新能源汽车领域中文术语抽取方法*

何　宇[1]　吕学强[1]　徐丽萍[2]
[1](北京信息科技大学网络文化与数字传播北京市重点实验室　北京 100101)
[2](北京城市系统工程研究中心　北京 100089)

摘要: 【目的】为提高新能源汽车领域中文术语抽取结果的准确率和召回率，提出一种适合该领域的术语抽取方法。【方法】在总结前人工作基础上，提出利用条件随机场模型作为抽取模型，选取词、词长、词性、依存关系、词典位置、停用词等特征作为特征模板。【结果】实验结果正确率为 93.12%，召回率为 90.47%。正确率比 Baseline 方法提高 7.73%。【局限】该方法只提高较短术语抽取结果的正确率。【结论】依存关系作为条件随机场模型的一项特征可以提高新能源汽车领域中文术语抽取结果的正确率和召回率。

关键词: 术语抽取　新能源汽车领域　条件随机场　依存句法关系
分类号: TP391.41

1　引　言

专利文献是指各专利管理机构在受理、审批、注册专利过程中产生的记述发明创造技术及权利等内容的官方文件及其出版物的总称[1]。有效地利用专利文献可以节省开发时间和研究经费。专利文献检索是快速有效利用专利文献的重要方法，并且专利文献中的术语又是文献检索的一个组成部分，因此术语抽取问题越来越受到相关研究者的重视。

目前，国内外相关学者对特定领域的术语抽取做了大量研究，主要有基于语言学规则的方法、基于统计的方法和两者相结合的方法。周浪等[2]根据术语的构词规律提出构词法，并根据构词法识别候选术语。基于语言学规则的方法主要缺点是识别的结果受制于规则模板的质量，不能灵活适应语料的变化。基于统计的方法分为基于统计量的方法和基于机器学习的方法。基于统计量的方法中，参数有频率(TF-IDF、词语在背景语料中的频率等)、似然比、信息熵、互信息等。梁颖红等[3]将 C-value 和互信息相结合构造了 C-MI 参数，通过与单独使用互信息和 C 值的方法比较，证明该方法提高了长术语识别的正确率。屈鹏等[4]对候选术语计算卡方检验、互信息、TF-IDF 值，然后根据特征值对候选术语进行排序，同时还提出生僻术语的识别算法，但在术语定义时限定术语长度为 2-3 词，术语覆盖范围不够大。董丽丽等[5]先使用停用词表去掉语料中的停用词，再利用互信息获取合成词串作为候选术语，最后利用似然比获取低频术语加入候选集合，但该方法的准确率和召回率受语料规模的影响较大。基于机器学习的方法，主要采用条件随机场模型进行术语抽取: 郭剑毅等[6]使用低层条件随机场模型以字为单位识别旅游景点名、特产小吃名等，然后使用高层条件随机场模型识别嵌套实体。该方法的识别结果优于仅使用单层条件随机场模型识别的结果，不足之处在于没有考虑语义信息，并且层叠条件随机场模型识别效率低于单层条件随机场模型。

通讯作者: 何宇，ORICD: 0002-8314-5525，E-mail: solocode@sina.com。

*本文系国家自然科学基金项目“基于本体的专利自动标引研究”(项目编号: 61271304)、北京市教委科技发展计划重点项目暨北京市自然科学基金 B 类重点项目“面向领域的互联网多模态信息精准搜索方法研究”(项目编号: KZ201311232037)和北京市科学技术研究院科技创新工程项目“基于 CGE-TIMES 模型的交通对大气环境综合影响评价方法研究”(项目编号: PXM2015_178215_000008)的研究成果之一。

施水才等[7]使用条件随机场模型以词和词性的组合为特征模板对专利文献中的术语进行抽取，虽然在实验中获得了较好的结果，但是缺乏对术语上下文信息的充分利用和对术语内部词语间的依存关系的考虑。将语言学规则和统计特征相结合的方法有：章成志[8]首先使用条件随机场模型抽取候选术语，然后使用规则模板对候选术语进行校正,该方法可以有效提高未登录词识别结果的召回率，但规则模板的制定依赖于语料，扩展不灵活；唐涛等[9]将语言学方法和统计方法相融合，综合考虑候选术语及其所在语句的术语度，将术语度作为条件随机场模型的一项特征进行术语抽取，但是对于边缘概率较高的候选术语不能有效识别。

针对新能源汽车领域的术语特点，在总结前人术语抽取研究成果的基础上，本文利用成熟的条件随机场模型，选取词、词长、词性、依存关系、词典位置、停用词等特征对新能源汽车领域中文术语进行抽取，探索了将依存句法关系应用到术语抽取问题即将其作为条件随机场模型的一项特征。实验结果证明这项特征可以提高术语抽取结果的正确率和召回率。

2　新能源汽车领域术语的特点

因国家标准中的术语较少，更新慢，所以本文在参考《GB/T 19596-2004 电动汽车术语》、《GB/T 24548-2009 燃料电池电动汽车术语》、《GB/T 28382-2012 纯电动乘用车技术条件》和《GB/T 20042.1 质子交换膜燃料电池术语》等国家标准的基础上，人工从专利文献中提取新能源汽车领域相关的9 644个术语进行学习和分析。

(1) 截至目前虽然国家颁布了新能源汽车领域的相关国家标准，但是标准中的术语少，术语的边界定义标准模糊，更新慢。

(2) 新能源汽车领域的术语组合方式多变，比如词长从两个字的“电机”到 10 个字的“电动汽车整车整备质量”，如“DC/DC 变换器”中英文混合术语和作者自创的术语。

(3) 领域术语的一个公共特点是存在嵌套(网状术语)，比如“燃料电池电动汽车”，其中“燃料电池”、“电动汽车”本身又分别作为术语出现。

(4) 专利文献作为具有法律效力的文本在写作时具有一些固定的特点，表现在为突出专利技术，专利文献中会使用一些常用的写作词语。例如“涉及”、“一种”这两个词后一般都会出现术语。

(5) 专利文献中的术语一般由名词、形容词、动词组成。而且结尾为名词的术语占 86.47%，结尾为动词 v 和语素 g 共占 9.45%，其他占 4.08%。术语的组成情况如表1所示，可以看出复杂术语一般由2-4个词组成，而且复杂术语的数量占术语总数的 83.43%。

表 1　术语的组成情况

单词数	比例
1 个词	4.53%
2 个词	29.32%
3 个词	35.28%
4 个词	18.83%
5 个词以上	12.04%

(6) 专利文献在写作时所用的语句，从词语间的依存关系角度考虑也具有一定的规律。从表 2 可以看出术语的首词和中心词之间的关系基本上为定中关系、主谓关系和动宾关系。从表 3 可以看出术语内部词与词之间的依存关系为定中关系、动宾关系、主谓关系、介宾关系和并列关系。

表 2　术语首词和中心词之间的关系

依存关系	比例
定中关系	46.89%
主谓关系	14.08%
动宾关系	18.43%
介宾关系	8.45%
并列关系	6.23%
其　　他	5.92%

表 3　两个词组成的术语中词与词之间的依存关系

依存关系	比例
定中关系	33.07%
动宾关系	29.40%
主谓关系	15.41%
介宾关系	8.60%
并列关系	7.19%
其　　他	6.33%

3　基于条件随机场的新能源汽车领域术语抽取

本文在前人研究成果基础上，探索将句法依存关

系用于提高术语识别结果的正确率。

3.1 条件随机场

条件随机场(Conditional Random Fields, CRFs)是一种统计模型。最早由 Lafferty 等[10]于 2001 年提出。CRFs 提出的时间晚于最大熵模型(ME)、隐马尔可夫模型(HMM)。在借鉴两个模型优点的基础上, CRFs 克服了隐马尔可夫模型强独立性假设条件的约束, 具有容纳上下文信息的能力, 而且其采用的全局归一化方法, 有效克服了最大熵马尔可夫模型的标记偏置问题。CRFs 是目前能够有效解决序列化数据分割与标注问题最好的机器学习模型之一, 在分词和命名实体识别等问题上已经得到了广泛的应用。

3.2 术语抽取模型

术语抽取任务是指当给定输入集合 $X=\{x_1,x_2,\cdots,x_n\}$, 经过模型处理可得到输出集合 $T=\{T_1,T_2,\cdots,T_m\}$, 集合 X 可数, 集合 T 是有穷集合且它的元素为抽取的术语。术语 T_i 是词的集合 $T_i=\{x_i,x_{i+1},\cdots,x_j\}(i=1,2,\cdots,n,j=1,2,\cdots,n且i\leq j)$。

序列标注是指对给定的可数输入序列 $X_1,X_2,X_3\cdots X_n$, 经过模型处理得到有穷输出序列 $y_1,y_2,\cdots,y_n$。其中, y_i 是对应于 x_i 的标记。

本文采用 BIO 组块表达法来标记术语。其中 B 表示术语的开头, I 表示术语的其他部分, O 表示不相关部分, 即通过上文的定义将术语抽取问题转化为基于 CRFs 的序列标注问题。例如: 根据序列标注结果“与/O 质子/B 交换/I 膜/I 结合/O”就可以得到术语为“质子交换膜”。

本模型的处理步骤如下:

(1) 获取专利文献, 对文献进行去噪、去重、分词、词性标注和依存句法分析等预处理操作;

(2) 选取合适的特征, 利用 CRFs 训练模型;

(3) 利用测试语料测试模型, 修正参数;

(4) 分析实验结果。

3.3 语言云

语言云(语言技术平台云 LTP-Cloud)是由哈尔滨工业大学社会计算与信息检索研究中心研发的云端自然语言处理服务平台。后端依托于语言技术平台, 语言云为用户提供了包括分词、词性标注、依存句法分析、命名实体识别、语义角色标注在内的丰富高效的自然语言处理服务[11]。

依存句法(Dependency Parsing, DP)通过分析语言单位内成分之间的依存关系揭示其句法结构[11]。简而言之, 就是可以通过句法分析识别出句子中的谓语、宾语、定语等成分, 通过这些依存关系即可清晰地看出词语与词语之间的关系。语言云平台提供了 14 种依存关系, 如表 4 所示:

表 4 依存关系

关系类型	标注	关系类型	标注
主谓关系	SBV	动补结构	CMP
动宾关系	VOB	并列关系	COO
间宾关系	IOB	介宾关系	POB
前置宾语	FOB	左附加关系	LAD
兼　　语	DBL	右附加关系	RAD
定中关系	ATT	独立结构	IS
状中结构	ADV	核心关系	HED

3.4 特征选取

特征的选取对 CRFs 模型非常关键, 特征选择的好坏将直接影响系统的性能。针对任务的不同应该选择具有代表性的特征。为了最大限度地利用已经发现的语言学规律, 可以通过不同的特征模板来筛选特征。模型中的模板可以把特定位置的上下文信息、词、词性、依存句法关系等信息综合考虑。在避免特征的碎片化的同时还可以最大程度地集成知识源, 这样可以将尽可能多的知识统一在特征模板中, 通过不同特征的组合达到识别效果的最大化。本文首先使用当前词、词性、词长、词典位置这 4 个特征, 然后通过分析新能源领域专业术语的统计特征和语言学特征提出停用词和依存句法关系这两个新特征。为了使用这些特征, 窗口大小设定为两个词, 即向上向下滑动两个词的距离。经过多次实验最终确定本文所使用的特征模板。

CRFs 模型的特征一般分为三类: 原子特征、复合特征以及全局变量特征。原子特征是单独的一项特征, 如 Word。复合特征为两项或两项以上特征复合, 如 $Word_i/POS_i$ 就是复合特征, 表示当前词本身和它的词性的组合。本文没有使用 CRFs 的全局变量特征。

下面介绍本文用到的 6 个特征:

(1) 词本身 Word

根据领域术语的统计信息可知, 有些词只出现在该领域内。所以, 词本身包含了候选词是否作为领域

术语的很多信息，因此使用词本身作为特征。例如“极片”这个词一般只在作为领域术语时出现。

(2) 词性 POS

分析新能源汽车领域中文术语的词性组合情况后发现，术语主要由名词短语构成。从表 5 可以看出，虽然构成术语的词性组合模式很多，但是术语词性变化不大，词性主要为名词、动名词这两种，共占 47.16%。因此，可以通过词性过滤掉很多标注为虚词或其他词性的词语和一些词性组合明显不是术语的词。

表 5　术语词性序列

词性序列	比例
n n n n	2.07%
n n vn n	2.29%
n n vn n	2.29%
n	4.33%
n n n	7.73%
n vn n	8.63%
n n	19.82%
其他	52.84%

(3) 词典位置 DicPos

本文用到的术语词典中有 9 644 个术语，其中有 2 305 个词。由上述数据可知，一些词不仅出现在一个术语中。根据统计复杂术语占 83.57%，单词在词典中的出现位置可以作为一项特征[12]。经统计，单词在词典中的出现位置有以下 7 种情况，如表 6 所示:

表 6　词典位置比例与取值[12]

词典位置	比例	特征值
单词型术语	2.35%	OS
术语首词	18.75%	DB
术语中部	3.95%	DI
术语尾词	14.94%	DE
位置不固定	43.03%	OD
构成复杂术语和单词型术语	16.98%	DS
非词典词	0%	O

根据以上分析，单词在词典中的位置可以记为“OS、DS、DB、DI、DE、OD、O”，其中 O 表示当前词不在词典中。

(4) 依存句法分析 Rel

术语虽然结构复杂，但是通过统计分析发现，术语内部词语与词语之间的依存关系存在一定的规律。从表 3 可以看出词语之间的依存关系主要是定中关系、介宾关系、主谓关系、动宾关系。通过过滤不可能组成术语的依存关系，可以提高识别术语的效果。术语词与词之间的关系可以使用依存句法关系的标注来表示。例如“一/ATT 种/ATT 用于/ATT 混合/VOB 动力/ATT 汽车/VOB”其中术语为“混合动力汽车”。“混合/VOB”表示“混合”是“用于”的宾语，那么“用于”就不是混合的一部分。“动力/ATT”表示“动力”是“汽车”的定语，“汽车/VOB”表示“混合”的宾语。

(5) 词长 Length

由于术语中有很多词是未登录词，所以分词工具会将这类术语切分为单个字。比如，“串/v 励/g 直流电/n 机/g”。因此，可以通过当前词的长度判断当前词是否为术语的组成部分。

(6) 是否为停用词 isStop

由于 CRFs 标注过程中会出现错误地扩大术语的前后边界，通过统计术语前后边界出现的错误信息，可以提取停用词表。通过停用词可以提高抽取的正确率。如果当前词是停用词记为 1，否则记为 0。停用词的提取规则为:

①术语前出现的数量词、介词，如“一个、一种、和”等。

②术语后出现的部分语气词、动词、数词、介词、方位词，如“了、输出、一个、在、左边”等。

4　实　验

4.1　实验数据

实验语料是某专利公司的 415 篇专利文献共计 25MB，其中 20MB(343 篇)作为训练语料，5MB(72 篇)作为测试语料。将 PDF 格式的专利文献转换为纯文本文件，并去除转换过程中产生的乱码，生成的文本内容作为语料。本文将语料按照比例分配后，使用 ICTCLAS 分词工具对语料进行分词和词性标注处理，再使用语言云平台提供的依存句法分析服务对已经分词的语料进行依存句法分析，抽取得到语料中词与词之间的依存关系，将依存关系标注作为 CRFs 训练的一项特征。最后，按照 CRFs 模型所要求的格式加入各项特征。训练语料中有术语 2 415 个(不重复)，重复术语共计 118 517 个；测试语料中有术语 1 074 个(不重复)，重复术语共计 22 796 个。训练语料中的术语长度的分布情况如表 7 所示。

表 7　训练语料中各长度的术语所占比例

词长	百分比
1	16.85%
2	46.46%
3	24.72%
4	8.41%
其他	3.56%

为了方便标注，以手工提取的 9 644 个术语为基础，自动标注专利文献中的相同词语。然后，手动校对该词语在上下文信息中表示的是否是与新能源汽车相关的意义，如果不相关就手工修改该标注。

标注方法如下:

(1) 由于新能源汽车领域的术语难以避免会出现一些汽车、化学等其他领域中的常见术语，因此将部分与新能源汽车领域相关的术语也定义为新能源汽车中的常用术语。如“坡道起步能力”、“电催化剂”等术语。

(2) 语料中有一部分术语是表达一些系统或结构，这些术语是与新能源汽车相关的系统、结构和对应的英文缩写。所以，这些术语也被视为新能源汽车领域的术语。如“燃料电池发电系统”、“膜电极组件”和“MEA”。

(3) 术语是领域知识的总结和概括，所以标注的术语应遵循尽量完整和详细的原则。如“电源输出的动态响应特性”、“最高车速(1km)”。

(4) 描述汽车品牌和型号的词不作为领域术语。

(5) 专利文献中如果存在中英文术语同时出现的情况，视为两个术语进行识别。如“质子交换膜(PEM)”标注为两个术语，分别为“质子交换膜”和“PEM”。

4.2　实验结果和分析

采用正确率(P)、召回率(R)、以及 F-Value 作为评价指标(术语数包含重复的个数)，计算方法如下:

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

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

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

使用 6 个特征进行领域术语抽取，为了验证特征的有效性，将各组特征分别加入到特征模板中进行测试，选取实验结果最好的组合作为最终的特征模板。实验结果如表 8 所示。

从表 8 可以看出，特征组合 5 使用当前词、词性、词长、停用词时正确率最高，达到 93.30%；特征组合 3 使用当前词、词性、词长、依存句法关系时 F-Value 最高，达到 91.78%，正确率略有降低。当加入停用词特征时召回率略有下降，分析原因是停用词造成识别结果中长术语数量减少，短术语正确数量增多进而使得召回率下降。分析实验数据可以看出，特征组合 3 在加入依存句法关系特征后的正确率比特征组合 2 的正确率提高了 2.24%。

表 8　不同特征组合的实验结果

编号	特征	P	R	F-Value
1	Word POS	85.39%	83.53%	84.45%
2	Word POS Length	90.88%	88.59%	89.72%
3	Word POS Length Rel	93.12%	90.47%	91.78%
4	Word POS Length Rel DicPos	92.99%	89.66%	91.29%
5	Word POS Length isStop	93.30%	87.95%	90.55%
6	Word POS Length Rel DicPos isStop	91.47%	88.88%	90.16%

表 9 中的百分比是指在表 8 采用特征组合 3 的特征模板情况下正确识别的术语(包含重复术语)占测试语料中该长度术语数的百分比。从表 9 可以看出简单术语的识别结果最好，其他(5 个以及 5 个词以上的术语)术语的识别结果最差。分析其原因可以发现，模型中的原子模板对术语识别的作用最好，复合模板的组合情况仍需要继续研究。

表 9　不同长度术语的识别结果

词长	P
1	87.53%
2	83.02%
3	71.31%
4	59.83%
其他	19.39%

在选取依存句法关系特征时，主要考虑的是词与词之间的依赖关系，从依赖关系入手，分析术语的首词和中心词之间的关系，术语内部词汇之间的关系，经过统计分析发现术语内部词汇之间的关系具有统计规律。在实验中通过尝试原子模板和复合模板，找到一种利用依赖关系的方式即将当前词、词性、当前词的依存句法关系标注三者复合在一起的模板对术语的抽取效果最好。经过实验证明依存句法关系可以作为一种特征来提高术语识别结果的正确率和召回率。

分析实验结果发现识别的错误举例:

(1) 由于分词错误导致的识别错误

由于术语“双极片”应该是一个词，但是分词工具有时将它分为两个词“双/m 极片/n”，有时分为三个词“双/q 极/d 片/g”，导致识别错误。

(2) 识别术语不全

例如术语“二次锂离子电池”，没有识别出“二次”；“负极片”只识别出了“负极”。术语的识别应该尽量详尽，所以上述识别的术语是错误的。

(3) 错误识别其他领域的术语

例如“纳米”是其他领域的术语，但不是新能源汽车领域的术语。

文献[7]在统计领域术语的词性组合概率基础上使用当前词和词性为特征模板识别领域术语，正确率是 89.03%, F-Value 是 88.759%。本文将文献[7]的方法作为 BaseLine 在本实验语料上进行实验。选取的特征与文献[7]相同，实验结果如表 10 所示:

表 10　本文特征组合 3 的实验结果与 BaseLine 的比较

指标	P	R	F-Value
BaseLine	85.39%	83.53%	84.45%
特征组合 3	93.12%	90.47%	91.78%

实验结果表明，在新能源汽车领域，通过选取有效的特征，建立了有效的术语抽取模型。

5 结　语

本文对新能源汽车领域的术语识别问题进行研究，选取当前词、词长、词性、依存句法关系、术语在词典中的位置和是否为停用词等多个有效特征作为特征模板, F-Value 最高达到 91.78%。本文的创新之处是从句法分析的角度考虑术语之间的依存关系并把它作为 CRFs 的一项特征加入到特征模板中，通过实验证明该特征提高了术语识别结果的正确率和召回率。通过对比实验可以看出，本文方法的 F-Value 比对比方法提高了 7.33%，明显优于对比方法。

本文提出的基于 CRFs 的新能源领域术语抽取的方法仍然有改进的空间。由于目前的方法识别出的结果中存在术语不完整的现象，因此需要进一步研究利用依存句法关系提高识别术语完整度的方法。此外，在未来的研究工作中，需要充分分析特征间的关系，进一步提高术语识别的效果。

参考文献：

[1] 国家知识产权局专利局专利文献部. 专利文献与信息检索[M]. 北京：知识产权出版社，2013. (The Patent Documentation Department of SIPO. Patent Documents and Information Retrieval [M]. Beijing: Intellectual Property Publishing House Co., Ltd., 2013.)

[2] 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 Shumin, 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.)

[3] 梁颖红，张文静，张有承. C 值和互信息相结合的术语抽取[J]. 计算机应用与软件，2010, 27(4): 108-110. (Liang Yinghong, Zhang Wenjing, Zhang Youcheng. Term Recognition Based on Integration of C-Value and Mutual Information [J]. Computer Applications and Software, 2010, 27(4): 108-110.)

[4] 屈鹏，王惠临. 面向信息分析的专利术语抽取研究[J]. 图书情报工作, 2013, 57(1): 130-135. (Qu Peng, Wang Huilin. Patent Term Extraction for Information Analysis [J]. Library and Information Service, 2013, 57(1): 130-135.)

[5] 董丽丽，李欢，张翔，等. 一种中文领域概念词自动提取方法研究[J]. 计算机工程与应用，2014, 50(6): 127-131. (Dong Lili, Li Huan, Zhang Xiang, et al. Method for Automatic Extraction of Chinese Domain Concepts [J]. Computer Engineering and Applications, 2014, 50(6): 127-131.)

[6] 郭剑毅，薛征山，余正涛，等. 基于层叠条件随机场的旅游领域命名实体识别[J]. 中文信息学报，2009, 23(5): 47-52. (Guo Jianyi, Xue Zhengshan, Yu Zhengtao, et al. Named Entity Recognition for the Tourism Domain Based on Cascaded Conditional Random Fields [J]. Journal of Chinese Information Processing, 2009, 23(5): 47-52.)

[7] 施水才，王锴，韩艳铧，等. 基于条件随机场的领域术语识别研究[J]. 计算机工程与应用，2013, 49(10): 147-149. (Shi Shuicai, Wang Kai, Han Yanhua, et al. Terminology Recognition Based on Conditional Random Fields [J]. Computer Engineering and Applications, 2013, 49(10): 147-149.)

[8] 章成志. 基于多层术语度的一体化术语抽取研究[J]. 情报学报，2011, 30(3): 275-285. (Zhang Chengzhi. Using Integration Strategy and Multi-level Termhood to Extract Terminology [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(3): 275-285.)

[9] 唐涛，周俏丽，张桂平．统计与规则相结合的术语抽取[J]. 沈阳航空航天大学学报，2011，28(5)：71-74. (Tang Tao, Zhou Qiaoli, Zhang Guiping. Term Extraction Based on the Combination of Statistics and Rules [J]. Journal of Shenyang Aerospace University, 2011, 28(5): 71-74.)

[10] Lafferty J D, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]. In: Proceedings of the 18th International Conference on Machine Learning (ICML'01). San Francisco: Morgan Kaufmann Publishers Inc., 2001: 282-289.

[11] 语言云[EB/OL]. [2014-08-25]. http://www.ltp-cloud.com/. (Language Technology Platform Cloud [EB/OL]. [2014-08-25]. http://www.ltp-cloud.com/.)

[12] 李丽双，党延忠，张婧，等．基于条件随机场的汽车领域术语抽取[J]. 大连理工大学学报，2013，53(2)：267-272. (Li Lishuang, Dang Yanzhong, Zhang Jing, et al. Automotive Term Extraction Based on Conditional Random Fields [J]. Journal of Dalian University of Technology, 2013, 53(2): 267-272.)

作者贡献声明：

吕学强：提出研究思路，设计研究方案；
何宇：研究过程的实施，包括获取数据，进行实验，起草论文；
徐丽萍：论文最终版本修订。

收稿日期：2015-01-29
收修改稿日期：2015-03-06

A Chinese Term Extraction System in New Energy Vehicles Domain

He Yu[1] Lv Xueqiang[1] Xu Liping[2]
[1](Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science & Technology University, Beijing 100101, China)
[2](Beijing Research Center of Urban System Engineering, Beijing 100089, China)

Abstract: [**Objective**] The problem of Chinese term extraction in new energy vehicles domain is a key problem which needs a special method to improve the precision and recall rate. [**Methods**] This paper uses conditional random fields model as extraction model, select the word, word length, part of speech, dependencies, dictionary location, stop words and other characteristics as the feature templates. [**Results**] Experimental results show that the precision and recall are 93.12% and 90.47% respectively. This method improves the performance by 7.73% when compared with the baseline in terms of accuracy. [**Limitations**] This method can only improve part of the accuracy of the results. [**Conclusions**] Dependency as one of the conditional random fields model features can improve the precision and recall rate in new energy vehicles domain.

Keywords: Term extraction New energy vehicles Conditional random fields Dependency syntactic relations