第 47 卷　第 4 期　　西南师范大学学报（自然科学版）　　2022 年 4 月
Vol. 47　No. 4　　Journal of Southwest China Normal University (Natural Science Edition)　　Apr.　2022

DOI：10.13718/j.cnki.xsxb.2022.04.001

基于深度神经网络的信息抽取研究综述①

代建华，　彭若瑶，　许路，　蒋超，　曾道建，　李扬定

湖南师范大学 语言与文化研究院/智能计算与语言信息处理湖南省重点实验室，长沙 410081

摘要：信息抽取任务旨在从非结构化的文本中抽取出结构化的信息，帮助将海量信息进行自动分类、提取和重构，提高信息的利用率. 目前，基于深度神经网络的信息抽取技术已经成为自然语言处理领域最重要的研究主题之一，它提供了分析非结构化文本的有效手段，是实现大数据资源化、知识化和普适化的核心技术，此外进一步为更高层面的应用和任务提供了支撑. 文章对基于深度神经网络的信息抽取相关研究进行了综述，首先，简要概述了信息抽取的任务定义、目标和意义，然后，回顾了信息抽取任务的发展历程，接下来，从实体抽取、实体关系抽取、事件抽取和事件关系抽取 4 个方面梳理了近几年关键技术的研究进展. 最后，文章对信息抽取领域的未来发展趋势进行了分析和展望.

关　键　词：信息抽取；深度神经网络；实体抽取；实体关系抽取；事件抽取；事件关系抽取

中图分类号：TP18　　**文献标志码**：A　　**文章编号**：1000－5471(2022)04－0001－11

A Survey of Information Extraction Based on Deep Neural Networks

DAI Jianhua，　PENG Ruoyao，　XU Lu，
JIANG Chao，　ZENG Daojian，　LI Yangding

Research Institute of Languages and Cultures / Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha 410081, China

Abstract: Information extraction aims at extracting structured information from unstructured text, realizing automatically classify, extracting and reconstructing massive information, and enhancing the use of information. Recently, information extraction technology based on deep neural network is one of the most significant research topics in the field of natural language processing. It creates an effective way of analyzing unstructured text, and facilitates the realize the resource, knowledge and universality of big data. In addition, it further provides support for higher-level applications and tasks. In this paper, the related research on information extraction based on deep neural network has been reviewed. First, the task definition, goals, and meanings of information extraction has briefly been described, followed by an analysis of

① **收稿日期**：2021－12－24
基金项目：国家自然科学基金项目(61602059)，湖南省自然科学基金项目(2020JJ4624)，国家社会科学基金项目(20&ZD047)，湖南省教育厅科研基金项目(19A020)，湖南师范大学语言与文化研究院青年培育项目 (2020QNP05).
作者简介：代建华，教授，主要从事人工智能、数据挖掘、机器学习、粗糙集理论、智能信息处理等方面的研究.

the development of the task. And then, the development of key technologies in recent years been summarized from four aspects: entity extraction, entity relation extraction, event extraction, and event relation extraction. Finally, the future development trends in the field of information extraction have been analyzed and looked forward to.

Key words: information extraction; deep neural network; entity extraction; entity relation extraction; event extraction; event relation extraction

信息抽取旨在从非结构化文本中抽取出结构化的信息，例如从病人的医疗记录中抽取出症状、检验结果等一系列信息，主要包括实体抽取、实体关系抽取、事件抽取和事件关系抽取等任务[1]. 实体抽取也称为命名实体识别(Named Entity Recognition, NER)，是指从文本中抽取出具有特定意义的实体，如人名、地名、机构名、专有名词等，命名实体识别在各种自然语言处理应用中发挥着重要作用. 实体关系抽取则是判断2个实体之间的语义关系. 事件抽取任务旨在识别特定类型的事件，并把事件中担任既定角色的要素以结构化的形式呈现出来，该任务可进一步分解为4个子任务：触发词识别、事件类型分类、事件论元识别和角色分类任务. 其中，触发词识别即识别出句子中促使事件发生的核心词，通常是名词或者动词. 例如，“2008年北京举办了奥运会”中的“奥运会”就是名词性触发词. 对于事件类型分类旨在判断句子中触发词所对应的事件类型，包括出生、结婚、死亡等. 事件论元识别即识别出事件中的论元，也就是事件的参与者，主要由实体、值、时间组成. 角色分类则是判断句子中触发词和实体之间的角色关系，如攻击者、受害者等. 事件抽取属于信息抽取领域中的深层次研究内容，它需要前述的几项研究作为基础，涉及自然语言处理、机器学习、模式匹配等多个学科的方法与技术，在信息抽取、情报学等领域都有很好的应用前景. 事件关系反映了事件之间的一种语义关系，可以为文本的深层理解提供关键线索，事件关系抽取的目的则是提取一段文本内容中2个事件可能存在的关系，例如“昨天突然降温导致小王感冒了，所以昨晚去了医院”，其中事件对(降温，感冒)、(感冒，去医院)存在因果关系. 事件之间存在多少种关系类型仍然是一个有争议的问题，目前事件关系抽取主要研究共指关系、因果关系和时序关系，此外关系文本的多样性和隐含性使得从文本中识别不同类型的事件关系面临巨大挑战.

自20世纪80年代信息理解会议(Message Understanding Conference, MUC)提出信息抽取任务以来，信息抽取一直是自然语言处理的研究热点. 早期主要采用基于规则的方法，该类方法依靠人工制定规则，其优点是可解释性强而且不需要太多已标注的语料库，但是总结规则模板耗时长且模板可移植性差，此外规则和词典的维护任务也很繁重. 由于该类方法的诸多限制，研究人员开始使用统计模型辅助机器学习算法来进行信息抽取，主要通过一些统计学的手段对文本的一些特征信息进行统计，然后利用机器学习拟合从输入到输出过程中的模型参数. 随着深度学习时代的来临，基于深度学习的信息抽取模型受到广泛关注，研究者主要聚焦于如何使用深度学习方法自动提取句子中的有效特征，不仅可以弥补传统工程的缺点，还可以避免使用传统自然语言处理工具抽取特征时存在的错误累积问题[2]. 随着研究的深入，特别是大规模预训练语言模型(Pre-trained Language Models, PLMs)的引入[3]，基于深度神经网络的信息抽取模型在公开数据集取得了不错的成绩. 信息抽取技术是中文信息处理和人工智能的核心技术，具有重要的科学意义. 通过将文本所表述的信息结构化和语义化，信息抽取技术提供了分析非结构化文本的有效手段，是实现大数据资源化、知识化和普适化的核心技术. 被抽取出来的信息通常具有结构化的形式描述，计算机可以直接处理，从而实现对海量非结构化数据的分析、组织、管理、计算、查询和推理，并进一步为更高层面的应用和任务(如自然语言理解、知识库构建、智能问答系统、舆情分析系统)提供支撑. 本文将对基于深度神经网络的信息抽取相关研究工作进行综述.

1 技术方法和研究进展

信息抽取的核心是将非结构化的自然语言映射为结构化表示，并转换为可供计算机处理的知识. 然而，

自然语言表达具有多样性、歧义性和结构性的特点，其中蕴含的知识也具有复杂性、开放性以及规模巨大的特点，进而导致信息抽取任务极具挑战性.

在早期，大部分信息抽取系统采用基于规则的方法，该类方法依靠人工制定抽取模板，优点是可预判和解释，但这种方法有其自身的局限性，如移植性差，很多场景很难甚至无法总结出有效的规则. 自 90 年代以来，统计模型成为主流方法，通常将信息抽取任务形式化为从文本输入到特定目标结构的预测，使用统计模型来建模输入与输出之间的映射，并使用机器学习方法来学习模型参数. 随着深度学习时代的来临，研究者开始探索如何使用深度神经网络自动学习有区分性的特征，进而避免使用传统自然语言处理抽取特征时存在的错误累积问题[2]. 随着研究的深入，大规模预训练语言模型[3]加下游任务微调的范式成为主流，基于深度神经网络的信息抽取模型性能得到很大提升.

1.1　实体抽取

实体抽取也称为命名实体识别，是自然语言处理领域的基础且重要任务之一，其目标是从非结构化的文本中抽取出具有特定意义或者指代性强的实体，实体抽取任务是文本理解、机器翻译、信息检索、知识库建设等众多自然语言处理任务的基础工具. 最早在第六届信息理解会议(MUC6)上提出“命名实体”并阐明了人员、组织和本地化的语义识别以及时间和数量等数字表达式的重要性. 自 MUC6 以来，研究人员对实体抽取的研究兴趣越来越浓厚. 传统的实体抽取方法主要分为三大类：基于规则的方法、基于无监督学习的方法、基于有监督学习的方法.

早期的实体抽取系统使用手工构造的基于规则的算法，这类算法通过选择标点符号、关键字、指示词等特征，采用专家构造的规则模板，通过模式和字符串匹配手段来识别实体，这类算法往往依赖于知识库或者词典. 1991 年 Rau 在第 7 届 IEEE 人工智能应用会议上首次发表了抽取和识别公司名称的研究成果，该成果主要采用启发式算法和手工编写规则的方法[4]. 随后，多个基于规则的命名实体识别系统被相继提出. 基于规则的方法不仅可解释性强而且不需要太多已标注好的语料库，但其局限性在于构建成本高，且规则针对特定领域，扩展性较差. 无监督学习的方法主要是基于聚类，通常根据上下文相似度，利用分布统计信息从未标记的文本中提取出实体. 随着大量标准语料的出现，研究者开发了许多基于监督学习的实体抽取系统，该方法通常将实体抽取任务转换为多分类任务或者序列标注任务，其主要步骤是从训练样本中学习出一个函数，通过这个函数预测出新样本的结果. 上述这些方法在很大程度上依赖于特征工程和特定的训练数据，当设计的特征不适合任务时，可能会导致误差传播的问题.

近年来，深度学习成为了主流，并受到了各大领域的广泛关注，在过去几年里，深度神经网络也被引入到实体抽取任务中来并取得了显著的成功，针对该任务提出了各种各样的深度神经网络模型. 深度神经网络具有强大的特征提取能力，能够自动学习任务所需的有效特征. 循环神经网络(Recurrent Neural Network，RNN)、门控循环单元(Gated Recurrent Unit，GRU)、长短期记忆网络(Long Short-Term Memory，LSTM)等模型在序列数据建模方面取得了显著成绩，这些神经网络模型中往往使用的都是词向量表示，对于词表外的词则无法得到很好的表示，因此许多研究人员通过添加字符向量来解决该问题. Li 等[5]提出了一种基于 RNN 的方法 WCP-RNN 来提取中文生物医学命名实体，不仅考虑了词向量还加入了字符向量来捕获正字法和词汇语义特征，此外还将 POS 标签作为先验词信息以提高最终性能. Huang 等[6]借助了条件随机场(Conditional Random Field，CRF)，提出利用 BiLSTM-CRF 架构来解决命名实体识别任务，BiLSTM 相较于普通的 RNN 对长序列具有更好的建模能力，CRF 也可以有效利用句子级的标注信息，此外该模型具有较好的鲁棒性. 随后，一系列基于 BiLSTM-CRF 的工作被提出[7-8]. Lample 等[8]利用 BiLSTM 来提取单词的字符级表示，提出了一种将 BiLSTM 与 CRF 结合用于英文序列标签预测的任务. RNN 对时间序列处理的很好，但是忽略了空间上下文的问题，因此许多研究人员将 CNN 引入实体抽取中获得了不错的性能. Ma 等[7]提出的 BiLSTM-CNN-CRF 模型在 BiLSTM-CRF 中加入了 CNN，能够提高模型提取词语上下文特征的能力. 注意力机制[9]在机器翻译、图像分类、语音识别等许多自然语言处理应用中发

挥了巨大的优势，一些研究人员试图将其应用于实体抽取任务中[10-11]，注意力机制有助于关注输入序列的相关部分，并捕捉较长序列的长期依赖，能够有效地从高维数据中提取特征. Zheng 等[11]将卷积神经网络与局部注意力机制结合形成卷积注意力层，用于捕捉局部上下文信息，并在双向门控循环单元(Bidirectional Gated Recurrent Unit，BiGRU)和 CRF 层中间应用全局注意力机制优化对句子级信息的处理. 通过大量研究结果表明，注意力机制可以提高 NER 模型的性能.

BERT 等大规模预训练语言模型在多项自然语言处理任务中获得了较高的性能，此外许多研究引入深度学习模型进行微调以更好完成实体抽取任务，整合或微调预训练语言模型的嵌入已成为深度神经网络的新范式. 一方面这些嵌入表示是随上下文变化的，并且可以和传统的嵌入表示相结合，在多种任务上取得了较好的效果. 另一方面，通过微调预训练语言模型能够迁移至其他各项任务中. Li 等[12]在未标记的中国临床记录上预训练了 BERT 模型，同时他们证明相较于只进行微调的 BERT 模型，在经过微调的预训练 BERT 模型之上添加 BiLSTM-CRF 层效果更好. 此外，BGRU-CRF 层在优化后的 BERT 模型上也取得了良好的效果[13].

1.2 实体关系抽取

实体关系抽取也称为关系事实抽取，也是自然语言处理领域的基础任务之一，其目标是在自然语言文本中识别出成对的命名实体，并抽取出实体对之间的关系，生成关系三元组. 即将非结构化文本转化为结构性的知识，在知识图谱的构建、融合等方面发挥重要作用[14-15].

早期的实体关系抽取方法可分为基于规则的方法和基于统计模型的方法. 基于规则的方法需要领域专家和语言学家之间的合作，依靠人工制定规则，构建基于单词、文本片段或语义的模式知识集. 有了这些语言知识和专业领域知识，可以通过将预处理的文本进行模式匹配来实现实体关系抽取. 其优点是可预判和解释，但面临移植性差的缺点，依赖于人工穷举规则，然而有的运用场景很难总结出有效的规则，更无法做到穷举. 基于统计模型的方法通常将实体关系抽取任务转为从文本输入到特定目标结构的预测，使用统计模型来建模输入与输出之间的关联，并使用机器学习方法来学习模型的参数. 可以细分为有监督方法、半监督方法、远程监督方法和无监督方法. 这一类方法相对于基于规则的方法有明显改进，但总体上仍依赖于人工提取的特征，领域适应性差，在实际的运用场景中效果欠佳，且使用传统工具抽取特征将不可避免存在误差传播的问题.

为了解决这些问题，研究者开始探索如何使用深度神经网络自动学习到有区分性的特征，进而避免使用传统自然语言的特征抽取方法，最初大多采用流水线方法进行关系抽取，该方法将关系抽取任务拆解为命名实体识别和关系分类 2 个任务，并为 2 个任务分别训练模型，即先对命名实体识别模块进行训练，再将已训练完成的命名实体识别模块的输出作为关系分类任务的输入，用以实现对关系分类模块的训练. 但流水线方式存在命名实体识别模块的错误向后续模块传递的问题，并且忽视了 2 个任务之间的关联性，捕获复杂语义关系能力较弱. 此外，它们假设句子中只有一个关系实例，而忽略了关系之间潜在的相互依赖关系. 由于流水线方法的固有缺陷，研究人员将目光转向了实体关系联合抽取. Getoor 等[16]首先指出 NER 和关系分类是密切相关的，他们使用单独的分类器识别句子中可能的实体和关系，这些分类器的输出用于使用线性规划计算最佳关系事实. 联合抽取方法主要分为两类，一类是表填充方法，另一类是序列到序列方法. Miwa 等[17]首先提出了表填充方法，设计了一个实体关系表，将关系抽取任务转化为表格填充. 对于表填充方法，Gupta 等[18]使用神经网络对实体和关系进行联合建模，并开发了一种上下文感知的循环方式来学习关系的相互依赖性. 但表填充方法只能为每个实体对预测唯一的一个关系，若一个实体对存在多种关系则无法被完整抽取，即无法解决实际运用场景中的三元组重叠问题. Bekoulis 等[19]添加了一个额外的 CRF 层来标记实体，并设计了一种新的表格方案，即多头选择，用以解决三元组重叠问题.

序列到序列方法将关系抽取任务重新定义为序列生成问题. 这种模型的主干是编码器-解码器结构. 编码器以一个句子作为输入，解码器需要自动生成序列结果，该结果可以进一步转换为关系三元组. 序列到

序列的方法最早由 Zeng 等[20]提出，该方法将关系抽取视为生成三元组的任务，其中实体从源句中复制，关系则由预定义的关系集预测. Zeng 等[21]进一步应用强化学习来学习提取顺序. 这 2 项探索性研究存在一个问题，即无法处理由多个词构成的实体. 为此，Zeng 等[22]在编码器部分添加了序列标注层，以帮助实体识别. 基于序列到序列的方法已经取得了很大的进展，但仍然存在一些缺陷. 首先是前向解码错误，解码器基于 RNN，以自回归方式从左到右生成关系三元组. 当前的三元组预测依赖于先前的三元组，一旦在某一步中出现解码错误，由于噪声左侧上下文的负面影响，后续预测将进一步积累误差. 其次是忽略了关系共现信息，来自同一句子的三元组之间的关系有很强的相关性，我们将其称为关系共现信息，这些信息可以用来预测一些仅考虑句子本身时难以预测的关系. 随着深度神经网络的不断发展，基于序列到序列方法的大规模预训练语言模型在关系抽取任务中取得了很好的效果. 但是，这些方法仍然难以满足实际的应用场景. 对于金融、医疗等垂直领域，缺失标注数据现象更为明显，甚至数据获取也很困难，而神经网络作为典型的“数据饥渴”模型，在训练样例过少时性能会受到极大影响. 针对小样本任务，Han 等[23]发布了小样本关系抽取数据集 FewRel，Gao 等[24]在 FewRel 数据集的基础上提出了 FewRel 2.0，增加了领域迁移和“以上都不是”检测. 利用海量无监督数据得到的预训练模型得到有效的语义特征是少量样本快速学习知识的代表性方法.

真实场景中的关系还面临着复杂的语境，例如，大量的实体间关系是通过多个句子表达的，同一个文档中的多个关系相互关联. 文档级的关系抽取最近也受到广泛的关注，代表性的方法是使用 GNN 融合分布在文档中不同位置的实体信息，并利用图算法进行信息的传递. Christopoulou 等[25]构建以实体、实体提及和句子为节点的文档图，并通过图上的迭代算法得到边的表示进行关系分类，之后有大量的研究者采用类似的方法对文档建模. 除使用图网络外，研究者也开始尝试直接使用大规模预训练语言模型建模文档[26-27]. Zhou 等[27]提出自适应阈值代替用于多标签分类的全局阈值，并直接利用预训练模型的自注意力得分找到有助于确定关系的相关上下文特征.

1.3　事件抽取

事件是指在特定时间和特定地点发生的某件事，涉及一个或多个参与者，通常可以描述为状态的变化. 事件抽取任务旨在识别特定类型的事件，并以结构化的形式把事件中担任既定角色的要素呈现出来，该任务可进一步分解为 4 个不同难度的子任务：触发词识别、事件类型分类、论元识别和角色分类任务. 事件抽取属于信息抽取领域中的深层次研究内容，它需要前述的几项研究作为基础. 在实际应用中，事件抽取在信息检索、问答、知识图谱构建等领域中得到了广泛的应用，具有重要的研究意义和实用价值.

最早的事件抽取方法主要是基于规则的方法，后来逐渐发展为模式匹配的方法，它首先构造一些特定的事件模板，然后通过各种模式匹配算法从文本中提取出符合模式约束条件的事件. 第一个基于模式的事件抽取系统来自 1993 年 Riloff 等人开发的用以提取恐怖事件的 autolog 系统[28]，autolog 利用了一组语言模板和一个手工标注的语料库来获取事件模式. 随后，也有研究者提出使用弱监督方法或自扩展技术，通过使用少量的预分类训练语料库或种子模式来自动获取更多的模板. 基于模式的事件抽取技术在许多工业领域中得到了广泛的使用，但由于其成本过高，因此各种基于机器学习的事件抽取技术得到了快速发展. 基于机器学习的方法提取事件本质上是将事件抽取作为一个分类问题，其核心在于分类器的构建和特征的选择. 其主要过程是从训练样本中学习分类器，然后应用分类器从文本中提取事件. 基于机器学习的方法能够有效地捕捉触发词、论元以及触发词之间关系的语义信息，具有较高的可移植性和灵活性.

特征工程是基于机器学习的事件抽取任务的主要难点，与经典的机器学习技术相比，近年来发展迅速的深度学习方法可以自动提取句子中的显著特征，不仅可以使特征较好地适应其他特定的领域，而且可以通过学习不断地自动更新特征表示. 研究人员将许多深度神经网络模型引入到事件抽取任务中，通常将单词表示作为输入，分类这些词是否为事件触发词. Nguyen 等[29]将事件抽取问题形式化为一个多分类问题，利用卷积神经网络自动从预训练的词嵌入、位置嵌入和实体类型嵌入中学习特征表示，克服了复杂的特征

工程和误差传播，但它依赖于其他监督模块来获取特征. Chen 等[30]提出了一种动态多池化卷积神经网络 DMCNN，根据事件触发词和论元使用一个动态多池层来保留更多关键信息，他们引入了一个单词表示模型来捕获词级别的语义信息，并采用基于 CNN 的框架来捕捉句子级别的特征. 基于 CNN 的事件抽取模型的缺点是不能很好地捕获远距离的词之间的语义依赖，而 RNN 结构理论上可以对任意距离的 2 个词进行建模表示. 因此，许多研究者尝试将循环神经网络应用于事件抽取中. 如 Sha 等[31]提出的 dbRNN 模型将 2 个 RNN 神经元的句法依赖连接添加到双向 RNN 中. 上述结构均难以处理图形数据结构，并且它们不能完全模拟词间的依赖关系. 研究者开始将 GNN 引入到事件抽取[32]中，其核心问题是为文本中的单词构建一个图. 随着 Transformer[9]的提出，自注意力机制发挥了巨大的作用，研究者们提出了许多基于自注意力机制来学习句子中每个单词的重要程度. Ahmad 等[33]提出的 GATE 框架引入了一种自注意力机制来学习不同句法距离单词之间的依赖关系，一方面 GATE 具有捕捉长距离依赖关系的能力，另一方面，GATE 使用句法距离来建模单词之间的成对关系从而使其适合在不同类型的语言之间转换.

随着 BERT 的成功，预训练语言模型也被用于事件抽取[34]中. 由于预训练语言模型使用了大量的未标记数据进行学习，相较于传统的神经网络，使用预训练语言模型进行特征学习有很大的改进. 而 Wadden 等[34]发现基于预训练语言模型的工作通常只专注于更好地微调，因此转而研究如何在大规模无监督数据中更好地利用丰富的事件知识来提高性能. 随着预训练语言模型体量的不断增大，对其进行微调的硬件要求、数据需求和实际代价也在不断上涨. 基于预训练提示(Prompt)学习范式[35]的方法允许语言模型在大量原始文本上进行预训练，通过定义一个新的提示模板，能够应用在少样本甚至零样本学习中. Si 等[36]第一次将基于 Prompt 的学习策略引入事件抽取领域中，自动利用输入和输出端的标签语义. 此外，现有事件抽取方法的研究大多集中在句子层面，即假设一个事件往往在一个句子中得以表示，通常使用句子级别的上下文中的局部信息，然而很多情况下事件信息分散在整个文档中，这种情况需要有更多的全局信息，因此如何提取文档级别的事件成为了重点研究对象，催生出文档级事件抽取的任务，文档级的事件抽取任务更具挑战性，需要考虑多事件表达等问题. 早期研究者用基于模式和基于分类器的方法来解决这个问题，直到最近几年，研究人员开始引入深度神经网络. 例如，Zhang 等[37]提出一种两步方法，通过检测句子中的隐含论证来连接论证；Li 等[38]扩展了该任务，提出一种基于条件文本生成的端到端神经事件参数提取模型.

1.4 事件关系抽取

事件关系反映了事件之间的一种语义关系，为文本的深层理解提供了关键线索. 事件关系抽取的目的主要是提取一段文本内容中 2 个事件可能存在的关系，它在文本理解、逻辑推理和知识图谱构建等众多应用中都发挥出了重要作用. 现有事件关系抽取研究主要包括共指关系抽取、因果关系抽取以及时序关系抽取.

1.4.1 事件共指关系抽取

共指关系抽取旨在确定文档中已识别的多个事件实例是否指向同一个事件. 共指关系可以当作一个分类任务处理，即看 2 个事件是否指向同一事件类型. 经典的机器学习模型如决策树算法、最大熵算法、支持向量机等被应用于共指关系抽取中. 基于机器学习的方法首先统计每个事件的上下文文本特征，比如词频特征、位置特征、句法特征、事件主题信息、语言特征等，然后利用机器学习方法进行二分类.

基于深度学习的共指关系抽取通常使用 CNN 或者 RNN 对事件的上下文信息进行特征提取，然后对所提取的信息进行动态池化整合，最后进行分类. Lee 等[39]提出端到端的模型，在不需要输入额外的特征的情况下，利用 BiLSTM 提取特征取得了好的效果. 使用 CNN 来提取出单词的上下文特征信息，只考虑了句子中单词与单词间的局部信息，并未注意到上下文对共指判断的影响，因此，注意力机制也被应用到事件关系抽取中. Bugert 等[40]提出一种多注意力机制的卷积神经网络模型，主要解决了事件特征难以获取的问题. 通过使用深层的 CNN 建立语言模型，自动地获取事件特征，并使用注意力机制进行加权，筛选重要

的特征，融合 2 个事件的特征，判断 2 个事件是否同指. 此外，论元兼容性经常被纳入事件共指关系抽取中作为判断依据，即若 2 个事件在任何一个论元角色中有不相容的论证，它们就不能是共指的. Huang 等[41]提出了一个迁移学习框架，利用大量的未标记数据来学习 2 个事件提及之间的兼容性.

1.4.2　事件因果关系抽取

事件因果关系抽取旨在识别文本中事件之间的因果关系，为逻辑推理、问题回答等 NLP 任务提供了关键线索. 现有的方法通常将事件因果关系抽取作为一项分类任务，通过判断 2 个句子的事件触发词来确定它们之间是否存在因果关系，或者进一步预测相应的因果关系类型. 早期的事件关系抽取方法主要使用特征工程的方法，为了提取表明事件因果关系的有效线索，研究人员探索了各种文本特征，例如，词汇和句法特征、时间模式等.

随着深度学习的发展，基于深度神经网络的方法被用于事件因果关系抽取任务中，Dasgupta 等[42]通过基于 LSTM 的模型从语言的角度确定了文本中因果关系的语言表达. 上述方法针对句子或者跨度不大的段落，而现实过程中往往需要对文档级别的文本内容进行因果抽取，此时就需要进一步考虑句子和句子之间、句子和段落之间以及段落和段落之间的关系. 近年来，研究者尝试利用图神经网络表示文档建模中各种不同粒度的信息，通过图卷积、随机游走等算法融合不同级别的节点信息，从而将局部信息和整体信息整合到一起，取得了较好的效果. 针对文档级别的事件因果关系抽取，Phu 等[43]提出了一个基于图卷积神经网络的模型，通过构建交互图来捕获输入文档中针对事件的重要对象之间的相关链接. 此外，为了增强对因果关系的表征，特别是当文本过短或者文本包含信息量过少时，可能没有充足的依据来判断，引入外部特征或者常识经验对事件因果关系的抽取发挥了促进作用，Liu 等[44]尝试将外部知识融入事件因果关系抽取任务中增强推理能力，还提出了一种模型泛化机制来学习事件无关的、上下文特定的模式，提高了模型的泛化能力. Cao 等[45]同时利用描述性知识和关系知识解决文本中缺乏明确因果线索的问题.

1.4.3　事件时序关系抽取

事件时序关系抽取任务旨在抽取事件之间的时间先后顺序关系. 近年来，时序关系抽取的主流研究主要基于 TimeML 格式[46]，它是标识事件、时间及其相互关系中使用最广泛的标注体系. 除了时间规则外，还使用一些通过统计上下文特征来构建基于机器学习的模型，这种方法通常根据事件的属性、语法等信息给出事件对的特征空间，利用机器学习算法给出关系抽取模型，并通过该模型预测事件对所属的时序关系类别. 最近，许多神经网络模型被用来捕捉时序，例如 RNN 方法[47]，以及利用神经网络和预训练语言模型构建的端到端系统[48].

2　发展趋势

信息抽取技术研究蓬勃发展，已经成为了自然语言处理和人工智能领域的重要分支. 这一方面得益于一系列国际权威评测和会议的推动，如消息理解系列会议，自动内容抽取评测(Automatic Content Extraction，ACE)和文本分析会议系列评测(Text Analysis Conference，TAC). 另一方面也是因为信息抽取技术的重要性和实用性，使其同时得到了研究界和工业界的广泛关注. 纵观信息抽取研究发展的态势和技术现状，本文认为信息抽取的发展方向包括：

1)高效的小样本学习能力

目前的小样本学习设定需要用一个巨大的训练集来训练，测试时只给出 N 个类别，每类 K 个样本，在这 $N*K$ 个样本上学习并预测. 真实场景下的小样本学习不存在巨大的训练集，从 GPT3 开始，Prompt 学习范式受到研究者的关注，该范式将下游任务也建模成语言模型任务，给出几条或几十条样本作为训练集，借助大规模预训练语言模型中蕴含的大量知识，取得了不错的小样本学习效果. 此外，相对于传统的 Pretrain+Finetune 范式，Prompt 可以摆脱指数级的预训练参数量对巨大计算资源的需求，能高效地利用预训练模型. 基于上述分析，本文认为信息抽取的发展方向之一就是利用预训练提示学习范式进行高效的

小样本学习.具体包括:①提示学习中信息抽取任务模板的设计;②模板的自动学习与挖掘;③预训练提示学习范式进行信息抽取的理论分析.

2)多模态信息融合

目前信息抽取主要针对的是纯文本数据,而常见的文档具有多样的布局且包含丰富的信息,文档以富文本的形式呈现,其中包含大量的多模态信息.从认知科学的角度来说,人脑的感知和认知过程是跨越多种感官信息的融合处理,如人可以同时利用视觉和听觉信息理解说话人的情感,可以通过视觉信息补全文本中的缺失信息等,信息抽取技术的进一步发展也应该是针对多模态的富文档.多模态信息的融合也是信息抽取的重要发展方向,具体包括:①多模态预训练模型的设计;②多模态信息抽取框架中跨模态对齐任务设计;③多模态信息的提取和表示.

3)数据驱动和知识驱动融合

现有神经网络信息抽取方法依靠深度学习以数据驱动的方式得到各种语义关系的统计模式,其优势在于能从大量的原始数据中学习相关特征,比较容易利用证据和事实,但是忽略了专家知识.单纯依靠神经网络进行信息抽取,达到一定准确率之后,就很难再改进.从人类知识获取方式来看,很多决策判断的同时要使用先验知识以及现有数据.数据驱动和知识驱动结合是模拟人脑进行信息抽取的关键所在.基于上述分析,本文认为构建数据驱动和知识驱动融合的抽取技术是信息抽取的发展方向,具体包括:①基于神经符号学习的信息抽取框架设计;②学习神经网络到逻辑符号的对应关系;③神经网络对于符号计算过程进行模拟.

3 结论

信息抽取技术是自然语言处理的核心技术,它将文本所表述的信息结构化和语义化,并将其作为计算机的输入,供机器识别并进行处理,实现了对海量非结构化数据的分析、组织、管理、计算、查询和推理,并进一步为更高层面的应用和任务提供支撑.本文介绍了信息抽取任务的研究概况,总结了近些年来信息抽取以及关键技术的研究进展,主要分析了3种信息抽取的方法,即基于规则和模板的抽取方法、基于传统机器学习的抽取方法以及基于深度神经网络的抽取方法,其中重点介绍了基于深度神经网络的方法.最后,总结了信息抽取领域的未来的发展方向,包括高效的小样本学习能力、多模态信息融合以及数据驱动和知识驱动融合等.

参考文献:

[1] LIU K. A Survey on Neural Relation Extraction [J]. Science China Technological Sciences, 2020, 63(10): 1971-1989.

[2] 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. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014: 2335-2344.

[3] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.

[4] RAU L F. Extracting Company Names from Text [C]// The Seventh IEEE Conference on Artificial Intelligence Application. Miami Beach, FL, USA: IEEE, 1991: 29-32.

[5] LI J Q, ZHAO S H, YANG J J, et al. WCP-RNN: a Novel RNN-Based Approach for Bio-NER in Chinese EMRs [J]. The Journal of Supercomputing, 2020, 76(3): 1450-1467.

[6] HUANG Z H, XU W, YU K . Bidirectional LSTM-CRF Models for Sequence Tagging [J]. CoRR, 2015, abs/1508.01991: 1-10.

[7] MA X Z, HOVY E. End-to-End Sequence Labeling via Bi-Directional LSTM-CNNS-CRF [C]//Proceedings of the 54th

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 1064-1074.

[8] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural Architectures for Named Entity Recognition [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA: Association for Computational Linguistics, 2016: 260-270.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need [C]// 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA: Curran Associates, Inc., 2017: 5998-6008.

[10] REI M, CRICHTON G K O, PYYSALO S. Attending to Characters in Neural Sequence Labeling Models [C]//Proceedings of {COLING} 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING 2016 Organizing Committee, 2016: 309-318.

[11] ZHENG K H, SUN L Y, WANG X, et al. Named Entity Recognition in Electric Power Metering Domain Based on Attention Mechanism [J]. IEEE Access, 2021, 9: 152564-152573.

[12] LI X Y, ZHANG H, ZHOU X H. Chinese Clinical Named Entity Recognition with Variant Neural Structures Based on BERT Methods [J]. Journal of Biomedical Informatics, 2020, 107: 103422.

[13] ALSAARAN N, ALRABIAH M. Classical Arabic Named Entity Recognition Using Variant Deep Neural Network Architectures and BERT [J]. IEEE Access, 2021, 9: 91537-91547.

[14] **应坚超，蒲飞，徐晨鸥，等. 基于互逆和对称关系补全的知识图谱数据扩展方法** [J]. **西南大学学报(自然科学版)**, 2020, 42(11): 43-51.

[15] **王红，卢林燕，王童. 航空安全事件知识图谱补全方法** [J]. **西南大学学报(自然科学版)**, 2020, 42(11): 31-42.

[16] GETOOR L, TASKAR B. Global Inference for Entity and Relation Identification Via a Linear Programming Formulation [J]. Introduction to Statistical Relational Learning, 2007: 553-580.

[17] MIWA M, SASAKI Y. Modeling Joint Entity and Relation Extraction with Table Representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1858-1869.

[18] GUPTA P, SCHUTZE H, ANDRASSY B. Table Filling Multi-Task Recurrent Neural Network for Joint Entity and Relation Extraction [C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING 2016 Organizing Committee, 2016: 2537-2547.

[19] BEKOULIS G, DELEU J, DEMEESTER T, et al. Joint Entity Recognition and Relation Extraction as a Multi-Head Selection Problem [J]. Expert Systems With Applications, 2018, 114: 34-45.

[20] ZENG X R, ZENG D J, HE S Z, 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). Melbourne, Australia: Association for Computational Linguistics, 2018: 506-514.

[21] ZENG X R, HE S Z, ZENG D J, et al. Learning the Extraction Order of Multiple Relational Facts in a Sentence with Reinforcement Learning [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: 367-377.

[22] ZENG D J, ZHANG H R, LIU Q Y. CopyMTL: Copy Mechanism for Joint Extraction of Entities and Relations with Multi-Task Learning [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 9507-9514.

[23] HAN X, ZHU H, YU P F, et al. FewRel: a Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 4803-4809.

[24] GAO T Y, HAN X, ZHU H, et al. FewRel 2.0: Towards more Challenging Few-Shot Relation Classification [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 Lin-

guistics, 2019: 6250-6255.

[25] CHRISTOPOULOU F, MIWA M, ANANIADOU S. Connecting the Dots: Document-Level Neural Relation Extraction with Edge-Oriented Graphs [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: 4925-4936.

[26] **陈珂, 陈振彬. 基于最短依存路径和BERT的关系抽取算法研究** [J]. **西南师范大学学报(自然科学版)**, 2021, 46(11): 56-66.

[27] ZHOU W X, HUANG K, MA T Y, et al. Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California, USA: AAAI Press, 2021: 14612-14620.

[28] RILOFF E. Automatically Constructing a Dictionary for Information Extraction Tasks [C]//Proceedings of the Eleventh National Conference on Artificial Intelligence. Washington D. C., USA: AAAI Press/MIT Press, 1993: 811-816.

[29] NGUYEN T H, GRISHMAN R. Event Detection and Domain Adaptation 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 2: Short Papers). Beijing, China: Association for Computational Linguistics, 2015: 365-371.

[30] CHEN Y B, XU L H, 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.

[31] 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 Thirty-Second Conference on Artificial Intelligence. New Orleans, Louisiana, USA: AAAI Press, 2018: 5916-5923.

[32] LAI V D, NGUYEN T N, NGUYEN T H. Event Detection: Gate Diversity and Syntactic Importance Scores for Graph Convolution Neural Networks [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 5405-5411.

[33] AHMAD W U, PENG N Y, CHANG K W. GATE: Graph Attention Transformer Encoder for Cross-Lingual Relation and Event Extraction [EB/OL]. (2020-10-6)[2022-2-15]. https://www.researchgate.net/publication/344529795__GATE_Graph_Attention_Transformer_Encoder_for_Cross-lingual_Relation_and_Event_Extraction.

[34] WADDEN D, WENNBERG U, LUAN Y, et al. Entity, Relation, and Event Extraction with Contextualized Span Representations [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: 5784-5789.

[35] LIU P F, YUAN W Z, FU J L, et al. Pre-Train, Prompt, and Predict: a Systematic Survey of Prompting Methods in Natural Language Processing [EB/OL]. (2021-7-28)[2022-2-15]. https://arxiv.org/abs/2107.13586.

[36] SI J H, PENG X T, LI C, et al. Generating Disentangled Arguments with Prompts: a Simple Event Extraction Framework that Works [EB/OL]. (2021-10-9)[2022-2-15]. https://arxiv.org/abs/2110.04525.

[37] ZHANG Z S, KONG X, LIU Z Z, et al. A Two-Step Approach for Implicit Event Argument Detection [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 7479-7485.

[38] LI S, JI H, HAN J W. Document-Level Event Argument Extraction by Conditional Generation [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: 894-908.

[39] LEE K, HE L H, LEWIS M, et al. End-to-End Neural Coreference Resolution [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguis-

tics，2017：188-197.

[40] BUGERT M，REIMERS N，GUREVYCH I. Generalizing Cross-Document Event Coreference Resolution across Multiple Corpora [J]. Computational Linguistics，2021，47(3)：575-614.

[41] HUANG Y J，LU J，KUROHASHI S，et al. Improving Event Coreference Resolution by Learning Argument Compatibility from Unlabeled Data [C]//Proceedings of the 2019 Conference of the North. Minneapolis，Minnesota：Association for Computational Linguistics，2019：785-795.

[42] DASGUPTA T，SAHA R，DEY L，et al. Automatic Extraction of Causal Relations from Text Using Linguistically Informed Deep Neural Networks [C]//Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue. Melbourne，Australia：Association for Computational Linguistics，2018：306-316.

[43] PHU M T，NGUYEN T H. Graph Convolutional Networks for Event Causality Identification with Rich Document-Level Structures [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，2021：3480-3490.

[44] LIU J，CHEN Y B，ZHAO J. Knowledge Enhanced Event Causality Identification with Mention Masking Generalizations [C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama，Japan：International Joint Conferences on Artificial Intelligence Organization，2020：3608-3614.

[45] CAO P F，ZUO X Y，CHEN Y B，et al. Knowledge-Enriched Event Causality Identification via Latent Structure Induction Networks [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1：Long Papers). Online：Association for Computational Linguistics，2021：4862-4872.

[46] PUSTEJOVSKY J，CASTANO J M，INGRIA R，et al. TimeML：Robust Specification of Event and Temporal Expressions in Text [M]//New Directions in Question Answering. Stanford，USA：AAAI Press，2003：28-34.

[47] LONG Y，LI Z J，WANG X，et al. XJNLP at SemEval-2017 Task 12：Clinical Temporal Information Ex-Traction with a Hybrid Model [C]//Proceedings of the 11th International Workshop on Semantic Evaluation (T-2017). Vancouver，Canada：Association for Computational Linguistics，2017：1014-1018.

[48] HAN R J，NING Q，PENG N Y. Joint Event and Temporal Relation Extraction with Shared Representations and Structured Prediction [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，2019：434-444.

责任编辑　崔玉洁