

实体抽取综述及其在中医药领域的应用*

孔静静, 于琦**, 李敬华, 于彤, 张竹绿, 田野, 祖雅琪

(中国中医科学院中医药信息研究所 北京 100700)

摘要:随着机器学习、深度学习与人工智能的兴起与发展,自然语言处理技术也不断革新,其中实体抽取作为自然语言处理的子任务,自其提出以来,实现了从基于匹配模式的抽取到基于传统的机器学习方法抽取,再到基于深度学习的抽取的进步与发展。在医疗信息抽取方面,自然语言处理技术也得到了很好地应用,医学信息得到了更深层次的处理和挖掘。面对中医药领域中灵活的、复杂的和海量的中医药信息,利用实体抽取技术可以将中医药领域更为重要的信息抽取,这也是进一步对中医药信息挖掘的关键一步,基础一步。本文概括了中医药各领域实体抽取技术的应用研究情况,为实体抽取技术在中医药领域的进一步应用发展提供参考,以期促进中医药的传承发展,守正创新。

关键词:中医药 命名实体识别 实体抽取 中医应用 综述

doi: 10.11842/wst.20210629009 中图分类号: R9-39 文献标识码: A

信息抽取技术(Information Extraction, IE)的目的是将非结构化的信息进行结构化抽取^[1]。自然语言处理技术(Natural Language Processing, NLP)可以把人类语言转化为机器语言,从而实现人机交互,将非结构化的文本转化为结构化信息,从而获得有用的、更加便于处理的信息^[2-3],结构化的信息更有利于知识发现和挖掘。命名实体识别(Named Entity Recognition, NER)属于自然语言处理的任务之一,又称为实体抽取,其目的是从信息中抽取的人名、地名、组织机构名等命名实体^[4]。Rau^[5]于1991年提出了从文本中提取公司名称的实体抽取任务,此外,在七届MUC(Message Understanding Conference)会议形成了沿用至今的抽取评价标准,大大推动了实体抽取发展的进程。随后,国际上举办众多有关NER发展的会议,涉及了英语、阿拉伯语、汉语等多种语言,包括消歧、评估等多任务方向的发展^[6]。上世纪90年代国内出版了规则库的识别方法的书用于识别人名^[7],由于

英文语言的天然分词结构及较为规范的语法表达方式,早期的结构化信息抽取主要面向英文文本,信息抽取技术在英文上得到了很好的应用和发展,而国内的信息抽取技术起步较晚,中文语法复杂,不能直接进行分词等特点,使实体抽取在中文上的应用面临一定的挑战^[8-9]。

随着大数据时代的到来,数据量的剧增,人们难以从海量的数据中获取有效的信息,实体抽取技术应运而生。实体抽取近年来飞速发展,为信息化的发展带来了新的契机。现如今实体抽取在生物医学、警情军事及农业渔业等方面均有广泛应用^[10]。在实际应用当中,由于不同的行业领域之间实体的相关性较弱,因此命名实体的确切含义需要根据具体应用来确定,例如在面向中医命名实体信息提取时,实体可为:中医证候、舌象、脉象、治法、中药和方剂等,在农业领域可能会把农作物、病虫、农药^[11]等作为实体。实体抽取作为自然语言处理的一个重要研究内容,在信息提

收稿日期:2021-06-29

修回日期:2022-02-25

* 中国工程科技知识中心建设项目(CKCEST-2020-1-18):中医学专业知识服务系统,负责人:黄璐琦、李宗友;中国中医科学院基本科研业务费自主选题项目(ZZ130303):中医传承工作站的系统集成与探索应用,负责人:于琦;国家自然科学基金委员会面上项目(81873200):基于深度学习的脾胃病关键诊疗要素发现与临床优化决策研究,负责人:李敬华。

** 通讯作者:于琦,副研究员,硕士生导师,主要研究方向:中医药信息学研究。

取、信息检索、主题分类、知识发现等方面应用广泛^[12],同时,知识图谱的构建也是以实体抽取为前提条件。实体抽取为更深层次的数据挖掘提供基础^[13]。

随着信息抽取技术不断的革新,在中文文本的应用领域不断拓宽,抽取效果不断改善。信息抽取技术在中医药领域的应用能为中医药信息更深层次地挖掘奠定基础,让中医药知识更好地服务于临床应用,有益于推动中医的继承和发展,使中医药与现代技术接轨。本文简要介绍了实体抽取概念及发展,阐述了其发展过程中常见的技术手段,从中医药角度出发,论述了实体抽取在中医药各领域中的研究应用情况,为中医药实体抽取的发展研究提供思路参考,以期拓宽实体抽取在中医药领域应用的范围,推动实体抽取技术的应用发展。

1 实体抽取技术

1.1 基于模式匹配的实体抽取

1.1.1 基于规则的抽取

基于规则的实体抽取是最早使用的方法,1991年Rau在第7届IEEE人工智能应用会议上提出基于人工编写规则的方法和启发式算法,研究如何从文本中抽取公司的名称^[5]。基于规则的方法由专家针对数据集进行人工构造规则模板,将需要处理的信息与已制定规则进行匹配,将符合的内容信息提取,并且在抽取过程中不断补充完善规则,以达到最优化的抽取效果,通常选用特征信息等方法^[14]。基于规则抽取的方法具有较好的灵活性,操作简单,但抽取效果高度依赖制订的规则,只适用于表达较为规范的文本和较小的数据集,且移植性较差,规则制作过程中耗费大量资源。

1.1.2 基于词典的抽取

基于词典的方法最早是和基于规则的抽取方法联合使用的,此方法需要构建特定的词典,然后按照需求将需要识别文本与所构建的字典进行匹配抽取^[10],并在过程中不断补充完善词典。早期就有学者利用大量的医疗知识词典结合既定的规则来建立专家系统,目前,针对不同领域已经构建了比较完备的词典,如基因词典、情感词典、姓名词典等。此方法词典是核心,因此依赖词典的准确性,适合于精确的搜索定位,对于词典范围外的实体难以准确识别,针对这一问题,尽管后来提出一种模糊字典匹配法,但

是仍然摆脱不了对词典的依赖。

基于模式匹配的实体抽取的两种方法在一定程度上实现了实体抽取,但是由于基于模式匹配两种方法的自身的局限性,很难再利用原方法对其改进提升,再有大数据时代的来临,信息量剧增,规则和词典的制定更加繁琐困难,因此,机器学习的应用逐渐兴起。

1.2 基于机器学习的实体抽取

1.2.1 隐马尔科夫模型 HMM

HMM最早由Rabiner等^[15]在统计学论文中提出,随后在语言识别中应用。HMM模型构建较为简单,在自然语言处理早期,在模式识别等领域得到广泛的应用,取得了不错的效果。HMM模型对特定位置和文本长度有所限制,所以该模型的适应性不理想。此外,有些学者通过建立高阶的HMM模型来优化模型抽取效果。

1.2.2 最大熵模型 ME

1957年E.T.Jaynes^[16]提出了最大熵原理。最大熵模型不对未知信息做任何假设,而是在已知特征的约束下使未知的信息分布去均匀分布,即熵最大^[17]。在最大熵方法中,就是找出一个特征的集合,并确定每条特征的重要程度,以此来进行抽取,最大熵模型可以集成各种特征与规则到一个统一的框架下^[18]。

1.2.3 支持向量机 SVM

支持向量机是由Cortes等^[19]在1995年提出并发表,SVM是指的一种二分类的统计模型,它是在特征空间中,使得两类样本间隔最大的线性分类器,它的分类方法就是最大化间隔^[20]。SVM算法分类思想简单,适用于小样本和中样本,大规模训练样本会耗费大量计算时间和内存,在解决多分类问题也存在困难。

1.2.4 条件随机场模型 CRF

CRF是21世纪初由Lafferty et al.^[21]首次提出发表,它是通过最大熵模型转换而来,条件随机场可以通过训练集的语料特征推断每个标签应有的类别标记。CRF即通过计算不同候选序列的联合分布概率,选择概率最大的序列作为输出结果^[22]。相比于SVM、HMM具有无标注偏见、可求得全局最优值、小规模数据可获得理想效果等优点^[23]。

实体抽取引入了机器学习方法,减少了对规则及词典的依赖,使得抽取效果进一步提升。机器学习方

法是从样本数据集中统计出相关特征和参数,以此建立识别模型对数据进行抽取^[14]。为提高抽取效果,机器学习可与规则库和词典方法结合使用来抽取,其抽取的效果也远远大于单一方法抽取^[24]。机器学习的模型抽取的方法虽然优于匹配模式的抽取方法,但是在构建模型时仍旧需要投入一定的资源用在人工标注数据和特征选择上且泛化能力不高^[25]。因此,随着算法的不断改进,引入了深度学习,对抽取进行了再度优化。

1.3 基于深度学习的实体抽取

1.3.1 Word2vec

2013年Mikolov等^[26]在谷歌团队的带领下提出了Word2vec,是用来生成词向量的模型,是一种浅层神经网络概率语言模型,一般分为CBOW和Skip-gram两个基础模型,可以将词语转换为包含语义的词向量,表达词内间关系,词语与文本前后内容的联系^[27]。Word2vec模型提出有效提升了NLP任务的处理效果。

1.3.2 Attention 机制

注意力机制是根据人类的注意力特点而提出的研究,可以实现高效分配信息资源。注意力机制能够通过调整权重值,来锁定所需的重要信息,同时可以过滤掉不重要的信息。具有很高的可扩展性和鲁棒性。在2017年,Vaswani^[28]提出了Transformer模型,随后,Kitaev等^[29]提出了新的Reformer模型,改进原来的Transformer模型。注意力机制与传统算法的结合,能大大提高算法系统的性能,提升抽取能力。

1.3.3 LSTM

长短期记忆网络(Long Short-Term Memory)本质上是一种RNN(循环神经网络)的递归神经网络结构,LSTM模型的3个门:“遗忘门”“记忆门”和“输出门”,可以筛选过滤出短序列,选择长序列处理成短序列,将含重要信息的进行传递^[30]。

1.3.4 BERT

BERT模型是在2018年提出的预训练模型,是基于Transformer的双向编码器表征,其结构是来自于Transformers模型的Encoder,在处理一个单词时,能够联系词语上下文,从而得到其含义。BERT被设计为用未标记的文本去训练出一个模型,它具有更深的层数,很好的泛化性和并行性,可以充分描述字符级、词级、句子级甚至句间关系特征^[31]。并且模型增加了对上下文的记忆。可以把BERT模型看作一个深层的

Word2vec模型,对于一些特定的任务,只需在BERT后面下接一些网络结构,就能很好的完成预设的任务。在文本挖掘领域,其优势巨大,BERT模型的出现是自然语言处理的一个巨大的进步,推动了NLP任务的发展。

深度学习(Deep Learning)是机器学习的分支^[32],是一种以人工神经网络为架构,对数据进行自动的特征学习的算法。深度学习具有学习能力强、覆盖范围广、适应力强和可移植性好的优点,但是,深度学习与机器学习相比较需要大的数据量,涉及的各类参数也更多,模型更加复杂。深度学习的发展,给实体抽取又提供了很大的改进方法,相比于基于规则或传统机器学习的方法,是当前最受欢迎的实体抽取方法。近期的命名实体识别方法不再单独利用单一的方法应用,通常会融入基于匹配模式、基于机器学习、深度学习中两种甚至多种方法,根据研究信息的特点选取抽取技术。还会在模型中加入注意力机制、迁移学习等来提高命名实体识别的效果。目前,BERT加双向长短期记忆神经网络和条件随机场是命名实体识别的热门方法。

2 抽取评估方法

对抽取任务进行抽取评价,是评价一个抽取方法或模型优劣的准则,为其他学者的研究提供参考价值。通常,在实体抽取任务中,一般采用3个评价指标包括:精确率(Precision)、召回率(Recall)和F1值。对于给定的测试数据集,精确率是指正确预测为正占全部预测为正的比例,召回率则是正确预测为正占全部正样本的比例,而F1值则是精确率和召回率的调和平均值,可以对系统的性能进行综合性的评价。通常综合3个指标情况,对抽取方法进行评估抽取效果。

3 中医领域中实体抽取技术的应用

3.1 中医药领域信息特点

中医药领域包含了海量的中医学知识,对其进行信息抽取意义重大。到目前为止,中医药信息还没有得到充分的挖掘和利用,中医药资源的海量性、散乱性和不规范性等特点,致使人们难以从海量的中医数据中精准获取自己所需的信息,对中医药不断地发掘和利用,有着巨大的医学价值和社会价值,同时中医药信息的特点也让实体抽取的应用面临更大的挑战。

从非结构化的中医信息中抽取有用信息,有助于中医信息的精准利用,避免人力以及资源的浪费,提高中医信息的利用率,对中医的进一步挖掘应用及传承发展有着十分重要的推动作用,从而有助于为中医临床研究提供参考价值和辨证思路。

中医药数据的特点可以概括为:①多样性。中医药数据来源广泛,由于中医药历史悠久,且国家面积广阔,形成了地域性的中医药文化,其涵盖了大量的中医药信息。且随着信息化时代的到来,中医药数据也越来越多样丰富。②复杂性。中医数据的表达呈现多样性,包含繁体字,民族语言,文言文等,使得抽取难度增加。③不规范性。例如对于同种疾病的描述可能会因人而异、因医院而异,中草药药物分布范围广泛,相同的一味中药名称的表达也不尽相同,进行统一的规范标准实行困难,这种没有进行统一规范化处理的数据信息,加深抽取的困难程度。

3.2 中医医案及中医电子病历领域的应用

中医医案包含大量信息,其专业性强,短句形式的特点给结构化抽取带来了巨大的挑战,非结构化的中医医案无法直接被计算机识别和利用。中医医案的知识丰富,实体抽取技术在中医医案中广泛应用,使医案结构化成为可能。

3.2.1 中医医案

刘博等^[33]针对统一阶隐马尔可夫模型在解决词性标注捕获上下文的信息有限问题上的不足,推导了二阶HMM的主要学习方法模型,该模型能更多地联系上下文,使得中医诊断文本的标注更为精确。但构建语料库较小,需要进一步优化。屈丹丹等^[34]对比了TFIDF与Word2vec方法抽取结果,随着抽取个数的不断增加,Word2vec在医案信息抽取过程中,由于考虑了其上下文之间的联系,使抽取结果的精确率和召回率较高。肖瑞等^[35]基于BiLSTM-CRF的中医文本命名实体识别研究,基于Word2vec的向量构建,从而进行模型训练迭代,利用CRF层,加强文本间信息的相关性,实现对中医医案文本进行命名实体识别,但训练数据量较小,具有一定的局限,有待进一步的优化。高佳奕等^[36]针对肺癌医案构建了LSTM-CRF混合模型,并且设计了多种算法,通过测试分析比较不同模型优劣,发现带有Peephole机制的双向LSTM识别效果最好,能够有效实现中医医案症状命名实体识别。但数据和标签等问题仍需进一步解决。Liu等^[37]基于

500份标注的中医医案数据,构建了对医案中症状和病机实体抽取的CRF训练模型,利用模型来抽取症状和病机,实验结果经过交叉验证,证明该模型的性能良好,F1达到了80%以上,适用于中医医案的信息提取利用,并且随着训练医案的增加,基于CRF的病机和症状实体识别性能稳步提升。Yang等^[38]针对中医医案中的分词、实体多样性和歧义性等难点,提出了一种基于BiLSTM-CRF的深度学习混合模型命名实体识别方法,获得较好的实体抽取效果,处方实体抽取精确率在90%左右,并且构建了知识图谱,可以提供检索、可视化等功能,帮助中医知识的学习和共享。

3.2.2 中医电子病历

刘一斌等^[39]的基于朴素贝叶斯和Word2vec对中医电子病历文本信息抽取,利用两种算法对中医电子病历长短文本进行了抽取,但是由于数据的限制以及供训练的中医电子病历的文本格式单一,可能造成了此方法的局限性。Jia等^[40]介绍了一种从中医临床记录中提取医疗实体的远程监督NER方法。它利用预处理的语言模型和简单的多层神经网络作为分类器来检测和分类实体,并且加入了过滤机制,利用该方法对中医临床记录的实体抽取,得到F1得分为77.34%,实验结果表明,该方法比其他基线方法具有更好的性能。

3.3 中医药文献领域的应用

中医文献按时间划分为古代文献、近代文献和现代文献。我国保留了大量的中医药文献,包含了众多名医的诊疗记录及中医药信息,体现了历代医家的中医思想,对其进行处理抽取可以便于更好的获取和利用中医文献中的知识,从而有助于药物搜索、药物挖掘等。但是很多中医文献往往知识体系复杂、句式复杂,存在大量的繁体字、通假字以及专业术语等,因此进行抽取困难。尤其是和民族医药相关的信息,处理难度更大。由于利用传统的方法难以实现,所以很多学者引入深度学习方法对中医文献进行实体抽取。

3.3.1 古籍文献

在中医古籍文献方面,叶辉^[41]通过条件随机场研究出从中医古籍中抽取症状和药物的方法。利用此方法对《金匮要略》内容进行有效提取,抽取效果较其他方法更好,为将来建立中医药搜索引擎及新药物发掘等方面提供了一种可行的方法。高甦等^[42]采用了BiLSTM-CRF的实体识别模型,对中医典籍《黄帝内

经》中的中医认识方法、中医生理、中医病理、中医自然、治则治法等5种实体进行识别,与其他方法对比获得了较好的识别效果。张艺品等^[43]选用中医典籍《备急千金方》、《千金翼方》、《神农本草经》作为语料,利用BiLSTM-CRF模型结构,对中医疾病、方剂、中草药等内容抽取,对比单独的HMM模型和CRF模型,结果表明BiLSTM-CRF模型算法抽取效果具有较高准确性。北京科技大学的学者^[44]发明了一种面向中医古籍文献的命名实体识别方法和装置专利,根据标注的中医古文语料训练模型Word Embedding词嵌入,并使用训练得到的AutoNER自动命名实体识别模型,对中医古文语料进行了实体识别。并结合现有的语言训练模型,研发了一种方法系统,用以解决中医命名实体的识别问题^[45]。基于小样本训练集,节省了人工标注的成本,提高了识别效果并且易操作,实现了对中医古籍文献的有效和更全面地利用。Qu等^[46]针对中医领域的文本进行命名实体识别,构建了基于BERT-BiLSTM-CRF模型,以《伤寒论》为训练集和测试集,对症状、疾病名称、时间、处方名称和药物名称五个实体进行抽取,由于BERT在学习上下文特征更明确,与其他模型结合提高了总体性能,其抽取结果优于BiLSTM-CRF及LSTM-CRF模型,且若提高样本的训练量,会取得更好的效果。

3.3.2 现代文献

现代文献方面,Zhu等^[47]针对中医教材,将语义本体与SVM分类和正则表达式匹配相结合,用于提取症状和诊断信息,把数据进行自动关联,构建了结构化知识库,为数字图书馆提供了服务。Deng等^[48]利用双向长短记忆网络与条件随机场相结合的方法对中药专利中的中药、疾病、症状等实体进行了识别抽取,整体F1值高于90%,相比于HMM、LSTM、BiLSTM模型,取得了良好的效果。Zhang等^[49]利用一种半监督嵌入式BERT-BiLSTM-CRF模型,以《中医诊断》为训练集和测试集,取得较好的抽取效果,精确率召回率和F1值达到80%以上,并且基于《中医诊断》一书中的实体建立起了简明的辅助诊断系统,提供了知识检索查询服务。

3.3.3 民族医药文献

民族医药方面,何家欢等^[50]提出了基于BiLSTM-CRF深度学习模型对藏药药理命名实体识别,BiLSTM模型相比LSTM模型能获取更佳特征表示,实现了对文献中藏药药理实体的有效抽取,为藏医药文献的数

据挖掘提供了新方法。郑光敏等^[51]提出了一种先抽取关系和后抽取实体的自动抽取知识模型,BERT-BiLSTM-CRF知识抽取模型能够有效组合BERT和传统模型的优势,可以充分提取文本特征,还进一步利用句子中相邻标签的关联性获得了全局最优的标签序列,改善了实体识别性能,较其他模型方法各方取得了最优,能够高效抽取《中国民族药辞典》中的实体和关系。并且还利用此模型抽取的中国民族药知识实体和关系构建了知识图谱,基于此知识图谱实现了中国民族药知识可视化和智能问答。

3.4 小结

中医药信息的多种特点让实体抽取技术在中医药领域的应用面临更大的挑战。目前,在中医药领域中的医案和临床电子病历、中医文献、和民族医药等多个方面都相应的运用了实体抽取技术进行处理信息,并且获得了一定的成果,在不断的改进发展下,实体抽取技术在中医药领域的应用,实现了对中医药信息的有效抽取,为进一步的中医数据挖掘打下了基础,有利于发现中医药的规律及隐藏知识,以便更好的应用中医药治疗人类疾病。

4 总结与展望

自然语言处理的实体抽取可以有效获取重要信息,信息抽取的数据形式由文本到图像、音频和视频等发展,面向更多的数据形式,便于多样的数据挖掘。中医药领域的复杂信息为实体抽出技术提供了大量数据,实体抽取也有效地处理了一部分中医药信息,两者之间得到了相互促进发展。实体抽取技术的应用,是对古代名医学术思想的继承挖掘和延续及对现代中医信息的进一步分析处理的重要基础。本文总结了实体抽取技术在中医药领域的应用情况,为中医药实体抽取的研究提供一定的参考,以期促进中医药的实体抽取研究。

本综述从中医药领域中实体抽取的实际研究情况出发,介绍了各技术的发展及中医药领域的应用情况。中医信息的抽取方法以及抽取模型,呈现出越来越多元化的研究发展趋势,对于中医领域的实体抽取的后续研究,在数据方面,需要足够大的、适合训练的数据量,并且需要更加精确具体的标注标签,在模型方面,要尽可能多方融合考虑各方法的优劣性,集成应用,力图使抽取结果更优化。抽取模型要尽可能提

高其适用性、鲁棒性、泛化性、并且能与中医领域的信息契合,以达到不断优化的抽取方法的目的,以便进行下一步的数据研究处理。

精准的获取利用医疗信息,能够有效促进医学的进步发展,造福于人类健康。在当前中医发展的最佳

时期,对中医领域包含的海量中医知识和学术思想等信息进行数据处理是重中之重,自然语言处理各种技术的出现与革新加速了中医领域实体抽取进程,为中医药的传承发展做出了极大的贡献,中医药资源也会在技术不断进步中得到更好的挖掘与研究。

参考文献

- 郭喜跃,何婷婷.信息抽取研究综述.计算机科学,2015,42(2):14-17.
- 李凡姝,姚登峰.自然语言处理中的文本表示和语言模型综述.中国计算机用户协会网络应用分会2020年第二十四届网络新技术与应用年会论文集.北京,2020:169-172.
- Pons E, Braun L M M, Hunink M G M, et al. Natural language processing in radiology: A systematic review. Radiology, 2016, 279(2): 329-343.
- 孙镇,王惠临.命名实体识别研究进展综述.现代图书情报技术,2010,(6):42-47.
- Rau L F. Extracting company names from text [1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application. Miami Beach, FL, USA. IEEE: 29-32.
- 袁清波,杨帆.命名实体识别研究综述.现代计算机,2021,(11):74-78.
- 宋柔.基于语料库和规则库的人名识别方法.北京:北京语言学院出版社,1993.
- 赵军.命名实体识别、排歧和跨语言关联.中文信息学报,2009,23(2):3-17.
- 陈德华,冯洁莹,乐嘉锦,等.中文病理文本的结构化处理方法研究.计算机科学,2016,43(10):272-276.
- 李嘉欣,王平.中文命名实体识别研究方法综述.计算机时代,2021,(4):18-21.
- 李想,魏小红,贾璐,等.基于条件随机场的农作物病虫害及农药命名实体识别.农业机械学报,2017,48(S1):178-185.
- 夏光辉,李军莲,阮学平.基于实体词典与机器学习的基因命名实体识别.医学信息学杂志,2015,36(12):54-60.
- 崔博文,金涛,王建民.自由文本电子病历信息抽取综述.计算机应用,2021,41(4):1055-1063.
- 杨锦锋,于秋滨,关毅,等.电子病历命名实体识别和实体关系抽取研究综述.自动化学报,2014,40(8):1537-1562.
- Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. P IEEE, 1989, 77(2): 257-286.
- Jaynes E T. Information theory and statistical mechanics[J]. Physical review, 1957, 106(4): 620.
- 赵伟,赵法兴,王东海,等.一种基于改进的最大熵模型的汉语词性自动标注的新方法.第二十三届中国数据库学术会议论文集(研究报告篇).广州,2006:185-189.
- 林亚平,刘云中,周顺先,陈治平,蔡立军.基于最大熵的隐马尔可夫模型文本信息抽取.电子学报,2005,33(2):236-240.
- Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273-297.
- 张敏.基于SVM和CRF的病症实体抽取方法研究.成都:成都理工大学硕士学位论文,2018.
- Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.
- 慕容伟波.基于GRU神经网络结合CRF的中文分词研究分析.广州:华南理工大学硕士学位论文,2018.
- 曹凯迪,施识帆,王忠民.基于CRF的入院记录中医院名称实体识别及应用.智慧健康,2019,5(32):5-7.
- 朱玲,朱彦,杨峰.基于中医疾病相关语义关系的正则表达式及知识抽取研究.世界科学技术-中医药现代化,2016,18(8):1241-1250.
- 冯云霞,衣鹏,韩正亮,等.基于CNN-BGRU-CRF的中文电子病历实体抽取方法.计算机与现代化,2020,(11):60-64.
- Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. 2013: arXiv: 1310.4546.
- 齐浩翔,马莉媛,朱翌民.基于Word2Vec的疫情虚假信息检测方法[J].智能计算机与应用,2021,11(10):134-138.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. NIPS. 2017.
- Kitaev N, Kaiser Ł, Levskaya A. Reformer: The efficient transformer. 2020: arXiv: 2001.04451.
- Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780.
- Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018: arXiv: 1810.04805..
- 翟剑锋.深度学习在自然语言处理中的应用.电脑编程技巧与维护,2013,(18):74-76.
- 刘博,杜建强,聂斌,等.基于二阶HMM的中医诊断古文词性标注.计算机工程,2017,43(7):211-216.
- 屈丹丹,杨涛,胡孔法.NLP在中医医案症状信息自动化抽取中的应用研究.软件导刊,2021,20(2):44-48.
- 肖瑞,胡冯菊,裴卫.基于BiLSTM-CRF的中医文本命名实体识别.世界科学技术-中医药现代化,2020,22(7):2504-2510.
- 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.中国中医药信息杂志,2021,28(5):20-24.
- Liu H L, Qin X N, Fu B. The symptoms and pathogenesis entity

- recognition of TCM medical records based on CRF2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops. Beijing, China. IEEE: 1479–1484.
- 38 Yang Y L, Zhong X Y, Xu L N. Research on entity recognition and knowledge graph construction based on TCM medical records. *J Artif Intell Res*, 2021, 4(1): 39–53.
 - 39 刘一斌, 叶辉, 易珺, 等. 基于朴素贝叶斯和 word2vec 的中医电子病历文本信息抽取. *世界科学技术-中医药现代化*, 2020, 22(10): 3563–3568.
 - 40 Jia Q, Zhang D Z, Xu H F, *et al.* Extraction of traditional Chinese medicine entity: Design of a novel span-level named entity recognition method with distant supervision. *JMIR Med Inform*, 2021, 9(6): e28219.
 - 41 叶辉, 姬东鸿. 基于多特征条件随机场的《金匱要略》症状药物信息抽取研究. *中国中医药图书情报杂志*, 2016, 40(5): 14–17.
 - 42 高甦, 金佩, 张德政. 基于深度学习的中医典籍命名实体识别研究. *情报工程*, 2019, 5(1): 113–123.
 - 43 张艺品, 关贝, 吕荫润, 等. 深度学习基础上的中医实体抽取方法研究. *医学信息学杂志*, 2019, 40(2): 58–63.
 - 44 北京科技大学. 一种面向中医古籍文献的命名实体识别方法和装置:CN201910340359.6[P]. 2019–10–11.
 - 45 北京科技大学. 基于中医古籍文献的中医命名实体识别方法及识别系统:CN201910367376.9[P]. 2019–08–16.
 - 46 Qu Q, Kan H, Wu Y, *et al.* Named Entity Recognition of TCM Text Based on Bert Model. 2020 7th International Forum on Electrical Engineering and Automation (IFEEA). IEEE, 2020: 652–655.
 - 47 Zhu W, Fu L, Xu L, *et al.* A TCM Diagnosis System Based on Textbook Information Extraction. 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing. *IEEE*, 2011: 483–487.
 - 48 Deng N, Fu H, Chen X. Named entity recognition of traditional Chinese medicine patents based on BiLSTM-CRF. *Wireless Wirel Commun Mob Com*, 2021, 2021: 6696205.
 - 49 Zhang M, Yang Z, Liu C, *et al.* Traditional Chinese Medicine knowledge Service based on Semi-Supervised BERT-BiLSTM-CRF Model. 2020 International Conference on Service Science (ICSS). *IEEE*, 2020: 64–69.
 - 50 何家欢, 刘勇国, 蒋羽, 等. 藏药药理命名实体识别. *医学信息学杂志*, 2020, 41(4):30–36.
 - 51 郑光敏, 易天源, 唐东昕, 等. 基于 BERT-BiLSTM-CRF 模型的中国民族药知识抽取. *武汉大学学报(理学版)*, 2021, 67(5): 10.

Summary of Entity Extraction and its Application in the Field of Traditional Chinese Medicine

Kong Jingjing, Yu Qi, Li Jinghua, Yu Tong, Zhang Zhulyu, Tian Ye, Zu Yaqi

(Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China)

Abstract: As machine learning, deep learning, and the emergence and development of artificial intelligence, natural language processing technology also constantly innovation, accordingly the entity extraction as child tasks of natural language processing, since its realized from extraction based on the pattern of matching to the machine learning method based on the traditional extraction, and based on the progress of the extraction of deep learning and development. In medical information extraction, natural language processing technology has also been well applied, and medical information has been processed and mined in a deeper level. In the face of flexible, complex and massive TCM information in the field of TCM, entity extraction technology can be used to extract more important information in the field of TCM, which is also a key and basic step for further TCM information mining. This paper summarizes the application and research of entity extraction technology in various fields of traditional Chinese medicine, providing reference for the further application and development of entity extraction technology in the field of traditional Chinese medicine, so as to promote the inheritance and development of traditional Chinese medicine and uphold innovation.

Keywords: Traditional Chinese medicine, Named entity recognition, Entity extraction, Chinese medicine application, Summary

(责任编辑: 刘玥辰, 审稿人: 王瑀、张志华)