doi:10.3772/j.issn.2095-915x.2018.05.006

基于条件随机场和篇章校对的有机物命名实体识别方法研究

谷威[1] 田欣[2]

1. 国家知识产权局专利局 北京 100088；
2. 中国专利信息中心 北京 100088

摘 要 有机物命名实体识别是生物医学等专利文本挖掘和机器翻译的关键步骤，只有正确地识别出有机物命名实体，才能准确、有效地完成专利挖掘和翻译。本文从有机物命名实体识别的自身构成特点出发，重点研究了有机物命名实体识别的流程、方法和特征，采用 CRF 算法和篇章校对结合的方法实现了有机物命名实体的自动识别，达到了较高的准确率和召回率。下一步的研究中将利用模板和 CRF 等多策略识别方法继续改进。

关键词： 有机物识别；命名实体；条件随机场；篇章校对

中图分类号： G35

开放科学（资源服务）标识码（OSID）

Research on Organic Matter Named Entity Identification Method Based on Conditional Random Field and Text Proofreading

GU Wei[1] TIAN Xin[2]

1. Patent Office of the State Intellectual Property Office, Beijing 100088, China;
2. China Patent Information Center, Beijing 100088, China

Abstract Organic matter named entity identification is the critical step of patent text mining (especially bio-medicine text) and machine translation, and only when the organic matter named entity is correctly identified, thepatent mining and machine translation can be effectively completed. From the aspect of self-

作者简介： 谷威（1978–），硕士，研究方向：专利信息化建设和信息服务，专利信息化发展规划制定、信息系统建设和管理等。田欣，研究方向：中国专利文献中英文数据处理，专利文献机器翻译的应用与研究。

structure features of organic matter named entity identification, this paper focuses on the research of the procedures, methods and features of organic matter named entity identification, and adopts the combined method of CRF and text proofreading to realize the automatic identification of organic matter named entities and achieve a rather high precision rate and recalling rate.

Keywords：Organic matter identification; named entity; conditional random field; text proofreading

1 引言

命名实体识别是文本挖掘的一项关键技术。它是实现信息抽取的第一步，同时也是信息检索、机器翻译、组块分析、问答系统等自然语言处理技术的重要基础[1]。命名实体通常包括人名、机构名、时间表达式和数字表达等[2]。在专利文献中，化学领域命名实体的识别还包括基因、蛋白质、疾病、化学方程式和有机物等。近几年，我国专利申请量逐年递增，年均增长率达到18.56%，2016年发明专利申请量达到133.9万件[3]。而以有机物结构、制备或用途为发明主题的专利，是化学类专利的重要组成部分。根据统计，包含有机物名称的化学类专利文献约占全部化学类专利文献的三分之一。这些有机物命名实体的识别对专利文献检索、分析具有重要的意义。而由于有机物命名方法的多样性、以及有机物名称组合的可变性，无法用常规收录的方法构建有机物名称词典。大量未登录词严重影响了有机物命名实体识别的精确率和召回率。因此，针对有机物命名实体识别的研究具有现实意义。

目前，关于中文命名实体识别，许多学者进行了有益的探索和尝试，并取得了一定的研究成果。研究中采用的方法可以归纳为基于词典的方法、基于规则的方法以及基于统计学习的方法。

①基于词典的方法。通过从文本中搜索出与给定词典中命名实体相同或者最相似的字符串来实现生物命名实体的识别。由于命名实体识别存在大量的变体名，使用字符串完全匹配的算法会导致极低的召回率。因此，基于词典的生物命名实体识别一般采用的都是字符串近似匹配算法，Yang[5]等使用的改进编辑距离算法就是字符串近似匹配算法的一种。

②基于规则的方法。通过分析命名实体的内部特征和外部特征，人工地或启发式地产生规则模板，来实现生物命名实体的识别。张冬梅等[2]利用模式匹配对人名、机构名等命名实体进行了识别，Fukuda[6]等提出的PROPER方法利用启发式规则进行命名实体识别。Olsson等[7]采用词法分析和句法分析技术对生物医学文献中的命名实体进行识别。

③基于统计学习的方法。通过从数据样本集合中统计出相关特征和参数，建立识别模型，来完成命名实体识别任务。周荣鹏[1]利用改进SVM算法对生物医药命名实体进行了识别，向晓雯[4]利用SVM进行命名实体识别，Wang等[8]综合了SVM、CRF等统计学习方法进行命名实体识别。

doi:10.3772/j.issn.2095-915x.2018.05.006

通过总结以往的研究，可以发现还存在以下几个方面的问题。

①现有化学领域命名实体识别研究基本是以基因、蛋白质、疾病等为主，有机物命名实体识别较少。

②现有命名实体识别一般都是针对期刊和论文的，没有针对专利文献的化学有机物命名实体研究。

③现有研究方法一般都集中在算法方面，没有针对科技文献的全文进行考虑，造成现有研究办法召回率低、一致性差，同一有机物被多次地分析、翻译，耗时费力。

针对目前的研究，鉴于没有充分把统计信息和全文信息有机结合起来的问题，本文拟从专利文献中的真实语料出发，通过对标注好的序列进行训练，通过篇章校对的方法实现对专利文献中的有机物自动识别。

本文以下内容中，首先介绍如何利用 CRF 算法实现对有机物的识别，接下来介绍关于篇章校对的方法，然后介绍试验结果及结果的分析，最后是对本次研究的总结。

2 基于 CRF 的有机物命名实体识别方法

2.1 基于规则的标引引擎研究

笔者前期在中国专利标引中即采用了基于规则的有机物标引方法搭建了通过已知的有机物词，形成一套规则，在下次再出现相同的词的时候就可以将这些有机物识别出来，这个方法叫做基于规则的有机物识别方法，目前基于规则的有机物识别方法准确率和召回率较低，急需一种方法进行有效识别。有机命名实体的片段分为以下七种类型：

WA[A] 基团

WA[B] 主链结构

WA[C] 数量

WA[D] 化学元素

WA[E] 位置

WA[F] 种类

WA[O] 其它

并且分别根据六种类型进行识别规则。制定此类规则的规则是：

（-1）{WA[A]}+（0）{CHN[，]}+（1）{WA[A]}=>TREE（-1，1）+PUT（fp，WA，B）

使用上述语义属性和识别规则来定义识别过程和算法。具体程序如下：

输入：中文句子

输出：复合识别结果

识别过程：

（1）中文单词的分割，复合属性的注释。

（2）逐句分析，看句子活动规律。 如果是这样，那么句子会转到第 2 步，否则会转到步骤 5。

（3）根据规定进行鉴定。 如果句子符合规定，则将该词合并起来，将新词归为一个复合词。

（4）当谈到最后的单词时，没有组合新的组合，然后转到步骤 5. 否则将转到步骤 2。

（5）辅助词的识别。

（6）结果输出。

识别规则识别“1，4-二取代-1，2，3-三氮唑类化合物”等有机命名实体的过程如下图1所示：

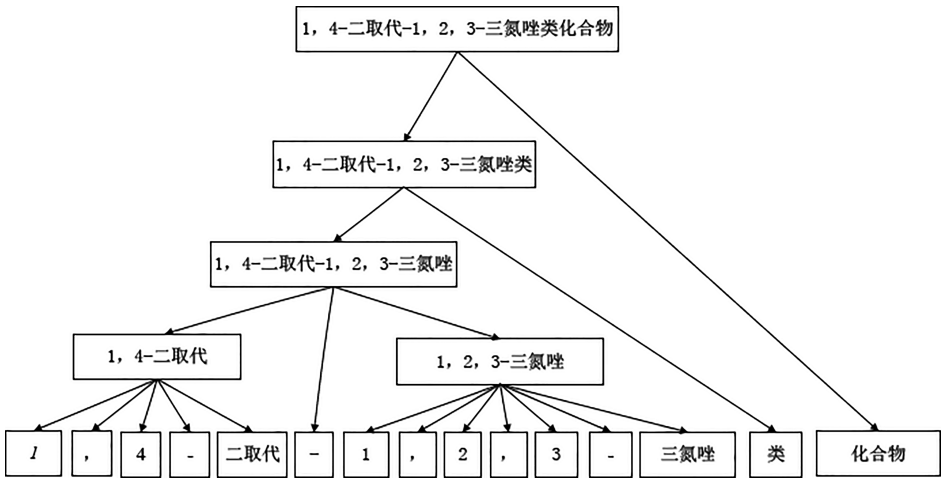

图 1 有机物命名实体识别树形结构图

2.2 基于CRF的有机物命名实体识别流程

根据有机物命名实体识别流程，如图 2 所示，把识别的流程分成命名实体模型训练和有机物命名实体识别，结合实际业务工作，分别对两部分内容进行说明。

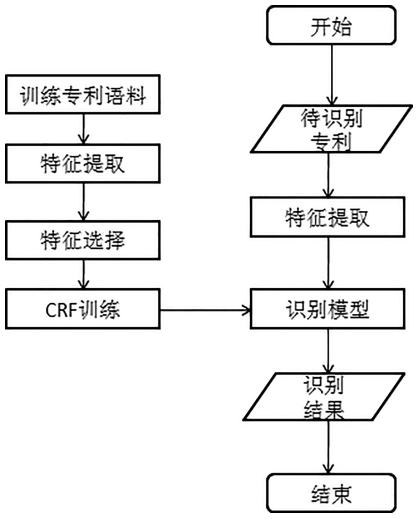

图 2 有机物命名实体识别流程图

①命名实体模型训练。首先从化工领域的相关专利文献中找出一些具有代表性的专利文献，按照预先要求的特征进行人工特征标注，包括对命名实体使用 NEOC 标签标注等。接下来将标注好的专利文献放入训练模型中进行训练，并按照要求不断调试特性，最终得到想要的特征模型。

②有机物命名实体识别。根据训练出的特征，将具备该特征的模型放入到实体识别的流程中，将待翻译文本文件输入系统后，经过系统的特征提取和识别模型的识别，最终得到识别结果。

2.3 基于CRF的有机物命名实体识别训练

2.3.1 训练流程

上节图 1 训练模型中所输入的训练专利语料，是在大量专利文献中搜寻出的与化工领域相关的专利文献，在这些专利文献中进行分词等操作，通过特征提取、特征添加将设定的特征进行人工标注，形成专利训练语料。特征提取模块的工作原理如下图 3 所示。

2.3.2 语料标注

在有机物命名实体的识别中，将标注的方法与训练的模型实现更好的融合，加入了一些规则，下面通过一个实例来说明其工作原理。如表 1 所示，是已经标注好的原始语料的格式。

doi:10.3772/j.issn.2095-915x.2018.05.006

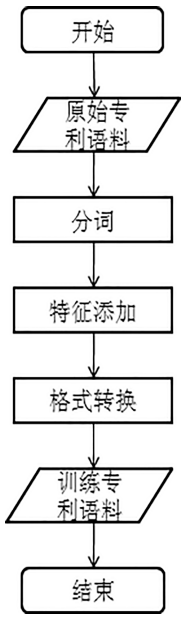

图 3 特征提取模块工作原理

表 1 原始语料标注格式表

【申请号】 201410255983.3
【发明名称】 5 - 氨基 - 4 - 氨甲酰咪唑核苷碳环类似物的合成方法
【摘要】 本发明公开了一种 <NEOC> 5 – 氨基 – 4 – 氨甲酰咪唑核苷碳环 </NEOC> 类似物的合成方法，以 <NEOC>（3aR，4S，6R，6aS）– 甲基 – 6 – 氨基 – 2，2 – 二甲基四氢 – 3a 氢 – 环戊［d］［1，3］二氧 – 4 – 甲酸酯 </NEOC>，原 <NEOC> 甲酸三乙酯 </NEOC> 和 <NEOC> 2 – 氨基 – 2 – 氰基乙酰胺 </NEOC> 为原料，回流反应合环得到中间体，再用硼氢化锂还原，与三氯氧磷反应后水解即可得到。通过上述方式，本发明的 <NEOC> 5 – 氨基 – 4 – 氨甲酰咪唑核苷碳环 </NEOC> 类似物的合成方法，反应步骤较少，反应条件温和，精制方便，操作简单，反应质量高，收率高，环保效果好。

2.3.3 文本预处理

运用规则合并法对命名实体词素进行合并。规则合并法是利用规则对相邻的有机物命名实体词素进行合并，最终得到完整有机物命名实体名称的过程。使用的规则可以为人工撰写的规则，也可以是利用机器学习预先从人工标注的有机物名称中学习而生成的规则。机器学习自动生成规则的方法可以为决策树方法、遗传算法、基于转换的错误驱动学习方法、SVM 法、神经网络算法、线性判别方法、K- 近邻算法等。以下以 TBL（transformation-based learning）为例说明自动生成规则的方法，但本文研究不限于这种方法。经过一次或一次以上规则合并后组成的字串临时标注为 CP。以为几个规则的示例。

{ 天干 }+{ 酮、醛、酯…}=>{CP}
{ 中文数字 }+{ 基团名 }=>{CP}
{ 天干 }+{ 基 }=>{CP}
{ 数字结构 }+{CP}=>{CP}
{ 正、异、新、伯、仲、叔、季 }+{CP}=>{CP}
{ 环 }+{CP}=>{CP}
{CP}+{ 酮、醛、酯…}=>{CP}
{ 氟、氯、溴、碘 }+{ 代 }=>{CP}

具体的规则合并方法为从上一步标注的候选名称种子出发，向前后扩展一个词（根据初次分词结果分出的词），分别进行规则匹配，匹配时如果符合多条规则，则优先匹配排名靠前的规则，重复该扩展过程直到所有规则均不匹配，则有机物命名实体识别完成，标注 NEOC 标签标明起始位置。

2.3.4 特征添加

通过预设置的特征，经过相应算法实现特征提取后，产生了如表 2 内容的词性标注表。其中，词性列标注出了每个词的词性（如 rz 指示代词、ule 助词等类型），位置列标注：O 表示独立，B 表示开始，I 表示中间，该列则说明了其所在的位置，即功能方面的信息，其中有机物特征信息如下所示：

①数字结构，例：10-、-2-、-2，3-

②中文数字，例：一、二

③天干地支，例：甲、乙、丙、丁

④希腊字母结构，例：α、β

⑤常见举例如下：化学常见词素：代、聚、缩、并、杂、亚、过、偏、次、酯、烯、醛

表 2 词性标注示例表

中文词	词性	位置
本	rz	O
发明	n	O
公开	v	O
了	ule	O
一	m	O
种	q	O
5 –	m	B
氨基	n	I
– 4 –	m	I
氨甲酰	n	I
咪唑	n	I
核苷碳环	n	I
类似物	n	I
的	ude1	O
合成	vn	O
方法	n	O

在特征添加的基础上，又加入有机物命名实体特征的标注，将其列举在表格的中间列，即有机物特征信息列，此处，O 是非有机物特征，B 是中文数字，A 是数字结构，F 是种类烷烃、烯烃、炔烃、烯炔、脂环烃，G 是其他化合物、类似物和有机物等。命名实体识别的特征信息标注示例如表 3 所示。

表 3 命名实体特征标注

中文词	词性	有机物特征信息	位置
本	rz	O	O
发明	n	O	O
公开	v	O	O
了	ule	O	O
一	m	B	O
种	q	O	O
5 –	m	A	B
氨基	n	F	I
– 4 –	m	A	I
氨甲酰	n	F	I
咪唑	n	F	I
核苷碳环	n	F	I
类似物	n	G	I
的	ude1	O	O
合成	vn	O	O
方法	n	O	O

2.3.5 特征选择

完成有机物命名实体训练语料的标注，还需要对有机物命名实体的特征进行提取。

CRF 是目前性能最好的序列标注器，已被广泛应用于中文分词、命名实体识别、浅层语法分析等多个自然语言处理任务当中。然而，模型的训练时间长、系统开销大，并且其训练时间和系统开销会随着语料规模和特征数量的增加而增加。基于有机物命名实体机器翻译系统在进行特征选取的时候，通过对有机物命名实体的特征模板进行特征上的反复试验，采用优化的特征提取的工具是 CRF++。在 CRF++ 的特征模板文件中，每行代表一个识别特征。例如，U02：%x[0，0] 代表了所选定的某分词结果中的词，以该词作为矩阵基点，定位其他词的位置标注的特征模板，因该方法的使用借

doi:10.3772/j.issn.2095-915x.2018.05.006

助了CRF++工具，此处不展开详述。如表4所示。

表 4 有机物命名实体识别的特征模板

```
# Unigram
U00:%x[-2, 0]
U01:%x[-1, 0]
U02:%x[0, 0]
U03:%x[1, 0]
U04:%x[2, 0]
U05:%x[-1, 0]/%x[0, 0]
U06:%x[0, 0]/%x[1, 0]

U10:%x[-2, 1]
U11:%x[-1, 1]
U12:%x[0, 1]
U13:%x[1, 1]
U14:%x[2, 1]
U15:%x[-2, 1]/%x[-1, 1]
U16:%x[-1, 1]/%x[0, 1]
U17:%x[0, 1]/%x[1, 1]
U18:%x[1, 1]/%x[2, 1]

U20:%x[-2, 1]/%x[-1, 1]/%x[0, 1]
U21:%x[-1, 1]/%x[0, 1]/%x[1, 1]
U22:%x[0, 1]/%x[1, 1]/%x[2, 1]

U30:%x[-2, 2]
U31:%x[-1, 2]
U32:%x[0, 2]
U33:%x[1, 2]
U34:%x[2, 2]
U35:%x[-2, 2]/%x[-1, 2]
U36:%x[-1, 2]/%x[0, 2]
U37:%x[0, 2]/%x[1, 2]
U38:%x[1, 2]/%x[2, 2]

U40:%x[-2, 2]/%x[-1, 2]/%x[0, 2]
U41:%x[-1, 2]/%x[0, 2]/%x[1, 2]
U42:%x[0, 2]/%x[1, 2]/%x[2, 2]

U50:%x[0, 1]/%x[0, 2]

# Bigram
B
```

2.4 有机物命名实体识别

有机物命名实体识别模块，是利用有机物命名实体训练得到的模型对文件进行识别。待识别专利首先进行预处理，然后进行特征提取和选择，将提取的特征输入到 test_files 文件，然后将文件提交给 CRF_test 进行识别，最后将识别结果保存到 result_file。识别模块的流程如图 4 所示。

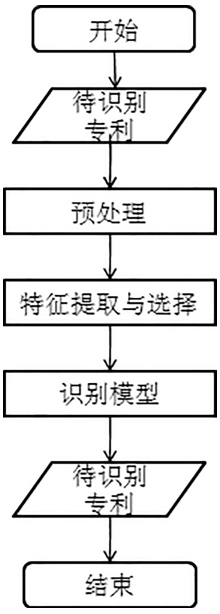

图 4 识别模块流程图

3 有机物篇章校对方法

由于现有方法中没有针对科技文献的全文进行考虑，造成现有研究办法召回率低，一致性差等问题。本节在 CRF 识别有机物的基础上，利用已经识别的有机物提出全篇有机物识别方法，在 CRF 有机物识别之后，将有机物命名实体存储在起来，通过有机物命名实体查找，找到未被识别的有机物命名实体，根据命名实体频率将只有一次的进行标记，输出最终得到有机物命名实体，如图 5 所示。

在表 5 所示专利摘要中，“2，2- 二溴 -2- 氰基乙酰胺”出现 3 次，如有一次不能识别，则把“2，2- 二溴 -2- 氰基乙酰胺”记忆在内存字典中，然后通过全文匹配，找到未识别的位置并对其进行标注，并将次数变为 3 次。

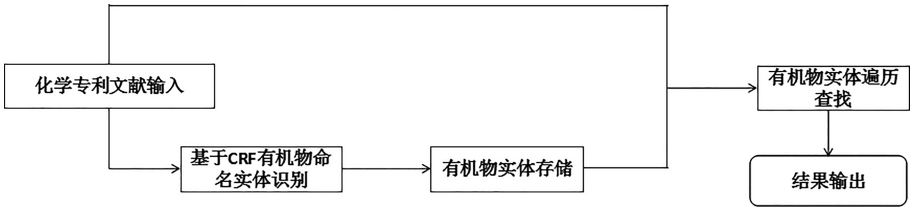

图 5 有机物篇章校对方法流程图

表 5 篇章校对信息表

本发明公开了一种 <NEOC>2，2- 二溴 -2- 氰基乙酰胺 </NEOC> 的合成方法。该方法以氰乙酰胺为原料，将氰乙酰胺和溶剂加入反应瓶中，在反应温度 15℃ ~45℃，搅拌下滴加液溴，当液溴的滴加量为其加入总量 1/3~4/5 时，开始滴加浓度为 30% 的双氧水，液溴和双氧水均滴加完后，继续反应 0.5h~6h 后，过滤得到 <NEOC>2，2- 二溴 -2- 氰基乙酰胺 </NEOC>；其中氰乙酰胺：溴：双氧水的摩尔比为 1 ： 1.1~1.3 ： 0.70~0.90；溶剂为水或滤液，其用量为每摩尔氰乙酰胺 252mL~420mL。本发明主要用于 <NEOC>2，2- 二溴 -2- 氰基乙酰胺 </NEOC> 的合成。

4 实验结果及分析

4.1 评测语料及评测指标

评测语料选取了 CPRS 检索引擎，在国际专利分类号（IPC）为 C07 和 C08 中含有有机物和组合物的专利文献 85847 篇，提取经过专利标引和数据深加工后结果中含有有机物的专利文献，采用专利说明书摘要作为训练语料、测试，为了测试篇章校对方法，特选择每篇标引有机物个数大于 2 作为评测语料，共计有 23241 篇，如表 6 所示。

表 6 有机物命名实体训练数据规模表

	所含摘要总数	所含句子总数	所含字数	所含有机物总数
训练语料	20000	143006	425486	68340
测试语料	3241	23199	1011205	11067

衡量一个系统翻译质量的高低，其中一个重要的因素是对命名实体识别的准确与否。而命名实体的识别，则关注在其范围的确定性和对类型的确定性两个方面。对于目前研究较多的人名、地名或组织名相关的命名实体识别来说，其在两方面的要求都会对最终的结果产生较大影响，而对本研究的方向来说，确定命名实体的边界范围和翻译的规范两个指标显得更为重要。结合各种特性，类似边界问题具有其共性。本文采用精确率、召回率和 F1 评测值对实验结果进行评测，具体公式如公式 1 至 3 所示。

精准率 =count（正确）/(count(正确)+count（虚假）) 公式（1）

召回率 =count（正确）/(count(正确)+count（丢失）) 公式（2）

F1=2* 精确率 * 召回率]/[精确率 + 召回率] 公式（3）

对公式 1 和 2 来说，其中“正确”表示识别的结果和标准结果相同；“丢失”表示没有识别出的结果而标准结果中有；“虚假”表示识别的有，但在标准结果中不存在。

4.2 评测结果及分析

评测分为训练和识别测试两个部分。

① 训练。对 CRF 模型进行训练，采用 CRF++ 训练模型训练过程中，根据标注语料和特征模板作为识别模型，CRF++ 根据特征模板进行参数计算，然后得出训练结果。CRF++ 的

doi:10.3772/j.issn.2095-915x.2018.05.006

训练命令行如下：

crf_learn template_file train_file model_file

识别测试。进行有机物命名实体识别，首先采用 CRF++ 进行识别，命令行如下：

% crf_test -m model_file test_files >> result_file

②利用已有的规则识别的办法进行验证；

③利用全文校对命令进行校对，实验结果表 7 所示。

表 7 实验结果

	有机物数量	识别有机物数量	识别正确数量	召回率	精准率	F1
规则识别	11067	12321	7635	68.99%	61.96%	65.28%
CRF 算法	11067	10609	9345	84.44%	88.08%	86.22%
篇章校对	11067	11631	10340	93.43%	88.89%	91.10%

由实验结果可以看出，有机物命名实体识别召回率为 93.43%，精准率为 88.89%，这表明已经可以在专利翻译和专利分析挖掘中进行使用。采用 CRF 算法和篇章校对结合的方法在召回率上有较大提高，在精准率上变化不大，分析原因在于篇章校对是对已经识别的有机物进行全篇召回，所以提高了召回率，而对没有识别的有机物无效。为了保证分类辞典准确，提高纠错的准确率，可以增加训练有机物句子的数量和结合模板等方法进行多策略识别。

5 结论

本文系统的研究了有机物命名实体的构成特点，总结了有机物命名实体的构成模式，利用 CRF 模型实现了对有机物命名实体的识别。在利用真实语料进行的测试中，该方法取得了较好的效果，得到了较好的准确率和召回率，这说明我们采用的基于 CRF 的有机物命名实体方法是一种有效的方法。下一步的研究中将利用模板和 CRF 等多策略识别方法继续改进。

参考文献

[1] 周荣鹏. 生物医学文献中命名实体的识别[D]. 大连: 大连理工大学, 2009.

[2] 张冬梅, 晋耀红. 基于模式匹配的中文命名实体识别[C]. 全国知识组织与知识链接学术交流会. 2011.

[3] 国家知识产权局. 国家知识产权局2016年年度报告[R]. 北京: 国家知识产权局, 2016.

[4] 向晓雯. 基于条件随机场的中文命名实体识别[D]. 厦门: 厦门大学, 2006.

[5] Yang Z H, Lin H F, Li Y P. Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature[J]. Computational Biology and Chemistry, 2008, 32(4):287-291.

[6] Fukuda K, Tsunoda T, Tamljra A, et al. Toward information extraction: identifying Protein names from biological papers[C]. Proceedings of the Pacific Symposium on Biocomputing, Hawaii, USA, 1998:705-716.

[7] Olsson F, Eriksson G, Franzen K, et al. Notions of correctness when evaluating protein name taggers[C]. Proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan, 2002:765-771.

[8] Wang H C, ZhaoT J, Tan H Y, et al. Biomedical named entity recognition based on classifiers ensemble[J]. International Journal of Computer Science and Applications, 2008, 5(2):1-11.

[9] 颜军. 基于条件随机场的中文分词研究与应用[D]. 武汉: 武汉理工大学, 2009.

[10] 何赛克, 王小捷, 董远,等. 归一化的邻接类别方法在基于条件随机场的中文分词中的应用[C]. 中国计算机语言学研究前沿进展. 2009.