北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

面向专利文献的汉语分词技术研究

岳金媛　徐金安†　张玉洁

北京交通大学计算机与信息技术学院, 北京 100044; † 通信作者, E-mail: xja2010@gmail.com

摘要　针对专利文献专业术语多、领域广的特点, 采用基于领域词典与统计相结合的方法探讨了专利文献的汉语分词问题。利用 NC-value 算法抽取专业术语, 使用条件随机场模型(CRF)提高专业术语识别率, 提高分词精度。实验结果表明, 提出的方法在开放测试下分词的准确率为 95.56%, 召回率为 96.18%, F 值为 95.87%, 大大提高了专利文献的分词精度。

关键词　汉语分词; 条件随机场; 专业术语提取

中图分类号　TP391

Chinese Word Segmentation for Patent Documents

YUE Jinyuan, XU Jin'an†, ZHANG yujie

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044;
† Corresponding author, E-mail: xja2010@gmail.com

Abstract　According to the characteristics of the patent documents, the authors present a statistics approach for Chinese word segmentation based on domain dictionaries. NC-value algorithm and conditional random fields model (CRF) are adopted for the domain terms extraction, to solve the unknown words recognition issue. The experimental results show that the proposed method can improve the efficiency of the word segmentation and the identification of the unknown words. For an open test, the precision of the experimental results is 95.56 %, the recall-rate is 96.18%, and F-measure is 95.87%.

Key words　Chinese word segmentation; conditional random fields (CRF); domain terms extraction

在社会信息化程度日益提高的今天, 专利文献已成为科学技术进步与创新的主要载体, 有效利用其所包含的大量信息, 可以避免重复研究, 减少开发时间, 降低开发成本。专利文献是了解新产品、新技术发展动向的窗口, 可以为开发新产品、新技术提供线索和借鉴[1]。目前, 随着各种技术资料文献的爆炸式增长, 汉语专利信息的处理需求日益增加, 而分词技术是其重要的基础工作, 专利检索、专利翻译的工作都离不开汉语专利文献的分词技术, 分词质量的高低直接影响专利文献应用的效率。

汉语分词和词性标注工作已经取得了非常丰硕的成果, 但是, 目前针对汉语专利文献分词研究的文献不多, 并且针对专利文献的分词工具也相对匮乏。在已有的传统技术中, 翟东升等[2]提出一种结合领域词典和规则的, 基于汉语专利权利要求书的分词方法, 能够将文本分割为有意义的特征实体, 取得了较好的分词效果。其问题在于系统的通用词典规模庞大、分词效率较低等。宋立峰[3]对比分析了基于词类的错误驱动学习、条件随机场和期望值最大等方法在专利文献分词方面的运用, 结果显示, 基于词类的错误驱动学习算法具有较高的领域适应能力, 其分词准确率达到 91%。张桂平等[4]提出一种统计与规则相结合的多策略分词方法, 利用专利文献中潜在的显、隐性切分标记和被切分文本的上

中央高校基本科研业务费专项资金(2009JBM027, 2010JBZ2007)、北京市重点学科共建项目(计算机应用技术)、中国科学院计算技术研究所智能信息处理重点实验室开放课题(IIP2010-4)和北京交通大学人才基金(2011RC034)资助

收稿日期: 2012–06–04; 修回日期: 2012–08–15; 网络出版时间: 2012-10-26 17:04

网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1704.005.html

下文信息，在此基础上进行最大概率分词，后处理中又利用术语的前后缀规律，显著提高了未登录词的正确识别率。但是，该方法存在对于出现频率较低的术语的识别精度不高等问题。

本文采用基于领域词典与统计相结合的分词方法，以期合理解决汉语专利文献的分词问题，旨在提高分词精度，降低人工劳动成本，也为专利文献的信息检索、机器翻译等提供技术基础。

1 专利文献的特点

大量未登录专业术语的密集出现[5]是专利文献的最大特点。专业术语是专利文献的重要组成部分，集中承载着学科领域的核心知识。由于专业术语的切分精度对诸如信息抽取、机器翻译等系统的整体性能的影响举足轻重[6]，因此，在进行专利文献的分词处理时必须重点研究专业术语的切分粒度。专利文献术语在用词上存在以下几个特点。

1) 术语用词遵循一定规则，语言严谨，用词一般较少出现歧义。比如“本发明的一个方面涉及”、“本申请所用的术语”等。

2) 存在大量的专业术语定语嵌套现象[7]。比如“免疫球蛋白”、“免疫球蛋白分子”、“免疫球蛋白分子编码序列”等。

3) 化学、纺织等学科的某些专业术语内部存在符号标记。比如“聚(α-甲基苯)乙烯”、“2，6-二叔丁基-4-甲基苯酚”等。

4) 专利文献带有很强的专业性，术语在某一特定专业领域内反复出现，而在其他领域则很少出现。比如“四氢呋喃”、“2-二甲基环丙醇”等，化学方面的专利会涉及这些词语，但这些词语在其他学科的专利文献的引用率则几乎为零。

5) 专利文献术语存在数据稀疏现象。比如“气相色谱法”，在整个训练语料中只出现过一次。

2 基于领域词典与统计相结合的专利文献分词方法

针对专利文献固有的特点，本文采用基于领域词典与统计相结合的分词方法。首先对训练语料进行预处理，规则与统计方法相结合最终识别出未登录的专业术语，在此基础上构建专利领域词典，应用中国科学院计算技术研究所 ICTCLAS 分词系统[8]，结合专利领域知识进行汉语分词。分词流程如图 1 所示。

2.1 预处理

对训练语料的预处理主要包括以下 3 个方面的内容：1) 文本规范化格式整理，将文本字符编码统一为 UTF-8 格式，标点统一为英文符号格式；2) 针对具体的训练语料，其中存在着一些与本分词实验无关的高频词语，且其大部分出现在每行的开头，如“1)”等分类标记、“CN200680009915.2”等专利号码，需将这些词语剔除；3) 应用 ICTCLAS 分词工具(使用自带的通用词典)，采用中国科学院计算技术研究所一级词性标记集对文本进行分词标注。

预处理之后的训练语料样式如例句所示：

各/*r* 向/*p* 异性/*n* 线/*n* 来自/*v* 于/*p* 逆变换/*v* ,/*w* 该/*r* 逆/*n* 变化/*v* 将/*p* 几何/*n* 基元/*n* 表面/*n* 上/*f* 的/*u* 点/*q* 映射/*v* 至/*p* 纹理/*n* 映射/*v* 中/*f* 的/*u* 点/*q* ./*w*

2.2 专业术语提取

汉语分词的两大难题为歧义词的切分和未登录词的识别[9]，通过对已预处理的训练语料的研究，这两大问题大部分都出现在专利文献术语的切分上，而专利文献用词规范严谨，出现歧义现象比较少，所以能够正确地识别出专业术语，大大提高未登录词的识别率，以便更准确地切分文本。

本论文专业术语提取流程如图 2 所示。分两步来提取专业领域术语：首先根据术语构词规律总结术语提取规则，应用 NC-value 算法和禁用词表，对候选术语进行评估筛选，得到初步的专业术语；其次，针对低频率的专业术语较难识别问题，将初步专业术语作为模板训练文本，运用条件随机场模型，构建术语抽取模板，再对已预处理的训练语料进行术语的标注识别，最终抽取出有意义的低频率专业术语，提高专业术语抽取精度。

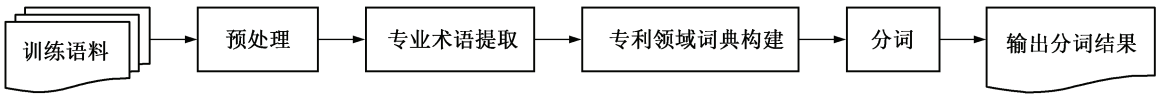

图 1 分词流程图

Fig. 1 Flow diagram of word segmentation

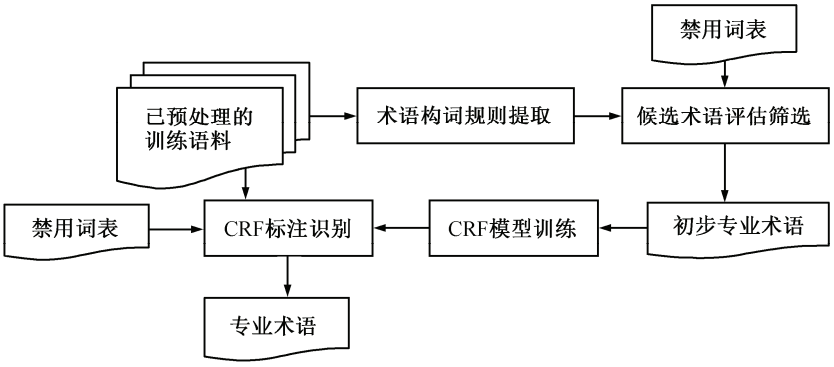

图 2　专业术语提取流程图

Fig. 2　Flow diagram of domain terms extraction

2.2.1　初步专业术语提取

术语的组成绝大多数为名词、形容词、动词和数量词, 且多为名词性短语, 长度一般为 2~8 个汉字, 且最后 1~3 个字大多是其中心词, 构词模式比较有限。研究发现, 语料中存在着明显的数字、标点、度量的基本单位、连接词等切分标记, 比如“术语‘XX’”、“XX 和 XX”、“XX、XX”、“XX135ml”等标记。鉴于此本文根据专业术语的构词特点、词性标注信息总结出术语提取的规则, 再结合上下文信息及语料中的切分标记, 经 NC-value 算法多次统计训练来进行专业术语的提取。总结出的术语的部分提取规则如表 1 所示。

表 1　术语提取的正则式实例

Table 1　Examples of regular expressions of terms extraction

术语提取模式	术语提取的正则式
名词+名词	$n^{+}n$
(形容词|动词|名词)+名词	$(a\|v\|n)^{+}n$
名词+(名词|动词|量词|未识别的词)	$n(n\|v\|q\|x)^{*}$

说明: *n* 表示名词, *v* 表示动词, *a* 表示形容词, *q* 表示量词, *x* 表示未识别的词, +表示出现一次以上, *表示出现零次以上。

对于语料中一些较长的化学术语, 其长度一般在 15~30 个字符之间, 且其内部存在多处符号标记, 一般均为“-”、“,”、“{}”、“[]”、“()”等符号, 较难准确抽取, 需对此种情况采用特殊的规则进行术语抽取。长化学术语的部分提取规则如表 2 所示。

NC-value 是 Frantzi 提出的一种领域独立的多词术语的统计抽取算法, 利用 NC-value 算法[10−11]在长术语识别、反映术语的上下文信息方面的优势, 本文采用 NC-value 算法来评估候选术语是否为有着实际意义的专业术语。计算候选术语的 NC-value 值

表 2　长化学术语提取的正则式实例

Table 2　Examples of regular expressions of long chemical terms extraction

错分样式例举	提取的正则式
二/*m* 异/*a* 氰酸/*n* 酯/*n*	$(a\|m\|x)^{*}nn^{*}(x\|v)^{*}$
2-/*m* 氟/*n* -3-/*m* 苄/*x* 氧/*n* 基苯甲醛/*n*	$Mnn^{*}(m\|x)^{+}nn^{*}$
甲苯/*n* 2/*n* ,/*w* 4-/*m* 二/*m* 异/*a* 氰酸/*n* 酯/*n*	$m^{*}nn^{*}w(a\|m\|x)^{+}nn^{*}$
4-/*m* (/*w* 甲/*m* 氧/*n* 基/*n* 羰/*x* 基/*n*)/*w* 苯甲醛/*n* 肟/*x*	$(m\|w)^{*}(a\|m\|x)^{+}nn^{*}(x\|w)^{*}$

说明: *n* 表示名词, *v* 表示动词, *a* 表示形容词, *m* 表示数词, *x* 表示未识别的词, +表示出现一次以上, *表示出现零次以上, *w* 表示“-”、”,”、”{}”、”[]”、”()”等符号。

的公式如下:

$$\text{C-value}(a)=\begin{cases}\log_2|a|f(a), & a\text{未被嵌套包含},\\ \log_2|a|\left(f(a)-\dfrac{1}{p(T_a)}\sum\limits_{b\in T_a}f(b)\right), & \text{其他},\end{cases} \tag{1}$$

$$\text{NC-value}(a)=\alpha*\text{C-value}(a)+\beta*\sum_{b\in T_a}f_a(b)*\frac{f(b)}{p(T_a)},\quad(\alpha+\beta=1), \tag{2}$$

其中, *a* 表示候选的字符串, |*a*|表示字符串 *a* 的长度, *f*(*a*)表示字符串 *a* 的词频, T_a 表示包含字串 *a* 的术语, *b* 表示 T_a 中任意的包含字串 *a* 的术语, $P(T_a)$表示包含字串 *a* 的术语总数, $f_a(b)$表示 *b* 在字串 *a* 的上下文中出现的次数。

因应用基于通用领域的 ICTCLAS 分词工具进行粗切分, 在本专利训练语料上会出现一定的分词标注错误, 相应的术语识别结果中也会存在部分错误, 很容易将整个专业术语或术语的一部分与其他词语划分成一个字串, 且多在术语前后出现此种词语粘连现象。为解决此类错误延伸问题造成的术语识别错误, 采用建立禁用词表的方式, 对经 NC-

value 算法评估后的术语, 通过前后缀禁用词过滤对术语识别结果进行校正, 以提高术语抽取的准确率, 实现对候选术语的评估抽取, 最终筛选出那些具有实际语义的词作为初步的专业术语。总结出的禁用词部分规则如表 3 所示。

表 3　禁用词规则

Table 3　Rule of the disable word

禁用类型	禁用词	不合格术语
术语前缀为数量词、介词	一个、为了、由于等	单根光纤布线
术语后缀为方位词、介词、数词、部分动词	上、下、除了、2、输入、表示等	品质因数上

2.2.2　低频率专业术语提取

鉴于专利文献术语数据稀疏的特点, 大量的专业术语出现频率较低, 词频并不占据主导地位, 通过第一步得到的初步的专业术语很少包含有意义的低频率术语。 条件随机场模型(CRF)是 Lafferty 等[12]于 2001 年在最大熵模型(MEM)和隐马尔科夫模型(HMM)的基础上提出来的, 是用于切分和标注序列化数据的统计模型, 其目标是在给定需要标记的观察序列的条件下, 使标记序列的联合概率达到最优。 条件随机场模型(CRF)具有表达字串长距离依赖性和交叠性的能力, 能较好地学习新的领域知识[13], 所以采用 CRF 模型来识别出现频率较低的术语。条件随机场定义如下:

$$P(y\mid x)=\frac{1}{Z(X)}\exp\left(\sum_i\sum_k\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_i\sum_k u_k s_k(y_i,x,i)\right), \tag{3}$$

其中, $t_k(y_{i-1},y_i,x,i)$为转移函数, 表示观察序列和标记序列在 $i-1$ 和 i 时刻的特征; $s_k(y_i,x,i)$为状态函数, 表示观察序列和标记序列在 i 时刻的特征; $Z(X)$为归一化因子; λ 和 u 为训练所得参数。

CRF 将术语抽取看做一个序列标注过程, 基于由字构词的理念, 利用词位信息来标记术语, 术语抽取的过程即为将字在字串中的特征进行标记的过程。本文采用四词位标注集, 四词位标注集的说明如表 4 所示。

特征模板的设置对术语标注识别的好坏起到关键的作用, 本文利用上下文的信息, 从训练语料中获得字特征, 主要采用当前字和其前后两个字及其词性信息作为特征。具体的特征模板的设置如表 5 所示。

我们将得到的初步专业术语的特征量化, 作为训练特征, 利用 CRF 模型训练出术语抽取模板, 利用该模板对经过 ICTCLAS 工具粗切分的训练语料进行标注抽取, 通过禁用词表对术语识别的错误情况进行修正, 识别出更多的专业术语。针对得到的专业术语研究发现, 有时会出现术语截取现象, 比如术语中有这样的两个词“阿茨海默/*n*”、“阿茨海默氏病/*n*”, 实际上“阿茨海默氏病”是一个专业术语, 需对术语截取方面的问题进行相关的后处理操作。

表 4　四词位标注集

Table 4　Four tag set

标注符号	不同词长的标注形式
B(术语的首字)	BE(2 字术语)
M(术语的中间字)	BME(3 字术语)
E(术语的尾字)	BM...ME(长度大于 3 的术语)
O(非术语)	O/O...O(非术语)

表 5　特征模板

Table 5　Feature template

特征类型	特征模板
Unigram(一元)	C_n, S_n　n=−2, −1, 0, 1, 2
Bigram(二元)	C_nC_{n+1}, S_nS_{n+1}　n=−2, −1, 0, 1

说明: C 表示当前字, S 表示词性特征。

2.3　专利领域词典的构建及分词

将 CRF 模型识别出的术语结果与初步的专业术语合并整理, 即为最终识别出的专业术语。整理后的专业术语根据 ICTCLAS 分词工具的词典格式[8]进行修改, 再将最终得到的专业词典加入到 ICTCLAS 分词工具中, 结合专利领域知识对训练语料进行词语切分, 得到切分序列。

3　实验

3.1　实验语料

本实验所用的语料是 NTCIR-9 会议的 Patent-MT 任务提供的汉语专利数据, 选取了 10 万句生语料作为汉语专利文献分词实验的训练集数据, 在训练语料中抽取 8000 句作为专业术语提取结果的抽样评价数据, 选取 103 篇专利文献(摘要、说明书、权利要求)作为分词实验的测试集数据(约 36000 句), 词性标注采用中国科学院计算技术研究所一级标注集, 标注集可参见《ICTPOS 汉语词性标记集》[8], 将按照专利术语标注标准[14]手工标注的测试语料作为标准结果集。实验数据的具体构成情况如表 6 所示。

表 6　实验数据统计信息

Table 6　Statistical information of the experimental data

数据种类	生活需要/%	化学和纺织/%	机械工程/%	物理/%	电子/%	医药/%	其他/%
专业术语提取结果抽样评价数据	22.12	7.27	6.67	10.91	14.12	20.00	18.91
汉语专利文献分词实验训练数据	20.12	15.33	5.36	7.00	17.58	17.64	16.97
汉语专利文献分词实验测试数据	29.13	14.56	6.80	7.77	11.65	16.70	13.39

3.2　评测方法

一般而言, 分词结果的评测有以下 3 个指标, 即分词的准确率、召回率、F 测度值。各指标用式(4)~(6)表示:

$$准确率P=\frac{正确分词的数量}{总的切分数量}\times 100\%, \quad (4)$$

$$召回率R=\frac{正确分词的数量}{标准结果集实有切分的数量}\times 100\%, \quad (5)$$

$$F测度值=\frac{2\times 召回率\times 正确率}{召回率+正确率}\times 100\%。 \quad (6)$$

3.3　实验结果及分析

本文采用较为成熟的 NC-value 算法和 CRF++工具[15]对训练语料进行训练, 以提取专利领域术语。实验使用 NC-value 算法时, 依据最小二乘法优化公式(式(2))中的参数, C-value 权重 α 和上下文信息权重 β 分别为 0.8 和 0.2, 通过对算法实验结果的分析将 NC-value 的阈值设为 0.0169, 即若 NC-value 值大于阈值, 就判定该词为一个专业术语。NC-value 算法和 CRF 算法专业术语识别的抽样统计结果如表 7 所示。

表 7　术语提取实验抽样结果

Table 7　Experimental results in terms extraction

术语提取方法	准确率 P/%	识别出的正确的术语/条	识别出术语/条
NC-value 算法	85.29	87	102
CRF 算法	78.85	179	227

由表 7 可见, 虽然通过 CRF 模型训练准确率相较 NC-value 算法有所下降, 但识别出了更多的专业术语。本实验中应用 NC-value 算法识别出了 24898 条初步的专业术语, 通过 CRF 模型训练相较 NC-value 算法词汇差异度(变化数量)为 236182 条。

本文通过规则的提取和统计学习来构建专利领域词典, 构建的领域词典共有 282657 项条目, 再应用 ICTCLAS 分词系统结合专利领域词典进行汉语分词。表 8 是在开放测试环境下针对测试语料全部单词进行切分的实验结果。

表 8　开放测试下实验结果

Table 8　Experimental results in an open test

分词方法	准确率 P/%	召回率 R/%	F 测度值/%
ICTCLAS 分词系统	71.65	85.42	77.93
NC-value 方法	88.75	93.80	91.21
本实验方法	95.56	96.18	95.87

说明: NC-value方法指仅应用NC-value算法构建专利领域词典的分词方法。

由表 8 可见, 通过在测试语料的多个领域的专利文献上进行分词实验, 本文提出的基于领域词典与统计相结合的分词方法与单纯使用 ICTCLAS 分词系统相比, 取得了较好的分词效果。

通过对切分序列的错误情况进行分析, 大多是因语料数据稀疏和上下文信息不同引起的术语识别不清。例如"光学脉冲"是在文本中出现较为多次的术语, 但"光学脉冲整形"只出现一次, 会被错误切分成"光学脉冲/*n* 整形/*v*"; "记录道频谱/*n*"被正确的作为一个专业术语识别出来, 但因词语的上下文信息差异, 在文本中也有极少的"记录道/*n* 频谱/*n*"的错误识别。

4　结语

本文在分析归纳现有方法的基础上, 根据专利文献自身的特点, 提出了一种领域词典与统计相结合的分词方法, 规则和统计方法并用, 与现有的 ICTCLAS 分词系统相比在专利领域内切分取得了很高的准确率与召回率, 且通过提取专业术语大大提高了未登录词的识别效率。在未来的研究中, 我们会进一步扩充语料的规模, 减少数据稀疏的影响, 挖掘语料中新的语言学规律, 在利用上下文信息方面进行加强改进, 以期更有效地识别出有实际意义的专业术语, 进一步提高分词精度。

参考文献

[1] 李绩. 专利文献的特点及利用. 中国科技成果, 2008(23): 27–29

[2] 翟东升, 马文姗. 中文专利权利要求书分词算法研究. 情报杂志, 2011, 30(11): 152–155

[3] 宋立峰. 中文分词算法在专利文献中的应用研究. 海峡科学, 2011(7): 9–11

[4] 张桂平, 刘东生, 尹宝生, 等. 面向专利文献的中文分词技术的研究. 中文信息学报, 2010, 24(3): 112–116

[5] 谷俊, 王昊. 基于领域中文文本的术语抽取方法研究. 现代图书情报技术, 2011(4): 29–34

[6] Tseng Y H, Lin C J, Lin Y I. Text mining techniques for patent analysis. Information Processing and Management, 2007, 43: 1216–1247

[7] 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. 计算机工程与应用, 2008, 44(23): 147–150

[8] ICTCLAS 简介[CP/OL]. (2012–04–16) [2012–05–30]. http://ictclas.org/

[9] 宗成庆. 统计自然语言处理. 北京: 清华大学出版社, 2008: 105–120

[10] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. Intl. Journal on Digital Libraries, 2000, 3(2): 115–130

[11] 梁颖红, 张文静, 周德富. 基于混合策略的高精度长术语自动抽取. 中文信息学报, 2009, 23(6): 26–30

[12] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data // Proceedings of ICML-01. Berkshires of western Massachusetts, 2001: 282–289

[13] He Y, Kayaal P M. Biological entity recognition with conditional random fields // Proceedings of AMIA Annual Symposium. Washington, DC, 2008: 293–297

[14] 国家技术监督局. 中华人民共和国国家标准 GB/T 13715—92 信息处理用现代汉语分词规范. 北京: 中国标准出版社,1993

[15] CRF++: Yet Another CRF toolkit [CP/OL]. (2012–05–30) [2012–08–21]. http://crfpp.googlecode.com/svn/trunk/doc/index.html