



车银超,董晨阳,熊淑萍,王顺,马新明,席磊.面向农药登记信息的实体关系联合抽取研究[J].中国农业大学学报,2025,30(05):214-229.
CHE Yinchao, DONG Chenyang, XIONG Shuping, WANG Shun, MA Xinming, XI Lei. Research on joint extraction of entities and relationships for pesticide registration information[J]. *Journal of China Agricultural University*, 2025, 30(05): 214-229.
DOI: 10.11841/j.issn.1007-4333.2025.05.18

面向农药登记信息的实体关系联合抽取研究

车银超^{1,2} 董晨阳¹ 熊淑萍^{2,3} 王顺^{1,2} 马新明^{2,3} 席磊^{1,2*}

(1. 河南农业大学 信息与管理科学学院, 郑州 450046;

2. 河南省农田环境监测与大数据分析工程研究中心, 郑州 450002;

3. 河南农业大学 农学院, 郑州 450046)

摘要 针对农药登记文本中信息密集、逻辑结构复杂、实体间跨度大以及实体长度异质性等特点,同时为克服传统联合抽取方法中面临的三元组重叠、曝光偏差和冗余计算问题,本研究提出一种多特征融合的单阶段实体关系联合抽取模型(Multi-feature fusion single-stage entity and relation joint extraction model, MF-SERel)。首先,在编码层,通过融合语义与句法特征,丰富字符向量表示,提升模型对复杂语料的表征能力;其次,在多维标注框架层,提出HT-BES多维标注策略,以解决重叠三元组问题。通过并行评分函数与细粒度分类组件,将实体关系联合抽取转化为了基于关系维度的多标签标注任务,该过程不包含相互依赖步骤,从而实现单阶段并行标注,避免了曝光偏差并降低了计算冗余;最后,在解码层依据细粒度分类预测标签,解码出实体关系三元组。将本研究提出的模型与GraphRel、CasRel和TPLinker等基线模型进行对比,在农药数据集(Pesticide registration dataset, PRD)和公开数据集(Dataset of unstructured information extraction, DuIE)上进行检测。结果表明MF-SERel模型在农药数据集PRD和公开数据集DuIE上具有良好的表现。在农药数据集PRD上,本研究提出的模型MF-SERel在推理速度上提升了20%,F1值提升了2.3%,说明MF-SERel模型在农药登记文本中具有良好的知识挖掘能力;在公开数据集DuIE上,MF-SERel模型在推理速度上提升了54%,F1值提升了1.7%,同样取得了较好结果,证明MF-SERel模型具有较好的泛化能力。综上,本研究提出的MF-SERel模型可为农药领域知识的结构化抽取提供新方法。

关键词 农药;实体关系抽取;标注策略;依存句法分析;图神经网络

中图分类号 TP391

文章编号 1007-4333(2025)05-0214-16

文献标志码 A

Research on joint extraction of entities and relationships for pesticide registration information

CHE Yinchao^{1,2}, DONG Chenyang¹, XIONG Shuping^{2,3}, WANG Shun^{1,2},
MA Xinming^{2,3}, XI Lei^{1,2*}

(1. College of Information and Management Sciences, Henan Agriculture University, Zhengzhou 450046, China;

2. Henan Engineering Research Center for Farmland Environmental Monitoring and Big Data Analysis, Zhengzhou 450002, China;

3. College of agronomy, Henan Agriculture University, Zhengzhou 450046, China)

Abstract In response to the characteristics of pesticide registration texts, such as information density, complex logical structure, large span between entities, and heterogeneous entity lengths, as well as to address the issues of triplet overlap, exposure bias, and redundant computation encountered in traditional

收稿日期: 2024-10-30

基金项目: 河南省重点研发专项项目(241111111500);国家自然科学基金项目(32372239)

第一作者: 车银超(ORCID:0009-0003-4246-4348),讲师,主要从事计算机应用技术研究,E-mail:cheyc@henau.edu.cn

通讯作者: 席磊(ORCID:0009-0001-9435-4189),教授,主要从事智慧农业研究,E-mail:xil@henau.edu.cn

joint extraction methods, a multi-feature fusion single-stage entity and relation joint extraction model (MF-SERel) is proposed in this study. Firstly, at the encoding layer, the model's ability to represent complex corpora is enhanced through integrating semantic and syntactic features and enriching the meaning of character symble. Secondly, at the multi-dimensional labeling framework layer, an HT-BES multi-dimensional labeling strategy is proposed to resolve the triplet overlap issue. Through parallel scoring functions and fine-grained classification components, the entity-relation joint extraction is transformed into a multi-label labeling task based on the relation dimension. This process does not involve interdependent steps, thereby enabling single-stage parallel labeling, avoiding exposure bias and reducing computational redundancy. Finally, at the decoding layer, labels are predicted based on fine-grained classification, and the entity-relation triplets are decoded. The proposed model is compared with the baseline models such as GraphRel, CasRel, and TPLinker, and evaluated on the pesticide registration dataset (PRD) and the public dataset of unstructured information extraction (DulE). The results show that the MF-SERel model performs well on the PRD and DulE. On the PRD, the proposed MF-SERel model improves inference speed by 20% and F1 score by 2.3%, demonstrating its strong knowledge-mining capability in pesticide registration texts. On the DulE, the MF-SERel model improves inference speed by 54% and F1 score by 1.7%, also achieves better resultsproves its strong generalization ability. In conclusion, the MF-SERel model proposed in this study provides a new method for the structured extraction of knowledge in the pesticide field.

Keywords pesticides; entity relationship extraction; annotation policies; dependency syntactic analysis; graph neural network

农药登记信息主要包括农药适用作物、防治对象、使用方法、使用限制等大量具有应用场景价值的信息,是农药知识库的重要组成部分^[1],这些信息与农药的化学和毒理学数据库^[2](如 Pesticide properties database, PPDB)等形成了有效互补,丰富了农药在实际应用中的数据维度,可为农业生产、农产品安全、农药监管与研发等领域提供重要的指导依据。由于农药登记信息包含大量的非结构化文本数据,使得这些信息在存储、检索和利用过程中变得困难复杂。农药实体是指具有独立意义和明确边界的具体对象或概念,通常是描述农药的具体信息、特性或构成部分;农药关系则是描述不同实体之间相互关联的语义信息^[3]。如何实现从非结构化文本中抽取实体关系,是自动化构建知识库的关键任务。农药实体关系抽取任务旨在发现非结构化农药登记数据中农药适用作物、防治对象、使用方法等相关实体之间的潜在关系,以生成结构化的实体关系三元组,为农药知识图谱构建、知识发现及问答应用提供关键性支撑^[4-5]。

通用语料通常涵盖了广泛的文本类型,包括新闻、文学、百科等内容,涉及多个领域的知识,其具有语言结构简单,且语境限制较少的特点。与通用语料相比,农药登记语料包含大量的专业术语,农药名称作为头实体与多个尾实体之间形成大量复杂重叠的三元组关系^[6];同时,由于文本通常较长,

造成实体间的跨度较大,削弱了上下文之间的依赖关系^[7],增加了抽取难度;此外,实体长度差异显著,最短实体仅为1个字,而最长实体达到9个字,实体长度的异质性增加了实体边界识别的难度。这些特性为从农药登记信息中自动抽取实体关系提出了多方面的挑战。

信息抽取的方法总体可分为基于生成预训练变换模型的方法和基于表示学习的方法。基于生成预训练变换模型的方法(如 ChatGPT、ChatGLM)将实体关系抽取视为文本生成任务,直接生成结构化实体关系三元组,尽管该类方法展现了强大的少样本学习能力,能够在不依赖大规模标注数据的情况下完成多种自然语言处理任务,但在解决复杂任务(如实体关系)时,因缺乏明确的推理过程和外部知识引导,难以实现推理平衡或解决幻觉问题^[8]。基于表示学习的方法在面对复杂的信息抽取任务时能够提供可解释的推理过程,并能更好地整合外部知识,因此,目前基于表示学习的方法仍是面对复杂信息抽取任务的主要方法。基于表示学习的方法又主要分为流水线法和联合抽取法两大类。流水线法^[9-12]将实体识别和关系抽取作为2个独立的子任务,子任务间存在交互缺失和误差累积的问题,后期一些研究提出联合抽取方法,实现了实体与关系的同步抽取^[13-31]。

联合抽取方法针对错误累积和子任务交互缺

失的问题进行了改进。如:Zheng等^[13]通过序列标注的解码策略实现了实体关系的联合抽取,但该方法受限于标签分配机制,无法有效处理关系重叠的现象;Chang等^[14]和Yuan等^[15]通过先抽取头实体,再基于头实体映射关系和尾实体,实现了实体关系的结构化抽取;Wei等^[16]和葛君伟等^[17]通过先抽取关系,再根据关系识别头尾实体,提高了模型的识别精度。尽管上述多阶段的联合抽取方法取得了良好的识别效果,但其实体识别结果与关系抽取结果之间相互依赖,早期步骤中的错误往往会传递至后续阶段,导致曝光偏差问题^[18]。为解决上述问题,已有研究提出了一系列单阶段的联合抽取方法,如基于全局矩阵和潜在关系矩阵的模型(PRGC)^[19]、表格填充模型(OneRel)^[20]、实体关系并行检测模型(PRE-Span)^[21]、基于握手标记方案的模型(TPLinker)^[22]等。尽管这些方法通过并行识别实体关系,减少了步骤间的相互依赖,降低了曝光偏差的风险,但因实体与关系识别模块的分离,在组合判断时仍存在冗余计算^[23]。此外,这些方法多依赖单一语义信息,难以全面解析文本的复杂逻辑结构。

随着句法信息与图神经网络的引入,模型在复杂文本特征解析和实体关系抽取任务中取得了显著进展。FU等^[24]提出一种多阶段的GraphRel模型,利用图神经网络融合句法结构信息,通过聚合邻居节点特征以增强文本表示,提高了重叠三元组的识别效果。苗琳等^[25]结合图注意力网络和语义依存分析技术,进一步增强了模型对复杂文本的理解和抽取质量。翟社平等^[26]提出了一种结合对抗训练、图注意力网络和依存分析的联合抽取模型,通过序列结构与单词语义特征的融合,改善了三元组重叠的问题。上述方法通过引入句法信息,增强了模型对复杂文本的表征能力,提升了模型的抽取精度以及对重叠三元组的处理能力,但因实体关系识别过程中相互依赖,仍面临着曝光偏差的挑战。

近年来,已有研究将实体关系联合抽取技术应用于农业领域提高了对农业知识的利用效率。吕东东等^[27]针对农产品标准数据分散问题,提出基于依存句法分析的联合抽取模型,通过实体关系抽取范式,实现知识自动化抽取。周俊等^[28]针对水稻施肥语料三元组重叠问题,采用RoBERTa编码和改进的CasRel模型,构建多阶段级联抽取框架,完成水稻施肥信息的整合。Jia等^[29]针对小麦种质信息

存在的关系重叠问题,提出了多阶段的深度字词融合模型,提高重叠关系的提取能力。吴赛赛^[30]针对小麦病虫害领域三元组重叠、知识共享困难等问题,设计了一种实体关系同步标注方案,有效解决单实体重叠问题。姜丽华等^[31]针对水产病害领域语料三元组重叠问题提出了一种基于BERT字嵌入的联合抽取模型,将实体关系抽取转化为序列标注任务,有效解决了单实体重叠的问题。上述研究针对农业领域不同语料的挑战,设计了多阶段的联合抽取模型,增强了数据特征提取和重叠三元组的处理能力。然而,多阶段的抽取方式和复杂标注策略导致了曝光偏差和冗余计算等问题。

现有的实体关系抽取研究多集中在通用领域和大规模知识图谱构建,在农药领域仍缺乏知识抽取的相关研究。因此,为应对农药登记文本中信息密集、逻辑结构复杂、实体间跨度大以及实体长度异质性所导致的信息抽取复杂问题,同时为克服联合抽取方法中面临的三元组重叠、曝光偏差和冗余计算问题,本研究拟提出一种多特征融合的单阶段实体关系联合抽取模型MF-SERel,并与现行常用的GraphRel、CasRel以及TPLinker方法进行了对比,在农药登记数据集(Pesticide registration dataset, PRD)和公共数据集(Dataset of unstructured information extraction, DuIE)^[32]上进行检测和验证,以期农药领域知识的结构化整合提供一种更准确、更高效的实体关系联合抽取方法。

1 试验材料与统计分析

1.1 农药登记数据获取

为保证数据来源的准确性和权威性,农药登记数据语料主要从世纪农药网(<http://www.nongyao001.com>)、中国农药信息网(<http://www.chinapesticide.org.cn>)和农业农村部种植业管理司(农药管理司)(<http://www.zzys.moa.gov.cn>)等3个网站进行获取。首先使用Scrapy爬虫框架从以上3个网站获取原始数据集。其次,对收集的非结构化数据以人工方式进行数据清洗,删除数据中的空格、空行和特殊字符等内容。消除句子中的主语指代问题。对长句子,根据产品性能、使用技术要求、注意事项、存储运输方法和急救措施等内容进行分类,并去除重复数据和无效数据。最终构建了包含适用作物、杂草防治、药剂类型等10种关系的1642条长句语料,字符数约为25万。

1.2 农药登记数据集的构建及统计分析

1.2.1 农药登记数据集的构建

本研究按照农药的产品性能、使用技术要求、注意事项、存储运输方法和中毒急救措施将语料分

为5类。在领域专家的指导下,依据农业农村部农药管理司发布的农药属性信息类型,最终将农药实体关系细粒度划分为11类实体和10类关系,详见表1。

表1 实体关系类别实例

Table 1 Examples of entity relation categories

头实体 Head entity	尾实体 Tail entity	关系 Relationship
草甘膦铵盐 Glyphosate ammonium salt	柑橘	适用作物
2-甲氯氟吡 2-Methylchloropyridine	鸭舌草	杂草防治
吡嘧磺隆 Pyrimisulfan	可湿性粉剂	药剂类型
乙氧氟草醚 Ethoxyfluoroacetate	除草剂	农药类别
二甲戊灵 Dimethenamid	天研施倍得	农药品牌
仲丁灵 Butralin	误服	中毒方式
噁草酮 Bromoxynil	呕吐	中毒症状
苄嘧磺隆 Bensulfuron-methyl	40~50 kg	亩用水量
氰氟草酯 Cyhalofop-butyl	手套	防护用品
灭草松 Acifluorfen	干燥	储存方式

对于实体关系的标注,在领域专家的指导下,基于标注指南使用标注工具 Doccano 对本研究 1 642 条语料进行3轮标注,并由专家进行结果检验。最终,生

成农药数据集(Pesticide registration dataset, PRD),其中包含 11 361 个实体,9 610 个关系,各个实体类型和关系类型的数量分布如图1所示。

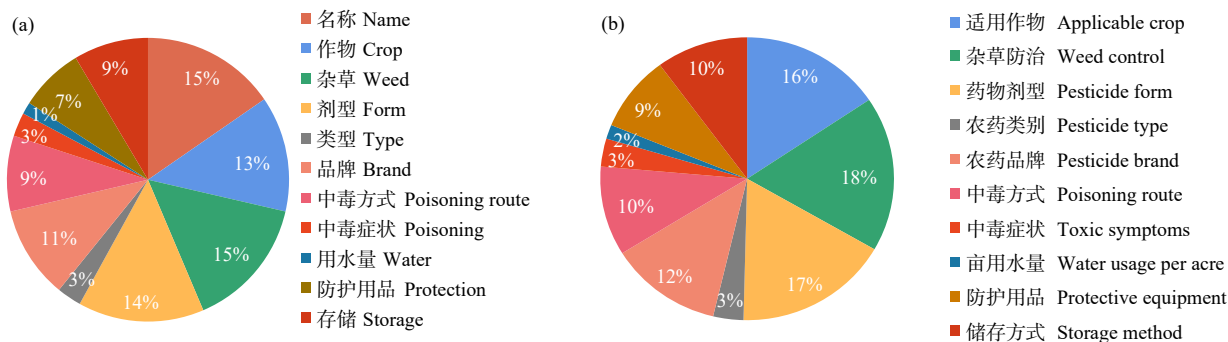


图1 不同实体(a)和关系(b)的所占比例分布情况

Fig. 1 Distribution of the proportions of different entities (a) and relationships (b)

1.2.2 实体长度统计分析

对自建数据集 PRD 中实体的长度进行统计分析,得到具体实体长度分布如图2所示。可知实体长度大多集中在2~5个字。其中:实体的最短长度为1个字,最长为9个字,实体间的长度存在差异。

2 模型构建

本研究提出的 MF-SERel 模型由编码层、多维

标注框架层和解码层三部分组成,其整体流程如下:首先,语料输入编码层,通过 RoBERTa 和多头自注意力机制提取每个字符的语义特征,同时利用图卷积网络(GCN)捕获句法特征,融合语义与句法信息生成字符的最终词向量表示;然后,将生成的词向量输入多维标注框架层,通过评分函数完成单阶段的并行评分,生成基于关系维度的评分矩阵,并依据该矩阵进行细粒度分类,得到关系标签分类

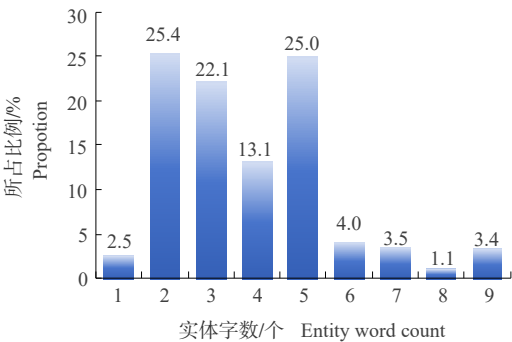


图2 实体不同长度所占比例分布情况

Fig. 2 The proportional distribution of entities of different lengths

矩阵,实现单阶段标签映射;最后,解码层依据关系标签分类矩阵,根据头尾实体关联分类标记,解码得到最终的三元组识别结果。本研究提出模型的框架如图3所示。

2.1 融合多特征的编码层

农药登记文本具有信息密集、逻辑结构复杂、实体跨度较大和实体长度异质性等特点,仅依赖预训练语言模型难以全面应对农药登记信息中的复杂性。因此,该模块结合RoBERTa预训练模型和多头自注意力机制,以增强对语义特征的捕捉能力;同时,利用依存句法分析和图卷积网络(GCN)学习句法结构特征,以提升对文本结构的深度理解。最终,将图卷积神经网络输出与

多头自注意力输出拼接得到每个字符最终特征向量,通过融合语义特征和句法特征丰富文本的特征表示,从而提升模型在复杂文本环境下对实体关系的识别能力。

2.1.1 基于RoBERTa的语义特征编码

RoBERTa是基于BERT进行改进的一种预训练语言模型。与BERT相比,RoBERTa在预训练过程中采用了更大规模的数据集和模型参数,并引入了动态掩码机制,在多个自然语言处理任务上表现出更强的性能和更好的泛化能力。对于给定的农药文本序列 $S=\{w_1,w_2,\dots,w_L\}$,序列 S 经过字符嵌入、分割嵌入和位置嵌入编码并进行融合得到最终的融合嵌入向量。将 S 的所有融合嵌入向量构成的序列送入RoBERTa模型,提取上下文语义信息,并生成农药文本序列的语义表示向量 $E=[e_1,e_2,\dots,e_L]$,计算公式如下:

$$[e_1,e_2,\dots,e_L]=\text{RoBERTa}[w_1,w_2,\dots,w_L]$$
 (1)

式中: e 为字符的特征向量; w 为文本中的字符; L 为序列的长度。

2.1.2 多头自注意力机制

多头自注意力机制能够捕获句子中字符之间的上下文依赖关系,根据字符在句内的重要程度,动态分配权重,提升关键信息的权重。本研究将RoBERTa的输出 E 作为多头自注意力计算的输入,式中的 Q 、 K 、 V 由上一层输出的向量分别与权重矩阵 W^Q 、 W^K 、 W^V 相乘所得,多头自注意力结构见图4。

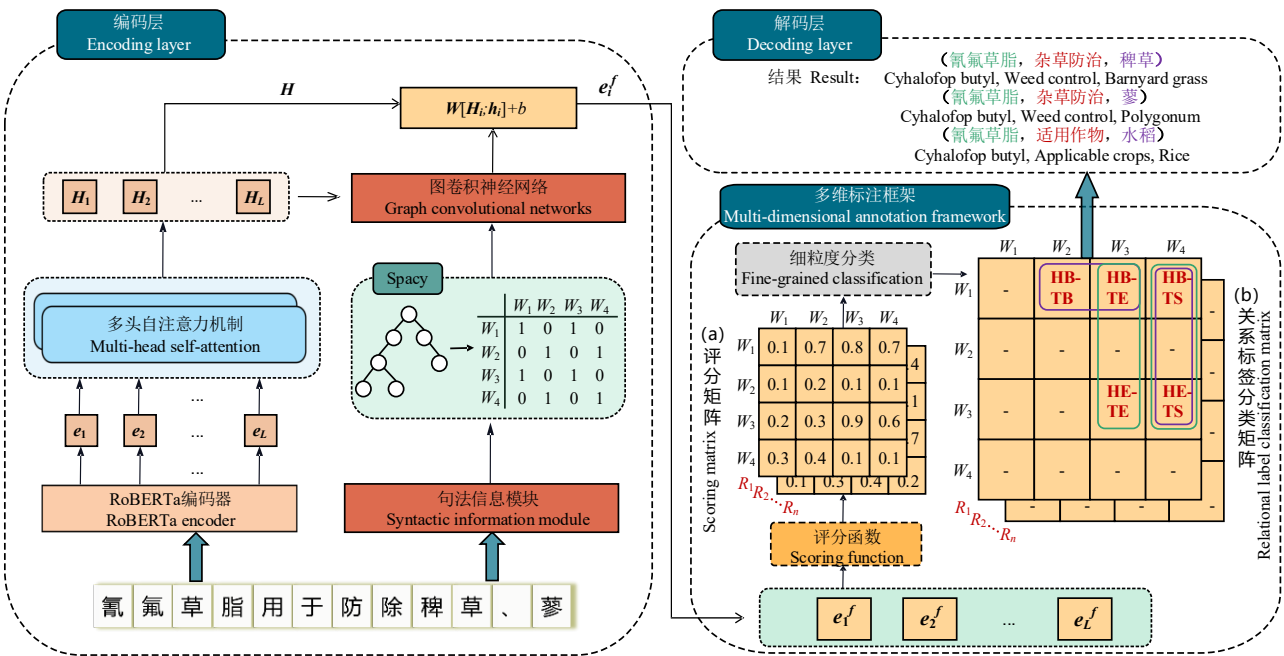


图3 多特征融合的单阶段实体关系联合抽取模型

Fig. 3 A single-stage entity relationship joint extraction model based on multi-feature fusion

$$H = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

式中： H 为每个字符在 Self-attention 层输出向量；

Softmax 为归一化表示，用于得到句中各字符间的相关性程度； QK^T 为字符与字符之间的相关性程度； d_k 为 Q 和 K 的向量维度； V 为值向量，用于计算最终结果。

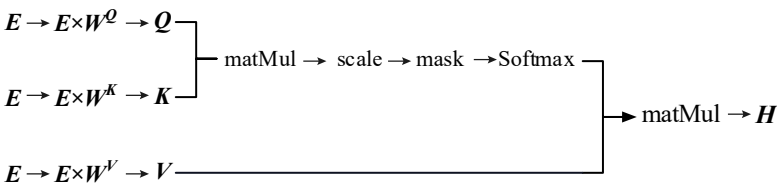


图 4 多头自注意力机制结构

Fig. 4 Structure of multi-headed self-attention mechanism

2.1.3 句法信息模块

本研究利用依存句法分析技术来深入解析句子结构。首先，通过构建句法依存树并利用依存矩阵进行存储，实现对句子结构的分析；再采用图卷积网络(GCN)来提取图上相邻节点之间的联系，以获取融合句法结构的文本特征表示。

句法依存分析。SpaCy 是一个由 Python 编写的开源自然语言处理库，广泛应用于分词、词性标注及依存句法解析等任务。为实现对文本的高效

处理，我们采用了 SpaCy 提供的 zh_core_web_sm 中文语言模型，对每条文本进行分词，并生成相应的依存句法树^[33]。对输入长度为 L 的句子 $S = \{w_1, w_2, \dots, w_L\}$ ，可以构建一个 $L \times L$ 邻接矩阵 A ，若两个词之间具有依存关系，则邻接关系置为“1”，否则置为“0”。此外为节点之间添加无向边和自邻接边，加强单词自身与相互之间的联系。例如，对句子“氰氟草酯用于防除稗草等禾本科杂草”进行依存分析结果如图 5 所示。

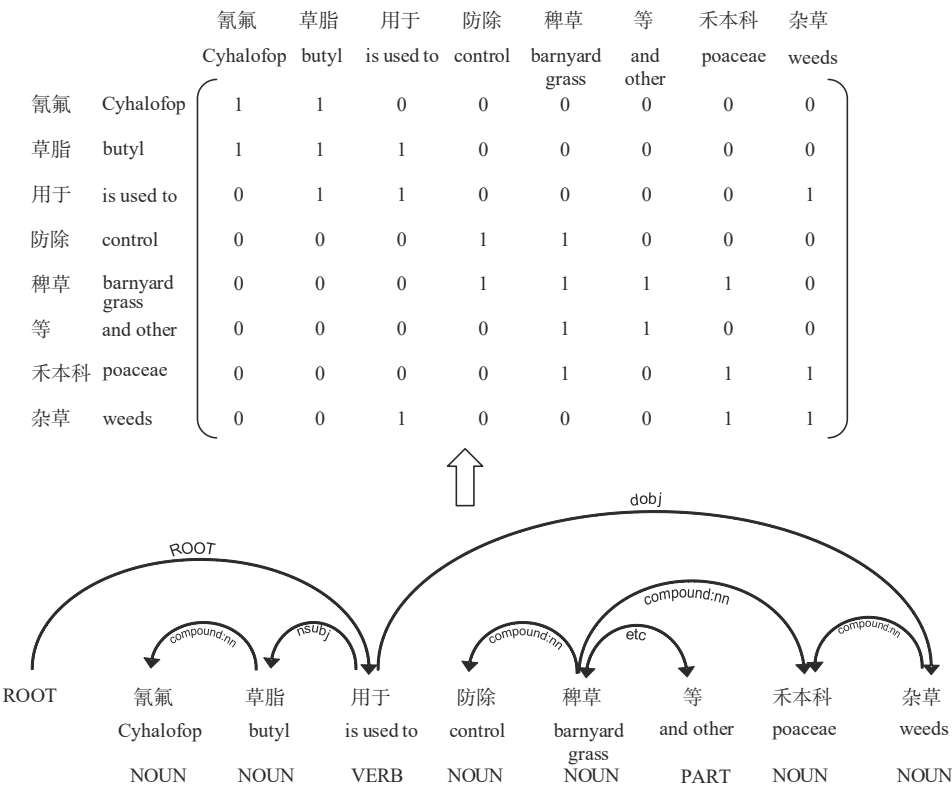


图 5 依存句法信息

Fig. 5 Dependent syntactic information

GCN图结构特征提取。通过句法依存结构图的构建,可以有效忽略单词之间的冗余信息,融合较远相关词之间的信息特征,以有效解决实体间跨度较大的问题。通常,一个图结构由节点集合和边集合构成,每个单词都相当于一个节点,而边集合由邻接矩阵中的词间依赖关系构成。对于邻接图中的每个节点,GCN通过聚合其邻居节点的特性来更新相应的嵌入节点。

在 N 层GCN中,以第 i 个节点的邻居 j 在第 $(n-1)$ 层的隐层向量 $h_j^{(n-1)}$ 作为输入。当 $n=1$ 时, h_i^0 则表示进行图卷积操作前第 i 个节点的初始隐层向量,即经过自注意力层输出的节点向量 H ,一个图卷积操作如下:

$$h_i^{(n-1)} = \sigma \left(\sum_{j=1}^{\tilde{n}} A_{ij} W^{(n)} h_j^{n-1} + b^{(n)} \right) \quad (3)$$

式(3)中: σ 为激活函数Relu; A_{ij} 为节点 i 与节点 j 之间的依赖关系; $W^{(n)}$ 为线性转换; $b^{(n)}$ 为偏置项。各节点在每一次进行图卷积计算时,汇集其相邻节点的特征信息。

最终,将图卷积神经网络输出 h_i 与多头自注意力输出 H_i 进行拼接,得到最终融合语义信息和句法信息的特征向量 e_i^f ,计算公式如下:

$$e_i^f = W_f [H_i; h_i] + b_f \quad (4)$$

式中: W_f 为线性转换矩阵; b_f 为偏置项。

2.2 多维标注框架层

2.2.1 HT-BES多维标注策略

HT-BES多维标注策略将实体关系联合抽取任务转化为基于关系维度的多标签标注任务来实现。依据头尾实体关联的思想,标注策略具体设计如下:

BES实体边界标记。其中,“B”表示实体的开始字符,“E”表示实体的结尾字符,“S”表示单字符实体。实体边界信息能够适应不同实体的长度变化,准确识别实体边界。

HT实体位置标记。使用“H”“T”对三元组中实体的位置进行标记,“H”代表头实体、“T”代表尾实体。位置信息可以精确地识别出头尾实体间的相对位置,对于处理实体间距较大的情况尤为重要。

头尾实体关联分类标记。为了反映头尾实体字符对间的关联,融合实体边界与位置标记,对头尾实体标记进行连接,标签种类及意义如表2所示,其中非实体标签被置为“-”。通过头尾实体关联标记反映了头尾实体的依赖性,可以准确地识别出三元组中的头尾实体。

表2 本研究标签种类

Table 2 Types of tags in this study

标记 Tag	描述 Description	标记 Tag	描述 Description
HB-TB	头实体开始-尾实体开始	HB-TS	头实体开始-尾单字实体
HB-TE	头实体开始-尾实体结尾	HE-TS	头实体结尾-尾单字实体
HE-TE	头实体结尾-尾实体结尾	HS-TB	头单字实体-尾实体开始
HS-TS	头单字实体-尾单字实体	HS-TE	头单字实体-尾实体结尾

关系标签分类矩阵。为解决三元组重叠问题,为每种关系构建一个独立的标签矩阵,矩阵中的每个元素表示特定关系下的字符对,若字符对之间存在关系,则在相应位置按照头尾实体关联分类标记进行标注。通过这种方式,实现对所有实体关系的整体化处理。

图6为HT-BES多维标注策略应用示例,图6中(a)矩阵包含2个三元组:<氟氟草脂,杂草防治,稗草>和<氟氟草脂,杂草防治,蓼>。对于第1个三元组,“氟”和“脂”分别是头实体“氟氟草脂”的开

始字符和结尾字符,“稗”和“草”分别是尾实体的开始字符和结尾字符。因此,依据头尾实体关联分类标记,(<“氟”,杂草防治,“稗”>)被指定为“HB-TB”、(<“氟”,杂草防治,“草”>)被指定为“HB-TE”、(<“脂”,杂草防治,“草”>)被指定为“HE-TE”。对于第2个三元组,由于“蓼”是尾实体且为单字符实体,依据头尾实体关联分类标记,(<“氟”,杂草防治,“蓼”>)被指定为“HB-TS”、(<“脂”,杂草防治,“蓼”>)被指定为“HE-TS”。

三元组<氟氟草脂,杂草防治,稗草>与<氟

氟草脂,杂草防治,蓼>共享“杂草防治”关系,因此分类标签存储在相同的“杂草防治”关系矩阵中。而图6中(b)矩阵的三元组<氟氟草脂,适用作物,水稻>

水稻>对应“适用作物”关系,标签应存储在“适用作物”矩阵中。不同关系分配到各自矩阵,有效解决了三元组重叠问题。

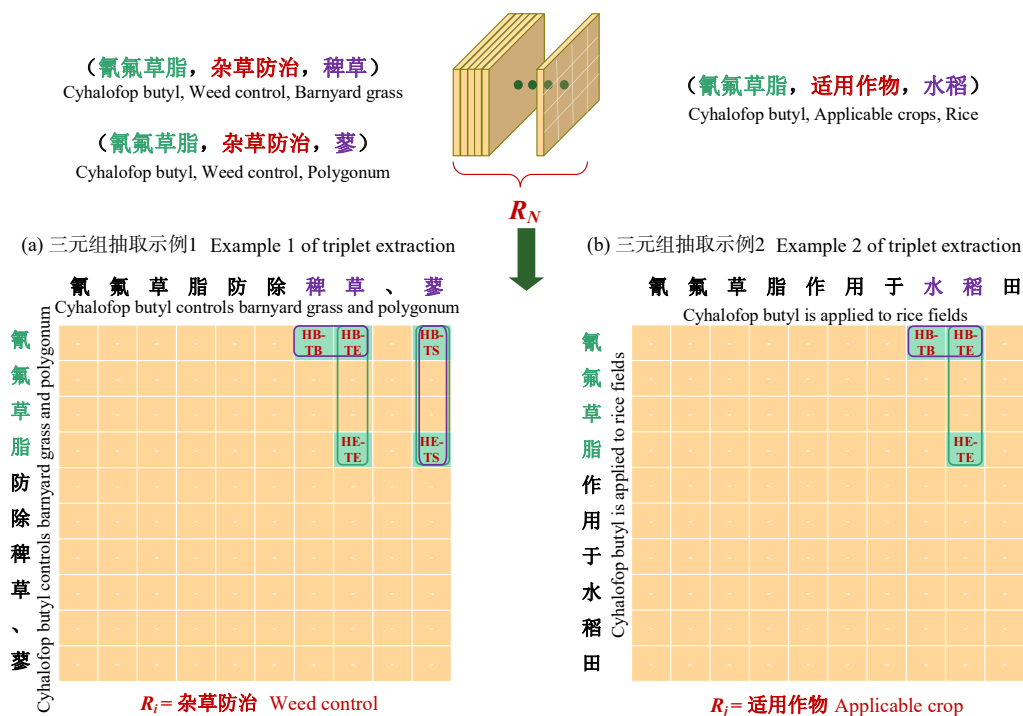


图6 HT-BES标注策略

Fig. 6 HT-BES annotation strategy

2.2.2 HT-BES 多维标注策略的实现

HT-BES 标注策略实现流程如下:首先,对输入句子中的所有字符对 (w_i, w_j) 及预定义关系 (r_k) 进行组合枚举;其次,通过评分函数组件计算每个组合 (w_i, r_k, w_j) 构成三元组的潜在得分,生成评分矩阵;最后,利用该评分矩阵通过细粒度分类组件,并依据头尾实体关联分类标记,生成关系标签分类矩阵。

1) 评分函数组件。

本研究基于知识图谱嵌入技术 HOLE 模型^[34]设计评分函数。该函数是一个可学习的映射函数,通过非线性的拼接投影,以评估 2 个字符 (w_i, w_j) ,通过某个关系 r_k 连接的可能性。评分函数公式如下:

$$v_{(w_i, r_k, w_j)} = R^T \left(\sigma \left(\text{drop} \left(W [e_i^f, e_j^f]^T + b \right) \right) \right) \quad (5)$$

式中: v 为得分向量; K 为预定义关系的数量; R 为关系向量表示矩阵,视为可训练的权重; σ 为激活函数 LeakyRelu; $\text{drop}(\cdot)$ 用于防止过拟合; W 为可

训练的权重矩阵; $[e_i^f, e_j^f]$ 为字符 i 和 j 的向量连接运算; b 为偏差项;矩阵 R 与 W 通过自适应学习,可以根据输入的字符特征动态调整评分函数,以便更好地评价字符对于预定义关系的依赖性。

通过评分函数生成的评分矩阵如图3中(a)矩阵所示,该函数通过2层全连接网络实现并行评分,从而将枚举 (w_i, r_k, w_j) 组合的评分计算复杂度从 $L \times K \times L$ 降低至 $L \times 1 \times L$,有效减少了冗余计算,提高了计算效率。

2) 细粒度分类组件。

细粒度分类的目标是依据评分矩阵,为预定义关系 r_k 下的字符对 (w_i, w_j) 分配高置信度标签,生成关系标签分类矩阵,分类结果如图3中(b)矩阵所示。计算公式如下:

$$P(y_{(w_i, r_k, w_j)} | S) = \text{Softmax}(v_{(w_i, r_k, w_j)}) \quad (6)$$

该过程通过 softmax 函数对得分向量 v 进行处理,将原始得分转换为概率分布,并依据 HT-BES 标注策略中的头尾实体关联分类中的标签种类,为

(w_i, w_j) 进行细粒度分类预测。

2.3 解码层

解码层的目标是依据头尾实体关联分类标记确定头实体和尾实体的边界,从关系标签分类矩阵中解码出实体关系三元组。实体关系三元组解码算法如下:

输入:原始文本 T ,预测的关系标签分类矩阵 M ,标签映射 TagId

输出:实体关系三元组集合 O

第1步:初始化实体关系三元组集合 O

第2步: relations, heads, tails = np. where ($M > 0$)

第3步: for $i \leftarrow 1$ to len(relations) do

第4步: $r_i \leftarrow \text{relations}[i]$, $h_b \leftarrow \text{heads}[i]$, $t_b \leftarrow \text{tails}[i]$ //获取当前关系 i 和实体的开始位置索引

第5步: if $M[r_i][h_b][t_b] == \text{TagId}['\text{HB-TB}']$ then //检查当前是否为 r_i 关系下头尾实体的开始标签 (HB-TB)

第6步: $t_e \leftarrow \text{tails}[i+1]$

第7步: if $M[r_i][h_b][t_e] == \text{TagId}['\text{HB-TE}']$ then

第8步: Find $M[r_i][h_b][t_e]$ (HE-TE) and add (H, R, T) to O // 检查当前位置是否是 r_i 关系下头尾实体的结束标签 (HE-TE),并根据获取到的标签位置信息提取三元组并添加到集合 O

第9步: if $M[r_i][h_b][t_b] == \text{TagId}['\text{HB-TS}']$ then

第10步: Find $M[r_i][h_e][t_e]$ (HE-TS) and add (H, R, T) to O

第11步: if $M[r_i][h_b][t_b] == \text{TagId}['\text{HS-}$

TB'] then

第12步: Find $M[r_i][h_e][t_e]$ (HS-TE) and add (H, R, T) to O

第13步: return O

实体关系三元组提取算法,首先根据细粒度分类预测的关系标签分类矩阵,定位可能存在的实体关系坐标。对于每个关系矩阵,如果当前矩阵单元格标识了头尾实体的开始标签,则向后搜索对应的实体结束标签,根据头尾实体的起止坐标 h_b, h_e 和 t_b, t_e ,即可确定头实体和尾实体的完整范围。随后,将提取的头实体、关系及尾实体组成三元组,并将其添加至集合 O ,最终生成并返回包含所有实体关系的三元组集合。

以图4中关系矩阵“杂草防治”为例,通过“HB-TE”与“HE-TE”的组合,观察“HB”和“HE”即可以识别出头实体“氟氟草脂”,通过“HB-TB”与“HB-TE”的组合,观察“TB”和“TE”可以识别出尾实体“稗草”。同样地,通过“HB-TS”和“HE-TS”标签的组合,可以解码出头实体“氟氟草脂”和尾单字实体“蓼”。基于此解码过程,最终提取出三元组(氟氟草脂,杂草防治,稗草)和(氟氟草脂,杂草防治,蓼)。

2.4 损失函数

为了有效训练 MF-SERel 模型并实现精确的实体关系联合抽取,本研究采用交叉熵损失作为优化目标。交叉熵损失函数在多标签分类任务中具有显著优势,能够有效衡量模型预测的类别分布与实际标签分布之间的差异。鉴于句子长度 L 和关系数量 K 的变化可能导致损失值的不稳定性,本研究在计算过程中引入了归一化因子,以对损失进行标准化处理。损失函数计算公式如下:

$$\text{Loss}_{\text{triples}} = -\frac{1}{L \times K \times L} \times \sum_{i=1}^L \sum_{k=1}^K \sum_{j=1}^L \log P\left(y_{(w_i, r_k, w_j)} = g(w_i, r_k, w_j) | S\right) \quad (7)$$

式中: $y(w_i, r_k, w_j)$ 为模型预测的分类标签; $g(w_i, r_k, w_j)$ 为真实分类标签; S 为输入的句子。

2.5 试验设置

2.5.1 试验环境和模型参数设置

本试验的环境定义: Windows10 操作系统, Python3.8, Pytorch 1.8.1, CUDA11.1, GPU 为 NVIDIA Tesla V100 16GB。本研究模型参数定义如下:预训练语言模型采用 RoBERTa-wwm-ext 版本,包括 12 层 Transformer 和 768 维隐藏层,使用 Adam 优化器,多头自注意力机制 Head 数定义为 3,

批尺寸 Batch_size 设置为 4,学习率 lr 定义为 $1e-5$, Dropout 定义为 0.2,最大序列长度定义为 200,图神经网络的隐藏层数设置为 768,激活函数为 Relu, SpaCy 中文模型 batch_size 设置为 256, Dropout 定义为 0.1,最大训练步数 max_steps 设置为 100 000。

2.5.2 评价指标

精确率 (P , Precision)、召回率 (R , Recall) 和 F1 值 ($F1\text{-value}$) 作为评价指标对本研究模型进行评估。其中:精确率 P 指被识别正确的样本所占比例;召回率 R 指所有样本中被识别出的个数; $F1$ 则代表

前两者的综合得分。3种评价指标均为值越高,性能越好。相关计算公式如下:

$$P = \frac{TP}{TP + FP} \quad (8)$$

$$R = \frac{TP}{TP + FN} \quad (9)$$

$$F1 = \frac{2PR}{P + R} \quad (10)$$

式中:TP表示预测正确的正样本个数;FP表示预测错误的正样本个数;FN表示预测错误的负样本个数。

3 结果与分析

3.1 模型在自建农药数据集PRD上的表现

为验证模型MF-SERel在农药数据集PRD上的实体关系抽取性能,选取近几年的SOTA模型GraphRel_{1p}、GraphRel_{2p}、CasRel、TPLinker作为对比模型进行试验,结果如表3所示。可见MF-SERel模型的各项指标均优于GraphRel、CasRel和TPLinker。其中:1)相较于GraphRel_{2p}模型,MF-SERel的F1提升23.3%,这是因为多阶段的GraphRel先抽取关系,再根据关系抽取头尾实体,

存在着曝光偏差的问题。2)与CasRel模型相比,MF-SERel的召回率和F1值分别提升了5.7%和3.4%,一方面,MF-SERel通过单阶段的联合抽取策略,有效避免CasRel存在的曝光偏差问题;另一方面,本研究在CasRel编码器BERT的基础上,使用了其变体RoBERTa,并通过融合语义特征和句法特征,丰富了对语料的表征能力,从而提升了三元组的抽取性能。3)与TPLinker相比,MF-SERel的F1值提升2.3%,这是由于TPLinker对实体长度的泛化能力较弱,无法解决在农药登记数据中存在的实体长度异质性问题,且其提取的文本特征单一,无法很好地捕捉长距离依赖关系,进而造成模型效果较差。本研究提出的HT-BES多维度标注策略,通过检测实体边界标记和位置标记,可以对各类三元组进行有效识别,同时在编码层引入注意力机制和句法结构信息,增强了特征表达能力以及远距离实体的捕获能力,进而提高了模型的识别精度。因此,本研究提出的MF-SERel模型可以有效地挖掘更深层次的语义特征和句法特征,具有了更强的表达能力,在农药领域文本信息密集、逻辑结构复杂的情况下,展现了良好的效果。

表3 不同模型农药数据集上的模型性能

Table 3 Performance of different models on pesticide datasets %

模型 Model	精确率 Precision	召回率 Recall	F1值 F1 value
GraphRel _{1p}	64.0	75.3	69.2
GraphRel _{2p}	68.8	78.3	73.3
CasRel	97.7	89.1	93.2
TPLinker	94.9	93.6	94.3
MF-SERel	98.4	94.8	96.6

3.2 MF-SERel对各类关系的抽取效果

MF-SERel模型对各类农药实体关系抽取的精确率、召回率和F1值如表4所示。可见:本研究模型针对各类关系的识别效果均较好,各关系类别的评价指标较为均衡,F1值均达到了90%以上。其中:“防护用品”和“适用作物”的识别效果相对较低,F1值分别为90.6%和90.8%;经过分析农药数据集发现,这一现象主要是由于“防护用品”类实体存在复杂的嵌套结构,例如“长衣”、“长裤”、“帽子”、“长衣长裤”及“长衣长裤帽子”等多种组合形

式,增加了模型的识别难度。“适用作物”类实体的多样化表述,如“小麦”“麦田”以及“小麦田”等,也导致了识别的复杂性;相比之下,“杂草防治”关系的识别效果最佳,F1值达到了97.1%。对农药数据集分析显示,杂草类实体通常密集出现,且各实体间以“、”符号分隔,上下文信息较为固定,使得模型能够更容易地进行识别和分类。

3.3 消融试验结果

为评估MF-SERel模型在获取语义和句法信息方面的表现,以及注意力机制对试验结果的影响,

表4 不同关系类别的抽取效果

Table 4 Extraction effects of different relationship categories			%
关系 Relation	精确率 Precision	召回率 Recall	F1值 F1 value
适用作物 Applicable crops	95.4	86.6	90.8
杂草防治 Weed control	98.3	96.0	97.1
药物剂型 Pesticide form	98.3	96.0	97.0
农药类别 Pesticide type	96.2	96.0	96.0
农药品牌 Pesticide brand	95.0	94.2	94.6
中毒方式 Poisoning way	96.3	91.0	93.6
中毒症状 Toxic symptoms	98.3	95.2	96.7
亩用水量 Water usage per acre	93.5	90.6	92.1
防护用品 Protective equipment	97.4	84.7	90.6
储存方式 Storage method	99.0	95.0	97.0

本研究在农药数据集 PRD 上对模型各模块进行了消融试验,结果如表 5 所示。可见:当 RoBERTa 被 BERT 替换时,模型的 F1 值下降了 3.3%,反映出 RoBERTa 由于其更大规模的训练数据和动态掩码策略,在处理农药领域的专业术语和长距离依赖关系时具有更优越的表现;在未采用多头自注意力机制的情况下,F1 值降低了 1.1%,表明多头自注意力机制在捕捉全局上下文信息方面起到了关键作用,

避免了在编码过程中引入冗余信息,从而提升了模型的整体性能;当未使用 GCN 进行句法信息提取时,召回率降低了 1.8%,F1 值下降了 1.2%,这表明仅依赖语义信息不足以充分挖掘实体的隐藏特征,而语义信息与句法结构信息的融合有助于提升模型的特征捕获能力,从而达到更好的识别效果。试验结果验证了各模块的引入对模型性能提升的重要性。

表5 去除不同模块的消融试验

Table 5 Ablation experiment with different modules removed			%
模型 Model	精确率 Precision	召回率 Recall	F1值 F1 value
-RoBERTa	94.8	91.9	93.3
-GCN	98.0	93.0	95.4
-ATT	97.1	94.0	95.5
MF-SERel	98.4	94.8	96.6

3.4 模型在公开数据集 DuIE 上的泛化试验效果

为验证 MF-SERel 模型在公开数据集上的泛化性能,本研究选择百度公司发布的中文信息抽取数据集 DuIE 进行试验,并与对应的 SOTA 模型 GraphRel_{1p}、GraphRel_{2p}、CasRel、TPLinker 进行比较,结果如表 6 所示。可见:MF-SERel 模型在精确率和 F1 值上均取得了最佳表现,相较于 TPLinker 模型,分别提升了 3.4% 和 1.7%。尽管 TPLinker 模型在召回率方面略有优势,但其较低的精确率和 F1 值表明其在正类别的预测上存在一定的精确度

不足。相比之下,MF-SERel 模型在提升精确率的同时,保持了与 TPLinker 模型相近的召回率,并在 F1 值上取得了更好的表现,反映出 MF-SERel 在整体预测性能上的优势。此外,MF-SERel 模型的 F1 值相较于多阶段的 GraphRel_{1p}、GraphRel_{2p} 和 CasRel 模型分别提升了 24.7%、17.7% 和 6.1%,进一步证明了单阶段的 MF-SERel 模型可以有效避免曝光偏差问题,不仅在农药数据集上表现出色,而且在公开数据集条件下也能维持较高的性能,具有良好的泛化能力。

表 6 在 DuIE 数据集上的泛化试验结果

Table 6 Generalization experiment results on DuIE %			
模型 Model	精确率 Precision	召回率 Recall	F1 值 F1 value
GraphRel _{1p}	40.2	49.0	44.2
GraphRel _{2p}	49.9	52.6	51.2
CasRel	63.1	62.5	62.8
TPLinker	62.2	73.0	67.2
MF-SERel	65.6	72.6	68.9

3.5 模型对重叠三元组的抽取效果 F1 值分析

为评估 MF-SERel 模型在提取关系重叠三元组方面的能力,本研究基于 DuIE 数据集,对 Normal、SEO (单实体重叠)和 EPO (实体对重叠)3 种不同重叠程度的数据集 F1 值进行了试验,结果如图 7 所示。由 F1 值的大小可知:在不同复杂程度的关系重叠数据集中,GraphRel、CasRel 等模型的抽取效果均有所下降,而 MF-SERel 模型依然保持了较好的性能;在复杂的 SEO

和 EPO 数据集上,模型效果得到了进一步提升。这种性能提升得益于本研究提出的 HT-BES 多维度标注策略。具体而言,标注策略为每一种关系构建了独立的关系子矩阵,在解码过程中,不同的关系被映射到各自的矩阵中,减少了重叠关系带来的混淆,充分体现了模型在复杂语境下的优势。综上所述,MF-SERel 模型具有较强的泛化性,能够适应各种复杂的语境,在关系重叠三元组的提取任务中展现出较好的表现。

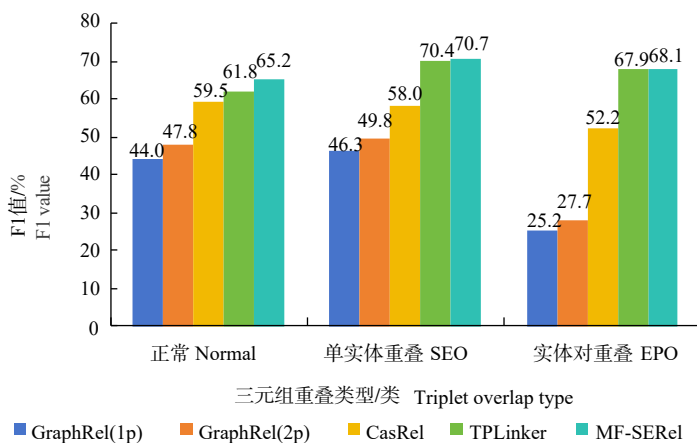


图 7 不同三元组重叠模式下的 F1 值比较

Fig. 7 Comparison under different triple overlap patterns of F1 scores

3.6 模型计算效率分析

为验证模型在解决冗余计算问题上的有效性,本研究使用农药数据集 PRD 和 DuIE 数据集,设置批尺寸为 4,最大句长为 200,对 MF-SERel 和对比模型的训练时间以及推理时间进行评估,结果如表 7 所示。可见:在农药数据集 PRD 以及公开数据集 DuIE 上,MF-SERel 的效率和准确性方面都展现了较好的效果。相较于 CasRel,MF-SERel 在 PRD 数据集上的推理速度提升了 20%。在 DuIE 数据集上,这一优势进一步扩大,推理速度提升了 54%,这

是因为在 DuIE 数据集中,关系种类较多,CasRel 分阶段先进行头实体的抽取,其次再遍历所有关系来抽取对应的尾实体,随着关系的增多,其推理时间会显著增加。相较之下,MF-SERel 模型通过并行处理预定义的关系,降低了对关系数量的敏感度,从而有效缩短了训练和推理时间,体现出较强的计算优势。TPLinker 模型相比 CasRel,需要构建和标注多个实体对之间的关系矩阵,这显著增加了其计算时间和资源消耗,但其能够更加全面考虑不同实体对及其关系,这使其在处理复杂关系时相比

CasRel具有较高的抽取精度。MF-SERel模型与TPLinker相比,MF-SERel无需依次遍历所有关系,不仅在计算效率上存在优势,同时,MF-SERel在特征表达方面也更为丰富,能够提取到文本的复

杂逻辑结构信息,在PRD和DuIE数据集上,F1值相比TPLinker分别提升了2.3%和1.7%,表明MF-SERel可以在处理复杂关系和冗余计算问题上达到良好的效果。

表7 模型的计算效率比较

Table 7 Comparison of the computational efficiency of the models

数据集 Data set	模型 Model	训练时间/s Training time	推理时间/s Inference time	F1值/% F1 value
PRD	GraphRel _{2p}	178	67	73.3
	TPLinker	162	57	94.3
	CasRel	133	41	93.2
	MF-SERel	116	34	96.6
DuIE	GraphRel _{2p}	1 006	395	51.2
	TPLinker	801	272	67.2
	CasRel	589	140	62.8
	MF-SERel	552	91	68.9

3.7 图卷积网络隐藏层维度设置分析

为验证图卷积网络隐藏层参数对模型性能的影响,试验分别将隐藏层维度设置为256、512、768和1 024。模型的精确率、召回率和F1值随隐藏层维度的变化趋势如图8所示。试验结果表明,当隐藏层数为768时,模型达到了最优性能,F1值达到了96.6%。相比之下,当模型的隐藏层为256时性能最差,F1值仅为95.0%,这表明隐藏层维度过小无法有效捕获文本的句法特征,导

致特征表达不足。在隐藏层数为1 024时,模型效果呈现降低趋势,F1值为95.6%。这种性能下降可能归因于过拟合现象,即模型在训练时过度依赖样本特征,学习到噪声数据,导致泛化能力减弱。综上所述,过少的隐藏单元无法充分学习文本特征,而过多的隐藏单元易导致过拟合,影响模型的泛化能力。因此,在本模型中,将隐藏层维度设置为768时,能够在特征学习和模型泛化之间取得良好的平衡。

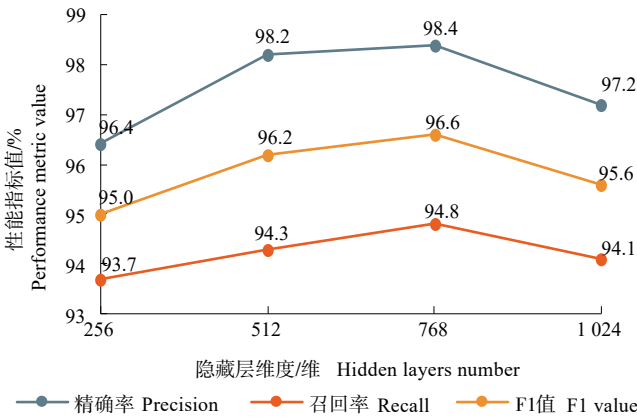


图8 不同隐藏层参数下模型的性能分析

Fig. 8 Performance analysis of the model under different hidden layer parameters

4 讨论与结论

本研究提出一种多特征融合的单阶段实体关

系联合抽取模型(MF-SERel)实现了农药信息的实体关系自动化提取,并与现有主流模型如GraphRel、CasRel和TPLinker进行对比,展现出了

较好的识别效果。与 GraphRel^[24]相比, MF-SERel 的核心优势在于其单阶段抽取策略,能够有效减少曝光偏差。GraphRel通过多阶段的联合抽取来处理实体关系,其多轮的推理过程可能导致实体与关系之间产生信息不一致或偏差。MF-SERel则通过单阶段抽取机制,将实体和关系的预测整合到同一推理过程,减少了信息损失和偏差,进而提升了抽取结果的一致性和准确性。相较 CasRel^[16]方法, MF-SERel在特征融合方面表现出了更强的能力。CasRel仅依赖语义信息进行特征提取,在面对复杂的特征和多层次的上下文时,特征提取的深度和广度受限。相比之下, MF-SERel通过多特征融合,结合了语义、句法结构的交互信息,显著提高了模型捕获复杂特征的能力。这种多特征融合的方式使得 MF-SERel在处理多重关系和复杂上下文时,能够展现出更强的表达能力和准确性。TPLinker^[22]方法的局限性在于其计算过程中的冗余,其需要遍历所有预定义关系进行三元组的整合,这一过程会引入不必要的计算开销,降低推理效率。本研究提出的 MF-SERel通过优化计算流程,减少了对冗余关系的计算,提升了推理速度和计算效率。特别是在大规模数据集上, MF-SERel能够显著减少计算成本,并保证较高的处理效率。

尽管本研究在准确性和效率上取得了显著进展,但仍存在一定的改进空间。当前模型对标注数据的依赖较高,且泛化能力尚有提升空间。未来的研究可以考虑将生成式模型(如 ChatGPT)与本方法结合,利用生成式模型在跨领域学习中的优势,以进一步提升模型的泛化能力和适应性。通过结合领域知识进行数据标注与预处理,可以进一步增强信息抽取任务的灵活性、鲁棒性和生成结果的可靠性。

综上所述,本研究针对农药登记文本中信息密集、逻辑结构复杂、实体跨度较大及实体长度异质性等特点,提出了一种多特征融合的单阶段实体关系联合抽取模型 MF-SERel。主要结论如下:

1) 本研究通过将语义特征与句法特征进行融合,提升了模型对复杂语料的综合理解能力,有助于解决传统联合抽取模型在特定领域特征提取不够充分的问题。

2) 本研究提出的 HT-BES 多维标注策略,将实体关系联合抽取转化为基于关系维度的多标签标

注任务,实现了单阶段标签分类,可以有效解决三元组重叠问题,缓解了曝光偏差并降低了计算冗余。

3) 本研究提出的 MF-SERel 模型能够在处理复杂逻辑结构及多三元组的情况下保持稳定的抽取性能,取得了较快的训练、推理速度以及较高的 F1 值,为农药领域知识的结构化提供了一种高效准确的抽取方法。

参考文献 References

- [1] 孙晓明, 卢海燕, 卞立平, 刘贤金, 孙爱东. 基于登记信息构建的农药使用风险智能预警模型研究[J]. 农药学报, 2021, 23(5): 1012-1023
Sun X M, Lu H Y, Bian L P, Liu X J, Sun A D. Study on intelligent early warning model of pesticide application risk based on registration information[J]. *Chinese Journal of Pesticide Science*, 2021, 23(5): 1012-1023 (in Chinese)
- [2] Lewis K A, Tzivilakis J, Warner D J, Green A. An international database for pesticide risk assessments and management[J]. *Human and Ecological Risk Assessment*, 2016, 22(4): 1050-1064
- [3] 魏昊, 周爱, 张益嘉, 陈飞, 屈雯, 鲁明羽. 深度学习生物医学实体关系抽取研究综述[J]. 计算机工程与应用, 2021, 57(21): 14-23
Wei H, Zhou A, Zhang Y J, Chen F, Qu W, Lu M Y. Review of deep learning-based biomedical entity relation extraction research[J]. *Computer Engineering and Applications*, 2021, 57(21): 14-23 (in Chinese)
- [4] Zhang J C, Zain A M, Zhou K Q, Chen X, Zhang R M. A review of recommender systems based on knowledge graph embedding[J]. *Expert Systems with Applications*, 2024, 250: 123876
- [5] 侯琛, 牛培宇. 农业知识图谱技术研究现状与展望[J]. 农业机械学报, 2024, 55(6): 1-17
Hou C, Niu P Y. Review of research status and prospects of agricultural knowledge graphs [J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2024, 55(6): 1-17 (in Chinese)
- [6] Zhao K, Xu H, Cheng Y, Li X T, Gao K. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction[J]. *Knowledge-Based Systems*, 2021, 219: 106888
- [7] 郝志刚, 刘冲, 秦丽. 基于中文字邻接图的食品抽检公告实体及关系联合抽取[J]. 农业工程学报, 2023, 39(14): 283-292
Hao Z G, Liu C, Qin L. Entity and relationship joint extraction model of food inspection announcement based on Chinese character adjacency graph [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2023, 39(14): 283-292 (in Chinese)
- [8] Choi H S, Song J Y, Shin K H, Chang J H, Jang B S. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer[J]. *Radiation Oncology Journal*, 2023, 41(3): 209-216
- [9] Liu Z R, Li H Z, Wang H, Liao Y L, Liu X G, Wu G J. A novel pipelined end-to-end relation extraction framework with entity mentions and contextual semantic representation [J]. *Expert Systems with Applications*, 2023, 228: 120435
- [10] Gao C, Zhang X, Liu H, Yun W, Jiang J H. A joint extraction model of entities and relations based on relation decomposition[J]. *International*

- Journal of Machine Learning and Cybernetics*, 2022, 13(7): 1833-1845
- [11] Ai X G, Kavuluru R. End-to-end models for chemical-protein interaction extraction: Better tokenization and span-based pipeline strategies [C]// *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. June 26–29, 2023, Houston: IEEE, 2023: 610-618
- [12] Yang C L, Deng J M, Chen X L, An Y. SPBERE: Boosting span-based pipeline biomedical entity and relation extraction via entity information [J]. *Journal of Biomedical Informatics*, 2023, 145: 104456
- [13] Zheng S C, Wang F, Bao H Y, Hao Y X, Zhou P, Xu B. 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 (Volume 1: Long Papers)*. Vancouver: USAACL, 2017
- [14] Chang H Y, Zan H Y, Guan T F, Zhang K L, Sui Z F. Application of cascade binary pointer tagging in joint entity and relation extraction of Chinese medical text [J]. *Mathematical Biosciences and Engineering*, 2022, 19(10): 10656-10672
- [15] Yuan Y, Zhou X F, Pan S R, Zhu Q N, Song Z L, Guo L. A relation-specific attention network for joint entity and relation extraction [C]// *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. January 7–15, 2021, Yokohama: ACM, 2021: 4054-4060
- [16] Wei Z P, Su J L, Wang Y, Tian Y, Chang Y. A novel cascade binary tagging framework for relational triple extraction [C]// *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: USAACL, 2020: 1476-1488
- [17] 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 the mixture of characters and words [J]. *Application Research of Computers*, 2021, 38(9): 2619-2623 (in Chinese)
- [18] 张梦赢, 沈海龙. DE-AA: 基于词对距离嵌入和轴向注意力机制的实体关系联合抽取模型 [J]. *计算机科学*, 2024, 51(12): 234-241
- Zhang M Y, Shen H L. Joint extraction of entities and relations based on word-pair distance embedding and axial attention mechanism [J]. *Computer Science*, 2024, 51(12): 234-241 (in Chinese)
- [19] Zheng H Y, Wen R, Chen X, Yang Y F, Zhang Y Y, Zhang Z H, Zheng N Y, Qin B, Ming X, Zheng Y F. 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 (Volume 1: Long Papers)*. Stroudsburg: USAACL, 2021: 6225-6235
- [20] Shang Y M, Huang H Y, Mao X L. OneRel: Joint entity and relation extraction with one module in one step [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(10): 11285-11293
- [21] Chen Z Q, Zheng Y J, Ge J K, Yu W C, Wang Z N. A parallel model for jointly extracting entities and relations [J]. *Neural Processing Letters*, 2024, 56(3): 165
- [22] Wang Y C, Yu B W, Zhang Y Y, Liu T W, Zhu H S, Sun L M. 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
- [23] 朱秀宝, 周刚, 陈静, 卢记仓, 向怡馨. 基于增强序列标注策略的单阶段联合实体关系抽取方法 [J]. *计算机科学*, 2023, 50(8): 184-192
- Zhu X B, Zhou G, Chen J, Lu J C, Xiang Y X. Single-stage joint entity and relation extraction method based on enhanced sequence annotation strategy [J]. *Computer Science*, 2023, 50(8): 184-192 (in Chinese)
- [24] 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*. Stroudsburg: USAACL, 2019: 1409-1418
- [25] 苗琳, 张英俊, 谢斌红, 李玉. 基于图神经网络的联合实体关系抽取 [J]. *计算机应用研究*, 2022, 39(2): 424-431
- Miao L, Zhang Y J, Xie B H, Li Y. Joint entity relation extraction based on graph neural network [J]. *Application Research of Computers*, 2022, 39(2): 424-431 (in Chinese)
- [26] 翟社平, 柏晓夏, 张宇航, 成大宝. 融合依存分析和图注意网络的三元组抽取 [J]. *计算机工程与应用*, 2023, 59(12): 148-156
- Zhai S P, Bai X X, Zhang Y H, Cheng D B. Triple extraction of combining dependency analysis and graph attention network [J]. *Computer Engineering and Applications*, 2023, 59(12): 148-156 (in Chinese)
- [27] 吕东东, 陈俊华, 毛典辉, 张青川, 赵敏, 郝治昊. 农产品标准领域知识图谱实体关系抽取及关联性分析 [J]. *农业工程学报*, 2022, 38(9): 315-323
- Lyv D D, Chen J H, Mao D H, Zhang Q C, Zhao M, Hao Z H. Entity relationship extraction and correlation analysis of agricultural product standard domain knowledge graph [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2022, 38(9): 315-323 (in Chinese)
- [28] 周俊, 郑彭元, 袁立存, 戈为溪, 梁静. 基于改进CASREL的水稻施肥知识图谱信息抽取研究 [J]. *农业机械学报*, 2022, 53(11): 314-322
- Zhou J, Zheng P Y, Yuan L C, Ge W X, Liang J. Knowledge graph information extraction for rice fertilization based on improved CASREL [J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2022, 53(11): 314-322 (in Chinese)
- [29] Jia X X, Zheng G, Dong C Y, Xi S Y, Shi L, Xiong S F, Ma X M, Xi L. Joint extraction of wheat germplasm information entity relationship based on deep character and word fusion [J]. *Scientific Reports*, 2024, 14(1): 10385
- [30] 吴赛赛, 周爱莲, 谢能付, 梁晓贺, 汪汇涓, 李小雨, 陈桂鹏. 基于深度学习的作物病虫害可视化知识图谱构建 [J]. *农业工程学报*, 2020, 36(24): 177-185
- Wu S S, Zhou A L, Xie N F, Liang X H, Wang H J, Li X Y, Chen G P. Construction of visualization domain-specific knowledge graph of crop diseases and pests based on deep learning [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2020, 36(24): 177-185 (in Chinese)
- [31] 姜丽华, 赵瑞雪, 董春岩, 常小燕, 马娟娟, 谢能付, 方松. 基于深度学习的水产病害可视化知识图谱构建与验证 [J]. *农业工程学报*, 2023, 39(15): 259-267
- Jiang L H, Zhao R X, Dong C Y, Chang X Y, Ma J J, Xie N F, Fang S. Construction and verification of the visual knowledge map of aquatic diseases based on deep learning [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2023, 39(15): 259-267 (in Chinese)
- [32] Li S J, He W, Shi Y B, Jiang W B, Liang H J, Jiang Y, Zhang Y, Lyu Y J, Zhu Y. Duie: A large-scale Chinese dataset for information extraction [C]// *Natural Language Processing and Chinese Computing: 8th CCF*

International Conference, NLPCC 2019, Dunhuang: Springer International Publishing, 2019: 791-800

- [33] Colic N, Rinaldi F. Improving spaCy dependency annotation and PoS tagging web service using independent NER services [J]. *Genomics & Informatics*, 2019, 17(2): e21

- [34] Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs[C]//*Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Phoenix: ACM, 2016: 1955-1961

责任编辑: 杨爱东



第一作者简介: 车银超, 博士研究生, 河南农业大学讲师。中国作物学会智慧农业专委会会员、中国中文信息学会会员。主要从事计算机应用、知识工程、农业信息技术等领域的教学科研工作。先后参与国家重点研发计划、河南省重点研发专项、河南省科技攻关等项目的研发工作。获得河南省科技进步二等奖 1 项, 河南省信息技术教育成果二等奖 1 项; 发表学术论文 10 余篇, 获得国家发明专利、软件著作权 6 项。获河南农业大学“优秀教师”“优秀班主任”等荣誉称号。



通讯作者简介: 席磊, 教授, 河南农业大学硕士生导师。现任河南省政府信息化专家咨询委员会委员, 中国仿真学会农业建模与仿真专委会副主任委员。负责河南省农田环境监测与大数据分析工程研究中心、农业信息认知计算与智能系统团队和平台建设。主要从事智慧农业技术的研究, 对粮食作物精准监测、知识挖掘与智能系统有着深入的研究。相关研究成果获河南省科技进步一等奖 1 项、二等奖 3 项, 三等奖 1 项, 河南省高等教育教学成果一等奖 1 项。获批河南省高等学校教学名师、河南省高层次人才等荣誉。先后主持、参加国家“863”计划、国家科技支撑计划、省重大科技专项、省重点研发计划等科研课题 19 项, 主持省级教育教学改革项目 3 项, 省级在线开放精品课程 1 门。发表学术论文 82 篇, 出版教材和专著 7 部, 授权国家发明专利 5 项。