科技文献算法嵌套实体识别

刘齐凯[1,2] 李鹏程[1,2] 陆伟[1,2] 程齐凯[1,2]
(1. 武汉大学信息管理学院，武汉 430072；2. 武汉大学信息检索与知识挖掘研究所，武汉 430072)

摘要：本文探讨了科技文献中算法实体的自动识别研究，着重研究嵌套型算法实体的识别优化问题。首先通过远程监督学习的方式构建算法实体训练语料，再引入数据增强技术扩充语料规模，最后应用BartNER模型实现科技文献中嵌套算法实体的自动识别。实验结果显示，在引用数据增强技术的基础上BartNER模型取得了76.66%的F1值，证明该方法对嵌套算法实体识别的有效性，同时证明数据增强策略能够有效提升嵌套算法实体的识别效果。

关键词：实体识别；嵌套实体识别；数据增强；BART
中图分类号：G35；TP391　DOI：10.3772/j.issn.1673-2286.2022.02.001
引文格式：刘齐凯，李鹏程，陆伟，等. 科技文献算法嵌套实体识别[J]. 数字图书馆论坛，2022 (2)：2-9.

命名实体识别（Named Entity Recognition，NER）作为自然语言处理的底层关键技术，旨在从非结构化的文本中抽取出具有特定含义的实体并将其分类为预定义的特定类型，常见实体包括人名、地名以及机构名等。命名实体识别构成了自然语言语义理解的重要基石，在事件抽取和问答系统等诸多下游任务中扮演了关键角色。随着学科领域的发展和课题研究的细化，面向特定领域的特定实体识别得到学者的关注。例如，面向医学领域的基因、疾病和药物的识别，以及面向计算机领域的算法、数据集和软件工具等实体的识别。

算法是计算机等领域的科技文献中广泛存在的一种命名实体，是一种预先定义好的，解决特定问题的一系列规则[1]。深度学习兴起以来，大量新的机器学习、深度学习算法和模型被逐一提出，一些经典算法如LDA、LSTM、BERT在不断改进的同时也被广泛用于不同类型的任务场景中。在科技文献中，算法通常是作为一种用于解决特定问题的方法和模型，如通过TextRank算法进行关键词抽取以及利用BiLSTM-CRF算法进行命名实体识别，这些算法在一定程度上揭示了当前文献的主要研究内容。因此，识别科技文献中使用或引证的算法实体对于大规模科技文献的文本理解与语义计算具有重要意义。本文将科技文献中引用、提出、应用的算法和模型统称为算法实体并对其进行自动化识别。

由于算法之间存在较为广泛的相互借鉴、改进和依赖，因此算法实体较其他类型的实体而言具有更加显著的嵌套现象。嵌套命名实体是一种特殊形式的实体，其内部包含其他实体。嵌套在里面的实体称为内部实体，外层的实体称为外部实体。嵌套实体在常见的实体识别数据集中普遍存在，GENIA数据集中含有嵌套实体的句子占比17%，ACE2004数据集和ACE2005数据集中包含嵌套实体的句子占比均达到了30%[2]。嵌套实体关系的存在对传统的单层普通实体识别带来了技术上的挑战。

当前，算法实体的自动识别研究已经得到学者的重视并取得一定进展，如Wang等[3]对自然语言处理领域的算法实体进行识别并分析了不同算法的影响力。然而，对于在结构上更加复杂的嵌套算法实体，现有研究缺乏足够的关注。为了更好地实现科技文献中算法实体的自动识别以及提升嵌套算法实体的识别效果，本文首先采用远程监督学习的方式构建标注语料，再引入数据增强的方法扩充训练语料，实现一种基于文本生成的实体识别框架BartNER，该方法最终实现76.66的F1值，有效解决了科技文献算法嵌套实体识别问题。

1 研究现状

传统命名实体识别主要从三个维度出发进行研究：基于规则的方法、基于传统机器学习的方法和基于深度学习的方法。基于规则的方法采用人工分析实体前后的语法语义规则，通过尽可能多地列出规则集合以达到识别出预测内容中具有相同模式的实体，因此需要较多的人力且推广性较差；基于传统机器学习的方法从已有的标注数据出发，通过对数据进行建模，利用概率统计模型识别标注数据的特征以期对未标注的数据做提取标注，但是特征表达能力较差；基于深度学习的方法不需要人工构建太多特征，而是通过复杂的网络结构表征之前方法无法表达的组合特征，更加适合大规模语料，因而可以取得更好的效果，此外基于深度学习的方法不需要特定领域的语料[4]，因此更具有可推广性。当前学者主要采用基于深度学习的方法进行实体识别研究。Ma等[5]利用CNN提取单词的字符级表示，采用LSTM-CNNs-CRF模型，在CoNLL 2003数据集上取得了91.21%的F1值。Wang等[6]基于强化学习的思想提出了一种自动组合词向量模型，通过对不同的任务选择不同的词向量组合，最终在CoNLL 2003数据集上取得94.6%的F1值。

除了对地名、机构名、人名、时间等常见的7个类别实体进行识别[6]外，学者还对生物学、药学、科技文献等领域实体进行了深一步的研究。随着文献数量的逐年增加，科技文献实体识别对于学者准确把握研究新方向有重要意义。余丽等[7]基于改进Bootstrapping方法，自动化构建了科研文献领域大规模的标注语料库，然后采用LSTM-CRF算法识别了论文摘要中研究方法、研究范畴、评价指标及取值、实验数据4类科技文献实体。在算法实体识别领域，AlgorithmSeer[8]作为CiteSeer的一部分，首次关注大规模学术文本中算法实体的抽取，该系统主要从文献中的伪代码和算法流程中抽取算法实体，采用基于规则和基于机器学习的方法，共从200万篇学术文献中抽取出20万条算法实体词。Safder等[9]聚焦与算法实体相关的元数据提取，抽取了算法相关的评价指标、数据集、算法复杂度等其他相关元数据，提出了一种基于Bi-LSTM算法实体识别模型，在3.7万条手工标注的文献句中实现81.0%的F1值。

随着深度学习的发展，Resnet、Transformers、BERT等预训练模型发挥越来越大的作用，学者基于先前研究的成果不断改进模型，提出新的算法，使得算法实体出现了越来越多的嵌套现象，而先前学者鲜有对嵌套实体识别进行研究。传统普通实体的识别方法一般被视作一个序列标注任务，只需要对句子中的每个词标上对应的标签即可，而嵌套实体因其相对复杂的结构使得序列标注方法无法直接使用。因此，学者在研究嵌套实体识别时主要运用以下两种方法。①基于超图（Hypergraph-based）的方法。超图（Hypergraph）是一种普通图的推广变体，其主要特征是一条边可以连接任意数量的顶点，其可以准确方便地描述存在多元关联的对象之间的关系，因而被广泛应用在机器翻译、句法解析、语义解析等领域[10]。超图可以完美刻画嵌套命名实体标签网络中一条边包含多节点的问题。Wang等[11]提出了一种新的神经网络分段超图表示方法，改善了先前超图方法在推理过程中普遍存在的结构歧义问题。该模型以O（cmn）的复杂度捕获到先前模型无法捕获的特征和特征交互，使用超图对嵌套实体的所有可能组合进行编码，在ACE2005数据集上取得了74.5%的F1值，体现了结构歧义消除的必要性。基于超图的方法因超图能够描述多元关联关系的特点为嵌套命名实体识别提供了一种较好的解决方法，但其存在结构上的表达歧义和伪结构的问题，同时因其需要构建可能的全部组合，耗费的时间较长。②基于区段（Span-based）的方法。基于区段的方法是将句子划分为多个子区段并对子区段分类来识别嵌套命名实体，由于内部实体和外部实体属于不同的子序列，可以相对容易地进行分类。Sohrab等[12]提出了深度穷举模型，主要思想就是列举出所有可能的区段作为潜在的实体，然后用深层神经网络进行分类，为了避免穷举带来的时间复杂度的指数增强，模型通过共享的Bi-LSTM来降低复杂度。Shen等[13]提出了一种借鉴计算机视觉思想的两阶段嵌套实体识别方法，将NER看作一个边界回归和区段分类的联合任务，基于IoU弹性地处理样本数据，而不是简单地把部分匹配的区段当作噪音样本。

除此之外，还有以下两种新的思路也实现了较好的效果。①基于状态转换的方法。Wang等[14]提出使用森林结构来建模分散在句子中的嵌套命名实体，定义了shift-reduce两种状态转换过程，利用分类器判别下一步要执行的操作，直到识别完成。②基于文本生成的方法。Yan等[15]提出将NER三类子任务描述为实体跨序列生成任务，通过一个统一的序列到序列（Seq2Seq）框架BartNER来解决。该统一框架利用预先训练的Seq2Seq模型来解决所有三种类型的NER子任务，而不

需要特别设计标记模式或枚举区段的方法。该模型基于Seq2Seq方法，与传统命名实体识别方法思想相似，在不需要像基于超图和基于区段的方法提出新的解决范式的前提下，能够同时识别普通实体、嵌套实体和不连续实体，具有较强的鲁棒性和可推广性。

2 研究方法

2.1 整体框架

本文提出的科技文献中算法嵌套实体识别的方案设计思路如图1所示。首先通过网络搜集并筛选算法实体词构建算法实体词词典，词典中包含普通算法实体和嵌套算法实体。然后通过对开源的ArXiv数据集提取文献摘要（文献限定在AI相关领域），得到待标注语料。结合算法实体词词典和摘要数据语料，通过远程监督的方法得到初始的训练语料，并结合实体词替换的数据增强方法得到扩充后的训练语料。最后实现一种基于文本生成的BartNER模型，该方法基于文本生成的思想，输出潜在实体的开始和结束位置下标以及对应的实体类型，巧妙地解决了嵌套实体的识别，同时还实现了另外两种具有代表性且效果较好的嵌套实体识别模型：基于超图的Neural Segmental Hypergraph（NSH）[11]和基于区段的Locate and Label[13]，并对这3个模型的效果进行对比分析。

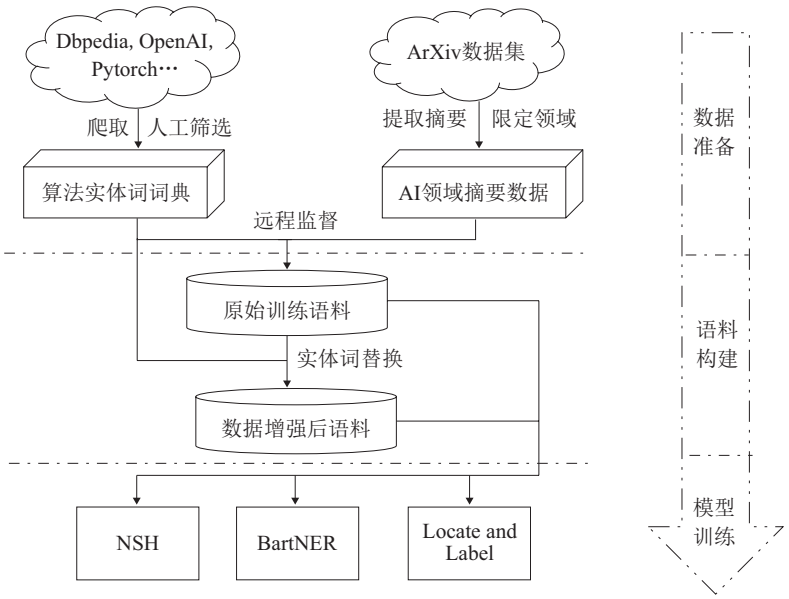

图1 技术路线图

2.2 基于远程监督的训练语料构建

2.2.1 实体词词典构建

机器学习是一门多学科交叉专业，涵盖概率论、统计学、近似理论和复杂算法知识，使用计算机作为工具并致力于真实、实时的模拟人类学习方式，并将现有内容进行知识结构划分来有效提高学习效率[16]。经过多年的发展，机器学习领域有大量成熟的算法被提出，一些经典的算法如Clustering、GBDT等仍然在学术界和工业界普遍使用。深度学习自兴起后各类新算法层出不穷，百家争鸣。为了得到更全、更新的机器学习和深度学习相关算法实体词，本文从多个渠道爬取了相关实体词并通过人工筛选得到最终的实体词词典。不同渠道的搜集策略：百科全书Dbpedia中“algorithm”相关的词条，Google和OpenAI提供的机器学习/深度学习术语手册，3个成熟的深度学习框架TensorFlow、PyTorch和Hugging Face提供的全部算法名称。人工筛选去重后取小写共得到490个算法实体词，构成本文的算法实体词词典。

2.2.2 标注语料构建

本文训练语料选用开源的arXiv数据集。arXiv是

一个预印本论文存缴、检索、发布和交流的开放平台，1991年由物理学家Paul Ginsparg创立，2001年转交美国康奈尔大学图书馆运营管理，目前已扩展至物理、数学、计算机科学、定量生物学、定量金融学、统计学、电子工程和系统科学、经济学8个学科领域。arXiv数据集共包含170万篇论文的PDF文件。本文选用2015年1月1日—2022年1月18日人工智能、自然语言处理、计算机视觉、机器学习、信息检索领域共17万篇论文的摘要数据，利用基于算法实体词词典的远程监督方法，得到24 606条原始数据集，其中包含嵌套算法实体的句子3 976条（占16%），包括299个算法实体词（占算法实体词词典总数的61%）。通过对原始数据集中算法实体词出现的频次进行统计，得到频次前五的算法实体词为Clustering、CNN、Reinforcement Learning、LSTM和Gradient Descent，可以看到传统的机器学习方法Clustering仍然是学者们使用较多的算法，而随着深度学习的兴起，CNN、LSTM、Reinforcement Learning等算法都被广泛使用，深度学习常用的优化方法Gradient Descent也被经常提及。

算法实体识别领域没有标准的数据集来验证训练出的模型效果，因此本文选取了100篇和训练数据不重合的文献进行人工标注，得到260条数据，其中包含嵌套实体的数据有49条，占比18.87%。

2.3 基于BartNER的嵌套算法实体识别模型

嵌套实体因其存在多元对象关联关系，识别过程无法像普通命名实体识别那样基于序列标注模型。常见的解决思路是基于超图的模型和基于区段的模型，同时学者也在探索新的思路和模型。本文实现了一种基于文本生成的BartNER[15]算法嵌套实识别模型。BartNER是一种基于BART预训练模型和pointer机制的Seq2Seq模型，模型能够输出潜在实体的开始位置和结束位置下标以及对应的实体标签，从而能够同时处理普通实体、嵌套实体和不连续实体3种情况，是一种命名实体识别的统一处理框架。

模型结构如图2所示[15]，主要由BART Encoder和BART Decoder组成，结合pointer机制生成潜在实体的开始位置和结束位置下标。其中，BART是一种由Transformers Encoder和Decoder堆叠而成，训练过程是将被破坏的原始文本恢复原样，pointer机制是对自注意力机制的一种简化。

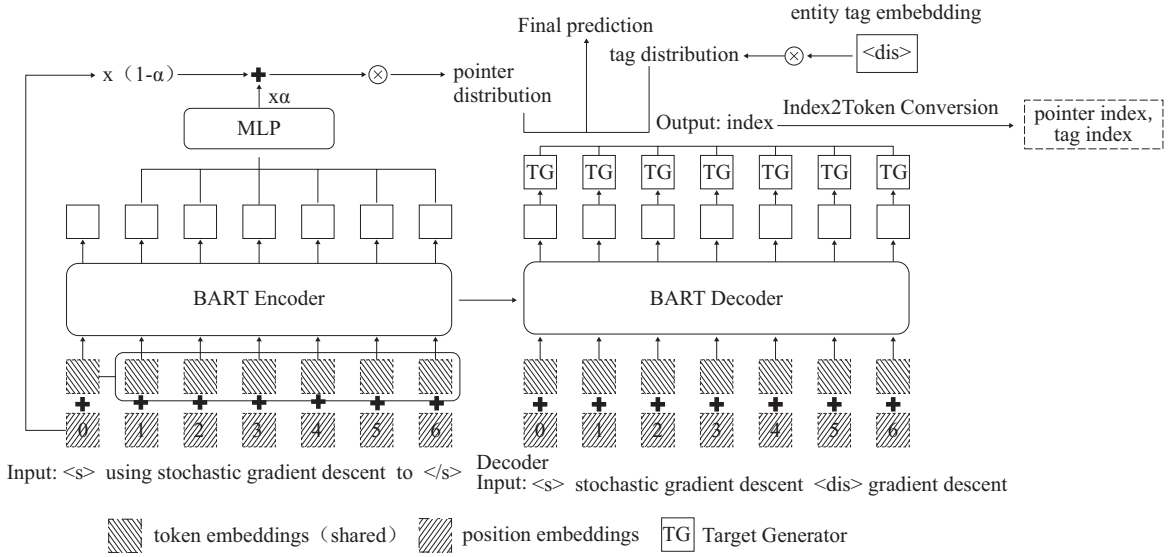

图2 BartNER模型框架

Encoder阶段主要是将句子的输入转化为向量，Decoder阶段则是通过计算下标的分布概率，得到相应实体开始位置和结束位置的下标以及对应的实体类型的索引，其中实体类型是通过实体类型词典中的索引得到。总的来说，模型接受输入X=[x1，x2，…，xn]，输出Y=[s11，e11，…，slj，elj，t1，…，si1，ei1，…，sik，eik，ti]，s和e分别代表一个实体的开始和结束位置下下标，t是一个实体的类型。因为嵌套现象的存在，一个词可能是多个实体的开始或结束，对于某一个特定实体来说，可以表示为[si1，ei1，ti]。因为模型能够输出同一个实体的多个开始位置和结束位置下标，所以模型能够识别出句子中潜在的实体和嵌套实体，是一种解

决嵌套实体的新思路和有效方法。

2.4 基于实体替换的数据增强策略

通过远程监督策略构建初始训练语料，容易因语料数量较少导致模型训练不足，造成过拟合现象，因此本文使用数据增强方法来扩充训练语料规模，探究数据增强在算法嵌套实体识别上的效果。数据增强（Data Augmentation），是指根据现有数据合成新数据的一类方法。当训练深度学习模型时，训练数据的规模和多样性会对模型效果产生巨大影响，数据在一定程度上决定了最终效果的上限，有了更多数据后可以提升效果、增强模型泛化能力、提高鲁棒性等。数据增强广泛应用于计算机领域，如图片剪切、旋转、增加噪点等，然而由于自然语言处理任务天然的难度，类似计算机视觉的裁剪方法可能会改变语义，既要保证数据质量又要保证多样性，使得在自然语言处理领域做数据增强时须十分谨慎。自然语言处理较常用的一种方法是词替换[17]，通过将词语替换成同义词或同类型的词，既能保持句子的语义不发生大的变化，也能够达到增加数据量的目的。考虑到本文构建的算法词典共包含490个算法实体词，其中在训练语料中出现的共有299个，占比61%，因此有39%的实体词未在训练语料中出现。本文采用词替换的数据增强方法，扩充了3种不同规模的语料，分别将原始训练语料的24 606条语句随机抽取出5 000条、10 000条和15 000条，将抽取出来的语句中的算法实体词替换成实体词词典中未在原始训练语料中出现过的其他算法实体词，具体实体替换步骤如图3所示，然后将替换实体词后的语料加入到原始24 606条语料中并打散，得到新的数据增强后共29 606条、34 606、39 606条训练语料。

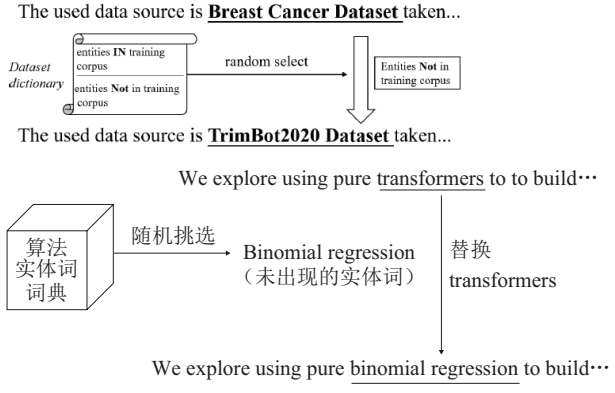

图3 实体替换示例

3 实验

3.1 实验设定

本文模型采用PyTorch框架，使用Nvidia 24G内存Ge-Force RTX-3090 GPU进行加速训练。本文使用BART作为模型的Encoder和Decoder，BART预训练模型有Base和Large两个版本，本文选用BART Large版本，有12层Transformers Encoder和Transformers Decoder，比同等规模的BERT模型约多10%的参数量。本文其他关键参数设置如表1所示。

表1 实验关键参数取值

参 数	取 值
Epoch	30
Warmup ratio	0.01
Learning rate	1e-5
Batch size	16
BART version	BART Large

3.2 实验评估指标

实体识别（包括普通实体和嵌套实体）主要有3个评估指标，即准确率、召回率、F1值，仅当模型识别出来的实体边界和类型都正确时才会标注该命名实体为正确识别。准确率代表了识别结果的准确程度，通过识别正确的实体总数除以模型识别出的实体总数得到；召回率代表了语料中全部正确样本被正确预测的程度，通过识别正确的实体总数除以语料中实体总数得到；F1值是准确率和召回率的加权调和平均，是两者的综合考量。

3.3 实验结果及分析

本文实现的3个嵌套实体模型都是目前在嵌套实体领域较典型且效果较好的模型，NSH、BartNER、Locate and Label模型在嵌套实体识别领域常使用的Genia数据集上取得的F1值分别为75.1%[11]、79.23%[13]、80.54%[15]，其中Locate and Label模型是GENIA数据集上效果最好的模型。

3个模型在原始训练语料和数据增强训练语料的结果如表2所示，BartNER模型在所有语料上都取得了最

好的F1值，分别为75.39%、76.66%、75.72%和75.56%。目前常见的嵌套实体集Genia的最优F1值为81.23%，最高76.66%的F1值与此接近，证明BartNER在科技文献算法嵌套实体识别上的有效性。NSH和Locate and Label模型的F1值最高分别为69.86%和71.53%，两者相差不大；但NSH模型的准确率最高达到了99.63%，远高于其他两个模型，不过NSH模型的召回率是3个模型中最低的，说明NSH模型过于保守，虽然能够准确识别算法实体，但识别出的数量较少，如果在追求较高准确率的场景下，可以尝试NSH模型。BartNER虽然在准确率和召回率上都不是3个模型中最优的，但其最终的F1值是最好的，说明BartNER平衡了模型识别的准确率和召回率。

表2 原始训练语料和数据增强语料实验结果

%

	原始训练数据			数据增强（+5 000）条训练数据			数据增强（+10 000）条训练数据			数据增强（+15 000）条训练数据		
	准确率	召回率	F1值	准确率	召回率	F1值	准确率	召回率	F1值	准确率	召回率	F1值
NSH	92.5	38.26	54.13	99.63	53.78	69.86	98.78	38.48	55.38	98.75	37.58	54.45
BartNER	86.6	66.74	75.39	88.82	67.43	76.66	87.88	66.51	75.72	87.84	66.28	75.56
Locate and Label	72.2	70.87	71.53	71.13	69.50	70.30	85.86	57.11	68.60	61.84	63.73	62.77

数据增强结果显示，NSH和BartNER在训练语料增强后都有一定程度的提高，BartNER虽然有提升，但提升不多，F1值都在76%左右；而NSH模型上数据增强结果效果较好，当增强5 000条数据时，F1值从54.13%提升到了69.86%，效果显著，证明数据增强方法在一定场景下可以取得较优的结果。需要注意的是，在Locate and Label模型上，数据增强不但没有取得正向提升，还导致F1值下降，尤其是在增强15 000条数据时，F1值下降将近10百分个点；同时在BartNER模型上，数据增强10 000条和15 000条数据时，F1值提升并没有数据增强5 000条的效果好。综上，数据增强方法是一种在自然语言处理任务中值得尝试的方法，方法成本较低，且在一定场景下能取得较优的结果，但因其本质上只是对原始语料进行了微小的变动，使得增强的文本所含的额外信息不多，如果加入过多重复信息，可能会对模型产生干扰，导致模型效果下降。

考虑到本文的训练语料通过基于实体词词典的远程监督方法得到，训练出的模型可能存在偏向仅在实体词词典中存在的词的偏差，本文对手动标注的100篇文献得到的260条测试语料进行区分，挑选出测试语料中存在但不在实体词词典中的实体词的语句，共计146条，其中包含嵌套实体的语句40条，嵌套实体句占比27%，3个模型在该146条测试语料上的实验结果如表3所示。

表3 非词典词测试语料结果

%

	准确率	召回率	F1值
NSH	86.67	29.09	43.56
BartNER	83.24	54.78	66.08
Locate and Label	70.64	56.62	62.86

可以看到，3个模型在新的测试语料上的F1值都有较大下降，都下降了10个百分点左右，说明模型对于全新算法嵌套实体的识别效果相对于在算法词典中存在的算法实体词较差。虽然F1值都有所下降，但BartNER模型的效果仍为3个模型中最佳，F1值达到66.08%，优于NSH模型在原始训练语料上54.13%的F1值，证明BartNER模型对于全新的算法实体词识别仍有不错的效果。考虑到一些经典的算法实体会在文献中不断出现，实际情况中算法实体词不会是全新的，BartNER模型在算法嵌套实体识别上仍是值得使用的。

综上所述，BartNER不仅取得了最好的F1值，同时作为一种统一的实体识别处理框架，除了可以识别普通实体和嵌套实体外，还可以处理另一类实体，即不连续实体的识别问题。不连续实体在算法实体中存在的情况较少，但在其他领域经常出现。此外基于预训练的文本生成方法使得模型的训练速度更快。本文证明了BartNER在具有较好的鲁棒性和可推广性的同时也有较高的精度，是一种可以有效应用于算法嵌套实体识别领域的模型。

4 总结

本文聚焦科技文献中算法实体这一类越来越重要的领域实体，构建了算法实体词词典，整理了适用的科技文献语料并通过远程监督方法得到科技文献算法实体的训练语料，同时结合数据增强方法，扩充原始的训练语料，该算法实体词词典和算法嵌套实体训练语料未来可以进一步使用。为解决科技文献算法实体识别的问题，着重考虑了算法实体识别过程中存在的嵌套实体现象，实现了一种基于文本生成BartNER模型，并与其他两个典型的嵌套实体识别模型进行对比，发现BartNER模型不仅具有在数据增强语料上达到了最好的76.66%的F1值，并且考虑到其作为命名实体统一框架以及高效的训练速度，认为这是一种有效的嵌套实体识别模型。不过本文仍存在一定不足。首先，本文构建的实体词词典是通过网络爬取筛选得到的，无法保证搜集的全面性；其次，本文使用的语料只选用了近8年开源的ArXiv数据库，其收录的论文有限，导致得到的原始语料库只有2万多条句子；最后，本文虽选用了3个在公开数据集上效果较好的模型，但没有根据领域实体识别任务的特殊性对模型细节进行优化，后续仍可在模型结构上加以调整，使其更好地适应算法嵌套实体识别任务。

参考文献

[1] SAFDER I，HASSAN S. Bibliometric-enhanced information retrieval：a novel deep feature engineering approach for algorithm searching from full-text publications [J]. Scientometrics，2019，119（1）：257-277.

[2] LUAN Y，WADDEN D，HE L，et al. A general framework for information extraction using dynamic span graphs [J]. NAACL-HLT，2019：3036-3046.

[3] WANG Y，ZHANG C. Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing [J]. Journal of informetrics，2020，14（4）：101091.

[4] YADAV V，BETHARD S. A survey on recent advances in named entity recognition from deep learning models [EB/OL]. [2022-01-01]. https://aclanthology.org/C18-1182.pdf.

[5] MA X Z，HOVY E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF [EB/OL]. [2022-01-01]. https://aclanthology.org/P16-1101.pdf.

[6] WANG X Y，JIANG Y，BACH N，et al. Automated Concatenation of Embeddings for Structured Prediction [EB/OL]. [2022-01-01]. https://faculty.sist.shanghaitech.edu.cn/faculty/tukw/acl21ace.pdf.

[7] 余丽，钱力，付常雷，等. 基于深度学习的文本中细粒度知识元抽取方法研究 [J]. 数据分析与知识发现，2019，3（1）：38-45.

[8] TUAROB S，BHATIA S，MITRA P，et al. AlgorithmSeer：a system for extracting and searching for algorithms in scholarly big data [J/OL]. Big Data IEEE Transactions on，2016 [2022-01-01]. DOI:10.1109/TBDATA.2016.2546302.

[9] SAFDER I，HASSAN S. Bibliometric-enhanced information retrieval：a novel deep feature engineering approach for algorithm searching from full-text publications [J]. Scientometrics，2019，119（1）：257-277.

[10] 余诗媛，郭淑明，黄瑞阳，等. 嵌套命名实体识别研究进展 [J]. 计算机科学，2021，48（S2）：1-10，29.

[11] WANG B，WEI L. Neural segmental hypergraphs for overlapping mention recognition [EB/OL]. [2022-01-01]. https://arxiv.org/pdf/1810.01817.pdf.

[12] SOHRAB M G，MIWA M. Deep Exhaustive Model for Nested Named Entity Recognition [C] //Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing，2018.

[13] SHEN Y，MA X，TAN Z，et al. Locate and Label：A two-stage identifier for nested named entity recognition [EB/OL]. [2022-01-01]. https://aclanthology.org/2021.acl-long.216.pdf.

[14] WANG B，LU W，WANG Y，et al. A Neural Transition-based Model for Nested Mention Recognition [EB/OL]. [2022-01-01]. https://aclanthology.org/D18-1124.pdf.

[15] YAN H，GUI T，DAI J Q，et al. A Unified Generative Framework for Various NER Subtasks [EB/OL]. [2022-01-01]. https://arxiv.org/abs/2106.01223.

[16] 李昊朋. 基于机器学习方法的智能机器人探究 [J]. 通讯世界，2019，26（4）：241-242.

[17] FENG S Y，GANGAL V，WEI J，et al. A survey of data augmentation approaches for nlp [EB/OL]. [2022-01-01]. https://arxiv.org/abs/2105.03075v1.

作者简介

刘齐凯，男，1996年生，硕士，研究方向：自然语言处理、数据挖掘。
李鹏程，男，1994年生，博士，研究方向：文本挖掘、深度学习。
陆伟，男，1974年生，博士，教授，研究方向：信息检索、知识管理、数据智能，E-mail：weilu@whu.edu.cn。
程齐凯，男，1989年生，博士，副教授，研究方向：自然语言处理、信息检索、机器学习。

Nested Algorithm Entity Recognition in Scientific and Technological Literature

LIU QiKai[1,2] LI PengCheng[1,2] LU Wei[1,2] CHENG QiKai[1,2]

(1. School of Information Management, Wuhan University, Wuhan 430072, P. R. China; 2. Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, P. R. China)

Abstract: The research of automatic recognition of algorithm entities in scientific literature is discussed, and the optimization of nested algorithmic entity recognition is emphatically studied. Firstly, the algorithm entity training corpus is constructed by means of distant supervision, then data augmentation is introduced to expand the corpus. Finally, the BartNER model is applied to recognize nested algorithm entities in scientific literature. The experimental results show that the BartNER model achieves an F1 value of 76.66% based on data augmentation, which proves the effectiveness of BartNER on the nested entity recognition problem, and also proves that data augmentation can effectively improve the recognition results of the nested algorithm entity.

Keywords: Entity Recognition; Nested Entity Recognition; Data Augmentation; BART

（收稿日期：2022-01-28）

书 讯

《汉语主题词表》

《汉语主题词表》自1980年问世以后，经1991年进行自然科学版修订，在我国图书情报界发挥了应有作用，曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要，由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，自2009年开始进行重新编制工作，拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表（工程技术卷）》共收录优选词19.6万条，非优选词16.4万条，等同率0.84，在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表（自然科学卷）》共收录专业术语12.4万条，包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域，收词系统、完整，语义关系丰富、严谨，每条词汇都有相应的学科分类号表现其专业属性，并与同义英文术语对应。同时，建立《汉语主题词表》网络服务系统，提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘，是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表（工程技术卷）》已于2014年由科学技术文献出版社出版，分为13个分册，总定价3 880元。《汉语主题词表（自然科学卷）》已于2018年5月由科学技术文献出版社出版，分为5个分册，总定价1 247元。两卷均可分册购买。