情报学报 2020年8月 第39卷 第8期
Journal of the China Society for Scientific and Technical Information, Aug. 2020, 39(8): 817-828

DOI: 10.3772/j.issn.1000-0135.2020.08.004

基于深度学习的情报学理论及方法术语识别研究

王 昊[1,2]，邓三鸿[1,2]，苏新宁[1,2]，官 琴[1,2]

（1. 南京大学信息管理学院，南京 210023；2. 江苏省数据工程与知识服务重点实验室，南京 210093）

摘 要 理论、方法的研究是学科不断发展前行的动力，了解掌握学科领域当前理论及方法的应用、发展情况是一项十分重要的工作。本文利用命名实体识别任务的分支——术语识别，对情报学理论方法进行研究，通过采集我国近20年来情报学领域相关文献20000篇左右，应用深度学习模型——Bi-LSTM-CRFs进行大规模语料训练与测试，通过实验验证其可行性并探究各实验变量对模型效果的影响，以求最大限度提高模型识别的效果。实验结果表明，对于理论方法术语等复杂实体，基于词切分的语料识别效果要优于基于字切分的语料；术语实体的长度对于识别效果也有一定影响，术语长度过大时（字数≥6），识别效果下降明显；同时，训练语料量与识别效果呈正相关关系，语料量越大，识别效果越好；实体的类型和数量直接影响识别结果，具有明显构词特征的实体识别效果较好；在特征引入实验中发现除拼音特征外，词性、词长以及词向量特征均能够对F1值有所提高，其中词向量和词性特征的提升效果最为明显。

关键词 情报学；术语识别；深度学习；Bi-LSTM-CRFs模型

A Study on Chinese Terminology Recognition of Theory and Method from Information Science: Based on Deep Learning

Wang Hao[1,2], Deng Sanhong[1,2], Su Xinning[1,2] and Guan Qin[1,2]

(1. *School of Information Management, Nanjing University, Nanjing* 210023;
2. *Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing* 210093)

Abstract: The study of theory and method is the driving force for the continuous development of any discipline. It is important to understand the application and development of the current theories and methods in the subject area. In this paper, terminology recognition which is a branch of the task of named entities is used to study the theoretical methods of information science. About 20000 articles in the field of information science in the past 20 years are collected, and as large-scale corpus to be trained and tested in Bi-LSTM-CRFs, a model of Deep Learning. The experiments verify the model's feasibility and explore the impact of each experimental variable on the model's effect, in order to maximize the effect of model recognition. The results show that for complex entities such as theoretical method terms, the corpus recognition based on word segmentation is better than the word segmentation-based corpus. The length of the term also has a certain influence on the recognition effect. When the length of the term is too long (word count ≥6), the recognition effect is obviously re-

收稿日期：2019-07-26；**修回日期**：2019-10-08

基金项目：国家社会科学基金重大招标项目“情报学学科建设与情报工作未来发展路径研究”（17ZDA291）；“江苏青年社科英才”人才培养项目；“南京大学仲英青年学者”人才培养项目。

作者简介：王昊，男，1981年生，博士，教授，博士生导师，主要研究方向为智能信息处理和检索、数据挖掘技术及其应用等；邓三鸿，男，1975年生，博士，教授，博士生导师，主要研究方向为知识图谱、科学评价等，E-mail：sanhong@nju.edu.cn；苏新宁，男，1955年生，教授，博士生导师，主要研究方向为智能信息处理和检索、期刊分析与评价等；官琴，女，1992年生，硕士研究生，主要研究方向为智能信息处理和检索。

duced. At the same time, the training corpus quantity is positively correlated with the recognition effect. Larger corpus quantities lead to better recognition. The type and quantity of the entity directly affects the recognition result. The entity recognition with obvious word formation features is better. In the feature introduction experiment, in addition to the pinyin feature, the part of speech, the length of the word, and the feature of the word vector can improve the F1 value. The improvement of the word vector and the part of speech features are obvious.

Key words: information science; terminology recognition; deep learning; Bi-LSTM-CRFs model

1 引 言

情报学科自1945年起正式成为一门独立的研究学科[1]，至今发展已有70多年的历史。相比于其他人文或理学学科，它仍是一门较为年轻的学科。包昌火教授等[2]指出，情报学学科地位不明确使得学界长期以来将情报与信息混淆，情报工作的方向被误导，同时也在一定程度上削弱了情报学的独立性。2017年，我国颁布了《中华人民共和国国家情报法》，将情报学再次拉回人们的视野，并且使得学者们重新审视情报学及其研究内涵。同年7月，第四届华山情报论坛在陕西西安举办，主题为“《国家情报法》与中国情报发展”，对《国家情报法》进行解读，同时，多位学者也提出自己对于情报学发展思考与建议。同年10月，“情报学与情报工作发展论坛（2017）”在南京大学召开，并形成《南京共识》。会议对于情报学科的发展目标、工作性质及主要作用进行了重新定义，并指出加强情报学理论、技术及方法的研究的重要性。不久之后，“情报学学科建设与情报工作未来发展路径研究”作为国家社会科学基金重大项目正式立项，该项目将会成为推动我国情报学学科迈向新发展的主要动力。在此背景下，情报学各个方向的研究也将迸发新的活力，尤其是情报学科发展的本根——理论和方法，将成为研究重点与热点。

在科学研究中，创建、发展与完善理论是至关重要的。理论是由在某一领域或学科中的有较高权威的专家提出创立的，其最终的目的必然是更好地服务于人类。Henshel等[3]在其提出的科学研究过程模型中指出两种典型的科学研究类型，一类是演绎式方法，从理论起始，进行逻辑思考与假设验证；一类是归纳式方法，以客观现象为切入点，运用实证方法进行概括总结并归纳为理论；二者虽然具有不同的研究思路，但毋庸置疑的是，理论在普遍性的科学研究中是极为重要的。与此同时，一个学科能够健康长久的发展，另一个至关重要的因素便是方法论[4]。在进行科学研究的过程中，对于方法的反复思考与审视是必不可少的，这可以使得研究变成更主动、更自觉的行为。而情报学的发展，更加离不开新问题的发现以及解决该问题时所采用的研究方法。

综上，对情报学理论和方法进行研究具有十分重要的意义，而情报学理论方法术语识别是研究中的基础工作和较为重要的一个环节。首先，可以由此梳理我国情报学的发展历程，了解不同历史时期各类理论方法的应用情况以及当时的研究重点，展现其演化过程；也能在研究过程中发现并暴露其中存在的一些问题，从而及时地进行修正，而不是一味蒙头向前。同时，在基础研究的基础上，能够构建出较为系统全面的理论方法体系，引导学科实现良性快速的发展。其次，将文献中使用的支撑理论方法或发展提出的新理论、新方法识别抽取出来，并且判断其主要的引用功能，在此基础上进行分类，能够对文献被引分析、引文内容分析提供一定的辅助补充功能[5]。再者，通过对识别出来的理论方法术语进行使用频次以及其重复率的计算，能够对其进行学科间交叉引用的分析，从而了解本学科理论方法的专属度以及与其他学科之间的融合渗透关系[6]。

2 相关研究

2.1 情报学理论方法相关研究

为了了解情报学理论方法研究与发展的现状，国内外众多学者采用多种方法从不同角度入手，进行深入探讨。在理论研究方面，Pettigrew等[7]收集了情报学6本期刊1993—1998年发表的1160篇文献，研究文献对情报学理论的应用情况，并统计所有理论的学科来源占比；其中社会科学占比最大，为45.4%，其次是情报学（29.9%）、自然科学（19.3%）以及人文学科（5.4%）。Jeong等[8]对韩国图书情报学的理论知识结构进行分析，利用1970年以来发表的645篇文献进行内容分析，揭示了理论的整体使用情况，包括理论的起源、使用程度以及

发展情况等，并归纳总结出常用的80余种理论。Kim等[9]通过对图书情报学20年（1984—2003年）共1661篇文献进行内容分析，发现文献中理论应用的比例逐年增加且呈现多学科融合的趋势。Kumasi等[10]利用2009—2011年发表的论文进行归纳总结，发现了理论之间的关联关系并进行了验证，同时讨论了各类期刊中对理论的使用情况。

在方法研究方面，van de Water等[11]在1976年对情报学领域相关文献中采用的研究方法进行了批判性分析，将1969—1971年发表的430篇论文与1974年发表的152篇论文进行对比分析，结果发现尽管文献数量一直保持稳定状态，但研究方法的种类和数量随时间而变化。Tuomaala等[12]在对1965—1985年图情领域文献进行内容分析时，总结并统计了文献中经常使用的研究方法，揭示了研究方法与研究主题之间的关系，并指出调查研究法是最常用的方法，占比为20%~23%。Chu[13]研究了2001—2010年情报学领域的1162篇文献，采用了定性定量相结合的方法，研究结果表明，研究方法种类和数量正在逐年增加，并且内容分析法、实验法以及理论分析法已经取代问卷调查法和历史研究法，成为图情领域首选的研究方法。Ferran-Ferrer等[14]对西班牙2012—2014年图情领域发表的580篇论文进行内容分析，总结了该领域常用的研究方法和技术以及其中涉及的主题，发现西班牙图情领域需要提高对实验性研究方法的重视。

国内近年来也开展了类似研究。其中，南开大学王芳团队对于情报学理论方法的研究较为全面，运用多种方法，从多个角度对于其进行研究探讨。2015年该团队利用内容分析法，对于《情报学报》2000—2013年的所有文献进行分析，通过编制各类编码表，对于我国情报学理论的应用情况进行了阐释，并指出对于理论的应用和创新方面，我国情报界仍需不断提高重视[15]；紧接着第二年，基于同样的方法（内容分析法），该团队扩大了数据范围[16]，研究了53本情报学领域期刊，时间跨度为1991—2015年，统计了情报学理论的应用频次，同时追根溯源，明确理论的来源学科，定义并计算了各理论的学科专属度，发现我国情报学所运用的各种理论，学科来源十分的广泛，不仅有人文社科类学科，还有自然科学类学科，包容性较好，但同时也暴露出情报学专属理论较少且应用不多的问题。王知津等[17]利用1990—2011年中国知网上关于情报学理论研究的期刊论文数据，采用Excel作为分析工具，对我国近20年来情报学理论研究的趋势进行统计分析；共选取1960篇文献，对发文量、作者分布、机构分布、期刊分布、关键词分布进行统计分析，得出我国情报学理论研究的主要趋势：知识化、人文化、定量化以及研究方法多元化。陈锋等[18]利用条件随机场模型，将理论识别当作命名实体识别中的子任务，基于语义泛化的基本思想并选取了理论的词性与词形作为标注特征，进行实验验证，证明了该方法具有可行性；但由于语料规模较小以及未做消歧处理，导致研究结果有一定的局限性。但是，陈锋等[18]的研究给众多学者提供了新的思路，即机器自动识别理论。

国内学者在方法领域也进行了探究。钱军等[19]利用聚类分析的方法将情报学方法分为基于数据的方法、基于知识的方法、基于文献的方法等，在此基础上还分析其各自的异同点和相互关系。杨锐[20]运用归纳法对情报学主要的研究方法进行总结分析，并建立了一整套基于情报学研究过程的系统方法论。王芳等[21]利用《情报学报》1999—2008年这10年间的文献数据，在对于1174篇文献逐一分析统计的基础上，采用计量分析的方法揭示了我国情报学研究中各类方法的使用情况，包括使用频率与变化趋势等；研究结果显示在情报学领域、计算机领域的相关研究方法占据较大的比例并且呈现出上升的趋势。接着，王芳等[22]又对我国情报学研究中混合方法的应用及分布进行分析，对2000—2014年《情报学报》发表的1950篇论文进行编码，采用统计分析的方法并结合可视化，呈现出当前情报学方法的混合运用频次与运用类型的变化情况，结论是我国计算机类方法的混合应用最为广泛，这也印证了情报学研究的逐渐趋向于实证型与技术型的方向。化柏林[23]利用知识抽取的相关技术方法，制定相关规则并与词表相结合，实现方法术语的提取，从而建立情报学研究方法语料库，这是研究情报学方法从人工统计分析到实现机器自动识别的一个重大进步。

通过文献调研发现，我国情报学领域，对于理论、方法的研究所采用的方法更多的是文献阅读法、归纳法、内容分析法以及统计分析等传统人工方法，这些方法不仅需要投入较多的人力，进行文献阅读、标记、统计，也会耗费大量的时间，同时对于研究人员的专业水平要求较高（需要依靠专业知识进行判断）。上述条件的约束，使得在进行研究分析时，文献数据量有限，在此基础上得到的分

析结果有可能会产生一定的误差。因此，本文提出借助深度学习模型——Bi-LSTM-CRFs，利用术语识别任务对情报学理论方法进行研究，对情报学领域近 20 年上万篇的文献数据进行大规模自动化的识别处理，通过实验分析确定其可行性，并且在此基础上进行计量分析，得出一般规律。本次研究能够给之后的领域理论方法研究带来新的方向，从人工转向自动化，从主观判断转向客观展示，同时也可以为后人的研究奠定一定的基础。

2.2 命名实体相关研究

命名实体识别是指从文本数据中提取出命名实体部分的过程[24]。随着命名实体识别研究的不断发展，其识别内容也不仅仅局限于人名、地名、机构名等范围，电子病历命名实体识别[25]、军事命名实体抽取[26]、金融领域术语识别[27]、专利术语识别[28]，以及理论术语抽取[29]等均是命名实体识别任务的分支。

命名实体识别主要研究方法包括基于规则方法、基于统计机器学习方法、基于混合方法以及基于深度学习方法。基于规则的方法是早期命名实体识别的主流方法，通过人工构建规则并进行字符串匹配，核心思想与人类思考的方法较为一致，简单易懂，且准确率较高；但其规则构建代价较大、可移植性不好，同时复杂的规则会影响识别效率。随着计算机技术的不断发展，数据量及数据复杂度都难以再用简单的规则处理，因此机器学习的兴起使得命名实体识别的发展进入了新的篇章。与基于规则的方法不同，基于统计的方法不需要利用人工制定的规则来进行实体识别，而是利用预先标注的语料进行训练学习，从而自动地从未标注语料中实现命名实体的识别。多种方法混合也是命名实体识别中经常采用的方法，取长补短使不同方法能够焕发新的活力。规则和统计相结合的主要优点在于既可以运用统计模型降低人工制定规则的成本，又可以利用统计模型自动抽取较为抽象且适用性较广的规则，与此同时，规则的加入也能够在一定程度上减少对大规模标注数据集的依赖。随着深度学习的热度不断攀升，许多学者将其最新的技术运用于命名实体识别中，并取得了不错的效果。将词向量作为特征向量，既可以实现词语的数字化表达，又能够包含更多的语义信息，对于命名实体的识别效果非常好[30]。Collobert 最早提出 NN（neural network）结构与 CNN（convolutional neural network，卷积神经网络）结构相结合的方法，从词级别与句子级别两方面入手分别进行模型训练，最终使得识别结果提升。后续研究在此基础上不断改进，从 RNN（recurrent neural network，循环神经网络）发展而来的 LSTM（long short-term memory，长短时记忆）[31]和 Bi-LSTM（bidirectional LSTM，双向长短时记忆）[32]经过验证，成为命名实体识别的主流方法。随后有学者将统计机器学习的方法 CRFs（conditional random fields，条件随机场）加入[33]，形成 Bi-LSTM-CRFs 模型，将 F1 值提升了近 5%。CNN 模型[34]、attention 机制[35]、主动学习机制[36]、迁移学习机制[37]的加入均使得 Bi-LSTM 模型在命名实体识别的过程中不断进行算法的优化，运算效率及准确率都得到很大的提升。

3 数据来源及模型构建

3.1 术语的定义

本文以理论术语和方法术语作为抽取对象，因此需要对这两类术语实体进行定义和说明，以确保探索对象和实验方向的准确性。

首先，笔者认为理论术语应具备如下特征：

（1）词特点：一般词中有定理、定律、理论、法则、论、原理、框架、机理、学说、假说、效应、猜想、模型等；

（2）上下文特点：前文一般会有基于、基础、采用、根据、用、提出、按照、引入、构建；

（3）示例：如表 1 所示。

至于方法术语，其在学术文献中的特征应归纳为：

（1）构词特征：一般词中包含法、方法、算法、技术；

（2）上下文特征：一般会有采用、运用、用、采取、根据、基于等；

（3）示例：如表 2 所示。

3.2 术语识别方法

情报学理论和方法术语识别主要包括 4 个步骤，具体如图 1 所示。

1）数据采集

由于此次实验所需数据均为情报学领域文献数据，包括题目、关键词、摘要等，数据较为规范且有较为便捷的下载方法，因此采用人工手动采集。

2）数据预处理

数据预处理过程包括数据清洗、数据预处理

表 1　理论术语示例

尾词	例句
理论	本文以**复杂适应系统理论**为基础，分析了危机信息管理的复杂特征
框架	信息管理流程、信息传播路径和信息系统建设三个维度建构了**危机信息管理体系框架**
定律	对各年段作者量进行**洛特卡定律**的拟合
原理	其面向宏观分析的技术系统进化法则和面向微观分析的**技术矛盾创新原理**在专利技术路线图的制定中都能够发挥重要的作用
学说	网络媒体的出现使**米哈依洛夫科学交流学说**受到了强烈的冲击
假说	结论一定程度支持了**牛顿假说**
机理	具体阐释了领域知识间的内在关系及**协同作用机理**

表 2　方法术语示例

尾词	例句
方法	运用**统计分析方法**、**类比法及路径分析**等方法
算法	运用Network X软件的**投影算法**

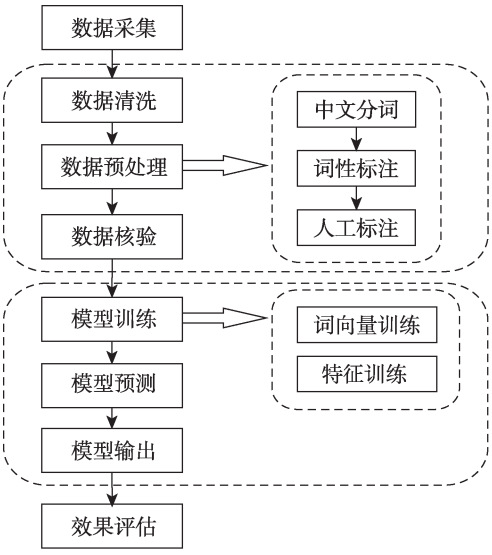


图 1　实验流程图

（中文分词、词性标注、人工标注）以及数据核验这几个步骤。其中数据清洗是为了去除无用数据；数据预处理是十分重要的环节，尤其是人工标注，标注量较大且标注人员需要具备一定的专业背景知识，同时标注的规范性与准确性对于模型训练效果影响较大；数据核验是为了对已经完成预处理的数据进行二次查验，纠正标注错误数据、补足未标注数据等，并且因为要采用五折交叉验证的方法保证实验结果偏差较小，因此需要将数据分成 5 份。

3）模型训练集预测

训练 Bi-LSTM-CRFs 模型并进行多次实验预测。在这个过程中均采用五折交叉验证，即随机选取数据集中的 1 份作为测试语料，其余 4 份作为训练语料，此过程重复 5 次，之后实验结果取 5 次的平均值。

4）实验效果评估

记录实验结果并进行详细分析。

不难发现，要完成上述过程，构建理论和方法术语的文本抽取模型并加以实验论证，需要解决如下关键性问题：数据采集、算法选择、语料标注、特征设置等。

3.3　数据来源

本文实验采用的文献信息包括文章标题、关键词、摘要、作者、单位等关键信息，其中标题、关键词以及摘要是全文的高度概括，一般文章的主要研究方法及理论支撑均会在其中有所体现。CNKI 期刊导航中，科技信息栏目中包含图书情报与数字图书馆专题，共有期刊 68 种，其中学术期刊共计 55 种。由于部分期刊中含有图书馆、计算机方向的文章，故需要进行限定条件的检索。

将 CNKI 作为检索数据库进行高级检索，检索条件为：中图分类号=G35（情报学、情报工作）；发表时间：1998 年 1 月 1 日至 2017 年 12 月 31 日；文献来源：55 种图书情报学术期刊名。需要注意的是，情报学著名期刊《情报学报》在 CNKI 中收录并不齐全，为此，笔者专门从万方数据库中采集了该杂志收录的相关学术文献。

将检索所得期刊论文提取题名、关键词及摘要等字段，并进行人工筛选，将会议通知、征稿启事、期刊索引、题录、精选文摘、英文摘要以及没有摘要的文献等进行剔除，最终选取期刊 48 种（将文献数少于 20 篇的期刊剔除），共计保留 18851 篇文献，包含 3610994 个字符。不可否认，从深度学习的机制而言，近 2 万篇的文献量的确偏小，不足以发挥出深度学习算法的优势。论文选择这样的一个数据量，主要基于以下考虑：

（1）就 G35（情报学、情报工作）类目而言，近 20 年（1997—2018 年）是其飞速发展的时期，

该时段期刊论文中蕴含了大量的理论和方法术语，而且该时段的数据相对来说比较完整、规范，比较适合作为本文实验对象。扩大年份范围，容易引入不规范的老旧数据；扩大学科范围，又不利于机器学习特征的聚焦。

（2）基于数学理论，大数据量显然有助于发挥深度学习的优势，但是也会明显拉大一般机器学习和深度学习算法之间的差距，不利于发现深度学习算法在具体场景应用中存在的问题或不足，进而影响分析一般机器学习算法和深度学习算法各自的适用场景或应用环境。

3.4 Bi-LSTM-CRFs模型介绍

领域术语识别作为一种特定类型的命名实体识别任务可以转化为序列标注问题[38]。而在序列标注问题上，传统的机器学习模型如HMM（hidden Markov model，隐马尔科夫模型）、MEMM（maximum entropy Markov model，最大熵马尔科夫模型）以及CRFs等均表现出了良好的性能。特别是CRFs，由于其可无限制扩充特征、引入之前标记状态、求解序列联合概率以解决标记偏置问题等特征和优势[39]，几乎已经成为解决不同场景序列标注问题的标配，在众多传统机器学习模型中表现出最佳的标注效果。CRFs利用观察对象本身和衍生特征、观察对象一定范围内的上下文特征，以及之前观察对象的标注结果等，来自动生成其标注状态。

深度学习则起源于人工神经网络（artificial neural network，ANN）的研究，原指含有多个隐藏层（hidden layer）的神经网络结构，而隐藏层的层数就是所谓的深度，后来逐渐应用于其他数学模型。从模型结构上讲，隐藏层越多即深度越大，模型越复杂，可适用的数据量也越大，自主特征学习也会越充分。RNN就是一种典型的深度学习算法，它允许隐藏层间的神经元相互连通，使得整个神经网络不仅考虑当前观察对象，也会对之前的学习内容具有一定的“记忆”功能，因此相对于CNN而言，RNN更适合对序列数据进行建模[40]。但是RNN短时“记忆”功能的特点，使得当输入较长序列时，RNN会遗忘之前的学习内容，甚至发生梯度消失或梯度爆炸现象。在这种背景下，LSTM模型作为一种改进的RNN被提出，用于解决后者无法适应的长距离依赖问题[31]。LSTM通过在神经元中设置记忆元件来实现对“前面”长距离特征的“记忆”或利用，同时为了能够利用“后面”的长距离特征，LSTM结合后向传播算法，进一步被改造为Bi-LSTM[32]。然而，标准的Bi-LSTM只能通过当前输入以及前后长距离特征预测分类标记，无法考虑序列的全局最优问题，换言之，之前的预测结果无法对当前的预测造成影响。为此，将Bi-LSTM和能够计算序列全局最优解的CRFs结合在一起，以Bi-LSTM的输出作为CRFs的输入，结合更多的特征（转移特征，即之前的标注结果）进行再计算，以获得最优的序列标注，最终形成了Bi-LSTM-CRFs模型[41]。

由此可见，Bi-LSTM-CRFs结合了深度学习算法Bi-LSTM和传统序列标注算法CRFs各自的优点，既具有适应大数据环境、算法结构更加复杂、自主获取序列长距离上下文特征等优点，又具备综合利用之前标注结果进而计算序列全局最优解的能力，从理论上讲应该能够表现出更优秀的标注效果，因此本文选用该模型进行后续的实验探究。但是也不难发现，Bi-LSTM-CRFs的本质或核心依然是Bi-LSTM，它并没有包含CRFs的全部功能或特点，特别是CRFs能够无限扩展特征并人为设置特征范围和来源的优点。因此在序列标注问题上，深度学习算法Bi-LSTM-CRFs和传统算法CRFs都是适用的，但是谁优谁劣，仅理论比较并无法得到确切结果，只能通过它们在不同实验或应用场景中的表现做进一步论证。

3.5 语料标注方法

本文采用的标注方法为BIO标注法[29]，即Begin, Intermediate, Other，用B-X表示此元素所在的片段属于X类型并且此元素在此片段的开头，I-X表示此元素所在的片段属于X类型并且此元素在此片段的中间位置，O表示不属于任何类型。

1）基于字标注

采用基于字的标注方法不需要对于语料进行分词处理，只需要将所有字符逐个分离，在标注时，用“B”标注术语的首字，用“I”标注术语内包含的其他字，用“O”标注非术语的字与符号等，如表3所示。

2）基于词标注

采用基于词的标注方法，首先需要对于语料进行分词处理，在标注时，用“B”标注术语的首词，用“I”标注术语内包含的其他词，用“O”标注非术语的词与符号等。若如该术语只有一个词，则用“B”标注该术语，如表4所示。

表3　基于字标注

处理	句子
原句	本文主要的研究方法为主成分分析法
切分句	本/文/主/要/的/研/究/方/法/为/主/成/分/分/析/法/
标注句	本/O 文/O 主/O 要/ O 的/O 研/O 究/O 方/O 法/O 为/O 主/B 成/I 分/I 分/I 析/I 法/I

表4　基于词标注

处理	句子
原句	本文主要的研究方法为主成分分析法
切分句	本文/主要/的/研究方法/为/主成分/分析法
标注句	本文/O 主要/O 的/O 研究方法/O 为/O 主成分/B 分析法/I

3.6　特征设置

深度学习模型对语料进行训练的过程其实就是不断捕获语料特征的过程，训练得到的模型对训练语料特征学习效果越好，那么在后面的自然语言处理任务中所得到的结果就越好。添加词向量、词长、词性、拼音等特征是为了用不同的语言单位形态来对语料加以表示，多角度、多层次地表示出训练语料以方便深度学习模型更好地训练，不同的向量化的手段可能让深度学习模型在训练过程中捕获到不同的训练语料的信息从而得到更好的模型结果。

1）词向量

词向量是将词转化为稠密向量表示，并且能够体现出词语的语义关系，对于相似的词语，它们的词向量也较为相近。在自然语言处理中，词向量一般都会作为深度学习模型的重要特征进行输入，对于模型的训练效果起着十分重要的影响。

2）词　长

词长是领域专业术语的一个重要特征，很多学者对此进行了研究，周浪[42]提到词组型术语是术语体系中的主流，且长度集中在2~6字。随后周浪等[43]对计算机领域的术语进行词长统计发现，2~6字的术语所占比例最大，达到了91.39%。因此，加入词语的词长作为特征辅助判断该词语是否为术语或属于术语的一部分。

3）拼　音

拼音作为中文汉字的一个重要属性，可以在一定程度上反映词语的特点，因此也可考虑加入作为其中的一个特征，用于补充其他特征不具备的信息。

4）词　性

词性在本质上就是用于反映词语的主要功能与其蕴含的语法信息，而很多的术语都有其一定的构词规律，并能通过词性体现出来。有学者对汽车术语进行统计分析发现，“名词+名词”、“动词+名词”以及“名词”为频率较高的3种术语词性组合，因此可以表明，词性是术语识别中一个非常重要的特征。本次实验采用结巴分词中的词性标注体系。

4　实验结果及分析

本文实验的评价指标为理论和方法术语识别的*P*值（precision，准确率）、*R*值（recall，召回率）以及F1值（F1-measure）。实验采用五倍交叉验证的方法，即将处理好的语料数据依照五倍交叉的方法分成5份，在实验过程中，将其中的4份作为训练集，剩余的1份作为验证集，并进行5次实验，最终的实验结果取5次实验结果的平均值，从而尽量减少由于数据的不平衡对实验结果产生的影响。

实验采用Bi-LSTM-CRFs作为深度学习算法用于训练标注模型，该算法的主要参数设置如表5所示。

表5　Bi-LSTM-CRFs算法的参数设置

参数名	参数值
epoch	15
Batch_size	20
hidden layer(隐藏层)	300(维)
learning rate(学习率)	0.001
dropout	0.5
learning rate_decay(学习率衰减)	0.9
learning rate_method	adam(算法)
epoch_no_improve	3

4.1　不同数据类别对比实验

本次实验采集的文献数据包括题目、关键词以及摘要信息，这3类数据中都包含着需要识别的术语实体，因此对这3类数据分别进行实验，通过实验结果来了解各类数据的术语识别情况，并确定后续实验所用数据。实验结果如表6所示。

表6　不同数据类别对比实验

数据类别	*P*	*R*	F1
题目	57.28%	67.25%	61.87%
关键词	62.40%	66.89%	64.57%
摘要	71.58%	63.51%	67.30%
全数据	75.41%	63.34%	68.85%

在题目、关键词以及摘要这3类数据中，利用摘要数据进行训练的F1值最高，也就是训练效果最好。但我们也发现，题目和关键词数据的召回率明显高于摘要数据，准确率则低于摘要数据，这可能与这两类数据中包含的术语实体数量较少有关。表7则统计了各类数据中包含的各类实体数量，可以看出题目与关键词数据中的术语实体数量远远少于摘要数据。而深度学习的模型需要大量的语料进行学习训练，术语实体越多，能够学习到的特征就越多，训练效果就越好，当使用全数据（包括题目、关键词、摘要）时，实验结果F1值最高，因此，在后面的实验中，均采用全数据，能够最大限度地提升模型的训练效果。

表7 不同数据类别中各类术语实体的数量

数据类别	理论术语	方法术语
题目	406	3945
关键词	762	5378
摘要	1824	12180
全数据	2992	21503

4.2 基于字标注与基于词标注的对比实验

本次实验的变量为文本的标注方式，即用于模型训练和测试的语料采用同一种标注方法，其余条件均保持一致；一种标注方式为基于字标注的，一种标注方式是基于词的标注。基于字的标注方式是在将语料按照单个字进行切分后进行人工标注，而基于词的标注方法是在将语料应用结巴分词代码进行分词处理后进行人工标注。结果如表8所示。

表8 基于字标注与基于词标注对比实验

	P	*R*	F1
基于字标注语料	72.66%	61.58%	66.66%
基于词标注语料	75.41%	63.34%	68.85%

观察表8结果可以看出，基于使用词标注语料的实验结果各项指标均高于基于字标注语料，*P*值高出2.75%，*R*值高出1.76%，F1值高出2.19%。可见，在没有其他特征的支持下，汉语中“词”比“字”具有更加丰富的语义，也使得前者比后者对标注对象角色类型的揭示力度更大更深刻。

此外，笔者对于识别出来的所有理论方法术语字数进行了统计，其不同字数长度术语的分布如图2所示。不难发现，4~6字术语占比近60%；而在之前的命名实体识别研究中，有研究结果表明，基于

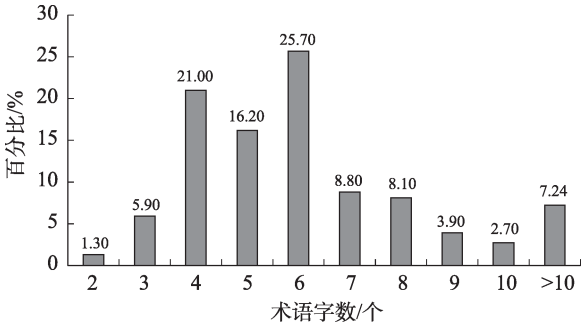


图2 被抽取术语的长度统计

字的方法效果要好于基于词的[44]，那是因为在传统的命名实体任务中，实体一般为人名、地名等短名词，而在领域术语识别中，术语长度一般较长，基于字的方法适用性较差。在基于词的方法中，由于分词系统的缘故，或多或少会产生一定的切分误差，尽管本文已经引入了自定义词典，但误差依然存在。但相比于分词产生的误差，由分词处理后带来的语义信息对于模型训练产生的优点更大，因此，在进行领域术语（尤其是长术语比重较大的情况）时，采用基于词标注语料的效果更佳。

4.3 不同比例训练集对比实验

为了探究训练集语料对于模型抽取性能的影响，本次实验设计了基于不同比例训练集的对比实验，将数据集随机分成5份，分别利用20%、40%、60%以及80%的语料量作为训练集，同时固定20%的语料作为统一的测试集，实验结果如表9所示。

表9 不同比例训练集对比实验

训练集占比	*P*	*R*	F1
20%训练集	57.33%	47.25%	51.80%
40%训练集	63.42%	53.42%	57.99%
60%训练集	70.40%	58.13%	63.68%
80%训练集	75.41%	63.34%	68.85%

从实验结果数据可以较为清晰地看出，*P*值、*R*值及F1值随训练集的增多而增大，呈正相关。因此，为了提升模型的抽取效果，需要尽可能多地提供训练集，从而使模型尽可能达到充分学习。

4.4 不同实体类型识别对比实验

在上文实验中，语料中同时包含标注了“理论术语实体”和“方法术语实体”两类术语。那么，为了验证实体类型及数量对于实验结果的影响，在本节实验中，笔者进行了两项控制：一是保证每次

实验的语料中仅包含一种类型的实体，以保证实验的专注度；二则是另外多增加了“模型术语实体”，以保证术语类型的多样性。实验结果图 3 所示。

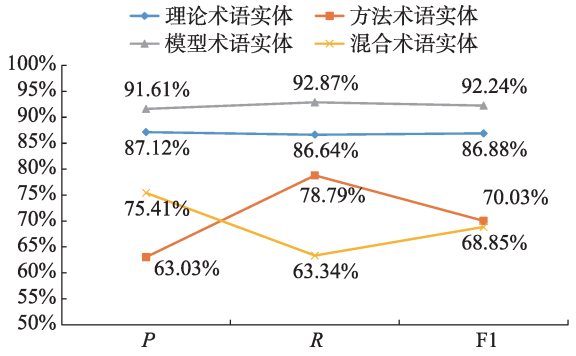


图 3　实体类型对实验结果的影响

可以看出，模型术语实体的识别结果最好，F1 值高达 92.24%；而方法术语实体的识别结果最差，F1 值仅为 70.03%，二者差距较大，达到 22.21%。笔者猜测这个差距可能是实体数量引起的，因为 F1 是由 *P* 值和 *R* 值共同决定的，因此笔者想探究一下实体的数量与 3 个指标之间的关系。通过统计各实体类型的数量（结果如表 10 所示）发现，在单一实体类型中，方法术语实体数量最多，为 6500 个；而理论术语实体数量最少，为 1541 个。观察图 3 中的 3 个单一实体的实验结果，笔者发现召回率以及准确率与实体数量并没有较为明确的关系，召回率最高值出现在模型术语实体中，最低值出现在方法术语实体中；准确率最高的为模型术语实体，最低的为方法术语实体；将混合术语实体一同加入比较发现，混合术语实体的数量为三种实体的总和，但其准确率并不是最低的。因此笔者可以得出结论，实体的数量并不能直接决定实体识别的结果。

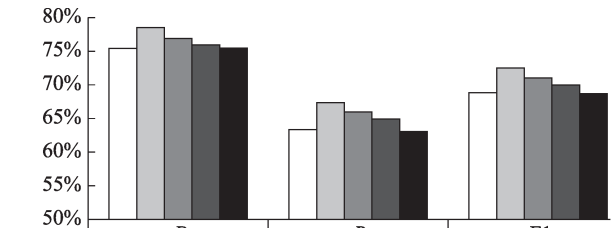
表 10　实验语料中各实体数量

实体类型	模型术语实体	理论术语实体	方法术语实体
实体数量	2079	1514	6500

为了探究为何模型术语实体的识别结果最好，笔者观察了文本语料，并发现模型术语实体的构词特征非常明显，即词尾均以“模型”二字结尾，同时通过大量的语料训练，模型很容易学习到这一特征，因此在识别过程中能够较为准确地识别出该实体。而理论术语尽管实体数量最少，但其构词特征没有模型术语明显，其词尾结束词包括“理论”、“论”、“机制”、“规律”、“现象”、“定律”等多种，因此识别结果略差于模型术语实体。

4.5　不同特征对比实验

本次实验引入了 4 个特征，分别是词向量特征（word2vec）、词性特征、词长特征以及拼音特征。在模型中分别加入这 4 个特征与原始无特征（即仅包含词本身）的模型进行对比实验，实验结果如图 4 所示。



	P	*R*	F1
□无特征	75.41%	63.34%	68.85%
□词向量	78.51%	67.38%	72.52%
■词性	76.89%	65.97%	71.01%
■词长	75.92%	64.93%	70.00%
■拼音	75.46%	63.05%	68.70%

图 4　不同特征对比实验

通过图 4 实验结果可以看出，加入词向量特征对模型训练效果的提升最为明显，F1 值提高 3.67%。本文采用的是百度百科数据，进行预训练得到的词向量特征（word2vec），能够较好地反映出词语之间的语义关系，具有更好的泛化能力[29]。词性特征的加入使得训练效果有明显的提升，其中 *P* 值、*R* 值以及 F1 值均有所提高，其中召回率提升最大，为 2.49%，这说明理论方法术语具有较好的句法特征，词性特征的加入能够使得模型更好地学习到较多的句法特征。词长特征对于模型的 F1 值有所提升，但效果不明显。拼音特征的加入略微降低了整体效果，其中主要是召回率略微下降导致最终 F1 值降低。

4.6　与传统 CRFs 模型的对比实验

为了多方面评估 Bi-LSTM-CRFs 模型的效果，本研究设计了与传统 CRFs 模型进行对比的实验。实验采用 CRFs++ 工具（https://sourceforge.net/projects/crfpp/），版本号为 0.58，参数定义为：Freq=1，eta=0.00010，C=1，shrinking size=20，定义的特征模板如表 11 所示。

本节对比实验仅仅对两个模型的基本性能进行对比，采用最基本语料数据，即仅包含分词结果与标注符号，不添加词性、词长等其他特征；实验方

表 11 CRFs 特征模板

编码	说明
U00:%x[-2,0]	左邻第二词
U01:%x[-1,0]	左邻第一词
U02:%x[0,0]	当前词
U03:%x[1,0]	右邻第一词
U04:%x[2,0]	右邻第二词
U05:%x[-1,0]/%x[0,0]	左邻第一词/当前词
U06:%x[0,0]/%x[1,0]	当前词/右邻第一词

法与上文相同，均采用五折交叉实验进行开放语料训练测试，并且训练语料占 80%，测试语料占 20%；又因为该实验建立在第 4.4 节实验的基础上，因此多增加了一类“模型”术语实体。最终结果如图 5 所示。

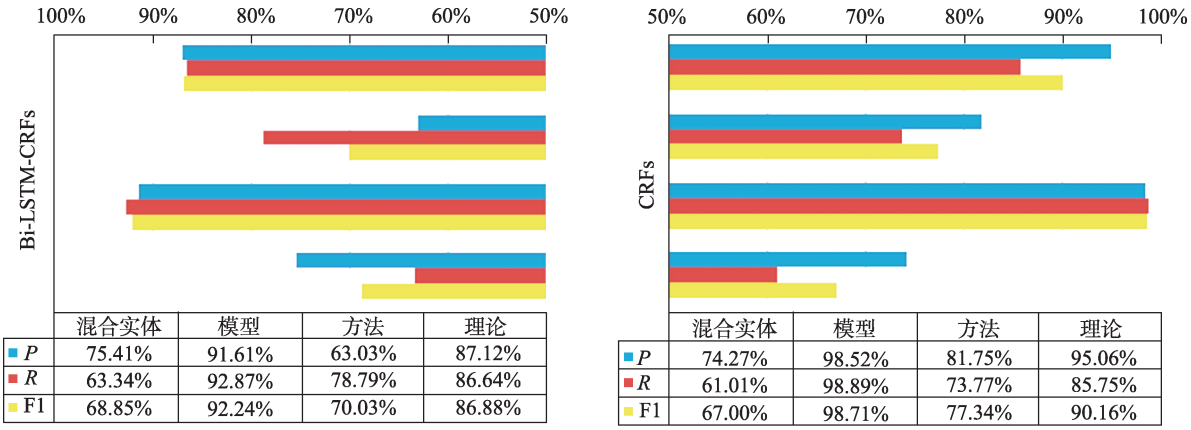
通过图 5 可以看出，从不同实体的识别效果来看，CRFs 的识别结果与 Bi-LSTM-CRFs 基本一致，即模型术语实体的识别效果最好，其次是理论术语实体，识别结果最差的是方法术语实体，再次印证了在命名实体识别中，实体的类型和数量对识别结果影响较大。同时，对比不同模型对同一种类实体的识别效果，可以看出，在识别单一实体的任务中，CRFs 模型的整体效果明显优于 Bi-LSTM-CRFs 模型，尤其体现在 *P* 值和 F1 值上，但 Bi-LSTM-CRFs 模型对于理论术语实体和方法术语实体的召回率要高于 CRFs 模型。通过对比混合实体的识别效果，我们可以看到，Bi-LSTM-CRFs 模型的准确率、召回率以及 F1 值均略微高于 CRFs 模型。

综上，在本文的实验数据环境中，在不附加其他特征的情况下，CRFs 模型对于单一实体类型的识别效果要优于 Bi-LSTM-CRFs 模型；但在处理包含多种实体类型的复杂语料中，Bi-LSTM-CRFs 的优势便凸显出来，如果能够扩大语料量并增加实体的复杂度，这个优势可能会更加明显。这是因为，多隐层的引入使得深度学习模型的神经网络结构比一般机器学习更为复杂，能够对数据特征进行自主识别和学习，因此也更能够适应数据量巨大且标注复杂的语料环境，而对于规模较小且标注简单的数据集来说，深度学习的优势便很微弱，甚至有可能效果还不如传统的独具特色的机器学习算法。在本次实验中，尽管数据量达到近 400 万字，但对于深度学习来说，数据量仍然是远远不够的，而通过混合实体的识别结果来看，在标注状态比较复杂（标注号多）的语料环境中，Bi-LSTM-CRFs 模型还是要优于传统的 CRFs 模型。

5 结 语

本文主要内容是利用 Bi-LSTM-CRFs 模型对情报学理论方法术语进行识别实验，通过多组对照实验研究不同变量对于实验结果的影响。实验结果表明，对于理论方法术语等复杂实体，基于词切分的语料识别效果要优于基于字切分的语料；术语实体的长度对于识别效果也有一定影响，术语长度过大时（字数≥6），识别效果下降明显。同时，训练语料量与识别效果呈正相关关系，语料量越大，识别效果越好；实体的类型和数量直接影响识别结果，具有明显构词特征的实体识别效果较好。本次实验也对 Bi-LSTM-CRFs 模型进行了一定的改进，在输入层中加入词向量、词长、词性以及拼音特征用于提高模型的识别效果，结果发现除拼音特征外，其余特征均能够使 F1 值有所提高，其中词向量和词性特征的提升效果最为明显。

针对本文的研究内容与研究成果，提出了以下



Bi-LSTM-CRFs	混合实体	模型	方法	理论
P	75.41%	91.61%	63.03%	87.12%
R	63.34%	92.87%	78.79%	86.64%
F1	68.85%	92.24%	70.03%	86.88%

CRFs	混合实体	模型	方法	理论
P	74.27%	98.52%	81.75%	95.06%
R	61.01%	98.89%	73.77%	85.75%
F1	67.00%	98.71%	77.34%	90.16%

图 5 Bi-LSTM-CRFs 和传统 CRFs 的对比实验

几个改进的方向：

（1）本文的实验语料选取了情报学领域相关文献，限定检索条件为“中图分类号=G35*”，但部分情报学文献的分类号并不是“G35*”，因此本次收集的数据可能不够全面，缺失部分数据；并且由于处理能力有限，仅选取了文献的题目、关键词以及摘要等数据；在后续研究中，可以扩展文献数据的范围，并收集全文数据，以提高模型的训练效果。

（2）本研究为了提高 Bi-LSTM-CRFs 模型的术语识别能力，在输入层加入了词性、词长以及拼音等特征，实验结果表明，特征的加入能够明显地提升训练效果，因此在后续的研究过程中，可以继续探索相关特征对训练效果的影响，如构词特征、上下文特征等。

（3）标注语料是模型训练的基础，人工标注语料首先需要投入大量的人力且对标注人员的专业背景有一定要求，其次会存在由于主观判断产生的分歧和失误，因此如何降低对人工标注的依赖是接下来的研究重点。

参　考　文　献

[1] Bush V. As we may think[J]. The Atlantic Monthly, 1945, 176: 101-108.

[2] 包昌火, 刘彦君, 张婧, 等. 中国情报学论纲[J]. 情报杂志, 2018, 37(1): 1-8

[3] Henshel R B R L, Wallace W L. The Logic of science in sociology[J]. Contemporary Sociology, 1972, 1(6): 520-521.

[4] 符福峘, 陆婷. 论情报学方法论体系的构建、发展和应用[J]. 情报理论与实践, 2007, 30(2): 149-153.

[5] 陆伟, 孟睿, 刘兴帮. 面向引用关系的引文内容标注框架研究[J]. 中国图书馆学报, 2014, 40(6): 93-104.

[6] 徐庶睿, 卢超, 章成志. 术语引用视角下的学科交叉测度——以 PLOS ONE 上六个学科为例[J]. 情报学报, 2017, 36(8): 809-820.

[7] Pettigrew K E, 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.

[8] Jeong D Y, Kim S J. Knowledge structure of library and information science in South Korea[J]. Library & Information Science Research, 2005, 27(1): 51-72.

[9] Kim S J, Jeong D Y. An analysis of the development and use of theory in library and information science research articles[J]. Library & Information Science Research, 2006, 28(4): 548-562.

[10] Kumasi K D, Charbonneau D H, Walster D. Theory talk in the library science scholarly literature: An exploratory analysis[J]. Library & Information Science Research, 2013, 35(3): 175-180.

[11] van de Water N, Surprenant N, Genova B K L, et al. Research in information science: An assessment[J]. Information Processing & Management, 1976, 12(2): 117-123.

[12] Tuomaala O, Järvelin K, Vakkari P. Evolution of library and information science 1965-2005: Content analysis of journal articles[J]. Journal of the Association for Information Science and Technology, 2014, 65(7): 1446-1462.

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

[14] Ferran-Ferrer N, Guallar J, Abadal E, et al. Research methods and techniques in Spanish library and information science journals (2012-2014)[J]. Information Research, 2017, 22(1): 1-8.

[15] 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.

[16] 王芳, 陈锋, 祝娜, 等. 我国情报学理论的来源、应用及学科专属度研究[J]. 情报学报, 2016, 35(11): 1148-1164.

[17] 王知津, 王璇, 韩正彪. 90 年代以来我国情报学理论研究期刊论文统计分析[J]. 图书馆理论与实践, 2012(1): 21-26.

[18] 陈锋, 翟羽佳, 王芳. 基于条件随机场的学术期刊中理论的自动识别方法[J]. 图书情报工作, 2016, 60(2): 122-128.

[19] 钱军, 杨欣, 杨娟. 情报研究方法的聚类分析[J]. 情报科学, 2006, 24(10): 1561-1567.

[20] 杨锐. 关于情报学方法体系建设的思考[J]. 情报探索, 2008(5): 126-128.

[21] 王芳, 王向女. 我国情报学研究方法的计量分析: 以 1999~2008 年《情报学报》为例[J]. 情报学报, 2010, 29(4): 652-662.

[22] 王芳, 祝娜, 翟羽佳. 我国情报学研究中混合方法的应用及其领域分布分析[J]. 情报学报, 2017, 36(11): 1119-1129.

[23] 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.

[24] 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.

[25] 杨红梅, 李琳, 杨日东, 等. 基于双向 LSTM 神经网络电子病历命名实体的识别模型[J]. 中国组织工程研究, 2018, 22(20): 3237-3242.

[26] 单赫源, 吴照林, 张海粟, 等. 结合词语规则和 SVM 模型的军事命名实体关系抽取方法[J]. 指挥控制与仿真, 2016, 38(4): 58-63.

[27] 梁晨. 金融领域术语识别的研究[D]. 大连: 大连理工大学, 2017.

[28] 杨双龙, 吕学强, 李卓, 等. 中文专利文献术语自动识别研究[J]. 中文信息学报, 2016, 30(3): 111-117, 124.

[29] 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.

[30] Cherry C, Guo H Y. The unreasonable effectiveness of word representations for Twitter nam ed entity recognition[C]// Proceed-

ings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2015: 735-745.

[31] Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling[C]// Proceedings of the 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, 2012: 601-608.

[32] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks[C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 6645-6649.

[33] Peng N Y, Dredze M. Improving named entity recognition for Chinese social media with word segmentation representation learning[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 149-155.

[34] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[35] Rei M, Crichton G K O, Pyysalo S. Attending to characters in neural sequence labeling models[C]// Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. The COLING 2016 Organizing Committee, 2016: 309-318.

[36] 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. Stroudsburg: Association for Computational Linguistics, 2017: 252-256.

[37] Yang Z L, Salakhutdinov R, Cohen W W. Transfer learning for sequence tgging with hierarchical recurrent networks[C]// Proceedings of the 6th International Conference on Learning Representations, Toulon, France, 2017: 234-253.

[38] Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[OL]. https://arxiv.org/pdf/1508.01991v1.pdf.

[39] 王昊, 苏新宁 . 基于CRFs的角色标注人名识别模型在网络舆情分析中的应用[J]. 情报学报, 2009, 28(1): 88-96.

[40] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]// Proceedings of INTERSPEECH 2010, the 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, 2010: 1045-1048.

[41] Rondeau M A, Su Y. LSTM-based NeuroCRFs for named entity recognition[C]// Proceedings of INTERSPEECH 2016, the 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 2016: 665-669.

[42] 周浪 . 中文术语抽取若干问题研究[D]. 南京: 南京理工大学, 2010.

[43] 周浪, 史树敏, 冯冲, 等 . 基于多策略融合的中文术语抽取方法[J]. 情报学报, 2010, 29(3): 460-467.

[44] 刘章勋 . 中文命名实体识别粒度和特征选择研究[D]. 哈尔滨: 哈尔滨工业大学, 2010.

（责任编辑　王克平）