

中医术语抽取方法研究进展^{*}

崔艺箫 翟 兴 钱亦昀 陈 悦 杨 颖 王 丽

(北京中医药大学 北京 100029)

〔摘要〕 介绍中医术语定义、特性、与一般术语的差异以及中医术语抽取研究现状, 阐述中医术语抽取方法及抽取结果评价方法, 对相关研究进行展望。

〔关键词〕 中医术语; 术语抽取; 研究进展

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.04.006

Study Progress of Traditional Chinese Medicine Term Extraction Methods CUI Yixiao, ZHAI Xing, QIAN Yiyun, CHEN Yue, YANG Ying, WANG Li, Beijing University of Chinese Medicine, Beijing 100029, China

〔Abstract〕 The paper introduces the definition, characteristics and differences between Traditional Chinese Medicine (TCM) terms and general terms and the study status of TCM term extraction, expounds the extraction methods of TCM terms and the evaluation methods of extraction results, and prospects the related study.

〔Keywords〕 Traditional Chinese Medicine (TCM) term; term extraction; study progress

1 引言

随着中医药国际化趋势加强, 中医药术语抽取与标准化工作重要性不断凸显^[1]。如何简洁、快速地筛选中医药专业术语并广泛应用于中医药文化推广、中医临床规范化诊疗、中医数字化、中医药翻译、中医信息检索等众多领域已成为中医药文化传播与发展

研究的重要方向^[2]。中医药理论起源和发展受中国古代哲学和历史因素影响, 其语言表述具有一定程度抽象性与模糊化特征, 且少部分用语具有文学色彩, 导致中医药术语难以被高效识别与区分^[3-4], 为中医药术语提取和区分带来较大困难。同时随着电子病历推行, 病案书写规范要求越来越高, 需要规范而标准的专业术语作为支撑。因此提高中医术语识别和区分水平是实现中医药文化继承和传播的基础^[5]。与其他术语识别领域相比中医术语识别研究较少, 本文针对中医术语抽取研究现状进行分析并对未来研究进行展望, 以期对相关研究提供参考。

2 术语定义与特性

2.1 定义

术语通常被称为“专有名词”, 但术语并不只由单独的名词或复合名词组成, 其可能是包含形容词、介词、名词、动词等的词组。术语是一类随着

〔收稿日期〕 2020-06-15

〔作者简介〕 崔艺箫, 本科生; 通讯作者: 王丽, 硕士, 讲师。

〔基金项目〕 国家自然科学基金青年科学基金项目“‘异病同治’视角下基于文本挖掘和复杂网络的证候生物学网络建模研究”(项目编号: 81603499); 中央高校基本科研业务费专项“基于网络药理学的经方活性成分筛选与作用靶点分析—以海藻玉壶汤为例”(项目编号: 2019-JYB-JS-031)。

各学科不断发展而产生的具有特定含义、专业知识性的概念^[6]。目前其定义尚未统一,各国语言学家与术语学家提出多种观点^[7]。例如《术语工作原则与方法》提出“术语是专业领域中概念的语言指称”^[8];冯志伟将术语定义为“通过语音或文字来表达或限定专业概念的约定性符号”^[9];Sager 认为“术语是概念的语言表征”。中医术语是中医药领域内利用特定语音或文字来表达或限定专业概念的约定性符号。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抽取和识别尤为重要^[10]。随着各学科交流不断加深,许多词汇被赋予新的含义。因此识别、抽取新术语成为各学科发展的必要条件。中医学文献具有数量较多、时间跨度较大等特点,高效识别中医药术语成为推动中医现代化的必由之路^[11-12]。

2.2 特性

术语是符合本民族语言习惯且和其所表达的特定领域具有较大相关性的一组词语。Kageura 将术语的特性归纳为单元性和领域性。单元性是指术语必须具有相对稳定的结构且能够独立表达完整含义;领域性是指数据在领域内的相关度^[13]。《术语工作 原则与方法》GB/T 10112-1999 对术语所具备特性进行了明确要求。(1)单名单义性。新术语是否具有相同含义的同义词,如果之前领域内无同义词,则可以命名新术语,存在同义词的情况下,应比较之后确定一个满足要求的术语。(2)顾名思义性。术语能够直接明了地理解其含义,不应难以理解。(3)简明性。术语应当简单方便使用,含义明确。(4)派生性。术语能够方便衍生,即在领域内扩展延伸词义。(5)稳定性。专业领域内使用较为频繁,认可度高,术语具有相对稳定性。(6)合乎本民族语言习惯。术语不应带有浓厚的感情色彩,应当中性^[14]。因此术语不仅具有特定的含义而且具有相对的专业性。

2.3 中医术语与一般术语的差异

中医术语具备一般术语的特性^[15],同时在实际应用中表现独有特点,造成这一现象的原因主要有:古今语法结构具有较大差异,古汉语特有的表

达方式易造成分词出现错误,部分术语不符合现代语言习惯;字词含义随时间改变而发生变化且具有多词同义和一词多义情况;有些中医名词含义过于抽象,大量借用其他学科词汇,所表达内涵与外延并不明确。例如在中医针灸学中,“扶突穴”又名“水穴”,“天窗穴”又名“窗笼穴”“窗聾穴”“天笼穴”。类似现象在中医药学领域普遍存在,而这些特性导致中医术语的识别与抽取与一般术语存在差异,具有复杂性。

3 术语抽取方法

3.1 概述

3.1.1 自动抽取方法和模型建立 通过建立自动抽取方法或模型可实现对文本或者资料中术语的识别、自动提取,从而完成术语获取。其中自动抽取方法和模型建立对于术语提取较为关键。信息抽取是指从给定文档集合中自动识别预先设定的实体、关系和事件等类型信息,对其进行结构化储存和管理的过程。信息抽取以分词和标准为基础。首先确立词性搭配模板进而自动抽取候选术语,然后对术语进行联合分析,从而进行合理方法分词和词性标准化。

3.1.2 主要方法 目前术语的抽取方法主要分为人工手动抽取和计算机自动抽取。手动抽取是依靠人工从大量文献中采集候选词,后经该领域专家辨别和筛选得到术语列表^[16]。计算机自动抽取最早由国外研究者提出,其研究始自 20 世纪 90 年代。国内自动术语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术语抽取技术^[17]。但中文术语自动抽取较拉丁语系术语更为复杂。以英文为例,英语词汇之间是有空格分隔的,而中文词汇之间没有分隔方式。因此中文术语在筛选前通常要进行分词^[18]。

3.2 中医术语分词方法

3.2.1 内涵 中医术语分词是指从中医药文献资料中识别中医药特有的且具有较高认可度的病名、症状、诊断方法、治疗方式、药品名称、解剖结构等术语实体,根据其词性进行分类^[19]。分词过程中

可以使用词性标注算法对词性进行标注。词性标注算法是指运用电子计算机,对分词后的每个词进行词性标注并根据不同词性筛选术语的过程。

3.2.2 中医术语分词研究难点 一是中文是黏合性语言,所有字符都是连接在一起的,而不像拉丁语系词汇之间有区隔。在对文本处理之前需先进行分词,而对新词的识别是分词难点。二是中文表述方式具有结构复杂多变、语法约束较弱的特点,导致中文术语具有多种表达结构。三是中文词语往往具有一词多义的特点,某些词语的动词词性和名词词性含义不同^[19]。

3.3 术语自动抽取方法

3.3.1 基于规则的方法 通过确定术语词性及其词法模式等语言学规则来进行抽取。首先从已知术语库中归纳出术语构成规律,根据该领域内术语构成规律确定术语组成规则,进而进行抽取。例如 Justeson 在 1995 年提出 Terms 系统,其工作原理是首先对文档进行标注并消除歧义,然后根据该领域语法特点制作语法模式表并根据模式表进行抽取,此系统在英文光谱分析领域语料库的抽取准确率达到 98%^[20]。此方法成立的基础在于已有术语集和对一个领域内术语规则的总结,因此其通用性比较差。在面对不同领域或语言时容易出现规则不兼容问题,相同领域或语言也可能因为术语规则复杂而导致不兼容。因此目前对该术语提取方法单独研究较少。

3.3.2 基于统计的方法 以统计学为基础,利用术语在语料库内的分布统计结果识别术语的方法。使用的统计方法主要可分为领域性度量和单元性度量。领域性度量主要衡量词频、TF-IDF 值、DR 和 DC 值。单元性度量主要衡量互信性、对比数偶然性。该方法的特点是规则可以在多个不同学术领域内通用,无视语法、语义对统计的干扰。但如果想提高“基于统计的方法”统计出的结果质量需要大规模的语料库作为数据支撑,依赖术语出现的概率,在使用时易忽略使用频率较低的术语^[21]。

3.3.3 多种方法混合 指利用两种及以上方法进行分析,通常是将基于规则和统计的方法进行混合使用,或在这两种方法基础上加入其他方法进行使

用。该方法具有多样性特点,选取几种较为常见的方法进行分析。(1) 基于规则和统计的方法。其基本思路为首先根据语料库和现有术语制定语法规则,然后根据规则进行候选术语筛选,最后运用统计方法对候选术语进行二次筛选。此方法结合了规则和统计优势,可解决语言和领域差异性^[22]。(2) 基于机器学习的方法。减少规则和公式的制定,但需要对大量标注语料进行学习,在面对不同领域时需要重新输入语料。所以其通用性较差,尚需进行进一步试验^[23]。(3) 基于术语部件扩展的方法。对术语部件库依赖性较强。当术语不包含术语部件库时抽取工作无法展开。术语部件库的构建费时费力且通用性较差,自动构建术语部件库的方法目前尚无法大规模开展应用。因此该方法有待于观察提高^[24]。综上,多种方法的相互融合能够形成优势互补,有效地改良了单一方法造成的缺陷,在使用中有良好效果。因此多方法融合应用将成为趋势。

3.3.4 中医领域术语自动抽取方法 利用以上术语自动抽取方法,结合中医领域术语特点,可以形成适合中医领域的自动抽取方法。例如通过对中医领域术语种子集进行有限次迭代,生成中医针灸领域术语构件集;进而以中医领域术语构建集作为字典,采用最大向前匹配算法对其进行切分并抽取候选术语,利用语言规则对术语进行处理,以筛选出中医领域专业术语。

4 术语抽取结果评价

目前常见的术语抽取效果评价方法主要有 3 种,即准确率(Precision)、召回率(Recall)和 F 值(F-Measure 或 F-Score)^[25]。在中医学术语研究中,首先对抽取结果信息进行统计,再选择相应方法进行评价,见表 1。通过参照标准术语表^[26],将标准术语表设置为 ST,提取的术语集合为 T,被抽取的术语个数为 $a = |ST \cap T|$; 准确率为 $P = a / (a + b) = |ST \cap T| / |T|$; 召回率为 $R = a / (a + c) = |ST \cap T| / |ST|$; F 值综合考虑了准确率和召回率,计算公式为 $F = (a^2 + 1) PR / (a^2 P + R)$ 。a 为可调节参数, $a = 1$ 准确率和召回率占 t

比相同， $a=2$ 准确率高于召回率， $a=0.5$ 则召回率高于准确率。国外学者针对术语抽取提出了较科学合理的评价方法，能够更加全面地评价术语抽取^[27]：假设 L 为 T 以某种度量排序的有序表，将标准术语表设置为 ST ，在 L 中前 n ($1, 2, \dots, N$) 中真正的术语个数 $CT(n) = ST \cap \{L(1), L(2), \dots, L(n)\}$ ；准确率为 $P(n) = |CT(n)|/n$ ；召回率为 $|CT(n)|/|ST \cap T|$ 。

表 1 术语识别结果统计

项目	被识别并提取的	未被识别并提取的
术语	a	b
非术语	c	-

5 中医术语抽取研究现状

5.1 概述

目前在中医术语抽取方面已经取得了一定成果，但是术语命名不一、内涵和外延模糊、描述精度不一等问题，给中医术语提取增加一定难度，且目前国内对于中医术语抽取文献有限，因此相较于现代医学术语标准化和标准化规范统计等要求还有一定差距，仍有较大发展空间。

5.2 条件随机场模型应用相关研究

目前利用数据挖掘、机器学习和人工智能等技术对中医术语进行针对性抽取，以结构化形式储存在数据库中，可以高效地获取中医相关知识。条件随机场（Conditional Random Fields, CRF）模型可以有效地解决最大熵等模型普遍存在的标记偏置问题，如灵活地引入多种特征则抽取效果较好，是目前常用的中医术语抽取方法。王世昆等将 CRF 方法与最大熵和支持向量机 3 种统计方法运用到明清古医案症状、术语的自动识别标注研究中，经比较发现 CRF 模型对明清古医案中症状、术语类中医命名实体识别具有明显优势^[28]。赵翔介绍了条件随机场、最大熵和隐马尔科夫 3 种基于统计的命名实体识别方法，讨论了条件随机场模型相对于其他模型

的特点，采用 CRF 方法在网页中进行疾病名称提取^[29]。范岩提出 Bubble – bootstrapping 算法与 CRF 模型相结合进行中医药文献疾病实体识别，但该实验对模型进行了简化，没有采用丰富的特征集，选取数据规模较小，在一定程度上制约了方法性能^[30]。张五辈等将中医领域术语抽取看作一个序列标注问题，将中医领域术语分布特征量化作为训练特征，利用 CRF 工具包训练领域术语模型，在此基础上进行术语抽取，实验结果表明术语抽取准确率为 83.11%，召回率为 81.04%， F 值 82.06%^[31]。

5.3 其他术语抽取方法研究

孙水华等建立了一种基于规则的领域术语抽取算法模型，该模型首先对中医针灸领域术语种子集进行有限迭代，生成中医针灸领域术语构件集；然后以术语构件集为领域词典，采用最大向前匹配算法对中文针灸医学文献中的句子进行切分并抽取候选术语；最后利用语言规则对候选术语进行过滤处理，筛选出中医针灸领域专业术语。分别以关键字集和中医词典为种子集进行实验，测试的 F 值分别达到 76.96% 和 35.59%^[32]。刘成帅等在分析空间向量模型和词频的基础上提出了一种改进的 TFIDF 方法，以此方法进行领域文本选取，在分析常用领域术语提取方法和贝叶斯推理基础上，将贝叶斯推理引入到领域术语提取中，准确率达到 75.5%，召回率 83.1%， F 值为 79.1%^[33]。

6 研究展望

6.1 概述

现有中医术语抽取方法相较于早期术语抽取研究已经取得较大突破。目前国内已形成多套中医术语体系并广泛应用于中医药文化传播、临床标准制定、中医药信息检索等方面。但目前专业针对中医术语的自动抽取技术尚不成熟^[34]。面对中医药语言独特的表达方式现行术语抽取系统存在一定弊端，需进行进一步优化。

6.2 完善中医术语评价体系

中医术语抽取不仅需要抽取系统开发和完善，

更需要建立完善的评价体系。目前针对中医术语的评价主要采用准确率、召回率和 F 值等集中统计学指标,缺乏专门针对中医术语语义的评价体系。在评价指标建立后需要对中医语料库进行合理筛选并建立中医体系各自的术语库评价指标^[35]。

6.3 中医术语抽取方法改进

无论是基于规则的方法、基于统计的方法或多种方法混合,术语抽取技术都在不断发展。特别是人工智能和机器学习的应用为术语自动抽取提供了崭新的研究途径。中医术语自动抽取研究需要在现有方法基础上进行优化,借鉴其他学科术语抽取经验^[36-37]。

7 结语

随着大数据不断发展,中医学应用大数据技术不断实现方法和理论创新。而深入研究中医术语抽取方法已经成为中医大数据发展的基础。本文对术语及中医术语的定义、特性、方法进行介绍,总结说明了现有中医术语抽取经验,以期为中医学研究提供参考。推进中医学术语规范化使用应当加快中医术语评价体系的完善和抽取方法的改进,从而有效促进我国中医学快速发展。

参考文献

- 1 王琼,刘亮亮,张晓如,等. 基于模式自动获取中医临床症状术语 [J]. 中国数字医学, 2018, 13 (3): 44-46.
- 2 孟洪宇,孟庆刚. 基于条件随机场的中医术语抽取方法及其应用探析 [J]. 中华中医药学刊, 2014, 32 (10): 2334-2337.
- 3 董洋溢. 面向智能决策应用的本体关键技术研究 [D]. 西安:西北工业大学, 2018.
- 4 袁劲松,张小明,李舟军. 术语自动抽取方法研究综述 [J]. 计算机科学, 2015, 42 (8): 7-12.
- 5 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.
- 6 陈睿. 基于深度学习的专业领域术语识别系统设计与实现 [D]. 北京:北京邮电大学, 2019.
- 7 于清,常乐,徐健,等. 基于汉维医疗平行语料的双语

- 术语抽取研究 [J]. 内蒙古大学学报 (自然科学版), 2018, 49 (5): 528-533.
- 8 Kageura K, Umino B. Methods of Automatic Term Recognition: a review [J]. Terminology, 1996, 3 (2): 259-289.
- 9 冯志伟. 基于经验主义的语料库研究 [J]. 术语标准化与信息技术, 2007 (1): 29-36, 39.
- 10 Sager J C, Dungworth D, McDonald P F. English Special Languages: principles and practice in science and technology [M]. Wiesbaden: Brandstetter, 1980.
- 11 师光达. 基于可比较语料库的术语提取方法研究 [D]. 北京:北京化工大学, 2015.
- 12 李勇. 基于聚类方法对特定领域术语的自动筛选 [J]. 计算机工程与科学, 2008 (2): 64-66, 134.
- 13 王密平. 汉语专利术语抽取及应用研究 [D]. 南京:南京大学, 2017.
- 14 朱惠,王昊,苏新宁,等. 汉语领域术语非分类关系抽取方法研究 [J]. 情报学报, 2018, 37 (12): 1193-1203.
- 15 王文君. 翻译实践中术语的提取和翻译 [D]. 上海:上海外国语大学, 2014.
- 16 杨洋. 基于中医术语规范化的中医英语教学探讨 [J]. 陕西教育 (高教), 2019 (12): 11, 21.
- 17 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.
- 18 叶晓. 基于多个中医术语英译标准的治湿诸法英译辨析 [J]. 中国中医基础医学杂志, 2019, 25 (9): 1307-1310.
- 19 刘平,张清怡,方廷钰. 从西医学术语中译标准化谈中医术语英译标准化 [J]. 中医药导报, 2019, 25 (14): 135-137.
- 20 成福春,朱亮,周嘉,等. 《中医术语编码与术语服务平台基本功能规范》的制定及意义 [J]. 上海中医药杂志, 2019, 53 (2): 6-9.
- 21 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.
- 22 贾李蓉,刘静,刘丽红,等. 中医临床术语系统 v2.0 病症分类体系构建研究 [J]. 中国中医药图书情报杂志, 2018, 42 (5): 8-12.
- 23 付璐,李思,李明正,等. 以清代医籍为例探讨中医古籍分词规范标准 [J]. 中华中医药杂志, 2018, 33 (10): 4700-4705.
- 24 闻永毅,王治梅. 中医文献语料库建设与顶层设计刍议 [J]. 西部中医药, 2018, 31 (7): 62-65.

(下转第 56 页)

已累计服务 300 万人次，交易额超 10 亿元。

4.3 云险

覆盖医疗全过程全场景，建立涵盖自费、医保、商保等统一多渠道的创新支付体系，同时提供智能保险风控服务。借助“互联网+”移动支付已优化 1 000 余家医疗机构院内结算模式，投放支付设备 9 000 多台，在线支付累计交易金额达 300 亿元，日均交易笔数超 30 万笔，日均交易金额超 1 亿元，已与太平人寿、中国平安、太平洋保险、蚂蚁金服等 50 余家保险公司系统对接，合作开展商保理赔业务。

4.4 云康

构建线上线下一体化的居民保健+慢病管理+就医导医+体检服务，已与全国 800 余家公立医院、1 500 余家体检机构开展合作，实现体检资源整合。

5 结语

以在线医疗为纽带促进在线经济发展，通过优质资源整合，实现覆盖诊疗全流程的线上线下医疗

同质化服务以及集约化建设和云化部署，打造全产业链高质量协同发展的生态模式，不断催生新业态、新模式、新产业。

参考文献

- 1 李华才. 互联网医疗的热点与痛点 [J]. 中国数字医学, 2017, 12 (8): 1.
- 2 胡泊洋. 年复合增长率超 30%, 互联网医疗成为医药健康产业发展新方向 [EB/OL]. [2020-07-01]. <https://www.cn-healthcare.com/article/20200701/content-538826.html>.
- 3 孟群, 尹新, 梁宸. 中国“互联网+健康医疗”现状与发展综述 [J]. 中国卫生信息管理杂志, 2017, 14 (2): 110-118.
- 4 王强, 易应萍. 临床医疗大数据治理和应用 [J]. 医学信息学杂志, 2018, 39 (8): 2-6.
- 5 舒婷. “互联网+”时代的患者隐私保护 [J]. 中国数字医学, 2016, 11 (5): 45-47, 87.
- 6 张宇清. “互联网+医疗服务”体系下患者隐私权保护研究 [J]. 医学信息学杂志, 2019, 40 (1): 6-11.
- 7 王洁. “互联网+医疗”信息安全问题及对策分析 [J]. 医学信息学杂志, 2018, 39 (3): 44-47.
- 8 周辉, 杨阳, 李卫东. 区块链在医疗信息化领域的应用展望 [J]. 中国数字医学, 2018, 13 (4): 110-111.
- 25 Vivaldi J, Rodriguez H. Evaluation of Terms and Term Extraction Systems: a practical approach [J]. Terminology, 2007, 13 (2) 225-248.
- 26 Zheng Y, Dou W, Wu G, et al. Automated Chinese Domain Ontology Construction from Text Documents [M]. Berlin: Springer, 2007: 639-648.
- 27 Korkontzelos I, Klapaftis I P, Manandhar S. Reviewing and Evaluating Automatic Term Recognition Techniques [M]. Berlin: Springer, 2008: 248-259.
- 28 王世昆, 李绍滋, 陈彤生. 基于条件随机场的中医命名实体识别 [J]. 厦门大学学报 (自然科学版), 2009, 48 (3): 359-364.
- 29 赵翔. 基于 Web 挖掘的中医知识发现研究 [D]. 北京: 北京交通大学, 2010.
- 30 范岩. 基于条件随机场模型的中医文献知识发现方法研究 [D]. 北京: 北京交通大学, 2009.
- 31 张五辈. 中医药知识库设计与实现 [D]. 沈阳: 沈阳航空航天大学, 2011.
- 32 孙水华, 黄德根, 牛萍. 中医针灸领域术语自动抽取研究 [J]. 中文信息学报, 2016, 30 (3): 118-124.
- 33 刘成帅. 中文领域术语自动获取方法的研究 [D]. 南京: 南京邮电大学, 2011.
- 34 张盼, 毛树松, 邓文萍. 中医药信息标准术语规范化研究 [J]. 医学信息学杂志, 2018, 39 (5): 12-15.
- 35 赵永红, 万莉莉, 杨具荣, 等. 中医方剂学科各部分基本术语英译规范研究 [J]. 中华中医药杂志, 2018, 33 (11): 5085-5087.
- 36 曾凡, 吴泽扬. 谈中医药术语标准化 [J]. 安徽文学 (下半月), 2017 (8): 109-111.
- 37 叶其松. 术语学核心术语研究 [D]. 哈尔滨: 黑龙江大学, 2010.

(上接第 34 页)