文章编号：1000-8608(2013)02-0267-06

基于条件随机场的汽车领域术语抽取

李丽双*[1]， 党延忠[2]， 张 婧[1]， 李 丹[1]

（1.大连理工大学 计算机科学与技术学院，辽宁 大连 116024；
2.大连理工大学 管理科学与工程学院，辽宁 大连 116024）

摘要： 中文领域术语抽取是中文信息处理领域的一项重要研究任务，在词典构建、领域本体构造等方面有重要的应用.采用条件随机场(conditional random fields，CRFs)，从汽车知识网站上爬取网页，预处理后得到纯文本，然后分析汽车领域的术语组成特点并制定相应的语料标注规则进行人工标注，对汽车领域进行了术语抽取.在使用词和词性特征的基础上增加了词典特征、领域词频和背景领域词频等特征，精确率、召回率和 F-值分别达到 84.61%、80.50%和 82.50%.与其他方法比较说明所提出的汽车领域术语抽取方法是有效的.

关键词： 信息抽取；领域术语抽取；汽车领域术语；条件随机场
中图分类号： TP391 **文献标志码：** A

0 引 言

术语是代表特定学科领域基本概念的语言单元，可以是词也可以是词组，在我国又称为名词或科技名词.术语抽取是信息处理领域中一项重要的研究任务，在词典编撰、领域本体构建[1]、机器翻译等领域都有重要的应用.

目前比较常用的术语抽取方法主要有三大类：一是基于规则的方法，主要是根据语言学及领域知识制定相应的规则模板，与规则模板匹配的视为术语，此方法受限于规则模板的质量，不够灵活.二是基于统计的方法，又分为基于统计量度和统计机器学习的方法.目前常用的统计量参数有频率、假设检验(t 检验、卡方检验等)、似然比、信息熵和互信息.文献[2]通过计算字串的互信息得到候选术语，最终取得 75%的 F-值.文献[3]提出一种基于质子串分解的算法，利用 C-value 和 F-MI参数来进行术语的抽取.由于没有大规模的标注语料，基于统计机器学习方法的中文领域术语抽取的研究不多，文献[4]和[5]基于条件随机场(CRFs)对科技术语和军事领域术语进行抽取，F-值分别达到 84.4%和 76.46%.文献[6]利用隐马尔可夫模型对计算机术语进行识别.文献[7]将语言学方法和统计方法进行一体化处理，同时考虑了词所在句子的术语度，利用 CRFs 进行计算机领域术语抽取，F-值为 79.64%.三是统计与规则相结合的方法，文献[8]首先利用语言学规则获取候选术语，再利用统计的方法进行过滤.文献[9]首先利用 C-value 和互信息获取候选术语，然后根据术语的词性规则和词典特征进行过滤，最终 F-值达到 42%.本文主要就汽车领域的术语抽取任务展开讨论，分析该领域术语的特点及抽取难点，利用目前较为流行的条件随机场(CRFs)模型，选取词、词性、词典及频率等特征进行汽车领域术语的抽取.

1 汽车领域的术语抽取

1.1 汽车领域术语

本文利用有监督的统计机器学习方法进行领域术语抽取，需要一定规模的带标签的训练语料.由于没有标注好的汽车领域标准语料，需要人工标注.目前缺少一个关于汽车领域术语的统一标准，本文对《汽车行业名词术语汇编》中和汽车零部件相关的 7 525 个术语进行了学习和分析，统

收稿日期： 2012-01-09； **修回日期：** 2013-01-15.
基金项目： 国家自然科学基金资助项目(71031002，61173101，61173100).
作者简介： 李丽双*(1967-)，女，副教授.E-mail：lilishuang314@163.com.

计得到单词型术语占9%,由两个单词组成的复杂术语占35%,三词术语占31%,四、五、六词术语分别占15%、6%、2%,七词及以上术语占2%,即复杂术语一般由2～4个单词组成,占全部术语的81%,符合中文术语的一般性特点.为了方便人工标注,本文分析了汽车领域术语的特点并借助前人对领域术语特点的研究成果,制定了一定的标注标准,凡是符合标注标准的词都被视为汽车领域的术语.标注标准如下:

(1)描述或表示汽车的词,一般是随着汽车领域的产生和发展而出现的,比如"轿车""两厢车"等,由于汽车领域外来词汇比较多,通常情况下人们会用外文直接描述,像类似于"SUV"(运动型多用途汽车)"RV"(休闲车)等英文单词或缩略词也归于汽车领域术语.

(2)表示汽车零部件或组成成分的词,如"底盘""后视镜",另外像"气门""活塞"等机械领域的词,虽然不是专属于汽车领域的,但也是描述汽车结构或功能所必需的,视为领域术语.

(3)与汽车相关的系统或结构,如"防抱死制动系统""高压共轨系统"等,相应的英文缩略词同样作为术语.

(4)一些词在通用领域也有应用,但是在汽车领域表示特定的含义,如"抬头""塌屁股"描述的是汽车的某种状态,可作为汽车术语.

(5)要遵循术语应尽可能详细和完整的原则,如类似"1.6升5缸发动机""四行程发动机缸内燃油直喷技术",要将其作为一个整体.

(6)描述汽车品牌及其型号的词语在本文中不作为领域术语,可单独作为一类词进行识别.

(7)文章中若出现英文缩写和中文译文联合使用的情况,按两个术语分别标注.如"ABS(防抱死制动系统)",标注为"ABS"和"防抱死制动系统"两个术语.

1.2 汽车领域术语抽取任务的特点

通过对汽车领域术语特点的分析可以看出领域术语在结构上比较复杂,所以与一般的命名实体识别相比,领域术语的自动抽取具有其特殊性,具体表现在:

(1)没有明确的关于领域术语的定义,不能清晰地界定术语的边界.目前已有的词典或是词表不足以涵盖全部的术语,而且随着技术的进步,新的产品或应用会不断增多,相应的术语表示也会不断丰富.比如"绿色汽车""零公里"是近几年提出的概念.

(2)由于汽车领域引入国外技术比较多,在表述时多采用音译词或是英文缩写,比如"皮卡"("pick-up"的音译)"RV"(休闲车),而且由于使用习惯等原因,在表述时使用的不同的名称代表同一事物,比如"皮卡"和"轿卡"就代表同一类型汽车,在使用时比较随意,没有特定的用法.

(3)汽车领域的术语模式多变,表现在长度、词性、组成模式等方面.例如,"悬架"和"综合电子控制动力转向系统"相差10个字长,还有类似于"可变预行程tics系统"和"D2T式制动器"的中英文混合术语.

(4)一般的命名实体(人名、地名或组织机构名等)通常会存在比较明显的特征词,上下文环境也相对规律,而就汽车领域术语而言很难找出比较统一的特点,而且中英文混用的现象明显.

(5)领域术语的一个公共特点就是存在嵌套(网状术语),比如"曲轴箱换气式二行程发动机",其中"曲轴箱""二行程发动机""发动机"本身又都分别作为术语出现.

2 基于CRFs的领域术语抽取

条件随机场是一种判别式图模型,由Lafferty等于2001年提出.CRFs同时具备最大熵模型(ME)和隐马尔可夫模型(HMM)的特点,不存在HMM那样严格的独立性假设,而且其采用的是全局归一化的方法,克服了最大熵马尔可夫模型的标记偏置问题,是目前处理序列化数据分割与标注问题最好的统计机器学习模型,在分词、命名实体识别等问题上已经得到广泛的应用.虽然领域术语和一般的命名实体在自身结构、所运用的环境等方面有很大的不同,但是就其识别任务而言也有一定的相似性,故本文将领域术语的识别任务转化为序列标注问题,利用CRFs进行汽车领域术语的识别.

汽车领域术语识别的基本流程是:

(1)获取语料,进行去噪、去重、分词和词性标注等一系列预处理.

(2)选取合适的特征,使用CRFs训练模型.

(3)在测试语料上用训练出来的模型进行识别.

(4)分析结果.

2.1 语料预处理

从网页上爬取一定规模的原始语料，去除HTML标签提取网页正文，获得纯文本.将获取的纯文本语料使用本实验室开发的分词工具对语料进行分词和词性标注处理.本文将术语识别任务转换为序列标注问题，采用目前比较流行的*BIO*短语组块标记方法来表示序列的标注结果，其中*B*表示术语的开始，即首词；*I*表示术语除首词以外的部分；*O*表示其他非术语词，如“鼓/*B* 式/*I* 制动器/*I* 一般/*O* 用于/*O* 后轮/*B*”.

2.2 特征选取

基于CRFs的术语抽取，选择合适的特征很关键.文献[4]使用词本身和词性作为特征，文献[5]选取了6个特征，即词本身、词性、左信息熵、右信息熵、互信息和*TF*/*IDF*.文献[7]将术语的统计信息融合到CRFs模型的特征中，并使用背景语料来强化词语的术语特性，即使用了词的频率、领域频率差、词频的*Rank*值，以及术语所在句子的信息.本文总结了前人的工作，并结合汽车领域术语的特点，选取了9个特征，分别介绍如下：

(1)词本身 *Word*

根据领域术语的特性可知，有些词只在本领域流通，故词本身包含了术语最大的信息，所以使用词本身作为特征.

(2)词性 *POS*

通过对已有的汽车术语资源分析可知虽然组成词性模式有很多种，但是大部分是名词性短语，统计得到前三位词性组合模式为“n+n”“v+n”“n”，可见词性对于术语的识别是一个重要特征.另外，汽车领域中一些术语由中英文搭配组成，用词性作为特征可以将此种情况考虑在内.

(3)词的长度 *WordLen*

领域术语中有一部分词是未登录词，通用的分词系统对于未登录词的处理办法通常是分成单个字，比如“排挡杆”被标记为“排/v　挡　/Ng　杆/Ng”，可以利用这个特性，通过考虑当前词的长度来判断其是否作为术语中的一部分.

(4)是否在已知词典中 *IsDic*

本文整理的词典中共7 525条术语，由3 109个词组成，可知一些词不止在一个术语中出现.由2.1的分析可知，复杂术语占80%以上，单词在复杂术语中出现的位置信息可以作为一项特征.经分析统计，词典中的3 109个词按在术语中的所处位置可分为以下6种情况：

(i)只作为单词性术语，如“外胎”，词典中不存在其出现在复杂术语中的情况.此类词共166个，占5.34%，记为OS；

(ii)可单独使用也可以作为复杂术语的一部分，占8.11%，记为DS；

(iii)只出现在复杂术语的开头，占14.09%，记为DB；

(iv)只作为复合词的结尾，占20.75%，记为*DE*；

(v)只出现在复合词的中间位置(针对由两个以上的词组成的术语)，占40.59%，记为*DI*；

(vi)只出现在复合词中，但其出现的位置不固定，占11.13%，记为*OD*.

根据以上分析，本文将词典特征分为7个值，分别为OS、DS、*DB*、*DE*、*DI*、*OD*、O，其中*O*为当前词不在词典中.

(5)当前词前后窗口大小范围内的词的词典特征 *WinDic*

文献[9]指出，一个候选术语，如果其前后窗口大小范围内的词中，已在词典中存在的词所占的比例大于一定阈值，则此候选术语也被视为术语.文献[10]分析得到一个领域通用词，如“是”，其周围的词通常是领域相关的.本文结合这两个特点，将上下文的词典特征分为3种类型：一是当前词窗口范围内的词在词典中出现的比例大于阈值且当前词也在背景语料中出现，其值为1；二是比例大于阈值，但是当前词不在背景语料中出现，值记为2；三是除去一、二外的情况，值记为3.

文献[7]将术语的统计信息融合到CRFs模型的特征中，并使用背景语料来强化词语的术语特性.本文借鉴文献[7]中采用的统计特征，在前文介绍的特征的基础上加入和频率有关的特征(6)～(9)：

(6)当前词在领域语料中的频率*DomainFreq*

记*C_word*为当前词在语料中出现的频次，*C*为语料中的总词数，则当前词的频率为

$$DomainFreq = C\_word/C$$

由于计算出的频率值是浮点数，不能直接用于CRFs的特征值，可以把浮点值按大小分为几类，本文按五类划分，即特征值取1到5.

(7) 当前词在背景语料中的频率 $ContrastFreq$

选用计算机语料作为背景领域语料,共 8 014 行,20 800 个词. 频率的计算方法和特征值的取值方法与汽车领域相同.

(8)当前词在两类语料中的频率差 $\Delta Freq$

(9)当前词所在句子中的所有词的语料频率差之和 $Sen\_\Delta Freq$

3 实 验

3.1 实验数据

使用 Heritrix 从"太平洋汽车网"的"汽车知识"版块爬取约 500 篇网页,去除 HTML 标签等噪音得到纯文本文档,进行去重处理,得到约 1 MB 的领域语料,共 529 651 字. 为了减少数据不平衡的影响,将语料分成 5 组,进行 5 倍交叉测试.

以第一组数据为例,测试语料中共 2 069 条术语(不包含重复),将分词后的组成成分的个数作为计算词长的标准,如"汽车发动机"分词后为"汽车/t 发动机",计其词长为 2. 经过分析可以看出本语料包含的术语在长度上基本符合一般领域术语的分布规律. 各长度所占比例如图 1 所示.

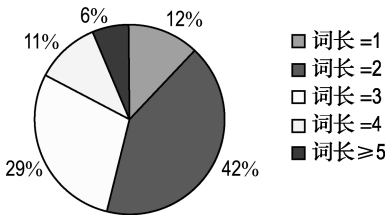

图 1 测试语料中各长度的术语所占比例

Fig. 1 The proportion of each length term in test corpus

3.2 实验结果

3.2.1 评价标准及结果 采用准确率(P)、召回率(R)以及 F-值作为评价指标(术语数包含重复个数),计算方法如下:

$$P = \frac{\text{正确识别的术语数}}{\text{识别的术语总数}} \times 100\%$$

$$R = \frac{\text{正确识别的术语数}}{\text{语料中的术语数}} \times 100\%$$

$$F\text{-值} = \frac{2 \times P \times R}{P + R} \times 100\%$$

本文采用了 9 个特征进行术语抽取,为了验证特征的有效性,将各组特征分别加入到特征集中,实验结果如表 1 所示,其中各组结果均为交叉测试得到的平均值.

表 1 不同特征的识别结果

Tab. 1 The results based on different features

所采用特征	P/%	R/%	F-值/%
$Word$, POS, $WordLen$	85.59	78.40	81.79
$Word$, POS, $WordLen$, $IsDic$, $WinDic$	85.18	79.63	82.31
$Word$, POS, $WordLen$, $IsDic$, $WinDic$, $DomainFreq$, $ContrastFreq$	84.61	80.50	82.50
$Word$, POS, $WordLen$, $IsDic$, $WinDic$, $DomainFreq$, $ContrastFreq$, $\Delta Freq$	84.36	80.56	82.41
$Word$, POS, $WordLen$, $IsDic$, $WinDic$, $DomainFreq$, $ContrastFreq$, $\Delta Freq$, $Sen\_\Delta Freq$	84.34	80.63	82.44

由表 1 可以看出,使用词本身、词性、词长时正确率最高,加入词典特征后正确率有所降低,召回率提高. 加入词典特征正确率反而降低,分析原因可能是有些字在有些词中属于术语的一部分,而在有些词中则不是,比如词典中的"差速器"分词后为"差/速/器",而在语料中,"差"这个字多用在"之/差"、"较/差"等词中,从而干扰了正确率. 在前 6 个特征的基础上加入词在领域语料和背景语料的频率特征后召回率增加,正确率略有降低,F-值达到 82.50%. 加入词所在句子的频率特征后召回率达到最高的 80.63%,但同时也导致正确率降低,F-值略有降低.

3.2.2 不同长度的词的识别结果 统计各个长度术语的识别的情况,结果见表 2.

表 2 各个长度的术语的识别情况

Tab. 2 The identification results of different lengths of terms

词长	识别正确百分比/%	词长	识别正确百分比/%
1	78.6	4	75.2
2	72.7	≥5	66.4
3	74.0		

其中百分比是指各长度正确识别的术语(不包含重复词)占测试语料中该长度的术语数的比

例.从表中可以看出,简单术语识别效果最好,5 词以上复杂术语的识别效果最差.

3.3　识别结果分析

以第一组为例分析实验结果,发现错误主要集中在以下几个方面:

(1)识别词语不全,如“多连杆悬架横梁”识别成了“连杆悬架横梁”,“双重防震悬架横梁”识别成了“悬架横梁”.

(2)由于分词错误导致的错误,如“定钳式盘式制动器”被识别成“下定钳式盘式制动器”,因为分词的结果是“装/v 下定/v 钳/Ng 式/k 盘/qr 式/k 制动器/n”,CRFs 模型共识别出 1 437 个术语(不包含重复),其中错误的占 323 个,有 17 个词是因为分词错误导致的.

(3)识别出的词比正确的术语多出一部分,除去因为分词错误的情况外,还有比如“车载 gps”识别成“车载 gps 价格”,“减速器”识别成“带有减速器”的情况.

(4)由于没有统一的标准,在标注上有一些歧义,比如根据标注规则,“3.2 升 fsi 发动机”被判定为一个术语,但是识别结果是“fsi 发动机”,类似的还有“车蜡”被识别成“高档车蜡”,这类词不能断定其错误,和术语判定标准有关.

(5)一些词不被认为是汽车领域的术语,但因其自身特点或其所处上下文环境和术语类似也可能被识别出来,比如“激光”“超声波”等.

(6)由于人工标注上不可避免的错误导致识别结果不正确.

由表 2 可知单词型术语识别效果最好,长术语较差.其中,单词型术语中诸如“SUV”“RV”等英文缩写词识别效果较差,分析原因可能是由于这类词所处的语言环境相对不固定,再加上语料稀疏.长术语识别效果较差可能是由于出现频次少,组成词串的各个词之间的联系不紧密.

4　与其他方法的比较

文献[7]用语言学和统计相结合的方法从计算机科学领域论文中抽取计算机术语,将语料的语言学特征和统计学特征综合起来作为 CRFs 训练的特征,进行术语抽取,其在计算机科学领域中抽取计算机术语的最高 *F*-值达到 79.64%.其采用的特征分别为词本身、词性、当前词在两类语料中的词频差 $\Delta Freq$ 和当前词所在句子中的所有词的语料频率差之和 $Sen\_\Delta Freq$.本文将文献[7]的方法在选用的语料上进行了实验.选取的特征与文献[7]相同,进行 5 倍交叉验证,实验结果如表 3 所示.从表 3 可以看出,本文的方法比采用文献[7]的模型其 *F*-值高 2.11%.实验结果表明,对汽车领域,本文通过选取有效的特征,建立了有效的术语抽取模型.

表 3　与文献[7]的比较

Tab.3　Comparison with Lit.[7]

	P/%	*R*/%	*F*-值/%
文献[7]	84.24	76.89	80.39
本文	84.61	80.50	82.50

基于 CRFs 的方法必须以标注语料为基础,人工标注语料费时费力,因此研究初期本文也采用了基于统计量的无监督方法在语料上进行了实验.将术语抽取分为候选术语抽取和术语确定两步,文本预处理上采用 Pat-tree 结构.Pat-tree 采用半无限长字符串,是一种压缩的二叉查询树,可以快速地得到任意长度的字符串及其在文本中出现的频次.候选术语利用计算词串内部关联度[9](SEF)和外部关联度[9](C-value)获取.术语的确定利用候选术语和候选词邻接词的词性信息.本文总结了正确术语的词性组合规律,从 3 000 行已标注语料统计术语前后的词性搭配情况,构建词性规则库,把不在规则库中的候选词串过滤,剩下的则是最终正确的汽车领域术语.用该方法在同样语料上测试的 *F*-值为 15.41%,要远低于基于 CRFs 的有监督的机器学习方法.这主要是由于语料的规模和数据稀疏问题导致统计信息不足,在很大程度上影响了无监督的统计方法在汽车领域语料上的术语抽取效果.

5　结　语

本文主要针对汽车领域进行术语抽取,将其转化为序列标注问题,使用 CRFs 模型将词、词性、词典、领域频率等多个有效特征整合,采用交叉验证的方法,最终的 *F*-值达到 82.50%,由于 CRFs 模型融合了多种有效特征,在汽车领域术语的抽取实验中取得了较好的效果.

参考文献:

[1]　温　春,王晓斌,石昭祥. 中文领域本体学习中术语的自动抽取[J]. 计算机应用研究, 2009, **26**(7): 2652-2655.

WEN Chun, WANG Xiao-bin, SHI Zhao-xiang. Automatic domain-specific term extraction in Chinese domain ontology learning [J]. **Application Research of Computers**, 2009, **26**(7): 2652-2655. (in Chinese)

[2] 张 锋,许 云,侯 艳,等. 基于互信息的中文术语抽取系统[J]. 计算机应用研究, 2005, **22**(5):72-73.
ZHANG Feng, XU Yun, HOU Yan, *et al*. Chinese term extraction system based on mutual information [J]. **Application Research of Computers**, 2005, **22**(5):72-73. (in Chinese)

[3] 何婷婷,张 勇. 基于质子串分解的中文术语自动抽取[J]. 计算机工程, 2006, **32**(23):188-189.
HE Ting-ting, ZHANG Yong. Automatic Chinese term extraction based on decomposition of prime string [J]. **Computer Engineering**, 2006, **32**(23): 188-189. (in Chinese)

[4] 刘 豹,张桂平,蔡东风. 基于统计和规则相结合的科技术语自动抽取研究[J]. 计算机工程与应用, 2008, **44**(23):147-150.
LIU Bao, ZHANG Gui-ping, CAI Dong-feng. Technical term automatic extraction research based on statistics and rules [J]. **Computer Engineering and Applications**, 2008, **44**(23): 147-150. (in Chinese)

[5] ZHENG D Q, ZHAO T J, YANG J. Research on domain term extraction based on conditional random fields [C] // **ICCPOL 2009, LNAI 5459**. Berlin: Springer-Verlag, 2009:290-296.

[6] 岑咏华,韩 哲,季培培. 基于隐马尔科夫模型的中文术语识别研究[J]. 现代图书情报技术, 2008(12):54-58.
CHEN Yong-hua, HAN Zhe, JI Pei-pei. Chinese term recognition based on hidden Markov model [J]. **New Technology of Library and Information Service**, 2008(12):54-58. (in Chinese)

[7] 章承志. 基于多层术语度的一体化术语抽取研究[J]. 情报学报, 2011, **28**(3):275-285.
ZHANG Cheng-zhi. Using integration strategy and multi-level termhood to extract terminology [J]. **Journal of the China Society for Scientific and Technical Information**, 2011, **28**(3):275-285. (in Chinese)

[8] 周 浪,史树敏,冯 冲,等. 基于多策略融合的中文术语抽取方法[J]. 情报学报, 2010, **29**(3):460-467.
ZHOU Lang, SHI Shu-min, FENG Chong, *et al*. A Chinese term extraction system based on multi-strategies integration [J]. **Journal of the China Society for Scientific and Technical Information**, 2010, **29**(3):460-467. (in Chinese)

[9] JI L, SUM M, LU Q, *et al*. Chinese terminology extraction using window-based contextual information [C] // **CICLing 2007, LNCS 4394**. Berlin: Springer-Verlag, 2007:62-74.

[10] YANG Y H, LU Q, ZHAO T J. Chinese term extraction using minimal resources [C] // **Proceedings of the 22nd International Conference on Computational Linguistics**. Manchester: [s n], 2008:1033-1040.

Automotive term extraction based on conditional random fields

LI Li-shuang*[1], DANG Yan-zhong[2], ZHANG Jing[1], LI Dan[1]

(1. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China;
2. School of Management Science and Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: Chinese domain term extraction is an important task in Chinese information processing, which has been applied to the construction of lexicography and ontology and so on. Term extraction based on CRFs (conditional random fields) in automotive field is discussed. Firstly, plain text is extracted from crawled web pages relating to automotive knowledge with preprocessing. Then, corpus is labeled manually with corresponding rules written by analyzing the characteristics of automotive terms. Therefore, domain corpus for term extraction is constructed. The features of dictionary, word frequencies in the domain and other domain corpora are used besides the features of word and part-of-speech. Experimental results show that the precision, recall and *F*-score are 84.61%, 80.50% and 82.50% respectively. The comparison with other methods illustrates that the established model for extracting automotive term is effective.

Key words: information extraction; domain term extraction; automotive term; conditional random fields