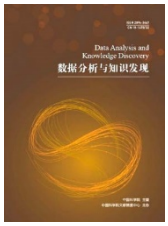

数据分析与知识发现
Data Analysis and Knowledge Discovery
ISSN 2096-3467,CN 10-1478/G2

《数据分析与知识发现》网络首发论文

题目：　基于机器学习的技术术语识别研究综述
作者：　胡雅敏，吴晓燕，陈方
网络首发日期：　2021-11-23
引用格式：　胡雅敏，吴晓燕，陈方．基于机器学习的技术术语识别研究综述[J/OL]．数据分析与知识发现.
https://kns.cnki.net/kcms/detail/10.1478.G2.20211123.1534.002.html

网络首发时间：2021-11-23 18:15:32
网络首发地址：https://kns.cnki.net/kcms/detail/10.1478.G2.20211123.1534.002.html

基于机器学习的技术术语识别研究综述

胡雅敏[1,2]，吴晓燕[1]，陈方[1,2*]

[1](中国科学院成都文献情报中心 成都 610041)

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

摘要：[目的]梳理机器学习算法在技术术语识别中的应用现状与前景。**[文献范围]**在 WOS 核心库和 CNKI 数据库中，以“technology term* recognition”、“技术术语识别”为检索词检索文献，并延伸阅读相关算法文献，共筛选 62 篇代表性文献进行述评。**[方法]**类比命名实体识别研究，归纳机器学习在技术术语识别中的应用和区别，从算法分类、一般流程、现存问题和下游应用 4 个方面进行梳理，并展望未来的应用前景。**[结果]**应用算法可分为单一的统计机器学习、单一深度学习和两者结合的混合算法，应用最广泛的是两者结合的混合算法，主流的模型代表是 BiLSTM-CRF 模型，迁移学习是未来重要的研究方向。**[局限]**深度学习快速发展，混合模型不断涌现，文中归纳的算法模型仅为应用较为广泛的算法，并未逐一列出。**[结论]**现有方法仍然有诸多待优化研究的问题，应加强细粒度的实体识别、特征表示方法、评估方法及开源工具包等方面的研究。

关键词：技术术语识别；机器学习；深度学习；模型

分类号：TP391.1

Review of the Technology Term Recognition Based on Machine Learning

HU Yamin[1,2], WU Xiaoyan[1], CHEN Fang[1,2*]

[1](Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041, China)

[2](Department of Library, Information and Archives Management, School of Economics and Management，University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [Objective] This paper reviews application status and prospects of technology term recognition based on machine learning. **[Coverage]** We searched “technology term* recognition” in Chinese and English with the Web of Science, CNKI, and extended the relevant algorithms literatures. A total of 62 representative literature were chosen for review. **[Method]** By analogy with the research of NER, this paper summarizes the application and differences of machine learning in technology term recognition, and combs it from four prospects: the algorithms classification, general process and existing problems and downstream application. Finally, we discussed the application prospect in the future. **[Results]** The application algorithms can be divided into single statistical machine learning, single deep learning and hybrid algorithms of both. The most widely used algorithm is the hybrid method, in which the mainstream model is the BiLSTM-CRF model. Transfer learning is an important research direction in the future. However, there are still many problems to be optimized. **[Limitations]** With the rapid development of deep learning, hybrid models are constantly emerging，the algorithm models summarized in this paper are only widely used algorithms, and were not listed one by one. **[Conclusion]** In the future, the research on fine-grained entity recognition, feature representation, evaluation and open source toolkits should be strengthened, and the future application prospects are also discussed.

Keywords: technology term recognition; machine learning; deep learning; models

---

作者简介： *陈方，ORCID: 0000-0001-9060-784X，Email: chenf@clas.ac.cn。

1 引言

词汇和术语是基本的认知概念，可以体现领域的知识，相关术语识别的技术被广泛应用在知识抽取研究中。命名实体识别（Named Entity Recognition，NER）可抽取文本中的命名实体，是信息检索、构建知识图谱和知识发现的基础和关键[1]。技术术语识别（Technology Term Recongnition，TTR）可识别专业领域文本中指代技术概念的词串[2,3]。一方面，大多 NER 研究仅包含时间、地点、人物等通用的命名实体，不足以体现研究领域的重要概念，而技术术语是一类粒度更细、更能体现知识语义的重要实体，且在科技领域的文本中相对规范，多为复合词；另一方面，随着科学技术的更新迭代，海量的科技文献不断报道新技术的发展和创新知识。因此，TTR 逐渐吸引了诸多研究者进行研究探索。TTR 挖掘出科技领域重要的技术词汇，有助于立体地了解领域的技术发展脉络和前瞻性的技术预见。

技术术语和命名实体都隶属于术语集合，且技术术语∈命名实体，如图 1 所示。术语是代表文档中重要的概念信息，并在语义上对文档进行表征[4]；命名实体是指被命名且具有实际意义的术语，除了 3 类通用类型，第 6 届 MUC 会议后续的研究逐渐增加预定义的实体类型，如“机构”、“货币”、“武器”、“产品”等 100 多种更细致的标签[5]；技术术语则是与技术概念相对应的术语[6]。如在“2020 年，中国农业科学院植物保护研究所在该专利中申请了一种用于基因组改造的杂合核酸序列”中，“中国农业科学院植物保护研究所”和“2020 年”是命名实体（分别为机构实体和时间实体），“基因组改造”是合成生物学领域的技术术语，而“杂合核酸序列”是普通术语。

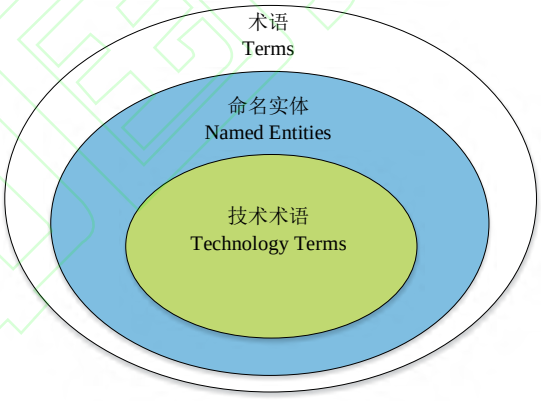

图 1 技术术语与命名实体的区别与联系

Fig.1 Differences and Relations Between Technology Terms and Named Entities

在图情档领域，传统的技术术语识别方法包括：对论文的著录信息（标题、摘要、关键词等），按照规则筛选出关键词汇，再对词汇进行主题聚类，最后归纳技术主题，如主题模型 LDA 方法[7] 等；对专利中的关键词汇、分类号，通过组合指标计算[8,9]、属性分类[10]、突变共词分析[11]、专利社会网络分析[12,13]等方法，对共性技术、关键技术、新兴技术、颠覆技术等不同活跃度的技术进行识别。该类方法是面向技术主题的识别，研究的知识单元是文本中的关键词汇[14]，最后依赖专家人工判读校验确定，没有真正考虑词语之间的语义关系。后继研究不断改进筛选词汇的方法，如利用词典式语义进行同义词聚类，增加信息熵、词语的位置、长度、密度等辅助信息[15,16]，或引入了 SPO（主语-谓语-宾语）、SAO

（主体-行为-客体）的语法分析，利用句法关系、语义结构提高筛选专利技术词汇的准确度[17,18]。然而这些改进方法都只是借助词性、语法结构等浅层语义信息，对词汇本身的语义信息利用仍然较少。

在机器学习快速发展的背景下，为了解决语义信息利用不足的问题，研究者开始在技术术语识别中使用基于内容语义的机器学习方法，尤其是结合深度学习的算法。机器学习的方法是将技术术语看作一种特殊的命名实体，迁移了 NER 相关方法进行技术术语识别[19]。当前关于 TTR 的综述，多基于传统文献计量方法的归纳，缺乏对机器学习相关算法的梳理，且鲜有从迁移 NER 方法角度的综述。因此，本文基于大量 NER 研究，总结 TTR 中应用的机器学习相关算法与区别，归纳应用机器学习进行技术术语识别的一般流程、存在问题及优化方案，以期为技术术语识别研究提供有用的方法参考。

2 应用机器学习的算法分类

技术术语识别应用的算法中，机器学习方法主要有 2 种：传统的统计机器学习算法和深度学习算法。为了凸显多种研究方法的不断改进和组合效果，本文将算法划分为 2 类：单一机器学习算法和混合机器学习算法。

2.1 单一算法

（1）单一的统计机器学习算法

统计机器学习算法，也被称为传统机器学习算法，主要包括隐马尔科夫模型（Hidden Markov Models，HMM）、支持向量机（Support Vector Machine，SVM）和条件随机场（Condition Random Fields，CRF）。统计机器学习算法进行实体识别的本质是丰富输入数据的特征，如词语的含义、上下文信息、词语位置等特征。本文总结了几种算法实现实体识别的本质区别和优缺点，如表 1 所示。

针对实体识别、序列数据标注分类任务，HMM 可看作一个有限状态自动机[20]，即有限个类别的转化过程。通俗地讲，HMM 建模过程是计算输入和输出同时发生的概率。输入一般是指输入的文本单元，具备人为可观察的外在特征，输出是指文本单元的内在所属类别标签。HMM 是将输入（外在特征）看作是输出（内在类别属性）经过马尔科夫随机过程生成的结果，属于生成式模型。岑咏华等[21]设计的基于双层 HMM 模型识别系统，在识别中文不同领域的学术论文中技术术语方面，表现出良好的性能。但 HMM 模型判定类别只考虑了当前观察的特征，没有考虑到实体的上下文特征。因此，HMM 广泛应用于词性标注、分词等场景，较少应用在技术术语识别中。

SVM 模型训练时，考虑了分类界限附近的样本点，即影响 SVM 分类决策的样本点是少数的结构支持向量，同时克服了维数灾难。Dodan 等[22]尝试使用 SVM 模型识别药物等实体。SVM 适用于少量样本的机器学习，但面对大规模样本或多分类问题时分类效果不理想[23]。因此，SVM 多应用在图像识别领域，较少应用在技术术语识别任务中。

为了解决 SVM 小样本效果不佳、HMM 考虑特征不足的弊端，CRF 应运而生。CRF 可以任意定义特征函数，因此提高了实体识别的精准度，是实体识别中应用最广泛的统计机器学习方法。CRF 将实体识别问题转化为判别式的序列标注分类问题。判别式模型是指在已知输入特征的条件下，判别其输出类别出现的概率，即条件概率。最早是由 Lafferty 等[24]提出使用 CRF 方法进行序列标注，

实验证明 CRF 的鲁棒性优于 HMM 等模型，并解决了标签偏差的问题。McCallum 等[25]使用 CRF 方法识别相关命名实体，在 CoNLL2003 数据集上的 F1 值达到 84.04%。McDonald 等人[26]扩展了数据集和实体类型，使用 CRF 识别生物医药文本中的基因和蛋白质实体，达到了 86.4%的精确度。黄菡等[27]在 CRF 的基础上，结合主动学习算法构造 AL-CRF 模型，识别法律术语时准确率和召回率达到 90%。但是，由于 CRF 模型复杂度高、训练代价大，一般结合深度学习算法作为类别标签解码器，较少独立地完成技术术语识别的任务。

表 1 主要统计机器学习算法的区别

Table1 Differences of Main Statistical Machine Learning Algorithms

模型	生成/判别式	优点	缺点
HMM	生成式模型	实际信息比判别式模型更丰富，单类问题灵活，充分利用了先验知识	输出独立性假设不合理，状态只考虑对应的观察序列；不能利用复杂特征；数据稀疏问题
SVM	判别式模型	小样本分类效果较好	大规模数据、多分类问题效果不佳
CRF	判别式模型	特征设计灵活，考虑了状态之间的关系	训练代价大，复杂度高

（2）单一的深度学习算法

统计机器学习算法是基于人工提取特征，工作较繁琐，且对训练语料要求较高，而深度学习算法可用于自动发现文本隐藏的特征。因此，深度学习算法在实体识别中的应用愈发受到关注。常用的深度学习算法有前馈神经网络（Feedforward Neural Network，FNN）、循环神经网络（Recurrent Neural Network，RNN）和卷积神经网络（Convolutional Neural Network，CNN），相应算法的特点及优缺点如表 2 所示。

CNN 主要用于特征学习，利用 CNN 对模型的输入层进行特征向量的训练，如字符级语义特征的捕获[28]。RNN 算法可用于特征向量学习和序列标注，其中长短期记忆网络（Long Short-Term Memory，LSTM），利用其门限机制（输入门、遗忘门、记忆门等）实现对长距离信息的更新和长时积累，目前已被广泛地运用到实体识别中。单向的 LSTM 能获取上文信息，保留了历史特征[29]。后来出现双向 BiLSTM（Bidirectional Long Short-Term Memory），通过补充下文信息以解决 RNN 长距离时梯度爆炸的问题。BiLSTM 模型既考虑了历史信息又兼顾了未来信息，已被证明是深度学习领域最适合做序列任务的模型[30]。

表 2 主要深度学习算法的区别

Table2 Differences of Main Deep Learning Algorithms

模型	特点	优点	缺点
CNN	用于训练输入层的特征向量	通过卷积核自动提取特征	卷积后末层神经元只得到了原始输入数据的少部分信息，无法解决长距离依赖问题，忽略局部和整体的关系
RNN	用于对文本序列进行编码	能够捕获序列单元之间隐藏的关系，即能捕捉长距离依赖关系	序列过长容易梯度消失；存在梯度爆炸问题
LSTM	获取了上文历史信息，使用 3 种门限机制控制记忆和遗忘	长序列输入效果更好，解决了梯度消失和梯度爆炸问题	不能处理更长的序列；训练时计算费时
BiLSTM	前向传播（历史信息）、后向传播（未来信息）	网络结构的记忆力记住全句的信息	并行计算的利用上不如 CNN

2.2 混合算法

在混合算法中，较多的是统计机器学习算法和深度学习算法结合使用，因此，该部分不再介绍单一种类的混合算法，聚焦于阐述统计机器学习和深度学习算法的混合算法应用。由于 CRF 可以对类别标签之间的关系进行约束，在神经网络标记类别后，输出层增加 CRF 层可以避免错误的标签输出。所以，应用比较多的混合方法都是与 CRF 算法的结合，主要有 CNN-CRF、LSTM-CRF、BiLSTM-CRF、BiLSTM-CNNs-CRF、BiLSTM-IDCNN-CRF、Att-BiLSTM-CRF，总结上述混合算法应用于技术术语识别的特点与区别，如表 3 所示。

表 3 主要的混合机器学习算法的区别

Table3 Differences of Main Hybrid Machine Learning Algorithms

模型	特点
CNN-CRF	CRF 用于提高标注的准确度，CNN 用于提取复杂的特征
LSTM-CRF	降低对语料的规范性要求
BiLSTM-CRF	获取双向信息，适用更复杂的语料
BiLSTM-CNNs-CRF	对 BiLSTM-CRF 的改进，适用于长句语料的识别
BiLSTM-IDCNN-CRF	对 BiLSTM-CRF 的改进，IDCNN 用于提高训练速度
Att-BiLSTM-CRF	加入注意力机制，突出重点，用于提高识别精度

基于 CNN 神经网络可提取复杂特征、CRF 便于标注的优点，曹依依[31]设计了一种基于 CNN-CRF 的实体识别算法框架，对中文电子病历进行对比，从病历中提取出身体部位、疾病、症状、检查及治疗方法五类实体，研究表明该算法的精确率和召回率都在 90%以上，效果较好。与 CRF 方法相比，LSTM-CRF 模型降低了对语料结构化、标准化程度的依赖。李明浩等[29]使用 LSTM-CRF 模型识别中医临床症状术语，并通过比较不同的 LSTM 单元的变体——GRU、CIFG 的识别效果，实验证明 CIFG 变体模型的 F1 值较高。

而 BiLSTM-CRF 是在 LSTM-CRF 模型的基础上改进的双向 LSTM 的联合模型，由 Lample 等[32]首次提出，是当前 NER 基于深度学习的最主流的方法。在技术术语识别应用中，袁慧[33]使用 BiLSTM 和 CRF 的联合模型从生态治理技术相关文献中识别生态治理的技术实体，验证了该模型的可行性。王昊等[34]在技术概念的基础上，融合了理论、方法等类型的技术术语，从我国近 20 年的情报学领域相关文献的标题和关键词信息中，识别出情报学的技术、理论、方法等专业术语，对比实验结果表明 BiLSTM-CRFs 模型比传统 CRFs 模型在标注复杂的语料环境下，识别的准确率、召回率以及 F1 值均略高于 CRFs 模型。面向国防科技领域的技术术语识别，王学峰等[35]为优化 BiLSTM-CRF 模型的输入层，增加了输入层的字向量表示，由领域内语料通过 Word2Vec 模型训练而得，该模型成功识别出了军事领域的 8 类技术术语，提高了识别精度。冯栾栾等[36]为进一步提升识别效果，将字符向量表示，结合词向量、句法等语言特征向量一起输入到模型中，识别了自定义的基础技术、综合技术、武器等军事技术术语，并提出了适用于技术术语识别的语言特征。刘宇飞等[30]扩展了技术术语识别的数据来源，从专利中识别数控系统领域的技术术语，F_1 值达到 99.63%，证实了 BiLSTM-CRF 在专利文本中的可行性，也可能说明专利文本中的技术术语密度高于论文类科技文献。

BiLSTM-CRF 模型在学习较长句子时，有可能因为模型容量问题丢弃一些重要信息，因此研究者进行了多种改进。BiLSTM-CNNs-CRF 模型中增加了 CNN

层，CNN 可用于提取当前词的局部特征，再将字符级向量输入到 BiLSTM[37]。Ma 等[38]利用此模型在 2003-CoNLL 数据集的实体识别实证中使 F1 达到 91.21%。Strubell 等人[39]为提高训练速度，改进了 CNN 算法，采用 IDCNN 模型（Iterated Dilated Convolutional Neural Network）识别实体。蒋翔等人[40]通过 BiLSTM-IDCNN-CRF 模型与其他主流模型的对比实验，识别生态治理领域的技术术语等实体，该模型 F1 值为 0.7207，均高于其他模型，证实了该方法在技术术语抽取中的有效性。

较新的技术术语识别方法是引入了注意力机制的 Att-BiLSTM-CRF 模型。Vaswani 等人[41]的研究表明，该模型能有效提高很多任务的性能，并提出了一个新的模型框架——多头注意力（Multi-head Attention）模型 Transformer，是在 Attention 的基础上实现对长距离位置关系进行建模，且并行计算提高了训练速度。注意力机制的重点在于通过权重分配，忽略噪音数据，获取更值得注意的关键元素，以便更精准地掌握全局信息，提高实体识别的性能。马千程等[42]使用 Att-BiLSTM-CRF 模型对商用飞机行业相关的互联网文本信息进行技术和机型两个重点关注的术语识别，研究结果证明加入注意力机制的模型比单向 LSTM-CRF 模型的准确率、召回率和 F1 值均提升 7%以上，比单一的 CRF 模型效果提升更明显。赵鹏飞等[43]在 BiLSTM-CRF 模型的基础上加入了文档级的注意力机制，通过获取实体相似度来确保实体标签的一致性，识别了农业领域的农作物、虫害和农药等术语，提高了模型识别性能。

为了解决新领域文本标注少、人工标注繁琐的问题，深度迁移学习逐渐进入实体识别研究者的视野。迁移学习的思想是利用领域的相似性，从数据更全面、更开放的源领域，转移到相似但样本稀疏标、注语料相对较少的目标领域，在源领域研究的基础上实现举一反三的效果。迁移学习模型可以共享模型隐藏特征和权重参数，节省了人力，提高了研究效率。同时，深度迁移学习还解决了跨语言训练的鸿沟问题[44]。BiLSTM-CRF 已被证明是深度学习领域最主流的序列任务模型，因此，深度迁移学习中的模型多为此模型。刘宇飞等[30]使用 BiLSTM-CRF 的迁移模型，以新闻领域公共数据为迁移的源数据，以数控系统领域专利文献为迁移的目标数据，实现了迁移实体识别。

综上所述，统计机器学习善于捕获输入层的特征，丰富表示方法，深度学习算法可以增强学习上下文语义特征，完成权重的自动优化，将两者结合的算法是当前技术与术语识别的主流方法。迁移学习让识别技术更高效地应用到其他领域，提高了深度学习的泛化能力。

3 机器学习进行技术术语识别的一般流程和存在问题

本文通过梳理机器学习算法在技术术语识别领域的应用，绘制了机器学习应用在技术术语识别中的一般流程，如图 2 所示。本节将解析每一环节中的重点工作并提出相应的难点，旨在确定影响识别性能的因素以及优化的途径，如表 4 所示。

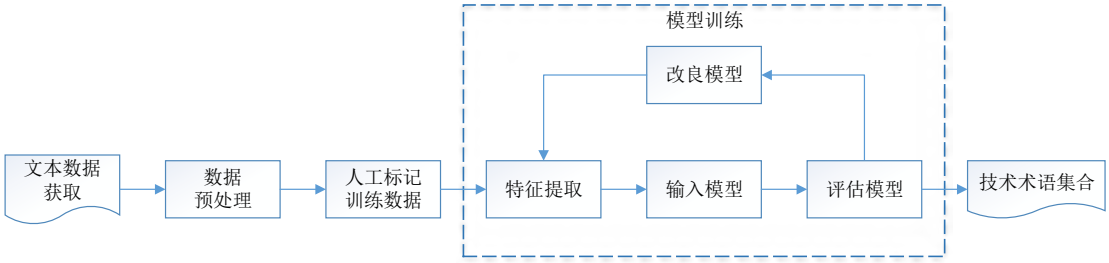

图 2 机器学习识别技术术语的一般流程

Fig.2 General Process of TTR Using Machine Learning

（1）文本数据获取阶段，面对众多的网络资源，如何获取更准确的文本数据、如何解决数据孤岛问题，从关联数据中获取更多的有效数据。如何将追溯性研究转变为前瞻性研究，对动态、实时的数据的获取与研究也是情报分析领域新的发展趋势[45]。

（2）数据预处理阶段，针对数据噪音太大的非结构化文本，需要进行更细粒度的过滤，以获取技术术语更密集的高质量文本。可在采集科技文献数据时，剔除文章中不包含技术术语的内容，保证语料的质量[36]。分词时可以使用 Word2Vec 模型预训练字向量的方法，缓解分词可能带来的错误问题。可采用的预训练模型还有 ELMO、BERT 模型及在此基础上改进的 BERT 系列模型，如 RoBERTa、ALBERT 等[46]。预训练采用的语料，除了通用的维基百科等，可根据研究领域的实际语料库进行预训练以增强输入数据的特征，如 Lin 等[47]在进行生物医学实体识别时，采用 PubMed 数据库进行训练。

（3）人工标记训练数据阶段数据标注的质量影响模型训练，但又不可避免会出现标记错误的情况。大多研究通过多次标记、计算标记一致性判断标记质量，一般一致性应达到 80%以上则达到标注规范。对于如何判断不同标记人员的标记一致性，可以通过不同标记人员标记数据的 F1 值判断[48]，具体以其中一名人员标记作为标准答案，计算 2 名标记人员的 F1 值。对于如何判断训练集和测试集的标注是否一致，Zeng 等[49]通过构造训练集中的 3 个互斥子集、1 个测试子集，两两组成 3 组新的训练集，验证 3 组不同的训练集识别一致性，如果 3 组数据预测结果不相似，则说明标记不一致。

同时，为了解决人工标记工作繁重的问题，除了使用迁移学习改善，研究者针对那些没有共性源领域的目标领域的研究，采用了半监督迭代深度学习的方式[50,51]。通常以少量标注数据训练初始模型，通过不断迭代学习预测生成更多的训练数据，直到达到阈值量。但是，如何判断迭代过程中的模型是否优化、伪标签训练数据如何降噪需要更深一步研究。标注实体方面，当前实体识别任务多以句子为研究单元，识别句子中的实体，而针对技术术语专业性较强的识别任务，只根据句子内容很难判断其是否为技术术语，可以引入篇章信息，对同一个短语标注时还强调识别的篇章一致性[36]。

（4）特征提取环节是为了获取更多的类别特征。在输入数据特征向量表示时，考虑不同粒度的特征向量表示可以改善识别精度，如词嵌入、字嵌入或混合的特征表示方法。蒋翔等[40]处理中文文本时，在分词信息中加入字向量，既借用了分词的信息，又使用字向量减小了错误分词的影响。也有研究表明字嵌入的效果优于词嵌入[52]，尤其是在预训练阶段的字嵌入识别效果更好，包含了字在全局范围内的上下文信息。Chiu 等[53]同时结合了词级特征（大小写、匹配词典）和字级特征（大写、小写、标点）进行技术术语识别。还有研究表明术语长度集

中在 2-6 字，术语的词长因素也会影响识别的效果[54]。技术术语识别也需结合研究领域的技术术语特点，如曾文等[55]在对科技政策文本中抽取术语时，总结了科技政策术语的 5 方面语言特点。因此，在输入特征向量表示时，可以考虑加入以上列举的有效附加特征，以制定更有针对性、准确性更高的特征表示方法。

（5）当前 NER 研究的传统评估指标包括：精确率（Accuracy），精确度 P 值和 F 值。但是这些指标只能比较不同模型的结果，无法识别不同模型之间的相对优势以及影响模型性能的特征因素，不能借此优化模型性能。Fu 等人[56]提出一种新的评估技术，核心思想是根据实体长度、密度等属性将数据划分为多个实体桶，然后分别在这些桶上评估模型，可以有效识别影响模型性能的因素。

（6）技术术语识别的目标是产生一个技术术语列表，该列表代表了研究领域关注的技术重点和专业信息。实体识别和关系抽取（Relation Classification，RC）是一体的研究，缺一不可。当前已有研究构造出基于神经网络的实体识别和关系抽取的联合学习模型[57]，其核心在于设计了一种特别的标记方案，将实体和关系提取转化为单个序列标注问题。

表 4 技术术语识别一般流程的现有问题及优化途径

Table4 Existing Problems and Optimization Approaches of General Process of TTR

流程	问题	现有优化途径	改进方向
文本数据获取	数据时滞性、数据规范、数据孤岛与关联	网络关联数据、实时数据库、爬虫获取	实时数据获取
数据预处理	去除噪音	人工去除不包含术语的噪音数据	自动化去除噪音；词汇、句子、篇章不同粒度的研究单元
训练数据处理	标记的质量 标记工作量巨大	计算不同样本标记一致性、预训练模型；迁移学习、半监督迭代学习	更多的预训练模型；迁移学习的有效性、迭代学习模型的优化验证
训练模型	如何封装、易懂	python 第三方库	如何开发出更简便的开源框架
评估和改良模型	调整参数复杂耗时；如何更精确地模型评价，相对优势不清楚、不具体	精确评估、宽松评估等不同评估标准	新的评估方法，具体到模型内部的评估指标
技术术语结果管理	更具体、更广泛的下游研究路径	实体识别+关系抽取的联合模型	细粒度、定制化的实体识别

4 技术术语识别的下游应用

实体抽取和关系抽取是众多研究任务的基础，而技术术语识别是实体抽取的一类重要内容。技术术语识别可能潜在的的下游研究方向及具体应用，如图 3 所示，主要包括三大类：情报决策方面、知识组织方面和自然语言处理方面的应用。对下游应用场景的探索和扩展，有助于指导技术术语识别方法的开发、迭代和优化。

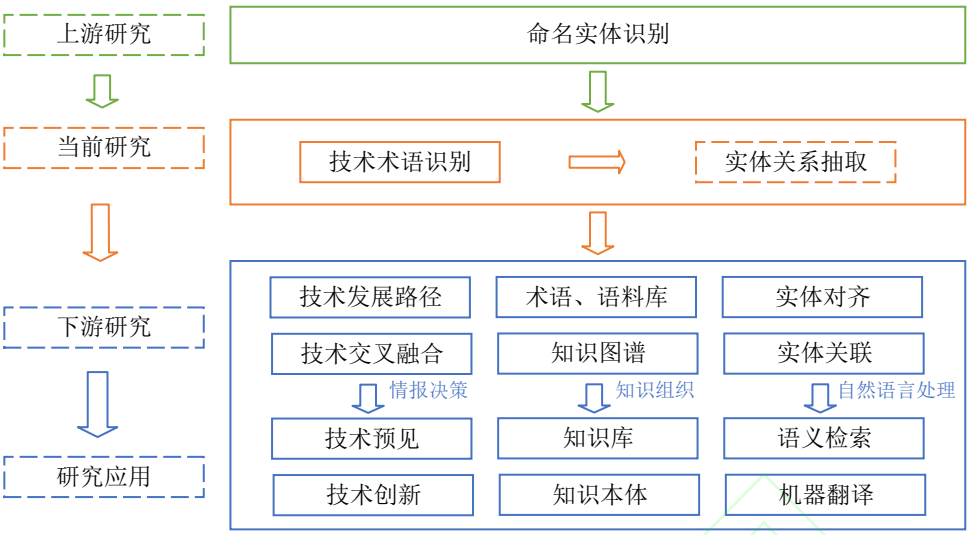

图 3 技术术语识别的后续研究方向和应用

Fig.3 The Follow-up Research Interests and Applications of TTR

（1）情报决策。技术预见是情报决策、科技创新中的重要实践活动。其中，技术的演化路径研究和技术交叉融合等研究是重要的研究课题。技术发展路径研究，是从技术系统的角度，对技术系统中技术构成的演化关系的研究[58]。因此，针对更细粒度的技术术语的技术预测，理论上更有针对性、效率更高，且能从深层次的技术术语的关系层次，整体、系统地表征各种类别技术系统随着时间的演化，从而进行更系统、更贴近实际的网络链路预测。

（2）知识组织。常见的知识表示和知识组织形式有本体、语料库[55,59]、知识库、语义网络和知识图谱[60]等，其基本的组成单元都是节点，节点即实体。本文所述的技术术语别后的术语集合，即可以用来构建专业的领域词典、语料库，从而构建本体，绘制领域知识图谱。

（3）自然语言处理。自然语言处理相关的诸多应用的基础都涉及到术语识别，如语义检索[61]、问答系统[43]和机器翻译[62]等。机器通过对专业术语的识别，依据术语的语义及其相似性，可以提高检索的查准率和查全率、提高问答系统的精确度，提高翻译的准确度和专业性等，进而提升机器对人类自然语言的理解准确性和处理速度。

5 结论与讨论

本文针对广义的机器学习算法在技术术语识别中的应用进展、一般流程和现存问题等方面进行了述评。从应用的算法分类入手，梳理各种算法的应用现状和优缺点，并归纳机器学习应用在技术术语识别中的一般流程、现存问题及优化途径。同时，总结了下游研究的应用前景。由研究结果可得出以下结论和展望。

（1）主流模型的迁移学习。机器学习算法在技术术语识别的应用可分为：单一的统计机器学习、单一深度学习和两者结合的混合算法。应用最广泛的是两者结合的混合方法，主流的模型代表是 BiLSTM-CRF 模型，迁移学习是未来重要的研究方向。同时，当前的应用流程中仍然有诸多待优化研究的问题，未来应继续加强细粒度的实体识别、丰富特征的表示方法等重点研究工作。当前研究不应固化于现有模型的调整、特征挑选和扩大语料库等方面，应将研究重点倾向于迁移学习方法，让技术术语识别方法得到更全面的发展。

（2）研究领域的扩展。NER 任务日益受到人们重视，但针对技术术语的识

别仍然较少，通过文献调研发现，相关研究主要分布在生物医学、军事科技领域，需要进一步扩展研究领域，探索能快速了解领域技术的研究途径。借助当前大数据的背景下，应考虑如何利用技术术语识别和知识发现技术，进行技术预见，作战略性、预测性研究。

（3）下游研究的集成。技术实体识别研究仅仅识别了文本中的实体对象，这和下游的关系抽取任务是分离的，即忽视了实体和关系之间的关联，也会影响实体抽取的质量。实体识别和关系一体化研究，有助于提升实体和关系抽取的质量，有待研究学者进行深入研究。结合 RC 研究，兼并实体识别和关系抽取的方法是未来的研究发展趋势。

（4）AI 开源工具的简便性。对于非计算机算法领域的研究人员，掌握各种复杂的机器学习算法相对困难，这是一直以来面临的问题。未来应考虑提高模型的普适性和简便性，利用当前深度学习框架（Tensorflow、Pytorch）开发简易操作的 TTR 开源工具包，让研究者使用工具包可以直接进行数据处理、输入表示、选择合适的模型并训练模型，让更多的研究工作者能够借助智能算法推动科学创新。

参考文献：

[1] 孙镇, 王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术, 2010 (6): 42-47.(Sun Zhen, Wang Huilin. Overview on the Advance of the Research on Named Entity Recognition[J]. New Technology of Library and Information Service, 2010 (6): 42-47.)

[2] Zadeh B Q, Handschuh S. Evaluation of Technology Term Recognition with Random Indexing[C]. In: Proceedings of LREC 2014 - Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland. Paris, France: ELRA, 2014: 4027-4032.

[3] 刘建华, 张智雄, 徐健, 等. 自动术语识别—对科技文献进行文本挖掘的重要技术方法[J]. 现代图书情报技术, 2008 (8): 12-17.(Liu Jianhua, Zhang Zhixiong, Xu Jian, et al. Automatic Term Recognitions—An Important Method for Text Mining on Scientific Literature[J]. New Technology of Library and Information Service, 2008 (8): 12-17.)

[4] Mima H, Ananiadou S, Nenadić G. The ATRACT Workbench: Automatic Term Recognition and Clustering for Terms[C]. In: Proceedings of the 4th International Conference on Text, Speech and Dialogue. Berlin, Germany: Springer-Verlag, 2001: 126-133.

[5] Linguistic Data Consortium. Entity Detection and Tracking:Phase 1—ACE Pilot Study Task Detection[EB/OL]. [2021-03-10]. https://www.ldc.upenn.edu/collaborations/past-projects/ace.

[6] Lan Y, Xu H G, Xu K, et al. Research on Named Entity Recognition for Science and Technology Terms in Chinese Based on Dependent Entity Word Vector[C]. In: Proceedings of the 14th International Conference on Anti-Counterfeiting, Security, and Identification. New York, USA: IEEE, 2020: 25-30.

[7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3 (4-5): 993-1022.

[8] 宋欣娜, 郭颖, 席笑文. 基于专利文献的多指标新兴技术识别研究[J]. 情报杂志, 2020, 39 (6): 76-81+88.(Song Xinna, Guo Ying, Xi Xiaowen. Research on Multi-Indicator Emerging Technology Identification Based on Patent Literature[J]. Journal of Intelligence, 2020, 39 (6): 76-81+88.)

[9] 王凌燕, 方曙, 季培培. 利用专利文献识别新兴技术主题的技术框架研究[J]. 图书情报工作, 2011, 55 (18): 74-78+23.(Wang Lingyan, Fang Shu, Ji Peipei. Using Patent Documents to Study the Technology Framework of Detecting Emerging Technology Topics[J]. Library and Information Service, 2011, 55 (18): 74-78+23.)

[10] 潘东华, 徐珂珂. 基于专利文献分类码的技术知识图谱绘制方法研究[J]. 情报学报, 2015, 34 (8): 866-874.(Pan Donghua, Xu Keke. Study on the Method of Mapping Technology Networks Based on Patent

Classification Codes[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34 (8): 866-874.)

[11] 刘忠宝, 康嘉琦, 张静. 基于主题突变检测的颠覆性技术识别——以无人机技术领域为例[J]. 科技导报, 2020, 38 (20): 97-105.(Liu Zhongbao, Kang Jiaqi, Zhang Jing. The disruptive technology of recognition based on topic mutation detection: With the drone technology as an example[J]. Science & Technology Review, 2020, 38 (20): 97-105.)

[12] 王海龙, 和法清, 丁堃. 基于社会网络分析的专利基础技术识别——以半导体产业为例[J]. 情报杂志, 2017, 36 (4): 78-84.(Wang Hailong, He Faqing, Dingkun. An Identifying Method of Industrial Essential Technologies Based on Social Network Analysis: Semiconductor Industry as a Case[J]. Journal of Intelligence, 2017, 36 (4): 78-84.)

[13] 吴颖文, 纪杨建, 顾新建. 基于专利技术共现网络的共性技术识别——以家电行业为例[J]. 情报探索, 2020 (3): 1-10.(Wu Yingwen, Ji Yangjian, Gu Xinjian. Generic Technology Identification Based on Technology Co-occurrence Network of Patents: Case Study of Household Appliance Industry[J]. Information Research, 2020 (3): 1-10.)

[14] 许海云, 王振蒙, 胡正银, 等. 利用专利文本分析识别技术主题的关键技术研究综述[J]. 情报理论与实践, 2016, 39 (11): 131-137.(Xu Haiyun, Wang Zhenmeng, Hu Zheng yin, et al. Review on Key Techniques of Technical Theme Identification Using Patent Text Analysis[J]. Information Studies:Theory & Application, 2016, 39 (11): 131-137.)

[15] 谷俊. 专利文献中新技术术语识别研究[J]. 现代图书情报技术, 2012 (11): 53-59. (Gu Jun. Study on New Technology Detection in Patents Documents[J]. New Technology of Library and Information Service, 2012 (11): 53-59.)

[16] Chang J S. Domain Specific Word Extraction from Hierarchical Web Documents: a First step Toward Building Lexicon Trees from Web Corpora[C]. In: Proceedings of the 4th SIGHAN Workshop on Chinese Language Learning, Jeju Island, Korea. Stroudsburg, USA: ACL, 2005: 64-71.

[17] 陈颖, 张晓林. 专利中技术词和功效词识别方法研究[J]. 现代图书情报技术, 2011 (12): 24-30.(Chen Ying, Zhang Xiaolin. Study on the Differentiating Method of Technical and Effect Words in Patent[J]. New Technology of Library and Information Service, 2011 (12): 24-30.)

[18] 曹国忠, 杨雯丹, 刘新星. 基于主体-行为-客体(SAO)三元结构的专利分析方法研究综述[J]. 科技管理研究, 2021, 41 (4): 158-167.(Cao Guozhong, Yang Wendan, Liu Xinxing. Review of Patent Analysis Methods Based on Subject-Action-Object Ternary Structure[J]. Science and Technology Management Research, 2021, 41 (4): 158-167.)

[19] 邱科达, 马建玲. 机器学习在术语抽取研究中的文献计量分析[J]. 图书情报工作, 2020, 64 (14): 94-103. (Qiu Keda, Ma Jianling. A Statistical Analysis of Literature on Term Extraction Based on Machine Learning[J]. Library and Information Service, 2020, 64 (14): 94-103.)

[20] Zhou G, Jian S. Named Entity Recognition Using an HMM-based Chunk Tagger[C]. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA. Stroudsburg, USA: ACL, 2002: 473-480.

[21] 岑咏华, 韩哲, 季培培. 基于隐马尔科夫模型的中文术语识别研究[J]. 现代图书情报技术, 2008 (12): 54-58.(Cen Yonghua, Han Zhe, Ji Peipei. Chinese Term Recognition Based on Hidden Markov Model[J]. New Technology of Library and Information Service, 2008 (12): 54-58.)

[22] Doan S, Hua X. Recognizing Medication Related Entities in Hospital Discharge Summaries Using Support Vector Machine[C]. In: Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China. Stroudsburg, USA: ACL, 2010: 259-266.

[23] Takeuchi K. Use of Support Vector Machines in Extended Named Entity Recognition[C]. In: Proceedings of

the 6th Conference on Natural Language Learning. Stroudsburg, USA: ACL, 2002: 1-7.
[24] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. In: Proceedings of the Nineteenth International Conference on Machine Learning, Sydney, Australia. San Francisco, USA: Morgan Kaufmann, 2002.
[25] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]. In: Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. Edmonton, Canada. Stroudsburg, USA: ACL, 2003: 188-191.
[26] McDonald R, Pereira F. Identifying gene and protein mentions in text using conditional random fields[J]. Bmc Bioinformatics, 2005, 6: 7.
[27] 黄菡, 王宏宇, 王晓光. 结合主动学习的条件随机场模型用于法律术语的自动识别[J]. 数据分析与知识发现, 2019, 3 (6): 66-74.(Huang Han, Wang Hongyu, Wang Xiaoguang. Automatic Recognizing Legal Terminologies with Active Learning and Conditional Random Field Model[J]. Data Analysis and Knowledge Discovery[J], 2019, 3 (6): 66-74.)
[28] Sahu S K, Anand A. Recurrent neural network models for disease name recognition using domain invariant features[C]. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany. Stroudsburg, USA: ACL, 2016: 2216–2225.
[29] 李明浩, 刘忠, 姚远哲. 基于 LSTM-CRF 的中医医案症状术语识别[J]. 计算机应用, 2018, 38 (S2): 42-46.(Li Minghao, Liu Zhong, Yao Yuanzhe. LSTM-CRF Based Symptom Term Recognition on Traditional Chinese Medical Case[J]. Journal of Computer Applications, 2018, 38 (S2): 42-46.)
[30] 刘宇飞, 尹力, 张凯, 等. 基于深度迁移学习的技术术语识别——以数控系统领域为例[J]. 情报杂志, 2019, 38 (10): 168-175.(Liu Yufei, Yin Li，Zhang Kai, et al. Deep Transfer Learning for Technical Term Extraction —A Case Study in Computer Numerical Control System[J]. Journal of Intelligence, 2019, 38 (10): 168-175.)
[31] 曹依依. 基于命名实体识别的医学术语发现及应用[D]. 重庆: 重庆邮电大学, 2019.(Cao Yiyi. Medical Terminology Discovery and Application Based on Named Entity Recognition[D]. Chongqing: Chongqing University of Posts and Telecommunications, 2019.)
[32] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[C]. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California. Stroudsburg: ACL, 2016: 260-270.
[33] 袁慧. 基于 Bi-LSTM 与 CRF 的命名实体识别研究——以生态治理技术相关实体为例[D]: 中国科学院兰州文献情报中心 2017.(Yuan Hui. Bi-LSTM+CRF-based Named Entity Recognition——Taking Ecological Management Technology as an Example [D]. Lanzhou: Lanzhou Information Center, Chinese Academy of Sciences, 2017.)
[34] 王昊, 邓三鸿, 苏新宁, 等. 基于深度学习的情报学理论及方法术语识别研究[J]. 情报学报, 2020, 39 (8): 817-828.(Wang Hao, Deng Sanhong, Su Xinning, et al. A Study on Chinese Terminology Recognition of Theory and Method from Information Science: Based on Deep Learning[J]. Journal of the China Society for Scientific and Technical Information, 2020, 39 (8): 817-828.)
[35] 王学锋, 杨若鹏, 朱巍. 基于深度学习的军事命名实体识别方法[J]. 装甲兵工程学院学报, 2018, 32 (4): 94-98.(Wang Xuefeng, Yang Ruopeng, Zhu Wei. Military Named Entity Recognition Method Based on Deep Learning[J]. Journal of Academy of Armored Force Engineering, 2018, 32 (4): 94-98.)
[36] 冯鸾鸾, 李军辉, 李培峰, 等. 面向国防科技领域的技术和术语识别方法研究[J]. 计算机科学, 2019, 46 (12): 231-236.(Feng Luanluan, Li Junhui, Li Peifeng, et al. Technology and Terminology Detection Oriented National Defense Science[J]. Computer Science, 2019, 46 (12): 231-236.)
[37] Li P H, Dong R P, Wang Y S, et al. Leveraging Linguistic Structures for Named Entity Recognition with Bidirectional Recursive Neural Networks[C]. In: Proceedings of the 2017 Conference on Empirical Methods in

Natural Language Processing, Copenhagen, Denmark. Stroudsburg, USA: ACL, 2017: 2664–2669.
[38] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C]. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany. Stroudsburg, USA: ACL, 2016: 1064–1074.
[39] Strubell E, Verga P, Belanger D, et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions[C]. In : Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, -Copenhagen, Denmark. Stroudsburg, USA: ACL, 2017: 2670–2680.
[40] 蒋翔, 马建霞, 袁慧. 基于 BiLSTM-IDCNN-CRF 模型的生态治理技术领域命名实体识别[J]. 计算机应用与软件, 2021, 38 (3): 134-141.(Jiang Xiang, Ma Jianxia, Yuan Hui. Named Entity Recognition in The Field of Ecological Management Technology Based on BILSTM-IDCNN-CRF Model[J]. Computer Applications and Software, 2021, 38 (3): 134-141.)
[41] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[C]. In: Proceedings of the 31st Conference on Neural Information Processing Systems(NIPS 2017), Long Beach, USA. California, USA: NIPS, 2017: 5998-6008.
[42] 马千程, 王崑声, 周晓纪. 基于深度学习的竞争情报命名实体识别研究[J]. 情报探索, 2020 (9): 1-7.(Ma Qiancheng, Wang Kunsheng, Zhou Xiaoji. Named Entity Recognition of Competitive Intelligence Based on Deep Learning[J]. Information Research, 2020 (9): 1-7.)
[43] 赵鹏飞, 赵春江, 吴华瑞, 等. 基于注意力机制的农业文本命名实体识别[J]. 农业机械学报, 2021, 52 (1): 185-192.(Zhao Pengfei, Zhao Chunjiang, Wu Huarui, et al. Research on Named Entity Recognition of Chinese Agricultural Based on Attention Mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52 (1): 185-192.)
[44] Ruder S, Peters M E, Swayamdipta S, et al. Transfer Learning in Natural Language Processing[C]. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, USA. Stroudsburg, USA: ACL, 2019: 15-18.
[45] 迟玉琢. 大数据背景下的情报分析[J]. 情报杂志, 2015, 34 (1): 18-22.(Chi Yuzhuo. Intelligence Analysis under Big Data Background[J]. Journal of Intelligence, 2015, 34 (1): 18-22.)
[46] 毛明毅, 吴晨, 钟义信, 等. 加入自注意力机制的 BERT 命名实体识别模型[J]. 智能系统学报, 2020, 15 (4): 772-779.(Mao Mingyi,Wu Chen, Zhong Yixin, et al. BERT Named Entity Recognition Model with Self-attention Mechanism[J]. CAAI Transactions on Intelligent Systems, 2020, 15 (4): 772-779.)
[47] Lin Y, Hong L, Yi L, et al. Biomedical Named Entity Recognition based on Deep Neutral Network[J]. International Journal of Hybrid Information Technology, 2015, 8 (8): 279-288.
[48] Hripcsak G, Rothschild A S. Agreement, the F-Measure, and Reliability in Information Retrieval[J]. Journal of the American Medical Informatics Association, 2005, 12 (3): 296-298.
[49] Zeng Q, Yu M, Yu W, et al. Validating Label Consistency in NER Data Annotation[EB/OL]. [2021-04-23]. https://arxiv.org/abs/2101.08698.
[50] 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.)
[51] Li Z, Ko B, Choi H-J. Naive Semi-supervised Deep Learning Using Pseudo-label[J]. Peer-to-Peer Networking and Applications, 2019, 12: 1-11.
[52] 高佳奕, 杨涛, 董海艳, 等. 基于 LSTM-CRF 的中医医案症状命名实体抽取研究[J]. 中国中医药信息杂志, 2021, 28 (5): 20-24.(Gao Jiayi, Yang Tao, Dong Haiyan, et al. Study on Named Entity Extraction of TCM Clinical Medical Records Symptoms Based on LSTM-CRF[J]. Chinese Journal of Information on TCM, 2021, 28 (5): 20-24.)

[53] Chiu J, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs[J]. Computer Science, 2015

[54] 周浪. 中文术语抽取若干问题研究[D]. 南京: 南京理工大学, 2010.(Zhou Lang. A study on the Chinese Term Extraction[D]. Nanjing: Nanjing University of Science & Technology, 2010.)

[55] 曾文, 李智杰, 王小玉, 等. 科技政策术语自动识别技术初探[J]. 中国科技资源导刊, 2017, 49 (3): 20-25.(Zeng Wen, Li Zhijie, Wang Xiaoyu, et al. Research on Automatic Recognition Technology of Science and Technology Policy Term[J]. China Science & Technology Resources Review, 2017, 49 (3): 20-25.)

[56] Fu J, Liu P, Neubig G. Interpretable Multi-dataset Evaluation for Named Entity Recognition[C]. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online. Stroudsburg, USA: ACL, 2020: 6058-6069.

[57] Zheng S, Wang F, Bao H, et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme[C]. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada. Stroudsburg, USA: ACL, 2017: 1227-1236.

[58] 李晓曼. 基于专利要素特征的技术演化分析[D]. 北京: 中国农业科学院, 2020.(Li Xiaoman. Technology Evolution Analysis Based on Patent Elements Features[D]. Beijing: Chinese Academy of Agricultural Sciences, 2020.)

[59] 冯鸾鸾, 李军辉, 李培峰, 等. 面向国防科技领域的技术和术语语料库构建方法[J]. 中文信息学报, 2020, 34 (8): 41-50.(Feng Luanluan, Li Junhui, Li Peifeng, et al. Constructing a Technology and Terminology Corpus Oriented National Defense Science[J]. Journal of Chinese Information Processing, 2020, 34 (8): 41-50.)

[60] 杨品莉, 谢志长. 基于 BiLSTM-CRF 的司法领域实体识别研究[J]. 现代计算机, 2020 (25): 3-8.(Yang Pinli, Xie Zhichang. Research on Named Entity Recognition in Legal Documents Based on BiLSTM-CRF[J]. Modern Computer, 2020 (25): 3-8.)

[61] Li J, Sun A, Han J, et al. A Survey on Deep Learning for Named Entity Recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, PP (99): 1-1.

[62] Babych B, Hartley A. Improving Machine Translation Quality with Automatic Named Entity Recognition[C]. In: Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Budapest, Hungary. Stroudsburg, USA: ACL, 2003: 1-8.

作者贡献说明：

胡雅敏：负责文献分析、论文撰写；

吴晓燕：负责论文修改与完善；

陈　方：提出研究选题，负责论文完善与最终版本修订。

利益冲突说明：

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：huyamin19@mails.ucas.ac.cn

[1]胡雅敏,陈方.Reference.rar.参考文献.