doi:10.3772/j.issn.2095-915x.2021.04.002

针对特定任务的方法实体评估研究

李小乐[1] 王玉琢[1] 章成志[1,2]

1. 南京理工大学经济管理学院信息管理系 南京 210094；
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 在科学的发展中，研究方法扮演着重要角色。收集并分析特定学科的方法实体，能够帮助学者更好地了解该领域的研究方法，并找到适合其自身研究的方法。目前已有针对方法抽取和评价的相关研究，但尚未针对特定任务开展知识实体抽取与评估研究。[方法 / 过程] 本文以命名实体识别（Named Entity Recognition，NER）任务为例，从 ACL Anthology 网站中收集相关论文，利用内容分析法对论文中作者使用的方法实体进行标注。本文从 426 篇学术论文中标注出 904 种方法实体。并基于使用次数和使用年代两个维度来评估方法实体影响力。[结果 / 结论] 条件随机场是 NER 任务中影响力最大的算法，神经网络算法在近五年发展迅猛；学者倾向于使用算法而不是现成的工具进行实体识别；在数据选择方面，经典数据集是学者的首选；F 值、正确率和召回率是影响力最大的评价指标。本文的标注结果能够帮助学者更好地理解该任务，提高科研的效率。实体评估的结果能够为初学者在选择具体研究方法时提供参考。

关键词：命名实体识别；实体影响力评估；全文内容分析

中图分类号：G35

Evaluation of Method Entities for a Special Task

LI Xiaole[1] WANG Yuzhuo[1] ZHANG Chengzhi[1,2]

基金项目 富媒体数字出版内容组织与知识服务重点实验室开放基金项目“富媒体数字出版内容中细粒度知识实体的抽取及关联与演化分析研究”（ZD2020/09-04）。

作者简介 李小乐（1997-），本科，研究方向为文本挖掘；王玉琢（1995-），博士，研究方向为自然语言处理与科学计量；章成志（1977-），教授，研究方向为信息组织、信息检索、自然语言处理与文本挖掘，E-mail：zhangcz@njust.edu.cn。

引用格式 李小乐 , 王玉琢 , 章成志 . 针对特定任务的方法实体评估研究 [J]. 情报工程 , 2021, 7(4): 13-26.

doi:10.3772/j.issn.2095-915x.2021.04.002

1. Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China;
2. Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, Beijing 100038, China

Abstract: [Objective/ Significance] In the development of science, research methods play an important role. Collecting and analyzing method entities of specific disciplines can help scholars better understand the research methods in this field and find methods suitable for their a own research. At present, there have been related researches on method extraction and evaluation, but no knowledge entity extraction and evaluation research has been carried out for specific tasks. [Methods/Process] This article takes the named entity recognition (NER) task as an example, collects relevant papers from the ACL Anthology website, and uses content analysis to annotate the method entities used by the authors in papers. We got 904 method entities from 426 academic papers. We evaluated the impact of the method entity based on the number of times of usage and the age of usage. [Results / Conclusions] The study found that conditional random field is the most influential algorithms in NER task and neural network learning algorithms have developed rapidly in the past 5 years; Scholars tend to use algorithms instead of ready-made tools for entity recognition; In terms of data selection, classic datasets are still the first choice of scholars; F-measure, recall and precision are the most influential indices and measurements. The annotation results in this article can help scholars better understand the task and improve the efficiency of scientific research. The results of entities' impact can provide a reference for beginners when choosing research methods.
Keywords: Named entity recognition; impact of method entity; full-text content analysis

引言

科学研究指的是通过需求分析提出问题，之后通过调查、验证、实验等方法进行综合分析，进而探索未知、解决问题的过程。在此过程中，研究方法的重要性不言而喻。研究方法的提出和使用推动了各学科的发展，它既可以是学者的研究对象，也可以是学者解决问题的重要工具。具体而言，在科学研究中，研究方法包括数据收集方法和数据分析方法[1]，具体表现形式可为数据集、算法模型、软件等实体。

作为知识传播的媒介，学术论文中包含了大量的研究方法，是学者们使用、学习和分析研究方法的重要资源。通过人工标注或自动抽取的方式从学术论文中获取方法实体，并利用不同特征对其开展学术影响力评估，可以为学者们特别是初学者提供方法参考，促进其对领域内方法的了解。因此，从学术论文中识别并评价不同类别的方法实体目前已成为学者关注的热点问题[2]。但当前的研究主要针对特定的学科和领域，所得结果无法直接应用于具体任务[3-5]。将抽取和评估工作限定在特定任务中，一方面能够充分了解方法实体在该任务的使用情况，反映该任务当前的发展程度和未来发展方向；另一方面评估的结果能够为初学者提供系统的研究方法体系，帮助其找到适合自身研究工作的方法。为此，本研究将面向具体任务，从解决同一任务的学术论文中识别出多类别方法实体，以期获得更有针对性的方法实体评价结果。以命名实体识别任务为例，我们将探究如下几个研究问题：

（1）在命名实体识别任务中，学者常用的

方法实体有哪些?

（2）学者在使用方法实体时，会有哪些使用特征？不同的方法实体如何评价其影响力?

考虑到目前并无研究工作对学术论文中的方法实体进行明确定义，本文拟从问题解决的角度出发，将研究方法实体定义为：作者在学术论文中为解决问题而使用到的算法模型、数据、工具、软件和评价指标。

本研究的意义在于，一方面标注获取得到的常用方法实体，能帮助构建研究方法体系，为学者在写论文时提供统一的参考，帮助建立使用规范；另一方面，通过不同维度评估实体影响力，能帮助学者深入理解该任务，提高科研的效率。未来，在命名实体识别任务中做好该项工作后，可以将标注结果作为基础数据，使用机器学习的方法扩充到其它相关领域。

1 相关工作概述

本文从学术论文全文出发，在人工标注的基础上，研究命名实体识别领域方法实体的使用情况，并在此基础上评估实体影响力。和本研究直接相关的工作包括方法实体的抽取和影响力评估，下面针对这两方面进行概述。

1.1 方法实体抽取概述

对于方法实体的抽取研究，学者们多采用人工抽取、规则匹配和机器自动抽取的方法。

人工抽取实体的方法主要借助内容分析法，通过人工阅读和标识从学术文本中获取方法实体。储荷婷 [6] 通过研究情报学领域论文，发现内容分析法、实验法和理论研讨法取代了过去占主导地位的问卷调查法和历史研究法。Zhao 等 [7] 通过内容分析法，发现数据集的提及和引用在各学科中差异很大，医学和健康科学是数据集使用率最高的学科 [3]。Behrang 和 Siegfried 在 ACL ARC 的基础上，通过手工和自动标注，将标注出的技术术语分为算法、方法和解决方案，为计算语言学提供了完整的注释语料库。Yang 等 [8] 为了研究软件在生物信息学中的重要性，借助外部资源形成词典，人工提取国内外生物信息学论文的软件实体，发现更好的科学软件有助于产生更好的科学结果。

规则匹配包括词典直接匹配和其他规则识别。词典匹配需要提前构建实体名称词典，通过将词典内容与王玉琢等 [9] 以十大数据挖掘算法为研究对象，通过在学术搜索引擎和在线资源中检索得到相应的算法缩写和别名，构建数据挖掘十大算法的名称词典，随后利用词典匹配法识别出 NLP 领域会议论文中提及的这十大算法及相关信息。其他规则包括文本中方法实体相关的引导词、固定句型、所处位置等。化柏林等 [10] 先通过词典匹配找出包含方法实体的句子，通过人工总结得到 5 大类的描述方法实体的句式规则，从而从情报学领域论文中识别方法实体。在生物医学领域，Tsuruoka[11] 通过文本相似规则生成实体的拼写变体，并依据变体名抽取并拓展了生物医学词典的实体术语表。

自动抽取的方法包括传统的机器学习方法和深度学习。机器学习的方法一般将实体识别看作分类任务或序列标注任务 [12]。Pan 等 [13] 提出一种改进的 bootstrapping 方法，并使用此方法从 2014 年 PLoS ONE 上发表的所有论文中学

doi:10.3772/j.issn.2095-915x.2021.04.002

习识别软件实体，最终获得超过 2000 种不同的软件实体。深度学习则能够从文本中自动学习特征。Ammar 等 [14] 使用 BiLSTM 和 CRF 相结合的模型从论文中识别“方法”和“数据集”等实体。

在上述实体识别的方法中，人工标注方法能较好地保证识别结果的准确性，但效率较低。自动抽取方法在处理数据规模和速度上表现较优，但仍然离不开人工标注的训练数据。因此，无论采用何种方法，人工标注获取方法实体都是重要的基础工作。因此本文仍然采用人工标注的方式从学术文本中识别方法实体。一方面保证所得结果的准确性，另一方面也为未来的自动抽取工作奠定基础。

1.2 方法实体评估概述

评估方法实体在领域内的影响力，常用方法包括专家评议和非专家评议两种方法。2006 年 12 月在 IEEE 国际数据挖掘会议（IEEE International Conference on Data Mining, ICDM）中，为了得到数据挖掘领域中最具影响力的算法，多名领域专家通过公开投票，从 18 个候选算法中得出数据挖掘十大算法 [15]。这是最早由学者归纳出的算法，在一定程度上能够体现这些算法的影响力。Chu 等 [6] 通过人工阅读，识别出图书情报学文献中的常用研究方法，并对不同方法在该领域中的影响力进行评价。专家评议的方式适合数据有限的情况，此方法对于实体的评估更加准确。但是专家评议的结果依赖学者的学识，使得评估结论带有主观性，而且当数据量较大时，处理的成本太高，耗费时间长，难以扩展。

当前，研究人员还依据使用次数、使用年代等非专家评议的方法来评估方法实体影响力。Wang 等 [5] 依据提及论文数、提及总次数、提及位置三个方面评估十大数据挖掘算法的影响力，发现 SVM 在其既定标准下的影响力最大。Settouti 等 [16] 选取 10 个使用不同分类算法的分类器，对特定的数据集进行分类，评价 10 种分类算法的效果。一些学者提出用文献计量的方法比如被引次数来评价数据、软件等知识实体的影响力 [17]。赵蓉英 [18] 通过专业网站 Depsy 获取软件下载量、软件复用次数、软件在文献中的引用情况三种指标，评价开源软件的学术影响力。Pan 等 [19] 考察学术论文中 CiteSpace、HistCite 以及 VOSviewer 三种软件的提及与引用情况，依据软件在文章、期刊、学科等层面上的扩散深度与速度，评估软件的影响力。杨波 [20] 通过生物信息学中软件实体的使用和引用情况，构建相关性指标来评估软件的影响力。科学研究离不开数据的支持，规范的数据集能够节约研究人员处理数据的时间。对科学数据进行评价，是促进数据共享与重用的基础 [21]。丁楠 [21] 通过 Web of Science 中 DCI 数据库的数据发布量、数据被引量等多个角度来衡量人口调查领域的数据集的影响力。Belter[22] 研究海洋学领域的三个典型数据集，使用被引次数评估数据集影响力。Ding 等 [23] 借助 PubMedCenter 的生物医学全文文档，根据生物实体的引用频次构建引文网络，评估不同实体之间的影响力。

本文采用非专家评议方法来评估方法实体的影响力，期望能促进针对特定任务的方法实体评估研究。

2 研究方法

如图 1 所示，本文通过人工标注的方法，识别命名实体识别任务相关文章中使用的研究方法实体，并探究不同类别方法实体的学术影响力。本文从开放数据平台上获取了 NER 相关论文全文内容，随后制订了标注规范开展预标注。根据预标注反馈的结果，本研究对标注方案进行优化，对全文内容中使用的方法实体进行正式标注。最后，通过分析标注结果中实体的使用情况和相关特征，对实体的学术影响力进行多维度评估。

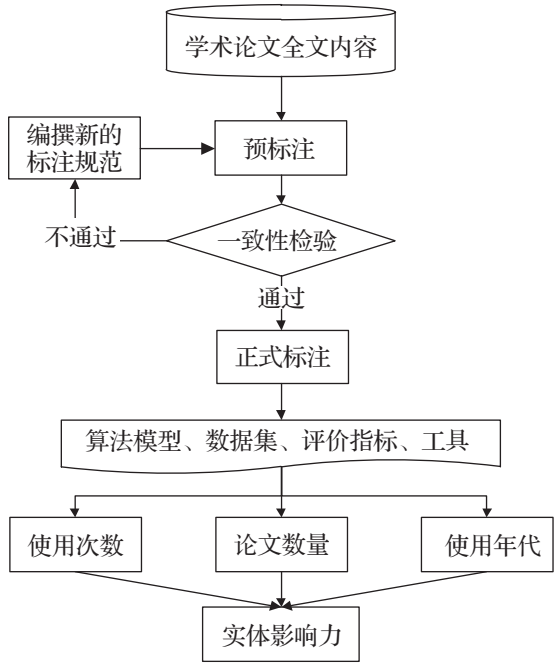


图 1　研究框架

2.1 数据获取

本研究所说的“特定任务”是指命名实体识别，以此为例分析研究方法实体在任务中的影响力。为了保证数据的可靠性，本文从 Association for Computational Linguistics（ACL）的开放数据平台中获取原始论文数据[①]，选择这个数据集的原因如下：首先，网站提供超 54000 篇的自然语言处理研究的论文全文，NER 则是自然语言处理的经典任务，该平台可保证本文能获取更为全面的论文和方法实体。其次，ACL 中收录了自然语言处理领域最知名的会议和期刊论文，从高质量论文中识别的研究方法实体更具代表性。

为了尽量获取更多的 NER 任务的相关论文，本文使用三类关键词从网站中检索，获得论文标题中含有“named entity recognition”或“named entity extraction” 或“named entity identification”（不区分大小写）的文章。通过网站给出的文献信息，提取出 450 条下载链接。通过人工的去重、删除了会议安排、特邀演讲、非英文文献、和文献综述等文章，最终得到 1998 年至 2019 年间 426 篇学术论文全文。

表 1　不同学术会议中获取的论文数

会议简称	论文数/篇	会议简称	论文数/篇
WS	162	EACL	11
ACL	43	ALTA	6
LERC	38	CONLL	4
EMNLP	36	SEMEVAL	4
IJCNLP	31	TACL	3
NAACL	23	HLT	3
COLING	22	ROCLING/ IJCLCLP	3
RANLP	20	ANLP	2
PACLIC	13	CL	2

表 1 展示了 426 篇学术论文在不同会议的

① https://www.aclweb.org/anthology/

doi:10.3772/j.issn.2095-915x.2021.04.002

数量分布（已按论文数量降序排列）。首先论文数量最多的是 WS（workshops），其次四大主会 ACL，EMNLP，NAACL，COLING 中论文数也明显比其他会议要高。

图 2 展示 426 篇论文的年代分布情况。该任务在 2003 年达到第一次顶峰，之后的十年基本上保持年发文量 18 篇的水平，近五年论文数量才表现出显著增长的趋势。

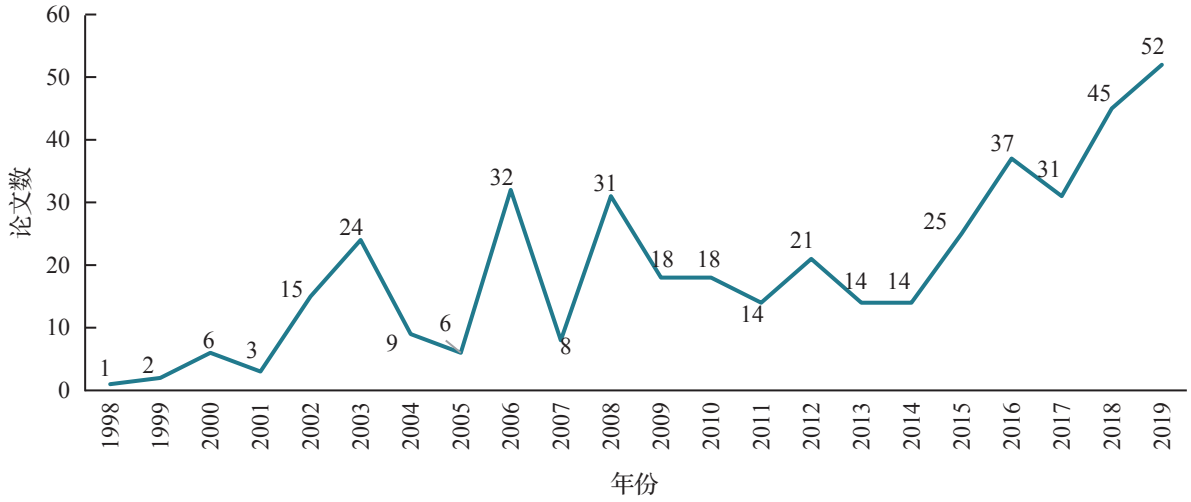


图 2　NER 任务中论文数量年代分布

2.2　数据标注

所有数据由两名信息管理与信息系统专业的大四学生②进行标注。在标注开始之前，根据相关工作制定标注规范，并邀请领域专家对标注规范进行完善。标注员依照标注规范对论文中的方法实体进行预标注后，根据标注员反馈对标注规范进行优化（优化后的标注规范见附录）。随后，基于新制订的标注规范，再次开展新的预标注，以了解标注结果的一致性。为保证标注结果的准确性，在预标注时，随机抽取 50 篇文献，由两位标注人员独立标注。本文将 Kappa 系数作为一致性检验的指标，评估二者标注的一致性[24]。Kappa 系数的计算公式如下：

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

其中，$P(A)$ 代表标注结果一致性的实际观测值，$P(E)$ 代表标注结果一致性的期望值。若 $K \geq 0.8$，则说明标注结果很可靠，若 $K \geq 0.69$，则说明标注结果较为可靠[25]。对标注人员独立标注的 50 篇文献的结果进行检验，随机 50 篇文献的 kappa 系数为 0.70，标注结果可以接受。

之后，标注人员对不一致的标注进行讨论分析，再次更新标注规范。依据最新的标注规范，二人平均分配剩下的学术论文，最终得到全部论文中作者使用的方法实体。标注结果包括 426 篇命名实体识别相关的文献中 3613 条记录，其中算法 1144 条，评价指标 1109 条，数据源 891 条，工具 469 条。具体标签定义与标注样例分别见表 2、表 3。

②两名标注人员分别为本文第一作者与陈仰。在此，感谢陈仰参加数据的标注工作。

表 2　人工标注的标签定义及示例

标签	含义	示例
ID	文献唯一标识符	373
Link	文献的下载链接	http://www.lrec-conf.org/proceedings/lrec2004/pdf/373.pdf
Title	论文题名	Corpus-based Learning of Lexical Resources for German Named Entity Recognition
Entity	文章内出现的方法实体	SVMs
Entity Type	实体类型	algorithm & model
Section Title	包含实体的句子所在的一级标题	Abstract
Entity Sentence	包含实体的句子	The approach uses linear SVMs and is based solely on an annotated corpus of reasonable size and a large amount of unlabeled data.

在引言部分已经定义本文的研究方法实体，表 3 给出具体示例。

表 3　四类方法实体的标注示例

实体类型	实体举例	实体所在的句子举例
算法模型	Conditional Random Fields model	A Conditional Random Fields model (Lafferty et al., 2001) annotates the entities components.
数据集	ACE 2005 Multilingual Training Corpus	We used the newswire section of the ACE 2005 Multilingual Training Corpus (128 documents, 66,015 tokens) for our experiments.
评价指标	recall、precision、F-measure	Performance was measured with the recall (R), precision (P), and F-measure (F = 2PR / (P+R)) scores.
工具	CRF++	For our baseline system, we used the CRF++ implementation of CRF sequence labeling.

2.3　方法实体的影响力评估

本研究通过使用次数、使用年代等维度，考察四类实体在学术论文中的使用情况并评估影响力。

2.3.1　基于使用次数的评估

论文数量是指某方法实体在全部数据中被使用在多少篇论文中。我们认为论文数越多，则实体的影响力越大。同时，我们借鉴学术论文影响力评价所使用的 count one[23] 方法，即一篇文章在不同位置多次说明使用同一实体，只记为 1 次。

（1）不同类别方法实体影响力分析

本文将使用了某一方法实体的论文数视为该方法的使用次数，我们按照数据源、算法模型、工具、评价指标对标注所得的方法实体进行统计，得到每类别中的方法实体数 N；每个方法实体 i 被使用的次数 ni 为使用该方法实体的论文数；每类方法实体被使用的总次数 UN 为使用该类方法的全部论文数量，即 $UN=\sum_{i=1}^{N} ni$；每类方法实体被使用的论文数 UA 为提及该类别方法的论文去重后所得的论文数；论文使用的平均实体数 $AN=\frac{UN}{UA}$。

（2）单一类别方法实体影响力分析

根据每类别种不同方法实体的被使用数排序，即可得到不同类别中方法实体影响力的排名。

doi:10.3772/j.issn.2095-915x.2021.04.002

2.3.2 基于使用年代的评估

本研究中，规定文章的发表年代即为实体的使用年代。对使用年代进行分析，考察实体使用情况演变趋势，能够得出在 NER 领域的各类实体的使用规律。本文的研究数据，每篇文献都有唯一的文章 ID（例如 Q15-1018，即为 2015 年发表的文章，有 28 篇文章无法直接从 ID 中看出年份，需要从下载链接中确定）。在人工标注过程中，记录下文章 ID 作为实体的使用年代信息，在此基础上统计各方法实体的论文数情况，以此进行年代分布分析。

3 结果分析

针对前文对数据的处理结果，本节将从使用次数、论文数量、使用年代三方面对命名实体识别任务中出现的方法实体及其使用情况进行分析并评估其实体的影响力。

3.1 基于使用次数的方法实体评估

经过人工标注及整理，最终得到 904 种不同的方法实体，其中包括 345 种数据源、251 种算法模型、235 种工具和 73 种评价指标。具体使用次数和涉及的论文数如下所示（表 4）。表 4 的使用总次数是指全部的某类实体标注得出的句子个数（比如全部的文章中，共 469 条使用工具的句子）。

从各实体的数量来看，识别出的实体中，数据源类型最多，评价指标数量最少。笔者分析，由于本课题研究的特殊性（即仅针对 NER 领域的文章），各个国家的研究人员在进行实体识别工作时可能都会采集不同的数据，进而形成不同的数据集，因此数据集的数量最多；而学术论文中最常见的评价指标有 F 值、P 值、R 值等，短时间内很难有别的更适合的指标，故该类别最少。

表 4 方法实体类型结果及论文数分布

实体类型	数量/种	占比	论文数/篇	使用总次数/次	平均每篇文献使用实体数量次/篇
数据源	**345**	38.1%	355	891	2.51
算法模型	251	27.8%	373	**1144**	3.08
工具	235	26.0%	361	469	1.30
评价指标	73	8.1%	**396**	1109	2.80

从表 4 的论文数量可知，评价指标会被绝大多数的文章所使用，占比超过 90%（396 篇 /426 篇）。其余三类实体在论文数量方面差别不大，且全都超过 350 篇。

通过每类方法实体的使用总数和其中涉及的论文数，可得每篇文献使用的实体数量。算法模型类实体的平均每篇文献出现该类实体数量最大，约为 3.08，也就意味着，在使用到算法模型的文章中，平均每篇会使用三种不同的算法。该标准下，工具类实体的平均每篇文献出现的实体仅约为 1.3，在 NER 任务中使用工具的现象并不普遍。

为了进一步分析实体的影响力，下面从具体的论文数量和使用年代两个方面评估实体影响力。

3.2 不同类别中的方法实体评估

经统计，得到四个类型的方法实体中排名前十的方法实体及使用的论文数量结果（表 5）。

表 5 各个类型实体中 **Top 10** 及使用论文数

	数据源	算法模型	工具	评价指标
1	CoNLL 2003(74)	CRF(194)	CRF++(40)	F-measure(371)
2	Wikipedia(74)	BiLSTM(72)	OpenNLP(11)	Precision(258)
3	Twitter(37)	SVM(50)	word2vec(11)	Recall(256)
4	CoNLL 2002(22)	ME(50)	Stanford CoreNLP(10)	cross validation(55)
5	MSRA(20)	Viterbi(49)	Twitter API(10)	Accuracy(34)
6	GENIA(17)	LSTM(35)	MALLET(9)	inter-annotator agreement(19)
7	ACE 2005(15)	HMM(31)	Brat(8)	Kappa(12)
8	People's Daily(14)	GloVe(29)	Stanford NER(8)	OOV rate(6)
9	OntoNotes 5.0(13)	SGD(28)	CRFsuite(7)	t-test(6)
10	MUC-6(12)	Adam algorithm(27)	LingPipe(7)	McNemar's test(5)

对于数据源来说，一些经典测评会议中产生的数据，如 CoNLL 02/03、ACE 05 会被学者广泛使用。CoNLL 系列的数据用于识别英语、德语和西班牙语中出现的地点、组织和人名。ACE05 则包括中英文和阿拉伯语的语料。此外，由于 Wikipedia 和 Twitter 平台上存在大量、多语言的语料，很多学者选择从这两个平台直接获取数据，进行 NER 任务。在生物医学领域，NER 的研究十分广泛。GENIA 语料库是为 GENIA 项目编写并标注的生物医学文献集合，该数据集包括生物医学中常用的术语、事件和共指关系等内容。时至今日，仍旧会被相关学者多次使用。关于 MSRA 和 People's Daily，前者是微软亚洲研究院提供的数据，后者是人民日报的数据，这两类数据常用于中文人名、地名和机构名的识别任务。

经典算法条件随机场（Conditional Random Field, CRF）排在算法类实体的首位。结合算法的发展历史和图 2 能够看出，自 2001 年 Lafferty 提出该模型之后，论文数量便有大幅提升。排名第二的深度学习算法 BiLSTM，虽然论文数量还不到 CRF 的一半，但已经超过了传统算法 SVM 和 HMM 等。作为老牌经典算法，SVM 算法理论基础坚实，是所有已知著名算法中最稳定且最精确的算法之一[7]，故使用 SVM 算法的论文数较多。值得注意的是，虽然 Viterbi 并不会直接应用到 NER 任务中，但是引入该算法能够得到可观察序列的最优可能的隐藏状态并且降低计算的复杂度，故 Viterbi 算法的排名较高。

由于命名实体识别属于 NLP 的子领域，因此论文中会出现一些 NLP 工具和一些机器学习的工具包。根据工具的用法可以将其分为三类：即在 NER 任务中被使用的仅支持 NER 类工具、在 NER 任务中起到转换作用的转换类工具和可执行多种任务的综合类工具。具体来

doi:10.3772/j.issn.2095-915x.2021.04.002

说，前十名中，仅有 CRF++ 和 CRFsuite 是专门针对 CRF 开发的工具，直接完成 NER 任务。word2vec 既是模型也是工具，其作为工具时，是 Google 在 2013 年推出的 NLP 工具，主要作用是将单词转换成向量形式，常用在深度学习中，起到转换词向量的作用。其他的工具则属于综合类，比如 OpenNLP 是一个机器学习工具包，用于处理自然语言文本，支持大多数常用的如：分词、分句、词性标注、NER 等 NLP 任务。Stanford CoreNLP 和 Stanford NER 均是斯坦福大学开发的 NLP 工具，用于完成词性标注、NER 等多种任务。

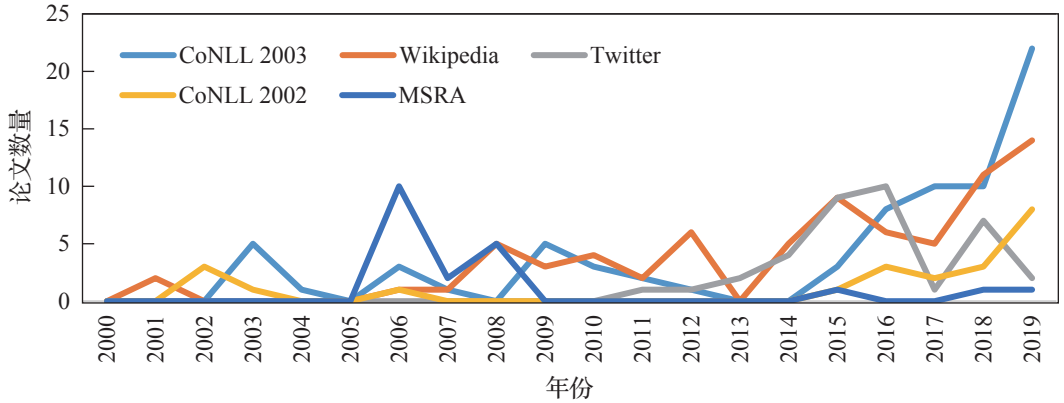
F 值、Precision 和 Recall 作为排名前三的评价指标，会被绝大多数的文章所使用。通过数据发现，这三种实体的数量不是一一对应的。具体来说，F 值的使用次数要比 P 值和 R 值要大得多。也就意味着学者在评估 NER 任务的好坏时，直接选择 F 值作为评价指标即可。cross validation 属于一种精度测试方法，用于评估模型的训练效果（常见的有 10 折交叉验证）。Accuracy 表示分类模型预测准确的比例，有时会跟在前三种指标的后面，有时也会单独使用。inter-annotator agreement 表示标注者间信度，是用来衡量一项任务中人类评分者意见一致的指标。如果意见不一致，则任务规范可能需要改进。

综上所述，四类方法实体中影响力最大的实体分别是 CoNLL 2003、CRF、CRF++、F-measure。Wikipedia 等开放平台上的数据成为研究热点，神经网络算法也受到更多学者的关注，学者最常用的工具实体大都是高校或互联网企业开发的，信息检索领域的评价指标（F 值、P 值、R 值）依然是影响力最大的指标类实体。

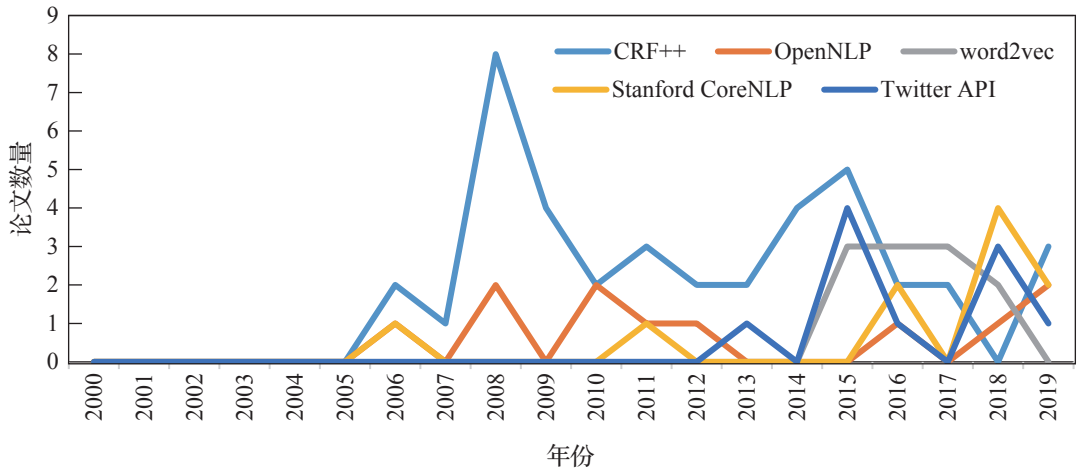
此外，数据源、算法模型、评价指标基本上符合“二八定律”，即排名前 20% 的方法实体会占据使用总次数的 80%，但工具不满足此规律。一方面，本来 Computer Science 领域直接使用工具就比较少，另一方面，在标注时，标注人员还发现虽然有些论文使用到某工具，但作者可能不会明确说明。

3.3 基于使用年代的方法实体评估

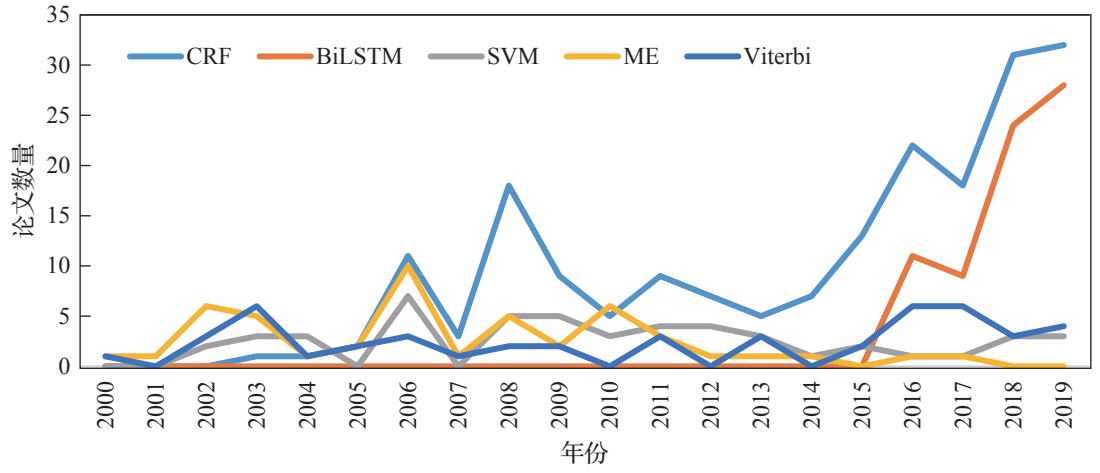
图 3 展示四类方法实体中排名前 5 的实体年代分布。总体来看，工具的使用次数最低（图 3（b））；算法类实体发展情况良好（图 3（c））；排名前三的评价指标类实体十分稳定（图 3（d））。



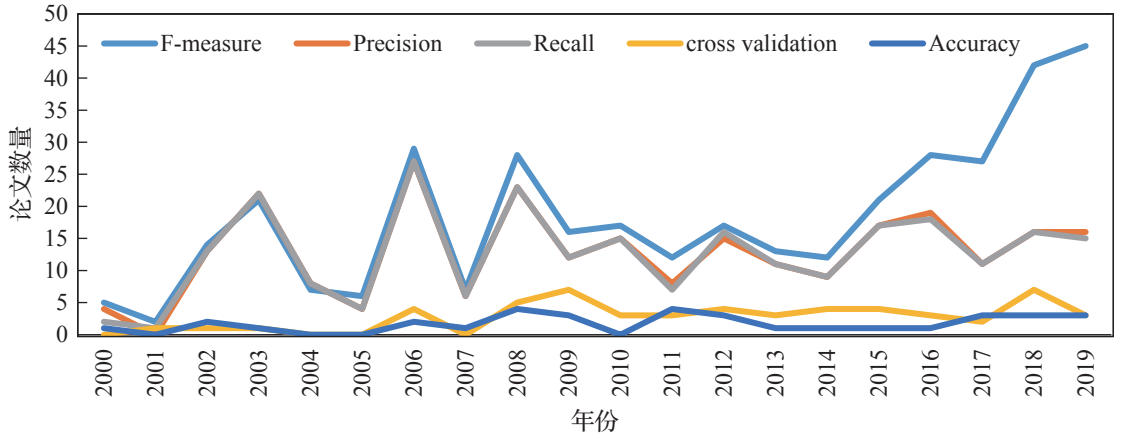
（a）排名前 5 的数据源及年份对应的论文数量图



（b）排名前 5 的工具及年份对应的论文数量图



（c）排名前 5 的算法模型及年份对应的论文数量图



（d）排名前 5 的评价指标及年份对应的论文数量图

图 **3**　四类方法实体中排名前 **5** 的实体年代分布图

doi:10.3772/j.issn.2095-915x.2021.04.002

图 3（a）显示，近年来经典数据集将被多次使用，这一点在近五年表现得更显著，并且 CoNLL 2003 的增长最为显著。同时，维基百科、推特上面丰富的语料也被广大研究人员使用。

在图 3（b）中，我们发现虽然工具的整体使用情况偏低，但在 2015 年之前被普遍使用，而近年来有所下降。相反，算法模型类实体的使用在 2015 年之后有了很大的提高（见图 3（c）），这表明学者们开始专注于算法本身来解决复杂的 NER 任务，而不是直接使用现成的工具。除此之外，本文发现 2013 年，算法的使用次数有明显的下降。研究认为，随着技术的发展，NER 领域引入了深度学习算法，研究人员有更多选择的机会来完成 NER 任务，但是此时大多数学者还处于观望状态。到 2015 年之后，深度学习算法 BiLSTM 被更多的学者使用，其影响力有了极大的提高。

如图 3（d）所示，F-measure 是依然是研究人员最常用的评价指标，P 值和 R 值已经趋于稳定。交叉验证（cross validation）最常用于评估一个模型在独立数据集上的概括能力，随着机器学习在 NER 任务的广泛应用，该实体的影响力也在不断变大。

4 结论与展望

研究表明，近 5 年，命名实体识别领域的论文逐年上升，说明研究人员还在不断改进算法，提高识别的准确性。在使用了评价指标的文章中，95% 的文章都会使用 F 值；使用次数最多的实体是算法模型类，该类别出现在 373 篇文献中，使用算法的文章平均会用到三种不同的算法；使用次数最少的是工具类实体，该类别仅出现在 361 篇文献中，其中超过半数的文章只会使用 1 种工具。当前 NER 领域最流行算法组合是 CRF+BiLSTM；研究人员最常在摘要和方法部分说明研究使用到的算法；工具作为一类重要的研究方法，研究人员对于工具的使用还不够重视；论文中会出现多个名称代表同一实体的情况，这给实体抽取带来一定的挑战，同时也不利于行业的规范发展。因此，制定一个统一的实体名称规范集合很有必要。

本研究仍存在不足之处，其一是人工标注的方法耗时耗力，扩展性不强；其二是对于标注人员的要求较高，不适应于非该领域的研究人员。其三是实体影响评估时，只通过次数分析影响力的广度，缺乏影响力深度的分析。未来，在本研究的基础上，使用机器学习的方法，自动识别文章中的研究方法实体，并进行更为细粒度的实体评估研究。

参 考 文 献

[1] Chu H T, Ke Q. Research methods: What’s in the name? [J] Library and Information Science Research, 2017(39): 284-294.

[2] 章成志 , 王玉琢 , 王如萍 . 情报学方法语料库构建 [J]. 科技情报研究 , 2020, 2(1): 30-45.

[3] Zhao M, Yan E, Li K. Data set mentions and citations: A content analysis of full-text publications [J]. Journal of the Association for Information Science and Technology, 2017(69): 32-46.

[4] 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.

[5] Wang Y, Zhang C. Using Full-Text of Research Articles to Analyze Academic Impact of Algorithms

[C]. In: International Conference on Information, England. Cham: Springer, 2018: 395-401.

[6] Chu Heting. Research methods in library and information science: A content analysis [J]. Library & Information Science Research, 2015, 37(1):36-41.

[7] QasemiZadeh B, Siegfried Handschuh S. The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics [C]. In: Proceedings of the 4th International Workshop on Computational Terminology, Dublin, Ireland. Association for Computational Linguistics and Dublin City University, 2014: 52-63.

[8] Yang B, Rousseau R, Wang X, et al. How important is scientific software in bioinformatics research? A comparative study between international and Chinese research communities[J]. Journal of the Association for Information Science and Technology, 2018, 69(9): 1122-1133.

[9] 王玉琢 , 章成志 . 考虑全文本内容的算法学术影响力分析研究 [J]. 图书情报工作 , 2017, 61(23): 6-14.

[10] 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.

[11] Tsuruoka Y, Tsujii J. Boosting precision and recall of dictionary-based protein name recognition [C]. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, Sapporo, Japan, 2003:41-48.

[12] 章成志 , 张颖怡 . 基于学术论文全文的研究方法实体自动识别研究 [J]. 情报学报 , 2020, 39(6): 589-600.

[13] 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.

[14] Ammar W, Peters M, Bhagavatula C, et al. The AI2 system at SemEval-2017 Task 10(ScienceIE):Semi-supervised end-to-end entity and relation extraction[C]. Proceedings of the 11th International Workshop on Semantic Evaluation. Stroudsburg:Association for Computational Linguistics, 2017:592-596.

[15] Wu X, Kumar V, Quinlan J R, et al. Top 10 algorithms in data mining [J]. Knowledge and Information Systems, 2008, 14(1):1-37.

[16] Settouti N, Bechar M E, Chikh M A, et al. Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classification Task [J]. International Journal of Interactive Multimedia and Artificial Intelligence, 2016, 4(1): 46-51.

[17] Ding Y, Song M, Han J, et al. Entitymetrics: measuring the impact of entities [J]. PLos ONE, 2013, 8(8): e71416.

[18] 赵蓉英 , 魏明坤 , 汪少震 . 基于 Altmetrics 的开源软件学术影响力评价研究 [J]. 中国图书馆学报 , 2017, 43(2): 80-95.

[19] Pan X, Yan E, Ming S, et al. Examining the usage, citation, and diffusion patterns of bibliometric mapping software: A comparative study of three tools [J]. Journal of Informetrics, 2018, 12(2): 481-493.

[20] 杨波 , 王雪 , 佘曾溧 . 生物信息学文献中的科学软件利用行为研究 [J]. 情报学报 , 2016, 35(11): 1140-1147.

[21] 丁楠 , 黎娇 , 李文雨泽 , 等 . 基于引用的科学数据评价研究 [J]. 图书与情报 , 2014(5): 95-99.

[22] Belter C W. Measuring the value of research data: A citation analysis of oceanographic data sets[J]. PLoS ONE, 2014, 9(3): e92590.

[23] Ding Y, Liu X Z, Guo C, et al. The distribution of references across texts: some implications for citation analysis [J]. Journal of Informetrics, 2013, 7(3): 583-592.

[24] 崔明 , 潘雪莲 , 华薇娜 . 我国图书情报领域的软件使用和引用研究 [J]. 中国图书馆学报 , 2018(3): 66-78.

[25] Carletta J. Assessing agreement on classification tasks: the kappa statistic [J]. Computational Linguistics, 1996, 22(2): 249-254

doi:10.3772/j.issn.2095-915x.2021.04.002

附录：针对特定任务的方法实体标注规范（以命名实体识别为例）

1 标注格式

ID	Link	Title	Entity	Entity Type	Entity Sentence
文章ID	文章链接	文章标题	实体名称	实体类型	实体句
见（1）	见（1）	见（1）	见（2）	见（3）	见（4）

说明：（1）点击文章链接，将下载之后的默认文件名作为文章 ID（是文献的唯一标识符），比如链接 https://www.aclweb.org/anthology/C12-1150.pdf 下载之后的文档默认名为 C12-1150，打开文件后文章标题为：Initial explorations on using CRFs for Turkish Named Entity Recognition。

（2）人工阅读文献全文，重点为摘要和方法论部分。识别其中作者使用的研究方法实体。

（3）实体类型分为四类，分别为：algorithm & model, tool, data source, index & measurement（现有一个方法词典可供参考）。

（4）实体句为文章明确提出使用了前面对应单元格里方法实体的句子。若一篇文章对于某个实体有多个句子，比如文章在摘要和方法论部分均提到使用了 CRF 模型。则仅标注最先使用的句子，即摘要里面那句话。

2 标注注意事项

（1）句子中含有 algorithm、model、approach、rules、grammar 等提示词属于算法模型的可能性比较大。

（2）句子中含有 package，parser，platform，tool，toolkit，API 和一些特殊的以“er”结尾的名词属于工具的可能性大。

（3）句子中含有 Wikipedia、corpus、dataset、corpora 和一些经典评测会议使用的数据集，比如 CoNLL 2002/2003、ACE 2005 等名词属于 data source 的可能性大。

（4）句子中含有 metrics，values，points，scores，test，rate 等词属于评价指标的可能性大。

（5）注意所有句子中全大写的单词，有可能是实体。