条件随机场与规则集成的专利摘要信息抽取*

□ 李鹏 桂婕 乔晓东 张兆锋 / 中国科学技术信息研究所 北京 100038

摘要：专利是一种重要的情报分析数据来源，由于专利使用的术语比论文更为抽象等原因，基于统计的信息抽取效果并不理想。文章利用文档结构的特点以及专利写作过程中的常用特色词汇，在利用条件随机场这种概率模型的基础上，提出了集成基于规则的专利摘要信息抽取方法。系统参加亚洲语言信息检索测评会议专利挖掘之技术趋势图谱子任务，取得较好的成绩，证实其系统的实用性与高效性。

关键词：信息抽取，条件随机场，规则
DOI：10.3772/j.issn.1673—2286.2010.09.002

1 引言

条件随机场（Conditional Random Fields，CRFs），是一种用来标记和切分序列化数据的统计框架模型，最早由Lafferty等人于2001年提出[1]，其模型思想主要来源于最大熵模型。但是，相比较而言，CRFs能够更好地模拟真实世界的数据。首先，CRFs突破了隐马尔可夫模型的强独立性假设条件的约束，具有容纳任意的上下文信息的优点。其次，CRFs能够计算全局最优输出节点的条件概率，克服了最大熵马尔可夫模型的长度偏置和标签偏置的缺点。因此，CRFs被广泛用于自然语言处理的研究领域，如组块分析[2]、汉语动宾搭配自动识别[3]、中文自动文摘[4]等任务。CRFs等统计方法覆盖面大，容易实现，缺点是容易丢失低频功能词和错误识别高频功能词。

基于规则的方法，主要是利用一些语言学的知识，以及特定的领域知识来制定一套规则，其优点是识别准确率高，主要缺点在于规则由手工编写，费时费力，规则覆盖面有限。

由于基于规则和基于统计的方法各有优缺点，将两者结合起来能够取长补短，如刘豹[5]等提出了一种使用条件随机场模型进行标注识别，并结合规则对错误识别结果进行后处理的科技术语识别方法。

本文选择CRFs作为专利摘要信息抽取的统计模型，同时结合专利文档结构的特点以及专利写作过程中的常用特色词汇，在利用条件随机场这种概率模型的基础上，提出了集成基于规则的专利摘要信息抽取方法。

2 CRFs与规则集成的专利摘要信息抽取

2.1 专利技术及其功效抽取的任务

亚洲语言信息检索评测会议（NACSIS Test Collections for IR，NTCIR）是由日本科技文部省下的情报信息研究所（NII）主办的多语言处理国际评测会议，主要关注中、日、韩等亚洲语种的相关信息处理。该评测会议的主要目的是增强信息获取技术的研究交流，包括信息检索、问答系统、文本摘要、信息抽取等。NTCIR-8的专利挖掘任务的目标[6]是从研究论文和专利数据集中创建技术趋势图谱。表1显示了技术趋势图谱的示例。在图谱中，研究论文和专利根据基

* 本文得到国家科技部“十一五”科技支撑计划（项目编号：2006BAH03B03）、中国科学技术信息研究所重点工作项目（项目编号：2009KP01—7—1）、中国科学技术信息研究所2009年度预研基金项目（项目编号：YY—200906）等项目的资助。

表1 技术趋势图谱示例

	功效1	功效2	功效3
技术1	[AA 1993] [US Pat. XX/XXX]		[BB 2002]
技术2	[CC 2000]		
技术3		[US Pat. YY/YYY]	[US Pat. ZZ/ZZZ] [JP Pat. WW/WWW]

本技术和其功效被进行了分类。

本文所用的语料为专利摘要语料，其目的是从专利标题及摘要中提取基本技术（Technology）及其功效（Effect）的内容[6]。技术，是指专利文献中涉及到的算法、工具、材料或者数据等；功效，是指技术与相应产生的功能或者效果。“技术”和“功效”的抽取内容被定义如下：

技术：由< TECHNOLOGY >标签中的内容表示，包括每项研究或发明中使用到的算法、工具、材料或数据。

功效：由< EFFECT >标签中的内容表示，包括一对属性和值的标签。

属性：由< ATTRIBUTION >标签中的内容表示，包括一项技术对应功效中所具有的属性。

值：由< VALUE >标签中的内容表示，包括一项技术对应功效中所具有的属性对应的取值。

抽取内容的示例如下：

Consequently, when <TECHNOLOGY>the driver ICs</TECHNOLOGY> are connected to each other, <EFFECT><ATTRIBUTE>the number of shift register bits</ATTRIBUTE> can be <VALUE>adjusted</VALUE></EFFECT>.

2.2 基于CRFs的抽取模型

基于CRFs的抽取模型能够综合利用对字、词、词性等多层次的资源。CRFs抽取模型的实现主要包括模型特征标记、构建训练模型、利用训练数据对模型进行训练、训练结果输出等环节。运行CRFs抽取模型使用的是开源软件CRF++[7]。

本文根据已标注语料将信息进行分类，定义了7类位置标签：S、B-TECHNOLOGY、I-TECHNOLOGY、B-TECHNOLOGY、I-TECHNOLOGY、B-VALUE、I-VALUE，分别对应无特定意义词、技术词首词、技术词首词外的其他词，以及功效属性、功效值对应的首词和其他词。

2.2.1 特征选择

本文定义了三个特征用于CRFs抽取模型中，具体内容如下：

特征1：词性特征。我们在分析了“技术”标签中的内容后发现，标注内容大多以名词或者名词短语的形式出现，因此，选择词性作为定义特征1。词性标注的任务就是根据一个词在某个特定句子中的上下文，为这个词标注正确的词性。词性标注工具较多，本文选取了POStagger[8]进行词性标注。

特征2：冠词特征。我们在分析了“技术”标签中的内容后发现，标注内容如果是以名词短语形式出现，则其开头的首字母多为a、an、the等冠词。系统对于冠词特征，进行二元分类，即是否为冠词。如果是冠词，则特征2的参数值为1，否则参数值为0。

特征3：词频特征。在选定词性特征、冠词特征后，本文也将词频作为一个重要特征引入。实际操作中，为了词频更加客观，对单复数、时态等进行了词干化处理。

2.2.2 特征模板定义

特征模板，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，它指导和限定了机器学习过程的空间范围。本文所用的特征模板定义如下：

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]
U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]
```

U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]
U30:%x[-2,2]
U31:%x[-1,2]
U32:%x[0,2]
U33:%x[1,2]
U34:%x[2,2]
U35:%x[-2,2]/%x[-1,2]
U36:%x[-1,2]/%x[0,2]
U37:%x[0,2]/%x[1,2]
U38:%x[1,2]/%x[2,2]
U40:%x[-2,1]/%x[-1,1]/%x[0,1]
U41:%x[-1,1]/%x[0,1]/%x[1,1]
U42:%x[0,1]/%x[1,1]/%x[2,1]
 # Bigram
B

2.3 基于规则的抽取模型

为了弥补基于统计抽取模型的不足，本文在CRFs的基础上，引入了基于规则的抽取模型，对其结果进一步抽取，以提高抽取效果。

专利文献作为一种具有法律效力的文本，其在写作中具有一些固定的特点，表现为在突出专利技术等内容时，经常使用一些常用写作词语。文章在分析标注为technology的词语时，有一些频率多次出现的词，比如：include、comprise、use、have以及by means of等。这些词出现后，一般会有大量的technology相关内容。

本文使用的规则包括基本规则，以及特殊规则，具体定义如下：

基本规则1：<技术>标签中的词为名词词组时，如：“冠词＋形容词＋名词”或者“名词”或者“形容词＋名词复数”，即要求此词组必须为以非动词开始的词组。

基本规则2：过滤掉a plurality of。如果A，B之前出现此词组，过滤掉。

（2）基于具体动词的抽取规则（此规则以动词include举例）

动词：includ（替代include\including）

规则1：includ A;B;C。解释：所有出现在includ之后的以分号分隔的均为<技术>标签中的词，即A，B，C都替换为：<tech>A</tech>形式；

规则2：includ A,B,and C。解释：所有出现在includ之后的以逗号分隔的均为<技术>标签中的词，即A，B，C都替换为：<tech>A</tech>形式，并去掉and；

规则3：includ A and B。解释：同规则2；

规则4：includ (a) A and (b) B　includ (1) A and (2) B；

规则4：not that/which/may includ。解释：排除以that which may开始的includ形式的句子；

规则5：plurality of A are included。解释：出现如此形式的句子中A必为<技术>标签中的词；

规则6：plurality of A included in B。解释：出现如此形式的句子中A，B均为<技术>标签中的词。

上述规则以动词include为例，其他具有相近含义的动词，在进行动宾句式结构抽取时都可参考此规则。

3 实验及测评结果

本实验的目的为检验本文提出的CRFs与规则集成抽取方法的有效性。

3.1 测评数据集

本文采用的测试数据集为NICIR-8提供的测评数据和评估数据，使用的是英文专利（摘要）数据，两类数据共计500篇。其中，主办方从两类数据中各随机挑选50篇作为演习（dry run）评估数据、200篇作为正式比赛（formal run）评估数据，剩下的各250篇文档作为训练数据供参赛队伍使用。表2显示了专利挖掘任务的专利数据集信息。

表2　NTCIR−8专利挖掘任务的专利数据集

数据集	数据源	数据条数
组1	演习的专利训练数据	250
组2	演习的专利评估数据	50
组3	正式比赛的专利评估数据	200

参赛队伍要求提交经过自动标注的文本，通过召回率、准确率和F值三个指标对结果进行测评。同时，组办方也向参赛队伍提供了评估工具，来计算这三个评估指标。在利用训练数据和评估数据进行抽取模型的训练实验后，本文采用了主办方提供的评估工具进行了实验结果的评估。

3.2 实验及正式测评结果

（1）实验系统设计及实验结果

在CRFs抽取模型和规则抽取模型构建完成后，本文利用组办方提供的训练数据和评估数据，设计了表3中的3组实验系统，每组实验系统使用了不同的训练数据集及抽取方法，通过数据集与抽取方法的组合，本文对专利抽取模型的适应性进行了实验。表4显示了3个系统的实验运行结果。

表3 基于演习训练与评估数据的实验系统设计

系统	训练数据集	抽取方法
系统1	组1	仅使用CRFs模型中的特征2和特征3
系统2	组1	CRFs 模型
系统3	组1+ 组3	CRFs 模型 + 规则模型

表4 基于组办方评估工具的实验系统评估

系统	召回率	准确率	F值
系统1	0.163	0.366	0.226
系统2	0.167	0.362	0.228
系统3	0.239	0.438	0.309

（2）正式比赛提交结果的评估

表5中的系统提交到NTCIR-8的正式比赛的专利文摘抽取系统，其评估结果为组办方返回的官方评估结果。相比较演习提交的抽取结果，由于我们引入了规则抽取的方法，整体结果有了较大的提高。这也说明使用CRFs抽取模型与规则抽取两种方法对改进抽取结果是有益的。

表5 正式比赛的专利评估结果

系统	召回率	准确率	F值
patent_ 1	0.239	0.438	0.309
patent_ 2	0.223	0.423	0.292

4 结语

本文利用文档结构的特点以及专利写作过程中的常用特色词汇，在利用条件随机场这种概率模型的基础上，提出了条件随机场与规则集成的专利摘要信息抽取方法，并进行了实验和提交给NTCIR-8进行测评，取得了相对满意的效果。通过对CRFs与规则集成及对比分析，得到以下结论：

（1）标注的质量对CRFs模型的抽取结果影响较大。

（2）对于“技术”类标签，抽取效果比较理想，而“属性”和“值”两类标签的抽取效果不理想，主要原因是这两类标签中的词性特征等特征不明显。

（3）基于规则的信息抽取模型引入后明显改善了抽取效果。

本文不足之处在于提出的规则由手工编写，费时费力，专业性、领域相关性比较强，导致通用性较差。下一步将研究进一步丰富与完善规则抽取方法，以更好地提高抽取效果。

参考文献

[1] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the International Conference on Machine Learning (ICML-2001), Williams MA, 2001:282-289.
[2] SHA F, PEREIRA F. Shallow parsing with conditional random fields[C]// Proceedings of Human Language Technology-NAACL, 2003.
[3] 程月,陈小荷. 基于条件随机场的汉语动宾搭配自动识别[J]. 中文信息学报,2009,23(1):9-15.
[4] 邓箴,包宏. 基于条件随机场的中文自动文摘系统[J]. 西安石油大学学报(自然科学版),2009,24(1):96-99.
[5] 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.
[6] NANBA H, FUJII A, IWAYAMA M, HASHIMOTO T. Overview of the Patent Mining Task at the NTCIR-8 Workshop[C]// Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2010.

[7] CRF++ -0.53 [CP/OL]. [2010-08-10]. http://sourceforge.net/projects/crfpp/files/.
[8] postagger.1.0 [CP/OL]. [2010-08-10]. http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/postagger/.

作者简介

李鹏(1979-)，硕士，助理研究员。研究方向：智能信息处理。通讯地址：北京市复兴路15号 中国科学技术信息研究所 信息技术支持中心 100038。E-mail: lipeng_cn@istic.ac.cn
桂婕(1976-)，博士，助理研究员。研究方向：专利分析和科技创新管理。通讯地址同上。E-mail: guij@istic.ac.cn
乔晓东(1965-)，硕士，研究员，研究方向：信息服务和信息资源管理。通讯地址同上。E-mail: qiaox@istic.ac.cn
张兆锋(1979-)，硕士，工程师。研究方向：信息系统和信息可视化。通讯地址同上。E-mail: zhangzf@istic.ac.cn

Information Extraction of Patent Summary Based on Integration of CRFs and Rule

Li Peng, Gui Jie, Qiao Xiaodong, Zhang Zhaofeng / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Patent is an important data source for Information Analysis. However, as patent terms are more abstract than that of research papers, most information extraction based on statistical results is not ideal. In this paper, information extraction of patent summary based on CRFs and rule is proposed by using the characteristics of the patent document structure and common characteristic words in the process of writing. The method was used in Subtask of Technical Trend Map Creation of NTCIR-8 Patent Mining Task. The evaluation and experiment results show that the method is practical and efficient.

Keywords: Information extraction, Conditional random fields, Rule

（收稿日期：2010-08-15）

业界动态

全版《牛津英文词典》不敌在线版或停印

据新华社电 牛津大学出版社高级管理人员8月29日说，鉴于对网络版的需求远远超过印刷版，《牛津英文词典》第三版将来可能不再印刷。

牛津大学出版社行政总裁奈杰尔·波特伍德告诉英国《星期日泰晤士报》记者，受互联网影响，《牛津英文词典》将来可能仅以电子版形式出现。“印刷版词典市场正在消失，每年缩水10%。”

波特伍德预计，随着电子图书和类似美国苹果公司平板电脑iPad等工具的普及，印刷版词典可能还有大约30年“货架寿命”。

眼下网络上流行的《牛津英文词典》为第二版，于2000年上传网络，每月点击率达200万次。用户每年需支付240英镑(约合373美元)订阅费。

按美联社说法，网络版《牛津英文词典》除方便用户查阅外，同时更便于及时更新大量新词汇。编辑人员大约每3个月更新一次。例如，今年3月，他们将类似superbug(超级细菌)等新词加入网络版。

波特伍德所提及的停印事宜主要针对完整版《牛津英文词典》。他说，出版社将继续出版更通俗的《牛津英文词典》，即书店常见的单卷版。

牛津大学出版社1879年开始准备编纂《牛津英文词典》，第一版于1884年开始分批印刷，最终于1928年完成全部印刷。

《牛津英文词典》面世前，第一本可辨认版的英文词典是1775年出版、由塞缪尔·约翰逊编纂的《英语词典》。

来源：http://it.people.com.cn/GB/42894/196086/12587877.html（查询时间：2010-09-02）