基于学术论文全文内容的特定领域算法实体抽取研究*

丁睿祎 王玉琢 章成志
（南京理工大学经济管理学院，南京 210094）

摘要：对学术论文中的算法实体进行研究，能够促进深入了解算法对科学研究的作用，而从全文数据中抽取算法实体是相关研究的基础。学术论文全文内容中算法实体的抽取可以看作一种特殊的命名实体识别。本文通过人工识别的方法，从4 641篇论文中抽取出977种算法实体并构建算法实体词列表，以此为基础构建标注语料，训练算法实体自动抽取模型，在剩余语料上抽取得到221种新算法实体，并将自动抽取结果与人工抽取结果进行整合得到全部算法实体1 198种。研究结果表明：人工抽取法的结果能够为自动抽取法构建一定数量的标注语料，所构建的算法实体自动抽取模型能够有效地抽取出人工方法中遗漏的新算法实体，同时还能够抽取出已有算法实体的全新表达形式，进一步对人工抽取结果进行扩充和完善。

关键词：学术论文全文内容；算法实体；实体抽取；学术文本挖掘
中图分类号：G250.2　DOI：10.3772/j.issn.1673-2286.2022.03.001
引文格式：丁睿祎，王玉琢，章成志. 基于学术论文全文内容的特定领域算法实体抽取研究[J]. 数字图书馆论坛，2022 (3)：2-14.

随着全文数据库越来越多地向用户免费开放，基于全文数据、针对文中学术实体的相关研究逐渐开展起来。Ding等[1]提出“实体计量”（Entitymetrics）的概念，将论文中的学术实体划分为评价实体与知识实体两种。其中，评价实体主要包括作者、机构、参考文献、期刊、作者国别等论文外部信息，常用于学术影响力的分析工作；知识实体则被定义为学术论文中知识单元的载体，主要包括研究理论、研究方法、数据集、软件、工具集等论文内部信息。抽取学术论文中的知识实体，可为知识发现、知识流动等问题提供新的研究思路，在一定程度上促进学术评价研究的深入发展。

算法实体是一种典型的知识实体[2]。目前，越来越多的科学研究与应用使用算法来解决研究问题。在与数据科学的相关领域中，数据的处理、问题的解决几乎都离不开算法的应用。此外，在人文社会科学领域，算法的应用也呈不断上升的趋势。尤其是在大数据环境下的科研领域，绝大多数数据驱动型的研究需要借助算法来完成特定任务。对于某一学科领域的科研工作者，特别是初学者而言，他们希望能快速了解本学科领域常用的算法，并学会如何根据研究任务来有效选择合适的算法。研究学术论文中的算法实体，对深入认识算法对于科学研究的作用、有效评估算法的科学价值等方面有重要意义。

因此，从学术论文中抽取出算法实体，是对算法实体在学术论文中的分布情况进行研究的基础。要科学、客观地研究特定领域学术论文中的算法实体，首先要高效、全面地进行算法实体抽取。学术论文中的算法实体的形态较为多变，并且不同学者的行文风格迥异，无法通过人工全面获取。因此，基于机器学习的方法更适合算法实体的抽取。本研究首先构建算法实体标注语料库，由此训练性能最优的算法实体自动抽取模型，并对标注语料进一步自动抽取，从而得到更加完整的算

* 本研究得到江苏省社会科学基金项目“多维视角下学术创新力评估与预测研究”（编号：18TQD003）资助。

法实体抽取结果。

1 相关工作概述

与本文相关的研究工作包括命名实体识别、知识抽取。下面对这两个方面的相关工作进行概述。

1.1 命名实体识别研究概述

命名实体识别是信息抽取领域的重要研究内容之一，主要是指识别出给定文本中具有特殊意义的实体成分[3]。算法实体的抽取从本质上看是一种特殊的命名实体识别。广义上的实体识别主要针对包括人名、机构名、地名等，随着数据集的丰富与科研实践的需要，更多种类的实体识别研究逐渐开展起来，包括学术论文中的数据集、软件工具、定义、概念，以及本文所要研究的算法等。

（1）基于规则的实体识别。在早期的研究中，学者多采用基于规则的方法进行实体识别。例如参与MUC-6评测的基于词性与语义规则的Proteus系统[4]、基于词形与词性的PLUM系统[5]、基于短语规则的MITRE系统[6]，参与MUC-7评测的基于语义规则的Lasie-II系统[7]等。也有学者使用基于规则的方法针对中文进行实体识别研究。孙茂松等[8]通过统计分析大量人名，构建一系列人名规则并基于此在新闻语料上进行人名识别，获得了很高的召回率。

基于规则的方法简单有效，在特定领域针对性强、效率高。但实体的变化多种多样，要想提高识别效率需要制定大量的规则，而人工制定的规则有限，且绝大多数情况下制定的规则只适用于小型数据集或某一特定领域，难以在大量数据或其他领域扩展。因此有学者开始使用更加灵活的基于统计机器学习的方法来进行实体识别。

（2）基于传统机器学习的实体识别。基于统计机器学习的方法从本质上看是对命名实体进行分类，适用于各种领域的语料。常用的方法主要有最大熵模型（ME）、隐马尔可夫模型（HMM）以及条件随机场模型（CRF）等[3]。Borthwick等[9]构建了一个基于最大熵模型的实体识别系统“MENE”，在MUC-7评测数据集上取得了较好结果，且该系统具有跨语言移植性，能够用于日语数据集中的实体识别。同样，Bikel等[10]设计了一个基于隐马尔可夫模型的系统“IdentiFinderTM”，用于对人名、地名、时间以及数值等实体进行识别，并在MUC-6与MUC-7评测数据集上取得了较好的结果。McCallum等[11]则将CRF模型应用到实体识别中，并在CONLL-2003评测中表现突出。

与基于规则的方法相比，传统机器学习方法适用于绝大部分情况且可移植性较好，能在不同领域或语料上发挥作用。但构建高效的识别系统需要人工标注足够的数据来训练模型，还需人工构建特征。随着神经网络的兴起以及其在自然语言处理领域展现出的优秀性能，学者们开始尝试使用基于神经网络模型的深度学习技术进行实体识别。

（3）基于深度学习的实体识别。相比传统机器学习，深度学习的技术能够自动学习词汇语义、上下文依赖等，减少人工设定特征的代价[12]。Cherry等[13]将词向量作为特征应用到推特文本的实体识别任务中，有效提高了识别性能。Huang等[14]将一系列基于LSTM的模型应用于序列标注任务中，包括单一的LSTM模型、双向LSTM模型（Bi-LSTM）、带有CRF层的LSTM模型（LSTM-CRF）以及带有CRF层的双向LSTM模型（Bi-LSTM-CRF），经过实验比较，Bi-LSTM-CRF表现出了最佳性能。Lample等[15]提出了LSTM和基于转换的两种神经网络模型，并在CoNLL-2002与CoNLL-2003评测数据集上均获得了目前最好的实体识别效果。由于LSTM模型性能优越，且具有很好的可移植性和跨语言性，越来越多的学者将该模型应用到不同的领域来解决实体识别相关的研究问题，包括军事文本[16]、生物医学领域[17]等。

1.2 知识抽取研究概述

知识抽取是指基于全文内容数据，对蕴含于学术论文中的知识经过识别、理解、筛选、格式化，从而把文献中的各个知识点抽取出来的过程[18]。在之前的计量研究中，由于技术的局限性以及大规模成熟语料的缺乏，有学者采取人工识别的方法来进行知识抽取[19-20]。人工识别的方法准确度较高，但耗时耗力，只能运用于数据量较小的研究中，且对标注人员的专业素养要求很高，具有很大的局限性。随着计算机技术的不断发展与结构化全文数据的丰富，知识抽取的方法逐渐向自动化方向发展。为了提高知识抽取的效率、降低人工标注成本，研究者们开始利用更加先进的技术进行知识抽取。综合现有的学术成果，知识抽取研究主要针对理

论、数据集以及软件或工具等对象。

（1）理论抽取相关研究概述。王芳等[20]采用人工识别的方法从《情报学报》2000—2013年发表的1 822篇文章中识别出586条理论；赵洪等[21]提出一种基于深度学习的理论术语抽取方法，以目前应用较为广泛的Bi-LSTM-CRF深度学习模型为基础框架，构建一个理论术语抽取模型，对经济学、管理学、社会学、物理学和计算机科学等多个学科领域的论文中术语实体进行抽取；化柏林[22]则选取小规模数据集，运用基于规则的方法对中文学术文献的情报学方法理论术语进行抽取研究。

（2）数据集抽取相关研究概述。Singhal等[23]提出一种结合学术搜索引擎和在线词典的方法，来挖掘学术论文中所用的数据集实体；王雪等[24]以生物信息学领域的论文全文为研究对象，采用人工识别的方式对论文中使用的数据集进行标注，在此基础上构建了一个基于引用行为的科学数据集，以数据集被引频次、下载量等为指标，对科学数据的引用行为及其影响力进行研究，提出文献质量受数据集的质量影响较大、数量影响较小的观点。

（3）软件或工具抽取相关研究概述。Howison等[25]对生物学领域学术文献全文中提及软件的方式进行分析，发现大部分的软件提及并不符合引用规范。Pan等[26-27]对全文数据中软件的提及、引用以及评价进行了一系列的研究，提出了一种改进的自举方法，以Bootstrapping作为基本方法，结合大写字母、软件版本号以及文本触发器等多个特征，从学术论文中“方法/方法论”章节抽取软件实体并进行统计分析，发现存在大量的软件不规范引用的情况；并进一步应用该方法抽取12个不同学科学术文献中的软件实体，分析不同学科对软件的提及与引用情况。Li[28]分析了学术文献中R软件以及其软件包等的提及与引用情况，发现核心软件、软件包以及软件功能均以不同的方式在学术研究中发挥作用。

此外，学术论文全文中的学术定义、学术概念属性、学科知识点、研究结果以及图表等的抽取与评价工作也逐渐开展起来[29-32]，目前少有针对算法实体的抽取。

2 研究方法

2.1 基本思路

本研究以自然语言处理领域（Natural Lanugage Processing，NLP）为例，以国际计算语言学年会（Annual Meeting of the Association for Computational Linguistics，ACL）主会论文集作为研究数据，采用人工抽取与自动抽取相结合的方法，从学术论文全文中将各算法实体抽取出来。

首先，从ACL Anthology网站（https://aclweb.org/anthology/）下载1979—2015年XML格式的所有主会论文全文4 641篇，通过浏览该论文集中的论文，人工抽取论文中所提及的算法实体，同时查阅相关专著和文献资料以及咨询专家补充领域内常见算法实体，构建算法实体词列表；其次，为弥补人工抽取可能造成的遗漏，提高算法实体抽取的全面性，在对原始论文集进行格式转换、数据清洗等预处理后，利用构建的算法实体词列表，通过算法词匹配的方法从已处理的论文全文数据集中抽取包含算法词的句子集，作为训练算法实体自动抽取模型的标注语料，并利用该模型在剩余的未标注语料上进行算法实体的自动抽取；最后，对自动抽取法所得的算法实体结果进行整理，并与人工抽取算法实体结果进行比较与合并，包括频次过滤、人工筛选等手段，最终得到ACL全部算法实体。

2.2 算法实体抽取方法描述

在本研究中，算法实体的抽取主要使用基于人工抽取的方法与基于机器学习的自动抽取方法两种方法共同完成。

2.2.1 人工抽取

本研究使用基于人工的抽取方法初步实现算法实体的抽取，并在此基础上构建自动抽取所需的标注语料。以本研究所用数据集中的每一篇论文为单位，首先，通过在论文中检索“algorithm”“model”等特殊后缀词语来初步查找算法实体；其次，考虑到自然语言处理领域研究的特性以及学术论文结构的特点，进一步以论文的摘要（Abstract）、方法（Method）、实验（Evaluation）以及结果（Result）章节为重点，通过人工浏览查找所提及的算法实体；最终将文章编号、文章标题以及所提到的算法词记录下来。为了获取自动抽取所需的标注语料，我们基于人工抽取结果，构建包含多种算法实体以及各算法种类下多种算法实体表达形式的实体词列表。

由于算法可能存在简称、别名等不同的书面表达形式，本研究以算法标准名为搜索词，在Google Scholar以及Wikipedia上搜索得到各算法的常用书面表达形式，更进一步对算法实体词列表进行扩充。此外，考虑到所研究领域的学科特性，本研究还参考《机器学习》[33]、《统计学习方法》[34]等该领域内权威学术著作，整理出常见的算法实体，对算法实体词列表进行进一步扩充与完善。进而对所获得的全部算法实体进行人工整理，将属于同一种算法实体的不同表达形式进行同类合并，如“svm”“svms”“support vector machines”均表示“Support Vector Machine”算法，因而合并为“Support Vector Machine”类，同时保留每种表达形式。算法实体的人工抽取过程由1名博士生与1名硕士生合作完成，并由1位研究方向为自然语言处理（NLP）的教授审核后得到最终结果。

2.2.2 自动抽取

考虑到人工抽取算法实体可能造成遗漏，本文还使用机器学习的方法进行算法实体的自动抽取，进一步对算法实体结果进行补充。在本研究中，我们将算法实体的自动抽取视为一种特殊的命名实体识别任务，因此，用于传统命名实体识别任务的机器学习模型同样适用于本文所要解决的算法实体自动抽取中。本研究选择命名实体识别领域经典的机器学习方法，即条件随机场模型（Conditional Random Fields，CRF）与长短时记忆网络模型（Long Short-Term Memory，LSTM）及其变体，进行算法实体抽取实验。

（1）条件随机场模型。条件随机场是一种判别式无向图模型，由Lafferty等[35]于2001年提出，主要原理是对多个变量在给定观测值的条件概率进行建模。目前，CRF模型广泛应用于自然语言处理领域的词性标注与命名实体识别任务中，尤其是线性链条件随机场。本研究所使用的是线性链条件随机场。

在本文研究的算法实体自动抽取中，CRF模型的输入应是学术论文全文中的每一个句子，输出是对应的被模型标记的句子，其中带有特殊标记的词为我们所需要抽取的算法实体。CRF模型很好地利用了待识别实体的上下文特征，从而确保了所构建模型的整体性能，在命名实体识别任务中取得了较好的效果。在本研究中，条件随机场的输入观测序列为语句（单词序列），输出标记序列为相应的实体标记（每一个实体所属的类型）。

（2）双向长短时记忆神经网络模型（Bi-LSTM）。LSTM模型是一种特殊的循环神经网络（Recurrent Neural Network，RNN）模型，与一般的RNN相比，LSTM通过引入特殊的“门”结构（gate）和记忆单元（memory cell），从而有选择性地保存输入序列的上下文信息[36]，能够很好地解决长序列输入时梯度消失的问题。“门”是一种可选的让信息通过的方式，输入是一个向量，输出为0~1之间的实数值，可以理解为使用激活函数进行转换计算的机制。LSTM的记忆单元内部由遗忘门、输入门、输出门3个门结构以及神经元状态组成，分别决定传输过程中需要丢弃哪些旧的信息、补充哪些新的信息以及输出什么样的结果。

引入门结构的LSTM模型能够有选择性地保存输入序列的相邻状态特征信息。考虑到本研究是基于词的算法实体自动识别，在模型训练中，不仅需要考虑当前词与前文的关系，同时也需要后文的特征信息。双向LSTM（Bi-LSTM）由两层并行的、反向的LSTM网络组成，能够同时存储前后文信息[37]，从而进一步提高实体识别的准确性。

（3）双向长短时记忆神经网络-条件随机场模型（Bi-LSTM-CRF）。虽然Bi-LSTM能够学习到上下文信息，较好地完成实体识别任务，但是在输出层仅以计算所得最大概率值的标记作为输出，并不会考虑前后输出结果之间的关系。例如，以BIEO为标签体系时，（B代表一个实体的首单词，I代表实体的中间单词，E代表实体的最后一个单词，O表示非算法实体），当某个单词对应的输出标签为B时，从实际情况考虑，紧随其后的不应当为B标签，Bi-LSTM模型不会考虑相邻的输出标签之间的这种约束关系，这就可能导致标注错误。为了解决这个问题，有学者进一步在Bi-LSTM的输出层后再添加一层CRF结构，通过该CRF层计算相邻输出标签之间的关系，获取相邻输出标签的约束性规则（如B标签后不应当再出现B标签），从而确保标注序列的正确性[14]。

（4）双向卷积-长短时记忆神经网络-条件随机场模型（Bi-LSTM-CNN-CRF）。有学者发现，利用卷积神经网络（CNN）捕获训练预料的字符级特征，能够提高序列标注等任务的性能[38]。卷积层使用一个大小是T的卷积核在单词的字符向量矩阵上进行卷积来提取局部特征，卷积核大小T决定了可以提取单词周围T个词的特征，然后通过池化获得单词的字符级特征向

量[17]。因此，本研究在Bi-LSTM-CRF的基础上，使用CNN捕获输入观测序列中的字符级特征，并结合词向量作为模型的输入。

与CRF模型类似，在本文所研究的算法实体自动抽取中，Bi-LSTM、Bi-LSTM-CRF以及Bi-LSTM-CNN-CRF模型的输入同样是学术论文全文中的每一个句子，不同的是在输入模型之前需要对这些句子进行预训练，训练为词向量形式后再输入，输出同样是对应的被模型标记的句子，其中带有特殊标记的词则为我们所需要抽取的算法实体。LSTM引入了“门”机制，对传统RNN的隐层进行了结构上的改进，可以解决传统RNN的长期依赖问题。本研究基于LSTM结构，训练Bi-LSTM、Bi-LSTM-CRF以及Bi-LSTM-CNN-CRF 3个模型并比较各模型的抽取结果。

2.3 算法实体自动抽取模型中使用的特征

特征对模型性能有着较大的影响。由于当前还缺乏基于机器学习的学术论文算法实体进行自动抽取研究，本文总结前人针对特定领域中特殊实体识别的相关研究成果，考虑算法实体的特性以及其在学术论文中的表达形式，最终选择算法实体词、算法实体词的词性以及算法实体词的大小写作为训练算法实体自动抽取模型的特征。其中，算法实体词是指作为构成算法实体的单词，与其他非算法实体的单词相互区分；算法实体词的词性是指该算法词属于哪一种词性类型；算法实体词的大小写是指该算法词构成字母的大小写情况，主要包括整个词语全大写、整个词语全小写以及仅开头为大写3种类型。词性特征由开源工具Natural Language Toolkit（NLTK）自动标注，该工具提供包含名词、形容词、副词、代词等36种词性；大小写特征由笔者编写程序自动标注，具体标注方法为：当算法词为全大写时将该特征值记为1，仅首字母大写时记为2，其他情况记为0，以此区分。

3 实体抽取模型选择实验与结果分析

3.1 实验数据概述

3.1.1 语料获取与预处理

算法广泛地应用于各个学术领域中，尤其是数据驱动的自然科学领域。NLP是一个以数据和技术为核心的研究领域，绝大多数学者都需要借助算法来完成相关研究任务。此外，有研究表明在计算机相关学科领域，与期刊论文相比，会议论文更具前沿性。因此，本研究选择NLP领域的顶级会议ACL公开的1979—2015年的论文集全文4 641篇（均为XML格式）作为算法实体自动抽取的数据集。

在进行算法实体自动抽取之前，需要对原始XML格式的4 641篇论文全文数据进行预处理，包括格式转换、正文内容提取、内容清洗；此外，由于后续自动抽取所用模型的输入是以句子为单位，因此本研究进一步对清洗后的全文数据进行分句处理，最终得到由句子为构成单位的、仅包含论文正文内容的TXT格式句子数据集。

3.1.2 实验标注语料构建

目前，少有针对学术论文全文内容中算法实体的自动抽取研究，更是缺少用于算法实体自动抽取的标注语料。为了解决这一问题，本研究使用人工抽取法所得的算法实体结果，通过算法词匹配法从ACL论文全文内容中抽取算法实体所在的句子，作为后续训练算法实体自动抽取模型的标注语料。

基于原始数据进行预处理后所获得的句子数据集，本研究利用人工抽取得到的算法实体结果，在句子数据集中进行算法词匹配，抽取算法实体所在的句子（以下简称“算法句”）、算法句的所在文章的编号，得到“文章编号-算法词-算法句”数据集，如表1所示；最终，得到算法句60 662条（涉及论文3 993篇）。

由前文可知，本研究所选择的算法实体自动抽取模型的输入均为句子。因此，在抽取出算法句之后，依据2.3节中描述的特征及其对应的表示方法，对每一个句子进行特征表示。此外，在命名实体识别任务中，需要对实体进行标签标注，从而区分实体词与非实体词。作为一种特殊的命名实体识别，需要明确算法实体所需的标注标签，将算法词与非算法词区分。考虑到算法实体可能由一个词或多个词组成，本研究选择统一的BIESO标注集作为模型的输出标记，具体标签及其对应的含义信息如表2所示。

对每一个算法句，初步进行分词处理后，基于BIESO标注集对句中每一个单词标注其所属的标签，从而区分算法实体词与非算法实体词。

表1 算法词匹配法抽取结果示例

文章编号	文章标题	算法实体	算法句
P04-1016	Attention Shifting for Parsing Speech	SVM	Support Vector Machine (SVM) was selected as the kernel-based classifier for training and classification.
P05-1002	A High-Performance Semi-Supervised Learning Method for Text Chunking	CRF	Most recent work on improving CRF performance has focused on feature selection.

表2 BIESO标注集及对应含义

	B	I	E	S	O
含义	算法词开头	算法词中间	算法词结尾	单个算法词	非算法实体

3.2 结果评价指标

本研究将算法实体的抽取看作命名实体识别任务，因此选择准确率（Precision）、召回率（Recall）以及*F1*值作为模型的性能评价指标。

准确率*P*用来衡量模型的查准性能，其中*TP*表示识别正确的算法实体的总个数，*TP*+*TN*表示识别正确的算法实体个数与识别错误的算法实体个数之和，即识别出来的算法实体总个数，两者的比值即为准确率*P*，见公式（1）。

$$P=\frac{TP}{TP+TN} \quad (1)$$

召回率*R*用来衡量模型的查全性能，其中*TP*表示识别正确的算法实体的总个数，*TP*+*FN*表示识别正确的算法实体个数与未被识别正确的算法实体个数之和，即真正正确的算法实体总个数，两者的比值即为召回率*R*，见公式（2）。

$$R=\frac{TP}{TP+FN} \quad (2)$$

*F1*值是结合准确率和召回率两个值进行评价的指标，考察的是模型的综合性能。由于在不同的任务中，对准确率和召回率的重视程度不同，因此分子中的β代表权重，用于在不同的任务中决定*P*值重要还是*R*值更加重要。在本研究中，准确率和召回率同样重要，因此权重β取值为1，此时*F*值又被称为*F1*值，见公式（3）。

$$F1=\frac{(\beta+1)P\times R}{\beta^2(P+R)} \quad (3)$$

3.3 模型实现

本研究选择CRF++工具包（https://www.findbestopensource.com/product/crfpp）作为条件随机场模型的实现工具，选择GitHub上Liu所开源的NER-LSTM-CRF工具包（https://github.com/liu-nlper/NER-LSTM-CRF）作为Bi-LSTM模型相关变体模型的实现工具。

CRF++是一个实现条件随机场模型的工具，最早由Lafferty等[35]用于自然语言处理任务中，具有良好的通用性。NER-LSTM-CRF是一个基于TensorFlow框架（https://tensorflow.google.cn）的Bi-LSTM及其相关变体模型的实现工具，由Liu开发并开源于GitHub，包括实现预训练、模型训练、模型预测等功能，能够用来实现Bi-LSTM及其相关变体模型，处理中英文命名实体识别任务。

为了保证模型的稳定性与可信度，得到更客观的实验结果，本研究使用十折交叉验证的方法，即将训练集等分为10份，取其中9份作为训练集，1份作为测试集，按照不同的特征组合与不同的模型进行实验，分别对每个模型进行10次训练，最后取10次训练评价指标结果的平均值，作为每个模型的最终性能评价标准。

3.4 实验结果分析

3.4.1 基于人工的算法实体抽取结果分析

在基于人工的算法实体抽取方法中，经过2.2节中所描述的处理后，对本文所选取的全部4 641篇论文进行人工抽取，最终得到算法实体977种，共1 839种算法词表达形式。由于篇幅限制，本节无法展示全部结果，因此对抽取的算法实体进行初步频次统计，仅展示所得结果中频次排名靠前的10种算法实体，共50余种算法词表达形式。具体如表3所示。

表3 人工抽取所得算法实体结果（以频次排名前10为例）

	算法名	其他表达形式	频 次
1	BLEU algorithm（双语评价替换算法）	max-BLEU	2 948
2	Support Vector Machine（支持向量机）	svm，svms，support vector machines，support-vector machines，support-vector machine	2 472
3	Expectation Maximization（最大期望算法）	EM，em-algorithm，em-style，expectation-maximization，expectation and maximization，expectation-maximisation，expectation and maximisation，ExpectationMaximization	1 724
4	Maximum Entropy（最大熵模型）	ME，maxent，max-ent，maximum-entropy	1 721
5	Hidden Markov Model（隐马尔可夫模型）	hmm	1 636
6	Conditional Random Field（条件随机场）	crf，crfs，conditional random fields	1 584
7	Context Free Grammar（上下文无关文法）	cfg，cfgs，context free grammars，context-free grammar，context-free grammars，contextfree grammar，contextfree grammars	1 554
8	Tree Adjoining Grammar（树邻接文法）	tag，tags，tree adjoining grammars，tree-adjoining grammar，tree-adjoining grammars，TreeAdjoining Grammars	1 428
9	Neural Networks（神经网络）	neural network，neural-network based	1 296
10	Probabilistic Context Free Grammar（概率上下文无关文法）	pcfg，pcfgs，probabilistic cfg，probabilistic cfgs，probabilistic context-free grammar，probabilistic context free grammars，probabilistic context-free grammars，probabilistic contextfree grammar	1 151

从表3中可以看出，基于人工的算法实体抽取结果中频次最高的为BLEU algorithm（双语评价替换算法），且频次排名靠前的10种算法实体中，每一种均存在多种表达形式。这一结果说明：在ACL主会论文集中，不同的作者对不同的算法均有自己的描写形式。因此，基于人工的算法实体抽取方法存在遗漏算法实体以及遗漏算法实体不同表达形式的可能性。这一结果也表明基于机器学习模型对算法实体进行自动抽取很有必要。

3.4.2 基于机器学习的算法实体抽取结果分析

在基于机器学习的算法实体自动抽取方法中，为了便于数据集的等量划分，本文结合不同的特征组合，从所得的60 662条已标注算法句中，随机选择60 000条作为数据集，并将该训练集等比例随机划分为10个等份数据集，并按照9∶1划分为训练集与测试集，训练CRF、Bi-LSTM、Bi-LSTM-CRF、Bi-LSTM-CNN-CRF 4个模型，并在测试集上进行算法实体自动抽取的十折交叉验证，并使用10次交叉验证的平均准确率、平均召回率和平均F1值作为模型评价指标，所得结果如表4所示。

可以看出，对于每一个模型而言，与不使用任何特征的结果相比，使用词性特征对于模型的性能有了较大的提升，F1值均提高了8%左右；而大小写特征并没有

表4 不同特征下的算法实体抽取模型整体结果

%

特 征	CRF			Bi-LSTM			Bi-LSTM-CRF			Bi-LSTM-CNN-CRF		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
单词本身	76.60	82.27	79.29	84.15	84.45	84.28	84.10	85.28	84.68	84.42	84.80	84.60
单词+大小写	76.62	82.03	79.19	83.66	84.91	84.26	84.66	84.83	84.72	84.27	84.98	84.60
单词+词性	88.26	87.10	87.64	93.82	92.99	93.40	93.87	93.21	93.53	94.05	93.01	93.52
单词+大小写+词性	88.43	91.22	89.80	93.48	93.60	93.52	94.34	94.17	94.24	94.32	94.00	94.15

带来相同的效果，但是同时使用词性特征与大小写特征时，模型的性能能够提升到最优情况，此时F1值提高了10%左右。这一实验结果表明：对本研究所训练的每一个模型而言，同时补充词性特征与大小写特征之后的模型性能最好。

此外，综合CRF、Bi-LSTM、Bi-LSTM-CRF、Bi-LSTM-CNN-CRF 4个模型最优情况下的结果，综合性能最好的为Bi-LSTM-CRF模型，其十折交叉实验所得的平均P值、平均R值以及平均F1值均达到最高，F1达到94.24%，略高于添加了CNN层从而捕捉字符级特征的Bi-LSTM-CNN-CRF模型的效果。这一实验结果表明：在英文学术论文全文数据集的算法实体自动识别任务中，Bi-LSTM-CRF模型取得了最好的效果。因此，本研究选择Bi-LSTM-CRF模型作为后续未标注语料上自动抽取算法实体的基本模型，同时选用单词本身、大小写、词性3种特征。

4 未标注语料上的算法实体自动抽取

本文选择在测试集上获得最好效果的模型与特征组合，即同时使用词性特征与大小写特征的Bi-LSTM-CRF模型作为算法实体的自动抽取模型。本文在ACL全部数据集的基础上，剔除标注语料，在剩余语料上进行算法实体的自动抽取。为了便于描述，本文将剔除了标注语料的剩余语料命名为ACL未标注语料。

4.1 算法实体自动抽取过程

ACL未标注语料上的算法实体自动抽取基本思路如图1所示。首先，以使用算法词匹配法所得的全部60 662条标注语料作为训练集，选用单词本身、大小写、词性3种特征组合，训练Bi-LSTM-CRF模型；其次，从经过预处理的ACL全部数据集中（共4 641篇论文）剔除标注语料，得到ACL未标注语料625 445句；再次，与前文处理步骤相同，对ACL未标注语料中的每个句子进行分词、词性标注、特征表示等进一步处理，并利用训练好的Bi-LSTM-CRF模型进行算法实体的自动抽取；最后，考虑到自动抽取的结果中存在较多噪音实体，因此需要对自动抽取出的结果进行人工整理与核对，剔除非算法实体、保留有效算法实体，并将其与人工抽取所得的算法实体进行比较与合并等人工整理，最终得到ACL全部算法实体结果。

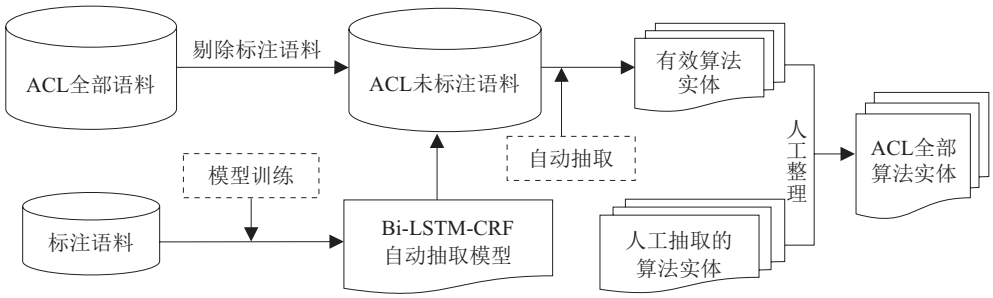

图1 ACL未标注语料上的算法实体自动抽取过程

4.2 算法实体自动抽取结果

4.2.1 抽取结果的整体分析

经过上述处理步骤后，本研究训练的Bi-LSTM-CRF模型最终抽取实体词51 884个，考虑到模型抽取结果存在噪音，且噪音基本存在于频次为1或2的实体词中，因此为了便于对抽取结果进行整理，本节首先对原始结果进行初步频次统计。

首先，经过初步频次统计，我们发现抽取的51 884个实体词中有28 015种实体，其中频次大于2的实体有2 811种。考虑到频次为1或2的实体词为噪音结果的可能性较大，且频次过低的实体词对本研究无意义，因此本节将频次为1或2的实体词直接剔除，仅对大于2的实体词进行整理与分析。

其次，进一步对频次大于2的实体进行人工筛选，保留属于算法实体的实体词、剔除非算法实体词。这一过程由1名具有NLP知识背景的硕士生和1名具有NLP知识背景的博士生完成，并由1位NLP领域教授审核后得到最后结果。我们通过判断抽取的实体词是否表示某种具有特定专属名称的算法来进行筛选，如实体词为“Monte Carlo Algorithm”（蒙特卡罗算法），经判断

该实体词为一种具有特定专属名称“Monte Carlo”的算法，因此在筛选过程中被认定为有效算法实体并保留；又如实体词为“Word Alignment Algorithm”（词对齐算法），经判断该实体词为用于解决某种研究任务的一类算法，不具有特定专属名称，因此在筛选过程中被认定为非算法实体被剔除。最终，经过人工筛选后得到有效算法实体词278个。限于篇幅原因在此仅展示所得结果中频次排名前10的算法实体，如表5所示。

表5 人工筛选后的算法实体自动抽取结果
（以频次排名前10为例）

	抽取出的实体词	频 次
1	Viterbi Algorithm（维特比算法）	125
2	Markov Model（马尔可夫模型）	87
3	Hidden Markov Model（隐马尔可夫模型）	76
4	Conditional Random Fields（条件随机场）	71
5	Bayesian Network（贝叶斯网络）	70
6	Brown Clusters（布朗聚类）	64
7	Beam-search（集束搜索）	46
8	Monte Carlo Algorithm（蒙特卡罗算法）	44
9	Vector Apace Model（向量空间模型）	35
10	Bag-of-words（词袋模型）	27

可以看出，在经人工筛选后的结果中，部分算法实体词已经存在于人工抽取的算法实体结果中，如“Hidden Markov Model”“Conditional Random Fields”等，这一结果表明在未标注语料中存在少量已存在于人工抽取结果中的算法实体的遗漏。原因可能在于：标注语料是通过算法词匹配法所构建的，而算法词匹配法对原数据集中的内容结构要求较高，原数据集中可能存在极少数的内容结构混乱情况，导致匹配失败，从而造成标注语料中存在少量算法句缺失，最终导致少量已存在于人工抽取结果中的算法实体的遗漏。而本文所使用的基于Bi-LSTM-CRF的算法实体自动抽取模型能够抽取出所遗漏的算法句，从而抽取出在人工抽取结果中已存在的算法实体，证明该自动抽取模型的整体抽取结果较为可靠，能够弥补人工抽取结果中算法实体的部分缺失。

除已存在于人工抽取结果中的算法实体外，从表5中还可以看出，部分算法词所属的算法实体种类在人工抽取的结果中已存在，但表达形式与人工抽取结果中的均不相同，如“Viterbi Algorithm”“Brown Clusters”等，分别属于人工抽取结果中的Viterbi类和Brown Cluster类，但经过观察发现Viterbi类和Brown Cluster类下并无“Viterbi Algorithm”和“Brown Clusters”这两种表达形式。这一结果表明：人工抽取的算法实体结果中存在同一种算法实体下表达形式不完整的情况，这可能因为算法实体本身的表达形式众多，且学者们在其论文中的描述方式各不相同，即使是查阅相关资料也难以将每种算法实体的所有表达形式全部获取。而本章所构建的基于Bi-LSTM-CRF的算法实体自动抽取模型能够抽取出这些遗漏的算法实体表达形式，从而丰富已有结果，证明该自动抽取模型的学习能力较强，能够补足同种算法实体下不同表达形式的不全，从而对算法实体抽取结果进行完善。

此外，表5所示的算法实体中还存在不属于原算法实体列表中任一种算法实体类的新算法实体，如“Markov Model”“Bayesian Network”“Monte Carlo Algorithm”等，这些抽取出的算法实体不在人工抽取结果中，也并非人工抽取算法实体的不同表达形式，是完全新种类的算法实体。这一结果证明：在人工抽取算法实体的过程中，可能存在由于个人知识水平的主观原因造成的算法实体遗漏情况。而本章所构建的基于Bi-LSTM-CRF的算法实体自动抽取模型能够抽取出人工未能抽取出的新算法实体，从而对算法实体抽取结果进行扩充，使得抽取结果更加全面。

综上，我们认为：在通过人工抽取的算法实体结果中，存在少量已有算法实体缺失、同种算法实体下算法表达形式不完全以及新算法实体遗漏的问题，而本文所构建的基于Bi-LSTM-CRF的算法实体自动抽取模型能够针对性地解决这些问题，从而对算法结果进行完善和扩充，说明本文所采用的算法实体的自动抽取方法可靠、有效。

4.2.2 抽取结果中的新算法实体分析

由前文可知，基于自动抽取所得的算法实体结果可分为3类：人工抽取结果中已有的算法实体、属于人工抽取结果中已有算法种类但表达形式不同的算法实体、不属于人工抽取结果中的新算法实体。我们将人工筛选后的自动抽取结果与人工抽取结果进行比较，发现在经人工筛选后的278个有效算法实体中，已存在于人工抽取结果中的算法实体有47个，约占所抽取出的有效算法实体的17%；属于人工抽取结果中已有算法实体种类但表达形式不同的算法实体27个，约占所抽取

出的有效算法实体词的10%；不属于人工抽取结果中的全新算法实体204个，约占所抽取出的有效算法实体词的73%，是抽取出的3种算法实体类型中数量最多的一种。这一结果表明：虽然本研究所构建的算法实体自动抽取模型抽取出的噪音较多，导致自动抽取的准确率不高，但该模型具有较高的召回率，能够抽取出较多的全新算法实体，弥补人工抽取中的遗漏，使得算法实体的抽取更全面。

由于新算法实体在抽取结果中所占比例最大，且对我们的算法实体抽取研究最具意义，因此本节进一步对新算法实体进行分析。与处理人工抽取所得算法实体的方法相同，首先对204个新算法实体进行类别合并。这一任务由1名具有NLP知识背景的硕士生完成，并由1位NLP领域教授审核，主要通过观察每一个有效算法实体的词语构成以及与其他算法实体的相似程度来进行，如遇到缩写词则结合算法实体所在算法句内容来进行判断。如“Dependency Tree-LSTMs”与“Dependency TreeLSTM”相似程度极高，仅存在符号以及词语单复数形式差异，因此在类别合并过程中被判断为同种算法，合并为Dependency TreeLSTM类。最终，得到新算法实体149种。由于篇幅限制在此仅展示频次排名前10的合并后新算法实体结果，如表6所示。

表6　合并后的新算法实体结果（以频次排名前10为例）

	新算法实体	频　次
1	Markov Model（马尔可夫模型）	181
2	Binary Tree（二叉树算法）	78
3	Bayesian Network（贝叶斯网络）	70
4	Cosine Distance（余弦距离算法）	65
5	Monte Carlo Algorithm（蒙特卡罗算法）	48
6	Adaptor Grammar（Adaptor语法）	47
7	Latent Variable Model（隐变量模型）	41
8	Reinforcement Learning（强化学习算法）	40
9	Binary Search（二进制搜索算法）	29
10	IBM Word Alignment（IBM词对齐算法）	19

由表6可看出，在自动抽取所得的新算法实体中，频次最高的是“Markov Model”，作为一种概率统计模型，其主要用于解决语音识别、音字转换等任务。第二位的则是“Binary Tree”，该算法是一种基础遍历算法，广泛应用于计算机相关的学科领域中。排在第三位的算法是“Bayesian Network”，其非常接近原有的算法词列表中的Bayesian Model。综合上述3种新算法实体的具体描述分析以及整体结果来看，我们发现人工抽取中遗漏了这些新算法的原因主要在于人工抽取方法中由于抽取人的知识水平有限，对NLP领域算法实体的了解还不够精深，误将本属于算法实体的词判断为非算法实体。这也从侧面反映出，本节所使用的基于Bi-LSTM-CRF的算法实体自动抽取模型具有较好的学习能力，能够发现由于主观因素造成的算法实体抽取遗漏，从而对总体抽取结果进行补充。

4.2.3　抽取结果中的错误分析

在对使用自动抽取方法对未标注语料进行算法实体抽取结果的分析中，我们注意到抽取出的错误结果即非算法实体较多。为了探究错误结果的产生原因，为未来进一步优化研究提供建议与参考，本节主要基于抽取出的实体所在句子以及标注语料对错误抽取结果进行深入分析。

在前文中，对频次大于2的抽取结果进行了人工筛选，剔除了非算法实体。考虑到抽取频次越高的错误实体词越能代表错误的典型性，在此从被剔除的非算法实体结果中，筛选出频次排名前10的非算法实体，如表7所示。对高频错误抽取结果进行分析，总结得到错误类型共有以下3种。

表7　错误抽取结果及频次（以频次排名前10为例）

	错误抽取结果	频　次
1	Brown（布朗）	739
2	art（无特定含义）	461
3	a grammar（无特定含义）	291
4	dependency trees（依存树）	268
5	word alignment（词对齐）	258
6	adjuncts（无特定含义）	201
7	the decoding（无特定含义）	197
8	the program（无特定含义）	181
9	TV（无特定含义）	168
10	dependency tree（依存树）	142

（1）大小写因素导致的错误。从表7可以看出，在错误抽取结果中，频次最高的实体词为“Brown”，达739次，远远高于其他实体。通过观察“Brown”所在的句子，发现抽取出的“Brown”表示的绝大部分均为人名含义的“布朗”。结合标注语料以及本研究所使用的特征，本研究认为“Brown”被抽取出来的主要原

因在于：“Brown”的首字母为大写，加上标注集中存在“Brown cluster”“Brown clustering”等与其较为相似的算法实体，导致抽取错误；同样，排在第9名的“TV”也是由于类似原因被当作算法实体而被抽取。

（2）上下文因素导致的错误。位列第2名的实体词是“art”，通过分析其所在的句子，发现被抽取出的主要原因之一为：其所在的常见词组“state of art”（中文意思为“目前为止最好的”）在论文撰写中较为常用。作为一个以数据与技术为核心的研究领域，NLP的大部分研究离不开算法以及具体实验。故在NLP领域的论文中，作者会把自己所提出的算法与他人所提出的算法相比较，在达到更好的结果时通常会在论文中用“state of art”来修饰自己的算法，因此“state of art”后通常会紧跟“algorithm”“model”等单词。同理，“a grammar”“word alignment”“adjuncts”“the program”这4个实体同样是因为包含或下文中多出现“algorithm”“model”“grammar”等常见算法实体后缀而被模型误抽取出。

（3）单词构成相似导致的错误。此外，还发现排在第4位的“dependency trees”与第10位的“dependency tree”十分相似，仅存在“trees”这一单词的单复数差异，由前文所述的同类合并原则来看，这两个实体词应为同一类实体。分析其被误抽取的原因，我们认为主要由标注语料中存在的“decission tree”（决策树算法）造成。“dependency tree”与“decission tree”拥有相同的后缀，单词构成十分相似，但“dependency tree”中文名为依存树，是NLP中常用的一种树结构并非算法实体。

通过对错误抽取结果进行分析，总结出在算法实体自动抽取过程中，造成错误抽取的原因主要有：一是大小写因素，由于很多算法实体词是首字母大写或全大写的结构，所训练的模型对大小写特征的记忆十分深刻，造成不少首字母大写或全大写的非算法实体被误抽取；二是上下文因素，由于很多算法实体词的后缀均为“algorithm”“model”“grammar”等，所训练的模型对前后文特征的感知十分敏感，造成不少后缀为“algorithm”“model”“grammar”的非算法实体被误抽取；三是构词因素，对于一些算法实体的特殊后缀如“tree”，所训练的模型同样也会学习到这一构词特征，从而将拥有类似词语构造的实体词误抽取出来。此外，由于算法实体本身形式多变，没有固定的构成模式，且在论文中的提及也并无规律可循，在一定程度上与一些非算法实体在论文中提及的特性相吻合，因此所训练的模型将符合这些特性的实体当作算法实体抽取出来。基于以上错误抽取结果的分析，本文认为，在算法实体自动抽取研究中，由于算法实体本身的特性，仅使用语法层面的特征容易带来较多噪音结果；对于这一问题，可进一步考虑加入算法实体本身以及所在句子上下文的语义层面特征进行模型优化，提高模型的准确率，从而优化自动抽取的结果。

最后，本研究结合人工抽取的算法实体与自动抽取的算法实体，对两个抽取结果进行人工整理与合并，得到最终的ACL全部算法实体1 198种。

5 结论与展望

本文以NLP领域为例，开展算法实体的抽取研究。本文基于CRF、Bi-LSTM、Bi-LSTM-CRF以及Bi-LSTM-CNN-CRF 4种模型，进行算法实体的自动抽取实验并进行性能的比较，结果证明在算法实体自动抽取任务中，表现最好的是基于Bi-LSTM-CRF的自动抽取模型。基于改进模型，本文在剩余未标注语料上进行算法实体的自动抽取，最终抽取出新算法实体221种。最后，通过频次过滤与人工筛选，整理人工抽取与自动抽取两种方法所抽取的算法实体结果，最终得到ACL全部算法实体集共1 198种。本研究表明，人工抽取法的结果能够为自动抽取法构建一定数量的标注语料，基于此训练的算法实体自动抽取模型能够有效地抽取出人工方法中遗漏的新算法实体，同时还能够抽取出已有算法实体的全新表达形式，对人工抽取结果进行扩充和完善，有效提高算法实体抽取的全面性。

本文研究还存在一些不足之处。首先，本研究仅以ACL主会的4 641篇论文作为研究数据，数据集较小，且人工抽取与自动抽取均在同一数据集上进行，未来将考虑扩大研究范围，充实训练集规模，并在ACL主会以外的数据集上实施自动抽取；其次，本研究的自动抽取实验中仅使用单词本身、词性、大小写3种语法层面的特征特征，所得抽取结果中存在较多噪音结果，未来可进一步考虑加入算法实体本身以及所在句子上下文的语义层面特征进行模型优化，优化自动抽取的结果；最后，由于缺乏黄金标准，本研究在分析自动抽取结果时仅将其与人工抽取的结果进行对比，未来将考虑人工标注部分生语料作为黄金标准，使得自动抽取结果的分析更加规范。

参考文献

[1] DING Y，SONG M，JIA H，et al. Entitymetrics-measuring the impact of entities [J]. PLOS ONE，2013，8（8）：e71416.

[2] WANG Y Z，ZHANG C Z，LI K. A review on method entities in the academic literature：extraction，evaluation，and application [EB/OL]. [2022-03-12]. https://doi.org/10.1007/s11192-022-04332-7.

[3] 刘浏，王东波. 命名实体识别研究综述 [J]. 情报学报，2018，37（3）：329-340.

[4] GRISHMAN R. The NYU System for MUC-6 or Where's the Syntax? [C] //Proceedings of the 6th Message Understanding Conference，San Francisco：Margan Kaufmann，1995：167-175.

[5] WEISCHEDEL R. BBN：description of the PLUM system as used for MUC-6 [C] //Proceedings of the 6th Message Understanding Conference，San Francisco：Margan Kaufmann，1995：55-69.

[6] ABERDEEN J，BURGER J，DAY D S，et al. MITRE：description of the alembic system used for MUC-6 [C] //Proceedings of the 6th Message Understanding Conference，San Francisco：Margan Kaufmann，1995：141-155.

[7] HUMPHREYS K，GAIZAUSKAS R，AZZAM S，et al. University of Sheffield：Description of The Lasie-II System as Used For Muc-7 [C] //Proceedings of 7th Message Understanding Conference，Fairfax：1998：1-12.

[8] 孙茂松，黄昌宁，高海燕，等. 中文姓名的自动辨识 [J]. 中文信息学报，1995（2）：16-27.

[9] BORTHWICK A. A Maximum Entropy Approach to Named Entity Recognition [D]. New York：New York University，1999.

[10] BIKEL D M，SCHWARTZ R，WEISCHEDEL R M. An algorithm that learns what's in a name [J]. Machine Learning，1999，34（1）：211-231.

[11] MCCALLUM A，LI W. Early results for named entity recognition with conditional random fields feature induction and web-enhanced [C] //CONLL'03：Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003，2003：181-191.

[12] 徐红霞，李春旺. 科技文献内容知识点抽取研究综述 [J]. 数据分析与知识发现，2019，3（3）：14-24.

[13] CHERRY C，GUO H Y. The unreasonable effectiveness of word representations for Twitter named entity recognition [C] //Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics，New York：Association for Computational Linguistics，2015：735-745.

[14] HUANG Z H，XU W，YU K. Bidirectional LSTM-CRF Models for Sequence Tagging [EB/OL]. [2022-01-01]. https://arxiv.org/abs/1508.01991.

[15] LAMPLE G，BALLESTEROS M，SUBRAMANIAN S，et al. Neural Architectures for Named Entity Recognition [EB/OL]. [2022-01-01]. https://aclanthology.org/N16-1030.pdf .

[16] 李健龙，王盼卿，韩琪羽. 基于双向LSTM的军事命名实体识别 [J]. 计算机工程与科学，2019，41（4）：713-718.

[17] 李丽双，郭元凯. 基于CNN-BLSTM-CRF模型的生物医学命名实体识别 [J]. 中文信息学报，2018，32（1）：116-122.

[18] 化柏林. 国内外知识抽取研究进展综述 [J]. 情报杂志，2008（2）：60-62.

[19] PETTIGREW K，MCKECHNIE L. The use of theory in information science research [J]. Journal of the American Society for Information Science and Technology，2001，52（1）：62-73.

[20] 王芳，史海燕，纪雪梅. 我国情报学研究中理论的应用：基于《情报学报》的内容分析 [J]. 情报学报，2015，34（6）：581-591.

[21] 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.

[22] 化柏林. 针对中文学术文献的情报方法术语抽取 [J]. 现代图书情报技术，2013（6）：68-75.

[23] SINGHAL A，SRIVASTAVA J. Data extract：Mining context from the web for dataset extraction [J]. International Journal of Machine Learning and Computing，2013，3（2）：219-223.

[24] 王雪，马胜利，余曾溧，等. 科学数据的引用行为及其影响力研究 [J]. 情报学报，2016，35（11）：1132-1139.

[25] HOWISON，J，BULLARD，J. Software in the scientific literature：Problems with seeing，finding，and using software mentioned in the biology literature [J]. Journal of the Association for Information Science and Technology，2016，67（9）：2137-2155.

[26] PAN X，YAN E，WANG Q，et al. Assessing the impact of software on science：a bootstrapped learning of software entities in full-text papers [J]. Journal of Informetrics，2015，9（4）：860-871.

[27] PAN X. Disciplinary differences of software use and impact in scientific literature [J]. Scientometrics，2016，109（3）：1593-1610.

[28] LI K. How is R cited in research outputs? Structure, impacts, and citation standard [J]. Journal of Informetrics, 2017, 11 (4): 989-1002.

[29] 丁君军，郑彦宁，化柏林. 基于规则的学术概念属性抽取 [J]. 情报理论与实践，2011，34 (12)：10-14，33.

[30] 温浩，温有奎，王民. 基于模式识别的文本知识点深度挖掘方法 [J]. 计算机科学，2016，43 (3)：279-284.

[31] GABB H A, LUCIC A, BLAKE C. A Method to Automatically Identify the Results from Journal Articles [C] //Proceeding of the iConference, 2015: 1-10.

[32] LEE P, WEST J, HOWE B. Viziometrics: Analyzing visual information in the scientific literature [J]. IEEE Transactions on Big Data, 2018, 4 (1): 117-129.

[33] 周志华. 机器学习 [M]. 北京：清华大学出版社，2016.

[34] 李航. 统计学习方法 [M]. 北京：清华大学出版社，2012.

[35] 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, San Fransisco: Morgan Kaufmann, 2001: 282-289.

[36] HOCHREITER S, SCHMIDHUBER J. Longs hort-termmemory [J]. Neuralcomputation, 1997, 9 (8): 1735-1780.

[37] GRAVES A, JÜRGEN S. Frame wise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 2005, 18 (5): 602-610.

[38] MA X Z, HOVY E. End-to-end sequence labeling via bi-directional lstm-cnns-crf [EB/OL]. [2022-01-01]. https://arxiv.org/abs/1603.01354.

作者简介

丁睿祎，女，1995年生，硕士研究生，研究方向：信息检索与数据挖掘。
王玉琢，女，1995年生，博士研究生，研究方向：文本挖掘与科学计量。
章成志，男，1977年生，博士，教授，博士生导师，通信作者，研究方向：信息组织、信息检索、数据挖掘及自然语言处理，E-mail：zhangcz@njust.edu.cn。

Extraction Algorithmic Entity from Full Text of Academic Articles in Special Domain

DING RuiYi WANG YuZhuo ZHANG ChengZhi
(School of Economics & Management, Nanjing University of Science & Technology, Nanjing 210094, P. R. China)

Abstract: The research on algorithmic entities in academic papers can promote an in-depth understanding of the role of algorithmic in scientific research, and extracting algorithmic entities from full-text of academic articles, which is regarded a special named entity extraction, is the basis of this research. Through the method of manual recognition, this paper extracts 977 algorithmic entities from full-text content of 4 641 papers and obtains a dictionary of algorithmic entity. Based on this, a labeled corpus is constructed and an automatic extraction model of algorithmic entities is trained. 221 new algorithmic entities are extracted from the remaining corpus. Finally, the automatic and manual extracting results are integrated to obtain a total of 1 198 algorithmic entities. The results show that the manual extraction method can build an annotated corpus for the automatic method, and the automatic model can extract the new algorithmic entities which are missed in the manual method effectively. What's more, the new expression form of the existing algorithmic entities are extracted, so as to further expand and improve the manual extraction results.

Keywords: Full Text of Academic Articles; Algorithmic Entity; Entity Extraction; Academic Text Mining

（收稿日期：2022-03-03）