基于深度学习的实体关系联合抽取研究综述

张仰森[1],刘帅康[1],刘 洋[2],任 乐[1],辛永辉[2]
(1. 北京信息科技大学智能信息处理研究所,北京 100192;2. 国家计算机网络应急技术处理协调中心,北京 100029)

摘 要: 实体关系抽取是信息抽取领域的核心任务. 从文本中抽取的实体关系三元组是构建大规模知识图谱的基础. 传统的流水线方法将实体关系抽取分解为独立的命名实体识别和关系抽取两个子任务. 首先,构建一个高效的命名实体识别器,从大规模非结构化文本语句中识别实体边界和类型. 然后,将该命名实体识别器识别的实体与类型作为关系抽取任务中所用数据的标注. 最后,通过关系抽取器得到两个实体之间的关系类别,进而组合成为结构化的实体关系三元组. 命名实体识别任务存在的误差会影响后续的关系抽取任务的性能,这使得流水线方法具有错误累积问题. 这是因为关系抽取任务中使用的标注数据来自于前面的命名实体识别任务,这会有一定的误差,进而影响关系抽取的结果质量. 此外,流水线方法减弱了两个子任务之间的特征关联,这会出现冗余实体的问题. 命名实体识别任务和关系抽取任务独立进行学习训练,导致这两个子任务间缺乏交互,使得文本信息没有得到充分利用,限制了流水线方法的性能瓶颈. 由于非结构化文本信息没有得到充分利用,流水线方法在抽取实体间长依赖关系时具有一定局限性,很难达到联合抽取模型的性能指标. 实际应用中,实体间往往存在多种关系,流水线方法无法充分使用全局文本信息,且命名实体识别会产生冗余实体,在抽取多元重叠关系时,该方法具有一定的局限性. 因此,在构建高准确率实体关系抽取模型时,流水线方法具有欠缺之处. 本文对实体关系联合抽取的研究发展全景进行了综述,简要阐明整数线性规划、卡片金字塔解析模型、概率图模型和结构化预测模型这四类基于特征工程的联合模型的共同缺点. 本文聚焦基于深度学习的实体关系联合抽取技术,根据近年来实体关系联合抽取前沿研究成果,总结了实体关系联合抽取模型的主流构建方法. 按照建模思想的特点总结为三种建模方法:多模块-多步骤、多模块-单步骤以及单模块-单步骤. 多模块-多步骤建模方法主要包含实体域映射关系域、关系域映射实体域和头实体域映射关系-尾实体域这三种类别. 这三类模型的共同特点都是将三元组的提取过程分为多个模块,通过共享参数的方式整合各个模块,逐步迭代得到三元组. 这种方法推动联合模型性能提升,初步解决了流水线方法存在的问题. 但每个步骤使用独立的解码算法,导致解码误差累积问题. 且共享参数整合各个模块的冗余误差会互相影响预测性能,从而产生级联冗余问题. 多模块-单步骤建模方法旨在构建一个最优化的联合解码算法,并对其求取最优解进而得到最优超参数. 这种方法设计了简单精确的联合解码算法,并加强了多个子模块间的交互性,减弱了因为逐步迭代导致的解码误差和级联冗余对联合模型性能的影响. 然而,模块的分离依然会产生冗余错误,具有一定局限性. 单模块-单步骤建模方法可以直接从文本语句中抽取三元组,有效缓解了多模块-多步骤和多模块-单步骤建模方法的级联错误和实体冗余等问题. 本文以前沿文献中具有代表性的联合模型为例,详细分析了这些模型的建模思路,剖析了各个模型的优缺点,将多个具有共同建模思路的经典模型进行归类,以阐述实体关系联合抽取模型的发展趋势. 本文将单模块-单步骤建模方法的代表模型在公开基准数据集上的模型性能与多模块-多步骤和多模块-单步骤的代表模型性能进行对比分析,阐明实体关系联合抽取模型的建模思路正在从基于多模块-多步骤和多模块-单步骤的复杂建模方法,逐渐向单模块-单步骤的高效建模方法转变的客观趋势. 最后,本文对三个实体关系联合抽取的研究方向进行了展望. 当下主流的联合模型聚焦于限定域的实体关系抽取任务,对于开放域问题研究得不够. 开放域实体关系联合抽取任务是未来的研究人员亟待解决的问题之一. 在实际工业应用中,文本语料包含多元信息,如时序信息. 而当前的实体关系联合抽取模型大多依据单一文本上下文信息进行特征抽取,从而忽略了时序信息. 若融入像时序信息这样的多元信息或能进一步提升联合模型性能,这是未来一项具有重大意义的课题. 此外,对于跨文本的实体关系联合抽取模型研究较少,这也是该领域未来的一个研究趋势. 本文旨在建立一个完整的基于深度学习的实体关系联合抽取领域研究视图,以对相关领域研究者有所帮助.

关键词: 信息抽取;知识图谱;深度学习;实体关系联合抽取;流水线方法
基金项目: 国家自然科学基金(No.62176023)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2023)04-1093-24
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221176

收稿日期:2022-10-17;修回日期:2023-03-20;责任编辑:朱梅玉

Joint Extraction of Entities and Relations Based on Deep Learning: A Survey

ZHANG Yang-sen[1], LIU Shuai-kang[1], LIU Yang[2], REN Le[1], XIN Yong-hui[2]

(1. *Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing* 100192, *China*;
2. *Computer Network Emergency Response Technical Team, Coordination Center of China, Beijing* 100029, *China*)

Abstract: Entity-relation extraction is a core task in the field of information extraction. Entity-relation triples extracted from text are the basis for building large-scale knowledge graphs. The traditional pipeline method decomposes entity-relation extraction into two subtasks: named entity recognition and relation extraction. First, an efficient named entity recognizer is built to identify the entity boundaries and types from large-scale unstructured text sentences. Then, the entities and types are used as labels for the data used in the relation extraction task. Finally, the relationship category between two entities is obtained through the relationship extractor and then combined into a structured entity-relation triplet. However, error in the named entity recognition task will affect the performance of the subsequent relation extraction task, which makes the pipeline method problematic because of error accumulation. This is because the labeled data used in the relation extraction task come from the previous named entity recognition task, which will include certain errors, and this will affect the quality of the relation extraction results. In addition, the pipeline method weakens the feature association between the two subtasks, which will lead to redundant entities. The named entity recognition task and relationship extraction task are independently learned and trained, which leads to a lack of interaction between these two subtasks. As a result, the text information is not fully utilized, which becomes the main reason the performance of the pipeline method is limited. Because unstructured text information is not fully employed, the pipeline method has certain limitations in extracting long dependencies between entities, and it is difficult to achieve high performance in the joint extraction model. In practical applications, there are often multiple relationships between entities, but the pipeline method cannot fully consider the global text information, and hence named entity recognition produces redundant entities, which has disadvantages when extracting multiple overlapping relationships. Therefore, when constructing a high-accuracy entity-relation extraction model, the pipeline approach has shortcomings. This paper reviews the research and development of the joint extraction of entity relationships. Furthermore, it briefly clarifies the common shortcomings of four types of joint models based on feature engineering: integer linear programming, card pyramid analysis models, probabilistic graph models, and structured prediction models. Focusing on the joint extraction techniques for entity relationships based on deep learning, the mainstream construction methods of these models are summarized according to the state-of-the-art results reported in recent years. According to the characteristics of the modeling idea, the modeling methods are categorized into three types: multi-module/multi-step, multi-module/single-step, and single-module/single-step models. Multi-module/multi-step modeling methods consist of three main types: entity domain mapping to the relationship domain, relationship domain mapping to the entity domain, and head-entity domain mapping to the relation-tail domain. The common feature of these three types of models is that they divide the extraction of triples into multiple modules, integrate each module by sharing the parameters, and gradually iterate to obtain triples. This approach improves the performance of the joint model and initially solves the problems of the pipeline method. However, because each step uses an independent decoding algorithm, it leads to the accumulation of decoding errors. Moreover, because the redundant errors of each module integrated with shared parameters affect the prediction performance of the others, this results in cascading redundancies. The multi-module-single-step modeling method aims to construct an optimal joint decoding algorithm and obtain the optimal solution to determine the optimal hyperparameters. This method designs a simple and accurate joint decoding algorithm and strengthens the interaction between multiple submodules. Therefore, the impact of decoding errors and cascading redundancies caused by gradual iterations on the performance of the joint model is weakened. However, the separation of the modules still produces redundancy errors, which cause certain limitations. The single-module/single-step modeling method can extract triples from text directly, which effectively alleviates the cascading error and entity redundancy problems of multi-module/multi-step and multi-module/single-step modeling methods. Taking the representative joint models in the high-impact literature as examples, this paper analyzes the modeling idea, advantages, and disadvantages of each model. It also classifies a number of classical models according to common modeling ideas to illustrate trends in the development of entity-relationship joint extraction models. This paper compares and analyzes the performance of the representative single-module, single-step modeling method with multi-module/multi-step and multi-module/

single-step models on a public benchmark data set. Moreover, it clarifies the objective trend that the modeling idea of joint extraction models is gradually changing from complex methods based on multi-module/multi-step and multi-module/single-step models to efficient single-module/single-step models Finally, this paper discusses the prospects of research directions in the joint extraction of three-entity relationships. The current mainstream joint model focuses on the entity-relationship extraction task of limited domains, and the open-domain entity-relationship joint extraction task is an urgent problem for future researchers to solve. In practical industrial applications, a text corpus contains multiple types of information, such as timing information. However, most current entity-relationship joint extraction models extract features based on single-text context information, thus ignoring time-series information. If multivariate information such as time-series information could be incorporated, the performance of the joint model would be further improved, and this is a topic of high importance for the future. In addition, there is little research on cross-text entity-relationship joint extraction models, which is also a future research topic in this field. This paper aims to establish a complete deep learning-based view of entity-relationship joint extraction research, which will be helpful to researchers in related fields.

Key words: information extraction; knowledge graph; deep learning; joint extraction of entities and relations; pipeline method

Foundation Item(s): National Natural Science Foundation of China (No.62176023)

1 引言

随着大数据时代的到来,建立可快速高效地从大量开放域、非结构化数据中抽取有效信息的模型,成为当前自然语言处理(Natural Language Processing,NLP)领域的一个重要问题. 作为信息抽取[1]的核心任务,实体关系抽取旨在通过对文本语句进行建模,以快速高效地抽取其中蕴含的实体对及其语义关系,进而获取句子中的结构化三元组信息<实体1,关系,实体2>. 获取的三元组信息在大规模知识图谱的构建[2]、机器阅读、文本摘要、问答系统[3]、机器翻译[4]、语义网标注等下游自然语言处理任务中具有奠基性意义. 近年来,随着信息抽取相关研究的兴起和深度学习的迅速发展,实体关系抽取问题的研究不断深入,产生了大量的优秀研究成果.

早期实体关系抽取被看作两个子任务,即命名实体识别(Named Entity Recognition,NER)[5,6]和关系抽取(Relation Extraction,RE)[7,8]. 针对这两个任务,研究人员最初以流水线(pipelined)方法对实体关系抽取进行研究,即首先利用人工特征提取和核函数的方法构建实体识别模型[9,10],然后在实体对的基础上构建能识别其语义关系的模型[11~13]以实现实体间关系的抽取. 随着近些年深度学习技术的迅速发展,一些端到端的深度学习模型在该领域逐渐占据了主导地位,基于深度学习的NER相关研究取得了丰硕的研究成果[14~20],而在RE领域,深度学习模型也取得了优异的研究成果[21~29],并在几个公开基准数据集上显示了它们的有效性.

流水线方法将实体关系抽取看作实体识别和关系抽取两个独立任务,虽然建模比较灵活,但也存在着错误累积[30],缺少子任务间的信息交互,信息冗余,实体间长依赖关系难以抽取等问题. 和流水线方法不同,实体关系联合抽取模型旨在利用单一模型抽取实体和关系. 它可以有效地整合实体识别和关系抽取两个子任务间的隐性关联特征,以期解决流水线方法存在的问题. 同时,联合模型可以更好地抽取实体间跨句、跨段和跨语义的层级性关联特征,使得该模型在抽取实体间长依赖关系和多元重叠关系时具有优越性.

随着研究者对实体关系联合抽取的深入研究,诸多优秀模型相继被提出. 分析联合模型发展路径,总结多种模型的特点,本文将基于深度学习的实体关系联合抽取分为三种建模方法,即在早期深度神经网络探索阶段提出的基于文本全局信息[30~32]的多模块-多步骤方法和多模块-单步骤方法[33,34],以及为解决模块之间、步骤之间存在级联冗余问题而提出的单模块-单步骤方法. 多模块是指将一个联合模型分为多个模块,以共享参数的方式整合各个模块,并从文本中抽取三元组. 按照是否是一次性将三元组预测成功,可将其分为多步骤和单步骤两种方法. 多步骤抽取三元组的方法具有级联冗余错误,为解决这一问题,研究者采用联合解码的方式一步抽取三元组,提升了联合模型性能. 单模块-单步骤是在文本上做细粒度的三重分类,直接一步预测语料库中的全部三元组,该方法减弱了模块分离和分步预测产生的级联冗余,可以取得比前两种建模方法更好的效果. 但单模块-单步骤方法诞生较晚,目前成熟的模型很少,是未来联合模型建模的新思路.

近年来,有多篇关于实体关系抽取方面的综述文献,但由于技术快速发展以及综述者的视角不同等因素,都存在着一些这样或那样的不足. 文献[35]对基于深度学习的实体关系抽取技术做了详尽的论述,但其着重介绍了基于流水线方法的实体关系抽取技术和远

程监督实体关系抽取技术，未包含近年实体关系抽取的前沿研究，对实体关系联合抽取技术介绍较少；文献[36]详尽介绍了重叠实体关系抽取技术，但大部分篇幅侧重于流水线方法如何解决重叠关系，未对实体关系联合抽取进行较为全面的论述；文献[37~39]综述了近些年实体关系抽取的发展全景，涵盖了最新的实体关系抽取方法，但对实体关系联合抽取方法介绍得较为简单，没有对联合抽取模型的发展形成一个系统性的认识；文献[40]虽然对实体关系联合抽取进行了系统性的论述，但其侧重于介绍基于特征工程的联合抽取方法，对基于深度学习的联合抽取方法论述得不全面. 本文则着重对基于深度学习的实体关系联合抽取研究进行系统性论述；对存在的问题进行分析，以建模思想的特点对模型进行分类说明并总结其优缺点；详尽梳理其中的关键问题和解决方法，整理评价指标以及本领域的发展情况与趋势，并对本领域的未来研究方向进行展望. 整体框架如图1所示.

具体而言，本文的贡献有以下几点：

(1)根据模型特点将联合模型建模方法分类为三种，阐明从基于多模块-多步骤和多模块-单步骤的复杂建模方法，逐渐向单模块-单步骤的简单建模方法转变的客观趋势；

(2)详细整理了实体关系联合抽取常用的公开数据集和评价指标，在各个数据集上总结了各个方法的性能差异并进行分析；

(3)基于前沿研究进展，总结三类建模方法下各个模型的优缺点，针对流水线方法存在的问题指明联合模型的优势和未来研究方向.

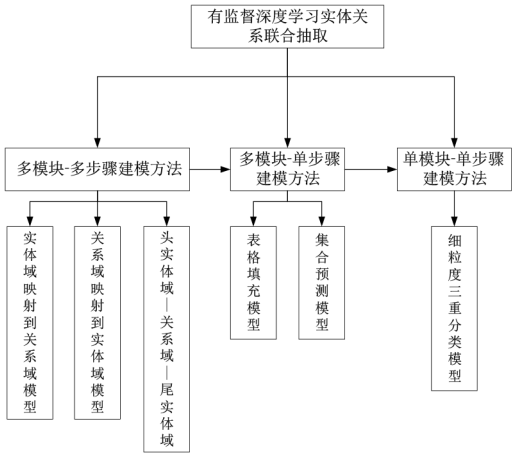


图1 有监督深度学习实体关系联合抽取方法分类

2 命名实体识别及关系抽取的基本概念及其实现方法

命名实体识别即识别文本中具有特定意义的单词或单词组，如人名、地名、组织名等. 其数学描述为：给定实体类别集合E，设给定句子$S=\{w_1,w_2,\cdots,w_n\}$，命名实体识别的目的是从句子S中识别所有的实体及实体类型$<w_i,w_j,e_k>$，其中$e_k\in E$，w_i和w_j分别是该实体的起始单词和结束单词.

关系抽取即抽取文本中实体之间的语义关系. 其数学描述为：已知关系类别集合R、实体类别集合E，对于给定的实体对$<h,e_1,t,e_2>$，其中，$e_1\in E$和$e_2\in E$分别表示h和t的实体类别，关系抽取输出h和t的关系类别$r\in R$.

实体关系联合抽取即将实体识别和关系抽取联合完成，直接从文本中获取实体关系三元组. 其数学描述为：已知关系类别集合R、实体类别集合E，给定句子$S=\{w_1,w_2,\cdots,w_n\}$，实体关系联合抽取利用建立的统一模型，抽取出句子S中的所有实体关系五元组$<h,e_1,r,t,e_2>$，其中，$r\in R$，$e_1\in E$，$e_2\in E$. 在做关系抽取时和流水线方法的显著性区别是没有预先标注给定的实体边界和类型，联合模型输出句子S的所有关系三元组$<h,r,t>$.

利用实体识别和关系抽取技术从非结构化文本中抽取<实体1，关系，实体2>结构化三元组是信息抽取领域研究的核心问题. 它为构建自然语言处理领域的下游任务提供了事实知识库的基础[41]，主要包含两种实现方法：流水线方法和联合抽取方法.

2.1 实体识别和关系抽取的流水线方法

流水线方法包含命名实体识别子任务和关系抽取子任务. 命名实体是指能够从某一元素集合中识别的具有相似属性元素的单词或单词组. 这一概念最早在1995年的MUC-6会议上被提出. 关系抽取子任务则最早在1998年的MUC-7会议上被提出，当时主要通过填充关系模板槽完成实体间三类关系的抽取.

基于流水线方法的实体识别和关系抽取存在以下问题.

(1)错误累积[30]. 流水线方法先构建命名实体识别器，从非结构化文本语句中识别所有已知种类的实体类别和边界. 识别得到的实体作为关系抽取子任务的标注特征共同参与构建关系抽取器. 识别实体的过程会产生误差，这些误差会影响关系抽取的结果质量.

(2)缺少子任务间的信息交互[34]. 流水线方法忽略了两个子任务之间隐性特征关联，导致两个子任务之间缺乏交互，使得文本信息没有得到充分利用，限制了流水线方法的性能瓶颈. 如实体类别和关系类别之间具有某种隐性特征关联. 在识别实体类型时，关系类别会影响实体类型的识别率；在对关系分类时，实体类别同样会影响关系分类的结果.

(3)命名实体识别具有冗余实体. 区别于错误累积问题中的命名实体识别任务在实体类别和边界存在误

差. 冗余实体是指从非结构化文本中识别的没有确定关系的实体. 冗余实体参与关系抽取过程会增加关系分类的错误率.

(4)实体间长依赖关系难以识别. 实体间长依赖关系包含三种:跨句、跨段和跨语义的层级性依赖关系. 这三种长依赖关系的识别需要文本全局信息特征抽取、流水线方法忽略子任务间特征关联关系,导致文本信息没有得到充分利用. 因此,流水线方法在抽取实体间长依赖关系时具有一定局限性,很难达到联合抽取模型的性能指标.

(5)多元重叠关系问题. 在实际应用中,实体间往往存在多种关系,如何准确识别这些多元重叠关系是实体关系抽取任务的一个难点. 流水线方法无法充分使用全局文本信息,且命名实体识别会产生冗余实体. 这些问题都表明,针对多元重叠关系问题,流水线方法具有局限性.

2.2　实体关系的联合抽取方法

为解决流水线方法的问题,提高实体识别和关系抽取两个子任务的交互性,研究人员构建了实体关系联合抽取模型. 这类模型包括基于特征工程的方法和基于深度学习的方法. 基于特征工程的方法主要有四类模型:整数线性规划模型、卡片金字塔解析模型、概率图模型和结构化预测模型[30,40~44]. 这四种模型都需要依赖大量人工提取的特征规则,具有高成本、低效率的缺点. 基于深度学习的联合模型解决了基于特征工程方法的问题,能更好地解决更具一般性的多重关系抽取任务,并不断在公开数据集上突破性能上限,是实体关系联合抽取方法当前和未来的研究热点. 基于深度学习的联合模型性能已经全面超越了基于特征工程的联合模型. 下文将着重对基于深度学习的实体关系联合抽取模型进行深入分析和综述.

3　基于深度学习的实体关系联合抽取

随着深度学习技术的发展,基于深度神经网络的端到端模型在实体关系联合抽取领域取得了丰硕成果[32,44~49]. 本节将对多模块-多步骤、多模块-单步骤和单模块-单步骤方法的建模思想以及它们的代表性模型进行详细分析,并对它们的优缺点进行综述评价.

3.1　多模块-多步骤建模方法

随着研究者对实体关系抽取深入地研究表明,联合模型在理论上优于流水线模型,但其难点在于如何加强实体识别模型和关系抽取模型之间的交互. 相关研究表明联合实体和关系抽取模型能够提取实体与关系的隐性特征关联,并有助于提升实体关系抽取模型的性能[30,31]. 该领域的研究重点在于如何构建能提高命名实体识别任务和关系抽取任务交互性的联合模型.

基于多模块-多步骤的建模方法,利用不同模块和相互关联的处理步骤,连续提取实体和关系. 其又可以分为三种形式. 第一种是先抽取出文本中全部的实体,然后对每个实体对做关系分类,最终得到三元组[32,44,50~56],这种模型被称为实体域映射到关系域模型. 第二种是先从文本语句中预测关系,然后基于这种关系去抽取头部实体和尾部实体[46,57~59],这种模型被称为关系域映射到实体域模型. 最后一种则先抽取头部实体,然后推断出对应的关系和尾部实体[49,60~62],这种模型被称为头实体域映射到关系、尾实体域模型. 多模块-多步骤建模方法采用共享参数并通过多个模块联合预测三元组,提高了命名实体识别和关系抽取两个子任务的交互性.

3.1.1　实体域映射关系域模型

在预训练模型还未被提出之前,一些基于循环神经网络(Recurrent Neural Network,RNN)及其变种网络的联合模型被相继提出[50~53]. 这些模型虽然在公开基准数据集上并未产生太大的性能提升,但为探索构建实体识别与关系抽取的联合模型方法做出了巨大贡献.

早期最具代表性的实体域映射关系域的联合模型是Miwa等[32]在2016年提出的基于端到端的树形结构实体关系联合抽取模型,如图2所示. 其中,左侧为实体识别模块,右侧为关系抽取模块. 实体识别模块中包括词嵌入层(embedding layer)、上下文语义特征提取层(Bi-LSTM layer)和实体边界与类别识别层(softmax layer). 关系抽取模块中包括Bi-TreeLSTM层和softmax层,Bi-TreeLSTM层共享实体识别模块的参数,并利用依存句法树(由外部NLP工具生成)的结构特征抽取实体间的关系特征,softmax层抽取出最终的实体关系. 该模型的优点是通过不同模块之间的参数共享,解决了流水线方法存在的错误累积问题以及基于特征的实体关系联合抽取模型的低效问题;其缺点是实体间关系抽取依赖于依存句法树的特征提取,这可能使得依存句法树的生成误差传递到关系抽取模型,另外,该模型的实体识别模块和关系抽取模块仍有先后次序的问题,没有做到真正同步进行实体关系的联合抽取.

这里要强调的是,多模块-多步骤建模方法虽然也是先识别实体,再抽取实体间的关系,但这两个模块之间通过共享参数实现了实体识别模块与关系抽取模块之间的交互,而前述提到的流水线建模方法分别对实体识别和关系抽取进行建模,二者之间没有交互.

针对文献[32]中依赖于依存句法树特征提取的联合模型的缺点,Katiyar等[54]提出了一种基于注意力机制的循环神经网络联合模型,实现实体与关系的联合抽取. 该模型进一步提升了ACE数据集上的性能.

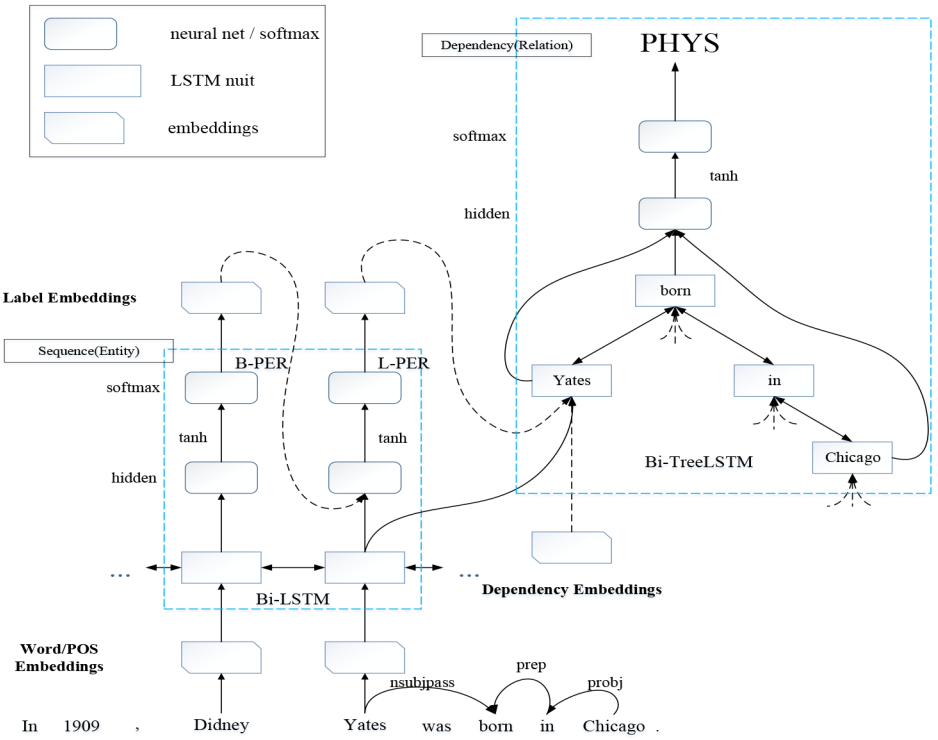


图2 端到端的树形实体关系联合抽取模型结构图[32]

由于实体间关系具有一定的复杂性,同一实体对之间可能具有多重关系,因此如何构建具有多重关系的实体关系联合模型就成为下一步要解决的难点. 2018年Giannis等[63]将实体关系联合抽取看作一个多头选择问题(multi-head selection),针对实体间的多重关系问题进行探讨,其构建的实体关系联合抽取模型在相关公开数据集上的性能表现超出前人研究成果,模型结构如图3所示. 该模型主要由五层构成:嵌入层(embedding layer)、上下文语义特征提取层(Bi-LSTM layer)、实体识别层(CRF layer)、标签嵌入层(label embedding layer)、关系抽取层(sigmoid scoring layer). 其中,嵌入层创新性地将词向量和字符向量通过图4所示的向量生成模型构建新的词向量,为图3所示模型的后续其他各层输入向量数据.

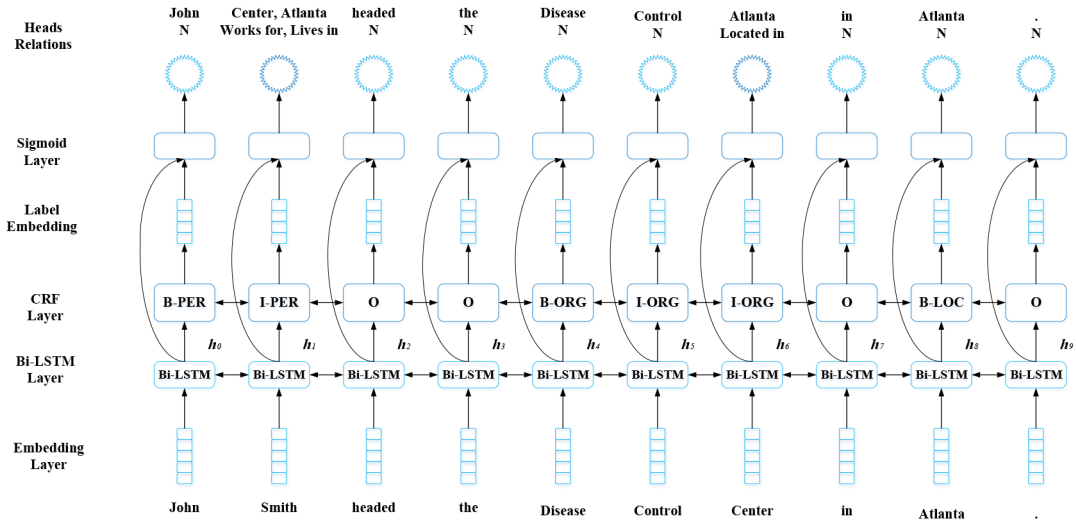


图3 多头选择模型结构图[63]

图4中,字符向量生成分为两个方向. 首先,从左向右利用LSTM网络针对每个字母生成表征该单词的右字符向量 $\boldsymbol{W}_{\mathrm{R\_chars}}$,然后从右向左做同样操作得到左字符向量 $\boldsymbol{W}_{\mathrm{L\_chars}}$,最后将这两个向量和词向量 $\boldsymbol{W}_{\mathrm{word2vec}}$

进行拼接得到最终表征该单词的输入特征向量 $\boldsymbol{W}_{\text{the}}$. 图3中,上下文语义特征提取层抽取输入文本的语义特征;实体识别层根据语义特征识别实体边界及类别;标签嵌入层生成预测的实体标签向量;关系抽取层标签向量;关系抽取层共享上下文语义特征提取层的参数,结合标签嵌入层的实体标签特征向量抽取实体间关系. 其中,关系抽取层对实体间每种关系做二分类预测,可以识别多重关系. 在预测某一实体与其他实体间的多重关系时,将每个实体都作为头实体,轮流预测与其他所有实体间的关系. 由于整个预测实体域都可作为头实体,因此这被看作一种多头选择问题. 需要强调的是,前人的研究大多将实体间的关系抽取当作一种多分类问题,这很难抽取实体间的多重关系. 该模型的优点是利用二分类和多头选择的思想初步解决了实体间多重关系的问题;缺点是在公开数据集上多重关系性能不高,具有实体冗余的问题. 由实体识别模块预测的实体域中,某些实体间并不存在关系,但关系抽取模型仍然需要抽取这类实体间的关系,会影响关系抽取的性能.

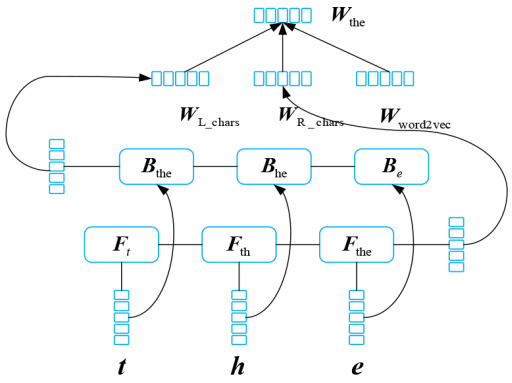


图4　基于字符向量和词向量相结合的词向量生成模型[63]

针对上述问题,Tan等[55]提出一种TME模型,该模型通过使用翻译机制对多重关系(也称关系重叠)识别结果进行重新排序,能同时自适应地抽取单句中多个三元组,在多重关系抽取任务上取得了更优的性能. Fu等[44]考虑重叠关系间的相互作用,采用图卷积神经网络(Graph Convolutional Network,GCN)进一步提升了联合模型性能. Liu等[56]为进一步解决关系重叠和实体边界识别等问题,提出一种行之有效的基于注意力机制的联合模型,在公开数据集上取得了优于前人工作的性能. 在面向中文语料的联合抽取研究过程中,研究者通常利用实体间单向关系语义特征,但仅依靠单向语义特征并不能完全利用实体间语义关系,会使得实体关系抽取的有效性受到影响. 针对这一问题,禹克强等[64]提出一种基于双向语义的中文实体关系联合抽取模型. 该模型先利用Roberta预训练模型获取具有上下文信息的文本字向量表征;再通过首尾指针标注识别中文语句中可能存在关系的实体;接着使用主客实体构建正负关系,利用两组全连接神经网络从正负两个角度抽取实体间的候选关系三元组;最后将正负关系下的概率分布序列与实体位置嵌入特征相结合来对候选三元组进行判别,以确定最终的关系三元组.

总体来说,将实体域映射到关系域的联合模型,首先识别语料库中的全部实体,然后预测实体间的关系. 其难点在于:(1)如何更准确地识别实体边界与类别;(2)如何更好地解决实体冗余问题;(3)如何更好地解决多重关系问题. 这类模型将实体识别和关系抽取分为两个独立模块,采用参数共享的方法进行连接以提高实体识别和关系抽取两个子任务的交互性. 在公开数据集上的性能表明,该方法相比流水线方法有一定提升,但多重关系问题仍然较为严峻,需待后续研究者提出更好的解决方案.

3.1.2　关系域映射实体域模型

实体域映射关系域的联合模型虽然理解起来较为直观,但也存在上面提出的一些问题. Miwa等[32]提出的端到端的树形结构联合抽取模型,虽然解决了流水线方法中的错误累积等问题,但该模型对外部NLP工具具有一定的依赖性,对模型的性能带来一定的影响. 2018年Zeng等[46]提出一种CopyRE的联合模型,该模型采用Seq2Seq框架,依次抽取关系、头实体、尾实体,被称为关系域映射到实体域类模型. 该类模型能降低实体边界与类别识别率对整体模型性能的影响,其模型框架如图5所示.

图5中,包括编码器(Encoder)和解码器(Decoder)两部分. 编码器采用Bi-LSTM抽取上下文语义特征向量 $\boldsymbol{o}_t^E$ 和包含文本隐性特征向量 $\boldsymbol{h}_t^E$:$\boldsymbol{o}_t^E, \boldsymbol{h}_t^E = f(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}^E)$,其中,$\boldsymbol{x}_t$ 为 t 时刻的输入,$\boldsymbol{h}_{t-1}^E$ 为 $t-1$ 时刻的隐藏变量. 解码器采用ATT-LSTM对编码器的特征向量解码得到预测关系特征向量 $\boldsymbol{o}_t^D$ 和隐藏特征向量 $\boldsymbol{h}_t^D$:$\boldsymbol{o}_t^D, \boldsymbol{h}_t^D = g(\boldsymbol{u}_t, \boldsymbol{h}_{t-1}^D)$,其中,$\boldsymbol{u}_t$ 为解码器 t 时刻的输入,且 $\boldsymbol{u}_t = [\boldsymbol{v}_t; \boldsymbol{c}_t] \cdot \boldsymbol{W}^u$,这里 $\boldsymbol{c}_t$ 是注意力向量,$\boldsymbol{v}_t$ 为前一步复制下来的实体或关系编码向量. 在做关系分类时,只需要将 $\boldsymbol{o}_t^E$ 直接送入SoftMax即可. 做头实体预测时,需要在当前解码步从 n 个token中选择一个作为实体.

具体来说:$\boldsymbol{q}_i^e = ([\boldsymbol{o}_t^D; \boldsymbol{o}_i^E] \cdot \boldsymbol{w}^e)$ 为每一个token的编码,加入当前解码的输出,然后根据 $\boldsymbol{p}^e = \text{softmax}([\boldsymbol{q}^e; \boldsymbol{q}^{\text{NA}}])$ 从 n 个token中选择概率最大的token作为头实体预测结果. 尾实体预测与头实体预测类似,只需要掩盖上一步预测的头实体,然后采取同样的操作方式即可.

该模型创新地采用先抽取关系,后逐步预测头实

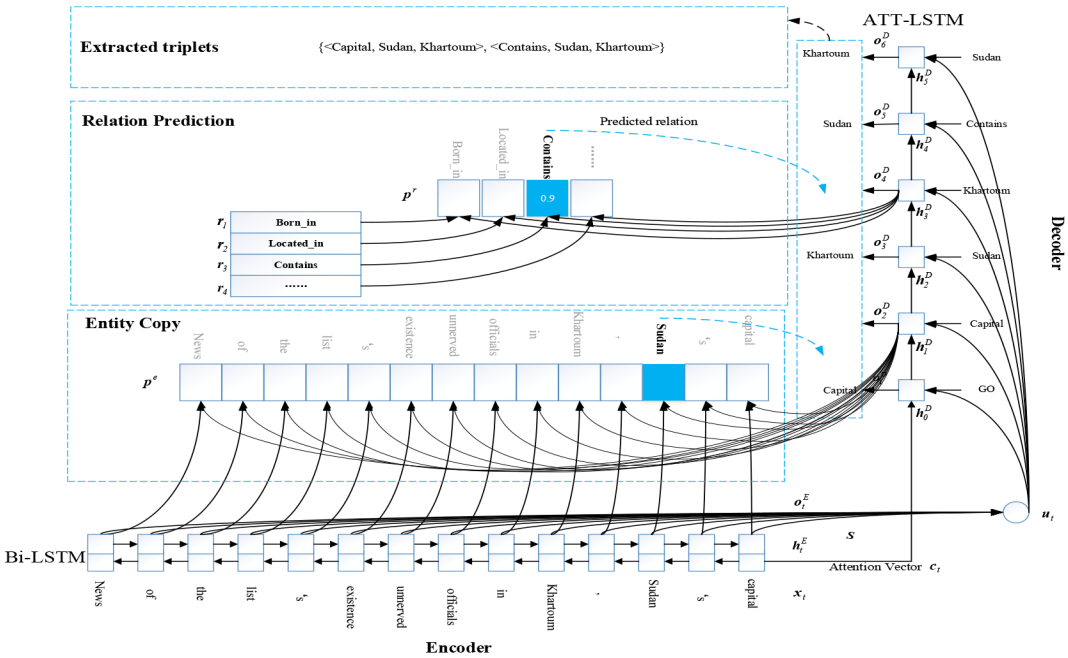


图 5 CopyRE 模型架构[46]

体和尾实体的方式构建一种高效的联合模型. CopyRE 模型采用 Seq2Seq 框架和复制机制对关系重叠问题进行了比较好的解决,在一些公开基准数据集上与前人相比提升了 31.1% 和 39.8%,进一步提高了两个子任务的交互性,开创了实体关系联合抽取建模方面的一种新思路. 但该模型仍然存在一些问题:(1)只考虑了一个 token 组成的实体,没有考虑多个 token 构成的实体,对实体边界的预测尚有欠缺;(2)CopyRE 模型对头、尾实体的区分不大,采用的是统一预测分布的方式,会影响实体间关系的预测性能. 因为,头、尾实体的前后顺序改变会产生相应的语义关系变化. 如:<中国,包含,台湾>中,“中国”和“台湾”分别是头实体和尾实体,其具有“包含”的语义关系. 一旦调换位置,<台湾,属于,中国>,其语义关系就变为“属于”了.

为解决 CopyRE 模型的问题,2020 年 Zeng 等[65]又提出一种 CopyMTL 的模型,它是对 CopyRE 模型的改进. CopyMTL 依然采用 Seq2Seq 框架依次抽取关系、头实体、尾实体. 其框架如图 6 所示.

图 6 中同样包含编码器和解码器两部分. 编码器使用 Bi-LSTM 建模语句上下文信息,融合复制机制生成多对三元组. 针对 CopyRE 只能抽取一个 token 的情况,CopyMTL 做出改善,能成功识别由多个 token 组成的实体. 解码器使用融合注意力机制的 LSTM 建模,采用全连接层获取输出.

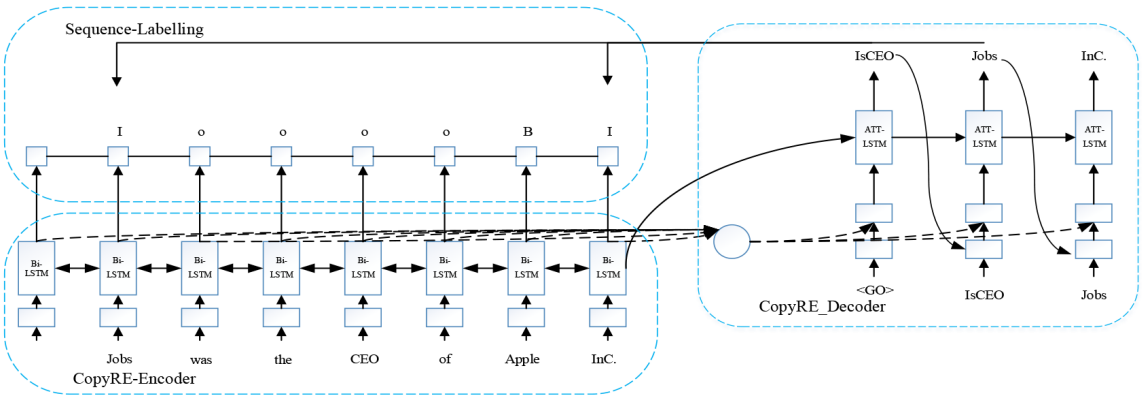


图 6 CopyMTL 模型架构[65]

CopyMTL 模型解决了 CopyRE 模型中实体复制性能不稳定的问题,以及 CopyRE 模型只能预测单个 token 的问题,在公开基准数据集上取得了更优秀的性能,进一步完善了关系域映射到实体域类模型. 尽管 CopyRE

与CopyMTL模型对联合模型性能产生了积极的推动作用,但其只是简单地利用分类结果作为识别实体的指导,忽略了更细粒度的语义关系和语句中单词的联系.为此,2021年Yuan等[57]在分析了关系域映射到实体域模型在公开数据集上效果超越实体域映射到关系域模型的原因后,考虑细粒度下单词对文本语义关系的影响,提出了一种RSAN的联合模型,以提升实体关系联合抽取模型的性能.随后,Zheng等[58]在前人研究基础上提出一种PRGC模型,极大缓解了冗余关系判断、基于广度的提取泛化能力差等问题.Ma等[59]采用基于双译码器模型首先检测文本语义关系,针对关系重叠问题进一步解决,并在公开数据集上取得了更优的性能.

总体来看,率先预测文本的语义关系,然后以此作为基础指导实体的预测,成为一段时间内实体关系联合抽取学界内的共识.从公开数据集上性能的对比来看,考虑全局文本信息的这类模型确实优于实体域映射关系域模型.本质上,该类模型依然是采取一种多模块-多步骤的建模方法,能更全面地使用文本信息.但由于模块分离和步骤分开,虽然使用共享参数的方法将其整合,但依然存在模块间和步骤间级联冗余的问题,并在一般性的应用中表现欠佳.如,如果单句中包含多个关系三元组时,这类模型性能较低.

3.1.3　头实体域映射关系、尾实体域模型

关系域映射到实体域模型从理论和实验上展现了一定的优越性,但在更一般的应用中具有一定局限性.研究者提出一类头实体域映射关系、尾实体域模型,以抽取单句中的多个三元组.

2020年,Wei等[49]创新性地提出了一种HBT(Hierarchical Binary Tagging)模型实现实体关系联合抽取.该模型在WebNLG和NYT数据集上取得了优于前人模型的成果,其模型结构如图7所示.

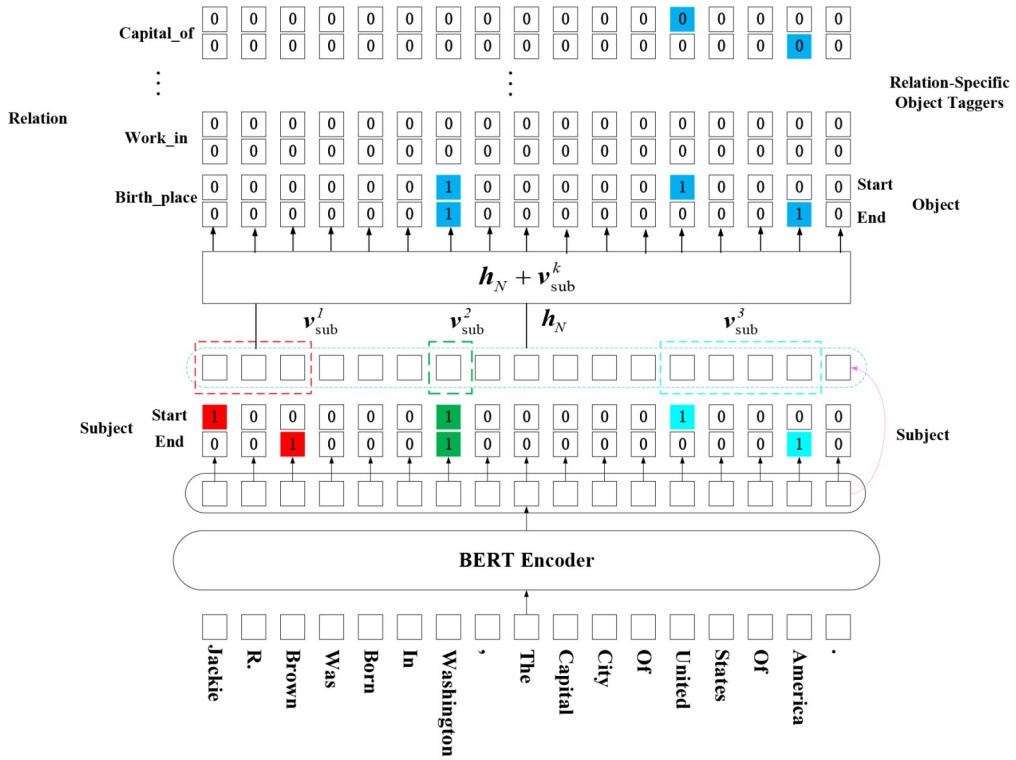


图7　Hierarchical Binary Tagging框架图[49]

图7共包含四层:词向量生成层(BERT encoder layer)对输入的词向量进行编码并抽取语义特征;头实体识别层(subject layer)抽取头实体;关系抽取层(relation layer)对关系域进行遍历,结合尾实体识别层(object layer)抽取关系和尾实体.需要强调的是,在头实体识别层对每个令牌(token)做两次二分类,由实体头(start)和实体尾(end)标记实体边界.在抽取关系时,输入词向量和预测头实体标签向量,对类关系在令牌层面做两次二分类,这与识别头实体类似,目的是抽取关系与尾实体.

实体域映射关系域类模型可用$f(s,o)->r$形式化描述,s是头实体,o是尾实体,r表示关系.这类模型将关系看作要分配给实体对的离散标签,在处理单一实体与多个实体间具有关系的问题时,分类建模较为复杂,是桎梏联合模型性能的一个重要原因.HBT模型将关系抽取看作两个实体间的映射关系寻找过程:(1)抽

取三元组的主实体;(2)针对每个$f_r(\cdot)$,抽取其对应的所有尾实体.其优点是:(1)减弱了实体冗余问题,只有头实体和尾实体之间存在某类关系时,针对该类关系的两个二分类器才会起作用,识别尾实体的边界与类别;(2)增强了实体识别和关系抽取的交互性,在抽取关系和尾实体时,融入了词向量和预测头实体标签向量;(3)在公开数据集,抽取多重关系时性能表现较为优异.

HBT模型虽然解决了更一般的实体关系抽取问题,并采用一种新的头实体域映射到关系、尾实体域模型,提升了联合模型性能.但该模型仍存在不足:(1)在训练时,头实体的选择具有随机性,没有将所有头实体统一进行抽取,这会使模型产生不稳定性;(2)文本信息没有充分利用,造成信息损失,HBT模型需要训练迭代较多次数才能充分利用文本全局信息,以达到更优的性能;(3)在实体类别不均衡的数据集中,该模型表现欠佳.

对于上述问题,Yu等[60]从更合理的分解策略入手,提出了一种能够充分捕捉到不同步骤之间语义依赖性的模型,进一步提高了实体识别和关系抽取的交互性.Zhao等[61]提出了一种异构图神经网络的联合模型,提高了联合模型性能.Ye等[62]使用批量动态注意掩蔽实现不同模块间的联合优化,并采用三重校正保证了三元组推断的可靠性,推进联合模型性能走向新高.实体识别和关系抽取之间采用共享跨度表示,并通过动态图传播层更新跨度表示能进一步提升联合模型性能[66~69].在此基础上,Lin等[70]延伸了基于DYGI++的跨度表示工作,其整合了全局文本信息,提高了实体识别和关系抽取的交互性.

在中文实体关系联合抽取领域,以先预测头实体后抽取关系和对应尾实体为建模思路的研究也取得了一些研究成果.王泽儒等[71]提出了一种基于指针级联标注的中文实体关系联合抽取模型.该模型采用以实体类型作为区分的指针标注策略来解决实体嵌套与预测实体类别的问题,并以关系模型作为评分函数,将语句中的头实体映射到尾实体以解决中文文本关系重叠问题.李代祎等[72]为加强中文命名实体识别和关系抽取两个子任务间的交互性,构建了一个包含头实体(subject)抽取、尾实体(object)和关系(relation)抽取的多模块联合抽取模型.该模型先利用BERT-BiLSTM-ATT模型识别中文语句中所有的头实体;然后,以头实体的BERT编码向量作为条件,使用条件正则化(Conditional Layer Normalization,CLN)方法对编码序列做归一化处理;最后,采用BiLSTM-ATT模型完成中文语句中关系和尾实体识别任务.这两种中文实体关系联合抽取模型都将头实体、尾实体和关系分别预测,充分利用了中文文本的全局信息,加强了两个子任务的信息交互,有效解决了中文文本语句三元组重叠问题.

以多模块-多步骤建模方法构建的这三类模型的共同特点都是将三元组的提取过程分为多个模块,然后通过共享参数整合各个模块,逐步迭代得到三元组.从联合模型的性能发展状况来看,多模块-多步骤的联合建模方法比流水线方法更优秀.但,这种建模方法依然存在诸多问题:(1)解码误差累积问题,即每个步骤使用独立的解码算法,导致解码误差累积;(2)级联冗余问题,即经共享参数整合的各个模块的冗余误差会互相影响预测性能.后续研究者针对这些问题采用联合解码的方法一步抽取三元组,进一步推进了联合模型性能提升.上述有关多模块-多步骤方法的联合模型总结如表1所示.

3.2 多模块-单步骤建模方法

多模块-多步骤建模方法具有一定局限性.为加强实体模型和关系模型的交互性,复杂的联合解码算法被提出.采用联合解码算法的多模块-单步骤方法具有一些挑战:(1)如何设计精确的联合解码算法比较困难——早期维特比联合解码算法[73],需要限制特征的阶数,增加了模型复杂度,导致训练效率太低;(2)如何设计加强子模块间交互性的联合解码算法较为困难——使用集束搜索的近似联合解码算法[74],可以抽取任意阶特征,但解码结果不精确,导致误差率太高.如何设计一个解决上述挑战的多模块-单步骤模型是一大难点.如何让多模块-单步骤建模方法同时考虑单句中所有实体与实体、实体与关系、关系与关系之间的交互性是另一大难点.随着研究者对这类方法的深入研究,大量优秀的多模块-单步骤模型相继被提出.本文将其总结为两类模型:表格填充模型和集合预测模型.

3.2.1 表格填充模型

早在2017年,Zhang等[75]为实现实体关系抽取的联合解码,提出了一种早期表格填充模型.该模型无法捕捉实体识别和关系抽取的特定信息,而在公开数据集上的性能较差,但这种表格填充的联合解码算法为后续研究者的深入研究奠定了基础.2019年Sun等[76]提出了一种基于图神经网络的表格填充模型实现实体关系联合抽取任务,该模型能够有效捕捉两个子任务的特定信息,但在公开数据集上的效果不好.

Wang等[77]针对早期表格填充模型的问题,提出了一种基于两个独立编码器的表格填充模型.该模型包含序列编码器与表格编码器,序列编码器可以提取实体识别任务的特定信息,而表格编码器可以提取关系抽取的特定信息.这两个编码器相互作用,可以很好地捕获实体识别和关系抽取两个子任务的特定信息.同时,基于两个独立编码器的表格填充模型利用BERT注

表1　多模块-多步骤方法联合模型总结表

模型	联合抽取顺序	描述
Miwa 等[32]	实体域映射关系域	首次提出了基于端到端的树形 LSTM 实体关系联合抽取模型
Katiyar 等[54]		提出了注意力机制和循环神经网络相结合的实体关系联合抽取模型
Giannis B 等[63]		采用多头选择方法对实体关系联合抽取问题进行建模,以解决关系重叠问题
Tan 等[55]		提出了一种采用翻译机制进行排序的 TME 模型,能自适应识别重叠关系三元组
Fu 等[44]		采用图卷积神经网络(GCNs)构建模型,提升了实体关系联合抽取模型的性能
Liu 等[56]		采用有监督的多头自注意力机制构建关系检测模型,提升了联合抽取模型的性能
禹等[64]		用双向关系语义信息建立句中主、客实体间正负关系,从正负两个角度确定三元组
Zeng 等[46]	关系域映射实体域	采用端到端框架,提出了一种 CopyRE 模型,提高了子任务间交互性
Zeng 等[65]		采用端到端框架,提出了一种 CopyMTL 模型以识别多个 token 组成的实体
Yuan 等[57]		提出了细粒度下单词对文本语义关系识别有影响的观点,并构建联合模型进行验证
Zheng 等[58]		所提出的联合模型关系抽取更加准确,且长距离实体间的关系抽取能力得到提升
Ma 等[59]		提出了采用基于双译码器模型检测文本语义关系的方法
Wei 等[49]	头实体域映射关系、尾实体域	提出了一种 HBT 框架,解决了单句中多个实体关系三元组及重叠关系三元组的问题
Yu 等[60]		所提模型能捕捉到不同步骤之间的语义依赖性,进而提升了子任务间的交互性
Zhao 等[61]		提出了一种基于异构图神经网络的实体关系联合模型
Ye 等[62]		将单个共享转换器模块引入端到端的生成模型,并采用对比学习提高模型的可信度
王等[71]		提出了基于指针级联标注策略的中文实体关系联合抽取模型,以解决实体嵌套问题
李等[72]		将实体关系建模为头实体映射到尾实体的函数,以解决三元组的重叠问题

意力权重中携带的字-字交互信息,更全面地学习了表格表示的信息,进一步提高了联合模型的性能. 但实体识别和关系抽取两个子任务的特征表示存在冲突,这反而会对模型预测结果带来不利影响. 文献[78]提出了一种 TPLinker(令牌对链接:Token Pair Linking,TPL)模型消除了两个子任务特征表示的冲突,进一步提升了联合模型在公开数据集上的性能,其模型框架如图8所示.

图8中,从下到上依次是示例样句、握手(handshaking)核函数编码器、实体边界令牌对(token pair)、每种关系下实体边界标记矩阵. TPLinker 模型将实体关系联合抽取看作一个令牌对链接问题,其认为,以某种关系 r 为条件,将头实体 h 和尾实体 e 的边界对齐即可确定三元组 $<h,r,e>$. 为在每种关系类型下都可对齐头实

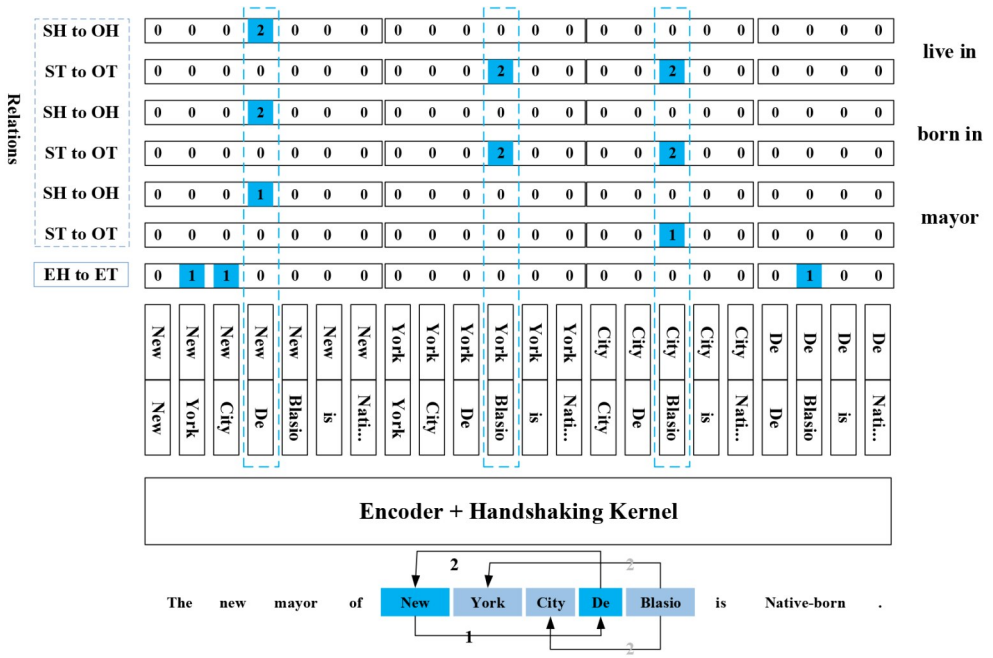


图8　TPLinker 模型架构图[78]

体 h 和尾实体 e 的边界，TPLinker 引入一种握手标记方案，将令牌对链接分为三种类型：EH-to-ET，SH-to-OH，ST-to-OT. EH-to-ET 是指实体的开头和结尾的令牌对标记，如示例中的“<New，York>”；SH-to-OH 是指头实体开头和尾实体开头的令牌对标记，如示例中“<New，De>”；ST-to-OT 是指头实体结尾和尾实体结尾的令牌对标记，如示例中“<York，Blasio>”. 在解码过程中，EH-to-ET 可直接解码出实体. 在确定关系种类时，根据 SH-to-OH 和 ST-to-OT 的令牌对标记可解码出三元组，标记为 1 时表示头尾实体顺序不用调换，标记为 2 时则需要调换. 如，在关系“born in”中，“<New，De>”被标记为 SH-to-OH 类型；“<York，Blasio>”和“<City，Blasio>”被标记为 ST-to-OT 类型，进而解码出具有多重关系的两个三元组：<De Blasio，born in，New York> 和 <De Blasio，born in，New York City>. TPLinker 模型在处理多重关系时具有一定优势，在多个公开数据集上均取得了超越前人模型的效果. 但该模型设计了较为复杂的联合解码算法，在抽取三元组时效率不高.

2021 年 Wang 等[47]采用一种效率更高的联合解码算法，提出了一种 UniRE 表格填充模型，其架构如图 9 所示. 该模型一共分为三个模块：(1)双仿射模块(biaffine model)可以抽取文本中长范围语义特征，采用深度双仿射注意力机制编码表中单词的方向信息；(2)概率张量模块(probability tensor model)填充实体的标签表格；(3)解码模块(decoding model)分三步抽取三元组——先解码实体或实体间跨度，接着解码每个实体跨度的类型，最后解码实体对应的关系类型. 该模型的优点是：(1)使实体识别与关系抽取共享编码器、解码器和标签空间，促进了两个子任务的交互；(2)模型训练的收敛速率与推理速率优于其他模型，同时参数量只有其他模型的一半；(3)在多个公开数据集上均取得了更优异的性能.

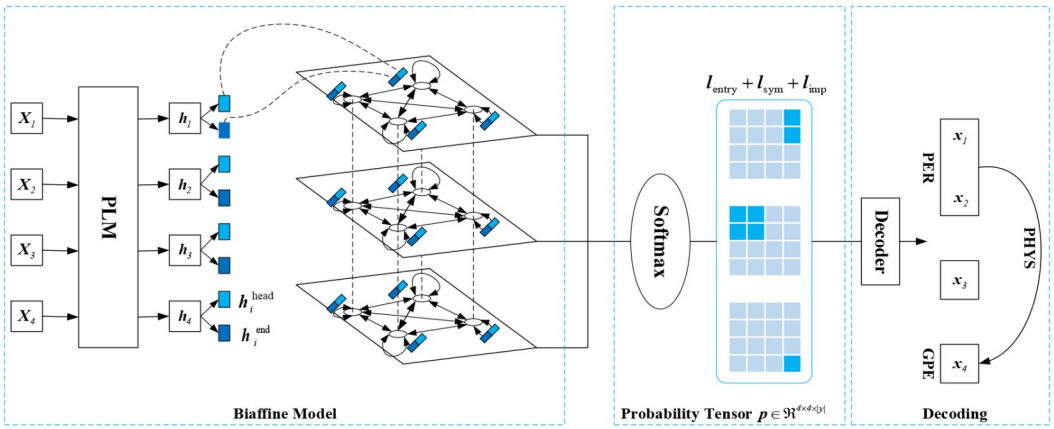


图 9 UniRE 模型架构图[47]

3.2.2 集合预测模型

集合预测模型采用多模块-单步骤建模方法，试图构建统一实体和关系的标注框架或集合预测框架，以实现联合解码抽取三元组. 2017 年 Zheng 等[45]将实体识别和关系抽取统一为一个序列标注任务，提出了一种基于统一标注策略的实体关系联合抽取模型 Novel-Tagging. 该模型针对流水线方法存在的问题，初步加强了两个子任务的交互性. 图 10 展示了统一标注策略的一个示例，共有三种标注信息块：(1)实体中词位置信息{B，I，E，S，O}分别表示{实体开始单词，实体内部单词，实体结束单词，单个单词，无关词}；(2)实体关系类型信息，需要根据实际情况自定义限定关系类型并编码，如{CF，CP，…}；(3)实体角色信息{1，2}，分别表示{实体 1，实体 2}.

基于统一标注策略的实体关系联合模型共有四层，其模型架构如图 11 所示. 嵌入层(embedding layer)生成词向量；编码层(encoding layer)采用 Bi-LSTM 抽取文本的上下文语义特征；解码层(decoding layer)采用 LSTMd 解码蕴含实体语义关系的特征向量，LSTMd 充分利用每一时刻的预测结果 $\boldsymbol{T}_t$，以提高解码准确率；实体关系预测层(softmax layer)分类出最终的预测标记. 为增强实体对之间的联系，该模型增

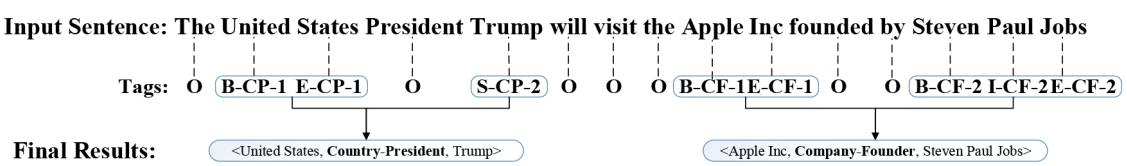


图 10 统一标注策略示例图[45]

添偏置损失函数,削弱了无效实体标签的影响力,进一步提高了模型性能. 其优点是:(1)统一标注策略加强了实体识别和关系抽取的交互性;(2)使用端到端模型进行联合解码,简化联合解码算法. 但缺点也很明显:(1)无法抽取实体间的多重关系;(2)没有充分利用文本全局信息,使其在公开数据集上的性能较差. 从图 11 可以看出,Novel-Tagging 模型使用 Bi-LSTM 模块抽取文本语句的实体特征,使用 LSTMd 模块融合预测实体特征,抽取出实体间语义关系特征,即利用多模块实现文本实体关系的语义特征抽取;同时,采用图 10 所示的实体关系统一标注策略实现一步解码得到三元组. 基于此分析,可以把 Novel-Tagging 模型看作一类采用联合解码方法的多模块单步骤的联合抽取模型.

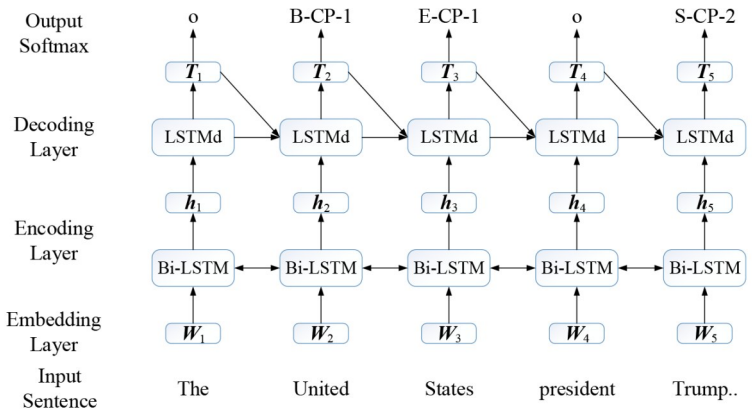


图 11　基于统一标注策略的实体关系联合模型架构[45]

2019 年 Dai 等人[79]针对 Novel-Tagging 模型存在的问题,提出了一种基于注意力机制的统一标注策略联合抽取模型,在公开数据集上取得了更优异的性能. Li 等[48]认为简单的三元组形式不能充分表现实体间的层级性依赖关系,结合强化学习和博弈论思想提出了一种多轮问答模型. 该模型能整合关系抽取任务所需的先验信息,不仅可以抽取简单的实体间多重关系,而且能抽取具有层级性依赖关系的三元组.

2020 年 Sui 等[80]提出了一种集合预测(Set Prediction Networks,SPN)模型,在公开数据集上取得了更优异的性能,其架构如图 12 所示. 该模型包含三大模块:(1)语句编码模块(sentence encoder model)采用 BERT 模型生成词向量作为输入;(2)非自回归解码模块(Non-Autoregressive Decoder Model, NAT)采用 N 层 transformer 块,对输入的词向量和来自于语句编码模块的词向量依次进行解码;(3)将 NAT 生成的解码向量送入 FFN 后经过 softmax 抽取三元组. SPN 模型将实体关系联合抽取任务看作一种集合预测问题,其优点是:(1)使用非自回归并行解码和二部匹配损失函数成功解决集合预测问题;(2)使用简化的联合解码算法抽取实体间的多重关系并在公开数据集上取得优异的性能.

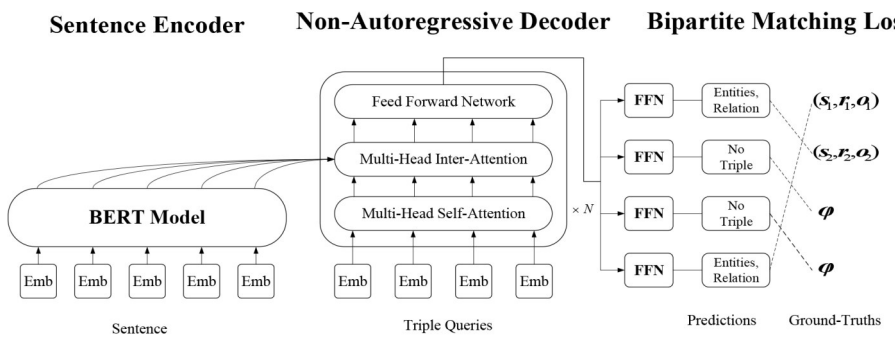


图 12　集合预测模型架构[80]

在中文实体关系联合抽取领域,张军莲等[81]为充分融合中文语句实体关系的复杂结构特征,提出一种基于图卷积神经网络的中文实体关系联合抽取模型. 该模型在双向长短时记忆网络抽取序列特征的基础上,利用 GCN 编码依存分析结果中的语法结构信息,借鉴改进的实体标注策略构建采用联合解码的多模块单步骤方法下的中文实体关系联合抽取模型. 图卷积神经网络能有效编码中文文本的先验词间关系并提升实

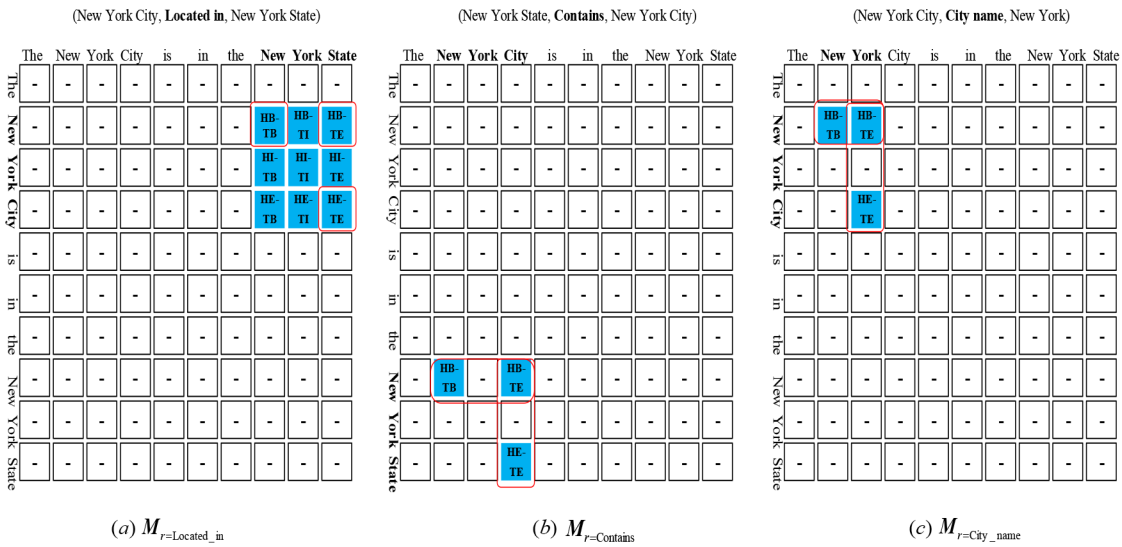
体关系抽取性能. 与英文不同,中文语料在做实体关系抽取任务时,需要先进行分词,边界切分出错会造成歧义问题. 为解决这一问题,葛君伟等[82]提出一种基于字词混合的联合抽取模型. 该模型在嵌入层的词向量基础上结合字向量,并增加位置信息来保证字与字之间的正确顺序,以解决分词边界的歧义问题;然后,引入混合扩张卷积神经网络在不同粒度、更远距离进行特征提取;最后,采用分层标注方法,融合主、客实体和关系信息抽取得到三元组. 这种模型引入中文字词向量,使得抽取中文重叠三元组时效果更好.

综上所述,表格填充模型和集合预测模型都需要解决一个全局最优化联合解码的问题. 各位研究者提出的模型虽然存在很大差异,但这些差异的根源在于从不同角度考虑流水线方法存在的实际问题,其本质特征是多模块-单步骤建模方法. 每个基于多模块-单步骤思想进行建模的模型,都需要构建一个最优化的联合解码算法,并对其求取最优解进而得到最优超参数.

3.3 单模块-单步骤建模方法

基于多模块-多步骤方法的联合建模思想在一些公开数据集上的性能表现优异,表明这一建模思路在未来依然具有丰富的研究前景,但模块分离与逐步迭代抽取三元组导致级联误差的问题出现. 基于多模块-单步骤方法采用联合解码算法减弱了多模块-多步骤方法的级联误差,但组合三元组时会出现冗余错误,依然存在局限性. 这两种建模方法具有的局限性是由模型分块和预测三元组分步的误差产生的. 因此,如何构建没有误差的基于单模块-单步骤的联合模型成为这一领域被重点关注的问题.

在3.2.2节提到的2017年由Zheng等[45]人提出的实体关系联合抽取模型Novel-Tagging,虽然可以利用多模块方法一次性从文本语句中直接抽取实体关系三元组,但所抽取的关系是一般关系,对语句中包含的重叠关系无法抽取,导致该模型具有一定的局限性. 但该模型在原始语句上对实体进行标注时,也同时把关系类别进行了标注,这种一个标注步骤就能完成实体关系三元组标注的方法,为后续相关研究者提供了启迪. 2022年,Shang等[34]提出的OneRel模型就借鉴了Novel-Tagging的这一思想,直接从文本语句中抽取三元组,在公开数据集上取得了超越前人研究的性能,其模型架构如图13所示. 具体来说,OneRel模型基于单模块-单步骤的建模思想,构建了一种基于细粒度的三重分类模型. OneRel模型认为对每种关系,在实体的token层面做细粒度三重分类即可抽取三元组. 对于每种关系,OneRel模型会生成一个如图13所示的标记矩阵. 样本句的每个token可被分为9种类别(图9(a)中标记的深蓝色9宫格),但只需要保留上三角的三个类别即可解码出三元组. 这三个类别分别是:(1)HB-TB表示头实体开头和尾实体开头;(2)HB-TE表示头实体开头和尾实体结尾;(3)HE-TE表示头实体结尾和尾实体结尾. 因此,在关系"Located in"可以抽取出头实体"New York City"和尾实体"New York State",进而组成三元组 < New York City, Located in, New York State >. 图13中(a),(b)和(c)是同一样本句在不同关系中的标记结果,这表明OneRel模型在解决EntityPairOverlap(EPO),SingleEntityOverlap(SEO)和HeadTailOverLap(HTO)等多重关系问题时具有优势.



(a) $\boldsymbol{M}_{r=\text{Located\_in}}$ (b) $\boldsymbol{M}_{r=\text{Contains}}$ (c) $\boldsymbol{M}_{r=\text{City\_name}}$

图13 OneRel模型架构[34]

OneRel模型是在单模块-单步骤建模思想指导下产生的联合模型,这种模型可以直接从文本语句中抽取三元组,有效缓解了多模块-多步骤建模方法和多模块-单步骤建模方法的级联错误和实体冗余等问题. 在

公开数据集上,OneRel模型也刷新了F_1性能,这表明,单模块-单步骤的指导思想下构建的联合模型是有效的.但OneRel模型只能从单句中抽取三元组,无法从段落中或篇章中抽取三元组.联合模型可以提高实体识别和关系抽取两个子任务之间的交互性,使得两个子任务的训练过程中可以交融二者的特征,以提高模型性能,但这会产生特征冲突的问题.特征冲突在单一模型中对性能影响较大,文献[33]在这方面进行研究,试图借助流水线方法减弱特征冲突带来的问题,且该流水线模型在公开数据集上取得了较好的效果,但该模型是总结前人联合模型的优点,在此基础上产生的效果.联合模型从多模块-多步骤、多模块-单步骤逐渐向单一模型的单模块-单步骤发展,确实可以解决流水线存在的问题,但自身又会显现一定的缺陷,而这种缺陷需要后续对单一联合模型进一步研究以解决特征冲突的问题.表2总结了基于深度学习的实体关系联合抽取模块的建模方法和优缺点.

表2　基于深度学习的实体关系联合抽取模型总结表

建模方法	优点	缺点	模型类型	文献	描述
多模块-多步骤	不同子模块能抽取丰富的特征信息	逐步抽取三元组使得信息交互不充分	实体域映射关系域	[32,44,54~56,63]	先进行命名实体识别,再根据其结果实现关系抽取
			关系域映射实体域	[46,57~59,65]	先抽取关系,再根据关系类型建模实体对的映射
			头实体域映射关系、尾实体域	[49,60~62,66~70]	先识别头实体,根据头实体抽取相应关系和尾实体
多模块-单步骤	采用联合解码器,实体关系两个子任务的信息得以充分交互	联合解码算法较为复杂,使得两个子任务局部特征抽取不够充分	表格填充模型	[47,75~77]	设计复杂的标签表格框架,标签需要同时表示实体及关系信息,采用联合解码抽取三元组
			集合预测模型	[45,48,80,79]	构建统一实体和关系的标注框架或集合预测框架,采用联合解码抽取三元组
单模块-单步骤	使得实体关系信息充分交互并减轻冗余误差	模型较少,存在实体识别和关系抽取之间的特征冲突问题	细粒度三重分类模型	[34]	在token层面构建三重分类模型,直接抽取三元组

4　数据集

在有监督领域,用作评估基于深度学习的实体关系联合抽取模型性能的公开数据集主要有ACE04,ACE05,CoNLL04,ADE,NYT,WebNLG这六种.

(1)ACE关系抽取数据集:关系抽取任务在2002—2007年被当作ACE会议的一个子任务,ACE会议提供的ACE04/ACE05数据集被认为是实体关系抽取领域的权威公开标准评测数据集.ACE04语料库来自于语言数据联盟(Linguistic Data Consortium,LDC),主要包含新闻专线和广播新闻两部分,共有451个文档和5 702个关系实例,其具有丰富的标注语句,为信息抽取领域的实体识别、指代消解以及关系抽取等子任务提供了基准的训练和测试语料库.ACE05数据集是对ACE04的进一步扩充,并做出了相应的修改和标注信息的完善.ACE04数据集的实体关系类型如表3所示.

(2)CoNLL04数据集:该数据集共有1 437个含有关系的语句样本.语句中的实体和关系类型都进行了标注.其一共包含5 336个实体,19 048个实体对(二元关系),共有四种实体类型和六种关系类型.其具体类别如表4所示.

表3　ACE04数据集

实体类型	关系类型
Person	PHYS
Organization	PER-SOC
Geographical Entities	PER/ORG-AFF
Location	ART
Facility	EMP-ORG
Weapon	GPE-AFF
Vehicle	DISC

(3)ADE数据集:ADE(Adverse Drug Events)数据集

表4　CoNLL04数据集

实体类型	实体数量	关系类型	关系数量
Person	1685	Located In	406
Location	1968	Work For	394
Organization	978	OrgBased In	451
Other	705	Live In	521
—	—	Kill	268
—	—	None	17 007

存在两种实体类型:Drug和Disease. 该数据集的目的是抽取Drug类型实体和Disease类型实体,同时抽取二者的关系. ADE数据集来源于1 644个PubMed的摘要信息,从摘要中选取至少存在一组实体类型为Drug-Disease且关系类型为ADE的句子. 其共有6 821条文本语句,包含10 652个实体和6 682个关系.

(4)NYT数据集:这是在远程监督实体关系抽取领域常用的数据集,也会被用在实体关系联合抽取领域做模型性能评估. 其主要是由Freebase知识库对纽约时报文本获取的数据集. NYT数据集的训练集由2005年、2006年文本获取,测试数据集为知识库对2007年文本获取. NYT数据集总计695 059条数据,其中训练集为522 611条数据,其中80%的句子标签为NA,测试集含有172 448条语句,共53种关系. 通过结合FreeBase对NYT语料做实体链接、关系对齐等操作进行标注,最终得到一个被广泛使用的关系抽取数据集.

(5)WebNLG数据集:WebNLG语料库由一组描述事实的实体关系三元组和以自然语言文本形式对应的事实组成. 该数据集包含216种关系类别,拥有5 019个训练样本,验证样本和测试样本数目分别是500和703个. 最初,该数据集被用于WebNLG自然语言生成挑战,这些挑战包含引用表达生成、聚合、词汇化、表面实现和句子分割等任务. 此外,该数据集还常用于三元组提取的反向任务. 近年来,WebNLG数据集已经成为评估三元组抽取模型效果最常用的通用领域数据集.

(6)DuIE数据集[83]:DuIE数据集来自百度2020语言与智能技术竞赛,是当前规模相对比较大的中文信息抽取数据集. 其包含458 184个三元组数据实例、214 739条中文语句以及包含50个预定义好的关系集合. 中文语句主要来自百度百科和百度新闻提要. 表5是对实体关系联合抽取常用公开数据集的总结.

表5 实体关系联合抽取常用公开数据集总结表

数据集	实体类型数	关系类型数	样本数	数据来源	网址
ACE04	7	7	6 800	语言数据联盟	https://catalog.ldc.upenn.edu/LDC2005T09
ACE05	7	6	10 500	语言数据联盟	https://catalog.ldc.upenn.edu/LDC2006T06
CoNLL04	4	6	1 400	国际文本信息检索会议	https://cogcomp.seas.upenn.edu/page/resource_view/43
ADE	2	1	6 800	美国国家医学图书馆	https://github.com/lavis-nlp/spert/tree/master/scripts
NTY	3	24	66 200	纽约时报	https://github.com/xiangrongzeng/copy_re
WebNLG	—	246	6 200	DBpedia	https://github.com/weizhepei/CasRel/tree/master/data/WebNLG
DuIE	—	50	214 739	百度	https://aistudio.baidu.com/aistudio/competition/detail/31?isFromCcf=true

5 评测指标

实体关系联合抽取领域采用3项基本评价指标:准确率(Precision,P)、召回率(Recall,R)和F_1值(F_1 Measure),其公式分别为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$F_1=\frac{2\cdot P\cdot R}{P+R} \tag{3}$$

其中,TP为真正例,FN为假反例,FP为假正例,TN为真反例. 除了使用综合评价指标F_1值对实体关系联合抽取模型进行评价之外,研究人员还常用PR(Precision-Recall curve)曲线或者AUC(Area Under Curve)值将自己的模型和其他模型进行比较.

表6汇总了实体关系联合抽取模型在NYT和WebNLG数据集上的性能结果对比,用"*"区别数据集的两个版本. 带有"*"是指只标注实体最后一个单词的数据集,反之则表示对实体内每个单词都进行标注的版本. 在该表中,模型名称前面的字母具有特殊含义,具体详见表7.

分析表6可以发现,NYT和WebNLG数据集的模型性能逐步提高. 在多模块-多步骤建模方法指导的模型中,关系域映射实体域模型和头实体域映射关系、尾部实体域模型要明显优于实体域映射关系域模型. 这表明,充分考虑关系信息与实体信息的交互有助于提高模型性能. 随后的多模块-单步骤模型又有所提高,这证明使用联合解码代替多步骤独立解码的方式有助于减弱多步骤之间的级联误差,最终提高性能. 最后的单模型-单步骤方法虽然研究较少,但从性能结果来看,整合多个子模块的方式也可降低模块间的级联误差,以提升性能. 从NYT和WebNLG数据集上联合模型性能的发展来看,实体关系联合抽取的建模方向正向着理想化的单模块-单步骤建模方法发展.

表8是对ACE数据集上模型性能的总结. 其中,"Ent"表示实体识别,"Rel"表示非关系抽取,"Rel+"表示严格关系抽取(实体边界和类型都必须正确). 表9是对CoNLL04和ADE数据集上模型性能的总结. 其模型名词前面的字母是对模型的分类,详见表7,没有字母

表6　NYT和WebNLG数据集联合模型性能对比表

模型	部分标注						完整标注					
	NYT*			WebNLG*			NYT			WebNLG		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
A-TME[55]	69.6	47.8	56.7	—	—	—	—	—	—	—	—	—
A-GraphRel[44]	63.9	60.0	61.9	44.7	41.1	42.9	—	—	—	—	—	—
A-MHSA[56]	**88.1**	**78.5**	**83.0**	**89.5**	**86.0**	**87.7**	—	—	—	—	—	—
B-CopyRE[46]	61.0	56.6	58.7	37.7	36.4	37.1	—	—	—	—	—	—
B-CopyMTL[65]	—	—	—	—	—	—	75.7	68.7	72.0	58.0	54.9	56.4
B-RSAN[57]	—	—	—	—	—	—	85.7	83.6	84.6	80.5	83.8	82.1
B-CasDE[59]	90.2	90.9	90.5	90.3	91.5	90.9	89.9	91.4	90.6	88.0	**88.9**	88.4
B-PRGC[58]	**93.3**	**91.9**	**92.6**	**94.0**	**92.1**	**93.0**	**93.5**	**91.9**	**92.7**	**89.9**	87.2	**88.5**
C-ETL-span[60]	84.9	72.3	78.1	84.0	91.5	87.6	85.5	71.7	78.0	84.3	82.0	83.1
C-CasRel[49]	89.7	89.5	89.6	93.4	90.1	91.8	—	—	—	—	—	—
C-RIFRE[61]	93.6	**90.5**	**92.0**	**93.3**	**92.0**	92.6	—	—	—	—	—	—
C-CGT[62]	**94.7**	84.2	89.1	92.9	75.6	**93.4**	—	—	—	—	—	—
D-TPLinker[78]	**91.3**	**92.5**	**91.9**	**91.8**	**92.0**	**91.9**	**91.4**	**92.6**	**92.0**	**88.9**	**84.5**	**86.7**
E-Bi-LSTM-bias[45]	0.693	0.414	0.495	—	—	—	—	—	—	—	—	—
E-PA-LSTM-CRF[79]	0.494	0.591	0.538	—	—	—	—	—	—	—	—	—
E-SPN[80]	**93.3**	**91.7**	**92.5**	**93.1**	**93.6**	**93.4**	**92.5**	**92.2**	**92.3**	—	—	—
F-OneRel[34]	92.8	**92.9**	**92.8**	**94.1**	**94.4**	**94.3**	93.2	**92.6**	**92.9**	**91.8**	**90.3**	**91.0**

表7　模型名词头字母含义表

字母	建模方法	模型类别
A	多模块-多步骤	实体域映射关系域模型
B	多模块-多步骤	关系域映射实习域模型
C	多模块-多步骤	头实体域映射关系、尾实体域模型
D	多模块-单步骤	表格填充模型
E	多模块-单步骤	集合预测模型
F	单模块-单步骤	细粒度三重分类模型

表明未被归为本文联合模型分类体系.

分析表8和表9可以发现,基于特征工程的Li等[30]的模型在不同的数据集上取得的性能都较低. 因为,基于特征工程的联合抽取模型需要针对特定领域的数据集人工构建特征,这会导致对文本信息使用不充分,无法获取深层次语义特征. 而基于神经网络的Giannis等[84]和Bekoulis等[85]的模型在不同的数据集上效果较好,超越基于特征工程的模型,这表明在提取文本深层次语义特征上,基于深度学习的实体关系联合抽取模型具有极大优势. 对于近些年较火的预训练模型,SPAN[86]和Wang等[77]的联合模型在不同数据集上都取得了更为优秀的模型. 总体来说,对于更具一般性的ACE,CoNLL04,ADE等数据集,基于深度学习的实体关系联合抽取模型具有更大优势,尤其是以预训练-微调为主的现代自然语言处理技术为统领的当下和未来.

表8　实体关系联合抽取典型模型在ACE数据集(部分)的性能结果对比表

模型	ACE05			ACE04		
	Ent(F_1)	Rel(F_1)	Rel+(F_1)	Ent(F_1)	Rel(F_1)	Rel+(F_1)
Li等[30]	80.8	52.1	49.5	79.7	48.3	45.3
A-Miwa[32]	83.4	—	55.6	81.8	—	48.4
A-Katiyar[54]	82.6	55.9	53.6	79.6	49.3	45.7
D-Zhang[75]	83.6	—	57.5	—	—	—
C-Luan[68]	88.4	63.2	—	87.4	59.7	—
Li等[48]	84.8	—	60.2	83.6	—	49.4
Dixit[87]	86.0	—	62.8	—	—	—
C-Wadden[69]	88.6	63.4	—	—	—	—
C-Lin[70]	88.8	67.5	—	—	—	—
D-Wang[77]	89.5	**67.6**	64.3	88.6	**63.3**	**59.6**
SPAN[86]	89.6	—	**65.2**	—	—	—
D-UniRE[47]	**90.2**	66.0	—	**89.5**	63.0	—

6　联合模型研究展望

基于深度学习的实体关系联合抽取模型已经在公开数据集上取得优异的性能. 为构建更具应用性的联合抽取模型,总结当下联合模型的研究进展,未来研究可聚焦以下几个方面.

(1) 开放域实体关系联合抽取

当下,主流的实体关系联合抽取模型大都在限定域关系类别集合里做关系分类任务. 对于关系类型OOV(Out Of Vocabulary)问题(开放域实体关系联合抽

表 9 CoNLL04 数据集和 ADE 数据集联合模型性能比对表

Data	Model	Ent(F_1)	Rel(F_1)
CoNLL04	A-Miwa[32]	80.7	61.0
	Giannis[84]	83.6	62.0
	Bekoulis[85]	83.9	62.0
	D-Zhang[75]	85.6	67.8
	Li 等[30]	87.8	68.9
	SpERT[88]	88.9	71.5
	Zhao 等[89]	88.9	71.9
	D-Wang[77]	90.1	73.6
	SPAN[86]	**90.2**	**74.3**
ADE	Li 等[30]	84.6	71.4
	Giannis 等[84]	86.4	74.6
	Bekoulis 等[85]	86.7	75.5
	SpERT[88]	89.3	79.2
	D-Wang[77]	89.7	80.1
	SPAN[86]	**90.6**	**80.7**

取问题),当前主流的联合模型解决的不好. 关系类型 OOV 问题是指抽取出不在预定义好的关系类别集合中的其他关系类别. 已有的联合模型框架无法准确预测出这种开放域的实体间关系类型. 虽然某些公开数据集中,引入 Other 类对不属于限定集合的关系类型的实例进行了描述,但这只是将可能存在的其他关系类型粗糙的划分为 Other 类,即使提升了模型的性能,仍然需要人工干预解决 Other 类关系类型难定义和模糊等问题. 因此,开放域的实体关系联合抽取问题是未来亟待解决的问题之一.

(2) 融入多元信息的实体关系联合抽取

当前的实体关系联合抽取模型大多是依据单一文本上下文信息进行特征抽取进而抽取出三元组. 在实际工业工程应用中,语料库中包含多元信息. 除文本信息外,对于含有时序信息的语料,其实体间的关系可能与时间具有某种关系而影响关系类别的变化. 如何在特征抽取时合理融合时序信息来提升联合模型鲁棒性仍有待研究. 此外,在包含事件的语料中存在大量的事件之间的因果关系,事件的变换发展会影响实体之间的关系类别. 因此,未来的联合模型需要同时考虑一对实体间相关联的不同事件,以提高联合模型的性能. 为构建更具有实用意义的联合模型,融入多元信息的实体关系联合抽取模型的研究是未来一项具有重大意义的课题.

(3) 跨文本的实体关系联合抽取

当下的联合模型主要集中在同一篇章的跨段、跨句和跨语义的层级性依赖方面. 对于在同一语料库中,不同文本之间的实体关系抽取研究较少. 受制于预训练语言模型 BERT 输入长度的限制,很难将多文本组合为长篇文本进行模型训练. 如何处理不同文本间的关系信息、不同关系间的关系信息,多个实体共指等复杂情况仍有待解决.

7 总结

基于深度学习的实体关系联合抽取方法的研究逐步解决了基于人工特征提取成本高、效率低和基于流水线方法错误积累的一系列问题. 随着近年来的研究发展,基于深度学习的实体关系联合抽取方法催生了一系列经典模型. 本文梳理了近十年自然语言处理顶会中与该领域相关的文章,详细阐述了实体关系联合抽取的研究进程中,针对实体关系错误累积、缺少子任务间的信息交互、冗余实体等问题的解决方案. 分析近些年人工智能与自然语言处理领域前沿学术文献,可将其划分为三种建模方法:以共享参数整合各个子模块的多模块-多步骤方法、以联合解码算法为主的多模块-单步骤方法、以细粒度三重分类为代表的单模块-单步骤方法. 本文对基于深度学习的实体关系联合抽取的这三类建模方法下产生的经典模型的优缺点进行了分析,总结了联合模型发展趋势,阐明了从基于多模块-多步骤方法和基于多模块-单步骤方法,逐渐向单模块-单步骤建模方法转变的客观趋势. 最后对实体关系联合抽取的未来研究趋势进行了探讨和展望. 本文尝试建立一个较为完整的基于深度学习的实体关系联合抽取领域研究视图,希望能对相关领域研究者有所帮助.

参考文献

[1] HAHN U, OLEYNIK M. Medical information extraction in the age of deep learning[J]. Yearbook of Medical Informatics, 2020, 29(1): 208-220.

[2] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600.
LIU Q, LI Y, DUAN H, et al. A survey of knowledge mapping construction techniques[J]. Journal of Computer Research and Development, 2016, 53(3): 582-600. (in Chinese)

[3] 徐健, 张智雄, 吴振新. 实体关系抽取的技术方法综述[J]. 现代图书情报技术, 2008, (8): 18-23.
XU J, ZHANG Z X, WU Z X. Review on techniques of entity relation extraction[J]. New Technology of Library and Information Service, 2008, (8): 18-23. (in Chinese)

[4] 甘丽新, 万常选, 刘德喜, 等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展, 2016, 53(2): 284-302.
GAN L X, WAN C X, LIU D X, et al. Chinese entity rela-

tionship extraction based on syntactic and semantic features [J]. Journal of Computer Research and Development, 2016, 53 (2): 284-302. (in Chinese)

[5] TJONG KIM SANG E F, DE MEULDER F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition[C]//Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. Edmonton: Association for Computational Linguistics, 2003: 142-147.

[6] RATINOV L, ROTH D. Design challenges and misconceptions in named entity recognition[C]//Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Boulder: Association for Computational Linguistics, 2009: 147-155.

[7] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction[C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002: 71-78.

[8] BUNESCU R C, MOONEY R J. A shortest path dependency kernel for relation extraction[C]//Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver: Association for Computational Linguistics, 2005: 724-731.

[9] FLORIAN R, HASSAN H, ITTYCHERIAH A, et al. A statistical model for multilingual entity detection and tracking[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. Boston: Association for Computational Linguistics, 2004: 1-8.

[10] FLORIAN R, JING H Y, KAMBHATLA N, et al. Factorizing complex models: A case study in mention detection [C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL. Sydney: Association for Computational Linguistics, 2006: 473-480.

[11] ZHOU G D, SU J, ZHANG J, et al. Exploring various knowledge in relation extraction[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor: Association for Computational Linguistics, 2005: 427-434.

[12] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the ACL Interactive Poster and Demonstration Sessions. Barcelona: Association for Computational Linguistics, 2004: 178-181.

[13] CHAN Y S, ROTH D. Exploiting sytactico-semantic structures for relation extraction[C]//The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: Association for Computational Linguistics, 2011: 551-560.

[14] HEINZERLING B, STRUBE M. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages [C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki: European Language Resources Association, 2018: 2989-2993.

[15] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

[16] 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: Association for Computational Linguistics, 2019: 4171-4186.

[17] MA L L, YANG J, AN B, et al. Medical named entity recognition using weakly supervised learning[J]. Cognitive Computation, 2022, 14(3): 1068-1079.

[18] 罗凌, 杨志豪, 宋雅文, 等. 基于笔画ELMo和多任务学习的中文电子病历命名实体识别研究[J]. 计算机学报, 2020, 43(10): 1943-1957.
LUO L, YANG Z H, SONG Y W, et al. Chinese clinical named entity recognition based on stroke ELMo and multi-task learning[J]. Chinese Journal of Computers, 2020, 43(10): 1943-1957. (in Chinese)

[19] 胡宇, 申德荣, 聂铁铮, 等. 面向生物医学实体链接的联合式学习方法[J]. 计算机学报, 2022, 45(4): 748-765.
HU Y, SHEN D R, NIE T Z, et al. A joint learning method for biomedical entity linking[J]. Chinese Journal of Computers, 2022, 45(4): 748-765. (in Chinese)

[20] 郜成胜, 张君福, 李伟平, 等. 一种基于混合神经网络的命名实体识别与共指消解联合模型[J]. 电子学报, 2020, 48(3): 442-448.
GAO C S, ZHANG J F, LI W P, et al. A joint model of named entity recognition and coreference resolution based on hybrid neural network[J]. Acta Electronica Sinica, 2020, 48(3): 442-448. (in Chinese)

[21] ZENG D J, LIU K, CHEN Y B, et al. Distant supervision

for relation extraction via piecewise convolutional neural networks[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015: 1753-1762.

[22] HUANG Y Y, WANG W Y. Deep residual learning for weakly-supervised relation extraction[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 1803-1807.

[23] VU N T, ADEL H, GUPTA P, et al. Combining recurrent and convolutional neural networks for relation classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016: 534-539.

[24] ZHANG Y H, QI P, MANNING C D. Graph convolution over pruned dependency trees improves relation extraction[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 2205-2215.

[25] ZHU H, LIN Y K, LIU Z Y, et al. Graph neural networks with generated parameters for relation extraction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 1331-1339.

[26] XIAO M G, LIU C. Semantic relation classification via hierarchical recurrent neural network with attention[C]// Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka: The COLING 2016 Organizing Committee, 2016: 1254-1263.

[27] 冯建周, 宋沙沙, 王元卓, 等. 基于改进注意力机制的实体关系抽取方法[J]. 电子学报, 2019, 47(8): 1692-1700.
FENG J Z, SONG S S, WANG Y Z, et al. Entity relation extraction based on improved attention mechanism[J]. Acta Electronica Sinica, 2019, 47(8): 1692-1700. (in Chinese)

[28] 李志欣, 孙亚茹, 唐素勤, 等. 双路注意力引导图卷积网络的关系抽取[J]. 电子学报, 2021, 49(2): 315-323.
LI Z X, SUN Y R, TANG S Q, et al. Dual attention guided graph convolutional networks for relation extraction [J]. Acta Electronica Sinica, 2021, 49(2): 315-323. (in Chinese)

[29] 赵超, 谢松县, 曾道建, 等. 融合预训练语言模型和标签依赖知识的关系抽取方法[J]. 中文信息学报, 2022, 36(1): 75-82.
ZHAO C, XIE S X, ZENG D J, et al. Combination of pretrained language model and label dependency for relation extraction[J]. Journal of Chinese Information Processing, 2022, 36(1): 75-82. (in Chinese)

[30] LI Q, JI H. Incremental joint extraction of entity mentions and relations[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014: 402-412.

[31] 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: Association for Computational Linguistics, 2014: 1858-1869.

[32] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016: 1105-1116.

[33] ZHONG Z X, CHEN D Q. 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. Stroudsburg: Association for Computational Linguistics, 2021: 50-61.

[34] SHANG Y M, HUANG H Y, MAO X L. OneRel: Joint entity and relation extraction with one module in one step [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022, 36(10): 11285-11293.

[35] 鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述[J]. 软件学报, 2019, 30(6): 1793-1818.
E H H, ZHANG W J, XIAO S Q, et al. Survey of entity relationship extraction based on deep learning[J]. Journal of Software, 2019, 30(6): 1793-1818. (in Chinese)

[36] 冯钧, 张涛, 杭婷婷. 重叠实体关系抽取综述[J]. 计算机工程与应用, 2022, 58(1): 1-11.
FENG J, ZHANG T, HANG T T. Survey of overlapping entities and relations extraction[J]. Computer Engineering and Applications, 2022, 58(1): 1-11. (in Chinese)

[37] 李冬梅, 张扬, 李东远, 等. 实体关系抽取方法研究综述[J]. 计算机研究与发展, 2020, 57(7): 1424-1448.
LI D M, ZHANG Y, LI D Y, et al. Review of entity rela-

tion extraction methods[J]. Journal of Computer Research and Development, 2020, 57(7): 1424-1448. (in Chinese)

[38] 刘辉, 江千军, 桂前进, 等. 实体关系抽取技术研究进展综述[J]. 计算机应用研究, 2020, 37(S2): 1-5.
LIU H, JIANG Q J, GUI Q J, et al. Review of research progress of entity relationship extraction[J]. Application Research of Computers, 2020, 37(S2): 1-5. (in Chinese)

[39] 王传栋, 徐娇, 张永. 实体关系抽取综述[J]. 计算机工程与应用, 2020, 56(12): 25-36.
WANG C D, XU J, ZHANG Y. Survey of entity relation extraction[J]. Computer Engineering and Applications, 2020, 56(12): 25-36. (in Chinese)

[40] 张少伟, 王鑫, 陈子睿, 等. 有监督实体关系联合抽取方法研究综述[J]. 计算机科学与探索, 2022, 16(4): 713-733.
ZHANG S W, WANG X, CHEN Z R, et al. Survey of supervised joint entity relation extraction methods[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(4): 713-733. (in Chinese)

[41] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2014: 601-610.

[42] YU X F, LAM W. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Beijing: Coling 2010 Organizing Committee, 2010: 1399-1407.

[43] REN X, WU Z Q, HE W Q, et al. CoType: Joint extraction of typed entities and relations with knowledge bases [C]//Proceedings of the 26th International Conference on World Wide Web. Perth: International World Wide Web Conferences Steering Committee, 2017: 1015-1024.

[44] FU T J, LI P H, MA W Y. GraphRel: Modeling text as relational graphs for joint entity and relation extraction[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 1409-1418.

[45] ZHENG S C, WANG F, BAO H Y, et al. Joint extraction of entities and relations based on a novel tagging scheme [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017: 1227-1236.

[46] 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. Melbourne: Association for Computational Linguistics, 2018: 506-514.

[47] WANG Y J, SUN C Z, WU Y B, et al. UniRE: A unified label space for entity relation extraction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 220-231.

[48] LI X Y, YIN F, SUN Z J, et al. Entity-relation extraction as multi-turn question answering[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 1340-1350.

[49] WEI Z P, SU J L, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 1476-1488.

[50] TAI KAI SHENG, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015: 1556-1566.

[51] XU Y, MOU L L, LI G, et al. Classifying relations via long short term memory networks along shortest dependency paths[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015: 1785-1794.

[52] LI J W, LUONG T, JURAFSKY D, et al. When are tree structures necessary for deep learning of representations? [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015: 2304-2314.

[53] XU K, FENG Y S, HUANG S F, et al. Semantic relation classification via convolutional neural networks with simple negative sampling[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015: 536-540.

[54] KATIYAR A, CARDIE C. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017: 917-928.

[55] TAN Z, ZHAO X, WANG W, et al. Jointly extracting multiple triplets with multilayer translation constraints [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019, 33(1): 7080-7087.

[56] LIU J, CHEN S W, WANG B Q, et al. Attention as relation: Learning supervised multi-head self-attention for relation extraction[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama: International Joint Conferences on Artificial Intelligence Organization, 2020: 3787-3793.

[57] YUAN Y, ZHOU X F, PAN S R, et al. A relation-specific attention network for joint entity and relation extraction [C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama: International Joint Conferences on Artificial Intelligence Organization, 2021: 4054-4060.

[58] ZHENG H Y, WEN R, CHEN X, et al. PRGC: Potential relation and global correspondence based joint relational triple extraction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6225-6235.

[59] MA L B, REN H M, ZHANG X L. Effective cascade dual-decoder model for joint entity and relation extraction[EB/OL]. (2021-06-21)[2022-10]. https://arxiv.org/abs/2106.14163.

[60] YU B W, ZHANG Z Y, SHU X B, Et al. Joint extraction of entities and relations based on a novel decomposition strategy[C]//24th European Conference on Artificial Intelligence. Santiago de Compostela: IOS Press, 2020: 2282-2289.

[61] ZHAO K, XU H, CHENG Y, et al. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction[J]. Knowledge-Based Systems, 2021, 219: 106888.

[62] YE H B, ZHANG N Y, DENG S M, et al. Contrastive triple extraction with generative transformer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021, 35(16): 14257-14265.

[63] BEKOULIS G, DELEU J, DEMEESTER T, et al. Adversarial training for multi-context joint entity and relation extraction[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 2830-2836.

[64] 禹克强, 黄芳, 吴琪, 等. 基于双向语义的中文实体关系联合抽取方法[J]. 计算机工程, 2023, 49(1): 92-99, 112.
YU K Q, HUANG F, WU Q, et al. Joint extraction method for Chinese entity relationship based on bidirectional semantics[J]. Computer Engineering, 2023, 49(1): 92-99, 112. (in Chinese)

[65] ZENG D J, ZHANG H R, LIU Q Y. CopyMTL: Copy mechanism for joint extraction of entities and relations with multi-task learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020, 34(5): 9507-9514.

[66] 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: Association for Computational Linguistics, 2017: 188-197.

[67] HE L H, LEE K, LEVY O, et al. Jointly predicting predicates and arguments in neural semantic role labeling[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 364-369.

[68] LUAN Y, WADDEN D, HE L H, et al. A general framework for information extraction using dynamic span graphs[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019: 3036-3046.

[69] 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: Association for Computational Linguistics, 2019: 5788-5793.

[70] LIN Y, JI H, HUANG F, et al. A joint neural model for information extraction with global features[C]//Proceedings of the 58th Annual Meeting of the Association for

Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 7999-8009.

[71] 王泽儒, 柳先辉. 基于指针级联标注的中文实体关系联合抽取模型[J]. 武汉大学学报(理学版), 2022, 68(3): 304-310.
WANG Z R, LIU X H. Joint model of Chinese entity-relation extraction based on a pointer cascade tagging strategy [J]. Journal of Wuhan University (Natural Science Edition), 2022, 68(3): 304-310. (in Chinese)

[72] 李代祎, 李忠良, 严丽. 一种面向中文的实体关系联合抽取方法研究[J]. 小型微型计算机系统, 2022, 43(12): 2479-2486.
LI D W, LI Z L, YAN L. Research on Chinese-oriented entity relation joint extraction method[J]. Journal of Chinese Computer Systems, 2022, 43(12): 2479-2486. (in Chinese)

[73] KATIYAR A, CARDIE C. Investigating LSTMs for joint extraction of opinion entities and relations[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016: 919-929.

[74] LI Q, JI H. Incremental joint extraction of entity mentions and relations[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore: Association for Computational Linguistics, 2014: 402-412.

[75] ZHANG M S, ZHANG Y, FU G H. End-to-end neural relation extraction with global optimization[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 1730-1740.

[76] SUN C Z, GONG Y Y, WU Y B, et al. Joint type inference on entities and relations via graph convolutional networks[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 1361-1370.

[77] WANG J, LU W. Two are better than one: Joint entity and relation extraction with table-sequence encoders[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 1706-1721.

[78] WANG Y C, YU B W, ZHANG Y 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: International Committee on Computational Linguistics, 2020: 1572-1582.

[79] DAI D, XIAO X Y, LYU Y J, et al. Joint extraction of entities and overlapping relations using position-attentive sequence labeling[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019, 33(01): 6300-6308.

[80] SUI D B, CHEN Y B, LIU K, et al. Joint entity and relation extraction with set prediction networks[EB/OL]. (2022-11-03)[2022-10]. https://arxiv.org/abs/2011.01675.

[81] 张军莲, 张一帆, 汪鸣泉, 等. 基于图卷积神经网络的中文实体关系联合抽取[J]. 计算机工程, 2021, 47(12): 103-111.
ZHANG J L, ZHANG Y F, WANG M Q, et al. Joint extraction of Chinese entity relations based on graph convolutional neural network[J]. Computer Engineering, 2021, 47(12): 103-111. (in Chinese)

[82] 葛君伟, 李帅领, 方义秋. 基于字词混合的中文实体关系联合抽取方法[J]. 计算机应用研究, 2021, 38(9): 2619-2623.
GE J W, LI S L, FANG Y Q. Joint extraction method of Chinese entity relationship based on mixture of characters and words[J]. Application Research of Computers, 2021, 38(9): 2619-2623. (in Chinese)

[83] LI S J, HE W, SHI Y B, et al. DuIE: A large-scale Chinese dataset for information extraction[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Dunhuang: Springer, 2019: 791-800.

[84] 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.

[85] BEKOULIS G, DELEU J, DEMEESTER T, et al. Adversarial training for multi-context joint entity and relation extraction[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 2830-2836.

[86] JI B, YU J, LI S S, et al. Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 88-99.

[87] DIXIT K, AL-ONAIZAN Y. Span-level model for rela-

tion extraction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 5308-5314.

[88] EBERTS M, ULGES A. Span-based joint entity and relation extraction with transformer pre-training[EB/OL]. (2019-09-17)[2022-10]. https://arxiv.org/abs/1909.07755.

[89] ZHAO T Y, YAN Z, CAO Y B, et al. Asking effective and diverse questions: A machine reading comprehension based framework for joint entity-relation extraction[C]// Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama: International Joint Conferences on Artificial Intelligence Organization, 2021: 3948-3954.

作者简介

张仰森 男,1962年6月出生于山西省运城市. 现为北京信息科技大学二级教授、博士生导师. 获教育部、北京市、山西省及中国中文信息学会科技进步奖等奖项7项. 在国内外发表学术论文200余篇. 主要研究方向自然语言处理与网络内容安全.
E-mail: zhangyangsen@163.com

刘帅康 男,1998年12月出生于河南省商丘市. 北京信息科技大学信息管理学院硕士研究生. 主要研究方向为自然语言处理与网络内容安全.
E-mail: 346266505@qq.com

刘 洋(通讯作者) 女,1983年7月出生于辽宁省沈阳市. 北京邮电大学计算机应用专业硕士. 高级工程师,曾获得中国通信学会科学技术一等奖,在国内外发表学术论文20余篇. 主要研究方向为信息安全、数据处理.
E-mail: lyang@cert.org.cn

任 乐 女,1999年12月出生于陕西省西安市. 北京信息科技大学信息管理学院硕士研究生. 主要研究方向为自然语言处理与网络内容安全.
E-mail: renlle2021@163.com

辛永辉 男,1990年2月出生于河南省漯河市. 中国科学院大学信号与信息处理专业博士. 中级工程师. 在国内外发表学术论文10余篇. 主要研究方向为信息安全、机器学习.
E-mail: xinyh@cert.org.cn