。

1)基于 CNNs

关系抽取最常用的模型是 CNN，如文献[54]首次将卷积神经网络应用于关系抽取中，首先使用同义词词典对输入词编码，再使用卷积神经网络进行词义特征的提取，在 ACE05 数据集中该方法比当时最新的模型在性能上提升了 9%。文献[55]利用深层卷积神经网络提取词汇和句子级别的特征，再将这两个层次的特征连接起来共同输入到 softmax 分类器中，用于直接预测两个名词之间的关系，取得了不错的效果。文献[56]提出了一个利用排序方式进行关系分类的网络(CR-CNN)，给定一个句子和两个目标实体，该模型使用 CNN 分别学习其分布式表示和每个类别的分布式矢量表示，比较后生成每个类的得分，该模型比文献[55]提出的模型达到了更好的效果。

在关系抽取中，每个单词对关系语义的贡献并不相等，因此学者们开始增加注意力机制以区分不同单词与关系的相关性，如文献[57]提出的 BiAtt-pooling-CNN 模型，该模型将注意力机制分别应用到输入序列和池化层中，以学习句子中各个词与目标实体的相关性以及各个词与关系的相关性，在公共数据集 SemEval2010 上的 F_1 值达到了 0.88。

2)基于 RNNs

由于 CNN 难以处理时序特征，且难以应对两个实体距离较远的问题，因此不少学者将 RNN 应用到关系抽取任务中。目前进行关系分类取得最好效果的是文献[58]提出的一种结合实体感知注意机制和潜在实体类型(Latent Entity Type，LET)的端到端循环神经模型(LET-BLSTM)。为了捕捉句子的上下文，该模型通过自我注意力机制来获得单词的表示，并用双向长短期记忆网络来构建循环神经结构。

3)基于 BERT

文献[59]提出了一种利用实体信息来丰富 BERT 预训练

语言模型的方法(R-BERT)，以进行关系分类，通过预先训练的体系结构定位目标实体并传递信息，并合并两个实体的相应编码。文献[60]提出了一种通用关系提取器($BERT_{EM}$+MTB)，将 BERT 与“匹配空白”任务相结合，仅从实体链接知识库文本中构建与任务无关的关系表示形式，实验证明，$BERT_{EM}$+MTB 大大优于 SemEval 2010 Task8 上的先前方法，展示了该模型如何在资源匮乏的情况下仍然有效，取得了目前的最高 F_1 值，为 89%。

(2)联合抽取法

由于流水线法会带来误差传播以及会忽略关系抽取的两个子任务之间的内在联系的问题，因此学者们考虑将实体抽取和关系抽取两个子任务联合建模，以便在统一的模型中进行共同优化。对于舆情信息来说，端到端的联合实体关系抽取也更适合实时变化的数据，直接在文本中抽取实体关系三元组。最近进行联合关系抽取任务的主要有 3 种方法：表格填充法(Table Filling)、标注法(Tagging)和 Sequence-to-Sequence 法。

1)基于表格填充法

Table Filling 的方法为每个关系维护一个表，表中的每一项用于指示一个标记对是否具有对应的关系，因此表格填充的关键是准确地填充关系表，然后根据填充的表提取三元组。如文献[61]提出了一个单阶段模型 TPLinker，基于单个标记对的填充历史提取局部特征，能够发现共享一个或两个实体的重叠关系，但是这种方法忽略了实体对和关系的全局联系；文献[62]提出了一个迭代模型 GRTE，将三元组之间的全局特征集成到表特征中，最后根据新的表特征提取所有与该关系相关的三元组，该模型在所有数据集上都取得了最好的效果，但是基于表格填充的方法需要列举所有可能的实体对，这会带来沉重的计算负担。

2)基于序列标注法

基于 Tagging 的方法把联合关系抽取任务转换为标签问题，仅标注存在关系的实体对，标签即为实体间的关系类型。文献[63]设计了一种特别的标注方法，用于分别标注实体中词的位置信息、实体关系类型和实体角色信息，使用 Bi-LSTM 编码后再使用 LSTMd 进行解码，最终输出标注好的实体-关系三元组，但是该模型存在不能抽取重叠关系的问题，即一对实体可能存在多种关系的情况。文献[64]针对这一情况，将同一句子标记多次以识别所有的重叠关系，增加了偏置损失函数，增强了相关实体之间的关系。

3)基于 Sequence-to-Sequence(Seq2Seq)法

基于 Seq2Seq 的方法通常将关系抽取任务转换为三元组生成任务，输入非结构化的文本并直接解码生成三元组。文献[65]提出了基于 Seq2Seq 和复制机制的 CopyRE 模型，该模型分为编码器和解码器两个部分。该模型使用 Bi-LSTM 作为编码器将文本转换成定长的语义向量，通过解码器根据该语义向量生成三元组，该模型最终能提取多个关系事实，但存在两个问题：1)不能区分主实体和客实体；2)不能识别具有组合性的实体，如“Steven Jobs”，CopyRE 只能识别“Jobs”。针对这些问题，文献[66]在文献[65]的基础上提出了模型 CopyMTL，该模型在编码前加入了序列标注层以协助实体识别，获取了具有多个标记的实体，同样使用 Bi-LSTM 建模上下文信息，最后使用注意力机制和 LSTM 配合全连接层进行输出，超越了 CopyMTL 的效果。

有监督的关系抽取模型的对比如表 5 所列。

表 5　有监督的关系抽取模型不同实验结果的对比

Table 5　Comparison of supervised relation extraction model results

	年份	模型	有监督		数据集	F_1/%
			流水线	联合		
文献[53]	2020	PURE	√	—	ACE05	64.8
文献[54]	2013	CNN	√	—	ACE05	83.8
文献[55]	2014	DNN	√	—	SemEval-2010	82.7
文献[56]	2015	CR-CNN	√	—	SemEval-2010	84.1
文献[57]	2016	CNN+Attention	√	—	SemEval-2010	88.0
文献[58]	2019	LET+Attention	—	√	SemEval-2010	85.2
文献[59]	2019	R-BERT	√	—	SemEval-2010	89.25
文献[60]	2019	$BERT_{EM}$+MTB	√	—	SemEval-2010	89.5
文献[61]	2020	TPLinker	—	√	NYT24 WebNLG	92.0 86.7
文献[62]	2021	GRTE	—	√	NYT24 WebNLG	93.1 89.4
文献[63]	2017	LSTM-LSTM-Bias	—	√	NYT	49.5
文献[64]	2019	ETL-Span	—	√	NYT-single/Multi WebNLG	59.0/78.0 83.1
文献[65]	2018	CopyRE	—	√	NYT WebNLG	67.1 55.3
文献[66]	2020	CopyMTL	—	√	NYT WebNLG	70.9 58.9

3.3.2　远程监督的关系抽取

有监督的关系抽取方法虽然取得了较高的准确率，但是会耗费大量的人力来制定规则或进行标注工作，泛化能力也不足，因此结合半监督和无监督学习的远程监督的关系抽取被提出，为关系抽取任务自动构建数据集。远程监督的关系抽取基于这样的假设：如果两个实体 h，t 在已知的知识图谱中存在关系 R，则所有包含 h，t 的句子 W 都将表达关系 R。但是这种方法构建的数据集存在错误标注以及数据长尾的现象，文献[67]针对这两个问题进行了论述，目前的研究方向也是针对这两个问题对远程监督的关系抽取进行改进。

3.4 基于深度学习的舆情事件抽取

基于深度学习的事件抽取的一般过程是建立一个神经网络，以词的嵌入为输入，输出每个词的分类结果，即该词是否是事件触发词或事件参数，如果是，则再识别它的事件类型或参数角色。如何设计一个高效的神经网络结构是基于深度学习的事件抽取的主要挑战，主要依赖于事件触发词的表示，目前触发词的表示有两种思路：1)使用触发词的上下文表示，因为在相似的上下文中出现的事件的触发词往往具有相似的意义；2)使用事件要素的语义表示，因为触发词的类型依赖于事件要素以及要素角色。最近许多学者利用卷积神经网络、循环神经网络、图神经网络(Graph Neural Networks，GNN)、Transformer 等深度学习架构显著改善了事件抽取任务，本节将沿用关系抽取的分类方法，将其分为有监督、半监督和无监督的事件抽取进行介绍，具体如表 6 所列。

3.4.1 有监督的事件抽取

有监督的事件抽取采取与关系抽取相同的模型设计思路，分别是流水线法和联合抽取法。基于流水线的方法将事件抽取的所有子任务视为独立的分类问题，包括 3 个阶段：识别触发词并分类，确定事件类型；识别事件参数，确定该词是否是事件的参数；将参数角色分类，从而确定参数的类别。采用流水线法的事件抽取最大的缺点是误差传播，如果触发词的识别出现错误，则会严重影响事件参数识别的准确率，因此流水线法常将触发词识别作为核心步骤。根据最新研究成果统计，事件抽取的联合抽取模型成果较多，是目前主流的事件抽取方法，而流水线法的成果较少，因此本节将不作流水线法和联合抽取法的分类，而是根据模型使用的不同神经网络进行介绍，具体如表 6 所列。

表 6 有监督事件抽取模型不同实验结果对比

Table 6 Comparison of supervised event extraction model results

	年份	有监督	联合抽取	模型	F_1/%		
					ETC	AI	AR
文献[68]	2016	√	—	DMCNN	69.1	59.10	53.50
文献[69]	2016	√	√	CNN+Bi-LSTM	64.5	—	46.90
文献[70]	2016	√	√	JRNN	69.3	62.80	55.40
文献[71]	2018	√	—	HNN	73.4	—	—
文献[72]	2018	√	—	DAG-GRU	71.1	—	—
文献[73]	2018	√	√	DBRNN	71.9	67.70	58.70
文献[74]	2020	√	√	Tree-LSTM+Bi-GRU	72.1	68.10	59.10
文献[75]	2018	√	√	JMEE	73.7	68.40	60.30
文献[76]	2019	√	—	MOGANED	75.7	—	—
文献[77]	2020	√	—	SDP-DMCNN	75.8	71.52	66.44

注：SDP 表示最短依赖路径，模型全部基于数据集 ACE2005 进行实验，其中 ETC(Event Type Classification)表示识别事件类型；AI(Argument Identification)表示参数识别；ARI(Argument Role Identification)表示角色识别

(1)基于 CNNs

现有的事件抽取任务严重依赖标注工具和实体抽取的结果，造成了误差传播，为了达到自动提取词义和语义信息的目的，文献[68]将 CNN 应用于事件抽取任务，为了捕获词汇级别和句子级别的特征，设计了一个动态的多层池化卷积神经网络 DMCNN，针对一个句子可能包含两个或两个以上的事件，以及一个要素可能在不同事件中扮演不同角色的问题，输入上下文特征、位置特征和事件类型特征进行要素角色的预测。在中文事件抽取中，文献[69]针对中文分词的差异性，选择序列标注策略来完成事件检测任务，将 Bi-LSTM 和 CNN 结合来共同捕获词汇级别和句子级别的信息，利用 CNN 捕捉局部词汇特征的能力来消除触发词的歧义问题，省去了词汇标注的步骤。

(2)基于 RNNs

RNN 可以建模序列信息，常被用于联合事件抽取框架中。文献[70]首次提出在联合抽取框架中使用双向的 RNN 的研究，该模型对上下文信息进行汇总并预测了触发词和事件参数，在当时取得了较好的性能。文献[71]提出了一个混合神经网络，用于捕获上下文中的特定序列和信息片段以及训练一个多语言的事件检测器，该模型使用 Bi-LSTM 获取需要识别的文档序列信息，然后使用 CNN 在文档中获取短语信息，最终识别触发词。

但是，这些模型并没有充分利用句法信息，它们仅将依赖关系作为输入的特征，而不是将其建模到体系结构中。文献[72]引入 DAG-GRU 建模句法信息，将与事件相关的两个单词之间的依存关系视为一个有向无环图(DAG)，通过双向阅读来捕捉上下文和语法信息。除此之外，该模型还通过注意力机制结合句法和序列信息，最后验证了该模型的有效性。文献[73]提出了基于依赖桥的递归神经网络 DBRNN 以进行事件提取，依赖桥即将一个句子中具有依存关系的单词连接为有向边，并将携带时序、因果、条件等信息的边赋予一个权重。该模型基于 Bi-LSTM 建立，将依赖桥的结构引入到 LSTM 单元中，以完成触发词和事件参数及角色的联合抽取。文献[74]通过 Tree-LSTM 来获取触发词和参数之间的依赖特征，同时使用 Bi-GRU 来获取候选事件句的上下文特征以辅助事件类型的判别。该模型基于句嵌入，直接通过识别候选事件的句子类型来确定事件类型，跳过了触发词的分类，提高了事件类型识别的准确率。

(3)基于 GCNs

为了抽取一个句子中可能存在的多个事件，基于 RNN 的序列建模在捕获长距离依存关系时效率非常低，因此许多学者开始考虑把 GCN 应用到事件抽取中，如文献[75]提出了一种新的联合多事件抽取的框架 JMEE，通过引入句法弧来增强信息的流动，改善多个事件在同一个句子中的情况。具体做法是使用 GCN 学习每个节点的句法上下文表示，然后通过自注意力机制聚合信息，保持多个事件之间的关联，共同提取触发词和参数。

随着研究的深入，研究者们发现，单个句法弧只能表示一阶句法关系，而在现实应用中可能需要多个弧才能最终确定单词的含义，文献[76]提出了多阶图卷积和注意力聚合机制的事件检测模型 MOGANED，该模型使用图注意网络(Graph Attention Network，GAT)来衡量相邻词在不同阶的句法图中的重要性，然后使用注意力聚合机制来合并它的多阶表示，使事件检测任务的准确率获得了很大的提升。然而，这些方法单独提取事件的每一个参数而忽略了参数之间的关系。为了提升事件抽取的效率，文献[77]引入最短依赖路径(Short Dependency Path，SDP)来消除句子中不相关的词，从而捕获长距离依赖关系。此外，该研究还提出了一种基于注意的图卷积网络，该网络沿着候选参数之间的最短路径携带语法相关信息，捕捉并聚集了参数之间的潜在关联。

3.4.2　半监督的事件抽取

有监督的事件抽取依赖于手工标注的数据进行训练，费时费力且大小有限，而常用的 ACE2005 语料库中只定义了 33 种事件类型，这在社交网络舆情事件中是远远不够的。近年来，为了缓解现有的研究中缺乏预定义事件模式的问题，学者们尝试使用远程监督、半监督以及弱监督的方法来扩展标记数据并训练模型，在领域内标注数据极少的情况下，研究者们提出了少样本学习（Few-shot Learning）和零样本学习（Zero-shot Learning）下的事件抽取，这是当前的研究热点。需要特别指出的是，远程监督需要链接的外部知识库的数据没有标记触发词，并且特定事件的事件元素在多个句子中都有提及，使生成的事件不可避免地具有噪声，因此远程监督并不适用于标记事件。

3.4.3　无监督的事件抽取

有监督和半监督的事件抽取模型难以应对新出现的事件类型和角色，无监督的事件抽取主要针对开放领域，常用的思路是通过聚类归纳事件模式，判定事件类型并命名，然后根据事件类型分配参数。本节总结了最常用的几种使用无监督事件抽取的方法。

（1）基于预训练语言模型的方法

近年来，随着 BERT 等预训练模型的发展，学者们开始将其用于事件抽取任务，如文献[78]提出了一个事件抽取模型，该模型把事件抽取看作一个两阶段的任务，包括触发词识别和参数识别；此外，针对训练数据不足的问题，提出了一种基于预训练语言模型的事件抽取器，它包括基于 BERT 的特征表示的触发词提取器和参数提取器，从而自动生成标记数据，再通过质量排序筛选生成的样本。文献[79]将事件抽取作为序列标记任务，使用双向 Transformer 获得词嵌入表示，根据文献[78]提出的方法检测到触发词后，再根据提出的 EE-DGCNN 模型分配参数角色。

这些研究存在的问题是只关注更好的微调预训练语言模型，而没有考虑建模事件特征，导致开发的事件抽取模型不能充分利用大规模无监督数据。文献[80]利用抽象语义表示[81]（Abstract Meaning Representation，AMR）结构构建自监督信号的面向事件的对比预训练框架 CLEVE，包括两个部分，即事件语义预训练和事件结构预训练，实现了从大规模无监督数据和它们的语义结构中学习事件知识。

（2）基于生成式对抗网络的方法

生成式对抗网络是应用于事件抽取中的新方法，旨在利用文档中的冗余信息提取事件的结构化表示。文献[82]提出了一种基于生成对抗性神经网络的事件抽取模型 AEM，AEM 的生成器用于学习与事件相关的词分布（实体、位置、关键字、日期）之间的投影函数，捕获与事件相关的模式，实现了从不同长度的文本中挖掘事件；AEM 的判别器用于区分重构的事件和原始事件，该研究在 FSD，Twitter，Google 这 3 个数据集上的效果都优于基线模型，在新闻文章这种长文本数据集上的改进效果更为明显，F_1 值增加了 15%。

（3）基于问答的方法

将事件抽取作为问答任务可以将问答任务的最新进展迁移到事件抽取任务中，也可以利用丰富的问答任务的数据集来缓解事件抽取任务的数据稀疏性问题。文献[83]明确将事件抽取转换为一个问答任务，先提取事件触发词，然后根据事件类型用无监督的方式生成事件要素的问题模板，最后以问答的形式进行事件要素抽取，将问题作为第一个序列，将原文本作为第二个输入序列，共同输入 BERT 模型中进行编码。

4　现存问题与未来发展趋势

4.1　现存问题

使用不同深度学习模型的信息抽取方法各有特点：基于卷积神经网络的模型可将不同长度的句子处理为统一长度的向量，有利于多特征的提取，但是基于 CNN 的方法难以提取时序特征；而基于循环神经网络及其变体的模型能充分考虑长距离词之间的依赖性，捕捉句子级别的特征，联合上下文提取特征；基于图卷积神经网络的模型用于表示实体间的关系，可以与 CNN，RNN 相结合来优化编码层；基于预训练语言的模型利用了大规模的语料库，有利于获取到具有深层语义信息的向量表示。虽然基于深度学习的信息抽取模型取得了很大的进展，但是应用于舆情信息抽取仍存在许多困难。

（1）社交网络舆情信息的数据量大、不规范、主题分布差异性较大，如何将抽取到的信息以更具效率的组织方式进行组织，以便下游知识图谱的构建、舆情主题的挖掘以及舆情分析等是需要解决的问题。

（2）当前的信息抽取技术的数据来源主要集中于文本方面，然而社交网络中还存在大量图片、视频、语音等其他模态的信息没有被利用，多模态的信息抽取成为了如今的研究趋势。

（3）目前信息抽取的研究多集中于特定领域预设的模板，而现实情况下的舆情信息更新迅速，缺乏标注数据，因此越来越多的研究者开始研究开放领域的信息抽取技术，期待能够实现自动发现实体、关系以及事件，但是这种方法的模型性能还有待提高，此外还缺少公认的评价体系，需要进一步完善。

（4）社交网络舆情信息语境复杂，通常同一实体或事件由多句语言、多种语言共同描述，因此应充分利用社交网络语言的多样性，研究跨语言的信息抽取技术。

4.2　未来发展趋势

4.2.1　舆情实体抽取方面

对于社交网络舆情实体抽取来说，由于其涉及众多领域，使用特定领域的数据，微调预训练语言模型是一种有效的办法。现阶段，实体抽取任务在嵌套实体抽取、细粒度实体分类方面还存在新的挑战。

（1）嵌套命名实体结构复杂多变，嵌套粒度与嵌套层数缺乏规律性[84]，如“中国首都北京”包含“中国”“中国首都”“北京”3 个内部命名实体，嵌套实体的存在极大地影响了语义的精准性，因此未来针对嵌套命名实体的研究将越来越重要。

（2）实体抽取任务不仅要将实体识别出来，还要对实体进行类型分类，更细粒度的划分会导致各实体类别在语义上的距离更紧密，使实体抽取任务更加精确。如今实体抽取研究工作追求更细粒度的分类，如何提升模型在实体分类中的区分效果是未来的研究趋势。

4.2.2 舆情实体链接方面

实体链接任务对知识图谱的构建十分重要，特别是在社交网络舆情这一领域，文本可能来自不同结构的不同数据源，例如来自新闻网站的新闻文档相对较长和正式，而来自社交平台的文本较为口语化和嘈杂，存在一词多义的现象，但是这些文本数据可能包含了大量可利用的知识。目前实体链接的研究主要集中在新闻文档和 Web 网页中的实体，由于不同类型的文本数据可能有不同的特征，现有的实体链接技术可能难以取得令人满意的链接效果。此外，由于目前相关研究所采用的数据集存在差异，难以对不同消歧算法的真实效果进行客观对比。相关组织机构可从评价框架、评测会议的角度提供规范的评测平台供学者交流，以进一步推动实体消歧研究的发展。

4.2.3 舆情关系抽取方面

舆情关系涉及实体关系和事件因果、时序关系，深度学习模型擅长处理单句语义信息，如今关系抽取多采用联合抽取模型，但是针对句级别的关系抽取可能捕获不到全面的语义关系。在舆情实体关系抽取方面，未来将从以下两个方面进行研究。

(1)篇章级关系抽取：如今的关系抽取任务大多还局限于单个句子，但是实际上一个事件可能涉及触发词和多种事件元素，触发词和所有事件元素出现在同一个句子中的理想情况并不常见，因此篇章级的事件提取非常有必要。

(2)重叠关系抽取：传统关系抽取更加关注于单实体对的关系，但是句子内包含不止一对实体且实体间存在重叠现象，未来针对重叠关系的研究将有助于挖掘更深层次的语义信息，推进信息抽取更加精准化。

4.2.4 舆情事件抽取方面

针对社交网络舆情事件的抽取已成为近年来的研究热点，研究者们采取的方法主要是基于深度学习的模型。目前，基于 BERT 的事件抽取方法已成为主流，然而事件抽取不同于 BERT 模型在预训练中的任务，不仅需要考虑事件参数角色之间的关系以提取同一事件类型下的不同角色，还需要针对事件抽取模型以学习文本的句法依赖关系。因此，为了全面准确地提取各个事件类型的参数，如何建立事件参数之间的依赖关系是一个迫切需要解决的问题。除此之外，未来事件抽取方面的研究将从以下两个方面开展。

(1)现有的预训练语言模型缺乏针对事件抽取任务的学习，社交网络舆情事件信息缺乏标注数据，因此大规模的事件库设计是未来的研究趋势。

(2)中文事件抽取工作目前还处于起步阶段，虽然在中文词语的形态结构和组合语义的利用方面取得了一些进展，但需要研究者们的努力，开发更具中文适用性的抽取模式。

结束语 近年来，随着社交网络舆情高频次大范围的爆发，舆情信息监管和分析成为了许多研究者和监管部门关注的问题。从海量的文本数据中抽取出结构化的信息是目前的研究热点，本文结合深度学习模型介绍了有关舆情信息抽取的方法，讨论了其适用性、局限性以及未来的发展趋势。从现有成果来看，虽然信息抽取各个子任务的性能得益于深度学习的发展并得到了很大的提升，但是能否顺利应用于舆情领域还需要结合舆情的时间、话题等属性和具体领域、具体事件的特点，实现舆情信息的全方位监管还需要不断优化信息抽取的精准性并探究复杂语境下的信息抽取。未来针对舆情信息的抽取应从多维度、多时期入手，不断丰富舆情领域的信息资源，将舆情信息的采集、组织以及分析系统化和流程化。

参 考 文 献

[1] MA Z K, TU Y. Online Emerging Topic Content Monitoring Based on Knowledge Graph[J]. Information Science, 2019, 37(2): 33-39.

[2] WANG X W, XING Y F, WEI Y N, et al. Research on the Topic Model Construction of Sentiment Classification of Public Opinion Users in Social Networks Driven by Big Data——Taking "Immigration" as the Topic[J]. Journal of Information Resources Management, 2020, 10(1): 29-38, 48.

[3] LIANG Y, LI X Y, XU H, et al. CLOpin: A Cross-Lingual Knowledge Graph Framework for Public Opinion Analysis and Early Warning[J]. Data Analysis and Knowledge Discovery, 2020, 4(6): 1-14.

[4] GUO X Y, HE T T. Survey about Reasearch on Information Extraction[J]. Computer Science, 2015, 42(2): 14-17, 38.

[5] HUANG W, XU Y J, HAN R X, et al. Study on Semantic Orientation Membership of Network Public Opinion[J]. Library and Information Work, 2015, 59(21): 27-32.

[6] QIAN S S, ZHANG T Z, XU C S. Survey of Multimedia Social Events Analysis[J]. Computer Science, 2021, 48(3): 97-112.

[7] ZHENG M, MA Y, ZHENG A, et al. Constructing method of public opinion knowledge graph with online news comments [C]// 2018 International Conference on Robots & Intelligent System(ICRIS). Changsha, China: IEEE, 2018: 404-408.

[8] ROSSI C, ACERBO F S, YLINEN K, et al. Early detection and information extraction for weather-induced floods using social media streams[J/OL]. International Journal of Disaster Risk Reduction, 2018, 30: 145-157. https://linkinghub.elsevier.com/retrieve/pii/S2212420918302735.

[9] NEMES L, KISS A. Information Extraction and Named Entity Recognition Supported Social Media Sentiment Analysis during the COVID-19 Pandemic[J]. Applied Sciences, 2021, 11(22): 11017.

[10] LI Z, DAI Y, LI X. Construction of sentimental knowledge graph of chinese government policy comments[J/OL]. Knowledge Management Research & Practice, 2021: 1-18. https://doi.org/10.1080/14778238.2021.1971056.

[11] WEI M Z, ZHANG H T, ZHOU H L. Research on the Management Thought of Chinese Ancient Library[J]. Information Science, 2021, 39(6): 10-18, 54.

[12] CHEN J Y, XIA L X, LIU X Y. Visual Analysis of Network Public Opinion Feature Evolution Based on Topic Map[J]. Information Science, 2021, 39(5): 75-84.

[13] SAHN X H, PANG S H, LIU X Y, et al. Research on Internet Public Opinion Event Prediction Method Based on Event Evolution Graph[J]. Information studies: Theory & Application,

2020,43(10):165-170,156.

[14] SHAN X H,PANG S H,LIU X Y,et al. Analysis on the Evolution Path of Internet Public Opinions Based on the Event Evolution Graph:Taking Medical Public Opinions as an Example[J]. Information studies: Theory & Application, 2019, 42(9): 99-103,85.

[15] WU F,ZHU P P,WANG Z Q,et al. Chinese Event Detection with Joint Representation of Characters and Words[J]. Computer Science,2021,48(4):249-253.

[16] QIU L Q,QU F S. Emotional map about emergency based on sentiment analysis and influence evaluation[J/OL]. Journal of Computer Applications. http://kns. cnki. net/kcms/detail/51. 1307. TP. 20210720. 1404. 003. html.

[17] LI T R,LIU M T,ZHANG Y J,et al. A Review of Entity Linking Research Based on Deep Learning [J]. Acta Scientiarum Naturalium Universitatis Pekinensis,2021,57(1):91-98.

[18] ZHUANG C Z,JIN X L,ZHU W J,et al. Deep Learning Based Relation Extraction:A Survey[J]. Journal of Chinese Information Processing,2019,33(12):1-18.

[19] HU Y,HUANG H,CHEN A,et al. Weibo-cov:a large-scale covid-19 social media dataset from weibo[J/OL]. arXiv: 2005. 09174,2020.

[20] WANG G,LIU S,WEI F. Weighted graph convolution over dependency trees for nontaxonomic relation extraction on public opinion information[J]. Applied Intelligence,2022,52(3):3403-3417.

[21] PENG N,DREDZE M. Named entity recognition for Chinese social media with jointly trained embeddings[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics,2015:548-554.

[22] ROSALES-MÉNDEZ H,HOGAN A,POBLETE B. Fine-grained evaluation for entity linking[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). 2019:718-727.

[23] HUANG Z,XU W,YU K. Bidirectional lstm-crf models for sequence tagging[J]. arXiv:1508. 01991,2015.

[24] ŽUKOV-GREGORI A,BACHRACH Y,COOPE S. Named entity recognition with parallel recurrent neural networks[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics, 2018:69-74.

[25] ELIGÜZEL N,ÇETINKAYA C,DERELI T. Application of named entity recognition on tweets during earthquake disaster:a deep learning-based approach [J]. Soft Computing, 2022, 26(1):395-421.

[26] QIU J,ZHOU Y,WANG Q,et al. Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field[J]. IEEE Transactions on Nano-Bioscience,2019,18(3):306-315.

[27] KONG J,ZHANG L,JIANG M,et al. Incorporating multi-level cnn and attention mechanism for chinese clinical named entity recognition[J/OL]. Journal of Biomedical Informatics, 2021, 116. https://linkinghub. elsevier. com/retrieve/pii/S1532046421000666.

[28] AGUILAR G,MAHARJAN S,LÓPEZ-MONROY A P,et al. A multi-task approach for named entity recognition in social media data[J/OL]. Proceedings of the 3rd Workshop on Noisy User-generated Text, 2017: 148-153. http://arxiv. org/abs/1906. 04135.

[29] RONRAN C,LEE S. Effect of character and word features in bi-directional lstm-crf for ner[C]// 2020 IEEE International Conference on Big Data and Smart Computing(BigComp). Busan, Korea(South):IEEE,2020:613-616.

[30] ALIFI M R,SUPANGKAT S H. Information extraction of traffic condition from social media using bidirectional lstm-cnn [C]// 2018 International Seminar on Research of Information Technology and Intelligent Systems(ISRITI). Yogyakarta, Indonesia:IEEE,2018:637-640.

[31] MOON S,NEVES L,CARVALHO V. Multimodal named entity recognition for short social media posts[J]. arXiv:1802. 07862, 2018.

[32] ARSHAD O,GALLO I,NAWAZ S,et al. Aiding intra-text representations with visual context for multimodal named entity recognition[J/OL]. arXiv:1904. 01356. http://arxiv. org/abs/1904. 01356.

[33] ASGARI-CHENAGHLU M,FEIZI-DERAKHSHI M R,FARZINVASH L, et al. A multimodal deep learning approach for named entity recognition from social media[J]. Neural Computing and Applications,2022,34(3):1905-1922.

[34] YAN H, DENG B, LI X, et al. TENER: adapting transformer encoder for named entity recognition[J]. arXiv: 1911. 04474, 2019.

[35] NIE Y,TIAN Y,WAN X,et al. Named entity recognition for social media texts with semantic augmentation[J]. arXiv:2010. 15458,2020.

[36] BAEVSKI A,EDUNOV S,LIU Y,et al. Cloze-driven pretraining of self-attention networks[J]. arXiv:1903. 07785,2019.

[37] LI X,SUN X,MENG Y,et al. Dice loss for data-imbalanced nlp tasks[J]. arXiv:1911. 02855,2020.

[38] LI X,YAN H,QIU X,et al. FLAT:Chinese NER using flat-lattice transformer[J]. arXiv:2004. 11795,2020.

[39] GUPTA N,SINGH S,ROTH D. Entity linking via joint encoding of types,descriptions,and context[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics,2017:2681-2690.

[40] PHAN M C,SUN A,TAY Y,et al. NeuPL:attention-based semantic matching and pair-linking for entity disambiguation [C]// ACM Conference on Information and Knowledge Management. Singapore:ACM,2017:1667-1676.

[41] SIL A,KUNDU G,FLORIAN R,et al. Neural cross-lingual entity linking[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[42] ESHEL Y,COHEN N,RADINSKY K,et al. Named entity disambiguation for noisy text[C]// Proceedings of the 21st Confe-

rence on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada: Association for Computational Linguistics, 2017: 58-68.

[43] MUELLER D, DURRETT G. Effective use of context in noisy entity linking[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 1024-1029.

[44] XUE M, CAI W, SU J, et al. Neural collective entity linking based on recurrent random walk network learning[C]// Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19). Macao, China, 2019: 5327-5333.

[45] YANG X, GU X, LIN S, et al. Learning dynamic context augmentation for global entity linking[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 271-281.

[46] HOU F, WANG R, HE J, et al. Improving entity linking through semantic reinforced entityembeddings [C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6843-6848.

[47] RAIMAN J, RAIMAN O. Deeptype: multilingual entity linking by neural type system evolution[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[48] LI T, YANG E, ZHANG Y, et al. Improving entity linking by encoding type information into entity embeddings[M]// Chinese Computational Linguistics. Cham: Springer International Publishing, 2021: 297-307.

[49] GANEA O E, HOFMANN T. Deep joint entity disambiguation with local neural attention[C]// Proceedings of the 2017 Conference on Empirical Methods in NaturalLanguage Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 2619-2629.

[50] PERSHINA M, HE Y, GRISHMAN R. Personalized page rank for named entity disambiguation[C]// Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics, 2015: 238-243.

[51] FRANCIS-LANDAU M, DURRETT G, KLEIN D. Capturing semantic similarity for entity linking with convolutional neural networks[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, 2016: 1256-1261.

[52] BAI L, JIN X L, XI P B, et al. A Surney on Distant Supervision Based Relation Extraction[J]. Journal of Chinese Information Processing, 2019, 33(10): 10-17.

[53] ZHONG Z, CHEN D. A frustratingly easy approach for entity and relation extraction[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 50-61.

[54] LIU C Y, SUN W B, CHAO W H, et al. Convolution neural network for relation extraction[C]// International Conference on Advanced Data Mining and Applications. Berlin: Springer, 2013: 231-242.

[55] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C]// Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014: 2335-2344.

[56] DOS SANTOS C, XIANG B, ZHOU B. Classifying relations by ranking with convolutional neural networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015: 626-634.

[57] WANG L, CAO Z, DE MELO G, et al. Relation classification via multi-level attention cnns[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 1298-1307.

[58] LEE J, SEO S, CHOI Y S. Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing [J/OL]. Symmetry, 2019, 11 (6): 785. https://www.mdpi.com/2073-8994/11/6/785.

[59] WU S, HE Y. Enriching pre-trained language model with entity information for relation classification[C]// CIKM '19: The 28th ACM International Conference on Information and Knowledge Management. Beijing China: ACM, 2019: 2361-2364.

[60] BALDINI SOARES L, FITZGERALD N, LING J, et al. Matching the blanks: distributional similarity for relation learning [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 2895-2905.

[61] WANG Y, YU B, ZHANG Y, et al. TPLinker: single-stage joint extraction of entities and relations through token pair linking [C]// Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: International Committee on Computational Linguistics, 2020: 1572-1582.

[62] REN F, ZHANG L, YIN S, et al. A novel global feature-oriented relational triple extraction model based on table filling[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 2646-2656.

[63] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme[J]. arXiv: 1706.05075, 2018.

[64] YU B, ZHANG Z, SHU X, et al. Joint extraction of entities and relations based on a novel decomposition strategy[J]. arXiv: 1909.04273, 2020.

[65] ZENG X, ZENG D, HE S, et al. Extracting relational facts by an end-to-end neural model with copy mechanism[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Australia: Association for Computational Linguistics, 2018: 506-514.

[66] ZENG D,ZHANG H,LIU Q. CopyMTL:copy mechanism for joint extraction of entities and relations with multi-task learning [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:9507-9514.

[67] YANG H Z,LIU Y X,ZHANG K W,et al. Survey on Distantly-Supervised Relation Extraction[J]. Chinese Journal of Computers,2021,44(8):1636-1660.

[68] CHEN Y,XU L,LIU K,et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing(Volume 1:Long Papers). Beijing,China:Association for Computational Linguistics,2015:167-176.

[69] ZENG Y,YANG H,FENG Y,et al. A convolution bilstm neural network model for chinese event extraction[M]// Natural Language Understanding and Intelligent Applications. Cham:Springer International Publishing,2016:275-287.

[70] NGUYEN T H,CHO K,GRISHMAN R. Joint event extraction via recurrent neural networks[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. San Diego,California:Association for Computational Linguistics,2016:300-309.

[71] FENG X,QIN B,LIU T. A language-independent neural network for event detection [J]. Science China Information Sciences,2018,61(9):1-12.

[72] ORR J W,TADEPALLI P,FERN X. Event detection with neural networks:a rigorous empirical evaluation[J]. arXiv:1808.08504,2018.

[73] SHA L,QIAN F,CHANG B,et al. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[74] YU W,YI M,HUANG X,et al. Make it directly:event extraction based on tree-lstm and bi-gru[J]. IEEE Access,2020,8:14344-14354.

[75] LIU X,LUO Z,HUANG H. Jointly multiple events extraction via attention-based graph information aggregation[J]. arXiv:1809.09078,2019.

[76] YAN H,JIN X,MENG X,et al. Event detection with multi-order graph convolution and aggregated attention[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong,China:Association for Computational Linguistics,2019:5765-5769.

[77] BALALI A,ASADPOUR M,CAMPOS R,et al. Joint event extraction along shortest dependency paths using graph convolutional networks[J]. Knowledge-Based Systems,2020,210:106492.

[78] YANG S,FENG D,QIAO L,et al. Exploring pre-trained language models for event extraction and generation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence,Italy:Association for Computational Linguistics,2019:5284-5294.

[79] KAN Z,QIAO L,YANG S,et al. Event arguments extraction via dilate gated convolutional neural network with enhanced local features[J]. IEEE Access,2020,8:123483-123491.

[80] WANG Z,WANG X,HAN X,et al. CLEVE:contrastive pre-training for event extraction[J]. arXiv:2105.14485,2018.

[81] BANARESCU L,BONIAL C,CAI S,et al. Abstract meaning representation for sembanking[C]// Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. 2013:178-186.

[82] WANG R,ZHOU D,HE Y. Open event extraction from online text using a generative adversarial network[J]. arXiv:1908.09246,2019.

[83] LIU J,CHEN Y,LIU K,et al. Event extraction as machine reading comprehension[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics,2020:1641-1651.

[84] YU S Y,GUO S M,HUANG R Y,et al. Overview of Nested Named Entity Recognition[J]. Computer Science,2021,48(S2):1-10,29.

WANG Jian, born in 1978, Ph. D, professor, is a member of China Computer Federation. Her main research interests include multimedia social networks, information security, trusted computing and usage control.

(责任编辑:喻藜)