DOI:10.12158/j.2096-3203.2020.06.023

基于无监督方法的电力文本专业词汇识别研究

朱婷婷[1], 杜一帆[1], 李睿凡[1,2], 熊永平[3]
(1. 北京邮电大学人工智能学院,北京 100876;2. 教育部信息网络工程研究中心,
北京 100876;3. 北京邮电大学计算机学院,北京 100876)

摘 要:电力专业词汇识别是面向变电运检文档进行深入语言理解和知识图谱构建等智能应用的基础。领域无关识别方法的效果不能令人满意,为此文中根据电力领域词汇的语言学特征提出一种面向电力领域的无监督专业词汇发现方法。首先以通用词典对电力文档语料分词,然后根据电力专业词汇的特征设置不同大小的滑动窗口,将之前分词结果的多种组合作为候选词;进一步计算邻接变化度、信息熵、点态互信息以及词频等 4 种候选词统计量;最后基于综合语言学特征和成词边界 3 种语法规则对候选词进行筛选形成专业电力新词。在公开数据集上与基线方法进行了对比实验,实验结果验证了文中提出方法的有效性。

关键词:领域词典;无监督学习;新词识别;滑动窗口;统计特征
中图分类号:TM930.9;TP391 **文献标志码**:A **文章编号**:2096-3203(2020)06-0159-07

0 引言

电力系统的运行维护工作中产生大量的文字材料,包括运行和检修、日志操作等非规范文件以及高精尖的技术报告等。这些文档的规模庞大,而且呈现指数型增加的趋势。这些不断增加的电力文档资料中包含大量的知识要点,由于技术资料管理体系不健全等因素,急需的知识点淹没在文档资料中,造成知识访问困难重重。如何利用好这些文档资料是提升电力系统运行和维护的智能化水平所面临的一个重要问题。

为精准和高效地利用电力系统运行维护的领域知识,需要建立以电力知识图谱等高层应用为导向的电力专业词汇的识别技术,从而形成规范的电力专业知识点及其网络化联系。为此,需要构建简洁高效的电力运检专业词汇识别方法。领域词汇发现是面向特定领域,基于自然语言处理技术深入分析理解语言文本的重要基础。领域词汇通常具有语义高凝炼的特点,是有效构建领域知识图谱[1—2]和专家问答系统[3—4]等高层应用的基础。因而,领域词汇发现的研究有着重要的研究意义与应用价值。新词一般是指在通用分词词典中未登录的词。而领域的新词发现就是找到通用分词词典中不存在的专业领域词汇的过程。

当前,电力领域的文本文档处理是电力人工智能的一个新热点。变电运检的过程文本文档亟待使用人工智能方法提升智能应用水平。因此,构建有效的电力领域新词发现方法是电力文本信息智能化的基础性工作。对于电力领域来说,标注专业性要求高、成本大、耗时耗力,所以无监督的新词发现[5—7]是一个非常有意义的研究。

文中提出面向电力领域的无监督新词发现方法,为后续建立变电运检的知识图谱等高层应用打下基础。以统计观察中的词和字的分布作为新词发现阈值的设置依据,首先提出以基础词典对文本语料分词,并使用不同大小的滑动窗口对分词结果进行组合,再使用多个统计量对特征进行筛选,最终得到领域新词。在公开数据集上与基线方法的对比评测验证了该方法的有效性。文中实验源码和结果数据已发布在公开网站提供给同行科研人员(https://github.com/bupt-mmai/electric-power)。

1 专业词汇识别方法概述

以下从 2 个方面总结与文中研究密切相关工作进展。一方面是有监督新词发现方法,另一方面是无监督新词发现方法。

目前监督新词发现方法大致有 2 种。一是将其转换成序列标注问题,即对大规模的领域文献进行新词标注,通过序列模型识别新词。文献[5]使用条件随机场发现开放领域的新词;文献[8]采用双向长短期记忆网络(bi-directional long short term memory,Bi-LSTM)与条件随机场,并结合词向量的前后信息进行实体识别;文献[9]使用将卷积神经网络加入 Bi-LSTM 与条件随机场组合模型中获得了更好的效果。二是将该问题转化为词语的相似

收稿日期:2020-06-05;修回日期:2020-07-18
基金项目:国家电网有限公司总部科技项目(5200-2019182-55A-0-0-00)

度计算问题。文献[10]使用百度百科的语料进行词的相似度计算;文献[11]则采用词相似度作为筛选新词的一种手段。

无监督发现新词方法需要根据相关统计指标对候选词进行筛选。其中的统计指标需要根据新词发现任务的特性,从词内和词外2个方面进行考虑。常用的指标包括点态互信息、信息熵、邻接变化度等[12—13]。文献[6]选取最佳统计指标筛选得到候选词,文献[14]使用互信息和邻接熵设置一系列过程对新词筛选。如何将统计指标与任务特性结合筛选新词是无监督领域新词发现的一个重要研究探索。在文本分类任务中,常见的关键词提取指标有词频与逆文档频率(term frequency-inverse document frequency,TF-IDF)和文本排序指标(TextRank)。文献[15]中提出了将文档长度作为TF-IDF的权重的关键词筛选方法;文献[16]结合典型词汇信息和TF-IDF进行文本分类;文献[17]通过双向正态分离(bi-normal separation,BNS)特征替换了IDF特征,结合BNS和TF作为新的关键词指标。TextRank的基本思想是将单词视作图中的节点,通过单词间的联系得到单词的重要程度。文献[18]提出了结合词向量和TextRank进行关键词提取的想法。最近,文献[19]结合TF-IDF和TextRank对生物医学领域的文献进行关键短语抽取。除以上传统统计量之外,词向量也经常被单独作为关键词提取指标。文献[20]将词向量融入了TextRank模型,文献[21]使用词向量抽取主题关键词。

2 电力专业词汇识别方法

2.1 电力词汇特点分析

专业领域不同,新词的特点往往不同。为更高效发现专业领域的新词,首先对该专业领域词汇进行分析,包括词长、词频、汉字分布及词汇分布等特征。电力领域词汇的词长、词元分布分别如图1、图2所示。

通过对电力领域词汇词典和典型通用词典的对比分析,发现该专业领域词汇有2个较明显的特征。一是专业词汇的词长较长,电力领域的词长范围在1~20之间,大多分布在2~10范围内,分布与图2对比较为均匀;二是专业领域词汇是由几个常规词汇组合构成,例如,“电压交流器”、“电压稳定”、“电压源”等都包含“电压”一词,再有“功角稳定”、“电压稳定”、“暂态稳定”等都包含“稳定”一词。分析发现电力词汇中包含词元(即常规词汇)个数的范围在1~10之间,且大多分布在1~4范围

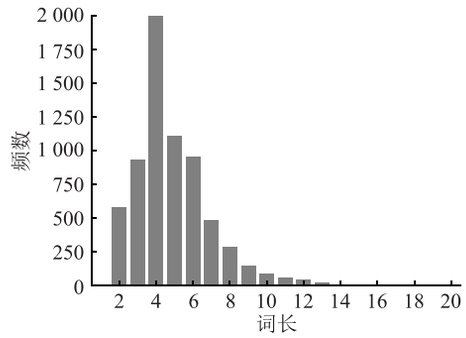


图1 电力领域词汇的词长分布

Fig.1 The word length distribution of power words

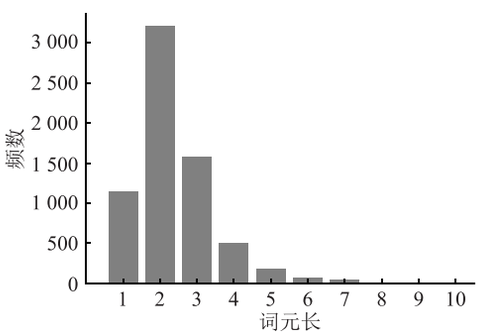


图2 电力领域词汇的词元分布

Fig. 2 The meta-words distribution of power words

内,且分布较为紧凑,如图2所示。典型通用词典的分析,词汇长度分布如图3所示。可以看到通用领域词汇大多分布在2~4范围内,词长较短,这与电力领域词汇有显著差异。

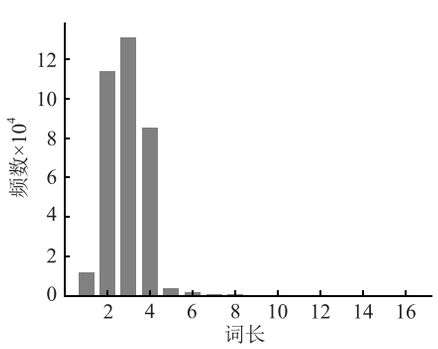


图3 通用词典的词长分布

Fig.3 The word length distribution of common words

2.2 发现方法

文中基于该领域词汇的特点,制定了如下的新词发现流程。首先对电力专业领域语料分词,然后选择合适大小的滑动窗口对窗口内的词进行组合,构建成候选词集。随后根据预定义的电力成词规则和统计特征对候选词进行过滤。整体的流程如图4所示。整体上,新词发现过程分为候选词生成和候选词过滤两大部分,其中候选词过滤又分为统计指标过滤和电力相关词典过滤两部分。

2.2.1 候选词生成

电力领域词汇的特点,即电力领域词汇是由多个词元组合而成,这是制定了候选词生成流程的重要依据。图 4 中,候选词生成过程主要包括预处理和滑动窗口组合 2 个模块;预处理模块包括文本的清洗、分词和去除停用词等功能;滑动窗口组合则是根据设置的滑动窗口大小将文本中相邻的窗口大小的词汇拼接起来。

对于常见的停用词,文中视其为非成词模块。为了能够更有效地减少候选词数量,在实际处理时将预处理中的去除停用词移至滑动窗口组合之后,即候选词中包含停用词,就将其过滤掉。

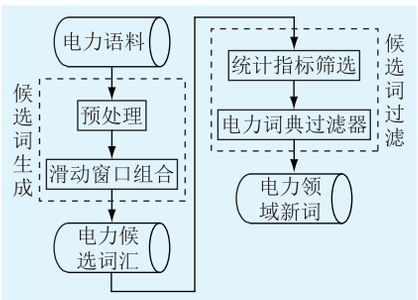


图 4 电力新词识别流程

Fig.4 Recognition flow of electric power new words

2.2.2 统计指标过滤器

统计指标过滤是指通过新词在语料中的统计特征,对新词进行过滤。文中使用如下 4 个指标。

(1) 邻接变化度(accessor variety, AV)。邻接变化度是衡量一个词串出现在不同语境中的可能性程度。如果该词适用于不同的语境,那么它串成词的可能性较大。在新词发现任务中,对于包含 n 个词的词串 $w=\{w_1,\cdots,w_n\}$,令 $V_L(w)$ 表示与该词串左相邻的不同字的个数, $V_R(w)$ 表示与该词串右相邻的不同字的个数。AV 的定义为:

$$V_G(w)=\log V(w) \tag{1}$$

其中, $V(w)=\min(V_L(w),V_R(w))$ 。文中对新词的筛选是通过统计指标的搜索。为保留更多的统计信息,文中采用 $V_L(w)$ 和 $V_R(w)$ 共同作为该文统计指标。

(2) 信息熵(information entropy, IE)。信息熵是用来衡量一个随机变量的不确定性。随机变量的信息熵越大,其不确定性就越大。在新词发现中,使用信息熵来衡量词串前后邻接字的不确定性,词串的前后信息熵越大,说明词串越有可能单独成词,否则其更有可能与前后文信息结合成词。

$$H_L(w)=\sum_{n=1}^{N}-P(w_n)\log P(w_n) \tag{2}$$

$$H_R(w)=\sum_{m=1}^{M}-P(w_m)\log P(w_m) \tag{3}$$

式中: w 为候选词串; N 为候选词串前邻接字的总数; M 为候选词串后邻接字的总数。与邻接变化度类似,保留了词串的前后信息熵作为统计指标。

(3) 点态互信息。该指标考察新词出现的可能。对于 2 个词 u 和 v,其点态互信息 $I_p(u,v)$ 定义为:

$$I_p(u,v)=\log\frac{P(u)P(v)}{P(u,v)} \tag{4}$$

其中, $P(u)$ 和 $P(v)$ 分别为词 u 和 v 在文档中出现的概率。 $P(u,v)$ 是由 u 和 v 组合成的新词在文档中出现的概率。如果互信息的值越大,则表明 u 和 v 组合成新词的可能性越高;反之,则表明 u 和 v 之间存在短语边界的可能性越高。

(4) 词频。该指标考查新词独立出现的可能性,此外根据 N-gram 组合特点,使用词频作为过滤指标,能过滤掉大量不相关的稀疏候选词。由于候选词和领域词的分布差异较大,为了使新词既满足领域词的分布,又能够得到较好的准确率,使用文中提出的生成候选词方法。上述统计指标在候选词和领域词的分布差异较为明显,如图 5、图 6 和图 7 所示。以往新词发现工作中,常根据多种统计指标的分布情况,选出 1 或 2 个指标作为筛选指标,再使用这种方法对候选词进行筛选。这类方法有较高的效率,但新词发现的结果往往不能令人满意。基于此,采用上述统计量共同作为筛选指标,对每个指标的阈值使用搜索的形式进行探索。具体而言,根据每个统计指标的分布情况设置阈值区间,以最大化某一函数作为搜索目标。

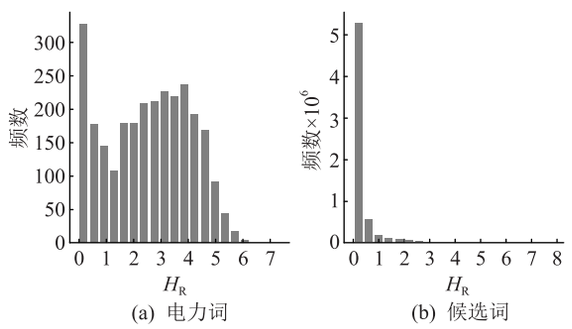


图 5 信息熵指标 H_R 在电力词和候选词的分布差异

Fig.5 The distribution difference of information entropy index H_R between domain words and candidate words

文中采用区间搜索方法。其中区间参数[a, b]可以依据思路设定。如果该统计指标的领域词典分布相对于候选词是右偏分布, a 是所有词该统计指标最小值, b 是领域词典统计指标的分位数;如果

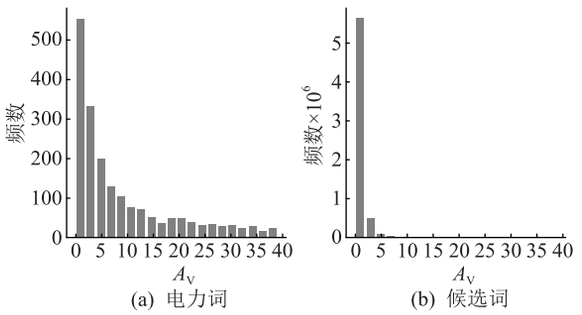


图 6 邻接变化度指标 A_V在电力词和候选词的分布差异

Fig.6 The distribution difference of adjacent change index A_V between domain words and candidate words

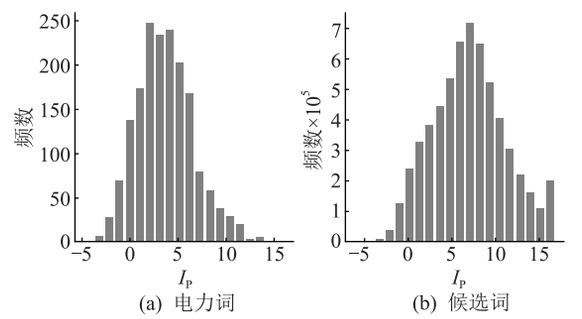


图 7 点态互信息指标 I_P在电力词和候选词的分布差

Fig.7 The distribution difference of point mutual information index I_P between domain words and candidate words

领域词典分布相对于候选词是左偏分布，b 是所有词该指标的最大值，a 是领域词典的分位数。对于分位数的选择，可以依据电力语料上的实验情况作出选择。文中最大化参数为精度和召回率的乘积。

2.2.3 电力词典过滤器

采用电力词典过滤器对候选词进行过滤。电力词典过滤器主要包括边位非成词字词典的过滤、人工筛选的非成词词典的过滤以及非电力领域词典的过滤 3 个方面，具体如下。

（1）基于语法规则的过滤。该词典记录了常见的非成词单字，包含起始词典和结束词典两部分。如果候选词的起始和结束位置包含对应词典中的任意一个字符，则该候选词无效。该词典包含“住”、“及”、“后”、“前”、“至”、“更”等词。

（2）基于电力领域非成词词典的过滤。该词典记录了电力领域常见的非成词单词，如果候选词包含该词典的词汇中任意一个字符，则该候选词无效。该词典包含“业务过程”、“互补间隙”、“仪测量”、“低温侧”等词。

（3）基于非电力领域词典的过滤。该词典记录了大量学术论文类和公司名称类等词汇。基于此，可以过滤明显不属于电力领域的词汇。该词典包含“参考文献”、“卷第”、“实验结果”、“实验研究”、“国家电网公司”、“浙江省电力公司”、“水利水电出版社”等词。

3 实验

3.1 数据集

为评估文中提出的方法，采用中国人工智能产业发展联盟（AIIA）比赛训练数据集以及公开电力词典作为实验评估数据。该公开电力词典共包含电力专业词汇 6 000 个。数据集 AIIA 可以从网址（https://www.datafountain.cn/competitions/320）下载。表 1 展示了该电力专业词典的 40 个词汇。

表 1 电力专业词典中词的展示

Table 1 Words demonstration in power dictionary

序号	词汇	序号	词汇	序号	词汇
1	功率	15	波特图	29	直接复位端
2	阻抗	16	继电器	30	移位寄存器
3	档位	17	原件设备	31	同步计数器
4	仿真	18	状态参数	32	感应式电机
5	工频	19	正序阻抗	33	永磁同步电机
6	波形	20	负序阻抗	34	三绕组变压器
7	转子	21	电压等级	35	双绕组变压器
8	感抗	22	硅二极管	36	道闸隔离开关
9	电容器	23	夹断电压	37	无刷直流电机
10	故障点	24	直流分量	38	交流环电动机
11	软启动	25	射极跟随器	39	锁定转子转矩
12	电动机	26	差动放大器	40	高压侧输电线
13	初相位	27	迟滞比较器		
14	角频率	28	方波发生器		

3.2 比较方法

采用 4 种发现新词的方法，包括基于筛选指标（screening indicators，SI）的方法、结合过滤非电力领域词（combining filtration of non-domain words，NF）的方法即 SI+NF、结合语法规则过滤词（filtering words with grammatical rules，GF）的方法即 SI+GF 以及结合非电力领域词过滤和语法规则过滤的方法即 SI+NF+GF。

3.3 评价指标

为评价提出方法的有效性，采用精度和召回率 2 个指标评估分别定义为：

$$P \triangleq \frac{N_{w_t}}{N_{w_c}} \tag{5}$$

$$R \triangleq \frac{N_{w_t}}{N_{w_d}} \tag{6}$$

式中：P，R 分别为精度和召回率；N_{w_t} 为发现的词在电力词典中的数目；N_{w_c} 为候选词的数目；N_{w_d} 为候选词中含电力词典中的数目。精度指标 P 考察提

出方法区分电力新词的能力,召回率 R 指标考察所提出方法涵盖电力新词的能力。

为给出一个单一指标来更好反映不同方法的性能,提出将精度和召回率乘积作为指标,称之为 F 指标,即:

$$F \triangleq P \times R \tag{7}$$

在区间搜索方法上,主要存在左偏和右偏 2 种情况。情况一,如果电力领域词的该统计指标的分布相对于候选词是右偏分布,如邻接变化度,左边界是所有词该统计指标的最小值,右边界是领域词典统计指标的分位数。情况二,如果电力领域的分布相对于候选词是左偏分布,如点互信息,右边界是所有词该统计指标的最大值,左边界是领域词典统计指标的分位数。对于分位数的选择,可以根据统计指标分布的实际情况选择,主要涉及到模型的泛化能力。文中建议的分位点如下:情况一选择 30%~50%,情况二选择 50%~70%。以点互信息为例,电力字典中相较于候选词为左偏分布,所以其搜索范围为[4.9, 16.8],4.9 为电力词典该统计指标的 70 分位点,16.8 为该指标最大值。

3.4 实验结果

4 种发现新词方法的精度、召回率和 F 指标如表 2 所示。为清晰起见,表 2 同时给出了新词数,且 3 个性能指标的结果用百分比表示。由表 2 可见,混合方法取得了最好的性能。而且过滤词结合统计指标相较于词频结合统计指标会有更好的结果,也证明了过滤词典的有效性。

表 2 不同电力新词发现方法的实验结果对比

Table 2 Comparison of experimental results with different electric power words discovery methods

方法	新词数	精度/%	召回率/%	F 指标
SI	29 865	4.02	40.4	1.62
SI+NF	29 856	4.02	40.4	1.62
SI+GF	24 401	4.8	39.42	1.89
SI+NF+GF	22 670	5.14	39.19	2.01

为进一步展示新词发现方法的特点,表 3 给出不同步骤得到的候选词数和相应包含在词典中的词数。由表 3 可见,随着不同处理方式的加入,候选词数和包含词典中词数均处于下降趋势,但候选词的下降趋势显然较大。此外,不同的处理方式都使候选词数明显下降,说明使用此过滤方法有效。

最好的混合方法与其他方法相比,在候选词数和包含词典中词数相较于基本的统计筛选方法如表 4 所示。可以看出,混合方法使候选词数大大降低。同时在词典中词数只是少量降低,混合方法的代价最低,效果最好。

表 3 实验中电力新词的变化

Table 3 Quantitative changing of electric power words with experiments

处理方式	候选词数	含词典中词数
滑动窗口组合	1.57×10^7	2 973
去除停用词	6.46×10^6	2 758
筛选指标	29 865	1 201
非电力词典过滤	29 856	1 201
语法规则过滤	22 670	1 165

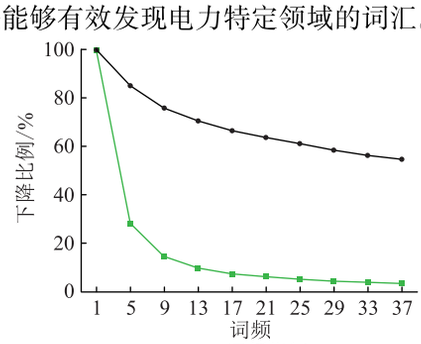
表 4 电力候选词变化比例对比

Table 4 Comparison of proportions of electric power word candidates %

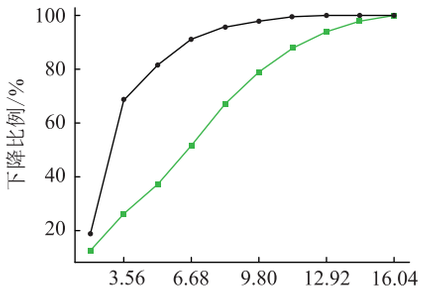
方法	候选词比例	含词典中词比例
SI	100.0	100.0
SI+NF	100.0	100.0
SI+GF	81.7	87.6
SI+NF+GF	75.9	97.0

针对统计指标的筛选,采用搜索方式,展示了在单指标筛选下的实验结果。与表 4 类似,以候选词数和包含词典中词数相较于初始阈值的下降百分比进行说明。如图 8 所示,词频、邻接变化度、信息熵等指标在电力词典中的分布相较于候选词均为右偏分布,因此其阈值从小到大,候选词和字典中词均处于下降趋势。而点互信息字典中词相较于候选词为左偏分布,因此阈值从小到大,候选词和字典中词均处于上升趋势。临界变化度和信息熵在初始阶段,字典中词下降趋势较慢,而候选词下降迅速,说明该指标在边缘部分筛选效果较好;信息熵则是在到达某一阈值后,候选词相较于字典中词下降明显;点互信息在整个区间熵字典中词和候选词下降趋势相差不大。不同的指标往往有不同的特性,多个指标间常存在某种联系,这样繁杂的关系也验证了使用参数搜索的必要性。

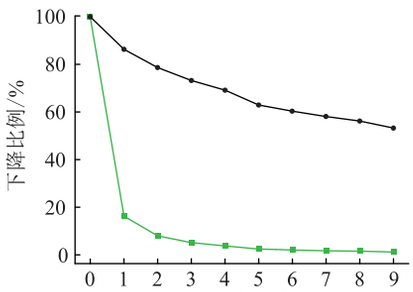
此外,该方法能够发现“耦合电容器”、“串联补偿装置”等国家电网公司变电运维检修管理办法中的典型词汇[22]。而这类词在给定的领域词典中是不存在的。也就是说,文中提出的电力领域新词发现方法能够有效发现电力特定领域的词汇。



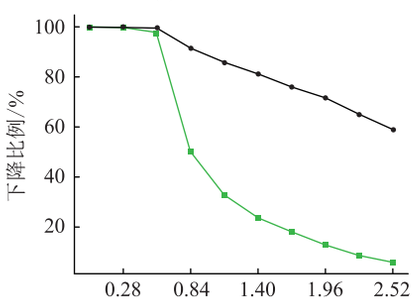
(a) 词频指标



(b) 点互信息指标



(c) 临界变化度指标



(d) 信息熵指标

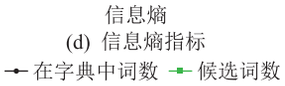


图 8 电力新词在 4 种指标下的变化比例对比

Fig.8 Comparison of proportions of candidate words under 4 indicators

4 结语

文中结合电力领域词汇的语言学特征提出一种面向电力领域的无监督新词发现方法。首先以基础词典对语料分词,然后以不同滑动窗口大小对分词结果的多种组合作为候选词,进一步基于 4 种词频统计量和电力相关词典过滤对候选词进行筛选,形成最终的专业新词。在电力公开数据集上的评测,文中提出的方法获得了较高的准确率和召回率。这将为后续领域知识图谱等应用工作奠定基础。

同时,文中提出的电力领域的新词发现方法还有进一步研究提升的空间。例如,可以考虑建立词向量,通过词向量的相似度来发现更多新词。再者,探索可以标注少部分数据,使用弱监督或者半监督的机器学习方法提升效率和性能。

参考文献:

[1] PALUMBO E, RIZZO G, TRONCY R, et al. Knowledge graph embed-dings with node2vec for item recommendation[C]//European Semantic Web Conference. 2018:117-120.

[2] 黄奇峰,杨世海,邓欣宇,等. 基于欠完备自编码器的用户用电行为分类分析方法[J]. 电力工程技术,2019,38(6):24-30.
HUANG Qifeng, YANG Shihai, DENG Xinyu, et al. Classification analysis method for electricity consumption behavior based on undercomplete autoencoder[J]. Electric Power Engineering Technogly,2019,38(6):24-30.

[3] WANG B, LIU B, WANG X, et al. Deep learning approaches to semantic relevance modeling for Chinese question-answer pairs [J]. ACM Transactions on Asian Language Information Processing,2011,10(4):1-16.

[4] 刘成民,戴中坚,陈轩. 基于 TensorFlow 框架的有源配电网深度学习故障定位方法[J]. 电力工程技术,2019,38(5):8-15.
LIU Chengmin, DAI Zhongjian, CHEN Xuan. A fault location method for active distribution network based on tensorflow deep learning[J]. Electric Power Engineering Technogly, 2019, 38(5):8-15.

[5] 陈飞,刘奕群,魏超,等. 基于条件随机场方法的开放领域新词发现[J]. 软件学报,2013(5):1051-1060.
CHEN Fei, LIU Yiqun, WEI Chao, et al. Open domain new word detection using condition random field method[J]. Journal of Software,2013(5):1051-1060.

[6] 张婧,黄锴宇,梁晨. 面向中文社交媒体语料的无监督新词识别研究[J]. 中文信息学报,2018,32(3):17-25.
ZHANG Jing, HUANG Kaiyu, LIANG Chen. Fusion of statistics and textrank for keyphrase extraction in biomedical literature [J]. Journal of Chinese information,2018,32(3):17-25.

[7] 丁祥武,张夕华. 医疗领域文本结构化[J]. 计算机工程与设计,2017,38(10):2873-2878.
DING Xiangwu, ZHANG Xihua. Text structuralization in medical field[J]. Computer Engineering and Design,2017,38(10):2873-2878.

[8] 陈伟,吴友政,陈文亮,等. 基于 BiLSTM-CRF 的关键词自动抽取[J]. 计算机科学,2018,45(S1):104-109,126.
CHEN Wei, WU Youzheng, CHEN Wenliang, et al. Automatic keyword extraction based on BiLSTM-CRF[J]. Computer Science,2018,45(S1):104-109,126.

[9] 顾孙炎. 基于深度神经网络的中文命名实体识别研究[D]. 南京:南京邮电大学,2018.
GU Sunyan. Research on Chinese named entity recognition based on deep neural network[D]. Nanjing: Nanjing University of Posts and Telecommunications,2018.

[10] 詹志建,梁丽娜,杨小平. 基于百度百科的词语相似度计算[J]. 计算机科学,2013,40(6):199-202.
ZHAN Zhijian, LIANG Lina, YANG Xiaoping. Word similarity measurement based on baidu baike[J]. Computer Science, 2013,40(6):199-202.

[11] 王欣. 微博新词发现及新词情感极性判断方法[D]. 重庆:重庆师范大学,2018.
WANG Xin. On the discovery of new words on Weibo and judgment method about new words′ sentimental polarity[D]. Chongqing:Chongqing Normal University,2018.

[12] FENG H,CHEN K,DENG X,et al. Accessor variety criteria for Chinese word extraction[J]. Computational Linguistics,2014,30(1):75-93.

[13] 欧阳冠宇. 基于组合频率的中文新词发现算法[D]. 北京:北京邮电大学,2018.
OUYANG Guanyu. Algorithm for discovering new Chinese words based on combination frequency[D]. Beijing:Beijing University of Posts and Telecommunications,2018.

[14] 刘伟童,刘培玉,刘文锋. 基于互信息和邻接熵的新词发现算法[J]. 计算机应用研究,2019,36(5):19-22.
LIU Weitong,LIU Peiyu,LIU Wenfeng. New word discovery algorithm based on mutual information and adbranch entropy[J]. Computer Application Research,2019,36(5):19-22.

[15] 贺科达,朱铮涛,程昱. 基于改进 TF-IDF 算法的文本分类方法研究[J]. 广东工业大学学报,2016,33(5):49-53.
HE Keda,ZHU Zhengtao,CHEN Yu. Research on text classification method based on improved TF-IDF algorithm[J]. Journal of Guangdong University of Technology,2016,33(5):49-53.

[16] ZHANG Y,GONG L,WANG Y. An improved TF-IDF approach for text classification[J]. Journal of Zhejiang University-Science A,2015,6(1):49-55.

[17] FORMAN G. BNS feature scaling:an improved representation over TF-IDF for SVM text classification[C]//Proceedings of the 17th ACM Conference on Information and Knowledge Management. 2008:263-270.

[18] 刘奇飞,沈炜域. 基于 Word2Vec 和 TextRank 的时政类新闻关键词抽取方法研究[J]. 情报探索,2018,248(6):26-31.
LIU Qifei,SHEN Weiyu. Research of keyword extraction of political news based on Word2Vec and TextRank[J]. Information Research,2018,248(6):26-31.

[19] 魏赟,孙先朋. 融合统计学和 TextRank 的生物医学文献关键短语抽取[J]. 计算机应用与软件,2017,34(6):27-30.
WEI Yun,SUN Xianpeng. Fusion of statistics and TextRank for keyphrase extraction in biomedical literature[J]. Computer Applications and Software,2017,34(6):27-30.

[20] 夏天. 词向量聚类加权 TextRank 的关键词抽取[J]. 现代图书情报技术,2017,1(2):28-34.
XIA Tian. Extracting keywords with modified Textrank model[J]. Modern Library and Information Technology,2017,1(2):28-34.

[21] 马晓军,郭剑毅,王红斌,等. 融合词向量和主题模型的领域实体消歧[J]. 模式识别与人工智能,2017,30(12):1130-1137.
MA Xiaojun,GUO Jianyi,WANG Hongbin,et al. Entity disambiguation in specific domains combining word vector and topic models[J]. Pattern Recognition and Artificial Intelligence,2017,30(12):1130-1137.

[22] 国网(运检/2)826—2017. 国家电网公司变电运维检修管理办法(试行)[Z]. 2017.
State Grid (Operation and Inspection/2) 826—2017. Management measures for substation operation maintenance and maintenance of state grid corporation (trial)[Z]. 2017.

作者简介:


朱婷婷

朱婷婷(1996),女,硕士在读,研究方向为自然语言处理(E-mail:lionztt@163.com);

杜一帆(1993),男,硕士在读,研究方向为自然语言处理;

李睿凡(1975),男,通信作者,博士,副教授,研究方向为多模态机器学习与自然语言处理(E-mail:rfli@bupt.edu.cn)。

An unsupervised approach to recognizing new words in power domain

ZHU Tingting[1], DU Yifan[1], LI Ruifan[1,2], XIONG Yongping[3]

(1. School of Artifical Intelligence,Beijing University of Posts and Telecommunications,Beijing 100876,China;
2. Engineering Research Center of Information Networks,Ministry of Education,Beijing 100876,China;
3. School of Computer Science,Beijing University of Posts and Telecommunications,Beijing 100876,China)

Abstract:The terminology word recognition in power domain lays the foundation for a profound language understanding of power documents and the intelligent knowledge graph construction. By incorporating the morphology of the power domain vocabulary, an unsupervised approach to recognizing new terminology words in documents is proposed. Firstly, the common dictionary is used to segment the corpus. Then segmented words are combined with terminology feature-based sliding window of different sizes constituting candidate words. Furthermore, four statistics including accessor variety, information entropy, point-wise mutual information, and word frequency are computed. Finally, based on the linguistics statistics and three types of word-formation grammatical rules, those words are screened generating the last electric new words. Experimental results on a public dataset demonstrate the effectiveness of our proposed method.

Keywords:domain dictionary;unsupervised learning;new word recognition;sliding window;statistical features

(编辑 钱悦)