结合半监督学习和规则校正的中文学术论文问题实体识别研究*

傅柱　邱畅唱　刘鹏
（江苏科技大学经济管理学院，镇江 212100）

摘要：为快速定位和识别学术论文中的研究问题，针对中文学术论文提出一种结合半监督学习和规则校正的问题实体识别方法。首先以条件随机场模型为基础框架，构建词性、指示词等有监督特征和相似度、重要度等无监督特征，然后对比不同特征组合下的模型识别效果，结合领域语言学规则对识别结果进行校对处理，最后以“共享经济”和“船舶建造”主题领域为例进行实证研究。所提方法的实体识别性能优于主流深度学习模型和大语言模型等预训练模型结果，在两个领域主题语料集上的F1值分别达到85.82%和86.38%，在1/2和1/4数据集上的性能优势进一步扩大，表明所提方法在不同领域小规模标注数据集上能较好识别中文学术论文的问题实体，呈现出良好的有效性和稳健性。

关键词：问题实体识别；条件随机场；半监督学习；规则校正；小规模数据
中图分类号：TP391；G250　DOI：10.3772/j.issn.1673-2286.2024.12.007
引文格式：傅柱，邱畅唱，刘鹏. 结合半监督学习和规则校正的中文学术论文问题实体识别研究[J]. 数字图书馆论坛，2024，20（12）：56-65.

科学问题是学术研究的起点和动力，推动科学研究的深入和发展，因而研究问题成为了学术论文的核心要素。然而，科研活动中研究人员通常需要花费大量时间和精力去研读学术论文才能发现所有的研究问题，如何方便快捷地定位和识别学术论文中的研究问题亟待解决。

命名实体一般是指预定义的文本中表述特定意义的文本或是以名称为标识的实体[1]，通常涵盖人名、地名、组织机构名、时间、日期、货币等类别[2]。命名实体识别是自然语言处理的一项重要任务，最早在1987年信息理解会议（Message Understanding Conference，MUC）上提出[3]。目前学术界尚未对问题实体有明确界定，实体类型通常取决于具体的应用需求和定义[4]，研究问题可以被视为一种阐述研究问题核心内容的特定实体类型。关于问题定义，有学者认为问题定义可分为直接问题定义和任务概念代替两种方式[5]：在直接问题定义中问题可描述为科研工作中需要解决的对象[6]；而在任务概念代替中问题是研究工作的主体并将其定义为任务[7]。也有学者根据描述规则将问题分类为不同类型，通过归纳分析学术论文研究问题的语步结构给出问题的线性结构和描述规则[8]。综合前述，命名实体强调实体具有特定的名称或标识，学术论文中的研究问题通常具有独特的描述性属性或标识，可从多维度进行描述和定义[9]，因此问题实体是具有“特定意义”的实体，具有重要识别价值[10]。本文将问题实体定义为一种特殊的用来阐述学术论文相关研究问题概念的实体类型。

当前中文学术论文问题实体识别研究存在预训练

收稿日期：2024-07-10
*本研究得到国家社会科学基金项目“面向AI4S的场景化智慧知识服务框架研究”（编号：24CTQ029）资助。

模型匮乏、标注数据不足、模型复杂度高、可操作性差等问题。为此，本文提出结合半监督学习和规则校正的中文学术论文问题实体识别方法，通过利用少量标注数据与大量未标注数据，构建包含有监督特征、无监督特征和语言学规则的方法框架，一定程度上能够降低传统机器学习和深度学习在标注数据不足时的局限性，提高中文实体识别方法在具体领域的可操作性。通过问题实体识别，能高效提取和聚合学术论文中的研究问题，提升文献检索与知识发现的效率，为数智时代信息资源建设提供底层基础和支撑。

1　相关研究

目前问题实体识别研究主要分为基于文本匹配和基于深度学习的信息抽取两类。基于文本匹配的研究主要依赖预定义规则、匹配模式、关键词等信息，通过将文本中的词语或短语与预先构建的词典、模板或规则库进行比对匹配，识别出与目标问题相关的实体。Rau[11]提出结合启发式算法和人工规则概念，以实现从文本中自动提取不同实体类型的目标。在规则的指导下，这种方法在实体提取中的逻辑性较强，但过度依赖规则的构建，适用性较低。Gupta等[12]通过设计匹配规则来识别问题，使用Bootstrapping方法匹配词汇模板进行问题实体匹配。该研究的优势在于规则的直观性和易用性，但难以处理复杂句式或未知领域。纪科等[13]通过关键实体和文本匹配算法来提高对研究问题的信息检索能力。这类研究需要构建特定的规则，随着数据规模的增加，这种方法将变得不再适用。

深度学习方法因其强大的特征学习和语义理解能力，在特征自动学习、命名实体识别等方面表现出显著优势[14]。张颖怡等[9]通过数据增强进行公式化脱敏，结合实体边界加强和深度学习模型对学术论文中的问题实体和方法实体进行识别，通过十折交叉验证，得出F1值高于基准模型3.69个百分点。Cohan等[15]介绍了一种基于引用信息的Transformer模型，该模型通过对学术文献进行文档级表示学习，捕获文献中的全局语义和结构信息，从而能够支持研究问题的识别和分类，并在14.7万篇学术论文的数据集上取得了80%的F1值。基于深度学习的方法能够自动学习文本特征，适用于复杂语境和多层次语义分析。然而，这类方法对标注数据依赖度较高，在标注稀缺的场景下表现有限。随着BERT、BioBERT等预训练模型的出现，基于预训练模型的研究问题分类和识别也被广泛应用。Lee等[16]展示了在大规模生物医学语料库上预先训练的特定领域的语言表示模型，其实体识别任务表现较BERT模型提高0.62个百分点，验证了BioBERT模型在生物医学领域研究问题文本挖掘中的有效性。何丽等[17]将问题分为ABBR、DESC、ENTY等6类，利用BERT模型的语义交互性来对研究问题进行分类并选择期望答案，在3个数据集上MAP和MRR指标都有明显提升。但深度学习方法前期训练往往需要大量高质量标注语料，预训练模型通常适用于通用领域，对于特定领域的针对性欠缺，且在较小文本或特定领域中易出现过拟合等现象[18]。

综上，问题实体识别研究已经积累一定的研究成果但仍存在不足，特别是在中文预训练模型匮乏的情况下，针对特定领域的问题实体识别存在标注数据不足、模型方法复杂等问题。本文提出一种结合半监督学习和规则校正的中文问题实体识别方法。以条件随机场（Conditional Random Field，CRF）模型为基准模型，使用少量标注数据和大量未标注数据进行训练，在有监督学习基础上充分利用无监督特征形成半监督学习，并结合语言学规则对半监督学习结果进行校对处理，提高小规模数据下中文问题实体识别的准确性，为缺乏标注数据的实体识别研究提供参考借鉴。

2　研究方法

2.1　研究框架

CRF模型可以使用字、词、词性等上下文特征，也可以将任意相关知识源融入文本特征以增益识别效果，提升实体识别的准确性和鲁棒性[19]，因此以CRF模型[20]为基础模型，结合半监督学习和规则校正，利用有监督特征词性特征和指示特征，同时引入无监督特征和规则，以实现问题实体识别的目标。在本研究中，未标注语料的数量通常远多于标注语料，且其构建成本低，因此适用于生成全局性、统计性强的无监督特征。而待标注语料通过人工标注提供了更加精准的监督信号，用于构建词性特征和指示特征。研究框架（见图1）主要分为5个部分。①无监督特征构建。进行语料词语词向量的模型训练和数据集词频-逆文档频率（Term

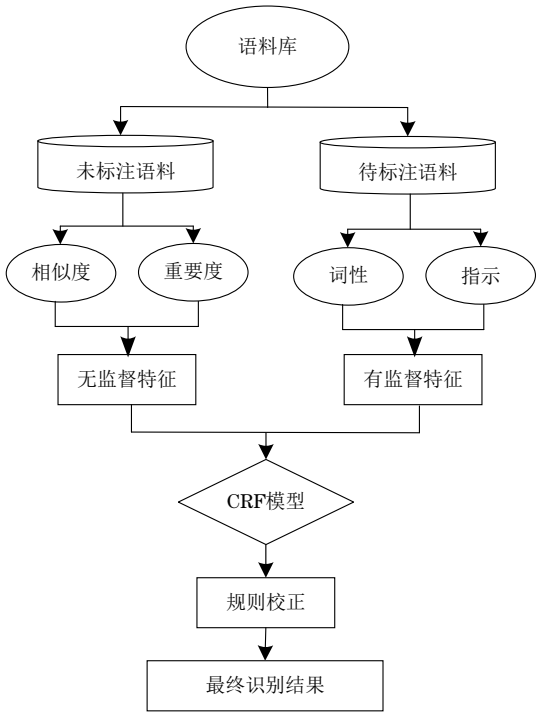

图1　研究框架

Frequency-Inverse Document Frequency，TF-IDF）的计算。词向量表达的是文本相似度，TF-IDF表达的是文本重要程度。②有监督特征构建。选取词性特征和指示特征作为有监督特征集。③CRF特征模板定义。融合不同特征和上下文信息来定义特征模板。④模型对比。添加不同外部特征的组合模型与基准模型以及其他模型进行对比分析。⑤规则库构建及结果校正。对中文学术论文文本信息的结构、语法和用词习惯进行总结和描述，进而得到文本的语言学规则结构，利用规则对问题实体进行定义和校正。

2.2　无监督特征构建

问题实体相较于一般命名实体有3个方面特点：①实体类型多样，问题描述的实体不限于常见的人名、地名、机构名等，还涉及事件、专业术语或词汇等；②实体边界模糊，由于特定领域的限制，容易出现较多的长实体，导致实体边界难以识别；③存在低频实体问题，在表述问题的实体中存在较多频率低但具有重要意义的实体，这些低频实体往往因样本不足或表达稀疏而难以准确识别。已有研究表明word2vec模型和TF-IDF算法在领域数据上的适用性[21-22]。因此，本研究从数据的领域特征出发，通过词向量计算语料词语与领域主题词的语义相似度，同时计算语料词语的TF-IDF值来表明语料词语的重要度，尝试基于结合特定领域内语义相似度和重要度的无监督特征来解决上述问题。

2.2.1　相似度特征

以待识别词语与领域主题词的语义相似度为无监督特征。相较于BERT模型需要较多计算资源和时间，word2vec的计算更简单高效，其中连续词袋（Continuous Bag of Words，CBOW）模型输入的是上下文中的词语，即根据上下文预测目标词语，更适合用于一般的语义表示[23]。因此，使用word2vec[24]作为词向量训练工具，使用Jieba工具进行分词、停用词等处理，后续经过训练得到词向量模型。为验证模型训练结果的可靠性，以训练好的“共享经济”领域词向量模型为例，提取“共享经济”主题词相似度前10的词语，结果见表1。

表1　“共享经济”主题词相似度前10词语

序号	词语	相似度	序号	词语	相似度
1	共享单车	0.966	6	Aribnb	0.952
2	分享经济	0.965	7	滴滴出行	0.947
3	商业模式	0.958	8	网约车	0.944
4	共享经济平台	0.955	9	共享平台	0.943
5	闲置资源	0.955	10	共享经济模式	0.940

通过计算待识别词$\boldsymbol{A}$与主题词$\boldsymbol{B}$的余弦相似度来测量词语间的语义相似度[25]，余弦值越大则语义越相似，$\boldsymbol{A}$与$\boldsymbol{B}$的语义相似度计算如式（1）所示。

$$\cos(\boldsymbol{A},\boldsymbol{B})=\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|}=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\sqrt{\sum_{i=1}^{n}B_i^2}} \quad (1)$$

式中：$\boldsymbol{A}$与$\boldsymbol{B}$分别代表待识别词和主题词向量，$\boldsymbol{A}=(A_1,A_2,\cdots,A_n)$，$\boldsymbol{B}=(B_1,B_2,\cdots,B_n)$，$n$代表词向量维度，$A_i$、$B_i$代表词向量在第$i$个维度上的分量。

2.2.2　重要度特征

词语重要度能有效传达该词的语义属性[26]，为判断其是否属于研究问题提供辅助信息。在本研究中，重要度特征通过TF-IDF计算，反映了词语在领域语料中的语义显著性，结合上下文的其他特征（如词性和相似度），可以更全面地评估词语是否为研究问题，计算方

法如式（2）所示。

$$\begin{aligned}\text{TF-IDF}(w) &= \text{TF}(d,w)\cdot\text{IDF}(w)\\ &= \text{TF}(d,w)\cdot\log\frac{N}{\text{DF}(w)+1}\end{aligned} \quad (2)$$

式中：TF(d,w)代表单词w在当前文本d中的词频，N代表文本集合中的文本总数，DF(w)代表文本集合中出现单词w的文本数。

2.3　有监督特征构建

从词性特征和指示特征两个方面构建有监督特征。其中，词性特征是描述输入序列中每个词的词性信息[27]，参考具有现代汉语词类标记规范的NLPIR词性标注集，对文本数据进行词性标注。指示特征是标识问题实体可能存在位置的指示信息。问题实体的语言描述存在一定的规律性，如“本文”“针对”“解决”“分析”等词语的上下文可能出现对应的问题实体。因此，构建问题实体的指示词词典，以期提高问题实体识别效果。在标注过程中，若该词语存在在指示词词典中，则标记为“Y”，若不存在则标记为“N”。

2.4　CRF特征模板定义

利用CRF++工具构建CRF模型，其中特征模板定义是整合不同特征以实现最佳识别效果的关键，定义的CRF特征模板见表2。

2.5　规则库构建

通过对数据文本的语言规律进行归纳总结，并参考其他学者的研究成果[5-7]，制定问题实体描述和问题实体校正两类规则。

2.5.1　问题实体描述规则

在索传军等[8]给定的问题描述框架基础上，参考其他问题定义表述，设计理解型、解决型、探究型3类问题实体描述规则，见表3。

表3　问题实体描述规则

问题类型	描述规则
理解型	本研究/本文+试图/关注/解决/回答/探索+（问题/任务）
	是否/是什么/为什么+旨在/具有/促进+（问题/任务）
解决型	针对/为了解决+（问题/任务）+提出/构建+方法/模型
	本文/本研究/+[拟/试图]+采用/运用/基于+方法/技术+探究/分析+（问题/任务）
探究型	本研究/本文/笔者+方法+揭示/探讨+关系/规律/机理/异同

2.5.2　问题实体校正规则

通过对预实验错误情况进行分析和总结，发现模型存在遗漏部分实体的问题，如“针对船舶建造过程中的构件流转和船舶舾装问题”句子中，“构件流转”和“船舶舾装”均是描述研究问题的实体，模型却只识别出“船舶舾装”为问题实体。为了进一步提升模型对问题实体的识别效果，根据摘要文本的语言结构规律设计了校正规则，如一些前向、并列、标点符号等标识信息会对问题实体界定起到一定的校正作用，根据这一语言结构规律并综合有监督特征中的指示特征构建指示词词典，见表4。

表4　指示词词典（示例）

标识类型	标识词
前向标识W_F	从、对、由于、基于、分析、研究、采用、针对、运用
并列标识W_C	和、与、及、乃至、或、并、同
标点标识W_P	顿号、书名号

表2　CRF特征模板

模板类型	模板	含义
一元模板	$t_i(-w<i<w,i,t\in z)$	当前词t_i及上下w窗口的词
	$F_{ij}(-w<j<w,i,j=1,t\in z)$	上下w窗口内词的特征
	$F_{(i-1)j}\|F_{ij}\|F_{(i+1)j}(-w<i<w,i,j=1,t\in z)$	当前词特征及上下w窗口词特征的组合
复合模板	$t_0\|t_i(-w<i<w,i,t\in z)$	当前词及上下w窗口内词的组合
	$t_0\|F_{ij}(i=0,j=1,2,3,4)$	当前词与某个特征的组合
	$F_{(i-1)j}\|F_{ij}\|F_{(i+1)j}(-w<i<w,j=1,2,3,4,t\in z)$	当前词的第j个特征的上下w窗口特征的组合
	$F_{ij1}\|F_{ij2}(-w<i<w,j_1,j_2=1,2,3,4,t\in z)$	当前词的某个特征与其他特征的组合

设定分词后的字符串W_q、W_d、W_h：W_q表示前一个词，W_d表示当前词，W_h表示后一个词，词串之间无字符级间隔。定义E_i为模型识别出的实体名称，“∈”表示包含于，“∧”表示且关系，“∨”表示或关系。制定如下校正规则。

规则1：$(W_q \in W_F) \rightarrow (W_d \in E_i)$

规则2：$(W_q \in E_i) \wedge (W_d \in W_C) \rightarrow (W_h \in E_i)$

规则3：$[(W_q \in E_i) \wedge (W_d \in W_P) \rightarrow (W_h \in E_i)] \vee [(W_q \in W_P) \wedge (W_h \in W_P) \rightarrow (W_d \in E_i)]$

3 实验结果及分析

3.1 实验数据与评价指标

3.1.1 语料库构建

当前中文问题实体识别领域尚未有公开数据，因此选择自建测试语料。学术论文的摘要部分是整篇文章的凝练，能帮助读者快速了解文献研究问题并评估文献价值[28]，因此选取学术论文摘要部分作为实验语料。以中国知网为数据源，以社会科学领域的“共享经济”和自然科学领域的“船舶建造”为实例主题，各选取3 000篇论文摘要作为未标注语料库，用于后续词向量训练和TF-IDF的计算。为保证数据的时效性和代表性，从2018—2022年的论文中选取，同时为验证小规模标注数据的主题有效性，从两个领域内各选取500篇学术论文摘要作为待标注语料，并根据问题描述规则对问题实体进行标注。

无监督特征标注中，在中国知网分别以“共享经济”和“船舶建造”为主题进行检索，得到两个领域词频排名前10的主题词（中国知网主题栏提供）作为主题词词典，通过word2vec向量化后分别计算待识别词与主题词词典中所有词的语义相似度，取最大值作为待识别词的相似度。CRF模型中的特征值通常是离散数值，因此需要将相似度值和重要度值映射到有限区域空间内，一般是0~9区间内的离散特征输入，0、6、9分别代表低、中、高3个区间差异[25]。在实验中对相似度和重要度结果进行统计分析发现，相似度大体以0.5和0.6为界分成3个类团，同时通过人工判断验证相似度小于0.5的词与主题弱相关，大于0.6的词语与主题高度相关，因此相似度以0.5和0.6为界分成3个区间；分析重要度分布后发现重要度低于5的词语对问题表述贡献较小，高于10的词语是领域高频核心词汇，因此重要度以5和10为界分成3个区间。综上，设计了特征值映射规则，见表5。

表5 特征值映射规则

相似度s		重要度k	
区间	离散值	区间	离散值
$s \geqslant 0.6$	9	$k \geqslant 10$	9
$0.5 \leqslant s < 0.6$	6	$5 \leqslant k < 10$	6
$s < 0.5$	0	$k < 5$	0

最终处理后的数据集部分标注数据见表6。其中：词性列的v、n、p和vn分别代表动词、名词、介词和动名词；词典列的Y和N代表文本词段是否存在在指示词词典中；相似度与重要度则是根据映射规则标记的，相似与重要程度从9、6到0依次递减；标签列的B-WT和E-WT代表实体的开头和结尾，M-WT代表实体的中间部分，O代表其他非实体文本。实验过程中将数据集按4∶1分为训练集和测试集。

表6 标注集样例

原文	词性	词典	相似度	重要度	标签
共	v	N	9	9	B-WT
享	v	N	9	9	M-WT
经	n	N	9	9	M-WT
济	n	N	9	9	E-WT
对	p	Y	0	0	O
绿	n	N	9	9	B-WT
色	n	N	9	9	M-WT
消	vn	N	9	9	M-WT
费	vn	N	9	9	E-WT

3.1.2 评价指标

采用实体识别模型主流的评价指标准确率P、召回率R和F1值来对实验结果进行评估[29]，其计算如式（3）~式（5）所示。

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \times 100\% \quad (3)$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \times 100\% \quad (4)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (5)$$

式中：N_{TP}表示模型正确识别出的样本数量；N_{FP}表示模型错误地识别为正样本的负样本数量；N_{FN}表示模型错误地识别为负样本的正样本数量。

3.2　结果分析

3.2.1　基准模型及窗口选取

上下文窗口w指的是当前词t及前后若干字符组成的w窗口内的字符串。窗口越大，上下文可利用的信息越多，但效率变低；窗口越小，对于特征的信息利用率可能越低，导致语义信息丢失。有研究指出中文语料信息中99%的词语是由5个及5个以下的字构成的[30]，考虑到中文词语较少以单个字的形式出现，在不添加特征的基础上验证不同窗口w取值的模型效果，结果见表7。当w为2时模型性能最好，因此模型窗口w取值为2。

表7　基准模型窗口选取

单位：%

窗口w取值	准确率	召回率	F1值
2	86.28	69.75	77.14
3	86.25	66.74	75.25
4	85.63	63.93	73.21
5	85.65	65.58	74.66

3.2.2　一元特征识别效果

引入词性、指示、相似度、重要度一元特征进行实验，其结果见表8。

表8　“共享经济”和“船舶建造”领域一元特征识别结果

单位：%

序号	特征	“共享经济”领域			“船舶建造”领域		
		准确率	召回率	F1值	准确率	召回率	F1值
1	基准	86.28	69.75	77.14	88.47	76.01	81.77
2	词性	85.35	75.74	80.26	87.83	79.41	83.41
3	指示	85.69	71.57	78.00	88.89	76.33	82.13
4	相似度	85.15	73.21	78.73	88.80	77.10	82.54
5	重要度	85.11	74.12	79.24	88.09	78.96	83.27

如表8所示，添加词性、指示、相似度和重要度一元特征的F1值均有提升。其中，添加词性特征的F1值最高。通过词性标注能够捕捉实体的语法角色，为识别提供额外线索，相比基准模型，添加词性特征后F1值分别提升3.12个百分点和1.64个百分点，可见词性特征对模型性能的影响最大。对相似度和重要度无监督特征进行分析发现，指示特征的提升幅度略逊于无监督语义特征，可能与特定领域指示词覆盖范围不足有关。无监督特征无需人工标注，通过词向量和词频计算，显著降低了标注数据不足的影响，说明在不进行人工标注的情况下，无监督特征可以实现甚至超越部分有监督特征的性能指标，可以极大节省人工标注的时间和成本。另外，相较于“共享经济”这一社会科学领域，由于“船舶建造”领域论文实体边界更清晰、歧义性小等，基准模型的识别效果优于“共享经济”领域，导致加入特征后“船舶建造”领域模型性能的提升幅度较小。

3.2.3　复合特征识别效果

在一元特征实验的基础上，以特征组合方式验证模型性能。鉴于在添加词性特征后获得了最佳识别性能，选择词性特征与其他特征的组合作为复合特征，通过调整特征模板得到不同复合特征组合的最佳识别效果，见表9。

如表9所示，复合特征模型的F1值均高于单一特征模型的F1值。“词性+指示”和“词性+相似度”的组合特征效果无明显差异，而“词性+重要度”相较于“词性+指示”的模型效果在两个领域中分别提高0.61个百分点和0.40个百分点；另外，“词性+重要度”组合特征的性能优于词性与其他特征的组合，说明词语重要度对问题实体有更强的区分能力，词性为模型提供了语法层面的基础支撑，而重要度通过区分词语的重要性，在领域范围内能对目标实体进一步区分。“相似度+重要度”无监督组合特征相比“词性+指示”有监督组合特征在

表9　不同领域复合特征识别结果

单位：%

序号	特征	“共享经济”领域			“船舶建造”领域		
		准确率	召回率	F1值	准确率	召回率	F1值
1	基准	86.28	69.75	77.14	88.47	76.01	81.77
2	词性	85.35	75.74	80.26	87.83	79.41	83.41
3	词性+指示	85.50	75.99	80.47	88.43	79.41	83.68
4	词性+相似度	85.96	75.43	80.35	88.09	79.76	83.71
5	词性+重要度	84.54	77.89	81.08	87.53	80.90	84.08
6	相似度+重要度	86.09	75.54	80.35	87.85	79.70	83.58
7	词性+指示+相似度+重要度	88.21	78.77	83.23	89.56	80.92	85.02

两个领域的F1值也仅降低了0.12个百分点和0.10个百分点，说明在模型任务中部分有监督特征能够被无监督特征所替代甚至超越，进而在减少标注工作量的同时，以更为出色的方式完成模型任务。融合4个特征时模型的识别效果最优，相较于基准模型在两个领域的F1值分别提高了6.09个百分点和3.25个百分点。

3.2.4　与其他模型识别效果对比

目前，主流深度学习组合模型BiLSTM-CRF[31]、BERT-BiLSTM-CRF[32]、BiLSTM-IDCNN-CRF[33]等和微调通用预训练模型BERT、ERNIE等在实体识别任务中通常表现出色。因此，在自建语料集上将上述模型与所提模型进行对比分析。另外，随着大语言模型的不断迭代进步，已有学者将大语言模型应用于实体识别任务，选取目前主流且先进的GPT-4o模型进行实体识别。所提模型与深度学习模型、微调预训练模型和GPT-4o的对比结果见表10。

如表10所示，在自建语料集上，由于标注数据规模有限或深度学习模型在领域数据上的特征分布差异，

表10　不同模型的性能对比

单位：%

方法	“共享经济”领域			“船舶建造”领域		
	准确率	召回率	F1值	准确率	召回率	F1值
所提模型	88.21	78.77	83.23	89.56	80.92	85.02
BiLSTM-CRF	61.17	55.17	58.02	68.46	59.65	63.75
BERT-BiLSTM-CRF	68.87	69.83	69.35	73.58	76.93	75.22
BiLSTM-IDCNN-CRF	81.34	75.06	78.37	86.11	82.24	84.22
微调ERNIE	71.50	65.17	68.19	72.15	70.12	71.12
微调BERT	73.86	71.13	72.47	77.27	76.87	77.07
GPT-4o	63.87	55.54	59.42	72.36	61.11	66.26

BiLSTM-CRF和BERT-BiLSTM-CRF未能实现较好的识别效果，BiLSTM-IDCNN-CRF模型的识别效果较好，但相对所提模型效果略差一些。基于训练集微调的BERT和ERNIE整体性能趋于稳定，但在特定任务上性能仍有欠缺，不及所提模型。在提供基础系统信息（任务说明、实体定义和输出格式规范）的情况下，GPT-4o识别效果也落后于所提模型。大语言模型虽具备强大的分析生成能力，但通常需要进行微调或融入情境学习才能更好地适应目标任务，尤其是在实体识别领域，仅在任务基础说明条件下，其结果往往落后于基准模型[34]。相较于选取的3个深度学习模型的最优性能，所提模型在两个领域的F1值分别提高了4.86个百分点和0.80个百分点，由此证明所提的半监督模型性能较优。

3.2.5　规则校正

规则校正可以有效地弥补半监督模型不能表达语言学特征的缺点。对半监督模型的错误结果进行分析，如：“区块链和经济税收策略”，半监督模型识别出“区块链”实体，却未识别出“经济税收策略”实体；“研究智慧城市、绿色消费”，模型识别出“智慧城市”实体，

未识别出“绿色消费”实体。利用前述问题实体校正规则对结果进行校正，实验结果见表11。

添加规则对模型结果修正后，“共享经济”领域的F1值提高了2.59个百分点，相比基准模型提高了8.68个百分点；“船舶建造”领域的F1值提高了1.36个百分点，相比基准模型提高了4.61个百分点。由此验证提出的结合半监督学习与规则校正的问题实体识别方法的有效性和较优性能。

表11　规则校正后的实验结果

单位：%

组合	“共享经济”领域			“船舶建造”领域		
	准确率	召回率	F1值	准确率	召回率	F1值
基准	86.28	69.75	77.14	88.47	76.01	81.77
词性+指示+相似度+重要度	88.21	78.77	83.23	89.56	80.92	85.02
词性+指示+相似度+重要度+规则	88.78	83.06	85.82	90.17	82.89	86.38

3.3　稳健性分析

已有研究在领域小规模数据指导下进行领域实体关系识别和抽取[35]。为验证所提模型在小规模领域数据下的适用性，分别取原训练集的1/2和1/4作为实验数据集，在两个领域数据上分别与深度学习模型和微调预训练模型进行对比，以验证所提方法在小规模数据上的稳健性能，结果见表12。

表12　“共享经济”和“船舶建造”领域稳健性分析

单位：%

领域	方法	全数据集			1/2数据集			1/4数据集		
		准确率	召回率	F1值	准确率	召回率	F1值	准确率	召回率	F1值
共享经济	BiLSTM-IDCNN-CRF	81.34	75.06	78.37	69.49	65.16	67.17	64.41	57.29	60.54
	微调BERT	73.86	71.13	72.47	70.23	66.23	68.17	62.38	60.18	61.26
	所提模型	88.21	78.77	83.23	86.00	76.16	80.78	83.71	70.60	76.60
船舶建造	BiLSTM-IDCNN-CRF	86.11	82.24	84.22	66.93	63.94	65.77	67.32	60.26	63.64
	微调BERT	77.27	76.87	77.07	73.61	72.77	73.19	66.82	63.71	65.23
	所提模型	89.56	80.92	85.02	86.19	78.32	82.06	86.78	69.77	77.35

在训练数据大幅缩减的情况下，深度学习模型的实体识别性能下滑严重，在1/2数据集上F1值下降约10个百分点，在1/4数据集上下降近20个百分点。深度学习模型依赖大规模、高质量标注数据，小规模数据集的特征分布稀疏问题导致模型难以学习到稳定内核。基于微调的预训练模型对通用语料进行预训练，性能趋于稳定，在1/2和1/4数据集上F1值仅下降约4个百分点和12个百分点，但在面对特定领域的数据时，可能需要更大规模的领域标注数据来充分适配。所提方法结合无监督特征，对标注数据依赖性较低，在小规模数据集上依然能够保持稳定性能，且规则校正进一步解决了模型因标注不足产生的错误识别问题，所提模型F1值仅下降了约3个百分点和7个百分点，性能差距进一步扩大。在“船舶建造”领域，问题实体边界较清晰，语义一致性较高，因此模型稳定性相对更强。相较于深度学习和预训练微调模型依赖大规模训练集，所提方法在小规模领域数据上依然能达到较优识别效果，领域规则校正进一步增强对小规模标注数据的适应性，可验证所提方法在小规模标注数据上具有稳健性和鲁棒性。

4　结语

本研究针对中文学术论文的问题实体识别需求，提出一种结合有监督特征与无监督特征的半监督学习模型，在小规模标注数据的场景下显著提升了问题实体识别性能。并且通过学术论文摘要的语言学特征制定规则校正方法，有效弥补了模型在实体边界识别上的不足，使最终识别效果进一步提升。实验结果表明，所提方法在“共享经济”和“船舶建造”两个领域的小规模数据集上具有较高的识别准确性和稳健性，其F1值优于主流深度学习模型和大语言模型，验证了方法的通

用性和实际应用价值。通过识别学术论文问题实体，可以快速定位研究问题句和问题实体，帮助研究人员检索和发现所需的文献信息，从而优化学术资源的利用效率，能够提升信息资源建设的质量和效果，为以问题为中心的信息检索和发现提供研究思路和基础。

后续研究中，将进一步对词向量的上下文感知性和多语言表示学习的领域自适应性进行探究，提升TF-IDF的聚类规模和精准度，进一步探索无监督算法在小规模领域数据中的实体识别研究，针对标注数据的高成本和稀缺性提供现实的解决方法。然而，本研究在大规模标注数据和多领域泛化方面仍存在局限性，未来研究可结合生成式预训练模型和跨领域迁移学习方法，进一步提升模型的适应性和鲁棒性，为细分领域的实体识别研究提供更强有力的支持。

参考文献

[1] 祁鹏年，廖雨伦，覃飙. 基于深度学习的中文命名实体识别研究综述[J]. 小型微型计算机系统，2023，44（9）：1857-1868.

[2] 王颖洁，张程烨，白凤波，等. 中文命名实体识别研究综述[J]. 计算机科学与探索，2023，17（2）：324-341.

[3] GRISHMAN R，SUNDHEIM B. Message Understanding Conference-6：a brief history[C]//Proceedings of the 16th Conference on Computational Linguistics. Morristown：Association for Computational Linguistics，1996：266-271.

[4] 焦凯楠，李欣，朱容辰. 中文领域命名实体识别综述[J]. 计算机工程与应用，2021，57（16）：1-15.

[5] 张颖怡，章成志，HE D Q. 学术论文中问题与方法识别及其关系抽取研究综述[J]. 图书情报工作，2022，66（12）：125-138.

[6] 程齐凯，李信. 面向语义出版的学术文本词汇语义功能自动识别[J]. 数字图书馆论坛，2017（8）：24-31.

[7] KOVAČEVIĆ A，KONJOVIĆ Z，MILOSAVLJEVIĆ B，et al. Mining methodologies from NLP publications：a case study in automatic terminology recognition[J]. Computer Speech & Language，2012，26（2）：105-126.

[8] 索传军，赖海媚. 学术论文问题知识元的类型与描述规则[J]. 中国图书馆学报，2021，47（2）：95-109.

[9] 张颖怡，章成志. 基于公式化表达脱敏与边界识别加强的学术论文研究问题与方法识别研究[J]. 情报学报，2024，43（6）：712-732.

[10] 赵继贵，钱育蓉，王魁，等. 中文命名实体识别研究综述[J]. 计算机工程与应用，2024，60（1）：15-27.

[11] RAU L F. Extracting company names from text[C]//[1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application. New York：IEEE Press，1991：29-32.

[12] GUPTA S，MANNING C D. Analyzing the dynamics of research by extracting key aspects of scientific papers[J]. IJCNLP 2011-Proceedings of the 5th International Joint Conference on Natural Language Processing，2011：1-9.

[13] 纪科，张秀，马坤，等. 基于关键实体和文本摘要多特征融合的话题匹配算法[J]. 郑州大学学报（工学版），2024，45（2）：51-59.

[14] 邓依依，邬昌兴，魏永丰，等. 基于深度学习的命名实体识别综述[J]. 中文信息学报，2021，35（9）：30-45.

[15] COHAN A，FELDMAN S，BELTAGY I，et al. SPECTER：document-level representation learning using citation-informed transformers[EB/OL]. [2024-11-12]. https://arxiv.org/abs/2004.07180v4.

[16] LEE J，YOON W，KIM S，et al. BioBERT：a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics，2020，36（4）：1234-1240.

[17] 何丽，张家铭，徐丽闪，等. 基于问题分类和深度模型的答案选择算法[J]. 计算机工程与设计，2023，44（5）：1412-1418.

[18] 张汝佳，代璐，王邦，等. 基于深度学习的中文命名实体识别最新研究进展综述[J]. 中文信息学报，2022，36（6）：20-35.

[19] 栗伟，赵大哲，李博，等. CRF与规则相结合的医学病历实体识别[J]. 计算机应用研究，2015，32（4）：1082-1086.

[20] JOHN L，ANDREW M，PEREIRA FERNANDO C N. Conditional random fields：probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the Eighteenth International Conference on Machine Learning. California：Morgan Kaufmann Publishers Inc，2001：282-289.

[21] 张清慧，陈谊，武彩霞. 基于词表示模型的领域文献数据可视分析方法[J]. 图学学报，2022，43（4）：685-694.

[22] 董伟，董思遥，王聪，等. 基于TF-IDF算法和DTM模型的网络学习社区主题分析[J]. 现代教育技术，2022，32（2）：90-98.

[23] MIKOLOV T，SUTSKEVER I，CHEN K，et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems，2013，26：3111-3119.

[24] TOMAS M. Word2vec project[EB/OL]. [2024-06-18]. https://code.google.com/p/word2vec/.

[25] 包振山，宋秉彦，张文博，等. 基于半监督学习和规则相结合的中医古籍命名实体识别研究[J]. 中文信息学报，2022，36（6）：90-100.

[26] 武永亮，赵书良，李长镜，等. 基于TF-IDF和余弦相似度的文本分类方法[J]. 中文信息学报，2017，31（5）：138-145.

[27] 孙静，李军辉，周国栋. 基于条件随机场的无监督中文词性标注[J]. 计算机应用与软件，2011，28（4）：21-23，46.

[28] 赵丹宁，牟冬梅，白森. 基于深度学习的科技文献摘要结构要素自动抽取方法研究[J]. 数据分析与知识发现，2021，5（7）：70-80.

[29] LI J，SUN A X，HAN J L，et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering，2022，34（1）：50-70.

[30] KIM J D，OHTA T，TSURUOKA Y，et al. Introduction to the bio-entity recognition task at JNLPBA[C]//Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications-JNLPBA’04. Morristown：Association for Computational Linguistics，2004：73-78.

[31] 张应成，杨洋，蒋瑞，等. 基于BiLSTM-CRF的商情实体识别模型[J]. 计算机工程，2019，45（5）：308-314.

[32] 谢腾，杨俊安，刘辉. 基于BERT-BiLSTM-CRF模型的中文实体识别[J]. 计算机系统应用，2020，29（7）：48-55.

[33] 蒋翔，马建霞，袁慧. 基于BiLSTM-IDCNN-CRF模型的生态治理技术领域命名实体识别[J]. 计算机应用与软件，2021，38（3）：134-141.

[34] WANG S H，SUN X F，LI X Y，et al. GPT-NER：named entity recognition via large language models[EB/OL]. [2024-06-16]. https://arxiv.org/abs/2304.10428.

[35] 陈果，许天祥. 小规模知识库指导下的细分领域实体关系发现研究[J]. 情报学报，2019，38（11）：1200-1211.

作者简介

傅柱，男，博士，副教授，研究方向：知识组织与挖掘。
邱畅唱，男，硕士研究生，研究方向：自然语言处理。
刘鹏，男，博士，副教授，通信作者，研究方向：复杂网络分析，E-mail：liupeng19821017@126.com。

Problem Entity Recognition of Chinese Academic Paper Combining Semi-Supervised Learning and Rule Correction

FU Zhu　QIU ChangChang　LIU Peng
(School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang 212100, P. R. China)

Abstract: In order to quickly locate and identify research problems in academic papers, this paper proposes a problem entity recognition method combining semi-supervised learning and rule correction for Chinese academic papers. First, based on the framework of conditional random field model, supervised features such as parts of speech and deixis and unsupervised features such as similarity and importance are constructed. Then, the model recognition effects under different feature combinations are compared, and the recognition results are proofread according to domain linguistics rules. Finally, the subject areas of “sharing economy” and “ship construction” are taken as examples for empirical research. The entity recognition performance of the proposed method is better than that of the mainstream deep learning model and the pre-trained model such as the large language model. The F1 score of the two domain subject datasets reaches 85.82% and 86.38%, respectively, and the performance advantage on the 1/2 and 1/4 datasets is further expanded. It shows that the proposed method can identify the problem entities of Chinese academic papers well on small-scale labeled datasets in different fields, and shows good validity and robustness.

Keywords: Problem Entity Recognition; Conditional Random Field; Semi-Supervised Learning; Rule Correction; Small Scale Data

（责任编辑：管清漾）