文章编号:1007-1423(2018)35-0012-07 DOI:10.3969/j.issn.1007-1423.2018.35.003

命名实体识别方法研究进展

黄晴雁,牟永敏

(北京信息科技大学计算机学院,北京 100101)

摘要:
命名实体识别是自然语言处理领域中非常重要的基础任务,近年来,由于机器学习的发展以及深度学习在文本处理方面的应用,命名实体识别的研究取得进一步的发展。为了了解近几年来命名实体识别方法的研究进展,介绍命名实体识别的研究内容与涉及领域,重点详细介绍近几年命名实体识别研究的主要技术方法并进行分析,对 BLSTM-CRF 模型进行实验。

关键词:
命名实体识别; 自然语言处理; 机器学习; 信息抽取

基金项目:
网络文化与数字传播北京市重点实验室开放课题(No.ICDD2017XX);北京市自然科学基金(重点研究专题项目)(No.Z160002);研究生教育(No.71D1811013)

0 引言

命名实体识别(Named Entity Recognition,NER),也称为实体提取,是指对文本中特定的实体进行识别并对区分其种类。近年来,深度学习在自然语言处理领域(Natural Language Processing,NLP)广泛应用,取得了良好的效果,命名实体识别作为基础任务得到了进一步的发展。作为信息抽取的子任务,从非结构化文本中识别并抽取结构化的数据,需要命名实体识别技术作为支撑。同时,随着人工智能的发展,对文本语义层面的研究得到了国内外学者的广泛关注,对命名实体识别的研究有助于理解语义层面的知识。

1 研究内容及应用

1.1 研究内容与领域

从语言分析的过程来看,命名实体识别属于词法分析中的未登录词识别,也就是识别文本中的命名实体(Named Entity,NE)。MUC-6 最早将命名实体作为你一个明确的概念和研究对象提出,以及后来的 MUC-7 规定了命名实体包括三大类(实体类、时间类和数字类)和七小类(人名、地名、机构名、时间、日期、货币和百分比)。ACE 将命名实体中的机构名和地名进行了细分,增加了地理-政治实体和设施实体,之后又增加了交通工具实体和武器实体。

实际早期对于命名实体识别的研究,主要集中于对一般“专有名词”[1]的识别,包括三类名词:人名、地名、机构名。后来随着研究的逐渐展开,研究者们将对命名实体识别的研究扩展到了更多的特定领域。张剑等[2]在农业领域进行了命名实体识别,采用基于条件随机场的方法,将农业命名实体分为病虫害、作物、化肥及农药 4 种命名实体。张磊[3]将命名实体识别的研究应用在了轨道交通领域,并且提出了一种基于条件随机场、半监督学习和主动学习相结合的方法,形成了一个统一的技术框架。余俊等[4]为了能快速、准确地将分散在 Web 网页中的音乐实体抽取出来,提出了一种规则与统计相结合的中文音乐实体识别方法,并实现了音乐命名实体识别系统。

在语言种类方面,命名实体识别对英语、中文、德语、日语、西班牙语、葡萄牙语等都有相应研究。最初的研究主要以英文为主,后来逐渐发展到对多语言和独立语言进行研究。2003 年举办的“963”测评最早将

汉语命名实体识别作为评测任务提出。2006 年 SIGHAN 正式将命名实体识别问题作为其评测比赛的一项任务。近几年，国内很多研究者对我国少数民族的语言进行了命名实体识别研究。金明等[5]对藏语进行了命名实体识别研究；吴金星[6]在蒙古语命名实体识别研究的基础上构建了蒙古语语料加工继承平台；塔什甫拉提·尼扎木丁[7]对维吾尔语文本中的人名命名实体进行了识别研究。

1.2 命名实体识别的应用

命名实体识别是多种自然语言处理技术的重要基础，对于句法分析、语法分析、语义分析等都有着极其重要的影响，主要应用在信息抽取、机器翻译、问答系统等方面。

文本信息抽取是在自然语言文本中抽取出指定类型的实体、关系、事件等事实信息，并形成结构化数据。赵军等[8]对开放式文本的信息抽取进行了研究，认为命名实体识别是信息抽取的基础，同时也是重中之重，并且对于知识库的构建、网络内容的管理、语义搜索等都具有重要的应用价值。

机器翻译，又称为自动翻译，利用计算机将一种自然语言转换为另一种自然语言。在机器翻译时，通常需要对专有名词如人名、地名、机构名等进行精确翻译。例如中国汉语人名翻译成英文时大多用拼音表示，且需要名在前姓在后，而其他普通词语则需要翻译成对应的英文。陈怀兴等[9]对命名实体的机器翻译等价对方法进行了研究，通过实体等价对对齐，得到了较高正确率的机器翻译结果。因此，准确而高效地识别出文本中的命名实体，对于提高机器翻译的准确率有重要意义。

问答系统是信息检索系统的一种高级形式，用准确、简洁的自然语言回答用户用自然语言提出的问题。周波[10]对面向问答系统的实体识别与分类进行了研究，认为实体识别是问答系统的关键技术之一，直接关系到问句类型的判断和答案的抽取。

2 主要研究方法

目前，关于命名实体识别的方法主要分为：基于词典和规则的方法、基于统计机器学习的方法、基于深度学习的方法等。而且，现在较为流行的是将其中两种方法结合甚至是三种结合，可以充分利用不同方法的优点，提高学习的准确度和效率。

2.1 基于词典和规则的方法

早期的命名实体识别工作大多都采用手工编写字典和规则的方法，并且由相关领域的专家来完成，其研究的重点是根据研究领域的特征构造词典并编写规则模板。一般来说，规则的构造需要考虑到该领域的关键字、指示词、中心词、前后缀等特征，依赖于已制定的词典和知识库，通过模式匹配或字符串匹配等方法来识别出命名实体。其中，词典负责已有词汇的识别，规则负责未登录词的识别。

早在 2000 年，Farmkiotou,D 等[11]提出了基于规则的用于希腊金融文本中的命名实体的识别算法。他们认为，典型的命名实体识别系统应是由词典和语法组成的。其中，词典是指研究领域中特有的词汇，而语法是指该领域语言所具有的特征。在新的领域进行研究时，该领域的词典应该是通过手工的方法或者机器学习技术根据其特点来制定的。因此，他们提出了一个基于人工构建词典的命名实体识别系统，并在希腊金融新闻语料库上进行了测试，取得了令人满意的效果。

近几年来，基于字典和规则的方法在学术研究上应用较少，且基本上是与基于统计的方法混合使用，而在实际产业中应用较多。一方面，基于字典和规则的方法精确度较高，往往可以满足实际应用中对准确率的要求，而且在工业中的应用仅限于固定的领域，即便是有新词，对识别系统的改动也不会太大；另一方面，由于语言的复杂性和灵活性，该方法中规则的编写费时费力且难以涵盖所有的语言现象，建设成本较高，并且该方法依赖于具体的领域、语言，可移植性不好，会遇到知识瓶颈问题。

图 1 为基于词典和规则的命名实体识别方法的基本处理过程，其中包括了新规则与新词的添加。

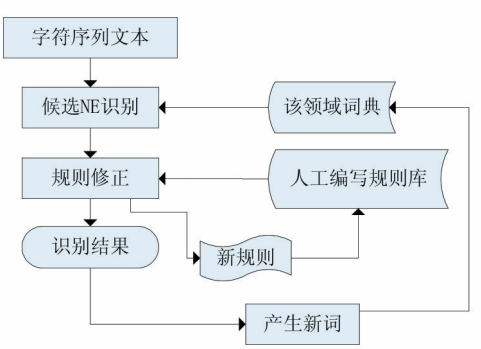

图1 基于规则和词典方法的基本流程

2.2 基于统计机器学习的方法

基于统计机器学习的方法将命名实体识别看做一个分类问题或者序列标注问题，需要利用经过人工标注的语料进行训练。目前该方法主要包括以下几种模型：隐马尔科夫模型（Hidden Markov Mode，HMM）、最大熵模型（Maximum Entropy，ME）、条件随机场（Conditional Random Fields，CRF）、决策树（Decision Tree）等。总的来说，该方法的步骤主要可以总结为：预处理语料、抽取特征并制定特征模板、训练模型、优化模型。

图 2 为基于统计机器学习的命名实体识别的流程。

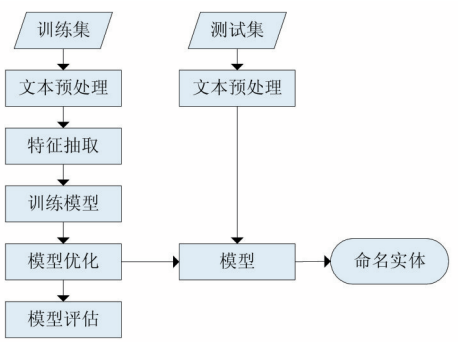

图2 基于统计机器学习方法的基本流程

近几年来，机器学习在命名实体识别方面取得了很大的进展，研究者们一直致力于设计识别效果更好、应用范围更广的算法，并取得了一定的成功。

2018 年，周法国等[12]提出了一种基于转移学习的中文命名实体识别算法，将命名实体识别看做分类任务，进行了中文人名、地名、组织机构名的识别。该算法有统计与规则相结合，利用初始标注语料及规则模板形成规则，对规则进行统计训练得到规则标注序列。所谓转移学习，主要是基于成功转换数据来更正数据，依据错误率获得较大的成功。其中心思想是开始以一些简单的结论应用于问题，然后在每个步骤应用转换，选择出每次转换的最优结论再次应用于问题，当选择的转换在足够的空间内不再修改数据时算法停止。实验验证，该模型获得了较好的结果。

高冰涛等[13]认为传统的生物医学领域命名实体识别标注数据代价较高，因此关注命名实体识别的迁移学习。他们在权值学习模型的基础上，构建了基于迁移学习的隐马尔可夫模型算法 BioTrHMM，其目的是降低生物医学文本中命名实体识别对目标领域标注数据的需求。BioTrHMM 算法在使用较少的目标领域数据的情况下，以相关领域数据为辅助数据集，利用数据引力的方法计算权值来评估辅助数据集的样本在目标领域——生物医学领域学习中的贡献程度，从而进行知识的迁移。该研究选取了 GENIA 语料库中的数据集，取得了较好的实验结果。

Yukun Chen 等[14]提出了一种基于主动学习的临床命名实体识别标注系统，任务是从临床笔记中提取问题、治疗和实验室相关实验的概念。该标注系统是基于已经标注的句子迭代地构建命名实体识别模型，并且选择下一个句子进行标注。系统的前端是一个用户推断界面，用户可以通过特定的查询引擎在系统提供的句子中标记临床命名实体。系统的后端会根据用户的注释对 CRF 模型进行迭代训练，并根据查询引擎选择最有用的句子。该系统的工作流程如图 3 所示：

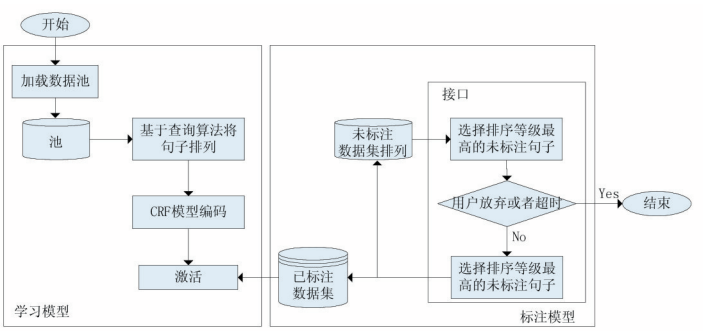

图3 主动学习模型

李刚等[15]将研究的关注点放在近年来发展迅速的微博等网络社交平台上，认为其独特的形式对传统的命名实体识别技术提出了新的挑战。因此，他们提出了一种基于条件随机场模型的改进方法，针对微博文本短小、语义含糊等特点，引入外部数据源提取主题特征和词向量特征来训练模型，针微博数据规模大、人工标准化处理代价大的特点，采取一种基于最小置信度的主动学习算法，以较小的人工代价强化模型的训练效果。研究选取了新浪微博数据集，并且考虑了中文的深层语义。实验证明，该方法与传统的条件随机场方法相比 F 值提高了 4.54%。

基于统计机器学习的方法对特征选取的要求较高，对语料库的依赖较大[2]。该方法的难点是构建特征工程，需要从语料文本中选取对研究任务有积极影响

的特征。而对于特征的构建，需要考虑选择的特征是否能有效地反映该类实体的特点，可以利用的特征包括字符、词性、词边界等。同时，组合特征可以表达出更复杂的含义[16]。

2.3 基于深度学习的方法

一般来说，深度学习是机器学习的一种。早期机器学习专家提出了人工神经网络（Artificial Neural Networks），与传统的统计机器学习算法不同。近几年来，随着科学技术的发展，基于神经网络的深度学习在机器学习领域掀起了一股热潮，同时也越来越多地将其应用到了自然语言处理上。近几年，比较通用的基础神经网络结构有 BLSTM-CRF、卷积神经网络（CNN）等，都取得了不错的识别效果。

Feng Y H 等[17]针对传统的命名实体识别方法需要构建特征工程和获取相关领域的知识，然而代价昂贵的问题，提出了一种基于 BLSTM（Bidirectional Long Short-Term Memory）的神经网络结构的命名实体识别方法。该方法利用基于上下文的词向量和基于字的字向量，前者表达命名实体的上下文信息，后者表达构成命名实体的前缀、后缀和领域信息；同时，利用标注序列中标签之间的相关性对 BLSTM 的代价函数进行约束，并将领域知识嵌入模型的代价函数中，进一步增强模型的识别能力。实验表明，所提方法的识别效果优于传统方法。

李丽双等[18]在生物医学领域进行了命名实体识别任务研究，提出了一种基于 CNN-BLSTM-CRF 的神经网络模型。首先利用卷积神经网络（CNN）训练出单词的具有形态特征的字符级向量，并从大规模背景语料训练得到具有语义特征信息的词向量，然后将二者进行组合作为输入，再构建适合生物医学命名实体识别的 BLSTM-CRF 深层神经网络模型。实验数据来自于 Biocreative Ⅱ GM 和 JNLPBA2004 生物医学语料，实验结果的 F-值分别为 89.09%和 74.40%。图 4 为该模型的结构框架。

2018 年，Yanyao Shen 等[19]提出了利用深度主动学习进行命名实体识别任务。将主动学习与深度学习相结合，可以利用少量的标注数据获得较高的学习准确度。由于主动学习的计算成本很高，因此他们提出了一个基于 CNN-CNN-LSTM 结构的轻量级模型。众所周知，在收集有标注的数据集的时候，需要依靠大量的人工标注，准确标注出正确的命名实体类别是非常耗时耗力的。因此，提出深度主动学习方法以便于减少标注量，降低数据标注的成本。实验表明，该模型能够迅速地对样本进行预测和评估不确定度。

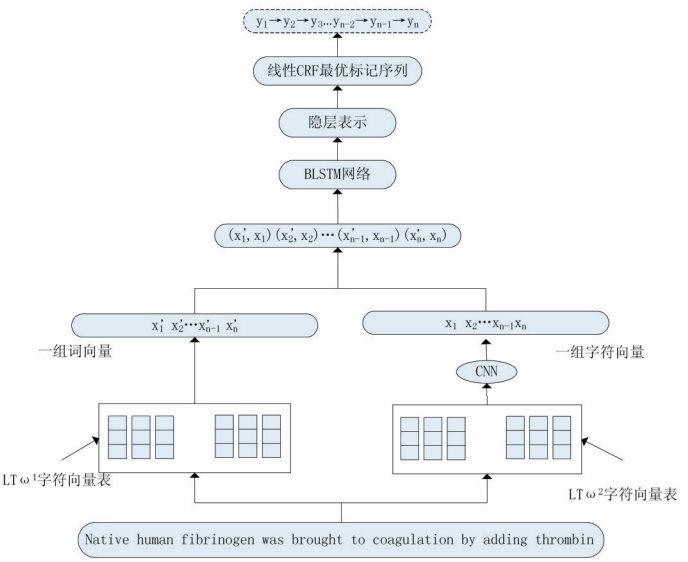

图4 生物医学命名实体识别的CNN-BLSTM-CRF模型

Akash Bharadwaj 等[20]提出了一种注意力神经模型（Attentional Neural Model）。该模型在原始的 BLSTM-CRF 模型上加入了音韵特征，并在字符向量上使用注意力机制来关注并学习更有效的字符。该模型可以快速地应用于有少量数据或没有数据的新语言领域，从而实现了跨语言的迁移学习。

深度学习使用词向量表示词语、字向量表示字，解决了传统命名实体识别方法需要花费大量精力构建特征工程的问题，甚至会人工构建特征工程包含更多的语义信息。虽然深度学习在命名实体识别研究上已经取得了较好结果，但仍有很多研究者致力于将新的技术应用到命名实体识别问题上。当前的研究趋势主要集中在两个方面：一是使用流行的注意力机制（Attention Mechanism）来提高模型的效果；二是致力于利用少量的标注训练数据进行研究。

3 实验部分

本文在前人研究的基础上对基于 BLSTM-CRF 的命名实体识别方法进行了实验，实验所采用的数据是

来自全国知识图谱与语义计算大会(China Conference on Knowledge Graph and Semantic Computing, CCKS) 2017 年任务二和 2018 年任务一的数据,均是来自于中文临床电子病历。

3.1 实验内容

本文实验采用的模型是 BLSTM-CRF 结构,并分别对两组数据进行了实验。实验对数据以字符为单位进行了标注,采用了 BIO 标注方法,即 B 表示实体的首字,I 表示实体的非首字,而 O 表示该字不属于实体的任何一部分。

2017 年评测大会的实验数据给出了疾病和诊断、身体部位、症状和体征、检查和检验以及治疗五类实体,本文用不同的标识符号分别对其进行了标识,并进行了统计,如表 1 所示。

表 1 CCKS 测评 2017 年 Task2 实验数据统计

标识符号	实体类别	示例	数量
D	疾病和诊断 (Disease)	左侧粗隆间骨折	502
B	身体部位 (Body)	右髋部、头部	8072
S	症状和体征 (Symptom)	疼痛肿胀、疼痛	6137
Te	检查和检验 (Test)	T、P、叩击痛	5732
Tr	治疗 (Treatment)	左大腿中下段截肢手术	694

从表 1 中可以看出,该任务给出的训练集中疾病和诊断这一类实体仅有 502 个,治疗类实体仅有 694 个,而身体部位这类实体有 8072 个,五类实体之间的数量有较大的差距。在 BIO 标注基础上,有如下标注例子:

肠 鸣 音 活 跃 , 双 下 肢 无 水 肿

B-Te I-Te I-Te O O O B-B I-B I-B O B-S I-S O

2018 年的数据将实体分为了手术、解剖部位、症状描述、独立症状以及药物五类。同样,对实验数据进行了统计,如表 2 所示:

表 2 CCKS 测评 2018 年 Task1 实验数据统计

标识符号	实体类别	示例	数量
OPE	手术 (Operation)	直肠癌根治术	1120
ANA	解剖部位 (Anatomic site)	直肠、肝、腹	7838
SYM	症状描述 (Symptom)	痛、胀、疼痛	2066
INS	独立症状 (Independent Symptom)	呕吐、发热、寒战	3055
MED	药物 (Medicine)	替加氟、奥沙利铂	1005

从表 2 中可以看出,2018 年该任务给出的训练集中除解剖部位类实体有 7838 个,其他四类实体的数量差距相对较小。

该模型的结构如图 5 所示:

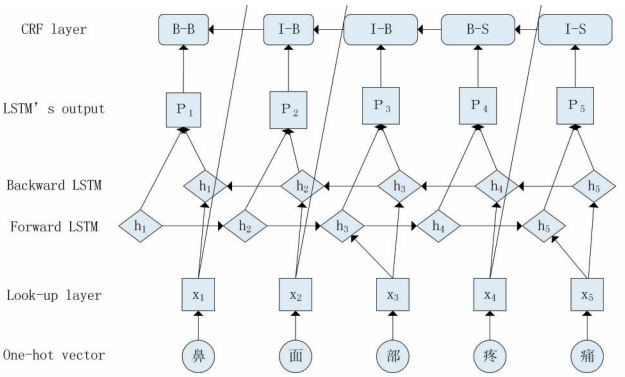

图5 BLSTM-CRF结构图

模型是以句子为单位进行输入,将一句话看作 n 个字符的序列 $(x_1,x_2,\cdots x_n)$。Look-up 层将句子中的每一个字符 x_i 映射为低维度稠密的字向量(character embedding) $x_i \in R^d$,其中, d 是字向量的维度。

BLSTM 结构对文本的上下文有记忆和过滤的能力,对长距离的信息能有效地运用,对序列数据所包含的信息能够动态捕获。将每个句子的字符序列 $(x_1,x_2,\cdots x_n)$ 作为 BLSTM 的输入,正向 LSTM 返回序列 $\overrightarrow{h_t}=(\overrightarrow{h_1},\overrightarrow{h_2},\cdots,\overrightarrow{h_n})$,反向 LSTM 返回序列 $\overleftarrow{h_t}=(\overleftarrow{h_1},\overleftarrow{h_2},\cdots\overleftarrow{h_n})$,直接拼接 $\overrightarrow{h_t}$ 与 $\overleftarrow{h_t}$,得到 BLSTM 在 t 时刻的输出,表示为 $h_t=[\overrightarrow{h_t};\overleftarrow{h_t}]$。

由于 CEF 是全局范围内统计归一化的条件转移概率矩阵,因此,CRF 层对文本进行了句子级别的序列标注,使模型可以学习到标签的上下文关系。

3.2 实验结果分析

通过调整模型的参数,得到较为理想的实验结果如表 3 所示:

表 3 实验结果

结果评价	实体类别				
2017 年数据	疾病和诊断	身体部位	症状和体征	检查和检验	治疗
F1 值	49.43%	92.57%	95.67%	93.99%	49.08%
2018 年数据	手术	解剖部位	症状描述	独立症状	药物
F1 值	78.55%	89.57%	85.44%	89.17%	78.55%

实验结果表明,训练语料的规模能够对识别结果产生较大的影响。总的来说,BLSTM-CRF 模型能取得较好的识别效果。2017 年数据的实验,对训练数据较多的身体部位、症状和体征、检查和检验三类实体分别取得了 92.57%、95.67%、93.99%的识别效果。然而对

于训练数据较少的疾病和诊疗、治疗这两类实体的识别效果就不理想，仅取得了 49.43%和 49.08%的识别效果。同样，对于 2018 年的实验数据来说仍是如此。但整体识别效果在 75%-90%之前。

4 结语

自然语言处理领域最为关心的技术问题之一是如何高效率地从不规范的非结构化文本数据中，获取并组织成结构化的文本数据。命名实体识别任务作为自然语言处理的基础任务，能够有目的性地对文本进行结构化处理。虽然，对于命名实体识别的研究已趋近于成熟，但是仍有很多学者认为该问题还未得到完善解决，对命名实体的外延和内涵的探讨还远未结束。目前，深度学习发展火热，仍将是命名实体识别研究最为关注的领域，减少语料数据的标注、扩展研究领域也将是命名实体识别研究的重点。

参考文献：

[1]Christine Thielen. An Approach to Proper Name Tagging for German. In Proc. Conference of European Chapter of the Association for Computational Linguistics. SIGDAT,1995.

[2]张剑,吴青,羊昕旖,等. 基于条件随机场的农业命名实体识别[J]. 计算机与现代化,2018(1):123-126.

[3]张磊. 特定领域的命名实体识别方法的研究[J]. 计算机与现代化, 2018(3).

[4]佘俊,张学清. 音乐命名实体识别方法[J]. 计算机应用,2010,30(11):2928-2931+2948.

[5]金明,杨欢欢,单广荣. 藏语命名实体识别研究[J]. 西北民族大学学报(自然科学版),2010,31(03):49-52.

[6]吴金星. 蒙古语语料库加工集成平台的构建[D]. 内蒙古大学,2015.

[7]塔什甫拉提·尼扎木丁. 维吾尔语文本信息中人名实体识别研究[D]. 新疆大学,2016.

[8]赵军,刘康,周光有,蔡黎. 开放式文本信息抽取[J].中文信息学报,2011,25(06):98-110.

[9]陈怀兴,尹存燕,陈家骏. 一种命名实体翻译等价对的抽取方法[J]. 中文信息学报,2008(04):55-60.

[10]周波. 面向问答系统的实体识别与分类研究[D]. 沈阳航空工业学院,2009.

[11]Farmakiotou D,Karkaletsis V,Koutsias J,et al. Rule-Based Named Entity Recognition for Greek Financial Texts[C]. Proc. of the International Conference on Computational Lexicography and Multimedia Dictionaries COMLEX 2000,2000:1-4.

[12]周法国,吴锡坤,孙泰,孙镇. 基于转移学习的中文命名实体识别[J]. 计算机工程与应用,2018,54(05):117-121.

[13]高冰涛,张阳,刘斌. BioTrHMM:基于迁移学习的生物医学命名实体识别算法[J/OL]. 2019,36(1). [2018-04-03]. http://www.aroc-mag.com/article/02-2019-01-035.html.

[14]Chen Y,Lask T A,Mei Q, et al. An Active Learning-Enabled Annotation System for Clinical Named Entity Recognition[J]. BMC Medical Informatics & Decision Making,2017,17(Suppl 2):82.

[15]李刚,黄永峰. 一种面向微博文本的命名实体识别方法[J]. 电子技术应用,2018,44(1):118-120,124.

[16]张祝玉,任飞亮,朱靖波. 基于条件随机场的中文命名实体识别特征比较研究[C]. 第 4 届全国信息检索与内容安全学术会议论文集,2008.

[17]Feng Y H,Hong Y U,Sun G,et al. Named Entity Recognition Method Based on BLSTM[J]. Computer Science,2018.

[18]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(1).

[19]Shen Y,Yun H,Lipton Z C,et al. Deep Active Learning for Named Entity Recognition[J]. 2018:252-256.

[20]Bharadwaj A,Mortensen D,Dyer C,et al. Phonologically Aware Neural Model for Named Entity Recognition in Low Resource Transfer Settings[C]. Conference on Empirical Methods in Natural Language Processing,2016:1462-1472.

作者简介：

黄晴雁(1993-)，女，山东潍坊人，硕士，研究方向为自然语言处理、软件理论与应用

牟永敏(1961-)，男，山东烟台人，博士，教授，CCF 会员，研究方向为大数据技术、软件理论与应用

收稿日期：2018-09-25　　修稿日期：2018-10-17

（下转第 22 页）

[11]周显春,肖衡. Spark 2.0 平台在大数据处理中的应用研究[J]. 软件导刊,2017,16(5):149-151.
[12]缪广寒. 关联规则 Apriori 算法在个性化学习系统中的应用研究[J]. 硅谷,2014(5):47-48.
[13]姜强,赵蔚,刘红霞,等. 能力导向的个性化学习路径生成及评测[J]. 现代远程教育研究,2015(6):104-111.
[14]闵宇锋. 网络教学系统中的可视化学习监控机制[J]. 现代教育技术,2010,20(1):92-96.

作者简介:
周显春(1974-),男,湖南常德人,硕士,讲师,研究方向为网络安全及数据挖掘
肖衡(1979-),女,湖南衡阳人,硕士,讲师,研究方向为计算机网络
高华玲(1980-),女,河北唐山人,硕士,讲师,研究方向为大数据语义分析
收稿日期:2018-10-18　　修稿日期:2018-10-30

Application of Apriori Algorithm in Personalized Learning

ZHOU Xian-chun, XIAO Heng, GAO Hua-ling

(School of Information and Intelligence Engineering, Sanya University, Sanya 572022)

Abstract:
To improve the effect of personalized learning, mainly improves the previous research from two aspects under the massive data of education and teaching. Firstly, parallelized Apriori in the Spark to improve the real-time performance of it; Secondly, the learning objectives and knowledge points modify the sequence of knowledge points recommended by Apriori, to make the learning contents or paths more targeted and improve the quality of personalized learning. It has been verified by specific teaching practice that the personalized learning recommendation path based on learning objectives and knowledge points can greatly improve learners´ learning efficiency, solve cognitive overload, network maze and other problems, and provide research ideas for the realization of such a platform.
Keywords:
Apriori; Individualized Learning; Knowledge Point; Learning Path; Study Analysis

(上接第 17 页)

Research Progress of Named Entity Recognition

HUANG Qing-yan, MU Yong-min

(School of Computer, Beijing Information Science and Technology University, Beijing 100101)

Abstract:
Named entity recognition is a very important basic task in the field of natural language processing. In recent years, with the development of machine learning and the application of deep learning in text processing, the research of named entity recognition has been further developed. In order to understand the research progress of named entity recognition in recent years, introduces the research contents and related fields of named entity recognition, and focuses on the main technical methods and analysis of named entity recognition in recent years, and tests the BLSTM-CRF model.
Keywords:
Named Entity Recognition; Natural Language Processing; Machine Learning; Information Extraction