

● 多学科研究 ●

引用:邓宇,张振铭,陈橙,匡浩铭,周琦人,杨张琪. 基于正则表达式的中医医案术语抽取方法研究[J]. 湖南中医杂志, 2023,39(5):202-207.

基于正则表达式的中医医案术语抽取方法研究

邓宇¹,张振铭¹,陈橙¹,匡浩铭²,周琦人¹,杨张琪²

(1. 湖南省中医药研究院,湖南 长沙,410006;

2. 湖南中医药大学,湖南 长沙,410208)

[摘要] 目的:探索从中医医案文本中识别出领域具有较高挖掘价值的术语实体,并根据其属性分类自动抽取,最终形成可直接利用和管理的结构化标准数据。方法:构建中医领域术语词典,采用 Python 语言编写正则表达式中文词语匹配算法和语义分析模块,通过词典中的术语与医案文本进行匹配和语法分析,实现中医术语的自动识别和分类抽取。结果:对 300 例中医肿瘤医案进行抽取测试,最终抽取术语 22540 个。经人工核对,发现抽取遗漏或错误术语 1352 个,约占抽取总术语的 6%。结论:现阶段的研究只是对中医医案术语抽取方法进行初步探索,今后的研究可在此基础上对医案术语词典作进一步补充,完善语义分析模块判断能力,使其能够更好地适应中医医案书写规则,为基于中医医案的文本大数据挖掘与利用奠定基础。

[关键词] 中医医案;知识抽取;Python;正则表达式

[中图分类号] R249, TP311.13 **[文献标识码]** A DOI:10.16808/j.cnki.issn1003-7705.2023.05.045

信息抽取(Information Extraction, IE)是一种能从海量数据中准确、快速抽取出人们感兴趣的事实信息,并将其以结构化形式存储起来的技术^[1]。国内信息抽取研究发展主要集中在近 20 年,以改进国外方法为基础,形成了适用于中文的信息抽取技术^[2]。中医术语是中医药领域内利用特定语言或文字来表达或限定专业概念的约定性符号^[3]。中医学术语具有特定属性且集中体现和承载了中医学领域的核心知识,因此,完善中医学术语识别和抽取尤为重要。从本世纪开始在中医药信息领域中引入了“行业术语标准”和“本体技术”,对该领域的概念和术语体系进行了系统性的表达^[4]。本体技术为文本信息抽取提供了领域背景知识,可将领域本体与文法分析技术相结合,从而改进文本信息抽取的效果^[5]。

中医医案作为历代中医医家临床诊疗经验的记录,蕴藏着历代医家的学术思想和辨证论治经验。然而中医医案是典型的医学术语密集型文本,其文献数量众多,文字记载复杂,行文风格各异,其保存也仅仅是简单的文档形式,且人工整理能力有限而机器识别度低。与一般的自然语言文本不同,中医医案文本是诊疗过程的真实记录,其遵循了一种半结构化的文本组织方式,例如:医案中常以“姓+左(右)”的模式来表示一个患者及其姓名、性别,如“张三,男”,表示一个姓张的男性患者;以“部位+描述”来表示一个症状,如“头痛”“腹胀”,其中“部位”已省略;以脏、腑、气、血、津液与八纲相互组合表示一个证型,如“气滞血瘀”;以“名词+数词+量词”的形式表示处方用药,如“桂枝 8 g”等。与此同时,在医案的复诊记录中,还常出现一些代表否

基金项目:湖南省中医药科研计划项目(2021231);长沙市自然科学基金项目(kq2202360、kq2202361);湖南省中医药研究院院级重点课题(202004)

第一作者:邓宇,男,高级工程师,研究方向:中医药信息与大数据

定和改善语义的词汇,如“未见失眠”“头痛改善”“去桂枝改黄芪 8 g”等^[6]。中医医案文献的这些特点,为正则表达式的应用提供了条件。因此,构建合适的正则表达式,可以快速、准确地从自然语言书写的医案文本中提取具有研究价值的术语信息。

本研究提出一种基于现有中医药本体知识库构建“中医医案语料库”,采用正则表达式的中文词语匹配算法,通过 Python 语言构建语义判断和术语抽取模型,对自然语言书写的中医医案文本进行术语识别和分类抽取,最终形成可分析、利用的结构化数据方法。

1 术语标准语料库的构建

1.1 资料来源 术语作为专业领域语料数据的重

表 1 主要语料库构建参考资料

标准名称	标准编号	发布日期	实施日期
中医舌象诊断信息分类与代码	T/CIATCM010—2019	2019 年 03 月 20 日	2019 年 05 月 01 日
中医脉象诊断信息分类与代码	T/CIATCM011—2019	2019 年 05 月 01 日	2019 年 03 月 20 日
中医临床基本症状信息分类与代码	T/CIATCM020—2019	2019 年 03 月 20 日	2019 年 05 月 01 日
中医证证分类与代码(疾病部分、证候部分、治法部分)	GB/T 15657-1995	1995 年 07 月 25 日	1996 年 01 月 01 日

1.2 语料库的类型 中医医案文献语料库按其功能分为术语词库和语法规则库,术语词库用于精准匹配医案文本中的术语词汇,主要包括中医病名、病因病机、临床症状、脉象、舌象、穴位、中医证型、中医治法、方剂、饮片 10 种类型。语法规则库用于识别医案文本中常用于表述特殊意义的词汇,主要包括否定词、改善词、饮片加减词、方剂类型词 4 种类型。

1.3 语料库构建流程 语料库的构建主要分为 3 个步骤,第一步:对收集到的标准文献进行数字化扫描或 ORC 文字识别,使其成为能够编辑的文本信息;第二步:语料的切词及标记,利用特征词定位或特殊符号标记作为文本切词标记的前期处理,在此基础上与北京大学计算语言所开发的通用领域语料的分词与词性标注工具相结合,初步形成结构化的 EXCEL 格式文件。与以往的人工标引工作相比,采用的“人机结合”方法较为客观,避免了因人而异产生的歧义,而且极大地减少了工作量。但受技术水平的制约,对误切分和误标注的部分,仍要采取人工方式进行审核;第三步:依据正则表达式的规则,形成术语的元字符集合。正则表达式定义一些特殊元字符,例如:“^”“\$”“*”“+”“?”“{”

要特征,对于命名实体识别有着重要支撑作用,构建领域术语词典是保证术语识别准确的重要手段之一^[7]。中医医案文本中涉及大量医学术语,而现有计算机中文分词系统对领域术语词汇的分割较容易产生歧义,比如方剂“麻黄桂枝汤”,其中“麻黄”“桂枝”“汤”都可分割为单独的词语,而此类词汇分隔歧义在中医医案中广泛存在。中医医案语料库从类型上属于专用性语料库,在语料的收集和处理上,应注重语料的专业性、系统性、合理性和代表性,尽可能全面地囊括所有的中医学学科门类。因此,本研究以现有的国际标准、国家标准及行业规范作为语料库构建的基础性资料,主要语料库构建参考资料见表 1。

“|”“\”“|”“()”,这些元字符不直接参与匹配,而是具有特殊语义,比如控制字符匹配的长度、类型、分组规则等^[8]。本文构建的语料库采用元字符“|”来分割单个术语词,例如“疏肝理气|疏肝健脾|疏肝明目”,“|”表示匹配条件是“或”的关系,且必须完全符合其中任何一个条件才算匹配成功。

1.4 语料库的规模 从理论上讲,如果语料库规模过小,覆盖率太低,不足以涵盖行文风格各异的中医医案文本所涉及的术语词汇,将导致术语识别的遗漏和语义规则判断错误。而语料库规模过大,又会影响程序匹配的精准性和运算效率。因此,前期语料库的规模不宜过于庞大,应在保障常用术语词汇识别准确率的基础上,通过医案文本的不断抽取测试,保持对语料库的长期补充和完善的过程。本次研究构建语料库术语词典分类及数据量统计情况见表 2。

表 2 语料库术语词典分类及数据量统计

实体类型	实体个数/个	实体类型	实体个数/个
中医病名	815	证型名称	4569
病因病机	805	治法名称	10856
脉象名称	304	症状名称	8096
舌苔名称	496	方剂名称	14705
穴位名称	393	饮片名称	1528

2 正则表达式匹配模型构建

2.1 医案文本切割 一个待抽取的医案文本往往记录了多条医案记录,而一个完整的个案的记录又以时间为主线,按先后顺序分别对病情变化、处方用药情况进行描述,在医案中常见于首诊、二诊、三诊、四诊等,且多诊记录中均会出现症状、处方、饮片等信息。在信息抽取之前,研究团队将这些文档进行简化,但又不丢失重要的信息,如此更有利于下一步的信息抽取。因此需要对记录多条医案的文本依据医案“个数”和诊疗“次数”进行“切割”,如计算每一个医案的开始和结束位置,以及每个个案中不同诊疗次数开始与结束的位置,从而保证抽取数据的完整性和多维性。

Python 的 re 模块提供了正则表达式匹配操作。其中 split() 方法可以使用正则模式分割一个字符串对象。因此,首先采用人工方法将记录多个医案的文本中用特殊符号在每个医案的起始部分进行标注,从而保证计算机对个案的准确区分。而个案诊疗次分割则需要依据医案文本书写风格,分别设定不同的诊次分割标签,具体文本切割主要代码如下:txt=open('YIAN.txt','r',encoding="utf-8"),表示将记录多条医案的文档 YIAN.txt 读入变量“txt”中;YIAN=str(txt.read()),表示将变量“txt”中的文本转换为字符串类型存入变量“YIAN”;pattern=r'@@@',表示将个案分割符号“@@@”读入变量“pattern”中;yianlist=re.split(pattern,YIAN),表示使用 re.split 函数,对文本按照个案进行分割,存入列表“yianlist”中;pattern=r'二诊|三诊|四诊|五诊|六诊|七诊',表示设置诊次切割标签;textlist=re.split(pattern,YIAN),表示将医案各诊次文本记录存入“textlist”列表中。

通过以上方法,实现了记录多条医案的文本,依据“个案”和“诊次”进行了文本切割,文本采用字符串的形式分别保存于列表中,其中“yianlist[0,1,2,3…]”依次保存完成个案文本,“textlist[0,1,2,3…]”依次保存个案的诊次文本。

2.2 正则表达式匹配模式 正则表达式,又称规则表达式(Regular Expression),是一个用正则符号写出的公式,程序对这个公式进行语法分析,建立一个语法分析树,再根据这个分析树结合正则表达式的引擎生成执行程序(这个执行程序我们把它称

作状态机,也叫状态自动机),可用于字符匹配^[9]。正则表达式引擎是一套核心算法,用于建立状态机,主要包括 DFA(确定有限状态自动机)和 NFA(非确定有限状态自动机)两种模式。本研究采用 Python 语言编写抽取程序,Python 自 1.5 版本起增加了 re 模块,可以直接调用 re.findall() 函数实现基于正则表达式匹配算法,该函数采用了 POSIX NFA 引擎的正则表达式,相较传统 NFA 引擎,它的特点主要是提供了“左侧最长匹配”模式,即遍历目标文档中所有字符串,并将最长匹配结果以列表的形式返回。例如正则表达式“疏肝理气|疏肝|理气”,在目标文档中匹配到“疏肝”还会继续匹配后续字符,一旦出现“疏肝理气”会优先匹配成功,进而丢弃之前或之后匹配到的“疏肝”和“理气”。以正则表达式 NFA 引擎在目标字符串“aababc”中匹配字符“abc|ab”过程为例,经过 15 个状态转换后,正则表达式完成全部匹配过程,成功匹配到“abc”和“ab”两个字符,按照“左侧最长匹配”规则,丢弃匹配到的字符“ab”,返回结果为“abc”,其过程见图 1、表 3。

2.3 正则表达式抽取模型设计 抽取模型设计是利用了正则表达式能够快速地从字符集合中搜寻特定字符串的机制,该模型通过将一系列的语料库与文本切割后的 yianlist、textlist 目标对象进行比较,判断目标对象中是否包含语料库中的各类术语词汇,识别并分类抽取字符串的过程。以在单个医案首诊记录中识别并抽取“证型”术语为例,其代码设计如下:1)读取语料库中证型术语词集合,以字符串的形式存入对应的变量中。代码示例:ZXkeyread=open('D:\ZX_key.txt','r',encoding="utf-8"),表示读取证型语料库;ZX_key=ZXkeyread.read(),表示将证型术语存入“ZX_key”变量中;2)使用 re.findall 函数,将证型术语集合变量与含医案首诊记录文本的列表进行匹配,再将匹配到的术语词去重后,存入证型抽取结果列表中。代码示例:zhengxing=re.findall(ZX_key,textlist[0]),表示匹配证型术语,存入“zhengxin”变量中;FT_zx=list(set(zhengxing)),表示证型术语去重后,存入“FT_zx”列表中。通过以上步骤,实现了对医案首诊记录中的证型术语精准匹配,并存入对应列表之中。通过 print(“首诊记录提取到的证型有:”,FT_zx)查看列表数据,显示如下:首诊记录提取到的证型有:['肝胆湿热','肝郁脾虚']。

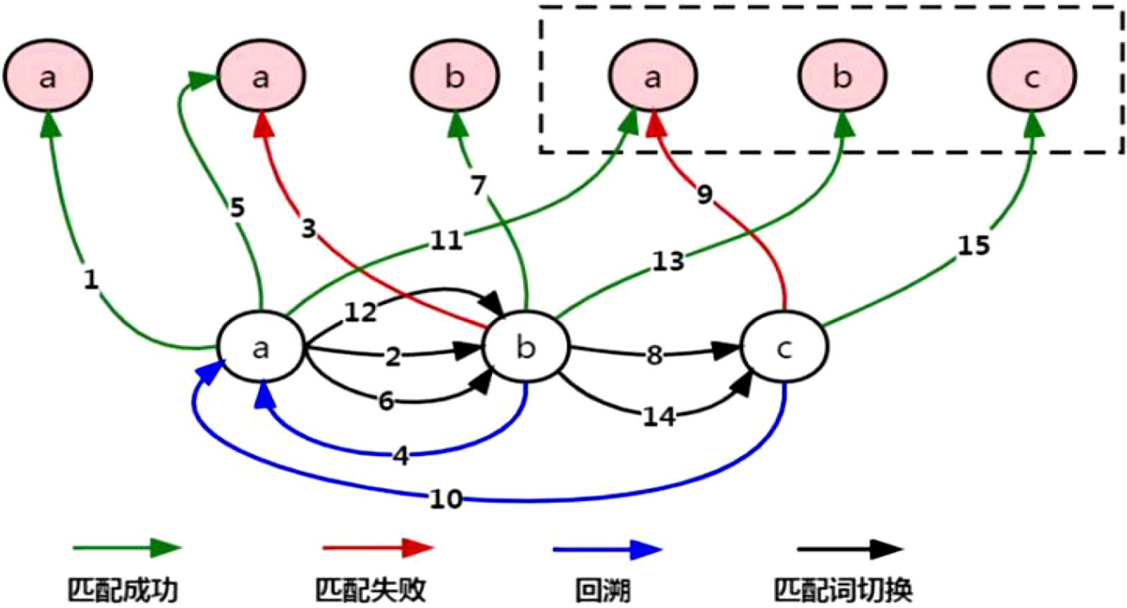


图 1 正则表达式匹配步态图

表 3 NFC 引擎匹配步态说明

状态号	步态注释	结果
1	正则表达式第 1 个字符“a”匹配目标字符串第 1 个字符“a”	成功
2	正则表达式跳转至第 2 个字符“b”	匹配词下移
3	正则表达式第 2 个字符“b”匹配目标字符串第 2 个字符“a”	失败
4	正则表达式回溯至第一个字符“a”	回溯
5	正则表达式第 1 个字符“a”匹配目标字符串第 2 个字符“a”	成功
6	正则表达式跳转至第 2 个字符“b”	匹配词下移
7	正则表达式第 2 个字符“b”匹配目标字符串第 3 个字符“b”	成功
8	正则表达式跳转至第 3 个字符“c”	匹配词下移
9	正则表达式第 3 个字符“c”匹配目标字符串第 4 个字符“a”	失败
10	正则表达式回溯至第一个字符“a”	回溯
11	正则表达式第 1 个字符“a”匹配目标字符串第 4 个字符“a”	成功
12	正则表达式跳转至第 2 个字符“b”	匹配词下移
13	正则表达式第 2 个字符“b”匹配目标字符串第 5 个字符“b”	成功
14	正则表达式跳转至第 3 个字符“c”	匹配词下移
15	正则表达式第 3 个字符“c”匹配目标字符串第 6 个字符“c”	成功

2.4 正则表达式语义判断模型设计 术语在医案中出现并不意味着对于术语概念所表达意义的肯定,特别是在复诊描述中常出现否定症状、改善症状、饮片加减等情况,其次中医方剂名称中常含有其他术语名称,例如“消肿活血汤”“麻黄桂枝汤”,直接基于语料库匹配进行抽取,会匹配到治法“消肿”“活血”和饮片“麻黄”“桂枝”,从而产生较多歧义。因此,本文收集了 500 例不同病种的中医医案总结出描述否定、改善以及饮片加减、方剂类型的常用词汇,构建语义后控词表(见表 4)。

同时本研究发现有部分否定词存在于术语内部,例如“无常”“无神”“无寒”“不寐”“不痛”等,以及一些非术语类固定词汇。例如“不论”“未来”等都会被误判为否定词。因此,直接利用后控词表来判断否定和改善语义,仍然会出现错误。基于以上情况,本研究构建的语义判断框架算法设定如下规则:①文档中对于症状的否定和改善以及饮片加减、方剂类型描述都可以归类到表 4 中;②否定词的作用范围不超出依据设定符号切割的子字符串;③增加饮片均带有数字表示的单位剂量。根据以上规则,构建后控词表语义判断模型规则(见图 2)。

具体步骤为:① 为了缩减语义判断的作用范围,在对医案文本依据诊次切割的基础上,再次依据逗号、句号、冒号对复诊字符串进行切割,形成描述相对单一的子字符串;② 通过正则表达式匹配术语词,以子字符串为边界,依次将匹配到的术语词替换为数字标号;③ 通过正则表达式匹配后控词,依据后控词对子字符串进行分类。(注:此时子字符串中的术语词已替换为数字标号,术语词中的“无、不、未……”不再干扰后控词的匹配;④ 依据子字符串的属性,将匹配到的术语词归类,对不含后控词子字符串中的术语词统一按照阳性症状归类,如“偶有心烦失眠”和舌脉信息等均归类为阳性症状;⑤ 可根据实际抽取需求,以单个术语词为单位进行抽取,也可以切割后的子字符串为单位进行抽取。

表 4 中医医案语义后控词表

否定词	改善词	饮片加减词	方剂类型词
未扪及 没有 未见 未述 否认 未闻 正常 不伴 排除 均可 无 未 不	减少 已减 解除 好转 康复 转佳 改善 稍增 痊愈 恢复 缓解 消失 已无 均无 消退 减退 缩小 转清 通畅 稳定 平稳 增加 自如 正常 成形 减轻 顺畅 停止 通利 通调 便调 尚可 寐可 已通 略减 未再 旋退 已能 已不 舒适 稍减 止	上方加 原方加 方中加 前方加 加上方去 上方减 方中去 原方去 去	汤 散 丸 剂 丹 方 饮 煎 酏

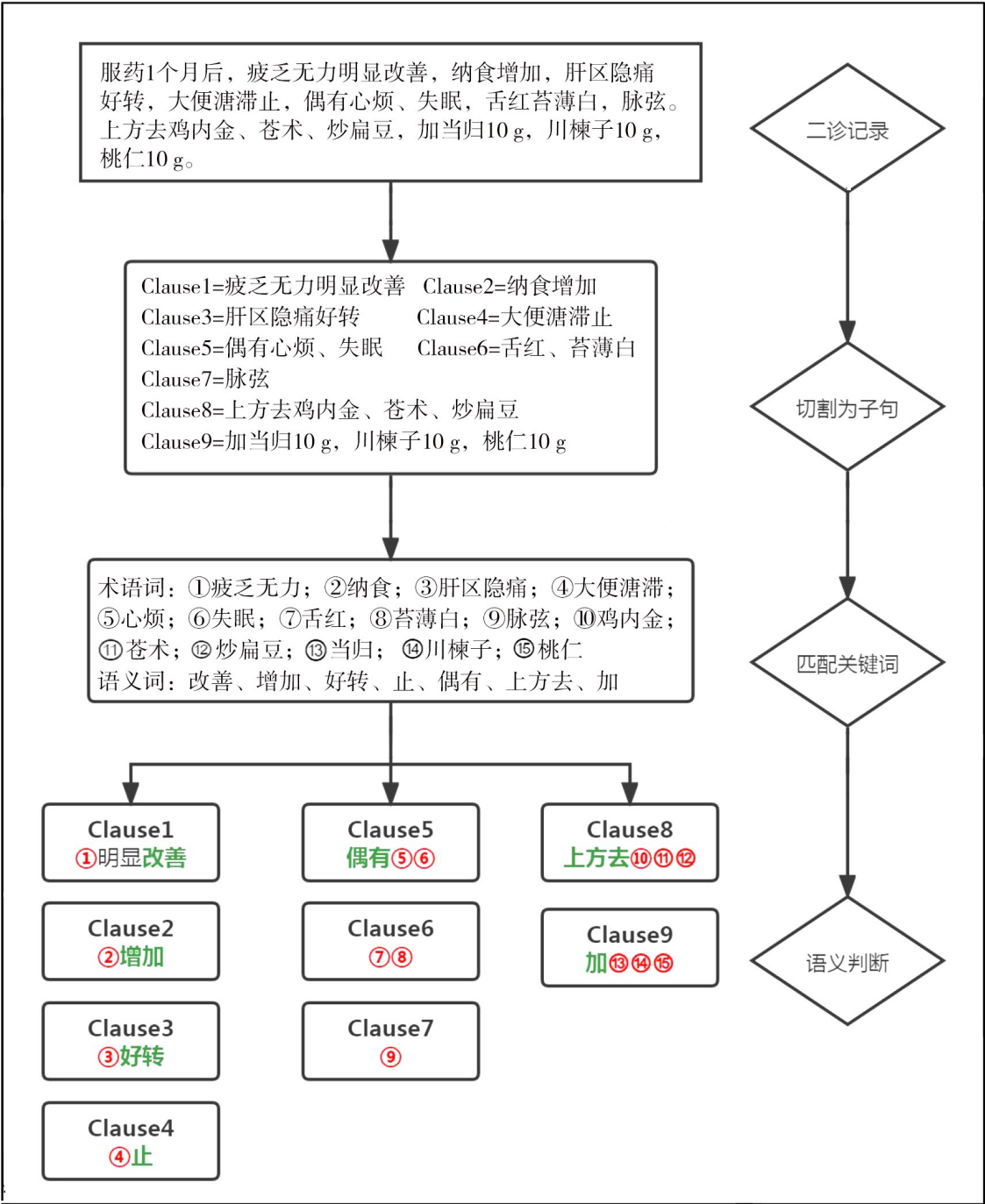


图 2 后控词表语义判断模型规则示意图

3 结 果

本研究选择中医肿瘤医案 300 例,共计 216600 字,利用本文构建的抽取系统进行抽取测试,最终抽取术语 22540 个。经人工核对,发现抽

取遗漏或错误术语 1352 个,约占抽取总术语量的 6%,错误原因主要涉及语料库未收录术语词、病名和症状术语混淆、子句划分错误、原文标点符号标注不准确等情况,其错误类型及原因分析见表 5。

表 5 错误样本示例

医案复诊记录	错误分析
补气健脾,通腑止痛,解毒散结	“通腑止痛”未纳入治法词典,导致术语未识别
无尿急	“无尿”和“尿急”都属于症状词库中的术语,导致错误识别为“无尿”
原方去半夏,黄芪,茯苓。患者无胸痛,胸闷等不适	原文标点符号使用不规范,导致语义判断错误
气促、咳嗽、恶心干呕、纳食少	“咳嗽”术语既是中医病名,又是症状词,导致“咳嗽”识别为中医病名
服上方半个月,头痛未见明显缓解	否定词“未见”和改善词“缓解”出现在同一个子句中,导致语义判断错误
较前(2009 年 5 月 4 日)大小稍缩小	症状描述不完整,导致抽取错误
犀角(已禁用,现用水牛角代)6 g(冲服)	语义判断错误,导致“犀角”和“水牛角”两味药同时提取
无明显诱因出现咳嗽、咳血	“无”和阳性症状出现在同一子句中,导致语义判断错误

4 小 结

本研究发现正则表达式可提供一种便捷的词语匹配方法,在最大匹配算法的前提下能够快速地从较长的医案文本中匹配到对应的术语词;其次现代中医医案书写趋于规范,所用术语词绝大部分能够归纳至标准语料库之中。但是也可以看到,基于语料库的抽取方法,需要构建能涵盖医案文本书写所应用的全部术语词汇。因此,语料库需要在不断地抽取测试,持续进行更新和完善。同时,无语义分析模块进行判断,会导致抽取出现较多错误,而语义分析和判断,需要根据所抽取的医案书写风格进行调整。本文仅通过正则表达式对中医医案术语抽取方法进行了初步探索,今后的研究可在此基础上对语料库作进一步补充,完善语义分析模块判断能力,使其能够更好地适应中医医案书写规则,为基于中医医案的大数据挖掘与利用奠定基础。

参考文献

[1] 周玉新. 信息抽取研究与发展综述[J]. 才智,2016

(27):262.
[2] 王曦,方廷钰. 现代化中医药术语英译的时代差异——基于近 30 年辞典语料库研究[J]. 上海翻译,2019(5):27-32,94.
[3] 崔艺箫,翟兴,钱亦昀,等. 中医术语抽取方法研究进展[J]. 医学信息学杂志,2021,42(4):30-34,56.
[4] 于彤,崔蒙,李敬华,等. 中医药本体工程研究现状[J]. 中国中医药信息杂志,2013,20(7):110-112.
[5] 丁晟春,刘逵逸,熊霞,等. 基于领域本体和语块分析的信息抽取的研究与实现[J]. 情报学报,2010,29(1):53-58.
[6] 王志飞,李晓君,郭霞珍,等. 正则表达式在中医文献研究中的应用初探[J]. 中国中医药信息杂志,2010,17(3):98-99.
[7] 周宏森,杨琴,朱懿敏. 基于语义 Web 技术的中医药术语知识库词典系统的研究[J]. 中医临床研究,2020,12(20):123-127.
[8] 徐荣飞. Python 正则表达式研究[J]. 电脑编程技巧与维护. 2015(9):45,49.
[9] 陈志宇. 基于优化 NFA 正则匹配的 Web 信息抽取技术及应用研究[D]. 杭州:杭州电子科技大学,2015.

(收稿日期:2022-11-20)

[编辑:徐霜俐]