基于条件随机场的专利术语抽取*

刘辉，刘耀
（中国科学技术信息研究所，北京 100038）

摘要：专利术语抽取是专利文献信息抽取领域的一项重要任务，有助于专利领域词表的构建，有利于中文分词、句法分析、语法分析等工作的进行。文章通过分析专利术语的特点并制定相应的语料标注规则进行人工标注，采用条件随机场（conditional random fields，CRFs）对标注后的数据进行训练和测试，实现了通信领域的术语抽取。标注方法采用基于字的序列标注，精确率、召回率和F值分别达到80.9%、75.6%、78.2%，优于将词和词性等信息作为特征的方法，表明所提出的专利术语抽取方法是有效的。

关键词：条件随机场；术语抽取；序列标注
中图分类号：TP391.1
DOI：10.3772/j.issn.1673—2286.2014.12.008

引言

专利文献是科技信息的载体，集中体现了科学技术的发展水平，有效利用专利可以提高国家和企业的发展速度。快速找出专利文献中相应的技术信息是有效利用专利文献的前提。在专利文献中专业术语是其核心内容和重要组成部分。对术语的分析研究是深入和有效应用专利的基础性工作。因此，研究专利文献术语的抽取技术越来越受到研究者们的关注，专利文献中的术语体现和承载了专利文献的技术信息。同时，通过所提取的专利文献术语，可以构建专利领域叙词表，有利于分词、句法分析、语法分析等工作的顺利进行，也可以进一步对专利文献进行分类，识别不同专利文献之间的相互关系。

目前，较为常用的抽取方法主要有三种：第一种是基于规则的方法，根据语言学知识制定相应的规则模板，按照规则模板对专利文献的术语进行匹配，匹配成功则抽取其中的术语部分。姚振军等运用正则表达式的字符串匹配功能对特定数据库中的汉英对照中国文化术语进行了抽取[1]；刘里等提出了一种领域现象术语的抽取方法，采用分隔符集合上下文术语进行候选领域现象术语的抽取[2]。但基于规则的术语抽取方法不够灵活，规则很难涵盖复杂的语言现象，尤其是随着现代科技的快速发展，新的术语层出不穷，人工来研究其语言学规律变得不可行。第二种是基于统计的方法，张锋等通过采用互信息计算字串的内部结合强度得到术语候选集，实验F值达到74.97%[3]；岑咏华等提出了一种基于双层隐马尔科夫模型的中文泛术语识别和提取的思路和系统框架[4]；刘豹等学者采用条件随机场对科技术语和军事领域术语进行了抽取，F值分别达到了84.4%和76.46%[5-6]。第三种是将规则与统计相结合的方法，唐涛等采用统计方法得出术语候选集，再采用规则的方法进行过滤和选取[7]；韩红旗等采用语言学规则得出可能的术语候选列表，再计算词语的术语度值选取候选术语[8]；章成志将语言学方法与统计方法进行并行融合，进行基于多层术语度的一体化术语抽取[9]。本文主要就专利文献的术语抽取任务展开讨论，分析专利文献中术语的特点及抽取难点，利用条件随机场

* 本研究得到“十二五”国家科技支撑计划项目“专利信息资源挖掘与发现关键技术研究”（编号：2013BAH21B02）资助。

模型对专利文献中的术语进行自动抽取。

1 条件随机场

条件随机场是一种以给定的输入节点值为条件来预测输出节点值概率的无向图模型，它是由Lafferty等在2001年提出的一种用于序列数据标注的条件概率模型，是一种判定性模型[10-11]。CRFs通过定义标记序列和观察序列的条件概率P(S|O)来预测最可能的标记序列。CRFs不仅能够将丰富的上下文特征整合到模型中，而且还克服了其他非产生性模型的标注偏差问题。线链CRFs的图形结构如图1所示。

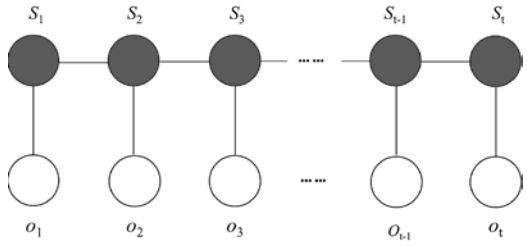

图1 线链CRFs的图形结构

设$O=\{O_1,O_2,\cdots\cdots,O_T\}$表示被观察的输入数据序列，例如有待标注的字或字串序列，$S=\{S_1,S_2,\cdots\cdots,S_T\}$表示被预测的状态序列，每一个状态均与一个词位标记相关联。这样，在一个输入序列给定的情况下，参数为$\Lambda=\{\lambda_1,\lambda_2,\cdots\cdots,\lambda_K,\}$的线链CRFs，其状态序列的条件概率为：

$$P_{\triangle}(S|O)=\frac{1}{Z_O}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t)\right) \quad (1)$$

其中，Z_O是归一化因子。它确保所有可能的状态序列的条件概率和为1，即它所有可能的状态序列的“得分”的和：

$$Z_O=\sum_S \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t)\right) \quad (2)$$

$f_k(s_{t-1},s_t,o,t)$是一个任意的特征函数，通常是一个二值表征函数。λ_k是一个需要从训练数据中学习的参数，是相应的特征函数$f_k(s_{t-1},s_t,o,t)$的权重，取值范围可以是-∞到+∞。特征函数$f_k(s_{t-1},s_t,o,t)$能够整合任何特征，包括状态转移$S_{t-1}\rightarrow S_t$特征，以及观察序列O在时刻t的所有特征。

给定一个由公式（1）定义的条件随机场，在已知输入数据序列O的情况下，最可能的标记序列可以由下式求出：

$$S^*=\arg\max P_{\triangle}(S|O) \quad (3)$$

最可能的标记序列可以由上式通过类似于隐马尔科夫模型中的韦特比算法动态规划求出。

2 基于CRFs的专利术语抽取

2.1 专利术语的特点

术语是代表学科领域基本概念的语言单元，可以是词也可以是词组，在我国又称为科技名词。目前对区分术语与普通词语并没有统一的标准。本文在标注中所规定的专利术语长度由一个字到十多个字不等，可以为两个字的词语“转子”，也可以为多字的“双频段宽带电台室内联试通信仿真系统”，术语必须具有较强的领域性，如“实现方法”、“成本”等通用词语则不纳入术语的认定范围内。专利术语有其自身的特点，主要有：长术语多，类似“外均衡高温超导线性相位滤波器”的长术语数量众多，字数的增加导致发生歧义的可能性增大；英文缩写术语多，如“CDMA”；英文缩写与中文词语合用，如“ISDN调度终端”。专利的用语较规范，其文本具有一定的结构性，若可以将术语进行有效的提取，剩下的文本可以视为规范的语言模板，如“本发明涉及一种……的方法”。

2.2 语料处理

采用CRFs做术语抽取时，通常的做法是将文本先进行分词，再进行词性标注，将词本身和词性作为主要的特征[4]，也可加入词频、互信息、左右信息熵等更多信息作为特征[5]。但由于专利文本的特点，分词可能造成许多问题，比如：专利文本中的长术语较多，分词一般无法将术语中的词语进行正确的切分，并且术语中存在不少的单个字表示词义的情况，词性作为特征无法对术语的识别提供较大帮助，另一方面，过多的特征数量容易造成过拟合问题，影响模型的效果。所以，我们采用基于字的标注方法，即使用B（术语首字）、I（术语中字）、E（术语尾字）、O（其他字）。基于字的术语抽取问题就是把术语抽取过程看作每个字的字位标注问题，如果一个汉字字串中每个字的字位都确定了，那么该字串中的术语也就识别完成了。例如：要对字串“所述编码单元持续将

媒体流向配置的组播组地址发送”进行术语识别，只需标注出该字串的序列（1），有标注结果就很容易得到对应的术语抽取结果（2）。

（1）标注结果：所/O述/O编/B码/I单/I元/E持/O续/O将/O媒/B体/I流/E向/O配/O置/O的/O组/B播/I地/I址/E发/O送/O。

（2）识别结果：编码单元 媒体流 组播地址。

2.3 模板定义

按照CRFs的要求设计相对应的特征模板，模板是对上下文环境中的特定位置和特定信息的考虑，反映了所要考虑的语言现象的选取标准，也可以理解为它指导和限定了机器学习过程的空间范围。特征模板文件中的每一行代表一个template。每一个template中，专门的宏%x[row,col]用于确定输入数据中的一个token，row用于确定与当前的token的相对行数，col用于确定绝对行数。有两种类型的模板文件，类型可由第一个字符来区分，第一种是Unigram template，第一个字符是U，当给出一个模板"U01:%x[0,1]"，CRFs会自动地生成一个特征函数集合(func1 . . .funcN)。另一种是Bigram template。第一个字符是B，这个模板用于描述bigram features。根据本文标注情况，编写了相对应的特征模板文件template，模板文件片断如下：

```
#Unigram
U000:%x[-2,0]
U001:%x[-1,0]
U002:%x[0,0]
U003:%x[1,0]
U004:%x[2,0]
```

图2 模板文件片断

3 实验

3.1 实验数据

数据采用通信领域的专利文献摘要1000篇，共702742字。首先对专利摘要中的术语进行人工标注，三人同时对1000篇语料进行标注，标注出的结果采用三份标注结果的交集，总共得到2216个术语。再采用计算机对标注后的语料进行序列标注，对术语的首字标上“B”，术语的尾字标上“E”，术语的内部字标上“I”，非术语文字和符号等全标上“O”。为了减少数据不平衡的影响，将语料分成五组，进行开放的交叉测试。

3.2 实验结果及分析

评价标准采用准确率(P)、召回率(R)以及F值作为评价指标，计算方法如下：

$$P=\frac{\text{正确识别的术语数}}{\text{识别的术语总数}}\times 100\% \quad (4)$$

$$R=\frac{\text{正确识别的术语数}}{\text{语料中的术语总数}}\times 100\% \quad (5)$$

实际评估时，应同时考虑P和R，但同时要比较两个数值，很难做到一目了然。所以常采用综合两个值进行评估的办法，综合指标F值就是其中的一种。计算公式如下：

$$F=\frac{(\beta^2+1)PR}{\beta^2 P+R}\times 100\% \quad (6)$$

其中，β决定对P侧重还是对R侧重，通常设定为1、2或1/2。本文β取值1，即对二者重视程度一样。表1为五组数据的测试结果。

表1 术语抽取结果

组号	P	R	F
一	0.802	0.763	0.782
二	0.818	0.774	0.795
三	0.836	0.752	0.792
四	0.783	0.739	0.76
五	0.807	0.752	0.779
平均	0.809	0.756	0.782

对数据进行开放测试后，准确率可以达到80%左右，可以识别出较为复杂的专利术语，如字数较多的术语“正交频分服用多载波无线通信系统”，中英文结合的专利术语，如“TTCAN网络时间主节点”。但同时还有一定的错误情况。将采用模型进行标注的结果与人工标注的结果进行比对，发现错误主要集中在以下几个方面：

（1）识别词语不全，如“等离子显示面板”识别成了“离子显示面板”，“物理混合自动重传请求指示符信道”识别成了“混合自动重传请求指示符信道”。

（2）识别出的词比正确的术语多出一部分，如“通信设备”识别成“测试通信设备”，“时钟同步消息”识

别成了“时钟同步消息状况”。

（3）将一些普通词语当作术语识别出来。如“传输方式”、“实现模式”。

（4）未识别出术语、误识别术语等其他错误。

3.3 与基于词标注方法的比较

对比的方法采用基于词的序列标注，对文本进行分词和词性标注等处理，实验采用不同的特征进行训练和测试，一种采用词本身、词性作为特征，另一种使用词本身、词性、词长和词频多个特征。

实验表明，采用以词为单位进行序列标注实验的结果不如基于字标注的实验，加上词长和词频等特征后，召回率提高了，但准确率却有所下降。这说明实验采用的特征对结果起着重要作用，特征并非越多越好，而是需要找到最适合数据要求的特征，并且过多的特征数量也容易造成数据的过拟合问题。

表2 与基于词标注方法的比较

特征	P	R	F
字	0.809	0.756	0.782
词、词性	0.713	0.641	0.675
词、词性、词长、词频	0.686	0.652	0.67

4 结语

本文主要针对专利文献的术语进行了抽取，将术语抽取问题转化为序列标注问题，使用CRFs模型对标注好的专利摘要进行训练和测试，采用交叉验证的方法进行开放测试，最终的准确率达到80%，并与基于词的实验进行了对比，实验表明，字标注的结果要好于采用词等特征的实验结果。

参考文献

[1] 姚振军,黄德根,纪翔宇,等.正则表达式在汉英对照中国文化术语抽取中应用[J].大连理工大学学报,2010,50(2):291-295.

[2] 刘里,刘小明.基于分隔符和上下文术语的领域现象术语抽取[J].华南理工大学学报:自然科学版,2011,39(7):146-149,155.

[3] 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):72-73,77.

[4] 岑咏华,韩哲,季培培,等.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.

[5] 刘豹,张桂平,蔡东风,等.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.

[6] ZHENG D Q, ZHAO T J,YANG J. Technical term automatic extraction research based on statistics and rule [C]// ICCPOL 2009, LNAI 5459. Berlin: Springer-Verlag, 2009: 290-296.

[7] 唐涛,周俏丽,张桂平,等.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.

[8] 韩红旗,朱东华,汪雪锋,等.专利技术术语的抽取方法[J].情报学报,2011,30(12):1280-1285.

[9] 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,2011,28(3):275-285.

[10] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]// Proceedings of 18th International Conference on Machine Learning. San Francisco, USA: AAAI Press, 2001: 282-289.

[11] Peng Fuchun, McCallum A. Accurate information extraction from research papers using conditional random fields [J]. Information processing and management, 2006, 42(4): 963-979.

作者简介

刘辉，1990年生，男，硕士研究生，研究方向：信息抽取、图储存，E-mail：liuhui2013@istic.ac.cn。
刘耀，1972年生，男，博士后，研究员，研究方向：知识工程，中文信息处理，E-mail：liuy@istic.ac.cn。

Patent Term Extraction Based on Conditional Random Fields

LIU Hui, LIU Yao
(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Patent term extraction is an important task in patent information extraction, which benefits the construction of lexicography, the work of word segmentation, and parsing. Corpus is labeled manually with corresponding rules written by analyzing the characteristics of patent terms. CRFs (Conditional Random Fields) is adapted to train and test labeled data. Sequence labeling is based on single Chinese characters. Experimental results show that the precision, recall and F-score are 80.12%, 74.2% and 76.9% respectively, which are superior to methods based on sequence labeling of words. Results illustrates that the established model for extracting patent term is effective.

Keywords: Conditional random fields; Term extraction; Sequence labeling　（收稿日期：2014-12-10）