基于特征融合的术语型引用对象自动识别方法研究*

马　娜[1,2]　张智雄[1,2,3,4]　吴朋民[5]

[1]（中国科学院文献情报中心　北京 100190）

[2]（中国科学院大学经济管理学院图书情报与档案管理系　北京 100190）

[3]（中国科学院武汉文献情报中心　武汉 430071）

[4]（科技大数据湖北省重点实验室　武汉 430071）

[5]（中国科学院自动化研究所　北京 100190）

摘要：【目的】设计特征融合和伪标签降噪策略，探索科技论文术语型引用对象自动识别方法。【方法】将术语型引用对象识别转换为序列标注问题，在BiLSTM-CNN-CRF输入层融合术语型引用对象的语言学和启发式两大类特征，增强引用对象的特征表示，设计伪标签学习降噪机制，采用半监督学习方法探究不同特征组合对识别效果的影响。【结果】本方法在术语型引用对象识别任务中最优F1值达到0.6018，比BERT模型实验结果提升8%。【局限】实验数据仅涉及计算机领域，在其他领域的可移植性有待考证。【结论】基于特征融合的深度学习方法在术语型引用对象的识别中有较好性能，伪标签学习方法解决了引用对象标注数据不足的问题，两者结合有效地探索了术语型引用对象自动化识别方法。

关键词：引用对象识别　特征融合　伪标签学习　BiLSTM-CNN-CRF

分类号：TP391

DOI：10.11925/infotech.2096-3467.2019.0869

引用本文：马娜，张智雄，吴朋民．基于特征融合的术语型引用对象自动识别方法研究[J]．数据分析与知识发现，2020，4(1)：89-98.(Ma Na, Zhang Zhixiong, Wu Pengmin. Automatic Identification of Term Citation Object with Feature Fusion[J]. Data Analysis and Knowledge Discovery, 2020, 4(1): 89-98.)

1　引　言

科技论文是公开表达展示科技创新活动、科研创新成果的重要载体之一，引用行为是科技论文的重要特征，反映了科研工作之间复杂而本质的关联关系。随着计算机技术的发展，研究者尝试利用文本挖掘技术和语义分析方法，深入引用内容，对作者引用学术著作时的态度、功能、目的等进行多角度分析，揭示学术著作间研究内容上的关联性。有学者认为，基于引用内容的引文分析将成为下一代引文分析的方向[1-2]。

引用对象是引用内容分析的重要研究单元，相比其他引用内容分析，引用对象直接地揭示了文献间的本质联系。引用对象概念雏形最早由Small[3]在1978年提出，即作者引用文献时提及引文的具体科学内容，以一种泛化的“概念符号”（Concept Symbols）描述引文中的概念或方法。引用对象是引

通讯作者：张智雄，ORCID：0000-0003-1596-7487，E-mail：zhangzhx@mail.las.ac.cn。

*本文系中国科学院基金项目“科技文献丰富语义检索应用示范”（项目编号：院1734）的研究成果之一。

文内容的一部分，但不包含作者对内容的主观评价。引用对象从内容层面展示引文的利用价值和学术贡献，与引文分类相比，可以更加直接地解释作者的引用行为，具有重要价值。通过识别引用对象，结合引用关系构建深度语义化的引文网络，可深入语义层面挖掘论文之间的潜在关联，从同行评议的角度帮助研究者快速发现论文的重要价值，为学术成果贡献分析、交叉学科链接点挖掘、知识发现、人才评价等提供重要维度，因此，引用对象的识别研究具有重要的理论意义和实际应用价值。

引用对象自动识别是一种基于非结构化数据的信息抽取任务，由于引用时语言模式较为灵活，许多研究者表示，引用对象自动化识别是一项比较困难的任务[4-6]。目前引用对象识别主要以人工标注识别[6-7]为主，用于证明和探索引用对象的重要作用，人工识别准确率较高，但无法满足大规模应用。机器自动化识别方面，Radoulov[8]、许德山[9]和Khalid等[10]从替代识别和直接识别两种不同角度进行了尝试，但识别效果有限。综上，引用对象自动化识别研究处于初期探索阶段，亟需在方法上进行新的探索和尝试，以满足大规模应用需要。

基于对引用对象的特征分析，本文尝试将引用对象识别转换为序列标注问题，提出基于特征融合的引用对象自动化识别方法。通过人工选取引用对象重要特征，设计特征向量表示和融合方法，为BiLSTM-CNN-CRF[11]神经网络模型提供更多的先验知识，同时为解决引用对象标注数据不足的问题，设计伪标签学习策噪音控制策略，提高算法准确性，最后通过探究不同特征以及组合特征对识别的影响，证明了本方法识别术语型引用对象的有效性。

2 相关研究

2.1 引用对象识别

针对引用对象识别，学者们已经开展了一定的研究工作。Radoulov[8]采用机器学习算法实现引用对象分类，算法融合了论文结构特征、引文位置特征、词性特征、句法特征和两类线索词，为每种对象类型单独训练朴素贝叶斯分类器，利用训练好的模型实现引文分类自动标注。许德山[9]通过计算引用句与引文原文内容的相似性，抽取引用句与引文共同含有的最大字符串直接识别引用对象。Khalid[10]则采用LDA主题识别替代直接识别引用对象的策略，集合每篇论文的全部引用句进行主题识别，利用词云为每个主题分配一个可以概括主要内容的高频术语词作为主题标签，并对相似主题进行合并，以最终的的主题标签作为引用对象。

根据上述研究，许德山[9]尝试在引用内容中抽取引用对象，但仅采用引文内容与引文标题进行字符串匹配的方法过于简单，识别效果并不理想。Khalid[10]探索利用LDA模型生成主题词代替直接识别引用对象，但该模型忽略了词与词之间的顺序，没有考虑主题词是否与引文真正相关，存在识别不准确的问题。综上，引用对象识别研究成果较少，并且在方法的通用性、扩展性和性能等方面需要较大提升，引用对象自动化识别需要更多的探索和尝试。

2.2 基于深度学习的序列标注

基于神经网络的深度学习方法在各类序列标注问题上表现出比传统机器学习方法更高效的性能。为更好地解决实际研究问题，学者们从扩展特征表示、改进模型架构和解决标注数据有限三个方面不断探索，提高模型的泛化能力。

以词向量表示文本中词语之间存在的相关关系[12]是深度学习算法的核心技术之一，学者们利用词向量表示词语的特征，弥补了人工提取特征的不足。在词向量基础上，Santos等[13]利用CNN对输入序列中每个字符进行编码，捕捉单词字符级特征，在英文语料的词性标注任务中取得97.32%的准确性。Rei等[14]在Bi-LSTM-CRF模型的基础上，使用字符向量和词向量的拼接解决序列标注问题中未登录词的问题，在向量拼接时尝试了两种不同的向量拼接方式并对比实验结果。赵洪等[15]利用Bi-LSTM-CRF模型作为基础模型，融入词性和理论术语实体特征，结合自训练方法，在理论术语抽取任务中F1值达到0.85。Zhang等[16]提出一种通过多信息实体增强语言表证的知识驱动模型，通过在大规模文本语料库和知识库上预训练语言模型，将知识信息整合进语言模型中，在自然语言处理任务中获得很好效果。

深度学习方法取得较好结果通常需要依赖于大规模高质量的标记数据，但受限于人工标注的成本，学者们尝试利用主动学习或联合学习解决标记数据

不足问题。Shen等[17]在命名实体任务上设计三种训练数据选择策略进行主动学习，结合CNN-CNN-LSTM轻量级模型，证明仅使用25%的训练语料即达到接近使用完整数据的效果。Ye等[18]在序列标注Bi-LSTM模型基础上，共用同一层词向量联合学习训练CRF层和HSCRF层，得到两个不同层级的预测标签，取Loss值较低的标签预测结果，通过实验证明了联合学习的有效性。

综上，深度学习方法在各种序列标注问题上得到广泛的发展，本文利用深度学习方法尝试解决术语型引用对象自动化识别问题，根据引用对象的特点，构建面向引用对象识别模型，设计半监督学习方法，解决标注数据不足问题。

3　识别方法

通过对引用对象的特征分析，按表现形式不同把引用对象划分为两种类型：术语型引用对象和事实型引用对象。术语型引用对象由引用句中连续的名词或名词短语组成，例如含有大写字母专有名词或缩略语，具体如Pharaoh、WordNet、BLUE；或由名词短语组成的关键短语，例如Alignment Error Rate Metric。术语型引用对象在字符、词性、语法、位置等方面具有显著特征，本文利用这些特征，构建引用对象抽取模型，通过与Baseline模型[11]和Bert模型[19]进行实验对比，证明方法的有效性。

3.1　模型选择

序列标注问题将引用句看作一个序列，输出一个等长的符号序列，其中每个符号都有特定的含义，引用对象以标签形式被符号所标记。传统的序列标注模型包括隐马尔科夫模型（Hidden Markov Model，HMM）[20]、条件随机场（Conditional Random Fields，CRF）[21]等方法，比较流行的是特征模板结合CRF的方法。特征模板通常是人工定义的一些特征函数，试图挖掘语句内部以及标注对象的语法、语义等特点，这种方式更多地依赖于特征和特征模板的选择。随着深度学习的发展，神经网络模型在解决传统自然语言处理任务中取得显著效果。神经网络模型以其强大的提取特征能力，避免了人工定义特征模版对准确性的影响。通过调研序列标注中常见的深度学习模型，最终采用能够捕捉序列中长距离上下文依赖和字符级特征的BiLSTM-CNN-CRF模型[11]作为本次实验的基础模型。长短期记忆模型（Long Short Term Memory，LSTM）是一种循环神经网络（Recurrent Neural Network，RNN），改进了RNN模型存在的梯度消失问题，用一个记忆单元替换RNN中的隐层节点。记忆单元由记忆细胞、输入门、遗忘门和输出门组成，记忆细胞主要负责记忆长时间的依赖关系，这种特殊结构可以捕捉长远的上下文信息，非常适合对基于上下文的序列数据建模。单纯的LSTM模型对句子建模时遵循从前到后的顺序，无法捕捉从后到前的信息特征，故选择双向BiLSTM以更好地捕捉前后完整的序列信息。

本文采用的基础模型BiLSTM-CNN-CRF整体框架如图1所示，由词嵌入结合字符表示作为模型的输入层，经过双向LSTM编码后，得到每个词所有标签的概率值，CRF层利用LSTM的输出以及转移概率矩阵作为输入，采用动态优化算法获得全局最优的输出序列即解决最终标签预测。

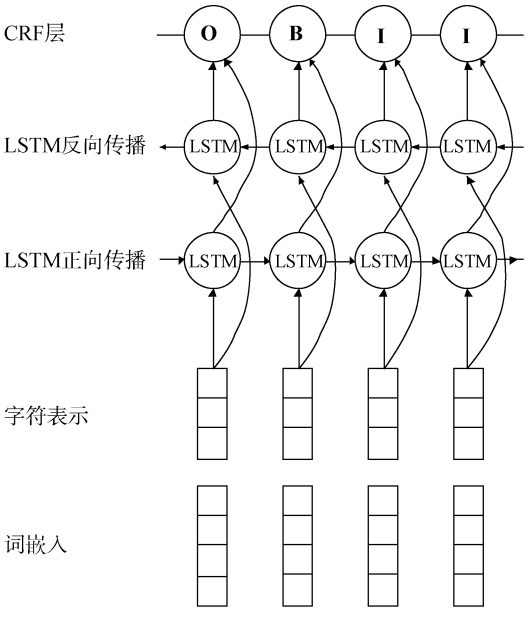

图1　BiLSTM-CNN-CRF模型

Fig.1　BiLSTM-CNN-CRF Model

模型中的字符表示层主要任务为抽取字符局部特征，例如字母大小写、前后缀等。字符表示由卷积神经网络CNN训练得到，句中单词的每一个字符通过卷积层提取数据局部特征，通过最大池化层决定

保留最具有代表性部分作为特征向量，如图2所示。字符级特征在实体识别领域有很好的辅助效果，因此引入CNN训练单词中每个字符的字符表示。

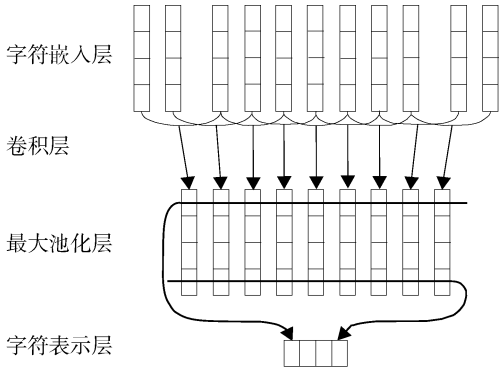

图2　CNN训练字符级特征

Fig.2　CNN Model for Extracting Char-level Representations

3.2　特征选择与表示

深度学习方法的优势在于无需特征工程，由模型自身复杂的结构完成句子级特征提取工作，但研究表明，人工特征的加入可以进一步提高模型准确性[22]。

本文在深度学习模型的基础上，融合术语型引用对象的语言学特征和启发式特征，研究这两类特征对引用对象识别的效果。术语型引用对象特征表示如图3所示，在BiLSTM的输入层分别融合了词特征向量、字符特征向量、词性特征向量、位置特征向量和标识词特征向量共计5种特征向量，增强引用对象的特征表示。基于此，引用句被表示为多元特征矩阵，再利用BiLSTM神经网络学习序列特征以及上下文语义依赖关系，使得模型对引用对象有更好的理解，进而提高预测准确性。

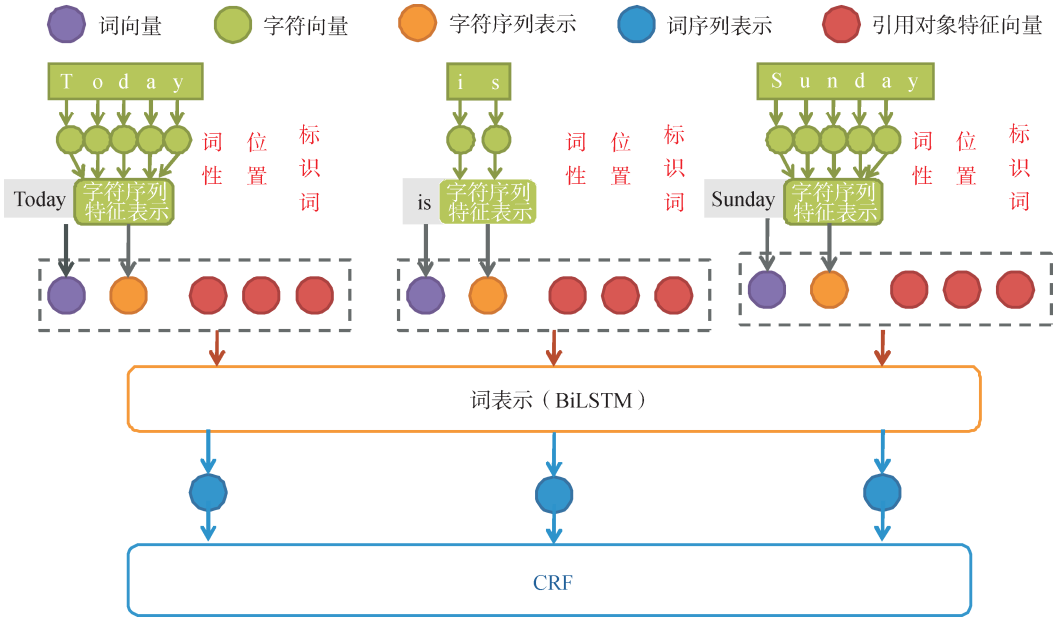

图3　术语型引用对象特征表示

Fig.3　Feature Representation of Term Citation Object

（1）语言学特征

语言学特征是指术语型引用对象在语言表达上具有的特点，主要包括词、字符以及词性三个特征。

① 词特征

词向量是将单词的语义信息分布式地表示为稠密的低维实值向量，以达到更丰富地表达单词之间语义信息关系的目的，词向量是词特征的有效表示方式，对于每一个单词，采用词向量 $W \in R^n$ 表示。已有研究结论显示利用LSTM-CNNs-CRF模型的序列标注任务中，使用同样模型的前提下，100维的GloVe词向量[23]比300维的Word2Vec效果要好[10]。本文使用GLoVe训练出来的词向量进行词向量表

示 $W_{glove} \in R^{d1}$，GLoVe 词表共计 400 000 个词，维度选择 100 维。

② 字符特征

为捕捉到更多单词本身的特征，例如大小写、前缀后缀和拼写规律等，训练字符级别特征 $W_{chars} \in R^{d2}$。字符特征向量使用卷积神经网络 CNN 进行训练，提取拼写层面特征。采用卷积核大小为 3×3，卷积核数量为 30，最终训练得到 30 维的字符级特征向量。

③ 词性特征

通过对引用对象的特征分析可以发现，术语型引用对象往往是名词或名词短语，并且具有一定词语搭配特点，如跟在动词、介词或介词短语之后，综合考虑加入词性特征 $W_{pos} \in R^{d3}$，用于捕捉引用对象本身词性特点以及前后词语搭配特征，特征选择随机初始化方式，维度为 20 维。

(2)启发式特征

在引用对象特征分析中发现，引用标签在引用对象判断时具有重要指示作用，术语型引用对象与引用标签距离也具有一定特点。鉴于此，本文提出两种启发式特征并观察其在该任务中的有效性。

① 标识符特征

无论引用标签在引用句中是否承担句法成分，引用标签作为引用对象识别的触发词和切入点，对引用对象识别起到一定的标识作用，故选取引用标签记为特征 $W_{ref} \in R^{d4}$，帮助模型聚焦引用标签及附近出现的词汇或短语。特征采用随机初始化向量的方式，维度为 20 维。

② 位置特征

通过观察术语型引用对象特征发现，引用对象常常出现在引用标签附近，即引用标签附近的词或词语成为引用对象的概率较大，位置特征可以帮助模型更好地捕捉全局特征，同样采用随机初始化的方式，维度为 20，表示为 $W_{dis} \in R^{d5}$。

通过上述特征描述，建立词向量、字符向量、引用标签向量、位置特征向量、词性向量 5 种不同描述角度的特征向量，对不同特征向量直接拼接，拼接顺序为词向量、字符向量、词性向量、位置向量和引用标签向量，向量维度为 100+30+20+20+20=190 维，表示为 $W=[W_{glove}, W_{char}, W_{pos}, W_{ref}, W_{dis}] \in R^{n}$，其中 $n=d_1+d_2+d_3+d_4+d_5$。这种向量直接拼接的方法是特征融合较为常用的方法，目的是人工加强句子的特征表示能力，以上特征中，词向量使用预先训练好的 GloVe 词表，字符向量由 CNN 训练获得，其他自建向量均采用随机初始化方式，在训练过程中不断优化表示。

3.3　伪标签学习

鉴于人工标注数据成本，采用伪标签学习(Pseudo-Label)[24-25]的半监督学习方法解决标注数据较少问题。根据引用对象特点提出伪标签降噪策略，利用数据间的相关性，用冗余对抗复杂，提高模型效果。伪标签学习可算作一种类 EM 方法[26]，利用信任度高的样本不断优化模型。伪标签学习利用未经过人工标注或确认的数据，经过模型预测得到近似标签的伪标签数据，结合标注数据和伪标签数据联合训练模型，提高模型的学习能力。伪标签学习主要流程如图 4 所示。

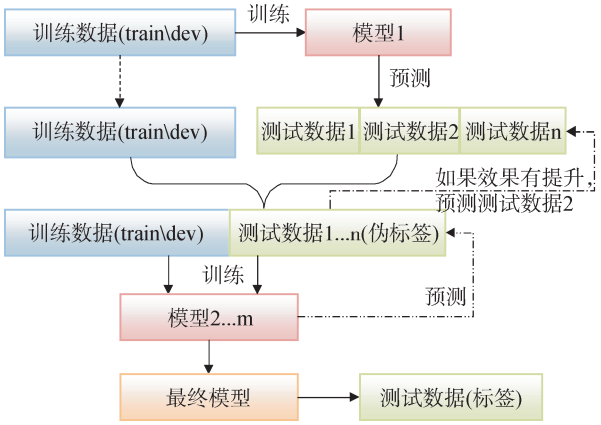

图 4　伪标签学习流程

Fig.4　Pseudo-label Learning Flowchart

伪标签学习中比较关键的问题是如何降低伪标签带来的噪声问题，如何有效降低伪标签噪声进而最大化利用廉价的未标注数据提升模型效果。由于本文采用的训练数据是随机采样获得的引用句，并没有预先经过人工判断，所以数据集中包含所有类型的引用对象。对于术语型引用对象识别任务来说，事实型引用对象和不含有引用对象的引用句会极大地影响模型训练效果，所以考虑从训练前和训练后两个方面联合限制伪标签数据带来的噪声。

噪声控制策具体略设计如下：对于无标签引用句集合A_{ul}，编写正则表达式查找引用标签位置，如果引用标签前后10个字符中包含名词或名词短语或符合术语型引用对象语法规则或包含大写字母单词、缩略语时，把这些引用句设为候选术语型引用句集合S_{ul}。使用预先训练好的模型W_l^i预测候选术语型引用句集合S_{ul}，得到伪标签数据集S_{pl}^i。选择与上一时刻伪标签集合S_{pl}^{i-1}中两次预测结果相同的n条数据（$n \leqslant \frac{1}{5} S_{ul}$）记为$S_{npl}^i$，则$S_{npl}^i \in S_{pl}^i \cap S_{pl}^{i-1}$。设$S_{l+npl}^{i+1} = S_1 \cup S_{npl}^i$作为训练集再次训练模型得到新模型$W_{1+npl}^{i+1}$，如果损失函数下降（$Loss = -\frac{1}{N}\sum_{i=1}^{N} \log P_i^l - \frac{1}{N}\sum_{i=1}^{N} \log P_i^{ul}$），则更新$W_l^i = W_{1+npl}^{i+1}$继续重复上述步骤，否则移除$W_{l+npl_l}^{i+1}$、$S_{pl}$重复上述步骤更新$S_{pl}$、$W_{l+pl_l}^i$。

在伪标签学习过程中主要依靠选择候选术语型引用句结合，限定伪标签采样规则和模型预测效果联合控制伪标签带来的噪音，当伪标签带来严重噪音导致模型预测能力下降时，去除这部分伪标签回溯到上一模型状态，继续迭代训练。通过不断的伪标签学习，改进训练数据集不足的问题，提高模型的泛化能力。

4 实　验

4.1 实验数据及预处理

由于目前引用对象识别暂无公开语料，本文随机选取计算机语言协会数据集（Association for Computational Linguistics Anthology Network，ANN）[27]为基础数据，采用人工标注引用对象的方法获得基础训练数据。搭建轻量级在线标注工具BRAT①，邀请两位图书情报领域研究生根据标注规范对引用对象进行标注。为检验标注一致性，采用Fleiss Kappa指标计算两位标注人员的标注一致性。Kappa系数是一个被广泛使用的一致性评价机制，其计算方法如公式（1）所示。

$$Kappa = \frac{P_0 - P_e}{1 - P_e} \tag{1}$$

其中，P_0为标注结果实际一致率，P_e为标注结果理论一致率。

随机选取5篇文献供两位标注者共同标注，对标注结果做交叉检验，结果显示Kappa一致性为0.51。根据Fleiss Kappa指标分布区间，Kappa大于0.75时已经取得相当满意的程度，小于0.40表示一致性差。Kappa值显示本次标注结果达到一个相对可靠的一致性水平。通过对比标注结果和与标注人员进行沟通，标注人员能够判断引用对象基本范围，但有些引用对象的边界存在差异，这是一致性指标不高的主要原因。通过人工标注最终构建术语型引用对象训练数据2 438条，未标注引用对象的引用句12 872条。

数据预处理阶段主要完成以下几项工作：

（1）对引用句中的引用标签进行规范化处理，根据论文遵循的哈佛引用格式标准，编写正则表达式匹配引用标签，使用“REF+文中参考文献序号”标签替代引用标签对数据对引用句进行预处理，如例1所示。既避免年和作者姓对分析带来的影响，又保留了引文在原文中的对照，为科研评价应用奠定基础。

例1：*The DSO corpus（**Ng and Lee，1996**）contains 192，800 annotated examples for 121 nouns and 70 verbs，drawn from BC and WSJ.*

转换为：*The DSO corpus **REF13** contains 192，800 annotated examples for 121 nouns and 70 verbs，drawn from BC and WSJ.*

（2）对引用句进行词性标注、引用标签标注、位置标注和序列标注标签标注，其中词性通过斯坦福大学提供的Stanford CoreNLP工具[28]获得；引用标签采用0，1标记，将引用句中的引用标签标记为1，其他词汇标记为0；位置特征采用数值标注，计算引用句中每个词语与引用标签的绝对距离，例如引用标签本身为0，左右第一个词均为1，左右第二个词均标记为2，以此类推；序列标注标签采用{B，I，O}标签格式，将引用句中的每个词标注为B-CIT、I-CIT和O，B-CIT表示引用对象的首词，I-CIT表示引用对象的中间词，O表示不属于引用对象的词，标注实例如例2所示。当引用句中存在多个引用对象与引文时，每次只标注一篇引文及其引用对象。标注后的

① http://brat.nlplab.org/.

实验数据遵循CoNLL2003命名实体识别任务[29]格式要求。

例 2：*REF6 describes a movie recommender system MadFilm where users can use speech and pointing to accept recommended movies.*

标注为：*REF6/NN/1/0/O describes/VBZ/0/1/O a/DT/0/2/O movie/NN/0/3/O recommender/NN/04/O system/NN/0/5/O MadFilm/NNP/06/B-CIT where/WRB/0/7/O users/NNS/0/8/O can/MD/0/9/O use/VB/0/10/O speech/NN/0/11/O and/CC/0/12/O pointing/VBG/0/13/O to/TO/0/14/O accept/VB/0/15/0 recommended/VBN/0/16/O movies/NNS/0/17/O.*

4.2　实验过程

为观察不同特征以及不同特征的组配方式对识别效果的影响，在实验过程中对词向量、字符向量、词性向量、引用标签向量、位置向量5种特征分别进行组配实验，有效对比特征选择的有效性和必要性。根据抽样法将数据按80%、10%、10%的比例随机划分为训练集、验证集和测试集，每次伪标签数据所占比例不超过标签数据的十分之一。

为避免算法过度拟合造成的识别结果不准确问题，实验过程中引入三种防止过拟合机制，分别为Dropout、L2正则、提前停止（Early Stopping）。Dropout机制设置Dropout使网络训练中以一定概率随机丢失神经节点，避免模型过于依赖局部特征，从而提高模型泛化能力。L2正则在目标函数后加入L2正则化项，使训练得到的权重变小，从而降低网络复杂度，避免过拟合。提前停止训练条件为，循环100次或当验证集的Loss在5次循环之内均不再下降。模型效果测评采用国际标准测评工具Conlleval。

实验时，根据本文所提特征表示规则，将引用句中每个单词拼接为190维向量，输入BiLSTM-CNN-CRF模型进行训练，实验中使用的某些参数的值如损失函数、优化器、学习率、LSTM层数等在不同任务的神经网络模型训练中会有所不同，最终参数选择是通过结合已有研究的经验以及实验过程中多次调参结果，如表1所示，其余神经网络参数如权重矩阵w和偏置b随机初始化，在神经网络训练时随之优化。

4.3　实验结果与分析

采用准确率（Precision）、召回率（Recall）和F1值对加入不同特征的模型识别效果进行评测，计算方法如公式（2）–公式（4）所示。

表1　参数说明

Table1　Parameters for Experiments

训练参数	值
LSTM层数	2
神经单元数量	100
学习率	0.015
Dropout率	0.5
损失函数	交叉熵损失函数
Batch_size	10
优化器	Adam
L2（权重衰减率）	1.0e-8
Char_max_len	20
卷积核大小	3*3
卷积核数量	30
句子最大长度	300

$$Precision=\frac{\text{正确识别出引用对象的数目}}{\text{自动识别引用对象数目}}\times 100\% \quad (2)$$

$$Recall=\frac{\text{正确识别出引用对象的数目}}{\text{实际引用对象的数目}}\times 100\% \quad (3)$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \quad (4)$$

各种特征组合在测试集上的测试结果如表2所示。

表2　各种特征组合在测试集上的测试结果

Table 2　Results with Different Features on Test Dataset

模型-特征	Precision	Recall	*F*1
BiLSTM-CNN-CRF（Baseline）	25.57%	8.17%	12.38%
BiLSTM-CNN-CRF（POS）	30.43%	15.26%	20.33%
BiLSTM-CNN-CRF（POS+REF）	60.42%	49.15%	54.21%
BiLSTM-CNN-CRF（POS+DIS）	61.18%	51.07%	55.67%
BiLSTM-CNN-CRF（REF+DIS）	61.71%	56.02%	58.73%
BiLSTM-CNN-CRF（POS+REF+DIS）	62.96%	57.63%	60.18%
BERT	52.13%	51.55%	51.94%

（注：POS表示词性特征，REF表示引用标签特征，DIS表示位置特征。）

作为Baseline的BiLSTM-CNN-CRF模型识别效果较差，证明对于表达形式灵活多变的引用对象，仅靠字向量和词向量特征无法达到识别要求，同时说明本文所提融合人类知识特征的识别方法的必要性。在词性、引用标签和位置三个特征中，词性特征结合位置特征或引用标签特征时，模型效果相似，但

加入引用标签特征和位置特征时模型F1值达到58.73%，比其他两种特征组合提高3-4个百分点，说明两个启发式特征对模型效果的提升起到了主要作用，分析原因不难发现术语型引用对象经常出现在引用标签附近，引用标签和位置两个特征帮助模型在建模时更好地捕捉这一特征。当模型融入所有人工特征时，F1值达到最佳水平60.18%，说明语言学特征和启发式特征在模型中能够显著提高模型效果，具有不可替代的作用。证明了本文所提人工特征融合方法对术语型引用对象识别任务的有效性。

为进一步验证本文识别方法的效果，与BERT（Bidirectional Encoder Representation from Transformers）[19]模型进行对比。本文所提多特征融合方法比单纯使用BERT模型的识别效果F1值提升8.24%，说明增强特征表示的方法有助于提高模型效果。反过来看，BERT在没有添加任何额外特征的情况下，模型的学习能力惊人，远超BiLSTM-CNN-CRF模型。

对识别结果进行分析，观察预测结果发现，存在引用对象遗漏和错误识别的情况，并且在同时预测多个引用对象时，尤其是句子较长且引用多篇论文时，会遗漏句子尾部的引用对象，另外识别出现一个引用对象被拆分造成表述不完整的情况，模型预测差异实例如表3所示。模型预测错误可能有两方面原因：

（1）训练数据的标注中缺少某些引用对象标注数据，导致模型学习效果有限，无法遍历所有引用句语言模式；

（2）词性特征的捕捉能力有限，比较明显的问题如有些名词词块型引用对象的识别边界错误，可尝试加入外部知识词典如IEEE术语词典[30]的形式，提高引用对象识别的完整性。

表3　标注结果与预测结果差异实例

Table3　Examples of Differences Between Labeled Results and Predicted Results

预测模型	预测结果
BiLSTM-CNN-CRF（POS+REF+DIS）	We have adopted the Conditional Maximum Entropy（MaxEnt）modeling paradigm as outlined in REF3 and REF19 To quickly（and approximately）evaluate this phenomenon，we trained the statistical IBM word-alignment model 4 REF7，using the GIZA ++ software REF11 for the following language pairs：Chinese-English，Italian-English，and Dutch-English，using the IWSLT-2006 corpus REF23 for the first two language pairs，and the Europarl corpus REF9 for the last one. In computational linguistic literature，much effort has been devoted to phonetic transliteration，such as English-Arabic，English-Chinese REF5，English-Japanese REF6 and English-Korean. Tokenisation，species word identification and chunking were implemented in-house using the LTXML2 tools REF4，whilst abbreviation extraction used the Schwartz and Hearst abbreviation extractor REF9 and lemmatisation used morpha REF12.

（注：表中蓝色表示人工标注引用对象，红色表示预测错误的引用对象，绿色表示正确预测的引用对象范围。）

5　结　语

本文探索了术语型引用对象自动化识别方法，提出基于特征融合的神经网络模型识别方法，利用语言学特征和启发式特征增强引用对象特征表示，设计伪标签学习噪音控制机制解决引用对象标注数据不足问题，通过实验证明了本方法的有效性。

虽然本方法有效地探索了术语型引用对象自动化识别问题，但对标其他自然语言处理任务模型F1值较低，存在预测结果错误和漏预测情况，距离模型大规模的应用仍有一段距离。未来工作可以从特征和模型两方面进一步探索研究：

（1）特征方面，尝试增加引用句中单词或词组是否在引文原文出现特征、语块特征和外部知识库特征，增强对引用对象及其边界的识别准确度；

（2）模型方面，由于半监督学习对标注语料仍存在一定依赖，未来识别模型可向无监督学习方向发展，降低对语料的依赖性，增强方法的可移植性，例如集合单篇引文的全部引用句，加入深度语义分析和语言模式识别策略，抽取与引文内容相关的关键术语或关键短语，从不同方向探索引用对象自动识别方法，提高识别性能。

参考文献：

[1]　Ding Y, Zhang G, Chambers T, et al. Content-based Citation

Analysis: The Next Generation of Citation Analysis[J]. Journal of the Association for Information Science and Technology, 2014,65 (9):1820-1833.

[2] 赵蓉英,曾宪琴,陈必坤. 全文本引文分析——引文分析的新发展[J]. 图书情报工作, 2014, 58(9): 129-135. (Zhao Rongying, Zeng Xianqin, Chen Bikun. Citation in Full-text：The Development of Citation Analysis[J]. Library & Information Service, 2014, 58(9): 129-135.)

[3] Small H G. Cited Documents as Concept Symbols[J]. Social Studies of Science, 1978, 8(3): 327-340.

[4] Qazvinian V, Radev D R. Scientific Paper Summarization Using Citation Summary Networks[C]// Proceedings of the 22nd International Conference on Computational Linguistics, Manchester. Association for Computational Linguistics, 2008: 689-696.

[5] Qazvinian V, Radev D R, Ozgur A. Citation Summarization Through Keyphrase Extraction[C]// Proceedings of the 23rd International Conference on Computational Linguistics, Beijing. Association for Computational Linguistics,2010:895-903.

[6] Jha R, Finegan-Dollak C, King B, et al. Content Models for Survey Generation: A Factoid-Based Evaluation[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, 1: 441-450.

[7] Anderson M H, Sun P Y T. What Have Scholars Retrieved from Walsh and Ungson (1991)? A Citation Context Study[J]. Management Learning,2010,41(2):131-145.

[8] Radoulov R. Exploring Automatic Citation Classification[D]. Waterloo: University of Waterloo，2008.

[9] 许德山. 科技论文引用中的观点倾向分析[D].北京:中国科学院文献情报中心，2012. (Xu Deshan. Sentiment Orientation Analysis for Evaluation Information of Citation on Scientific & Technical Paper[D]. Bejing: National Science Library, Chinese Academy of Sciences,2012.)

[10] Khalid A, Khan F A, Imran M, et al. Reference Terms Identification of Cited Articles as Topics from Citation Contexts [J].Computers and Electrical Engineering, 2019, 74: 569-580.

[11] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany. Association for Computational Linguistics, 2016:1064-1074.

[12] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3:1137-1155

[13] Santos C D, Zadrozny B. Learning Character-Level Representations for Part-of-Speech Tagging[C]// Proceedings of the 31st International Conference on Machine Learning, Beijing. Association for Computational Linguistics, 2014:1818-1826.

[14] Rei M, Crichton G K O, Pyysalo S. Attending to Characters in Neural Sequence Labeling Models[C]// Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan. Association for Computational Linguistics,2016:309-318.

[15] 赵洪, 王芳. 理论术语抽取的深度学习模型及自训练算法研究[J]. 情报学报, 2018, 37(9):923-938. (Zhao Hong, Wang Fang. A Deep Learning Model and Self-Training Algorithm for Theoretical Terms Extraction[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(9): 923-938.)

[16] Zhang Z Y, Han X, Liu Z Y, et al. ERNIE: Enhanced Language Representation with Informative Entities[OL]. arXiv Preprint. arXiv: 1905.07129.

[17] Shen Y Y, Yun H, Lipton Z C, et al. Deep Active Learning for Named Entity Recognition[C]// Proceedings of the 2nd Workshop on Representation Learning for NLP, Vancouver, Canada. Association for Computational Linguistics, 2017: 252-256.

[18] Ye Z X, Ling Z H. Hybrid Semi-Markov CRF for Neural Sequence Labeling[OL]. arXiv Preprint. arXiv: 1805.03838.

[19] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[OL]. arXiv Preprint. arXiv: 1810.04805.

[20] Bikel D M, Miller S, Schwartz R, et al. Nymble: A High-Performance Learning Name-finder[C]// Proceedings of the 5th Conference on Applied Natural Language Processing, Washington. Association for Computational Linguistics, 1997: 194-201.

[21] Lafferty J, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Proceedings of the 18th International Conference on Machine Learning, Williamstown, USA. Morgan Kaufmann Publishers Inc, 2001:282-289.

[22] Ma C, Zheng H F, Xie P, et al. DM_NLP at SemEval-2018 Task 8: Neural Sequence Labeling with Linguistic Features[C]// Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, USA. Association for Computational Linguistics, 2018:707-711.

[23] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar. Association for Computational Linguistics, 2014:1532-1543.

[24] Lee D H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks[C]// Proceedings of the 30th International Conference on Machine Learning, Atlanta, USA. 2013.

[25] Li Z, Ko B S, Choi H J. Naive Semi-supervised Deep Learning Using Pseudo-label[J]. Peer-to-Peer Networking and Applications,

2019, 12(5): 1358-1368.

[26] Dempster A P, Larird N M, Rubin D B. Maximum Likelihood from Incomplete Data via the EM Algorithm[J]. Journal of Royal Statistical Society: Series B, 1977, 39(1): 1-38.

[27] Radev D R, Muthukrishnan P, Qazinian V, et al. The ACL Anthology Network Corpus[J]. Language Resources and Evaluation, 2013, 47(4): 919-944.

[28] Manning C, Surdeanu M, Bauer J, et al. The Stanford CoreNLP Natural Language Processing Toolkit[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, USA. Association for Computational Linguistics, 2014: 55-60.

[29] Sang E F, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition[OL]. arXiv Preprint. arXiv: 0306050.

[30] IEEE Thesaurus [EB/OL]. [2019-07-12]. https://www. ieee. org/publications/services/thesaurus.html.

作者贡献声明：

马娜：提出研究思路，设计研究方案，进行特征融合实验，起草并修改论文；

张智雄：优化研究方案及论文最终版本修订；

吴朋民：进行BERT实验，参与论文修改。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储,E-mail:mana@mail. las. ac. cn。

[1] 马娜. citobjdata_train. txt. 术语型引用对象训练数据集.

[2] 马娜. citobjdata_test. txt. 术语型引用对象测试数据集.

收稿日期：2019-07-23

收修改稿日期：2019-10-11

Automatic Identification of Term Citation Object with Feature Fusion

Ma Na[1, 2] Zhang Zhixiong[1, 2, 3, 4] Wu Pengmin[5]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](School of Economic and Management, University of Chinese Academy of Sciences, Beijing 100190, China)

[3](Wuhan Library, Chinese Academy of Sciences, Wuhan 430071, China)

[4](Hubei Key Laboratory of Big Data in Science and Technology, Wuhan 430071, China)

[5](Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] This paper explores methods automatically identifying term citation objects from scientific papers, with feature fusion and pseudo-label noise reduction strategy. [**Methods**] First, we converted the identification of term citation objects into sequential annotation. Then, we combined linguistic and heuristic features of term citation objects in the BiLSTM-CNN-CRF input layer, which enhanced their feature representations. Finally, we designed pseudo-label learning noise reduction mechanism, and compared the performance of different models. [**Results**] The optimal F1 value of our method reached 0. 6018, which was 8% higher than that of the BERT model. [**Limitations**] The experimental data was collected from computer science articles, thus, our model needs to be examined with data from other fields. [**Conclusions**] The proposed method could effectively identify term citation objects.

Keywords: Citation Object Identification Feature Fusion Pseudo-Label Learning BiLSTM-CNN-CRF