机械设计领域的命名实体识别研究

陈秋瑗[1,2],程 光[1,2],李 迪[1,2],张 建[1,2]
CHEN Qiuyuan[1,2], CHENG Guang[1,2], LI Di[1,2], ZHANG Jian[1,2]

1.北京联合大学 机器人学院,北京 100020
2.北京市智能机械创新设计服务工程技术研究中心,北京 100020
1.College of Robotics, Beijing Union University, Beijing 100020, China
2.Beijing Engineering Research Center of Smart Mechanical Innovation Design Service, Beijing 100020, China

CHEN Qiuyuan, CHENG Guang, LI Di, et al. Named entity recognition for mechanical design and manufacturing area. Computer Engineering and Applications, 2017, 53(20):100-104.

Abstract: Named entity recognition has wide applications in the area of natural language processing, but the common methods cannot accurately identify the proper nouns in manufacture area. In order to solve the named entity recognition for mechanical design and manufacturing area, this paper proposes a new machine learning based method. It carefully finds some statistical features, and then uses the logistic regression algorithm to calculate the tightness between two adjacent strings. The proposed method can recognize proper nouns more accurately and efficiently for mechanical design and manufacturing area.

Key words: named entity recognition; mechanical design; logistic regression; tightness

摘 要:命名实体识别技术在自然语言处理技术中占有重要的地位,通用的方法不能很好地解决机械领域的识别问题。基于字符串之间紧密相邻程度等统计特征,定义不同词之间紧密相连的程度,从而识别机械领域的领域词。通过计算特征值,用逻辑回归的方法确定相邻字串的紧密相邻程度,从而发现新词。该方法对比通用的方法准确率和召回率得到了提高,更好地识别机械领域的领域词。

关键词:命名实体识别;机械领域;逻辑回归;紧密相邻
文献标志码:A **中图分类号:**TP391.1 doi:10.3778/j.issn.1002-8331.1604-0231

1 引言

近年来,搜索引擎、网络新闻作为互联网的基础应用,使用率均在80%以上。为了方便人们在机械设计与制造领域更好地搜集到想用的信息,如何从海量的网页中抽取到感兴趣的内容是亟待解决的问题。命名实体识别是一个领域信息检索[1]的基础,所以研究命名实体识别变得非常必要。有研究表明[2-6],约60%的分词错误都是由新词导致的,所以有效识别新词是中文词法分析领域亟待解决的问题。

命名实体(Named Entity,NE)是文本中基本的信息单位,是文本中的固有名称、缩写及其他唯一标识,是正确理解文本的基础。命名实体识别[7]是信息提取、问答系统、句法分析、机器翻译、元数据标注等应用领域的重要基础性工作,在自然语言处理技术走向实用化的过程中占有重要地位。

命名实体识别有代表性的方法主要分为以下两种[8]:基于规则的方法和基于统计的方法。基于这两种方法,有人提出了基于规则和统计相结合的新方法。Zhang等人[9]引入角色标注未登录词识别概念,首先用贝叶斯公式和Viterbi算法进行设定角色标注,然后用模

基金项目:科技部创新方法工作专项(No.2015IM020100);2015年智能制造专项智能制造新模式项目。
作者简介:陈秋瑗(1991—),女,硕士研究生,主要研究方向:信息检索;程光(1964—),通讯作者,男,博士,教授,主要研究方向:先进制造与工业工程、现代设计方法与理论,E-mail:chengguang@buu.edu.cn;李迪(1992—),男,硕士研究生,主要研究方向:机电一体化;张建(1991—),男,硕士研究生,主要研究方向:机电一体化。
收稿日期:2016-04-18 **修回日期:**2016-08-29 **文章编号:**1002-8331(2017)20-0100-05
CNKI **网络优先出版:**2016-12-23, http://www.cnki.net/kcms/detail/11.2127.TP.20161223.1703.002.html

式匹配方法来获得新词候选集,最后用规则过滤实现新词集合。贺敏等人通过上下文邻接分析,位置成词概率以及双字耦合度来进行新词过滤[10];施水才等人通过频率比,互信息以及概率比的方法对新词过滤[11];林自芳等通过从语料库中统计词的内部模式,结合互信息和位置成词概率对新词进行过滤[12]。基于统计的方法虽然能找到大量的新词,但是同时也产生了大量的垃圾词串,即基于统计的方法在保证召回率的同时却也降低了准确率。霍帅等将统计方法与词法信息相结合,提出引用词关联性信息的迭代上下文熵算法,得到候选新词列表,并用词法特征与统计特征相结合的方法进行过滤[13]。

针对机械设计与制造领域,常见的命名实体识别方法不能有效识别领域词汇,即专有名词(也称为新词)。比如:热处理,很容易把它拆分成两个词,热和处理。因此,对命名实体识别的研究具有重要的理论和现实意义。本文针对机械设计与制造领域的文本,提出一种新的命名实体识别方法,运用信息熵中左右熵算法,可以算出每个字符的左右熵。接下来可以用词法特征与统计特征相结合的方法定义紧密度公式。最后提出一种基于紧密度的新词识别方法,使其能更高效更快速地发现新词,从而更好地将信息检索应用于机械设计与制造、产品设计的领域。

2　背景知识

在自然语言处理中,对于字串来说左右熵是很重要的统计特征,它体现了字串的上下文活跃程度。如果某个字词左右熵都比较高,说明它上下文搭配丰富,所以说左右熵可以代表字串的上下文丰富程度。在术语抽取、新词检测[14]领域有着非常广泛的应用。信息熵的直观意义:

表1、2列举了信息熵应用的直观例子。变量 X 代表 A 和 B 的比赛结果,两种不同的情况下,由于概率分布不同,导致熵的结果不同。一般来说,如果熵越高,结果更随机。

表1　A 和 B 比赛输赢的概率分布及熵

变量 X	A 赢	B 赢
概率	0.9	0.1
信息熵 H	0.325	

表2　A 和 B 比赛输赢的概率分布及熵

变量 X	A 赢	B 赢
概率	0.5	0.5
信息熵 H	0.693	

变量的不确定性越大,熵也就越大,把它搞清楚所需要的信息量也就越大。一个系统越是有序,信息熵就越低;反之,一个系统越是混乱,信息熵就越高。所以,信息熵也可以说是系统有序化程度的一个度量[14]。

在信源中,考虑的不是某一单个符号发生的不确定性,而是要考虑这个信源所有可能发生情况的平均不确定性。若信源符号有 n 种取值:$U_1U_2\cdots U_i\cdots U_n$,对应概率为:$P_1P_2\cdots P_i\cdots P_n$,且各种符号的出现彼此独立。这时,信源的平均不确定性应当为单个符号不确定性 $-\mathrm{lb}\,P_i$ 的统计平均值(E),可称为信息熵[14],即

$$H(U)=E\left[-\mathrm{lb}\,p_i\right]=-\sum_{i=1}^{n}p_i\mathrm{lb}\,p_i \tag{1}$$

式中对数单位为比特,并且当熵越大的时候说明比特越高,此时表明越混乱。

3　基于紧密度的新词发现方法

本文针对机械设计与制造领域的文本,提出一种新的命名实体识别方法,运用信息熵中左右熵算法,可以算出每个字符的左右熵。接下来可以用词法特征与统计特征相结合的方法确定紧密度公式。

3.1　左右熵的计算

基于信息熵的概念,用左右熵来解决特定领域新词发现问题。左右熵可以作为相邻两个字之间是否可以组成一个词的重要参考维度。左右熵的计算公式[15]:

$$LE=-\sum_{i=0}^{n}p_i(xAB|AB)\mathrm{lb}\,p(xAB|AB) \tag{2}$$

其中 x 为所有出现在 AB 左边的词。

$$RE=-\sum_{i=0}^{n}p_i(ABy|AB)\mathrm{lb}\,p(ABy|AB) \tag{3}$$

其中 y 为所有出现在 AB 右边的词。

左右熵主要用于度量一个词左右两边出现不同词的丰富程度。通过计算得到每个词项的左右侧出现的其他词的概率,从而得到左右熵的值。基于此,提出左右熵计算方法,具体步骤如下所示:

步骤1　建一个表Table,将分词结果中每个字并遍历一遍,将遍历的每个字 w_i 存到Table中并记下出现的次数 n。

步骤2　建第二张表Table1,找到 w_i 左(右)相邻的位置出现的字 $T_i(v_i)$,并记录下它们出现的次数 $A_i(B_i)$。将结果存入到Table1中。

步骤3　计算出 $T_i(v_i)$ 出现在 w_i 左右两边的概率。

步骤4　根据左右熵的计算公式,计算出左熵和右熵。

3.2　似然比

似然比(Likelihood Ratio,LR)是反映真实性的一种指标,属于同时反映灵敏度和特异度的复合指标。在

其最早应用中是为体现有病者中得出某一筛检实验结果的概率与无病者得出这一概率的比值。

从参考文献中,可以得到似然比的计算公式[16]如下。首先定义变量 p,p_1,p_2 ,把词 w^1,w^2,w^{12} 在文档中出现的次数分别记作 c_1,c_2,c_{12} 可得:

$$p=\frac{c_2}{N},p_1=\frac{c_{12}}{c_1},p_2=\frac{c_2-c_{12}}{N-c_1} \tag{4}$$

假定二项分布:

$$b(k;n,x)=\binom{n}{k}X^k(1-x)^{(n-k)} \tag{5}$$

在实际观察中得到 w^1,w^2和w^{12} 可能的出现次数,下面两种情况分别代表了相互独立情况下和有依赖关系的情况下的发生 N 次事件中 w^2 出现 c_2 次的概率:

$$L(H_1)=b(c_{12},c_1,p)b(c_2-c_{12},N-c_1,p) \tag{6}$$

$$L(H_2)=b(c_{12},c_1,p_1)b(c_2-c_{12},N-c_1,p_2) \tag{7}$$

从文章中可以知道似然比的公式表示[16]如下所示:

$$\mathrm{lb}\,\lambda=\mathrm{lb}\frac{L(H_1)}{L(H_2)}=\mathrm{lb}\frac{b(c_{12},c_1,p)b(c_2-c_{12},N-c_1,P)}{b(c_{12},c_1,p_1)(c_2-c_{12},N-c_1,P_2)} \tag{8}$$

将该理论应用到本文的新词发现中,做出如下的假设及推理过程:

AB 两个词同时出现的次数用 a 表示,A 没有出现但 B 出现的次数用 b 表示,B 没有出现但 A 出现的次数用 c 表示,B 没有出现且 A 也没有出现的次数用 d 表示。那么对应的就有 c_1 为 A 出现的次数,c_2 为 B 出现的次数。

$$P=\frac{a+b}{N} \tag{9}$$

$$P_1=\frac{a}{a+c} \tag{10}$$

$$p_2=\frac{b}{b+d} \tag{11}$$

假设一个二项分布:

$$b(k;n,x)=\binom{n}{k}x^k(1-x)^{n-k} \tag{12}$$

在这里:

$$L(k,n,x)=x^k(1-x)^{n-k} \tag{13}$$

把公式(9)(10)(11)带入公式(8)可知:

$$\mathrm{lb}\,\lambda=\mathrm{lb}\,L\left(a,a+c,\frac{a+b}{N}\right)+\mathrm{lb}\,L\left(b,b+d,\frac{a+b}{N}\right)-\mathrm{lb}\,L\left(a,a+c,\frac{a}{a+c}\right)-\mathrm{lb}\,L\left(b,b+d,\frac{b}{b+d}\right) \tag{14}$$

根据公式(13)和(14),可得:

$$\mathrm{lb}\,\lambda=\mathrm{lb}\left(\frac{(a+b)\times(a+c)}{N\times a}\right)^a+\mathrm{lb}\left(\frac{(c+d)\times(a+c)}{N\times c}\right)^c+\mathrm{lb}\left(\frac{(a+b)\times(b+d)}{N\times b}\right)^b+\mathrm{lb}\left(\frac{(c+d)\times(b+d)}{N\times c}\right)^d \tag{15}$$

由参考文献[16]可推知似然比公式如下:

$$LLR(x,y)=A=\mathrm{lb}\,\lambda$$

$$LLR(x,y)=2(a\times\mathrm{lb}\frac{a\times N}{(a+b)\times(a+c)}+b\times\mathrm{lb}\frac{b\times N}{(a+b)\times(b+d)}+c\times\mathrm{lb}\frac{c\times N}{(c+d)\times(a+c)}+d\times\mathrm{lb}\frac{d\times N}{(c+d)\times(b+d)}) \tag{16}$$

3.3 紧密度的计算方法

接下来引入紧密度的概念,作为新词识别的主要标准。AB 两个词之间的紧密度表示 A 和 B 成为一个新词的概率。紧密度与下面五个特征紧密相关。通过逻辑回归的方法可以确定五个特征的权重,从而得到紧密度的计算公式。基于紧密度的结果就可以识别出新词。紧密度需要考虑的特征如下。

(1)两个词紧密相邻的情况的概率(相邻概率):紧密出现的次数/总的统计样本数。

(2)两个词在定长窗口内顺序出现的概率(相近概率):在定长窗口出现的次数/总的统计样本数。

(3)两个词逆序出现的概率(单调概率):逆序出现的次数/总的样本数。

(4)左右熵:通过熵的公式计算一个词的左熵和右熵。熵值越高,与其他字组合在一起的样式就越多,就越不可能组成一个词组:

$$LE=-\sum_{i=0}^{n}p_i(xAB|AB)\mathrm{lb}\,p(xAB|AB) \tag{17}$$

$$RE=-\sum_{i=0}^{n}p_i(ABy|AB)\mathrm{lb}\,p(ABy|AB) \tag{18}$$

(5)似然比:首先构建似然比假设如表3。

表3 似然比假设

	A出现的次数	A没有出现的次数
B出现的次数	a	b
B没有出现的次数	c	d

似然比(Likelihood Ratio)的计算公式定义为(其中 $N=a+b+c+d$):

$$LLR(x,y)=2(a\times\mathrm{lb}\frac{a\times N}{(a+b)\times(a+c)}+b\times\mathrm{lb}\frac{b\times N}{(a+b)\times(b+d)}+c\times\mathrm{lb}\frac{c\times N}{(c+d)\times(a+c)}+d\times\mathrm{lb}\frac{d\times N}{(c+d)\times(b+d)}) \tag{19}$$

综上所述,紧密度主要依赖于相邻概率、相近概率、单调概率、左右熵和似然比。相邻概率和相近概率从文本中共同出现的概率角度来衡量两个词一起出现的概率。单调概率从顺序的角度来确保两次词是以固定的顺序出现来组成一个新词。左右熵从概率的角度确保两个词相对其他词更容易组成一个新词。似然比从假设检验的角度来衡量能够准确识别一个词的概率。假

设五个特征值分别是 w_1,w_2,w_3,w_4,w_5 ,那么

$$紧密度=w_1\times f_1+w_2\times f_2+w_3\times f_3+w_4\times f_4+w_5\times f_5 \quad (20)$$

其中, w 表示权重, f 表示特征值。

基于紧密度的概念,本文提出新词识别方法,主要包括以下三个步骤:

(1)对指定文本进行预处理并分词。

(2)针对分词结果计算对应的五个特征值。

(3)利用人工标注结果对五个特征进行逻辑回归,计算五个特征构成紧密度的权重;并确定构成一个新词的紧密度阈值;在第二步中,需要计算五个特征对应的取值。其中相邻概率、相近概率和单调概率三个特征可以通过遍历文本统计得到。在遍历的过程中,记录下两个词共同出现、逆序出现、和定长序列中出现的次数。计算复杂 度与文本长度线性相关。左右熵的特征值通过3.1节的方法进行计算。似然比的计算过程与左右熵相似,对于任意两个共同出现的词 A 和 B ,在计算左右熵的过程中,已经统计出来他们共同出现的次数。 A 出现且 B 没有出现的次数,可以通过A出现的总次数减去 AB 共同出现的次数得到。同理,可以得到似然比需要的其他值,从而计算出似然比的值。

在第三步中,利用标注结果进行逻辑回归,从而确定各个特征的权重。标注的结果如 (f_1,f_2,f_3,f_4,f_5,R) ,其中 R 取值为0或者1,表示是否是一个新词。将标注的结果集利用逻辑回归进行训练,从而得到权重结果,从而获得紧密度的计算公式。得到紧密度的计算公式后,可以按照公式计算出所有标注结果中的 R 值,利用决策树的方法确定紧密度阈值,从而得到是否是一个新词的过滤条件。

4　实验结果分析

一般文本中会包含许多不规范的词或符号,这些词或符号对命名实体的识别没有实际意义,对文本进行规范化处理,既不会对文本整体语义产生影响,又能减少大量干扰信息。所以,本文首先对爬到的文本内容进行了预处理,剔除了一些常用的分隔符、标点符号和助词。

训练过程如下。

计算特征:对于候选集的词汇,首先计算它们对应的特征概率。

标注数据:标注两个词 AB 是否应该组成一个词组:1为应该,0为不应该。

在计算特征的时候,相近概率特征的定长窗口在本文中取值为5。在实际计算过程中左熵和右熵作为两个特征进行考虑。利用这些训练数据通过逻辑回归进行训练,确定各个特征的权重。整个实验的流程如图1所示:事先训练抓取的数据文本,然后对预处理的文本做相应的处理,接着计算每个字串的左右熵和其他特征的值,最后用逻辑回归的算法计算得到想要的结果。整个流程为了直观反映做了下边的流程图,实验流程图如图1所示。

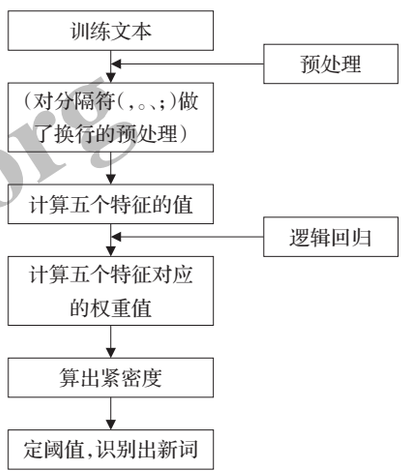

图1　实验流程图

本文抓取了300个页面进行实验,新词总数为195个,本文方法中识别出新词156个,正确的新词数有147个,准确率高达94.42%,召回率75.38%,检测出的新词有:球墨铸造、灰口铸铁、数控技术等词,对比中科院NLPIR汉语分词系统(又名ICTCLAS2013)(链接:http://ictclas.nlpir.org)把300个页面放入系统中,系统新词发现栏列举出的新词,识别出的新词80个,正确的新词75个,准确率达93.75%,召回率达41.02%,现做如下对比,如表4所示。

表4　对比表　　%

方法	准确率	召回率	F-measure
本文方法	94.42	75.38	83.89
汉语分词系统NLPIR结果	93.75	41.02	57.07

接下来分析实验结果数据,在如表5中,第一列列出了一些新词,第二列代表两个词之间的紧密度计算结果,中括号里面的值显示了左边词与右边词的紧密度。第三列列出用中科院NLPIR汉语分词系统中新词的识别结果,小括号表示分词结果的分隔符号。第四列列出了对比结果,显示为“都可以识别或者只有本文方法可以识别”。从实验结果中可以看到,对比中科院NLPIR汉语分词系统方法,本文方法可以很有效地识别领域内的新词。

5　结束语

本文提出了一个基于紧密度的命名实体识别方法。通过词法特征与统计特征的抽取,确定了紧密度的计算公式,从而较为准确地进行新词发现。实验结果表

表5 新词识别表格

新词	紧密度	中科院新词识别结果	结果对比
型壳	型[1.000 000]壳	型壳	都可以识别
熔模铸造	熔[0.729 412]模[0.729 412]铸造	熔模铸造	都可以识别
专家系统	专家[0.729 412]系统	专家系统	都可以识别
熔模	熔[0.733 333]模	熔模	都可以识别
模锻	模[0.776 471]锻	模锻	都可以识别
石状断口	石[0.776 471]状[0.776 471]断口	石状断口	都可以识别
高温合金	高温[0.729 412]合金	高温合金	都可以识别
护环	护[1.000 000]环	护环	都可以识别
灰口铸铁	灰口[0.917 647]铸铁	(灰口)(铸铁)	只有本文方法可以识别
螺旋压力机	螺旋[0.917 647]压力机	(螺旋)(压力机)	只有本文方法可以识别
固溶处理	固[0.917 647]溶[0.917 647]处理	(固溶)(处理)	只有本文方法可以识别
淮解理断口	淮[0.776 471]解理[0.776 471]断口	(淮)(解)(理)(断口)	只有本文方法可以识别
低温回火	低温[0.917 647]回火	(低温)(回火)	(低温)(回火)
火耗	火[1.000 000]耗	(火)(耗)	只有本文方法可以识别
精密锻件	精密[0.780 392]锻件	(精密)(锻件)	只有本文方法可以识别
冷隔	冷[0.917 647]隔	(冷)(隔)	只有本文方法可以识别
铸造技术	铸造[0.917 647]技术	(铸造)(技术)	只有本文方法可以识别
球墨铸铁	球墨[0.729 412]铸铁	(球)(墨)(铸铁)	只有本文方法可以识别
热分解	热[0.917 647]分解	(热)(分解)	只有本文方法可以识别
铝模	芯轴[0.776 471]拔[0.776 471]长	(铝)(模)	只有本文方法可以识别
芯轴拔长	芯轴[0.776 471]拔[0.776 471]长	(芯)(轴)(拔)(长)	只有本文方法可以识别
萘状端口	萘[0.776 471]状[0.776 471]断口	(萘)(状)(端口)	只有本文方法可以识别

明新的方法在机械制造领域取得了较好的应用结果,为进一步全网搜索打下了基础。

参考文献:

[1] 李飒.推动新型工业化与信息化相互融合研究[D].沈阳:辽宁大学,2014.

[2] Sproat R,Emerson T.The first international Chinese word segmentation bake off[EB/OL].[2013-03-10].http://acl.ldc.upenn.edu/W/W-03/W03-1719.pdf.

[3] Zhang K,Liu Q.Automatic recognition of Chinese unknown words based on roles tagging[C]//Proc of the 1st SIGHAN Workshop on Chinese Language Processing,2002:71-78.

[4] Chen K J,Ma W Y.Unknown word extraction for Chinese documents[C]//International Conference on Computational Linguistics,2002.

[5] 邹纲,刘洋,刘群,等.面向Internet的中文新词语检测[J].中文信息学报,2004(6):1-9.

[6] 杨绪明,杨文全.当代汉语新词新语探析[J].汉语学习,2009(1):97-104.

[7] Bheganan P,Nayak R,Xu Y.Thai word segmentation with hidden Markov model and decision tree[M]//Advances in knowledge discovery and data mining.Berlin/Heidelberg:Springer,2009:74-85.

[8] 黄轩,李熔烽.博客语料的新词发现方法[J].现代电子技术,2013,36(2):144-146.

[9] Li H Q,Huang C N,Gao J F,et al.The use of SVM for Chinese new word identification[C]//First International Joint Conference on Natural Language Processing,2004:723-732.

[10] 贺敏,龚才春,张华平,等.一种基于大规模语料的新词识别方法[J].计算机工程与应用,2007,43(21):157-159.

[11] 施水才,俞鸿魁,吕学强,等.基于大规模语料的新词语识别方法[J].山东大学学报:理学版,2006(3):89-91.

[12] 林自芳,蒋秀凤.基于词内部模式的新词识别[J].计算机与现代化,2011(11):162-164.

[13] 霍帅,张敏,刘奕群,等.基于微博内容的新词发现方法[J].模式识别与人工智能,2014(2):141-145.

[14] 张海军,彭成,栾静.基于外部排序的字串左右熵快速计算方法[J].计算机工程与应用,2011,47(19):18-20.

[15] Mccallum A,Li W.Early results for named entity recognition with conditional random fields,feature induction and web-enhanced lexicons[C]//Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL,2003:188-191.

[16] Chris M,Hinrich S.Foundations of statistical natural language processing[M].London,England:MIT Press,1999:171-173.