ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.10, No.28, October 2014
E-mail: eduf@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

中文领域术语自动抽取方法进展研究

闫琪琪,张海军

(新疆师范大学 计算机科学技术学院,新疆 乌鲁木齐 830054)

摘要:论文梳理总结了目前术语自动抽取的研究现状,分析讨论了术语自动抽取的研究方法,通过对术语抽取方法剖析和比较,提出了目前研究中存在的问题和发展趋势,这对后续的中文领域术语自动抽取的研究具有一定的指导意义。

关键词:术语自动抽取;中文信息处理;研究现状

中图分类号:TP18 文献标识码:A 文章编号:1009-3044(2014)28-6716-03

The Research Progress of Chinese Domain-Term Automatic Extraction Methods

YAN Qi-qi, ZHANG Hai-jun

(School of Computer Science and Technology Xinjiang Normal University ,Urumqi 830054,China)

Abstract: This paper summarized current research status in ATE (Automatic Term Extraction) studies, analyzed and discussed the characteristics of term, as well as methods of ATE. Through the analyses and comparisons of term extraction methods, current problems existing in the researches and development tendency were presented, which has some significance to the further researches.

Key word: Automatic Term Extraction; Chinese Information Processing; Research Statu

DOI:10.14004/j.cnki.ckt.2014.0563

术语是人类智慧在语言中的结晶,它凝聚了领域知识的精髓。领域术语使用过程中,由于术语标准化工作没有及时对新产生术语进行规范化处理,导致各领域术语混乱,领域内部和领域间的科学交流困难重重。因此,开展术语库自动构建和术语规范化已迫在眉睫,利用计算机手段开展术语抽取和规范化工作已成为术语学研究和自然语言处理中的重要问题[1]。研究将从领域术语自动抽取方法、术语抽取研究中存在的问题及术语抽取研究发展趋势几个部分展开。

1 中文领域术语抽取研究现状及发展趋势

自动术语抽取是从特定的领域文本中抽取体现领域核心术语词汇的过程。目前中文术语抽取的研究中通常综合考虑术语的语言特征和术语领域特征。主要体现在候选术语提取和候选术语过滤阶段的工作中,术语抽取的一般流程如下图:

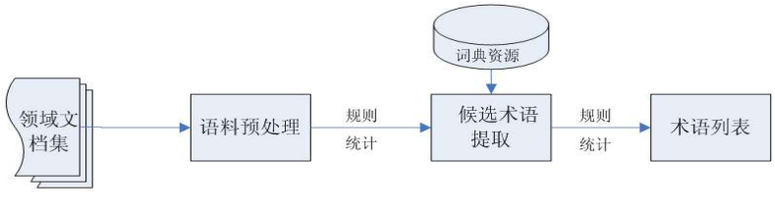

图1 术语自动抽取流程

术语抽取流程反映了术语抽取工作开展的一般步骤。选择合适的领域语料是开展术语抽取工作的必要条件;预处理的处理方式与术语抽取采用的方法有直接关系,主要是生语料的格式转化、去噪、分词及词性标注等;采用统计或规则的方法从语料中提取候选术语,通过统计参数或规则过滤候选术语以获取领域术语列表。候选术语提取阶的词典资源主要有,普通词语前缀后缀词典、停用词词典等。

1.1 领域术语自动抽取方法研究

1.1.1 基于词典与规则的方法

基于词典的方法就是利用现有术语资源定位术语在文本中的出现,一定程度上来说,术语本身就是术语最基本的语言特征,它本身包含着术语的最大信息。基于规则的方法就是利用术语语言特征进行术语抽取的方法。研究[2]借助一般词典和种子扩展

收稿日期:2014-08-16
基金项目:国家自然科学基金项目(NO.61163045,61263044);新疆维吾尔自治区自然科学基金(NO.2012211A057);新疆师范大学重点学科招标课题(NO.12XSXZ0601);新疆师范大学研究生创新金项目(NO.20131201)

方法自动识别单词术语,实验表明该方法是有效的,但召回率较低。研究[3]借助早期的语法过滤器,较之前研究使用的语法规则宽松,能够匹配更多不同语言结构的术语,提高了术语抽取的召回率,但降低了准确率。研究[4]运用正则表达式的字符串匹配功能对特定数据库中的术语实现抽取,证明了简化正则表达式规则能提高特定应用的需求匹配效率,研究将抽取效率提高1倍左右。

词典与规则的方法对特定领域和特定类型的术语抽取具有良好的效果。此类方法有准确率高、处理过程简单、计算量小等特点,但术语构词规则灵活、表达方式复杂且存在术语变体和领域新词等问题,致使术语的语言规则难以把握,术语规则库构造困难。目前国内外纯基于规则的术语抽取研究很少,它在术语抽取研究中多用于低频术语抽取和准确率提高。

1.1.2 基于统计的方法

基于统计的方法[5-7]以统计理论为基础,从概率意义上衡量多字单元是否为术语。术语的统计特征有两类,一是术语单元性即术语作为独立的语言单位具有稳定的语言结构;二是术语领域特性即测度词汇单元与特定领域之间的相关程度。

表1 典型统计方法比较

统计指标	方法名	优势	不足
术语单元性(Unithood)	互信息	计算候选术语字串内部结合强度过滤术语	针对多词术语此方法对分词结果具有较强的依赖性;无法识别字符串结合强度低于阈值的术语
	左右熵	估算候选术语中的词语左右熵过滤候选术语干扰碎片,有效实现干扰项过滤,提升术语抽取准确率	与互信息类似,使用左右熵时需要设置过滤阈值,无法识别熵值高于过滤阈值的术语
	LCS	能完整识别术语语义单元,避免未登录术语问题	无法识别候选术语集中的嵌套问题
术语领域特性(Termhood)	词频分布	统计词频分布信息获取术语,对低频属于和高频普通词串识别具有良好的效果	易受测试语料规模、词性标注效果的影响
	TF-IDF	充分考虑术语在领域文档集中的分布特征,能有效过滤在特定文件内高频但在整个文档集中分布低的词汇,能有效过滤常见词汇	实验文档的规模会影响到术语的召回
	DC+DR	领域相似度和领域一致度方法引入了支撑领域文档集,通过无关领域文档集过滤掉不再停用词表中的一些常用词,同时能过滤掉仅在局部文档中出现的高频词	对分词效果具有很强的依赖性

融合多统计特征的统计模型是目前主流的统计方法,选择符合领域术语特征的统计参数是对术语抽取研究的有效尝试。基于统计的方法适用于大规模语料、容易实现自动化且对不同领域的适应性很强,但存在依赖分词结果、易受测试语料规模影响、缺乏语义逻辑等问题。

1.1.3 规则与统计相结合的方法

规则与统计相结合的方法又称混合方法,此类方法是从经验主义和理性主义两方面对术语进行量度的,即采用了统计方法适用于大规模语料的特征,又融合了语言规则精确度高的特征用于提取领域术语。混合方法,特别是统计机器学习模型,是目前领域术语抽取研究的重点和热点。研究[8]提出的基于质子串分解的方法,使用参数F-MI抽取简单质词,质子串分解方法抽取复杂结构合词,有效的提高术语抽取的准确率。研究[9]的研究中采用的IC-value方法从逆文档频率、公共破碎字串和术语长度三个方面改进了C-value方法,实验证明500词内的抽回术语准确率和召回率分别为77.8%和29.81%,此算法能有效识别长术语和公共破碎字串,但对低频术语的识别能力较差。条件随机场(CRFs)兼具最大熵模型(ME)和隐马尔科夫模型(HMM)的特征,是目前标注和切分序列数据效果最好的机器学习模型。研究[10]以CRFs为依托,融合了词性、词典、领域频率等术语特征,并采用交叉验证方法确定模型训练参数,准确率、召回率分别为84.61%、80.5%。但此方法需要合适的训练集对模型参数进行训练,而训练集构建耗费大量的时间和人力,且不同领域训练集也不同,这就导致了训练模型的可移植性很差。

混合方法是当前术语抽取研究的主流方法。此类方法吸取统计方法适用于大规模语料处理的特征并融合了规则方法抽取精度高等优点,在对领域语料整理、领域概念和领域特征分析的基础上,选择符合领域特征的统计参数与语言规则,有效提高了术语抽取的准确率和召回率。

1.2 领域术语抽取工作中存在的问题

1) 依赖分词及词性标注的准确度

由于专业领域词汇的缺乏,在分词过程中,专业领域词汇常会被错误的切分成多个单词或形成单词碎片。目前的一些研究直接对分词结果进行统计作为候选术语,忽略了可能存在的分词错误对术语单元性和领域性造成的破坏。

2) 过分依赖前景知识(领域词典)

Krauthammer(2004)曾对词典术语抽取方法进行实验,结果表明由于词典易受到灵活的语言表达和术语变体的影响,此类方法不但领域移植性较差而且术语识别率较低。针对术语抽取词典方法中存在的问题,研究[11]提出了不依赖领域词典的术语抽取算法,取得了一定的效果。

3) 重视领域特征而忽视了术语的单元性特征

术语单元性和领域性是术语的两个基本统计指标,为了有效的提高领域术语抽取的召回率和准确率,术语抽取研究应对术语单元性和领域性两方面给予同等的关注。研究[12]表明集成术语的单元特征和领域特征能有效提高术语抽取的准确率。

1.3 领域术语抽取研究发展趋势

多策略融合无疑是提升术语抽取效果的有效途径,其基本思想即不同术语抽取策略间的补充。目前基于多策略术语抽取方法主要有两个方面:一是融合多种规则和术语统计参数的多策略融合术语抽取方法;二是统计机器学习方法融合多种术语特征。多种统计特征结合术语构词规则的术语抽取方法已成为术语抽取研究的主要方向,研究[13]提出的NC-value参数和互信息结合的方法,集中识别三字以上的长术语,实验表明此方法在准确率和召回率均获得了一定的提升。而研究[14]提出一种双层HMM算法,利用HMM有效的解决语法规则的概率存在和穷举局限性问题,实验表明此方法具有良好的性能。

关注自然语言处理各领域中的最新研究动态,将相关领域的研究策略向术语自动抽取进行有效的迁移是对术语自动抽取研究的一种有意义的探索。此外领域术语抽取是从领域文本中获取代表领域核心概念的词语集合,如果术语抽取能够在抽取术语的同时构建术语的内涵和外延,实现从领域术语短语数据到知识的价值转化,将对术语抽取相关研究具有极其重要的意义。

2 结束语

领域术语抽取的研究与实现是一个复杂的过程,从领域术语研究的整体视角对各类方法和关键技术进行探讨,并对目前研究中存在问题的反思,对于不同特征识别算法的有效融合具有重要的理论意义。

参考文献:

[1] 冯志伟.现代术语学引论[M].北京:商务印书馆,2011.
[2] 段国成.基于CCD的术语抽取研究[D].郑州:郑州大学,2007.
[3] Sui Z, Chen Y, Wei Z. Automatic recognition of Chinese scientific and technological terms using integrated linguistic knowledge[C]// Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on. IEEE, 2003: 444-451.
[4] 姚振军,黄德根.正则表达式在汉英对照中国文化术语抽取中应用[J].大连理工大学学报,2010,2:140-144.
[5] 周浪,冯冲,黄河燕.一种面向术语抽取的短语过滤技术[J].计算机工程与应用,2009,45(19):9-11.
[6] 潘虹,徐朝军.LCS算法在术语抽取中的应用研究[J].情报学报,2010,29(5):853-857.
[7] 周浪,张亮,冯冲,等.基于词频分布变化统计的术语抽取方法[J].计算机科学,2009,36(5):177-180.
[8] 何婷婷,张勇.基于质子串分解的中文术语自动抽取[J].计算机工程,2006,32(23):188-190.
[9] 胡阿沛,张静,刘俊丽.基于改进C-value方法的中文术语抽取[J].现代图书情报技术,2013,(02):24-29.
[10] 李丽双,党延忠.基于条件随机场的汽车领域术语抽取[J].大连理工大学学报,2013,53(2):267-272.
[11] 王卫民,贺冬春,符建辉.基于种子扩充的专业术语识别方法研究[J].计算机应用研究,2012,29(11):4105-4107.
[12] Kang Jingjing, Liu Tao, Hu He. Discovering Chinese compound term using termhood and unithood measure[C]//IEEE 2011 Sixth Annual China Grid Conference Dalian,2011:60-67.
[13] 梁颖红,张文静.基于混合策略的高精度长术语自动抽取[J].中文信息学报,2009,23(6):26-30.
[14] 岑咏华,韩哲.基于隐马尔科夫模型的中文术语识别研究[J].现代图书情报技术,2008,12:54-58.