软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006040]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

自动术语抽取研究综述*

张 雪[1,2], 孙宏宇[1,2], 辛东兴[1,2], 李翠平[1,2], 陈 红[1,2]

[1](中国人民大学 信息学院,北京 100872)

[2](数据工程与知识工程教育部重点实验室(中国人民大学),北京 100872)

通讯作者: 李翠平, E-mail: licuiping@ruc.edu.cn

摘 要: 自动术语抽取是从文本集合中自动抽取领域相关的词或短语,是本体构建、文本摘要、知识图谱等领域的关键基础问题和研究热点.特别是,随着近年来对非结构化文本大数据研究的兴起,使得自动术语抽取技术进一步得到学者的广泛关注,取得了较为丰富的研究成果.本文以术语排序算法为主线,对自动术语抽取方法的理论、技术、现状及优缺点进行研究综述:首先概述了自动术语抽取问题的形式化定义和解决框架.然后围绕"浅层语言分析"中基础语言信息和关系结构信息两个层面的特征对近年来国内外的研究成果进行分类,系统总结了现有自动术语抽取方法的研究进展和面临的挑战.最后对术语抽取使用的数据资源及实验评价进行分析,并对自动术语抽取未来可能的研究趋势进行了探讨与展望.

关键词: 自动术语抽取;术语识别;文本处理;机器学习

中图法分类号: TP391

中文引用格式: 张雪,孙宏宇,辛东兴,李翠平,陈红.自动术语抽取研究综述.软件学报. http://www.jos.org.cn/1000-9825/6040.htm

英文引用格式: Zhang X, Sun HY, Xin DX, Li CP, Chen H. Research survey on automatic term extraction. Ruan Jian Xue Bao/Journal of Software, (in Chinese). http://www.jos.org.cn/1000-9825/6040.htm

Research Survey on Automatic Term Extraction

ZHANG Xue[1,2], SUN Hong-Yu[1,2], XIN Dong-Xing[1,2], LI Cui-Ping[1,2], CHEN Hong[1,2]

[1](School of Information, Renmin University of China, Beijing 100872, China)

[2](Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education (Renmin University of China), Beijing 100872, China)

Abstract: Automatic term extraction is to extract domain-related words or phrases from document collections. It is a core basic problem and research hotspot in the fields of ontology construction, text summarization and knowledge graph. In particular, under the rise of unstructured text studies in big data, automatic term extraction technology has been further concerned by researchers and has obtained rich research results recently. With the terminology sorting algorithm as the main clue, this paper surveys the basic theories, technologies, current research works, advantages and disadvantages of automatic term extraction methods. First, the formalized definition and solution framework of automatic term extraction problem are outlined. Then, based on the features of the basic language information and the relational structure information in the "shallow parsing", the latest study results are classified, research progress and major challenges of existing automatic term extraction methods are summaried systematically. Finally, some available data resources are listed, evaluation approaches are analyzed and this paper predicts possible research trends in the future.

Key words: automatic term extraction; term recognition; text processing; machine learning

随着大数据、移动互联网和社交媒体等技术的迅猛发展,使得网络空间中所蕴含的文本数据量呈指数级增

* 基金项目: 国家自然科学基金(61772537, 61772536, 61702522, 61532021); 国家重点研发计划(2018YFB1004401)

Foundation item: National Natural Science Foundation of China (61772537, 61772536, 61702522, 61532021); National Key Research and Development Program of China (2018YFB1004401)

收稿时间: 2019-09-17; 修改时间: 2020-02-09; 采用时间: 2020-04-12; jos 在线出版时间: 2020-04-21

长.因此,如何对这些文本数据进行分析并挖掘出最有价值的内容(例如术语、实体、关系、语义图等)成为当前备受关注的重要研究领域.其中,从大型文本集合中抽取出描述某一特定领域(例如科技文献、社交推文等领域)的术语(term,包括单词或短语)是文本挖掘和信息抽取的首要步骤,也是本体构建[1,2]、文本分类[3]、文本摘要[4,5]、机器翻译[6,7]、知识图谱[8]等领域的关键基础问题和研究热点.特别是,随着近年来对非结构化文本大数据研究的兴起,术语抽取问题得到更加广泛的关注和深入研究,一些最新的研究成果出现在信息检索[9]、自然语言处理[10]、数据库[11,12]、人工智能[13]、数据挖掘[14,15]等相关领域的顶级国际会议和期刊上.

自 20 世纪 30 年代初期奥地利术语学博士 Eugen Wuister 教授正式创立“术语学”起至今 80 余年,大量学者对术语相关领域展开了广泛的研究.最初,借助于术语学者和领域专家的背景知识人工进行术语识别及抽取,形成特定领域的术语库,供学术界和工业界使用.但这一时期的术语抽取严重依赖于专家知识,抽取工作繁重、耗时长且效率低,属于人工术语抽取阶段.

之后,伴随着计算机技术的迅猛发展,自动术语抽取(Automatic Term Extraction,简称 ATE)越来越受到关注,大量的自动术语抽取方法、框架和工具不断涌现,这些方法取得了一定的成绩和较好的效果.这一阶段(属于经典方法阶段)的自动术语抽取方法主要分为基于语言学、基于统计学和两者混合的抽取方法三类.基于语言学的术语抽取方法主要是制定可涵盖领域语言特征的规则集合,然后通过形式化定义的规则集合来抽取术语.如 Bourigault 等人[16]通过词性标注(Part-of-speech tagging,简称 POS tagging)来标记目标语料库中的所有文档,根据已有的术语集使用有限状态机技术自动学习出文档中的规则集合.基于语言学的术语抽取方法准确率很高,但依赖于特定语言规则,可移植性较差,不能跨领域迁移使用,局限性很大.基于此,继而提出了基于统计学的术语抽取方法和两者混合的术语抽取方法来解决语言无关性和模型通用性问题.如 Justeson 等人[17]最早提出了基于词频的术语抽取方法.Frantzi 等人[18]首次将语言学和统计学方法进行融合,提出了 C-value 方法.

经典方法在自动术语抽取过程中只考虑了术语本身特征及其在目标语料库中的词频特征,使得术语抽取效果深受目标语料库规模和质量的影响.因此,学者逐渐将外部知识(例如维基百科、WordNet 等)、语义信息、图结构、主题模型及深度学习等技术应用到自动术语抽取任务中.这一阶段(属于拓展方法阶段)的术语抽取方法不再局限于“浅层语言分析”中的基础语言信息:即术语本身的构词特征和词频特征.而是考虑较深一层的关系结构信息:包括术语与常用词之间的频率分布差异、术语与术语之间的语义关联以及更多类型特征的融合等,因此拓展阶段的自动术语抽取方法分为基于外部知识的术语抽取、基于语义相关的术语抽取、基于机器学习的术语抽取、基于深度学习的术语抽取、基于图的术语抽取和基于主题模型的术语抽取.如 Vivaldi 等人[19]于 2010 年使用维基百科辅助抽取术语,Astrakhantsev 等人[20]于 2014 年结合术语候选词与领域关键概念共同计算语义相似度进行术语抽取及排序,同年 Judea 等人[10]使用特征工程及条件随机场模型 CRF 来抽取专利术语,之后 Wang 等人[21]将深度学习模型引入自动术语抽取任务中,Lossio-Ventura 等人[22]首次将图结构应用到生物医学领域进行术语抽取,Bolshakova 等人[23]利用主题建模技术(例如聚类,LDA)对特定领域的术语进行抽取,并证明主题信息可以有效提高术语抽取质量.除此之外,自动术语抽取还结合其他领域的思想来提高抽取效果,如 Liu 等人[12]于 2015 年首次将短语分割思想与术语抽取相结合,提出 SegPhrase 模型,Shang 等人[14]于 2018 年在 Liu 的基础上添加远程监督技术和 POS 指导的短语分割技术,提出 AutoPhrase 模型,有效避免了额外的手动标记工作并增强术语抽取效果.

本文不同于已有综述文献[24,25],将所有 ATE 方法按照术语特征进行分类[24]或者按照术语抽取的关键技术进行分类[25],而是创新性得提出利用“浅层语言分析”中基础语言信息和关系结构信息两个层面的特征对近年来国内外的研究成果进行分类总结,并详细描述各个类别中包含的术语抽取模型、使用特征及优缺点.这样做的好处是可以从更基础更全面的角度对现有 ATE 解决方案进行了解,有助于综合已有的高效方法、较新的方法及引入有用的外部资源,进而提出更加高效的自动术语抽取特征及解决方法.与之前的综述论文相比,本文主要围绕“浅层语言分析”来建立一个尽可能完整的领域术语语义图,然后根据语义图中不同类型的信息对现有 ATE 方法进行分类:1)术语特征,即基础语言信息(类似于图中顶点)和 2)术语间的语义关系,即关系结构信息(类似于图中的边).这种分类方法补充了之前综述论文均忽略的角度:术语间的语义关系,使得本文更加清晰.

除本节外,本文第 1 节概述了术语抽取问题的形式化定义以及通用解决框架.第 2 节详细总结了现有文献所使用的术语抽取方法,并对其进行分类;系统分析各类自动术语抽取方法的研究现状及面临的挑战.第 3 节归纳分析了自动术语抽取常用的数据集、工具、评价方法及评价指标,便于学者开展实验评估.第 4 节对自动术语抽取未来可能的研究趋势进行了探讨与展望,并总结全文.

1 自动术语抽取的问题定义及解决框架

1.1 问题定义

自动术语抽取任务的目标是从文档集合中抽取并排序与领域相关度高的词或短语,其形式化定义如下:

输入: 给定包含 n 篇文档的目标语料库 $D=\{d_1,d_2,\ldots,d_n\}$,每篇文档 $d_i \in D$ 是由 N_d 个单词 w_{ij}, $j=1,2,\ldots N_d$ 序列组成的集合,则文档表示为 $d_i=\{w_{i1},w_{i2},\ldots,w_{ij}\ldots,w_{iN_d}\}$ 自动术语抽取的基本流程如图 1 所示:

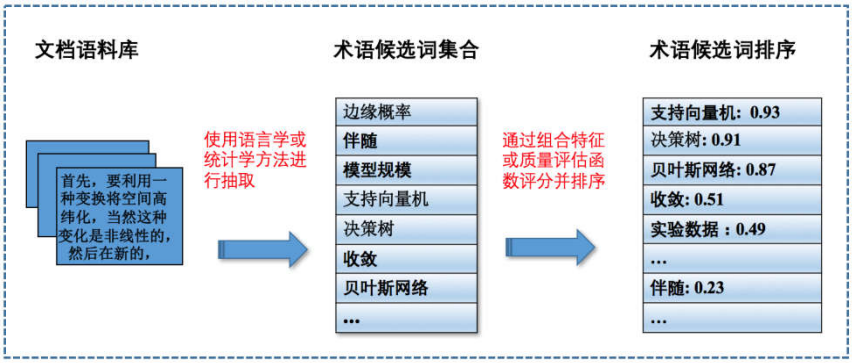

Fig.1 The basic process of automatic term extraction

图 1 自动术语抽取基本流程

(1) 使用语言学或统计学工具,生成术语候选词(包括词或短语)集合 $T=\{t_1,t_2,\ldots,t_i\ldots,t_m\}$;

(2) 使用组合特征或组合方法评估术语候选词的质量,将术语候选词 $t_i \in T$ 映射到某个评分,然后根据术语候选词的评分对集合 T 进行排序.

输出: 按照候选词质量评分降序排列的术语列表.

用户想要获取机器学习领域术语时,需要先对所有文档进行预处理,然后使用语言解析器或 *n*-gram 得到术语候选词集合.从图 1 可以看出,该集合中包含质量参差的候选术语,如高质量术语“支持向量机”、常用词“实验数据”及错误短语“伴随”等.为了从候选术语集合中抽取真正术语,需利用 ATE 方法来度量每个候选术语的质量,并为其分配一个[0,1]之间的分数,如“支持向量机”为 0.93,“实验数据”为 0.49,“伴随”为 0.23.随后按照候选词的评分降序排列,得到如图 1 右边所示的术语列表.最后按照用户需要,从术语列表中抽取前 N 个或大于评分阈值 k 的术语表返回给用户.

1.2 解决方案框架

解决自动术语抽取问题的框架和具体步骤如图 2 所示:

(1) 确定语料库文档.即确定目标语料库中要抽取的文档类型.

根据语料库中可用标注数据的量级,分为通用文档和特定文档,例如新闻类文档与科技文献类文档.在通用文档语料库中,有大量公共标注数据集可供使用;而在特定文档语料库中,只有少量标注数据可用.因此可以考虑是否借助外部知识库,例如维基百科、百度百科、HowNet 通用知识库等来扩充语料库中的标注数据.

根据语料库文档是否遵循常规语法,分为规范文档和非规范文档,例如新闻类文档与微博类文档.规范文档通常使用严格定义的语法及符号,便于抽候选术语,而微博、推特等非规范文档则使用较为宽泛的语法来组织词汇、图像和符号,如大写,缩写等,增加了术语抽取的难度.因此可以考虑使用文档预处理来过滤所有无关的内

容.

(2) 生成术语候选词集合.

术语候选词集合的生成是自动术语抽取方法的基础重要步骤,因为候选术语质量的好坏直接影响术语抽取的最终结果.

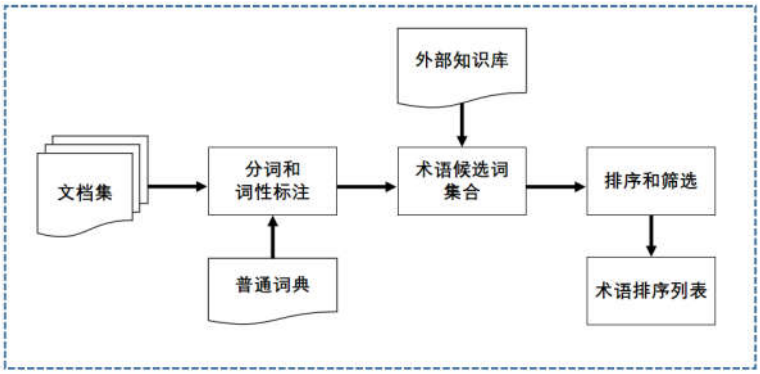

Fig.2 The framework of automatic term extraction

图 2 自动术语抽取的具体步骤及解决框架

通常,文档集合使用语言处理器或基于统计的启发式规则来抽取候选术语,生成术语候选词集合.首先,对目标语料库中的文档集合进行预处理,包括分词、词干化、词性标注等;同时进行文档切分,以标点符号作为分隔符分割文档,得到文本串片段(segment).然后,使用启发式规则(如 *n*-gram 过滤)[12]保留频率较高的 *n* 元单词序列,或者使用语言处理器中词性标注 POS tagging 标记单词的词性,根据预定义的词类模式[18,26,27],如名词短语块、动词短语块等进行浅层分块解析,得到多元词组.最后,为了减少术语候选词集合中的噪声数据,需要进行额外的过滤处理:(1)根据词频过滤:出现次数少于 2 次或 3 次的候选词会被过滤;(2)根据预设的停用词表过滤[27]:一些单词很少包含在真正术语中,如“good”、“interesting”等,但在语料库中出现频率很高(例如 good method).因此,需要将这些单词纳入停用词表进行过滤;(3)根据候选词的长度或候选词是否包含特殊符号过滤[28]:候选词长度大于 6 或候选词包含非字母符号或候选词是由一个字母组成的单词等情况均会被过滤.

(3) 对术语候选词排序并筛选.

术语候选词排序算法是自动术语抽取方法中最重要、最复杂的步骤[29,30],术语排序算法通常是不同类别自动术语抽取方法的重要区别,即不同的抽取方法采用不同的术语排序算法.如前所述,本文主要围绕“浅层语言分析”中基础语言信息和关系结构信息两个层面的特征对近年来国内外文献提出的自动术语抽取方法进行详细的分类总结,具体见第 2 章,将自动术语抽取方法细分为 9 种类别.术语筛选则通常采用启发式算法对已排序的术语候选词进行判断筛选,分为两种方法:(1)按照阈值设定,将大于术语评分阈值的候选词确定为真正术语;(2)按照数量要求(前 *N* 个),将得分高的候选词确定为真正术语.

值得注意的是,本文中绝大部分 ATE 方法采用图 2 所示的解决步骤和框架,但仍有少部分 ATE 方法存在例外.一种情况是 ATE 方法直接产生最终术语集合,但不对集合中的术语评分排序.例如基于语言解析器[34]或基于序列标注的深度学习方法[90,91]均属于此种情况.另外一种情况是在产生候选术语集合并对其进行排序后，再融合其他方法对之前的排序结果进行重新排序(re-ranking).例如 Segphrase 方法[12]和 AutoPhrase 方法[14]均结合短语分割模型对术语排序结果进行重新排序,以及 Zhang 等人[103]提出的通用方法 SemRe-Rank,是在现有 ATE 方法产生排序术语列表的基础上构建图模型,而后根据图中术语(顶点)的语义相关性对术语进行重新排序,得到一个重排后的术语列表.其中,第一种情况中的术语集合可能会包含领域相关度不高的短语.

2 自动术语抽取方法的对比与分析

自动术语抽取方法的研究工作分为两个阶段:经典阶段和拓展阶段.在研究初期,一些经典的自动术语抽取

方法如基于语言学的方法、基于统计学的方法及两者混合的抽取方法被广泛使用,这一阶段(属于经典方法阶段)以不断总结语言特征规则和尝试各种经典统计学方法为主,取得了一定的成绩和较好的抽取效果.但经典方法在自动术语抽取过程中只考虑术语本身特征及目标语料库中候选术语的词频特征,使得术语抽取效果受目标语料库规模和质量的影响很大.

为了克服这一问题,学者逐渐将外部知识(例如维基百科、WordNet、参考语料库等)、语义信息、图结构及主题模型等方法应用到自动术语抽取任务中.这一阶段(属于拓展方法阶段)的术语抽取方法不再局限于“浅层语言分析”中的基础语言信息:即术语本身的构词特征和词频特征.而是考虑较深一层的关系结构信息:包括术语与常用词之间的频率分布差异、术语与术语之间的语义关联以及不同类型特征之间的融合等.因此,在研究的拓展阶段,以加入新兴的特征和方法为主,可以将自动术语抽取方法分为基于外部知识的方法、基于机器学习的方法、基于深度学习的方法、基于语义相关的方法、基于图的方法以及基于主题模型的方法.

本文所提出的自动术语抽取分类方法,不同于已有综述文献中按照术语特征分类[24]或者按照术语抽取方法的关键技术分类[25],而是基于文档分析层次中“浅层语言分析”来划分自动术语抽取方法的分属类别,填补了已有综述中均忽略的角度:术语间的语义关系,使得本文更加详细清晰.具体来说,自动术语抽取过程中对目标文档的分析分为两个层次:浅层语言分析和深层语义分析.浅层语言分析主要包括:(1)对文档中术语候选词的词性、词频等基础语言信息进行分析;(2)对术语和术语之间的共现关系、语义相似等关系结构信息进行分析.深层语义分析则是链接术语在现实世界中对应的实体,方便更深层次理解术语的完整语义.

简单来说,可以将目标文档集合看作一个或多个语义图,术语表示语义图中的顶点,术语之间的关联关系表示语义图中的边:例如语义相似关系、同义关系、上下位关系、整体-部分关系等.因此,浅层语言分析中的基础语言信息等同于顶点的属性(即顶点特征),关系结构信息等同于边的权重(即边的特征).而深层语义分析则是通过链接关系构建术语与外部知识库实体之间的映射,进而使用整个知识库进行术语含义的消歧、扩展以及深层语义理解.经过广泛调研,基于深层语义分析的 ATE 方法较少[98,99,100].因此本文主要围绕“浅层语言分析”中基础语言信息和关系结构信息两个层面的特征对近年来国内外的研究成果进行分类总结,将上文中提到的 9 种自动术语抽取方法进行如下分类,详见表 1.(注:将少量基于深层语义分析的 ATE 方法归入“基于语义相关的方法”)

Table 1 Method category of automatic term extraction

表 1 自动术语抽取方法分类

术语抽取分类	详细类别	使用特征
基础语言信息类	基于语言学的方法	词形特征,语义特征,词法特征
	基于统计学的方法	词频特征
	混合方法	语义特征,词法特征,词频特征等
	基于外部知识的方法	候选词在特定领域与其在通用领域的对比特征
	基于机器学习的方法	语义特征,词法特征,词频特征,外部资源特征,分布式特征等
	基于深度学习的方法	分布式特征
关系结构信息类	基于语义相关的方法	候选词之间的相似性
	基于图的方法	候选词之间的关系特征:共现关系、语义相似等
	基于主题模型的方法	候选词在主题上的分布特征

这样分类的好处是可以从更基础更全面的角度对现有 ATE 解决方案进行了解,有助于对比不同方法之间的关联关系,综合已有的高效方法、较新的方法以及引入有用的外部资源,进而提出更加高效的自动术语抽取特征及解决方法.在下文中,将依次对各类 ATE 方法做详细介绍.

2.1 基于语言学的抽取方法

基于语言学的自动术语抽取方法主要利用词法模式、词形特征、语义信息等基础语言知识从目标语料库中抽取术语.其基本思想:术语常以特定的语言结构和模式出现,通过发现符合术语模式的字串,构建一套较完整的词法规则集合,自动抽取出领域术语.

在研究初期(20 世纪 90 年代),自动术语抽取主要是基于语言学知识,使用词或词组的词性标注和分块技术等来确定术语候选词的前后边界,利用语言学专家手工构造的规则模板确定候选词是否为领域术语.这一时期,自动术语抽取通常应用于翻译和搜索领域来协助提高这些任务的效率.例如 FASTR 系统[31,32]、Terms 系统[33]、TERMINO 系统[34]、NODALIDA 系统[35]、LEXTER 系统[36]、Naulleau-98 系统[37]均使用术语构造规则对候选术语进行筛查,所使用的规则数量从七十条到上百条不等,包含大量启发式规则.其中术语抽取效果最好的是 NODALIDA 系统[35],在英文宇宙学语料中准确率达到 95%～98%,召回率达到 98.5%～100%,但由于测试阶段使用的语料库太少而无法大范围拓展使用.最具特点的是 TERMINO 系统[34],首次引入 synapsy 检测器的概念,synapsy 检测器采用启发式方法来构造词法规则,并将多个连续词汇构成的核心成分标记为候选术语.Terms 系统[33]则在构造规则的基础上,增加了术语平均长度(为 1.91)特征及技术术语的构词特征,然后对文档进行候选词抽取排序,在光谱分析语料库中准确率达到 96%,但生成结果中噪声较大.同样,抽取结果中含噪声数据较多的还有 LEXTER 系统[36],该系统首先利用最长名词短语作为术语候选词,然后用抽取的候选词构造术语网络,最后由内部学习机制来筛选术语.虽然部分术语抽取系统的结果中包含噪声,但大多数抽取系统不依赖于大型词典,且在特定领域的术语抽取准确率极高,被较早应用在工业领域.

上述基于语言学开发的术语抽取系统或方法大多受到手动规则和噪声数据等方面的限制,很难适应其他领域.针对这一问题,学者们提出自动学习领域语言规则的模型,借助模型对大规模语料中术语的词性规则进行抽取,并为这些词性规则定制优先级.如 1994 年,Punyakanok 等人[16]通过词性标注(POS tagging)来标记目标语料库中的所有文档,根据已有的术语集使用有限状态机技术自动学习出文档中的规则集合.Koo 等人[38]则使用更为复杂的 NLP 技术(依赖解析器)进一步提高术语抽取的准确率.2010 年,Foo 等人[39]采用有监督的机器学习算法 Ripper 来学习目标语料库中的语言学规则,分别使用基于语言学(词性标记、形态-语法、语法功能、语义信息等)和统计学(归一化词频等)2 类 10 种术语特征来获取完善的规则集合,实验表明机器学习算法可以生成术语规则,且比人工归纳快很多倍.最新研究中,Li 等人[40]于 2018 年提出了面向基础教育领域的术语抽取模型 DRTE,首次在语言学抽取中考虑术语间的关系.该模型首先使用术语定义及术语关系模板来获取术语候选词,然后综合构词规则与边界检测方法来最终确定术语.该模型能够有效抽取领域低频术语,$F1$ 值达到 82.7%.

基于语言学的自动术语抽取方法主要利用语言专家对特定领域的术语进行识别,归纳总结出该领域的语言规则集合.理论上,只要在特定领域提取足够多的语言规则并为其定制良好的优先级,则该类方法在术语抽取的准确率上有极大的优势,还能有效识别出低频术语.但是基于语言学的 ATE 方法也存在着以下三个缺点:(1)过于依赖专家知识及 POS 标注器,使得抽取规则集合的模型不具有泛化性并导致标记错误向下游应用累积传播;(2)人工编写的规则不能覆盖领域中所有语言学特征;(3)针对某一领域的规则集合很难迁移到其他领域,导致该类方法的可移植性不强.因此,目前很少使用纯语言学方法进行自动术语抽取的研究,主要将其作为术语抽取的预处理步骤用以生成术语候选词集合,如文献[4,9,22,64,67,99]均使用语言规则作为自动术语抽取的第一步.

2.2 基于统计学的抽取方法

基于统计学的自动术语抽取方法利用目标语料库中词或词组的分布频率来抽取术语.相比较基于语言学的方法,这类方法简单高效,不需要领域专家、人工标注数据和外部词典.其基本思想:文档集合经过预处理后,可使用简单的统计方法进行过滤,比如词频、TF-IDF 等,生成术语候选词集合;然后按照阀值设定将大于术语评分阈值的候选词确定为真正术语或者按照数量要求(前 N 个)将得分高的候选词确定为真正术语.

基于统计学的抽取方法通常将术语特性归结为两个便于度量的原则[41]:单元性度量和领域性度量.第 2.2.1 节和第 2.2.2 节根据不同度量原则,对其常用方法进行介绍.

2.2.1 单元性度量

单元性度量(unithood):衡量术语候选词(长度>=2)内部的搭配强度和粘合程度,只针对多字术语(Multi-word Terms, MWT),又称单词关联度量.单元性度量最显著的特征是词频,词频越高,候选术语内部结构越稳定.

单元性度量的假设基础:如果一个单词序列频繁地出现在一起,它可能表达了一个独立完整的语言含义,需要有效的方法验证该单词序列是否具有稳定的内部结构.常用方法:z 检验[42],t 检验[43], χ^2 检验[44],对数似然比[45],点互信息[46].

(1) z 检验(z test).z 检验[42]是一种基于均值的统计检验,根据独立性假设来检验构成候选术语的单词内部是否存在关联.主要用于样本量较大,数据满足正态分布的语料库,公式如下:

$$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}=\frac{TF(t)/N-p}{\sqrt{p(1-p)/N}}$$

其中,t 表示候选术语,N 表示目标语料库中单词的总个数,$TF(t)$表示候选术语 t 的词频.z 检验假设语料库是一个伯努利过程,遇到候选术语 t 记为一次“成功”,其他是“失败”.根据独立性假设,构成术语 t 的单词 w_1 , w_2 相互独立,则术语 t 的概率可表示为 $p=p(w_1w_2)=p(w_1)\cdot(w_2)$;然后使用最大似然估计,得到单词频率替换后的术语概率 $p=TF(w_1)/TF(w_2)/N$;最后计算 z 值,检验样本均值 $\bar{x}$ 与总体均值 μ 之间的差异是否显著.当 z 值大于查表所得的阀值 α 时,认为差异性显著,即构成候选术语 t 的单词内部存在关联.

(2) t 检验(t test)[43]与 z 检验原理相似,主要用于样本量较小(例如 n<30),总体标准差 σ 未知的正态分布语料库.与 z 检验相比,t 检验的优点是不需要提前知道总体方差,可以使用样本方差替代总体方差.因此,在实践中 t 检验比 z 检验更为常用.

(3) χ^2 检验(Chi-square test).χ^2 检验[44]是一种基于方差的统计检验,根据观察值和期望值之间的偏差程度来检验候选术语内部单词是否相互独立,公示如下:

$$\chi^2=\sum_i\frac{(O_i-E_i)^2}{E_i}$$

其中,O_i 表示候选术语 t 中某个类别 i 的观察词频, E_i 表示基于语料库计算出的期望词频.由 χ^2 公式求出候选术语 t 的 χ^2 值,然后从卡方界值表找到特定阀值 α,当 χ^2 值大于 α 时,认为差异性显著,且两者差值越大,说明术语候选词内部关联性越强,结构越稳定.χ^2 检验的优点是不需要假设目标语料库满足正态分布,缺点则是当语料库的规模较小时,其检验结果不具备说服力.

(4) 对数似然比(Log Likelihood Ratio,简称 LLR).对数似然比[45]旨在检测候选术语序列是否只是文档集合中的简单偶然事件.根据 Dunning 的研究[45],需要先制定两个假设:

$$h_1: P(w_1\mid w_2)=P(w_1\mid\neg w_2)$$
$$h_2: P(w_1\mid w_2)\neq P(w_1\mid\neg w_2)$$

其中, w_1w_2 表示二元候选术语,h_1 表示独立性假设 1,即 w_2 的出现独立于 w_1 的出现.h_2 表示相关性假设 2,即 w_2 的出现和前面 w_1 的出现是相关的.给定 c_1 、 c_2 、 c_{12} 表示 w_1 、 w_2 、 w_1w_2 在目标语料库中的出现次数,N 表示语料库中的单词总数,使用最大似然估计方法,得到不同假设条件下的概率 p、p_1、p_2,其计算公式如表 2 所示:

Table 2 Probability calculation formula

表 2 概率计算公式

	h_1	h_2
$P(w_2\mid w_1)$	$p=\frac{c_2}{N}$	$p_1=\frac{c_{12}}{c_1}$
$P(w_2\mid\neg w_1)$	$p=\frac{c_2}{N}$	$p_2=\frac{c_2-c_{12}}{N-c_1}$

根据表 2,对数似然比的公式为:

$$LLR=-2*\log\frac{L(H_1)}{L(H_2)}=-2*\log\frac{b(c_{12};c_1,p)b(c_2-c_{12};N-c_1,p)}{b(c_{12};c_1,p_1)b(c_2-c_{12};N-c_1,p_2)}$$ 其中, $b(k,n,x)=x^k(1-x)^{(n-k)}$

当对数似然比值较大时,说明比较符合假设 2 的预期,即候选术语 w_1w_2 不是偶然出现的,其内部具有相关性.

(5) 点互信息(Pointwise Mutual Information,简称 PMI).点互信息[46]主要用来衡量两个随机事件之间的相关程度.在术语抽取中,PMI 作为验证两个单词之间是否相关的度量准则,通过计算候选术语内部结合的紧密强

度,达到抽取术语候选词的目的.公示如下:

$$PMI = \log\frac{p(w_1w_2)}{p(w_1)p(w_2)} \approx \log\frac{N\cdot TF(w_1w_2)}{TF(w_1)TF(w_2)}$$

其中,p 表示概率, $p(w_1w_2)$ 表示单词 w_1 与 w_2 同时出现的概率,当语料库数据充足时,可以使用归一化词频表示.由独立性假设可知:如果单词 w_1 与 w_2 不相关,则 $p = p(w_1w_2) = p(w_1)p(w_2)$;如果两词相关性越大,则 $p(w_1w_2)$ 与 $p(w_1)p(w_2)$ 比值越大.因此,PMI 值越大,说明单词 w_1 与 w_2 相关性越大,可以组成一个短语.Pecina[47]通过实验证明:衡量二元候选词关联度的 90 多种方法中,使用 PMI 计算的二元词组搭配效果较好.但点互信息的缺点是过于强调罕见事件,在其他条件都相等的情况下,由低频词组成的候选术语的点互信息要大于高频词组成的候选术语.

虽然单元性度量在自动术语抽取中起着不可或缺的作用,如大量文献[11,12,14,48,49]均使用 unithood 来进行术语候选词的度量.但是 Loukachevitch 等人[50]及 Wong 等人[51]的研究表明,单元性度量本身不足以评估术语候选词的有效性,只能确定该单词序列是固定搭配,不能验证是否是真正的术语.因此,术语候选词即使通过了单元性度量检验,还是需要度量其与领域的相关性.

2.2.2 领域性度量

领域性度量(termhood):衡量术语候选词与特定领域的相关程度.领域性度量主要根据目标语料库中词或词组的分布统计数据(例如词频、TF-IDF 值等)来计算术语与领域的关联程度,可同时度量单字术语(Single-word Terms,简称 SWT)和多字术语(Multi-word Terms,简称 MWT).常用方法分为两类:基于词频的方法和基于文档频率的方法.

(1) 基于词频的方法

词频(Term Frequency,简称 TF)是指候选术语 t 在文档集合中出现的总次数.Justeson 等人[17]于 1995 年最早提出了基于词频的术语抽取方法,其假设基础:如果单词序列特定于某个领域,那么它会经常出现在该领域的文本集合中.词频通常被用作术语候选词的初始过滤器,如果候选词的词频非常低会被过滤掉.这样做有助于减少大量的噪声数据,提高术语抽取准确率.为了研究方便,一些论文[15]会使用归一化词频 NTF 来代替总词频 TF.

平均词频(Average Term Frequency,简称 ATF)[52]使用候选术语 t 的总词频 TF 除以该候选术语所出现的文档数量,比值表示每个文档中候选术语 t 的平均出现次数.公式见表 3.

领域共识(Domain Consensus,简称 DC)[53]识别在语料库中均匀分布的术语.候选术语 t 在语料库中分布越均匀,说明候选术语在更多的文档中出现,DC 值就越高.Liu 等人[54]使用 DC 作为判断术语领域性的重要特征.

(2) 基于文档频率的方法

逆文档频率(Inverse Document frequency,简称 IDF)[55]用来衡量候选术语 t 所出现的文档数量占整个语料库文档数量的逆比重.Liu 等人[12]在所提出的 SegPhrase 模型中引入 IDF 作为衡量短语信息性的特征之一.

词频-逆文档频率(Term Frequency-Inverse Document frequency,简称 TF-IDF)[55]结合词频和逆文档频率来衡量术语候选词的领域性.相较于度量候选术语均匀分布的领域共识 DC,TF-IDF 是有偏重的,在少量文档中频繁出现的单词序列会得到更高的评分.由于 TF-IDF 是领域性度量中最富有成效的方法,一些研究者在 TF-IDF 的基础上进行改进.如 Zhou 等人[56]针对术语和非术语在各文档中词频分布情况的不同:即术语在不同文档中的词频差异较大,非术语的词频相对平稳,提出一种 TF-IDF 和词频方差相结合的领域相关性计算方法.Yan 等人[57]则在 Web 资源领域中引入新词发现算法及 TF-IDF 筛选进行术语抽取,最终在 3 类语料库中均取得较好的效果,其中基于锚文本语料中术语抽取准确率最高.此外,Lossio-Ventura 等人[58]将 TF-IDF 与 C-value 方法相结合提出了 F-TFIDF-C 方法,应用在生物医学领域进行特定术语抽取.

残差 IDF(Residual-IDF,简称 RIDF)[59]主要用来度量候选术语 t 的实际 IDF 得分与 t 在泊松分布上的预测 IDF 得分之间的偏差.RIDF 最初由 Church 等人[59]提出,用来抽取文档关键词增强 IDF 功能.后被 Zhang 等人[52]修改用来抽取术语候选词,其度量假设是基于泊松分布的 IDF 模型与基于目标语料库观察到的实际 IDF 之间是有偏差的,这种偏差使得领域术语更容易被发现.因为领域术语的偏差值通常高于非领域术语的偏差值.

领域性度量在自动术语抽取中是非常重要的,可以将真正的术语与常用短语进行区分.但现有的领域性度量方法还比较基础,主要基于频率进行度量,忽略领域低频术语的抽取,不能满足多种类型术语的抽取需求.

Table 3 The methods of termhood

表 3 领域性度量方法

类别	方法名称	计算公式
基于词频	词频	$TF(t)=\sum_{d\in Docs} f_d(t)$, $f_d(t)$表示候选术语 t 在文档集合 $Docs$ 中某一文档 d 中出现的次数
	归一化词频	$NTF(t)=\frac{TF(t)}{N}$, N 表示文档集合 $Docs$ 中包含的单词总数
	平均词频	$ATF(t)=\frac{TF(t)}{DF(t)}$, $DF(t)$表示文档集合 $Docs$ 中包含候选术语 t 的总文档数
	领域共识	$DC(t)=-\sum_{d\in Docs}\frac{TF_d(t)}{TF(t)}\log_2\frac{TF_d(t)}{TF(t)}$, $TF_d(t)$表示候选术语 t 在文档集合 $Docs$ 中某一文档 d 中出现的次数
基于文档频率	逆文档频率	$IDF(t)=\log\frac{\lvert Docs\rvert}{\lvert\{Docs:t\in Docs\}\rvert}$, $\lvert\{Docs:t\in Docs\}\rvert$表示文档集合 $Docs$ 中包含候选术语 t 的总文档数
	词频-逆文档频率	$TF\text{-}IDF(t)=TF(t)\cdot IDF(t)$
	RIDF	$RIDF(t)=IDF(t)-\log(\frac{1}{1-p(0;\lambda)})$, p 为泊松分布,参数 λ 为候选术语 t 的平均词频, 1-$p(0;\lambda)$为文档中至少一次出现候选术语 t 的泊松概率

基于统计学的自动术语抽取方法主要利用词频、文档频率等概率统计信息来抽取符合阀值的词或词组作为领域术语.该类方法简单易实现,通用性较强,不需要领域专家、语言学规则、语义信息,不需要标注数据和外部知识库,也不受领域限制.但是基于统计学的 ATE 方法依然存在以下两个缺点:(1)严重依赖目标语料库的规模和质量,若目标语料库规模较小,术语抽取效果直线下降.Li 等人[60]已通过实验证明了这一问题;(2)使用数据的信息力度较粗,对低频术语和单字术语的抽取效果不理想,且经常抽取到意义不完整的单词序列和常用词,导致输出列表中包含较多的噪声数据,召回率较低.目前,纯统计学方法已较少使用,多与其他方法相结合.

2.3 混合的抽取方法

混合术语抽取方法在研究初期多是结合语言学方法和统计学方法进行自动术语抽取,其中较早且有代表性的是 C-value 方法和 NC-value 方法.在拓展研究阶段,则以结合多种方法取其优势为主.

早在 2000 年,Frantzi 等人[18]观察到基于频率的术语抽取方法无法正确度量以下两种情况中的术语:(1)嵌套候选术语应具有与被嵌套术语相同或更高的频率;(2)两个不同长度的候选术语,在语料库中出现次数均为 n 次,较长候选术语应比较短候选术语更为重要.基于此,Frantzi 等人提出了可以合理度量嵌套术语的 C-value 方法,其基本思想:首先利用词法规则生成术语候选词集合,然后使用统计信息对集合中的术语进行过滤,其公式如下:

$$C-value(t)=\begin{cases}\log_2|t|\cdot f(t) & \text{if t is not nested}\\ \log_2|t|\cdot(f(t)-\frac{1}{|T_t|}\sum_{b\in T_t}f(b)) & \text{otherwise}\end{cases}$$

其中,t 表示抽取的某一候选术语,$|t|$表示候选术语 t 的长度,$f(t)$表示 t 在目标语料库中的词频,T_t 表示包含候选术语 t 的嵌套术语集合.C-value 方法首先计算术语候选词 t 的词频及其长度,然后根据嵌套 t 的较长候选术语的词频进行调整.如果候选术语 t 经常嵌套在较长词串中,其重要性(即 C-value 评分)会被降低.针对 C-value 方法只能处理多字术语(MWT),Barrón-Cedeno 等人[27]提出了 C-value 方法的变种,通过使用 $C(t)=i+\log_2|t|$(i 为常数)来替换公式中的 $\log_2|t|$,将 C-value 方法扩展到单字术语(SWT).C-value 及其变种方法因考虑术语长度及嵌套信息,在抽取长术语方面效果较好.

不少研究在 C-value 方法的基础上进行改进,最新的几种方法如 RAKE[61]、Basic[62]和 ComboBasic[63]也采用类似 C-value 方法的思路.RAKE 方法[61]支持抽取更长的多字术语,主要根据两部分来联合计算候选术语 t 中

每个单词 w_i 的评分:一部分倾向于单词嵌套在较长的候选术语,一部分有利于频繁出现的单词;然后将组成候选术语的单词评分相加.Bordea 等人[62]于 2013 年提出 Basic 方法,认为较长词串中的嵌套术语数量也应作为候选词领域性度量的一部分,通过扩展嵌套术语来修改 C-value 方法.Basic 方法仅适用于度量较长的嵌套术语.

$$Basic(t) = |t|\log f(t) + \alpha e_t$$

其中 e_t 表示包含候选术语 t 的嵌套术语数量.与 C-value 方法一样,Basic 方法仅适用于多字术语;不同的是,Basic 方法对于经常嵌套在较长词串中的候选术语 t,不会降低其评分,反而会增加 t 的重要性评分.

2015 年,Astrakhantsev[63]在 Basic 方法的基础上进一步进行修改,提出了可以定制度量候选词领域性的 ComboBasic 方法,抽取出更具体的术语,公式如下:

$$ComboBasic(t) = |t|\log f(t) + \alpha e_t + \beta e_t'$$

其中 e_t'表示候选术语 t 中所包含候选词的数量.通过增加 β 值,可以提取更具体的术语.ComboBasic 方法之所以引入 e_t',是因为通过简单减小 Basic 方法的参数 α,无法达到定制衡量候选词领域性的目的.Basic 较多关注提取频繁和较长词串,忽略特定于领域的候选术语;而 ComboBasic 方法解决了这一问题.

Frantzi 等人[18]提出的 NC-value 方法则是通过引入'术语上下文'的概念来扩展 C-value 方法,其基本假设:(1)特定领域的语料库通常有一系列出现在术语附近的“重要”单词;(2)在这些词汇背景下出现的候选术语应该被赋予更高的权重.因此,NC-value 方法首先计算语料库中术语候选词的 C-value 评分并排序,然后对每个候选术语生成'语境词列表',且每个语境词都有权重.最后,根据该候选术语的 C-value 评分和'语境词列表'计算 NC-value 值.NC-value 方法在抽取高频术语方面比 C-value 方法表现更好,准确率达到 75.70%.

之后,研究者逐渐尝试结合多种术语抽取方法取其优点的混合策略.2011 年,You 等人[64]使用词性规则自动生成算法产生规则模板来获取术语候选词集合,之后利用 C-value,TF-IDF,TermExtractor 三种方法的结果进行加权投票并排序候选术语,实验表明多特征融合优于单一特征抽取.2012 年,He[65]提出结合候选术语分布度、活跃度以及主题度的多策略术语抽取方法,该方法能够不增加计算量,同时提高领域术语的抽取质量.2014 年 Lossio-Ventura 等人[22]人提出使用语言规则及 IDF、C-value 混合的 LIDF-value 方法,克服了候选术语频率信息不足的缺点.2015 年,Li 等人[66]提出了结合信息熵和词频分布变化的术语抽取方法,应用在汽车领域语料库抽取出 1300 个术语,准确率达到 73.7%,对低频术语也有较好的抽取效果.

2016 年,Stanković 等人[67]通过分析塞尔维亚语的特点,在使用语言及统计方法混合的基础上,借助外部电子词典和常见句法结构来提高术语抽取准确率.2017 年,Dong 等人[68]针对特定领域规则更新速度慢、文本特征考虑不足等问题,提出一种基于文本特征和复合统计量(TF-IDF 和信息熵)结合的中文术语抽取方法.同年,Li 等人[13,15]提出解决嵌套术语不合理分割和消除术语次序敏感的策略,确保所抽取短语的恰当性及完整性.与 Liu 等人提出的 SegPhrase 方法[12](见 2.5.1 章节)不同的是,Li 等人采用轻量级单元性度量方法结合短语分割技术,即在度量短语内部粘合程度时,使用动态规划组合所有顺序,避免生成不完整的术语;在结合短语分割模型时,额外考虑短语间的'内部间隔(inter-isolation)',保证术语可以恰当划分.实验表明,在前人的思想基础上加入此两点创新,明显提升了术语抽取效果.

混合自动术语抽取方法主要利用语言学、统计学、主题信息等方法的不同特征组合抽取术语,兼具多种方法的优点,具有较好的领域独立性和语言无关性,进一步提高了术语抽取的准确率和召回率,如 Pazienza 等人[69]已验证:混合抽取方法可以提升术语召回率.但是该类方法通常使用投票算法或启发式算法线性组合各类术语特征,算法理论过于单薄,没有考虑特征间的非线性关系,缺乏特征组合深度,使得抽取效果远差于基于机器学习的 ATE 方法,Fedorenko 等人[26]通过对比实验验证了这一结论(详见 3.3 章节).

2.4 基于外部知识的抽取方法

基于外部知识的术语抽取方法主要利用外部资源,如参考语料库、维基百科等来提高术语抽取的准确率.其基本思想:某一特定领域的术语候选词在该领域中的分布一般与通用领域(general domain)的分布有明显的不同,候选术语在特定领域的出现次数比在通用领域的出现次数更加频繁.因此可以使用外部资源作为参考,通

过对比词或词组在目标语料库和在外部资源中出现频率的差异,将术语候选词与常用词、无意义的词串区分开,从而达到术语抽取的目的.

其中,参考语料库(Reference Corpus)是指包含通用领域或其他领域的文档集合、电子书、新闻集以及语言学家创建的语料库,例如开放的美国国家语料库(Open American National Corpus,简称 OANC)和英国国家语料库(British National Corpus,简称 BNC)等.

Ahmad 等人[70]于 1999 年提出 Weirdness 方法,该方法认为术语候选词在特定领域语料库中的分布不同于在参考语料库中的分布,因此将特定领域语料库中术语候选词 t 的归一化频率与 t 在参考语料库(例如 BNC)中的归一化频率进行比较,证明在目标语料库中频繁出现的候选术语具有更高的'领域特异性'(domain specificity),更有可能是真正的术语.此外,Weirdness 方法在识别低频术语方面富有成效.公式如下:

$$Weirdness(t)=\frac{TF_{target}(t)\cdot\left|Corpus_{reference}\right|}{TF_{reference}(t)\cdot\left|Corpus_{target}\right|}$$

其中,$TF_{target}(t)$表示特定领域语料库中候选术语 t 的词频,$TF_{reference}(t)$表示参考语料库中 t 的词频.Weirdness 方法还提供了处理词汇表外(Out of Vocabulary,简称 OOV)候选术语的方法,即通过获取候选术语 t 中每个单词 w_i 的 Weirdness 评分,然后逐个相加得到候选术语 t 的 Weirdness 评分.

不少学者在 Weirdness 方法基础上进行研究改进.Relevance 方法[71]是通过增加候选术语 t 出现的文件数量 $DF_{tartget}(t)$来扩展 Weirdness 方法,这一改变使得三类术语候选词的评分受到影响:(1)目标语料库中较少出现的候选词;(2)出现在文档中的候选词;(3)参考语料库中频繁出现的候选词.

$$Rel(t)=1-\frac{1}{\log_2\left(2+\frac{TF_{target}(t)\cdot DF_{target}(t)}{TF_{reference}(t)}\right)}$$

Domain Specificity 方法[72]利用改进的 Weirdness 评分除以候选术语的长度实现归一化处理,优势是能够较好识别 OOV 候选术语,公式如下:

$$DomainSpecificity(t)=\sum_{w_i\in t}\log\frac{P_d(w_i)}{P_c(w_i)}\Big/|t|$$

其中,$|t|$表示候选术语 t 包含单词的个数,$P_d(w_i)$表示候选术语 t 中各个单词 w_i 出现在目标语料库中的概率,$P_c(w_i)$表示候选术语 t 中各个单词 w_i 在参考语料库中的概率.

之后开发的 GlossEx 系统[72]和 TermEx 系统[73]分别从不同维度扩展了 Weirdness 方法.2002 年,Park 等人[72]提出了 GlossEx 系统,该系统在改进 Weirdness 方法的同时,增加了词汇单元性度量.具体来说,GlossEx 系统基于两种启发式方法:(1)评估候选术语'领域特异性'(domain specificity)程度的方法,使用 Domain Specificity 公式度量;(2)度量候选术语内部凝聚程度的方法,类似 2.2.1 章节提到的 unithood 度量.2007 年,Sclano 等人[73]通过线性组合的方式对 GlossEx 系统进一步扩展,提出了 TermEx 系统.TermEx 系统额外增加了'领域共识 DC'(见 2.2.2 章节)的度量,使得目标语料库中均匀分布的候选术语具有更高的权重,有助于抽取出分布均匀的术语.总体来说,GlossEx 系统和 TermEx 系统综合考虑 unithood 度量、termhood 度量以及参考语料库三个维度的特征来抽取术语,并取得了较好的抽取效果.

最新研究成果中,Lopes 等人[74]提出 tf-dcf(term frequency-disjoint corpora frequency)方法,该方法认为候选术语评分应与其在多个参考语料库的出现频率成反比,并以此评估候选术语的领域特异性.2018 年,Mykowiecka 等人[75]针对已抽取的术语排序列表仍包含非领域短语的问题,提出了对比多个参考语料库并通过术语上下文过滤不相关短语的方法,最高准确率达到 75%.

另一个较为重要的外部知识是维基百科(Wikipedia),不仅支持多语言,涵盖众多领域,而且知识内容和条目持续更新扩充,同时能够满足各种规模目标语料库的需求.尤其对于较小规模语料库非常实用,因为较小规模语料库自身的统计信息不足以区分术语和非术语,需要使用维基百科来提供特定领域的统计信息.

Vivaldi 等人[19,76]于 2010 年较早使用维基百科作为语义知识资源来抽取特定领域中的术语.首先,对于每个术语候选词,找到它所对应的所有维基百科文章(由于一词多义,每个候选词可能对应多篇文章);然后,确定每篇文章所属的所有类别;之后,对于每个类别,递归遍历类别图(仅跟踪到达顶级类别的链接),直到达到指定的域边界或最顶层类别.最后,使用所找到的路径数量来对术语候选词进行评估.该方法的优点是易扩展,能够较为容易得应用到维基百科覆盖的领域及语言;缺点也很明显,不能正确度量没有出现在维基百科中的候选术语.

2014 年 Astrakhantsev 等人[20]提出了 LinkProbability 方法,使用维基百科作为参考语料库,并将候选词的概率标准化为候选术语 t 在维基百科中以超链接标题出现频率与其在维基百科中出现总频率的比率.其中比值特别小或没有出现在维基百科中的候选术语,LinkProbability 分数设置为 0,进行过滤处理.此方法对于区分常用词或词组非常有效,因为它们不特定于识别的领域,LinkProbability 分数较小.

2017 年,Haque 等人[77]提出融合外部知识库的双语术语抽取模型,该模型使用 n-gram 过滤及统计学方法对源端及目标端进行候选术语抽取排序,之后借助外部知识库--维基百科跨语言链接数据库提升源端重要术语的排序位置,实验表明该模型在英语到西班牙语的 8 个测试数据集中术语抽取效果均达到最先进水平.

基于外部知识的自动术语抽取方法主要通过对比词或词组在目标语料库和外部资源中出现频率的显著差异进行术语抽取.该类方法有助于弥补因目标语料库质量不佳或统计信息不足造成术语抽取效果差的缺陷,通常借助外部知识来获取目标语料库之外的有效特征,解决低频术语抽取问题,提高术语抽取准确率.但美中不足的是并非所有领域都可以使用外部知识资源,一些特定专业领域并无可用的外部资源.

2.5 基于机器学习的抽取方法

基于机器学习的自动术语抽取方法可分为 3 类:有监督方法,弱监督方法,和远程监督方法.三类抽取方法都需要先标注数据后进行有监督学习,区别在于每类方法所需人工标注数据的规模不同.其基本思想:在给定训练数据的情况下,基于机器学习的抽取方法通常会将训练实例转换成一个特征空间,特征空间融合多种自然语言特征来提高术语抽取的准确率.这些特征可以是基于语言学的特征(例如 POS 模式、特殊字符的出现等),也可以是基于统计学的特征或者是两者的组合特征,还可以是来自外部知识库的特征.其中,基于统计学的特征通常使用统计学自动术语抽取方法(例如 TF,TF-IDF)作为指标来计算训练实例的分数.

2.5.1 有监督方法

有监督方法将术语抽取看作是二分类问题,判断语料库中的词串(词或短语)是或者不是术语.这种方法必须先提供已标注好的术语作为训练集;然后利用训练集来训练一个术语抽取模型;最后将训练好的模型应用到所有术语候选词中,得到每个候选术语的类别分数,再将其分为术语或非术语.

2009 年,Zheng 等人[78]最先使用随机条件场(conditional random fields,简称 CRF)模型来抽取领域术语,模型采用六种组合特征:POS 标签,语义信息,左信息熵,右信息熵,互信息和 TF-IDF,在军用材料领域进行测试,其准确率,召回率和 $F1$ 值分别达到 79.63%,73.54%和 76.46%.2010 年,Zhang 等人[79]同样采用 CRF 模型,将经过处理的句法信息作为新特征加入到语料中.当训练集大小是测试集 8 倍时,该方法在新术语上抽取效果最好.2011 年,Zhang 等人[80]使用综合语言学特征和统计学特征的 CRF 模型,同时考虑候选词所在句子的术语度,采用一体化方法来抽取术语.实验结果表明多特征融合较单个特征更为有效.Loukachevitch[81]则认为术语抽取模型需要基于更多类型的特征,因此提出了使用三类不同特征来提取双字术语的模型:包括基于特定领域的特征,基于搜索引擎的特征以及基于同义词库的特征.2013 年 Conrado 等人[82]提出一种利用丰富特征集合进行自动术语抽取的机器学习方法,使用的特征分为两类:(1)从目标语料库中获取语言学、统计学及混合知识的特征;(2)获取目标语料库与通用语料库之间的对比特征.该方法提升了 3 个测试语料的准确率和 $F1$ 值,证明了加入不同层级的特征可以有效改进 ATE.2017 年,Yuan 等人[83]针对跨领域跨语言的术语抽取问题,提出使用 10 种统计学方法作为特征的机器学习方法,通过 6 种机器学习算法在不同领域语料库上进行评估,验证 Yuan 的方法在通用 ATE 任务具有稳健性和较高效率.同年,Liu 等人[54]针对较长术语易被错误切分的问题,提出了一种基于术语长度和语法特征的自动抽取方法.首先利用支持向量机 SVM 结合约束规则抽取出术语候选词集合,然后使用词长比、领域相关性、领域共识 3 种 termhood 特征加权计算出候选词评分,过滤出真正的术语.

除了上述多种特征融合方法外,Liu 等人[12]针对术语原始词频不能正确评估术语真实质量的缺陷,于 2015 年提出 SegPhrase 方法,首次将短语分割的思想与基于机器学习的术语抽取方法相结合,取得了不错的抽取成果.SegPhrase 方法的创新点一:定义了什么是高质量短语,从四个维度给出度量方法,使得术语质量衡量完善.

(1) 普遍性:高质量短语应当多次在文档中出现.普遍性本质上是指短语的词频,因此文中使用 *n*-gram 过滤(n<=6)+词频大于 30 来初步筛选术语候选词;

(2) 一致性:高质量短语的出现次数要高于普通词的平均出现次数.一致性主要指短语内部的固定搭配程度,文中使用点互信息、KL 距离两个 unithood 特征来进行度量;

(3) 信息性:高质量短语在特定领域中表示有意义的词组(例如,this paper 则不具备信息性).信息性主要指短语特定于领域的程度,文中使用去除停用词、IDF、候选词大小写 3 个特征来进行度量;

(4) 完整性:高质量短语在句子中应表示一个完整的语义单元,不是机械的切分.文中采用短语分割技术来进行句子的最优分割,从而获取语义完整的短语.

SegPhrase 方法的创新点二:形成了一个整体可迭代可裁剪的框架,可伸缩性很强(如图 3).框架流程如下:

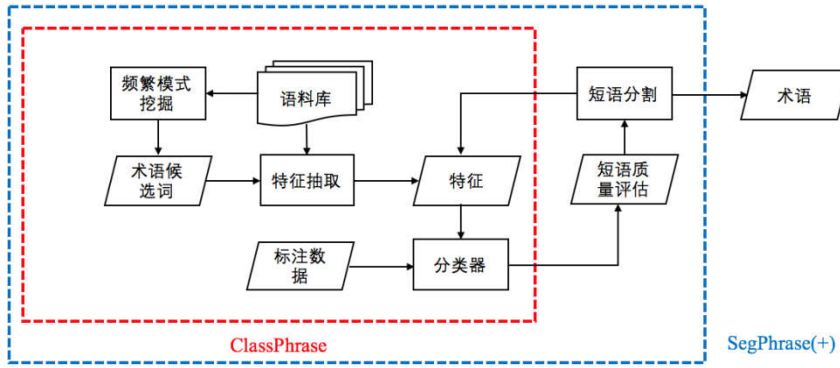

Fig.3 The basic framework of SegPhrase method[12]

图 3 Segphrase 方法的基本框架[12]

(1) 频繁模式挖掘.生成频繁术语候选词集合

(2) 短语特征的提取.将提取好的特征(一致性及信息性共 5 个特征)输入分类器中,得到一个预估的术语候选词质量,即短语质量评分(大于 0.5 为高质量短语,小于 0.5 为劣质短语);

(3) 短语分割阶段.使用第(2)步生成的短语质量分数来进行短语分割,得到短语纠正后的频率;

(4) 短语纠正特征的提取.短语分割后可提取出 2 个关于纠正频率的特征放入分类器的特征集合(feature set)中,提升分类器下次预估的准确性;

(5) 过滤低纠正频率的短语.按照纠正后的分数排序输出短语列表.

其中,步骤(1)~(3)对应图 3 中红框部分 ClassPhrase 方法,可以裁剪成常规的机器学习方法使用;步骤(1)~(5)对应图 3 中蓝框部分 SegPhrase 方法,可以多次迭代步骤(4)(5),得到准确率更高的短语列表.本篇论文提出的 SegPhrase 方法在大型语料库中只需极少的标注数据便可达到与领域专家相同水平的准确率,且所用抽取时间极少,与语料库规模成正比.后续 Li 等人的方法[13,15]及 AutoPhrase 方法[12]都是在此方法的基础上提出的.

2.5.2 弱监督方法

有监督自动术语抽取方法需要大量的标注数据,但是获取带标注的数据集耗时长,成本昂贵,不易实现.因此,近年来,研究者更多将注意力转移到半监督和弱监督的术语抽取方法上,以期解决训练数据的标注问题.弱监督方法不像有监督方法需要大量标注好的训练数据,只需要少量的标注数据作为训练集,利用这些数据训练抽取模型,然后使用抽取模型再进行未标注候选词集合的术语抽取,人工或自动对抽取结果进行甄别,最后将结果正确的标注数据加入训练集中,再继续训练模型.

Yang 等人[84]于 2010 年提出使用容错学习和联合训练的方法,从噪声数据中迭代构建种子训练集.首先基

于两个无监督抽取算法(TF-IDF 和基于分割符的算法)自动生成两组候选词种子集合,种子集合包含排名最高的 500 个术语作为正样例集和排名最差的 500 个非术语作为负样例集,然后使用两组不同的种子集进行后续的有监督学习分别训练两个分类器,其中,分类器采用支持向量机 SVM 算法,选择五个术语特征:候选词词频 TF,POS 标签,词分隔符,候选词中第一个词和最后一个词特征.接着使用训练好的分类器对所有候选词进行标注,抽取两组具有最大和最小置信度的实例组进行双重检查后(当两个分类器将不同的标签分配给同一候选词时,该候选词从实例组中移除)分别添加到对应的种子集中,用作下一次迭代训练的正负样例集合.Yang 的方法不需要领域知识及提前标注训练数据,因此比传统有监督抽取方法或基于领域知识的方法更容易迁移到不同领域.相较于有监督的 SVM 算法,容错学习方法显著提高了术语提取的准确率.

Astrakhantsev 等人[20]遵循类似 Yang 等人的思路:使用种子方法 ComboBasic 抽取前 100～300 个术语候选词作为正样例集,并将所有其他术语候选词视为未标记的实例,来构建基于 PU(Positive-unlabeled,PU)学习算法的分类器模型,选择 C-value,DomainCoherence,Relevance 作为术语特征来训练分类器,之后将分类器应用于每个术语候选词获得置信度分数,并加入相应样例集.此外,Maldonado 等人[85]提出一种针对在线增量语料库的再训练方法,该方法将领域专家的验证纳入弱监督学习循环,使用新的训练数据迭代训练分类器,新训练数据结合了手动标记的示例(通过专家验证)和先前训练模型中已标记的实例.Aker 等人[86]将弱监督方法应用于双语术语抽取任务,使用平行语料库将已从源语言资源提取的术语投影到不同的目标语言,训练目标端的术语抽取模型.

2014 年,Judea 等人[10]提出一种无监督方式自动生成高质量训练数据的术语抽取模型,采用启发式算法生成专利技术领域的正负样例集:正样例集是指在专利文档中直接提到且分布在数字标记前的词或短语,而负样例集是指只在一个专利文档中出现的词或短语、专利引用及测量单位(例如 3cm).之后采用 74 个特征来训练有监督分类器(逻辑回归和条件随机场):包括 POS 标签、上下文特征、候选词出现次数以及基于字符串度量的特征等.最终在自动生成训练集上训练两种分类器,其击败了最先进的基线方法,并取得了 $F1$ 值超过 75%的好成绩,说明 Judea 提出的自动标记方法能够生成高质量的训练数据,不足之处是用于标记正面样例的方法只适用于专利领域,不能迁移到其他领域或语言上使用.论文还发现:术语抽取准确率在很大程度上取决于术语边界的正确识别,可以通过改进候选术语的识别来大大提高抽取效果.

2016 年 Wang 等人[87]分析钢铁冶金领域中文术语的基本特征,提出了基于字角色标注的机器学习术语识别模型.该模型使用自动构建核心词汇库代替人工标记数据,解决语料库中训练数据不足的问题.以 CRF 模型获取新术语为基础,使用增量迭代方式重复 CRF 术语抽取过程;并根据合成规则构造的新术语,经领域专家确认后添加至核心词汇库,最终获取大量的训练数据,术语抽取效果 F1 值达到 94%.

2.5.3 远程监督方法

远程监督方法不需要人工标注的训练数据,主要利用远程对齐外部知识库(例如维基百科,WordNet 等)来对术语候选词集合中的候选词进行自动标注,得到大量的正负样例,形成训练集.远程监督方法在自动术语抽取领域的应用还比较少,最新的研究是 Shang 等人[14]于 2018 年在 Liu 等人[12]SegPhrase 方法的基础上,提出了支持多种语言无需人工标注训练数据的 AutoPhrase 方法,解决术语抽取需要专家来设计规则或是标记数据的问题.图 4 描述了该论文用于术语抽取的远程监督体系结构.

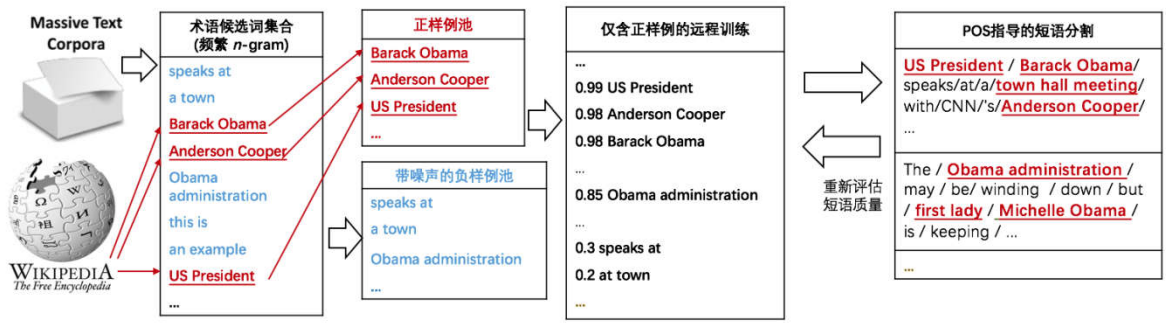

Fig.4 The basic framework of AutoPhrase method[14]

图 4 AutoPhrase 方法的基本框架[14]

与 SegPhrase 方法不同的是,AutoPhrase 方法引入两种新技术:(1)远程训练技术.使用通用知识库(例如维基百科,Freebase 等)来标记候选词集合中的正样例,形成正样例池(positive pool);剩下的候选词自动构成带噪声的负样例池(negative pool),之后通过分类器集合来降低噪声数据的影响.这一过程可使架构摆脱额外的手工标记工作,实现真正的术语抽取自动化;(2)POS 指导的短语分割技术.在 SegPhrase 方法短语分割的基础上,利用 POS 标签中浅层句法信息来指导短语分割模型,更准确地定位短语的边界(图 4 最右部分),从而提高模型抽取效率.最终实验证明:AutoPhrase 方法在性能和效率方面均超越 SegPhrase 方法,不需要手动标记训练集,并增加了单字术语的抽取,将术语抽取召回率提升约 10%～30%.

基于机器学习的自动术语抽取方法是目前术语抽取领域的研究热点,主要利用多种混合特征及分类器来抽取术语.该类方法中有监督抽取方法依赖于人工标注得到训练集,准确率较高,无需人工制定规则,具有较好的实验价值.但人工标注的数据集耗时长,成本昂贵,标注数据量有限,可扩展性低,削弱了模型的领域独立性,使其跨领域泛化能力较差.而远程监督抽取方法则是采用远程对齐外部知识库自动标注数据集,极大节约了人力成本,增强了领域通用性,但是远程标注也带来了很多错误标注数据,导致错误标签的误差逐层传播,影响术语最终抽取效果.相较于有监督方法和远程监督方法,弱监督抽取方法是目前使用较多的术语抽取方法,具有明显的优势,只需少量的标注数据便可得到增量扩展的训练集;并在人力成本可控条件下不断优化训练模型,得到最先进的术语抽取效果;同时,少量标注使得模型的通用性更强,跨领域迁移能力更大.总体而言,基于机器学习的 ATE 方法虽取得了不错的成果,但是还不够成熟,仍依赖人工筛选术语特征及标注数据,需要更进一步得研究探索.

2.6 基于深度学习的抽取方法

基于深度学习的自动术语抽取方法主要结合最新的深度学习技术来进行自动术语的抽取,是一种数据表示的特殊机器学习方法,可解决抽取术语中人工挑选最佳特征工程的问题.其基本思想:通常将候选术语或整个句子的词嵌入表示(word embedding representation)作为输入,喂给特定的深度学习模型(例如深层神经网络 DNNs、深层信念网络 DBNs、递归神经网络 RNNs、深层递归神经网络 DRNNs),然后由多个处理层组成的深度计算模型学习出具有多个抽象级别的候选术语表示,最后对该表示进行术语类别划分.

近年来,深度学习技术为各种 NLP 任务提供了多种解决方案以及接近专家水平的准确率.因此,深度学习模型在自动术语抽取任务中得到了应用.最早将深度学习引入自动术语抽取领域的研究者[21],将术语抽取看作是二分类问题,需要先抽取大量的候选术语并进行判断.2016 年,Wang 等[21]人提出了一种深度学习模型--弱监督的联合训练方法,使用两个深度学习模型 LSTM 和 CNN 作为分类器,分别学习候选术语的不同表示,无需手动选择特征.首先,两个分类器在少量标记数据上进行训练,然后独立对未标记数据子集进行预测,最后将置信度最高的 n 个候选术语添加到训练集中以重新训练分类器,迭代 k 轮后结束.实验结果表明:即便在训练数据有限的情况下,结合深度学习的联合训练方法达到了与有监督机器学习相当的准确率,表现了该模型具有较强的性能.基于此,Khosla 等人[88]在 2019 年选择同样的深度学习联合训练模型来抽取术语.不同的是,该模型在输入层新添加了字符级的 n-gram 嵌入,使用 CNN 和全连通网络作为分类器.但这类模型存在训练耗时长和添加错误候选术语的问题,会加重标注数据不均衡现象.

同年,Gao 等人[89]为了抽取嵌套术语,提出了深度学习端到端(End-to-End)模型来学习候选术语的向量表示,然后将候选术语向量表示喂入分类器,得到每个候选术语的评估分数,将其分为术语或非术语.其中,术语向量表示融合了多种类型信息,包括术语拼接表示信息、重要词表示信息、术语头部尾部信息、句子表示信息等.

另一些研究者将自动术语抽取转化为序列标注问题,如图 5 所示,先对句子中的每个单词进行字符向量表示和词嵌入向量表示,将这些向量表示拼接之后喂入深度模型(例如 LSTM、GRU),最后经过 CRF 层的处理得到每个单词对应的标签.2018 年,Zhao 等人[90]将术语抽取看作是序列标注问题,提出了 Bi-LSTM-CRF 深度学习

模型,抽取中文文档中每个字的词向量特征、词性特征和实体特征作为输入数据,经过双向多层隐藏层处理后,使用 CRF 方法将字映射为{B,I,O,E,S}标签之一.并针对所在领域不存在大规模成熟语料,论文使用增量自训练算法 Viterbi 来降低人工标注代价,提升术语抽取效率.同年, Kucza 等人[91]构建了一个基于深度学习术语抽取系统,通过在字序列上使用 BILOU 标签方案执行序列标记来识别术语,同时使用了不同类型的递归神经网络和字嵌入方法来测试抽取效果.但这类方法的最大缺点是不能输出术语排序列表,无法同传统抽取方法进行对比.

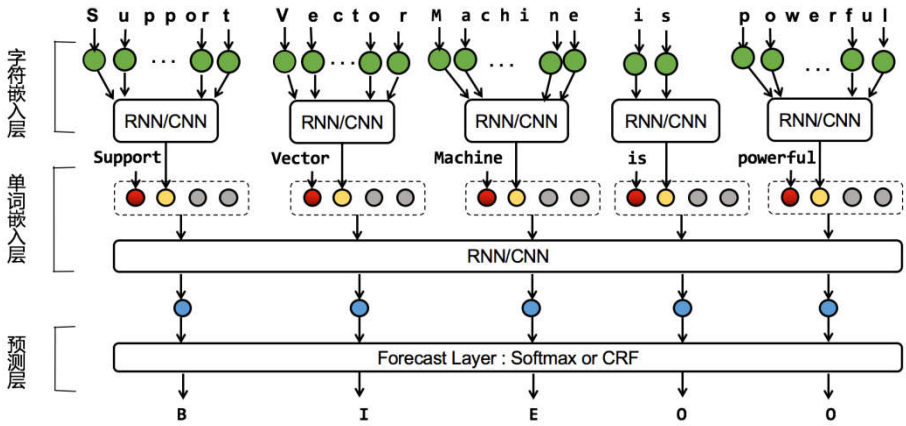

Fig.5 The basic framework of sequence labeling issue

图 5 序列标注问题的基本框架

基于深度学习的自动术语抽取方法主要利用深度学习模型来抽取术语.该方法无需人工筛选术语特征,减少了昂贵的人工成本,并有助于将候选术语和上下文信息结合起来,以词嵌入向量表示融合更多类型的特征,从而达到较好的术语抽取效果,尤其适合超大文档集合.但该方法的缺点也很明显,依赖于复杂的深度学习模型,需要非常大量的标注数据或标注句子(对应序列标注方法),较长的训练时间,且模型的跨领域泛化能力较弱.当前,一些论文结合 *n*-gram 过滤和标准术语表来进行训练数据的标注,但会带来正负样例不均衡现象(正样例仅占 10%左右).即便存在诸多问题,但基于深度学习的自动术语抽取方法仍表现出了显著的术语抽取性能.毫无疑问, 基于深度学习的 ATE 方法将会成为下一个研究热点.

2.7 基于语义相关的抽取方法

基于语义相关的自动术语抽取方法主要利用词组间的语义关系,来改进语义相关术语的排名,达到提高术语抽取准确率的目的.

词与词之间的语义关系包括语义相似性(semantic similarity)和语义相关性(semantic relatedness),语义相似性关系例如汽油和柴油之间的关系,语义相关性关系例如鼠标和键盘之间的关系.所以,量化术语之间的关联程度,不仅要考虑语义相似性,还要考虑术语和术语之间可能存在的语义相关性关系.详细来说,可以从两个角度来度量术语之间的关联程度[92](两个角度略有重叠,并非完全独立):

- 浅层语义度量: 根据 Bikel 等人的论述[93],浅层语义其内涵仅涉及句子问题的两个方面:词义和部分词项的语义关系等.因此,将浅层语义度量分为领域关键信息度量和上下文相似性度量.

  1) 领域关键信息度量:提取领域关键信息(例如一组概念,或一组种子术语等)作为特定领域的表征, 通过计算候选术语与领域关键信息的相关性进行度量;

  2) 上下文相似性度量:通过术语与周围其他词在一定距离窗口内的同现频率(co-occurence)或分布式表示来进行度量;

- 深层语义度量: 以知识库为基础度量,使用知识资源(例如同义词典,语义网络或分类法)度量术语间的相似及相关关系[94].深层语义度量是通过链接关系构建术语与外部知识库实体之间的映射,进而使

用整个知识库进行术语含义的消歧、扩展以及深层语义理解.

(1) 浅层语义度量

在领域关键信息度量方面,Astrakhantsev 等人[20]认为领域是由一组含义紧密相关的特定概念表征,而一个真正的术语应该与领域关键概念间具有高度的语义相关性.因此,于 2014 年提出 KeyConceptRelatedness(KCR)方法,利用领域关键概念来衡量候选术语的质量.该方法首先利用 El-Beltagy[95]提出的方法提取 N 个领域关键概念(N=200);然后筛选术语候选词集合,通常选择超过某个评分阈值的前 n 个术语候选词.之后,对每个候选术语,使用 Dice 系数算法计算其与 N 个领域关键概念间的语义相关性,术语的最终得分是前 k 个(k<n)高相似性分数的平均值.

与 KCR 思路相似,Bordea 等人[62]提出了领域一致性方法 PostRankDC,该方法使用自动构建的域模型来替换 KCR 方法中的'关键概念',域模型由候选术语上下文中重要的单词和短语组成.使用“标准化 PMI”计算候选术语与域模型中高排序词之间的语义相关性.实验表明 PostRankDC 方法相比基于参考语料库的 ATE 方法,更加适合抽取具体的术语.

2018 年,Yu 等人[96]针对专利领域中无法过滤高频非术语及抽取低频术语的问题,提出了融合术语部件相似信息的术语抽取方法.在抽取术语候选词集合的基础上,利用与候选术语有相同术语部件的相似候选术语信息,评估候选术语成为术语的可能性.实验表明,该方法有效提高了专利术语抽取的准确性(准确率达到 83.56%,提升约 32%).同年,Lahbib 等人[97]将术语相关性的思想应用到双语术语抽取领域,首先使用 TF-IDF 方法抽取出术语种子集合,然后计算候选术语与种子术语集之间的语义相关性度量分数,最终抽取到特定于领域的源端术语.

在上下文相似性度量方面,Li 等人[98]针对低频术语抽取效果差的问题,于 2018 年提出了基于术语嵌入向量的抽取方法.该方法主要根据术语的上下文来衡量术语的质量，首先使用多种 ATE 方法生成候选术语集合,然后将每个候选术语的出现上下文信息汇总到嵌入向量表示中,最后利用术语嵌入向量来评估可能概念的质量,并提出了四种度量候选术语相似性的方法:上下文共性、上下文纯度、上下文泛化以及上下文链接能力.其中,前三种方法属于浅层语义度量方法,最后一种方法属于深层语义度量方法(详见本章节第(2)部分).

上下文共性(Context commonness): 衡量一个候选术语 c 的上下文与其他术语 c'出现上下文的的共性.术语“支持向量机”通常与其他机器学习算法,如“随机森林”或“逻辑回归”等出现在共同的上下文词句中.而一个没有意义的候选术语(例如向量支持机)可能只出现在少量上下文中,且这些上下文与许多术语不相关.公式如下:

$$\sum_{c'\in V_c, c'\neq c} \mathbb{I}(\langle \theta_c, \theta_{c'} \rangle > \kappa)$$

其中 c 表示将度量的候选术语,V_c 表示候选术语集合,$\langle \theta_c, \theta_{c'} \rangle$ 表示使用余弦相似函数度量术语嵌入向量间的相似性, k 表示相似性阀值.因此,上下文共性是指与术语 c 相似的其他术语个数.值越高,代表与 c 相似上下文关联的候选术语越多,因此术语 c 的上下文具有共性.

上下文纯度(Context purity): 衡量候选术语 c 上下文的的内部差异.对于具有明确含义的术语(例如查询优化器),其使用的上下文相对具体,且常与相同类型的候选术语相关联.而一般性短语(例如性能)的使用上下文将更加多样化,因此术语纯度降低.

$$\sum_{c'\in V_c, c'\neq c} \langle \theta_c, \theta_{c'} \rangle \Big/ \sum_{c'\in V_c, c'\neq c} \mathbb{I}(\langle \theta_c, \theta_{c'} \rangle > \kappa)$$

上下文纯度是指术语 c 与其他候选术语之间的平均相似性.值越高,代表术语 c 的使用上下文约可能与特定类型的术语相关联.

上下文泛化(Context generalizability): 衡量候选术语 c 的整个字串是否能够泛化表示更具体的术语.假设识别出两个或两个以上的术语(例如“模糊支持向量机”、“孪生支持向量机”、“一类支持向量机”),它们的使用上下文都与“支持向量机”相似,则进一步证实“支持向量机”是一个真实术语,而非随机组合的词序列,公式如下:

$$\max_{\theta}\left(\sum_{c'\in V_c, c\prec c'} \mathbb{I}(\langle\theta_c,\theta_{c'}\rangle>\kappa)\text{-}1,0\right)$$

上下文泛化是指与术语 c 相似的其他术语个数,且 c 与其他术语部分字段相同.

此外,Lossio-Ventura 等人[22]使用基于同现频率的 Dice 系数函数来计算候选术语之间的语义相关性.Khan 等人[99]则结合 C-value 和 TF-IDF 方法提取排名靠前的术语候选词子集,利用基于词嵌入模型的余弦定理相似性度量术语间的语义相似评分.

(2) 深层语义度量

2016 年,Conde 等人[100]提出了一种结合 Wikipedia 实体链接的术语抽取方法 LiTeWi,使用 Wikipedia 作为语义知识库来抽取教育领域的重要术语.LiTeWi 方法首先利用 TF-IDF、C-value 及浅层解析器等多种 ATE 方法来抽取候选术语,构建一个大型术语集合;然后将每个术语映射到一个或多个维基百科文章,无映射文章的术语被删除;随后对多语义的术语(对应多个维基百科文章)对进行语义消歧,挖掘出候选术语最匹配的语义,并识别映射到同一语义(即同一维基百科文章)的术语进行合并;最后,过滤掉与领域无关的术语.

LiTeWi 方法最大的特点是利用 Wikipedia 作为语义知识库对多语义的术语进行消歧,合并同一语义的术语.实验表明,候选术语集合中约 25%的术语存在多种语义,使用知识库语义消歧后,改进了语义相关术语的排名,提高了领域术语抽取的准确率.基于此,Khan 等人[99]采用相同的语义消歧思路,使用领域同义词术语表作为语义知识库,对语义图中的同义术语(即顶点)进行合并,从而极大改善了低频术语的排名.

2018 年,Li 等人[98]提出了一种基于术语嵌入向量的方法来抽取高质量术语,该方法在评估术语质量时利用外部知识库的实体集合来确认候选术语.即知识库中的实体被认为是“高质量术语”,若候选术语与知识库中的一个或多个实体具有高相似性,则认为该候选术语具有较高质量.

上下文链接能力(Context link-ability):衡量候选术语 c 上下文与知识库中高质量实体的相似程度.假设 Runina Murina(一种蟾蜍)通常出现在类似于 frog(青蛙)的上下文中,而 frog 存在外部知识库内,是一个高质量术语,则进一步证实 Runina Murina 也是一个高质量术语.公式如下:

$$\sum_{c'\in V_{ext}, c'\neq c} \mathbb{I}(\langle\theta_c,\theta_{c'}\rangle>\kappa)$$

其中,V_{ext} 表示知识库中高质量实体集合.上下文链接能力是与术语 c 相似的知识库中高质量实体个数.

基于语义相关的自动术语抽取方法主要利用术语之间的关联关系(边的特征)来抽取术语.该类方法考虑了术语与术语之间语义关系,可以融合更多的其他特征,因此取得了较好的抽取效果.但是,术语间的语义关系依赖于领域关键概念的获取或分布式表示的学习以及外部知识库的构建,若特定领域的关键概念选取失败或不存在特定领域的知识库,则影响术语抽取质量.因此,常采用分布式相似度量作为基于语义相关的自动术语抽取方法的质量评估算法.

2.8 基于图的抽取方法

基于图的自动术语抽取方法是最近几年开始在术语领域流行的一类无监督抽取方法.该类方法的灵感来源于 PageRank 中网页重要度的排序方法.2004 年,Mihalcea 等人[4]最先将 PageRank 思想应用于自然语言处理领域,提出可以抽取关键单词的 TextRank 方法.

基于图的术语抽取基本步骤如下:

(1) 文档图形化表示.可以将语料库中的所有文档表示为一个图,也可以将每个文档单独表示为一个图.其中,顶点表示文档预处理后生成的单词或短语,边表示单词或短语在滑动窗口中的共现关系,或者表示单词或短语间的语义相似关系.

(2) 评分函数定义.使用不同的排序方法对图中的顶点进行评分.

在基于图的抽取方法中,最常使用 PageRank 分数作为排序指标.PageRank 算法[101]采用随机游走思想利用图形结构递归地计算图中各顶点的重要性,即先模拟用户通过点击链接随机访问图中各顶点的行为,然后计算稳定状态下各顶点的随机访问概率.

对于图结构中的任一顶点 V_i,其 PageRank 重要性分数计算方法如下:

$$S(V_i)=(1-d)+d*\sum_{j\in In(V_i)}\frac{1}{|Out(V_j)|}S(V_j)$$

其中,d 为阻尼因子,通常取 0.85;顶点 V_i 的重要性分数取决于顶点 V_i 邻居顶点的数量及每个邻居顶点的重要性分数.

受 PageRank 算法的启发,Mihalcea 等人[4]于 2004 年首次将该算法引入到自然语言处理领域,并提出了 TextRank 方法.TextRank 方法尝试构建基于词间共现频率的文本语义图,迭代基于图的排序算法后计算每个单词(即顶点)的重要性分数,公式如下:

$$WS(V_i)=(1-d)+d*\sum_{V_j\in In(V_i)}\frac{w_{ji}}{\sum_{V_k\in Out(V_j)}w_{jk}}WS(V_j)$$

公式在 PageRank 的基础上增加了边的权重,语义图中每个单词的重要性分数通过相连单词的数量和它们的重要性来评估.单词按照重要程度排序后选取前三分之一,合并相邻的单词,抽取出关键术语.TextRank 模型是通过词间共现频率构建的无向加权图,忽略了词与词之间的语义相关性,也未考虑上下文信息和其他辅助信息.

Khan 等人[99]于 2016 年在 TextRank 方法的基础上进行改进,提出了 Term Ranker 方法.原始 TextRank 算法依赖词之间的共现关系构建语义图,存在两点限制:(1)受到滑动窗口内不相关频繁词出现频率的影响,导致不重要边的权重增加;(2)相关词之间距离太远,使用滑动窗口机制无法正确捕获.为了缓解这些问题,Term Ranker 方法采用术语嵌入来学习候选术语的语义表示,利用该语义表示捕获术语之间的相似性及关系强度,之后在顶点之间添加边的关系和权重,从而构建一个无向加权图.此外,Term Ranker 方法还融合了同义词术语表,将表示同义术语的顶点进行合并,增加语义图上中心节点的数量,进而提高低频术语的评分.2017 年,Pan 等人[102]采用相似的思路来解决 MOOC 领域低频术语引发的问题—与课程相关但不常见的概念极易被忽略.不同的是 Pan 提出了新型图传播算法,根据学习的词间语义相似性,投票术语数量及投票术语的质量对候选术语进行排序.

2018 年,Zhang 等人[103,104]思考到目前还不存在一种 ATE 方法可以在所有领域都超越其他 ATE 方法,因此他改变思路,提出一种通用的 SemRe-Rank 方法对现有的术语抽取方法进行增强.即在现有的术语抽取方法之上执行 PageRank 排序,提高现有方法抽取术语的准确率.首先,使用现有的 ATE 方法抽取术语候选词集合并评分;然后为每个文档构建一个图(类似 Term Ranker 中语义图构建),使用个性化的 PageRank 算法在图中进行迭代直至收敛,最终为每个候选术语计算修订后的重要性分数,达到重新排序的目的.实验是在 4 种数据集上对 13 种 ATE 方法进行测试,结果表明 SemRe-Rank 对所有 ATE 方法都有较大的改进,准确率的提升最高达到 15%.

基于图的自动术语抽取方法考虑语料库中术语和术语之间的共现关系(或语义相似性关系,即存在语义关联),依赖于重要性分数且能够融合更多的顶点特征信息,取得了较好的抽取效果.同时,该类方法不需要花费昂贵的人力成本来标注数据,能够弥补以词频为主的统计学方法所带来的缺点,即容易遗漏低频但重要的术语.但是该方法对图规模及边的疏密较为敏感,如何快速有效进行图的传播收敛仍是研究者需解决的问题.

2.9 基于主题模型的抽取方法

主题模型是以无监督学习方式对文本集合的隐含语义进行聚类的概率模型,旨在根据主题描述文本,确定每个文本与哪些主题相关以及每个主题由哪些单词(或短语)构成.事实上,每个主题可以表示为一组经常出现的单词(或短语)集合,该组单词按照对主题的相关程度降序排列.

基于主题模型自动抽取方法的理论基础:大多数术语可以表示成与特定领域子主题相关的概念,最新研究表明[9,11,23,105]在文档集合中划分主题,然后根据主题抽取术语,可以提高自动术语抽取的质量.

基于主题模型的术语抽取基本步骤如下:

(1) 使用主题建模技术(例如聚类,LDA)将目标语料库映射到由多个主题组成的语义空间;

(2) 使用词的主题概率分布来对术语候选词进行评分.

Bolshakova 等人[23]在 2013 年将多种主题建模技术(例如聚类,LDA)用于自动术语抽取,通过对比实验得出主题信息确实可以提高术语抽取的质量.该方法首先对给定语料库中的文本集进行主题划分,然后按照 Loukachevitch[81]提出的方法抽取术语候选词,使用基于主题的概率分布特征对候选词进行排序.Bolshakova 等人在主题建模的基础上开发了 7 个基于主题的术语抽取特征,但是只针对单字术语.这些特征将候选术语 t 的标准词频替换为候选词所属主题下的概率总和,将文档频率替换为包含候选词出现的总主题数量来适应 TF 和 TF-IDF,详见表 4.

同年,Li 等人[9]针对术语词频不能反应术语所承载的语义等问题,引入领域背景、领域主题、领域特定单词三个维度的语义信息到新型术语抽取方法 NovelRopicModel 中.该方法的核心思想是:某一高质量的术语候选词应该由代表某个主题的典型词汇组成.首先将目标语料库映射到潜在的多主题语义空间中,包括领域主题、背景主题和特定于文档的主题;然后使用词的主题概率分布来对候选词进行评分,具体如下:(1)使用 i-SWB 主题模型计算出所有词汇在三类主题下的概率;(2)抽取每一类主题最相关的 200 个词汇;(3)每个候选术语的评分可表示为构成候选词中每个单词的最大概率之和.最终,NovelRopicModel 方法在 4 个领域的数据集中均达到了最先进的性能.

Table 4 The topical features of term

表 4 术语的主题特征

方法名称	计算公式
Term Frequency(TF)	$\sum_{i=1}^{K} P_i(t)$, $P_i(t)$表示主题 i 中候选术语 t 的概率,K 表示主题的数量.
TF-IDF	$TF(t)\cdot\log\frac{\lvert K\rvert}{\lvert DF(t)\rvert}$, K 表示主题的总数量,$DF(t)$表示包含候选术语 t 出现的总主题数量
Domain Consensus	$-\sum_{i=1}^{K}(P_i(t)\times\log P_i(t))$,表示熵相关特征,$P_i(w)$表示主题 i 中候选术语 t 的概率
Maximum TF	$\max P_i(w)$,表示候选术语 t 在各种主题中的最大概率
Term Score(TS)	$\sum_{i=1}^{K} P_i(t)\log\frac{P_i(t)}{(\prod_{i=1}^{K} P_i(t))^{\frac{1}{K}}}$,表示 TF-IDF 扩展特征
TS-IDF	$TS(t)\times\log\frac{K}{TF(t)}$,表示 TF-IDF 扩展特征
Maximum TS	$\max_i TS_i(t)$,表示主题最大术语评分,选取候选术语 t 在各种主题中的最大 TS 值

El-Kishky 等人[11]认为主题可以被建模为术语上的多项分布,且与主题相关的频繁术语概率分值应较大.因此,于 2014 年提出了一种基于短语词袋(Bag-of-phrases)的短语挖掘架构 ToPMine 来抽取高质量的术语,并定义了高质量术语的 3 个要求:频繁性、搭配性和完整性.为达到这 3 个要求抽取出高质量短语,El-Kishky 设置候选术语的词频阀值来满足频繁性,使用"t-统计量"来衡量搭配强度,借助短语分割来达成完整性,进而将原始文档上划分成"短语词袋(Bag-of-phrases)"形式.在此基础上增加限制条件"将术语中的每个单词分配到同一主题"来构建主题模型 Phrase LDA,协助抽取出子主题下的高质量术语.

同年 Sun 等人[105]提出并实现了结合主题信息的无监督双语术语自动抽取方法,该方法主要基于以下两个基本假设:(1)领域指示性强的候选术语在相应主题语料中的频度排名要高于在其他主题语料中的频度排名;(2)领域指示性差的候选术语在不同主题语料中的频度排名比较接近.因此,在使用常规方法(短语对齐技术和 CRF 组块分析技术)获取双语术语后,引入领域主题信息来计算候选词的术语性并排序.实验结果表明使用主题信息有效提高了双语术语的抽取效果,准确率高达 94%.

2016 年,Li 等人[106]针对通用主题抽取方法会遗漏特定于子域的术语(指术语在整个语料库频率低,在子域中频率高),在 El-Kishky 等人[11]的基础上提出一种基于子域的迭代主题短语挖掘框架 CITPM,增加了"子域"的概念,通过使用持续迭代的聚类方法,聚类具有相似主题分布的文档构成子域,在子域中抽取出高质量的术语.

2017 年,Arora 等人[107]为了解决 ATE 工具构建术语表时产生较差术语召回率,丢失较多词汇表术语问题,

提出了利用聚类方法来抽取术语.首先使用 POS 标签来标记语料库,抽取名词短语作为术语候选词集合,然后利用候选术语之间句法和语义相似性计算出术语相似性矩阵,最后根据相似性矩阵对候选术语进行聚类.实验表明该方法比目前通用的术语抽取工具更为有效,能够在保证准确率的同时,提高召回率(高达 20%).Arora 等人进一步论证得出结论:聚类技术可以为自动术语抽取提供切实的帮助.

基于主题建模的自动术语抽取方法不同于基于频率的统计学抽取方法,主要使用词的主题概率分布来对术语候选词进行评分,可以兼顾术语的语义信息,提高低频术语的评分.但该类方法依赖于主题划分的准确性,在术语抽取领域的应用还不太成熟.

2.10 总结

2.10.1 宏观对比分析

上述章节详细介绍了各类自动术语抽取方法的基础理论、关键技术以及研究现状.总体而言,基于语言学的 ATE 方法相对简单易行,但大多数方法基于规则，需要人工归纳总结,不利于跨领域迁移使用;基于统计学的 ATE 方法不受领域限制,通用性较强,但严重依赖于目标语料库的规模和质量;基于混合的 ATE 方法兼具多种方法的优点,具有良好的领域独立性和语言无关性,但特征组合算法过于单薄,其抽取效果差于基于机器学习的 ATE 方法;基于外部知识的 ATE 方法可以获取到目标语料库外的有效特征,弥补因语料库质量和统计信息不足造成术语抽取效果差的缺陷,美中不足的是一些特定领域并无可用的外部资源;基于机器学习的 ATE 方法术语抽取准确率较高,但需要人工筛选特征集,且模型对人工标注的训练集有较强的依赖性;基于深度学习的 ATE 方法无需耗时设计特征工程,便能达到与机器学习相当的抽取准确率,但需要大量的标注数据,模型训练耗时长;基于语义相关的 ATE 方法则无需人工标注,但依赖领域关键概念的获取,应用较少;基于图的 ATE 方法属于无监督方法,可减少大量的人工干预,通过术语间的关联关系提高低频重要术语的评分,但图模型对图规模及边的疏密较为敏感,如何快速、有效进行图的传播收敛仍是研究者需解决的问题;基于主题建模的 ATE 方法通过划分主题对术语进行分类,兼顾术语的语义信息,使得术语抽取效率得到进一步提升,难点在于如何将候选术语划分到正确的主题类别下.

综上,表 5 从技术特点、主要优缺点及发展趋势等宏观角度对 8 类自动术语抽取方法进行对比分析.其中,优缺点主要从通用性(跨领域能力)、是否需要人工标注数据、术语抽取准确率等侧面进行说明.

Table 5 Comparison of 8 types of automatic term extraction methods

表 5 各类自动术语抽取方法的对比分析

方法	技术特点	优点	缺点	代表性成果	发展趋势
基于语言学	利用词法模式、词形特征、语义信息	准确率高	通用性差,需人工标注	LEXTER	长期研究方向
基于统计学	利用词频、文档频率等概率统计	通用性强,无需标注数据	准确率依赖于目标语料库的规模和质量	TF-IDF	主流方法
混合方法	利用语言学、统计学、主题信息等抽取方法的特征,兼具多种方法优势	通用性强,无需标注数据,准确率较高	特征组合算法简单，缺乏组合深度	C-value	应用较多,长期研究方向
基于外部知识	获取目标语料库外的有效特征,作为对比使用,如参考语料库、维基百科等	通用性较强,无需标注数据,可抽取低频术语	引入噪声数据,一些特定领域并无可用外部资源	Weirdness	主流方法,多领域应用
基于机器学习	利用分类器融合多种特征,包括术语的语言特征、统计特征、外部知识库特征等	准确率高,无需人工制定规则	需大量的标注数据,跨领域能力弱,目前还不成熟	SegPhrase, AutoPhrase	研究热点
基于深度学习	利用深度学习模型,结合分布式特征(词嵌入表示)	准确率高，无需复杂的特征工程	需大量标注数据(远大于机器学习所需),通用性弱	/	研究热点
基于语义相关	利用术语之间的关联关系(边的特征),包括语义相似性和语义相关性	无需标注数据	对领域关键概念的抽取依赖性强	KeyConcepts Relatedness	小范围应用
基于图	将文本图形化表示,点表示术语(点的特征),边表示术语之间的关联特性(边的特征)	通用性较强,无需标注数据,解决低频词问题	图中边比较稀疏,收敛慢,应用不成熟	TextRank	多领域应用

基于主题模型	将文档、主题、术语(词或短语)三方面的信息综合考虑	术语聚簇化,无需标注数据,提高抽取准球率	依赖于主题划分的准确性,应用不成熟	ToPMine	主流方法

2.10.2 微观对比分析

除了 2.9.1 章节中 9 类 ATE 方法对比之外,每种类别下的 ATE 方法也各有特点.表 6 从抽取方法、使用模型、使用算法及所用特征等微观角度对 2015～2019 年最新 ATE 方法的多项属性进行详细对比分析.由于每种方法所使用数据集(部分使用封闭式数据集)及术语评价方法各个不同,使得术语抽取结果(准确率、召回率及 *F*1 值)无法在表中进行较为公正的比较.后续 3.3 章节将介绍现有文献对部分 ATE 方法实验结果的对比分析.

Table 6 The list of analysis of automatic term extraction methods

表 6 自动术语抽取方法分析列表

抽取方法	类别	模型	使用算法	语言特征	统计特征	对比特征	分布式特征	语义关系特征	主题特征	外部资源
ComboBasic,2015[63]	混合法	Basic	语言规则过滤+ComboBasic 公式	✔	✔					
Li,2015[66]	混合法	DV-termhood	信息熵和词频分布变化混合+语言规则过滤	✔	✔					
SegPhrase,2015[12]	有监督	随机森林	短语分割模型+随机森林算法+迭代循环		✔					
RIDF,2016[52]	统计法	IDF	RIDF 算法[58]		✔					
Stanković,2016[67]	混合法	/	基于语法的语言规则+4 种统计方法+外部电子词典	✔	✔					✔
tf-dcf,2016[74]	外部知识	TF-IDF	tf-dcf 算法		✔	✔				
Wang,2016[87]	弱监督	CRF 模型	CRF 模型+构建核心词汇库	✔						✔
Wang,2016[115]	深度学习	联合模型	联合模型+深度学习模型 LSTM 和 CNN 作为分类器				✔			
LiTeWi,2016[100]	语义相关	/	wikipedia 作为语义知识库+多语义术语消歧合并	✔	✔	✔				✔
Term Ranker,2016[99]	图方法	TextRank	相似关系语义图+TextRank+外部同义词表	✔	✔		✔	✔		✔
CITPM,2016[106]	主题方法	Phrase LDA	词频过滤+Phrase LDA 主题模型+文档聚类		✔				✔	
Dong,2017[68]	混合法	/	语言规则过滤+基于文本特征和复合统计量	✔	✔					
Li,2017[13,15]	混合法	unithood	术语 unithood 度量+完整短语挖掘+短语分割		✔					
Haque,2017[77]	外部知识	WikiPedia	*n*-gram 过滤+Dice 系数结合 LL 排序+维基百科重排		✔		✔			✔
Yuan,2017[83]	有监督	机器学习	6 种机器学习算法+10 种统计特征		✔					
Liu,2017[54]	有监督	SVM 模型	SVM 模型抽取规则+3 种领域特征加权评分	✔	✔					
Pan,2017[102]	图方法	PageRank	相似关系语义图+新型图传播算法	✔			✔	✔		
Arora,2017[107]	主题方法	聚类	语言规则过滤+相似性聚类方法	✔				✔	✔	
DRTE,2018[40]	语言法	词性规则	构词规则+边界检测算法	✔				✔		
AutoPhrase,2018[14]	远程监督	随机森林	外部知识库+POS 指导短语分割+随机森林算法	✔	✔					✔
Zhao,2018[90]	深度学习	Bi-LSTM-CRF	增量自训练算法+基于 Word2Vec 和 POS 特征的 Bi-LSTM-CRF 深度学习	✔			✔			
Kucza,2018[91]	深度学习	RNN 模型	LSTM 模型 / GRU 模型				✔			
Li,2018[98]	语义关系	分布式相似	术语嵌入向量+4 种相似性度量方法	✔	✔		✔	✔		✔

Yu,2018[96]	语义关系	STC-value	通用词作分割符+相似候选术语发现+STC-value 排序	✔	✔					
Lahbib,2018[97]	语义关系	共现频率	TF-IDF 抽取种子术语集+候选词与种子术语关系度量		✔			✔		
SemRe-Rank,2018[103]	图方法	PageRank	相似关系语义图+个性化 PageRank		✔		✔	✔		
Khosla,2019[116]	深度学习	联合模型	联合模型+CNN 和全连通网络作为分类器				✔			
Gao,2019[89]	深度学习	RNN 模型	术语向量表示+CNN+分类器进行排序	✔			✔			

2.10.3 面临的挑战

自动术语抽取已得到广泛的研究,取得了一定的成绩和较好的效果.但是,现有的自动术语抽取方法仍处于较为初期的阶段,离问题的真正解决还有很长的距离,亟待进一步提升术语抽取的效率和质量,并克服面临的诸多挑战:

(1) 目标语料库缺乏标注数据.

现阶段,目标语料库中需处理的文档大多是特定领域的文档,主要特点是文本稀疏,缺乏标注数据.使用手动标注或创建领域知识资源,代价高,耗时长,可行性较低.

(2) 抽取效果不理想,无法过滤噪声数据.

噪声数据通常是在生成术语候选词时引入,如 POS 标注器的错误标注、词性过滤规则过松等.若这些噪声数据在术语排序算法中没有得到有效的处理,对抽取结果的准确率和召回率影响很大.此外,并非文档中所有的词串都可以作为术语候选词,如何有效降低噪声数据的数量,提高候选术语抽取的质量,是学者必须面对的问题.

(3) 遗漏低频重要术语.

现有大部分自动术语抽取方法无法抽取低频重要的领域术语,因为没有足够的统计信息来保证低频术语的抽取.例如一些与领域相关性很大但在整个语料库中出现次数很少的术语,很容易被忽略.

(4) 评价体系不够完善.

大多数自动术语抽取研究自成体系,评价方法及使用数据集(部分使用封闭人工标注数据集)各不相同.很难将所有抽取方法放在一起进行评价,阻碍了自动术语抽取研究的更好发展.

3 数据资源及实验评估

3.1 可用数据资源

3.1.1 数据集

为了方便学者更好得研究自动术语抽取任务及获取数据集,本文整理了现有研究工作经常用到的公开数据集及其 URL 链接,详见表 7.该表详细列举了论文中的常用的 10 个数据集:

- GENIA 数据集[108]:生物医学领域文献集合,是测试术语抽取中最常用的数据集之一.最新版本是 3.02 版,包含来自 PubMed 的 1,999 个 Medline 摘要,434,782 个单词及 33,396 个真实术语,且术语被分成多种类别(例如 DNA 域),每个类别均有语言和语义信息注释.
- FAO 数据集[109]:包含 780 篇手工标记的粮食及农业组织报告,每份报告提供两个已确认的真正术语.
- Krapivin 数据集[110]:包含 2304 篇关于信息学的论文,其标准术语列表由论文中的关键词构成.
- ACL RD-TEC 数据集[111]:包含 1965 年至 2006 年间在计算机科学领域发表的 10,922 篇文章,和三个手动注释的术语列表:把排名位于前 82,000 个候选词人工标记为有效术语或无效术语;然后将有效术语进一步标记为技术术语和非技术术语.通常会使用有效术语列表作为“标准术语表”进行测试.
- ACL RD-TEC 2.0 数据集[112]: ACL RD-TEC 数据集的扩展版,由 ACL Anthology Reference Corpus 的 300 篇摘要组成,包含 1,384 个有效句子,32,921 个单词和 3,059 个注释术语,标记术语被分为七个类别.

- TTC 类数据集[113]:TTC 是术语提取、翻译工具和可比较语料库的缩写,旨在为双语术语获取和翻译提供资源.其中,风能领域 TTC-wind(TTCw)和移动技术领域 TTC-mobile(TTCm)数据集较为常用.两个数据集均是通过网站爬取构建的,提供了人工过滤的“标准术语表”.
- EuroParl 数据集[114]:是从欧洲议会程序中提取的平行语料库,标准术语列表由 Eurovoc 词库提供.

除了上述可以提供“标准术语表”的公开数据集外,还有一些目前常用的数据集(例如 DBLP,Adamedic,Yelp 等)不提供可参考的术语列表,需要借助领域专家进行术语标注后,方可进行术语评价.

Table 7 Open dataset details list

表 7 公开数据集详情列表

数据集	文档数	单词数(K)	标准术语表大小	术语来源	URL 链接
GENIA[108]	1,999	435	33,396	人工标注	http://www.geniaproject.org/
FAO[109]	780	26,672	1,554	人工标注	http://www.fao.org/global-perspectives-studies/resources/dataset/en/
Krapivin[110]	2,304	21,189	8,766	文章的关键词	http://dit.unitn.it/~krapivin/
ACL[111]	10,922	41,202	21,543	人工标注	https://github.com/languagerecipes/the-acl-rd-tec
ACL 2.0[112]	300	33	3,059	领域专家标注	https://github.com/languagerecipes/acl-rd-tec-2.0
TTCw[113]	103	801	287	网站爬取过滤	http://www.lina.univ-nantes.fr/?Reference-Term-Lists-of-TTC.html
TTCm[113]	37	305	254	网站爬取过滤	
Europarl[114]	9,672	63,279	15,094	Eurovoc 词库	http://eurovoc.europa.eu/drupal
DBLP	/	/	/	无	https://dblp.uni-trier.de/
Academia	/	/	/	无	http://aminer.org/billboard/AMinerNetwork
Yelp	/	/	/	无	https://www.yelp.com/academic_dataset

3.1.2 自动术语抽取工具

研究者已开发了很多 ATE 软件工具.但由于以下三个原因,使得真正能被用户使用的 ATE 工具很少.(1)ATE 工具中一部分是针对特定领域开发的(例如 BioTex,FlexiTerm)或仅限于学术用途(例如,TerMine);(2)大部分 ATE 工具是建立在整体式架构中,这种架构只有非常有限的定制性，可扩展性低;(3)大多数 ATE 工具只能提供一种 ATE 算法,局限性较大.因此,ATR4S 和 JATE 2.0 作为高度可扩展和模块化的 ATE 工具,被较多应用.

- ATR4S[29],基于 Scala 编写的 ATE 开源软件,包含超过 15 种 ATE 算法,高度可扩展,模块化和可配置的工具,支持自动缓存.ATR4S 作为目前最新的 ATE 工具,新增的自动术语抽取算法包括:PU-ATR, KeyConceptRelatedness, NovelTopicModel 和 Basic.
- JATE 2.0[52],是在 Apache Solr 框架内开发的 Java 自动术语提取工具包,是 JATE(2008 版)[115]的升级版本.同时,JATE 2.0 是一个免费开源,具有高度模块化,适应性强且可扩展的 ATE 库,拥有 10 种已实现的 ATE 算法.
- TermSuite[116],基于 JAVA 语言及 UIMA 框架的 ATE 工具,采用多语言设计,可扩展,可处理术语变体.该工具主要通过 Weirdness 方法对术语候选词进行排序,侧重于使用句法和形态模式识别术语变体.
- TBXTools[117]是基于 Python 编写的免费 ATE 工具,实现一种结合语言和统计方法的多字词术语抽取算法,可以在任何流行的操作系统下工作.
- BioTex[118],是一个 Web 应用程序,仅用于生物医学领域提取术语,提供在线测试和评估,也可在任何程序中用作 Java 库(库中不包括 POS 标记器).
- FlexiTerm[119],仅应用于生物医学领域的术语抽取工具,提供比 C-value 方法更为灵活有效的候选术语比较方法. FlexiTerm 功能强大,适用于非正式的结构化文档.
- TOPIA,是一个广泛使用的 Python 库,提供一种基于 POS 标注和简单统计度量(例如频率)混合的术语抽取方法.但该 ATE 工具自 2009 年以来没有再更新.

Table 8 Comparision of seven ATE tools

表 8 比较 7 种 ATE 工具

术语抽取工具名称	编写语言	发布时间	实现的方法	是否开源	使用方式	URL 链接
ATR4S[29]	Scala	2018	13 种 ATE 算法: ATF, ResidualIDF, C-value, Basic, ComboBasic, PostRankDC, Relevance, Weirdness, LinkProbability, NovelRopicModel, KeyConceptRelatedness, Voting, PU-ATR	免费开源	调用 API 接口	https://github.com/ispras/atr4s
JATE 2.0[52]	Java	2016	10 种 ATE 算法: TF, ATF, TF-IDF, RIDF, χ^2, C-value, Weirdness, GlossEx, TermEx, RAKE	免费开源	(1)嵌入模式; (2)插件模式:作为 Solr 插件	https://github.com/ziqizhang/jate
TermSuite[116]	Java	2016	1 种 ATE 算法: Weirdness	免费开源	(1)Java API;(2)命令行 API;(3)图形用户界面.	https://github.com/termsuite/termsuite.github.io
TBXTools[117]	Python	2015	1 种 ATE 算法:结合语言和统计学方法	免费非开源	Python 库	https://sourceforge.net/projects/tbxtools/
BioTex[118]	Java	2014	8 种 ATE 算法	非开源	(1)Web 在线使用;(2)Java 库	http://tubo.lirmm.fr/biotex/
FlexiTerm[119]	Java	2013	1 种 ATE 算法:基于 POS 标注和 C-value 方法	免费开源	独立软件	http://www.cs.cf.ac.uk/flexiterm
TOPIA	Python	2009	1 种 ATE 算法:基于 POS 标注和简单统计度量	免费非开源	Python 库	https://pypi.python.org/pypi/topia.termextract

3.2 实验评价方法及指标

目前,自动术语抽取的评价方法主要分为两种方案:

(1) 人工评价方式:在领域专家的帮助下对抽取术语列表进行人工评价.

(2) 术语参考表评价方式:提前预设一个术语参考表,即形成一个“标准术语表(golden standard)”.按照此标准对抽取术语列表进行评价.

两种评价方法的优缺点显而易见:第一种方法借助领域专家的知识提供最准确的评估,可操作性强但主观性也很大,会产生认识分歧、复杂术语的组合分歧;第二种方法提供了可重现的实验结果、可调整的参数以及可以使不同方法在一个数据集上的比较.

自动术语抽取效果的评价指标通常借鉴信息检索模型中的 3 个基本评价指标,包括准确率 P(Precision)、召回率 R(Recall)、综合指标 $F1$ 值(F-score).

$$P=\frac{|Correct\ \cap Retrived[1:N]|}{N},\qquad R=\frac{|Correct\ \cap Retrived[1:N]|}{|Correct|},\qquad F1=\frac{2PR}{P+R}$$

其中,$Correct$ 表示标准术语表集合,$Retrieved[1:N]$表示被评估的术语抽取方法抽取出排名前 N 个术语集合.

此外,部分论文[23,26,29,76,81]还采用平均准确率(Average Precision,简称 AvP)作为术语抽取的评价指标:

$$AvP=\sum_{i=1}^{N}P(i)(R(i)-R(i-1))$$

其中,$P(i)$、$R(i)$分别表示抽取出的术语位于第 i 位时的准确率 P 和召回率 R.

3.3 实验结果对比分析

因自动术语抽取评价体系不够完善,各种文献中术语抽取方法的实验评估完全不同,在语料库选择(例如领域,规模)、评价方法(例如人工评价方式,术语参考表评价方式)和候选术语选择范围(例如整个抽取结果,前 N 个最佳结果)等方面各成体系,并且部分论文[9,10,26,54,80,87,99]在自己标注的封闭数据集上进行实验评估,导致很难将所有术语抽取方法的实验结果进行统一公正对比.

目前,有少量文献对自动术语抽取方法及其使用的特征进行实验对比.其中,Zhang 等[115]人于 2008 年对比了 TF-IDF、Weirdness、C-value、GlossEx 和 TermEx 五种术语抽取方法,得出混合方法比单一的领域性度量方法效果好,且单字术语在特定领域占据较大的比例,忽略单字术语会造成抽取效果差.Nokel 等人[50]于 2013 年比较了信息检索领域抽取单字术语和双字术语的特征效果:包括语言特征、词频特征、参考语料库特征、主题模型特征等,得出基于主题模型的特征对于抽取单字术语效果最佳,基于上下文的特征对于抽取双字术语非常重要.2014 年,Fedorenko 等人[26]在语言学特征(预定义的词性模式)、统计学特征(TF-IDF、对数似然比等)、外部知识特征(Weirdness、Relevance 等)的基础上对比基于机器学习的抽取方法和基于投票算法的混合抽取方法.结果表明机器学习方法仅在少量训练集的情况下便超越投票算法的最优效果.从侧面反映,虽然混合抽取方法和机器学习抽取方法都能够使用多种类型术语特征,但使两者组合特征的深度不同,最终抽取效果有较大的差异. 2016 年,Verberne 等人[120]从语料库规模、参考语料库及多字术语的重要性 3 个方面出发对术语抽取方法进行评估,得出如下结论:(1)较大规模的目标语料库能抽取质量更高的术语;(2)所有抽取方法在小规模语料库(低于 1000 字)上效果都很差;(3)抽取单字术语和多字术语效果最好的是混合方法,因为其融合了多种特征.2018 年,Astrakhantsev 等人[29]对 ATR4S 术语抽取工具中的 13 种 ATE 方法在 7 个公开数据集上进行比较,如表 9 所示.实验结果表明:没有一种自动术语抽取方法能够在所有数据集上都表现最佳;但多种特征相结合可以抽取质量更好的术语,例如混合抽取方法和机器学习抽取方法,其中机器学习方法表现尤为稳定(4 个数据集中最佳).

现实应用中的自动术语抽取会受到多方面因素的影响,包括目标语料库规模、人工标注数据、抽取特征、术语排序算法的选择、参考语料库、噪声数据等.但现有术语方法的实验结果对比多是假设大部分条件存在并确定,然后对比 1～2 个方面因素对于抽取效果的影响.例如在语料库和标注数据确定的情况下,评测分析不同类型术语抽取特征及方法的优劣.因此,如何对实际应用中的 ATE 方法进行评测也是研究者需解决的问题.

Table 9 Comparison of 13 ATE methods over 7 datasets(by average precision)[29]

表 9 13 种 ATE 方法在 7 个数据集上的比较(使用 *AvP*)[29]

类别	方法名称	GENIA	FAO	Krapivin	Patents	ACL	ACL2.0	Europarl
统计学方法	ATF	0.7105	0.0415	0.1107	0.5397	0.0682	0.6802	0.1689
	ResidualIDF	0.7047	0.0133	0.1063	0.5268	0.0645	0.6774	0.1302
混合方法	C-value	0.7283	0.3845	0.4009	0.6452	0.4304	0.7879	0.3213
	Basic	0.6444	0.3795	0.3912	0.5548	**0.5393**	0.6966	**0.3917**
	ComboBasic	0.6440	0.3797	0.3913	0.5526	**0.5391**	0.7013	**0.3920**
基于外部知识	Relevance	0.7410	0.1504	0.2988	0.5044	0.4782	0.7530	0.2139
	Weirdness	0.7672	0.1478	0.3315	0.5422	0.4797	0.7579	0.2270
	LinkProbability	0.7071	0.0068	0.1024	0.4571	0.0980	0.7185	0.0851
基于主题模型	NovelRopicModel	0.7138	0.0598	0.1081	0.6003	0.2484	0.7958	0.2076
基于语义相关	KeyConceptRelatedness	0.6758	**0.4671**	0.3384	0.6190	0.3227	0.7124	0.3408
	PostRankDC	0.6655	0.4138	0.4068	0.5033	0.4577	0.6471	0.3784
基于机器学习	Voting	0.7582	0.1326	0.2683	0.6243	0.3353	0.7871	0.2617
	PU-ATR	**0.7823**	0.4429	**0.4210**	**0.6821**	0.4938	**0.8028**	0.3688

4 未来研究方向探讨和总结

通过对现有的术语抽取研究工作进行总结,未来可以从以下几个方面展开相关研究:

(1) 借助外部知识库,协助抽取术语

针对目标语料库缺乏标注数据的问题,可以借助于外部知识库进行解决.随着互联网的快速发展,诸如维基百科、百度百科、WordNet 知识库、搜索引擎或同义词典等外部知识库的资源越来越多,可以借助外部资源对目标语料库进行自动标注,有助于提高自动术语抽取的效率,如 Shang 等人[14]提出使用维基百科对语料库进行远程监督标注,取得比人工标注更好的效果.因此,借助外部知识库扩充并标注目标语料库,提高术语自动抽取的效果,将是一种主流的研究方向.

(2) 多维异质特征融合,提升术语抽取效果

针对现有抽取方法效果差,过滤噪声术语不理想的问题,考虑将多维异质特征融合到同一种方法中.由第 2 章自动术语抽取方法的分类,可以看出术语包含语言结构特征、分布特征、领域特征、上下文特征、语义相似

特征、术语间的关系特征(基于图的方法)、主题特征等多个维度的信息.只使用一两个维度进行术语抽取,得到的效果较为一般,噪声数据也较多.因此自动术语抽取应由只考虑单维特征转向将现有的多维异质特征融合到统一的模型中.例如 Khan 等人[99]提出基于图的术语抽取方法 Term Ranker 架构,融合了语言过滤器、C-value、TFIDF、语义相似特征、术语之间的关系以及同义词典,提高术语抽取效果.同时,Term Ranker 架构指出了术语抽取的研究方向,即将多维异质特征、多种方法有机融合,是自动术语抽取研究的发展趋势.

(3) 尝试将自动术语抽取与语义知识相结合

针对已有的抽取方法遗漏重要低频术语问题,考虑融合术语的语义关系信息进行抽取.基于词频的术语抽取方法无法对低频术语正确评分,导致其排序靠后被遗漏.因此需要转换角度考虑低频术语:低频术语并非“孤岛”,领域内的术语之间存在多种语义关联关系,如同义词关系、上下位关系、整体-部分关系等.解决低频术语问题,需要重视术语之间的语义关系信息,将术语在语义上关联性较强的其他术语或关系识别出来,形成术语语义网.对术语语义网的研究,给自动术语抽取提供了新的研究方向和思路.

(4) 术语评价体系的完善

目前,自动术语抽取属于刚起步阶段,评价体系还没有切实有效的验证标准.对于术语评价体系的完善,不仅需要落地可行的效果评价方法和评估指标,还需要坚实可靠的理论体系进行支撑.所以,完善自动术语抽取的基础评价体系是一个长期的研究方向和目标.

随着大数据、移动互联网和社交媒体等技术的迅猛发展,导致本文数据量激增,使得作为文本挖掘中的基础性工作--自动术语抽取变得尤为重要和迫切.本文在简要介绍了自动术语抽取问题定义和解决框架的基础上,围绕“浅层语言分析”中基础语言信息和关系结构信息两个层面的特征对自动术语抽取方法进行分类,对主流的技术和方法进行了对比与分析.此外,还对术语抽取未来可能面临的挑战和研究方向进行了探讨与展望.

References:

[1] Rani M, Dhar AK, Vyas OP. Semi-automatic terminology ontology learning based on topic modeling. Engineering Applications of Artificial Intelligence, 2017,63:108-125. [doi: 10.1016/j.engappai.2017.05.006]

[2] Wong W, Liu W, Bennamoun M. Tree-traversing ant algorithm for term clustering based on featureless similarities. Data Mining and Knowledge Discovery, 2007,15(3):349-381. [doi: 10.1007/s10618-007-0073-y]

[3] Uysal AK. An improved global feature selection scheme for text classification. Expert Systems with Applications, 2016,43:82-92. [doi: 10.1016/j.eswa.2015.08.050]

[4] Mihalcea R, Tarau P. Textrank: Bringing order into text. In: Proc. of the EMNLP. Stroudsburg: ACL, 2004. 404-411.

[5] Baralis E, Cagliero L, Mahoto N, Fiori A. GRAPHSUM: Discovering correlations among multiple terms for graph-based summarization. Information Sciences, 2013,249:96-109. [doi: 10.1016/j.ins.2013.06.046]

[6] Bouamor D, Semmar N, Zweigenbaum P. Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In: Calzolari N, Choukri K eds. Proc. of the LREC. Istanbul: European Language Resources Association, 2012. 674-679.

[7] Yuan Y, Gao Y, Zhang Y, Sharoff S. Cross-lingual Terminology Extraction for Translation Quality Estimation. In: Calzolari N, Choukri K eds. Proc. of the LREC. Miyazaki: European Language Resources Association, 2018. 3774-3780.

[8] Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web, 2017,8(3):489-508. [doi: 10.3233/sw-160218]

[9] Li S, Li J, Song T, Li W, Chang B. A novel topic model for automatic term extraction. In: Proc. of the SIGIR. New York: ACM, 2013. 885-888. [doi: 10.1145/2484028.2484106]

[10] Judea A, Schütze H, Brügmann S. Unsupervised training set generation for automatic acquisition of technical terminology in patents. In: Proc. of the COLING. Stroudsburg: ACL, 2014. 290-300.

[11] El-Kishky A, Song Y, Wang C, Voss CR, Han JW. Scalable topical phrase mining from text corpora. Proceedings of the VLDB Endowment, 2014,8(3):305-316. [doi: 10.14778/2735508.2735519]

[12] Liu J, Shang J, Wang C, Ren X, Han JW. Mining quality phrases from massive text corpora. In: Proc. of the SIGMOD. Victoria: ACM, 2015. 1729-1744. [doi: 10.1145/2723372.2751523]

[13] Li B, Yang X, Wang B, Cut W. Efficiently mining high quality phrases from texts. In: Singh SP, Markovitch S eds. Proc. of the AAAI. Palo Alto: AAAI Press, 2017. 3474-3481.

[14] Shang JB, Liu J, Jiang M, Ren X, Voss CR, Han JW. Automated phrase mining from massive text corpora. IEEE Transactions on Knowledge and Data Engineering, 2018,30(10):1825-1837. [doi: 10.1109/TKDE.2018.2812203]

[15] Li B, Yang X, Zhou R, Wang B, Liu C, Zhang Y. An efficient method for high quality and cohesive topical phrase mining. IEEE Transactions on Knowledge and Data Engineering, 2019,31(1):120-137. [doi: 10.1109/TKDE.2018.2823758]

[16] Chen K, Chen H H. Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation. In: Proc. of the ACL. Stroudsburg: ACL, 1994. 234-241. [doi: 10.3115/981732.981764]

[17] Justeson J S, Katz S M. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering, 1995,1(1):9-27. [doi: 10.1017/S1351324900000048]

[18] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the c-value/nc-value method. International journal on digital libraries, 2000,3(2):115-130. [doi: 10.1007/s007999900023]

[19] Vivaldi J, Cabrera-Diego L A, Sierra G, Pozzi M. Using Wikipedia to Validate the Terminology found in a Corpus of Basic Textbooks. In: Calzolari N, Choukri K eds. Proc. of the LREC. Istanbul: European Language Resources Association, 2012. 3820-3827.

[20] Astrakhantsev N. Automatic term acquisition from domain-specific text collection by using Wikipedia. Proceedings of the Institute for System Programming, 2014,26(4):7-20. [doi: 10.15514/ISPRAS-2014-26(4)-1]

[21] Wang R, Liu, W, McDonald C. Featureless domain-specific term extraction with minimal labelled data. In: Proc. of the Australasian Language Technology Association Workshop. 2016. 103-112.

[22] Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M. Yet another ranking function for automatic multiword term extraction. In: Proc. of the 9th International Conference on NLP. Switzerland: Springer, 2014. 52-64. [doi: 10.1007/978-3-319-10888-9]

[23] Bolshakova E, Loukachevitch N, Nokel M. Topic models can improve domain term extraction. In: Proc. of the European Conference on Information Retrieval. Moscow: Springer, 2013. 684-687. [doi: 10.1007/978-3-642-36973-5]

[24] Astrakhantsev NA, Fedorenko DG, Turdakov DY. Methods for automatic term recognition in domain-specific text collections: A survey. Programming and Computer Software, 2015,41(6):336-349. [doi: 10.1134/S036176881506002X]

[25] Yuan JS, Zhang XM and Li ZJ, Survey of automatic terminology extraction methodologies. Computer Science, 2015,42(8):7-12 (in Chinese with English abstract).

[26] Fedorenko D, Astrakhantsev N, Turdakov D. Automatic recognition of domain-specific terms: an experimental evaluation. Proceedings of the Institute for System Programming, 2014,26(4):55-72. [doi: 10.15514/ISPRAS-2014-26(4)-5]

[27] Barrón-Cedeno A, Sierra G, Drouin P, Ananiadou S. An improved automatic term recognition method for Spanish. In: Proc. of the CICLing. Mexico: Springer, 2009. 125-136. [doi: 10.1007/978-3-642-00382-0]

[28] Bordea G. Domain adaptive extraction of topical hierarchies for expertise mining [Ph.D. Thesis]. Galway, Ireland: National University of Ireland, 2013.

[29] Astrakhantsev N. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in scala. Language Resources and Evaluation, 2018,52(3):853-872. [doi: doi:10.1007/s10579-017-9409-4]

[30] Korkontzelos I, Klapaftis IP, Manandhar S. Reviewing and evaluating automatic term recognition techniques. In: Ranta A, Nordstrom B eds. Proc. of the GoTAL. Berlin: Springer, 2008. 248–259. [doi: 10.1007/978-3-540-85287-2_24]

[31] Jacquemin C. Recycling terms into a partial parser. In: Proc. of the fourth conference on Applied natural language processing. Stuttgart: ACL, 1994. 113-118. [doi: 10.3115/974358.974384]

[32] Jacquemin C. Syntagmatic and paradigmatic representations of term variation. In: Dale R, Church KW eds. Proc. of the 37th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 1999. 341-348. [doi: 10.3115/1034678.1034733]

[33] Dagan I, Church K. Termight: Identifying and translating technical terminology. In: Proc. of the fourth conference on Applied natural language processing. Stuttgart: ACL, 1994. 34-40. [doi: 10.3115/974358.974367]

[34] Lauriston A. Automatic recognition of complex terms: Problems and the TERMINO solution. Terminology, 1994,1(1):147-170. [doi: 10.1075/term.1.1.11lau]

[35] Arppe A. Term Extraction from unrestricted text. In: Proc. of the 10th Nordic Conference of Computational Linguistics. 1995.

[36] Bourigault D, Gonzalez-Mullier I, Gros C. LEXTER, a Natural Language Processing tool for terminology extraction. In: Proc. of the 7th EURALEX International Congress. Sweden: Novum Grafiska AB, 1996. 771-779.

[37] Naulleau E. Profile-guided terminology extraction. In: Proc. of the TKE. 1999.

[38] Koo T, Carreras X, Collins M. Simple semi-supervised dependency parsing. In: Proc. of the 46th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2008. 595-603.

[39] Foo J, Merkel M. Using machine learning to perform automatic term recognition. In: Proc. of the LREC. European Language Resources Association, 2010. 49-54.

[40] Li SL, Xu B and Yang YJ, DRTE: A Term Extraction Method for K12 Education. Journal of Chinese Information Processing, 2018,32(3):101-109 (in Chinese with English abstract).

[41] Kageura K, Umino B. Methods of automatic term recognition: A review. Terminology, 1996,3(2):259-289.

[42] Montgomery DC, Runger GC. Applied Statistics and Probability for Engineers. 7th ed., NJ: Wiley, 2018. 208-211.

[43] Church K, Gale W, Hanks P, Hindle D. Using statistics in lexical analysis. In: Uri Z ed. Lexical acquisition: Exploiting on-line resources to build up a lexicon. Hillsdale: Lawrence Erlbaum Associates, 1991. 115-164.

[44] Pearson KX. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1900,50(302):157-175. [doi: 10.1080/14786440009463897]

[45] Dunning T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 1993,19(1):61-74.

[46] Church K W, Hanks P. Word association norms, mutual information, and lexicography. Computational Linguistics, 1990,16(1):22-29.

[47] Pecina P. An extensive empirical study of collocation extraction methods. In: Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2005. 13-18.

[48] Song SK, Choi YS, Chun HW, Jeong CH, Choi SP, Sung WK. Multi-words terminology recognition using web search. In: Proc. of the International Conference on U-and E-Service, Science and Technology. Berlin: Springer, 2011. 233-238. [doi: 10.1007/978-3-642-27210-3_29]

[49] Chaudhari D L, Damani O P, Laxman S. Lexical co-occurrence, statistical significance, and word association. In: Proc. of the EMNLP. Stroudsburg: ACL, 2011. 1058-1068.

[50] Loukachevitch N, Nokel M. An experimental study of term extraction for real information-retrieval thesauri. In: Proc. of the TIA. 2013. 69-76.

[51] Wong W. Determination of unithood and termhood for term recognition. In Handbook of research on text and web mining technologies. USA: IGI Global, 2009. 500-529.

[52] Zhang Z, Gao J, Ciravegna F. Jate 2.0: Java automatic term extraction with apache solr. In: Calzolari N, Choukri K eds. Proc. of the LREC. Portoro: European Language Resources Association, 2016. 2262-2269.

[53] Navigli R, Velardi P. Semantic interpretation of terminological strings. In: Proc. of the 6th International Conference on Terminology and Knowledge Engineering. 2002:95-100.

[54] Liu L and Xiao YY. A statistical domain terminology extraction method based on word length and grammatical feature. Journal of Harbin Engineering University, 2017,38(9):1437-1443 (in Chinese with English abstract).

[55] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 1988,24(5):513-523. [doi: 10.1016/0306-4573(88)90021-0]

[56] Zhou L, Shi SM, Feng C and Huang HY, A Chinese term extraction system based on multi-strategies integration. Journal of the China Society for Scientific and Technical Information, 2010,29(3):460-467 (in Chinese with English abstract).

[57] Yan XL, Liu YQ, Fang Q, Zhang M, Ma SP, Ru LY. Domain-Specific terms extraction based on web resource and user behavior. Ruan Jian Xue Bao/Journal of Software, 2013,24(9):2089-2100 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4358.htm [doi: 10.3724/SP.J.1001.2013.04358]

[58] Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M. Biomedical terminology extraction: A new combination of statistical and web mining approaches. In: Proc. of the JADT'14. 2014. 421-432.

[59] Church K, Gale W. Inverse document frequency (idf): A measure of deviations from poisson. In: Natural language processing using very large corpora. Dordrecht: Springer, 1999. 283-295. [doi: 10.1007/978-94-017-2390-9_18]

[60] Li LS, Dang YZ, Zhang J, Li D. Domain term extraction based on conditional random fields combined with active learning strategy. Journal of Information &Computational Science, 2012,9(7):1931-1940.

[61] Rose S, Engel D, Cramer N, Cowley W. Automatic keyword extraction from individual documents. In: Text mining: applications and theory, 2010. 1-20.

[62] Bordea G, Buitelaar P, Polajnar T. Domain-independent term extraction through domain modelling. In: Proc. of the 10th international conference on terminology and artificial intelligence. 2013.

[63] Astrakhantsev N. Methods and software for terminology extraction from domain-specific text collection[D]. Ph.D. thesis, Institute for System Programming of Russian Academy of Sciences, 2015.

[64] You HL, Zhang W, Shen JY, Liu T. A weighted voting based automatic term recognition method. Journal of Chinese Information Processing, 2011,25(3):9-17 (in Chinese with English abstract).

[65] He L. Domain ontology terminology extraction based on integrated strategy method. Journal of the China Society for Scientific and Technical Information, 2012,31(8):798-804 (in Chinese with English abstract).

[66] Li LS, Wang YW, Huang DG. Term extraction based on infomation entropy and word frequency distribution variety. Journal of Chinese Information Processing, 2015,29(1):82-87 (in Chinese with English abstract).

[67] Stanković R, Krstev C, Obradovic I, Lazic B. Rule-based automatic multi-word term extraction and lemmatization. In: Calzolari N, Choukri K eds. Proc. of the LREC. Portoro: European Language Resources Association, 2016. 507-514.

[68] Dong YY, Li WH, Hu H. Domain term extraction method based on hierarchical combination strategy for Chinese web documents. Journal of Northwestern Polytechnical University, 2017,35(4):729-735 (in Chinese with English abstract).

[69] Pazienza MT, Pennacchiotti M, Zanzotto FM. Terminology extraction: an analysis of linguistic and statistical approaches. In: Knowledge Mining. Berlin: Springer, 2005. 255-279. [doi: 10.1007/3-540-32394-5_20]

[70] Ahmad K, Gillam L, Tostevin L. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In: Proc. of the TREC. 1999. 1-8.

[71] Peñas A, Verdejo F, Gonzalo J. Corpus-based terminology extraction applied to information access. In: Proc. of Corpus Linguistics. 2001. 458-465.

[72] Park Y, Byrd RJ, Boguraev BK. Automatic glossary extraction: beyond terminology identification. In: Proc. of the COLING. Stroudsburg: ACL, 2002. 1-7. [doi: 10.3115/1072228.1072370]

[73] Sclano F, Velardi P. Termextractor: a web application to learn the shared terminology of emergent web communities. In: Proc. of the 3th International Conference on Interoperability for Enterprise Software and Applications. London: Springer, 2007. 287-290.

[74] Lopes L, Fernandes P, Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf. Knowledge-Based Systems, 2016,97:237-249.

[75] Mykowiecka A, Marciniak M, Rychlik P. Recognition of irrelevant phrases in automatically extracted lists of domain terms. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 2018,24(1):66-90.

[76] Vivaldi J, Rodríguez H. Using Wikipedia for term extraction in the biomedical domain: first experiences. Procesamiento del Lenguaje Natural, 2010,45:251-254.

[77] Haque R, Penkale S, Way A. TermFinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. Language Resources and Evaluation, 2018,52(2):365-400. [doi: 10.1007/s10579-018-9412-4]

[78] Zheng D, Zhao T, Yang J. Research on domain term extraction based on conditional random fields. In: Proc. of the ICCPOL. Berlin: Springer, 2009. 290-296.

[79] Zhang X, Song Y, Fang AC. Term recognition using conditional random fields. In: Proc. of the 6th International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2010. 1-6.

[80] Zhang ZC. Using integration strategy and multi-level termhood to extract terminology. Journal of the China Society for Scientific and Technical Information, 2011,28(3):275-285 (in Chinese with English abstract).

[81] Loukachevitch N V. Automatic Term Recognition Needs Multiple Evidence. In: Calzolari N, Choukri K eds. Proc. of the LREC. Portoro: European Language Resources Association, 2012. 2401-2407.

[82] Conrado MD, Pardo TA, Rezende SO. A machine learning approach to automatic term extraction using a rich feature set. In: Proc. of the 2013 NAACL HLT Student Research Workshop. Stroudsburg: ACL, 2013. 16-23.

[83] Yuan Y, Gao J, Zhang Y. Supervised learning for robust term extraction. In: Proc. of the International Conference on Asian Language Processing. IEEE, 2017. 302-305.

[84] Yang Y, Yu H, Meng Y, Lu Y, Xia Y. Fault-tolerant learning for term extraction. In: Proc. of the 24th Pacific Asia Conference on Language, Information and Computation. Institute for Digital Enhancement of Cognitive Development, 2010. 321-330.

[85] Maldonado A, Lewis D. Self-tuning ongoing terminology extraction retrained on terminology validation decisions. In: Proc. of Conference on Terminology and Knowledge Engineering, 2016. 91-101.

[86] Aker A, Paramita M, Gaizauskas R. Extracting bilingual terminologies from comparable corpora. In: Proc. of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2013. 402-411.

[87] Wang H, Wang MP, Su XN. A study on Chinese patent terms extraction for ontology learning. Journal of the China Society for Scientific and Technical Information, 2016,35(6):573-585 (in Chinese with English abstract).

[88] Khosla K, Jones R, Bowman N. Featureless Deep Learning Methods for Automated Key-Term Extraction. 2019.

[89] Gao Y, Yuan Y. Feature-Less End-to-End Nested Term Extraction. In: Proc. of the CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer, 2019. 607-616.

[90] Zhao H, Wang F. A deep learning model and self-training algorithm for theoretical terms extraction. Journal of the China Society for Scientific and Technical Information, 2018,37(9):923-938 (in Chinese with English abstract).

[91] Kucza M, Niehues J, Zenkel T, Waibel A, Stüker S. Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In: Proc. of the Interspeech. Hyderabad: ISCA, 2018. 2072-2076.

[92] Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M. Biomedical term extraction: overview and a new methodology. Information Retrieval Journal, 2016,19(1-2):59-99. [doi: 10.1007/s10791-015-9262-2]

[93] Bikel D, Zitouni I. Multilingual natural language processing applications: from theory to practice. IBM Press, 2012.

[94] Yang K, Ding X, Zhang Y, Chen L, Zheng B, Gao Y. Distributed Similarity Queries in Metric Spaces. Data Science and Engineering, 2019,4(2):93-108.

[95] El-Beltagy S R, Rafea A. Kp-miner: Participation in semeval-2. In: Proc. of the 5th international workshop on semantic evaluation. Stroudsburg: ACL, 2010. 190-193.

[96] Yu Y, Zhao NX. Patent term extraction based on generic words and term components. Journal of the China Society for Scientific and Technical Information, 2018,37(7):742-752 (in Chinese with English abstract).

[97] Lahbib W, Bounhas I, Slimani Y. A possibilistic approach for Arabic domain terminology extraction and translation. In: Proc. of the International Symposium on Computer and Information Sciences. Cham: Springer, 2018. 231-238.

[98] Li K, Zha H, Su Y, Yan X. Concept mining via embedding. In: Proc. of the 2018 IEEE International Conference on Data Mining. Singapore: IEEE Computer Society, 2018. 267-276.

[99] Khan M T, Ma Y, Kim J. Term ranker: A graph-based re-ranking approach. In: Proc. of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference. Florida: AAAI Press, 2016. 310-315.

[100] Conde A, Larrañaga M, Arruarte A, Elorriaga JA, Roth D. LiTeWi: A combined term extraction and entity linking method for eliciting educational ontologies from textbooks. Journal of the Association for Information Science and Technology, 2016,67(2):380-399.

[101] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.

[102] Pan LM, Wang XC, Li JZ, Tang J. Course concept extraction in MOOCs via embedding-based graph propagation. In: Proc. of the 8th International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing, 2017. 875-884.

[103] Zhang Z, Gao J, Ciravegna F. Semre-rank: Improving automatic term extraction by incorporating semantic relatedness with personalised pagerank. ACM Transactions on Knowledge Discovery from Data, 2018,12(5):1-41. [doi: 10.1145/3201408]

[104] Zhang Z, Petrak J, Maynard D. Adapted textrank for term extraction: a generic method of improving automatic term extraction algorithms. In: Proc. of the 14th International Conference on Semantic Systems. Elsevier, 2018. 102-108.

[105] Su MS, Li L, Liu ZY. Unsupervisied bilingual terminology extraction algorithm for Chinese-English parallel patents. Journal of Tsinghua University (Science and Technology), 2014,54(10):1339-1343 (in Chinese with English abstract).

[106] Li B, Wang B, Zhou R, Yang X, Liu C. Citpm: A cluster-based iterative topical phrase mining framework. In: Proc. of the International Conference on Database Systems for Advanced Applications. Switzerland: Springer, 2016. 197-213.

[107] Arora C, Sabetzadeh M, Briand L, Zimmer F. Automated extraction and clustering of requirements glossary terms. IEEE Transactions on Software Engineering, 2017,43(10):918-945.

[108] Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus - a semantically annotated corpus for bio-textmining. In: Proc. of the Eleventh International Conference on Intelligent Systems for Molecular Biology, 2003. 180-182

[109] Medelyan O, Witten I H. Domain‐independent automatic keyphrase indexing with small training sets. Journal of the American Society for Information Science and Technology, 2008,59(7):1026-1040. [doi: 10.1002/asi.20790]

[110] Krapivin M, Autaeu A, Marchese M. Large dataset for keyphrases extraction. DISI-09-055: Italy, DISI, University of Trento, 2009.

[111] Handschuh S, QasemiZadeh B. The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics. In: Proc. of the 4th international workshop on computational terminology. Stroudsburg: ACL, 2014. 52-63

[112] QasemiZadeh B, Schumann A K. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In: Calzolari N, Choukri K eds. Proc. of the LREC. Portoro: European Language Resources Association, 2016. 1862-1868.

[113] Blancafort H, Daille B, Gornostay T, Heid U, Sharoff S, Méchoulam C. TTC: Terminology extraction, translation tools and comparable corpora. In: Proc. of the 14th EuraLex International Congress. 2010. 263–268.

[114] Koehn P. Europarl: A parallel corpus for statistical machine translation. MT summit. 2005,5:79-86.

[115] Zhang Z, Iria J, Brewster C, Ciravegna F. A comparative evaluation of term recognition algorithms. In: Calzolari N, Choukri K eds. Proc. of the LREC. Portoro: European Language Resources Association, 2008. 28-30.

[116] Cram D, Daille B. TermSuit: Terminology extraction with term variant detection. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016. 13-18. [doi: 10.18653/v1/P16-4003]

[117] Oliver A, Vàzquez M. TBXTools: a free, fast and flexible tool for automatic terminology extraction. In: Proc. of the International Conference Recent Advances in Natural Language Processing. 2015. 473-479.

[118] Lossio-Ventura J A, Jonquet C, Roche M, Teisseire M. BIOTEX: A system for biomedical terminology extraction, ranking, and validation. In: Proc. of the 13th International Semantic Web Conference. Italy: CEUR-WS.org, 2014. 157-160.

[119] Spasić I, Greenwood M, Preece A, Francis N, Elwyn G. FlexiTerm: a flexible term recognition method. Journal of Biomedical Semantics, 2013,4(1):27-43. [doi: 10.1186/2041-1480-4-27]

[120] Verberne S, Sappelli M, Hiemstra D, Kraaij W. Evaluation and analysis of term scoring methods for term extraction. Information Retrieval Journal, 2016,19(5):510-545. [doi: 10.1007/s10791-016-9286-2]

附中文参考文献:

[30]袁劲松,张小明,李舟军.术语自动抽取方法研究综述.计算机科学,2015,42(8):7-12.

[40]李思良,许斌,杨玉基.DRTE:面向基础教育的术语抽取方法.中文信息学报,2018,32(3):101-109.

[54]刘里,肖迎元.基于术语长度和语法特征的统计领域术语抽取.哈尔滨工程大学学报,2017,38(9):1437-1443.

[56]周浪,史树敏,冯冲,黄河燕.基于多策略融合的中文术语抽取方法.情报学报,2010,29(3):460-467.

[57]闫兴龙,刘奕群,方奇,张敏,马少平,茹立云.基于网络资源与用户行为信息的领域术语提取.软件学报,2013,24(9):2089-2100. http://www.jos.org.cn/1000-9825/4358.htm[doi:10.3724/SP.J.1001.2013.04358]

[64]游宏梁,张巍,沈钧毅,刘挺.一种基于加权投票的术语自动识别方法.中文信息学报,2011,25(3):9-17.

[65]何琳.基于多策略的领域本体术语抽取研究.情报学报,2012,31(8):798-804.

[66]李丽双,王意文,黄德根.基于信息熵和词频分布变化的术语抽取研究.中文信息学报,2015,29(1):82-87.

[68]董洋溢,李伟华,于会.文本特征和复合统计量的领域术语抽取方法.西北工业大学学报,2017,35(4):729-735.

[80]章成志.基于多层术语度的一体化术语抽取研究.情报学报,2011,28(3):275-285.

[87]王昊,王密平,苏新宁.面向本体学习的中文专利术语抽取研究.情报学报,2016,35(6):573-585.

[90]赵洪,王芳.理论术语抽取的深度学习模型及自训练算法研究.情报学报,2018,37(9):923-938.

[96]俞琰,赵乃瑄.基于通用词与术语部件的专利术语抽取.情报学报,2018,37(7):742-752.

[105]孙茂松,李莉,刘知远.面向中英平行专利的双语术语自动抽取.清华大学学报:自然科学版,2014(10):1339-1343.